PhD Notebook

Patrick Li

2021-04-06

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 - Hadley Wickham, Dianne Cook, Heike Hofmann and Andreas Buja (2010)

3.1.1 Section 1 - Introduction

In this section, the paper argues that much research in Infovis and most statistical methods are on two extremes. One focuses on finding relationships. Another focuses on checking whether a relationship really exists.

Some previous attempts of graphical inference are mentioned in this section. The 2009 paper which develops the statistical concepts of graphical inference is also mentioned.

It then summarises the structure of the paper.

3.1.2 Section 2 - What is inference and why do we need it?

This section starts by stating the two components of statistical inference, which are testing and estimation. It then defines the meaning of inference in graphics.

It uses the criminal justice system as an example to explain the principles of hypothesis testing, and points out the difference between traditional testing and visual testing, which are the test statistic and the mechanism of computing similarity.

To be able to produce a visual test, null datasets which are samples from the null distribution are required. A null plot is a plot of a null dataset.

It ends the section by making an important argument that the introduction of the visual test is not meant to replace traditional tests. “Traditional statistical tests are well studied, well-formulated and work best when data is well-behaved, following a known distribution in relatively simple scenarios.” However, traditional statistical tests do not cover all of the aspects we want to test. On the other hand, visual tests can be used in complex data analysis setting when traditional tests are unavailable.

3.1.3 Section 3 - Protocols of graphical inference

This section introduces two protocols for graphical inference: the “Rorshach” and the “line-up”.

The Rorschach protocol is used to calibrate our vision. In most of the imte, it contains only null plots.

The line-up protocol consists of a plot of the true data and n-1 null plots. N is typically set to be 19.

3.1.4 Section 4 - Examples

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

It provides a list of questions that we would like to answer in commonly seen data plots.

The null data can be generated in two ways in most cases: resampling and simulation.

One of the benefits of using permutation is it preserves the marginal distribution of each variable.

3.1.5 Section 5 - Power

The power of the visual test depends on many factors. An appropriate choice of plot is important to increase the power. For large datasets, aggregation is one of the considerations. It can significantly increase the power of the test.

3.1.6 Section 6 - Use

One can implement these two protocols using the R package `nullabor`.

3.1.7 Section 7 - Conclusion

"Graphical inference is important because it helps us to avoid false convictions."

3.1.8 Bib

```
@article{Wickham2010,
abstract = {How do we know if what we see is really there? When visualizing data, how do we avoid
author = {Wickham, Hadley and Cook, Dianne and Hofmann, Heike and Buja, Andreas},
doi = {10.1109/TVCG.2010.161},
issn = {10772626},
journal = {IEEE Transactions on Visualization and Computer Graphics},
keywords = {Statistics,data plot,null hypotheses,permutation tests,visual testing},
number = {6},
pages = {973--979},
pmid = {20975134},
title = {{Graphical inference for infovis}},
volume = {16},
year = {2010}
}
```

3.2 Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods - William S. Cleveland and Rovert McGill (1984)

```
@article{Cleveland1984,
abstract = {The subject of graphical methods for data analysis and for data presentation needs a
author = {Cleveland, William S. and McGill, Robert},
doi = {10.2307/2288400},
issn = {01621459},
journal = {Journal of the American Statistical Association},
month = {sep},
number = {387},
pages = {531--554},
publisher = {JSTOR},
title = {{Graphical Perception: Theory, Experimentation, and Application to the Development of Gr
volume = {79},
year = {1984}
}
```

3.3 Others

```
@book{wickham2016ggplot2,
author = {Wickham, Hadley},
file = {:Users/patrickli/Desktop/papers/(Use R!) Hadley Wickham - ggplot2_ Elegant Gra
publisher = {springer},
title = {ggplot2: elegant graphics for data analysis},
year = {2016}
}

@book{Wilkinson2005,
abstract = {The grammar of graphics (GoG) denotes a system with seven classes embedded
address = {New York},
author = {Wilkinson, Leland},
booktitle = {The Grammar of Graphics},
doi = {10.1007/0-387-28695-0},
edition = {2},
file = {:Users/patrickli/Desktop/papers/(Statistics and Computing) Leland Wilkinson, D
publisher = {Springer-Verlag},
title = {{The Grammar of Graphics}},
year = {2005}
}
@article{Zhao2013MindRU,
abstract = {Graphical statistics plays a very important job in research and industry. A
author = {Zhao, Y and Cook, D and Hofmann, H and Majumder, M and Chowdhury, N R},
file = {:Users/patrickli/Desktop/papers/Mind Reading Using an Eyetracker to See How Pe
journal = {International Journal of Intelligent Technologies and Applied Statistics},
number = {4},
pages = {393--413},
title = {{Mind Reading: Using an Eye-Tracker to See How People are Looking at Lineups}}
volume = {6},
year = {2013}
}
@article{Hofmann2012,
abstract = {Lineups [4, 28] have been established as tools for visual testing similar t
author = {Hofmann, Heike and Follett, Lendie and Majumder, Mahbubul and Cook, Dianne},
doi = {10.1109/TVCG.2012.230},
file = {:Users/patrickli/Desktop/papers/Graphical Tests for Power Comparison of Competi
issn = {10772626},
journal = {IEEE Transactions on Visualization and Computer Graphics},
keywords = {Efficiency of displays,Lineups,Power comparison,Visual inference},
number = {12},
pages = {2441--2448},
title = {{Graphical tests for power comparison of competing designs}},
volume = {18},
```

```
year = {2012}
}

@article{RoyChowdhury2015,
abstract = {Statistical graphics play an important role in exploratory data analysis, model check
author = {{Roy Chowdhury}, Niladri and Cook, Dianne and Hofmann, Heike and Majumder, Mahbubul and
doi = {10.1007/s00180-014-0534-x},
file = {:Users/patrickli/Desktop/papers/RoyChowdhury2015_Article_UsingVisualStatisticalInferenc.p
issn = {16139658},
journal = {Computational Statistics},
keywords = {Data mining,Lineup,Projection pursuit,Statistical graphics,Visualization},
month = {nov},
number = {2},
pages = {293--316},
publisher = {Springer Verlag},
title = {{Using visual statistical inference to better understand random class separations in hig
url = {https://link.springer.com/article/10.1007/s00180-014-0534-x},
volume = {30},
year = {2015}
}
@article{Loy2017,
abstract = {The complexity of linear mixed-effects (LME) models means that traditional diagnostic
archivePrefix = {arXiv},
arxivId = {1502.06988},
author = {Loy, Adam and Hofmann, Heike and Cook, Dianne},
doi = {10.1080/10618600.2017.1330207},
eprint = {1502.06988},
file = {:Users/patrickli/Desktop/papers/Model Choice and Diagnostics for Linear Mixed Effects Mod
issn = {15372715},
journal = {Journal of Computational and Graphical Statistics},
keywords = {Lineup protocol,Model diagnostics,Model selection,Statistical graphics,Visual inferen
month = {jul},
number = {3},
pages = {478--492},
publisher = {American Statistical Association},
title = {{Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street
url = {https://www.tandfonline.com/doi/abs/10.1080/10618600.2017.1330207},
volume = {26},
year = {2017}
}
@article{Majumder2013,
abstract = {Statistical graphics play a crucial role in exploratory data analysis, model checking
author = {Majumder, Mahbubul and Hofmann, Heike and Cook, Dianne},
doi = {10.1080/01621459.2013.808157},
file = {:Users/patrickli/Desktop/papers/Validation of Visual Statistical Inference Applied to Lin
issn = {01621459},
```

journal = {Journal of the American Statistical Association},
keywords = {Data mining,Effect size,Exploratory data analysis,Lineup,Nonparametric tes
number = {503},
pages = {942--956},
publisher = {Taylor & Francis Group},
title = {{Validation of visual statistical inference, applied to linear models}},
url = {https://www.tandfonline.com/doi/abs/10.1080/01621459.2013.808157},
volume = {108},
year = {2013}
}
@article{Chowdhury2018,
abstract = {Graphics play a crucial role in statistical analysis and data mining. Being
author = {Chowdhury, Niladri Roy and Cook, Dianne and Hofmann, Heike and Majumder, Mahl
doi = {10.1080/10618600.2017.1356323},
file = {:Users/patrickli/Desktop/papers/Measuring Lineup Difficulty By Matching Distanc
issn = {15372715},
journal = {Journal of Computational and Graphical Statistics},
keywords = {Cognitive perception,Data mining,Data science,Data visualization,Distance m
month = {jan},
number = {1},
pages = {132--145},
publisher = {American Statistical Association},
title = {{Measuring Lineup Difficulty By Matching Distance Metrics With Subject Choices
url = {https://www.tandfonline.com/doi/abs/10.1080/10618600.2017.1356323},
volume = {27},
year = {2018}
}
@article{Buja2009,
abstract = {We propose to furnish visual statistical methods with an inferential framew
author = {Buja, Andreas and Cook, Dianne and Hofmann, Heike and Lawrence, Michael and L
doi = {10.1098/rsta.2009.0120},
file = {:Users/patrickli/Desktop/papers/Statistical inference for exploratory data anal
issn = {1364503X},
journal = {Philosophical Transactions of the Royal Society A: Mathematical, Physical an
keywords = {Cognitive perception,Permutation tests,Rotation tests,Simulation,Statistica
month = {nov},
number = {1906},
pages = {4361--4383},
pmid = {19805449},
publisher = {Royal Society},
title = {{Statistical inference for exploratory data analysis and model diagnostics}},
url = {https://royalsocietypublishing.org/doi/10.1098/rsta.2009.0120},
volume = {367},
year = {2009}
}

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