Significance of patterns in data visualizations

Rafael Savvides

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.
Helsinki 29.7.2019

Supervisor

Prof. Aristides Gionis

Advisor

Assoc. Prof. Kai Puolamäki

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Rafael Savvides

Title Significance of patterns in data visualizations

Degree programme Computer, Communication and Information Sciences

Major Acoustics and Audio Technology **Code of major** ELEC3030

Supervisor Prof. Aristides Gionis

Advisor Assoc. Prof. Kai Puolamäki

Date 29.7.2019 **Number of pages** 62+1 **Language** English

Abstract

When a data analyst explores data visually and observes a pattern, how can he or she determine whether the pattern is real or just a random artefact of the data? This thesis addresses the problem of evaluating visual patterns observed during visual data exploration by developing a statistical significance testing framework for visual patterns.

Traditionally, patterns observed during data exploration are not evaluated with statistical testing. The reason is that any hypotheses to be tested about the data must be formulated prior to viewing the data, else there is a risk of false discoveries (Type I errors). A naive solution for combining visual exploration with statistical testing involves pre-specifying all possible hypotheses about observable patterns and then applying a multiple testing correction. However, the sheer number of potential patterns results in an overly strict multiple testing correction, resulting in low statistical power. This means that true patterns in the data may fail to be discovered, i.e., there is a risk of false negatives (Type II errors).

The framework proposed in this thesis is a principled statistical significance testing procedure that controls Type I errors and is not overly conservative. The framework is based on improving statistical power by leveraging the data analyst's knowledge and by utilising multiple testing corrections that are suitable for visual exploration.

An empirical investigation of the framework is performed on real and synthetic tabular data and time series, using different test statistics and null distributions. The investigation shows that the proposed framework allows the significance of visual patterns to be determined during exploratory analysis.

Keywords exploratory data analysis, statistical significance testing, visualisation, visual analytics

Preface

This thesis was carried out at the Department of Computer Science at the University of Helsinki and was financially supported by the Academy of Finland.

I thank my advisor Kai Puolamäki for suggesting an interesting topic and for his guidance throughout the development of the ideas in this thesis. I am also grateful for his helpful comments during the writing process.

I thank my collaborators Andreas Henelius and Emilia Oikarinen for their contributions to the scientific paper based on the thesis and their comments during the writing process of the thesis. I also thank my collaborator Anton Björklund for his comments for the writing of the thesis.

I thank my supervisor Aristides Gionis for supervising the thesis and for helping me finalise this work.

Helsinki, 29.7.2019

Rafael Savvides

Contents

Symbols and abbreviations

Abbreviations

CDA	Confirmatory data analysis
EDA	Exploratory data analysis
MTP	Multiple testing problem
MTC	Multiple testing correction
FWER	Family-wise error rate
FDR	False discovery rate
PCA	Principal component analysis
i.i.d.	Independently and identically distributed

Symbols

$\Pr(X)$	Probability of X
$\Pr(X \mid Y)$	Conditional probability of X given Y
$\mathbb{E}[X]$	Expectation of X
$[n]$	$\{1, \ldots, n\}$
Ω	Sample space
ω	Data sample from Ω
ω_0	Observed dataset
H_0	Null hypothesis
H_0^C	Complete null hypothesis
H_{0i}	Null hypothesis for pattern i
$T_i(\omega)$	Test statistic for pattern i calculated on data sample ω
α	Significance level
$\tilde{p}_i$	MTC adjusted p-value for pattern i
p_i^t	p-value for pattern i at iteration t

List of Figures

List of Tables

1 Introduction

One of the first steps in analysing data is exploratory data analysis (EDA). A data analyst uses tools such as summary statistics (e.g., averages) and visualisations (e.g., histograms and scatterplots) to understand the data. Any resulting insights from EDA inform possible directions of research and create ideas about statistical modelling in later analyses.

An important problem in EDA is assessing discoveries during exploration and preventing misinterpretation or overinterpretation of the data [1]. Although the results of EDA are not directly used for inference, it is inevitable that any exploration will influence the analyst's perception of the data and instil some bias in later analyses. This is especially relevant when using data visualisations in EDA (visual exploration). A fundamental question in visual exploration (and the motivation behind this thesis) is: *Suppose a data analyst visualises some data and observes a pattern. Is that pattern real or is it due to chance?*

Determining whether an observed effect is due to chance is traditionally assessed with statistical significance testing (or confirmatory data analysis, CDA). A hypothesis about the data is formed and then the data are used to reject or not reject the hypothesis. The problem of determining whether a pattern in a data visualisation is due to chance can thus be formulated as a statistical significance test. However, in significance testing a hypothesis is supposed to be formulated *before* looking at the data, else there is a risk of hindsight bias and post-hoc fallacies [19]. Any hypotheses must be formulated strictly before any testing. In contrast, in EDA (and especially in visual exploration) a data analyst, by definition, looks at the data *first* and only then formulates hypotheses. EDA is an iterative process in which hypotheses are formulated constantly. Combining statistical testing with visual exploration is thus non-trivial and several subtleties of statistical testing need to be addressed.

Traditionally EDA is thought of as any method of data analysis that does not involve formal statistical modeling or inference. This simplified notion has been challenged in recent years and the lines between EDA and CDA are being blurred. Yu [2] claims that there are widespread misconceptions about EDA, due to the fact that EDA is a philosophy or mentality, rather than a fixed set of methods [3] and due to the emergence of new methodologies, such as data mining and resampling, that change the nature of EDA. Data mining has even been viewed as an extension to EDA, since the goal is to discover patterns that are present in the data, without any preconceived assumptions or prior hypotheses [4].

In the past, the early and messy work of preliminary data analyses were not addressed in traditional statistical (i.e., CDA) training [5]. It is in these initial stages of data analysis that the analyst refines her hypotheses and aligns her mental model with the reality of the data. Ignoring the preliminary work can negatively impact the course of the data analysis process. This discrepancy has been called the hypothesis testing myth [6].

Already in 1972, Tukey [7], who coined the term EDA, suggested that data analysis is a continuum between EDA and CDA. He called the area in between them *rough confirmatory analysis*. In EDA, multiple hypotheses are generated by

looking at the data with or without any theoretical reasoning. The objective is a "rich description of the data". EDA was likened to "detective work": the analyst explores the data in multiple ways to collect evidence that suggests the presence of patterns or a narrative. In CDA, specific hypotheses are tested using strict, state-of-the-art statistical methodologies. If EDA is detective work and collecting evidence, then CDA is the trial, in which findings and arguments are evaluated and assumptions are challenged. Rough CDA answers the question "With what accuracy are the appearances already found to be believed?" by assessing the generated hypotheses in a crude manner with probabilistic methods. According to this simple categorization, the problem of assessing visual patterns during data exploration is thus situated in rough CDA.

This thesis presents a statistical framework that solves the problem posed by the question: "Is the pattern I see in a data visualisation real or due to chance?". The problem is formulated as a statistical significance test for visual patterns and a multiple testing procedure is developed that is suitable for visual exploration.

A scientific paper based on the thesis has been accepted for publication in the conference proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2019 [8]. The idea in the paper was initiated by Associate Professor Kai Puolamäki. The development of the idea and the writing process was carried out together with all the authors, while the author of this thesis carried out all of the experiments.

Organisation The remainder of this thesis is organised as follows:

- Section 2 describes relevant concepts related to confirmatory data analysis (CDA), such as statistical testing, p-values, resampling methods, multiple testing, multiple testing corrections.
- Section 3 describes relevant concepts related to exploratory data analysis (EDA), such as data visualisation, dimensionality reduction, visual exploration, and reviews previous work related to statistically testing visual patterns.
- Section 4 describes the developed framework and the proposed solutions.
- Section 5 contains several use-cases on tabular and time series data that empirically evaluate the framework.
- Section 6 contains conclusions and future work.

2 Confirmatory data analysis

Confirmatory data analysis (CDA) is a general term that refers to methods of statistical inference in the context of data analysis. CDA evaluates the evidence that was gathered during exploratory data analysis (EDA).

Inference involves drawing conclusions on the basis of evidence and reasoning. In formal logic, the evidence is a set of premises, while in statistical inference the evidence is a sample from a population. The aim of statistical inference is to infer properties of the population.

Statistical inference can be approached from a frequentist or a Bayesian perspective. This thesis uses frequentist methods, in which the main tool is *statistical hypothesis testing* (also referred to as significance testing or statistical testing).

2.1 Statistical hypothesis testing

The idea behind statistical testing is that a hypothesis is formulated about the data and then tested to determine whether it is likely to have occurred by chance. The hypothesis concerns a property of the data generating mechanism (i.e., the population from which the data was sampled). The truth or falsehood of a hypothesis is determined by gathering "evidence" against it.

Modern statistical hypothesis testing is a mixture of two approaches developed by Fisher and Neyman-Pearson. There are epistemological differences between them and it has not been established which method is valid. To complicate things further, modern hypothesis testing is a mixture of the two, a result of simplifications over the years. [9] Nevertheless, statistical testing forms the fundamental basis of experimental science. It is typically used in multiple fields, such as physics, biology, psychology, economics, and sociology, to assess the significance of experimental results. Statistical testing is especially important in fields such as psychology, in which it is not possible to compare experimental data with a predicted value from a model.

A statistical test is formulated using two hypotheses: the null and the alternative [10]. The null hypothesis H_0 expresses the case in which an effect is absent, while the alternative H_1 expresses the opposite case. We gather evidence in order to falsify H_0 (similarly to a trial, in which one is innocent unless proven guilty). If the evidence shows that H_0 (an effect is absent) is false, then H_0 is rejected and H_1 is accepted (an effect is present). Specifically, a hypothesis H_1 is assumed valid if its "counter-hypothesis" H_0 is improbable. The null hypothesis is "never proved or established, but is possibly disproved, in the course of experimentation" [11].

For example, suppose that the problem is to determine whether being married affects income level. The problem is formulated as a statistical test as follows:

$$H_0 : \text{income level is independent of marital status}$$

$$H_1 : \text{income level is dependent on marital status}$$

Evidence is then gathered in the form of data, e.g., through a survey on a sample of 100 people, shown in Table 1. If H_0 is true and marital status is independent of

income level then it is expected that married and not-married persons would have the same income distribution, as is the case in Table 1a. If marital status is perfectly correlated with income level then it would be expected to see the results of Table 1b.

	> 50k	< 50k	Total
Married	49	21	70
Not-Married	21	9	30
Total	70	30	100

(a) Independent

	> 50k	< 50k	Total
Married	70	0	70
Not-Married	0	30	30
Total	70	30	100

(b) Dependent

Table 1: Statistically testing whether marital status is correlated with income level.

The evidence that disproves H_0 is traditionally gathered in the form of a *test statistic*. A suitable test statistic is defined so that it distinguishes the null from the alternative hypothesis. Suppose that the dependent data are observed from the above example. If the observed counts in each cell are significantly different from the expected ones (as in Table 1a), then H_0 is rejected and marital status is determined to be correlated with income level. One possible test statistic for testing this hypothesis is the following:

$$T = \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{1}$$

where *observed* is the value of a cell (not including totals) in Table 1b, and *expected* is the value of the same cell in Table 1a.

The observed test statistic t_{obs} is determined to be "significant" by considering the random variable T's distribution under the assumption that H_0 is true (Figure 1a).

(a) Analytical (b) Empirical

Figure 1: Probability distribution of the test statistic T under the null hypothesis (chi squared distribution). The observed t_{obs} is denoted with a vertical dotted line. The black region is equal to the (one-sided) p-value.

If the probability p that $T \geq t_{obs}$ is small when H_0 is true, then t_{obs} is unlikely to have occurred by chance. The probability p quantifies the evidence against the null and is called the *p-value*. The p-value is defined as the probability of observing a value of the test statistic as extreme or more extreme as the one that was observed, under the assumption that the null hypothesis is true:

$$p = \Pr(T \text{ more extreme than } t_{obs} \mid H_0 \text{ is true}) \tag{2}$$

In the previous example, a more extreme value for T means a value larger than the observed value t_{obs} ($T \geq t_{obs}$). This is called a one-sided (or one-tailed) test. A test is two-sided (or two-tailed) when the extreme refers to the distance from a reference value, such as the mean of the null distribution or zero ($|T| \geq |t_{obs}|$). The threshold for how "small" the p-value must be to claim that the observed effect is unlikely to have occurred by chance, is called the *significance level* α. Typical thresholds are 0.01, 0.05 and 0.10. If $p \leq \alpha$, then t_{obs} is significant at the α level.

The above is an example of the chi-squared test of independence. The test statistic T in Equation (1) is a random variable that follows the chi-squared distribution when the variables in Table 1b are independent (i.e., under the null hypothesis of independence) [12]. The distribution of the test statistic under H_0 is called the *null distribution* or the null model.

The significance level α expresses the threshold for what is considered unlikely. It can thus be interpreted as the probability of falsely rejecting H_0. When H_0 is rejected, the validity of H_1 is proven, which typically means that a discovery is made. If H_0 is falsely rejected, then a false discovery is made. Finding something that is not there is called a *false positive* or a *Type I error*. Accordingly, a *false negative* or a *Type II error* occurs when H_0 is false and is *not* rejected. The rate of false negatives (Type II error rate) is usually denoted by β. False negatives are related to the *statistical power* of a test through $\texttt{power} = 1 - \beta$. The power of a test expresses the ability of the test to make discoveries and is defined as the probability that the test rejects H_0 when H_1 is true, i.e., $\texttt{power} = \Pr(\text{reject } H_0 \mid H_1 \text{ is true})$.

2.2 Resampling

It is not always possible to have an analytical formulation of the null distribution, such as the chi-squared distribution above. Nevertheless, it may still be possible to approximate the process of drawing data samples from the null distribution and compute the test statistic on the samples. In this case, the samples of the test statistic are drawn from an empirical distribution (Figure 1b) and an *empirical p-value* [13] is computed:

$$p = \frac{k+1}{n+1} \tag{3}$$

where n is the number of drawn samples, and $k = |\{j \in [n] \mid t_j \geq t_{obs}\}|$ is the number of drawn samples j in which the test statistic t_j is more extreme that the observed value t_{obs}.

The empirical p-value is the probability, when H_0 is true, of sampling a test statistic at least as extreme as the observed value. The added ones in Equation (3) protect against underestimating small p-values. The numerator ensures that the smallest p-value is $1/(n+1)$ (and not zero), which means that in order to obtain a small p-value, a large number of samples is needed from the empirical distribution. The denominator can be interpreted as including the observed test statistic t_{obs}, which is also drawn from the null distribution [13].

Computing a p-value from an empirical distribution is a *resampling* method. In general, resampling methods refer to methods in which the observed data are used

repeatedly to provide inferences. These include the *bootstrap*, *permutation tests*, and *Monte Carlo methods*.

Resampling methods are based on empirical distributions. Consider a sample X of i.i.d. observations $\{x_i\}_{i=1}^n$ that follow a distribution F. The *empirical distribution* $\hat{F}$ places probability mass $1/(n+1)$ on each data point and is defined as:

$$\hat{F}(x) = \frac{\#\text{ of } x_i \leq x}{n+1} = \frac{\sum_{i=1}^n I(x_i \leq x)}{n+1}$$

where $I(\square)$ is the indicator function which equals unity if $\square$ is true and zero otherwise.

Bootstrap The bootstrap is a type of statistical methods that are based on the fact that the sampling distribution of a test statistic can be estimated by sampling with replacement from the empirical data distribution [14]. In other words, m bootstrap datasets $X_j,\ j = [m]$ are generated from a dataset $X = \{x_i\}_{i=1}^n$, by sampling n elements of X uniformly at random with replacement m times. During this process, the bootstrap samples X_j are in effect random variables with a distribution equal to the empirical distribution $\hat{F}$ of X. If the test statistic is computed on each bootstrap sample, then the result is an estimate of the *sampling* distribution of the test statistic. Resampling methods do not estimate the underlying population distribution of the test statistic. The observed dataset is a sample from a population and, put simply, the sample is assumed to *be* the population.

Permutation test Sampling from X *without* replacement is equivalent to permuting the data. In a *permutation test*, the null distribution of a test statistic is obtained by calculating all possible values of the test statistic under shuffling of the data labels in the observed data. The permutation test computes an empirical probability (similarly to introductory probability exercises with balls drawn from an urn), rather than a probability calculated from an analytical distribution under certain assumptions (e.g., normality).

As an example, suppose the null hypothesis is that the datasets A and B originate from the same distribution. One possible test statistic for testing this hypothesis is the difference between their means $T = \bar{x}_A - \bar{x}_B$. In a permutation test, the distribution of T under the null hypothesis is obtained by calculating the difference of means in all possible ways of splitting $A \cup B$ into two datasets.

Similarly, a permutation test for the independence of marriage status and income level in Table 1 can be formulated by considering all possible ways that a table can occur with the same group totals, i.e., the same marginal distributions. However, in this case the number of possible permutations is very large ($100! = 9.33 \times 10^{157}$), which makes the procedure of the previous example practically unfeasible. Although for this particular test the p-value can be efficiently computed using the hypergeometric distribution, this is not generally the case for more complex distributions.

As an alternative to computing all possible permutations, the permutation distribution can be approximated by randomly sampling a very large number of permutations, to perform an *approximate permutation test* (also known as Monte Carlo permutation test or random permutation test) [15].

2.3 Criticism of statistical testing

Statistical testing and p-values are often misunderstood and misused. Such a misuse of p-values has been called "data dredging" or "p-hacking" and has received widespread criticism [16], even resulting in some journals banning the use of p-values [17].

The most striking consequence of the misuse of p-values is the replication crisis in science [18]. Many scientific studies are difficult to replicate or reproduce, due to researchers analysing their data in multiple ways to obtain significant results, when there is no real underlying effect. If many hypotheses are tested simultaneously, there is a higher probability of false positives. This is referred to as the multiple hypotheses or multiple comparisons or multiple testing problem (see Section 2.4).

A related problem is that given enough data, any null hypothesis can be proven false. As an example, suppose that a test is performed to determine whether there is a difference between the means of two groups A and B. The null hypothesis is that $\mu_A = \mu_B$ and the alternative is $\mu_A \neq \mu_B$. What is being tested is essentially the sign of the difference $\mu_A - \mu_B$. If the alternative hypothesis is stated as $x_A > x_B$ or $x_A < x_B$, then with a large enough sample a significant result occurs 50% of the time (i.e., either a positive or a negative sign). If the direction of the test is not stated, then a significant result occurs 100% of the time (with a large enough sample). This is because in the real world (as opposed to experimental settings), the null hypothesis of zero difference is always wrong: in nature there always exists some difference between two groups (Meehl's conjecture) [19]. This partly explains the phrase: *statistical significance does not imply practical, real-world significance*. Often a statistically significant result is a very small effect. This small effect may indeed be significant in terms of the statistical test, but it may not provide useful knowledge in terms of decision making.

Another problem with statistical test is the arbitrary choice of significance level α and its interpretation. When α was introduced, it was stated that the significance level should be set before any data collection [20] and according to the specific circumstances of the experiment and the field of study [21]. Nevertheless, a "convenient" threshold was suggested by Fisher [21] as $\alpha = 0.05$, which then persisted due to convention. As a result, many publications, e.g., in experimental psychology, report p-values rounded to often used significance levels (0.05, 0.10 and 0.01) [22]. To complicate things further, the interpretation of the significance level and the whole hypothesis testing procedure is different in the Fisher and the Neyman-Pearson versions [19].

The results from a statistical test are often not easily interpretable and carry the danger of post-hoc explanations. Post-hoc explanations may occur, for instance, with the rejection of a null hypothesis whose alternative hypothesis is stated after the data have been gathered (Feynman's conjecture) [19]. Since statistical testing can be interpreted as fitting a model on the data, the problem of post-hoc explanations can be viewed as an overfitting problem. If a model is fit on the observed data, then the predictive power of the model cannot be evaluated using the same data. Nevertheless, e.g., in the field of psychology, fitted models are rarely validated on new data [19].

Misconceptions about p-values are also very common [23], mainly due to the

subtleties in their interpretation. For example, the Type I error rate was defined in a previous section as equal to the significance threshold (e.g., $\alpha = 0.05$). However, this is true only when the null hypothesis is true; in other cases the error rate is between 0 and α. In contrast to common misconceptions, a p-value does not measure the probability that a hypothesis is true or false, or the probability that the observed data were produced by random chance. Rather, a p-value expresses how compatible a dataset is with a hypothesis. [16]

Despite the criticism and misconceptions surrounding p-values, they remain a valid tool for determining whether a statistical model is incompatible with the observed data [16]. Hence, p-values are used in this thesis, since the proposed framework determines whether the assumptions of a data analyst are compatible with an observed visual pattern.

Many problems with the use of p-values (e.g., p-hacking) originate from the multiple testing problem. These problems can be mostly alleviated with suitable corrections, which are presented in the next section.

2.4 Multiple testing

When a hypothesis is tested, there is a chance of a false discovery (Type I error). If multiple hypotheses are tested simultaneously, then the probability of obtaining at least one false discovery is higher.

As an illustrative example, consider a null hypothesis H_0 that is tested against an alternative at a significance level of 5%. Then there is a 5% chance of incorrectly rejecting H_0 when it in fact true, i.e., $\Pr(\text{reject } H_0 | H_0 \text{ is true}) = \Pr(\text{false discovery}) = 5\%$. If 20 independent true null hypotheses are tested simultaneously, then the probability of at least one erroneous rejection is:

$$\Pr(\text{at least one false discovery}) = 1 - \Pr(\text{no false discoveries}) = 1 - 0.95^{20} = 64\%$$

This is referred to as the *multiple hypotheses or multiple comparisons or multiple testing problem* (MTP). In general, the MTP occurs when multiple inferences are considered simultaneously. Suppose that there are m null hypotheses $H_i, i = [m]$ and that for each hypothesis a statistical test is performed. A hypothesis is rejected if the computed p-value is below some significance threshold, else it is not rejected. Each hypothesis is either true or false and the performed test is either significant or non-significant. This results in the four scenarios of Table 2 that define four random variables: false positive (FP), true positive (TP), true negative (TN) and false negative (FN) counts.

	Significant (H_0 rejected)	Non-significant (H_0 not rejected)
True H_0	FP	TN
Non-true H_0	TP	FN

Table 2: Possible outcomes of a statistical test

2.4.1 Multiple testing corrections

In order to control the rate of false discoveries and to avoid erroneous inferences, statistical techniques have been developed. These techniques are called *multiple testing corrections* (MTC) and are also known as correcting procedures, adjustment procedures or simultaneous testing procedures.

MTCs recompute the p-value of each considered hypothesis to account for the total number of hypotheses being tested. The resulting p-values are called *adjusted p-values* (denoted by $\tilde{p}$) and the original p-values are called *unadjusted* or *raw p-values* (denoted by p). MTCs can control several error rates; here the focus is on the family-wise error rate (FWER) and the false discovery rate (FDR). For a review of other possible error rate measures, see [24].

The family-wise error rate (FWER) is defined as the probability of at least one false positive (Type I error) [24]:

$$\text{FWER} = \Pr(\text{FP} > 0) \tag{4}$$

Controlling FWER is ideal for cases in which false positives are prioritised over false negatives, i.e., it is preferable to not discover anything than to discover something false. As the number of hypotheses increases, the probability of at least one false positive also increases. If FWER is controlled, the procedure loses statistical power and it is more difficult to make *any* discoveries.

An alternative error measure is the false discovery rate (FDR) [25] which is defined as the expected proportion of false positives:

$$\text{FDR} = \mathbb{E}[\frac{\text{FP}}{\text{FP} + \text{TP}}] \tag{5}$$

When $R = \text{FP} + \text{TP} = 0$, FDR is defined to be zero. An alternative definition that includes the case $R = 0$ is $\text{FDR} = \Pr(R > 0)\mathbb{E}[\frac{\text{FP}}{\text{FP+TP}} \mid R > 0]$ [25]. Controlling FDR is ideal for cases in which statistical power is important, i.e., it is preferable to discover something false than to possibly not discover something.

In most scenarios there is a trade-off between Type I and Type II error rates. If multiple hypotheses are tested, then there is a risk of false discoveries (Type I error). Type I errors are controlled by applying a MTC. If the correction is too strict, then there is a risk of not discovering true patterns present in the data (Type II error). Balancing this trade-off depends on the application. For example, in medicine the priority may be to avoid false negatives: it may be preferable to find that the patient has a disease (even if they do not) and to assign them more tests, instead of not discovering a true underlying disease.

A MTC is said to have *weak* control over an error rate, if it controls the error rate only when all of the considered null hypotheses are true, i.e., when the *complete* null hypothesis $H_0^C = \bigcap_{i \in [m]} H_i$ is true. Conversely, *strong* control is provided when the error rate is controlled regardless of which subset M of the considered null hypotheses are true, i.e., when a *partial* null hypothesis $\bigcap_{i \in M \subset [m]} H_i$ is true. Although strong control is generally preferable, proving it is very difficult without knowledge or assumptions about the joint distribution of p-values for the considered hypotheses (nulls and their alternatives) [26].

The joint distribution of the random p-values P_i (and the random test statistics T_i) satisfies the *subset pivotality* property when it is independent of which hypotheses are actually true, i.e., when it is identical under the complete null hypothesis and under any subset of true null hypotheses M [26]. A convenient consequence of this property is that sampling from the complete null distribution is identical to sampling from the partial null distribution (i.e., the subset M of null hypotheses). This consequence simplifies computations for the minP and maxT procedures, described below.

A MTC can be a single-step or a stepwise (also known as levelwise or sequential) procedure. A *single-step* MTC adjusts all tests equivalently, regardless of the sequence of the hypotheses or the order of the p-values. A *stepwise* MTC allows for a different adjustment to each hypothesis depending on the ordering of the hypotheses (e.g., based on their p-values) [26]. A *step-down* MTC sorts the p-values from lowest to highest and sequentially tests hypotheses until all hypotheses are rejected or a hypothesis is not rejected. If a hypothesis is not rejected, then the remaining hypotheses are also not rejected. A *step-up* MTC sorts the p-values from highest to lowest and sequentially tests hypotheses until a hypothesis is rejected. Once a step-up MTC rejects a hypothesis, the remaining hypotheses are also rejected.

Stepwise MTCs depend on whether the considered hypotheses satisfy the *free combinations* or *restricted combinations* conditions. If the simultaneous truth of any subset of considered hypotheses and falsehood of the remaining ones is a plausible event, then the considered hypotheses satisfy the free combinations condition, else they satisfy the restricted combinations condition [27]. In other words, in the restricted combinations condition, certain combinations of true hypotheses necessarily imply truth or falsehood of other hypotheses [26].

Stepwise MTCs are instances of *closed testing procedures* [28]. For a set of hypotheses $H_i,\ i \in [m]$, the *closed testing principle* guarantees strong control of FWER at level α for a hypothesis H_i, if all subsets of hypotheses that contain H_i are rejected at level α with a suitable local test. A local test rejects an *intersection hypothesis* $H_I = \bigcap_{j \in I} H_j,\ i \in I \subset [m]$ (i.e., a subset of hypotheses that contain H_i), if $p_I \leq \alpha$. For example, the intersection hypothesis H_{123} is rejected, if any of the MTC adjusted p-values $\tilde{p}_1, \tilde{p}_2, \tilde{p}_3$ is significant.

Next, a description follows of several MTCs used in this thesis:

Bonferroni correction The most well-known MTC is the single-step Bonferroni correction, which provides strong control over FWER [24]. The procedure simply multiplies the unadjusted p-values by the number m of the considered null hypotheses (setting them equal to one, if they become larger than one) to obtain the adjusted p-values:

$$\tilde{p}_i = \min(1, mp_i) \tag{6}$$

Applying the Bonferroni correction is equivalent to setting the significance threshold of the unadjusted p-values to α/m (instead of α). The Bonferroni correction is general (i.e., does not require strict assumptions) but it is overly conservative and hence lacks statistical power.

Holm-Bonferroni correction An improvement to the Bonferroni correction proposed by Holm [27] is called the Holm-Bonferroni correction (or Holm's sequential Bonferroni correction). The procedure is step-down (i.e., p-values are sorted from lowest to highest) and has an improved power over the simple Bonferroni for any scenario. It provides strong control over FWER for both free and restricted combinations.

The Holm-Bonferroni correction is applied by sorting the raw p-values from lowest to highest such that $p_{r_1} \leq \ldots \leq p_{r_m}$ (where r_i is the rank of p_i) and then applying the following steps:

$$\begin{aligned} \tilde{p}_{r_1} &= m p_{r_1} \\ \tilde{p}_{r_2} &= \max(\tilde{p}_{r_1}, (m-1)p_{r_2}) \\ &\vdots \\ \tilde{p}_{r_m} &= \max(\tilde{p}_{r_{m-1}}, p_{r_m}) \end{aligned}$$

where $p_{r_1} \leq \ldots \leq p_{r_m}$ are the ordered raw p-values, r_i is the rank of p_i, and $\tilde{p}_i$ are the adjusted p-values.

If any p-value is larger than one, then it is set equal to one. The procedure is then summarised for the general case as:

$$\tilde{p}_{r_i} = \max\left[\tilde{p}_{r_{i-1}}, \min((m - r_i + 1)p_{r_i}, 1)\right] \tag{7}$$

The Holm-Bonferroni correction is equivalent to sorting the unadjusted p-values from the lowest to the highest and setting for each one the significance threshold to $\alpha/m,\ \alpha/(m-1),\ \ldots,\ \alpha/1$. In other words, once H_1 is rejected with a Bonferroni threshold α/m, then H_1 is false, which leaves $m-1$ hypotheses that are possibly true, hence the Bonferroni threshold for H_2 is $\alpha/(m-1)$, and so on.

The Holm-Bonferroni correction, under the free combinations condition, is a closed testing procedure. A hypothesis H_i is rejected, if every intersection hypothesis $H_I = \cap_{j \in I} H_j,\ i \in I \subset [m]$ is rejected with the Bonferroni correction [28]. For instance, suppose 3 hypotheses H_1, H_2, H_3 are considered. Then H_1 is rejected if all intersection hypotheses H_{123}, H_{12}, H_{13} and H_1 are rejected at level α with a Bonferroni correction. In other words, H_{123} is rejected if any of $p_1, p_2, p_3 \leq \alpha/3$, H_{12} (and H_{13}) are rejected if any of $p_1, p_2 \leq \alpha/2$ (any of $p_1, p_3 \leq \alpha/2$), and H_1 is rejected if $p_1 \leq \alpha$.

Single-step minP and maxT The statistical power of a multiple testing correction can be improved by directly incorporating the exact correlation structure of the test statistics. The correlations present in the data are naturally taken into account with resampling procedures, such as the *minP* and *maxT* procedures [26].

The minP procedure is based on the following remarks about the distribution of p-values. Consider m null hypotheses H_i and their m p-values P_i (capital letter denotes that it is a random variable instead of an observed value). Each individual p-value P_i is uniformly distributed in $[0, 1]$ under H_i and under assumptions of normality, independence and homoscedasticity. The probability of an *individual* test

being significant at level α is then α. The probability of declaring *at least one* test significant is equal to the probability of the smallest p-value being significant:

$$\Pr(\text{at least one test is significant} \mid H_0^C) = \Pr(\min_i P_i \leq \alpha \mid H_0^C) \tag{8}$$

If the probability of the smallest p-value being significant is controlled, then the probability of false discoveries is also controlled. The single-step minP correction is defined as follows:

$$\tilde{p}_i = \Pr(\min_{1 \leq j \leq m} P_j \leq p_i \mid H_0^C) \tag{9}$$

where p_i is the observed p-value for the hypothesis H_i, P_j is the random variable for the p-value of the hypothesis H_j, $H_0^C = \bigcap_{i \in [m]} H_i$ is the complete null hypothesis, and $\tilde{p}_i$ is the minP adjusted p-value.

Computing the probability of Equation (9) requires knowledge of the distribution of the minimum p-value, which in turn requires knowledge about the joint distribution of the p-values P_j under the complete null hypothesis H_0^C. If there is dependence between P_j, then this distribution is usually intractable [26]. In these cases, resampling methods are used to approximate the empirical joint distribution of P_j and the distribution of the minimum p-value. In practice, data samples are drawn from the null distribution and approximate p-values p_i^* are computed on the resampled data. The resulting joint distribution of p_i^* is approximately the same as the theoretical distribution of P_j. The empirical single-step minP adjusted p-values of Equation (9) are then computed as the proportion of resampled data in which the minimum p-value is smaller than the observed p_i:

$$\tilde{p}_i = \frac{k_i + 1}{n + 1} \tag{10}$$

where n is the number of drawn samples, and $k_i = |\{j \in [n] \mid \min_j p_j^* \leq p_i\}|$ is the number of drawn samples j in which the minimum p-value is smaller that the observed one p_i.

The resampling minP procedure provides strong control over FWER, if the subset pivotality property is satisfied. However, when the p-values are estimated, e.g., with resampling methods, then FWER is only *approximately* controlled. The accuracy of the approximation depends on the accuracy of the resampling-based estimates. [26]

The maxT procedure is equivalent to the minP, since the maximum test statistic corresponds to the minimum p-value. Nevertheless, in practice they have certain differences. For instance, in maxT the test statistics are assumed to have the same magnitude (which may cause problems, if both one-sided and two-sided tests are considered), while in minP the test statistics are scaled to the interval $(0, 1)$ [26].

Step-down minP and maxT An improvement over the single-step minP and maxT procedures is obtained with their step-down versions. The *step-down minP*

procedure is defined (similarly to the Holm-Bonferroni) as:

$$\begin{aligned}
\tilde{p}_{r_1} &= \Pr(\min_{j \in \{r_1,\ldots,r_m\}} P_j \leq p_{r_1} \mid H_0^C) \\
\tilde{p}_{r_2} &= \max\left[\tilde{p}_{r_1}, \Pr(\min_{j \in \{r_2,\ldots,r_m\}} P_j \leq p_{r_2} \mid H_0^C)\right] \\
&\vdots \\
\tilde{p}_{r_m} &= \max\left[\tilde{p}_{r_{m-1}}, \Pr(P_{r_m} \leq p_{r_m} \mid H_0^C)\right]
\end{aligned}$$

where $p_{r_1} \leq \ldots \leq p_{r_m}$ are the ordered raw p-values, r_i is the rank of p_i, and P_j is the random variable for p_j, and $\tilde{p}_i$ are the minP adjusted p-values.

Step-down adjustment makes the adjusted p-values uniformly smaller, while retaining the same FWER control. Instead of adjusting all p-values according to the minimum p-value distribution, only the minimum p-value is adjusted using this distribution. The remaining p-values are then successively adjusted according to smaller sets of p-values, resulting in a less strict correction and an increased statistical power. The step-down minP and maxT procedures strongly control FWER when the considered hypotheses satisfy the free combinations condition. [26]

Similarly to the Holm-Bonferroni procedure, the step-down minP is, under the free combinations condition, a *closed test procedure*, in which all intersection hypotheses $H_I = \cap_{j \in I} H_j,\ i \in I \subset [m]$ for a hypothesis H_i are tested with the single-step minP procedure.

As in the single-step version, the theoretical joint distribution of P_j is usually not known and is approximated with resampling to obtain the approximate p-values p_i^*. The approximate p-values p_i^* are then sorted from lowest to highest such that $p_{r_1}^* \leq \ldots \leq p_{r_m}^*$ (where r_i is the rank of p_i) and the *step-down minP adjusted p-values* are obtained with the following algorithm [26]:

1. The successive minima are defined:

$$\begin{aligned}
q_m^* &= p_{r_m}^* \\
q_{m-1}^* &= \min(q_m^*, p_{r_{m-1}}^*) \\
&\vdots \\
q_1^* &= \min(q_2^*, p_{r_1}^*)
\end{aligned}$$

2. The adjusted p-values are computed:

$$\tilde{p}_i = \frac{k_i + 1}{n + 1}$$

where n is the number of drawn samples, and $k_i = |\{j \in [n] \mid q_j^* \leq p_i\}|$ is the number of drawn samples j in which the successive minimum of the sorted p-values q_j^* is smaller that the observed p_i.

3. Monotonicity is enforced on the adjusted p-values through successive maximisation:

$$\begin{aligned}\tilde{p}_1 &= \tilde{p}_1 \\ \tilde{p}_2 &= \max(\tilde{p}_1, \tilde{p}_2) \\ &\vdots \\ \tilde{p}_m &= \min(\tilde{p}_{m-1}, \tilde{p}_m)\end{aligned}$$

The monotonicity is enforced to ensure that the adjusted p-values have the same ordering as the raw p-values. If the adjusted p-values were ordered differently than the raw p-values, then given $p_1 < p_2$, it would be possible for $\tilde{p}_1 > \tilde{p}_2$, i.e., it would be possible for the adjustment to reduce the evidence that p_2 has against the null hypothesis.

In closed testing procedures, if the number of hypotheses m is very large, then testing every single intersection hypothesis H_I becomes computationally infeasible, since there are $O(2^m)$ intersection hypotheses. The minP (and maxT) procedure, however, eases the computational burden to $O(m)$ in two ways. First, since each H_I is tested with the minP test, only the minimum p-value is compared with the significance level, instead of all $p_i,\ i \in I,\ \forall I$. Second, if the subset pivotality property is assumed, then the null distribution can be sampled from the complete null hypothesis $H_0^C = \bigcap_{i \in [m]} H_i$ once, instead of drawing samples from every intersection hypothesis $H_I = \bigcap_{i \in I \subset [m]} H_i$ separately.

3 Exploratory data analysis

Exploratory data analysis (EDA) refers to methods that familiarise the user with the data and provide a general picture of possible patterns, before detailed analyses. During EDA, the data are investigated with summary statistics and visualisations with the goal of understanding the data. Statistical modelling and inference are typically not performed in high detail but merely act as aids to the exploration or illustration of possible structure in the data.

EDA was first promoted by Tukey as an alternative to CDA [29]. EDA was likened to "detective work". The analyst explores the data in multiple ways to collect evidence that suggests the presence of patterns or a narrative. EDA widened the scope of data analysis and statistics from hypothesis testing to hypothesis generation, i.e., "to look for patterns in data beyond the expected" [30]. The preface of Tukey's seminal book [29] begins with the sentence: "It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it." Traditionally, there is a divide between EDA and CDA. However, the distinction between them has become increasing blurred, as mentioned in the Introduction.

The tools of EDA help the analyst investigate the data to detect and describe patterns and relations. These tools have changed since Tukey first coined EDA, due to computing advancements and the ease with which tasks such as data visualisation and regression analysis are now performed. Nevertheless, the overall idea is the same: the data are explored through summaries, transformations and visualisations that reveal patterns (e.g., trends, outliers) in the data from as many perspectives as possible. EDA helps the analyst to generate hypotheses, which in the traditional workflow are then validated using CDA. This thesis is concerned with a major part of EDA: data visualisation and visual exploration.

3.1 Data visualisation

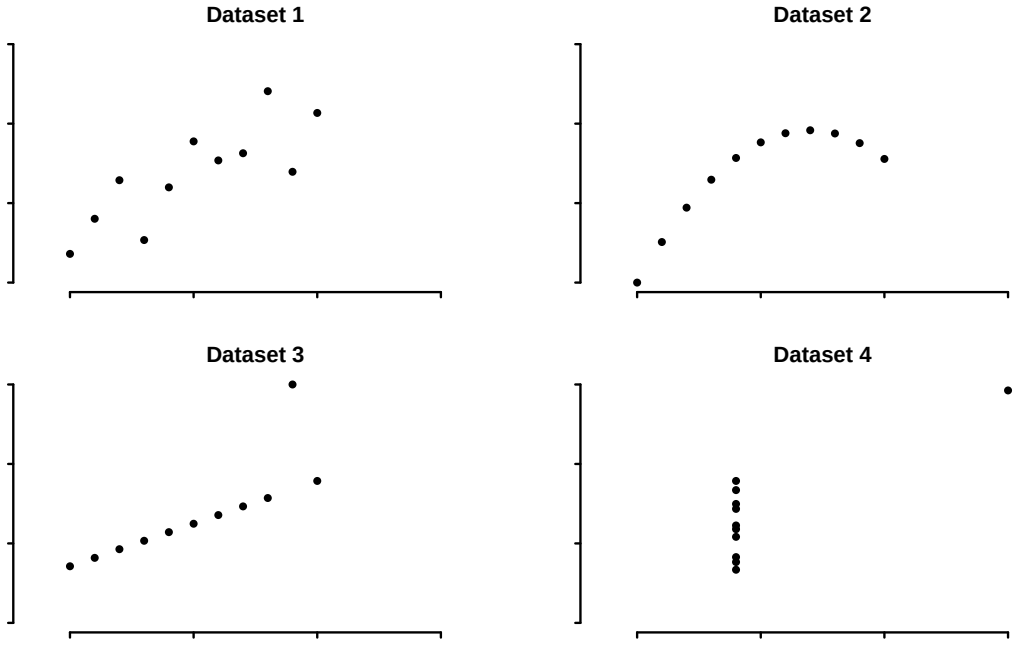
Data visualisation refers to methods that communicate data through visual objects (e.g., points, lines). The goal of data visualisation is to enhance the analyst's understanding of the data by exploiting their visual perception and pattern finding skills. Data visualisation combines the analyst's cognitive reasoning processes with perceptual ones, enabling them to understand and analyse large amounts of information more efficiently [31].

Visualisation is an essential component of EDA and the data analytic process. Its importance is demonstrated with the well-known example of Anscombe's quartet [32]. The four datasets in Table 2a have (nearly) the same summary statistics (mean, standard deviation). However, when displayed in a visualisation they are very dissimilar (Figure 2b). The differences between the four datasets and the patterns present in each one are immediately obvious in a visualisation. Although the same information is presented in a table of numbers, it is not readily available to be discovered. Datasets similar to Anscombe's quartet (identical statistics and dissimilar visualisations) can now be generated with automated methods [33, 34].

Anscombe's quartet is a rather straightforward case of discovering unexpected

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

(a)



(b)

Figure 2: Anscombe's quartet

patterns through visualisation. In many cases, choosing the right visualisation that reveals the most interesting aspects of the data is not an easy task. Cleveland's well-known study [35] showed that certain visualisations are better than others in conveying the same information. This raises the issue of the quality of a data visualisation. Tukey stated that "the greatest value of a picture is when it forces us to notice what we never expected to see" [29]. This is an ideal to strive for but difficult to attain in practice. Further complications arise from the fact that the quality of a visualisation depends directly on quality of the data representation [31].

Tufte's seminal book on statistical graphics [36] introduced several principles for designing good graphics and avoiding misleading graphics. Such principles for

good graphics include avoiding redundant visual elements and ensuring that the depicted visual elements have a perceived magnitude that is directly proportional to the numbers they represent. In contrast, bad graphics mislead the viewer with non-informative and non-representative visual elements.

As an illustration of the ease with which a visualisation can mislead the viewer, consider the scatterplots of Figure 3. The same data are presented in both scatterplots; the only difference is the scaling of the axes. Although both visualisations contain the same information, they are perceived differently by a human viewer.

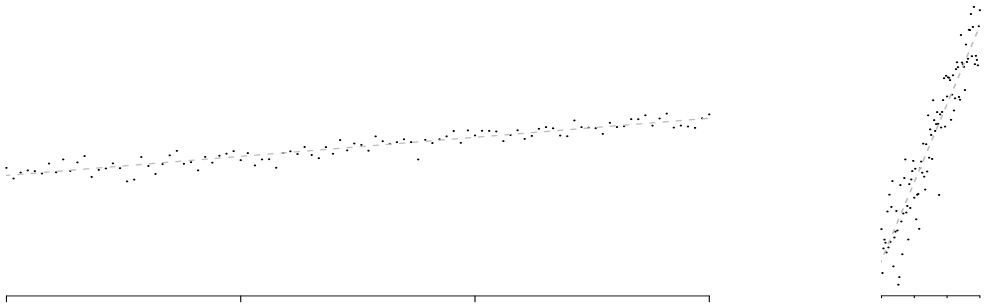


Figure 3: The same data are presented in both scatterplots. Changing the aspect ratio, e.g., by resizing the output window in the statistical software, results in a different perception of the slope.

If the scatterplots of Figure 3 are perceived differently, then it is evident that visualisations can create false impressions and mislead the viewer about the true nature of the data and the underlying phenomenon that the data describe. Although this is a known issue when viewing visualisations made by others, the dangers of misleading graphics also apply to self-made visualisations. Such self-made visualisations include the ones typically created during EDA, in the intermediate steps of the data analysis process, when the analyst seeks to understand the data. Misleading self-made visualisations are especially relevant if one considers modern statistical software, in which a visualisation is often displayed in a separate window. The window can be resized at will, changing its aspect ratio and consequently the perceived form and shape of the enclosed visual elements (e.g., exaggerating the slope of a line in Figure 3). These alterations in the form of the visualisation can lead to false impressions about the depicted data. Consequently, if the visualisation is part of an analytical reasoning process, this can result in misinterpretation of the data and errors of judgement that propagate throughout the data analytic process and eventually into the decision making process.

A well-known example of a poorly designed visualisation that led to catastrophic results is the O-ring data of the failed Challenger space shuttle launch. Tufte claimed that the engineers presented the performance data of the O-rings in a poorly designed visualisation that failed to communicate crucial information to decision makers [37]. This misinformation then contributed to the decision to launch the space shuttle in sub-optimal weather conditions, which resulted in its disintegration and the deaths of the whole crew. Although Tufte's simplified treatment was later criticised [38],

the example still illustrates potential dangers of poorly designed visualisations.

The design choices regarding good visualisations should be motivated by human perception [39], since in visual methods the communication medium is the visual perceptual channel. However, most principles related to data visualisations are based on graphic design (such as Tufte's aforementioned principles), while there is a lack of principles based on experiments in visual perception (such as Cleveland's study [35]) [40]. This kind of knowledge is required to understand what makes certain visualisations effective, and to inform the development and evaluation of new representations for new classes of data [40]. The lack of quantitative assessment methods for visualisations also presents challenges for the experimental evaluation of this thesis.

3.2 Dimensionality reduction for visualisation

Data visualisations often depict high-dimensional data. In order to be able to display the data in a 2D image, their dimensionality must be reduced. During this reduction, information is unavoidably lost resulting in errors, called embedding distortion. Since a partial view of the data is presented, there are more possibilities for judgement errors and false impressions.

The performance of dimensionality reduction techniques can be summarised with various metrics, such as recall (points that are near in the projected space are also near in the original space) and precision (points that are near in the original space are also near in the projected space) [41].

There are a multitude of dimensionality reduction techniques, both linear and non-linear, based on both global and local features of the high-dimensional data manifold. The focus in this thesis is on linear dimensionality reduction methods, which are intuitively thought of as projections.

Principal Component Analysis Principal Component Analysis (PCA) [42] is a linear transformation that attempts to transform the data into a new coordinate system. The new coordinates (called the principal components) are the directions of greatest variance in the data and they are sorted such that projecting to the first coordinate results in the direction of the greatest variance in the data. This new coordinate system is then used to visualise the two directions of greatest variance in the data, which results in a 2D visualisation that retains most of the variance in the data, see e.g., Figure 4.

Note that PCA is based on a global feature of the data (variance), as opposed to a local feature (e.g., distance of each point to its neighbours). Furthermore, there is no interactivity involved or additional (explicit) parameters to tune. Ideally, the user would be able to guide the exploration process according to his or her knowledge using local features of the plot.

Projection Pursuit One way to introduce local features to dimensionality reduction and enable guided exploration is through *projection pursuit* [43]. In projection pursuit, an objective function (called a projection index) is defined for each direction

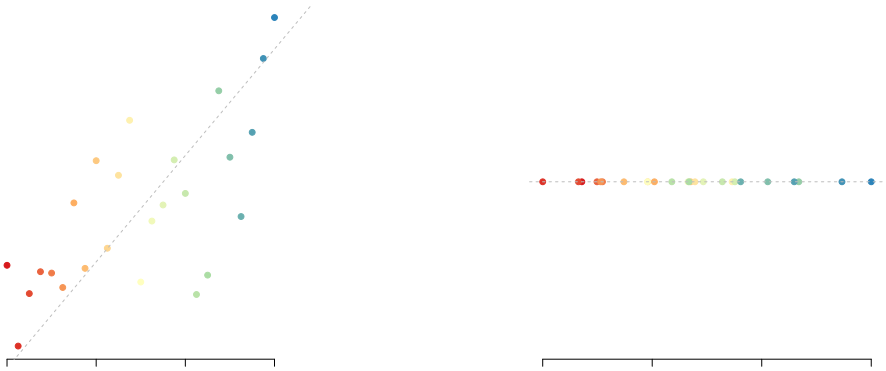


Figure 4: Principal Component Analysis is used in visualisation for projecting data to lower dimensionality (here 2D to 1D) while retaining as much of the variance as possible. The dotted line denotes the direction of maximum variance in the data. The right figure is obtained from the left one by projecting the data on the direction of maximum variance.

in the high-dimensional space. If the index measures the "interestingness" or "usefulness" of a direction, then maximising that index computes the most "interesting" or "useful" projections of the data. In PCA, the index is the variance of the data. When the data are projected onto one "interesting" direction, the process can be repeated to compute interesting directions on the lower dimensional data; this is the *pursuit* aspect.

Note that the interestingness is entirely dependent on the projection index, which is an objective measure. Defining the index can be difficult, especially if one wishes to incorporate prior knowledge or previously discovered patterns. This motivated research on guided exploration and *subjective interestingness* [44, 45, 46]. The idea behind subjective interestingness is that the data analyst can steer the exploration according to his or her interests. For example, in a projection pursuit method based on subjective feedback [47], the data analyst explores the data in a projection pursuit workflow and provides subjective feedback in the form of observed or known patterns in the data. The feedback is assimilated in the projection index, which is then used to compute new interesting projections. This procedure is implemented through the *background distribution*, which is described in Section 4.

3.3 Related work

The framework proposed in this thesis can be used to assess the significance of visual patterns. Relevant fields include visual analytics [48], visual data exploration [49], and graphical (or visual) inference [1, 50, 51, 52].

Visual analytics Visual analytics is a multidisciplinary field that focuses on the analytical reasoning process through interactive visualisations of data [53, 54, 55]. In other words, visual analytics combines analytical reasoning and decision making with interactive data visualisations [31].

Visual analytics helps an analyst to explore data in a fluid manner. In such contexts, it is important to communicate uncertainty in the data [54, 56, 57] and errors stemming from projecting multivariate data to two-dimensional plots [58, 59, 60].

A user study [61] demonstrated that over half of the user insights obtained through visualisations were false, due to the effect of multiple comparisons. This indicates a need for assessing visual discoveries during exploration. Furthermore, it has been noted that methods are needed that represent data quality and reliability throughout the data analytic process [31]. An example of such a method is the use of bootstrap confidence intervals to communicate the sampling variability of the visualised data, such as the confidence intervals around the locally estimated scatterplot smoothing (LOESS) curve in the ggplot2 R package [62]. The sampling variability in various plots has also been proposed to be communicated visually through bootstrap samples that are maximally different based on a visually motivated distance metric [63].

Visual inference Visual inference (or graphical inference) is a recent area of study that combines visual exploration with statistical testing [1, 50]. Visual inference [1] formalises visual patterns as test statistics and the discovery of a pattern as a rejection of a null hypothesis. The statistical test is the user's cognition: the user is presented with $n - 1$ plots of simulated data and one plot of the real data. If the user correctly identifies the plot of the real data from the other plots and explains which feature distinguishes it, then that feature is statistically significant at level $1/n$. The procedure is called the *line-up protocol*, similarly to how criminals are lined up for identification by a witness. The line-up protocol was later used to understand high dimension, low sample size data [64] and to match distance metrics with human choices by crowd-sourcing through Amazon's Mechanical Turk [65].

In the line-up protocol, valid visual inference may be limited to distributional assumptions and general dependencies [1]. The approach presented in this thesis is similar to the line-up; however, there are two key differences. First, the user is presented only with a visualisation of the real data, instead of multiple plots, and second, the statistical test is quantitative, instead of being based on the user's cognition. The visual pattern to be tested is specified beforehand according to the user's knowledge and a test statistic is explicitly quantified as a function of the plot that measures the strength of the visual pattern.

Visual inference is not the first instance of combining visual tools with statistical testing. A confirmatory data analysis methodology is combined with exploratory tools whenever data visualisations are used as comparisons to reference distributions [30]. The reference distribution is either derived from a fitted model (explicit comparison) or from assumptions of the analyst (implicit comparison), e.g., independence. If the visual inspection of the data reveals patterns that do not generally appear in the data sampled from the model, then the exploratory analysis has indicated a potential misfit of model to data [30]. An example is the quantile-quantile (QQ) plot which is often used to showcase differences between the real data and the model-generated data.

Significance testing in data mining Significance testing has been used in data mining to search for and assess patterns in large datasets [66]. In pattern mining, the search space of patterns is typically reduced through the use of constraints. The constraints specify which patterns are interesting to the data analyst and they are expressed through *interestingness measures* (also known as utility or quality measures and score functions) [67, 68]. However, even after constraining the search space, the number of discovered patterns can be very large, which leads to the risk of false discoveries, i.e., patterns that satisfy the defined constraints by chance alone. The risk can be reduced by applying significance tests and discarding non-significant patterns. The null hypothesis is typically stated, e.g., as H_0: the pattern is an uninteresting, random artefact of the data. Significance testing has also been used in iterative data mining through randomisation tests which take into account previously discovered patterns [69] and in non-interactive settings to find the most informative set of patterns according to a global test statistic [70].

There are various definitions for patterns in data mining, e.g., frequent itemsets and association rules [66], clusterings, and so on. Each definition determines the particular significance test that assesses the results of discovered patterns. In a general sense, a pattern i is present in the observed dataset x, if x is a subset of all possible datasets that include the pattern i [44]. In 0-1 data, patterns such as correlations, frequent itemsets and clusterings have been tested with randomisation tests [71]. If a pattern is defined as a region with a high density of objects, then patterns are discovered by rejecting the null hypothesis of a uniform data distribution. In this case, the null hypothesis can be rejected by estimating local peaks of an empirical probability density function that is based on nearest neighbours for each point [72]. Similarly, statistically significant high-dimensional regions (hyper-rectangles) can be discovered by comparing the enclosed count of points to the expected count under the uniform distribution [73]. For more related work on assessing data mining results with significance testing see [71, 74].

Statistically sound pattern discovery is in its infancy and there are numerous open problems [74]. Webb noted that pattern mining may have to be accepted as statistically unsound by its nature, since it either leads to high false discovery rates or to no discoveries due to multiple testing corrections [75]. A solution he proposed is the *holdout* approach [76, 75], which splits the dataset into an exploratory and a holdout set. The exploratory set is used for searching for patterns, while the holdout set is used to evaluate any discovered patterns. The multiple testing correction is then based only on the discovered patterns, instead of the whole search space. Although the holdout approach is a faster alternative to multiple testing corrections, its performance depends on how the dataset is partitioned [77].

Visual patterns One of the tasks of this thesis is to quantify visual patterns. Visual features have been previously described through score functions to perform interesting tasks.

Scagnostics [78] reduce a scatterplot into a number through a score function that describes visual features of the 2D point cloud, e.g., its clumpiness or its skewness. The idea of scagnostics was initially outlined by Tukey [79] and later implemented into

9 graph-theoretic measures (Outlying, Convex, Skinny, Stringy, Straight, Monotonic, Skewed, Clumpy, Striated) on geometric graphs of the 2D point cloud (convex hull, alpha shape, minimum spanning tree) [78]. The 9 scagnostics were later examined and evaluated in terms of consistency under different sample sizes, homogeneity for various point distributions, sensitivity to outliers and their dimensionality [80]. The measures were conceived to differentiate between many types of point distributions, while being small in number, having comparable scales and distributions, and being efficiently computable. Since a scagnostic is essentially any score function that reduces a scatterplot into a number, the aforementioned evaluative approaches can be applied to any score functions that are based on visual features. It has been noted that more scagnostics scores need to be devised [81].

Scagnostics have been applied to various tasks: interactive exploration [82, 83], ranking subplots in scatterplot matrices [84], discovering visual patterns in subspaces of high-dimensional data [85], detecting interesting subsequences in multivariate time series [86] and discovering "hidden" visual patterns [87]. While scagnostics describe the global visual features of a scatterplot, scores for local visual features have also been developed using motif discovery methods [88].

Ideally, score functions that attempt to quantify visual patterns would be perceptually motivated. An interesting question in that direction is what kinds of patterns are perceived visually by a data analyst. A related study defined visual aggregation [89] as the ability to extract higher order statistics from visualisations, such as when a person estimates the mean of a point cloud in a scatterplot. The same study also noted that visual aggregation may increase the trust and ownership in conclusions derived from the data, compared to automated systems. This suggests that it is especially important to control and correct biases that may occur during visual exploration.

4 Significance testing of patterns in data visualisations

This section describes the proposed statistical significance testing framework for visual patterns. It is organised as follows: first, the problem is described and defined, then examples are presented of test statistics and null distributions, then the proposed solutions are described, and finally the framework is summarised.

Problem motivation Suppose that a data analyst iteratively views different visualisations of the data. If a pattern is observed, how can the data analyst determine whether it is significant or just a random artefact of the data? As a solution, a statistical significance test for visual patterns can be formulated by defining a test statistic and its distribution under the null hypothesis (null distribution). The test statistic is a function of the elements in the visualisation and measures the "strength" of the visual pattern. The null hypothesis is defined in a broad sense as the assumptions of the data analyst about the data generating process, e.g., that the data originate from a Gaussian distribution or that certain attributes in the data are independent. If it is possible to sample from the null distribution, then an empirical p-value is computed as the proportion of sampled test statistics (or of test statistics calculated on sampled datasets) that are more extreme than the observed test statistic. The significance of the observed visual pattern is then expressed through this computed empirical p-value.

This approach of testing visual patterns, as described above, is not statistically valid and requires certain modifications. In a visual data exploration, the data analyst typically views the data *first* and then formulates a hypothesis. Testing a hypothesis after viewing the data can lead to false discoveries, as discussed in Sections 2.3 and 2.4. Hence, the procedure described above must be modified to control the risk of false discoveries (Type I errors). A naive solution is to formulate all possible hypotheses prior to visualising the data and then apply a multiple testing correction (MTC) whenever a pattern is tested. This solution, however, fails in a visual exploration. The number of possible hypotheses is typically very large, since a hypothesis is formulated for essentially every visual pattern that may potentially be observed. Testing multiple hypotheses and then applying a MTC typically reduces the statistical power of the procedure, as the number of hypotheses increases. For instance, if 1 million hypotheses are considered, then applying a Bonferroni correction corresponds to multiplying the p-values by 1 million, which makes it more unlikely that any p-value is deemed significant. This means that true patterns in the data may not be discovered, i.e., Type II errors may occur due to low statistical power.

The problem, therefore, is to define a suitable procedure that controls both Type I and Type II errors when using significance testing on visual patterns during exploration.

Problem formulation Specifically, the objective is to define a statistical significance testing procedure for visual patterns that has high statistical power and that

provides an upper bound for the probability of even one false discovery by a given α, i.e., a procedure that controls the family-wise error rate (FWER).

A false discovery is made when a pattern i is a random artefact, yet is found to be significant, i.e., $p_i \leq \alpha$. The procedure should ensure that the probability of even one false discovery is less than α, i.e., FWER $\leq \alpha$. In other words, if the visual pattern is a random artefact, then its adjusted p-value $\tilde{p}_i$ should be at most α with a probability that is bounded from above by α:

$$\Pr(\tilde{p}_i \leq \alpha \mid \text{pattern } i \text{ is a random artefact}) \leq \alpha \tag{11}$$

In the remainder of this thesis $\alpha = 0.1$ is used, unless otherwise stated.

Definitions The notation that follows is adapted from [70]. Let Ω denote the sample space, which includes all possible data samples, and let $\omega_0 \in \Omega$ denote the observed dataset. The probability of a single data sample $\omega \in \Omega$ is denoted with $\Pr(\omega)$.

Let n_T denote the number of pre-defined *test statistics*, which correspond to *hypotheses* to be tested. Each test statistic is defined by a function $T_i : \Omega \mapsto \mathbb{R}$, where $i \in [n_T]$ and corresponds to a distinct *visual pattern*, denoted by i. The visual pattern i is defined as a set of n_c constraints $j \in [n_c]$ on the data. Let $C_j \subseteq \Omega$ denote the set of data samples that satisfy constraint j. Then the visual pattern i is present in ω_0, if ω_0 satisfies all the constraints of i, i.e., if $\omega_0 \in C_j,\ \forall j \in [n_c]$.

The *null distribution* is the probability distribution of the test statistic if the null hypothesis is true, i.e., $\Pr(T_i(\omega) \mid H_i)$. Since the null distribution is obtained through resampling, the term is also used to refer to the probability distribution of the *data* under the null hypothesis, i.e., $\Pr(\omega \mid H_i)$, depending on the context.

The assumption is that the user iteratively views different visualisations of the data, where views are denoted by $t = 1, 2, \ldots$, and attempts to find all test statistics that do not obey the distribution given by $T_i(\omega)$ when $\omega \sim \Pr(\omega)$, i.e., he or she attempts to find all test statistics that are "significant".

The *unadjusted or raw p-value* that corresponds to the test statistic $T_i,\ i \in [n_T]$ in an iteration t is defined as $p_i^t = \Pr(\Omega_i^+)$, where $\Omega_i^+ = \{\omega \in \Omega \mid T_i(\omega_0) \leq T_i(\omega)\}$ and where either T_i or ω is sampled from the null distribution. In other words, the raw p-value at an iteration t is the probability of observing values of the test statistic T_i at least as high as in the observed data in the iteration t.

4.1 Examples of visual test statistics

This section summarises several illustrative examples of test statistics used in this thesis.

Tabular data are often visualised using scatterplots. A usual visual pattern in scatterplots is a *cluster*. If the user draws on the plot a polygonal region over a dense region of points, then the *number of points inside that region* can be used as a test statistic for a cluster (see Figure 5 in Section 4.3.1).

As described in Section 3.3, *scagnostics* measure the overall form and global visual features of a scatterplot. A test statistic based on scagnostics can be used to

quantify, e.g., the skewness, stringiness or clumpiness of a scatterplot (see Figure 9 in Section 5.2).

In a time series visualisation, visual patterns may correspond to individual time instances or time intervals. As an example of the former, a test statistic based on the visual pattern of a *peak value* can be encoded as *the value of the time series* at a particular time (see Figure 6 in Section 4.3.2). As an example of the latter, a test statistic based on the visual pattern of an *increase in an interval* $[t_0, t_1]$ of the time series x can be encoded as the *difference* $x(t_1) - x(t_0)$ (see Figure 12 in Section 5.5). In general, a test statistic for a visual pattern can be any function that is defined on a time instance or an interval, e.g., groups of outliers, level change-points, volatility change-points, cycles, trends, changes in trends, turning-points and gaps [90]. Furthermore, the time series data mining literature can be utilised (see [91] for a review), e.g., to define patterns as motifs [92] or time series shapelets [93].

4.2 Examples of null distributions

This section describes several approaches for approximating the null distribution. The null distribution encodes the assumptions that the data analyst wishes to be disproven by an observed pattern. The choice of a suitable null distribution depends on the data, the expected visual pattern and the application domain.

Although the distribution of the test statistic under a specific null hypothesis is unknown in the general case, it can be approximated with resampling methods. This thesis utilises the *method of surrogate data* [94] to empirically estimate the sampling distribution of the test statistic under the null hypothesis. The null distribution is approximated by generating an ensemble of surrogate datasets that are consistent with the null hypothesis. The generated surrogates are then used to compute an empirical p-value by comparing the value of the test statistic in the original data to the values in the surrogate data. There are two main approaches for generating surrogate data [95, 96]: *typical realisations* and *constrained realisations*.

Typical realisations surrogates Typical realisations surrogates are obtained by constructing a model of the data and then generating data from the model. For instance, if the data are assumed to follow a Gaussian distribution with a particular mean and standard deviation, then the null distribution is obtained by sampling from that particular Gaussian distribution. When an observed visual pattern suggests that the data do not follow the null distribution, then the test statistic that corresponds to that visual pattern is significant.

For time series data, an example of typical realisations surrogates is *Gaussian processes* (see Figure 11 in Section 5.4), which resemble normality assumptions in statistical tests. A Gaussian process is a stochastic process in which all finite subsets follow a multivariate Gaussian distribution [97]. It is a generalisation of the multivariate Gaussian distribution that determines a probability distribution over functions. A Gaussian process is defined completely by a mean function and a covariance function, often called a *kernel function*, that specifies the similarity between any two points. A convenient aspect of Gaussian processes is that constraints

can be added on the null distribution, such that the sampled surrogates always pass through certain time points (see Figure 11b in Section 5.4).

Another example of typical realisations surrogates for time series data is *historical surrogates*, which involve sampling from the time series itself (see Figure 12 in Section 5.5). For instance, if time series data of past measurements are available, then it is possible to directly sample datasets from the same distribution as the observed dataset. An important assumption in this case is that the underlying distribution of the data does not change over time, which may be stated as a null hypothesis of stationarity of a certain property. As another example of historical surrogates, consider the pattern of sudden fluctuations in the price change of a particular stock on a particular day. Instead of generating surrogates to investigate a hypothesis, the actual historical data can be used as surrogates, i.e., time series representing the price change of the stock from different days.

A similar approach to historical surrogates is to sample from *related* time series. For the previous example of stocks data, this involves randomly sampling surrogates from similar stocks (e.g., stocks from the same sector), instead of randomly sampling different days of the same time series. However, the use of related time series as surrogates depends heavily on the domain knowledge of the user and is most likely not feasible in practice, since many complicated implicit assumptions are made about the data generating process of the time series.

Constrained realisation surrogates Constrained realisation surrogates are constructed such that certain properties are exactly present in the surrogates. For instance, when tabular data are permuted column-wise, the generated permuted datasets are surrogates with the same marginal distributions as the original data.

One example of constrained surrogates used in this thesis is the *background distribution* of the user during visual exploration (see Figure 10 in Section 4.3.3), as defined in [98]. The background distribution models the knowledge of the user and is defined by a probability function Pr over the sample space Ω. In the beginning of the exploration, the user is assumed to have no knowledge about the data besides the marginal distributions. At this stage, samples are drawn from the background distribution by permuting the columns of the data matrix. If the user has knowledge of a pattern in the data or discovers a pattern (e.g., a correlation or a cluster), then this pattern can update the background distribution. If the background distribution is updated, it models the user's knowledge of the pattern and enables the process of discovery to be repeated ($\Pr(\omega) \leftarrow \Pr(\omega \mid \text{discovered pattern})$). The observed pattern updates the background distribution by acting as a *tile constraint* on the permutation process. A tile is a subset of rows and columns of the data matrix. A tile acts as a constraint by using the same permutation on the elements inside the tile, thus preserving the relations between them, while breaking any relation with other elements. After the distribution has been updated, the p-values for the test statistics that are used as constraints can no longer be significant. The background distribution thus models the knowledge of the user by preserving the relations between the elements of discovered or known patterns. Note that during visual exploration, the background distribution can be different from the null distribution. The user's

background distribution is used to find the most interesting views and the null distribution is used to test observed patterns.

Constrained surrogates also exist for time series data. The simplest case of permuting a time series is equivalent to sampling from time series with the same values and in which the autocorrelation structure is broken. Using permutations of a time series as a null distribution corresponds to a null hypothesis that the data are white noise. If this null hypothesis is not rejected, then the data have not been determined to be significantly different from white noise and any subsequent analyses may be of no use [99]. Another example of constrained surrogates in time series, as mentioned above, is obtained from Gaussian processes that have been constrained to always pass through certain points (see Figure 11b in Section 5.4).

An important point to consider when using surrogates in statistical testing is that an observed result is not tested for significance against a set of surrogates. Rather, the only statistically meaningful interpretation is that a null hypothesis is rejected or not rejected [96]. For a survey of null hypotheses for testing, e.g., non-linearity and independence in time series, see [100].

4.3 Solutions for multiple testing problem

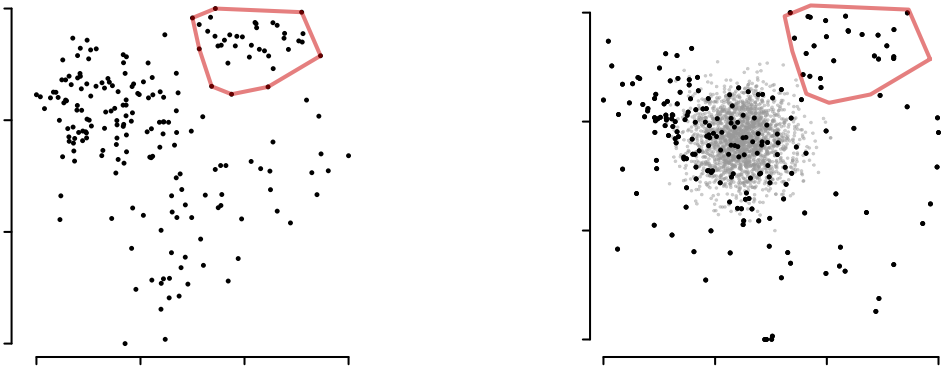
When a statistical test is performed, the null hypothesis must be formulated prior to viewing the data. When a data analyst explores the data visually, he or she views the data first, which presents a problem if hypotheses are to be tested. The naive solution of specifying all possible hypotheses and applying a multiple testing correction fails, due to the sheer number of possible visual patterns. The more hypotheses are tested, the less powerful the test becomes and the more likely it is that true effects are not discovered in the data. In other words:

- If no multiple testing correction is applied, then there is a risk of Type I errors (false discoveries), i.e., the analyst discovers a pattern, when it is only random noise.
- If an overly strict multiple testing correction is applied (e.g., Bonferroni) then there is a risk of Type II errors (false negatives), i.e., the analyst fails to discover a pattern that is truly present in the data.

This thesis presents three solutions related to this problem: splitting the data into visualisation and validation sets, leveraging the analyst's knowledge to reduce the hypothesis space, and a suitable multiple testing correction for iterative exploration.

4.3.1 Splittable data

If the dataset is *splittable* into two conditionally independent parts, given the generating model (e.g., i.i.d. data), then the first part can be used to formulate the hypothesis to be tested, while the second part of the data is used for the actual test. If the dataset is *non-splittable* (e.g., time series data), then the test statistics must be chosen prior to viewing the data and the workflow of the next section (4.3.2)



(a) The visualisation set is used for viewing the data and discovering patterns (such as the marked region).

(b) The validation set is hidden from the user and is used for testing patterns discovered in the visualisation set. The grey points are surrogate data generated by independently permuting the columns of the data.

Figure 5: An example of the workflow with splittable data.

is followed. The approach of splittable data is similar to the *holdout* approach in pattern mining [75] (described in Section 3.3).

Figure 5 presents the workflow with splittable data on a PCA projection of the `German` data to the first two principal components. The data are independently split into two: the *visualisation set* is used for visual exploration, while the *validation set* is hidden from the user and is used for validating patterns observed in the visualisation set. In this example, the test statistic is the number of points inside the marked region and the null hypothesis is that the attributes of the data are independent. Surrogate datasets are sampled from the null distribution by independently permuting the columns of the data matrix. A p-value is computed by counting the proportion of surrogate datasets that have more points in the marked region than the observed dataset. This example is continued and explained in more detail in Section 5.3.

4.3.2 Leveraging the analyst's knowledge (within-iteration correction)

This thesis proposes two solutions to alleviate the problem of prior specification of all possible hypotheses: *leveraging the analyst's knowledge* and the choice of a *suitable multiple testing correction for visual exploration.*

The first solution relies on the fact that a data analyst does not need to pre-specify all possible hypotheses about potentially observable visual patterns. If he or she has knowledge about which visual patterns are likely to occur or are likely to be interesting, then the hypotheses can be formulated on this subset of all possible visual patterns. In other words, the hypotheses are specified prior to viewing the data and their number is reduced from the space of all possible visual patterns, to a space that is constrained by the user's knowledge or prior experience.

The second solution involves the choice of multiple testing correction, which is warranted since the analyst formulates multiple hypotheses. The previous solution

already reduces the number of considered hypotheses, thereby increasing the statistical power of the procedure. However, the number of considered hypotheses is still typically large enough that a traditional Bonferroni correction has very little power and may lead to no significant patterns. Nevertheless, even though the number of visual features is very large, the features are often correlated, which means that the effective number of considered hypotheses is smaller. A multiple hypotheses correction that takes into account these correlations is hence more suitable. As explained in Section 2.4, resampling methods naturally incorporate the correlation structures present in the data. Therefore, as a multiple testing correction for visual patterns, this thesis uses the *step-down minP procedure* (henceforth referred to as *minP*), as described in Section 4.3.

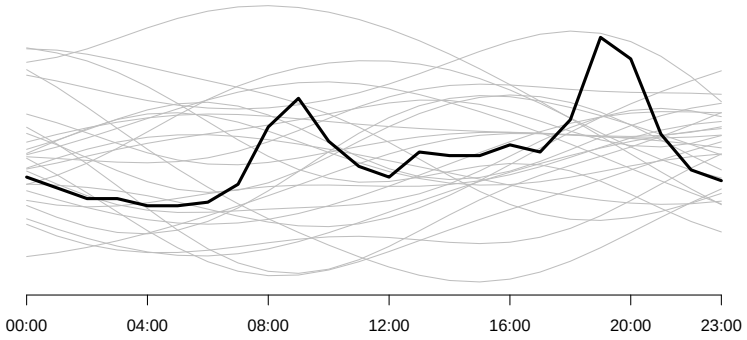


Figure 6: A time series is an example of non-splittable data. The grey lines denote the null distribution, i.e., the assumptions of the data analyst about the data. See Figure 11 in Section 5.4 for details.

As an illustrative example, consider the time series in Figure 6 of the hourly level of carbon monoxide (CO) for a single day (March 25$^{\text{th}}$, 2004) from the UCI [101] `Air quality` dataset [102]. Suppose that an analyst wishes to visually explore time series of CO concentration, one day at a time. The analyst may have knowledge (from experience or expertise) about the typical behaviour of CO over one day and this knowledge informs the questions she asks and the patterns in which she is interested. A typical question may be whether the values of the time series at certain given time instances are extreme compared to her expectations. Before starting the visual exploration, the analyst can prepare a statistical hypothesis test for testing whether an observed peak value is extreme compared to her expectations. The test is formulated by selecting a test statistic that corresponds to the visual pattern of interest (peak value) and a null distribution of time series that reflects the analyst's assumptions about CO concentration. A natural choice for a test statistic in this case is simply the value of the time series at a given time instance. The choice of null distribution is more difficult, since it typically requires expertise about the particular data; here the null distribution is considered to be a prior distribution of a Gaussian process with a squared exponential kernel and a length scale of 6 hours (denoted as grey lines).

When viewing the data, the analyst obviously observes that the value at time 19:00 is the largest and thus chooses to test it using the above mentioned test statistic and null distribution. The resulting raw p-value for the tested time point (19:00) is then 0.02. The choice of testing that particular time instance, however, was made after seeing all the other time instances. The result is not valid, since, out of the 24 available time instances, the "best" option is selected (the one with the maximum test statistic or minimum p-value) which confirms the user's assumptions and leads to risk of false discoveries (Type I errors).

The proposed solution is to instead formulate a test for each possible observable peak value, which results in 24 tests (as opposed to naively specifying every possible observable visual pattern which would result in a much higher number of tests). For each test, the calculated p-values are then adjusted for multiple testing. If the adjustment is a Bonferroni correction, then the adjusted p-value for the highest value at 19:00 is 0.5, i.e., it is not significant. This may be a false negative (Type II error), since it is obvious from the visualisation that the value is high and the Bonferroni correction is known to be conservative. In contrast, if the adjustment is a minP correction, then the adjusted p-value for the considered point is 0.1, which is significant at the significance level $\alpha = 0.1$. This provides a more suitable correction for visual exploration, since the correlation structure in the data and the considered visual patterns is taken into account, resulting in a more realistic adjustment.

The example in Figure 6 is continued and explained further in Section 5.4.

4.3.3 Iterative exploration (between-iteration correction)

In an iterative exploration, such as projection pursuit methods, multiple "views" of the data are iteratively shown to the user. Each view provides information about possibly varying sets of test statistics. If a large enough number of these views are shown, then eventually false positives occur due to chance alone.

A multiple testing correction is warranted, in addition to the previous one, that applies an adjustment on the p-values based on the sequence of views. Again, a standard Bonferroni correction is unrealistic, since the number of views is not known beforehand and early views give more information about the data than views later in the exploration.

This thesis proposes a weighted Bonferroni procedure [103] as a suitable correction for iterative exploration: the p-values in each view are multiplied by a factor of $1/w_t$ where t denotes the order of the view, and the weights w_t sum to unity, i.e., $\sum_{t=1}^{\infty} w_t = 1$. Choosing $w_t = 2^{-t}$ means that it is possible to have an unlimited number of iterations in which the first view has the most statistical power and subsequent views have exponentially decreasing power. In this way, FWER is controlled for the whole sequence of iterations at the chosen level.

The final adjusted p-value at iteration t is then:

$$\tilde{\tilde{p}}_i^t = \min(1, w_t^{-1}\tilde{p}_i^t) \tag{12}$$

where $\tilde{p}_i^t$ is the minP adjusted p-value for the pattern i at iteration t.

An example of iterative exploration is presented in Section 5.3.

4.4 Visual pattern significance framework

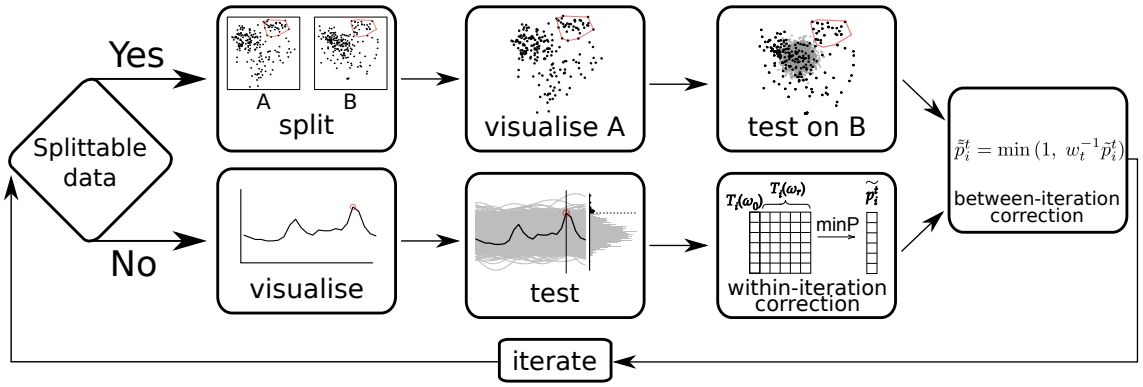


Figure 7: Summary of the visual pattern significance framework

The proposed framework is summarised in Figure 7. The following steps describe a typical workflow for visual exploration using the framework:

1. If the observed data ω_0 are i.i.d. splittable, then split the data into a visualisation set ω_{0_A} and a validation set ω_{0_B}. If the data are non-splittable, then use ω_0 for both visualisation and validation, i.e., $\omega_{0_A} = \omega_{0_B} = \omega_0$.

2. Define the test statistics T_i for the visual patterns of interest $i \in [n_T]$ and the null distribution from which surrogate datasets are sampled.

3. At iteration t, show the user a visualisation of the dataset ω_{0_A} containing k_t patterns.

4. For each pattern, determine the value of the respective test statistic $T_i(\omega_{0_B})$, where $i \in [k_t]$.

5. Sample R surrogate datasets ω_r, where $r \in [R]$, from the null distribution and calculate the test statistics $T_i(\omega_r)$ for each dataset ω_r and $i \in [k_t]$.

6. Compute minP adjusted p-values $\tilde{p}_i^t$ using the observed test statistics $T_i(\omega_{0_B})$ and their respective values on the surrogate datasets $T_i(\omega_r)$.

7. Adjust the p-values for the current iteration by multiplying the p-values by $w_t^{-1} = 2^t$. The final p-values are then $\tilde{\tilde{p}}_i^t = \min(1,\ w_t^{-1}\tilde{p}_i^t)$, which are deemed significant if $\tilde{\tilde{p}}_i^t \leq \alpha$.

8. Increment t by one and repeat the process from step 1 until the exploration is over.

5 Empirical evaluation

In this section, the framework is empirically evaluated by presenting several use-cases on tabular and time series data. First, a simulation is performed to assess how the knowledge of the analyst affects the the significance of visual patterns. The framework is then applied in testing visual patterns in tabular data and time series. In each example, two baseline solutions are presented: no MTC adjustment (risk of Type I error rate) and Bonferroni adjustment (risk of Type II error rate). The experiments were run using R (version 3.5.2) [104]. For the sake of reproducibility, all code used for the experiments and for generating the figures in this thesis is publicly available from https://github.com/rafael-savvides/visual-pattern-significance.

Datasets The experiments use three real-world datasets and one synthetic dataset. (1) The `German` socioeconomic dataset [105, 106] is a tabular dataset containing data from 412 German administrative districts. Each district is represented by 46 socioeconomic, political and geographic attributes.[1] The data are preprocessed as in [98], resulting in 32 real-valued attributes and two class attributes *Type* (Urban, Rural) and *Region* (West, South, East, North). (2) The UCI [101] `Air Quality` dataset [102] contains time series originating from sensors measuring air quality (e.g., carbon monoxide, nitrogen oxides and Benzene). Only the hourly concentration of carbon monoxide (CO) is used. For preprocessing, values with missing time stamps are removed. In both datasets, the real-valued variables are scaled to zero mean and unit variance. (3) The `SMEAR` dataset [107] contains atmospheric measurements on aerosol particle size distributions that originate from SMEAR stations.[2] The SMEAR (Stations for Measuring the forest Ecosystem–Atmosphere Relationships) collect comprehensive data on aerosol and trace gas concentrations, biosphere–atmosphere interactions, aerosol formation and growth, and the biogenic background for processes leading to aerosol formation. Only the aerosol particle size distributions are used. (4) A synthetic dataset is described in Section 5.1.

5.1 Simulated user study

A user with prior knowledge about the data has a high chance of asking the right questions. This experiment is a simulated user study, in which the framework is applied in a simple scenario with synthetic data. The data consist of n numbers, one of which is different (e.g., x_1). The user has access to a test that can be used to determine whether a particular number is different. Using this test, the task of the user is to discover true patterns in the data, i.e., that x_1 is different. The assumption is that an expert user is more likely to correctly select x_1, while a non-expert may need to perform multiple tests before correctly selecting x_1. If multiple numbers are tested, then a multiple testing correction is applied. If the number of tested numbers is overly large, then the correction causes the test to not determine that x_1 is different, i.e., it results in a Type II error.

[1] Available from http://users.ugent.be/~bkang/software/sica/sica.zip
[2] Available through a public API at https://avaa.tdata.fi/web/smart/smear/api

This experiment demonstrates that experts using the framework are more likely to discover that x_1 is different even when there are many numbers, and that non-experts using the framework are also likely to discover x_1 when there are few numbers but fail to do so for increasing n due to the multiple testing correction. In a visual exploration, the numbers are replaced by visual patterns. The expert knows which patterns are likely to be significant and can specify which ones to test before looking at the visualisation. This improves statistical power, i.e., the ability to detect true patterns and reduce false negatives.

The experiment is described in more detail as follows: The synthetic data consist of n real-valued numbers $X_n = \{x_1, \ldots, x_n\}$. The significant test statistic x_1 is sampled from a Gaussian distribution with mean $\mu = 3$ and standard deviation 1, i.e., $x_1 \sim \mathcal{N}(\mu, 1)$, while the other x_i's are sampled from $\mathcal{N}(0, 1)$. The null hypothesis H_i for each data element x_i is that $x_i \sim \mathcal{N}(0, 1)$. The test statistic is the value of each data element x_i. Therefore, the assumption is that only the test statistic for x_1 should be significant.

The analyst's knowledge is simulated with a simple single-parameter model defining the analyst's level of *expertise*. The parameter k describes the probability of knowingly choosing the correct test statistic (representing the *knowledge* of the analyst). If the correct test statistic is not chosen (probability $1 - k$), then the test statistic is chosen uniformly at random among all test statistics. In other words, an expert (high k) is able to select the correct test statistic with high probability, while a non-expert (low k) is more likely to make a random selection.

In addition, the user may select multiple test statistics to increase the probability of choosing the correct one. Suppose that the number of chosen test statistics is m and that an analyst with expertise k wishes to guarantee that the probability of choosing x_1 correctly out of the n test statistics is at least β. Then the required number m^* of chosen test statistics to ensure success at probability at least β is calculated (in Appendix A) as:

$$m^* = \lceil \frac{\log(1-\beta)}{\log(1-\frac{1}{n}) + \log(1-k)} \rceil \tag{13}$$

where $\lceil \square \rceil$ denotes the ceiling function that maps $\square$ to the least integer greater than or equal to $\square$.

The simulation is run for varying numbers of test statistics n and knowledge levels of the analyst k to compute whether the analyst correctly finds the test statistic for x_1 to be significant. The analyst chooses $m*$ test statistics which correspond to m^* tests for which an adjusted p-value is computed using the minP correction (all the chosen test statistics are evaluated simultaneously). The simulation is run for 1000 replications and the mean adjusted p-value for x_1 is presented in Figure 8. If the test statistic for x_1 is not among the test statistics evaluated, then the p-value is taken to be 1.00.

Figure 8 illustrates that high knowledge ($k \geq 0.9$) results in the analyst finding the significant test statistic (i.e., results in low p-values), regardless of the number of test statistics n. On the other hand, an analyst with low knowledge ($k < 0.2$) does

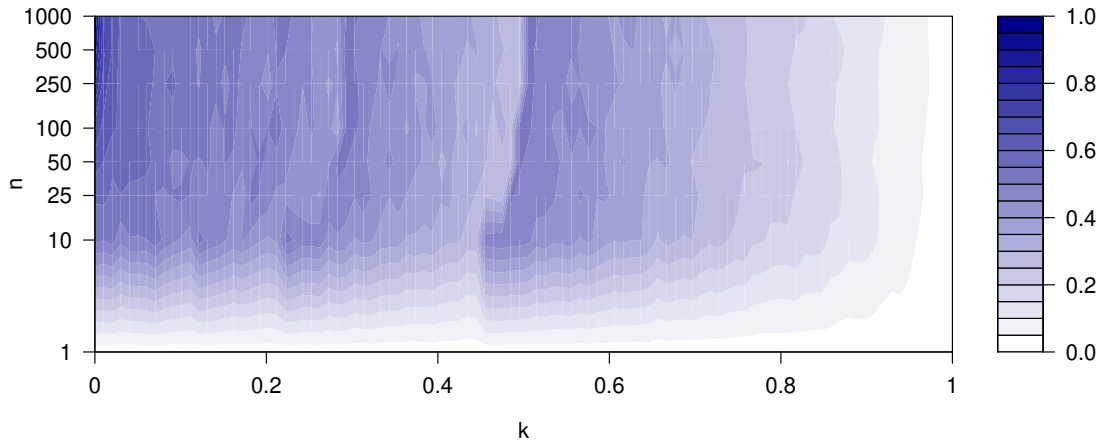


Figure 8: Experiment with synthetic data. The parameter k (on the x-axis) models the experience ("knowledge") of the analyst and n (on the y-axis) is the number of test statistics. The colour shows the mean adjusted p-value of the test statistic for x_1 over 1000 replications. The mean for x_1 is $\mu = 3$ ($x_1 \sim \mathcal{N}(3, 1)$) and the desired level of confidence is $\beta = 0.5$.

not benefit from the procedure, since "guessing" the test statistic is likely to fail due to the applied correction, even if n is relatively low.

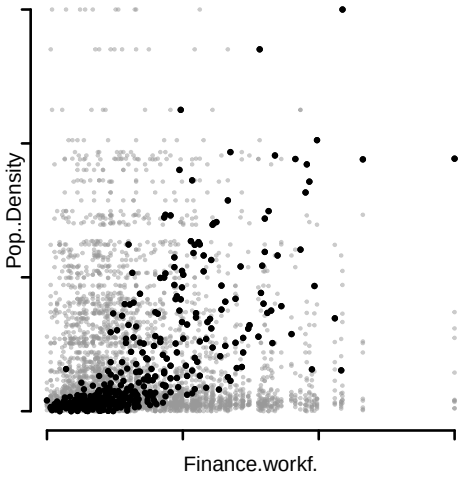
5.2 Scatterplot with scagnostics and permutations

This example uses scagnostics as test statistics and column-wise independent permutations as the null distribution. The scatterplot in Figure 9a presents the attributes `Finance Workforce` and `Population Density` in the `German` dataset.

Suppose that the analyst assumes these attributes to be independent (the null hypothesis) and wishes to investigate how this assumption is reflected in the scagnostics. After deciding to test for scagnostics, the data is visualised (Figure 9a). The analyst observes that the plot appears *skewed* and *monotonic* and computes these scagnostics (column T in Table 9b). Are these values significant, if the assumption is that the attributes are independent, or are they likely to have occurred by chance?

The significance of the scagnostics is determined by following the steps in Section 4.4 for one iteration ($t = 1$) using the 9 scagnostics measures as test statistics. The null distribution corresponds to uniformly sampling datasets from a distribution over all datasets having the same marginal distributions as the original dataset, with the requirement that all attributes are independent. This distribution is realised by permuting each attribute independently.

Table 9b presents the raw, Bonferroni adjusted and minP adjusted p-values for each scagnostic. Notice that most of the Bonferroni adjusted p-values are insignificant, while the raw p-values are much smaller and the minP adjusted p-values are in between. According to the observed minP adjusted p-values, the null hypothesis of independence of `Finance Workforce` and `Population Density` is rejected for the scagnostic *monotonic* ($p_{minP} \leq 0.10$). In other words, if the attributes `Finance`



(a) Scatterplot showing attributes `Finance Workforce` and `Population Density` in the `German` data. Grey points are surrogates generated by permuting each column of the data independently.

Scagnostic	T	p_{raw}	p_{bonf}	p_{minP}
Outlying	0.37	0.31	1.00	0.90
Skewed	0.81	0.57	1.00	0.95
Clumpy	0.03	0.17	1.00	0.74
Sparse	0.05	0.22	1.00	0.80
Striated	0.05	0.69	1.00	0.95
Convex	0.46	0.36	1.00	0.91
Skinny	0.46	0.81	1.00	0.95
Stringy	0.35	0.70	1.00	0.95
Monotonic	0.63	0.00	0.01	0.01

(b) Significance of scagnostics computed for the scatterplot in Figure 9a. The columns show the value of the test statistic (T) and the corresponding raw, Bonferroni adjusted and minP adjusted p-values.

Figure 9: Scagnostics example

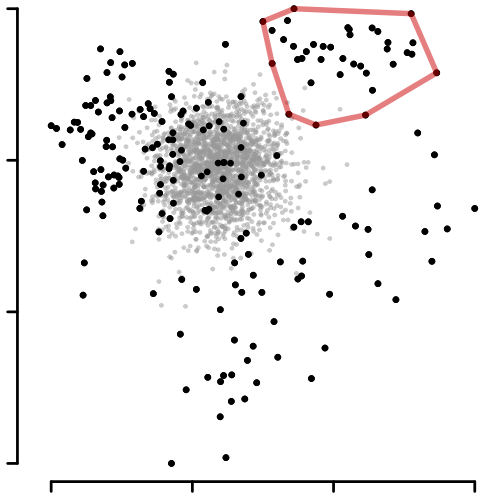
`Workforce` and `Population Density` are independent and have these particular marginal distribution, then it is unlikely that this value or higher would be observed for the *monotonic* measure. In contrast, for the *skewed* measure the null hypothesis of independence fails to be rejected. This means that this value or higher for the skewed measure is not unusual, when the attributes are independent and have these particular marginal distributions.

The results do not mean that the figure is not skewed (which it seems to be in this visualisation). Rather, it implies that the skewness of the visualisation is not statistically significant, when the assumption is that the attributes `Finance Workforce` and `Population Density` are independent and have these particular marginal distributions.

In addition, the minP adjusted p-values in this example are conservative (many are over 0.90) and are comparable with the Bonferroni adjusted ones. The reason for this may be that the 9 scagnostics measures were devised to measure different aspects of the scatterplot, i.e., there are not many dependencies between them. Since the improved performance of the minP correction over the Bonferroni is mainly based on leveraging correlations between the test statistics, the absence of dependencies may explain the results of Table 9b.

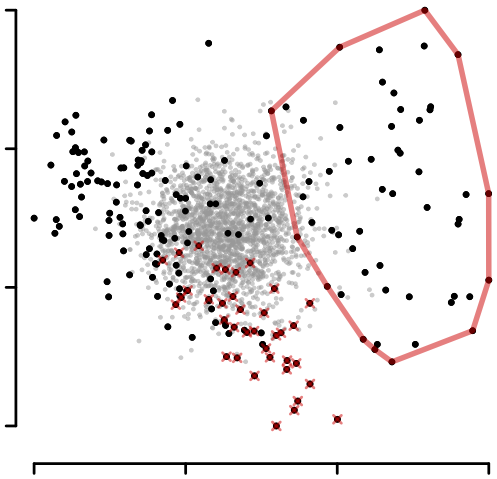
5.3 Scatterplot with number of points and permutations

In this example, a test statistic is used for testing clusters in the scatterplots of Figure 10. The data are assumed to be i.i.d. splittable, therefore one part is visualised and the other part is used for testing the observed visual patterns. The



	x	w_x	y	w_y
1	Unemploy.	0.30	GREEN.2009	-0.29
2	Unempl..Youth	0.29	Finance.workf.	-0.29
3	Children.Pop.	-0.28	Constr..workf.	0.28
4	LEFT.2009	0.28	Mid.aged.Pop.	-0.27
5	FDP.2009	-0.25	Old.Pop.	0.27

(a) In this projection onto the first two principal components, a pattern marked by the analyst as a polygonal region is found to be significant.



	x	w_x	y	w_y
1	Children.Pop.	-0.28	Elderly.pop.	-0.40
2	X.cars	-0.27	Unemploy.	-0.35
3	Service.workf.	0.25	No.school.degree	-0.31
4	Prod..workf.	-0.24	Mid.aged.Pop.	0.30
5	LEFT.2009	0.24	SPD.2009	-0.29

(b) Projection of the data onto the *most informative view* [98] which is computed using the pattern in Figure 10a as a constraint on the background distribution. The data points enclosed in the polygon in Figure 10a are here marked with red crosses. Another pattern that is marked by the user with a polygonal region is also found to be significant.

Figure 10: Projected views of `German` data. The tables below the scatterplots provide the attributes with the five largest (in absolute value) weights on the projection axes. The grey points are surrogate data sampled from the null distribution.

iteration correction is also applied, since the second figure is generated based on information from the first figure (using the background distribution to compute the most informative view [98]).

Suppose that an analyst wishes to visually explore the `German` data in an *iterative* workflow and evaluate any observed clusters. The data are randomly split into a visualisation and a validation set. Figure 10a displays a projection of the data onto its first two principal components. A dense region is observed (marked with a red polygonal line) which contains *rural* districts in the *East*. Is this cluster a true pattern in the data or is it just a random artefact?

Following the steps in Section 4.4, a null distribution and a test statistic are required. The test statistic is the number of points inside the marked region. The null hypothesis is that the pattern is explained by the user's background knowledge, which at this stage is only the marginal distributions of the data. Samples can be drawn

from the null distribution by permuting the attributes of the data independently (similarly to the previous scagnostics example). The permutations are performed on the validation set.

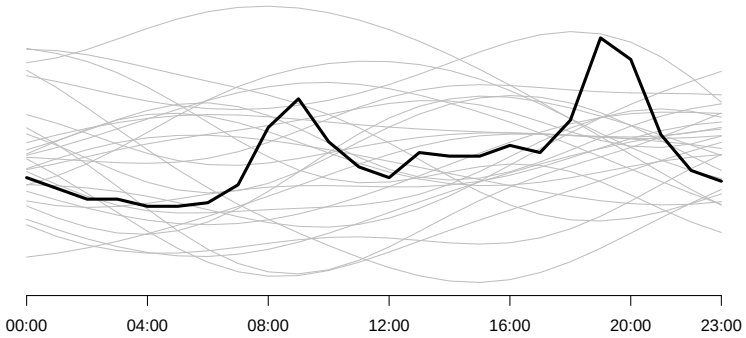
Using this test statistic and null distribution, the procedure for testing the significance of the observed cluster is as follows: samples are drawn from the null distribution, then they are projected onto the same view as Figure 10a and the number of points is counted inside the region for each sample, resulting in the null distribution. No MTC is applied for multiple test statistics, since the observed pattern is validated using independent data. Nevertheless, an iteration correction is warranted, since the data are iteratively explored.

In this first iteration step, the pattern inside the marked region is indeed significant, since its iteration adjusted p-value for the iteration $t = 1$ is $\tilde{p}^{t=1}_{\texttt{RuralEast}} = 0.002$. This means that the observed cluster is not present in datasets in which the attributes are independent and have the same marginal distributions as the observed data, i.e., it is not explained by the knowledge of the user (as modelled with by background distribution).

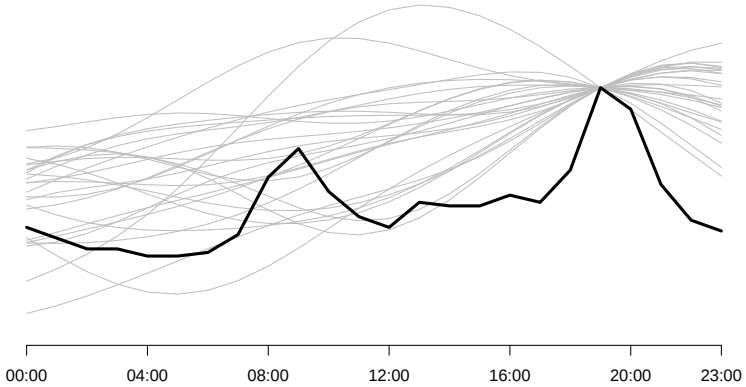
A second iteration step is taken by computing the *most informative view* with respect to the background distribution (as described in [98]), shown in Figure 10b. This view is computed by adding the pattern observed in the previous step as a constraint to the *background distribution*. The pattern thus acts as a *tile constraint* in the randomisation process that generates samples from the distribution of datasets with the same marginal distributions. The constraint effectively alters this distribution in such a way that the observed cluster is always present in the resampled versions of the data. The tile constraint retains the interaction between certain attributes for a subset of the data by permuting them independently from the rest of the data. In this case, the interaction is retained between all attributes in the subset defined by the marked region. The previous pattern is displayed in the second view with red crosses and is no longer significant with respect to the background distribution.

Suppose that a new pattern is observed (marked polygonal region in Figure 10b), that corresponds to *urban* districts. Is this pattern explained by the background distribution (i.e., the knowledge of the user)? The same test statistic is used (number of points inside the marked region) and the null distribution is the background distribution, updated with the previously observed pattern. The iteration adjusted p-value for the new pattern in Figure 10b is then $\tilde{p}^{t=2}_{\texttt{Urban}} = 0.004$. This means that the pattern is significant and is not explained by the knowledge of the user (as modelled by the background distribution).

Note that if the previous pattern `RuralEast` is tested using the null distribution constrained by `RuralEast`, then it is deemed non-significant ($p_{\texttt{RuralEast}} = 1$), since it appears in all the drawn samples. Similarly, the second pattern `Urban` is also found non-significant using a null distribution constrained by `RuralEast` and `Urban`. Patterns that are observed and deemed significant, are no longer significant under the updated knowledge of the user.



(a) Gaussian process priors



(b) Gaussian process posteriors, constrained on the time point with a significant value in Figure 11a. After the constraint, the pattern is no longer significant $p_{t=19:00} = 1$.

time	00	01	02	03	04	05	06	07	08	09	10	11
p_{raw}	0.70	0.76	0.83	0.82	0.86	0.86	0.84	0.76	0.35	0.17	0.45	0.67
p_{bonf}	1	1	1	1	1	1	1	1	1	1	1	1
p_{minP}	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.78	0.51	0.88	0.96
time	12	13	14	15	16	17	18	19	20	21	22	23
p_{raw}	0.75	0.55	0.58	0.57	0.49	0.53	0.32	0.02	0.05	0.39	0.68	0.74
p_{bonf}	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	1.0	1.0	1.0	1.0
p_{minP}	0.98	0.93	0.94	0.94	0.90	0.92	0.75	0.10	0.21	0.83	0.97	0.98

(c) p-values (raw, Bonferroni adjusted and minP adjusted) for each time instance of Figure 11a.

Figure 11: Time series of the hourly level of carbon monoxide (CO) for a single day. The grey lines are surrogate data sampled from a Gaussian process with a squared exponential kernel.

5.4 Time series with peak value and Gaussian process

This section continues the example of Figure 6 in Section 4.3. Figure 11a shows the hourly level of carbon monoxide (CO) for a single day from the UCI [101] `Air quality` dataset [102].

In this example, time series are explored visually. Time series data are not

splittable and the visual pattern significance procedure requires pre-specification of the test statistic and the null distribution. The visual pattern is a peak value, which is pre-specified by using as a test statistic the value of the time series at each time instance. The null distribution is also specified beforehand and it expresses the assumptions of the analyst about the data. The null distribution is in this case obtained as typical realisations surrogates that are sampled from a Gaussian process with a squared exponential kernel and a length scale of 6 hours. In other words, the assumption of the analyst is that the data are generated by a Gaussian process with a squared exponential kernel and he or she wishes to test whether any observed values of the time series are significantly higher than expected.

Table 11c compares 3 p-values for each test: raw, Bonferroni adjusted and minP adjusted p-values. Notice that there are no significant Bonferroni adjusted p-values, while the minP adjusted p-value for the value at 19:00 is still significant at level $\alpha = 0.1$, even though it is adjusted to account for the probability of even one Type I error. Although it may appear that the results do not differ noticeably whether an adjustment is made or not (i.e., the raw p-values and the minP adjusted p-values are not noticeably different), the minP adjusted values provide certain statistical guarantees (control of probability of even one false positive, i.e., FWER control).

As an additional step, a constraint is added on the null distribution to demonstrate that the observed pattern is no longer significant, similarly to the background distribution in the previous example in Section 5.4. Gaussian processes are convenient in this regard, since a posterior distribution which passes through the observed peak value can act as a constrained null distribution. In Figure 11b, a constraint is added on the observed peak value which updates the null distribution and results in the peak being non-significant, i.e., $p_{t=19:00} = 1$.

5.5 Time series with interval and historical surrogates

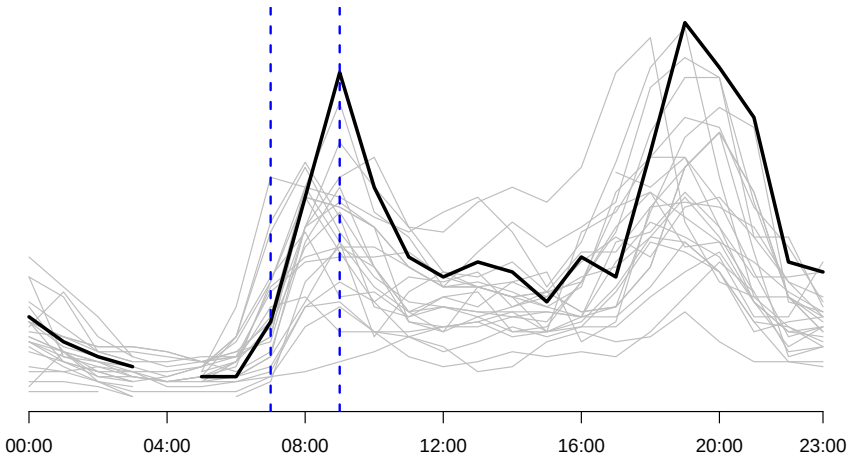
A second example with time series is presented using a different test statistic and *historical surrogates* as a null distribution.

Figure 12 shows the level of carbon monoxide (CO) on Tuesday, March 17$^{\text{th}}$, 2004. Suppose an analyst wishes to visually explore the data and evaluate observed *level changes* in the CO concentration. An unusually large increase is observed in the level of CO during 07:00–09:00 which may be due to, e.g., increased traffic during the morning rush. Is the observed increase an unusual phenomenon in the observed day or is it just a random artefact that is not significantly different from other days? The null distribution is generated by sampling work days as historical surrogates. The test statistic is defined as the difference between the level of CO at 09:00 and 07:00. Since all possible two-hour intervals are observed by the analyst simultaneously, a minP multiple testing correction is applied by considering the difference in the CO level for all two hour intervals in the data.

The resulting raw, Bonferroni adjusted and minP adjusted p-values are displayed in Table 12b. The raw p-values indicate that there are several significant intervals, while the Bonferroni adjusted p-values are all non-significant. The former are not valid due to the multiple testing, while the latter are too conservative to be realistic.

The minP adjusted p-values present a middle solution, as evidenced in its values being in between the other two. Although the minP adjusted p-values are also not significant, they are less conservative than Bonferroni. In this example, there is not enough evidence to reject the null hypothesis i.e., that the observed increase is unusually high on that particular day. If only the raw p-values are considered, the opposite conclusion is reached, i.e., that there is a significant increase in the considered interval.

Notice that the time series in Figure 12 contains a missing value. The sampled historical surrogates may also contain missing values. In this case, the calculations of the empirical p-values and the minP adjusted p-values require a modification. In this example, the missing values are assumed to be insignificant and are replaced by the least significant value, i.e., $-\infty$, since a one-sided test is performed.



(a) An interesting interval is observed between times 7 and 9 in the morning, in which there appears to be an unusually high increase in the level of CO.

interval	00-02	01-03	02-04	03-05	04-06	05-07	06-08	07-09	08-10	09-11	10-12
p_{raw}	0.54	0.40	0.45	0.59	0.44	0.40	0.12	0.02	0.24	0.78	0.73
p_{bonf}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.53	1.00	1.00	1.00
p_{minP}	0.89	0.86	1.00	0.89	1.00	0.86	0.77	0.32	0.85	0.89	0.89
interval	11-13	12-14	13-15	14-16	15-17	16-18	17-19	18-20	19-21	20-22	21-23
p_{raw}	0.49	0.34	0.72	0.31	0.51	0.22	0.01	0.06	0.56	0.78	0.80
p_{bonf}	1.00	1.00	1.00	1.00	1.00	1.00	0.22	1.00	1.00	1.00	1.00
p_{minP}	0.89	0.86	0.89	0.86	0.89	0.85	0.14	0.58	0.89	0.89	0.89

(b) p-values (raw, Bonferroni adjusted and minP adjusted) for each two-hour time interval of Figure 12a.

Figure 12: Time series of the hourly level of CO in the `Air Quality` dataset on March 17$^{\text{th}}$, 2004. The x-axis shows the time of the day. The grey lines are a random sample of other days (historical surrogates). The hourly average value data point for 4 is missing.

5.6 Banana plot

A third example with time series presents an application for multivariate time series visualised in a heatmap. Figure 13a shows a banana plot of the `SMEAR` data for a single day (July 2nd, 2010). The banana plot is a heatmap visualisation of atmospheric aerosol particle size distributions over time, in which the x axis is time, the y axis is particle size and the z axis (encoded as colour) is the particle concentration. A banana is a visual pattern that indicates the formation of new particles (new particle formation, NPF, or nucleation). During NPF, small aerosol particles cluster together over time to form larger particles, resulting in a linearly rising and growing shape of high particle concentrations that resembles a banana. In Figure 13a, a banana starts at 15:00 and continues until the end of the day. Atmospheric scientists visually inspect banana plots of different days to classify them as event or non-event days depending on whether there is NPF occurring or not [108].

The significance of observed bananas can be determined by defining a test statistic and a null distribution. A suitable null distribution in this case is historical surrogates, due to the sheer volume of available data (the measurements have been made every 10 minutes since 1996). A suitable test statistic is chosen here based on the assumption that the banana shape in the heatmap plot is implicitly encoded in the user's perception into another visual representation that resembles a rising line segment. Hence, the test statistic is defined as a function computed on a line segment drawn by the user on the visualisation. A line segment is significant if it resides on a banana, i.e., if a large portion of the line length resides in a high concentration region. In this example, a line is assumed to be significant if most of its values (75%) are sufficiently high, hence the test statistic T is defined to be the 25% quantile of the particle concentration values z on a line segment L.

Similarly to the previous examples, the user simultaneously views all the visual patterns present in the observed data, therefore a multiple testing correction is warranted. In the univariate time series of Section 5.5, a test statistic was defined on an interval and a multiple testing correction was applied by considering all possible intervals of certain length. Accordingly, in this multivariate time series heatmap visualisation, a test statistic is defined on a line segment and the correction is applied by considering all possible lines with certain properties. Since computing all possible lines is not practical, the number of tests is reduced through resampling and through incorporating domain knowledge about the specific visual pattern. Resampling is performed by sampling a number n_L of random lines on the visualisation, instead of considering all possible line segments. A line segment L is parametrized by its starting and ending points (x_0, y_0, x_1, y_1), which are sampled from an $n \times m$ grid, where n, m are the number of points in the x, y axes respectively.

The space of possible lines is reduced further by incorporating domain knowledge. The domain knowledge in this case originates from decision rules that determine whether a banana is a nucleation event [108]. The decision rules state that the observed banana has to: (1) start from particles of size smaller than 25 nm, (2) persist for at least 1 hour, (3) show signs of growth and (4) be distinct mode of new particles. Rules 1-3 can be encoded into constraints for the sampled lines as: (1) y_0

$< 25\,\text{nm}$, (2) $x_1 - x_0 > 1\,\text{h}$, (3) $y_1 > y_0$. This results in improved statistical power, since irrelevant lines are not considered, such as non-rising line segments or lines that begin from large particles.

Figure 13 showcases an example of the procedure. The user views a banana plot and observes a potential banana shape (Figure 13a at time 15:00). The user draws a line on the banana shape to determine whether it is significant (black line in Figure 13b). The significance of the visual pattern is then determined as follows. A number ($n_L = 1000$) of randomly sampled lines is drawn on the banana plot (hidden from the user). The line sampling procedure is constrained by domain knowledge, as described above. The test statistic is defined as the 25% quantile on each line, resulting in 1000 test statistics (plus the user selected line). The null distribution is historical surrogates sampled from previous non-event days, i.e., days in which there are no banana shapes observed. This test statistic and this null distribution are used to compute minP adjusted p-values for each line, determining their significance.

Table 13c presents the raw, Bonferroni adjusted and minP adjusted p-values for the 13 lines with the lowest minP adjusted p-values. The Bonferroni adjusted p-values are all insignificant, since the number of test statistics is very high. The raw p-values indicate that 100 lines out of 1000 are significant (not shown in Table 13c for brevity). Since the raw p-values are not adjusted for false discoveries, it is likely that a number of these significant lines are false discoveries. The proposed solution of a minP adjustment results in 13 significant lines (including the user selected line), which are denoted with brown in Figure 13b.

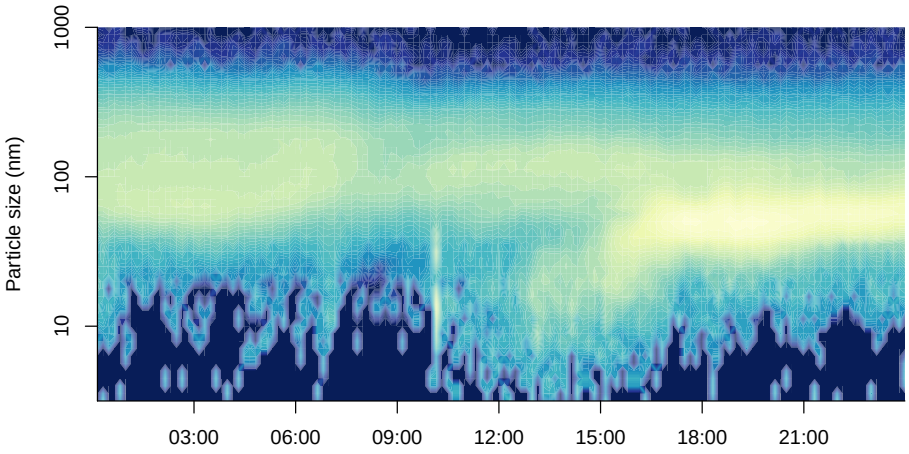
Although a statistically valid interpretation of these results is not straightforward, they can be summarised informally as follows: out of the possible lines that the user could have drawn on the plot, these 13 lines are unusual when compared to other non-event days (i.e., days without a banana shape) using a test statistic that is based on the percentage of high values on the line. This example illustrates that the proposed procedure is both statistically robust (controlling FWER) and not overly conservative (resulting in discoveries), even for more complicated visualisations and hypotheses.

5.7 Scalability

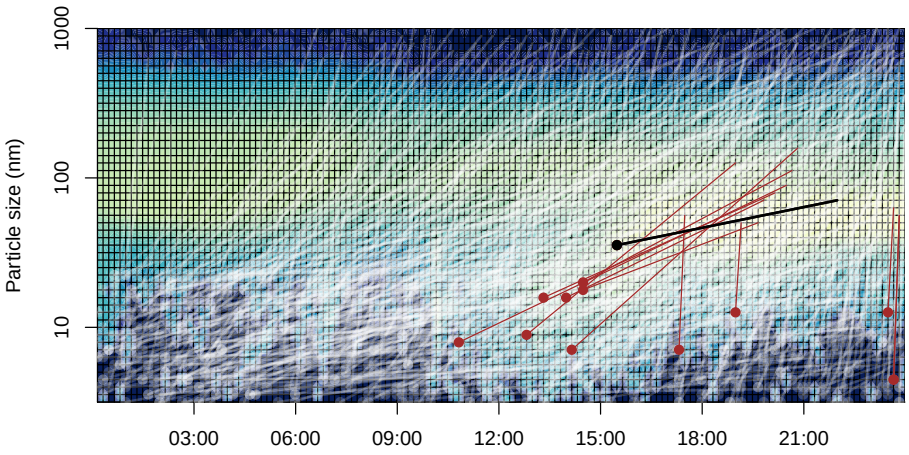
A typical use-case for the proposed significance testing framework is interactive visual exploratory data analysis. During exploration, the data analyst must be able to test hypotheses in a fluid manner, hence the procedure must be fast enough.

All experiments in this thesis can be run in less than 15 minutes using a standard Apple MacBook Pro with a 3.1 GHz Intel Core i5. The most time consuming experiments are the simulated user model and the banana plot, while the others are an order of magnitude faster. For instance, the analysis of the time series example in Fig. 11 requires a few seconds, which is fast enough for interactive use. As a rough estimate, the time to test a single pattern is $R \times (T_T + T_S) + T_C$ where R is the number of surrogates, T_T and T_S are respectively the computation time for the test statistic T and sampling a surrogate dataset S, and T_C is the time for applying the multiple testing corrections (minP and iteration adjustment).

Note, however, that the complexity of generating surrogates depends entirely on the null hypothesis. For certain complex hypotheses it is possible that, e.g., Markov Chain Monte Carlo (MCMC) methods must be utilised to generate surrogates (e.g., [109]), which may be computationally demanding.



(a) A banana plot for a single day (July 2nd, 2010). x-axis: time (10 minute measurements), y-axis: particle sizes in log-scale (nm), z-axis (colour, brighter is higher): particle concentration in log scale (cm^{-3}).



(b) The user draws a line (in black) on the observed banana shape. Then, 1000 random lines are drawn on the plot (hidden from the user). The lines are sampled from the underlying grid and are constrained by domain knowledge (see text for details). The starting point of each line is denoted with a circle. The user selected line is deemed significant. The other significant lines are drawn in brown.

L	user	#412	#454	#455	#457	#518	#527	#549	#576	#767	#787	#840	#783
q_{75}	8.72	8.19	7.76	6.96	7.59	7.44	6.71	7.62	7.81	8.24	7.26	7.81	6.71
p_{raw}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
p_{bonf}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p_{minP}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10

(c) p-values (raw, Bonferroni adjusted and minP adjusted) for the 13 lines with the lowest minP adjusted p-values in Figure 13b. Note that any $p_i = 0.00$ is in fact $p_i = 1/(1000 + 1)$ (see definition of empirical p-value in Section 2.2).

Figure 13: An example of applying the framework on multivariate time series visualised as a heatmap. Test statistic: 25% quantile of z on a line. Null distribution: historical surrogates of non-event days.

6 Conclusions

This thesis presented a principled framework for evaluating the significance of visual patterns during exploration. The significance of a visual pattern is interpreted as how likely it is that this pattern would have occurred by chance, given that a certain distribution is assumed for the data. The visual pattern is quantified through a test statistic (e.g., the number of points inside a region) and the assumptions of the user about the data are encoded through the null distribution (e.g., the independence of certain attributes).

An empirical evaluation of the framework demonstrated how the knowledge of the user influences the significance of visual patterns (Section 5.1) and how the framework can be used in the analysis of both tabular (Sections 5.2, 5.3) and time series (Sections 5.4, 5.5, 5.6) data using different types of null distributions and test statistics. Furthermore, the significance of visual patterns was also evaluated during iterative data exploration (Section 5.3).

Data exploration and visualisation are important processes in data analysis. The insights gained from exploring the data have a significant impact on further analyses and modelling of the data, e.g., through machine learning algorithms. The proposed framework represents an important contribution in exploratory data analysis by making it possible to directly determine the significance of visually observed patterns during exploration.

Future Work A natural direction for future work is to study different visualisations, visual test statistics, null distributions and how to apply the framework in practice. Furthermore, the effect of various parameters in the framework was not studied, such as the number of required surrogates for a given confidence level and the internal parameters of test statistics and null distributions.

Since visual patterns are tied to human visual perception, the visual test statistics should be motivated by human perception. A detailed study of the human perception of visual patterns is outside of the scope of this thesis and is an interesting future direction. For example, which test statistics best encapsulate a specific visual pattern? Is the number of points in a region a good descriptor of a cluster observed by a data analyst? One approach for studying whether a test statistic corresponds to a visual pattern that is perceived by an analyst is through the use of Amazon's crowdsourcing tool Mechanical Turk, which has been previously used to study the line-up protocol [51].

The statistical validity of the framework is not thoroughly discussed in this thesis. There are implicit assumptions about the data that are not stated in each case, which can be addressed in future work. Furthermore, the use of p-values in this thesis resembles the traditional null hypothesis testing procedure that results in a binary decision of significance or non-significance. The framework may be adapted to instead interpret p-values as a continuous indicator of whether the data are described by a model that corresponds to the user's assumptions about the data.

Since the framework is intended for interactive use by data analysts, future work can study how it can be incorporated into data analysis workflows of specific data

types, such as time series and networks, and for specific applications, such as the banana plot example in Section 5.6. The current form of the proposed framework can be described as a Do-It-Yourself (DIY) statistical testing procedure for visual patterns, since the weight of the burden is on the user of the framework. The user specifies the visualisation, the test statistic and the null distribution and these choices depend on a number of factors that are application-specific and non-trivial. This presents a challenge, since many data analysts are not necessarily well-versed in statistical theory. A solution may be provided by an automation of the workflow into a software or an automated machine learning (AutoML) framework which selects parameters without user intervention, and which includes application-specific considerations and a suitable user interface. In any case, user studies should be used to evaluate the workflow of the framework, since the task of data analysis itself is tied to human perception.

References

[1] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham, "Statistical inference for exploratory data analysis and model diagnostics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, pp. 4361–4383, oct 2009.

[2] C. H. Yu, "Exploratory data analysis in the context of data mining and resampling," *International Journal of Psychological Research*, vol. 3, p. 9, jun 2010.

[3] F. Hartwig, *Exploratory Data Analysis*. SAGE Publications, Inc, 1980.

[4] J. Luan, "Data mining and its applications in higher education," *New Directions for Institutional Research*, vol. 2002, no. 113, pp. 17–36, 2002.

[5] J. T. Behrens, "Principles and procedures of exploratory data analysis," *Psychological Methods*, vol. 2, no. 2, pp. 131–160, 1997.

[6] W. J. McGuire, "A perspectivist approach to the strategic planning of programmatic scientific research," in *Psychology of science* (B. Gholson, W. R. S. Jr., R. A. Neimeyer, and A. C. Houts, eds.), pp. 214–245, Cambridge University Press, 1989.

[7] J. W. Tukey, "Data analysis, computation and mathematics," *Quarterly of Applied Mathematics*, vol. 30, pp. 51–65, apr 1972.

[8] R. Savvides, A. Henelius, E. Oikarinen, and K. Puolamäki, "Significance of patterns in data visualisations," *KDD*, 2019.

[9] J. Lenhard, "Models and statistical inference: The controversy between Fisher and Neyman–Pearson," *The British Journal for the Philosophy of Science*, vol. 57, pp. 69–91, jan 2006.

[10] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. Springer New York, 2005.

[11] R. A. Fisher, *The design of experiments*. Oliver & Boyd, 1935.

[12] A. Agresti, *An Introduction to Categorical Data Analysis*. John Wiley & Sons Hoboken, 2007.

[13] B. North, D. Curtis, and P. Sham, "A note on the calculation of empirical p values from Monte Carlo procedures," *The American Journal of Human Genetics*, vol. 71, pp. 439–441, aug 2002.

[14] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, 1984.

[15] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Human Brain Mapping*, vol. 15, pp. 1–25, nov 2001.

[16] R. L. Wasserstein and N. A. Lazar, "The ASA statement on p-values: Context, process, and purpose," *The American Statistician*, vol. 70, pp. 129–133, apr 2016.

[17] C. Woolston, "Psychology journal bans p-values," *Nature*, vol. 519, pp. 9–9, feb 2015.

[18] H. Pashler and E. Wagenmakers, "Editors' introduction to the special section on replicability in psychological science," *Perspectives on Psychological Science*, vol. 7, pp. 528–530, nov 2012.

[19] G. Gigerenzer, "Mindless statistics," *The Journal of Socio-Economics*, vol. 33, pp. 587–606, nov 2004.

[20] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 231, pp. 289–337, jan 1933.

[21] R. A. Fisher, "Statistical methods for research workers," Oliver and Boyd, 1925.

[22] M. Krawczyk, "The search for significance: A few peculiarities in the distribution of P values in experimental psychology literature," *PLOS ONE*, vol. 10, p. e0127872, jun 2015.

[23] S. Goodman, "A dirty dozen: Twelve p-value misconceptions," *Seminars in Hematology*, vol. 45, pp. 135–140, jul 2008.

[24] S. Dudoit, J. P. Shaffer, and J. C. Block, "Multiple hypothesis testing in microarray experiments," *Statistical Science*, vol. 18, pp. 71–103, feb 2003.

[25] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc.*, vol. 57, pp. 289–300, 1995.

[26] P. Westfall and S. Young, *Resampling-based multiple testing: Examples and methods*. Wiley, 1993.

[27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, p. 65–70, 1979.

[28] A. Dmitrienko, F. Bretz, P. H. Westfall, J. Troendle, B. L. Wiens, A. C. Tamhane, and J. C. Hsu, "Multiple testing methodology," in *Multiple testing problems in pharmaceutical statistics*, pp. 53–116, Chapman and Hall/CRC, 2009.

[29] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[30] A. Gelman, "Exploratory data analysis for complex models," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 755–779, 2004.

[31] J. Thomas and K. Cook, eds., *Illuminating the Path: A visual analytics agenda*, vol. 26, ch. Visual Representations and Interactions Technologies, pp. 69–104. IEEE Press, 2006.

[32] F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[33] S. Chatterjee and A. Firat, "Generating data with identical statistics but dissimilar graphics," *The American Statistician*, vol. 61, no. 3, pp. 248–254, 2007.

[34] J. Matejka and G. Fitzmaurice, "Same stats, different graphs," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, ACM Press, 2017.

[35] W. S. Cleveland, *Visualizing Data*. Hobart Press, 1993.

[36] E. R. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[37] E. R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.

[38] W. Robison, "Representation and misrepresentation: Tufte and the Morton Thiokol engineers on the Challenger," *Science and Engineering Ethics*, vol. 8, no. 1, pp. 59–81, 2002.

[39] C. R. J. Charles D. Hansen, *Visualization Handbook*. Elsevier Science, 2011.

[40] R. A. Earnshaw, *Illuminating the Path: A Research and Development Agenda for Visual Analytics*, ch. Visual Representations and Interaction Technologies, pp. 69–104. 2005.

[41] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castrén, "Trustworthiness and metrics in visualizing similarity of gene expression," *BMC Bioinformatics*, vol. 4, no. 1, p. 48, 2003.

[42] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, vol. 24, no. 7, pp. 498–520, 1933.

[43] J. Friedman and J. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers*, vol. C-23, no. 9, pp. 881–890, 1974.

[44] T. D. Bie, "Subjective interestingness in exploratory data mining," in *Advances in Intelligent Data Analysis XII*, pp. 19–31, Springer Berlin Heidelberg, 2013.

[45] B. Kang, K. Puolamaki, J. Lijffijt, and T. D. Bie, "A constrained randomization approach to interactive visual data exploration with subjective feedback," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.

[46] K. Puolamäki, E. Oikarinen, B. Kang, J. Lijffijt, and T. D. Bie, "Interactive visual data exploration with subjective feedback: An information-theoretic approach," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, IEEE, apr 2018.

[47] K. Puolamäki, B. Kang, J. Lijffijt, and T. D. Bie, "Interactive visual data exploration with subjective feedback," in *Machine Learning and Knowledge Discovery in Databases*, pp. 214–229, Springer International Publishing, 2016.

[48] D. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Tenth International Conference on Information Visualisation (IV'06)*, IEEE.

[49] K. Puolamäki, P. Papapetrou, and J. Lijffijt, "Visually controllable data mining methods," in *2010 IEEE International Conference on Data Mining Workshops*, IEEE, dec 2010.

[50] H. Wickham, D. Cook, H. Hofmann, and A. Buja, "Graphical inference for infovis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 973–979, nov 2010.

[51] M. Majumder, H. Hofmann, and D. Cook, "Validation of visual statistical inference, applied to linear models," *Journal of the American Statistical Association*, vol. 108, pp. 942–956, sep 2013.

[52] H. M. Widen, J. B. Elsner, S. Pau, and C. K. Uejio, "Graphical inference in geographical research," *Geographical Analysis*, vol. 48, no. 2, pp. 115–131, 2015.

[53] P. C. Wong and J. Thomas, "Visual analytics," *IEEE Computer Graphics and Applications*, vol. 24, pp. 20–21, sep 2004.

[54] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Lecture Notes in Computer Science*, pp. 76–90, Springer Berlin Heidelberg, 2008.

[55] G. E. F. M. Daniel Keim, Jörn Kohlhammer, ed., *Mastering the Information Age Solving Problems with Visual Analytics*. Eurographics Association, 2010.

[56] A. M. MacEachren, "Visual analytics and uncertainty: It's not about the data," 2015.

[57] K. Potter, J. Kniss, R. Riesenfeld, and C. Johnson, "Visualizing summary statistics and uncertainty," *Computer Graphics Forum*, vol. 29, pp. 823–832, aug 2010.

[58] J. Stahnke, M. Dork, B. Muller, and A. Thom, "Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 629–638, jan 2016.

[59] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 240–249, jan 2016.

[60] M. de Ridder, K. Klein, and J. Kim, "A review and outlook on visual analytics for uncertainties in functional magnetic resonance imaging," *Brain Informatics*, vol. 5, jul 2018.

[61] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska, "Investigating the effect of the multiple comparisons problem in visual analysis," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, ACM Press, 2018.

[62] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[63] R. Menjoge, *New procedures for visualizing data and diagnosing regression models*. PhD thesis, Massachusetts Institute of Technology, 2010.

[64] N. R. Chowdhury, D. Cook, H. Hofmann, M. Majumder, E.-K. Lee, and A. L. Toth, "Using visual statistical inference to better understand random class separations in high dimension, low sample size data," *Computational Statistics*, vol. 30, pp. 293–316, nov 2014.

[65] N. R. Chowdhury, D. Cook, H. Hofmann, and M. Majumder, "Measuring lineup difficulty by matching distance metrics with subject choices in crowd-sourced data," *Journal of Computational and Graphical Statistics*, vol. 27, pp. 132–145, jan 2018.

[66] D. J. Hand, *Principles of Data Mining*. MIT Press, 2001.

[67] L. Geng and H. J. Hamilton, "Interestingness measures for data mining," *ACM Computing Surveys*, vol. 38, pp. 9–es, sep 2006.

[68] H. Yao, H. Hamilton, and L. Geng, "A unified framework for utility based measures for mining itemsets," *Second International Workshop on Utility-Based Data Mining*, 01 2006.

[69] S. Hanhijärvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila, "Tell me something i don't know: randomization strategies for iterative data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 09*, ACM Press, 2009.

[70] J. Lijffijt, P. Papapetrou, and K. Puolamäki, "A statistical significance testing approach to mining the most informative set of patterns," *Data Mining and Knowledge Discovery*, vol. 28, pp. 238–263, dec 2014.

[71] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas, "Assessing data mining results via swap randomization," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, pp. 14–es, dec 2007.

[72] R. Bolton and D. Hand, "Significance tests for patterns in continuous data," in *Proceedings 2001 IEEE International Conference on Data Mining*, IEEE Comput. Soc.

[73] Moise, Gabriela; Sander, Joerg, "Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, University of Alberta Libraries, 2008.

[74] W. Hämäläinen and G. I. Webb, "A Tutorial on Statistically Sound Pattern Discovery," *arXiv e-prints*, p. arXiv:1709.03904, Sept. 2017.

[75] G. I. Webb, "Discovering significant patterns," *Machine Learning*, vol. 68, pp. 1–33, apr 2007.

[76] G. I. Webb, "Preliminary investigations into statistically valid exploratory rule discovery," in *Proc. of the Australasian Data Mining Workshop (AusDM03), Canberra, Australia*, 2003.

[77] G. Liu, H. Zhang, and L. Wong, "Controlling false positives in association rule mining," *Proceedings of the VLDB Endowment*, vol. 5, pp. 145–156, oct 2011.

[78] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, IEEE, 2005.

[79] J. W. Tukey and P. Tukey, "Computer graphics and exploratory data analysis: An introduction," in *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics85* (N. C. G. Association, ed.), 1985.

[80] L. Wilkinson and G. Wills, "Scagnostics distributions," *Journal of Computational and Graphical Statistics*, vol. 17, pp. 473–491, jun 2008.

[81] A. Anand, *Visual Pattern Detection in High-dimensional Spaces*. PhD thesis, University of Illinois at Chicago, Chicago, IL, USA, 2012. AAI3551364.

[82] L. Wilkinson, A. Anand, and R. Grossman, "High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 1363–1372, nov 2006.

[83] T. N. Dang and L. Wilkinson, "ScagExplorer: Exploring scatterplots by their scagnostics," in *2014 IEEE Pacific Visualization Symposium*, IEEE, mar 2014.

[84] A. Anand and J. Talbot, "Automatic selection of partitioning variables for small multiple displays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 669–677, jan 2016.

[85] A. Anand, L. Wilkinson, and T. N. Dang, "Visual pattern discovery using random projections," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, oct 2012.

[86] T. N. Dang, A. Anand, and L. Wilkinson, "TimeSeer: Scagnostics for high-dimensional time series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 470–483, mar 2013.

[87] T. N. Dang and L. Wilkinson, "Transforming scagnostics to reveal hidden features," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, pp. 1624–1632, dec 2014.

[88] L. Shao, T. Schleicher, M. Behrisch, T. Schreck, I. Sipiran, and D. A. Keim, "Guiding the exploration of scatter plot data using motif-based interest measures," in *2015 Big Data Visual Analytics (BDVA)*, IEEE, sep 2015.

[89] M. Correll, *Improving Visual Statistics*. PhD thesis, University of Wisconsin-Madison, 1210 W. Dayton, Madison, WI., aug 2015.

[90] A. Unwin, "If you can't see the pattern, is it there?," in *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*, pp. 63–76, 2002.

[91] T. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 164–181, feb 2011.

[92] J. Lin, E. Keogh, S. Lonardi, and P. Patel, "Finding motifs in time series," in *Proc. of the 2nd Workshop on Temporal Data Mining*, pp. 53–68, 2002.

[93] L. Ye and E. Keogh, "Time series shapelets," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 09*, ACM Press, 2009.

[94] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: The method of surrogate data," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1–4, pp. 77–94, 1992.

[95] J. Theiler and D. Prichard, "Constrained-realization Monte-Carlo method for hypothesis testing," *Physica D: Nonlinear Phenomena*, vol. 94, pp. 221–235, jul 1996.

[96] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D: Nonlinear Phenomena*, vol. 142, pp. 346–382, aug 2000.

[97] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[98] K. Puolamäki, E. Oikarinen, and A. Henelius, "Guided visual exploration of relations in data sets," *CoRR*, vol. abs/1905.02515, 2019.

[99] D. Kugiumtzis, "Surrogate data test on time series," in *Modelling and Forecasting Financial Data*, pp. 267–282, Springer US, 2002.

[100] G. Lancaster, D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska, "Surrogate data for hypothesis testing of physical systems," *Physics Reports*, vol. 748, pp. 1–60, jul 2018.

[101] D. Dua and C. Graff, "UCI machine learning repository," 2019.

[102] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750–757, 2008.

[103] C. R. Genovese, K. Roeder, and L. Wasserman, "False discovery control with p-value weighting," *Biometrika*, vol. 93, no. 3, pp. 509–524, 2006.

[104] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[105] M. Boley, M. Mampaey, B. Kang, P. Tokmakov, and S. Wrobel, "One click mining—interactive local pattern discovery through implicit preference and performance learning," in *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA)*, pp. 27–35, ACM, 2013.

[106] B. Kang, J. Lijffijt, R. Santos-Rodríguez, and T. De Bie, "Subjectively interesting component analysis: Data projections that contrast with prior expectations," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1615–1624, ACM, 2016.

[107] H. Junninen, A. Lauri, P. Keronen, P. P. Aalto, V. Hiltunen, P. Hari, and M. Kulmala, "Smart-SMEAR: on-line data exploration and visualization tool for SMEAR stations," in *Boreal Environment Research*, vol. 14, 2009.

[108] M. Dal Maso, M. Kulmala, I. Riipinen, R. Wagner, T. Hussein, P. Aalto, and K. Lehtinen, "Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiälä, Finland," *Boreal Environment Research*, vol. 10, no. 5, pp. 323–336, 2005.

[109] A. Henelius, J. Korpela, and K. Puolamäki, "Explaining Interval Sequences by Randomization," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, vol. 8188 of *LNCS*, pp. 337–352, Springer, 2013.

A Appendix

Equation 13 from Section 5.1 is derived. The parameter k describes the probability of knowingly choosing the correct test statistic (representing the *knowledge* of the analyst). If the correct test statistic is not chosen (probability $1-k$), then the test statistic is chosen uniformly at random among all test statistics. The user may select multiple test statistics to increase the probability of choosing the correct one. Suppose that the number of chosen test statistics is m and that an analyst with expertise k wishes to guarantee that the probability of choosing x_1 correctly out of the n test statistics is at least β. Then the required number m^* of chosen test statistics to ensure success at probability at least β is derived as follows.
Suppose that:

$$\begin{aligned} A &= \{\text{knowingly choose } x_1\}, \\ B &= \{\text{randomly choose } x_1\} \\ C &= \{\text{choose } x_1\} = A \cup (B \mid \text{not } A) \end{aligned}$$

Then:

$$\begin{aligned} \Pr(A) &= k \\ \Pr(B) &= \frac{1}{n} \\ \Pr(C) &= \Pr[A \cup (B \mid \text{not } A)] = k + (1-k)\frac{1}{n} \end{aligned}$$

The analyst wishes to guarantee that the probability of choosing x_1 correctly out of the n test statistics is at least β:

$$\begin{aligned} \Pr(\text{at least one C in m trials}) &\geq \beta \\ 1 - \Pr(\text{no C in m trials}) &\geq \beta \\ 1 - [1 - \Pr(C)]^m &\geq \beta \\ 1 - [1 - k - (1-k)\frac{1}{n}]^m &\geq \beta \\ &\vdots \\ m &\geq \frac{\log(1-\beta)}{\log(1-\frac{1}{n}) + \log(1-k)} \end{aligned}$$

The minimum number of chosen test statistics to ensure that x_1 is chosen correctly with probability at least β is then:

$$m^* = \lceil \frac{\log(1-\beta)}{\log(1-\frac{1}{n}) + \log(1-k)} \rceil$$