Randomization Inference When $N = 1$

Tengyuan Liang [1] Benjamin Recht [2]

[1]The University of Chicago, Booth School of Business

[2]University of California Berkeley, EECS

Plan

- Motivation and Literature
  - RCTs vs. individualization
  - causal inference: interference
  - system identification: impulse response
  - problem setup
  - our contributions
- Theory and Methodology
  - convolution models
  - estimand and estimator
  - moments: formulas and estimates
  - inference: asymptotic normality
- Empirics

Motivation and Literature

Motivation

Individualized Inference:

How can you be the treatment and the control group?

Motivation

Individualized Inference:

How can you be the treatment and the control group?

- Manage mental health
- Physical therapy
- Learn an instrument/language
- Treat a chronic condition

Motivation

Individualized Inference:

How can you be the treatment and the control group?

THE DESIGN OF CLINICAL EXPERIMENTS*

D. D. Reid
M.D. Aberd., Ph.D. Lond.
READER IN EPIDEMIOLOGY AND VITAL STATISTICS IN THE UNIVERSITY OF LONDON, AT THE LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

Patient	Order 1	2	3	4	5	6
A	X	Y	Z	Z	Y	X
B	Y	Z	X	X	Z	Y
C	Z	X	Y	Y	X	Z

Evaluating three treatments for arthritis
Lancet. 1954. 264(6852):1293-6.

CLINICAL COMPARISON OF DIAMORPHINE AND PHOLCODINE AS COUGH SUPPRESSANTS BY A NEW METHOD OF SEQUENTIAL ANALYSIS

Order of administration	Preference for	
	Lipect	Placebo
Lipect before placebo	11	4
Placebo before lipect	11	2
Total ..	22	6
	Lipect	Heroin
Lipect before heroin	11	2
Heroin before lipect	6	8
Total ..	17	10
(χ^2 with continuity correction = 3·41 ; P = 0·07)		
	Placebo	Heroin
Placebo before heroin	2	10
Heroin before placebo	2	8
Total ..	4	18

Armitage and Snell.
Lancet. 1957. 272(6974):860-2.

The Patient as his own Control

In some instances it may be better to design the trial so that each patient provides his own control—by having various treatments in order. The advantages and disadvantages of that procedure will need careful thought.
By such means ...

"The Patient as his Own Control" in Bradford Hill's
Principles of Medical Statistics, 1961

Clinical and biomedical research: N-of-1 trials

discovery of Vitamins

THE NOBEL PRIZE

Nobel Prizes & Laureates | Nomination | Alfred Nobel | News & insights | Events | Educational

The Nobel Prize and the discovery of vitamins

by Kenneth J. Carpenter*

Introduction

In the course of the 19th century, chemists and physiologists studying the composition of foods and the nutritional requirements of humans and animals found that our diets needed to include the complex nitrogenous compounds called "proteins" (that, with water, form the bulk of our lean tissues), together with fats, starch and sugars that all provide usable energy during their oxidation in the body. It was also realized that bones contain high concentrations of lime (calcium oxide) and phosphate salts and the body, generally, has a variety of other necessary mineral salts, though it was felt that mixed diets normally supplied adequate quantities of all these without any need for special precautions.

With hindsight, we can see repeated early observations indicating that we also had a need for some other nutrients. Thus, sailors after 10-12 weeks on dry foods, during long sailing ship voyages before the days of refrigeration, typically developed scurvy, a disease characterized by weakness, pains in the joints, loose teeth and blood spots appearing all over the body, and finally sudden death "in the middle of a sentence" from the bursting of a main artery. However, desperately ill men would recover in 10 days or so after reaching land where they could be given fresh fruit or salad greens.

Nobel Laureates and their work with vitamins

Nobel Prize in Physiology or Medicine	
Discovery of vitamins	
Christiaan Eijkman (1929)	Vitamin B_1
Sir Frederick Gowland Hopkins (1929)	Growth Stimulating Vitamins
George Hoyt Whipple (1934)*	Vitamin B_{12}
George Richards Minot (1934)*	Vitamin B_{12}
William Parry Murphy (1934)*	Vitamin B_{12}
Henrik Carl Peter Dam (1943)	Vitamin K
Isolation of vitamins	
Adolf Otto Reinhold Windaus (1928)*	Vitamin D
Albert von Szent-Györgyi Nagyrapolt (1937)	Vitamin C
Richard Kuhn (1938)	Vitamin B_2 and B_6
Edward Adelbert Doisy (1943)	Vitamin K

Casimir Funk was nominated for the Nobel Prize four times but never received it.

Motivation

Online platforms and targeting: sequential A/B testing

Netflix's interleaving strategy

Interleaving at Netflix

At Netflix, we use interleaving in the first stage of experimentation to sensitively determine member preference between two ranking algorithms. The figure below depicts the differences between A/B testing and interleaving. In traditional A/B testing, we choose two groups of subscribers: one to be exposed to ranking algorithm *A* and another to *B*. In interleaving, we select a single set of subscribers who are exposed to an interleaved ranking generated by blending the rankings of algorithms *A* and *B*. This allows us to present choices side-by-side to the user to determine their preference of ranking algorithms. (Members are not able to distinguish between which algorithm recommended a particular video.) We calculate the relative preference for a ranking algorithm by comparing the share of hours viewed, with attribution based on which ranking algorithm recommended the video.

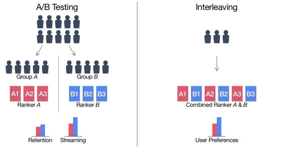


Fig. 3: A/B Testing vs. Interleaving. In traditional A/B testing, the population is split into two groups, one exposed to ranking algorithm A and another to B. Core evaluation metrics like retention and streaming are measured and compared between the two groups. In contrast, interleaving exposes one group of members to a blended ranking of rankers A and B. User preference for a ranking algorithm is determined by comparing the share of viewing hours coming from videos recommended by rankers A or B.

For example, imagine that LinkedIn develops a new algorithm for matching job seekers with job openings. To measure its effectiveness, LinkedIn would simultaneously expose all job postings and seekers in a given market to the new algorithm for 30 minutes. In the next 30-minute period, it would randomly decide to either switch to the old algorithm or stay with the new one. It would continue this process for at least two weeks to ensure that it sees all types of job search patterns. Netflix's interleaving strategy is a special application of this more general methodology.

Bojinov, Saint-Jacques, and Tingley, HDSR 2020

causal inference **system identification/control**

causal inference system identification/control

Causal inference

Neyman, 1923: *On the Application of Probability Theory to Agricultural Experiments. Essay on principles. Section 9.*

Potential Outcomes for Field Experiments

- Unknown potential yields indexed by varieties $\times$ plots
- If randomize, mean outcome in treatment vs. control is estimable
- Formula for the variance (measurement precision) of the difference between average observed yields of two varieties
- Probability theory can be used even when yields from different plots do not follow Gaussian law.

Causal inference

Neyman, 1923: *On the Application of Probability Theory to Agricultural Experiments. Essay on principles. Section 9.*

Potential Outcomes for Field Experiments

- Unknown potential yields indexed by varieties × plots
- If randomize, mean outcome in treatment vs. control is estimable
- Formula for the variance (measurement precision) of the difference between average observed yields of two varieties
- Probability theory can be used even when yields from different plots do not follow Gaussian law.

Opens up the Potential Outcome Framework for randomization inference: **RCTs as measurement device for effects with uncertainty quantification**.

What do we learn from RCTs

- Measures treatment effect: how a policy works in a population
- Removes cofounding factors (identifiability), reasonable estimation of measurement precision (error bar)
- In particular, RCTs do **NO**T inform us: how a policy works for an individual? what happens over time?

What do we learn from RCTs

- Measures treatment effect: how a policy works in a population
- Removes cofounding factors (identifiability), reasonable estimation of measurement precision (error bar)
- In particular, RCTs do **NO**T inform us: how a policy works for an individual? what happens over time?

Individual treatment effect? Ignorability assumption: effectively RCTs conditioning on covariates x.

Potential outcome framework

Cross-sectional data:

- Potential outcomes: $y_i(0), y_i(1) \in \mathbb{R}$ for $i = 1, \ldots, n$.
- Observed outcome: $y_i = y_i(1) \cdot x_i + y_i(0) \cdot (1 - x_i)$, binary treatment variable $x_i \in \{0, 1\}$, equiv.

$$y_i = \frac{y_i(1)+y_i(0)}{2} + \frac{y_i(1)-y_i(0)}{2} \cdot (2x_i - 1) \ .$$

Neyman, 1923, Fisher, 1937, Imbens and Rubin, 2015

Potential outcome framework

Cross-sectional data:

Estimand: average treatment effect

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

Potential outcome framework

Cross-sectional data:

Estimand: average treatment effect

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

Estimator: Horvitz-Thompson

$$\widehat{\tau} := \frac{1}{n}\sum_{i=1}^{n}(2x_i - 1)\cdot 2y_i$$

Horvitz and Thompson, 1952

Potential outcome framework

Cross-sectional data:

Estimand: average treatment effect

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

Estimator: Horvitz-Thompson

$$\widehat{\tau} := \frac{1}{n}\sum_{i=1}^{n}(2x_i - 1) \cdot 2y_i$$

Horvitz and Thompson, 1952

Unbiasedness, variance and asymptotic normality: design-based randomness, first moment $\mathbb{E}[2x_i - 1] = 0$ and $(2x_i - 1)^2 \equiv 1$

Stable Unit Treatment Value Assumption (SUTVA): no interference

y_i only depends on x_i

Potential outcomes for time-series? Interference

Time-series data: SUTVA is violated. N-of-1 clinical trials and macroeconomic studies. Granger, 1969, Sims, 1972

Interference: outcomes at time t depend on treatments assigned before, namely, the treatment path $x_0, x_1, \ldots, x_t$.

Potential outcomes for time-series? Interference

Time-series data: SUTVA is violated. N-of-1 clinical trials and macroeconomic studies. Granger, 1969, Sims, 1972

Interference: outcomes at time t depend on treatments assigned before, namely, the treatment path $x_0, x_1, \ldots, x_t$.

One approach extending Neyman-Fisher-Rubin framework: include all interactions $x_S := \prod_{s \in S} x_s$, for all subsets $S \subseteq \{0, 1, \ldots, t\}$

$$y_t = \sum_{S : S \subseteq \{0,1,\ldots,t\}} \alpha_S^{(t)} \cdot x_S \ .$$

Potential outcomes for time-series? Interference

Time-series data: SUTVA is violated. N-of-1 clinical trials and macroeconomic studies. Granger, 1969, Sims, 1972

Interference: outcomes at time t depend on treatments assigned before, namely, the treatment path $x_0, x_1, \ldots, x_t$.

One approach extending Neyman-Fisher-Rubin framework: include all interactions $x_S := \prod_{s \in S} x_s$, for all subsets $S \subseteq \{0, 1, \ldots, t\}$

$$y_t = \sum_{S : S \subseteq \{0,1,\ldots,t\}} \alpha_S^{(t)} \cdot x_S \ .$$

Such representation is with full generality, analysis of Boolean functions O'Donnell, 2014

However, curse of dimensionality 2^t prevents meaningful statistical analysis

Granger-Sims causality

Granger-Sims causality framework: testing correlations in x's that can explain y Granger, 1969, Sims, 1972, Angrist and Kuersteiner, 2011

$$y_t = \alpha_\emptyset + \alpha_0 \cdot x_t + \alpha_1 \cdot x_{t-1} + \ldots + \alpha_t \cdot x_0 + \mathrm{error}_t \ .$$

Granger-Sims causality

Granger-Sims causality framework: testing correlations in x's that can explain y Granger, 1969, Sims, 1972, Angrist and Kuersteiner, 2011

$$y_t = \alpha_\emptyset + \alpha_0 \cdot x_t + \alpha_1 \cdot x_{t-1} + \ldots + \alpha_t \cdot x_0 + \mathrm{error}_t \ .$$

Granger causality leverages time-invariance and linearity to provide practically useful answers to whether the time-series x forecast time-series y.

causal inference system identification/control

system identification and control

Control theory: model the input-output behavior of a dynamical system, design feedback policy

$$s_{t+1} = f_t(s_t, x_t, \epsilon_t)$$
$$y_t = h_t(s_t, x_t, \epsilon_t)$$

x_t input, y_t output

ϵ_t exogenous noise, s_t state of the dynamical system

system identification and control

Control theory: model the input-output behavior of a dynamical system, design feedback policy

$$s_{t+1} = f_t(s_t, x_t, \epsilon_t)$$
$$y_t = h_t(s_t, x_t, \epsilon_t)$$

x_t input, y_t output

ϵ_t exogenous noise, s_t state of the dynamical system

Policy evaluation and optimization

Find appropriate estimates of the function f_t and h_t so that inputs x_t can be planned to steer y_t to desired values.

linear dynamical systems

f_t and h_t must not be too complicated to be identifiable: linear class

Ljung, 1998

linear dynamical systems

f_t and h_t must not be too complicated to be identifiable: linear class

Ljung, 1998

For linear dynamical systems, the input-output map as

$$y_t = \sum_{s \in [t]} G_s^{(t)} x_s + e_t$$

where $G_s^{(t)}$ are scalars and e_t are linear functions of the $\epsilon_s, s \in [t]$ and s_1.

linear dynamical systems

f_t and h_t must not be too complicated to be identifiable: linear class

Ljung, 1998

For linear dynamical systems, the input-output map as

$$y_t = \sum_{s \in [t]} G_s^{(t)} x_s + e_t$$

where $G_s^{(t)}$ are scalars and e_t are linear functions of the $\epsilon_s, s \in [t]$ and s_1.

further restriction: time-invariant linear dynamical systems: $f_t \equiv f$ and $h_t \equiv h$

$$y_t = \sum_{s \in [t]} g_{t-s} x_s + e_t$$

Here the sequence g is called the impulse response function, modeling interference.

Bakshi, Liu, Moitra, and Yau (2023), Oymak and Ozay (2019), and Simchowitz, Boczar, and Recht (2019)

causal inference **system identification/control**

we model interference by impulse response function

Problem setup: N-of-1 trial or interleaving design

- Two type of actions: A or B
- Each time t, pick one action $\mathbf{x}_t \in \{A, B\}$, try it out, document outcome $\mathbf{y}_t$
- At the end, infer the effect of A vs. B by "correlating" **time-series** $\mathbf{y}$ to $\mathbf{x}$

Problem setup: N-of-1 trial or interleaving design

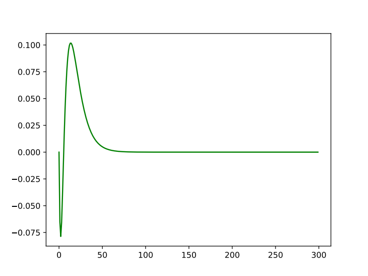
- Two type of actions: A or B
- Each time t, pick one action $\mathbf{x}_t \in \{A, B\}$, try it out, document outcome $\mathbf{y}_t$
- At the end, infer the effect of A vs. B by "correlating" **time-series** $\mathbf{y}$ to $\mathbf{x}$

Challenging: **interference**! $\mathbf{y}$ depends on the whole path of $\mathbf{x}$

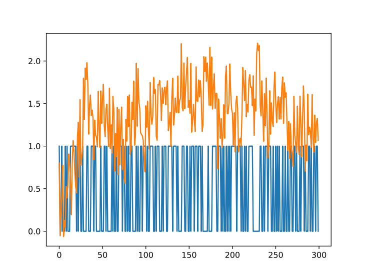
Problem setup

Treatment/Control variable $\mathbf{x}$, Observed response $\mathbf{y}$

Estimate/Inference: linear functionals of impulse response function g



impulse response g

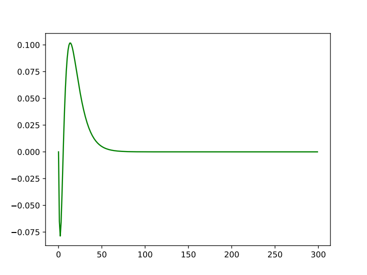


treatment $\mathbf{x}$, response $\mathbf{y}$

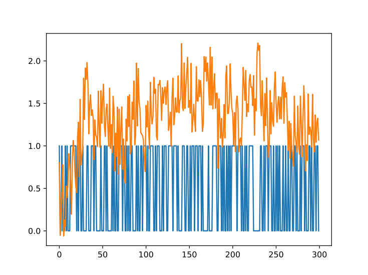
Problem setup

Treatment/Control variable **x**, Observed response **y**

Estimate/Inference: linear functionals of impulse response function g



impulse response g



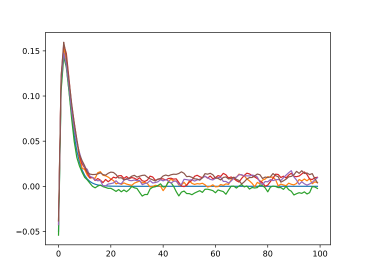
treatment **x**, response **y**

Robust to arbitrary oblivious error

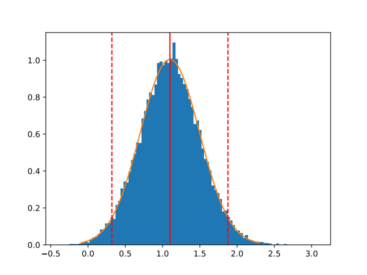
What we will show

counterfactual reasoning and control of the system

depends on the impulse response g



(a) identification of g



(b) inference of $\langle g, q \rangle$

Our contributions

Robust to arbitrary error sequence : extends the Granger-Sims

Potential outcomes for time-series : generalize Neyman, interference

New unbiased estimator : generalize Horvitz-Thompson to time-series

Asymptotic inference : new to the system identification and control

Theory and Methodology

convolution and interference

$$\mathbf{y}_t = (\mathbf{x} * g + e)_t$$

- Linear convolution $(x * g)_t := \sum_{s=0}^{t} x_s g_{t-s}$
- **adversarial error**: $e \in \mathbb{R}^T$ is any error oblivious to the randomization $\mathbf{x}$

Convolution models the **interference** effect.

Estimand and estimator: time-invariant

- Consider time-invariant impulse response $g \in \mathbb{R}^T$.

Estimand and estimator: time-invariant

- Consider time-invariant impulse response $g \in \mathbb{R}^T$.
- **Estimands**: linear functionals indexed by a time-invariant $q \in \mathbb{R}^T$

$$\Delta(q) := \langle q, g \rangle .$$

Estimand and estimator: time-invariant

- Consider time-invariant impulse response $g \in \mathbb{R}^T$.
- **Estimands**: linear functionals indexed by a time-invariant $q \in \mathbb{R}^T$

$$\Delta(q) := \langle q, g \rangle \ .$$

- **Estimators**: convolution estimator

$$\widehat{\Delta}(q) = \frac{1}{T} \langle (2\mathbf{x} - 1) * q^{\circ}, 2\mathbf{y} \rangle \ .$$

Special case: lag-K effect

Special cases include the cumulative lag-K effects, $K \in [1, T] \cap \mathbb{Z}$, where vector $q = 1_{<K} := (\underbrace{1, \ldots, 1}_{K}, 0, \ldots, 0)$ is plugged in,

$$\Delta_K := \Delta(1_{<K}) = \sum_{k=0}^{K-1} g_k \ .$$

The estimator for the cumulative lag-K effects Δ_K is

$$\widehat{\Delta}_K := \widehat{\Delta}(1_{<K}) = \tfrac{1}{T} \langle (2\mathbf{x} - 1) * 1^{\circ}_{<K}, 2\mathbf{y} \rangle \ .$$

Estimand and estimator: time-variant

- More generally, we may consider a broader class of time-variant $g^{(t)} \in \mathbb{R}^t$'s

$$\mathbf{y}_t = (\mathbf{x} * g^{(t)} + e)_t \ .$$

$(\mathbf{x} * g^{(t)})_t := \sum_{s=0}^{t} x_s g^{(t)}_{t-s}$

Estimand and estimator: time-variant

- More generally, we may consider a broader class of time-variant $g^{(t)} \in \mathbb{R}^t$'s

$$\mathbf{y}_t = (\mathbf{x} * g^{(t)} + e)_t \ .$$

- For a sequence of vectors $q^{(t)} \in \mathbb{R}^t$, define the estimand

$$\tau := \frac{1}{T} \sum_{t \in [T]} \langle g^{(t)}, q^{(t)} \rangle \ ,$$

Estimand and estimator: time-variant

- More generally, we may consider a broader class of time-variant $g^{(t)} \in \mathbb{R}^t$'s

$$\mathbf{y}_t = (\mathbf{x} * g^{(t)} + e)_t \ .$$

- For a sequence of vectors $q^{(t)} \in \mathbb{R}^t$, define the estimand

$$\tau := \tfrac{1}{T} \sum_{t \in [T]} \langle g^{(t)}, q^{(t)} \rangle \ ,$$

- The corresponding unbiased estimator

$$\widehat{\tau} := \tfrac{1}{T} \sum_{t \in [T]} \left(2\mathbf{x} - 1\right) * q^{(t)}\Big)_t \cdot 2\mathbf{y}_t \ .$$

If $g^{(t)} = (g_0^{(t)}, 0, \ldots, 0)$, $q^{(t)} = (1, 0, \ldots, 0)$, the estimator reduces to the classic Horvitz-Thompson.

Estimand and estimator: time-variant

- More generally, we may consider a broader class of time-variant $g^{(t)} \in \mathbb{R}^t$'s

$$\mathbf{y}_t = (\mathbf{x} * g^{(t)} + e)_t \ .$$

- For a sequence of vectors $q^{(t)} \in \mathbb{R}^t$, define the estimand

$$\tau := \frac{1}{T} \sum_{t \in [T]} \langle g^{(t)}, q^{(t)} \rangle \ ,$$

- The corresponding unbiased estimator

$$\widehat{\tau} := \frac{1}{T} \sum_{t \in [T]} \left(2\mathbf{x} - 1\right) * q^{(t)} \Big)_t \cdot 2\mathbf{y}_t \ .$$

  If $g^{(t)} = (g_0^{(t)}, 0, \ldots, 0)$, $q^{(t)} = (1, 0, \ldots, 0)$, the estimator reduces to the classic Horvitz-Thompson.

Generalize classical potential outcomes, unbiasedness and var formula, space vs. time

Some properties of the estimator

Moments: formulas and estimates

First moment (L. and Recht, '23)

$$\mathop{\mathbb{E}}_{\mathbf{x}}\left[\widehat{\Delta}(q)\right] = \Delta(q) = \langle q, g \rangle \ .$$

Moments: formulas and estimates

First moment (L. and Recht, '23)

$$\mathop{\mathbb{E}}_{\mathbf{x}}\left[\widehat{\Delta}(q)\right] = \Delta(q) = \langle q, g\rangle \ .$$

Unbiasedness of the estimator.

Moments: formulas and estimates

Second moment (L. and Recht, '23)

$$\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\widehat{\Delta}(q)-\Delta(q)\right)^2\right]$$
$$=\frac{1}{T}\left(\|g*q^\circ\|_2^2+\langle g*g^\circ,q*q^\circ\rangle-2\langle q,g\rangle^2\right)+\frac{1}{T^2}\|(1*g^\circ+2e)*q^\circ\|_2^2\,.$$

Moments: formulas and estimates

Second moment (L. and Recht, '23)

$$\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\widehat{\Delta}(q)-\Delta(q)\right)^2\right]$$
$$= \tfrac{1}{T}\left(\|g * q^\circ\|_2^2 + \langle g * g^\circ, q * q^\circ\rangle - 2\langle q, g\rangle^2\right) + \tfrac{1}{T^2}\|(1 * g^\circ + 2e) * q^\circ\|_2^2\,.$$

The formula for the variance of the treatment effects: measurement accuracy.

Depends on the functional forms of g and q.

Moments: formulas and estimates

Second moment (L. and Recht, '23)

$$\mathop{\mathbb{E}}_{\mathbf{x}}\left[\left(\widehat{\Delta}(q)-\Delta(q)\right)^2\right]$$
$$=\tfrac{1}{T}\left(\|g*q^\circ\|_2^2+\langle g*g^\circ,q*q^\circ\rangle-2\langle q,g\rangle^2\right)+\tfrac{1}{T^2}\|(1*g^\circ+2e)*q^\circ\|_2^2\,.$$

The formula for the variance of the treatment effects: measurement accuracy.

Depends on the functional forms of g and q.

If $g=(g_0,0,\ldots,0)$, $q=(1,0,\ldots,0)$

$$\tfrac{1}{T}\left(\|g*q^\circ\|_2^2+\langle g*g^\circ,q*q^\circ\rangle-2\langle q,g\rangle^2\right)=0$$

$$\tfrac{1}{T^2}\|g_0\cdot 1+2e\|_2^2=\text{var formula of HT}$$

Asymptotic normality?

Non-trivial due to interference.

Asymptotic normality

Assume

$$\frac{\|(|g| * |q|^\circ) * (|g| * |q|^\circ)^\circ\|_2^2}{\|g * q^\circ\|_2^4} = o(T)$$

Asymptotic normality

Assume

$$\frac{\|(|g| * |q|^\circ) * (|g| * |q|^\circ)^\circ\|_2^2}{\|g * q^\circ\|_2^4} = o(T)$$

True if $|g_t| \precsim 0.99^t, |\mathrm{supp}(q)| = K \precsim \mathrm{polylog}(T)$.

Asymptotic normality

Assume

$$\frac{\|(|g| * |q|^\circ) * (|g| * |q|^\circ)^\circ\|_2^2}{\|g * q^\circ\|_2^4} = o(T)$$

True if $|g_t| \lesssim 0.99^t$, $|\mathrm{supp}(q)| = K \lesssim \mathrm{polylog}(T)$.

Asymptotic normality (L. and Recht, '23)

$$\frac{\sqrt{T} \cdot \mathbf{H}_T}{\sqrt{\mathcal{V}_Q}} \Rightarrow \mathcal{N}(0,1), \quad \text{as } T \to \infty\ ,$$

where $\mathcal{V}_Q := \|g * q^\circ\|_2^2 + \langle g * g^\circ, q * q^\circ \rangle - 2\langle q, g\rangle^2$.

Asymptotic normality

Assume

$$\frac{\|(|g| * |q|^\circ) * (|g| * |q|^\circ)^\circ\|_2^2}{\|g * q^\circ\|_2^4} = o(T)$$

True if $|g_t| \lesssim 0.99^t$, $|\mathrm{supp}(q)| = K \lesssim \mathrm{polylog}(T)$.

Asymptotic normality (L. and Recht, '23)

$$\frac{\sqrt{T} \cdot \mathbf{H}_T}{\sqrt{\mathcal{V}_Q}} \Rightarrow \mathcal{N}(0,1), \quad \text{as } T \to \infty \, ,$$

where $\mathcal{V}_Q := \|g * q^\circ\|_2^2 + \langle g * g^\circ, q * q^\circ \rangle - 2\langle q, g \rangle^2$.

Variance $\to$ Distribution: non-trivial for temporally dependent problem

Moments: formulas and estimates

Decomposition: denote the centered randomization vector $\mathbf{z} := 2\mathbf{x} - 1 \in \mathbb{R}^T$, and $h = 1 * g^\circ + 2e \in \mathbb{R}^T$,

$$\begin{aligned}\widehat{\Delta}(q) &= \tfrac{1}{T}\langle (2\mathbf{x}-1) * q^\circ, (2\mathbf{x}-1) * g^\circ + 1 * g^\circ + 2e \rangle \ ,\\ &= \tfrac{1}{T}\langle \mathbf{z} * q^\circ, \mathbf{z} * g^\circ \rangle + \tfrac{1}{T}\langle \mathbf{z} * q^\circ, h \rangle \ .\end{aligned}$$

Moments: formulas and estimates

Decomposition: denote the centered randomization vector $\mathbf{z} := 2\mathbf{x} - 1 \in \mathbb{R}^T$, and $h = 1 * g^\circ + 2e \in \mathbb{R}^T$,

$$\begin{aligned}\widehat{\Delta}(q) &= \tfrac{1}{T}\langle (2\mathbf{x}-1) * q^\circ, (2\mathbf{x}-1) * g^\circ + 1 * g^\circ + 2e\rangle \ , \\ &= \tfrac{1}{T}\langle \mathbf{z} * q^\circ, \mathbf{z} * g^\circ\rangle + \tfrac{1}{T}\langle \mathbf{z} * q^\circ, h\rangle \ .\end{aligned}$$

$\mathbf{W}_T$ denotes the difference between the estimator and the estimand,

$$\mathbf{W}_T := \widehat{\Delta}(q) - \Delta(q) = \boxed{\frac{1}{T}\sum_{i \neq j \in [T]} \mathbf{z}_i \mathbf{z}_j H_{ij}} + \frac{1}{T}\sum_{i \in [T]} \mathbf{z}_i L_i \ ,$$

where

$$H_{ij} := \sum_{t \in [T]} q^\circ_{t-i} g^\circ_{t-j}, \ \ L_i := \sum_{t \in [T]} q^\circ_{t-i} h_t \ .$$

Moments: formulas and estimates

Fourth moment (L. and Recht, ’23)

Denote

$$\mathbf{H}_T := \boxed{\frac{1}{T} \sum_{i \neq j \in [T]} \mathbf{z}_i \mathbf{z}_j H_{ij}}$$

where $H_{ij} := \sum_{t \in [T]} q^\circ_{t-i} g^\circ_{t-j}$.

Then

$$\left| \frac{\mathbb{E}[\mathbf{H}_T^4]}{\left(\mathbb{E}[\mathbf{H}_T^2]\right)^2} - 3 \right| \leq \frac{4}{T} + \frac{16}{T} \frac{\|(|g| * |q|^\circ) * (|g| * |q|^\circ)^\circ\|_2^2}{\left(\|g * q^\circ\|_2^2 + \langle g * g^\circ, q * q^\circ \rangle - 2\langle q, g \rangle^2\right)^2} .$$

Moments: formulas and estimates

Fourth moment (L. and Recht, '23)

Denote

$$\mathbf{H}_T := \boxed{\frac{1}{T} \sum_{i \neq j \in [T]} \mathbf{z}_i \mathbf{z}_j H_{ij}}$$

where $H_{ij} := \sum_{t \in [T]} q^{\circ}_{t-i} g^{\circ}_{t-j}$.

Then

$$\left| \frac{\mathbb{E}[\mathbf{H}_T^4]}{\left(\mathbb{E}[\mathbf{H}_T^2]\right)^2} - 3 \right| \leq \frac{4}{T} + \frac{16}{T} \frac{\|(|g| * |q|^{\circ}) * (|g| * |q|^{\circ})^{\circ}\|_2^2}{\left(\|g * q^{\circ}\|_2^2 + \langle g * g^{\circ}, q * q^{\circ} \rangle - 2\langle q, g \rangle^2\right)^2} .$$

$\mathbf{H}_T$ is **quadratic** in $\mathbf{z}$'s.

Eighth moment calculations in $\mathbf{z}$'s.

Higher moments and asymptotic normality

- Rademacher chaos $\mathbf{H}_T := \boxed{\frac{1}{T} \sum_{i \neq j \in [T]} \mathbf{z}_i \mathbf{z}_j H_{ij}}$ with

  $H_{ij} := \sum_{t \in [T]} q^{\circ}_{t-i} g^{\circ}_{t-j}$ for generic g, q.

Higher moments and asymptotic normality

- Rademacher chaos $\mathbf{H}_T := \boxed{\frac{1}{T}\sum_{i\neq j\in[T]} \mathbf{z}_i\mathbf{z}_j H_{ij}}$ with $H_{ij} := \sum_{t\in[T]} q^\circ_{t-i} g^\circ_{t-j}$ for generic g, q.
- Martingale differences: Central Limit Theorem?

Higher moments and asymptotic normality

- Rademacher chaos $\mathbf{H}_T := \boxed{\frac{1}{T} \sum_{i \neq j \in [T]} \mathbf{z}_i \mathbf{z}_j H_{ij}}$ with

  $H_{ij} := \sum_{t \in [T]} q^{\circ}_{t-i} g^{\circ}_{t-j}$ for generic g, q.
- Martingale differences: Central Limit Theorem?
- Not true (for generic g, q)!

Higher moments and asymptotic normality

- Rademacher chaos $\mathbf{H}_T := \boxed{\frac{1}{T}\sum_{i \neq j \in [T]} \mathbf{z}_i \mathbf{z}_j H_{ij}}$ with $H_{ij} := \sum_{t \in [T]} q^\circ_{t-i} g^\circ_{t-j}$ for generic g, q.
- Martingale differences: Central Limit Theorem?
- Not true (for generic g, q)!
- True when $\frac{\|(|g|*|q|^\circ)*(|g|*|q|^\circ)^\circ\|_2^2}{\left(\|g*q^\circ\|_2^2\right)^2} = o(T)$. Higher moments calculation implies $\left|\frac{\mathbb{E}[\mathbf{H}_T^4]}{\left(\mathbb{E}[\mathbf{H}_T^2]\right)^2} - 3\right| \to 0$ as $T \to \infty$.

Higher moments and asymptotic normality

- Rademacher chaos $\mathbf{H}_T := \boxed{\frac{1}{T}\sum_{i\neq j\in[T]} \mathbf{z}_i\mathbf{z}_j H_{ij}}$ with $H_{ij} := \sum_{t\in[T]} q^{\circ}_{t-i} g^{\circ}_{t-j}$ for generic g, q.
- Martingale differences: Central Limit Theorem?
- Not true (for generic g, q)!
- True when $\frac{\|(|g|*|q|^{\circ})*(|g|*|q|^{\circ})^{\circ}\|_2^2}{\left(\|g*q^{\circ}\|_2^2\right)^2} = o(T)$. Higher moments calculation implies $\left|\frac{\mathbb{E}[\mathbf{H}_T^4]}{\left(\mathbb{E}[\mathbf{H}_T^2]\right)^2} - 3\right| \to 0$ as $T \to \infty$.

If $g * q^{\circ}$ is approx. supported on top K-elements, where $K \ll T$.

How to construct confidence intervals based on data?

Inference: confidence intervals

But how to estimate the variance?

$$\mathcal{V}_Q = \|g * q^\circ\|_2^2 + \langle g * g^\circ, q * q^\circ \rangle - 2\langle q, g \rangle^2 \ ,$$
$$\mathcal{V}_L = \frac{1}{T}\|(1 * g^\circ + 2e) * q^\circ\|_2^2 \ .$$

Inference: confidence intervals

But how to estimate the variance?

$$\mathcal{V}_Q = \|g * q^\circ\|_2^2 + \langle g * g^\circ, q * q^\circ \rangle - 2\langle q, g \rangle^2 \ ,$$
$$\mathcal{V}_L = \tfrac{1}{T}\|(1 * g^\circ + 2e) * q^\circ\|_2^2 \ .$$

Plug-in estimate of the variance based on the formula? Denote

$$g \leftarrow \widehat{g}_{<K} = (\overbrace{\widehat{\Delta}(\delta_0), \dots, \widehat{\Delta}(\delta_{K-1})}^{\text{first } K \text{ elements}}, 0 \dots, 0)$$
$$e \leftarrow \widehat{e} := \mathbf{y} - \mathbf{x} * \widehat{g}^\circ_{<K}$$

and define $\widehat{\mathcal{V}}_Q, \widehat{\mathcal{V}}_L$ accordingly.

Inference: confidence intervals

But how to estimate the variance?

$$\mathcal{V}_Q = \|g * q^\circ\|_2^2 + \langle g * g^\circ, q * q^\circ \rangle - 2\langle q, g\rangle^2 ,$$
$$\mathcal{V}_L = \tfrac{1}{T}\|(1 * g^\circ + 2e) * q^\circ\|_2^2 .$$

Plug-in estimate of the variance based on the formula? Denote

$$g \leftarrow \widehat{g}_{<K} = (\overbrace{\widehat{\Delta}(\delta_0), \dots, \widehat{\Delta}(\delta_{K-1})}^{\text{first } K \text{ elements}}, 0 \dots, 0)$$
$$e \leftarrow \widehat{e} := \mathbf{y} - \mathbf{x} * \widehat{g}^\circ_{<K}$$

and define $\widehat{\mathcal{V}}_Q, \widehat{\mathcal{V}}_L$ accordingly.

Use non-asymptotic concentration inequalities to derive that with $K = K(T) = \Theta(\log(T))$, as $T \to \infty$

$$\widehat{\mathcal{V}}_Q \stackrel{a.s.}{\to} \mathcal{V}_Q, \quad \widehat{\mathcal{V}}_L \stackrel{a.s.}{\to} \mathcal{V}_L .$$

Inference: confidence intervals

Proof technique: Hanson-Wright inequalities

Uniform consistency (L. and Recht, '23)

For integer $k \in [0, K)$ with $K \ll T$, denote $\delta_k := (0, \dots, \overbrace{1}^{\text{k-th element}}, \dots, 0)$. Assume $|e_t| \leq M$, $\forall t \geq 0$ and $\|g\|_1 \leq C$ for absolute constants $C, M > 0$. Then with probability at least $1 - 4T^{-2}$

$$\sup_{k<K} \left| \widehat{\Delta}(\delta_k) - g_k \right| \lesssim \frac{\log(T)}{\sqrt{T}} .$$

Inference: confidence intervals

Data-driven confidence interval with coverage $1-\alpha$: choose $K = K(T) = \Theta(\log(T))$, let $q = 1_{<K}$, and

$$\widehat{\Delta}(q) \pm C_\alpha \sqrt{\frac{\widehat{\mathcal{V}}_Q + \widehat{\mathcal{V}}_L}{T}}\ .$$

Inference: confidence intervals

Data-driven confidence interval with coverage $1-\alpha$: choose $K = K(T) = \Theta(\log(T))$, let $q = 1_{<K}$, and

$$\widehat{\Delta}(q) \pm C_\alpha \sqrt{\frac{\widehat{\mathcal{V}}_Q + \widehat{\mathcal{V}}_L}{T}} \,.$$

Compared to Neyman-Fisher-Rubin on the estimation of the variance:

> *This variance involves the variance over all plots of the potential yields and the correlation coefficient r between the potential yields of the two varieties on the same plot. Since it is* **impossible to estimate** *r directly, Neyman advises taking r = 1, observing that in practice this may lead to using too large an estimated standard deviation, when comparing two variety means.*

Dabrowska and Speed, 1990

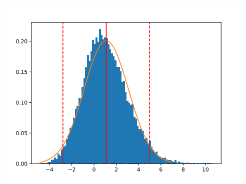
We leverage time-invariance to construct asymptotically valid confidence intervals.

Empirics

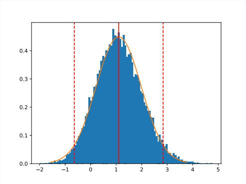
Experiment setup

- Impulse response function:
  $g_t = 1.0 * 0.65^t - 1.6 * 0.50^t + 0.75 * 0.48^t$.
- Estimand $\Delta(1) = \langle 1, g \rangle$.
- Estimator $\widehat{\Delta}(q) := \frac{1}{T}\langle (2\mathbf{x} - 1) * 1_{\leq K}, 2\mathbf{y} \rangle$ with
  $1_{\leq K} := (\underbrace{1, \ldots, 1}_{K}, 0, \ldots 0) \in \mathbb{R}^T$ with $K = 25$.

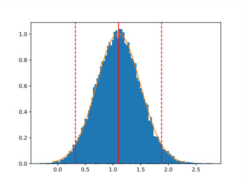
When does asymptotic normality kick in



$T = 200$



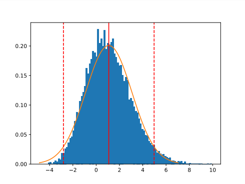
$T = 1000$



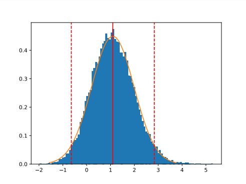
$T = 5000$

Circular convolution model

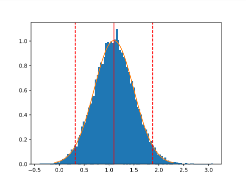
When does asymptotic normality kick in



$T = 200$

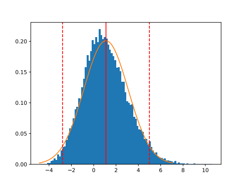


$T = 1000$

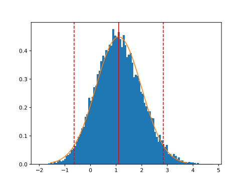


$T = 5000$

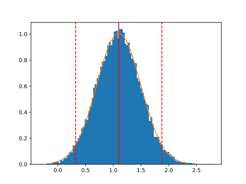
Linear convolution model



$T = 200$

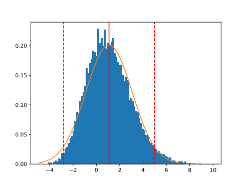


$T = 1000$

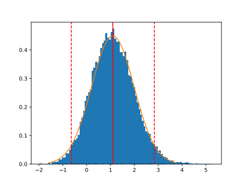


$T = 5000$

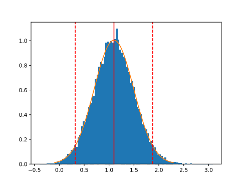
Circular convolution model



$T = 200$



$T = 1000$



$T = 5000$

Linear convolution model

Future directions

- Math works out, but it remains a difficult problem in practice.
  A measurement device for practitioner to tease out the interference effect in N-of-1 trials.
- Interference: just one sample (even when $T = \infty$), hard problem.
- Interference over networks: can the convolution analysis be generalized?
- Experimental design?
- Sequential decision making?

Thank you!

Tengyuan Liang, Benjamin Recht. Randomization Inference When N Equals One. arXiv:2310.16989, 2023.

linear dynamical systems

x_t matters as this input design must make all g_t of interest identifiable.

linear dynamical systems

x_t matters as this input design must make all g_t of interest identifiable.

The most popular choice in theory and practice chooses x_t to be an i.i.d. sequence of zero mean random variables.

Mareels, 1984; Overschee and Moor, 1994; Verhaegen and Dewilde, 1992

linear dynamical systems

x_t matters as this input design must make all g_t of interest identifiable.

The most popular choice in theory and practice chooses x_t to be an i.i.d. sequence of zero mean random variables.

Mareels, 1984; Overschee and Moor, 1994; Verhaegen and Dewilde, 1992

Statistical error rates for the estimation of the impulse response g from a single input sequence $\{x_t\}$: only recently determined.
nonparametric, infinite-dimensional problem

Bakshi et al., 2023; Oymak and Ozay, 2019; Simchowitz, Boczar, and Recht, 2019

References i

- Angrist, Joshua D and Guido M Kuersteiner (2011). **"Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score".** In: *Review of Economics and Statistics* 93.3, pp. 725–747.
- Bakshi, Ainesh et al. (2023). **"A New Approach to Learning Linear Dynamical Systems".** In: *arXiv preprint arXiv:2301.09519*.
- Fisher, Ronald Aylmer (1937). **The design of experiments.**
- Granger, Clive WJ (1969). **"Investigating causal relations by econometric models and cross-spectral methods".** In: *Econometrica: journal of the Econometric Society*, pp. 424–438.
- Horvitz, Daniel G and Donovan J Thompson (1952). **"A generalization of sampling without replacement from a finite universe".** In: *Journal of the American statistical Association* 47.260, pp. 663–685.

- Imbens, Guido W and Donald B Rubin (2015). **Causal inference in statistics, social, and biomedical sciences.** Cambridge University Press.
- Ljung, Lennart (1998). **System Identification. Theory for the user.** 2nd. Upper Saddle River, NJ: Prentice Hall.
- Mareels, Iven (1984). **"Sufficiency of excitation".** In: *Systems & control letters* 5.3, pp. 159–163.
- Neyman, Jerzy (1923). **"On the application of probability theory to agricultural experiments. Essay on principles".** In: *Ann. Agricultural Sciences*, pp. 1–51.
- O'Donnell, Ryan (2014). **Analysis of boolean functions.** Cambridge University Press.
- Overschee, Peter Van and Bart De Moor (1994). **"N4SID: Subspace algorithms for the identification of combined deterministic– stochastic systems".** In: *Automatica* 30, pp. 75–93.

- Oymak, Samet and Necmiye Ozay (2019). **"Non-asymptotic identification of lti systems from a single trajectory".** In: *2019 American control conference (ACC)*. IEEE, pp. 5655–5661.
- Simchowitz, Max, Ross Boczar, and Benjamin Recht (2019). **"Learning Linear Dynamical Systems with Semi-Parametric Least Squares".** In: *Conference on Learning Theory (COLT)*.
- Sims, Christopher A (1972). **"Money, income, and causality".** In: *The American economic review* 62.4, pp. 540–552.
- Verhaegen, Michel and Patrick Dewilde (1992). **"Subspace model identification".** In: *International Journal of Control* 56.5, pp. 1187–1210.