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

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 Artificial intelligence and predictive modelling

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 on 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. They 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 upon the famous

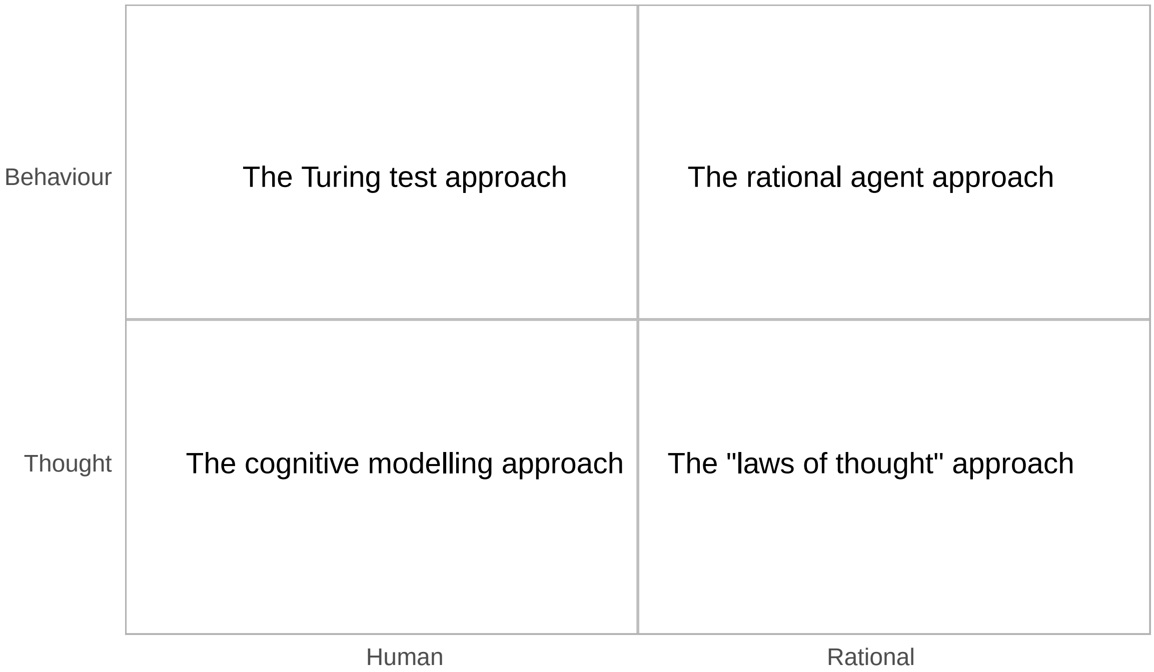

Figure 1.1: *test*

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**, **knowledge representation**, **automated reasoning** and **machine learning.** Some researchers argued that 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 new requirements to the list, including **computer vision**, **speech recognition** and **robotics** (Russell and Norvig, 2002). Notably, all 6 required capabilities have become major subfields of AI today.

The rational agent approach

which impacts the version of AI pursued by researchers.

due to the definition of intelligence,

It has been pursued in many different ways by the researchers. The major separations between different approaches are the reasoning and thoughts, focus on the results or the action. Act and think humanly or rationally.

Two dimensions

machine think humanly, think rationally, acting humanly, acting rationally

A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome.

versions due to

human performance

The cognitive modeling approach

The "laws of thought" approach

The rational agent approach

human vs. rational2 and thought vs. behavior

Some have defined intelligence in terms of fidelity to human performance, while others prefer an abstract, formal definition of intelligence called rationality—loosely speaking, doing the "right thing." The subject matter itself also varies: some consider intelligence to be a property of internal thought processes and reasoning, while others focus on intelligent behavior, an external characterization.

As required by the Turing

As defined by Turing test,

The Turning test is well known

The field of artificial intelligence, or AI, is concerned with not just understanding but also building intelligent entities—machines that can compute how to act effectively and safely in a wide variety of novel situations.

1.2 AI -> predictive modelling

1.3 Predictive modelling -> Model diagnostics

1.4 Limitation of MD/EDA (before 2009)

1.5 Visual inference

1.6 Limitation of visual inference

1.7 Automatic visual inference -> Computer vision

Chapter 2

Automatic Visual Statistical Inference, with Application to Linear Regression Diagnostics

2.1 Abstract

2.2 Introduction

2.2.1 Model Diagnostics

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. In these diagnostics, though numeric summaries are mostly available and some are even endorsed by finite or asymptotic properties, graphic 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.

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

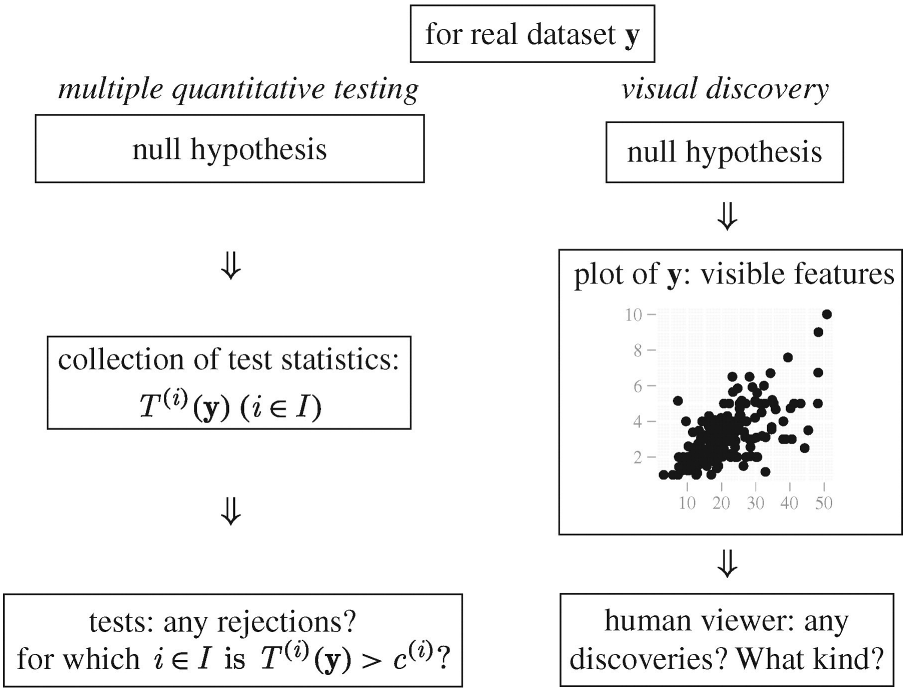

Figure 2.1: *Parallelism between multiple quantitative testing and visual discovery (Buja et al., 2009).*

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).

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 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., 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 2.2.

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).

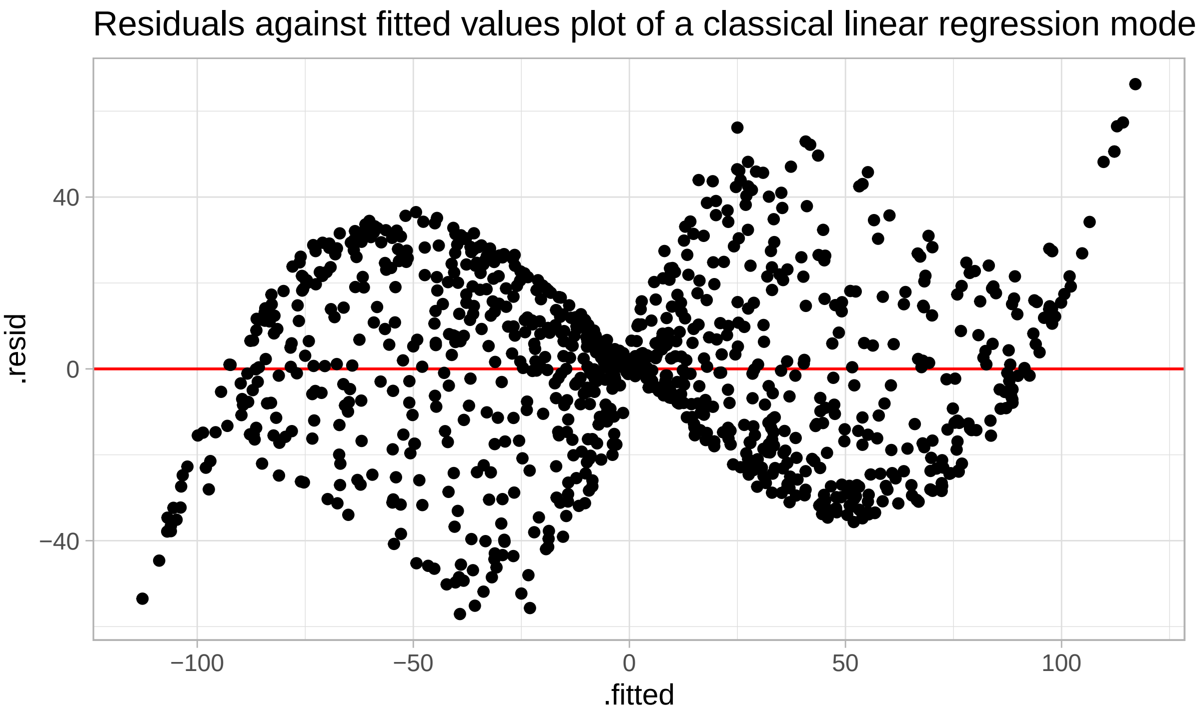

Figure 2.2: *test*

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

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). Publisher: Royal Society, 4361–4383. (Visited on 01/24/2022).

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). Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1984.10478080, 531–554. (Visited on 01/25/2022).

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

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).

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.

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.