Title	A modified depth function for outlier detection in multivariate data with applications
Author(s)	Abedin, Md Jaynal
Publication Date	2020-09-24
Publisher	NUI Galway
Item record	http://hdl.handle.net/10379/16188

Doctoral Thesis

A Modified Depth Function for Outlier Detection in Multivariate Data with Applications

Md Jaynal ABEDIN

September 22, 2020

External Examiner	Supervisor	Internal Examiner
Dr. Paulo J. G. Lisboa	Prof. John Newell	Dr. Andrew Simpkin

Insight Centre for Data Analytics
College of Science and Engineering, National University of Ireland, Galway

ABSTRACT

Data Science is the new and exciting interdisciplinary response that has emerged as a consequence of the staggering amounts of data generated in many new forms from digital images to audio to text. It is an interdisciplinary field involving Statistics, Computer Science and Mathematics. It involves the study of data, how they are collected, stored, accessed, visualised, modelled and ultimately used to inform decision making by turning data into intelligence.

Despite this 'data revolution' and the development of Data Science as a consequence, the aim of any data analysis is still the same, to make inference about unknown population parameters using sample statistics. One fundamental challenge in inference is the identification of outliers.

Such oddities, or atypical observations, could be indicative of poor data management or biased sampling. In this situation the presence of such outliers are considered a negative aspect and efforts are needed to account for them (e.g. correct data entry errors) accordingly to avoid introducing bias in parameter estimation. On the other hand, finding an outlier may be the key focus of the exercise as an outlier may represent something new and novel.

Many statistical methods have been developed to identify outlying data points and robust methods developed to account for outliers in statistical models. A central property of all such methods is that an observation is classified as an outlier or not (i.e. a binary decision); being able to quantify an observations 'outlyingness' is clearly an attractive alternative.

In this thesis, a novel method is presented for outlier detection in multivariate data based on the idea of a statistical depth function. The proposed approach enables outlier detection in multivariate data while taking into consideration the local geometry of the underlying probability distribution.

CONTENTS

DECLARATION

I declare that this thesis, titled "*A Modified Depth Function for Outlier Detection in Multivariate Data with Applications*", is composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Galway, September 22, 2020

Md Jaynal ABEDIN

ACKNOWLEDGEMENTS

This work was supported by many people over the last couple of years to whom I would like to express my deepest gratitude and sincerest appreciation. First of all, I would like to thank Prof. John Newell for giving me the opportunity to work on the fascinating research topics presented in the thesis. I would also like to thank him for his continuous guidance throughout the journey with his ever smiling friendly gesture. Prof. Newell, not only an academic mentor, is a person from whom I am continuously learning many things that affect academic and personal life in a positive way. I would like to thank him for all his support. I would also like to thank Prof. Paulo J. G. Lisboa and Dr. Andrew Simpkin for very useful contributions as PhD examiners.

I would like to thank my former co-supervisor Prof. Dietrich Rebholz-Schuhmann who guided me at the beginning of my PhD journey. I am also grateful to the faculty members of the School of Mathematics, Statistics and Applied Mathematics at the National University of Ireland Galway (NUI Galway) and the members of the Insight Centre for Data Analytics (now known as the Data Science Institute) for their feedback in various weekly meetings. I would like to express sincere gratitude to my Graduate Research Committee members; Prof. Cathal Seoighe, Prof. James Duggan and Dr. Colm O'Riordan for their continuous assessment and constructive feedback throughout the journey.

I gratefully acknowledge and appreciate Science Foundation Ireland (SFI) and Orreco for their funding to support my work. I would like to extend my thanks to all my colleagues at The Insight Centre for Data Analytics for their friendly support. I am also grateful to the administrative team members of the Insight Centre for Data Analytics and NUI Galway who supported my work in various ways.

At this point I would like to take the opportunity to thank the local community of Galway for their friendly attitude towards the international student community. Indirectly, they also supported my work so, I would like to thank you all.

I would like to extend and express my gratitude towards my former colleagues in Bangladesh who constantly inspired me to do something good for the academic community as well as for society as a whole. I like to thank specially Dr. Faisal Zaman who forward the circular and encouraged me to apply for this PhD offering and Dr. Yushuf Sharker who encouraged and inspired me every single day.

It is beyond words to express my parent's contribution. Throughout the time since my birth their unconditional love and never ending mental and emotional support have made everything possible. They never gave up during crisis moments, without their support I would not be here. Whatever I say in expressing my gratitude towards them is not enough. I would like to add my sincere thanks to my sisters, their families and extended family for their unconditional support.

I would like to sincerely thank my wife “Mita”; the person who has sacrificed a lot for my success to be realised and for her unconditional love and support for me in every possible way.

In the last but definitely not the least, I want to thank my friends for their never ending inspiration and emotional support. I don’t want to mention all of your names individually but I would like to use a word to represent you, the word is "Prithibi"; a Bangla word, which means "the World". Yes, you were there and I strongly feel that you will always be there to support me emotionally and morally, to believe in me and motivate me to do something good for society. Thank you "Prithibi" for supporting me and being beside me on my journey.

Finally, I would like to dedicate this thesis to Professor Mohammed Nasser, Department of Statistics, University of Rajshahi, Bangladesh; the person is a lifetime mentor for me. Prof. Nasser passed away in 2016 but he was, is and will be a true inspiration to me every single day.

——————————————————————————————-Jaynal Abedin

1 INTRODUCTION

1.1 INTRODUCTION

Data Science is the new and exciting interdisciplinary response that has emerged as a consequence of the staggering amounts of data generated in many new forms from digital images to audio to text. It is an interdisciplinary field involving Statistics, Computer Science and Mathematics. It involves the study of data, how they are collected, stored, accessed, visualised, modelled and ultimately used to inform decision making by turning data into intelligence.

Despite this 'data revolution' and the development of Data Science as a consequence the fundamental challenges in data analysis still remain. One such challenge is outlier detection.

Identifying outliers in data is a fundamental component of all data analyses. Since the development of the discipline of Statistics numerous methods have been presented to detect and visualise outliers from a univariate and multivariate perspective and statistical models developed that are robust to outliers.

In this thesis outlier detection is revisited with an overall goal to evaluate, through simulation, a newly proposed statistical algorithm to identify outliers in multivariate data. Methods to use the proposed algorithms in data visualisation are given by embedding the outlying score in classical data visualisations so that potential outlying points can be identified visually in a novel and an intuitive manner.

In addition a new and novel application of Data Science, involving topic modelling and cluster analysis, is presented in the form of a tool to aid in literature reviews when the number of published article is vast.

To conclude, a case study in Data Science is given where modern computational techniques in Statistics and Computer Science are used to predict severity of Knee Osteoarthritis.
In the next few sections the chapter specific objectives and new contributions to the field of Data Science developed in this thesis are summarised.

1.1.1 Chapter-2: Data Science Approach to Literature Analysis

A literature review is a comprehensive summary of previous published research on a topic to i) give a theoretical base for new research, ii) acknowledge the work of previous researchers in the domain and iii) summarise the developments in the field to date. With ongoing advancements in technology, publication rates continue to increase rapidly creating challenges for completing ef-

ficient reviews. Due to the large number of papers in a particular domain it is almost impossible to manually review all available published papers on a certain area [BGC10; Cha03]. A semi-automated approach has the potential to overcome the challenges [ARO+09].The use of such a semi-automated approach to analyze literature search results gives more opportunity to perform the analysis faster and produce more evidence by being able to summarise a large number of titles and abstracts in a time efficient manner [ARO+09].

In this chapter, the difficulties of manual reviews of a large number of scientific publications are presented to summarise a large collection of abstracts and to identify latent research themes. To overcome these, a new workflow is presented which will augment the traditional literature review process. The proposed workflow consists of steps of data collection; key words searches from indexed electronic databases (e.g. PubMed, ScienceDirect, Web of Science etc.), data cleaning and tidying followed by topic modelling.

The proposed workflow is described using three different examples from three different application domains. The proposed workflow can be used in any research domain where the objective is the synthesize existing body of knowledge and identify latent research themes from a large collection of available published studies.

Using the steps of the proposed workflow an open source application (shiny app) was developed, so that it can be used by any non-technical user to augment their literature review. The user will be able to interact with the collection of abstracts to uncover underlying latent research themes and visualise them in a meaningful way. Moreover, users can identify if there are any clusters of note among the abstracts.

The proposed workflow, along with the open source shiny app, will be a companion tool for literature reviews of large collections of published studies that is not feasible to do manually. For a large number of published studies performing an exhaustive manual review is nearly impossible but the proposed workflow along with the shiny app will play an important role in mitigating the difficulties of manual reviews.

1.1.2 Chapter-3: Outlier Detection in Multivariate Data

In the statistics community the term "outlier" was used as early as 1969 by Grubbs [Gru69]. Though the terminology has been used for many years there is still a lack of a unified definition of outliers [CBK09]. It is more challenging to define and detect outliers in multivariate data in comparison to univariate data. Having the ability to define an outlying score from multivariate data point is clearly attractive.

In this chapter the concept of a univariate 'outlyingness' score is discussed, along with limitation of existing methods. One such idea, introduced by Tukey [Tuk75], was a statistical depth function, an extension of order statistics in multivariate data. A new algorithmic approach, a modified Mahalanobis depth function, that overcomes previous limitations is presented in this thesis, along with an evaluation of the proposed approach using simulation experiments. The proposed approach is evaluated using benchmark and real-world datasets.

The primary goal of the proposed modified Mahalanobis depth function is to define an outlyingness score for each data point to numerically and visually identify potential outliers for further investigation. The notion of 'outlyingness' is presented as a continuous scores rather than using a cutoff to use it as a binary indicator variable. The continuous score represents the degree of outlyingness compared to the centre of the distribution or compared to a nearest neighbour if the objective is the identify local outlying points.

1.1.3 Chapter-4: Visualising Multivariate Data

To present meaningful insight from raw data, visual representations are more useful than numerical summaries. In this chapter, various ways of visualising multivariate data are presented along with their limitations in visualising potential outlying points.

Classical approaches for data visualisation of multivariate data are presented using a case study involving motion tracking data in elite socce where traditional statistical graphs are augmented through the inclusion of the proposed modified Mahalanobis depth function introduced in Chapter 3. The new contribution of this chapter is the ability to visualise potential outlying points from a multivariate data using any type of classical visualisation by incorporating the modified Mahalanobis depth function.

1.1.4 Chapter-5: Predicting the Severity of Knee Osteoarthritis: A Data Science Case Study

This chapter could be considered an outlier as the focus is not on outlier detection rather a showcase of modern methods of data science in the area of knee injury through a collaborative project across the Insight Centre for Data Analytics.
In this project, a statistical model will be presented to predict severity of Knee Osteoarthritis based on patient symptoms and other characteristics as an alternative to a model developed using knee X-ray images.

1.2 NOVELTY

The novel contributions of this thesis are:

- A new workflow to perform the analysis of a collection of articles from the scientific literature to uncover latent research themes
- A web tool developed for the workflow to analyse the scientific literature to uncover latent research themes and visualise the findings
- A novel approach using a modified Mahalanobis depth function in multivariate data while taking into consideration the local geometry of the underlying probability distribution
- The use of an outlyingness score to visualise multivariate data to indicate potential candidate outliers in classical visualisation techniques such as the scatter plot.

1.3 SUMMARY

The aim of this thesis is to develop new methods for outlier detection in multivariate data that are relevant and useful in Data Science. As part of the journey of discovery classical methods in Statistics have been revisited, adapted and merged with methods in Computer Science to create useful tools for outlier detection, knowledge discovery in literature reviews and aid in the treatment of Knee Osteoarthritis.

2 DATA SCIENCE APPROACH TO LITERATURE ANALYSIS

2.1 INTRODUCTION

A literature review is a comprehensive summary of previous published research on a topic to i) give a theoretical base for new research, ii) acknowledge the work of previous researchers in the domain and iii) summarise the developments in the field to date. It is assumed that the author has read, evaluated, and critiqued the work in the context of the arguments they present.

The collection of published research in popular domains could form a large corpus of text documents and manually collating and summarising those is time consuming and not efficient. As the number of online scientific journals and corresponding publication rates continue to increase rapidly, it is almost impossible to complete an exhaustive and efficient traditional literature review [BGC10] and [Cha03]; the use of semi-automated approaches such as text mining and topic modelling have the potential to address these challenges [ARO+09].

In this chapter a new approach will be presented (and a shiny application showcased) which uses text mining and topic modelling to augment a 'classical' literature review allowing an automated search across all relevant articles to identify

- the most prevalent research themes/sub-themes in a specific application domain of interest
- new research themes that are emerging
- the evolution of a specific research theme over time

A workflow is presented which consists of the steps of data collection (through keyword searches in electronic databases), cleaning and tiding followed by text mining tool to extract research themes.

The steps presented are as follows:

- Data collection
- Data cleaning
- Exploratory analysis
- Text mining and topic modelling
- Visualisation and communication

An example will be given on the topic of outlier detection from a statistical perspective followed by an example relating to injuries in elite soccer. The tool presented in this chapter could be used in any domain where the objective is to summarise previous published research with an aim to identify common and emerging research themes.

In each case the target population consists of all scientific papers published in English over time on a specific domain of interest.

2.2 TEXT MINING & TOPIC MODELLING

A classical statistical analysis involves the application of statistical models to make inference on unknown population parameters using a random sample from the study population of interest. Making inference from unstructured text data requires an additional step as the raw text data need to be transformed into structured form suitable for analysis. The analysis of such data is often referred to as text mining or text data mining [FD95; Hea97].

Text mining is the umbrella term used in Computer Science for natural language processing (NLP); the process of deriving high-quality information from plain text. Text mining typically involves finding patterns of term/word use within a collection of documents (i.e. a corpus), an analysis of term frequency provides information on the overall theme of a document. Approaches and applications of text mining include, but are not limited to, document clustering[CTL17], opinion mining and sentiment analysis[PL+08], identifying protein interactions[PPT+15; SMC+16] and protein disease association[LMS+18].

Tan [Tan+99] presented a framework for text mining that consisted of two phases; 1) text refinement where the unstructured text are transformed into an intermediate form typically known as a Document-Term-Matrix and then 2) knowledge extraction phase from the intermediate form of the text documents. Based on the type of intermediate form, the knowledge extraction phase differs; the document based intermediate form leads to document clustering, categorisation and visualisation whereas, content-based intermediate forms lead to predictive modelling, associative discovery and visualisation.

The aim of Topic modelling is to discover latent themes from a large number of documents (i.e. a corpus) using suitable statistical models. Each document is considered to contain a mixture of topics representing latent themes, governed by term frequency. Topic modelling was first described in 1998 by Papadimitriou, Raghavan, Tamaki and Vempala[PRT+00]. In 1999, Thomas Hofmann described topic modelling as Probabilistic Latent Semantic Analysis (PLSA)[Hof99], generalised in 2003 as Latent Dirichlet Allocation (LDA)[BNJ03], becoming one of the more popular and most commonly used technique in topic modelling. Text mining & topic modelling have been used to find research trends in marketing[ACR+18], abbreviation identification[YH17], identification of studies by automatically expanding the search queries and document clustering to support systematic reviews[OTM+15; ARO+09] and personalisation and customization of research literature[SB12].

The novelty of this chapter is the use of topic modelling to analyse the content of the scientific Journal articles; especially abstracts to uncover latent research theme which was absent in the previous attempts [OTM+15; ARO+09]

In the next section the workflow for applying topic modelling and text mining in the analysis of text in literature reviews is outlined.

2.3 LITERATURE ANALYSIS WORKFLOW

A comprehensive workflow is presented in Figure 2.1, consisting of the steps needed to search and filter relevant literature from well known scientific databases and to pre-process data for modelling and visualisation.

2.3.1 Search and Retrieval

The top block of Figure 2.1 relates to the search & retrieval of relevant abstracts. The aim is to perform a search in well known scientific databases using a combination of key words relevant to the research domain of interest. Initially the download is driven by search results which contain the title of the paper, abstracts and other related information. Duplicate titles are then removed and a corpus of abstracts is created based on the inclusion criteria (e.g. abstracts in the English language only, original or systematic review papers and availability of abstracts).

2.3.2 Text Processing

The upper middle section of Figure 2.1 relates to text processing and the conversion of the raw unstructured text into a a Document-Term-Matrix: a more structured form suitable for statistical analysis.

The scientific abstracts are variable in length, containing various forms of the same words, punctuation, numerals and Unicode characters. In order to uncover latent research themes and how such themes evolve over time, a pre-processing step is needed to represent the unstructured text data (abstracts) in a structured layout i.e. each document represented a row and each word a column. Initially all 'general words' are removed from the abstracts to reduce potential bias in subsequent analyses. The list of 'general words' included but not limited to: *introduction, background, aim, objective, method, study, result, conclusion, and discussion*. The "terms", or a combination of "terms", used for the search are also removed due to their low discriminatory power (as they appear in every abstracts in the corpus) between themes. All text is then converted to lowercase or uppercase to mitigate case-sensitivity of the "terms". Following this, all numerals, punctuation, Unicode characters, and common English stop-words e.g. auxiliary verbs, prepositions, conjunctions etc., are removed.

Due to grammatical structures employed, abstracts may contain the same "term" in different forms and "terms" may appear with similar meaning. To address this, these type of words are

reduced into common inflectional forms and derivationally related forms of "terms"; known as lemmatization [GBK+09; MCF+15; BG18]. A document-term matrix (DTM) is then created, where each row represents one document (an abstract here) and each column represents one "term" from the abstract and the entries of the matrix is the number of occurrence of the "terms" in a document. A seperate DTM is constructed using unigram (taking one "term" at a time) and bigram (taking two consecutive "terms" at a time considered as a single instance) approaches. (Figure 2.1).

Different words may not be equally informative in uncovering latent themes from a collection of abstracts where some "terms" might appear in only a handful of abstracts and are deemed less useful. To identify only meaningful "terms" that could aid in uncovering latent themes from a corpus, sparse "terms" are removed by calculating the "Term Frequency-Inverse Document Frequency" (TFIDF) representing how important a "term" is to a document in a corpus. TFIDF is defined as:

$$\mathrm{TFIDF}(i) = \frac{\mathrm{Frequency}(i)}{N_i} \times \log \frac{N}{N_{di}} \tag{2.1}$$

where $\mathrm{Frequency}(i)$ is the frequency of $i-th$ "term" in a document, N_i is the number of "terms" in the document (i.e. length of the document from which i-th "term" comes), N is the total number documents in the corpus, and N_{di} is the number of document with "term" i [RU11].

Once these steps are completed successfully the original unstructured data are ready for analysis.

2.3.3 Exploratory Analysis

In the exploratory analysis step the number of publications per year is calculated to assess the growth of relevant publications over time. An attempt to identify the forum/journal was made with the largest number of publications focused on the domain of interest. The top k single "term" (unigram) used with highest frequency of occurrence is identified followed by the top k two consecutive "terms" (bigrams). The unigram and bigram frequency is then used to provide insight on the overall research theme and its evolution over time.

2.3.4 Modelling & Visualisation

Latent Dirichlet Allocation (LDA)[BNJ03] is one of the popular and most commonly used techniques in topic modelling. LDA is a hierarchical Bayesian modelling approach which can learn from a set of latent topic based on the "terms" that occur together in a document. The primary idea of LDA is that the *documents* are represented as a random mixture of latent topics, where each topic is characterized by a distribution of "terms". There are two underlying assumptions:

- Exchangeability for *terms* in a document, that is the order of occurrence of a "term" does not affect the underlying latent topic. This is also known as the Bag-of-Words (BoW) assumption; the grammatical structure is completely ignored and the count of individual

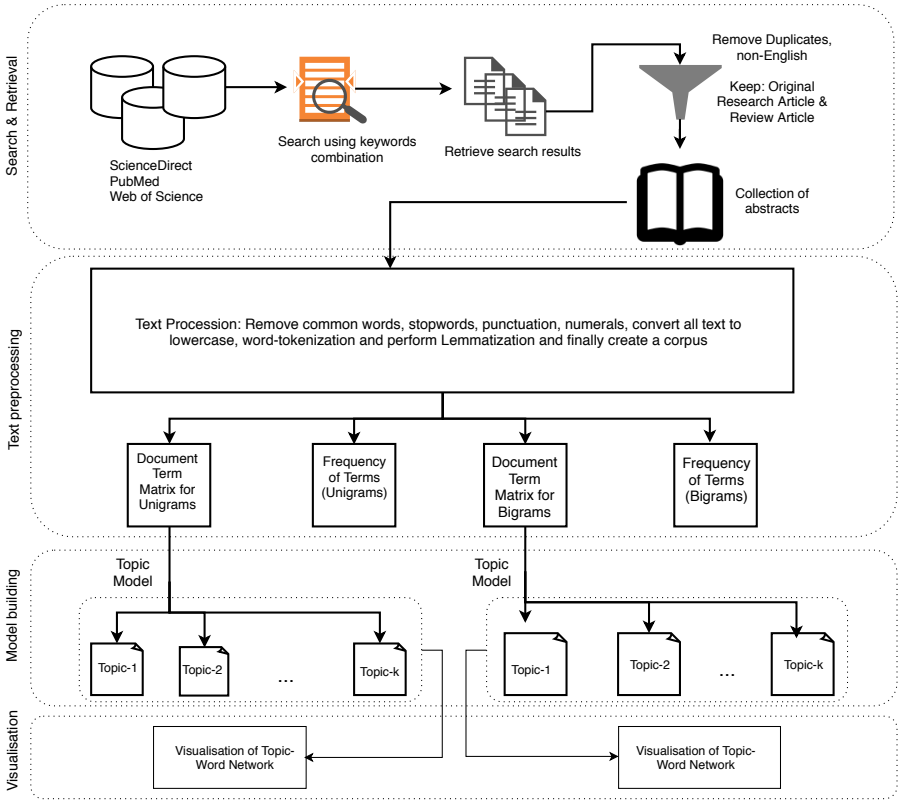


Figure 2.1: Schematic work-flow of literature search and topic modelling

"terms" are the only things that matter here. The order of the "terms" do not affect the underlying latent topic rather the co-occurrence in a document.

- Exchangeability for *documents* in a corpus, that is the order of documents do not affect the underlying latent topics. That is the temporality of the document does not affect the underlying latent topic.

The LDA model uses the observed conditional probability of a "term" for a given document ($P(term|document)$) to calculate the probability of a latent theme given a document ($P(topic|document)$) and also to calculate distribution of "terms" for each topic as ($P(term|topic)$).

In the LDA model the number of topics needs to be pre-specified. To decide on the optimal number of topics, a tuning step was performed using the *ldatuning* an R library. A range of potential values of the number of topic k was plotted against different metrics as suggested by Griffiths [GS04], Cao [CXL+09], Arun [ASM+10] and Deveaud [DSB14].

- **Griffiths2004** [GS04]: A range of values of the number of topic k is plotted against the likelihood of the data for a given model. In this case, $P(term|k)$ is plotted against k, where $P(term|k)$ is approximated by the harmonic mean of a set of values of $P(term|topic, k)$. The resultant curve initially shows an increasing pattern and then reaches a steady state and thereafter starts to decrease. The optimum value of k is the value for which the curve attains its maximum.

- **CaoJuan2009** [CXL+09]: Average cosine distance between a pair of topic is plotted against a range of values for the number of topic k. The value of k for which the average cosine distance is minimum is chosen as the optimal number of topics.

- **Arun2010** [ASM+10]: A symmetric KL-Divergence is calculated for each value of the number of topic k and plotted. The optimum value of k is the value for which the curve attains its minimum.

- **Deveaud2014** [DSB14]: A dissimilarity between pairs of topics is being calculated and plotted against the number of topic k. Here the information divergence between pairs of topics is being used as a measure of dissimilarity between topics. The optimum value of k is the value for which the dissimilarity between attains its maximum

The four different matrices could give different values of k which helps get an idea of the range of optimal number of topics to consider in LDA model fitting. A higher value of k will result in too granular a level of topics identified which might not be of interest. To keep the balance between topic granularity and information content, we have chosen the smallest value of k from the range obtained by using the four metrics.

Once the optimal number of topics k is determined, the LDA algorithm is run to extract the composition of topics and the distribution of "terms" within a topic based on the estimated value of ($\beta = P(topic|document)$). The top p "terms" are extracted from each of the k topics and then visualised in a bar chart. The distribution of "terms" in a topic are visualised in a series of bar charts where the size of the bar represents the probability of the "term" appearing in the

topic $P(\text{term}|\text{topic})$.

If all theses steps are carried out correctly, what started as a large corpus of documents that was prohibitively large to read, should now be represented graphically in a manner that informs on the most popular research themes present in the corpus including information on their change over time.

2.4 EXAMPLES

Three examples are now presented following the literature analysis workflow presented in Figure 2.1. The objective is to demonstrate how the workflow could be used on literature from any domain of interest to find latent themes from a large number of abstracts. The examples covers three different research domains, namely, a) outlier detection, b) statistical depth function and c) injuries in elite soccer.

2.4.1 Outlier Detection

In Statistics the terminology "outlier" was defined in the early 60s as *an observed value of a random variable that is deviated remarkably from the majority of the values of that random variable* [Gru69]. Since the introduction of the terminology it has been synonymous with atypical observation, extreme points, unusual observation, anomalous points; for a comprehensive overview see [BL74; Haw80; RL05; BC83; HA04; CBK09; GU16].

Though two terminologies "outlier" and "anomaly" are used interchangeably the corresponding application domains are often quite different. In this example, an analysis of the relevant literature is presented based on a keyword search in an electronic database.

Search & Selection

A combination of keywords as *"outlier detection OR anomaly detection"* was used to perform a search in *ScienceDirect;* an indexed scientific database. The search result was then filtered based on the following inclusion criteria:

- Language "English" only
- Original research paper or reviews or systematic reviews
- Availability of abstracts

A total of 1966 abstracts was included for further analysis.

Text Processing

Following the procedure outlined within the workflow Figure 2.1 the downloaded abstracts were cleaned and converted into a Document-Term-Matrix (DTM). In the processing steps, numerals, special characters, English stop words (such as, auxiliary verbs, prepositions etc.) were removed.

Moreover, the "general words" that appear in many abstracts were removed e.g. *"introduction", "methods", "method", "objective", "objectives", "results", "result", "conclusions", "conclusion", "study", "detection", "detect","data","datum", "analysis","program", "model", "anomaly", "outlier"*. The DTM contained 1966 documents representing rows of the DTM while 10651 unique "terms" representing the columns of the DTM.

Exploratory Analysis

The earliest result was 1975 with very few publications in the 1970s found by the search. The search result did not pick up an early paper where Grubbs [Gru69] gave the first definition of "outlier". This exclusion could be due to the search performed only in one electronic database; *ScienceDirect*. The number of publications increased in the 1990s and there was a sharp increase from the year 2004 onward. (Figure 2.2)

The name of the journal could provide useful information on the broader research domain in question. The frequency count of the top 10 journals was plotted (Figure 2.3). The journal names could be broadly categorized into two domains, a) statistical computing (covering broad are of computational statistics and data analysis) and b) computer science (covering computer networks and information science).

Similar to the journal name, the most frequently used "terms" in the corpus also informs on the underlying research domains. A count of top 50 "terms" was displayed as a bar chart (Figure 2.4). By looking at the most frequently used "terms" it can be assumed that those "terms" can be represented as two broad research domains, a) statistics and b) machine learning & computer science. To uncover the research themes in a more granular level, an LDA model was fitted using the DTM of the corpus.

modelling & Visualisation

To uncover latent research themes from the corpus of abstracts, an LDA model was fitted on the DTM where the number of topic needs to be pre-specifed before fitting the algorithm. To decide on the optimal number of topics, a tuning step was performed by plotting four different metrics (**Griffiths2004**, **CaoJuan2009**, **Arun2010**, **Deveaud2014**) against k; the number of topics, as mentioned in section 2.3.4, and choosing the lower limit of the range of probable values of k in order to maintain balance in granularity level while avoiding excessive topic aggregation (Figure 2.5). The dashed vertical line shows the possible range of values for the number of topics k. For this analysis $k = 40$ was chosen to fit the final LDA model.

The distribution of "terms" for each of the 40 topics extracted is presented in Figure 2.6, 2.7, 2.8 and 2.9. Based on the composition of the words within a topic, it can be interpreted subjectively that each of the topics represents a different application domain such as fraud detection in energy sector (topic-1), intrusion detection in a system (topic-2), statistical algorithms (topic-5) and so on.

The over count of "terms" in Figure 2.4 indicates that there appears to be two major research domains emerging with several sub domains present in the corpus.

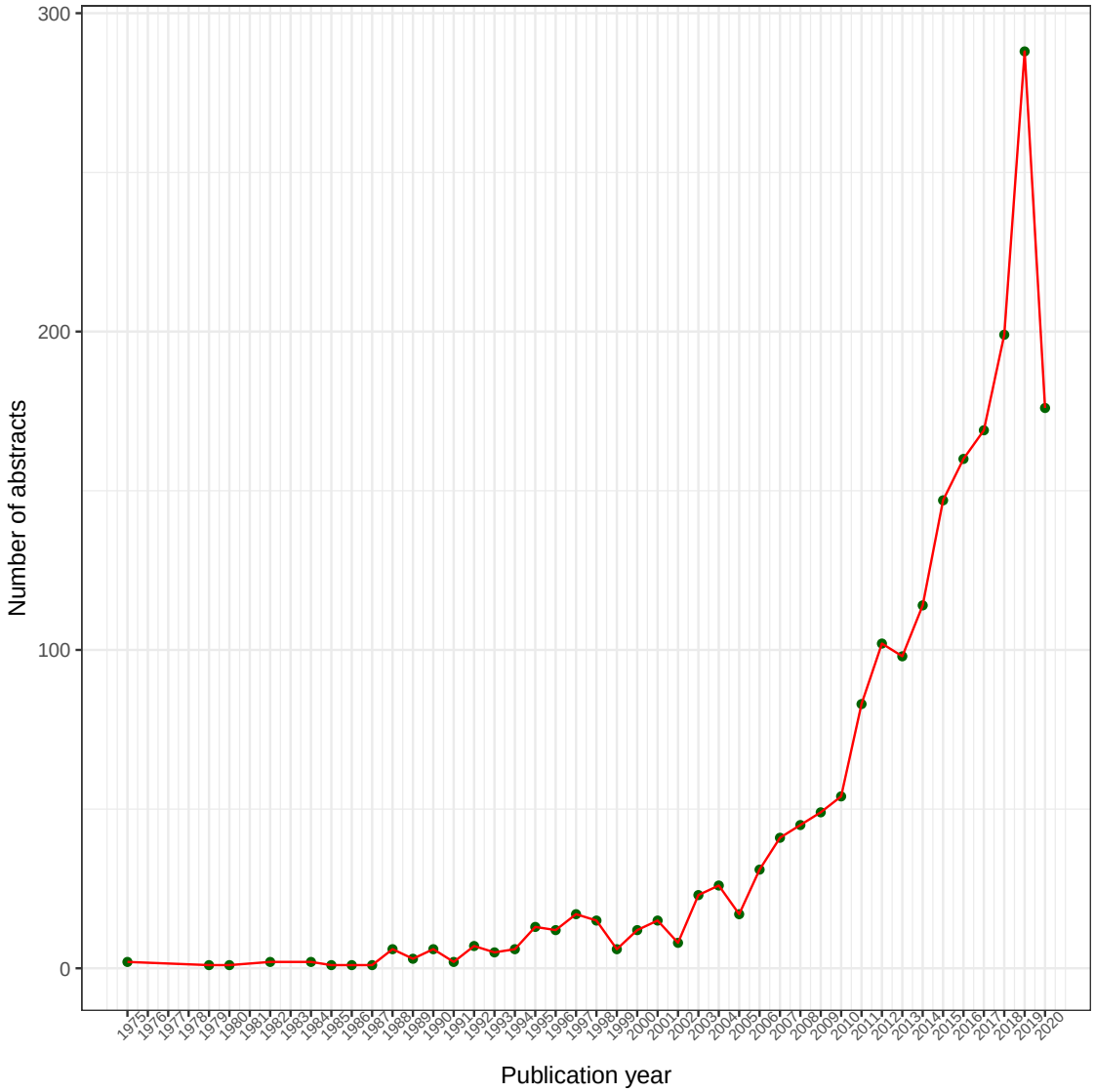


Figure 2.2: Number of abstract per year

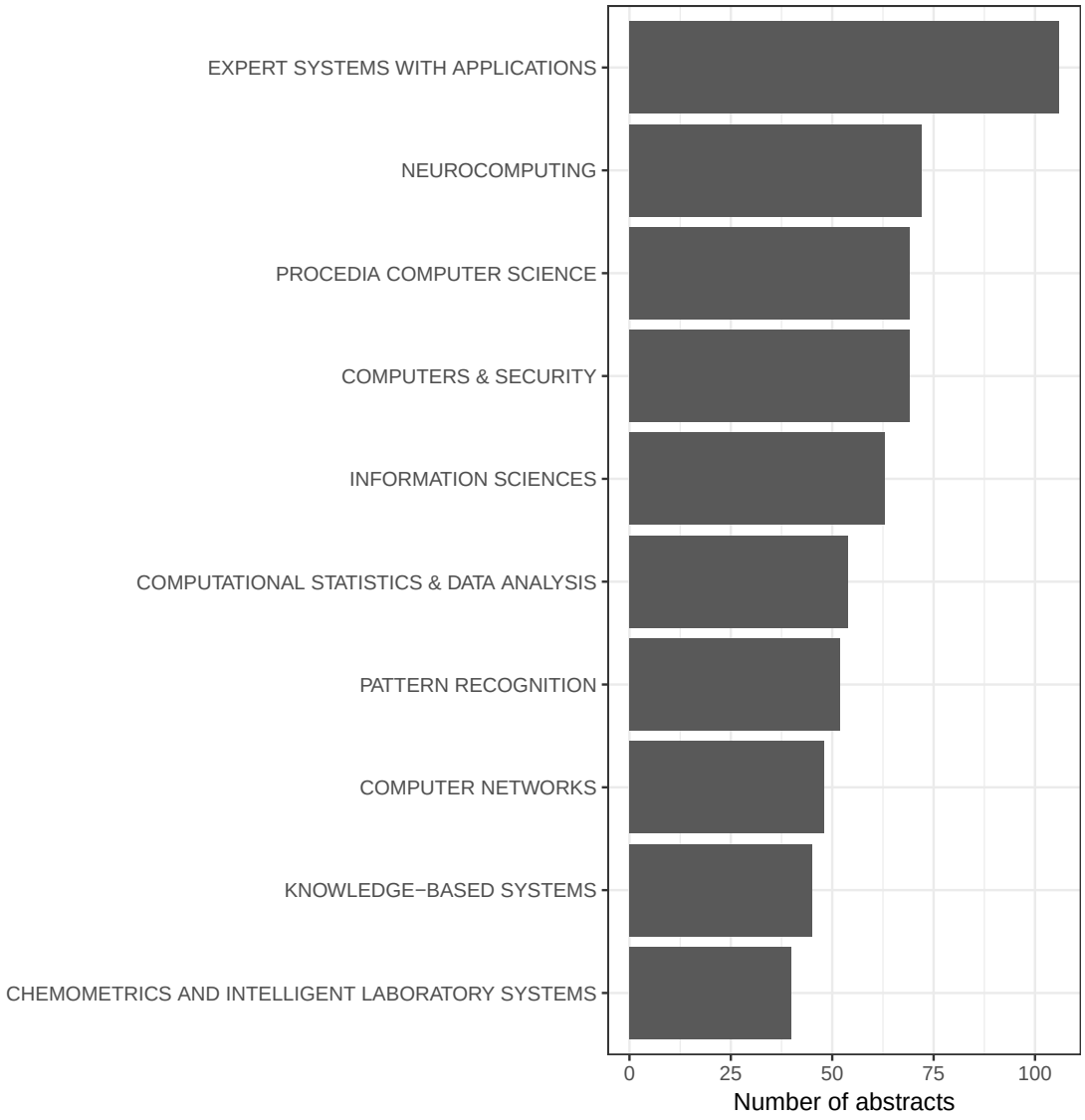


Figure 2.3: Name of Journals with highest number of publications

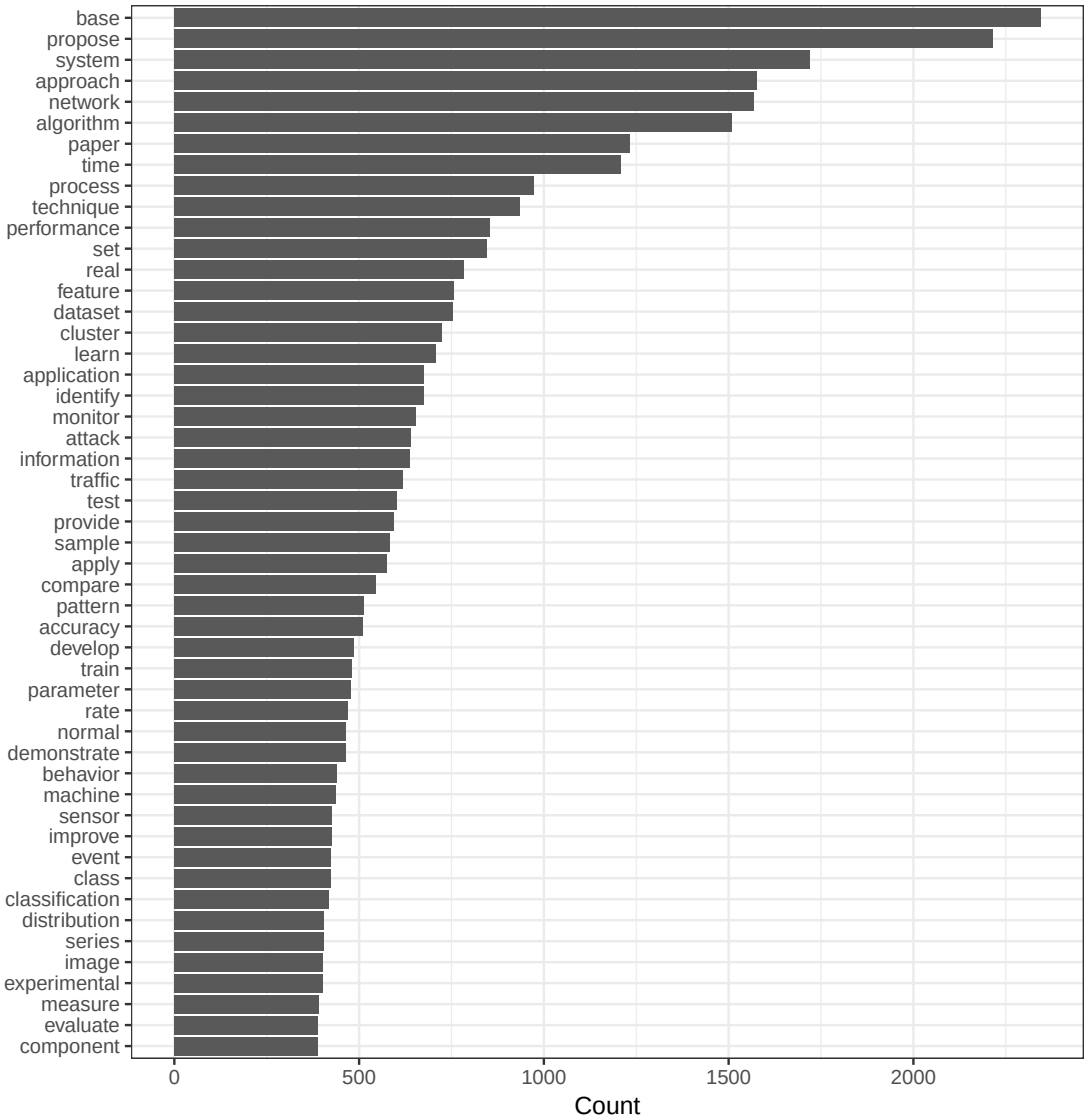


Figure 2.4: Top 50 Frequently Used Terms (Number of "terms" in this graph could exceed 50 because of same frequency for multiple "terms")

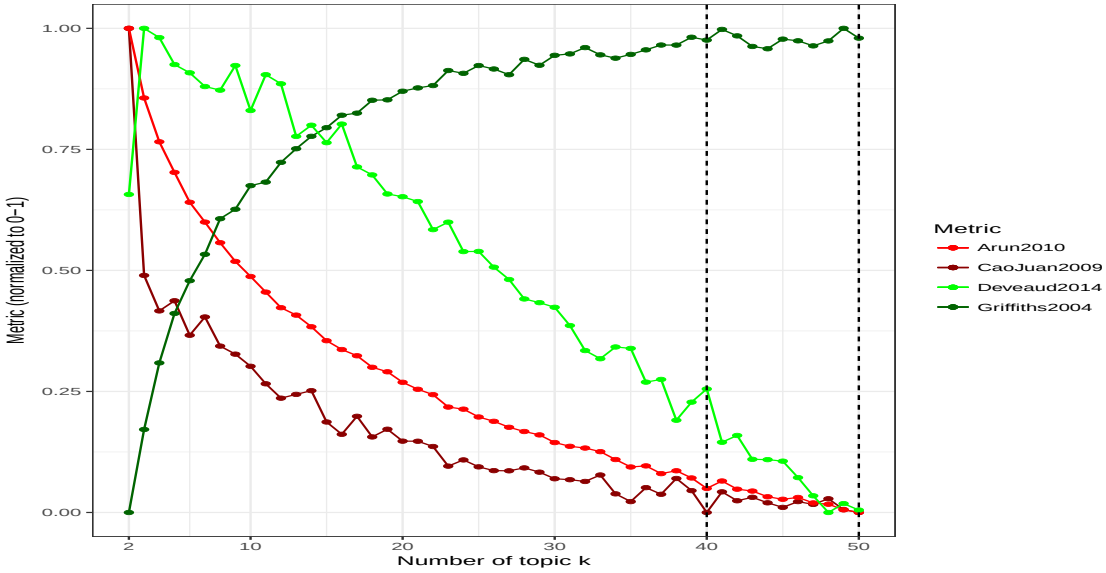


Figure 2.5: Tuning LDA model for selected metrics where vertical lines represents possible lower and upper limit of k

In the next example, a more focused research domain is explored to see whether the tool proposed can identify sub-domains within focused search results.

2.4.2 Statistical Depth Function

A statistical depth function is a function used to generalise the concept of order statistics in multivariate data. Tukey (1975) [Tuk75] in his seminal article introduced the concept of data-depth to define ad hoc ordering in multivariate random variables. A statistical depth function is a mapping from $\mathbb{R}^d$ to $\mathbb{R}$, where d is the number of variables i.e. an approach to generate a univariate score for multivariate data. Such a score can then be used to study properties of the distribution of multivariate random variables e.g. multivariate-location [DG+92] and confidence regions around multivariate-location [YS97]. The statistical depth function has been used in various statistical applications such as clustering [Hob00; JVZ02; JCS+16; Jör04] and classification [GC05; MH06; CLY08; DG11; LCL12; DG12] problems.The concept of data-depth has been extended to identify outliers in functional data [LR09; FGG08].

A more detailed description of a statistical depth function will be given in the next chapter, in particular its role in outlier detection. To explore the concept of a depth function further, the methods presented in this chapter are now used to search and explore the research landscape involving *statistical depth functions*.

Search & Selection

An electronic search was performed in ScienceDirect; an indexed scientific database using a combination of keywords: *("statistical depth" OR "depth function" OR "data depth" OR "data-depth")*.

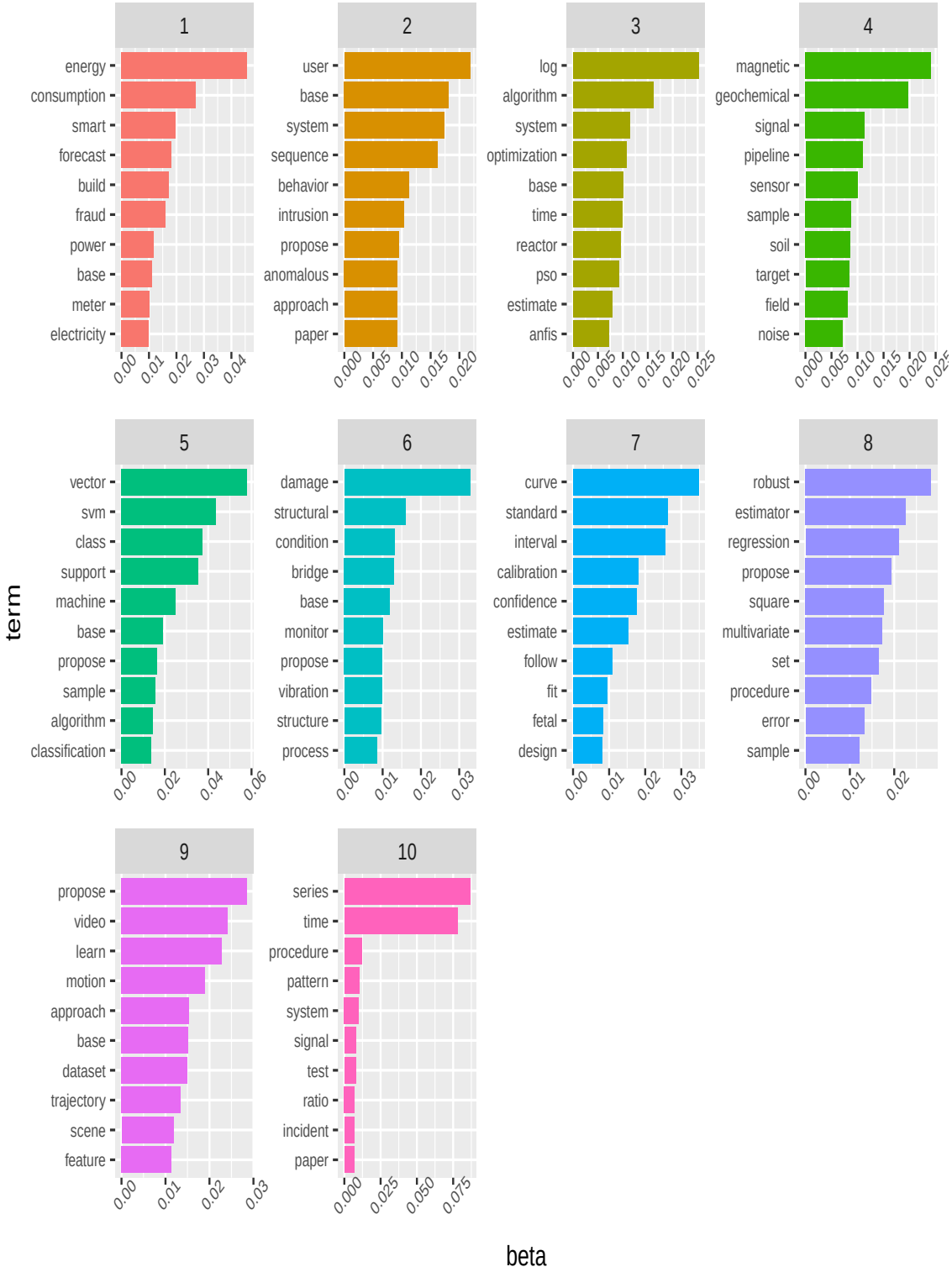


Figure 2.6: Probability of top 10 terms for topics 1-10 (x-axis is the probability of a "term" given a topic i.e. P(terms|topic))

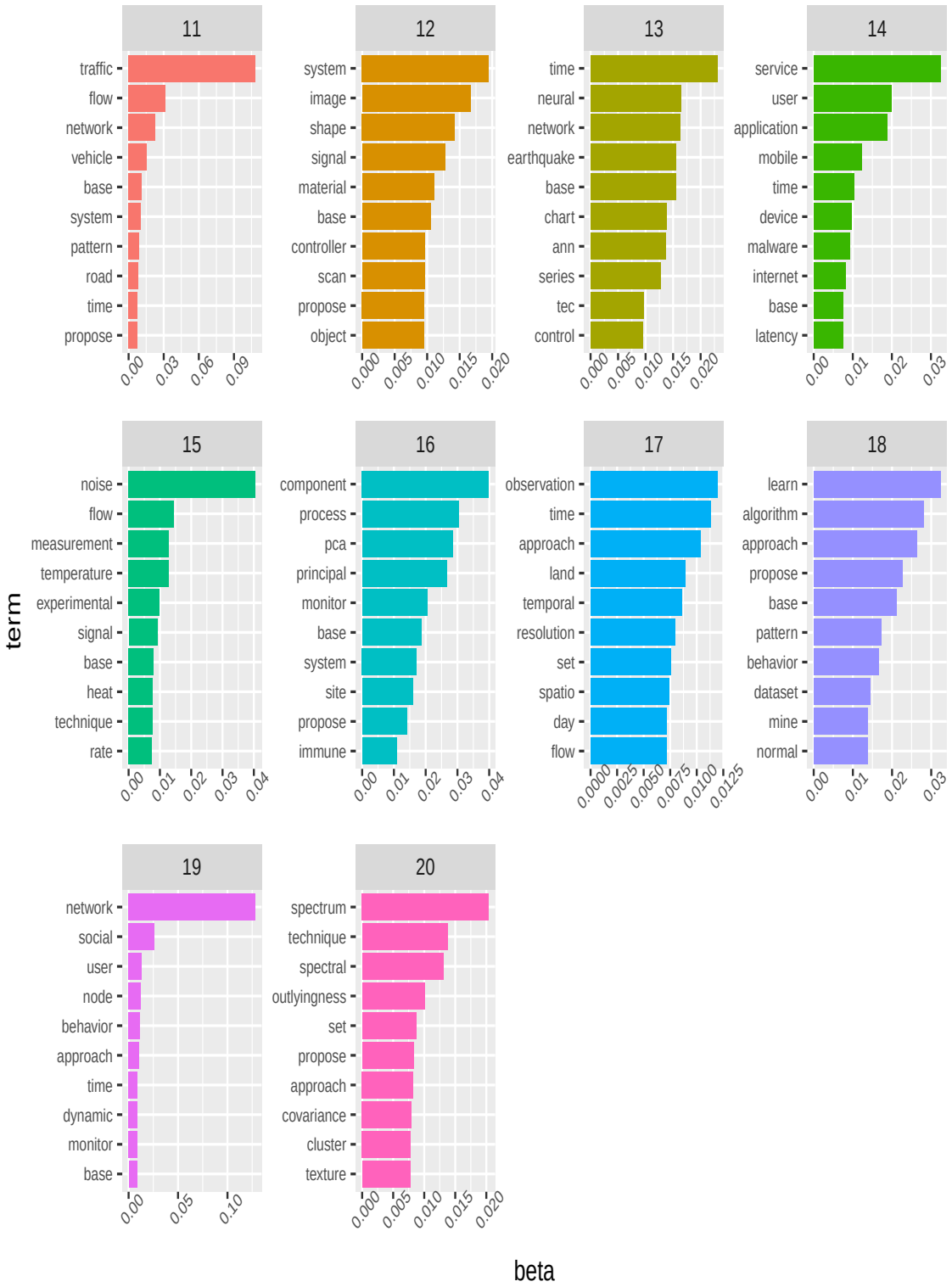


Figure 2.7: Probability of top 10 terms for topics 11-20 (x-axis is the probability of a "term" given a topic i.e. P(terms|topic))

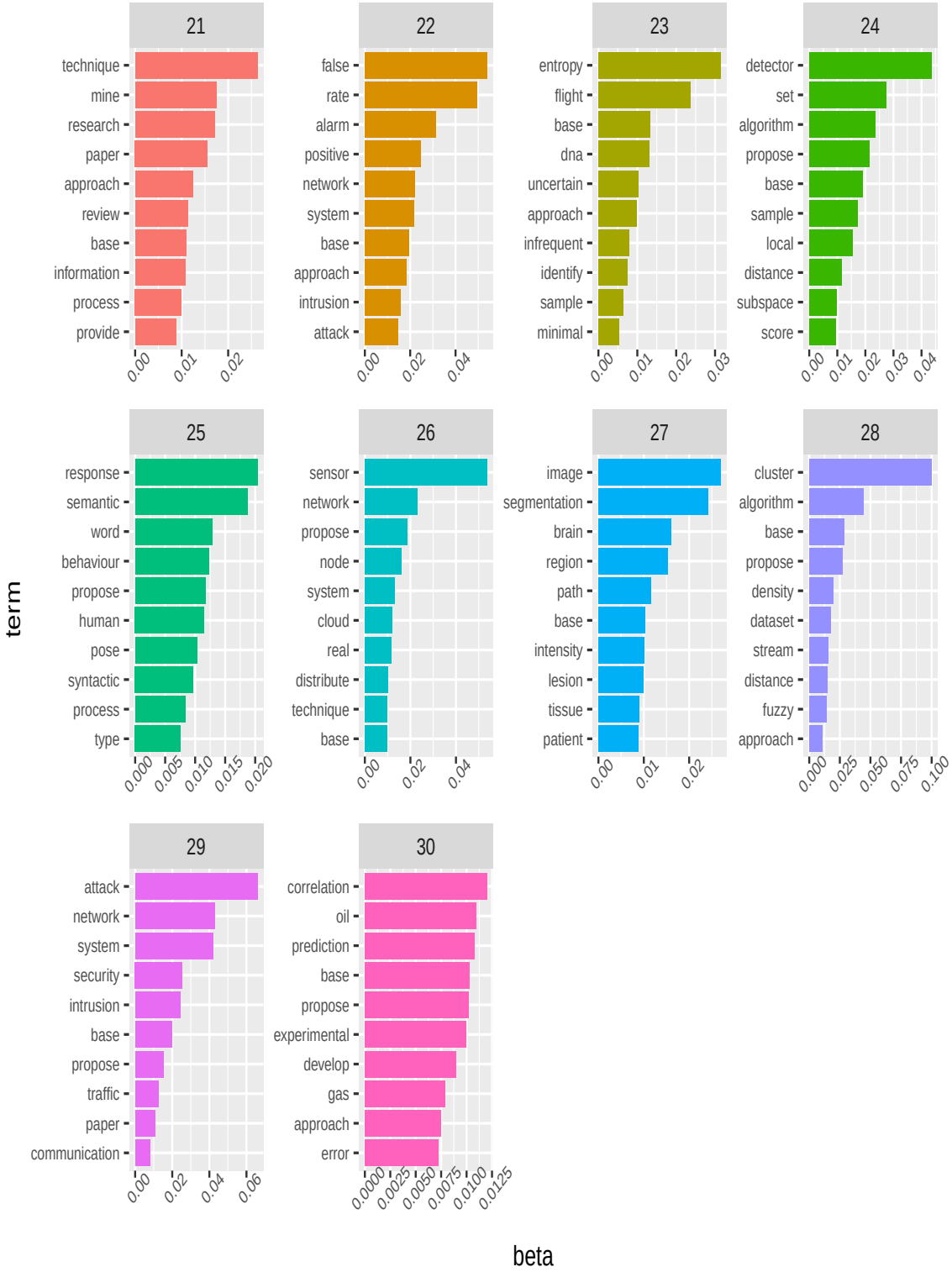


Figure 2.8: Probability of top 10 terms for topics 21-30 (x-axis is the probability of a "term" given a topic i.e. P(terms|topic))

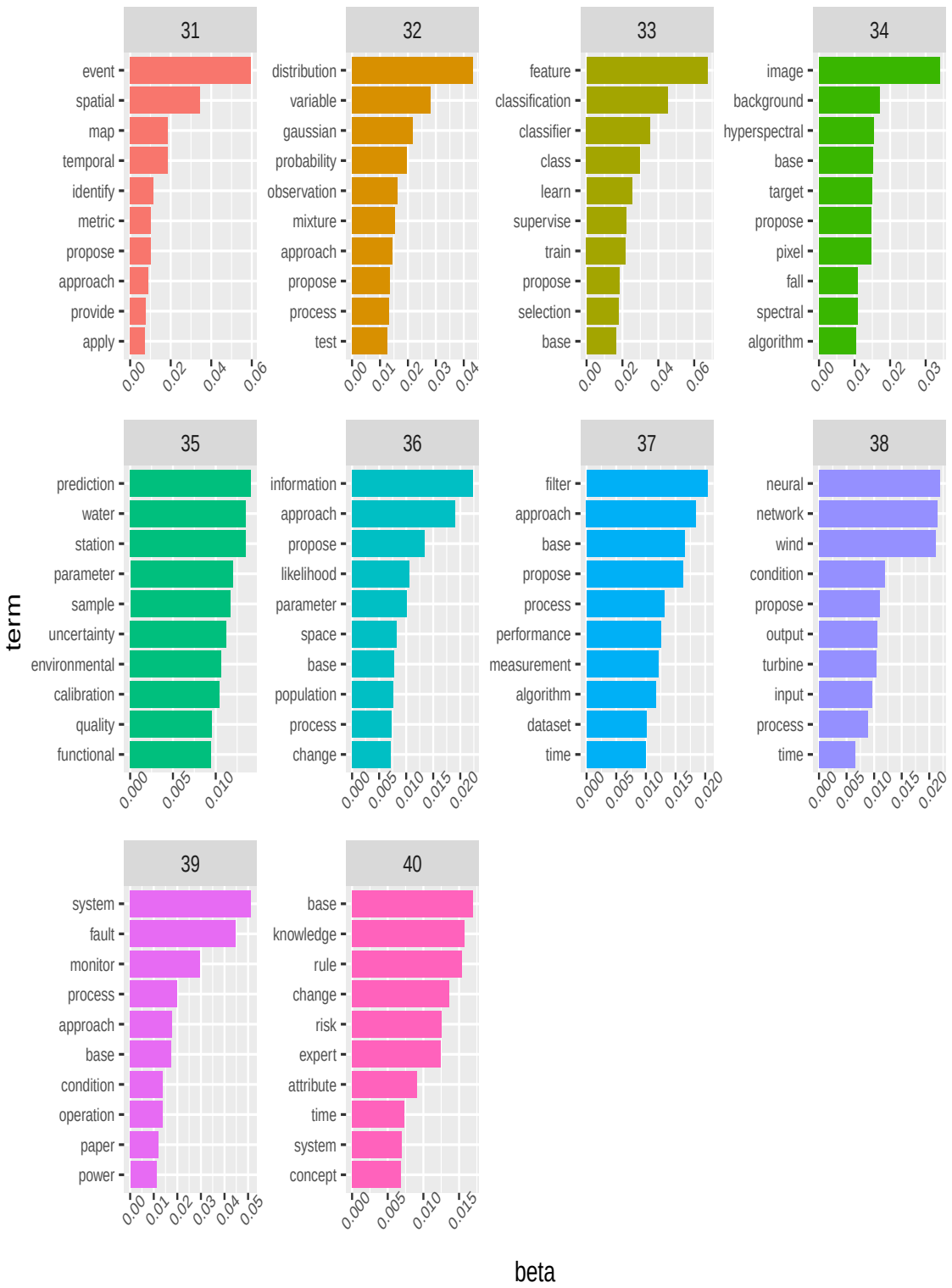


Figure 2.9: Probability of top 10 terms for topics 31-40 (x-axis is the probability of a "term" given a topic i.e. P(terms|topic))

The search results were then downloaded for further analysis. An initial inclusion criteria to select potential abstracts for further analysis was:

- Language "English" only
- Original research paper or reviews or systematic reviews
- Availability of abstracts

A total of 226 abstracts were downloaded for analysis. The earliest year of publication was 1972. The progression of the number of publications over time was calculated for further analysis.

Text Processing

The text of each abstract was cleaned by removing numerals, special characters, Unicode characters, common stop words in English. Some of the common words that could affect the frequency because of their appearance in every abstract were removed for further analysis. The 'general words' that were removed were: *"introduction"*, *"methods"*, *"method"*, *"objective"*, *"objectives"*, *"results"*, *"result"*, *"conclusions"*, *"conclusion"*, *"study"*, *"data"*,*"datum"*, *"analysis"*, *"program"*, *"model"*, *"depth"*, *"function"*. The abstracts were then used to create a corpus of text documents, where each abstract represented a document. To perform a frequency count of each "term" and to perform topic modelling the corpus needed to be converted into a Document-Term-Matrix. Each of the rows of DTM corresponded to an abstract and each column corresponded to a "term" in that abstract; the entries of the matrix is the count of occurrence of the words within an abstract.

Exploratory Analysis

There is a sharp increase in number of abstracts from the late 1990's onwards, (Figure 2.10). The increase in the number of abstracts in recent years in the use of statistical depth function could be due to the advancement of computational power. A barchart (Figure 2.11) was generated to identify the popular publication forums (name of journal) where the majority of the papers have been published. The name of journal itself indicates that there are two domains, namely statistics and environmental science, based on the search results.

The distribution of high frequency "terms" across all abstracts indicates two major research domains, namely, statistics and environmental science (Figure 2.12). To know whether there are sub themes embedded within these two major domains, the results of topic modelling are presented in next section.

Modelling & Visualisation

An LDA model was fitted using the DTM constructed in previous subsection. To identify the optimal number of topics from the corpus, a search was performed using the *ldatuning* R library and a range of potential values of k were plotted against different quality metrics on the extracted topics (Figure 2.13). The value of $k = 20$ was chosen and the distribution of words for each of the 20 topics extracted is presented in Figure 2.14. Based on the composition of the topic, it can be interpreted that topic numbers 2 and 13 are primarily focused on the domain of Statistics, particularly the use of depth functions in multivariate data analysis. All other extracted topics are indicative of various sub themes within broader domains of environmental science.

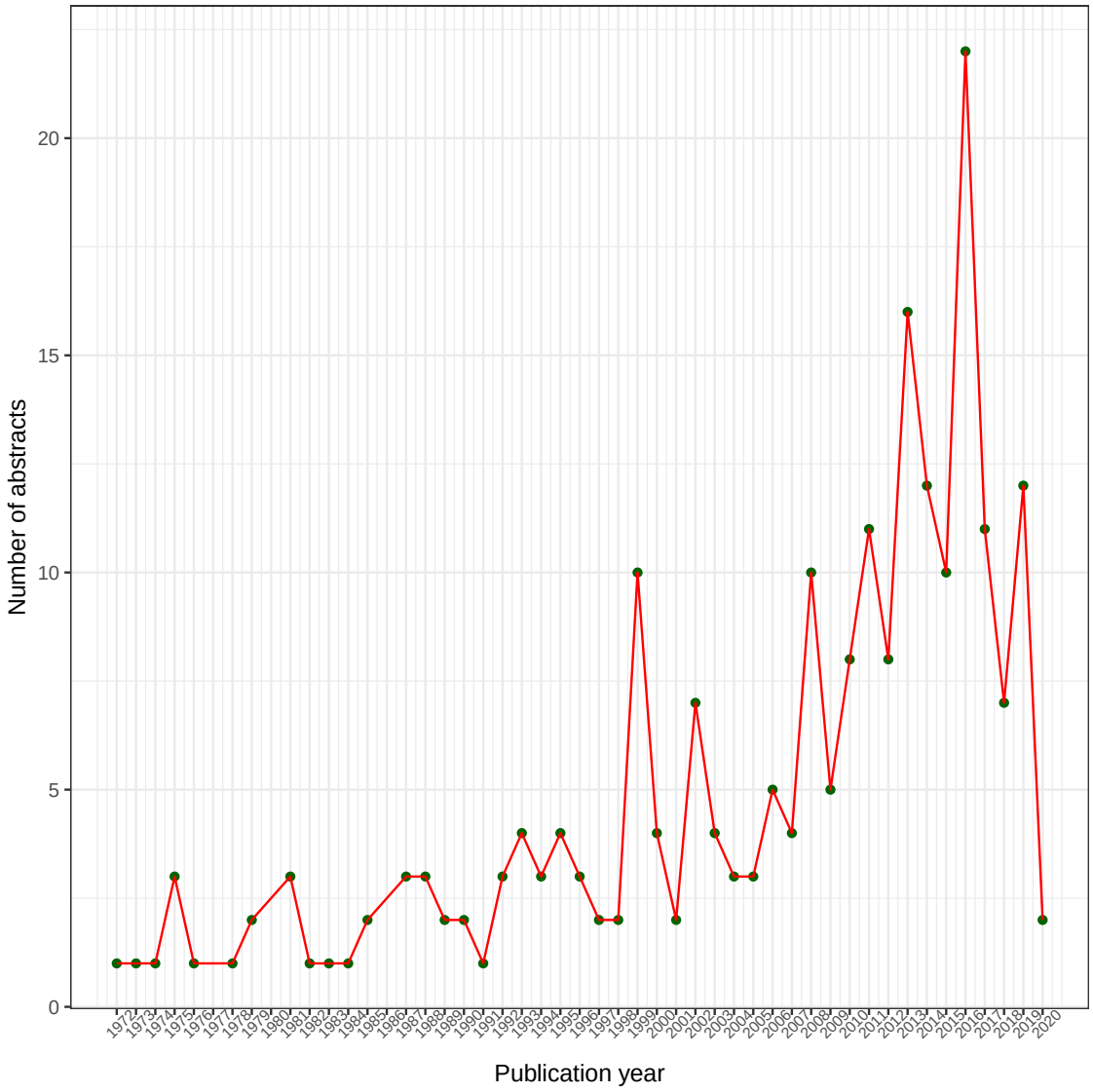


Figure 2.10: Number of abstracts per year

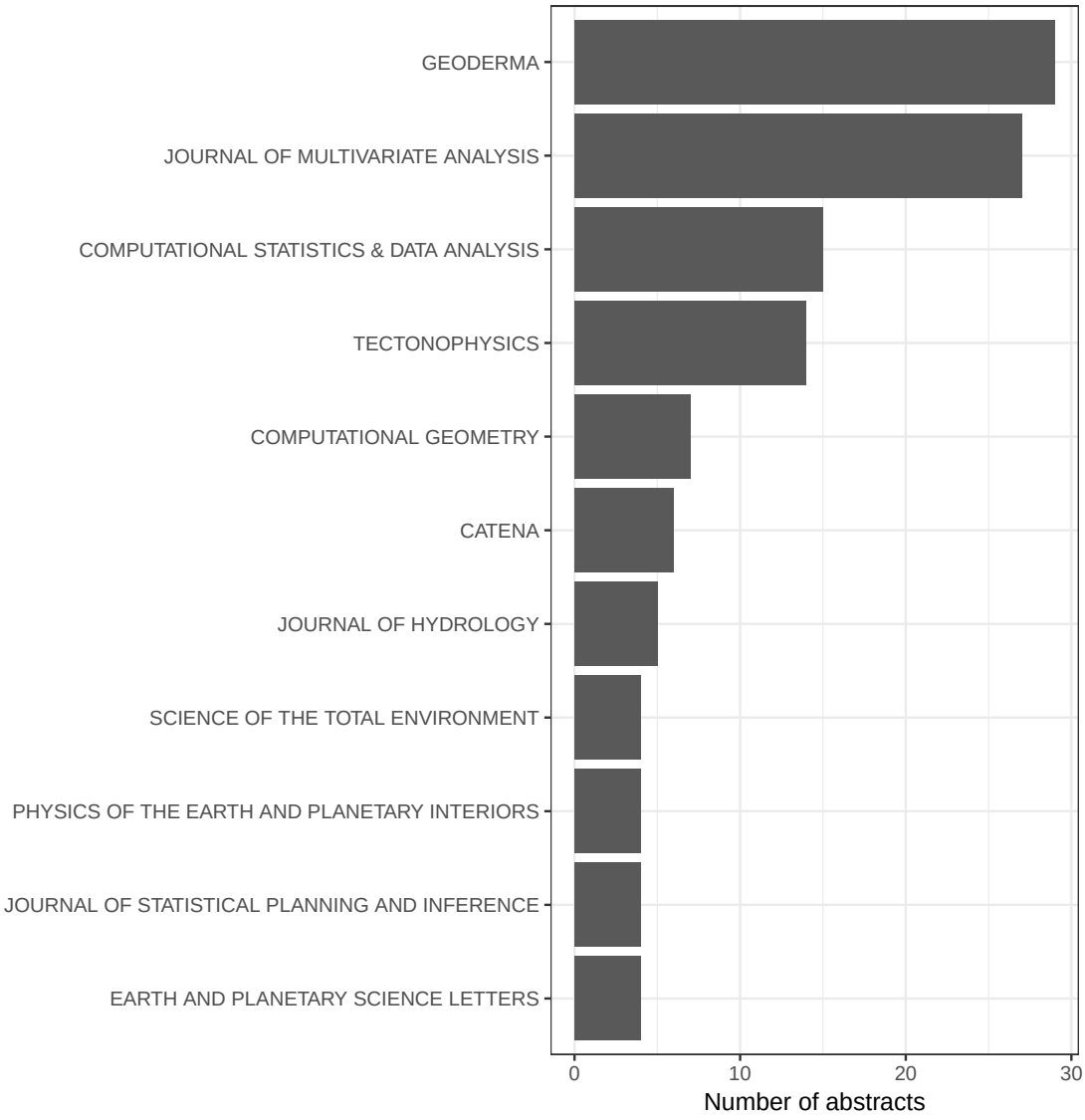


Figure 2.11: Top 10 Journals

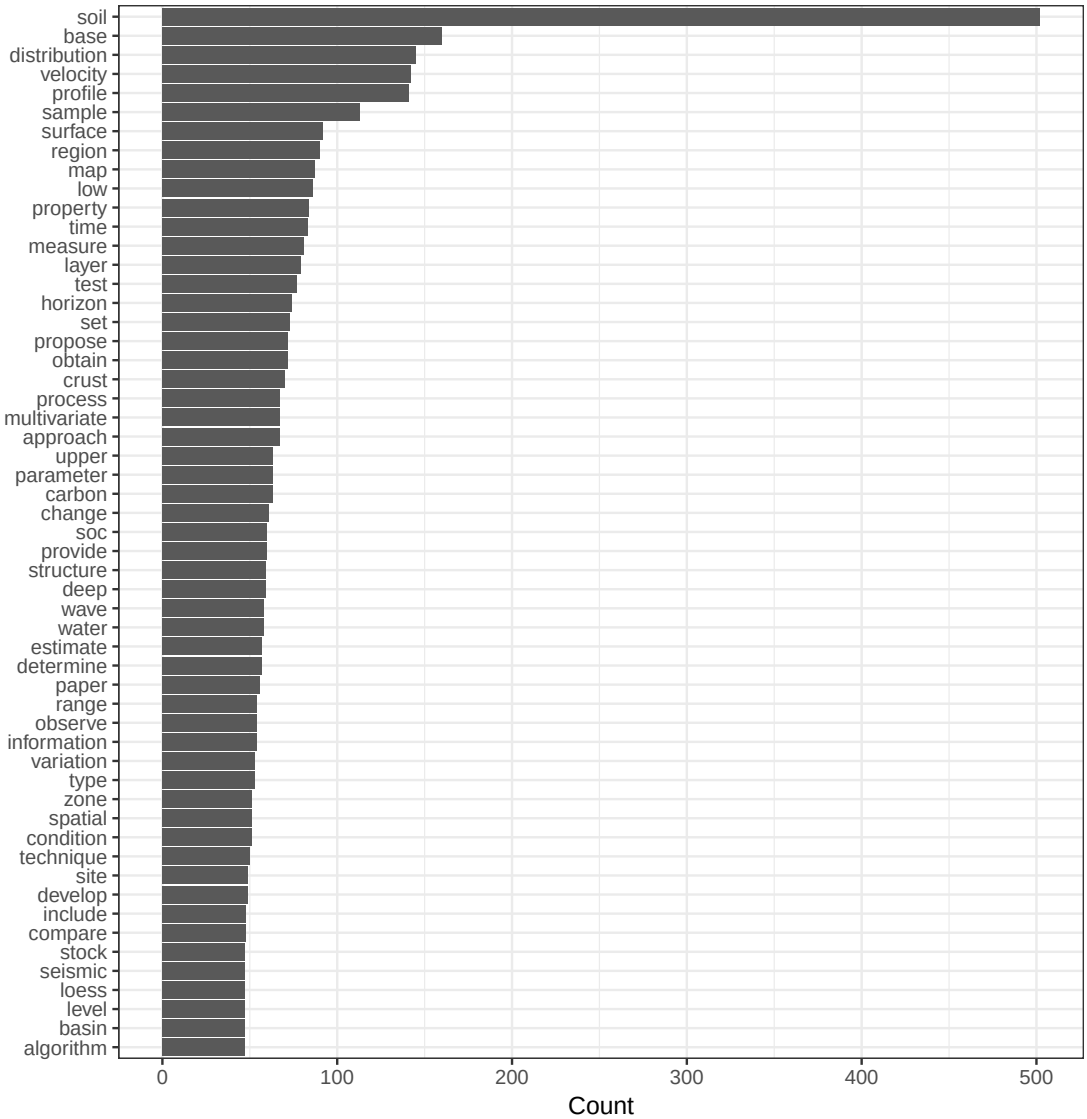


Figure 2.12: Top 50 Frequently Used "Terms". The number of "terms" presented in this figure could exceed 50 because of same frequency for multiple "terms"

In the example just presented it was plausible to read all the abstracts identified through the initial search. What was evident however is in doing so a lot of abstracts may not be of relevance to the research domain of primary interest (e.g. statistics). Although the initial search was conducted with the aim of finding research papers focused only on application of statistical depth function in statistics, the results contained many applications primarily focused on environmental science. The usefulness of the approach presented is further highlighted through the identification of these less relevant abstracts.

In the next example, a domain with a large number of papers is considered to show the usefulness of the approach in domains where it is not feasible to read all abstracts identified so that the main research themes and their evolution can be identified to augment a 'classical' literature review.

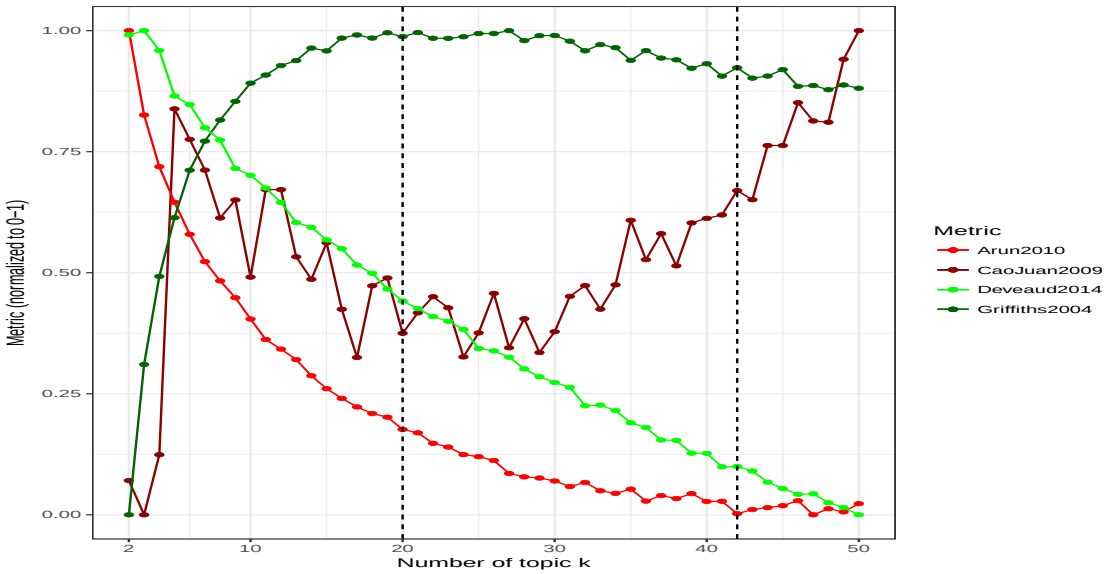


Figure 2.13: Tuning LDA model using select metrics (green and red lines) to select optimal k (vertical lines represents possible lower and upper limit of k)

2.4.3 Injuries in Elite Soccer

In the concluding example in this chapter, an example is given in the domain of sports and exercise science relating to the study of injuries in elite soccer. As a consequence the research domains will be very broad including physiology, bio-mechanics and exercise science in relation to the ability of the human body to adapt to motion, movement and physical activity.

When considering elite athletes, the advancements of athlete monitoring coupled with the technology to capture and store the vast amounts of data generated (e.g. GPS, blood bio-markers, wearable) has enabled sports scientists to utilise these data to develop optimal training strategies to maximize performance, prevent injuries, aid in recovery and prolong careers. As a consequence the number of scientific publication is growing rapidly with many new research themes emerging over the last decade.

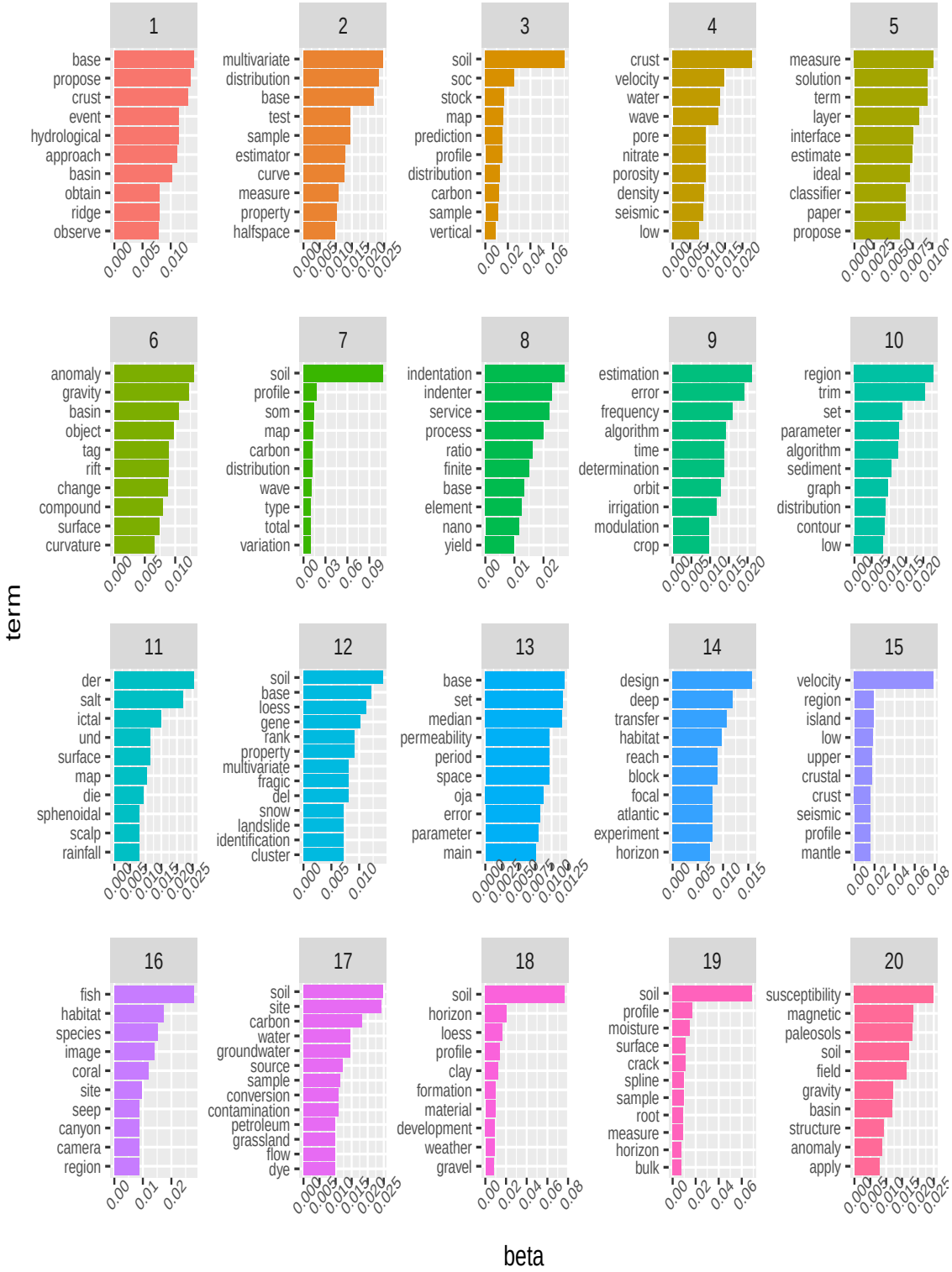


Figure 2.14: Probability of top 10 "terms" in each topic (x-axis is the probability of a word given a topic i.e. $P(words|topic)$)

Search & Selection

To identify as many articles as possible that are relevant on research in soccer (also, known as football), a search was conducted in three well-known electronic databases - ScienceDirect, PubMed and Web of Science - using the following inclusion criteria:

- Language "English" only
- Original research paper or Reviews or systematic reviews
- Availability of abstracts

To include the entire range of publication years, no filtering on the year of publication was employed. In the search result only original research articles and review article including systematic reviews were retained. The search query for the three databases were:

- ScienceDirect: *Title, abstract, keywords: ("soccer" OR "football")* with article type "Review Article" and "Research Article"
- PubMed: *"soccer"[Title/Abstract] OR "football"[Title/Abstract] AND ((Journal Article[ptyp] OR Review[ptyp] OR systematic[sb]) AND hasabstract[text] AND "humans"[MeSH Terms] AND English[lang])*
- Web of Science: *(TS=("soccer" OR "football") OR TI=("soccer" OR "football")) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Review)*

After doing an initial search using the keyword combinations, the results were retrieved from all databases and downloaded for further processing and analysis.

Text Processing & Exploratory Analysis

The name of the database, query terms used in the search and number of papers are listed in (Table 2.1). In the initial screening phase only accessible abstracts written in English relating to original research, review and systematic review articles were considered. The Web of Science database contained the majority of the results followed by PubMed and ScienceDirect. A paper could appear in multiple databases creating duplicate records in the combined results. To remove such duplicates, the search results were combined and duplicate titles were removed. After removing duplicate results, 28,115 abstracts were available to create a corpus for analysis.

The search results included a few early publications from the 1930s where the primary area of research concerned the treatment of fractures [Hen33] in general and not specifically in sports. Research appearing in the 1950's [Ats57] focused on the coordination between the coach and team physician to determine whether a player should play or be substituted, and explored treatment approaches for various injuries such as temporary fixation by metallic lag screws to repair ligament structures [MGT57], treatment of elbow joint dislocation of athletes [Lip58] and the treatment of shoulder girdle injuries in football [Pat60].

Repositories	Searched Queries	Papers
ScienceDirect	Title, abstract, keywords: ("soccer" OR "football")	2801
PubMed	"soccer"[Title/Abstract] OR "football"[Title/Abstract] AND ((Journal Article[ptyp] OR Review[ptyp] OR systematic[sb]) AND hasabstract[text] AND "humans"[MeSH Terms] AND English[lang])	9382
Web of Science	(TS=("soccer" OR "football") OR TI=("soccer" OR "football")) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Review)	25071
Total		**37254**

Table 2.1: Number of articles searched in different repositories (including duplicate titles)

There was a noticeably rapid increase in publications from 2000 to 2010 where the search results exceeded 1500 papers per year (Figure 2.15). This rapid increase is due to ongoing technological advancements in the development of wearable devices used in athlete monitoring. Such devices include GPS monitors and motion sensor cameras to track and quantify movement, sleep sensors to monitor sleep hygiene and analysers for point of care testing (e.g. blood and saliva biomarkers). The ability to capture and store player monitoring data in real time has created new research opportunities and subsequent publications in sports, as evidenced by the emergence of specialised journals in sports science. This is particularly evident in research relating to soccer, one of the most popular sports worldwide, where research is primarily focused on player monitoring to optimise performance while minimising the incidence of soft tissue injuries.

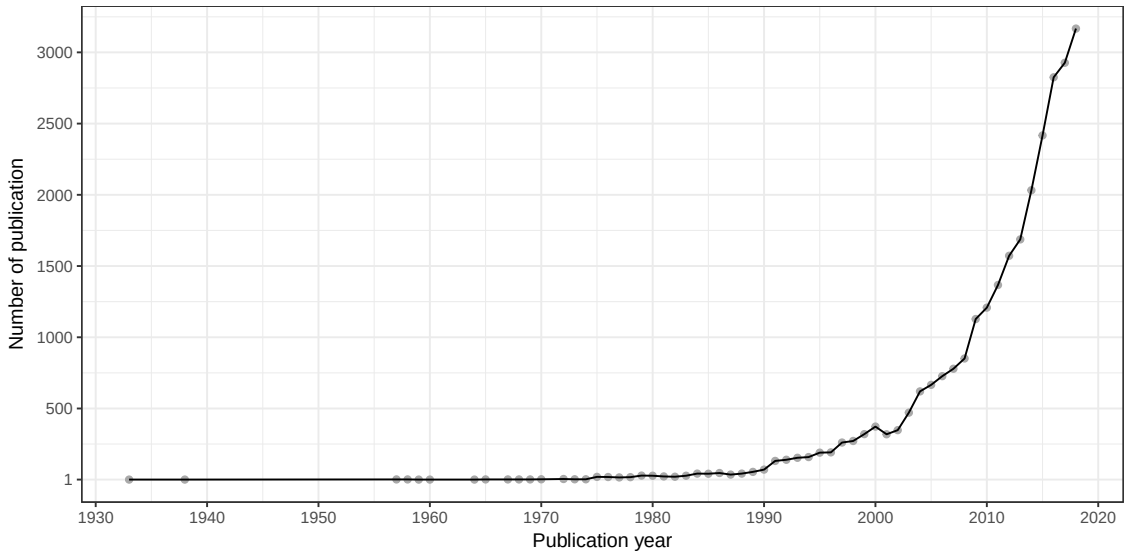


Figure 2.15: Number of paper per year

The top 10 Journals, in terms of the number of papers in the search results, are given in (Figure 2.16). A Journal with over 1000 papers (Figure 2.16) appeared in the search results are the ones focused on sports medicine but also included journals that focus on training, performance and sports psychology.

After combining the initial search results and removing duplicate entries a total 28115 abstracts were available for further analysis. Clean texts were obtained (as discussed in *Text processing and exploratory analysis section*) by removing numerals, stop-words, common words and sparse words. The collection of 'clean' abstracts were then converted into a document term matrix for

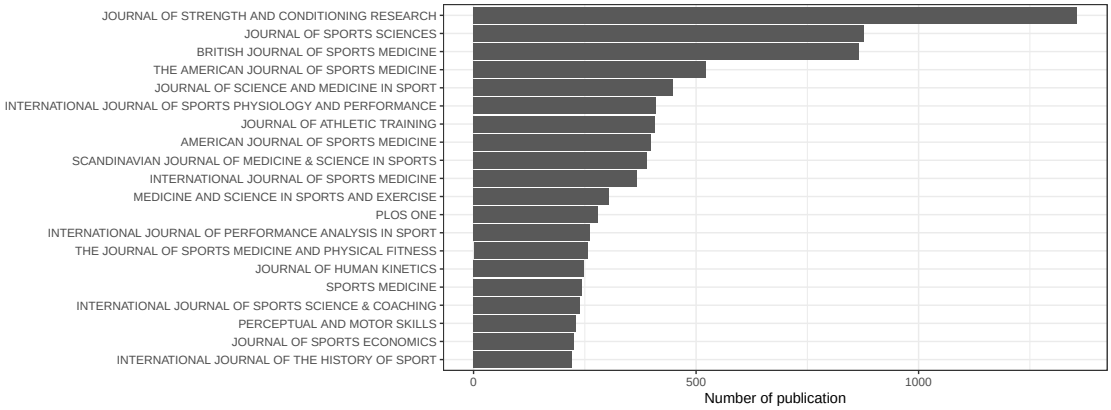


Figure 2.16: Top 10 Journals (based on frequency of papers in the search result)

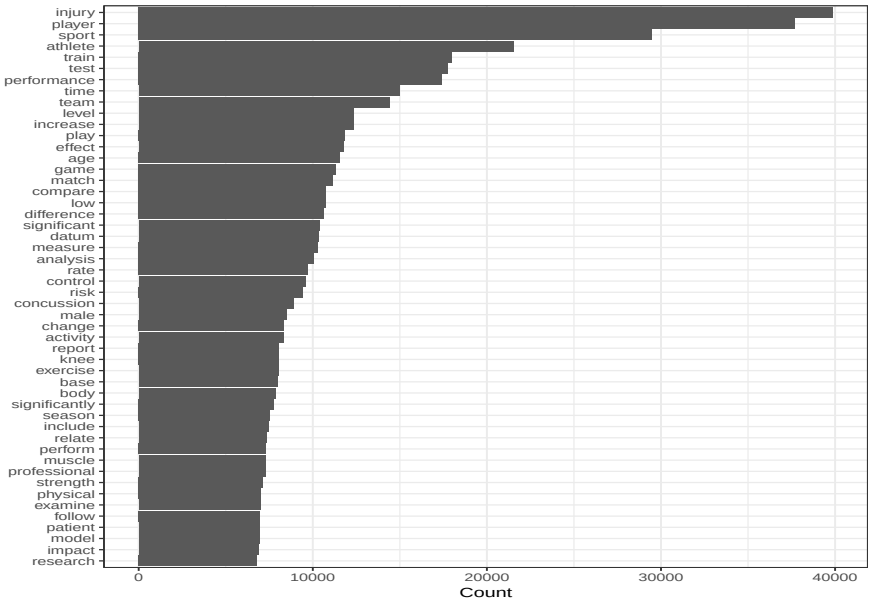
unigram and bigram analyses to identify the most frequent words used in the abstract corpus. The top 50 words in the unigram highlighted the depth of research activity in *injury, performance analysis and training*(Figure 2.17a). Top 50 bigram statistics indicated the more granular level of injury detail. Interestingly, *significant difference* (Figure 2.17b) stood out as the top terms used in the abstract corpus suggest an excessive reliance on statistical significance based on reporting p-values as opposed to interval estimation (Figure 2.17).

The document-term matrix (DTM) for unigram and bigram contained 50856 and 1151815 words; that is, for unigram there were 50856 unique words (column of DTM), and for bigram a total of 1151815 columns. After removing sparse words (words with zero frequency for almost every document in the corpus) from the DTM 5057 and 210 words remained for unigram and bigram to use in topic modelling. Based on the frequency of word use in the abstracts over time, it is evident that the majority of research reported was focused on injury, injury treatment and sometimes injury prevention. In recent years, research on performance analysis is emerging as the frequency of words related to performance analysis has continued to increase (Figure 2.18 and 2.19).

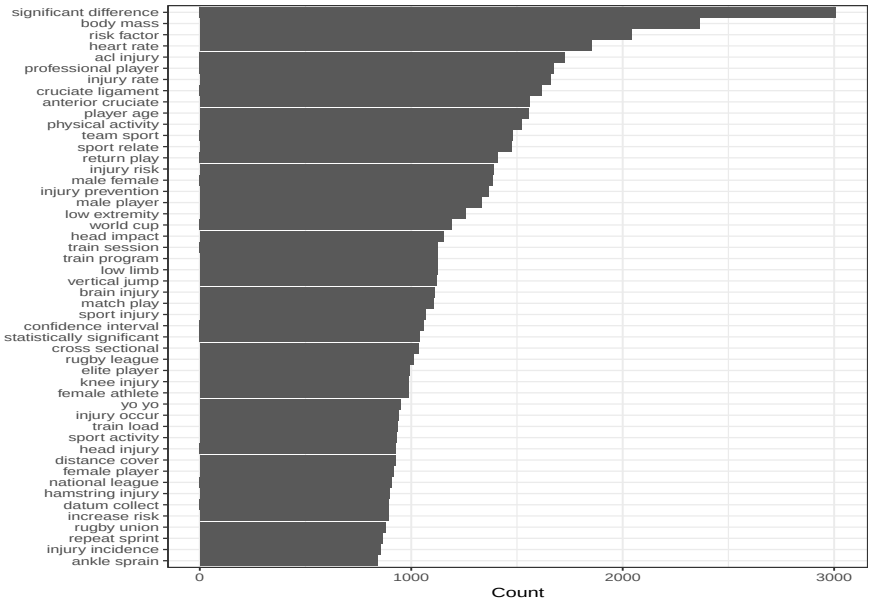
Before performing topic modelling on the document term matrix, a visualisation of the high dimensional matrix as a two-dimensional scatter plot was created to investigate whether any obvious clusters existed among the abstracts. A t-distributed Stochastic Neighbour Embedding (t-sne) [MH08] analysis was used to reduce the dimensionality followed by k-means clustering algorithm to explore potential cluster memberships of the abstracts. The k-means algorithm was tuned using average Silhouette resulting in 48 potential clusters (Figure 2.20).

modelling & visualisation

In the LDA modelling step, tuning was performed as discussed in section 2.3.4 to select the value of k. The optimal number of topics k is selected from the lower limit of the range of probable values in order to maintain balance in granularity level while avoiding excessive topic aggregation. A value of $k = 48$ was used to fit the final LDA model. The x-axis (Figure 2.21) represents the number of topics and the y-axis represents the LDA model evaluation metric. For this analysis $k = 48$ was chosen and the topics based on this choice were extracted from the abstract corpus.



(a) Top 50 "terms" (unigram)



(b) Top 50 "terms" (bigram)

Figure 2.17: Frequently used words (unigram and bigram)

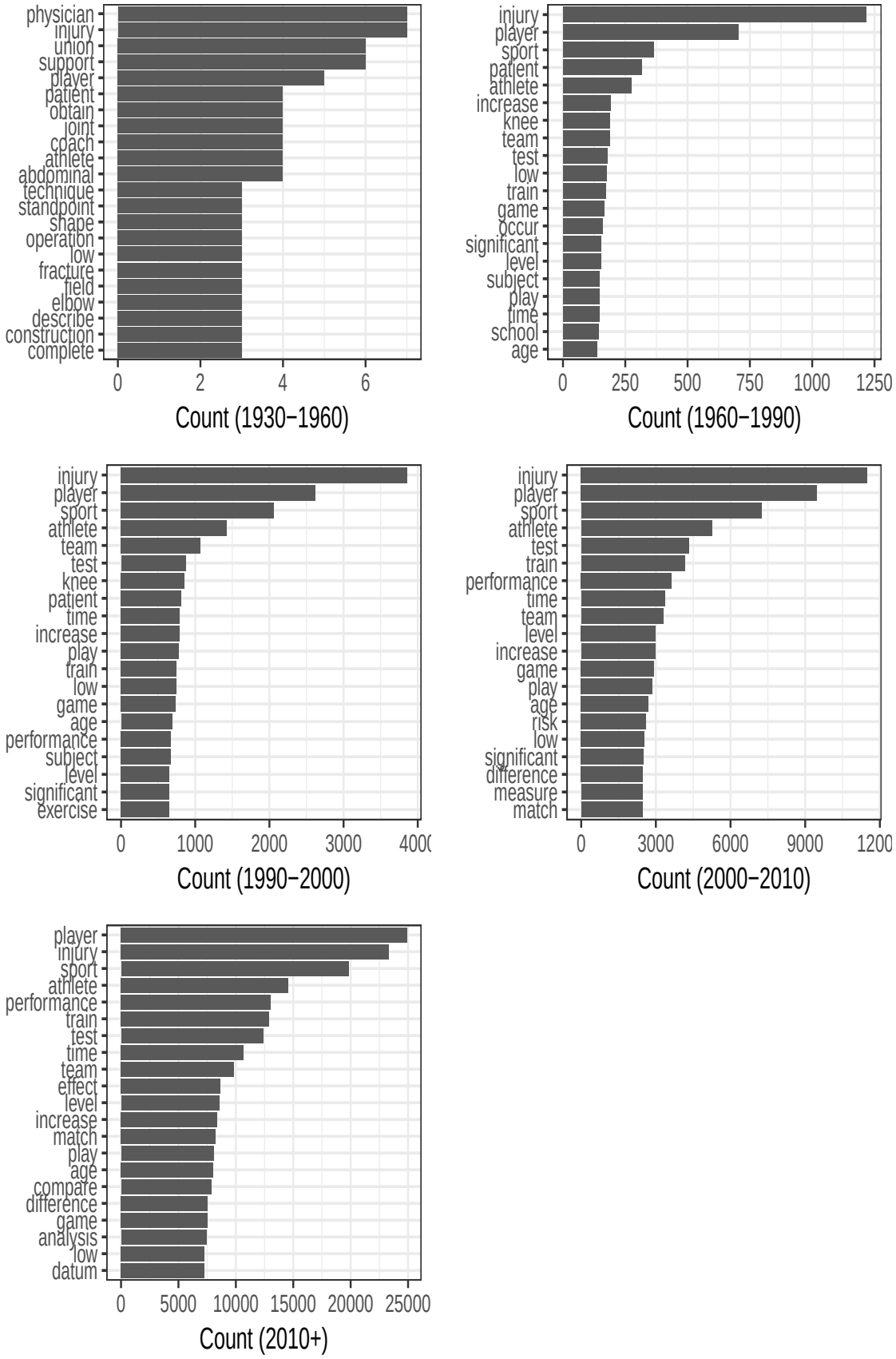


Figure 2.18: Frequently used unigram "terms" over time

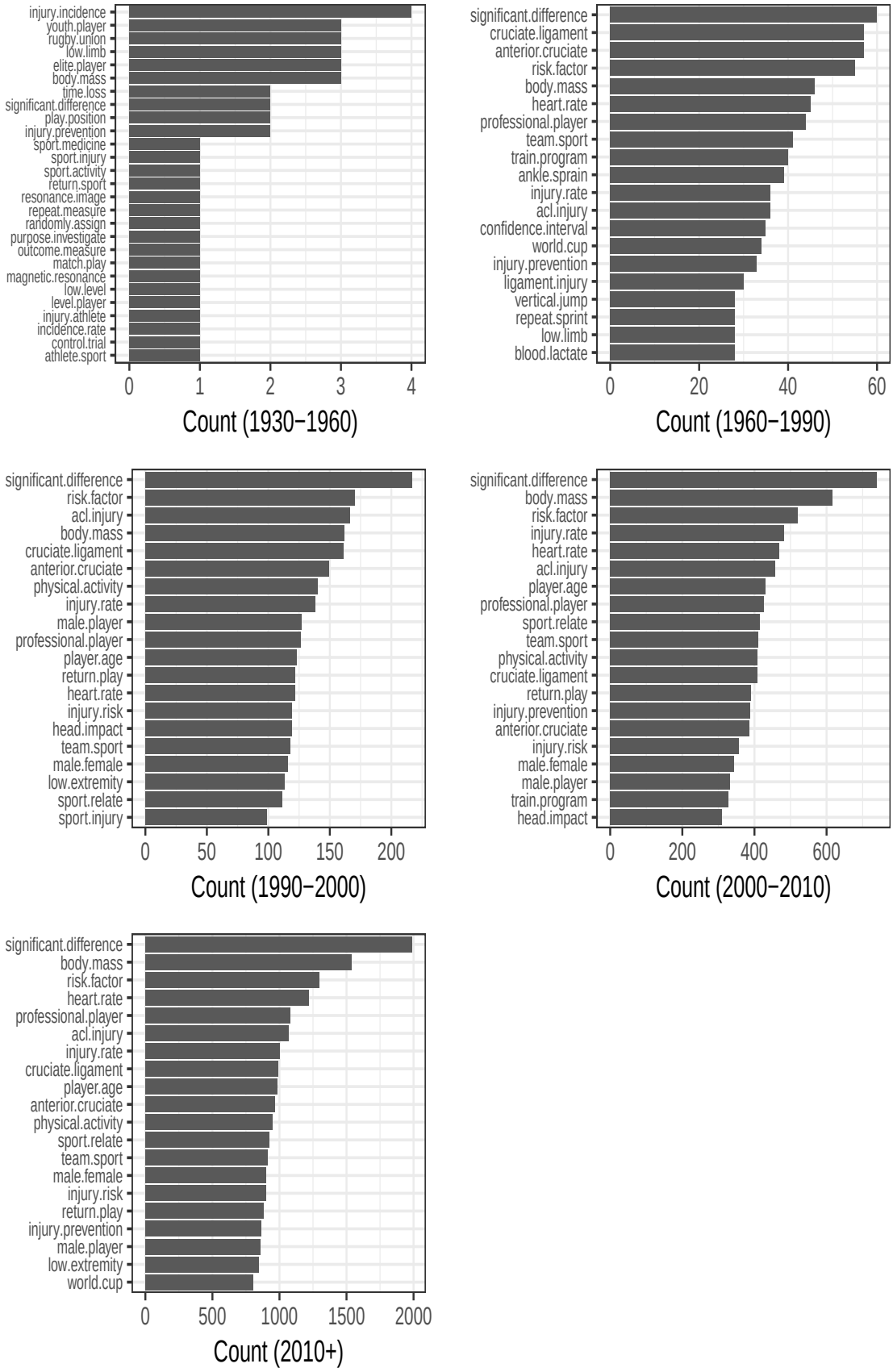


Figure 2.19: Frequently used bigram "terms" over time

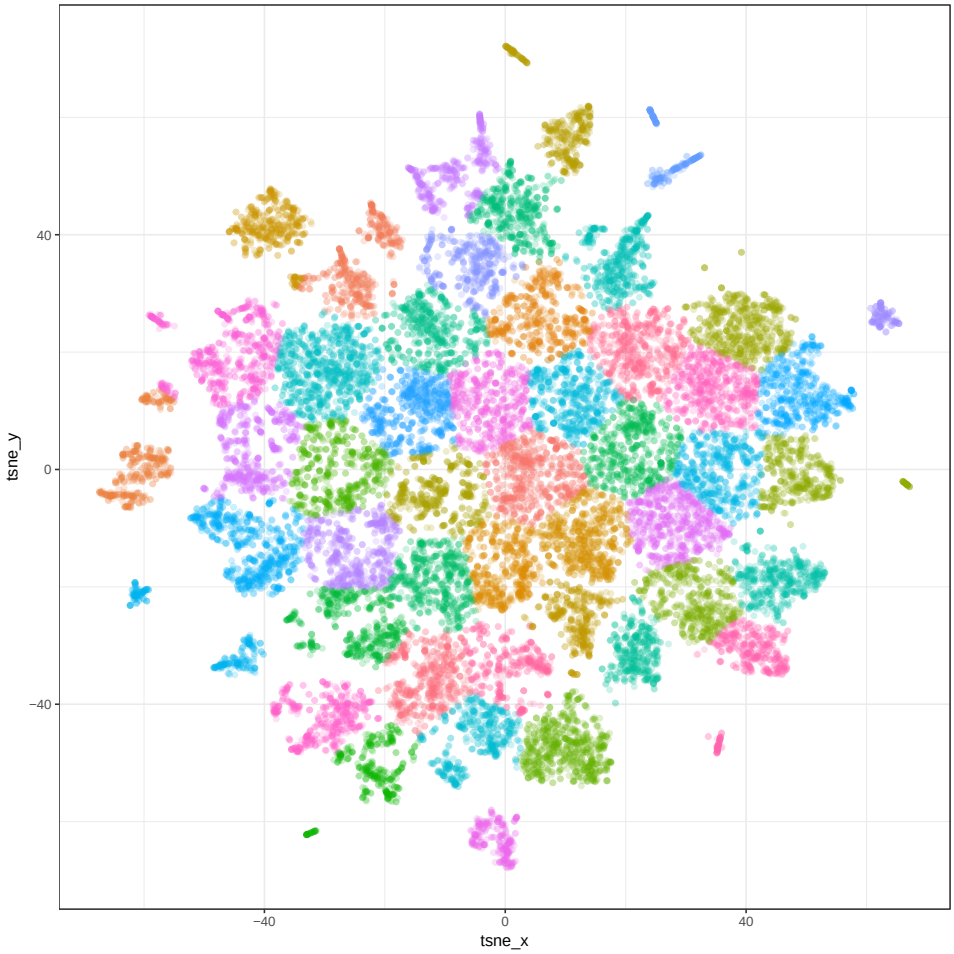


Figure 2.20: Clustering of the abstracts based on unigram DTM (x-axis and y-axis are derived by using t-sne algorithm, here x and y both are the projected axis from the original DTM)

Upon selecting the number of topics based on tuning, a topic model using LDA was fitted and the underlying topics from the DTM extracted. The probability of each "term" for a given topic was then estimated. The "terms" with highest probability for each topic were then visualised and interpreted based on the composition of the "terms" within each topic. There was a total of 48 topics identified and top 10 "terms" from each topic was visualised in Figure 2.22, 2.23, 2.24 and 2.25.

Each of the topics can be interpreted based on the composition of "terms" within each topic e.g. topic 1 can be considered as knee injury and performance related whereas topic 2 is more about player's strength training. On the other hand topic 10 is more about head injury and topic 22 about hamstring injuries.

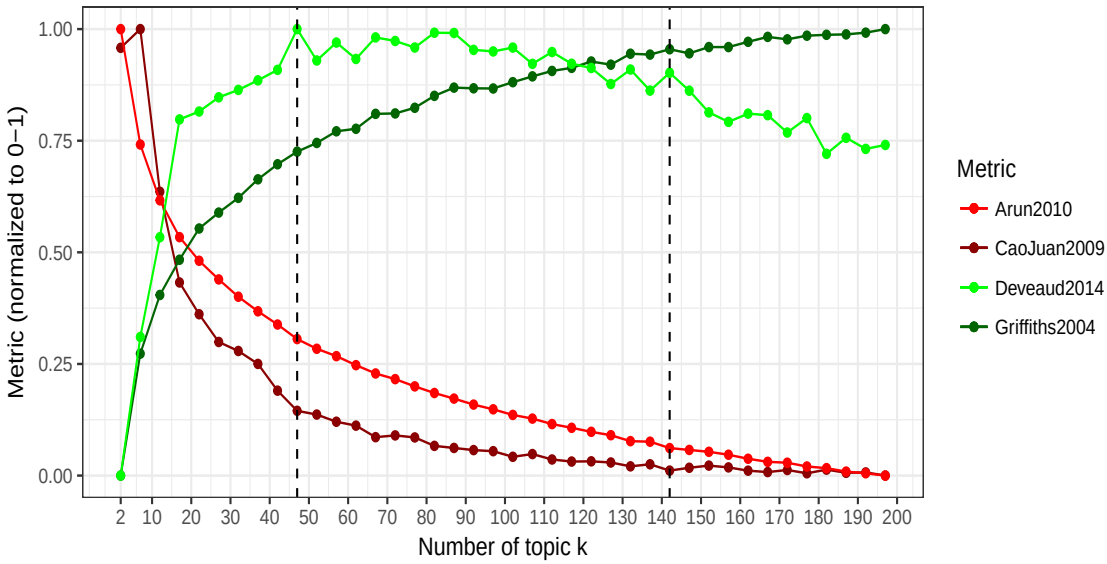


Figure 2.21: Tuning LDA model for select metrics where vertical lines represents possible lower and upper limit of k

The previous examples have highlighted the usefulness of text mining and topic modelling to augment a traditional literature review. In order to make this process accessible to researchers in general, a tool, in the form of a web application, has been developed and deployed as outlined in the following section.

2.5 A SHINY APP

A web application has been developed to automate the approach discussed in this chapter which can be used to analyse collections of abstracts, apply topic modelling and visualise the results to identify latent research themes and how the topics have changed over the time. Users will be able to investigate the growth rate of research in any area of interest where electronic publications are available . This R shiny app has the potential to be a valuable tool when undertaking a literature review in any domain of interest. The shiny app has the following functionality;

- to pre-process plain text;

Figure 2.22: Probability of top 10 "terms" for topics 1-12 (x-axis is the probability of a word given a topic i.e. $P(words|topic)$)

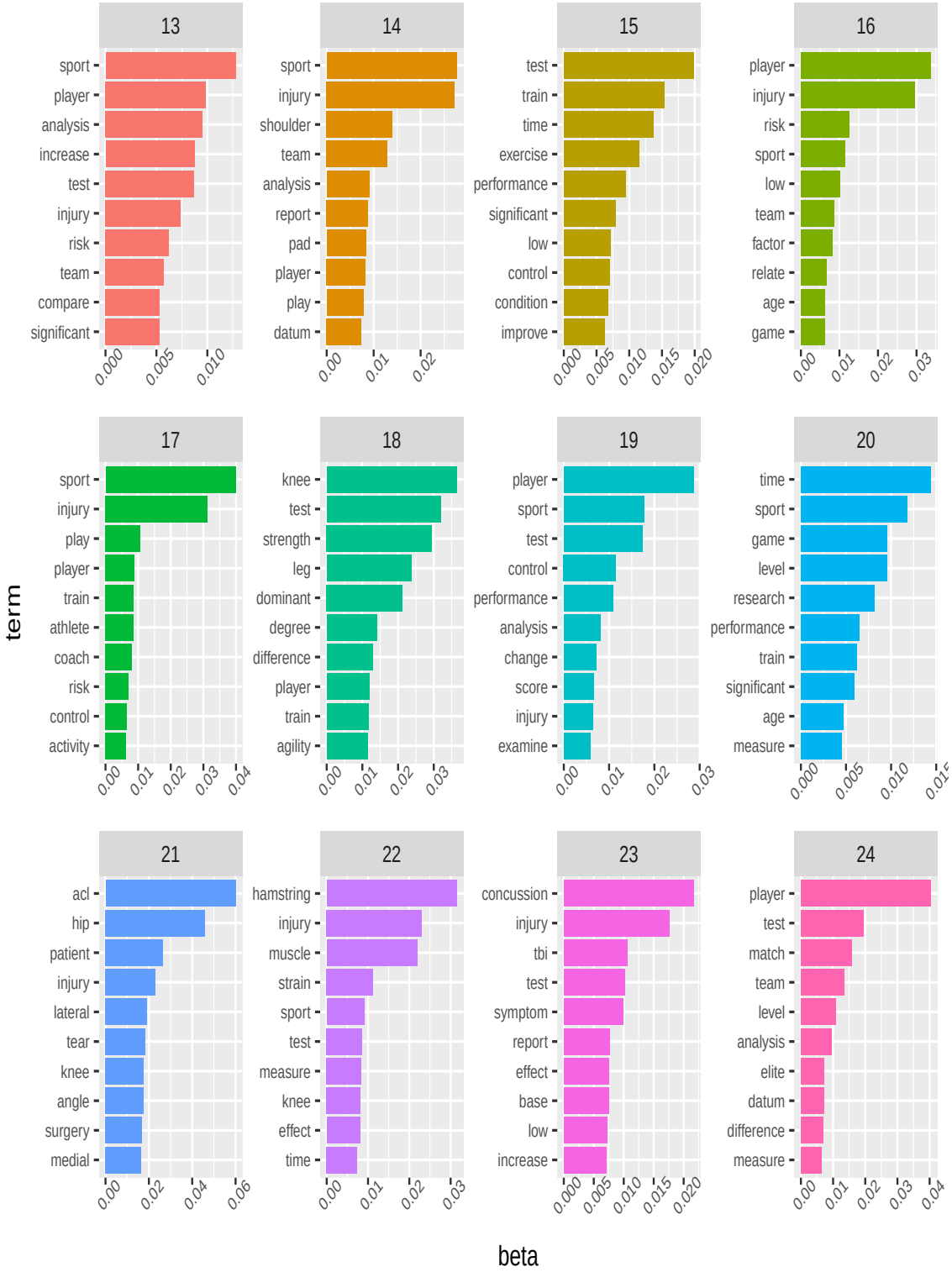


Figure 2.23: Probability of top 10 "terms" for topics 13-24 (x-axis is the probability of a word given a topic i.e. $P(words|topic)$)

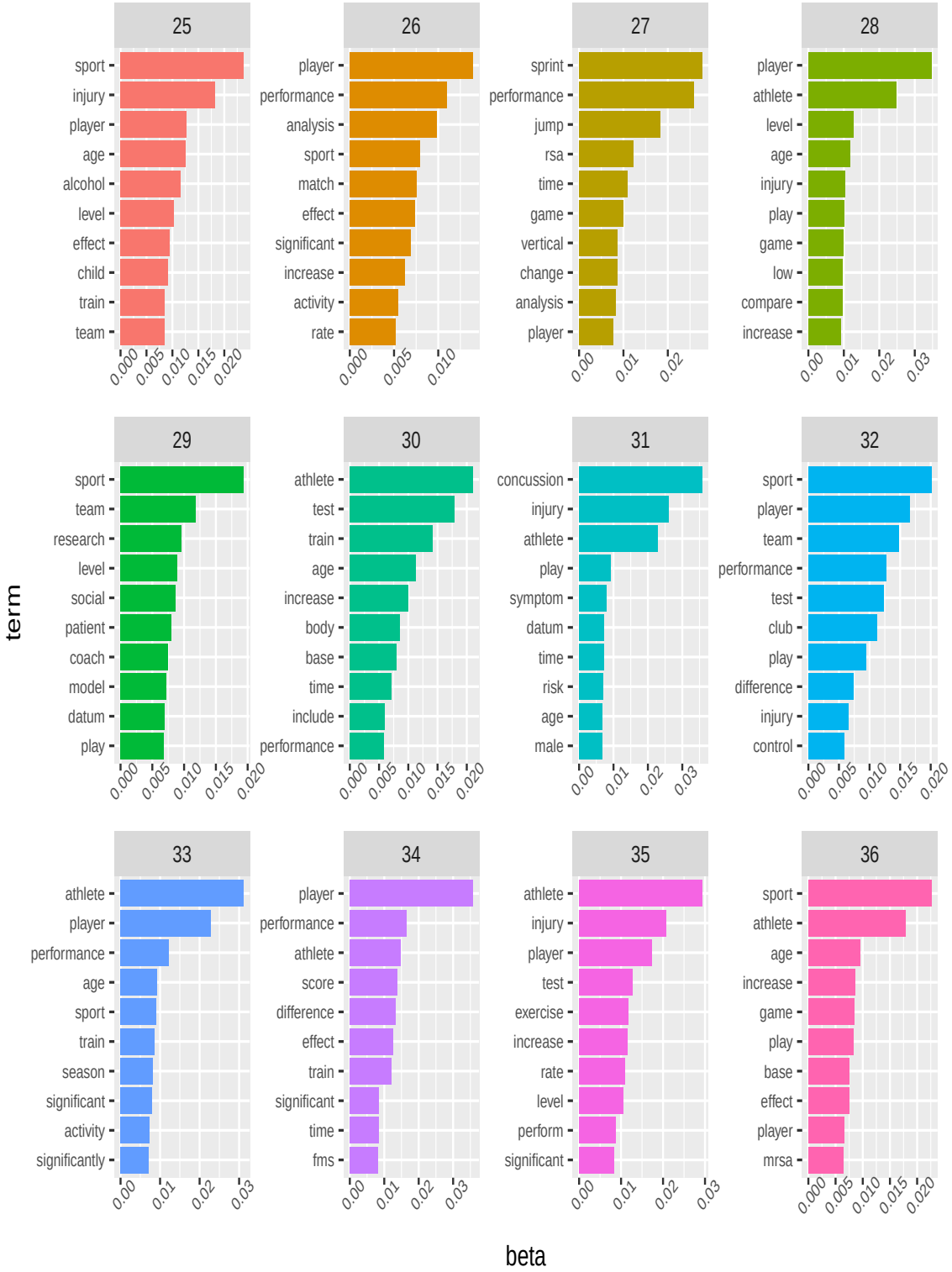


Figure 2.24: Probability of top 10 "terms" for topics 25-36 (x-axis is the probability of a word given a topic i.e. $P(words|topic)$)

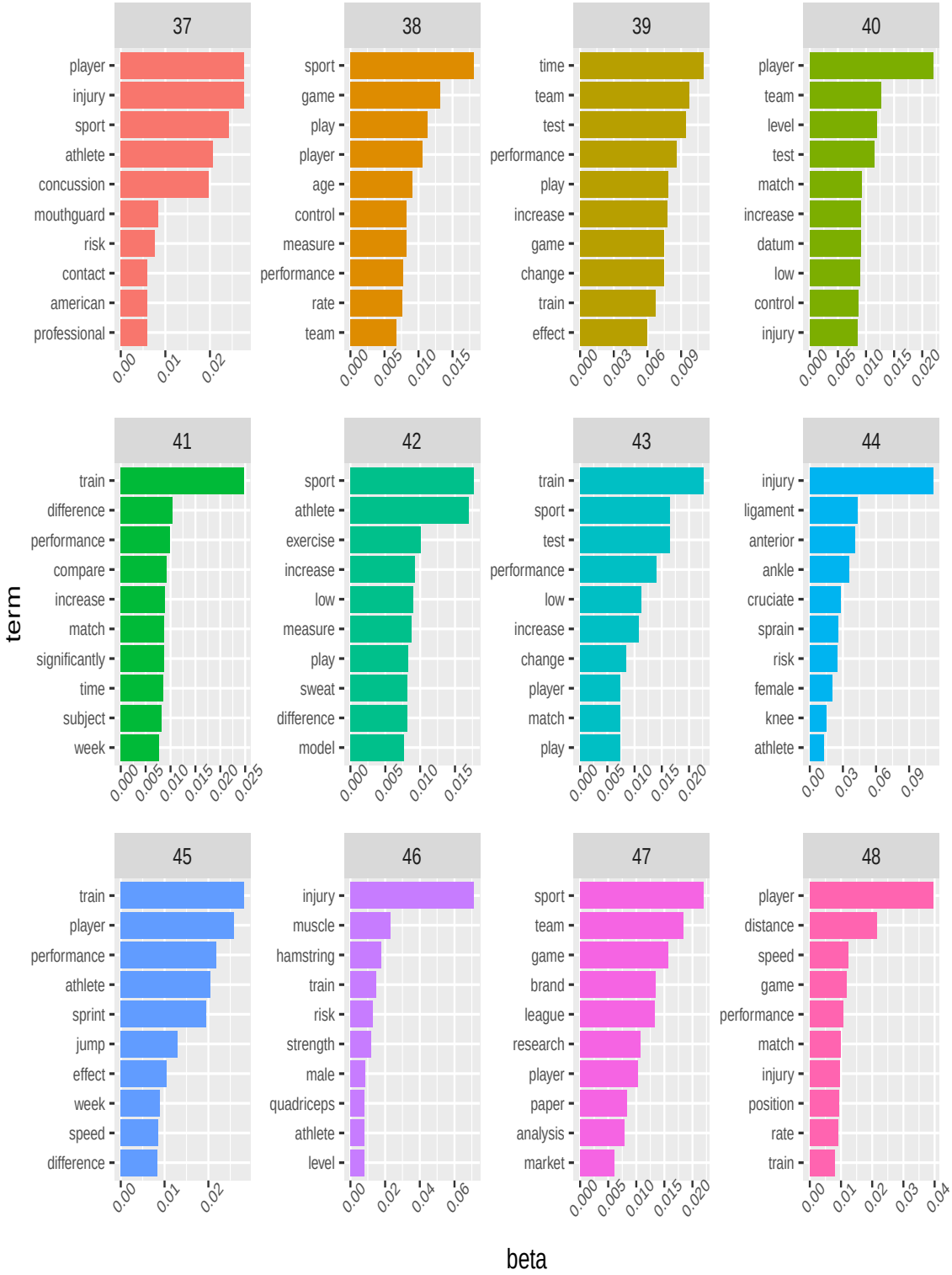


Figure 2.25: Probability of top 10 "terms" for topics 37-48 (x-axis is the probability of a word given a topic i.e. $P(words|topic)$)

- to explore the term frequencies of uni-gram and bigram;
- to fit an LDA model (with tuning) to find an optimal number of topics;
- to visualise the composition of words for each topic;
- to perform document clustering;
- to visualise the network of words and topic to see how each topic are connected to each other based on the words within each topic.

2.5.1 litReview: Shiny App Demo

Screen shots of the litReview app are now given to highlight the functionality of the app and the options available to investigate different configurations.

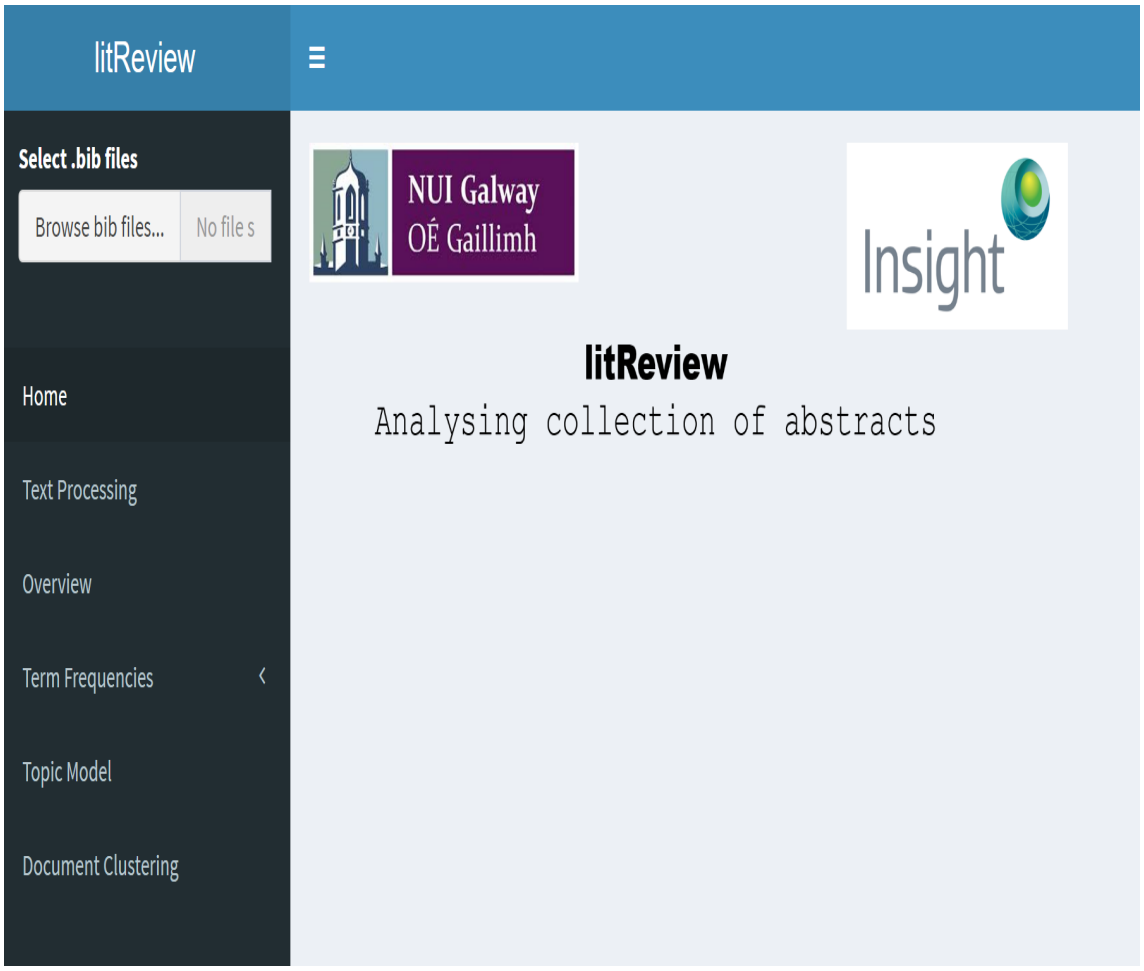


Figure 2.26: litReview: Shinyapp screen-1 (Users can import bibtex files into the app for further analysis, note that the bibtex files must contain abstracts)

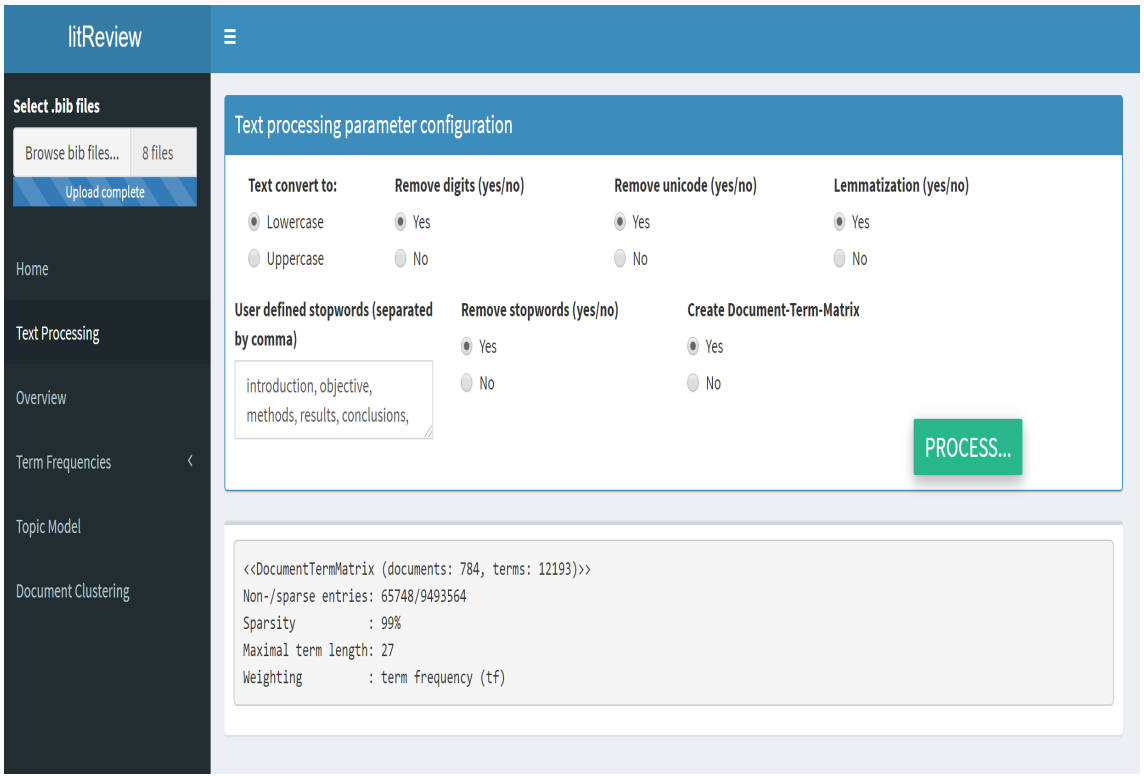


Figure 2.27: litReview: Shinyapp screen-2 (Options to configure text pre-processing and removing user defined common words, and then create Document Term Matrix)

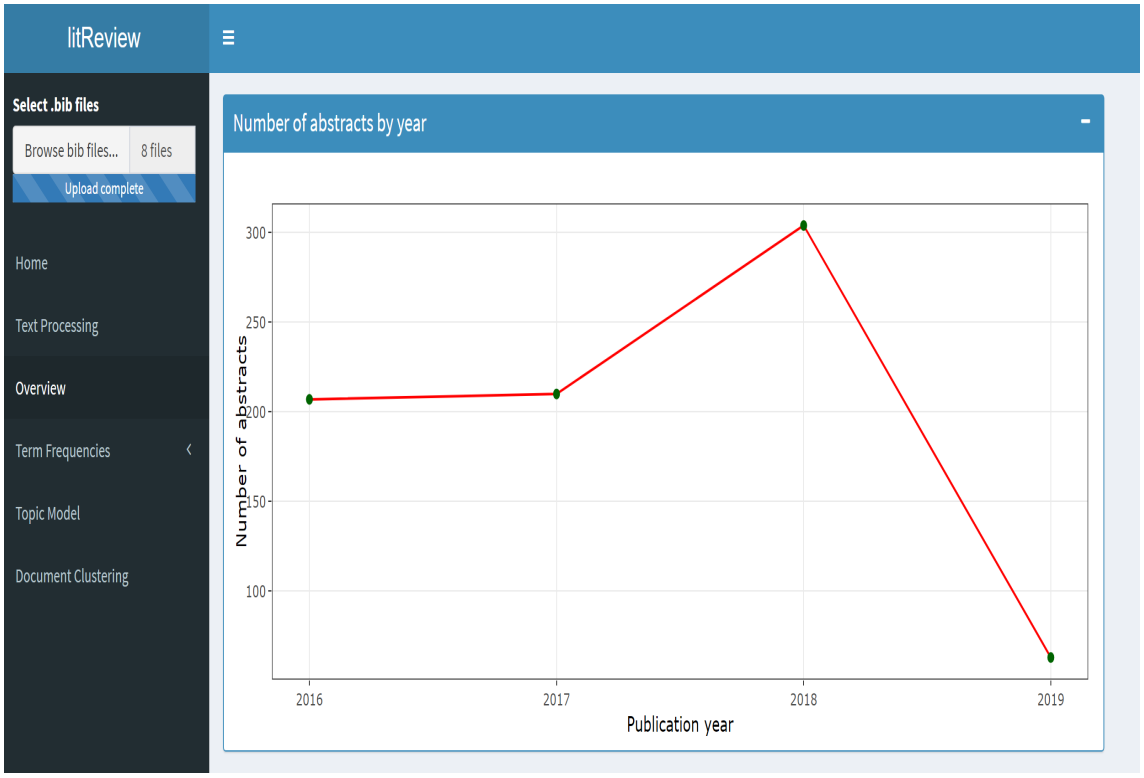


Figure 2.28: litReview: Shinyapp screen-3 (Plotting number of abstracts against publication years)

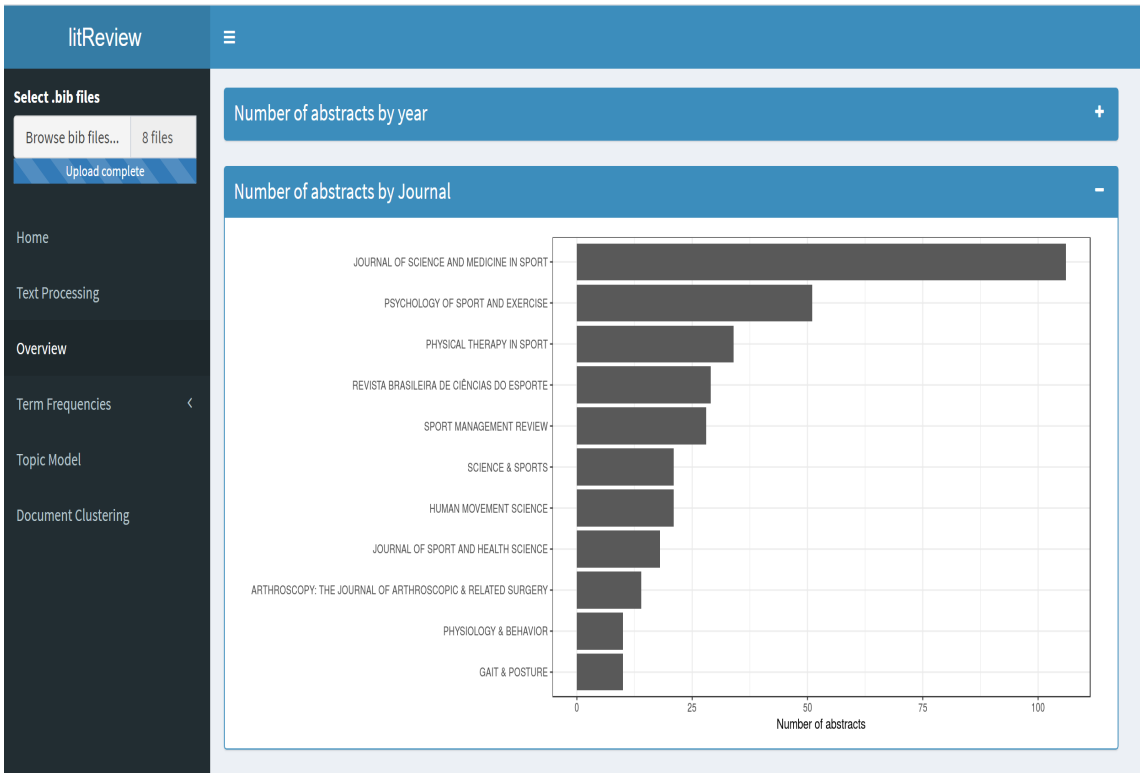


Figure 2.29: litReview: Shinyapp screen-4 (Finding top publication sources)

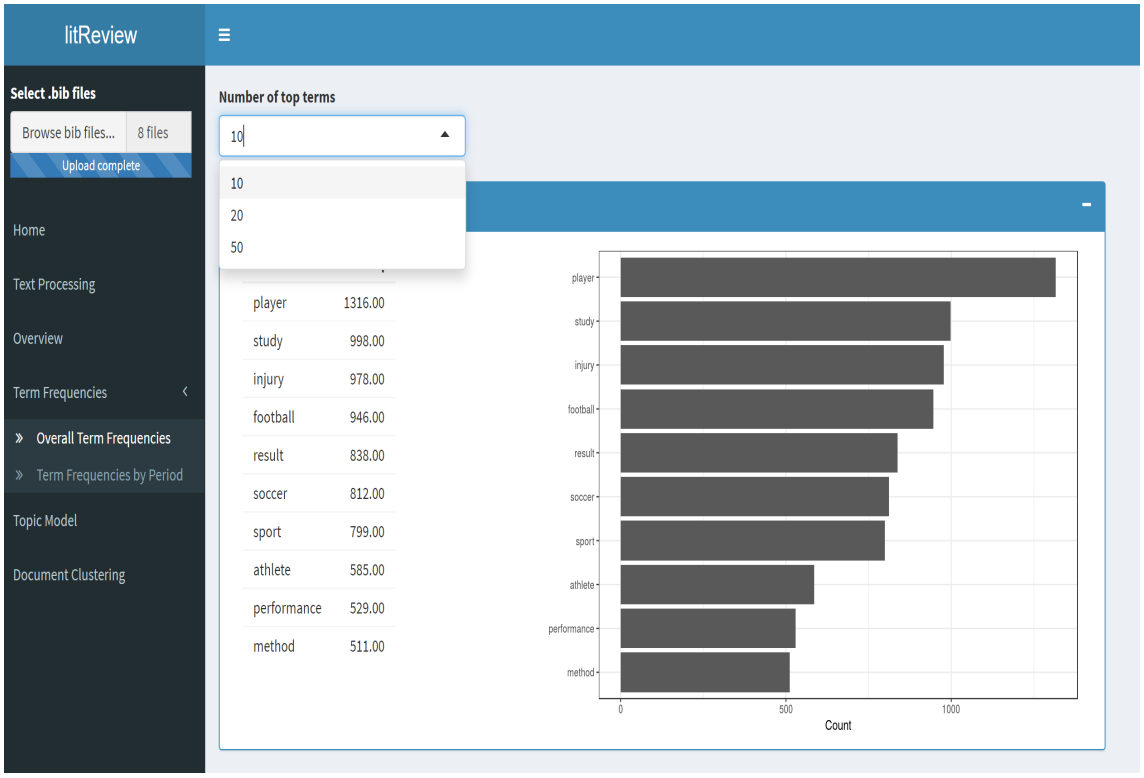


Figure 2.30: litReview: Shinyapp screen-5 (Most frequently used words across all abstracts)

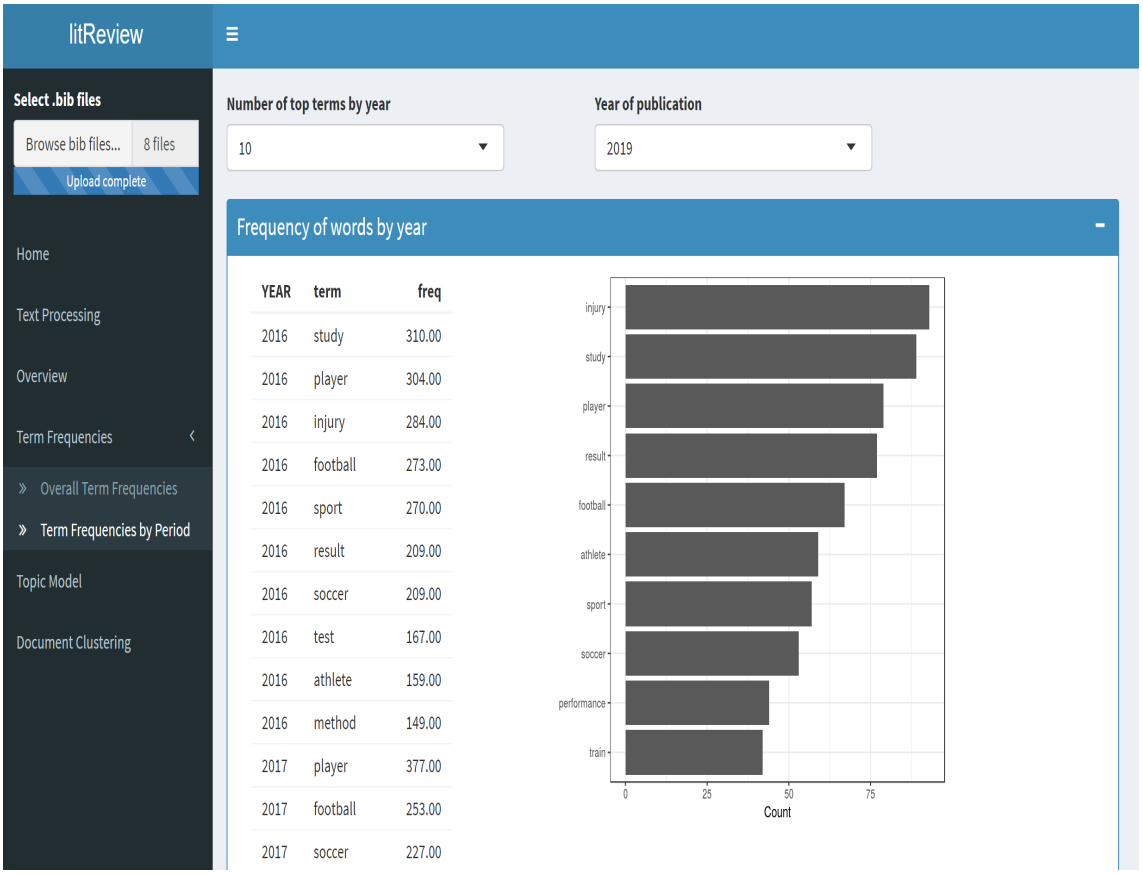


Figure 2.31: litReview: Shinyapp screen-6 (Most frequently used words across all abstracts across different years)

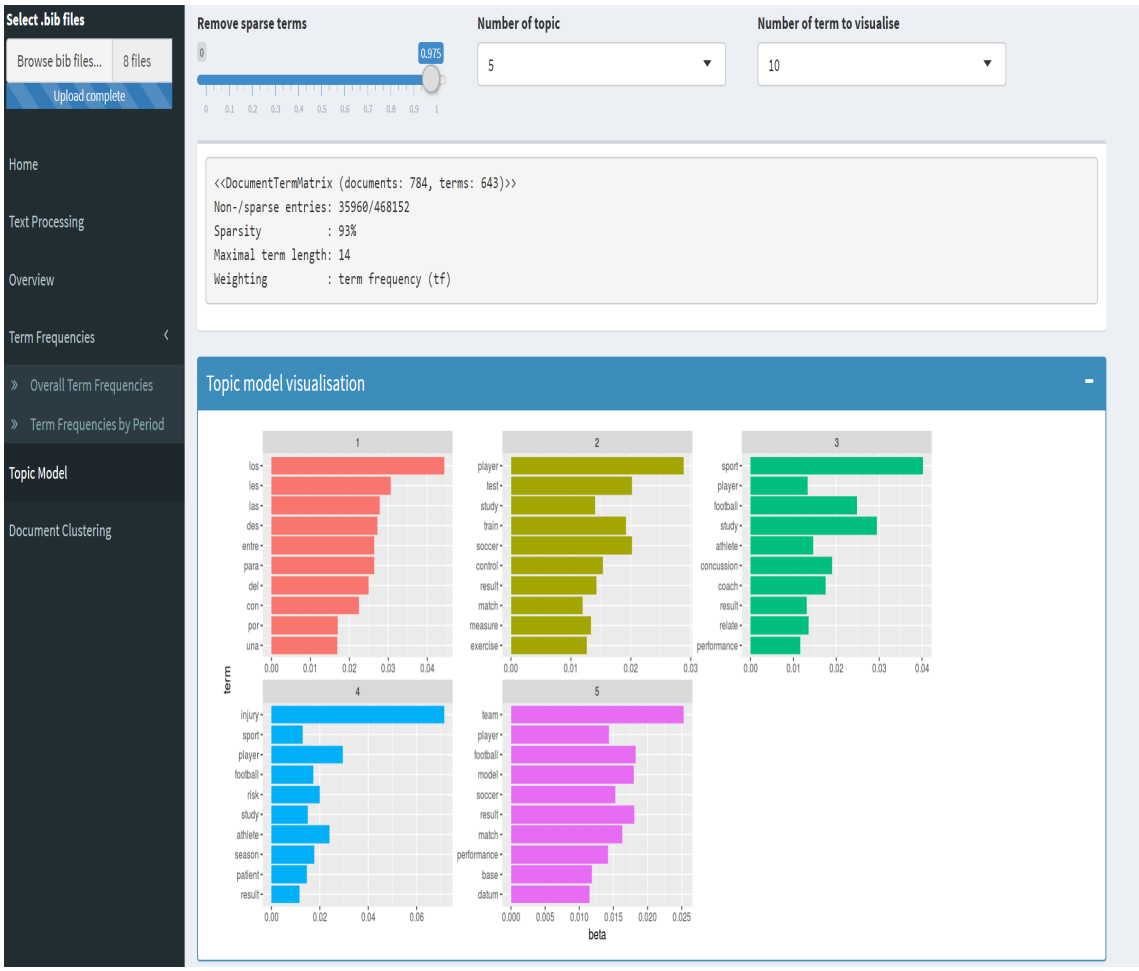


Figure 2.32: litReview: Shinyapp screen-7 (LDA modelling with options to interact with the number of topics, number of words to visualise and sparsity parameter selection to remove sparse words)

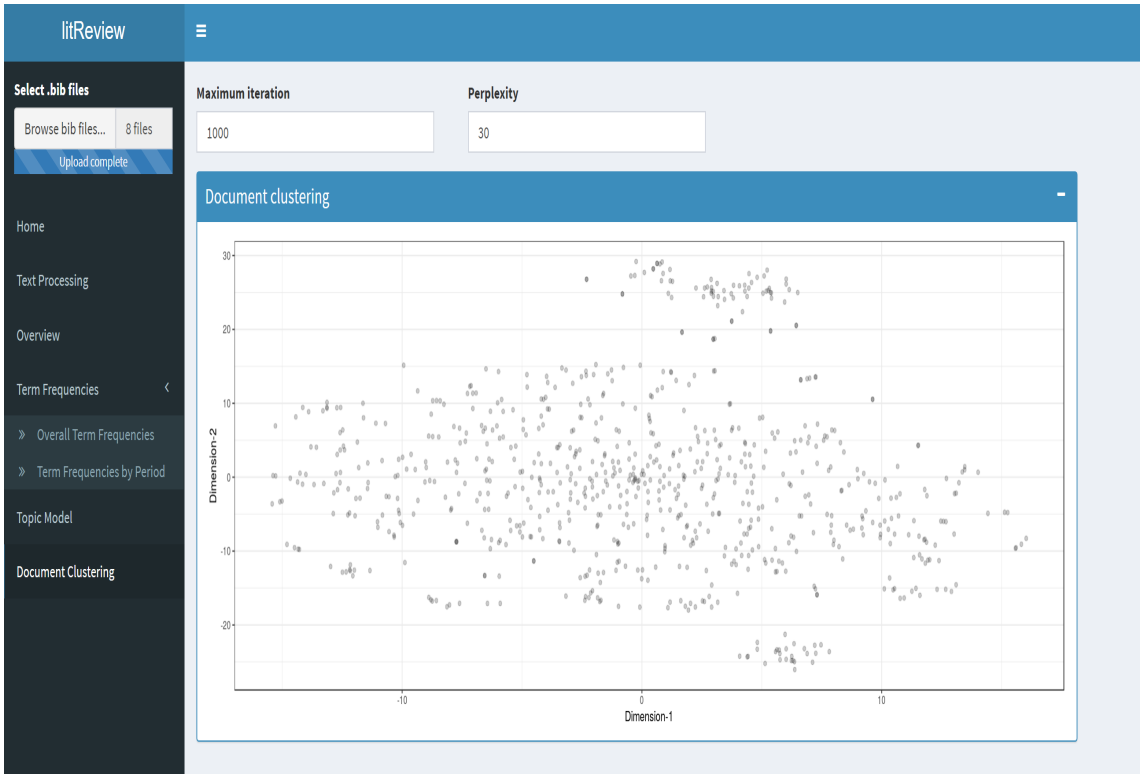


Figure 2.33: litReview: Shinyapp screen-8 (Example of document clustering using t-sne algorithm)

2.6 SUMMARY

A new approach to perform a literature review using topic modelling is presented in this chapter including a publicly available web application. In theory the approaches presented in this chapter can be used to augment a traditional literature review in any research domain where publications are available in electronic format.

The developed web application will enable non-technical users to quickly summarise a collection of many abstracts, possibly thousands. The summarised results will give an indication of the common sub research themes of relevance in the specific domain based on the composition of words within all the abstracts in the final set. This approach, coupled with domain knowledge and of course reading the key papers of interest, should make the difficult task of dealing with the vastly increasing number of published research more manageable.

The concept of a statistical depth function for detecting outliers in multivariate data was introduced briefly in this chapter and an overview of the literature relating to this topic was presented. In the next chapter statistical depth functions will be explored in more detail and a novel extension will be presented in the context of outlier detection.

3 OUTLIER DETECTION IN MULTIVARIATE DATA

3.1 INTRODUCTION

There is a natural tendency to be attracted towards things that appear as 'strange' or 'different'. For example, in sports spectators usually notice a player who played particularly well or particularly poorly compared to the squad as a whole. When watching a game these phenomena are visible to the viewers. Identification of atypical (good or bad) player performance is more challenging when considering player data alone (e.g. running speed, distance covered with and without possession of the ball). The level of complexity in identifying an atypical performance increases as the numbered of measured characteristics increases.

Data visualisation is an intuitive and effective way to represent data but it is challenging to visualise multivariate data in a meaningful way when the size of the data is vast. Dimension reduction techniques are a popular and useful technique allowing visualisations of a smaller number of composite scores derived to represent as much of the variability in the larger set. If multivariate data could be converted into a single score (e.g. a single player performance score) then it should be possible to distinguish atypically good and poor performance based on visualisations of the marginal distribution of this single variable.

In this chapter a new outlier detection method is presented which is an extension of a classical data depth approach for outlier detection in multivariate data. The chapter begins with a discussion on outliers and anomalies followed by an introduction to the concept and properties of a statistical depth function. A newly proposed modified depth function is then presented based on Mahalanobis Distance.

The results of a comprehensive simulation study, using benchmark data constructed to represent real-world data, are given to show the performance of the proposed approach.

3.2 OUTLIERS FROM A STATISTICAL AND COMPUTER SCIENCE PERSPECTIVE

As outlined in Chapter 2, in Statistics extreme, atypical, unexpected of unusual observations are termed as "outliers" while in Computer Science the term "anomaly" is more popular. For a comprehensive overview see [BL74; Haw80; RL05; BC83; HA04; CBK09; GU16].

Though these two terminologies are used interchangeably in the literature, there is a conceptual distinction between them. In statistics, one of the earliest definitions of "outlier" was 'an observed value of a random variable in a dataset that deviated remarkably from the majority of the points present in the same variable' [Gru69]. The noun *anomaly* comes from the Greek word *anomolia*. The literal meaning of it is *uneven* or *irregular* and is used in a more general sense than outlier, which is primarily a statistical concept.

An outlying observation could be due to an extreme manifestation of random variability associated with the variable of interest. Alternatively, the outlying observations could arise due to an experimental procedure or recording error or an error in the calculation of the values (e.g. a miscoded GPS signal due to poor calibration). If the outlying observation occurs due to random variability, the data point should be retained for all subsequent analyses, whereas, if it is due to a recording error (non-random) then it can be discarded from a subsequent analysis [Gru69].

From a statistical perspective, the definition of an outlier by Grubbs, implies the notion of *distance* when defining and identifying outlying observations in data. In the case of data with one random variable, the distance from its centre is easy to calculate and to understand how far an individual data point is from the centre of the distribution of the random variable. Let's consider X to be a random variable from a random sample of size with n with observations $x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \cdots \leqslant x_{(n)}$ to be the ordered value of the random variable X. If the largest value $x_{(n)}$ is a doubtful value (potential outlier), then according to Grubbs [Gru69], the test criterion T_n:

$$T_n = \frac{x_{(n)} - \bar{x}}{s} \tag{3.1}$$

where s is the sample standard deviation, can be compared with a theoretical value under a null hypothesis that $x_{(n)}$ is not an outlier at a pre-selected level of significance α , (e.g. $0.05, 0.01$ etc. If the calculated value is larger than theoretical value under the null hypothesis, then $x_{(n)}$ considered as an outlying observation. On the other hand if the smallest point $x_{(1)}$ is the observation under consideration then, the test criterion is defined as:

$$T_1 = \frac{\bar{x} - x_{(1)}}{s} \tag{3.2}$$

and the same procedure applies as of $x_{(n)}$. As per the Grubbs criterion, the outlying observation could only occur in either of one extremes (smallest or largest). If the distribution of the underlying random variable has multiple modes, then this method fails to identify outlying data points with respect to a local mode.

Later on in the mid 70's Tukey proposed a graphical tool, namely the *boxplot* [Tuk77] for exploratory data analysis in the univariate case. The *boxplot* is well known and a popular method for visualising and exploring the sample distribution along with highlighting potential outlying data points of a univariate random variable. Typically a *boxplot* is constructed by displaying a five number summary: minimum, maximum, median, first quartile (Q1) and third quartile (Q3). Apart from these five numbers, potential outlying observations are also marked based on a measure of distance namely $1.5 \times \text{IQR}$; IQR is the inter quartile range $(Q3 - Q1)$. In a *boxplot*, the lower edge of the box is the first quartile and upper edge is the third quartile. The median is

placed within the box, and a line drawn from upper quartile to the maximum data point within $1.5 \times \text{IQR}$ distance, a similar line is drawn from first quartile to the minimum data point within $1.5 \times \text{IQR}$ distance. All other points beyond $1.5 \times \text{IQR}$ are marked (displayed individually) as potential outliers Figure 3.1. The region $(Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR})$ contains 99.3% of the data points which is equivalent to the 3σ rule for Gaussian data.

The construction of a boxplot uses the distance from the centre (median) of the distribution

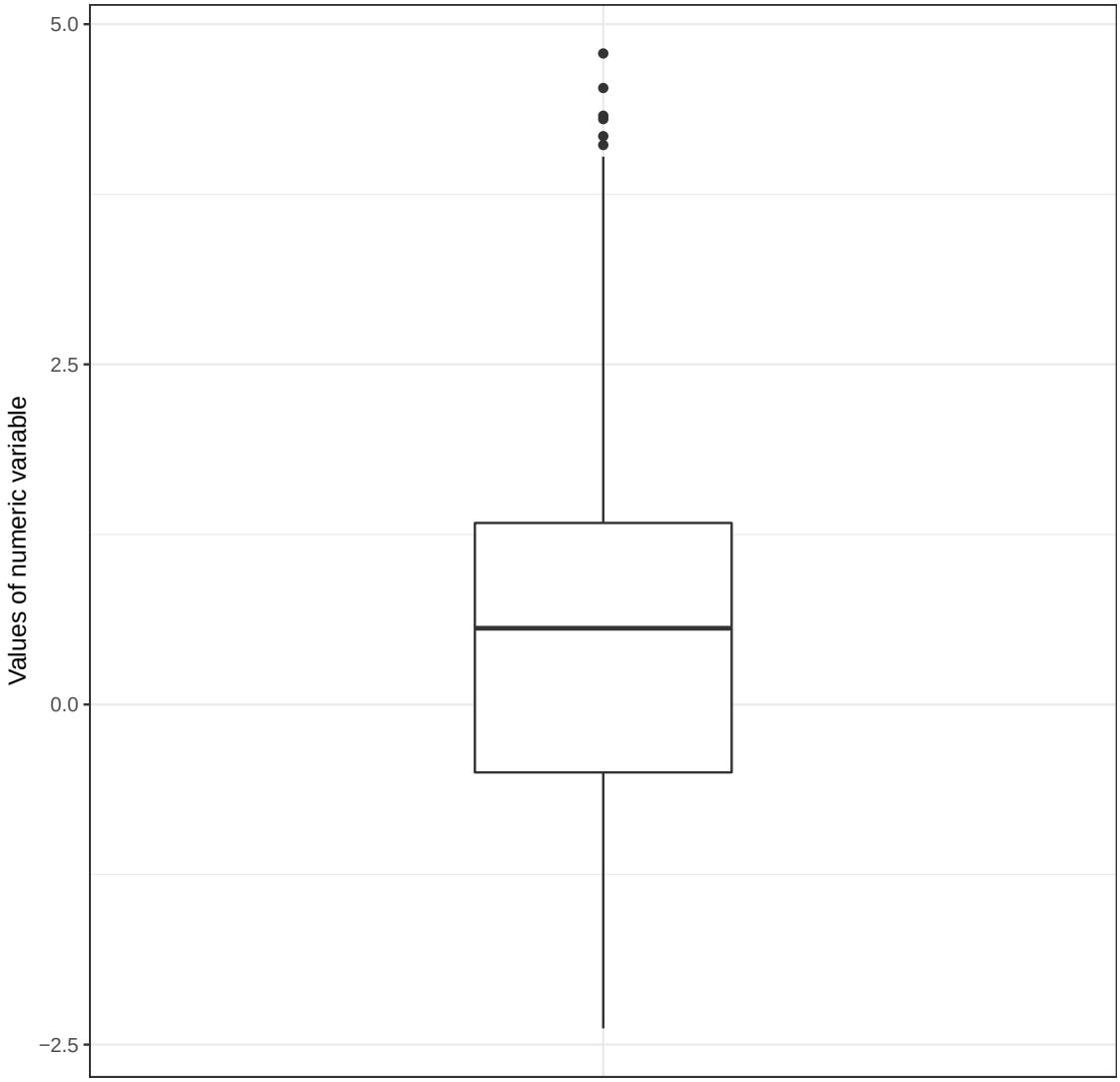


Figure 3.1: Tukey's Univariate Boxplot

as a method to identify potential outlying data points. In the case of a single random variable, measuring this distance is straightforward and uses the concept of *order statistics*; i-th smallest value of a continuous random variable is defined as i-th *order statistics*. In a univariate *boxplot* approximately 50% of the data are contained inside the box. The box goes from $[n/4]$-th quartile to $[3n/4]$-th quartile with the centre of box is at $[n/2]$ which is the median. Though there is no natural generalization of order statistics from univariate to bivariate (or multivariate) the notion of *halfspace depth*; where *halfspace depth* is defined for a point $x \in \mathbb{R}^d$ is the smallest number of

points contain in a closed half space through the point x [Tuk75] is considered as a generalization of order in multivariate data. Roussseeuw et.al. [RRT99] used this notion of *halfspace depth* to propose a bivariate version of the *boxplot* and named it is a *Bagplot*. The primary component of a *Bagplot* is the *bag* which contains at least 50% of the data points, a *fence* which is a boundary between the inlier and outlier and a *loop* indicating data points outside the *bag* but inside the *fence*. The *bag* is essentially a convex polygon containing at least 50% of the points and inflating the *bag* by a factor of 3 to obtain the *fence*. The points that are outside the *fence* are marked as potential outliers. The outlying points indicated in red in Figure 3.2. The classification of outlying points is intuitive in the case of bivarate random variable and can be highlighted visually using a *Bagplot*.

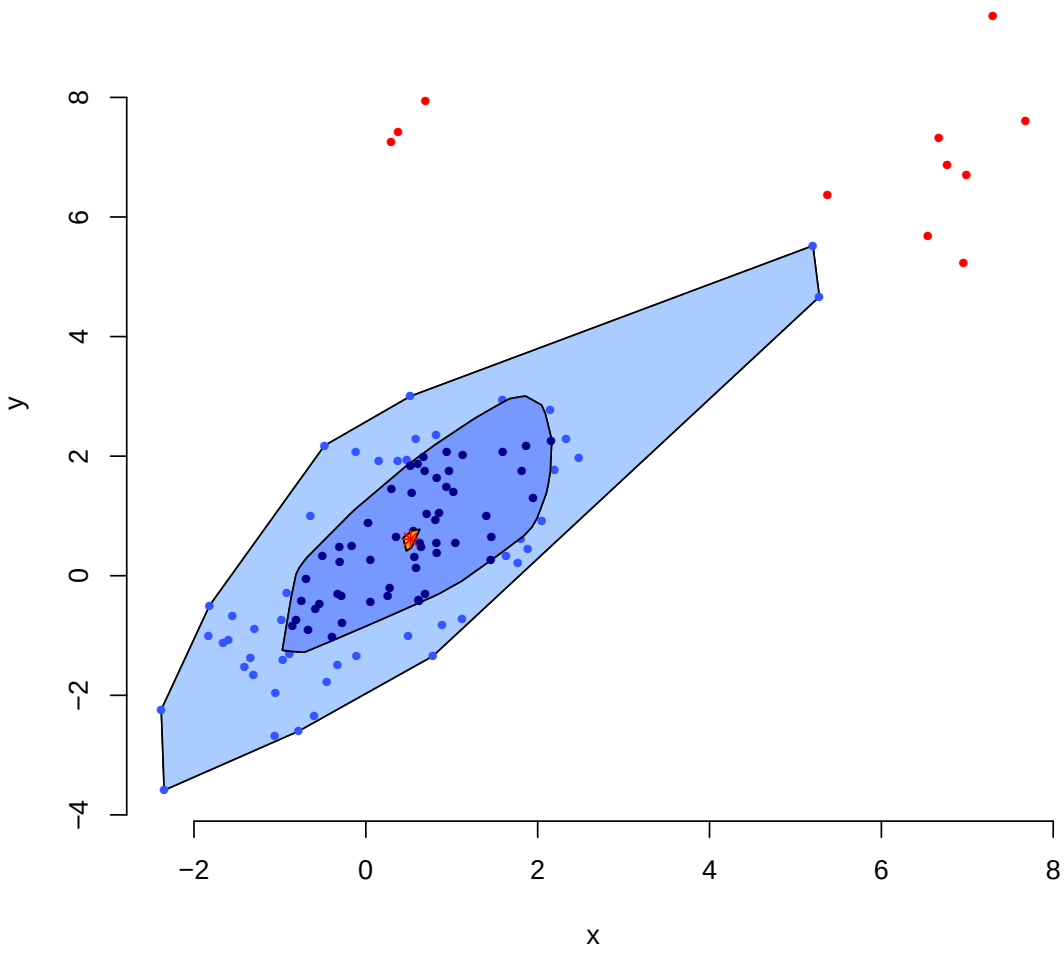


Figure 3.2: An example of a Bagplot using simulated bivariate data

The central idea in defining an outlying point is again the notion of distance; this distance could be simple Euclidean distance of a point from a specified location of the underlying probability distribution, or it could be a probability measure of a certain point based on an assumed

probability distribution for the random variables in question. If a data point falls in a very low probability density region then that point could raise a suspicion that it could have been generated from a different probability distribution [Haw80].

From the prospective of Computer Science, methods for anomaly detection can be categorized into six different approaches (Figure 3.3), each with its own set of underlying assumptions [CBK09].

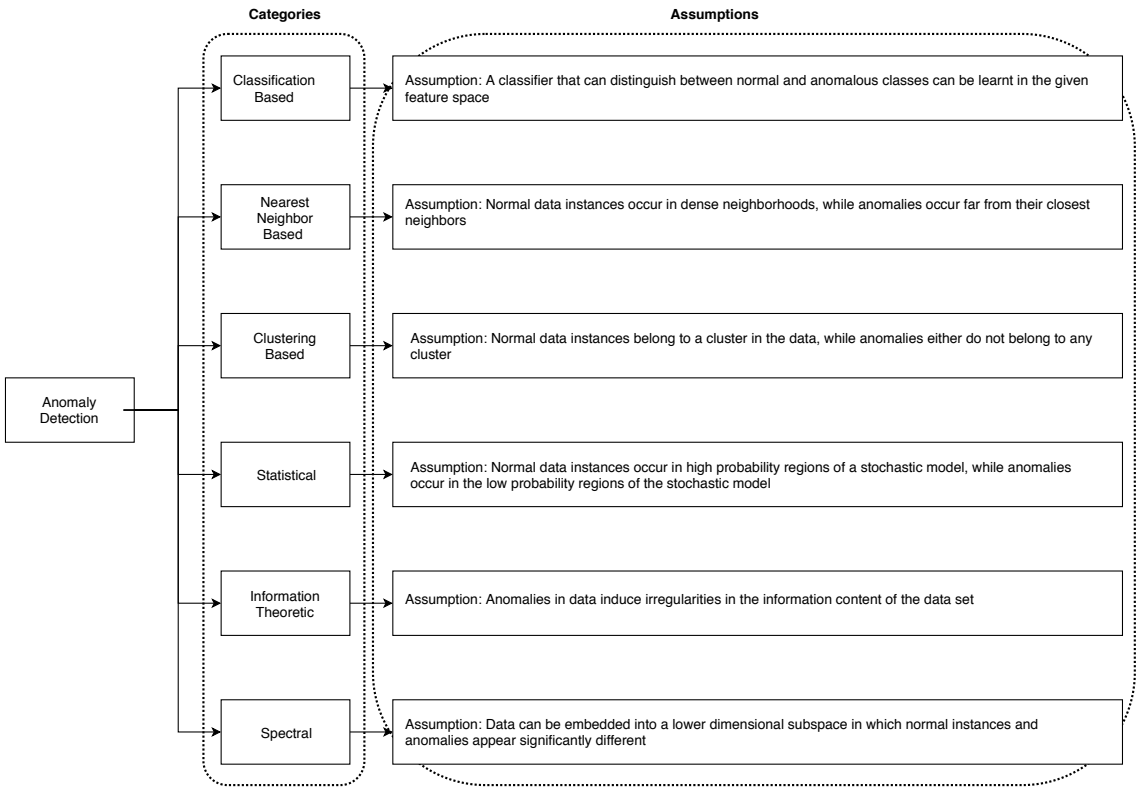


Figure 3.3: Categorisation of Anomaly Detection Methods [CBK09]

The definition of an anomaly in Computer Science is not unique and varies based on the application domain. Though it is more common that anomalous instances are considered as rare observations when compared to the dataset as a whole, in some application domains anomalous instances are those that occur more frequently in number. The assumption of rare instances might be appropriate in certain application domains. As a result, an anomaly detection technique developed for one application might not be applicable to other domain without modification or adjustment. The assumptions presented by [CBK09] might not always be true and there could be deviations. High density regions and short distance might not always indicate a non-anomaly, the local geometry of the distribution along with domain knowledge should be takes into consideration. Moreover, the choice of distance metric is challenging for mixed data types and measuring density for high dimensional data is not straightforward.

Visually identifying an outlying data point is often limited to three or fewer dimensions. In a d-variate random variable if outlyingness is checked for each component marginally, this will not be adequate; one observation could be outlying with respect to one of the d variables whereas the observation could be within the central region of another observation (e.g. normal body-

weight but high cholesterol).

In a multivariate data an outlying data point could be attributed to one variable only or a combination of several variables or it could be due to all variables in the dataset. Being an outlying data point in one variable does not necessarily indicate the same data point will be outlying with respect to another variable. A data point that is outlying only with respect to a subset of variables from a multivariate data could be categorised as "component-wise" outlier whereas an outlier with respect to all variable can be categorised as "Structural" outlier [AVY+09].

A "component-wise" outlier is assumed to be an independently contaminated data point from other variables in the dataset, in 2016 Rousseeuw and Van Den Bossche introduced "cell-wise" outliers where the individually independent contamination assumption is not required [RV16].

Approaches are needed therefore that take the underlying geometry of the random variable into account, moreover, approaches that could take local geometry into account while maintaining the global structure. In the remainder of this chapter a new algorithmic approach combining Mahalanobis distance and k nearest neighbours is developed to address this need.

Using this approach, a score derived from a modified statistical depth function is calculated for each observation of each multivariate random variable as a measure of *outlyingness* relative to its nearest neighbours and the direction of variability associated with the random variable of interest.

3.3 STATISTICAL DEPTH, OUTLYINGNESS FUNCTIONS AND PROPERTIES

A statistical depth function, a mapping from $\mathbb{R}^d$ to $\mathbb{R}$ (where d is the number of variables), is a method to generate a univariate score for multivariate data. Such a score can then be used to study properties of the distribution of multivariate random variables such as multivariate-location [DG+92] and confidence regions around multivariate-location [YS97].

Different types of depth functions and their mathematical properties are covered by Mosler [Mos13] and reviews on depth functions by Serfling [Liu06; Ser06] and Cascos [Cas10]. Statistical depth functions have been used in various statistical application such as clustering [Hob00; JVZ02; JCS+16; Jör04; DDP+07] and classification [GC05; MH06; CLY08; DG11; LCL12; DG12] problems, outliers detection, anomaly detection [CDP+08; DS10; Ven11]. The original data are mapped into a kernel induced feature space and then the spatial depth function has been used to detect outliers from multivariate data [CDP+08]. The complexity or limitation of this approach is that there is a need to choose a proper kernel to map the data into the feature space. Moreover each kernel function has its own hyper-parameters which also need to be tuned. In 2010, Dang [DS10] applied depth functions in outlier detection and discussed their robustness properties but they did not highlight the need to capture the local geometry of the distribution. The concept of data-depth has been extended to include outlier detection in functional data [LR09;

FGG08].

To visualise multivariate data the depth function has been also used in rainbow plots, bagplots and plots of functional data [HS10]. Depth functions have also been used in supervised classification problems [LCL12; PMD16].

More formally, let X be a univariate random variable from a random sample $x_1, x_2, ..., x_n$ of n observation. The ordering of the values of X is clearly understood and unambiguous. In the case of a continuous random variable the values can be ordered by increasing order of magnitude such as $x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant ... \leqslant x_{(n)}$. The subscript with parenthesis indicates the ordered value of the random variable X. In the univariate case the ordered values $x_{(i)}, i = 1, 2, 3, ..., n$ are defined as *order statistics*. The ordering of univariate values can also be done with reference to a specific value e.g. a; where a is any real number, and the absolute distance from that specific value to all other values. Consider a be a reference point, the value of X can then be ordered as $a \leqslant x_{(1)} \leqslant x_{(2)} \leqslant ... \leqslant x_{(n)}$. This type of reference point is useful in the multivariate context. Having the ability to order the values of a random variable, the intrinsic features of data [Bar76] can be explored in more detail. The order directly reflects 'extremeness' which leads to some of the well-known treatments of robust methods such as trimming and winsorization (e.g. trimmed mean, trimmed regression, winsorized mean, winsorized regression) [Bar76]

The generalisation of univariate order statistics to a multivariate context is not straightforward however. Tukey (1975) [Tuk75] in his seminal article proposed an alternative approach to define order in multivariate random variables where he introduced the concept of 'data-depth' as and ad hoc approach for ordering. The data-depth, or simply depth function, is a bounded non-negative function that maps from $\mathbb{R}^d \to \mathbb{R}$; (here $\mathbb{R}$ represents the set of real numbers), providing a centre-outward ordering of multivariate observations. In the univariate case, the order of the data points follows a linear pattern where the values can be organised easily by increasing order of magnitude, eventually corresponding to the cumulative distribution function (cdf) F of the corresponding random variable X. Thus for the univariate case, the ordering of X's and corresponding $p - th$ quantile function can be defined as,

$$Q(p) = inf\{x \in \mathbb{R} : p \leqslant F(x)\} \tag{3.3}$$

This definition of a quantile function cannot be generalised to $\mathbb{R}^d$ with $d \geqslant 2$. By introducing or orienting the center of the data cloud this can be compensated, which results in a centre-outward ordering of multivariate points with nested contours. For example, Tukey (1975) [Tuk75] defined location depth, also known as halfspace depth (HD) for a d dimensional multivariate data point $x \in \mathbb{R}^d$ as the smallest number of points contained in a closed halfspace through the point x, i.e.

$$HD(x; F) = D(x; F) = inf\{P(H) : H \text{ a closed halfspace }, x \in H\}, x \in \mathbb{R}^d \tag{3.4}$$

Here, HD; stands for halfspace depth, D; represents depth, F is the cumulative distribution function of a random variable X, and P is the probability measure over F. The above definition can be simplified for the univariate case where the location or HD of a univariate point x is given by:

$$min\{\#\{x_i \leqslant x\}, \#\{x_i \geqslant x\}\} = min\{F(x), 1 - F(x)\} \tag{3.5}$$

The depth-function therefore provides a centre-outward ordering of the values of multivariate random variable. The value of the depth-function decreases as the distance of a points increases from its central value. The maximum depth value is attained at the center, which is also known as a point of *maximal depth* or deepest point. The most central region or points also serve as a *median* of a multivariate random variable.

In summary:

- **Depth Function:** Let X be a d-variate random variable on $\mathbb{R}^d$ having a probability distribution function F. A non-negative bounded function $D(x, F)$ for each point $x \in \mathbb{R}^d$ is a mapping from $\mathbb{R}^d \to [0, 1]$ that provides a centre-outward ordering of the points is termed a depth-function, here $x \in \mathbb{R}^d$. The larger values of $D(x, F)$ represents the higher centrality of the points with the points having maximal depth value is considered as the center of the distribution.

- **Outlyingness Function:** A function $O(x, F)$, is an equivalent of depth-function $D(x, F)$ of a multivariate random variable X on $\mathbb{R}^d$, that maps a point $x \in \mathbb{R}^d \to \mathbb{R}^+$ which also provides a center-outward ordering of the multivariate data points. The higher value of $O(x, F)$ represents higher outlyingness. Mathematically $O(x, F) = 1 - D(x, F)$

There have been several types of depth-functions proposed such as, convex hull peeling depth [Bar76], Oja depth [Oja83], simplical depth [Liu+90] with the aim to define order in multivariate observations and to study other properties of the distribution of multivariate random variables. Though there have been many depth functions proposed there was a lack of clear definition of the unified properties of a such functions. Zuo & Serfling [ZS00] outlined four criteria that a depth-function should possess, namely:

1. P1. **Affine invariance**: The depth $D(x, F)$ of a point $x \in \mathbb{R}^d$ should not depend on the underlying scale of measurement of the random variable X. That is the deepest point should not change based on the coordinate system of the random variable X

2. P2. **Maximality at Centre**: In a distribution F of a random variable X with uniquely defined centre, the depth $D(x, F)$ should be a maximum at the centre. The central point usually serves as the point of symmetry of the distribution.

3. P3. **Monotonicity relative to deepest point**: If the point $x_m \in \mathbb{R}^d$ with maximal depth value of $D(x_m, F)$, then for any $\alpha \in [0, 1]$, $D(y, F) \leqslant D(x_m + \alpha(y - x_m), F)$ for $y \in \mathbb{R}^d$. That is, if a point moves away from the centre then the corresponding depth-value of that point decreases.

4. P4. **Vanishing at infinity**: This is related to property P3; for a point $x \in \mathbb{R}^d$ the depth value should go towards zero while $||x|| \to \infty$

Based on the proposed criteria P1 to P4, a general definition of depth-function, known as *statistical depth function*, [ZS00] is given as:

- **Statistical depth function:** Let $\mathcal{F}$ be the family of distribution on $\mathbb{R}^d$, and F_X be the distribution of a random variable X, if the mapping $D(.;F) : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}$ is bounded non-negative and satisfies properties P1 to P4, the mapping $D(.;F)$ is called a statistical depth function. The empirical version can be written as $D(.;,\hat{F})$ where $\hat{F}$ is the empirical distribution function.

The general properties of a statistical depth function and its definition enables the construction of various types of statistical depth functions. For a more detailed discussion refer to Serfling [ZS00]. In this thesis only the *Type C* statistical depth function based on Serfling [ZS00] is considered and a specific version of it is highlighted based on the use of distance between points. Compared to other types of statistical depth function constructors the *Type C* is more intuitive to understand and to define the distance metric for a multivariate observations X.

- **Type C depth function:** Let X be a random variable on $\mathbb{R}^d$ with a probability distribution F and a function $O(x,F) : \mathbb{R}^d \to \mathbb{R}^+$ is the outlyingness of a point $x \in \mathbb{R}^d$ with respect the centre of the distribution F. A corresponding bounded function $D(x,F)$ defined by

$$D(x,F) = (1 + O(x,F))^{-1} \tag{3.6}$$

  is called *Type C* depth function. Based on the type of distance metric used in $O(x,F)$ the specific version of *Type C* is different.

Projection Depth: Let X be a random variable on $\mathbb{R}^d$ with a distribution function F, then for a point $x \in \mathbb{R}^d$ is the worst case outlying point with respect to the univariate median with any one-dimensional projection of X with u a projection vector of unit norm; $||u|| = 1$, that is,

$$O(x,F) = \sup_{||u||=1} \frac{|u^t x - Med(u^t X)|}{MAD(u^t X)} \tag{3.7}$$

The corresponding bounded function $PD(x,F)$ is known as projection depth function [ZS00]:

$$PD(x,F) = \frac{1}{1 + O(x,F)} \tag{3.8}$$

where PD stands for Projection Depth, $Med(x)$ is the median of X and $MAD(x)$ is the median absolute deviation from the median. The projection depth function is straightforward for univariate random variables. When considering multivariate random variables however, the direction of projection could be any random direction and the depth of a point is not uniquely defined. In practice multiple depth-values have been calculated and then averaged over different directions. The projection depth function satisfies the properties of a statistical depth function P1 to P4 [ZS00; Mos13].

L_2-Depth: For a random variable X on $\mathbb{R}^d$ with a probability distribution function F. The average depth of a $x \in \mathbb{R}^d$ is defined as:

$$D_{L_2}(x,F) = \frac{1}{1 + E||x - X||} \tag{3.9}$$

Here E represents the mathematical expectation of a random variable. The corresponding sample version can be written as:

$$\hat{D}_{L_2}(x, F) = \frac{1}{1 + 1/n\Sigma_{i=1}^{n}\|x - x_i\|} \tag{3.10}$$

where n is the sample size. The L_2-depth function does not satisfy the property P1 (affine invariance) moreover the orders induced by this depth do not produce a sensible ordering of the dispersion [ZS00; Mos13]. L_2-depth does satisfy other properties P2 to P4 but under certain transformation an affine invariance version of L_2-depth can be constructed.

Mahalanobis Depth: Instead of using Euclidean distance between two points of a multivariate random variable X, Mahalanobis [Cha+36] defined a new distance metric between two points by taking the direction of the variance of the random variables into consideration. For any two points x and y in $\mathbb{R}^d$, the distance between these two points with respect to a positive definite matrix M of order $d \times d$ is:

$$d_M^2(x, y) = (x - y)^t M^{-1}(x - y) \tag{3.11}$$

Consider a bivariate random variable X with the following characteristics:

$$\mu = (0.5, 0.5); \Sigma_{2\times 2} = (0.3, 0.2; 0.2, 0.3)$$

Also consider two more points $B = (0, 1)$ and $C = (1.5, .5)$. We are interested in calculating the distance of B and C from the mean vector μ. The data along with the points B, C and μ are represented as green, blue and red respectively in Figure 3.4. The Mahalanobis distance between two points in multivariate space takes into account the direction of maximum variance whereas Euclidean distance ignores this property.

Using Mahalanobis distance between two points an equivalent statistical depth function is defined by Zuo and Serfling [Liu92; ZS00] as:

$$mahD(x, F) = \frac{1}{1 + d^2_{\Sigma_F}(x, \mu(F))} \tag{3.12}$$

Here, F is the distribution function of a d-variate random variable X with mean vector $\mu(F)$ and covariance matrix $\Sigma(F)$. In 1993 Liu and Singh [LS93] noted that the mean vector and covariance matrix is not robust and fails to attain the maximum at the centre of the distribution. To overcome this non-robust property, a robust location vector and covariance matrix could be used such as minimum volume ellipsoid (MVE) or minimum covariance determinant (MCD) [RL87; LR+91]. The Mahalanobis depth function satisfies all the required properties P1-P4 to be a statistical depth function

The depth function enables the properties of distribution of a multivariate random variable to be studied; detailed theoretical discussions on depth functions are covered by Zuo and Serfling [ZS00; Mos02]. In the following section, an overview of the use of depth function in anomaly (outlier) detection will be given.

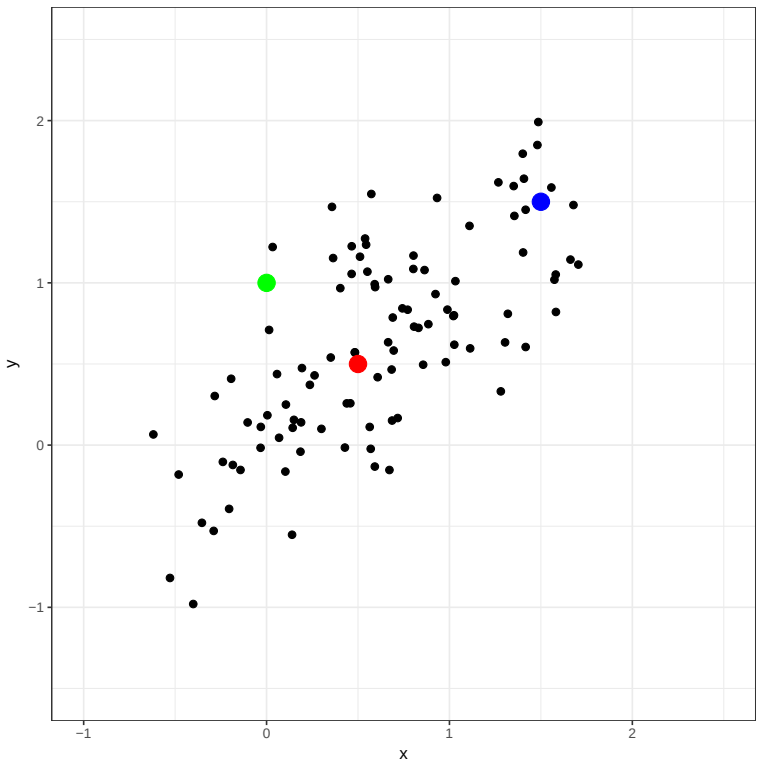


Figure 3.4: Demonstration of the difference between Euclidean distance and Mahalanobis distance. The Euclidean distance of the point B; "green" to the sample mean μ; "red" is 0.71 and the distance of the point C; "blue" to the sample mean μ; "red" is 1.4. Whereas the Mahalanobis distances of points B and C from mean are 5 and 4 respectively

3.3.1 Statistical Depth Function in Anomaly Detection

It is not always straight forward to generalize univariate outlier/anomaly detection methods to multivariate data. Moreover, extending univariate methods to a multivariate context creates challenges due to lack of natural ordering and limited visualisation methods for such data. The first principal component score from the principal component analysis (PCA) has been used in social science and public health research to construct a univariate score from multivariate data: specifically PCA has been used to construct an asset index to understand the socio-economic categories based on the index and it is one of the popular methods to do so [FP01]. In a simulation experiment to assess whether a PCA based index constructed from multivariate data can retain the order of individual data points, the results showed that only less than 5% of the individual data preserve their actual order. This simulation study suggests that a PCA based index from multivariate data may not be suitable and should be used with caution [SNA+14].

Having a univariate depth score from multivariate data points enables us to define the outlyingness of that point. The points that have very small values of depth score are considered as outlying points with respect to the centre of a multivariate random variable. This approach works based on computational geometry where a layer is constructed based on a convex hull of each multivariate point, the outer layer is more outlying than the inner layer [TIR98].

A depth function such as halfspace depth that satisfies properties P1-P4 [ZS00] of a statistical depth function but produce more false positive outlying points compared to other depth functions [DS10]. The definition of halfspace depth involves infinitely many directional projections, the exact computation for a random variable on R^d with $d > 2$ is computationally expensive. Moreover, as halfspace depth is calculated based on an implicit assumption of a global centre (uni-modal distribution) of a multivariate random variable, it can only identify outlying points with respect to that centre. If a random variable has multiple local groups then halfspace depth is unable to capture the local structure.

The Mahalanobis distance based statistical depth function has been used in outlier detection because of it satisfying mathematical properties P1-P4 above. The Mahalanobis depth function based approach produces low false positive outlying points compared to other depth functions [DS10]. The covariance matrix used in Mahalanobis distance is however highly affected by extreme points present in the dataset. To overcome the effect of extreme points in estimating the location and scale parameter for a random variable, a variety of robust approaches have been proposed [DG93; Atk94; RW96; RD99]. As pointed out by Zuo & Serfling [ZS00]; a statistical depth function should possess four properties, the Mahalanobis distance based depth function should satisfy those properties and the fast and robust covariance matrix estimation make it feasible to use this method for high dimensional data. Like the halfspace depth, Mahalanobis depth also calculates the depth of a data point from the center of the distribution. As a result, this approach can only detect outlying points that are far away from the center of the distribution and completely ignores points that are outlying with respect to its neighbouring groups i.e. it produces centre-outward ranking of multivariate data points which completely ignores local groupings. A point could be outlying with respect to its neighbouring groups but might not be outlying with respect to the global centre of the distribution. The existing depth functions

that are used in anomaly detection work better with respect to global centre of the distribution ignoring local grouping [Liu92; ZS00; Zuo+03], spatial depth [VZ00; Ser02].

Mahalanobis distance is used in outlier detection without being transformed into a statistical depth function. In 2016 Goldstein [GU16] discussed the use of Mahalanobis distance in outlier detection in a comparative study of unsupervised outlier detection algorithms. In particular, Mahalanobis distance was used to calculate an outlyingness score of a multivariate data point from its nearest cluster; the clusters were identified by applying k-means clustering algorithm at the first stage. The potential limitation of this approach is that at the first stage a clustering algorithm needs to be chosen and then the covariance matrix estimated for each of the clusters separately in order to estimate the outlyingness score of a point.

In the following section, a new modified depth function is introduced which will be able to capture local outlyingness while calculating depth of a multivariate observation. To do so one of the four desired properties of a statistical depth function needs to be modified to reflect the local structure of the distribution.

3.4 PROPOSED MODIFIED MAHALANOBIS DEPTH IN ANOMALY DETECTION

A new algorithmic approach is presented by modifying the original definition of Mahalanobis depth that is able to detect outliers in multivariate data with respect to the local neighbouring center of the distribution. Let X be a d variate random variable with probability distribution F_X, where n is the number of observations in the dataset, m_k is the mean of the k nearest neighbour points of query point x_q and S the covariance matrix of X with respect to location m. The classical definition of Mahalanobis depth is:

$$MD_{x_q} = \frac{1}{1 + (x_q - m)^t S^{-1} (x_q - m)} \tag{3.13}$$

which measures the depth of a query point x_q from the centre of the data. As a result it ignores the local structure and will fail to detect local anomalies. The range of MD_{x_q} is between 0 to 1, the smaller the value the higher the indication of outlying point. To take the local structure into account, the above definition of Mahalanobis depth is modified using the mean of the k nearest points as:

$$kMMD_{x_q} = \frac{1}{1 + (x_q - m_k)^t S^{-1} (x_q - m_k)} \tag{3.14}$$

This modified Mahalanobis depth will consider the global and local structure of the data. The covariance matrix has been calculated using the global mean whereas the distance of the point in question has been calculated from the local mean. The range of the modified Mahalanobis depth also lies between 0 and 1, with smaller values indicating anomalous points. The steps for calculating the modified Mahalanobis depth are as follows:

- For each data point $x_q; q = 1, 2, 3, \cdots, n$ find the k nearest points based on Euclidean distance
- Calculate the mean m_k of the k-nearest points
- Estimate the covariance matrix S from the data using the global mean m
- Calculate the modified Mahalanobis depth kMMD for each data point with respect to m_k and S

The proposed modified Mahalanobis depth function satisfies all four desired properties of a statistical depth function proposed by Zuo & Serfling [ZS00] but with respect to the local centre of the distribution. That is:

- P1: Affine invariance; this property does not depend on the centre of the distribution
- P2: Maximality at the local centre; that is multiple deepest point
- P3: Monotonicity relative to local deepest point
- P4: Vanishing at infinity with respect to the corresponding local centre of the distribution

In the proposed approach we do not need to calculate clusters beforehand which mitigates the steps to select clustering algorithms and their respective hyper-parameter tuning. Also, as we do not need to calculate clusters, we are no longer required to estimate multiple covariance matrix for each clusters, rather only one covariance matrix for the entire data will adequate. There is no need to use k-mean algorithms as we are not applying the clustering steps, rather we are using a mean of k-nearest neighbour points. These above points are primary differences with the approach discussed in [GU16]; especially with clustering based approach where Mahalanobis distance is being used to detect outliers.

3.5 EVALUATION OF PROPOSED APPROACH

To evaluate the proposed approach, several datasets were considered including simulated data, artificial benchmark data and benchmark data derived from a real world dataset. In all of the evaluations the outlier status of each observation is known, i.e. an observation is or is not an outlier. For each of the algorithms, we have used the Area Under the ROC curve (AUC) as an evaluation metric because to calculate the AUC we do not need a specific cut-off to decide whether a point is an outlier or not. A higher AUC value indicates better performance of an algorithm.

3.5.1 Simulation Study

To evaluate the proposed kMMD approach to detect outliers, simulated datasets have been generated to mimic configurations with global and local outliers. The kMMD approach has been evaluated on each of the datasets and compared with Mahalanobis depth (mahD) and k-nearest neighbour (kNN).

- Scenario-1: A bivariate dataset was generated with outliers and non-outliers present. The non-outliers were generated from a bivariate Gaussian distribution as: $N(n_1, [0,0]^t, I)$, $N(n_2, [0.5, 0.5]^t, I)$, $N(n_3, [-0.5, 0.5]^t, I)$, $N(n_4, [-0.5, -0.5]^t, I)$, $N(n_5, [0.5, -0.5]^t, I)$ where I is an identify matrix of order 2. A total of n_6 outliers was then generated from a bivariate Uniform distribution over the range of -10 to +10. That is, the outliers were generated from a bivariate Uniform distribution over the region of $[-10, 10]x[-10, 10]$. The total sample size of this dataset is $n = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$. Here N represents *normal distribution* and the n_i's represents the sample size of corresponding portion of the data.

- Scenario-2: A bivariate dataset was generated with outliers and non-outliers present. The non-outliers were generated from bivariate Gaussian distribution as: $N(n_1, [0,0]^t, I)$, $N(n_2, [3,3]^t, I)$, $N(n_3, [-3,3]^t, I)$, $N(n_4, [-3,-3]^t, I)$, $N(n_5, [3,-3]^t, I)$ where I is an identify matrix of order 2. A total of n_6 number of outliers were then generated from a bivariate Uniform distribution over the range of -10 to +10.

- Scenario-3: A bivariate dataset was generated with outliers and non-outliers present. The non-outliers were generated from a bivariate Gaussian distribution as: $N(n_1, [0,0]^t, I)$, $N(n_2, [5,5]^t, I)$, $N(n_3, [-5,5]^t, I)$, $N(n_4, [-5,-5]^t, I)$, $N(n_5, [5,-5]^t, I)$ where I is an identify matrix of order 2. A total of n_6 number of outliers was then generated from a bivariate Uniform distribution over the range of -10 to +10.

- Scenario-4: A bivariate dataset was generated with outliers and non-outliers present. The non-outliers were generated from a bivariate Gaussian distribution as: $N(n_1, [0,0]^t, I)$, $N(n_2, [7,7]^t, I)$, $N(n_3, [-7,7]^t, I)$, $N(n_4, [-7,-7]^t, I)$, $N(n_5, [7,-7]^t, I)$ where I is an identify matrix of order 2. A total of n_6 outliers was then generated from a bivariate Uniform distribution over the range of -10 to +10.

For each of the datasets, five different groups of points were generated from a bivariate normal distribution with a different location vector while the covariance matrix was fixed at the identity. Specifically, 200 points were generated from a bivariate normal distribution for each location vector (i.e. $n_1 = n_2 = n_3 = n_4 = n_5 = 200$) and 30 outliers (i.e. $n_6 = 30$) generated from a Uniform distribution. Each dataset contains a total of 1030 points where approximately 3% represents outlying observations. The distribution of both non-outliers and outliers are displayed in Figure 3.5 where the red points are the known outliers and the objective is to identify those points by applying the proposed kMMD approach.

For each of the simulated datasets, the area under the ROC curve (AUC) has been calculated for the proposed kMMD, Mahalanobis depth and kNN and the calculation has been repeated one thousand times to estimate the uncertainty in the AUC by calculating a bootstrap confidence interval for the true AUC for each scenario. For the proposed approach kMMd, the mean and 95% confidence interval of mean AUC for each of the simulated dataset has been reported in the table 3.1.

Figure 3.6 shows the distribution of AUC for each of the simulation scenarios and for the different algorithms. The raincloud plot in Figure 3.6 suggests that the value of AUC corresponding to Mahalanobis depth reduces when moving from scenario-1 to scenario-4. Note that scenario-1 represents the outliers from a global perspective whereas the other datasets represents outliers arising from local structure in the data 3.5. The distribution of AUC for the proposed kMMD is

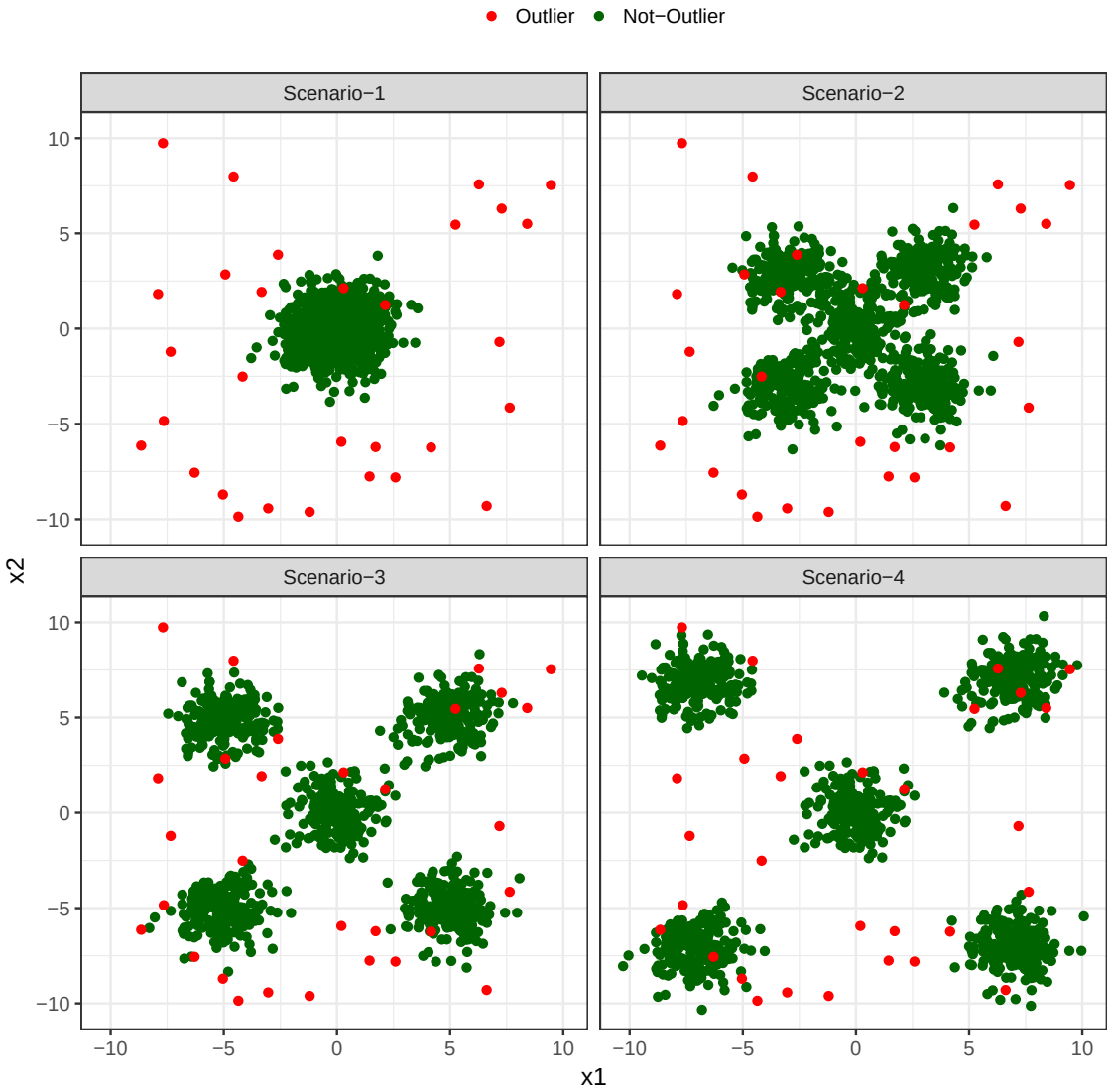


Figure 3.5: Simulated Dataset: Outliers are in Red

Simulated Data	Mean AUC	Standard Deviation of AUC	95% CI
Scenario-1	0.9776	0.01954	(0.976383, 0.978808)
Scenario-2	0.9156	0.03631	(0.91333, 0.91784)
Scenario-3	0.9117	0.03579	(0.90950, 0.91394)
Scenario-4	0.9145	0.03614	(0.91224, 0.91673)

Table 3.1: Area Under the ROC Curve (AUC)

consistent irrespective of the nature of the structure of the simulated data. As the local structure gets stronger, the Mahalanobis depth performs consistently poorer than the proposed kMMD approach. The kNN approach also follows the kMMD for this simulated experiment but might perform poorly when applied to other data as shown by Goldstein [GU16]. The raincloud plot also reveals that the anomaly detection technique that takes the local structure into account could produce more false positives, this is also true for the proposed kMMD approach as indicated in the top left panel of Figure 3.6.

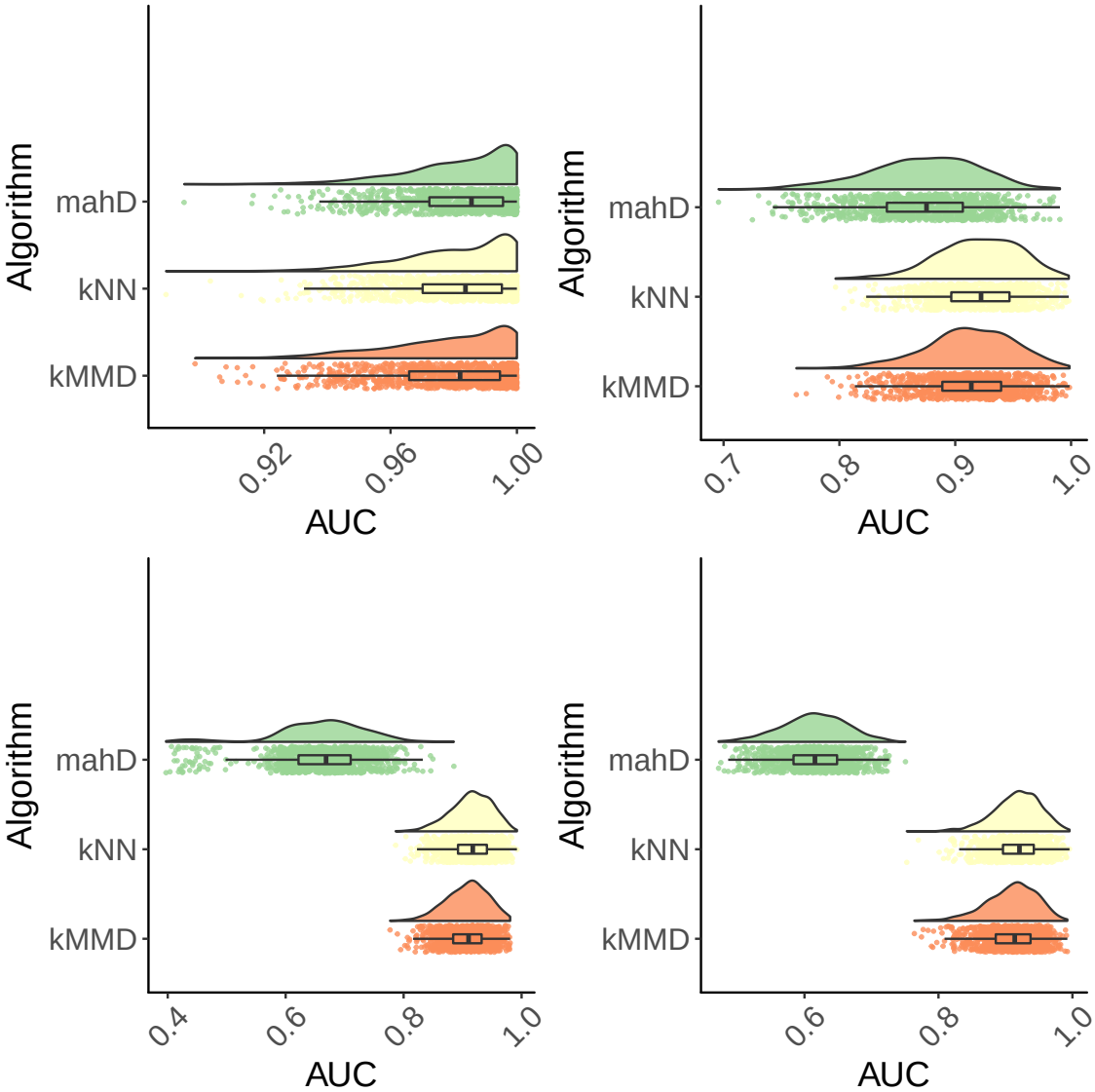


Figure 3.6: Distribution of AUC for simulated dataset. Here mahD is the Mahalanobis depth, kNN is the k-nearest neighbour approach and kMMD is the proposed modified Mahalanobis depth

3.5.2 Artificial Benchmark Data

The second experiment involved data consisting of two continuous variables, generated from four Gaussian clusters where outliers where again generated from a Uniform distribution. This artificial dataset (displayed in (Figure 3.7) has been used by Goldstein [GU16] to evaluate unsu-

pervised anomaly detection algorithms as it contains both global and local anomalies along with micro clusters. The AUC for the proposed kMMD method was 0.9999 whereas the Mahalanobis depth based approach produces an AUC of 0.7397. This result is similar to and aligns with the results presented based on the simulated data. The AUC for kNN was 0.994 which was similar to the proposed kMMD approach. In the next section benchmark data created from real-world data will be considered and then the proposed approach evaluated.

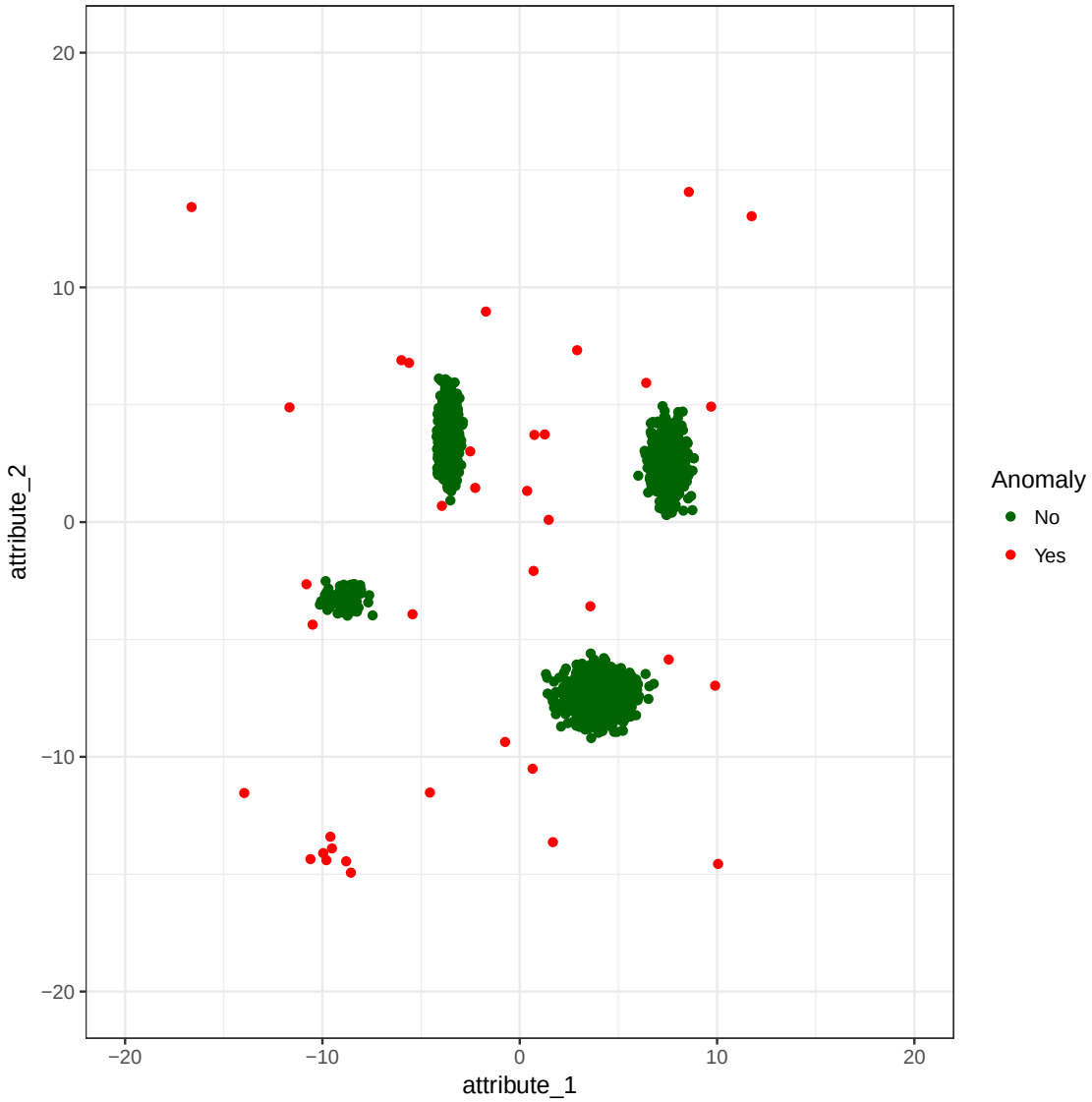


Figure 3.7: Artificial Dataset with global and local outliers (in red) along with micro clusters)

3.5.3 Benchmark Data Derived from Real World Data

The simulated data presented before might not resemble real world scenarios as data generated from an assumed probability distribution may not reflect the correct distribution of a real-world phenomena. For example, outliers may not be just those observations occurring in the extreme tail of a certain distribution, rather it could come from a completely different probability distribution (e.g. a mixture). In 2013, Emmott et.al [EDD+13] pointed out four necessary requirements

for a benchmark dataset that is used in evaluating algorithms for anomaly detection. According to the criteria, an anomaly detection benchmark data should satisfy the following points:

- A1: Non-outliers should be drawn from a real-world data generating process.
- A2: Outliers should also be from a real-world process that is semantically distinct from the process that generates non-outliers.
- A3: Many benchmark datasets are needed.
- A4: Benchmark datasets should be characterised in terms of well defined and meaningful problem dimensions that can be systematically varied.

Goldstein's [GU16] in 2016 proposed benchmark datasets to evaluate anomaly detection algorithms that satisfies A1 and A4.

Based on the criteria A1-A4, Emmott et.al [EDD+13] proposed an algorithmic steps to construct benchmark dataset derived from real-world data. The proposed approach for constructing benchmark data that satisfy the above-mentioned four requirements (A1-A4) can be summarized as:

1. Fit a binary classification model $P(y|x)$ with $y = 1$ represents a non-outlier and $y = 0$ represents an outlier
2. Calculate $P(y = 1|x)$ for the candidate outliers with $y = 0$
3. Take a random sample from the candidate outliers with outlier proportion $0.001, 0.005, 0.01, 0.05$ and 0.1. That is the number of outliers will be $n \times$ "$\mathtt{outlier_proportion}$"
4. Use the pre-defined difficulty score based on $P(y = 1|x)$ for $y = 0$ for each value of the proportion, that is a total of 20 possible combinations of difficulty level and proportion of outliers
   - Easy: $P(y = 1|x) \in (0, 0.16); y = 0$
   - Medium: $P(y = 1|x) \in [0.16, 0.30); y = 0$
   - Hard: $P(y = 1|x) \in [0.3, 0.5); y = 0$
   - Very Hard: $P(y = 1|x) \in [0.5, 1); y = 0$
5. To ensure semantic variations and clusteredness select candidate outliers that are either close to each other based on Euclidean distance

There are few limitation of the proposed approach, the model used to calculate $P(y|x)$ should represent the data well, that the model should be able to estimate the probability as correctly as possible, but there is no such assessment of model quality presented. Moreover, some the data points that originally belong to the non-outlier class could exhibit a very small estimated probability if the model is not a good fit. This lower probability for non-outliers will affect the benchmark data and will compromise the very first criteria.

To overcome the smaller probability of non-outliers in the benchmark data, such observations

were dropped in the process of generating benchmark data keeping only those observations that has $P(y = 1|x) >= 0.5$ with $y = 1$. Also, as the objective is the estimate the probability as correctly as possible, an over-fitted classification model or complex ensemble model could be used.

A series of benchmark dataset has been created from **MAGIC Gamma Telescope** data [DG17] using the method and configuration discussed above. AUC has been calculated for all three algorithm mahD, kNN and kMMD.

The following configuration was used to create a series of benchmark dataset:

- Anomaly proportion = 0.001; proportion of data points out of the total sample size
- Level of difficulty = easy (0.15), medium (0.15-0.30) and hard (0.30-0.50)
- Number of benchmark datasets generated = 1000

Based on the above configurations, three thousands benchmark datasets were constructed, one thousands benchmark dataset for each difficulty level. The proposed kMMD approach applied to each of the datasets and anomaly score was calculated. The AUC was then calculated for each dataset for the Mahalanobis and kNN depth approaches. The following raincloud plot depicts the distribution of the AUC for each configuration.

In Figure 3.8, the top left panel shows that the distribution of AUC for the Mahalanobis depth approach is quite flat with individual AUC values stretched over the x-axis. On the other hand the distribution of AUC for the proposed kMMD and kNN methods are concentrated on the right side of x-axis indicating better performance. As the level of difficulty goes up, the performance of all approaches falls downward. For the medium difficulty level the proposed new approach performs better than Mahalanobis counterpart. In the case of the hard difficulty level the performance of all models is similar. If the difficulty level is hard it indicates that the input variables do not have enough signal to differentiate between outlying points and non-outlying observations.

3.6 CONCLUSION

In this chapter, outlier detection techniques have been presented from a statistical point of view and the notion of a statistical depth function introduced. The use of depth functions in detecting outliers in multivariate data has been presented along with a discussion of potential limitation of existing approaches. A new approach has been introduced by modifying the classical existing Mahalanobis depth function and the (empirical) performance compared using a variety of simulation experiments. The proposed approach shows promising results in the experimental scenarios under investigation.

In the next chapter methods will be presented to incorporate the proposed outylingness score

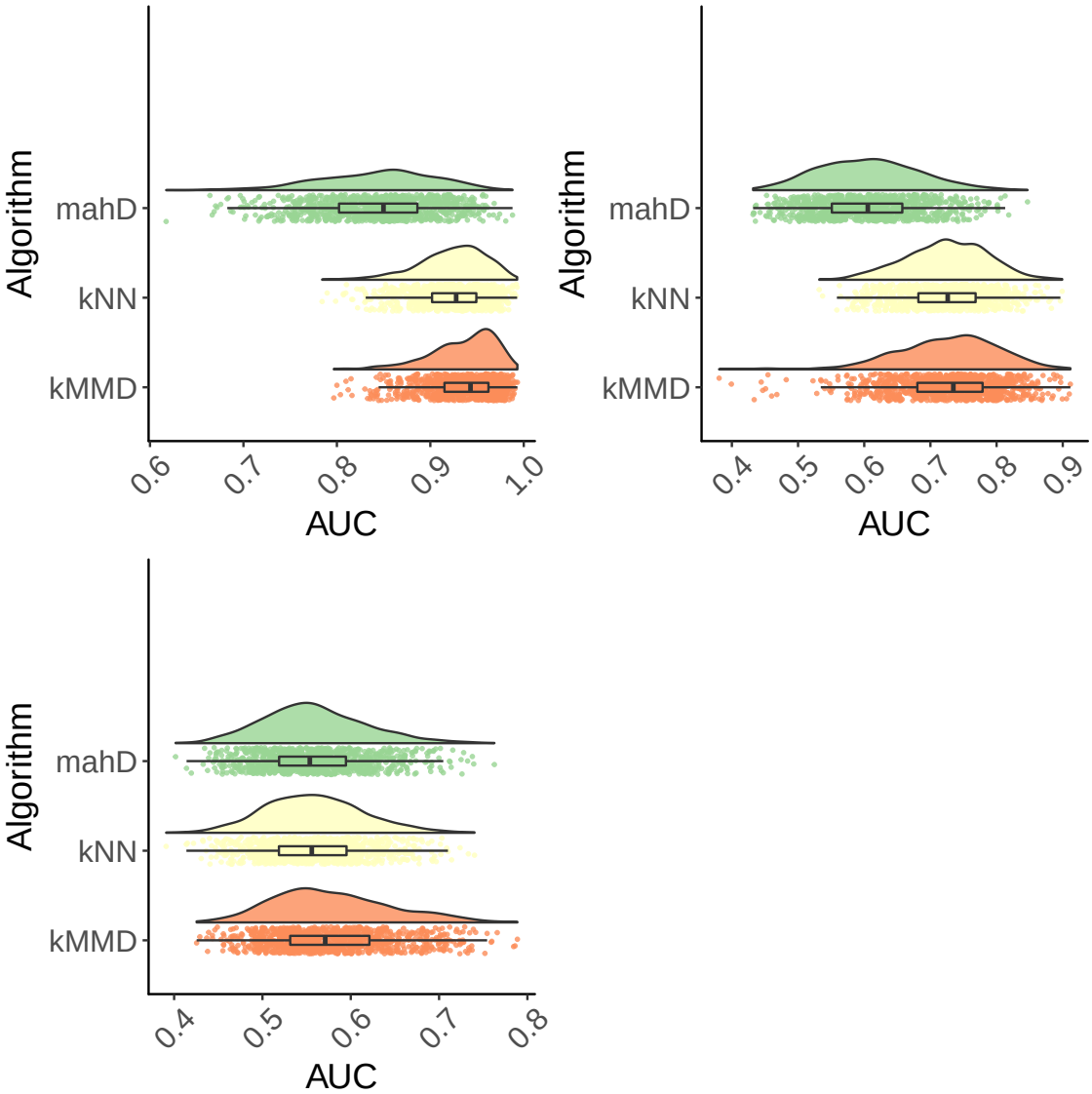


Figure 3.8: Distribution of AUC for Benchmark data derived from Real Data (Top left corner represents the configuration of easy (0.15), top right corner represents the configuration of medium (0.15-0.30), and lower left panel represents hard (0.30-.050) difficulty level)

into visualisations of multivariate data in order to visually identify potential outliers in an efficient manner.

4 VISUALISING MULTIVARIATE DATA

4.1 INTRODUCTION

The primary aim of this chapter is to present graphical approaches to identify potential outliers while incorporating the modified Mahalanobis depth function for multivariate data introduced in this thesis.

To achieve this aim, classical and modern methods to visualise multivariate data are presented and adaptations to these methods for outlier detection are presented that can augment many of the typically used graphical techniques in data analysis.

A case study relating to motion tracking in elite soccer is presented. A random sample of 181 players from a particular season was chosen and data on 22 continuous variables extracted (Table 4.1). The aim was to use data visualisation to identify players that had an atypical pattern (outlier) across the set of motion tracking variables recorded. An outlier in this context could be considered a positive or negative outcome (e.g. impressive high intensity running or fatigue due to excess distance covered) that occurred during a game.

4.2 CASE STUDY: MOTION TRACKING IN ELITE SOCCER

In this chapter a case study using motion tracking data is used as an exemplar. The dataset is comprised of 22 variables (Table 4.1) collected in game from a random sample of 181 soccer players across a season. The players names and positions have been removed from the dataset. The objective is to visualise the dataset in a way to identify players whose movement (e.g. distance covered, top speed, high intensity running) is atypical and to identify which variables are responsible for this outlyingness.

4.3 VISUALISING UNIVARIATE AND MULTIVARIATE DATA

Uncovering meaningful insight from data is one of the primary objectives in all phases of a statistical analysis from exploratory analysis to model checking to model translation. For example, in an exploratory data analysis numerical summaries are often inadequate when conveying valuable information on the underlying structure of the data; visualisation plays a critical role for effective communication of information that is inaccessible through the use of summary statistics alone. The Anscombe Quartet is a celebrated example of this case in point. The quartet is

Variable Name	Description	Variable Name	Description
v1	Distance	v12	High Speed Run Distance Team in Possession (TIP)
v2	Standing	v13	Sprint Distance TIP
v3	Walking	v14	No. of High Intensity Runs TIP
v4	Jogging	v15	Distance (OTIP)
v5	Running	v16	High Speed Run Distance Other Team In Possession (OTIP)
v6	High Speed Running	v17	Sprint Distance OTIP
v7	Sprinting	v17	No. of High Intensity Runs OTIP
v8	No. of High Intensity Runs	v19	Distance Ball Out of Play (BOP)
v9	Top Speed	v20	High Speed Run Distance BOP
v10	Average Speed	v21	Sprint Distance BOP
v11	Distance Team In Possession (TIP)	v22	No. of High Intensity Runs BOP

Table 4.1: List of variables of a multivariate dataset

comprised of four different datasets each containing two numerical variables. The number of data points in each dataset is 11.

Property	Value	Accuracy Level
Mean of x's	9	Exact
Mean of y's	7.50	Two decimal places
Sample Variance of x's	11	Exact
Sample Variance of y's	4.125	Plus/Minus 0.003
Pearson correlation between x & y	0.816	Three decimal places
Linear regression line	$y = 3.00 + 0.500x$	to two and three decimal places
R^2 of linear regression	0.67	Two decimal places

Table 4.2: Summary statistics of Anscombe Quartet

Based on the summary statistics alone, it could be concluded that all four dataset show the same pattern as the mean and variance of x and y are same as is the sample correlation up to two decimal places. Moreover, the simple linear regression equation is the same for each dataset as is the coefficient of variation. The actual relationships in each dataset are only evident once the data are presented graphically.

The message behind the Anscome Quartet was extended further in 2017 by Matejka and Fitzmaurice [MF17] in the Datasauras Project where they created 12 dataset with different structures despite each having identical summary statistics. (Figure 4.2). Once again the within dataset mean, standard deviation and Pearson correlation coefficients are same up to two decimal places but a simple scatter plot highlights the different, and novel, relationships evident in each dataset.

Visualizing multivariate data is a well studied topic in Statistics with excellent references such

as Tufte [Tuf01], Grammar of Graphics [Wil12], Unwin [Unw15; Unw19].

Using visualisation to identify an outlier is not straightforward however and becomes increasingly difficult when the number of variables is vast.

For example a scatter plot using different colours, shapes and sizes for data points could be used to visualise 5 random variables concurrently. In the situation of multivariate data with more than 5 variables alternative visualisations are needed to represent the data in a manner such that outliers can be identified more easily.

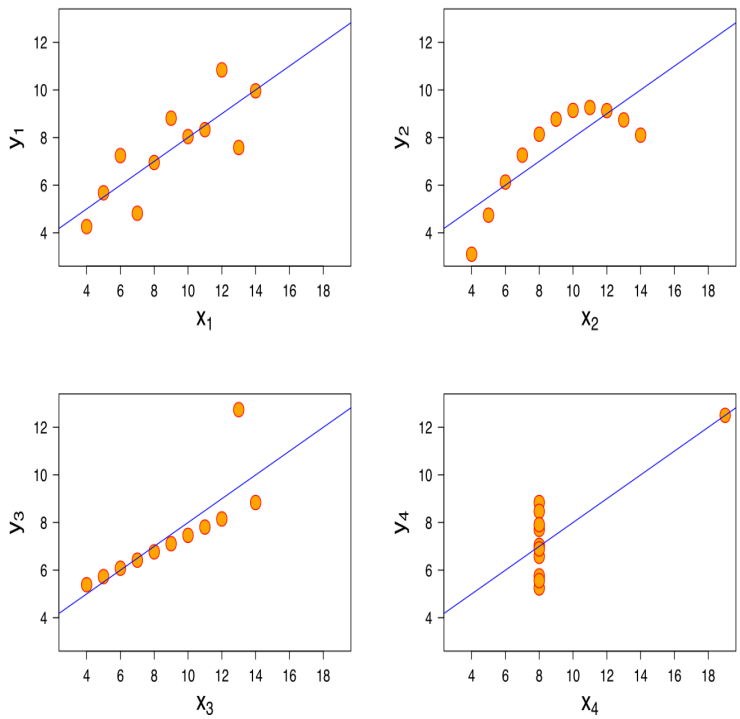


Figure 4.1: Anscombe's Quartet: The dots represents (x,y) pair in the dataset and the line is the best fit linear regression line

One approach is to visualise each variable marginally and look for potential outlying points in each separate plot. This approach is not ideal as an outlying observation in one variable might not be an outlier in another variable. The boxplot is one such example.

4

4.3.1 Boxplot

The *boxplot* [Tuk77] is a well known and a popular method for visualising and exploring a univariate random variable with a mechanism to highlight potential outliers. An observation is deemed to be a potential outlying observation if its distance from the median is greater than $1.5 \times \text{IQR}$; where IQR is the inter quartile range $(Q3 - Q1)$.

For example, boxplots of the 22 variables from the Case Study data are given (Figure 4.3). The

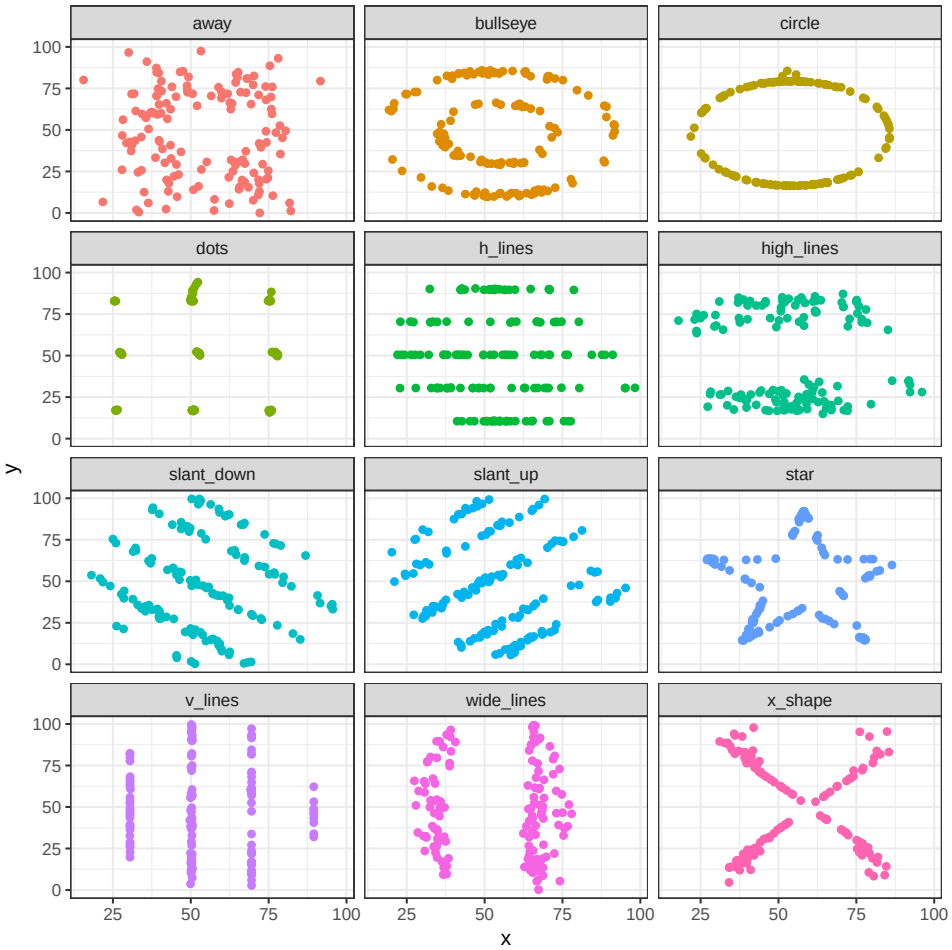


Figure 4.2: Twelve dataset with various shape while their summary statistics similar up to 2 decimal places

red points are potential outliers based on the marginal distribution of each variable. A boxplot is clearly useful when considering the univariate distribution of a random variable, it is less effective in identifying outlying points in a multivariate context nor does it give detail about the shape of the underlying probability distribution for the variable in question or the size of the sample.

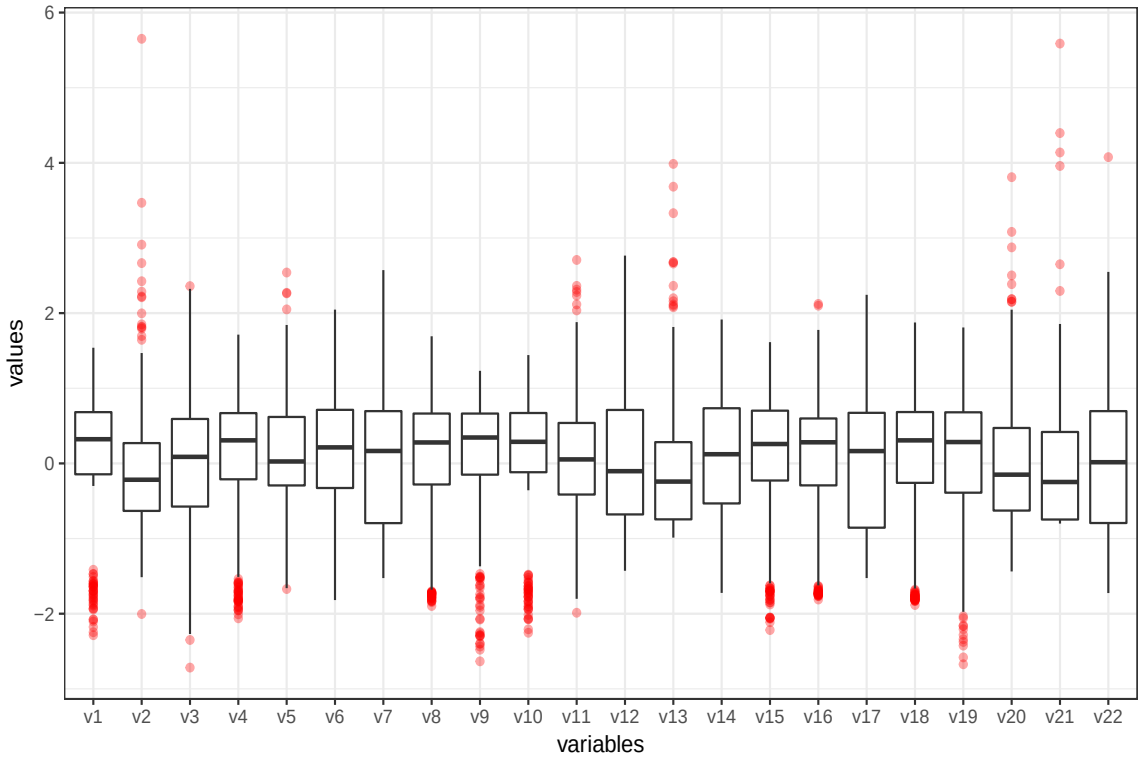


Figure 4.3: A boxplot of each of the 22 variables. Red points represents potential outliers from marginal point of view

4.3.2 Raincloud Plot

A density plot (using a suitable smoother) is a popular method to visualise the distribution of a random variable in order to infer the likely data generating mechanism (i.e. population distribution). The Raincloud [APW+19] plot is a recent extension to a density plot as in addition to providing information about the likely probability distribution an indication is also given as to potential outlying points from a marginal perspective.

A Raincloud plot is presented at Figure 4.4 to visualise the likely probability distribution of each of the 22 variables in the case study. By including a boxplot underneath the density plot information on the classic '5-number summary' is given in addition to the density. The jittered dots overlaid on the boxplot gives an indication of the sample size and shows the position of each individual data point with potential outlying points either far to the left or right of the centre of the distribution. Like a boxplot however a Raincloud plot also only considers a variable marginally.

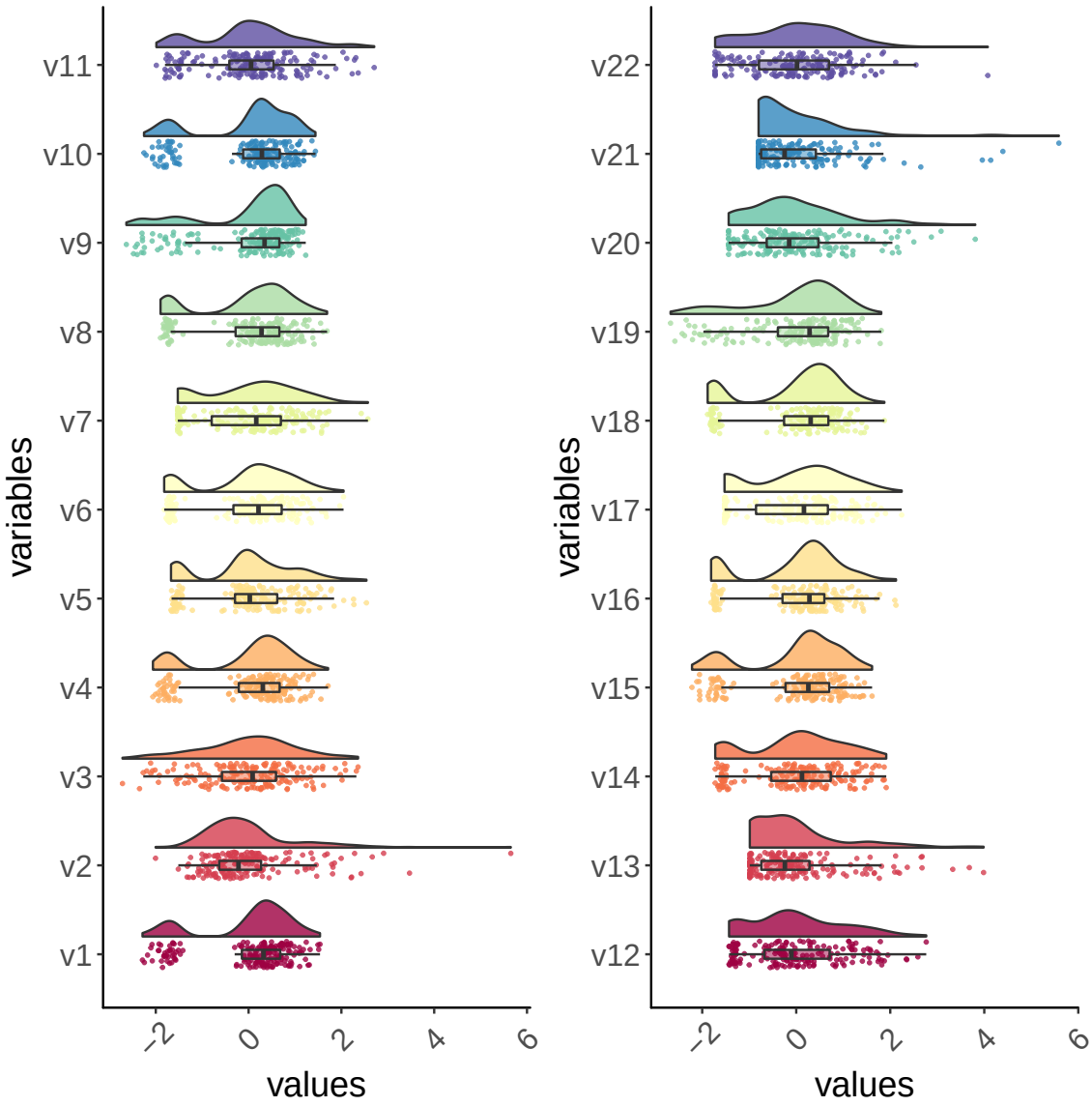


Figure 4.4: A Raincloud plot of motion tracking data with 22 variables, displaying the empirical probability distribution of each variable along with individual data points

4.3.3 Scatter plots

A scatter plot is an easy to understand visualisation of a pair of continuous variables. Using a scatter plot it is easy to see if there are any outlying points, to summarise the likely relationship between the pair of variables and to identify possible clusters of observations. When there are a large number of variables this approach is clearly limited as the number of scatter plots can increase dramatically. For example there will be a total of 231 (i.e. (22*21)/2) scatter plots needed for the Case Study data which is not a feasible exercise. A subset of scatter plots is presented in Figure 4.5, 4.6, and 4.7.

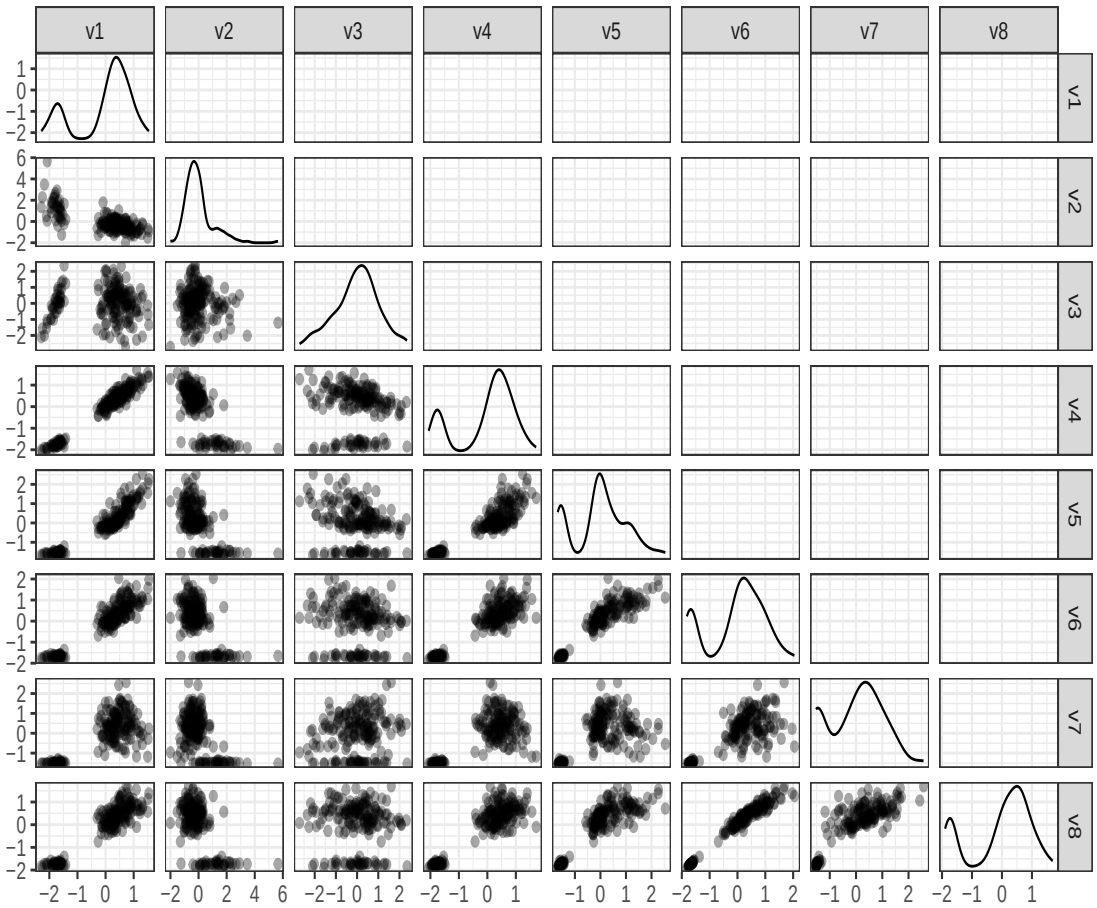


Figure 4.5: Scatter plot of a subset of variables (Part-1)

4.3.4 Bagplot

To overcome the limitation of a univariate boxplot, a bivariate extension of univariate boxplot which is known as **Bagplot** [RRT99] was presented in 1999.

The primary component of a **Bagplot** is the *Bag* which contains at least 50% of the data points, a *fence* which is a boundary between inlier and outlier and a *loop* indicating data points outside the *Bag* but inside the *fence*. The *Bag* is essentially a convex polygon containing at least 50% of

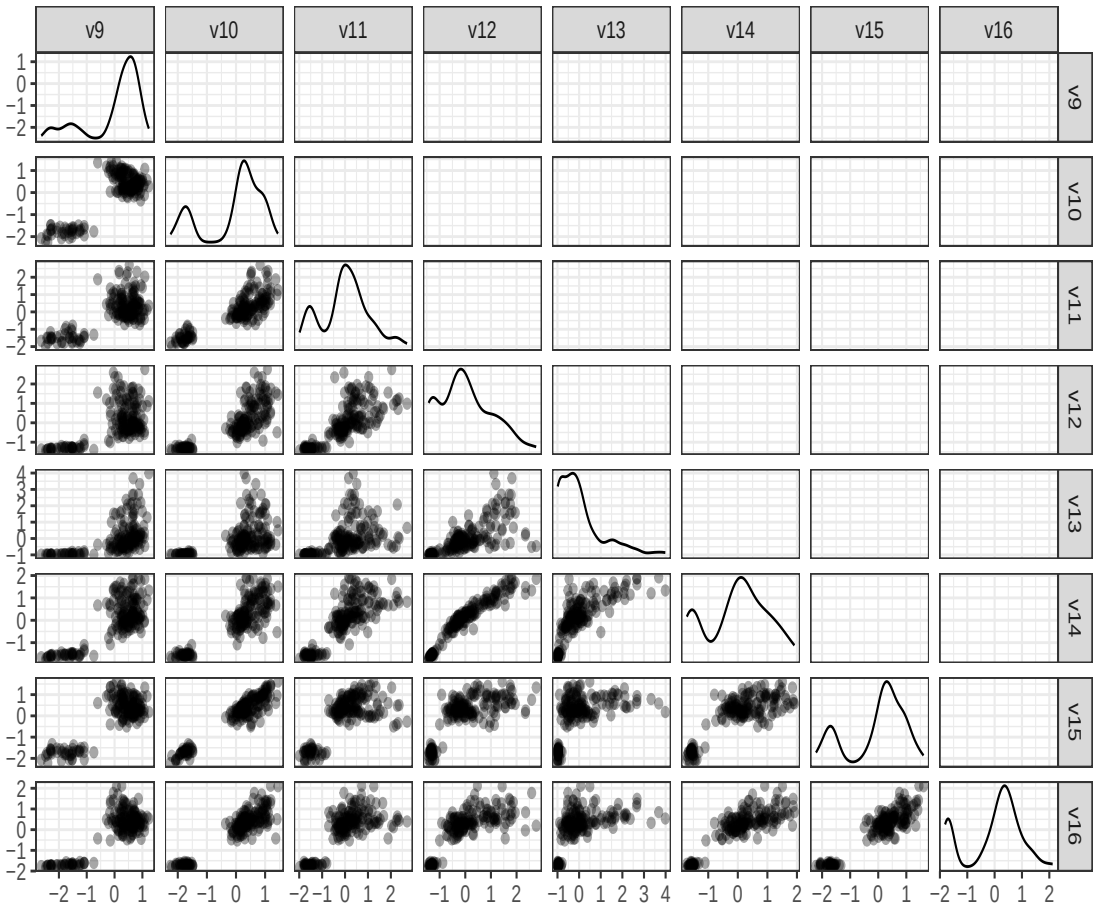


Figure 4.6: Scatter plot of a subset of variables (Part-2)

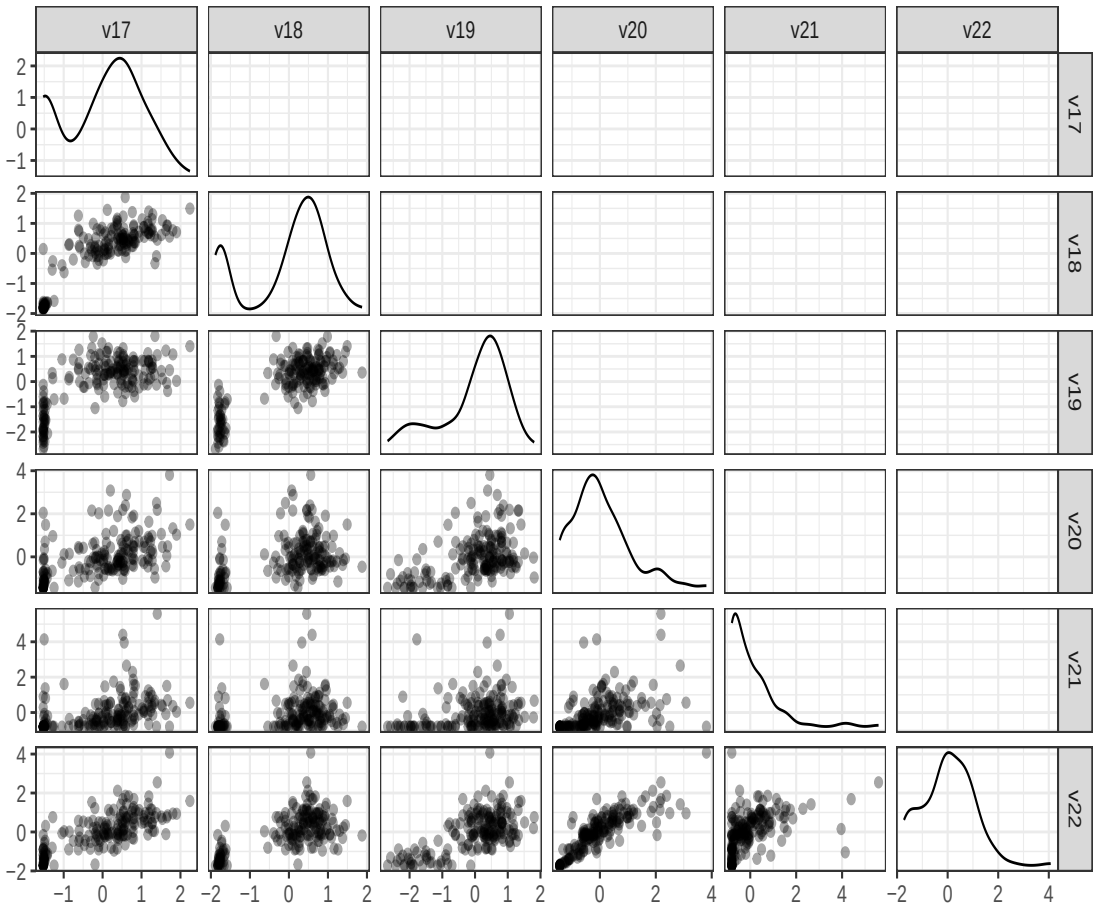


Figure 4.7: Scatter plot of a subset of variables (Part-3)

the data points and by inflating the *Bag* by a factor of 3 a *fence* is constructed. The points that are outside the *fence* are marked as outliers (Figure 3.2).

The generalization of a univariate boxplot to a Bivariate **Bagplot** provides a valuable visual tool to explore the joint distribution of two continuous variables. Such a graph displays several important and interesting characteristics such as depth median (bivariate median), spread (i.e. size of the Bag), correlation between variable represented by the orientation of the *bag*, skewness represented by the shape of the *bag* and *loop* and the identification of outliers with respect to the two variables represented. A subset of **Bagplot** is presented using the variables in the case study (Figure 4.8, 4.9 4.10).

Though a **Bagplot** is useful and intuitive to understand in a bivariate context, extensions are needed for multivariate data with many variables as the construction of a **Bagplot** for each pair faces the same challenges as a scatter plot i.e. 231 **Bagplots** are needed. To produce the **Bagplots** the R package *aplpack* and *ggplot2* has been used.

The set of scatter and Bagplots presented suggest the presence of two clusters in the data distinguised by values relating to distance covered and the presence of potential outlying points in some of the bivariate distributions considered. The next step is to provide graphical summaries that help identify outliers in general across all variables collected.

4.3.5 Andrew's Curve

An Andrew's curve is one of the early attempts to visualise multivariate data that can be easily understood while preserving key properties of the data. The idea is to map a multivariate data point to a curve and then plot each of the data points as a function over negative π to positive π. Mathematically, the multivariate data $x_1, x_2, \cdots, x_p$ is projected on

$$(1/\sqrt(2), \sin(t), \cos(t), \sin(2t), \cos(2t), ...)$$

as:

$$f_x(t) = x_1/\sqrt(2) + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + x_5\cos(2t) + \cdots \tag{4.1}$$

This function then plotted on $-\pi < t < \pi$. The representation of multivariate data points into this function has some useful and important properties.

- The function preserves the mean of multivariate data point. That is the function corresponding to the mean vector of the multivariate data is the point-wise mean of the function, i.e. $f_{\bar{x}}(t) = \frac{1}{n}\Sigma_{i=1}^{n} f_{x_i}(t)$
- The function representation also preserves the distance between data points, specifically, the distance between two functions is proportional to Euclidean distance between two data points.
- This function representation also yields a projection on to a single dimension

The one dimensional representation is useful in identifying patterns in multivariate data. Since the function preserves most of the properties, anything derived from the projected data is a true

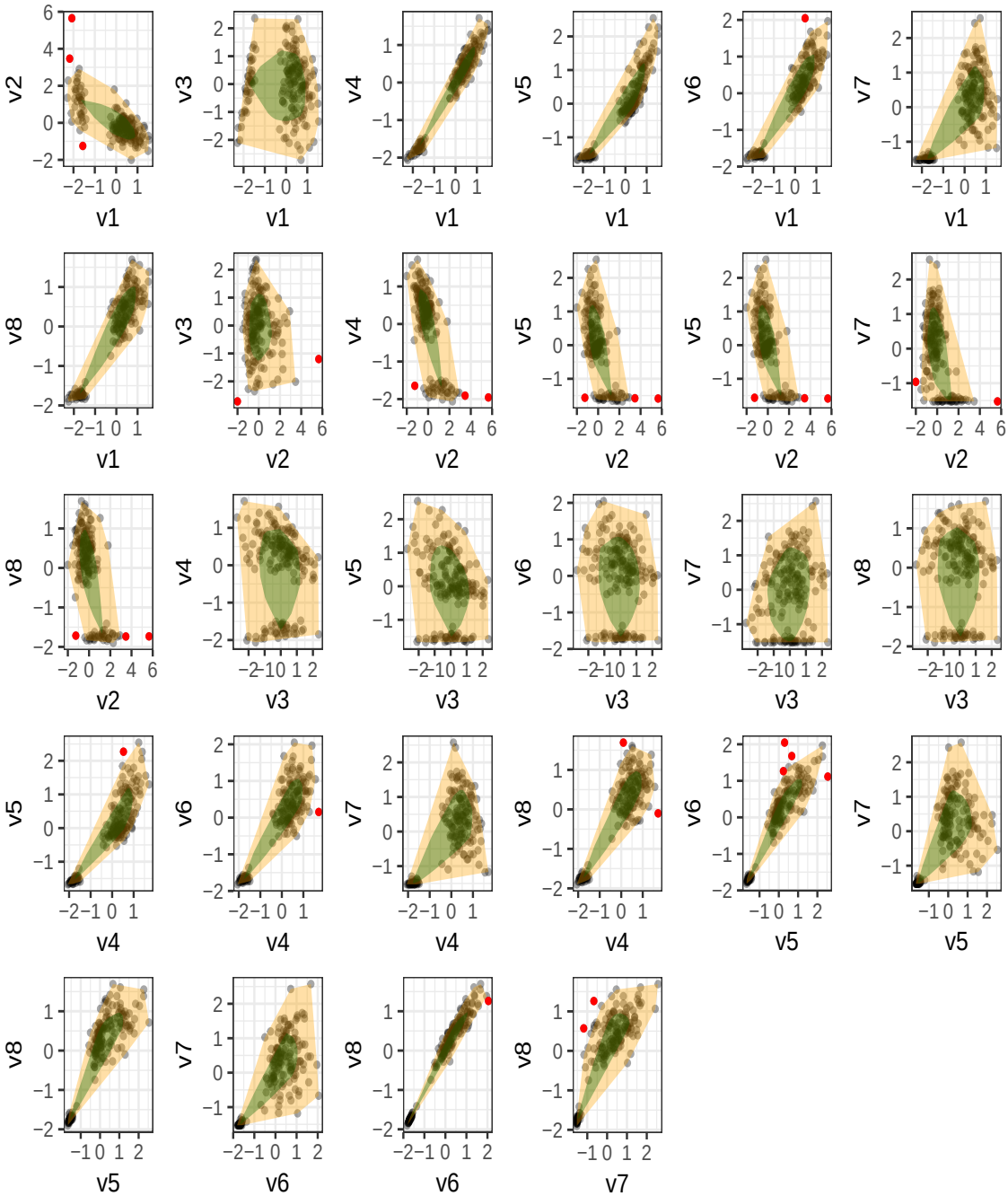


Figure 4.8: Bagplots of a subset of variables. The red points are the potential outliers with respect to the two variables (Part-1)

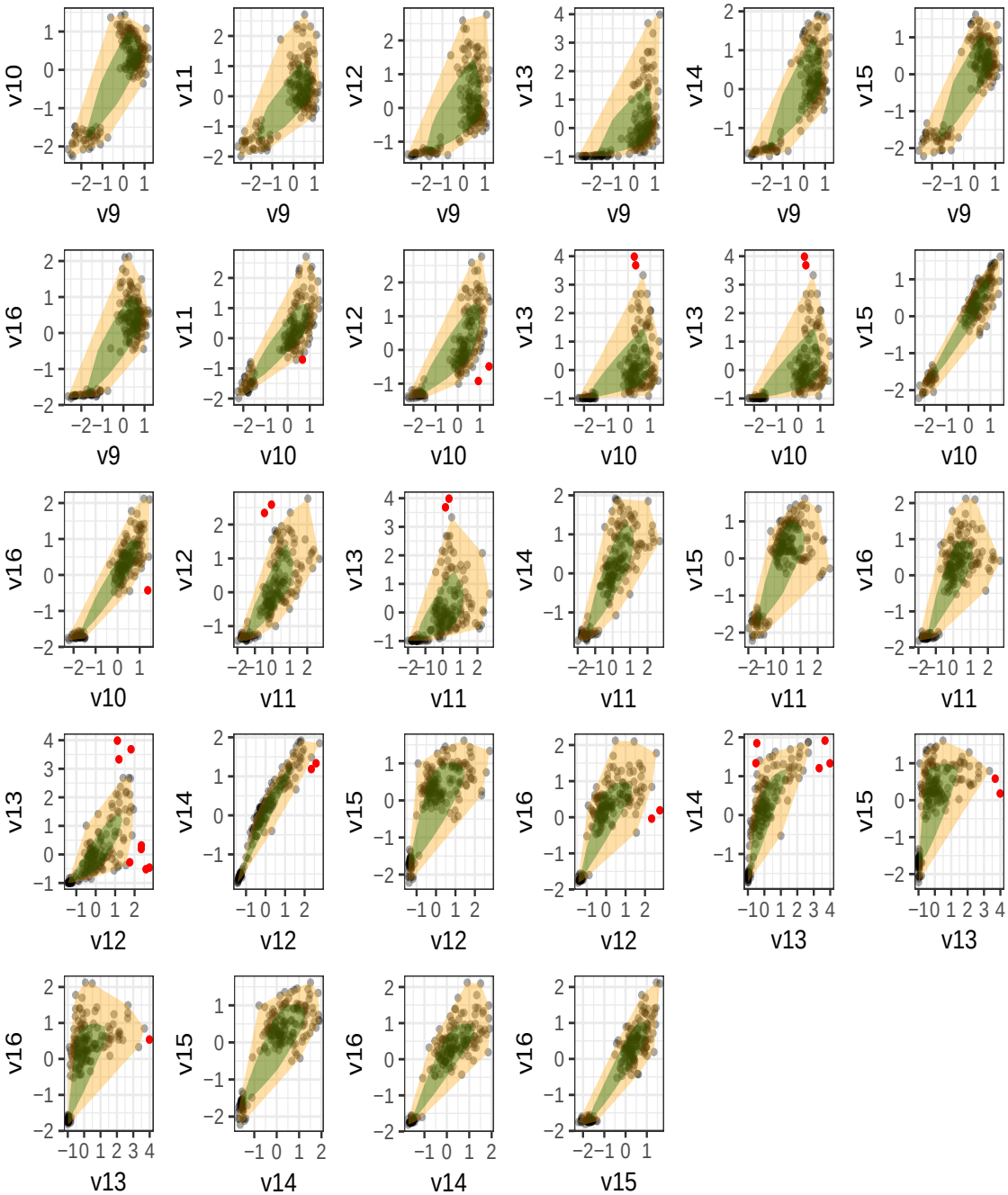


Figure 4.9: Bagplots of a subset of variables. The red points are the potential outliers with respect to the two variables (Part-2)

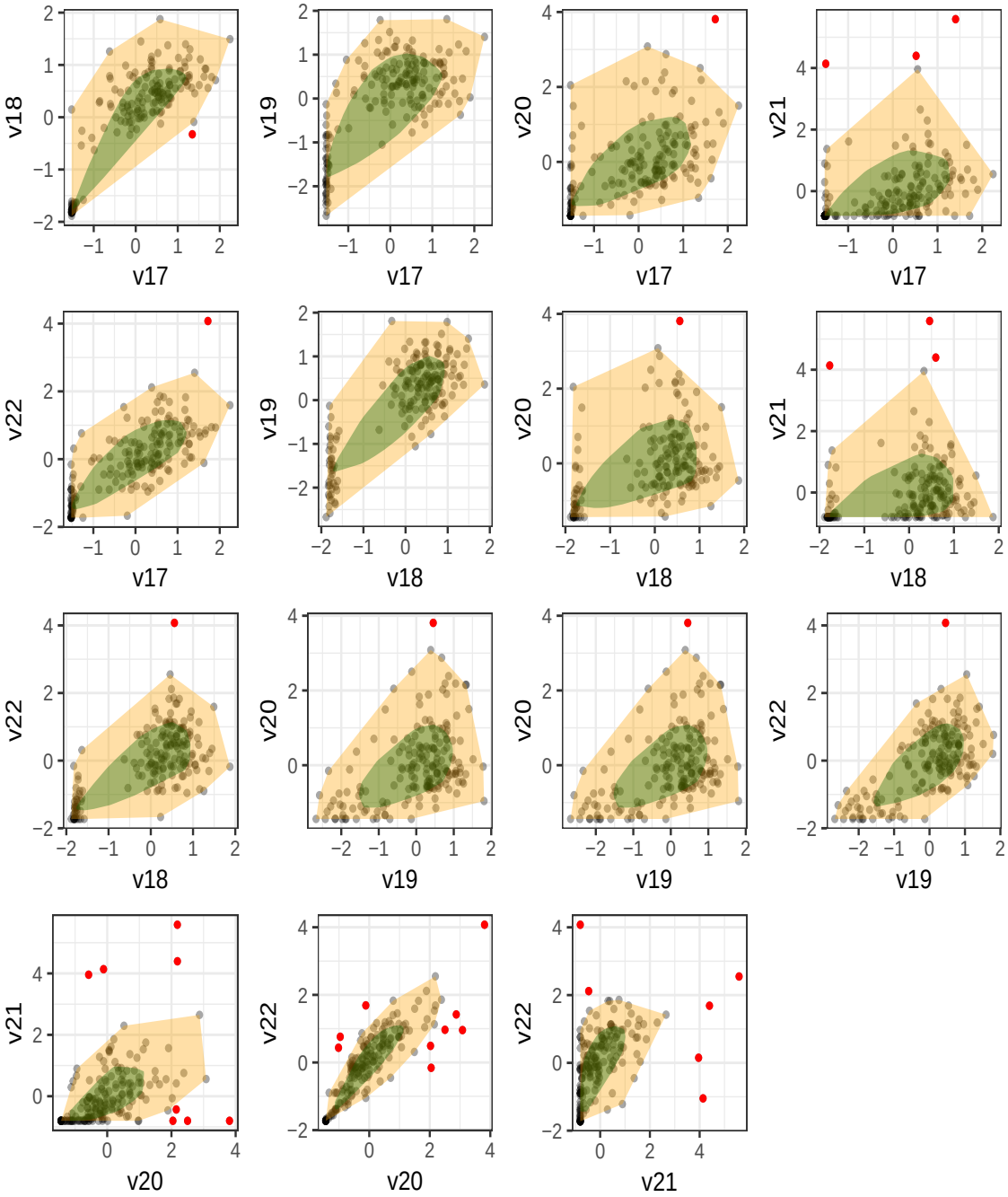


Figure 4.10: Bagplots of a subset of variables. The red points are the potential outliers with respect to the two variables (Part-3)

representation of the original multivariate data.

The Figure 4.11 is the Andrew's curve corresponding to the case study data. Looking at the curve, we may infer that there are two clusters in the data but it is difficult to identify outlying points if there are any. The advantage of this curve is that it takes all the available variables in the dataset and converts it into a function. The groups visible in this curve is a representation of grouping in a multivariate context as well.

The limitation of the Andrew's curve is that if the number of data points is large then it can be difficult to identify structure within the plot.

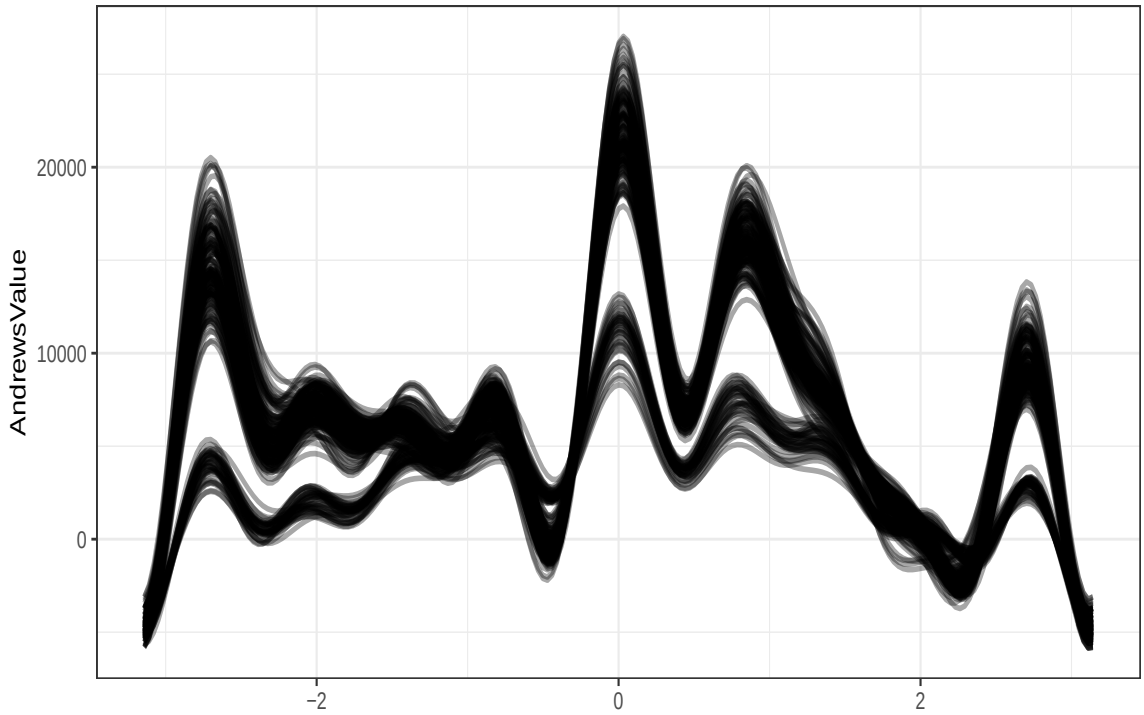


Figure 4.11: Andrew's curve of a dataset containing 22 variables. Each curve corresponds to one row of the data

4.3.6 Parallel Coordinate Plot

In mid 80's Inselberg proposed a new way of visualising multivariate data [Ins85; ID87]. The variables are represented as parallel axes instead of using perpendicular axes. Each data point is then plotted on each of the vertical axis line and then connected the points with a line for each of the individual data points. The plot using parallel axes, is know as parallel coordinate plot.

This plot can reveal patterns that exist in the dataset. For example, the parallel coordinates plot in Figure 4.12 displays the case study data containing 22 continuous variables. There is again clear indication that there are two major groups in the dataset. Moreover, a parallel coordinate plot can reveal which variables are responsible for grouping of the individual data points.

In this case of Figure 4.12 variables number 4 to 19 (except 13) are those that appear to distinguish the two cluster in the dataset.

Though parallel coordinate plots gained popularity in visualising high dimensional data they are of limited use in identifying 'real' patterns in data if the number of variables is too high. One solution is to reduce the dimensionality in the data and then generate parallel coordinate plots on the smaller set of composite variables.

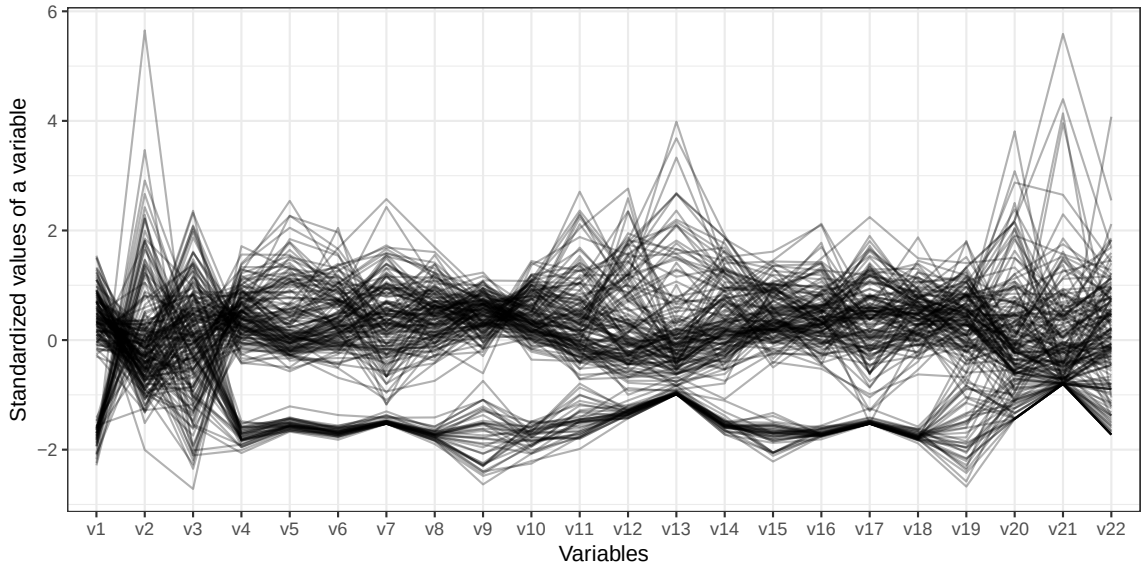


Figure 4.12: Parallel coordinate plot of a dataset containing 22 variables, each of the variables is being standardized before constructing the plot

4.3.7 Principal Component Analysis

Principal Component Analysis (PCA) is one of the most popular techniques to reduce the dimensionality in large data. The primary objective of PCA is to reduce dimensionality of a dataset with correlated variables, while retaining maximum variance with a few linearly uncorrelated variables from the source data by applying orthogonal transformations. PCA was originally proposed by Karl Pearson in the early 20th century [Pea01] and later on it was independently developed and named by Hotelling [Hot33].

Though the primary objective of PCA is to reduce dimensionality often time it has been used as an unsupervised algorithm to identify clusters [YR01] or patterns in data by examining plots of the first few components. In particular, scatter plots and biplots of PC1 and PC2 has been used to visualise the data to uncover patterns such as clusters and outliers.

Apart from visualisation, a first principal component score has been used as a composite score representing the rank of multivariate data; in social science research the PC score has been used to represent asset index [FP01].

Though PCA is one of the popular techniques for dimensionality reduction and index construc-

tion, the first PC score does not always preserve the order of an individual data point in a multivariate context [SNA+14]. The figure below explains how PCA works in reducing dimensionality and visualizing multivariate data to uncover patterns.

The Figure 4.13 indicates that two variables are linearly correlated and the variance is spread in both dimensions. After applying PCA on this data, the axes are rotated in such a way that the correlation between the rotated axes is zero. The direction of the first Principal component is in the direction of maximum variance. The Figure 4.14 shows that the rotated axes are uncorrelated and the maximum variance is in the direction of first PC. It can be clearly seen from Figure 4.14 that the larger variation is in the direction of first principal component and there is smaller variation in the second principal component. When the number of input variables is larger than 2, the first few principal components account for most of the variation in the data whereas the last few principal components account for the least amount of the variance in the original data.

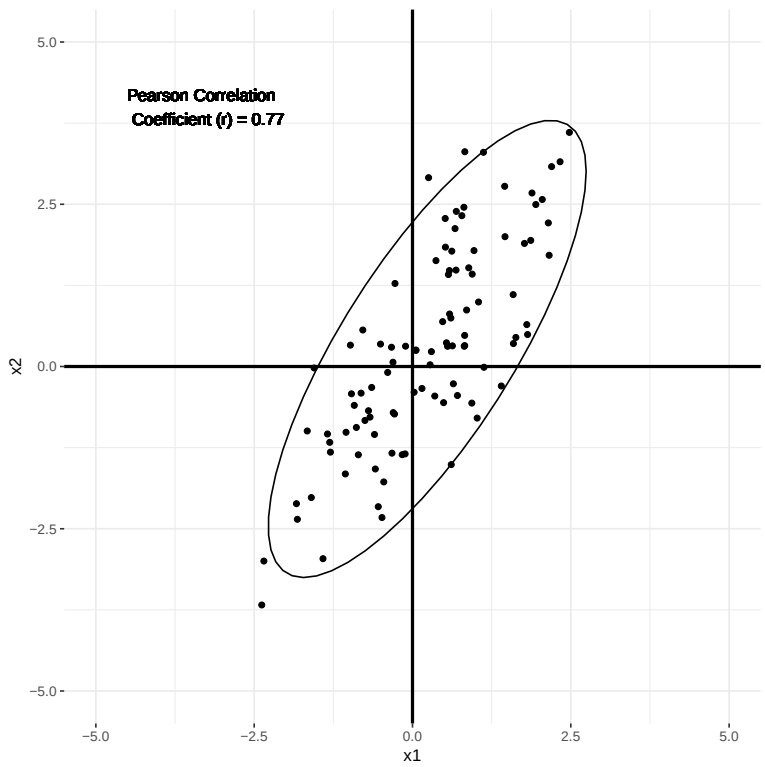


Figure 4.13: Scatter plot of two linearly correlated variables

The Figure 4.15 is a scatter plot of first two principal components derived from the case study dataset of 22 variables. We can see there are two obvious clusters in the dataset and some of the individual points are far away from majority of points which is an indication of outlying points.

The principal component analysis is optimised to find the linear combination of original variables that maximises the variance in the data, that is finding a few linear combinations that contain most of the variation in the original variables. PC is not optimised to find clusters but if there are inherent groups present in the data then plotting the first few PC could reveal this structure. That said, a PCA cannot be used to identify exceptions or anomalies in data, as it is not designed to do so. Moreover, though the first few PC could reveal inherit hidden grouping

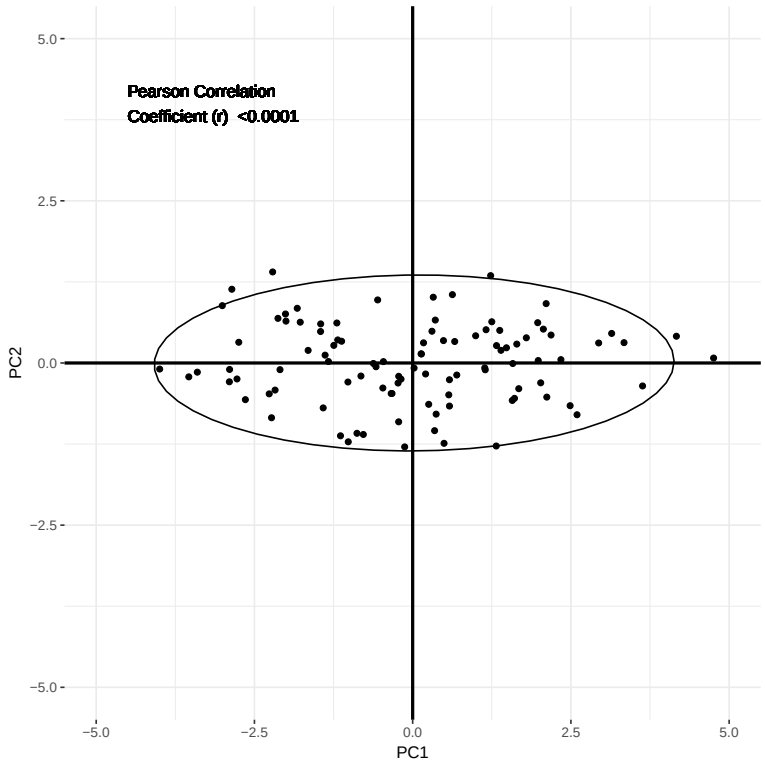


Figure 4.14: Scatter plot of two PC scores

but it cannot provide a single score that could determine whether the point should be consider as an anomalous observation.

On the other hand, PCA is designed to handle linear relationships between input variables. If the input variables are not linearly correlated then the first few PC are not enough to capture the major variation in the data. Also, if the input variables are not correlated then PCA is not useful in visualising the dataset. The classical PCA is designed to handle numeric variables only, to accommodate mixed type of variables polychoric PCA [KÁ04] could be used but again this is only maximise the variance along the first few principal components.

4.3.8 Generalized Low Rank Models

In real world applications a dataset could contain various types of variables ranging from numeric, binary, nominal and ordinal. Uncovering hidden patterns in such complex data is usually challenging and the visualisation of such data is not straight forward. While a PCA could be used to reduce the dimensionality of a dataset, while retaining maximum variance, PCA however is not suitable for data with mixed types of variables. Despite this PCA is one of the most popular techniques in dimensionality reduction and often used as an unsupervised learning algorithm to find groups in data. Moreover, if there are any missing values in any one of the variables then the corresponding data points are dropped from the analysis. A more general framework for low rank approximation has been developed as an extension to the 'classical' PCA known as Generalized Low Rank Models (GLRM) [UHZ+16]. In GLRM arbitrary types of variables can be handled along with techniques to handle missing data. The original PCA could

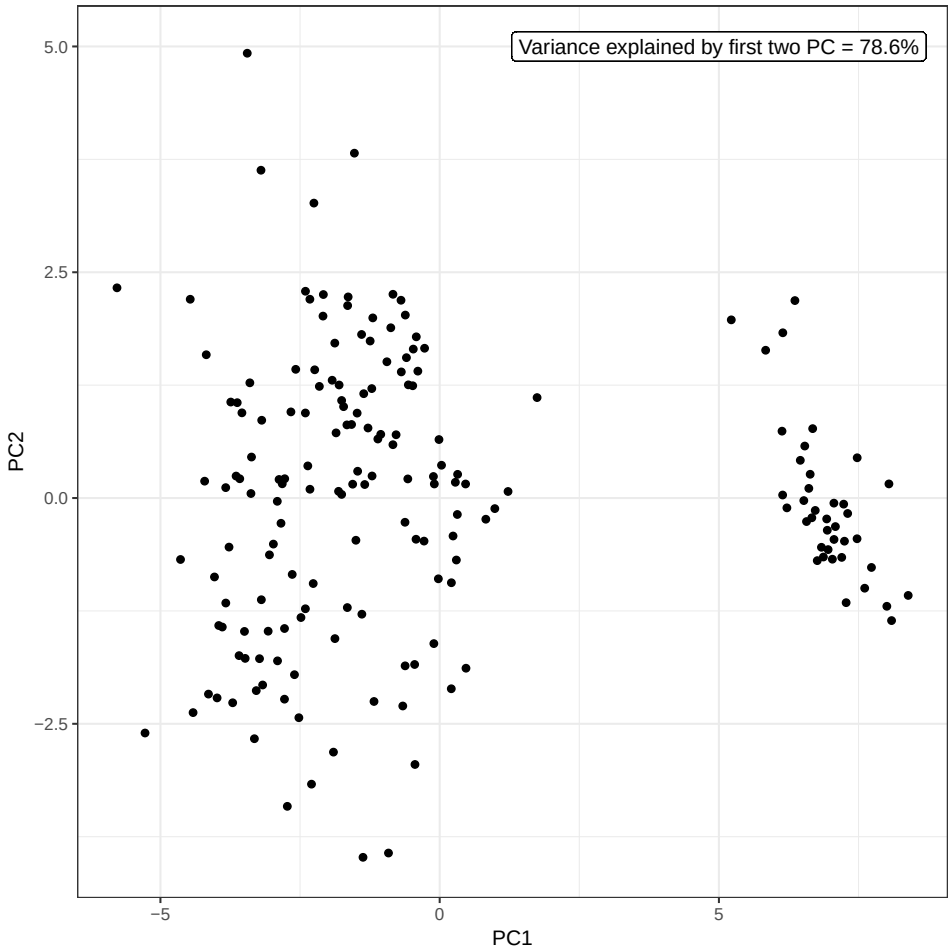


Figure 4.15: Scatter plot of first two PC derived from a dataset of 22 variables

be defined as a special case of GLRM.

In GLRM a new loss function has been used that is appropriate for data with mixed variables, by replacing the least-squares error loss function. Apart from the new loss function a regularisation of the low dimensional factors has been introduced to uncover structure such as sparsity in low dimensional factors. The new optimisation problem consists of an approximation error along with regularisation of low dimensional factors, it produces the similar low dimensional representation as that of PCA and it can be used to visualise high dimensional data in lower dimensional spaces. The proposed optimisation is non-convex and cannot be solved globally and efficiently. In GLRM, the original input data matrix is factored into two matrices so that their product can approximate the original data matrix as close as possible. If X is a data matrix with p input variables and n data points, then the data matrix X can be written as a product of two other matrices of $n \times k$ and $k \times p$, as:

$$X_{n \times p} \approx Y_{n \times k} \times Z_{k \times p} \tag{4.2}$$

Here k is user defined number of dimensions on the projected space and is usually small in number compared to p. The rows of Z are the new variables, called archetypal features, derived from the columns of original input matrix X. The rows of Y corresponds to the rows of input matrix X but projected onto a smaller dimensional space $k << p$. The matrices Y and Z reconstruct the original input matrix X approximately with a rank k lower than p. The projection is similar to the projection in an original PCA but the linear combination in this case is not orthogonal. If the linear combination used here is orthogonal then a GLRM is the same as a PCA and this explains why a GLRM is an extension of PCA and alternatively PCA is a special case of GLRM.

The lower dimensional projection of the original data matrix is used to visualise high dimensional data into lower dimension. Since this is an approximation of the original data matrix, the proximity between data points on the projected space is proportional to the proximity of the data points in the original high dimensional space.

A scatter plot of a lower dimensional projection of the case study data is presented in Figure 4.16. In Figure 4.16 it is evident that there are two broad clusters present along with a few potential outliers with respect to the clusters. As there is no outlyingness score associated with the points in a GLRM it it not possible to distinguish outliers from non-outliers. An outlyingness score would be useful here to visually identify outlying points.

4.3.9 t-SNE

The primary aim of any dimensionality reduction technique is to preserve the structure of data as much as possible on the lower dimensional projected space. As discussed earlier, the biggest limitation of PCA is that it while it maps data into lower dimensional spaces using the (global) covariance matrix based on a linear projection, it cannot capture non-linear structure of data. An alternative approach could be to map a high dimensional point into lower dimensional space in such a way that that the distance between points in the high dimensional space is as similar as possible as in the lower dimensional projection. In 2008 Maaten and Hinton [MH08] proposed

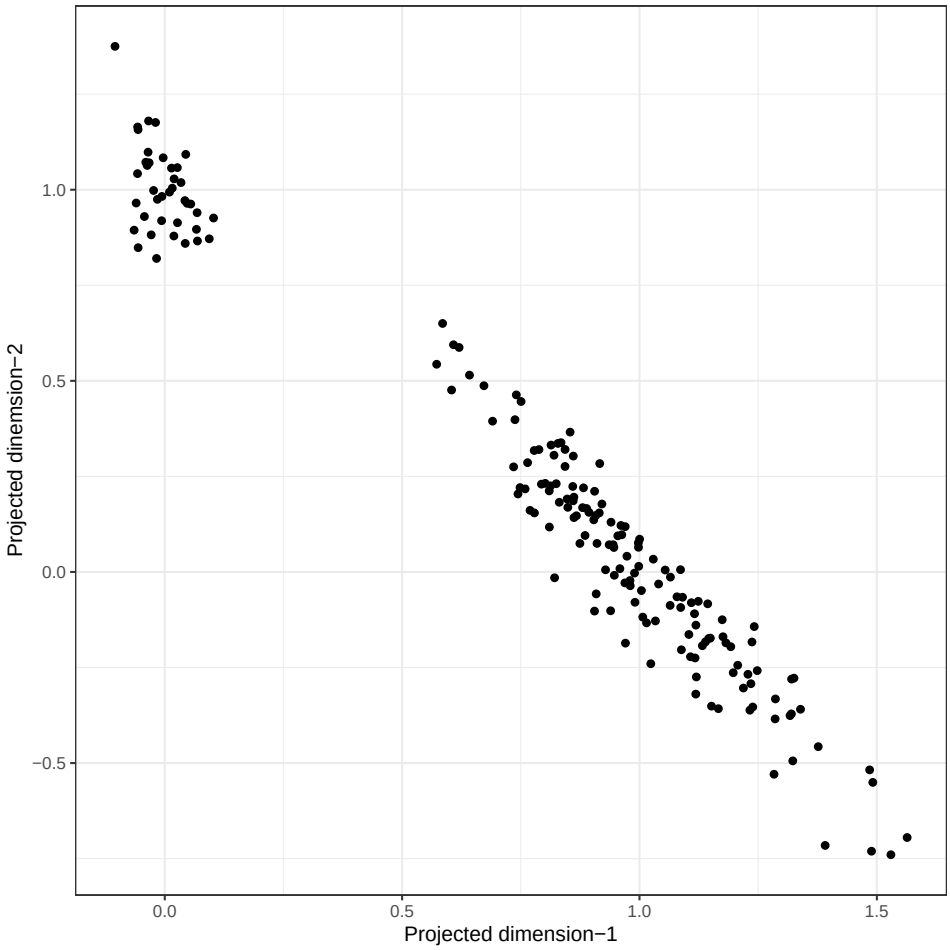


Figure 4.16: Low rank projection of case study data of 22 variables

a new technique to visualise high dimensional data into lower dimensional spaces namely a t-Distributed Stochastic Neighbour Embedding (t-SNE) based on non-linear projection of the input data into lower dimensional space while keeping local structure intact.

Their approach can be considered an extension of the original Stochastic Neighbour Embedding (SNE). A SNE suffers from a problem termed as the *crowding problem* and the associated cost function is difficult to optimise. In high dimensional data projecting points into lower dimensional space tend to squeeze points next to each other due to curse of dimensionality. In the original SNE technique a Gussian distribution has been used to recreate the lower dimensional space. Since the Gaussian distribution is not a long tail distribution the projected points squashed and deteriorated the structure of the data which ultimately caused such crowding. In a t-SNE the crowding problem is easily handled by incorporating a long tail distribution, such as a t-distribution briefly explained as follows:

In the first stage of t-SNE a probability distribution is assumed and the parameters estimated based on the neighbouring points. In the second stage the distribution is re-created using a long tail Student's t-distribution. For example, if we pick a single data point x_i, the probability of picking a point from the same data which will be a neighbouring point is written as:

$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\Sigma_{k\neq l} exp(-||x_k - x_l||^2/2\sigma_i^2)} \tag{4.3}$$

This is a Gaussian distribution with mean at x_i and variance σ_i^2. The value of σ_i^2 is chosen in such a way that the number of neighbours for each data points is roughly the same. Alternatively, we can say that the value of σ_i^2 chosen as a smaller value for highly dense areas and larger for sparse regions. This flexibility of choosing σ_i^2 gave the ability to keep a balance between the number of neighbouring points corresponding to each point in the dataset. Once the probability of picking neighbouring points in higher dimensional space is determined, the probability distribution in a lower dimensional projected space can be written as:

$$q_{ij} = \frac{exp(-||y_i - y_j||^2/2\sigma_i^2)}{\Sigma_{k\neq l} exp(-||y_k - y_l||^2/2\sigma_i^2)} \tag{4.4}$$

Here y_i are the points in lower dimensional space corresponding to the point x_i in the original higher dimensional space. This lower dimensional representation of the probability density still uses a Gaussian distribution and it will suffer from the crowding problem in lower dimensional space. To overcome this, a Student's t-distribution has been used with one degree of freedom which is essentially a Cauchy distribution. The new representation is:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\Sigma_{k\neq l}(1 + ||y_k - y_l||^2)^{-1}} \tag{4.5}$$

Once the lower dimensional distribution is defined then the optimisation is done using gradient descent on the KL-divergence between two distributions (lower dimension and high dimension), q and p respectively. The gradient descent is of the form:

$$\frac{\delta C}{\delta y_i} = 4\Sigma_j(p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{4.6}$$

The updated gradient function can represents the strength and direction of attraction/repulsion between two points. A positive gradient represents an attraction, whereas a negative gradient represents a repulsion between the points. This "push-and-pull" eventually makes the points settle down in the low-dimensional space. One important thing to note in the t-SNE technique is that it does not have any parameter like classical PCA rather it directly optimises the embedding on to lower dimensional space. In effect a t-SNE gives better lower dimensional embedding and may produce better visualisation it cannot be used to train a model for future data points.

The cost function of a t-SNE is a non-convex function with multiple local optimum and therefore there is a risk of getting stuck in a local optimum. Moreover, a t-SNE is non-deterministic in nature, so every time the technique is applied on the same data different results will result. Even though a t-SNE is a non linear projection of high dimensional data it still poses implicit assumptions of linearity in the local manifold and the distance between points defined by Euclidean distance. A t-SNE could be used to describe a dataset as a whole but it cannot be used as a model for future data points and it cannot be used as a way to find out 'exceptional' data points. A lower dimensional embedding could indicate that some of the points are atypical compared to the majority of the data but still it cannot provide a single score of outlyingness multi-dimensional data.

The t-SNE algorithm was applied to the case study data to extract and visualise a 2-dimensional projection (Figure 4.17. The Figure 4.17 displays the original case study data on a lower dimensional projection. It is again clear that there are at least two/three major groups of individuals in the dataset. There are a few points that seems isolated from the majority of the data and these isolated points could be outlying data point in a multivariate context.

As the primary objective is to create a visualisation to identify outlying data points from a multivariate point of view, a t-SNE although useful for identifying potential clusteres in a dataset it is not very useful in identifying outlying data points. Moreover, the non-deterministic nature of a t-SNE creates another level of difficulty in maintaining the same clustering over multiple run with the same parameter settings.

4.3.10 Multivariate Outliers and O3 Plot

An outlying data point in one variable does not necessarily indicate the same data point will be outlying with respect to another variable. Various method of outlier detection may produce different results but why a certain data point is outlying and in which variable is contributing to the outlyingness is crucial information.

In 2019, Unwin [Unw19] introduced a new approach to visualise multivariate outliers called an Outlier O3 plot. Such a plot is used to identify outliers for every possible combination of variables in a dataset by applying the HDoutliers [Wil16] technique and then displaying only those combinations of variables that has at least one outlier along with the case number of the outlying data point. An O3 plot is useful to explain why a certain data point is outlying and what are the variables responsible for the outlyingness. The major limitation of this plot is that for a large dataset it is necessary to investigate every possible combination of variables which

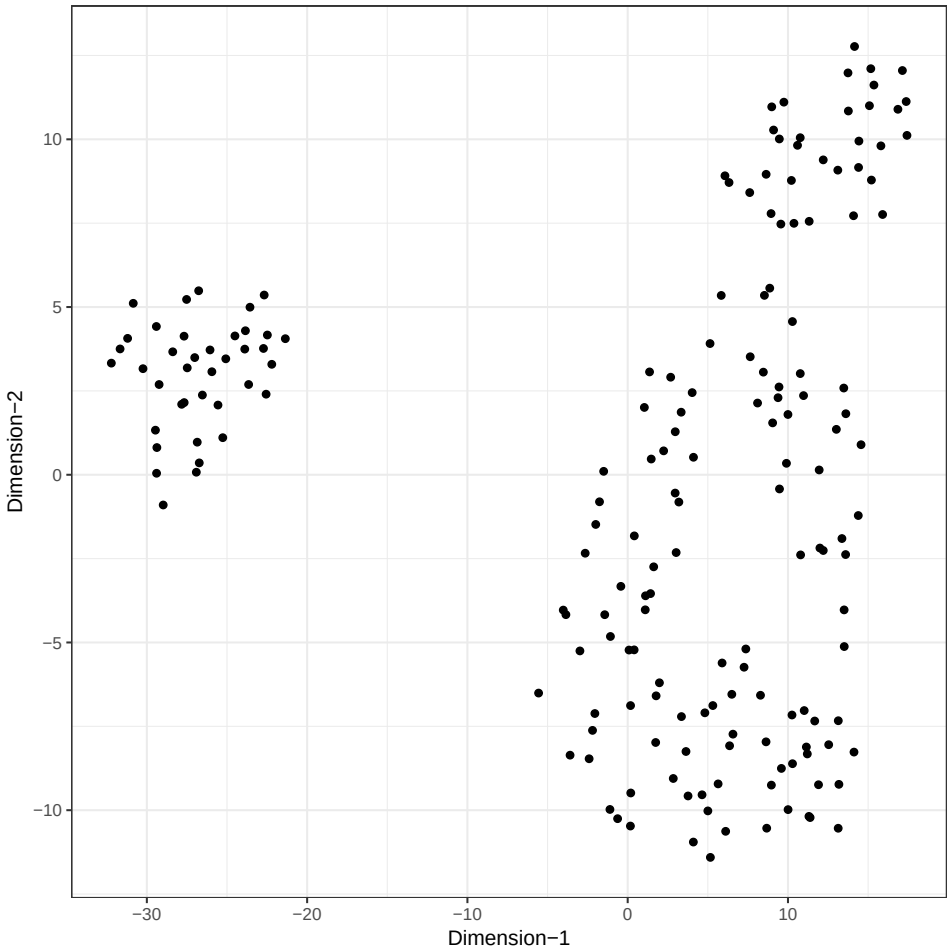


Figure 4.17: A 2-dimensional embedding using t-SNE of the *sports data* with 22 variables

is often not feasible. For example, in the case study with 22 variables an evaluation of several hundred thousand combinations of variables is required to identify outlying points. One option proposed in this thesis is to project the multivariate data onto a lower dimensional space using a PCA [Hot33], GLRM [UHZ+16] or t-SNE [MH08], and then generate an O3 plot to identify which projected component scores are responsible for a point being considered an outlier.

An O3 plot using the first three principal component scores of the case study data is presented in Figure 4.18 where the individual players who are identified as outliers can be identified quite easily for the first three PC scores. To produce the Figure 4.18 *OutliersO3* R package has been used.

Figure 4.18: O3 plot of case study data. Each row represents the combination of variables for which at least one outlier was found; dark gray colour represents the presence of the variable in the combination. Red cells are the outlying data point with corresponding individual ID represented by X*

4.4 APPLICATION OF DEPTH-FUNCTION IN VISUALISING OUTLIERS

As described in the previous chapters, a statistical depth function is a mapping from $\mathbb{R}^d$ to $\mathbb{R}$ (where d is the number of variables), is a method to generate a univariate score for multivariate data. Such a score can is used to study properties of the multivariate distribution.

In Chapter 3, a novel *modified Mahalanobis depth (kMMD)* was presented and its properties investigated. In this chapter, the proposed kMMD is applied to the case study data in order to

create an 'outlyingness' score which is then used to augment classical visualisation to identify potential outlying observations.

By definition of a depth function, a higher value indicates a point that is surrounded by the majority of other points within close proximity with a gradual reduction in the value of the depth function as a point moves away. For ease of visualisation, the reciprocal of the depth function is calculated and its probability distribution inspected (Figure 4.19) such that the higher the value of the reciprocated depth function the higher the outlyingness.

From the distribution of the outlyingness scores, a right tail distribution, values at the tails are values corresponding to potential outlying data points. Now if this outlying score is incorporated into any of the 'classical' visualisations used the score can be used to indicate outlyingness that is not necessarily evident in the plot at hand.

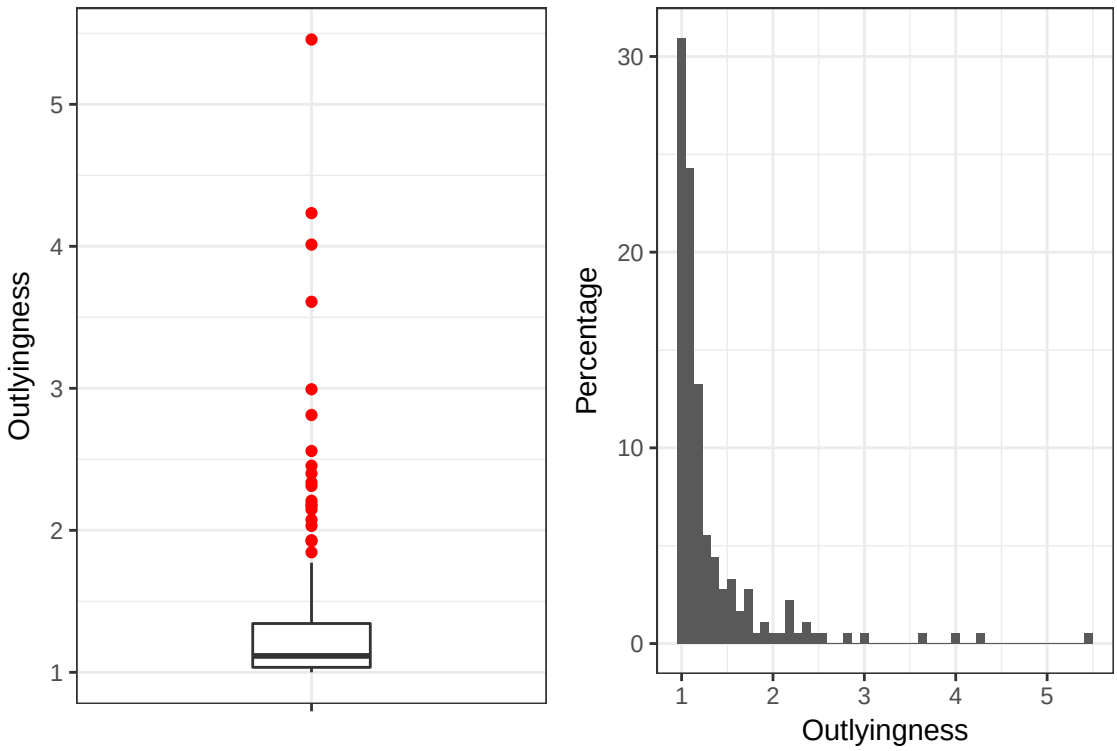


Figure 4.19: Probability distribution of outlyingness score (kMMD) calculated from the case study data with 22 variables. The red points in the boxplot indicate potential outliers based on the boxplot 'rules', these points can be investigated further to see why they are showing high outlyingness scores

For example, Figure 4.20 is a bagplot of two of the original variables from the case study data where this time each point is sized by its corresponding outlyingness score while red points are potential outliers based on the Bagplot algorithm. It is clear from the Figure 4.20 that there are potential outliers, based on the outlyingness score, that the Bagplot fails to identify, in particular points at the boundary of the *fence*. Moreover, some of the points inside the loop have a higher outlyingness score as those points are away from their closed neighbours in a multivariate context.

Using the outlying score in this way, potential outlying points can be easily identified that might not be visible in a two-variable plot (Figure 4.21) but when all variables are taken into consideration using the outlying score, outliers in general are easily identified. Using the outlying score in this way makes it easy to isolate observations with relatively high outlyingness scores, along with their k-nearest neighbours, and do further investigation a either numerically or visually. One possible way could be to highlight these observations in the corresponding parallel coordinate plot and to calculate summary statistics and compare them with the summary statistics to observations small outlyingness scores.

The observation with the highest outlyingness score was extracted along with its k-nearest neighbours and parallel coordinate plot was created (Figure 4.22). The red line is the individual with the highest outlyingness score and the dark-green lines are its 15-nearest neighbours. It can be seen that for variables "v2 and "v21" the values are extreme compared to the neighbours and these two variables are 'driving' the higher outlyingness score of the individual. In Figure 4.22, we can see what variables are responsible for an individual observation to be considered an outlying point in a multivariate context.

In the Bagplot Figure 4.21 the point with highest outlyingness score seems to surrounded by other points and cannot be easily distinguished as an outlying point or not based on this bivariate view. When all variables are considered and the outlyingness score calculated, it is clearly visible that the point is far away from its nearest neighbour and needs further investigation. The points with higher outlyingness scores also coincided with the points identified by the O3 plot using a PCA projected space.

The advantage of embedding the outlyingness score in a classical scatter plot or Bagplot is that potential outliers, in a multivariate context, can be easily identified. The outlyingness score is larger if a data point is outlying in any combination of variables from a multivariate data and, unlike the O3 plot, it is not required to evaluate every possible combination to calculate the score.

4.5 SUMMARY

Data visualisation is one of the most important and effective way of communicating information in data. The ability to identify and visualise a point that is atypical in the context of multivariate data is of particular relevance, in particular when considering 'big data'. In this chapter, different visualisation approaches were presented from single variable plots to visualisation for multivariate data. Though all of them convey useful information they lack the ability to visually identify atypical data points in an efficient manner.

The limitation of existing classical visualisation were presented with an example using a case study involving 22 variables representing player's movement during a soccer game. The example started with a visualisation using a univariate boxplot followed by a two variable scatter plot and a bagplot. Dimension reduction techniques such as PCA, GLRM and t-SNE were then used and the results visualised. A more recent visualisation approach outliers O3 was then presented

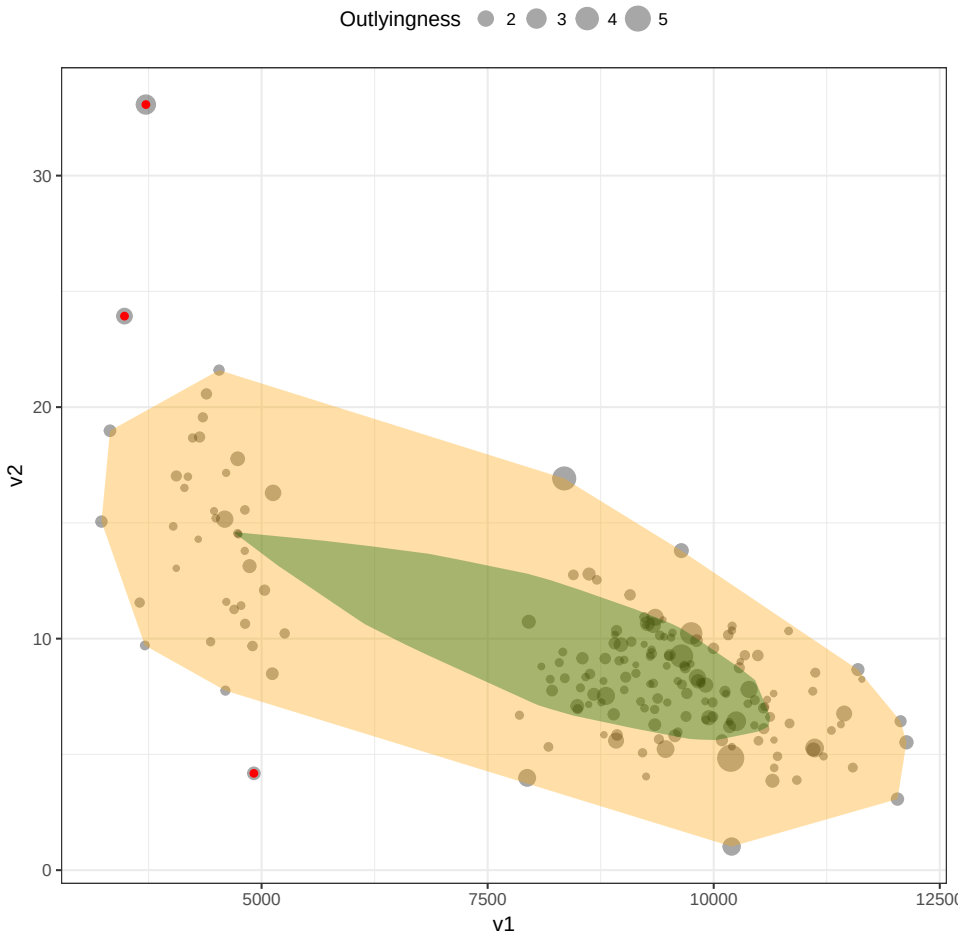


Figure 4.20: A bagplot of two selected variables from *sports data*. The size of the points are the outlying score; higher value indicate high potential to be outlying point in multivariate context

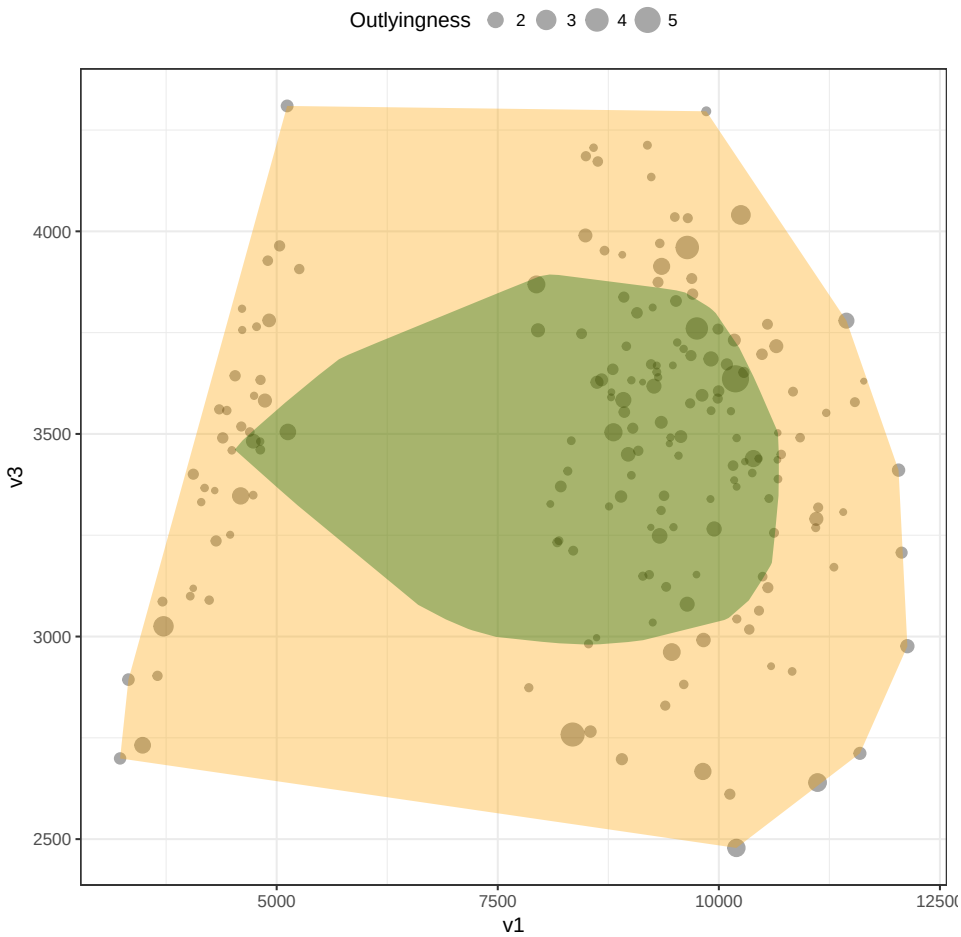


Figure 4.21: A bagplot of two selected variables from *sports data* where not outlying points were marked. The size of the points are the outlying score; higher values indicate high potential to be an outlying point in a multivariate context

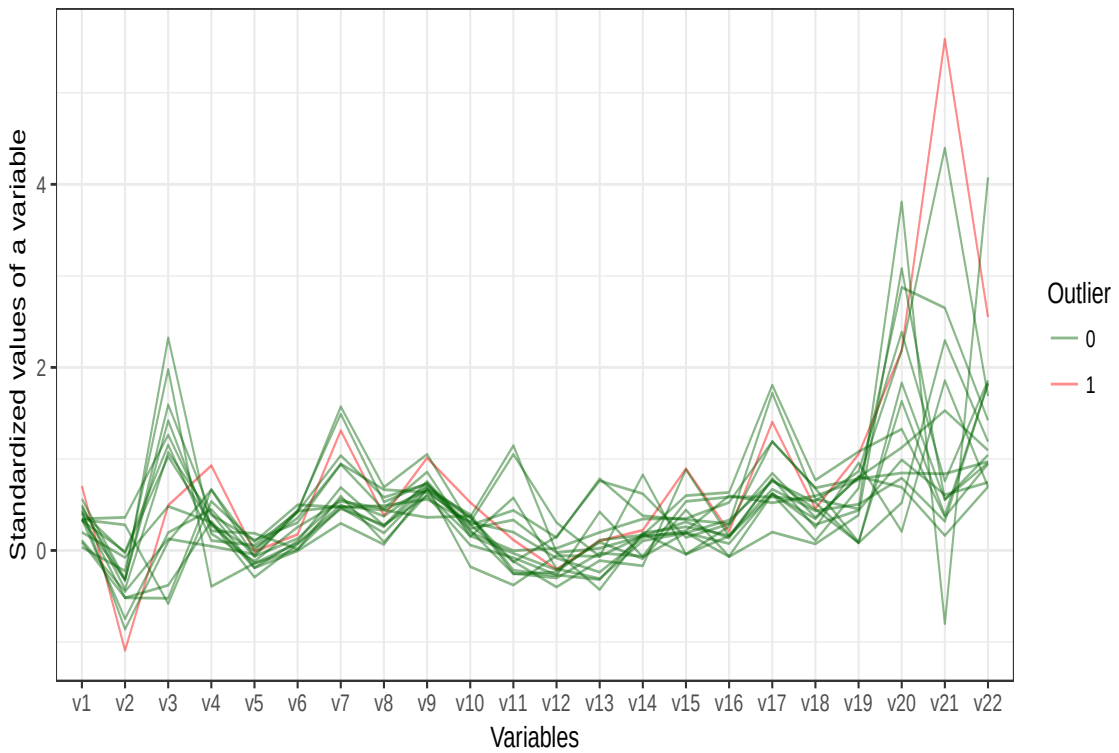


Figure 4.22: A parallel coordinate plot of potential outlying points and non outlying points filtered based on outlyingness scores

along with a discussion of its limitation in multivariate data when a large number of variables are present and a potential work-around was proposed.

The use of embedding the proposed statistical depth function outlying score into data visualisations was proposed in order to identify potential outlying points in general. Using this score as an additional graphical parameter, outliers can be identified more easily for further investigation.

In the chapters to date examples of the application of data science have been given to identify patterns and anomalies in textual and numerical data. The remaining chapter continues this theme where modern approaches in statistics and computer science are used again, this time in the domain of clinical research.

5 PREDICTING THE SEVERITY OF KNEE OSTEOARTHRITIS: A DATA SCIENCE CASE STUDY

5.1 INTRODUCTION

Osteoarthritis (OA) is the result and observable status of inflammatory processes in a joint leading to functional and anatomical impairments. The resulting status often shows irreversible damages to the joint cartilage and the surrounding bone structures [ABC+14; Eyr04].

Swelling, joint pain, and stiffness are the prominent symptoms among others, such as restrictions in movement including walking, stair climbing, and bending [Hei11]. The symptoms worsen over time and elderly patients are affected more frequently than patients in other age groups. The presence of knee OA (KOA) reduces activity in daily life and eventually leads to disability, which can incur high costs related to loss in productivity [Alt10].

Considering the impact of KOA on disability and the subsequent unavoidable economic burden, there is a need to quantify the severity of KOA during the early stages of development. KOA severity level helps in determining appropriate treatment decisions and for the monitoring of disease progression [BG12].

In this chapter the results of a collaborative project, involving investigators across the Insight nationally, are given to investigate where statistical modelling could be used to quantifying KOA severity. In a previous study[AMM+17], we introduced a fully convolutional neural network (FCN) was introduced to automatically detect and extract the knee joints from an X-ray image, and trained CNNs from scratch to predict the KOA severity in both discrete and continuous scales using classification and regression respectively[AMO+16; AMM+17].

The primary goal of this study is to extend this work to develop a predictive model of KOA severity using patient questionnaire data only while addressing the following questions:

- To what level of accuracy can a statistical model predict the severity score of knee osteoarthritis using patients' data without using X-ray images of knee?
- What is the prediction accuracy of the proposed statistical model to predict the severity score of knee osteoarthritis of the same patient using an X-ray image only?
- Are the predictions from patient data alone comparable to the predictions made using X-ray images?

In this chapter information is given on the available relevant data sources, the statistical approaches used and an assessment of the suitability of the proposed final model. Several statistical approaches are considered to predict the severity of KOA using patient questionnaire data. Furthermore, a convolution neural network (CNN) model is proposed to predict the same outcome using corresponding X-ray images for the same patients. The performance of both the approaches has been compared using the calculated root mean squared error on a validation set. As a secondary goal, key variables with the strongest predictive ability were identified, which may be useful to monitor a patient over time and design early interventions for prolonging healthy life in patients of concern.

5.2 KNEE OSTEOARTHRITIS

The knees are the most commonly affected joints in the human body and knee osteoarthritis (KOA) is more prevalent in females aged 60 years or more compared to males of the same age (13% vs 10%) [MLP+11]. Severity of KOA amongst females aged 55 and over is higher compared to their male counterparts and the severity of KOA is typically higher compared to other types of OA [ZJ10; PMC01]. Approximately one in every six patients consult with a general practitioner in their first year of an OA episode [ZJ10; PMC01]. The incidence of KOA has a positive association with age and weight and the prevalence is more common in younger age groups, particularly those who have obesity problems [BC09].

It is estimated that functional impairment of the knee and the hip are the eleventh highest disability factors [CSH+14] contributing to considerable socio-economic burden with an estimated cost per patient per year of approximately 19,000 Euro [PZ15]. The estimated prevalence of disability due to arthritis is expected to reach 11.6 million individuals by the year 2020 [DC+94], which is greater than the estimated risk of disability attributable to cardiovascular diseases or any other medical condition [GFA+94]. Total joint replacement surgery is the most favorable option to treat advanced stage OA. However, diagnosing the status of KOA at an early stage and providing behavioral interventions could be beneficial for prolonging a healthy life for a patient [KML+16].

The classical way of quantifying KOA severity is by inspection of X-ray images of the knee by a radiologist who then grades the images according to the KL scale (from 0 for "normal" up to 4 for "severe" stage) [KL57]. This approach suffers from high levels of subjectivity as there is no gold standard grading system: the semi-quantitative nature of the KL grading scale creates ambiguity, thus giving rise to disagreements between raters (for details please refer to [KL57; GJM+08; SCM+15]).

To reduce the influence of subjectivity in quantifying KOA severity from X-ray images, computer-aided diagnosis has been very helpful [DH89]. To date the sample size of available images has been the main limiting factor to train an efficient model [SFF+10; WPS+12; SLS+09; TOF+15]. The Osteoarthritis Initiative (OAI) [EMH07] and the Multi-centre Osteoarthritis Study (MOST) [SNG+13]

mitigated this small sample size limitation by making thousands of patients' data and X-ray images available. Recently, several researchers have used these resources to develop an automatic approach for quantifying KOA severity by analyzing X-ray images [SLS+09; OMA+08; AMO+16; AMM+17; TTR+18]. Although there have been multiple attempts to quantify KOA severity based on an automated analysis of X-ray images, so far there has been no attempt to build a predictive model on a patient's assessment data such as signs, symptoms, medication and other characteristics about a patient (later on referred as patient's questionnaire data) and to compare this approach against the X-ray based prediction. Developing predictive models using patient data other than X-ray images offers additional advantages such as identifying those variables that contribute strongest towards predicting the severity of KOA. A good predictive model based on patient's questionnaire data could reduce treatment costs and could also contribute to a prolonged healthy life of a patient due to early behavioral intervention.

In a review of possible risk factors of KOA, Heidari [Hei11] concluded that age, obesity, gender (i.e., female), repetitive knee trauma and kneeling are the most common risk factors for KOA. The common symptoms include pain, functional impairment, swelling and stiffness. The severity of KOA and the pain status is measured based on the Kellgren and Lawrence (KL) scale of 0 to 4 by visual inspection of the knee X-ray images [KL57].

5.3 OSTEOARTHRITIS INITIATIVE

The Osteoarthritis Initiative (OAI) is a multi-center longitudinal study for men and women sponsored by the National Institute of Health (NIH) to better understand KOA. Data collected through the OAI can provide useful information about the marginal distributions of relevant patient characteristics, their demographics, signs & symptoms and medication history. To date, there are more than 200 scientific publications that have used data collected through the OAI including several attempts to automate the KL grade quantification using X-ray images. But to date no study (or publication) has tried to predict KOA severity based on patient questionnaire data. Our primary goal is to compare the prediction accuracy of a statistical model based on patient questionnaire data to the prediction accuracy using X-ray images only.

5.3.1 Obtaining & Tidying Dataset

The OAI data used in this study are available for public access at `http://www.oai.ucsf.edu/`. The specific dataset used is labeled 0.2.2; data from the multi-center longitudinal and prospective observational study of KOA. The baseline dataset has been used for this work. The description of each variable used in our analysis is given in our corresponding published paper [AAM+19].

The baseline dataset contains a large number of variables related to patients' characteristics, their vital signs, symptoms of KOA, medication history, and functional impairment. At the early stage of the analysis, a manual inspection of each of the variables was carried out and a subset of candidate variables selected that were clinically relevant and previously reported risk factor for

KOA [Hei11; HMK08]. The completeness of the data was inspected in terms of missing values. The amount of missing values (in percent) was calculated for each variable. The variables that had at least 85% non-missing values were kept for further analysis. If it can be assumed that missing data are missing at random (i.e. missingness is explained by the covariates available) a multiple imputation step is unnecessary if a linear mixed model is used as the likelihood is correctly specified under this assumption. Moreover, we excluded categorical variables with very low discriminatory power, for example, the variables with very low frequency in one of the categories compared to the rest within the same variable. The reduced set of candidate variables was used for further processing and analysis. The dataset was split into two parts, training and validation sets, by taking a random sample of 70% of the data for training and the remaining 30% for validation. To make valid comparisons, the same validation set was used in the models developed on patient's questionnaire data and the model developed using X-ray images [AMO+16; AMM+17]. The data pre-processing steps are summarized in (Figure 5.1).

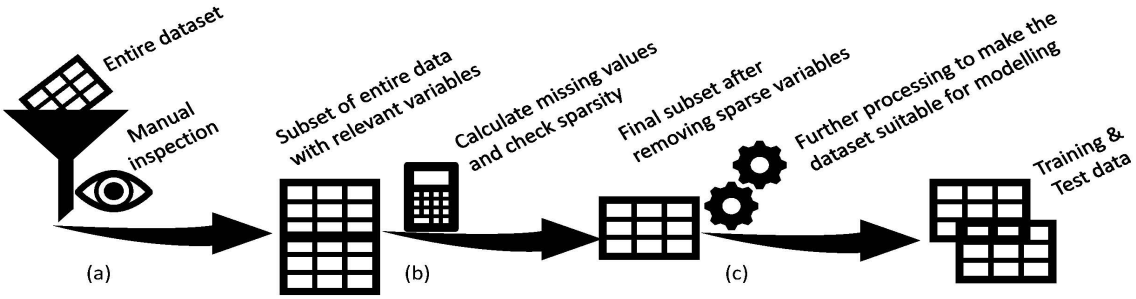


Figure 5.1: Data pre-processing work flow: (a) Inspection of the entire dataset manually to identify subsets of relevant candidate variables, (b) calculation of the percentage of missing values for each variables and also the sparsity of categorical variables. Drop variables that had more than 15% missing values or very low e.g. less than 5% into one category in a binary variable, (c) creation of dummy variables from multi-category variables and then splitting the dataset into training and test data for predictive model building.

To summarise and explore the explanatory variables, descriptive statistics were calculated: mean and standard deviation for numeric variables, frequency and percentage for categorical variables. The relationship among predictors was also explored by calculating the Pearson correlation between numerical variables, polyserial correlations between numerical and categorical variables and polychoric correlation between categorical variables [KÁ04].

The KL grade score was recorded on an ordinal scale from 0: normal to 4: severe. To model an ordinal outcome, ordinal logistic regression [McC80; And84] is the typical approach used. An ordinal logistic regression model was fitted, but the prediction performance was poor. In this analysis the severity score was also treated as a continuous response to investigate if this would improve predictive ability. Moreover, the data are hierarchical in structure; each patient had data for both knees. To capture this structure appropriately a linear mixed effect model incorporating a random effect at the subject level [LW82]was used .

5.3.2 Exploratory Analysis

The OAI dataset contains data for 4,796 individuals. After initial pre-processing, 2,951 patients with sufficient data on potential candidate explanatory variables were selected, representing 62% of the original patients. The remaining 38% of individuals did not have enough data for the potential explanatory variables and were not included in the analysis. The list of candidate variables, their labels and type (binary, numeric and categorical) has been outlined in a paper [AAM+19].

To train and validate the predictive models a training and validation data split was used as shown in Table 5.1 (roughly a 70% - 30% split). To make valid comparisons, the same validation set was used in the models developed using the patient questionnaire data and the model developed using X-ray images [AMO+16; AMM+17]. The validation set contained data for both knees for 846 patients, i.e. 1,692 data points for a knee in total. The training set for the predictive models included data from 2,105 patients, i.e. 4,210 knees.

Severity level	Training: Freq (%)	Validation: Freq (%)	Total: Freq (%)
Level 0	1818 (43.2)	685 (40.5)	2503 (42.4)
Level 1	728 (17.3)	312 (18.4)	1040 (17.6)
Level 2	1045 (24.8)	416 (24.6)	1461 (24.8)
Level 3	503 (12.5)	237 (14.0)	740 (12.5)
Level 4	115 (2.7)	42 (2.5)	157 (2.7)

Table 5.1: Distribution of KOA severity between training and validation data

The validation data that contains 30% of the original patients data are the same patients information that has been used in X-ray image based modelling. To make the results comparable with X-ray based prediction the same validation patients information was used, although cross validation was used to check sensitivity of the entire analysis. The cross validation result is consistent with the original 70-30 split.

Relevant summary statistics for patient characteristics are given in Table 5.2 for the entire dataset and for the training and the validation subsets. Good balance is evident when comparing the mean and variability for each patient characteristic across the training and validation data and it is plausible that they can be considered as representative samples taken from the same overall population. Maintaining a similar distribution of patient characteristics between the training and validation data is paramount for making reliable inference. An individuals recorded occupation was dropped from the patient characteristics table and from subsequent analysis as this variable had more than 30% missing data.

Characteristics	Training: Mean (SD)	Validation: Mean (SD)	Total: Mean (SD)
Age	60.3 (9.2)	61.1 (8.9)	60.5 (9.1)
Female (Freq. %)	1177 (56.0)	454 (53.7)	1631 (55.3)
Height (mm)	1685.2 (93.2)	1687.3 (92.6)	1685.8 (93.0)
Weight (kg)	80.7 (16.3)	80.5 (15.7)	80.6 (16.1)
BMI (kg/m^2)	28.3 (4.8)	28.2 (4.6)	28.3 (4.7)
Systolic	123.3 (15.9)	123.7 (16.7)	123.3 (16.1)
Diastolic	75.5 (9.8)	75.4 (9.6)	75.5 (9.8)

Table 5.2: Summary statistics of patient characteristics between complete, training and validation data

The box-plot (Figure 5.2) displays the distribution of several patient characteristics. Minor displacement quantity (jitters) were introduced along the horizontal axis and alpha-blending was used to make the display of the distribution of the points clearer. The level of KOA severity appears to be higher among elderly people. Height, weight and BMI show similar patterns but in contrast the distribution of blood pressure measurements does not indicate any strong obvious pattern with severity score. (Figure 5.2).

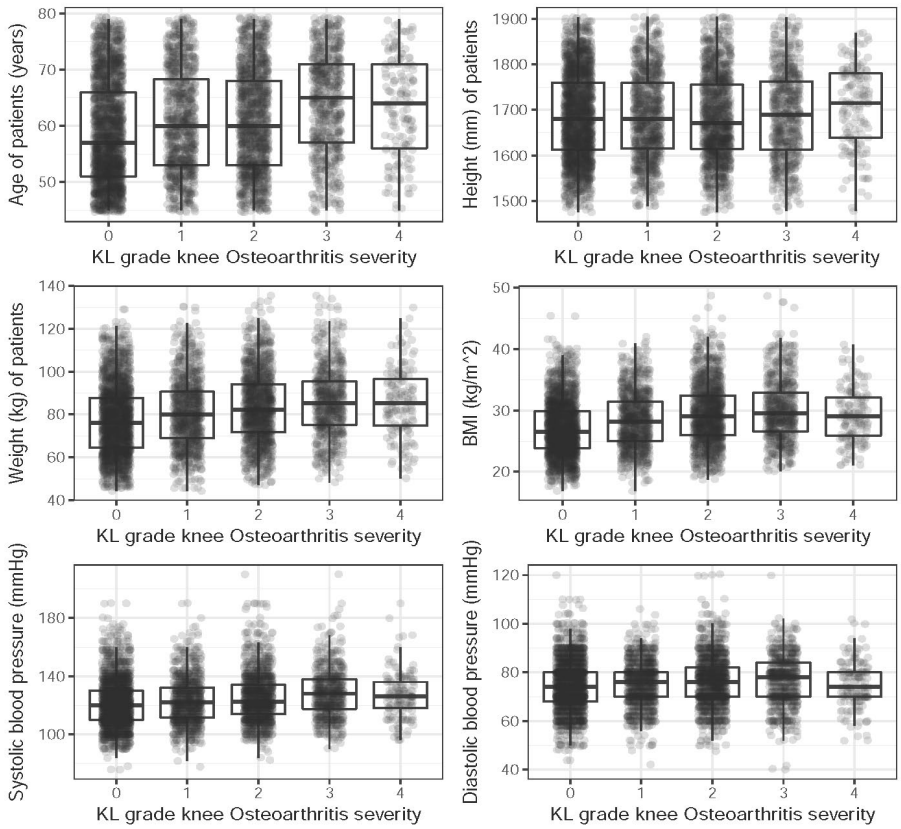


Figure 5.2: Boxplots of patient characteristics by Knee Osteoarthritis Severity

The data contain a mixture of continuous, binary, and categorical predictors. To observe the relationships among such a collection of predictor variables the Pearson correlation was calculated between continuous predictors, polyserial correlations between continuous and categorical predictors and polychoric correlation between categorical predictors [KÁ04]. The correlation matrix (Figure 5.3) depicts the relationship among the predictor variables of interest where a higher colour intensity indicates a stronger correlation between variables. The blue colour indicates a positive correlation and red colour indicates a negative correlation. We can observe that the predictor variables are positively correlated with each others to a moderate degree. Patients sex, height and weight shows weak negative correlation with other variables but only sex and height show a strong negative correlation. The upper block represents correlation among signs and symptoms in the left knee whereas the lower right block represents correlation among signs and symptoms of the right knee. The lower left block is the correlation between signs and symptoms of left to right knee. Other than the three blocks of correlation there are some variables that represent neither of the knees; rather, those variables represents medication history and other characteristics (Figure 5.3). What is clear from Figure 5.3 is the large number of candidate predic-

tors and the presence of multicollinearity amongst predictors which will have to be accounted for accordingly in any subsequent model.

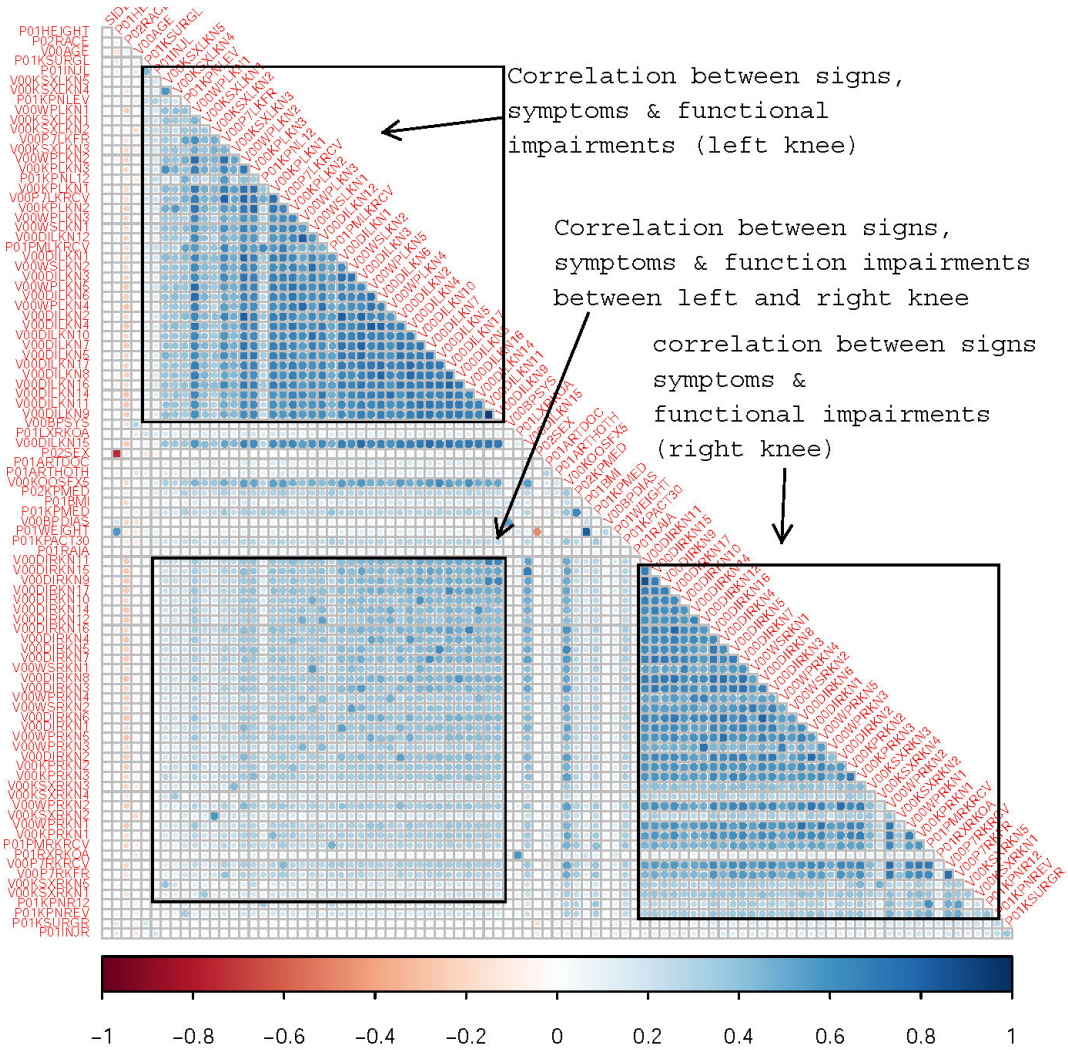


Figure 5.3: Correlation among predictors (dark colour indicates stronger correlation)

Among the five levels of KOA severity there were very few (2.7%) patients with severity level 4 (KL grade) in comparison to other categories. Overall 42% of patients were from severity level 0 indicating the normal knee followed by mild severity level 2 with 24% and doubtful severity level 1 with 17%. The distribution of severity level frequencies across the training and the validation data is well balanced indicating that it is plausible that the training and validation data came from same underlying population (Table 5.1.)

5.3.3 Statistical modelling

Several approaches were used build the predictive models needed and the merits of each are discussed below.

Elastic-net Regression

Elastic Net regression is a combination of ridge regression and LASSO, and this model is appropriate in the presence of correlated predictors [ZH05]. We denote the outcome variable: KL grade score by Y (considered as a continuous variable) and all predictors by $X_1, X_2, ..., X_p$. The Elastic Net regression linearly combines L_1 and L_2 penalties as follows:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda[(1-\alpha)\sum_{j=1}^{p} \theta_j^2 + \alpha \sum_{j=1}^{p} |\theta_j|] \tag{5.1}$$

Here $\hat{y}_i$ is calculated using linear regression with θ_i be the regression coefficients. The L_1 penalty is defined as the sum of absolute value of the regression coefficients and the $L_1 = \sum_{j=1}^{p} |\theta_j|$ and L_2 penalty is defined as the sum of squared values of the regression coefficients: $L_2 = \sum_{j=1}^{p} \theta_j^2$. The amount of mixing between two penalty terms is controlled by a mixing parameter α. If the value of $\alpha = 0$ then it leads to a ridge regression whereas a value of $\alpha = 1$ leads to a LASSO regression. The hyper-parameter λ controls the amount of shrinkage of regression coefficients for various values of α. A higher value of λ leads to shrink the regression coefficients towards zero and a very small value of λ has little effect on the regression coefficients. Using both the L_1 and L_2 penalty enables the selection of appropriate variables that have higher predictive power by shrinking some of the regression coefficient to zero using an appropriate value of hyper-parameter λ. To estimate the most suitable value for the shrinkage parameter λ repeated cross validations were performed with fixed values of the mixing parameter $\alpha = 0.5$ and the value of λ that minimizes the root mean squared error (RMSE) chosen. Figure 5.5, shows the cross-validation results while selecting λ.

Random Forest

Random Forests (RF) are an ensemble method that combine the predictive ability of multiple decision tree models. The RF model is an extension of the original work of Tin Kam Ho [Ho95] who developed the algorithm for random decision forests. Leo Breiman [Bre01] used the idea of bagging (bootstrap aggregating) and random variable selection. The principle of a random forest is to combine multiple tree based models to form a single model that can achieve better accuracy compared to its individual counterparts. This method takes a random sample with replacement from the original data, then builds a decision tree model based on a random selection of variables at each branch in the tree. This process is repeated for multiple trees and stores the prediction from each tree. The predicted value is then the mode (for a categorical response) for the mean (for a continuous response) across the forest. The random forest model is popular because it can reduce the variance of single tree models and also overcomes the problem of correlated predictors as it takes only a subset of candidate predictor variables in each of the individual trees.

Linear Mixed Effect Model

There is a clear hierarchical structure in the dataset as we have patient level data along with knee level data. A linear mixed effect model (LMM) [LW82] is an extension of a linear model that accounts for such hierarchical structure in data. The primary benefit of using a LMM in this analysis is that the uncertainty in knee level prediction is now correctly adjusted for through

the introduction of a suitable random effect. This approach will account for measurements on both knees collected for each subject correctly. The Intra-class correlation has been reported that indicates how much variation in the dependent variable is due to the random effect component in the LMM model. A random effect model can be formulated as:

$$y_{ij} = x_{ij}^t \beta + u_{ij}^t \gamma_i + \epsilon_{ij}; i = 1, 2, \cdots, m; j = 1, 2, \cdots, n_i \tag{5.2}$$

Here y_{ij} is the KL grade of i-th knee of j-th patient, x_{ij} the covariate of vector of j-th member of cluster i for fixed effects; u_{ij} covariate vector of j-th member of cluster i for random effects; γ_i is the random effect parameter, m is the the number of clusters (in our case $m = 2$ representing left and right knee), β is the regression coefficient of the fixed effect covariates.

Convolution Neural Network

In a machine learning based approach to automatically assess the KOA severity, the first step is to localize the region of interest (ROI), that is to detect and extract the knee joint regions from the X-ray images, and the next step is to classify the localized knee joints based on KL grades. In our previous study[AMM+17], we introduced a fully convolutional neural network (FCN) to automatically detect and extract the knee joints, and trained CNNs from scratch to predict the KOA in both discrete and continuous scales using classification and regression respectively[AMO+16; AMM+17]. Baseline X-ray images from the OAI dataset were used to train the CNN model. After testing different configurations, the network in Table 5.3 was found to be the best for classifying knee images. The network contained five layers of learned weights: four convolutional layers and a fully connected layer. Each convolutional layer in the network is followed by batch normalisation and a ReLU activation layer. After each convolutional stage there is a max pooling layer. The final pooling layer (maxPool4) is followed by a fully connected layer (fc5) with output shape of 1024 and a softmax dense (fc6) layer with output shape of 5 representing five level of KOA severity. To avoid overfitting, a drop out layer with a drop out ratio of 0.25 is included after the last convolutional (conv4) layer and a drop out layer with a drop out ratio of 0.5 after the fully connected layer (fc5). Also, a L2-norm weight regularization penalty of 0.01 is applied in the last two convolutional layers (conv3 and conv4) and the fully connected layer (fc5). Applying a regularisation penalty to other layers increases the training time whilst not introducing significant variation in the learning curves. The network is trained to minimize categorical cross-entropy loss using the Adam optimizer with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$. The inputs to the network are knee images of size 200×300. This size is selected to approximately preserve the aspect ratio based on the mean aspect ratio (1.6) of all the extracted knee joints.
After training, this network achieves an overall root mean-squared error 0.771 on the test data. Figure 5.4 shows the learning curves whilst training this network. The learning curves show proper convergence of the training and validation losses with consistent increase in the training and validation accuracy until they reach constant values.

5.3.4 Model building, evaluation, and comparison

Initially an Elastic Net regression [ZH05], a weighted combination of LASSO and Ridge regression, was fitted. An Elastic Net regression model can be used to select variables with high pre-

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	32×100×150
maxPool1	–	3×3	2	32×49×74
conv2	64	5×5	1	64×49×74
maxPool2	–	3×3	2	64×24×36
conv3	96	3×3	1	96×24×36
maxPool3	–	3×3	2	96×11×17
conv4	128	3×3	1	128×11×17
maxPool4	–	3×3	2	128×5×8

Table 5.3: Best performing CNN for classifying the knee images

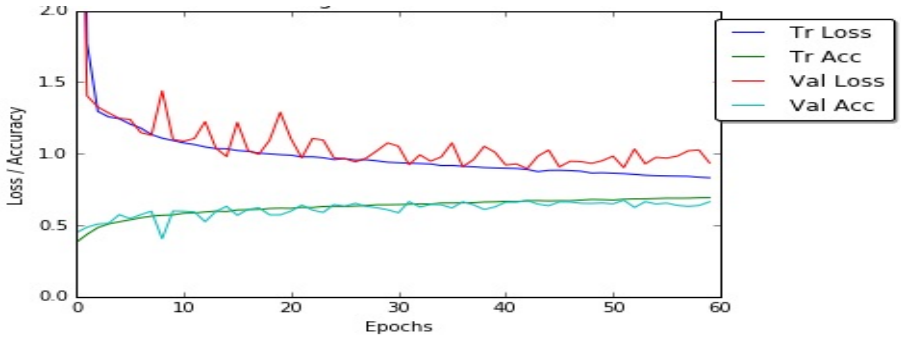


Figure 5.4: Learning curves: training and validation losses, and accuracy of the fully trained CNN.

dictive power. The weighting is controlled by the mixing parameter α that controls the amount of mixing between LASSO and Ridge penalties, whereas the parameter λ controls the amount of shrinking in the regression coefficients. To estimate a suitable value for the shrinkage parameter λ repeated cross validation using a fixed $\alpha = 0.5$ was performed, which corresponded to the minimum cross-validation RMSE. Using this value of α the value of λ that also minimizes the RMSE (Figure 5.5) was selected. The contribution (i.e. direction and magnitude) of each predictor variable has been extracted from the corresponding estimated regression coefficients (Figure 5.6).

A Random Forest [Bre01] regression model was then fitted using differing numbers of trees where the RMSE was calculated for each scenario. Based on these evaluations, it was found that using 100 trees produced the lowest RMSE in the validation set. Those predictors with highest variable importance Were identified in terms of improved predictive ability of the final forest.

The overall RMSE for the Elastic Net regression model is 0.97 and the RMSE for the random forest model is 0.94. Both models gave higher accuracy for the prediction of the severity levels 1 and 2 in contrast to the other categorical levels. The RMSE from the X-ray image based CNN model is 0.77, which is slightly lower than the RMSE from the Elastic Net regression and the Random Forest model. The advantage of using Elastic Net regression over a Random Forest regression model and X-ray image based CNN model is that the variables that have high predictive power can be easily identified as can the direction of the contribution of each variables by looking at the magnitude and sign of their regression coefficients (Figure 5.6).

The Elastic Net regression model produced higher prediction accuracy for severity level 1 and 2,

in comparison to other levels. A similar result is noted in the predictions by the Random Forest model. The overall RMSE of Elastic Net regression and Random Forest regression models are 0.974 and 0.943. The overall accuracy of the CNN model is higher than Elastic Net and Random Forest regression. The performance of each of the three models show their lowest outcome for the KOA severity level 4 as there is less data available in that category. Relatively higher accuracy in predicting KOA severity using an X-ray based CNN model has been observed however, the margin of difference between the RMSE of the predictions from the X-ray image based CNN model in comparison to the predictions from the patient's questionnaire data models is considerably small. Table 5.4 shows the RMSE for the models trained with patient data (Elastic Net and Random Forest)and the model trained with X-ray images (CNN regression).

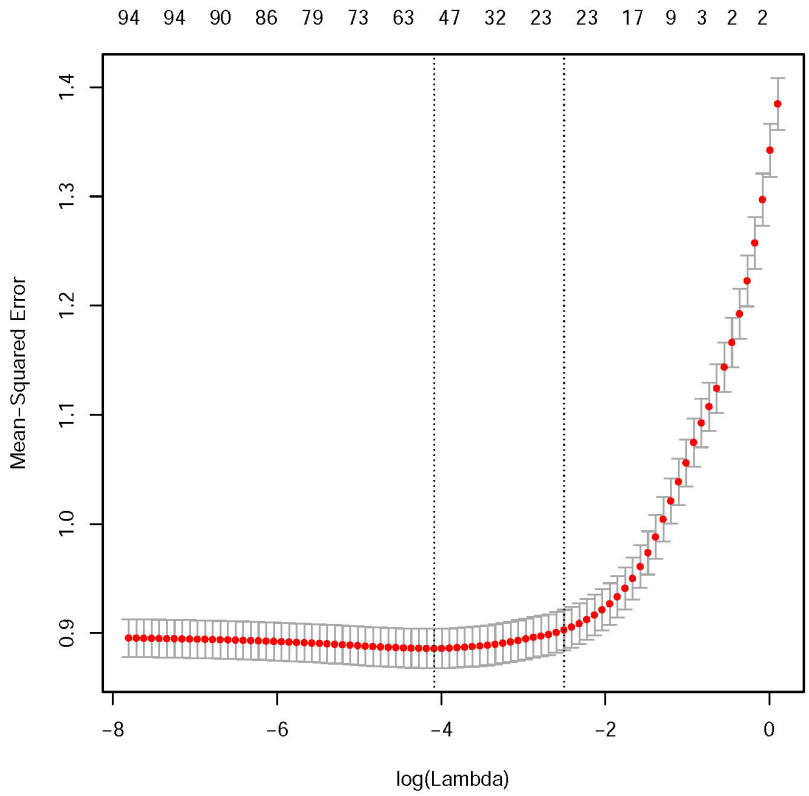


Figure 5.5: Estimation of hyper-parameter λ for Elastic net regression model. The Y-axis displays the RMSE for different values of λ whereas the upper horizontal axis represents the number of predictors. The RMSE increases as the number of predictors decrease but stabilizes after a certain number of predictors are added. The red points represents RMSE and the gray line segments represents a 95% confidence interval corresponding to each RMSE. The optimal value of λ is the minimum value corresponding to a steady-state RMSE.

Both the Elastic Net and Random Forest models allow the variable importance of the individual predictors on overall predictive ability to be calculated. There are some variables commonly identified by both these models with higher contribution towards the final predictions. However, the variables identified by the Elastic Net have more interpretable properties than the variables selected by the Random Forest model. The sign of the regression coefficients in an Elastic Net regression allows provide an understanding of the direction of the contribution; whether it increases the severity score or reduces it. A negative sign indicates a reduction in the overall

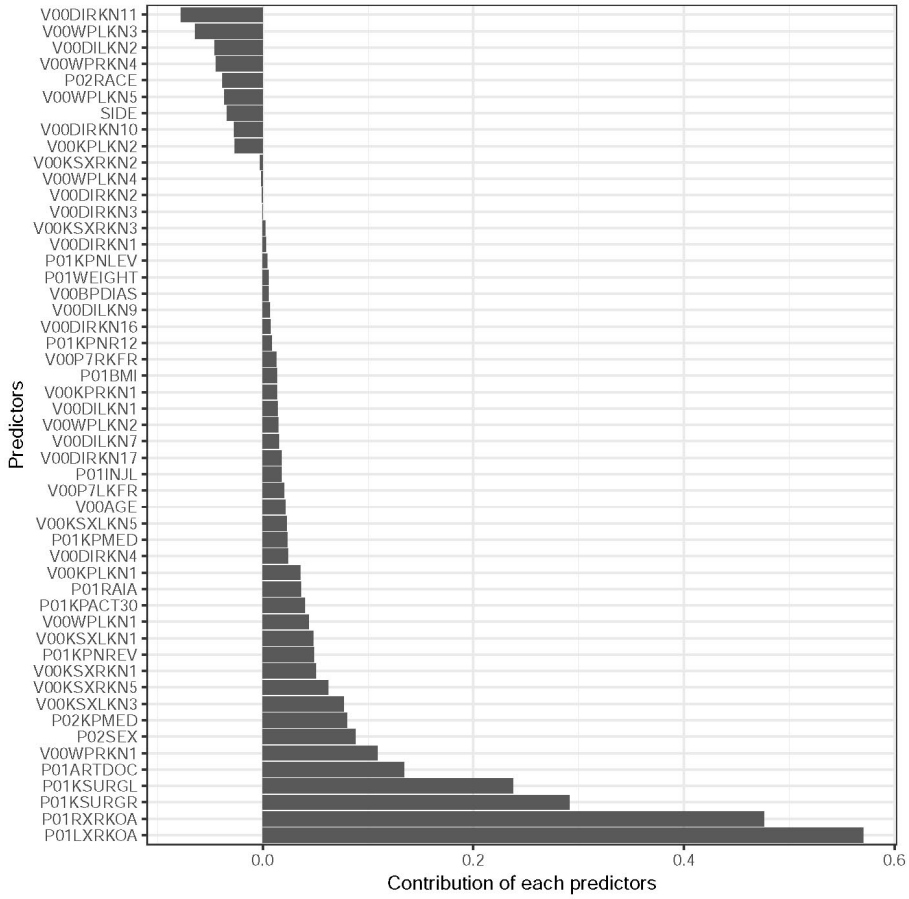


Figure 5.6: Contribution of each variable on KOA severity score prediction by Elastic Net Regression

Severity Level	Elastic Net Regression	Linear Mixed Model (LMM)	Random Forest Regression	CNN Regression
Level 0	0.917	0.920	0.909	0.816
Level 1	0.563	0.591	0.511	0.485
Level 2	0.881	0.895	0.853	0.840
Level 3	1.320	1.320	1.270	0.795
Level 4	2.140	2.10	2.02	0.846
Overall	0.973	0.978	0.943	0.770

Table 5.4: Estimated RMSE from different models for each level of KOA severity level

severity score for increasing values of the predictor, whereas a positive sign indicates an increase in the severity score for increasing values of the predictor. The direction of the contribution by predictors selected by the random forest model is unclear as it gives similar importance to both directions. Figure 5.6 shows the sign and magnitude of the contribution for each of the selected variables. The positive sign indicates an increase in the severity score whereas a negative sign indicates a decrease in the severity score. The identified variables could be a proxy indicators of patient knee's anatomical structure which ultimately indicates the level of severity.

As the data have a hierarchical structure (i.e. knee nested within patient) the number of replicates at the individual level is more appropriately modeled using a mixed effects model with a random effect to capture the correlation between knees within an individual. To explore the

random effect of patient level information, a linear mixed effects model was fitted using the predictor variables initially selected from the Elastic Net regression. There is clear evidence for the need of random effects due to the study design. In addition a small p-value (less than 0.001) was evident for the test for the need of the random effect term due to subject level within knee correlation in the model. The intra-class correlation coefficient is 0.265 which indicates that the proportion of the variance explained by the random effect component (patient level information) in the population 26.5%. The overall RMSE for the linear mixed effects model is 0.978, which is almost the same as the RMSE from the Elastic Net regression. However, the predicted severity levels, and more importantly the corresponding uncertainty, is correctly adjusted to account for the within patient correlation.

5.4 SUMMARY

Judging the impairment for patients with KOA requires a thorough understanding of the disease condition. Expert radiologists or clinicians assess the functional knee impairments and the KOA severity level from the X-ray images. Ideally, the image analysis should give an objective measure of the impairments; however, in reality not all functional impairments show up in anatomical transformations of the knee, and the patho-physiological evaluation relies on the subjective perception of the patient and the physician jointly.

The primary goal was to explore whether the prediction accuracy of a statistical model based on patient's questionnaire data is comparable to the prediction accuracy based on X-ray image based modelling to predict KOA severity. It has been demonstrated that statistical models, using patients' questionnaire data, could predict KOA severity level with a good level of accuracy (RMSE: 0.974 & 0.943). The prediction performance of the statistical models presented in this chapter are comparable to models using X-ray image data based on model performance as assessed by RMSE measures [AMO+16; AMM+17; TTR+18]. In particular it has been demonstrated that functional impairment at severity levels 1 and 2 can be predicted by our statistical models (Elastic Net & Random Forest and LMM) trained from the patients' assessment data to a level of accuracy similar to the accuracy achieved on the basis of CNN model trained on X-ray images. There are very subtle structural variations in the knee joints (minimal joint space narrowing (JSN) and osteophytes formation) belonging to grade 0 and grade 1, and these are not fully reflected in the KL grades. Also, there are relatively large overlaps in the JSN measurements for KL grades 0 and 1 compared to the other grades [HS03]. These factors make them challenging to distinguish by inspecting the X-ray images. Also, patients share almost similar distribution on their characteristics, signs, symptoms and functional impairments. Due to very subtle differences of the predictors between KOA levels the prediction accuracy gets affected.

The key variables that contributed most to the predictive ability in the models can be monitored over time to assess the progression of KOA severity. The strong indicator variables are reporting on knee baseline radiographic OA status for the right or left knee (P01LXRKOA, P01RXRKOA) and on treatments such as surgery on the right or left knee (P01KSURGR, P01KSURGL) as well as other reasons to see the doctor (P01ARTDOC). A patient's sex also plays an important role in predicting KOA severity. The next indicator variables cover medication (P02KPMED) and functional

impairments, pain or other symptoms to the right or left knee (V00WPRKN1, V00KSXLKN3, V00KSXRKN5, V00KSXRKN1, V00KSXLKN1, V00WPLKN1, P01KPNREV, P01KPACT30). An additional variable of interest is whether a doctor "ever said you have rheumatoid arthritis or other inflammatory arthritis" (P01RAIA). The predictors variable found as important predictors of KOA severity were also those important risk factors reported in previous studies [Hei11; HMK08].

Importantly, an early behavioral intervention could be developed based on the identified variables to prolong the healthy life of a patient. By observing the identified variables that have higher predictive ability to predict KOA severity, subjects can be identified who are currently taking medication for pain relief and facing functional difficulty in their daily life. Variables representing limited knee functions in particular are the potential indicators for quantifying KOA severity that could lead to developing targeted interventions for further treatment and medications.

When making predictions the LMM is favored as it is the only approach that correctly adjusts for the hierarchical structure present in the data. It is interesting that the severity levels 1 and 2 can be predicted with good accuracy in all the four models (EN, RF, LMM, and CNN), while the other levels of severity are more challenging to predict. For higher severity levels, i.e. levels 3 and 4, this could be due to the lack of patient data, i.e. the sample sizes at these levels are smaller than for levels 1 and 2 (Figure 5.2 & table 5.2).

As a conclusion based on the results in this paper, it can be concluded that patient questionnaire data can predict KOA severity level with good accuracy and it is comparable with the prediction based on X-ray images. Patient assessment data also enables an identification of some of the key variables that can be used to design early interventions and monitor the patients over the treatment period. The accuracy of the model developed using patient's assessment data is almost comparable to the CNN model. Moreover, the statistical models have an edge over the CNN model by identifying key variables that helps the physicians to design interventions and helps the patients for further treatment.

There is at least one potential limitation in developing any statistical model to predict KOA severity, that is the KL grade score itself is not a gold standard and suffers from subjectivity. The KL grade is dependent on the perception of the radiologist who is inspecting the X-ray images. i.e. a quasi-gold standard outcome. Considering this potential limitation, one way to improve the prediction accuracy could be to build a model of the X-ray image data in combination with the patients' assessment data. The prediction of KOA severity based on patients data shows comparable accuracy, it would be interesting to see the performance of prediction based on a statistical model combining both patient's questionnaire data and with X-ray images.

6 SUMMARY, CONCLUSIONS & FURTHER WORK

In this thesis the overall goal was to use modern approaches in data science to develop and evaluate a new approach to identify and visualise outliers in multivariate data. The proposed approach was evaluated using simulated and benchmark data. Examples from specific real world applications were also presented. In this chapter, a summary of each chapter is presented while highlighting the new scientific contribution in each.

6.1 SUMMARIES PER CHAPTER

6.1.1 Chapter-2: Data Science Approach to Literature Analysis

In the process of scientific discovery or in any research project, one of the most important steps is to look back in time and evaluate and critique published evidence. This step is usually achieved by conducting a systematic review. This step is becoming increasingly complex in domains with a large number of previous research studies and a manual review to cover all published material is not a feasible option. A semi-automated analysis of the texts retrieved from a scientific publications could be a useful companion tool and it could enhance the knowledge synthesize process.

In this chapter, the difficulties of manual systematic reviews of a large number of scientific publications is presented. To overcome the difficulty, a workflow is presented to augmented undertaking a literature review under these constraints. The proposed workflow is presented with specific discussion on data collection, cleaning, exploratory analysis, text mining and visualisation of the results.

One of the major steps in the proposed workflow is the analysis of text data using topic modelling. A brief description of classical topic modelling was presented. The primary goal is to uncover research sub themes which might not be easy to uncover by manual review. The hidden themes could give an indication of emerging research areas within a research application or could indicate common themes that most of the research studies are focusing on.

An open source web application (shiny app) is being developed that follows the workflow presented in this chapter. The web app will enable non-technical users to quickly and easily summarise collection of many research studies, especially abstracts. Using the web application a user can interact with the collection of abstracts to uncover underlying latent research themes and to visualise them in a meaningful way. Moreover, users can identify if there are any clusters among the abstracts.

In conclusion, the workflow and the web application presented in this chapter can be used to enhance traditional literature reviews by allowing the researcher gain a better understanding of constructs within the complete corpus that may be missed when reviewing manually. Furthermore, the tool can be used to explore the existing body of knowledge which ultimately helps to generate new research hypotheses. This chapter concludes with a small introduction of statistical depth function which is the primary topic of next chapter. An analysis of existing literature related to injuries in elite sport and on the use of depth functions and outlier detection is presented using the method and tool presented in this chapter.

6.1.2 Chapter-3: Outlier Detection in Multivariate Data

The concept of order statistics is an easy to understand statistical concept in the univariate setting. For a situation with more than one variable, the concept of order is not straightforward. The ability of extend the idea of ordering in multivariate data opens up many areas of study of the properties of multivariate distributions. In this chapter, a discussion of such an idea is presented as motivation for the chapters that follow including potential areas of application.

A statistical depth function, a mapping from $\mathbb{R}^d$ to $\mathbb{R}$ (where d is the number of variables), is a method to generate a univariate score for multivariate data and provides an ordering of multivariate data points. The order is defined from the centre of the distribution; a point that moves away from the centre of the distribution has a low depth value with the the deepest point at the centre. Such a score can be used to study properties of multivariate distributions, especially in identifying potential outlying points that might not be easy to identify in a marginal context.

In this chapter, the concept of statistical depth function is presented and its potential application in outlier detection discussed along with potential limitations. A new novel algorithmic approach is proposed, a modified Mahalanobis depth (kMMD), which is then evaluated using simulated data to assess it's ability to identify outlying data points in multivariate data.

The simulation experiments shows that the proposed approach works well and perform better than original existing algorithms. This new algorithm also been evaluated using benchmark dataset where the status of a data point is known (outlying or non outlying) beforehand. The proposed algorithm performed better in the simulated data as well as in benchmark data in terms of identifying outliers in multivariate data.

The proposed modified Mahalanobis Depth function has the potential to be used in any domain to identify probable outlying data points in multivariate data. The score that is being generated using the depth function can enhance any classical visualisation as a method to display the outlyingness of an observation. The application of the proposed depth function in data visualisation is presented in Chapter 4.

6.1.3 Chapter-4: Visualising Multivariate Data

Visually representing information from raw data is one of the most important aspects in data analysis and presentation. In the case of a single variable scenario, visualisation is easy but the difficulty arises as the number of variables increases. In this chapter, different visualisation approaches are presented starting from a single variable to the multivariate context with the goal to identify potential outlying data points (in multivariate context) visually while using classical plots. Such plots are augmented by the inclusion of the outlying score, calculated using the proposed statistical depth function, to enhance the ability to display the potential outlying points in a multivariate context regardless of the graphical approach used.

Visualising potential outlying data points is important to identify observations that may be influential in any statistical analysis or that have a very distinct characteristics compared to the majority of the data points. In a univariate situation, a potential outlying point can be easily visualised using a boxplot but the same idea is not generalizable to multivariate data. A Bivariate boxplot, namely a Bagplot, is an extension of a univariate boxplot and can display potential outliers in a bivariate context. When the number of variables variables is large this method for visualising outliers is not feasible however.

The limitation of existing classical visualisations in outlier detection is presented with an example using *sports data*; a dataset containing 22 variables measuring player movement during a soccer game. The example starts with univariate boxplots continuing to two variable scatter plots and Bagplots. Plots of composite variables using dimensionality reduction techniques such as PCA, GLRM and t-SNE are then considered. Each of these plots is useful in the sense of visualising data either for a marginal distribution (univariate plot) or using multivariate plots but are again limited in terms of outlier detection.

To overcome the limitation of these visualisation approaches, an outlying score using statistical depth function was calculated for each observation and was included as a graphical parameter to identify potential outlying points. The outlyingness score can be used in this way in any existing visualisation technique which displays the raw (or composite) data. Using this additional graphical parameter, users can identify and isolate atypical data points for further investigation.

6.1.4 Chapter-5: Predicting the Severity of Knee Osteoarthritis: A Data Science Case Study

In this chapter a collaborative project in developing and evaluating a predictive model to predict knee osteoarthritis severity is presented.

Expert radiologists or clinicians assess the functional knee impairments and the KOA severity level from the X-ray images. Ideally, the image analysis should give an objective measure of the impairments; however, in reality not all functional impairments show up in anatomical transformations of the knee, and the patho-physiological evaluation relies on the subjective perception of the patient and the physician jointly.

It has been demonstrated that the prediction performance of the statistical models presented

in this chapter are comparable to models using X-ray image data based on model performance as assessed by RMSE measures [AMO+16; AMM+17; TTR+18].

As a conclusion based on the results, it can be concluded that patient questionnaire data can predict KOA severity level with good accuracy and it is comparable with the prediction based on X-ray images. Patient assessment data also enables an identification of some of the key variables that can be used to design early interventions and monitor the patients over the treatment period.

6.2 FUTURE DIRECTIONS

In this thesis several novel ideas are presented with very good potential of practical application. Each of the ideas presented in this thesis could be further improved.

In Chapter 2, a new workflow along with an open source application (shiny app) for performing literature reviews is presented. The workflow could be further improved by incorporating network analysis of highly cited papers along with author network so that users can identify a new collaborative research network. The topic modelling section of the workflow could be further improved by including the most recent modelling techniques and the ability to perform topic network analysis to further enhance the applicability of the proposed workflow. Adding more recent document clustering algorithms in the document clustering part of the web application will provide the flexibility to investigate more rigorously and visualise the results to get further insights. A way to filter out the non-relevant topic and re-run the entire analysis could be added in a future version so that researcher can deep dive into more granular topic within a broader topic.

In Chapter 3, a new statistical depth function is proposed and evaluated. The proposed approach can be further improved in several aspects. Currently, the proposed approach is applicable to independent and identically distributed data. In future work, this could be extended to handle longitudinal data and functional data (i.e. the datum is a function) .

In the current proposed modified Mahalanobis depth function, one important property a depth function is relaxed. A rigorous mathematical investigation is required so that the effect of relaxing this assumption can be studied in more detail from mathematical perspective.

In terms of computational cost, the primary cost is to invert the co-variance matrix which is a once off calculation. This can be improved by incorporating more efficient algorithmic approaches.

For visualisation, an open source web application to implement the proposed modified Mahalanobis depth function in data visualisation could be useful.

In Chapter 5, an approach to modelling knee osteoarthritis severity is presented, where patient baseline characteristics and symptoms were used as potential predictors of severity level. It would be interesting to see if the proposed depth function could identify certain patients that

are 'different' from other patients and to investigate whether using the depth score to achieve this could improve the performance of the predictive models.

BIBLIOGRAPHY

[AAM+19] J. Abedin, J. Antony, K. McGuinness, K. Moran, N. E. O'Connor, D. Rebholz-Schuhmann, and J. Newell. "Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images". In: *Scientific reports* 9.1 (2019), pp. 1–11.

[ABC+14] N. Arden, F. Blanco, C. Cooper, A. Guermazi, D. Hayashi, D. Hunter, M. K. Javaid, F. Rannou, F. Roemer, and J.-Y. Reginster. *Atlas of osteoarthritis*. Springer, 2014.

[ACR+18] A. Amado, P. Cortez, P. Rita, and S. Moro. "Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis". In: *European Research on Management and Business Economics* 24.1 (2018), pp. 1–7.

[Alt10] R. D. Altman. "Early management of osteoarthritis." In: *The American journal of managed care* 16 (2010), S41–7.

[AMM+17] J. Antony, K. McGuinness, K. Moran, and N. E. O'Connor. "Automatic Detection of Knee Joints and Quantification of Knee Osteoarthritis Severity using Convolutional Neural Networks". In: *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer. 2017, pp. 376–390.

[AMO+16] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran. "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks". In: *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE. 2016, pp. 1195–1200.

[And84] J. A. Anderson. "Regression and ordered categorical variables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.1 (1984), pp. 1–22.

[APW+19] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, and R. A. Kievit. "Raincloud plots: a multi-platform tool for robust data visualization". In: *Wellcome open research* 4 (2019).

[ARO+09] S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas. "Supporting systematic reviews using text mining". In: *Social Science Computer Review* 27.4 (2009), pp. 509–523.

[ASM+10] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy. "On finding the natural number of topics with latent dirichlet allocation: Some observations". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2010, pp. 391–402.

[Atk94] A. Atkinson. "Fast very robust methods for the detection of multiple outliers". In: *Journal of the American Statistical Association* 89.428 (1994), pp. 1329–1339.

[Ats57] R. F. Atsatt. "The high school football Team Physician". In: *California medicine* 87.4 (1957), p. 263.

[AVY+09] F. Alqallaf, S. Van Aelst, V. J. Yohai, R. H. Zamar, et al. "Propagation of outliers in multivariate data". In: *The Annals of Statistics* 37.1 (2009), pp. 311–331.

[Bar76] V. Barnett. "The ordering of multivariate data". In: *Journal of the Royal Statistical Society: Series A (General)* 139.3 (1976), pp. 318–344.

[BC09] H. Bliddal and R. Christensen. "The treatment and prevention of knee osteoarthritis: a tool for clinical decision-making". In: *Expert opinion on pharmacotherapy* 10.11 (2009), pp. 1793–1804.

[BC83] R. J. Beckman and R. D. Cook. "Outlier..... s". In: *Technometrics* 25.2 (1983), pp. 119–149.

[BG12] H. J. Braun and G. E. Gold. "Diagnosis of osteoarthritis: imaging". In: *Bone* 51.2 (2012), pp. 278–288.

[BG18] T. Bergmanis and S. Goldwater. "Context sensitive neural lemmatization with lematus". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 1391–1400.

[BGC10] H. Bastian, P. Glasziou, and I. Chalmers. "Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?" In: *PLoS medicine* 7.9 (2010), e1000326.

[BL74] V. Barnett and T. Lewis. *Outliers in statistical data*. Wiley, 1974.

[BNJ03] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[Bre01] L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[Cas10] I. Cascos. "Data depth: multivariate statistics and geometry". In: *New perspectives in stochastic geometry* (2010), pp. 398–423.

[CBK09] V. Chandola, A. Banerjee, and V. Kumar. "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3 (2009), p. 15.

[CDP+08] Y. Chen, X. Dang, H. Peng, and H. L. Bart. "Outlier detection with the kernelized spatial depth function". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.2 (2008), pp. 288–305.

[Cha+36] M. P. Chandra et al. "On the generalised distance in statistics". In: *Proceedings of the National Institute of Sciences of India*. Vol. 2. 1. 1936, pp. 49–55.

[Cha03] I. Chalmers. "Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations". In: *The Annals of the American Academy of Political and Social Science* 589.1 (2003), pp. 22–40.

[CLY08] X. Cui, L. Lin, and G. Yang. "An extended projection data depth and its applications to discrimination". In: *Communications in Statistics—Theory and Methods* 37.14 (2008), pp. 2276–2290.

[CSH+14] M. Cross, E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C. L. Hill, et al. "The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study". In: *Annals of the rheumatic diseases* (2014), annrheumdis–2013.

[CTL17] W. L. Chang, K. M. Tay, and C. P. Lim. "A new evolving tree-based model with local re-learning for document clustering and visualization". In: *Neural Processing Letters* 46.2 (2017), pp. 379–409.

[CXL+09] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang. "A density-based method for adaptive LDA model selection". In: *Neurocomputing* 72.7-9 (2009), pp. 1775–1781.

[DC+94] C. for Disease Control, P. (CDC, et al. "Arthritis prevalence and activity limitations–United States, 1990." In: *MMWR. Morbidity and mortality weekly report* 43.24 (1994), p. 433.

[DDP+07] Y. Ding, X. Dang, H. Peng, and D. Wilkins. "Robust clustering in high dimensional data using statistical depths". In: *BMC bioinformatics*. Vol. 8. 7. BioMed Central. 2007, S8.

[DG+92] D. L. Donoho, M. Gasko, et al. "Breakdown properties of location estimates based on halfspace depth and projected outlyingness". In: *The Annals of Statistics* 20.4 (1992), pp. 1803–1827.

[DG11] S. Dutta and A. K. Ghosh. "On classification based on Lp depth with an adaptive choice of p". In: *Preprint* (2011).

[DG12] S. Dutta and A. K. Ghosh. "On robust classification using projection depth". In: *Annals of the Institute of Statistical Mathematics* 64.3 (2012), pp. 657–676.

[DG17] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[DG93] L. Davies and U. Gather. "The identification of multiple outliers". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 782–792.

[DH89] J. Dacree and E. Huskisson. "The automatic assessment of knee radiographs in osteoarthritis using digital image analysis". In: *Rheumatology* 28.6 (1989), pp. 506–510.

[DS10] X. Dang and R. Serfling. "Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties". In: *Journal of Statistical Planning and Inference* 140.1 (2010), pp. 198–213.

[DSB14] R. Deveaud, E. SanJuan, and P. Bellot. "Accurate and effective latent concept modeling for ad hoc information retrieval". In: *Document numeric* 17.1 (2014), pp. 61–84.

[EDD+13] A. F. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. "Systematic construction of anomaly detection benchmarks from real data". In: *Proceedings of the ACM SIGKDD workshop on outlier detection and description*. ACM. 2013, pp. 16–21.

[EMH07] F. Eckstein, T. Mosher, and D. Hunter. "Imaging of knee osteoarthritis: data beyond the beauty". In: *Current opinion in rheumatology* 19.5 (2007), pp. 435–443.

[Eyr04] D. R. Eyre. "Collagens and cartilage matrix homeostasis." In: *Clinical orthopaedics and related research* 427 (2004), S118–S122.

[FD95] R. Feldman and I. Dagan. "Knowledge Discovery in Textual Databases (KDT)." In: *KDD*. Vol. 95. 1995, pp. 112–117.

[FGG08] M. Febrero, P. Galeano, and W. González-Manteiga. "Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels". In: *Environmetrics: The official journal of the International Environmetrics Society* 19.4 (2008), pp. 331–345.

[FP01] D. Filmer and L. H. Pritchett. "Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India". In: *Demography* 38.1 (2001), pp. 115–132.

[GBK+09] N. Green, P. Breimyer, V. Kumar, and N. F. Samatova. "WebBANC: Building Semantically-Rich Annotated Corpora from Web User Annotations of Minority Languages". In: *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*. 2009, pp. 48–56.

[GC05] A. K. Ghosh and P. Chaudhuri. "On maximum depth and related classifiers". In: *Scandinavian Journal of Statistics* 32.2 (2005), pp. 327–350.

[GFA+94] A. A. Guccione, D. T. Felson, J. J. Anderson, J. M. Anthony, Y. Zhang, P. Wilson, M. Kelly-Hayes, P. A. Wolf, B. E. Kreger, and W. B. Kannel. "The effects of specific medical conditions on the functional limitations of elders in the Framingham Study." In: *American journal of public health* 84.3 (1994), pp. 351–358.

[GJM+08] L. Gossec, J. Jordan, S. Mazzuca, M.-A. Lam, M. Suarez-Almazor, J. Renner, M. Lopez-Olivo, G. Hawker, M. Dougados, and J. Maillefert. "Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI–OMERACT task force". In: *Osteoarthritis and cartilage* 16.7 (2008), pp. 742–748.

[Gru69] F. E. Grubbs. "Procedures for detecting outlying observations in samples". In: *Technometrics* 11.1 (1969), pp. 1–21.

[GS04] T. L. Griffiths and M. Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.

[GU16] M. Goldstein and S. Uchida. "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data". In: *PloS one* 11.4 (2016), e0152173.

[HA04] V. Hodge and J. Austin. "A survey of outlier detection methodologies". In: *Artificial intelligence review* 22.2 (2004), pp. 85–126.

[Haw80] D. M. Hawkins. *Identification of outliers*. Vol. 11. Springer, 1980.

[Hea97] M. A. Hearst. "Text data mining: Issues, techniques, and the relationship to information access". In: *Presentation notes for UW/MS workshop on data mining*. Vol. 1. 1997, p. 997.

[Hei11] B. Heidari. "Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I". In: *Caspian journal of internal medicine* 2.2 (2011), p. 205.

[Hen33] G. Hendon. "The treatment of fractures of the shaft of the femur". In: *The American Journal of Surgery* 20.3 (1933), pp. 542–554.

[HMK08] D. J. Hunter, J. J. McDougall, and F. J. Keefe. "The symptoms of osteoarthritis and the genesis of pain". In: *Rheumatic Disease Clinics of North America* 34.3 (2008), pp. 623–643.

[Ho95] T. K. Ho. "Random decision forests". In: *Document analysis and recognition, 1995., proceedings of the third international conference on*. Vol. 1. IEEE. 1995, pp. 278–282.

[Hob00] R. Hoberg. "Cluster analysis based on data depth". In: *Data Analysis, Classification, and Related Methods*. Springer, 2000, pp. 17–22.

[Hof99] T. Hofmann. "Probabilistic latent semantic analysis". In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, pp. 289–296.

[Hot33] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417.

[HS03] D. Hart and T. Spector. "Kellgren & Lawrence grade 1 osteophytes in the knee—doubtful or definite?" In: *Osteoarthritis and cartilage* 11.2 (2003), pp. 149–150.

[HS10] R. J. Hyndman and H. L. Shang. "Rainbow plots, bagplots, and boxplots for functional data". In: *Journal of Computational and Graphical Statistics* 19.1 (2010), pp. 29–45.

[ID87] A. Inselberg and B. Dimsdale. "Parallel coordinates for visualizing multi-dimensional geometry". In: *Computer Graphics 1987*. Springer, 1987, pp. 25–44.

[Ins85] A. Inselberg. "The plane with parallel coordinates". In: *The visual computer* 1.2 (1985), pp. 69–91.

[JCS+16] M.-H. Jeong, Y. Cai, C. J. Sullivan, and S. Wang. "Data depth based clustering analysis". In: *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. 2016, p. 29.

[Jör04] R. Jörnsten. "Clustering and classification based on the L1 data depth". In: *Journal of Multivariate Analysis* 90.1 (2004), pp. 67–89.

[JVZ02] R. Jörnsten, Y. Vardi, and C.-H. Zhang. "A robust clustering method and visualization tool based on data depth". In: *Statistical Data Analysis Based on the L1-norm and Related Methods*. Springer, 2002, pp. 353–366.

[KÁ04] S. Kolenikov and G. Ángeles. *The use of discrete data in principal component analysis with applications to socio-economic indices. CPC*. Tech. rep. MEASURE Working Paper No. WP-04-85, 2004.

[KL57] J. Kellegren and J. Lawrence. "Radiological assessment of osteoarthritis". In: *Ann Rheum Dis* 16.4 (1957), pp. 494–501.

[KML+16] M. Karsdal, M. Michaelis, C. Ladel, A. Siebuhr, A. Bihlet, J. Andersen, H. Guehring, C. Christiansen, A. Bay-Jensen, and V. Kraus. "Disease-modifying treatments for osteoarthritis (DMOADs) of the knee and hip: lessons learned from failures and opportunities for the future". In: *Osteoarthritis and cartilage* 24.12 (2016), pp. 2013–2021.

[LCL12] J. Li, J. A. Cuesta-Albertos, and R. Y. Liu. "DD-classifier: Nonparametric classification procedure based on DD-plot". In: *Journal of the American Statistical Association* 107.498 (2012), pp. 737–753.

[Lip58] A. B. Lipscomb. "Observations on posterior dislocation of the elbow joint in athletes". In: *The American Journal of Surgery* 96.3 (1958), pp. 393–395.

[Liu+90] R. Y. Liu et al. "On a notion of data depth based on random simplices". In: *The Annals of Statistics* 18.1 (1990), pp. 405–414.

[Liu06] R. Y. Liu. *Data depth: robust multivariate analysis, computational geometry, and applications*. Vol. 72. American Mathematical Soc., 2006.

[Liu92] R. Liu. *Data depth and multivariate rank tests. In* L_1-*Statistical Analysis and Related Methods (Y. Dodge, ed.) 279–294*. 1992.

[LMS+18] D. A. Liem, S. Murali, D. Sigdel, Y. Shi, X. Wang, J. Shen, H. Choi, J. H. Caufield, W. Wang, P. Ping, et al. "Phrase Mining of Textual Data to Analyze Extracellular Matrix Protein Patterns Across Cardiovascular Disease". In: *American Journal of Physiology-Heart and Circulatory Physiology* (2018).

[LR+91] H. P. Lopuhaa, P. J. Rousseeuw, et al. "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices". In: *The Annals of Statistics* 19.1 (1991), pp. 229–248.

[LR09] S. López-Pintado and J. Romo. "On the concept of depth for functional data". In: *Journal of the American Statistical Association* 104.486 (2009), pp. 718–734.

[LS93] R. Y. Liu and K. Singh. "A quality index based on data depth and multivariate rank tests". In: *Journal of the American Statistical Association* 88.421 (1993), pp. 252–260.

[LW82] N. M. Laird and J. H. Ware. "Random-effects models for longitudinal data". In: *Biometrics* (1982), pp. 963–974.

[McC80] P. McCullagh. "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

[MCF+15] T. Müller, R. Cotterell, A. Fraser, and H. Schütze. "Joint lemmatization and morphological tagging with lemming". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 2268–2274.

[MF17] J. Matejka and G. Fitzmaurice. "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 1290–1294.

[MGT57] A. Meier, W. Grannis, and J. Tanner. "ACROMIOCLAVICULAR DISLOCATIONS—Open Reduction with Screw Fixation". In: *California medicine* 87.4 (1957), p. 261.

[MH06] K. Mosler and R. Hoberg. "Data analysis and classification with the zonoid depth". In: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 72 (2006), p. 49.

[MH08] L. v. d. Maaten and G. Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[MLP+11] S. L. Murphy, A. K. Lyden, K. Phillips, D. J. Clauw, and D. A. Williams. "Subgroups of older adults with osteoarthritis based upon differing comorbid symptom presentations and potential underlying pain mechanisms". In: *Arthritis research & therapy* 13.4 (2011), R135.

[Mos02] K. Mosler. *Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach*. Vol. 165. Springer Science & Business Media, 2002.

[Mos13] K. Mosler. "Depth statistics". In: *Robustness and complex data structures*. Springer, 2013, pp. 17–34.

[Oja83] H. Oja. "Descriptive statistics for multivariate distributions". In: *Statistics & Probability Letters* 1.6 (1983), pp. 327–332.

[OMA+08] H. Oka, S. Muraki, T. Akune, A. Mabuchi, T. Suzuki, H. Yoshida, S. Yamamoto, K. Nakamura, N. Yoshimura, and H. Kawaguchi. "Fully automatic quantification of knee osteoarthritis severity on plain radiographs". In: *Osteoarthritis and Cartilage* 16.11 (2008), pp. 1300–1306.

[OTM+15] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. "Using text mining for study identification in systematic reviews: a systematic review of current approaches". In: *Systematic reviews* 4.1 (2015), p. 5.

[Pat60] R. Patton. "Football injuries to the shoulder girdle". In: *The American Journal of Surgery* 99.5 (1960), pp. 633–635.

[Pea01] K. Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[PL+08] B. Pang, L. Lee, et al. "Opinion mining and sentiment analysis". In: *Foundations and Trends in Information Retrieval* 2.1–2 (2008), pp. 1–135.

[PMC01] G. Peat, R. McCarney, and P. Croft. "Knee pain and osteoarthritis in older adults: a review of community burden and current use of primary health care". In: *Annals of the rheumatic diseases* 60.2 (2001), pp. 91–97.

[PMD16] O. Pokotylo, P. Mozharovskyi, and R. Dyckerhoff. "Depth and depth-based classification with R-package ddalpha". In: *arXiv preprint arXiv:1608.04109* (2016).

[PPT+15] N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou, and I. Iliopoulos. "Protein–protein interaction predictions using text mining methods". In: *Methods* 74 (2015), pp. 47–53.

[PRT+00] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. "Latent semantic indexing: A probabilistic analysis". In: *Journal of Computer and System Sciences* 61.2 (2000), pp. 217–235.

[PZ15] J. Puig-Junoy and A. R. Zamora. "Socio-economic costs of osteoarthritis: a systematic review of cost-of-illness studies". In: *Seminars in arthritis and rheumatism*. Vol. 44-5. Elsevier. 2015, pp. 531–541.

[RD99] P. J. Rousseeuw and K. V. Driessen. "A fast algorithm for the minimum covariance determinant estimator". In: *Technometrics* 41.3 (1999), pp. 212–223.

[RL05] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Vol. 589. John wiley & sons, 2005.

[RL87] P. Rousseeuw and A. Leroy. "Robust regression and outlier detection". In: (1987).

[RRT99] P. J. Rousseeuw, I. Ruts, and J. W. Tukey. "The bagplot: a bivariate boxplot". In: *The American Statistician* 53.4 (1999), pp. 382–387.

[RU11] A. Rajaraman and J. D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[RV16] P. Rousseeuw and W. Van den Bossche. "Detecting anomalous data cells". In: *arXiv preprint arXiv:1601.07251* (2016).

[RW96] D. M. Rocke and D. L. Woodruff. "Identification of outliers in multivariate data". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 1047–1061.

[SB12] A. Sunikka and J. Bragge. "Applying text-mining to personalization and customization research literature–Who, what and where?" In: *Expert Systems with Applications* 39.11 (2012), pp. 10049–10058.

[SCM+15] L. Sheehy, E. Culham, L. McLean, J. Niu, J. Lynch, N. A. Segal, J. A. Singh, M. Nevitt, and T. D. V. Cooke. "Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the Multicenter Osteoarthritis Study (MOST)". In: *Osteoarthritis and cartilage* 23.9 (2015), pp. 1491–1498.

[Ser02] R. Serfling. "A depth function and a scale curve based on spatial quantiles". In: *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Springer, 2002, pp. 25–38.

[Ser06] R. Serfling. "Depth functions in nonparametric multivariate inference". In: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 72 (2006), p. 1.

[SFF+10] L. Shamir, D. T. Felson, L. Ferrucci, and I. G. Goldberg. "Assessment of osteoarthritis initiative–kellgren and Lawrence scoring projects quality using computer analysis". In: *Journal of Musculoskeletal Research* 13.04 (2010), pp. 197–201.

[SLS+09] L. Shamir, S. M. Ling, W. Scott, M. Hochberg, L. Ferrucci, and I. G. Goldberg. "Early detection of radiographic knee osteoarthritis using computer-aided analysis". In: *Osteoarthritis and Cartilage* 17.10 (2009), pp. 1307–1312.

[SMC+16] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, et al. "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible". In: *Nucleic acids research* (2016), gkw937.

[SNA+14] M. Y. Sharker, M. Nasser, J. Abedin, B. F. Arnold, and S. P. Luby. "The risk of misclassifying subjects within principal component based asset index". In: *Emerging themes in epidemiology* 11.1 (2014), p. 6.

[SNG+13] N. A. Segal, M. C. Nevitt, K. D. Gross, J. Hietpas, N. A. Glass, C. E. Lewis, and J. C. Torner. "The Multicenter Osteoarthritis Study: opportunities for rehabilitation research". In: *PM&R* 5.8 (2013), pp. 647–654.

[Tan+99] A.-H. Tan et al. "Text mining: The state of the art and the challenges". In: *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. Vol. 8. sn. 1999, pp. 65–70.

[TIR98] J. Theodore, K. Ivy, and T. Raymong. "Fast Computation of 2d depth contours". In: *ACM SIG KDD* (1998), pp. 224–228.

[TOF+15] J. Thomson, T. O'Neill, D. Felson, and T. Cootes. "Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 127–134.

[TTR+18] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala. "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach". In: *Scientific reports* 8.1 (2018), p. 1727.

[Tuf01] E. R. Tufte. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT, 2001.

[Tuk75] J. W. Tukey. "Mathematics and the picturing of data". In: *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*. Vol. 2. 1975, pp. 523–531.

[Tuk77] J. W. Tukey. *Exploratory data analysis . ch. 2*. 1977.

[UHZ+16] M. Udell, C. Horn, R. Zadeh, S. Boyd, et al. "Generalized low rank models". In: *Foundations and Trends® in Machine Learning* 9.1 (2016), pp. 1–118.

[Unw15] A. Unwin. *Graphical data analysis with R*. Vol. 27. CRC Press, 2015.

[Unw19] A. Unwin. "Multivariate outliers and the O3 Plot". In: *Journal of Computational and Graphical Statistics* 28.3 (2019), pp. 635–643.

[Ven11] O. Vencálek. "Concept of data depth and its applications". In: *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* 50.2 (2011), pp. 111–119.

[VZ00] Y. Vardi and C.-H. Zhang. "The multivariate L1-median and associated data depth". In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1423–1426.

[Wil12] L. Wilkinson. "The grammar of graphics". In: *Handbook of Computational Statistics*. Springer, 2012, pp. 375–414.

[Wil16] L. Wilkinson. *Visualizing outliers*. 2016.

[WPS+12] T. Woloszynski, P. Podsiadlo, G. Stachowiak, and M. Kurzynski. "A dissimilarity-based multiple classifier system for trabecular bone texture in detection and prediction of progression of knee osteoarthritis". In: *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 226.11 (2012), pp. 887–894.

[YH17] D. Yang and J. Hong. "Performing literature review using text mining, Part II: Expanding domain knowledge with abbreviation identification". In: *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE. 2017, pp. 3297–3301.

[YR01] K. Y. Yeung and W. L. Ruzzo. "Principal component analysis for clustering gene expression data". In: *Bioinformatics* 17.9 (2001), pp. 763–774.

[YS97] A. B. Yeh and K. Singh. "Balanced confidence regions based on Tukey's depth and the bootstrap". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.3 (1997), pp. 639–652.

[ZH05] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[ZJ10] Y. Zhang and J. M. Jordan. "Epidemiology of osteoarthritis". In: *Clinics in geriatric medicine* 26.3 (2010), pp. 355–369.

[ZS00] Y. Zuo and R. Serfling. "General notions of statistical depth function". In: *Annals of statistics* (2000), pp. 461–482.

[Zuo+03] Y. Zuo et al. "Projection-based depth functions and associated medians". In: *The Annals of Statistics* 31.5 (2003), pp. 1460–1490.