DSO 699: Statistics Theory
Special Topics in Data Sciences and Operations

Week 6

Adel Javanmard

USC Marshall
Department of Data Sciences and Operations

September 28, 2023

Announcements

- HW1 is due today (please email your work if not already)
- Solutions to HW1 will be posted on BB tomorrow.
- Midterm exam next week.
  - ✓ in-class exam
  - ✓ you can have access to slides
  - ✓ no internet
  - ✓ **Show up on time**

Recap from the previous class

We talked about:

- False discovery rate (FDR), false discovery proportion (FDP) and false discovery exceedance
- Benjamini-Hochberg procedure
  - ✓ for independent p-values
  - ✓ for arbitrarily dependent p-values
- Storey's procedure (improving BH by estimating the fraction of nulls)

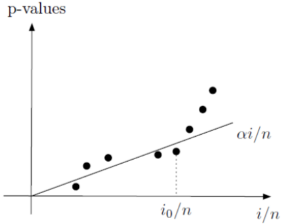
Outline for today

1. Online control of false discovery rate: problem formulation
2. Examples/Applications
3. Methods for online control of FDR
4. Some particular rules: LOND and LORD

Online control of false discovery rate: problem formulation

Benjamini-Hochberg (BH) procedure

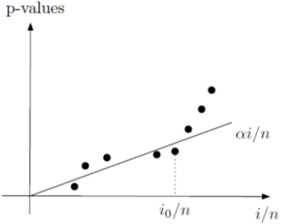
1. Sort p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$
2. Let $i_0 := \max\{i \in [n] : \; p_{(i)} \leq \frac{i}{n}\alpha\}$
3. Reject all p-values $p_{(i)}$ for $i \leq i_0$ (or $p_j \leq p_{(i_0)}$)



BH cutoff is shown by dashed line.

Benjamini-Hochberg (BH) procedure

1. Sort p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$
2. Let $i_0 := \max\{i \in [n] : \ p_{(i)} \leq \frac{i}{n}\alpha\}$
3. Reject all p-values $p_{(i)}$ for $i \leq i_0$ (or $p_j \leq p_{(i_0)}$)



BH cutoff is shown by dashed line.

Some observations:

- BH cutoff depends on all p-values
- i_0 : total $\#$ of rejections, cutoff:= $\frac{\alpha i_0}{n}$

Some concerns about the BH procedure

- Cutoff is a function of α, n and *all* p-values
- If we get a new test(p-values), need to rerun the procedure
- Computational complexity aside, we may need to alter outcomes of previous test!

Offline procedures: requires to know all p-values upfront!

Online setting

Null hypotheses:

$$H_{0,1}, H_{0,2}, \ldots, H_{0,M}$$

Sequence of p-values: one at each time

$$p_1, p_2, p_3, \ldots$$

Ground truth:

$$\theta_1, \theta_2, \theta_3, \ldots [H_{0,i} : \theta_i = 0]$$

Test ouput ($\boldsymbol{p}_1^t = (p_1, \ldots, p_t)$):

$$T_1(\boldsymbol{p}_1^1), T_2(\boldsymbol{p}_1^2), T_3(\boldsymbol{p}_1^3), \cdots \in \{0, 1\}$$

[Foster, Stine, 2007]

Online setting

Null hypotheses:

$$H_{0,1}, H_{0,2}, \ldots, H_{0,M}$$

Sequence of p-values: one at each time

$$p_1, p_2, p_3, \ldots$$

Ground truth:

$$\theta_1, \theta_2, \theta_3, \ldots [H_{0,i} : \theta_i = 0]$$

Test ouput ($p_1^t = (p_1, \ldots, p_t)$):

$$T_1(p_1),\, T_2(p_2;\, T_1),\, T_3(p_3;\, T_1,\, T_2), \cdots \in \{0, 1\}$$

[Foster, Stine, 2007]

Online control of FDR

- $V(n) \equiv$ False discoveries up to time n
- $R(n) \equiv$ Total number of discoveries up to time n

$$\mathrm{FDR}(n) \equiv \mathbb{E}\Big\{\frac{V(n)}{\max(R(n),1)}\Big\}$$

Want $\mathrm{FDR}(n) \leq \alpha$ for all n, θ

Examples/Applications

A/B testing

- Assume (!) that I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - ✓ Users are more likely to click on the first search result
  - ✓ Users are more likely to on top right ads
  - ✓ Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

A/B testing

- Assume (!) that I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - ✓ Users are more likely to click on the first search result
  - ✓ Users are more likely to on top right ads
  - ✓ Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

Example

Idea: *Users click more on the first search result than on the second*

Null H_0: Users are equaly likely to click on first and second

Data:

- n events
- n_1 clicks on the *first* result
- $n_2 = n - n_1$ clicks on the *second* result

p-value

$$H_0 \quad \Rightarrow \quad z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0,1) \quad \Rightarrow \quad p = 1 - \Phi(z) \sim \mathrm{Uniform}([0,1])$$

Example

Idea: *Users click more on the first search result than on the second*

Null H_0**:** Users are equaly likely to click on first and second

Data:

- n events
- n_1 clicks on the *first* result
- $n_2 = n - n_1$ clicks on the *second* result

p-value

$$H_0 \quad \Rightarrow \quad z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0,1) \quad \Rightarrow \quad p = 1 - \Phi(z) \sim \mathrm{Uniform}([0,1])$$

Company policy

Collect p-values every day, and run BH

Problems:

- Centralized
- Controls end-of-day FDR
  Not end-of-year FDR

$\rightarrow$ Online FDR control

Company policy

Collect p-values every day, and run BH

Problems:

- Centralized
- Controls end-of-day FDR
  Not end-of-year FDR

$\rightarrow$ Online FDR control

Anomaly detection

Another example of online FDR control is for anomaly detection in streaming, real–time applications.

NYC taxi dataset:

- records counts of NYC taxi passengers every 30 minutes from July 1, 2014 to January 31, 2015
- There were five known anomaly in this period:
  (the NYC marathon, Thanksgiving, Christmas, New Years day and a snow storm)

[Numenta Anomaly Benchmark (NAB) repository]

NYC taxi dataset

Figure 5: NYC Taxi passenger count time series from July 1st 2014 to Jan 31st 2015. Blue lines are Loess smoothed time series indicating the overall trend change.

[Gang, Sun, Wang 2020]

Formulating anomaly detection as hypothesis testing

Consider a sequence of hypotheses:

$$H_{0,t} : \text{no anomaly at time } t, \qquad H_{1,t} : \text{otherwise.}$$

Hypotheses arrive sequentially!

Constructing p-values:

1. Use **S**easonal-**T**rend decomposition using **L**oess (STL) to write:

$$\text{taxis count} = \underbrace{\text{trend} + \text{seasonal}}_{bias} + \underbrace{\text{remainder}}_{\text{new effects}}$$

```
data_stl <- stlplus(data$Value, n.p =12, s.window = 13,
                    sub.labels = substr(month.name,1,3) , sub.start = initial_month)
```

Formulating anomaly detection as hypothesis testing

Consider a sequence of hypotheses:

$$H_{0,t} : \text{no anomaly at time } t, \qquad H_{1,t} : \text{otherwise.}$$

Hypotheses arrive sequentially!

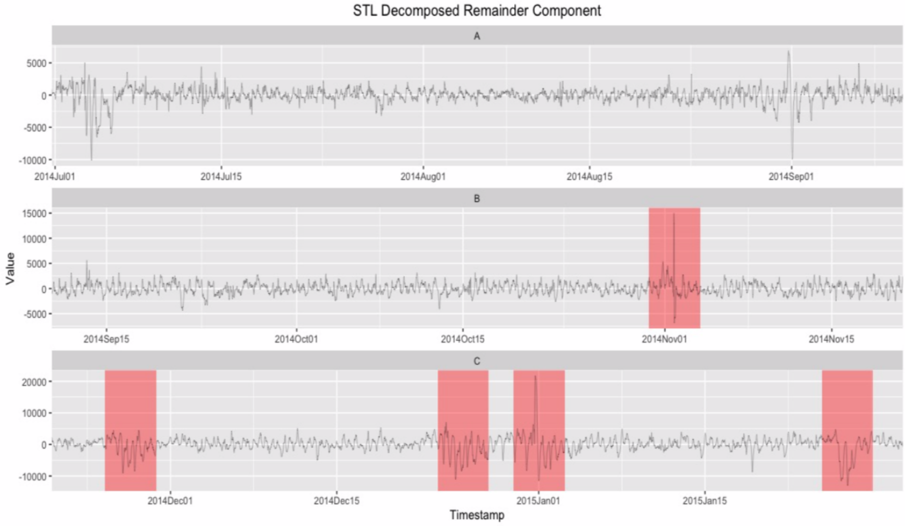
Constructing p-values:

1. Use **S**easonal-**T**rend decomposition using **L**oess (STL) to write:

$$\text{taxis count} = \underbrace{\text{trend} + \text{seasonal}}_{bias} + \underbrace{\text{remainder}}_{\text{new effects}}$$

```
data_stl <- stlplus(data$Value, n.p =12, s.window = 13,
                    sub.labels = substr(month.name,1,3) , sub.start = initial_month)
```

Remainder component

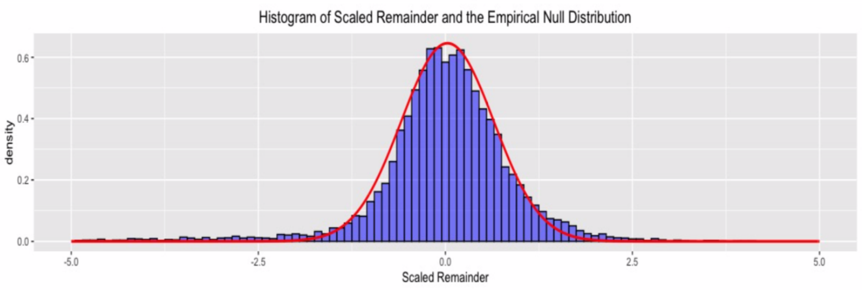


[Gang, Sun, Wang 2020]

Formulating anomaly detection as hypothesis testing (cont'd)

Constructing p-values:

2. Looking at the distribution of the remainder:



Model it as a mixture distribution:

$$r_j \sim \pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0) N(\mu_j, \sigma_j^2), \quad 1 \leq j \leq n$$

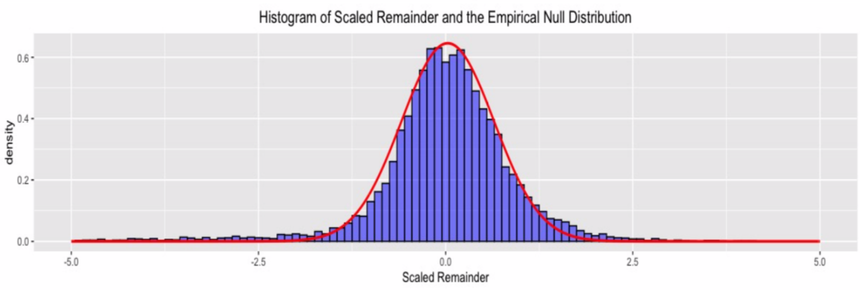
estimate π_0, μ_0, σ_0^2 to obtain the null distribution $F_0 := N(\mu_0, \sigma_0^2)$.

[Jin& Cai 2007]

Formulating anomaly detection as hypothesis testing (cont'd)

Constructing p-values:

2. Looking at the distribution of the remainder:



Model it as a mixture distribution:

$$r_j \sim \pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0) N(\mu_j, \sigma_j^2), \quad 1 \leq j \leq n$$

estimate π_0, μ_0, σ_0^2 to obtain the null distribution $F_0 := N(\mu_0, \sigma_0^2)$.

[Jin& Cai 2007]

Formulating anomaly detection as hypothesis testing (cont'd)

Constructing p-values:

3. Construct p-values as $p_i = 2F_0(-|r_i|)$

Of course, all these steps should be carried out using historical data and not the future data!

Formulating anomaly detection as hypothesis testing (cont'd)

Constructing p-values:

3. Construct p-values as $p_i = 2F_0(-|r_i|)$

Of course, all these steps should be carried out using historical data and not the future data!

Summary.

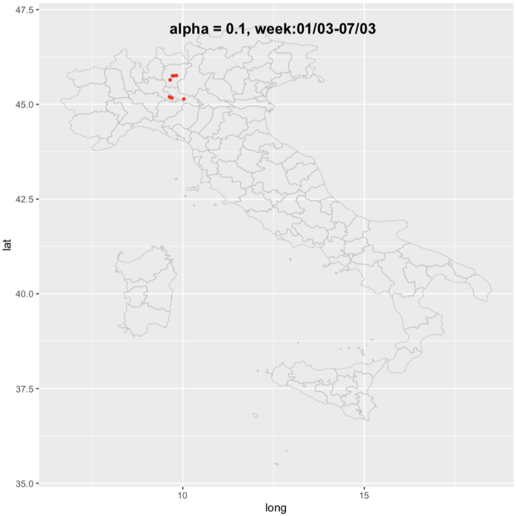
- We have a stream of p-values, $p_1, p_2, \ldots, \ldots$.
- Small $p_i \Rightarrow$ evidence for anomaly
- We formulated anomaly detection as an online hypotheses testing problem!

Another example: COVID outbreak detection

It is another example of anomaly detection:

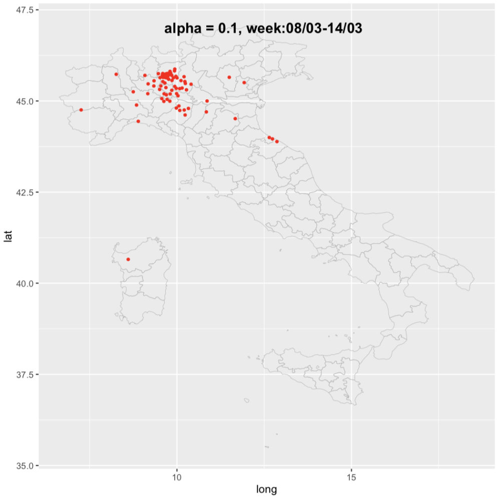
- Working with death rates and would like to detect anomalies
- Seasonal and trend components should be removed
  (eg., Winters have higher death rates)
- Would like to declare emergency if there is significant jump in death rates.
- Any false rejections may trigger an alert system and cause unnecessary lockdowns.
- Another example of online hypotheses testing

An illustration



Performance of LORD algorithm on Italy's death rates. Neighborhoods with significant outbreaks are shown by red marks.

An illustration



Performance of LORD algorithm on Italy's death rates. Neighborhoods with significant outbreaks are shown by red marks.

Methods for online control of FDR

Trivial approach

- $\mathrm{FD}(n) \equiv$ False discoveries up to time n
- $\mathrm{D}(n) \equiv$ Total number of discoveries up to time n

$$\mathrm{FDR}(n) \equiv \mathbb{E}\left\{ \frac{\mathrm{FD}(n)}{\max(\mathrm{D}(n), 1)} \right\}$$

Trivial approach

- $\mathrm{FD}(n) \equiv$ False discoveries up to time n
- $\mathrm{D}(n) \equiv$ Total number of discoveries up to time n

$$\mathrm{FDR}(n) \equiv \mathbb{E}\Big\{ \frac{\mathrm{FD}(n)}{\max(\mathrm{D}(n), 1)} \Big\}$$

Bonferroni:

- Choose $\beta_i \in [0, 1]$, $\sum_{i=1}^{\infty} \beta_i \le \alpha$
- Set

$$T_i = \begin{cases} 1 & \text{if } p_i \le \beta_i, \\ 0 & \text{otherwise.} \end{cases}$$

Trivial approach

- $\mathrm{FD}(n) \equiv$ False discoveries up to time n
- $\mathrm{D}(n) \equiv$ Total number of discoveries up to time n

$$\mathrm{FDR}(n) \equiv \mathbb{E}\Big\{\frac{\mathrm{FD}(n)}{\max(\mathrm{D}(n),1)}\Big\}$$

Bonferroni:

- Choose $\beta_i \in [0,1]$, $\sum_{i=1}^{\infty} \beta_i \le \alpha$
- Set

$$T_i = \begin{cases} 1 & \text{if } p_i \le \beta_i, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed

$$\mathrm{FDR}(n) \le \mathbb{E}\{V(n)\} \le \sum_{i:\,\theta_i=0} \mathbb{P}(p_i \le \beta_i) = \sum_{i:\,\theta_i=0} \beta_i \le \alpha$$

Very conservative!

A better approach?

What rules are we allowed to use?

Online rule:

- Compare p-value p_i with the threshold α_i:

$$T_i = \mathbb{I}(p_i \leq \alpha_i)$$

- thresholds α_i can only be functions of *previous* outcomes

$$\alpha_i = f(T_1, T_2, \ldots, T_{i-1}).$$

How to choose rule f to ensure $\mathrm{FDR}(n) \leq \alpha$ for all n, θ?

Generalized Alpha-investing

A game between statistician and Nature

- Initial alpha-wealth to "invest" in the tests

$$W(0) = \alpha$$

- At step j Statistician pays ϕ_j; if rejects H_j earns ψ_j.

$$W(j) - W(j-1) = -\phi_j + T_j \psi_j \,.$$

- Stops when the wealth becomes negative ($W(j) < 0$).

Generalized Alpha-investing

A game between statistician and Nature

- Initial alpha-wealth to "invest" in the tests

$$W(0) = \alpha$$

- At step j Statistician pays ϕ_j; if rejects H_j earns ψ_j.

$$W(j) - W(j-1) = -\phi_j + T_j \psi_j \,.$$

- Stops when the wealth becomes negative ($W(j) < 0$).

$$\phi_j = \phi_j(T_1, \ldots, T_{j-1}) \qquad \psi_j = \psi_j(T_1, \ldots, T_{j-1})$$

Assumptions on ϕ_j and ψ_j

We want the followings hold for all j:

- Don't bet the money you don't have!

$$\phi_j \leq W(j-1)$$

- At each step, the wealth changes by at most α:

$$\psi_j \leq \phi_j + \alpha$$

- An upper bound on the threshold α_j:

$$\psi_j \leq \frac{\phi_j}{\alpha_j} + \alpha - 1$$

- For all j, if $W(j-1) = 0$, then $\alpha_j = 0$.

A theorem

Definition (monotonicity of a rule)

For $x, y \in \{0,1\}^n$, we write $x \preceq y$ if $x_j \leq y_j$ for all $j \in [n]$.
We say that an online rule is monotone if for all j, α_j is monotone-decreasing w.r.t to this partial ordering:

$$x \preceq y \Rightarrow \alpha_j(x) \leq \alpha_j(y) \quad \forall j.$$

A theorem

Definition (monotonicity of a rule)

For $x, y \in \{0, 1\}^n$, we write $x \preceq y$ if $x_j \le y_j$ for all $j \in [n]$.
We say that an online rule is monotone if for all j, α_j is monotone-decreasing w.r.t to this partial ordering:

$$x \preceq y \Rightarrow \alpha_j(x) \le \alpha_j(y) \quad \forall j.$$

Theorem [Javanmard, Montanari 2015]

Suppose the null p-values are jointly independent, and independent from the non-null p-values. Them, for any monotone generalized alpha investing rule

$$\sup_{n,\theta} \mathrm{FDR}(n) \le \alpha \,.$$

Some particular rules: LOND and LORD

A simple rule

LOND (Levels based On Number of Discoveries)

- Choose non-increasing sequence of $\gamma_j \in [0,1]$, such that $\sum_{j=1}^{\infty} \gamma_j = 1$.
- $R(j-1) \equiv$ Number of rejection by time $j-1$
- Start with wealth $W(0) = \alpha$ and set

$$
\begin{aligned}
&(\text{Bet}) \quad\ \ \phi_j = \alpha\gamma_j(R(j-1)+1) \\
&(\text{Reward})\ \psi_j = \alpha
\end{aligned}
$$

- Significance levels are set as $\alpha_j = \phi_j$

A simple rule

LOND (Levels based On Number of Discoveries)

- Choose non-increasing sequence of $\gamma_j \in [0, 1]$, such that $\sum_{j=1}^{\infty} \gamma_j = 1$.
- $R(j-1) \equiv$ Number of rejection by time $j-1$
- Start with wealth $W(0) = \alpha$ and set

$$\begin{aligned} (\text{Bet}) \quad & \phi_j = \alpha\gamma_j(R(j-1)+1) \\ (\text{Reward}) \quad & \psi_j = \alpha \end{aligned}$$

- Significance levels are set as $\alpha_j = \phi_j$

LOND is a generalized alpha investing rule for any such sequence of $\{\gamma_i\}_{i\in\mathbb{N}}$

(*Why?*)

• An inherent positive feedback (more discoveries at the beginning would lead to more future discoveries!)

Another online rule

LORD (Levels based On Recent Discovery)

- Choose non-increasing sequence of $\gamma_j \in [0,1]$, such that $\sum_{j=1}^{\infty} \gamma_j = 1$.
- $\tau_j \equiv$ Time of the last discovery before j
- Set

$$\begin{aligned} (\text{Bet}) \quad & \phi_j = \gamma_{j-\tau_j} W(\tau_j) \\ (\text{Reward}) \quad & \psi_j = \alpha \end{aligned}$$

- Significance levels are set as $\alpha_j = \phi_j$

Another online rule

LORD (Levels based On Recent Discovery)

- Choose non-increasing sequence of $\gamma_j \in [0, 1]$, such that $\sum_{j=1}^{\infty} \gamma_j = 1$.
- $\tau_j \equiv$ Time of the last discovery before j
- Set

$$\begin{aligned} (\text{Bet}) \quad & \phi_j = \gamma_{j-\tau_j} W(\tau_j) \\ (\text{Reward}) \; & \psi_j = \alpha \end{aligned}$$

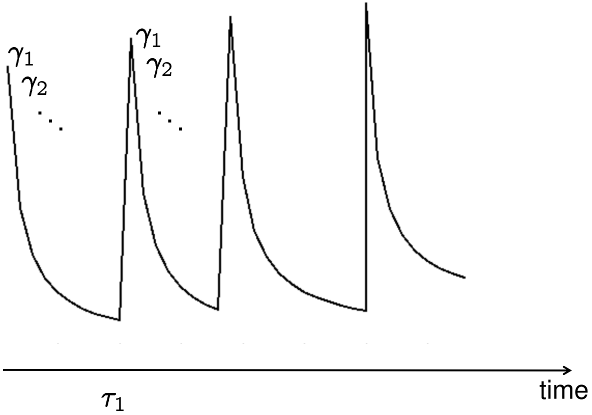
- Significance levels are set as $\alpha_j = \phi_j$

LORD is a generalized alpha investing rule for any such sequence of $\{\gamma_i\}_{i \in \mathbb{N}}$

(*Why?*)

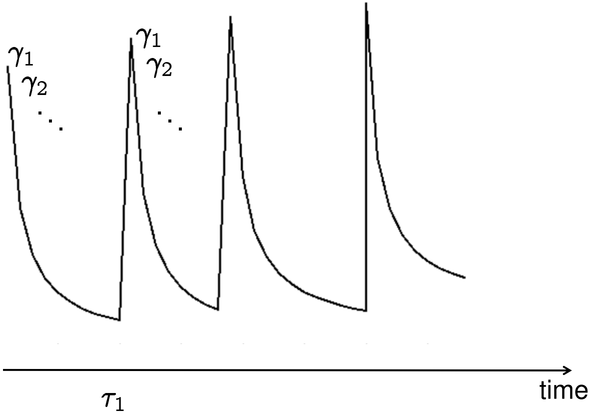
An illustration of LORD

- trend of test levels α_j



An illustration of LORD

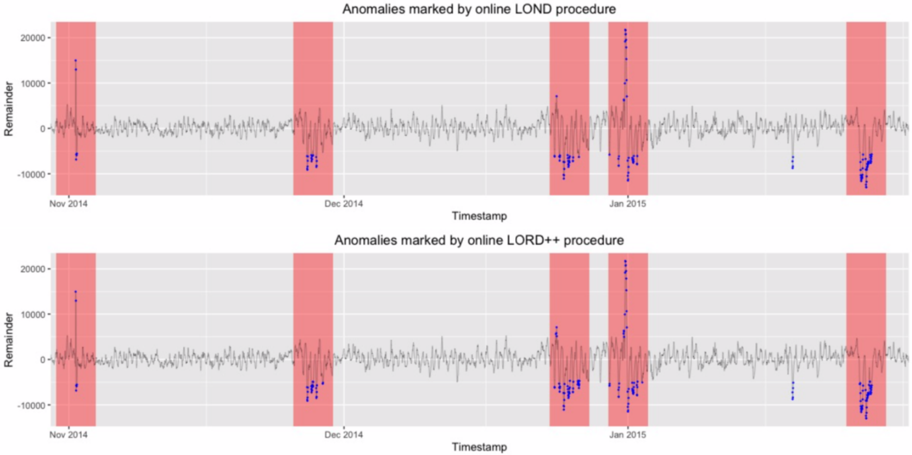
- trend of test levels α_j



- Each discovery increases chance for further discoveries.
- This special structure helps to have batched discoveries (why?)

Back to NYC taxi example

Performance of LOND and LORD++



Anomaly points detected by LOND and LORD++ algorithms. Both detect the most anomaly points within the labeled window marked by red rectangles. Nominal significance level chosen as 0.0001.

Recap

We talked about

- Online false discovery rate
  (deciding on tests which arrive sequentially without knowledge about the future tests/p-values, no possibility to retract your previous decisions)
- Applications
  - ✓ A/B testing
  - ✓ Anomaly detection
- Methodology (Generalized alpha investing rules)
  - ✓ LOND
  - ✓ LORD

Next week

- In-class midterm exam

FIGHT ON!