MONASH University

Advances in Artificial Intelligence for Data Visualization: Developing Computer Vision Models to Automate Reading of Data Plots, with Application to Predictive Model Diagnostics

Weihao Li

B.Comm. (Hons), Monash University

A thesis submitted for the degree of Doctor of Philosophy at

Monash University in 2022

Department of Econometrics and Business Statistics

Contents

Copyright notice **v**

Abstract **vii**

Declaration **ix**

Acknowledgements **xi**

Preface **xiii**

1 Introduction **1**

1.1 AI: Four Approaches 1
1.2 Predictive Modelling and Visual Diagnostics 3
1.3 Visual Inference 4
1.4 Pre-specification of Visual Discoverable Features 5
1.5 Lineup Protocol 7
1.6 Applications of Lineup Protocol 9
1.7 Limitations of visual tests 9
1.8 Automatic visual inference -> Computer vision 9
1.9 Discussion of potential methods 9

2 Automatic Visual Statistical Inference, with Application to Linear Regression Diagnostics **11**

2.1 Abstract 11
2.2 Introduction 11

A Additional stuff **19**

Bibliography **21**

Copyright notice

(Choose one of the following notices.)

(Notice 1)

© Weihao Li (2022).

The second notice certifies the appropriate use of any third-party material in the thesis. Students choosing to deposit their thesis into the restricted access section of the repository are not required to complete Notice 2.

(Notice 2)

© Weihao Li (2022).

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Abstract

The abstract should outline the main approach and findings of the thesis and must not be more than 500 words.

Declaration

(Standard thesis)

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

(Thesis including published works declaration)

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes (insert number) original papers published in peer reviewed journals and (insert number) submitted publications. The core theme of the thesis is (insert theme). The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the (insert name of academic unit) under the supervision of (insert name of supervisor).

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.) Remove this paragraph for theses with sole-authored work

In the case of (insert chapter numbers) my contribution to the work involved the following:

Thesis chapter	Publication title	Status (published, in press, accepted or returned for revision)	Nature and % of student contribution	Co-author name(s), nature and % of co-author's contribution	Co-author(s), Monash student Y/N
2	xx	xx	xx	xx	N
3	xx	xx	xx	xx	N
4	xx	xx	xx	xx	N
5	xx	xx	xx	xx	N

I have / have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: Weihao Li

Student signature:

Date:

Acknowledgements

I would like to thank my pet goldfish for …

Preface

The material in Chapter 1 has been submitted to the journal *Journal of Impossible Results* for possible publication.

The contribution in Chapter **??** of this thesis was presented in the International Symposium on Nonsense held in Dublin, Ireland, in July 2015.

Chapter 1

Introduction

1.1 AI: Four Approaches

Artificial intelligence (AI) is the field of research concerned with understanding and building machines who can demonstrate intelligence. As discussed in Russell and Norvig (2002), historically, there are disagreements among researchers about the definition of intelligence, which is caused by two critical questions:

1. Should AI act and think humanly or rationally?
2. Without the thought process and reasoning, are behaviours sufficient to demonstrate intelligence?

Based on the answer to the above questions, four major approaches to pursue AI have been established. These approaches can be summarized into a two by two table as shown in figure 1.1, where the row is "Human" vs. "Rational", and the column is "Behaviour" vs. "Thought". Positioning at the top right cell, the **rational agent approach** aims to build agent that perform mathematically perfect acts such that the best expected outcome can always be achieved. In contrast, the **"laws of thought" approach** focus on understanding the logic behind the rationality. Closely related to **cognitive science**, the **cognitive modelling approach** attempts to express theories of human cognition as computer program to mimic the thought process of human. Lastly, the **Turing test approach** is built

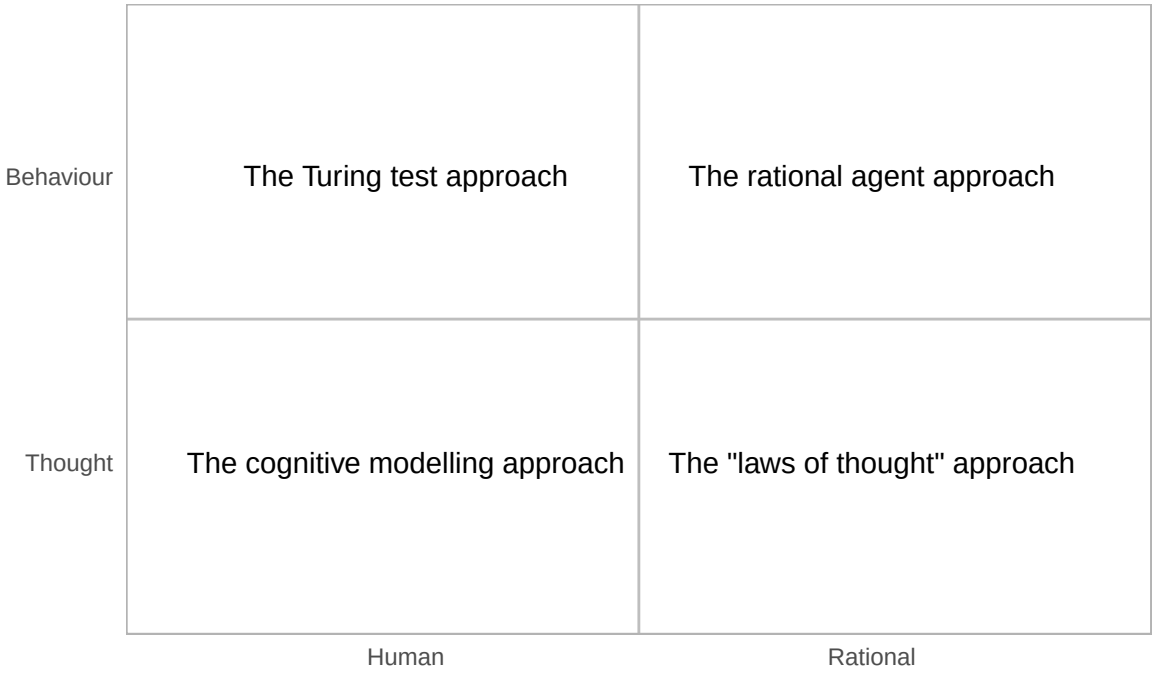

Figure 1.1: *Four possible approaches to pursue AI based on the two dimensions in AI research - human vs. rational and thought vs. behavior (Russell and Norvig, 2002).*

upon the famous **Turing test** proposed by Turing and Haugeland (1950). The test can be roughly described as, whether a human can distinguish another human from a computer with written communications only. To pass the test, several capabilities of computer are required. This includes **natural language processing** for communication with human, **knowledge representation** for encoding knowledge, **automated reasoning** for derivation of conclusions and **machine learning** for improving AI automatically through experience and data. Some researchers argued that written communication is insufficient to demonstrate intelligence, and some degree of physical simulation of a person is still necessary. One such example is the **total Turing test** proposed by Harnad (1991). It adds 3 new requirements to the list, including **computer vision**, **speech recognition** and **robotics**, which are response for interactions with the physical world. Notably, all 7 required capabilities have become major subfields of AI today. And their development has made AI one of the fastest-growing fields in the 21st century (Russell and Norvig, 2002).

With the development of AI, mature AI technologies, such as facial recognition and web recommendation system, have profoundly affected the way modern society operates and

citizen's daily life. This is largely as a consequence of the huge investment in AI industry by the financial market in recent years. Further, the increasingly cheap computing cost and the massive amount of accessible e-commerce data produced in the Internet age provide the possibility for applying data-intensive AI models, which enables AI performance to reach new heights in history (Jordan and Mitchell, 2015). Some AI systems have already been remarkably better than human in certain areas, e.g., game playing. AlphaGo and AlphaZero developed by the Google DeepMind team surpass all human Go players (Silver et al., 2018).

1.2 Predictive Modelling and Visual Diagnostics

Behind the success of AI, a great propotion of AI systems rely on the predictive modelling framework. Donoho (2017) in its summary of data science stated that the concept of this modelling culture could be traced back to an article written by Breiman (2001). In contrast to the generative modelling culture, which aims to develop stochastic models to make inferences about the data generating process, predictive modelling emphasizes the ability of the model to make accurate predictions. Most AI tasks are complex prediction problems where the data mechanism is mysterious, or at least, partly unknowable. Breiman (2001) suggests that generative models are obviously not applicable in these scenarios, while the predictive modelling seeks only an accurate approximated function $f(\boldsymbol{x})$ to describe the relationship between the features $\boldsymbol{x}$ and the responses $\boldsymbol{y}$.

Predictive models are primarily evaluated by predictive accuracy with the use of validation and test data, but in predictive model diagnostics, especially model testing and tuning, data plots play an irreplaceable role. In these diagnostics, though numeric summaries are mostly available and some are even endorsed by finite or asymptotic properties, graphical representation of data is still preferred, or at least needed by researchers, due to its intuitiveness and the possibility to provide unexpected discoveries which may be abstract and unquantifiable.

However, unlike confirmatory data analysis built upon rigorous statistical procedures, e.g., hypothesis testing, visual diagnostics relies on graphical perception - human's ability to interpret and decode the information embedded in the graph (Cleveland and McGill, 1984),

which is to some extent subjective. Further, visual discovery suffers from its unsecured and unconfirmed nature where the degree of the presence of the visual features typically can not be measured quantitatively and objectively, which may lead to over or under-interpretations of the data. One such example is finding a separation between gene groups in a two-dimensional projection from a linear discriminant analysis where there is no difference in the expression levels between the gene groups (Roy Chowdhury et al., 2015).

1.3 Visual Inference

Visual inference was first introduced by Buja et al. (2009) as an inferential framework to extend confirmatory statistics to visual discoveries. This framework redefines the test statistics, tests, null distribution, significance levels and p-value for visual discovery modelled on the confirmatory statistical testing. Figure 2.1 outlines the parallelism between conventional tests and visual discovery.

In visual inference, a visual discovery is defined as a rejection of a null hypothesis, and the same null hypothesis can be rejected by many different visual discoveries (Buja et al.,

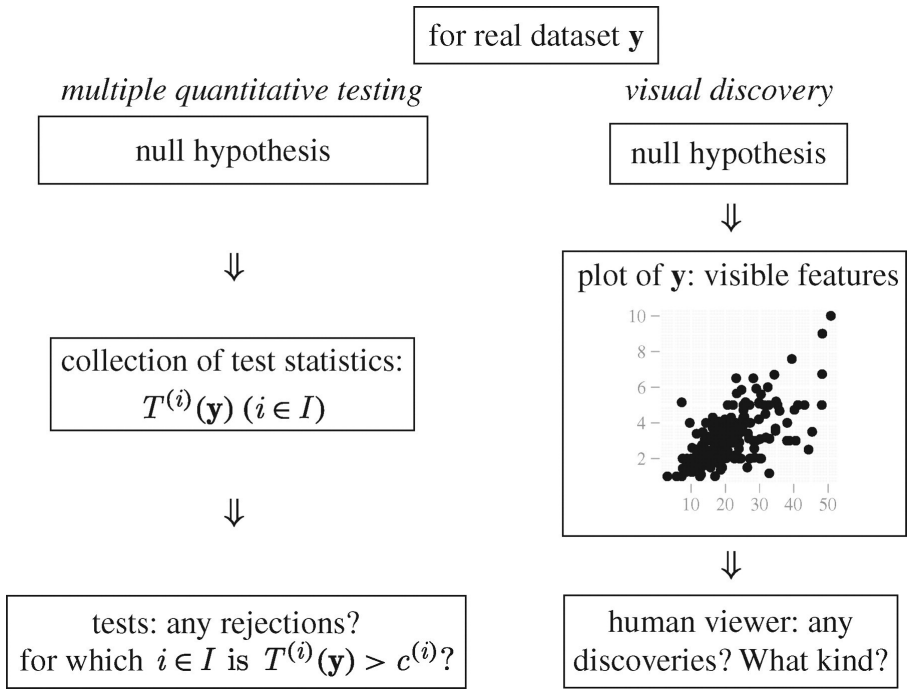

Figure 1.2: *Parallelism between multiple quantitative testing and visual discovery (Buja et al., 2009). Visible features in a plot are viewed as a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$, and any visual discoveries are treated as evidence against the null hypothesis.*

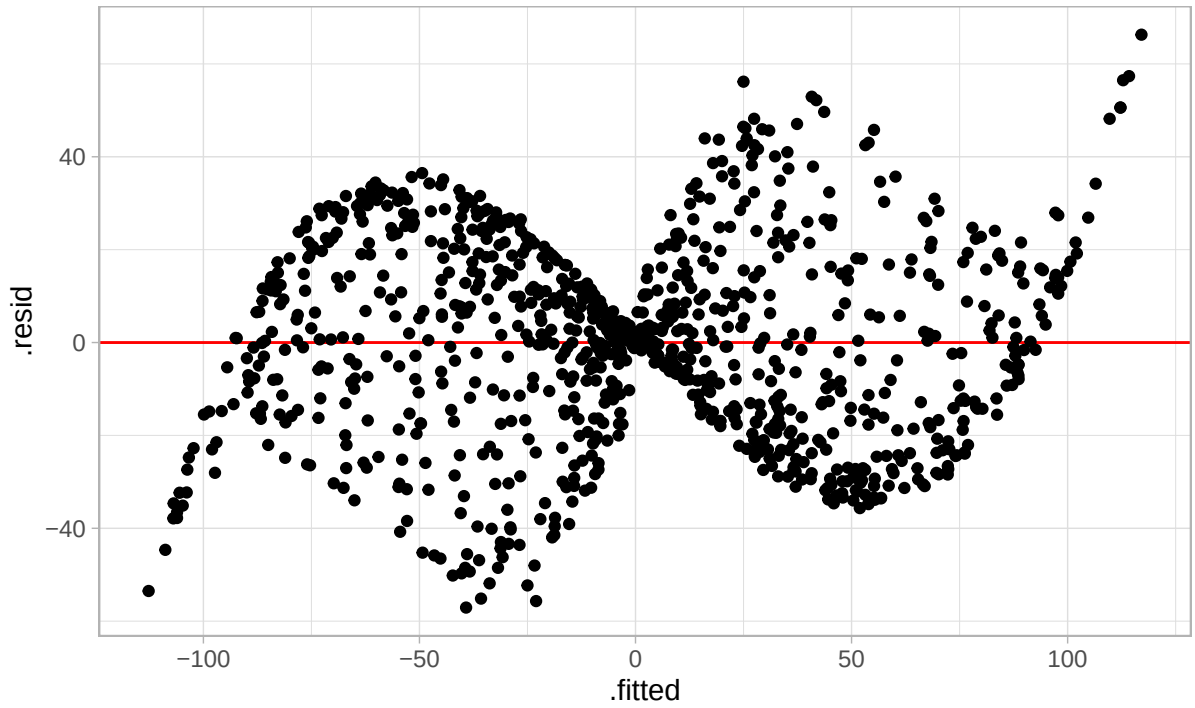

Figure 1.3: *Residuals vs. fitted values plot for a classical linear regression model. The residuals are produced by fitting a two-predictor multiple linear regression model with data generated from a cubic linear model. From the residual plot, "butterfly shape" can be observed which generally would be interpretd as evidence of heteroskedasticity. Further, from the outline of the shape, nonlinear patterns exist. Both visual discoveries are evidence against the null hypothesis, though heteroskedasticity actually does not exist in the data generating process.*

2009). For model diagnostics, the null hypothesis would be the assumed model, while the visual discoveries would be any findings that are inconsistent with the null hypothesis. The same assumed model, such as classical linear regression model, can be rejected by many reasons with residual plot, including nonlinearity and heteroskedasticity as shown in figure 1.3.

1.4 Pre-specification of Visual Discoverable Features

As discussed in Buja et al. (2009), in the practice of model diagnostics, the range of possible visual discoveries is not pre-specified. In other words, people do not explicitly specify which one or more visual features they are looking for before the read of the diagnostic plot. This is concerning since conventional hypothesis testing always requires the pre-specification of the parameter space Θ of the parameter of interest $\theta \in \Theta$ to form a valid

inferential procedure. To address this issue, a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ is defined, where $\mathbf{y}$ is the data and $\boldsymbol{I}$ is a set of all possible visual features. Buja et al. (2009) described each of the test statistics $T^{(i)}(\mathbf{y})$ as a measurement of the degree of presence of a visual feature. Alternatively, Majumder, Hofmann, and Cook (2013) avoids the use of visual features and defined the visual statistics $T(.)$ as a mapping from a dataset to a data plot. Both definitions of visual test statistics are valid, but in the rest of the paper the first definition will be used as it covers some details needed by the following discussion.

The size of the collection $T^{(i)}(\mathbf{y})$ $(i \in I)$ depends on the size of the set I. Thus, if one can define I comprehensively, i.e, pre-specify all the visual discoverable features, the validity issue will be solved. Unfortunately, to our knowledge, there is no such a way to list all visual features. In linear regression diagnostics, possible visual features of a residual plot may be outliers, shapes and clusters. But this is an incomplete list which does not enumerate all the visual features.

Similarly, Wilkinson, Anand, and Grossman (2005) proposed the work called graph theoretic scagnostics, which adopted the idea of "scagnostics" - scatter plot diagnostics from (can't find the 1984 citation). It includes 9 computable scagnostics measures defined on planar proximity graphs: "Outlying", "Convex", "Skinny", "Stringy", "Straight", "Monotonic", "Skewed", "Clumpy" and "Striated" which attempts to describe outliers, shape, density, trend and coherence of the data. This approach is inspiring but it still does not give the complete list of visual discoverable features. In fact, it is possible that such a list will never be complete as suggested in Buja et al. (2009).

Thinking out of the box, Buja et al. (2009) argued that there is actually no need for pre-specification of visual discoverable features. In model diagnostics, when the null hypothesis is rejected, the reasons for rejecting the hypothesis will also be known. This is because observers can not only point out the fact that visual discoveries have been found, but also describe the particular visual features they observed. Those features will correspond to the subset of the collection of visual test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ which resulted in rejection. This argument helps justifies the validity of visual inference.

1.5 Lineup Protocol

With the validity of visual inference being justified, another aspect of hypothesis testing that needs to be addressed is the control of false positive rate or Type I error. Any visual statistic $T^{(i)}(\mathbf{y})$ needs to pair with a critical value $c^{(i)}$ to form a hypothesis test. When a visual feature i is discovered by the observer from a plot, the corresponding visual statistic $T^{(i)}(\mathbf{y})$ may not be known as there is no general agreement on the measurement of the degree of presence of a visual feature. It is only the event that $T^{(i)}(\mathbf{y}) > c^{(i)}$ is confirmed. Similarly, if any visual discovery is found by the observer, we say, there exists $i \in I : \ T^{(i)}(\mathbf{y}) > c^{(i)}$ (Buja et al., 2009).

Using the above definition, the family-wise Type I error can be controlled if one can provide the collection of critical values $c^{(i)}$ $(i \in I)$ such that $P(\text{there exists } i \in I : T^{(i)}(\mathbf{y}) > c^{(i)}|\mathbf{y}) \leq \alpha$, where α is the significance level. However, since the quantity of $T^{(i)}(\mathbf{y})$ may not be known, such collection of critical values can not be provided.

Buja et al. (2009) proposed the lineup protocol as a visual test to calibrate the Type I error issue without the specification of $c^{(i)}$ $(i \in I)$. It is inspired by the "police lineup" or "identity parade" which is the act of asking the eyewitness to identify criminal suspect from a group of irrelevant people. The protocol consists of m randomly placed data plots, where 1 plot is the actual data plot, and $m - 1$ null plots are produced by plotting data simulate from the null distribution which is consistent with the null hypothesis. Then, an observer who have not seen the actual data plot will be asked to point out the most different plot from the lineup.

Under the null hypothesis, it is expected that the actual data plot would have no distinguishable difference with the null plots, and the probability of the observer correctly picks the actual data plot is $1/m$ due to randomness. If we reject the null hypothesis as the observer correctly picks the actual data plot, then the Type I error of this test is $1/m$.

This provides us with an mechanism to control the Type I error, because m - the number of plots in a lineup can be chosen. A larger value of m will result in a smaller Type I error, but the limit to the value of m depends on the number of plots a human willing to view (Buja

et al., 2009). Typically, m will be set to 20 which is equivalent to set $\alpha = 0.05$, a general choice of significance level for conventional testing among statisticians.

Further, if we involve K independent observers in a visual test, and let X be a random variable denoting the number of observers correctly picking the actual data plot. Then, under the null hypothesis $X \sim \text{Binom}_{K,1/m}$, and therefore, the p-value of a lineup of size m evaluated by K observer is given as

$$P(X \geq x) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^{i} \left(\frac{m-1}{m}\right)^{k-i},$$

where x is the realization of number of observers correctly picking the actual data plot (Majumder, Hofmann, and Cook, 2013).

The multiple individuals approach avoids the limit of m, while provides visual tests with p-value much smaller than 0.05. In fact, the lower bound of p-value decreases exponentially as K increases. With just 4 individuals and 20 data plots in a lineup, the p-value could be as small as 0.0001.

Compared to the conventional test, whose power only depends on the parameter of interest θ, several studies (see Hofmann et al., 2012; Majumder, Hofmann, and Cook, 2013, 2014; Roy Chowdhury et al., 2015; Loy, Follett, and Hofmann, 2016) have shown the power of the visual test is subject-specific. Thus, to be able to account for individual's ability, an individual is required to evaluate multiple lineups (Majumder, Hofmann, and Cook, 2013).

Assumes that individuals have the same ability and a lineup has been evaluated by multiple individuals, under the alternative hypothesis, the estimated power for a lineup can be expressed as $\hat{p} = x/K$, the estimated probability of identifying the actual data plot from the lineup. If the individual skill needs to be taken into account, and L lienups have been evaluated by K individuals, Majumder, Hofmann, and Cook (2013) suggests that mixed effects logistic regression model can be fit as:

$$g(p_{li}) = W_{li}\delta + Z_{li}\tau_{li},$$

where $g(.)$ is the logit link function $g(p) = log(p) - log(1-p); 0 \leq p \leq 1$. $W_{li}, 1 \leq i \leq K$, $1 \leq l \leq L$, is the covariate matrix including lineup-specific elements and demographic information of individuals, and δ is a vector of parameters. Z is the random effects matrix, and τ is a vector of variables follow $N(\mathbf{0}, \sigma_\tau \boldsymbol{I}_{KL \times KL})$.

Then, the estimated power for lineup l and individual i can be calculated as $\hat{p}_{li} = g^{-1}(W_{li}\hat{\delta} + Z_{li}\hat{\tau}_{li})$ (Majumder, Hofmann, and Cook, 2013).

1.6 Applications of Lineup Protocol

Lineup protocols are

List applications, fields, significance here.

1.7 Limitations of visual tests

1. infeasible in a large scale
2. unfriendly to vision-impaired people
3. high finical cost and human cost
4. time consuming

similar to Handicraft in pre-industrial society

1.8 Automatic visual inference -> Computer vision

relieve people's workload by automating repeating tasks, and provide standard result in a control environment

the use of technology and machinery to enable mass evaluation of visual tests

1.9 Discussion of potential methods

1.9.1 different approaches of AI

- Not aim for understanding the thought process
- Not aim for mimicking the human vision mechanism

- May be able to define distance metrics to measure difference between data plots for making mathematically prefect decisions
- May be able to use computer vision model to approximate how people evaluate lineups

computer vision model:

- Use human data to train model with human selection as target -> mimic the human behaviour
- Use simulated lineup to train model with actual data plot as target -> assume the actual data plot is the most different one
- Use simulated data plot to train model with null or not null as target -> limited to the alternatives given to the model, and Type I error can not be controlled
- Use simulated data plot to train model, then let the model evaluate each plot of a lineup -> leads to multiple selections in a lineup, and the model does not compare different plots in a lineup
  - Use a very large m to ensure the null hypothesis can be rejected by a classifier slightly better than a random selector. The problem becomes how large the m should be
  - Use multiple models as multiple individuals to evaluate lineup -> Which models should be used? How many models should be used?
- Let the model select multiple plots from a simulated lineup while comparing different plots in a lineup -> how to build such a system?
  - The input is m plots, what is the output? The probability of being the most different one?
  - Then how many of plots should be selected? Top 5? Or cumulative probability greater than a threshold?
  - How to select the threshold?

Chapter 2

Automatic Visual Statistical Inference, with Application to Linear Regression Diagnostics

2.1 Abstract

2.2 Introduction

2.2.1 Model Diagnostics

[ET: suggestion: A model can be fitted to data with no guarantee of a meaningful interpretation. Model diagnostics play an important role in assessing the appropriateness of the model. The assessment can involve examining the goodness of fit, checking if there are potential violations of model assumptions and validating models with external information.]

Model diagnostics is the part of data analysis, preceded by the fit of a model, whose primary objectives are to examine the goodness of fit and reveal potential violations of model assumptions.

[ET: model fit is more than just goodness of fit and checking model assumptions. It's too reductive to say this because you could have a model with good fit (according to some metric) and no model violation but if the way that the data were collected was flawed then scientific interpretation of the model is rendered useless. Diagnostics involve also the domain knowledge checks as well.]

In these diagnostics, though numeric summaries are mostly available and some are even endorsed by finite or asymptotic properties, graphical representation of data is still preferred, or at least needed, due to its intuitiveness and the possibility to provide unexpected discoveries which may be abstract and unquantifiable.

[ET: Generally, your sentences are too long. Break it up into smaller sentences. The generally idea is to make sentences that *continue an idea from the preceding sentence* so it leads from one to the next naturally. E.g. Today is cloudy. Cloudy days are rare. Perhaps today will be different to other days!]

However, unlike confirmatory data analysis built upon rigorous statistical procedures, e.g., hypothesis testing, visual diagnostics relies on graphical perception - human's ability to interpret and decode the information embedded in the graph (Cleveland and McGill, 1984), which is to some extent subjective. Further, visual discovery suffers from its unsecured and unconfirmed nature where the degree of the presence of the visual features typically can not be measured quantitatively and objectively, which may lead to over or under-interpretations of the data. One such example is finding a separation between gene groups in a two-dimensional projection from a linear discriminant analysis where there is no difference in the expression levels between the gene groups (Roy Chowdhury et al., 2015).

[ET: Confirmatory data analysis is well characterised by a rigorous procedure to discern particular hypothesis. The most widespread form of confirmatory data analysis are the use of p-values in testing hypothesis in a frequentist framework. More specifically, hypothesis testing in a frequentist framework involve stating a well-defined hypotheses; summarising the evidence as a test statistic; and calculating the probability of observing a statistic as extreme as the observed test statistic under the null hypothesis. This rigour, however, is often not present when inferences are drawn from a plot.

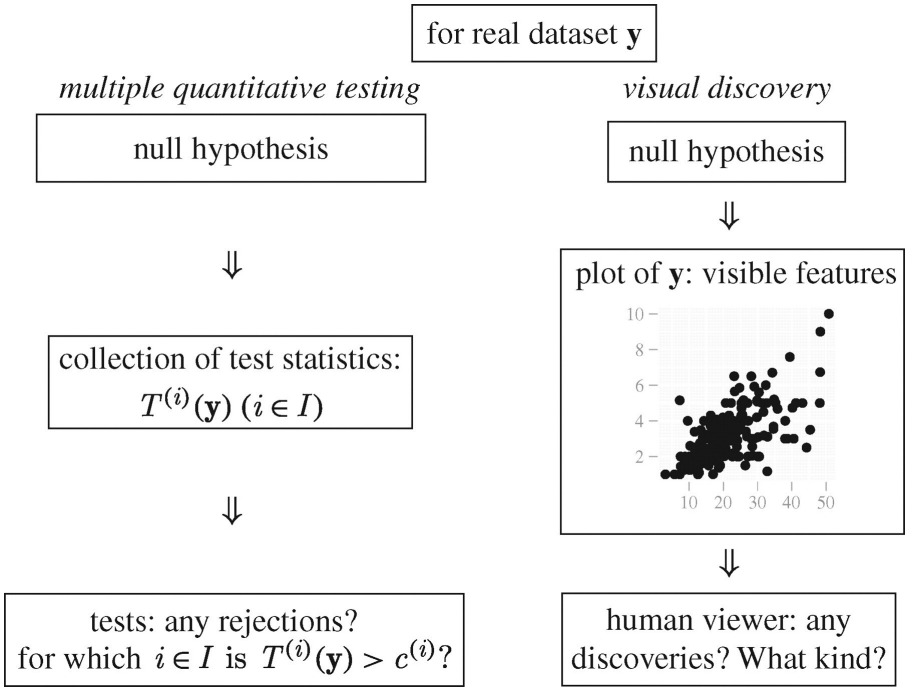

Figure 2.1: *Parallelism between multiple quantitative testing and visual discovery (Buja et al., 2009). Visible features in a plot are viewed as a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$, and any visual discoveries are treated as evidence against the null hypothesis.*

Something about model diagnostics with a plot being the common thing to do..]

2.2.2 Visual Inference

Visual inference was first introduced by Buja et al. (2009) as an inferential framework to extend confirmatory statistics to visual discoveries. This framework redefines the test statistics, tests, null distribution, significance levels and p-value for visual discovery modelled on the confirmatory statistical testing. Figure **??** outlines the parallelism between conventional tests and visual discovery.

[ET: note use Rmd syntax for figures references! So it's easy to use it for HTML later if you want to do something like Earo and have PDF + HTML thesis.]

Parallelism between multiple quantitative testing and visual discovery (Buja et al., 2009). Visible features in a plot are viewed as a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$, and any visual discoveries are treated as evidence against the null hypothesis. Parallelism between multiple quantitative testing and visual discovery (Buja et al., 2009).

In visual inference, a visual discovery is defined as a rejection of a null hypothesis, and the same null hypothesis can be rejected by many different visual discoveries (Buja et al., 2009). For model diagnostics, the null hypothesis would be the assumed model, while the visual discoveries would be any findings that are inconsistent with the hypothesis. The same assumed model, such as classical linear regression model, can be rejected by both nonlinearity and heteroskedasticity with the residual plot as shown in Figure 1.3.

2.2.3 Pre-specification of Visual Discoverable Features

As discussed in Buja et al. (2009), in the practice of model diagnostics, the range of possible visual discoveries is not pre-specified. In other words, people do not explicitly specify which one or more visual features they are looking for before the read of the diagnostic plot. This is concerning since conventional hypothesis testing always requires the pre-specification of the parameter space Θ of the parameter of interest $\theta \in \Theta$ to form a valid inferential procedure. To address this issue, a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ is defined, where $\mathbf{y}$ is the data and $\boldsymbol{I}$ is a set of all possible visual features. Buja et al. (2009) described each of the test statistics $T^{(i)}(\mathbf{y})$ as a measurement of the degree of presence of a visual feature. Alternatively, Majumder, Hofmann, and Cook (2013) avoids the use of visual features and defined the visual statistics $T(.)$ as a mapping from a dataset to a data plot. Both definitions of visual test statistics are valid, but in the rest of the paper, the first definition will be used as it covered some details needed by this work.

The size of the collection $T^{(i)}(\mathbf{y})$ $(i \in I)$ depends on the size of the set I. Thus, if one can define I comprehensively, i.e, pre-specify all the visual discoverable features, the validity issue will be solved. Unfortunately, to our knowledge, there is no such a way to list all visual features. In linear regression diagnostics, possible visual features of a residual plot may be outliers, shapes and clusters. But this is an incomplete list which does not enumerate all the visual features.

Similarly, Wilkinson, Anand, and Grossman (2005) proposed the work called graph theoretic scagnostics, which adopted the idea of "scagnostics" - scatter plot diagnostics from (can't find the 1984 citation). It includes 9 computable scagnostics measures defined on planar proximity graphs: "Outlying", "Convex", "Skinny", "Stringy", "Straight",

"Monotonic", "Skewed", "Clumpy" and "Striated" which attempts to describe outliers, shape, density, trend and coherence of the data. This approach is inspiring but it still does not give the complete list of visual discoverable features. In fact, it is possible that such a list will never be complete as suggested in Buja et al. (2009).

Thinking out of the box, Buja et al. (2009) argued that there is actually no need for pre-specification of visual discoverable features. In model diagnostics, when the null hypothesis is rejected, the reasons for rejecting the hypothesis will also be known. This is because observers can not only point out the fact that visual discoveries have been found, but also describe the particular visual features they observed. Those features will correspond to the subset of the collection of visual test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ which resulted in rejection. This argument helps justifies the validity of visual inference.

2.2.4 Lineup Protocol

With the validity of visual inference being justified, another aspect of hypothesis testing that needs to be addressed is the control of false positive rate or Type I error. Any visual statistic $T^{(i)}(\mathbf{y})$ needs to pair with a critical value $c^{(i)}$ to form a hypothesis test. When a visual feature i is discovered by the observer from a plot, the corresponding visual statistic $T^{(i)}(\mathbf{y})$ may not be known as there is no general agreement on the measurement of the degree of presence of a visual feature. It is only the event that $T^{(i)}(\mathbf{y}) > c^{(i)}$ is confirmed. Similarly, if any visual discovery is found by the observer, we say, there exists $i \in I : \ T^{(i)}(\mathbf{y}) > c^{(i)}$ (Buja et al., 2009).

Using the above definition, the family-wise Type I error can be controlled if one can provide the collection of critical values $c^{(i)}$ $(i \in I)$ such that $P(\text{there exists } i \in I : T^{(i)}(\mathbf{y}) > c^{(i)}|\mathbf{y}) \leq \alpha$, where α is the significance level. However, since the quantity of $T^{(i)}(\mathbf{y})$ may not be known, such collection of critical values can not be provided.

Buja et al. (2009) proposed the lineup protocol as a visual test to calibrate the Type I error issue without the specification of $c^{(i)}$ $(i \in I)$. It is inspired by the "police lineup" or "identity parade" which is the act of asking the eyewitness to identify criminal suspect from a group of irrelevant people. The protocol consists of m randomly placed data plots, where 1 plot is the actual data plot, and $m - 1$ null plots are produced by plotting data

simulate from the null distribution which is consistent with the null hypothesis. Then, an observer who have not seen the actual data plot will be asked to point out the most different plot from the lineup.

Under the null hypothesis, it is expected that the actual data plot would have no distinguishable difference with the null plots, and the probability of the observer correctly picks the actual data plot is $1/m$ due to randomness. If we reject the null hypothesis as the observer correctly picks the actual data plot, then the Type I error of this test is $1/m$.

This provides us with an mechanism to control the Type I error, because m - the number of plots in a lineup can be chosen. Further, if we involve K independent observers in a visual test, and let X be a random variable denoting the number of observers correctly picking the actual data plot. Then, under the null hypothesis $X \sim \text{Binom}_{K,1/m}$, and therefore, the p-value of a lineup of size m evaluated by K observer is given as

$$P(X \geq x) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^{i} \left(\frac{m-1}{m}\right)^{k-i},$$

where x is the realization of number of observers correctly picking the actual data plot (Majumder, Hofmann, and Cook, 2013).

2.2.5 Visual Inference Applied to Linear Regression

How people used visual inference in linear regression?

2.2.6 Limitation of the Visual Inference

What are the limitations?

2.2.7 Computer Vision Model

What is computer vision model?

2.2.8 Contribution

What has been done by this paper?

2.2.9 Structure of This Paper

What is the structure of the paper?

Model diagnostics is the part of data analysis whose primary objectives are to examine the goodness of the model fit and reveal potential violations of the assumptions. Graphical approaches

For regression diagnostics, it may includes the needs of

Linear regression is an modelling approach to describe the relationship between an response variable and one or more explanatory variable. It has been widely used for both generative modeling and predictive modelling.

Regression diagnostics is needed

1. to check whether the assumptions has been violated
2. to check whether the line fit the data

Model diagnostics for linear regression is well developed

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.

Bibliography

Breiman, L (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* **16**(3), 199–231.

Buja, A, D Cook, H Hofmann, M Lawrence, EK Lee, DF Swayne, and H Wickham (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4361–4383.

Cleveland, WS and R McGill (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association* **79**(387), 531–554.

Donoho, D (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics* **26**(4), 745–766.

Harnad, S (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* **1**(1). Publisher: Springer, 43–54.

Hofmann, H, L Follett, M Majumder, and D Cook (2012). Graphical Tests for Power Comparison of Competing Designs. *IEEE Transactions on Visualization and Computer Graphics* **18**(12). Conference Name: IEEE Transactions on Visualization and Computer Graphics, 2441–2448.

Jordan, MI and TM Mitchell (2015). Machine learning: Trends, perspectives, and prospects. en. *Science* **349**(6245), 255–260.

Loy, A, L Follett, and H Hofmann (2016). Variations of Q–Q Plots: The Power of Our Eyes! *The American Statistician* **70**(2). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2015.1077728, 202–214. (Visited on 01/23/2022).

Majumder, M, H Hofmann, and D Cook (2013). Validation of Visual Statistical Inference, Applied to Linear Models. *Journal of the American Statistical Association* **108**(503). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2013.808157, 942–956. (Visited on 01/21/2022).

Majumder, M, H Hofmann, and D Cook (2014). Human Factors Influencing Visual Statistical Inference. *arXiv:1408.1974 [stat]*. arXiv: 1408.1974.

Roy Chowdhury, N, D Cook, H Hofmann, M Majumder, EK Lee, and AL Toth (2015). Using visual statistical inference to better understand random class separations in high dimension, low sample size data. en. *Computational Statistics* **30**(2), 293–316. (Visited on 01/23/2022).

Russell, S and P Norvig (2002). Artificial intelligence: a modern approach.

Silver, D, T Hubert, J Schrittwieser, I Antonoglou, M Lai, A Guez, M Lanctot, L Sifre, D Kumaran, T Graepel, T Lillicrap, K Simonyan, and D Hassabis (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. en. *Science* **362**(6419), 1140–1144. (Visited on 02/09/2022).

Turing, AM and J Haugeland (1950). Computing machinery and intelligence. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, 29–56.

Wilkinson, L, A Anand, and R Grossman (2005). Graph-theoretic scagnostics. In: *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* ISSN: 1522-404X, pp.157–164.