PhD Notebook

Patrick Li

2021-07-08

Contents

Chapter 1

Welcome

I am Patrick Li.

Chapter 2

Introduction

This note consists of:

1. records of weekly meetings
2. literature review
3. TO-DO list
4. milestones
5. links to resources

Chapter 3

Literature

3.1 Graphical Inference for Infovis

BibTex:

```
@article{wickham2010graphical,
  title={Graphical inference for infovis},
  author={Wickham, Hadley and Cook, Dianne and Hofmann, Heike and Buja, Andreas},
  journal={IEEE Transactions on Visualization and Computer Graphics},
  volume={16},
  number={6},
  pages={973--979},
  year={2010},
  publisher={IEEE}
}
```

3.1.1 Keywords

Statistics, visual testing, permutation tests, null hypotheses, data plots.

3.1.2 Introduction

Infovis focuses on uncovering new relationships by **tools of curiosity**, but most statistical methods focuses on examining relationships by **tools of skepticism**. Neither extreme is good. Hence, graphical inference try to fill the gap between them. It claims that this kind of inference can provide a tool for skepticism that can be applied in a curiosity-driven context.

3.1.3 What is inference and why do we need it?

Inference is about drawing conclusions about the population from the sample. There are two components of statistical inference, testing and estimation. For graphical inference, the focus is to test whether what we see in a plot of the sample is an accurate reflection of the entire population or not. The test statistic in visual inference is a plot of the data. A **null dataset** is a sample from the null distribution, and a **null plot** is a plot of a null dataset. The benefit of visual inference is that it can be used in complex data analysis settings that do not have corresponding numerical tests.

3.1.4 Protocols of graphical inference

3.1.4.1 Rorschach

Rorschach protocol is used to calibrate our vision to the natural variability in plots in which the data is generated from scenarios consistent with the null hypothesis.

3.1.4.2 Line-up

The line up is consisted of $n-1$ decoys and 1 plot of the true data. If we set $n = 19$, then under the null hypothesis, there is only 5% chance to pick the plot of the true data. If we recruit K observers, then under the null hypothesis, the p-value is $P(B(K, 0.05) \geq k)$.

3.1.5 Examples

To use the line-up protocol, we need to: 1. Identify the question the plot is trying to answer 2. Characterize the null-hypothesis 3. Figure out how to generate null datasets

There are two techniques that can be applied in many caeses:

1. Resampling. Permutation and bootstrapping.
2. Simulated data from a assumed model.

3.1.5.1 Tag clouds

A tag cloud can be used to visualize frequency of words in a document. Words are arranged in various ways, often alphabetically, with size proportional to their frequency. The null hypothesis for a comparison tag cloud is that the two documents are equivalent, the frequency of words is the same in each document. Null data can be generated by randomly permute the one of the column.

3.1.5.2 Scatterplot

A scatterplot displays the relationship between x and y. A strong null hypothesis is that there is no relationship between x and y variables.

3.1.6 Power

The power of a statistical test is the probability of correctly convicting a guilty data set. The capacity to detect specific structure in plots can depend on the perceptual properties.

3.1.7 Use

An R package: nullabor

Chapter 4

Meetings

4.1 March 10, 2021 - Week 2

1. Human subject experiment and Monash permission
    - There is a workshop in May
2. Use simulated data to set up an experiment
3. Check out Gallery of graphs (a name yan … not sure)
4. The experiment could start from residual plot
5. Q-Q plot could also be considered
6. A revelant research - residual calculation by kaiwen - master project
7. Use Appen survey to collect data
8. Build a Github to-do list, meeting record and summary of the literature
9. There may be some development in the theory by Nancy Reid - theoretical statistician (recent work)
10. Consider using Kears to build a computer vision model

4.2 March 17, 2021 - Week 3

1. Read Susan Vanderplas's personal website to find additional information
2. Check out NUMBAT residual plot comparision - summer-vis-inf : Aarathy Babu - code examples
3. Read human subject premission examples (sent by Di)
4. Check out top-up application
5. Consider to use `Edibble` to set up the experiment
6. Build the PhD repo
7. Consider to use non-shiny framework

4.3 March 21, 2021 - Week 4

1. A short meeting - late for 30 minutes
2. Aarathy introduces her repo

4.4 March 28, 2021 - Week 5

1. Discuss the options of building a alternative hypothesis in residual plot
2. AR, heterogeneity of variance, endogeneity, skewness, exp distribution, poisson distribution and missing covariance
3. Choice of plot design (loess or $y = 0$ line), number of lineup (5~15), number of observations
4. Use flow chart to illustrate the choices
5. Literature review - check previous designs
6. Build bookdown to track records

Chapter 5

TODO

5.1 Week 3

- ☒ Check human subject experiment and Monash permission materials
  - https://www.intranet.monash/researchadmin/start/ethics/human
- ☒ Check Kaiwen's project & paper
- ☐ Check Gallery of graphs - unclear author

5.2 Week 4

- ☒ Build prototype of html webpage to collect data
- ☒ Send data to google sheet
- ☒ Check summer-vis-inf
- ☒ Read examples sent by Di
- ☐ Build PhD repo
  - ☒ meetings
  - ☐ paper
  - ☒ TODO
- ☐ Check Susan Vanderplas's website
  - ☐ paper
  - ☐ talks
  - ☐ posts

5.3 Week 5

- ☒ modify the webpage to be able to select multiple plots
- ☒ Attempt to generate data from one assumed model
- ☐ draft Human Ethics Application Form

5.4 Week 6

- ☐ Do the literature review of previous design
- ☐ Draw a flow chart to illustrate the design

Chapter 6

Milestones

Chapter 7

Works

Visual inference is an alternative to the traditional statical tests. It is particularly useful when there is no suitable numerical tests. However, one of the drawbacks of visual inference is it needs to be conducted by human, which limits its use cases.

The aim of this work is to provide a tool for automating the test procedure of visual inference. In other words, provided a lineup, the agent developed by us would need to be able to, at least, pick up the most different plot.

7.1 S1 2021

Considering the lineup protocol is a graphical representation of the data, we can define the choice function f for a perfect observer on a collection of nonempty sets

$$X = \{\{p_{1,1}, p_{1,2}, ..., p_{1,M}\}, \{p_{2,1}, p_{2,2}, ..., p_{2,M}\}, ..., \{p_{L,1}, p_{L,2}, ..., p_{L,M}\}\}$$

and assigns one element of each set S in this collection to S by $f(S)$, where $p_{l,m}$ is the ith plot of the lth lineup for $1 \leq l \leq L$ and $1 \leq m \leq M$, M is the size of the lineup and L is the number of lineups. Thus, f is a perfect selector which always select the visually most different plot.

We could further assume there exists a real-valued function g such that

$$f(S) = \sup\{g(s_1; S\ \ \{s_1\}), g(s_2; S\ \ \{s_2\}), ..., g(s_M; S\ \ \{s_M\})\},$$

where s_i is the ith element of the set S for $1 \leq i \leq M$. Then, there is a natural explanation of $g(s_i; S\ \ \{s_i\})$, which is the visual difference between s_i and the rest of the plots or the strength of the signal of s_i. Finding g is not trivial. There are some previous researches that attempt to find such a function with manually defined distance metrics under relatively simple graphical settings and show promising results. However, those distance metrics are case-specific, which means new distance metrics needs to be defined, if they could be defined, for new cases. Obviously, the function g is the linkage between visual test and numerical tests. Unless we can properly define "visual difference" and accurately quantify it, we may never find g, not to mention the possibility of g does not exist.

An alternative will be to find the selector f directly. Unfortunately, perfect observer does not exist, but the selector can still be approximated. Given a lineup and K imperfect observers with probability $p_i \geq 1/M$ to pick the most different plot for $1 \leq i \leq K$, the most frequently chosen plot is a reasonable estimator of the output of the selector f.