



Pattern Discovery In Adverse Event Data

by

Zhicheng Zhang

A thesis submitted for the degree of Doctor of Philosophy of the
University of London and for the Diploma of Imperial College

Department of Mathematics

Section of Statistics

Imperial College London

University of London

180 Queen's Gate, London SW7 2AZ, United Kingdom

March 2007

Abstract

Pharmacovigilance is an area where associations need to be detected in large (often sparse) contingency tables, containing counts of drug and adverse drug events. The data are collected in spontaneous reporting databases, where voluntary reporting is the only source of information. Because of this, problems often arise due to the unreliability of the data and especially the baseline information, which in turn makes most measures of association unreliable.

Some of the standard methods in Pharmacovigilance attempt to remedy this problem by deriving a smoothed version of some measure of association, while others try to ignore this all together.

In this thesis, we firstly review some of the methods previously used in Pharmacovigilance, including the odds ratio and more recent methods whereby different forms of smoothing are applied to the basic forms of association measures. We outline some of problems associated with these methods with some illustrative examples.

Secondly we outline a new approach to the problem, which is based on extracting structural information from the data. This information can be derived from various sources both inherent in the data and external to the data. This is achieved by making use of the similarities among drugs and adverse drug events according to their profiles. We discuss various distances suitable for measuring these similarities.

Thirdly we introduce a new algorithm for exploring high-dimensional data sets, which we call Peaker. We discuss the various properties of Peaker when used in an exploratory sense, in terms of the two facets of pattern detection and pattern verification. We introduce tests for the purposes of pattern verification. We illustrate Peaker using examples, both simulated and real.

Finally we apply the new profiling approach and the Peaker algorithm to the problem of signal detection in Pharmacovigilance. In this case we apply the methods

to real data sets used in the Pharmacovigilance area to illustrate the improvements and differences to the older methods.

Acknowledgements

I would like to thank my supervisors Professor David Hand and Dr Steven Barrett who both provided untiring sources of ideas, advice and guidance throughout my studies at Imperial College. I would also like to thank them for the encouragement and generous donation of time and effort they gave me during the time of my studies.

I would also like to thank Professor Xiaohui Liu and Dr Duncan Gillies, my external and internal examiners, for all their constructive comments and recommendations during my Viva.

At Imperial, I would like to thank Dr Niall Adams for advice I received and discussions which we shared, Dr Richard Bolton and Dr Nick Heard for their advice during the time of my Ph.D. I would also like to thank Adrien, Ajay, Georgia, Kitty, Matt, Patrick and Ross for all the help they provided during my Ph.D.

Over the years, this thesis has benefited from constructive comments and suggestions by the participants at the conferences and seminars, which included the IDA 2005 conference, Madrid, Spain and at GlaxoSmithKline *etc.*

The financial support from an EPSRC CASE award in conjunction with GlaxoSmithKline is gratefully acknowledged. This award provided me the opportunity to work on this project and write this thesis. I would also like to thank GlaxoSmithKline for the use of their equipment and database. The people at GlaxoSmithKline also made me feel very welcome and provided a warm working atmosphere.

Finally, my special thanks go to my parents for their constant support and encouragement during the writing of this thesis.

Contents

1	Introduction	15
1.1	Background and Motivation	15
1.2	Shortcomings of Clinical Trials	16
1.3	Pharmacovigilance and Post Marketing Surveillance	18
1.3.1	Introduction	18
1.3.2	The Need for Post Marketing Surveillance	20
1.3.3	Issues in Post Marketing Surveillance	21
1.4	The Role of Data Mining in Pharmacovigilance	23
1.4.1	Introduction to Data Mining	24
1.4.2	The Need for Data Mining	24
1.4.3	Potential Improvements	25
1.5	Brief Description of the Data	27
1.6	Thesis Outline	28
2	Summary of Data	30
2.1	Background	30
2.2	Spontaneous Reporting Database	31
2.2.1	Overview	33
2.2.2	Coding Dictionary	35
2.2.3	Strengths and Weaknesses of Spontaneous Reporting Systems	36

2.3	Processing the Raw Data	40
2.4	The GPRD Database	41
2.4.1	Overview	41
2.4.2	Strengths and Weaknesses Associated with the GPRD	42
2.4.3	Mining the GPRD Database for the COX-II Data	44
2.5	Data Issues	45
3	Review of Methods	47
3.1	Notation	47
3.2	Measures of Association	48
3.2.1	Discrepancy Measures	48
3.2.2	Ratio Measures	50
3.3	Bayesian Approach	51
3.3.1	Empirical Bayes and Shrinkage Estimators	52
3.3.2	Gamma Poisson Model	53
3.3.3	DuMouchel's Gamma Poisson Shrinker	54
3.3.4	Beta Binomial Model and the Bayesian Confidence Propagation Neural Network	55
3.3.5	Bayesian Nonparametrics	56
3.4	Application	58
3.5	Discussion	60
3.6	Other Issues	62
4	Alternative Approaches	64
4.1	Motivation	64
4.2	Improvements	65
4.3	Profiling and Expert Knowledge	66
4.4	Outline of Remainder of Thesis	67

5	Profiling	69
5.1	Basics	69
5.2	Uncertainty	70
5.3	DuMouchel's Approach	71
5.4	Distance Metric	71
5.4.1	Euclidean Distance	71
5.4.2	χ^2 Distance	72
5.4.3	Jensen-Shannon Divergence	72
5.4.4	Bayes Factor	73
5.4.5	Comparisons of Distances	74
5.5	Multinomial Formulation	77
5.5.1	Method	77
5.5.2	Prior Specification	79
5.5.3	Distance	79
5.5.4	Application	82
5.6	Application and Discussion	84
5.6.1	Conclusion	89
6	Data Mining Software	91
6.1	Features of a Modern Data Mining Package	92
6.1.1	User Interface	93
6.1.2	Toolset	94
6.1.3	Visualisation and Graphics	96
6.1.4	Customisation	97
6.1.5	Database Extraction	99
6.1.6	Deployment and Exporting	99
6.2	Comparison of Software	100

6.2.1	Advantages and Disadvantages of Data Mining Packages . . .	100
6.3	Examples	104
7	Peaker	106
7.1	Method	107
7.2	Density Estimation and Bandwidth Selection	109
7.3	Identification of Significant Peaks	112
7.3.1	Setting M	115
7.3.2	Sharpness of a Peak	118
7.3.3	Probability Mass Contained in Peaks	120
7.3.4	Testing for Large Peaks	122
7.3.5	Size of the Local Neighbourhood	123
7.4	Simulated Study	124
8	Anomaly Detection Under the New Approach	131
8.1	Detecting Groups of Similar Drugs	131
8.2	Comparison of Signals Using Simulated Data	133
8.2.1	Simulation	133
8.2.2	Discussion	138
8.3	Combining Distances	139
8.4	Anomaly Detection	141
8.5	Example	142
8.6	Conclusion	148
9	Conclusion and Future Research	149
9.1	Data Mining	151
9.2	Summary of Thesis	154
9.3	Conclusion	156
9.4	Future Work	158

A Tables of Signals	160
B Cluster Analysis Results	166

List of Figures

2.1	MEDWATCH Voluntary Report Form	34
4.1	Flow Diagram of Remainder of Thesis	67
5.1	Plots of Four Beta Distributions	80
5.2	Plots of 1st v 2nd Principal Component of Distance Matrix Obtained by Multinomial Approach	83
5.3	Plots of 1st v 2nd Principal Component of Distance Matrix Obtained by Basic Profiling Method	88
6.1	Workspace for the Statistica Package	93
6.2	Stream Workspace for the Statistica Package	94
6.3	Customisation in Statistica	98
6.4	Model and Script Exporting in Statistica	100
6.5	KXEN User Interface	102
6.6	KXEN User Interface	102
6.7	KXEN Model Output Options	103
6.8	KXEN Lift Curves	103
6.9	Model Curves in KXEN for Adult Data Set	105
6.10	Variable Contribution in KXEN for Adult Data Set	105
7.1	Example of a Small Local Peak	110

7.2	A Maximum	113
7.3	A Maximum Peak	114
7.4	A Non-Maximum Peak	114
7.5	Simulation Results for Convergence	119
7.6	First Two Principal Components of 10-dimensional Data With Three Modes	125
7.7	Histograms of Peaks Corresponding to Distances From True Gaussian Means	127
7.8	Histograms of Peaks Corresponding to Distances From True Gaussian Means (Continued)	128
8.1	First Two Principal Components Drugs Data Set Showing Peaks of the Common Drugs	132
8.2	First Two Principal Components Drugs Data Set Showing All Peaks	133
B.1	Cluster Analysis Plot of DuMouchel's Method (Single Linkage)	166
B.2	Cluster Analysis Plot of DuMouchel's Method (Complete Linkage) . .	167
B.3	Cluster Analysis Plot of Kullback-Liebler Distance (Single Linkage) .	167
B.4	Cluster Analysis Plot of Kullback-Liebler Distance (Complete Linkage)	168
B.5	Cluster Analysis Plots of χ^2 Distance (Single Linkage)	168
B.6	Cluster Analysis Plots of χ^2 Distance (Complete Linkage)	169
B.7	Cluster Analysis Plots of Bayes Factor Distance (Single Linkage) . . .	169
B.8	Cluster Analysis Plots of Bayes Factor Distance (Complete Linkage) .	170

List of Tables

- 2.1 The MedDRA Hierarchy 35
- 3.1 Table of Presence of Drug and ADE 50
- 5.1 Three Vectors 70
- 5.2 Ordering of Distances (M1(high) denotes vectors of high counts generated under Multinomial distribution 1, *etc.*) 76
- 5.3 Kullback-Liebler Distances Between Four Distributions 81
- 5.4 Prior Specification Based on the Mean 82
- 5.5 Four Vectors 87
- 6.1 Summary of Available Analyses 95
- 6.2 Summary of Visualisation Tools 97
- 8.1 Strength of Simulated Signals With GPS 135
- 8.2 Strength of Simulated Signals After Reducing Size of Table With Peaker 137
- 8.3 Strength of Simulated Signals Within Each Group Picked by Peaker . 138
- 8.4 Highest 20 Odds Ratios 145
- 8.5 Highest 20 GPS Signals 146
- 8.6 20 Signals (Peaks) Detected Using New Approach (Using GPS as the Baseline Measure) 147
- A.1 Highest 40 Odds Ratios 161

A.2 Highest 40 GPS Signals	162
A.3 Highest 40 Gamma-Poisson Measures	163
A.4 Highest 40 Beta-Binomial Measures (Method of Moments)	164
A.5 Highest 40 BCPNN Signals (Uniform Prior)	165

List of Acronyms

ADE(s) - Adverse Drug Event(s)

AERS - Adverse Event Reporting System

AMISE - Asymptotic Mean Integrated Squared Error

BCPNN - Bayesian Confidence Propagation Neural Network

CSM - Committee on Safety of Medicines

FDA - Food and Drug Administration

GPS - Gamma-Poisson Shrinker

GPRD - General Practice Research Database

GSK - GlaxoSmithKline

HLGT(s) - High Level Group Term(s)

HLT(s) - High Level Term(s)

ISE - Integrated Squared Error

JSE - James-Stein Estimator

MCA - Medicines Control Agency

MDP - Mixture of Dirichlet Process

MDS - Multidimensional Scaling

MedDRA - Medical Dictionary for Regulatory Activities

MHRA - Healthcare Products Regulator Agency

MISE - Mean Integrated Squared Error

MLE - Maximum Likelihood Estimator

MM - Method of Moments

OR - Odds Ratio

PMML - Predictive Model Makeup Language

PRR(s) - Proportional Reporting Ratio(s)

PT(s) - Preferred Term(s)

RR - Reporting Ratio

SOC(s) - System Organ Class(es)

SRD(s) - Spontaneous Reporting Database(s)

SRS(s) - Spontaneous Reporting System(s)

WHO - World Health Organization

Chapter 1

Introduction

1.1 Background and Motivation

If one looks at the packaging of any drug, the one thing that is instantly disconcerting is the number of side effects associated with it. Take Aspirin for instance, there is a huge list of potential side effects ranging from the trivial (drowsiness or indigestion) to the serious (vomiting and seizures) and it is suspected that deaths can also occur as a result of a severe reaction. Whilst drugs can save lives, they can also kill. The number of incidents of side effects to drugs is huge. In the United States alone, it has been estimated (as of 1998) that the number of instances of death attributed to drug side effects is in the region of some 100,000 each year (the estimate is that some 3 in every 1000 hospital patients die from an adverse reaction). A similar number of deaths can be attributed to medical mistakes. The situation can be far worse in other countries, especially in many third world countries where drug monitoring is backward and moreover the use of illegal drugs can be profligate. Of course, the statistic is not as shocking as it initially appears since, in many cases, the patients are suffering from serious conditions and need drastic treatments which have known serious side effects. But in many cases severe adverse reactions can be caused by

innocent looking drugs and in this case it is imperative that the user is aware of the associated risks. This should impress upon the reader the magnitude of the problem associated with drug safety and the need for a thorough procedure for drug monitoring.

It is generally the case that most drugs have potentially serious side effects, especially if taken inappropriately. The importance of adequately detecting these side effects cannot be stressed too highly. Moreover, it is just as important (some may argue more important) to have enough power in the detection process so that the rare side effects (which as a rule tend to be more serious) can also be picked up.

For the remainder of this chapter, we will discuss the area of safety detection which is specifically designed to detect and act upon side effects of drugs. We will give the advantages and disadvantages of various methods of detection (we will refer to these methods as safety detection or adverse event detection). We will discuss some issues and difficulties associated with this area, possible improvements and the need for improved methodologies.

1.2 Shortcomings of Clinical Trials

The task of detecting side effects initially rests with the pharmaceutical industry. They are required to carry out the necessary clinical trials for the evaluation of the risks and benefits of a treatment before a drug is permitted to be marketed. Clinical trials are split into four parts which are:

- Phase I - The initial stages of a trial involving a small population (tens of people). Subjects are studied extensively, generally in an inpatient clinic and phase I trials normally include dose-range studies.
- Phase II - A larger population is selected (some 100–300) to assess the efficacy of the treatment. This usually includes a continuation of the phase I trial in

a larger population.

- Phase III - A randomised control trial is designed to further test for the efficacy of a treatment, especially in comparison to other alternative treatments.
- Phase IV - This trial generally involves the post-marketing assessment of a drug in the phase III population.

Phase IV trials are designed for pharmaceutical companies to discover the long term risks and benefits of a drug. Large and carefully designed cohort studies can be quite effective in picking up the majority of side effects, particularly the majority of commonly occurring side effects. However, many issues arise that question the effectiveness of such trials, of which the following are a few:

- The population used in clinical trials tends to be much more limited when compared to the wider population at risk. Typically clinical trial populations average only a few thousand people which, while effective for the assessment of efficacy, can be lacking when the issue of safety is at stake. Moreover clinical trials tend to consist of healthier and younger populations whereas the real world populations at risk to drugs tend to be the opposite (special patient groups such as children and the elderly are generally not represented in clinical trials). Additionally, long term effects may not be studied well in a clinical trial. In particular, it is often the case that, due to the restriction of a clinical trial population, many serious but rare side effects cannot be effectively detected (consider a side effect with an incidence rate of one in a million). This leads to the situation that experience about drugs can be very limited at the time of marketing and this is verified by the fact that new side effects are constantly discovered after marketing.
- Most pharmaceutical companies tend to place more emphasis on drug efficacy

rather than drug safety. Phase IV clinical trials often are barely carried out as a mandatory necessity rather than with the aim of genuine risk prevention.

- In many cases there are various pressures to market a drug as soon as possible. These can include financial reasons (clinical trials can be very expensive to maintain) as well as the simple need to save lives. As such, safety considerations can often take somewhat of a back-seat.
- It is generally perceived that there is a shroud of secrecy when it comes to drug safety in the pharmaceutical industry and often healthcare professionals are kept somewhat in the dark concerning many safety issues. This can be specially true in certain countries. Moreover, certain companies will try to minimise or deny the side effects of their drugs and this can very much complicate matters in terms of effective drug usage.

Given the points mentioned above, it seems clear that other processes are needed in complement to the phase IV trials when the issue of drug safety is at stake. In particular, it seems very much necessary that a follow-up process to phase IV trials is needed to effectively detect the rare and serious side effects. In the next section we will describe a process of drug monitoring which has been very effective in picking up side effects that clinical trials tend to miss.

1.3 Pharmacovigilance and Post Marketing Surveillance

1.3.1 Introduction

Pharmacovigilance is a French word, which has been described by Begaud as

“A discipline involving detection, evaluation and prevention of undesirable effects of medicine”.

The United Kingdom Medicines and Healthcare products Regulatory Agency (MHRA) defines Pharmacovigilance as

“...the process of a) monitoring medicines as used in everyday practice to identify previously unrecognised or changes in the patterns of their adverse effects; b) assessing the risks and benefits of medicines in order to determine what action, if any, is necessary to improve their safe use; c) providing information to users to optimise safe and effective use of medicines and d) monitoring the impact of any action taken”.

The World Health Organization (WHO) defines Pharmacovigilance as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems...”.

In particular, we are interested in the area of post-marketing surveillance which refers to the constant monitoring of adverse drug events (ADEs) that is carried out after the marketing of a drug to ensure the safety of the drug in question, both in the short term and the long term. Trontell (2001) defined an ADE as any adverse event associated with the use of a marketed drug, whether or not that event is considered drug related¹. This is necessary to ensure that the rare and often serious ADEs undetected at the phase IV clinical trials stage can be adequately detected. Much of the surveillance relies on voluntary reports by healthcare professional and users of drugs. Much of current practice in the area of Pharmacovigilance uses reports generated via spontaneous reporting systems (SRSs). These include the WHO database and the United States Food and Drug Administration (FDA) database as well as

¹In Trontell (2001) a serious ADE is defined as the occurrence of any of a set of specific outcomes, regardless of dose, including death, life-threatening ADEs, hospitalization or the prolongation of hospitalization, persistent or significant disability or incapacity, congenital anomaly or birth defect, or other situations in which intervention is required to prevent the occurrence of one of the preceding events. All other ADEs are defined as nonserious.

many other national regulatory agencies. The aim is to detect and act upon as quickly as possible any potential ADEs. There have been constant advances made in this area due to the rising interest in drug safety. Improvements in handling and storage of reports and the advent of electronic data transfer have also vastly improved the ease with which analyses can be made. The FDA uses a network of data resources to supplement detection of ADEs via the Adverse Event Reporting System (AERS). These data resources can be used to amplify, validate, and quantify ADE signals and then compare them to their expected background occurrence in the population. Use of additional databases and perspectives will improve the ability to detect ADEs in all settings, including the pediatric population, and to monitor risk management efforts to curtail the occurrence of known ADEs.

Many good accounts of Pharmacovigilance have been published, including Lindquist et al. (1999), O'Neill & Szarfman (2001), Szarfman et al. (2002), Jones (2001), Amery (1999*b*) and Amery (1999*a*).

1.3.2 The Need for Post Marketing Surveillance

The main role of post-marketing surveillance in general is to complement clinical trials and attempt to discover the rare and serious side effects which are not labelled. Most monitoring systems tend to be designed to achieve this objective. Various guidelines have been put forward to make the practice as scientific and ethically minded as possible. We have already mentioned many of the problems associated with clinical trials and the need for an effective system of drug safety monitoring. We now summarise some of the issues giving rise to the need for post-marketing surveillance:

- There is a deficiency of clinical trials and the need for assessment in the population at risk, in particular the long term assessment of drug safety and the

assessment of rare events. An ongoing surveillance program is necessary to cover these aspects.

- There is a need for an inexpensive and effective way of maintaining a monitoring system. This system should be carried out independent of the pharmaceutical companies and as such is free of the pressures that can affect the efficiency of clinical trial results. Clear guidelines can be suggested and maintained in this case for the best interest of science and ethics.
- There is a need for a way for users and medical practitioners to effectively report side effects and to change the method of prescription in the event of any serious risk of treatment. Many instances of historical disasters could have been avoided with the introduction of a system of rapid alerts.
- There is a need for an international safety surveillance system which is not based on companies and practices. This will require the support of the pharmaceutical industry and moreover the introduction of a more integrated pharmaceutical market. In Europe there are some prospects of this occurring although there remain many difficulties.

1.3.3 Issues in Post Marketing Surveillance

Although it seems an intrinsically important idea to monitor ADEs caused by drugs after marketing, the level of reporting throughout Europe has been relatively low. This has been well documented and the estimate is that only somewhere between 10 – 25 percent of reactions are reported and these tend to consist mostly of well known ADEs. This can be attributed to various reasons, including the following:

- Lack of public awareness about the role of Pharmacovigilance and ADEs. Many people simply do not realise that they can report suspected ADEs or

have no viable means to do so. This is especially true in certain countries where drug safety awareness is very low and there are no organised bodies to push for more public education in this area.

- In Europe, the lack of a single unified pharmaceutical market means that disparate regions have different practices and guidelines. International safety surveillance bodies find this situation quite difficult since one has to take into account different incidence rates which cannot be easily merged without considering some hidden factors.
- The notification of ADEs by healthcare professionals can be unsystematic. This leads to many people being kept somewhat in the dark on crucial issues of drug safety. In certain countries this lack of good communication can be particularly severe.
- Many countries simply do not maintain a spontaneous reporting system or such systems are inadequately maintained. This is clearly an undesirable situation but it is rather common. Moreover, as a result of this, international monitoring organisations have considerable difficulty in integrating data from disparate geographical regions.
- The emphasis placed on drug safety can vary considerably across countries and practices. Many countries do not even acknowledge this major issue and consequently any necessary actions are neglected. There have been some initiatives in pushing for a more effective international drug monitoring system but the situation at present is not ideal.

1.4 The Role of Data Mining in Pharmacovigilance

Data mining methods have emerged to define associations in many types of databases. Specific methods include artificial neural networks, Bayesian probability approaches, genetic algorithms, decision trees, nearest neighbour methods, rule induction, and new data visualization techniques. Much of the attention in post-marketing surveillance has been paid to the process of signal detection, with some recent advances as well as improvements on the older methods². The signal detection process is vital to safety surveillance. To evaluate the safety of a drug after detection, vast resources are required and this can be both expensive and time consuming. While in contrast the signal detection stage is relatively inexpensive and fast (one advantage of a spontaneous reporting system is that it is relatively cheap to maintain). Consequently improvements made here can lead to a significant reduction in follow-up work.

The key to reducing the amount of follow-up work is to reduce the number of false positive signals (but hopefully not at the expense of detection power). It can be difficult to do this via an improvement in the current methodology since any method is highly dependent on the data available (which in this case can be quite poor). A viable alternative is to invest in data mining methods which will improve the understanding of the data and thus lead to improved ways of detection. In particular, it is to be hoped that data mining can significantly improve the understanding of the baseline data (structures intrinsic in the data) and so overcome the shortcomings of the data quality. We will now discuss this in further detail in the following sections.

²Evans (2000) accounted that “Statistical methods for utilizing these data have been described for a long time, but have not been well used”.

1.4.1 Introduction to Data Mining

Data mining has enjoyed a surge in popularity in recent years, due to the fact that it has been able to find solutions in many diverse areas of science. In particular, areas such as physical sciences, social science and computational sciences have seen many techniques emerge that have been able to solve problems previously proven impossible to investigate. Along with this emergence much literature have emerged which seek to quantify data mining more precisely. Some of the most important ideas can be found in Hand et al. (2001), Hand et al. (2002), Hand (1998) and Hand et al. (2000).

1.4.2 The Need for Data Mining

To better detect signals, it is imperative to better understand the data. To this end data mining is ideal for making use of information in the data to explore structures. We can give the following reasons for the suitability of data mining to the subject area:

- The data sets tend to be large and standard analyses are difficult to utilise. There is scope for many data reduction techniques. In particular, in the case of contingency table analysis (which will be described later), the table in this case can be very sparse and so traditional hypothesis generation methods are not always applicable.
- Data quality is often quite poor. Many types of analyses will not work well against such poor data. Data mining contains many techniques for handling poor data quality and it is often the case that a rather *ad hoc* method (but with good understanding of the data) will work better than a more theoretically sound method.

- Extra information is available for safety data (simple examples would be classifications of drugs and ADEs). Much of this information can be utilised to discover interesting structures (both local and global) about the data. Data mining techniques are ideally suited to explore these structures. It should give a lot of insight into the data to perform certain types of exploratory analysis to help to better understand the data.
- External forces can often affect the way in which the data behaves. In keeping with the last note, this is another case of interesting structures and trends in the data, but from a different standpoint. Again data mining techniques can be used to discover these structures.

Recently there has been markedly more interest in data mining techniques to explore the safety databases. Lilienfeld (2004) delved into several interesting issues in this regard. In particular, there has been a call to establish a more systematic approach to this area and to discover whether a general data mining approach has more to offer than the current collection of rather restricted methods.

A thorough discussion on data mining can be found in Hand et al. (2001)³.

1.4.3 Potential Improvements

In practice, the safety detection analysis involves a hypotheses generation process. The problem is to look for ADEs which exceed some threshold (anomalously high count) beyond which action needs to be taken. How such thresholds are set tend to be rather arbitrary and many practices have individuals requirements based on experience and preference.

An alternative way to look at this is that we are looking for unknown (unsupervised) anomalous patterns. In particular we are interested in small, localised

³Chen & Liu (2004) gave an account of the aspects of data mining applied to information science.

anomalies since it is likely that larger patterns (or the more obvious patterns) will tend to point to ADEs which are common and already known. It is here that data mining can be an invaluable tool in identifying and analysing these anomalies. In common with most data mining applications, we are looking for localised patterns of interest in the data.

While there have now been marked improvements made in the collection and storing of reports on ADEs, the development of statistical techniques to explore these databases is still somewhat in its infancy. Although there have been some recent advances in Bayes and Empirical Bayes based methods, as well as the odds ratio (OR) based methods, much still relies on examining case reports.

Furthermore, as mentioned, we are typically faced with relatively large data sets in post-marketing surveillance. Data mining is particular well suited for exploring such large databases. Hand et al. (2001) described data mining as “searching for global and local patterns of interest that exist within large databases”, such that conventional statistical techniques are not well equipped to deal with. This is precisely the objective here.

Much work is still to be done in this area. Current practices are varied and there does not seem to be an underlying principle behind this practice. Furthermore, there do not seem to be any rules regarding good practice in the area and as a result various practices use various methods. This has led to a proliferation of methods which leaves one rather confused as to how to approach the problem at hand. The methods themselves do not seem to significantly improve on one another at a practical level and so much of the decision making is left to one’s preference.

We believe that in searching for such patterns of interest, the use of ratio based methods (described in section 3.2) can have some disadvantages. In particular there seems to be no specific way of making the full use of all information available, some of which can be very useful. The lack of more detailed information precludes the

possibility of developing a more systematic approach to the subject area. So as a preliminary comment, we believe that we need to establish some global or local structure with which we would be able to establish richer models. This does not necessarily mean more complicated models and indeed it is preferable to look for simpler models due to the size and poor quality of the data. That said, however, many techniques are available in data mining and it will not be at all straightforward to decide on the appropriate path. Herein lies our challenge, and the scope of such work presents much excitement. This will be discussed further in the subsequent chapters.

1.5 Brief Description of the Data

In this section we will give a brief introduction of the data sets that are available, a full review of the data will be given in the next chapter.

The data comes from the following sources: AERS which consists of MEDWATCH reported events in the United States, the General Practice Research Database (GPRD) which consists of events reported via patient visits to the general practitioner, Oceans which is a GlaxoSmithKline (GSK) internal reporting system, the HIF (Health claims by plan members registered with specific HMOs, 17M patients).

The adverse event reporting system collects data through spontaneous reports by users of all drugs. It is particularly well configured to detect rare events, where the background incidence is typically in the range of a few cases per million people. However, it does suffer from problems of under-reporting (van der Heijden et al. (2002)) and the lack of accurate baseline information. It must also be said that reports to AERS may not contain actual diagnoses, but rather the assumptions of the patients themselves, so that we cannot place implicit reliance on them alone. Trontell (2001) gave a comprehensive description of AERS.

The Oceans database has some 4 million case reports. It is of better quality than the AERS database. While in the AERS database reporters are given the option of indicating the drug(s) they think are associated with an event, the Oceans databases require that one drug only is indicated with each event (still by the patient's own assumption, rather than concrete diagnosis). This means that it is much easier to associate drugs with events and leads to considerably simpler analysis.

The GPRD database is the registered patient medical record database. Currently it contains some 4 million patients and some 80 million events. This database is considerably richer than the others in that it does not rely on voluntary reporting of events and therefore does not suffer from many of the problems associated (under-reporting *etc.*). However, the database is not oriented towards the purposes of research and great care must be taken when extracting data from this database. In particular, it can be difficult to associate drugs with events.

1.6 Thesis Outline

We have introduced the reader to the area of Pharmacovigilance and in particular post-marketing surveillance. We have mentioned the objective of the field and at the same time many of the shortcomings. Improvements are certainly needed and we hope to illustrate in the rest of the thesis how data mining techniques can be applied to this end.

The rest of the thesis is organised as follows. Chapter 2 will give an in-depth discussion of the data used in the area. We describe the traditional form of the data used to detect ADEs and we will also describe a more recent development using a new kind of database to the traditional approaches.

In chapter 3 we will describe the various methods used in the area of safety signal detection. We will describe the ratio based methods and then give a comparison of

these methods. We will also apply these methods to a real example and compare the signals generated.

In chapter 4 we introduce a new approach to detecting signals. We discuss the motivation behind the new approach, and the steps needed to arrive at practical conclusions.

Chapter 5 will describe the basis and ideas of the new approach, concentrating on the generation of a distance metric. We describe the ideas of profiling a drug or an ADE such that we can measure the similarities among drugs and ADEs. We will describe the basic ideas of the profiling approach and discuss the various considerations we need to take into account. We will then describe and compare a variety of distance metrics for the purposes of measuring the similarities in our particular type of data.

Chapter 6 will discuss data mining software. These include some major packages which have been maintained over many years and are consequently very mature, some more recent packages as well as some very application specific packages.

In chapter 7 we will discuss new data mining techniques which can be applied in the new approach. Mainly a pattern discovery tool called Peaker is introduced to identify modes in a data set. Peaker has the advantage of being simple to compute, intuitively clear and easy to extend to higher-dimensional space.

In chapter 8 we will utilise all the methods discussed in the previous chapters in order to perform the anomalous signals detection using the new approach. Under the new approach, we will use distance metric obtained using the chemical properties of drugs and analytic properties of drugs and ADEs. We will use Peaker to detect the anomalously high concentrations of side effects and also compare the signals obtained under the new approach to that obtained using the traditional methods.

Finally the summary of the results and options for further research will be discussed in chapter 9.

Chapter 2

Summary of Data

In this chapter we describe the traditional form of the data used to detect ADEs. We will also describe a more recent development using a new kind of database to the traditional approaches. The data is variously known as Pharmacovigilance, post-marketing or safety data. Collection of the data is through consumer reports on post-marketed drugs. As such, there are many problems associated with this form of data collection and we will outline most of the difficulties in the ensuing sections (Lilienfeld (2004)). Many of the major issues in the area of Pharmacovigilance arise as a direct result of the data collection process and the different variants of this process.

2.1 Background

The spontaneous reporting system was introduced in the United States in the early 1950s, expanded in the early 1960s and has remained largely unchanged to the present day. In the UK, a similar scheme (known as the Yellow Card Scheme) was introduced in 1964 in response to the thalidomide tragedy, which collects voluntary reports of suspected ADEs submitted by various sources. This is currently maintained by the MHRA, which replaced the Medical Devices Agency and the MCA

in 2003, and is responsible for monitoring the safety of medicines in the UK. Over 400,000 reports have been submitted since the advent of the Yellow Card Scheme. In many European countries, however, the importance attached to safety monitoring is somewhat neglected and recently attempts have been made to harmonise the reporting methods across Europe (which is one of the problems facing the unification of a single European pharmaceutical market). One advance in this regard is the advent of the Medical Dictionary for Regulatory Activities (MedDRA) which is an international medical terminology now in wide use. We will describe the MedDRA in section 2.2.2.

While the spontaneous reporting system has been undoubtedly useful in the monitoring of suspected ADEs, there have been various shortcomings raised (we will describe these in the succeeding sections). More recently there has been a great deal of interest devoted to an alternative database called the GPRD. The GPRD is currently world's largest database of anonymised longitudinal medical records from primary care and comprises of some 3 million patients. The GPRD is also maintained by the MHRA and offers a genuine alternative to the spontaneous reporting databases (SRDs). It is both richer and more reliable in terms of the data collected, although there are many difficulties in making good use of such data for the purposes of adverse event signal detection. We will describe the issues associated with the GPRD in section 2.4.

2.2 Spontaneous Reporting Database

Spontaneous reports are useful aids for signal generation, but their volume can be overwhelming, and it is important to assess them using a numerical and objective approach. The usual solution is to obtain usage data for the drug in order to assess incidence. Statistical methods have been used to assess reports in this context.

There are problems with this approach in that sales data do not necessarily measure usage, and even with prescription data being available accurately there are a number of biases associated with spontaneous reporting. The rate of reporting is very variable over time, with the highest rates occurring when a drug is newly introduced. Another approach is graphical, where the so-called safety profile is a graph showing the numbers of reactions for a particular drug in medical categories called system organ classes (SOCs). While this is helpful for prioritisation in some instances, it tends to overemphasize the reactions that occur most frequently. It is not itself numeric and does not have an obvious comparison group. The use of proportional reporting ratios (PRRs) can be seen as a numerical version of the safety profile. A similar approach is used in classical epidemiology with death data, where denominators are unknown or uncertain proportional mortality ratios. The basic idea is to use a 2×2 table. The rationale for using this is that the proportion of reactions that are reported on Yellow Cards is relatively constant over time. Since the 1960s when Yellow Cards were introduced, the number of suspected ADEs reported per year has risen, with a peak in 1992. The UK database of these reports is called ADROIT (Adverse Drug On-line Information and Tracking system), and is maintained by the MCA. While the number of reports has changed, the proportion in different system organ classes has remained relatively stable. While the use of a test of statistical significance should not be taken as having an exact meaning in this context, it does aid interpretation in allowing for the likely effect of chance on the PRR. It also helps in the screening process for identifying signals that should be investigated as a priority when there is a small number of reports received. The PRR is of greatest use in monitoring drugs, especially those which have recently been introduced to the market, since safety information on them is inherently limited. The number of statisticians who have applied their creativity and effort to the problems of assessing drug safety is notably fewer than the numbers doing so for studies of

efficacy. Large randomized pragmatic trials will undoubtedly help, but the motives for funding these do not exist in companies, and the regulatory authorities do not have a role as funders of such studies.

2.2.1 Overview

The conventional form of data collection is through that of a spontaneous reporting database, in which data is received in terms of reports filled in by drug users and medical practitioners who suspect an ADE. There are many separate forms of this database across different geographical regions, of which the main ones are:

- The AERS maintained by the FDA in the United States.
- The WHO database which is maintained by the Uppsala Monitoring Center.
- The database on the Yellow Card Scheme maintained by the MHRA.
- Oceans which is an internal database maintained by GSK.

The basic idea is the same in most of these systems, with a few semantic changes. To illustrate, Figure 2.1 shows a MEDWATCH voluntary report form as collected by the AERS database.

The MEDWATCH form allows two suspect drugs (primary and secondary) to be entered. This can lead to quite messy data in terms of drug-event association. The problem is that whilst it is in principle desirable to list a number of the suspected drugs, in practice it is difficult to look for such interactions and often such multiple interactions complicate the analysis. Moreover, a mixture of single and multiple drugs interactions gives very unreliable and confusing baselines. This has led to many practitioners only using the primary suspect drug. The Oceans database only allows one suspect drug to be entered on each report and is therefore a much cleaner

database. On the other hand, having two suspect drugs does permit investigation into drug interactions which constitute a prominent question in Pharmacovigilance.

Each database has its own set of guidelines (for a comprehensive description of the guidelines of the AERS database see Trontell (2001)). Bortnichak et al. (2001) discussed the use of spontaneous reporting databases in safety surveillance.

2.2.2 Coding Dictionary

When using a spontaneous reporting database, the encoding of ADEs is an important component. As already mentioned, the most widely used coding dictionary is the MedDRA although other older dictionaries are still in use which include the WHO Adverse Reaction Terminology and the Read-OXMIS dictionaries (although the Read-OXMIS has now been converted to MedDRA, there are still some terms which have not been mapped). We will center our discussion on the MedDRA given its international acceptance.

The MedDRA includes five levels of hierarchy, which are (from the top down) the System Organ Classes (SOCs), High Level Group Terms (HLGTs), High Level Terms (HLTs), Preferred Terms (PTs) and the Lower Level Terms (LLTs). The following table contains the number of terms at each level, as published in Brown (2002). The

SOC	26
HLGT	333
HLT	1685
PT	14287
LLT	51083

Table 2.1: The MedDRA Hierarchy

MedDRA is multi-axial, which means that child leaf terms may be contained in more than one parent leaf term. This means that a distinct classification of ADE groups according to the higher level class is impossible and as such it is difficult to carry

out hierarchical analysis. When looking for specific ADE signals it is most common to look at only one level of MedDRA and analyse using the terms at this level. Additionally MedDRA is not set up for research purposes but gives more leverage to other considerations such as convenience of terminology for medical practitioners. In this way it is again not suitable for grouping families of ADEs, especially if one expects these families to exhibit some inherent similarities with regards to their drug response.

The most important issue with regards to the area of safety signal detection concerns the level of the dictionary which one should use to mine for associations. Clearly this is a crucial aspect since the generality of ADEs which one uses plays a decisive role in determining the signals which one can detect. The whole issue is as yet unresolved and it is rather a matter of choice as to precisely which level one should use for specific types of signal detection. Most practices tend to use the PT level for any analyses which is widely considered to offer the most flexibility in terms of a balance between being too general and too specific. Moreover, many practitioners attempt to use customised levels derived from the MedDRA using some certain set of criteria. We will see in the remainder of the thesis that in a way this is precisely the problem that we try to overcome, by designing a more flexible method which automatically and adaptively determines the choices that we would otherwise have to make in the above.

2.2.3 Strengths and Weaknesses of Spontaneous Reporting Systems

The spontaneous reporting system can be particularly suitable for the purposes of drug safety monitoring. We will now list some of the strengths below:

- First and foremost the SRS takes into account all patient groups in contrast

to clinical trials. This is one of the main advantages of the SRS set up and most other databases cannot accomplish the same.

- It has relatively low set up and running cost since it relies on voluntary reporting. Compared to a clinical trials study, which uses cohort populations which can be very expensive to set up, the SRS is considered very cheap. Moreover, the set up allows for constant and long term monitoring of all post-marketed drugs.
- The ideal scenario for the SRS is to allow for rapid (in a temporal sense) drug alerts. One would hope that reports would be received at the advent of a new ADE and quickly lead to follow-ups. Moreover, the SRS is unlikely to cause or influence prescribing behaviour (unlike clinical studies) and so it is reasonably safe to assume a constant incidence rate over time. There are some special circumstances for this to be not true but it can be taken as a general rule.

However, there are many difficulties associated with the system. We have already mentioned some of the problems associated with the spontaneous reporting databases in the above sections. Some of these may appear rather intuitive but there are many other problems which are not at all obvious. Note that we are only concerned with the problems arising as a direct result of the data collection process itself and consequently the database system itself. We do not mention the problems of applying specific methods to the data. We will look into these inadequacies of the data in the next section. However, it is worthwhile to keep in mind that many of the method specific problems can be traced to the issues mentioned in this section.

We will now list and explain in more detail the various difficulties which can arise when using the spontaneous reporting databases:

- There is a general lack of a usable baseline (number of patients exposed to a drug) for the reports collected and this severely hampers the signal detection

process. In simple terms there is generally no sensible estimate of the number of prescriptions which underlie the numbers of reports received. Clearly, without such baselines, it is difficult to precisely quantify the severity of any ADRs.

- As we can see from the MEDWATCH report form, some reports are only based on patients' suspicions. This clearly leads to inaccurate reports and much follow-up work has to be done to verify the claims. For rare events in particular, where there may only be a few reports collected, it is difficult to be sure of the validity of each report. Of course in the majority of cases the reports come as a result of medical diagnosis, but the proportion of non-diagnosis is by no means insignificant. With most practices, the attitude is to take into account every report if it contains a rare and severe event. As such, a rather arduous process is required to back up the initial data collection.
- Duplicate reports can and often do arise. The most common occurrence of this problem is attributed to the same ADE being reported by multiple sources (the patients, doctors and pharmacists), although other reasons (multiple reports by the same source) are also possible but generally much more infrequent. For rare or severe events this can be a major problem since only a few reports are required to constitute a signal. Of course, as mentioned in the previous point, a validation process can pick up such inconsistencies but any such false signals cause waste of valuable resources.
- Because of the reporting nature of the data, it is often the case that external forces (such as media coverage) can cause a dramatic surge in the number of reports filed. Often this can cause enormously disproportionate levels of the signals detected. As yet this is a rather unexplored area of research and the precise extent of the differences this can cause is difficult to quantify.

Precise temporal aspects of the data are also rather difficult to obtain in some databases and this further complicates the analysis.

- The problem of under-reporting is another concern. Whilst erroneous reports need careful validation, it is much more common for patients to simply not report anything at all. The general attitude tends to be that this is not as serious as receiving a false report. But clearly this leads to rare events being undetected. Research into this area is as yet somewhat undeveloped and the precise effect of under-reporting unexplored. van der Heijden et al. (2002) gave a quantitative account of the under-reporting problem.
- We also need to consider the level of the coding dictionary to use such that it best reflects on the detection of rare events. Currently the PT level is the most widely used in Pharmacovigilance, although this is considered by some to be too fragmentary whilst the HLT level is too general. Brown (2002) gave an in-depth account of the effect of coding dictionary on the signal generation process.
- Practices are different across geographical regions and institutions. This can lead to very different signals as varying background rates arise. The full effect of this on the signal detection process is as yet unclear. It is also the case that different practices operate under different guidelines (for instance, this can be in terms of methods employed, thresholds or data preparation) and signals detected under one practice may well be lost under another. Some effort has been devoted in achieving a more unified approach to signal detection across these different practices and in researching into the differences involved but this is as yet a rather grey area.

2.3 Processing the Raw Data

The raw data itself would initially be in report form. This is subsequently processed into a tabular form which comprises of each report crossed by a variety of attributes (most of these attributes are patient and report information such as sex, gender and time of report).

To perform signal detection, the traditional type of analysis (we will describe a review of these methods in the next chapter) is to look for abnormally high counts for combinations of drug and ADE. This can be done in a number of ways but by far the most common is to arrange the data in a large categorical contingency table consisting of drugs against ADEs. Each cell of the contingency table would then represent the number of times (number of reports) that an ADE is associated with a drug. As mentioned earlier, in certain SRSs, the associations are clearly defined in that only one suspected ADE is permitted to be entered. In these instances, there is a clear definition of a single count (single association) for that drug-ADE cell and the task of setting up the contingency table is clearly straightforward. In other SRSs, multiple suspected ADEs and drugs can be reported and here the task of looking for associations is more complicated. In the remainder of this section we will describe how to transform the reports of multiple associations to that of a categorical contingency table.

In the instance when there exist multiple associations in the reports, one has to count all the instances of a particular ADE being associated with a particular drug. To do a brute search for such associations would be computationally quite expensive, especially if one is also interested in higher-order associations.

The general problem comes under the heading of Association Analysis. The initial application was that of the market basket problem, whereby one looks for combinations of items that tend to be bought together in large supermarket databases.

There has been a great deal of effort made to invent fast algorithms to tackle this problem. The *apriori* algorithm is one example of a fast algorithm developed to handle large datasets (for literature on association analysis see Agrawal et al. (1993), Agrawal & Sikrant (1994) and Hipp et al. (2000)).

2.4 The GPRD Database

2.4.1 Overview

The GPRD currently contains some 3 million active patients and is the largest database of its kind in the world. In particular it contains the prescription and event history of each patient and thus is a longitudinal database as opposed to the spontaneous reporting databases. The GPRD contains much richer information than the spontaneous reporting databases. It contains the exact baseline for each drug which is an advantage over all traditional spontaneous reporting databases. Moreover, the information is collected through medical practitioners which makes the GPRD more reliable. Other information is also available through the GPRD which ranges from the basic (such as gender, age, registration detail and other patient specific data) to the very detailed (such as free-text which allows the medical practitioner to input manual text).

Undoubtedly there is potential in the GPRD for it to be usefully employed in the context of mining for safety data. Currently the GPRD is mainly used for population-based cohort studies of various drugs. Its use as a Pharmacovigilant database is yet to be clearly defined. This is partly due to the difficulties that one can encounter when mining the GPRD in this way (we will describe this in the next section).

2.4.2 Strengths and Weaknesses Associated with the GPRD

In this section we will discuss the main strengths and weaknesses of the GPRD, as compared to the SRSs. Wong (1999) and Waller et al. (1996) gave introductions to the GPRD. We will list some of the strengths below:

- The database holds exact baseline for each prescribed drug as well as the exact number of patients experiencing an ADE. This by itself is already enough to justify the use of the GPRD. Moreover, the rate of associations is much higher since we no longer rely on voluntary reporting which, as mentioned, can be very low.
- Potentially it is possible to obtain quicker and more up to date reports, since it is likely that healthcare professionals will report quickly given their improved awareness of drug safety.
- Other information is also available to supplement simple associations. Some of this information can be quite useful, such as dosage levels and time of prescriptions since these can be used as stratifying variables. Moreover, the GPRD contains all medical records over time and so can be used to study various factors (examples are multiple associations, the effects of external forces and changing levels of prescriptions).

However, the GPRD has a new set of problems associated with it. Clearly the way GPRD is set up makes it best suited to cohort studies and to use it for the purposes of post-marketing surveillance requires some careful thought. Below we will list some of the difficulties associated with GPRD:

- The initial difficulty lies in how to associate drugs with causing particular events. Unlike the SRSs, there are no suspected drug-ADE associations so that any associations are speculative at best. Given that we have an enormous

amount of data for each patient, this is by no means an easy task. There are also various other considerations to take into account with regards to associations. We will list some of these below:

- Intervals of association - Even if a patient is taking only one drug for a sustained period of time, there is still the question of how much of the time period we need to take into account when looking for the events caused.
 - Switching between drugs - Some patients switch between drugs and it is difficult to have a distinctive cut-off point for the switch. In particular, the problem lies in when one decides that a patient is clear of the previous drug and onto the next. This involves obtaining knowledge with regards to the incubation periods of various ADEs and this is clearly not at all an easy task and is in practice impossible to implement.
 - Discontinuous prescription - Truncated courses of treatment occur regularly through each patient history. A central issue for such information is whether it is advisable to mine events in all these disparate time intervals or to take one time interval only. Whilst it may seem desirable to take into account all the history of a patient, it is highly likely to give rise to duplicate events.
- The GPRD applies only to primary care patients and does not apply to hospital care. There can often be poor communication between primary care records and hospital records and certain information can be lost. Hospital drugs or treatments are also not available so that not all products are encompassed.
 - There are various amounts of prescription and dosage levels (as well as the inherent free text contained in GPRD, which is not shown here). This is

obviously important in determining the length of time a patient is taking a drug. We can also see the disparity of the information among records. Therefore we need some way of mining for the text information in these records.

2.4.3 Mining the GPRD Database for the COX-II Data

It is possible to mine for very pure drug event associations so as to disregard certain suspected drug interactions, take only continuous courses of prescription and to ignore events which have already occurred in the patients' previous histories. In this section we will outline an attempt to mine the GPRD database. The data used is the records of 1,000,000 patients who have been prescribed the COX-II drug. First we define a continuous course of treatment as one for which the patient is taking one drug (one variant of COX-II). The time at which each prescription expires before the next treatment is calculated as $\frac{\text{prescription period}}{\text{dosage}}$. Additionally we set a 28 day tolerance period (the majority of the prescription periods being 28 days). So we say that a patient is on a continuous treatment if there are no breaks in the treatment such that the time between two prescriptions exceeds $28 + \frac{\text{prescription}}{\text{dosage}}$ days. The level of dosage is obtained from the Dosage column.

For each patient, we take the largest contiguous block of treatment and ignore others. This allows us to leave out carry-over effects whereby the events occurring for one treatment can be attributed to the previous treatment. Additionally we will ignore switching for the time being and investigate pure treatments containing one drug only.

Recent research has revealed a connection between COX-II drugs and cardiac events and these constitute the major aspects of interest. In this case, there is a marked difference in terms of taking into account patients that are on a prescribed course of Aspirin since Aspirin is known to suppress cardiac events. We will therefore stratify patients who are on a consistent course of Aspirin and those who are not.

Additional factors to take into consideration are stratification according to dosage levels of COX-II drugs. This is somewhat complicated by the different levels of dosage throughout one contiguous course of treatment. We will therefore take the mean dosage level in any one treatment. There are three main levels of dosage around approximately 25, 100 and 200 milligrams so we will stratify with these three levels and discard other levels.

All this leaves approximately 55% of patients from the initial 100,000 who qualify for with our criteria of taking COX-II drugs. We can then mine for the ADEs associated. In this particular instance, it will be interesting to initially restrict the ADEs to cardiac events since these are the events of interest. We can also add a temporal element to these events and mine for ADEs at 2-month intervals. This can reveal interesting tendencies in the ADEs since it is known that some associated events have particular incubation periods.

2.5 Data Issues

As described in section 1.5, the data sets involved in this project are typically large, involving contingency tables with millions of cells. This presents many problems of its own, such as follows:

1. Computational efficiency - This is particularly relevant and we have to decide carefully which techniques are computationally viable. In particular, a large amount of programming is required as most existing software packages are not well equipped to deal with such vast data sets. Such programming must always be carried out with efficiency in mind.

2. Dimensionality - As the size of the dataset increases, so typically do the dimensions used to represent the data. The curse of dimensionality is well known and this can cause very different interpretations of the data in terms of distance

between points. Points are further away from each other in higher dimensional space.

3. Visual exploration - Large data sets can be difficult or impossible to inspect. Many simple plots which can be extremely useful in giving an initial impression of the data may simply not be adequate. It is also difficult to pick out individual points among a mass of other points.

4. Data formatting and checking - Putting the data into the right format for analysis can be time consuming and may require some careful thought. Moreover, any corrupted data is hard to pick up and in many cases we may have an idea only after an analysis has been carried out.

We need to give careful thought to all these issues before carrying out each analysis.

Chapter 3

Review of Methods

In this chapter we will describe the various methods used in the area of safety signal detection. Since the advent of the SRS, there have been some advances in methodology. However, many of the techniques still revolve around making use of ratio based measures to give meaning to the degree of severity of each drug-ADE combination. In this, there has been no significant improvement over the years. Moreover, there is still a sense that most methods cannot be relied upon too heavily and stringent follow up work is needed in the case of each signal.

For the rest of this chapter, we will first describe the ratio based methods and then give a comparison of these methods. We will also apply these methods to a real example and compare the signals generated.

3.1 Notation

We first define some notations¹, which we will use for the remainder of the chapter.

Assume a two-way contingency table of r rows and c columns:

- n_{ij} - count for cell (i, j) , $i = 1, \dots, r$, $j = 1, \dots, c$.

¹For a thorough discussion of some methods mentioned in this chapter, see Fedorov et al. (2003).

- $n_{i\cdot}$ - marginal count for row i .
- $n_{\cdot j}$ - marginal count for column j .
- $n_{\cdot\cdot}$ - total count in the table.
- $\pi_{ij} = \frac{n_{ij}}{n_{\cdot\cdot}}, \pi_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}}, \pi_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}$.
- $p_{ij}, p_{i\cdot}$ and $p_{\cdot j}$ are the true probabilities for cell count, row marginal count and column marginal count respectively.

3.2 Measures of Association

As mentioned, most of the analyses in safety signal detection are based on ratio measures. This stems directly from the traditional association measures which are used in many areas, in particular (as an area closely linked to our problem) epidemiology and more generally traditional contingency table analysis. Classically, measures are derived which in some way estimate the ratio between the actual and expected number of incidences based on some initial assumptions about the data. In the following sections we will list some of the most widely used ratio measures.

3.2.1 Discrepancy Measures

In traditional contingency table analysis, most of the interest centers on testing the hypotheses of independence within the table and homogeneity between rows and columns. These can be presented in the following null hypotheses:

1. Independence

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \tag{3.1}$$

2. Homogeneity

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad (3.2)$$

These tests are equivalent (assuming the absence of “zero probabilities”), therefore we will only consider the test for independence.

For the test of independence, the well-known χ^2 test gives

$$X^2 = \frac{\sum_i \sum_j \frac{(\pi_{ij} - \pi_{i\cdot}\pi_{\cdot j})^2}{\pi_{i\cdot}\pi_{\cdot j}}}{n_{..}}, \quad (3.3)$$

and the Kullback Liebler divergence gives

$$KL = \frac{\sum_i \sum_j \pi_{ij} \ln \frac{\pi_{ij}}{\pi_{i\cdot}\pi_{\cdot j}}}{n_{..}}, \quad (3.4)$$

with X^2 and $2KL$ following a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

In the case that we are only interested in the contribution of particular cells to the totals (3.3) and (3.4), we can define each cell contribution for the two tests respectively as

$$\theta_{ij}^2 = \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}},$$

and

$$KL_{ij} = p_{ij} \ln \frac{p_{ij}}{p_{i\cdot}p_{\cdot j}},$$

which can be estimated by replacing p_{ij} , $p_{i\cdot}$ and $p_{\cdot j}$ by their maximum likelihood estimators π_{ij} , $\pi_{i\cdot}$ and $\pi_{\cdot j}$ to give

$$\hat{\theta}_{ij}^2 = \frac{(\pi_{ij} - \pi_{i\cdot}\pi_{\cdot j})^2}{\pi_{i\cdot}\pi_{\cdot j}},$$

and

$$\hat{KL}_{ij} = \pi_{ij} \ln \frac{\pi_{ij}}{\pi_{i\cdot}\pi_{\cdot j}}.$$

The above can be used to indicate the measures of association between a particular row and column. In the case of adverse event detection this is naturally applicable since we are only interested in each cell and not the overall table.

3.2.2 Ratio Measures

We can adjust the discrepancy measures of the last section to suit the purposes of adverse event detection. Many measures of association have been proposed and used, leading essentially to a similar premise². The odds ratio is a commonly used measure of association (Agresti (1990)). The OR is easy to interpret and represents the odds of an event occurring in group 1 to that of group 2.

For a drug-ADE combination, if we have the following table, then the OR takes

	Drug present	Drug not present
ADE present	a	b
ADE not present	c	d

Table 3.1: Table of Presence of Drug and ADE

²For literature in this context see van Puijenbroek et al. (2002).

the form

$$OR = \frac{ad}{bc}. \quad (3.5)$$

Other ratio measures commonly in use are the Reporting Ratio (RR) in DuMouchel (1999) and the PRR in Evans (2002), defined by

$$RR_{ij} = \frac{\pi_{ij}}{\pi_{i.}\pi_{.j}}, \quad (3.6)$$

and

$$PRR_{ij} = \frac{\pi_{ij}(1 - \pi_{.j})}{\pi_{.j}(\pi_{i.} - \pi_{ij})}. \quad (3.7)$$

All these ratio measures reduce to essentially the same form when the table is large. The generic form is the ratio of the observed counts to the expected counts. Of course, there is a marked difference when the table is sparse, since it can often be the case that one cell can predominate in an entire row or column.

3.3 Bayesian Approach

A few of the more recent publications (DuMouchel (1999) and Orre et al. (2000)) in adverse event signal detection have used Bayesian ideas to estimate p_{ij} , $p_{i.}$ and $p_{.j}$. In particular, Empirical Bayes estimation is used to produce smoothed estimates. In this section we will give a general overview of Empirical Bayes and Shrinkage Estimators methodology as well as some of the methods applied to adverse event signal detection. In particular, there are generally two recognised models currently being applied to safety signal detection. The first is the Gamma-Poisson model (or a similar model called the Gamma-Poisson Shrinker (GPS) given in DuMouchel

(1999)). The GPS is widely used and currently is applied to the AERS database. The second is the Beta-Binomial Model (or a similar model called the Bayesian Confidence Propagation Neural Network (BCPNN) given in literature including Bate et al. (1998) and Bate et al. (2002)). The WHO database extensively uses the BCPNN.

3.3.1 Empirical Bayes and Shrinkage Estimators

There is a large body of literature on the subjects of Empirical Bayes and Shrinkage Estimators, which are closely linked. Indeed, one common and widely used method of obtaining a Shrinkage Estimator is through the use of Empirical Bayes methodology.

The most famous example of a Shrinkage Estimator is the James-Stein Estimator (JSE). James & Stein (1961) demonstrated that the Maximum Likelihood Estimator (MLE) for the means of a multivariate normal distribution is inadmissible (*i.e.*, it is possible to construct an estimator with smaller risk). The point of Shrinkage Estimators is that they incorporate information from many different entities to describe a single entity, thereby having smaller average risk than the MLE. Shrinkage Estimators, in particular the JSE, are discussed in Efron & Morris (1973), Gruber (1998) and Morris (1983).

The Empirical Bayes approach has largely been stimulated by Robbins. It involves both Bayesian and Frequentist approaches to statistics. One specifies a prior parametric form for the parameters of interest and proceeds to estimate the prior parameters through the data via the marginal distribution. Empirical Bayes Estimators are known to have asymptotically optimal properties (for a discussion see Deely & Lindley (1981) and Morris (1983)).

In our context, the Empirical Bayes approach smoothes the values of the association measures to allow more reliable estimates for smaller sample sizes. The

Empirical Bayes process is repeated separately for different measures in the cases of higher order associations.

3.3.2 Gamma Poisson Model

In this section we will describe the Gamma Poisson model used in adverse event signal detection. The premise of the model is that data is assumed to follow a Poisson distribution (Po) with a prior which follows a Gamma distribution (Ga).

Assume $n_{ij} \sim Po(\mu_{ij})$ and $\mu_{ij} \sim Ga(\alpha_{rc}, \beta_{rc})$. In the posterior $\mu_{ij} \sim Ga(\alpha_{rc} + n_{ij}, \beta_{rc} + 1)$.

The marginal distribution of n_{ij} follows a negative binomial distribution of the form

$$P(n_{ij}|\alpha_{rc}, \beta_{rc}) = \frac{\Gamma(n_{ij} + \alpha_{rc})}{\Gamma(\alpha_{rc})n_{ij}!} \left(\frac{1}{1 + \beta_{rc}}\right)^{n_{ij}} \left(\frac{\beta_{rc}}{1 + \beta_{rc}}\right)^{\alpha_{rc}}. \quad (3.8)$$

To estimate α_{rc} and β_{rc} , one can use the method of maximum likelihood, which would require the estimation of each parameter by some numerical approximation method. It is far more practical to use the method of moments (MM) estimators³. The expectation and variance of n_{ij} are

$$E(n_{ij}) = \frac{\alpha_{rc}\beta_{rc}}{1 + \beta_{rc}}, \quad (3.9)$$

and

$$V(n_{ij}) = \frac{\alpha_{rc}(1 + \beta_{rc})}{\beta_{rc}^2}. \quad (3.10)$$

Substituting the sample mean and variance, $\bar{n}_{ij}$ and s^2 , for $E(n_{ij})$ and $V(n_{ij})$ gives the MM estimates.

³In the sense of computational efficiency, although in some cases it is viable to use the MLE.

The posterior estimate (or Empirical Bayes estimate) for p_{ij} is then given by

$$\hat{p}_{ij} = \frac{n_{ij} + \hat{\alpha}_{rc}}{n_{..}(1 + \hat{\beta}_{rc})}. \quad (3.11)$$

Similarly estimates for $p_{i.}$ and $p_{.j}$ can be obtained.

3.3.3 DuMouchel's Gamma Poisson Shrinker

The Gamma-Poisson model used in the last section is the basic form of such a model. DuMouchel (1999) used an alternative formulation of the Gamma Poisson Model, termed the GPS. We will now describe this alternative model.

Assume $n_{ij} \sim Po(\mu_{ij})$ and $\frac{\mu_{ij}}{E_{ij}} \sim Ga(\alpha_{ij}, \beta_{ij})$. DuMouchel actually used a mixture of two Gamma distributions for the prior, but we will just use a single Gamma distribution to reduce notation.

Again we can calculate the estimates $\hat{\alpha}_{ij}$ and $\hat{\beta}_{ij}$ through the marginal distribution, which is a negative-Binomial distribution.

The posterior expectations of λ_{ij} and $\log(\lambda_{ij})$ are given by

$$E(\lambda_{ij}|N_{ij} = n_{ij}) = Q_{n_{ij}} \frac{\alpha_1 + n_{ij}}{\beta_1 + E_{ij}} + (1 - Q_{n_{ij}}) \frac{\alpha_2 + n_{ij}}{\beta_2 + E_{ij}} \quad (3.12)$$

and

$$E(\log(\lambda_{ij})|N_{ij} = n_{ij}) = Q_{n_{ij}} \psi(\alpha_1 + n_{ij}) \log(\beta_1 + E_{ij}) + (1 - Q_{n_{ij}}) \psi(\alpha_2 + n_{ij}) \log(\beta_2 + E_{ij}) \quad (3.13)$$

where ψ is the Digamma function.

DuMouchel advocated the use of $E(\log_2(\lambda_{ij})|N_{ij})$ and $2^{E(\log_2(\lambda_{ij})|N_{ij})}$ as ranking measures for signals generated by each drug-ADE combination, which he termed $EBlog2_{ij}$ and $EBGM_{ij}$ respectively.

There is also an extension to the GPS, termed Multiple Gamma-Poisson Shrinker,

which is used to find signals in sets of size 3 or larger. The details can be found in DuMouchel & Pregibon (2001).

The GPS has an inconsistency, which is that although the data are modelled by a Poisson distribution, it assumes fixed marginal totals in the prior⁴.

3.3.4 Beta Binomial Model and the Bayesian Confidence Propagation Neural Network

The Beta-Binomial model assumes that $n_{ij} \sim Bn(n_{..}, p_{ij})$, $n_{i.} \sim Bn(n_{..}, p_{i.})$, $n_{.j} \sim Bn(n_{..}, p_{.j})$ and $p_{ij} \sim Be(\alpha_{ij}, \beta_{ij})$, $p_{i.} \sim Be(\alpha_{i.}, \beta_{i.})$, $p_{.j} \sim Be(\alpha_{.j}, \beta_{.j})$, where Bn denotes the binomial distribution and Be denotes the Beta distribution.

Orre et al. (2000)⁵ suggested that $p_{i.}$ and $p_{.j}$ both follow uniform distributions over the unit interval, and to set α_{ij} and β_{ij} so that

$$\alpha_{i.} = \beta_{i.} = \alpha_{.j} = \beta_{.j} = \alpha_{ij} = 1, \beta_{ij} = \frac{1}{\hat{p}_{i.}\hat{p}_{.j}}. \quad (3.14)$$

Alternatively we can again estimate the prior parameters via Empirical Bayes. To do this, we assume that $\alpha_{i.} = \alpha_r, \beta_{i.} = \beta_r, \alpha_{.j} = \alpha_c, \beta_{.j} = \beta_c, \alpha_{ij} = \alpha_{rc}, \beta_{ij} = \beta_{rc}$. The marginal distribution for n_{ij} is given by

$$P(n_{ij}|\alpha_{rc}, \beta_{rc}) = \binom{n_{ij}}{n_{..}} \frac{B(n_{ij} + \alpha_{rc}, n_{..} - n_{ij} + \beta_{rc})}{B(\alpha_{rc}, \beta_{rc})}, \quad (3.15)$$

where $B(\)$ is the beta function.

We can use the MLE or the MM estimator. Again for practical purposes the

⁴A fixed marginal total in this case points to the fact that the data are distributed as a Multinomial distribution, which is inconsistent with the Poisson assumption.

⁵The BCPNN described in Orre et al. (2000) uses a slightly different formulation of the Beta Binomial Model. This difference is similar to that between the Gamma Poisson Model and the GPS.

MM estimator is preferred. The expectation and variance of 3.15 are given by

$$E(n_{ij}) = n_{..} \frac{\alpha_{rc}}{\alpha_{rc} + \beta_{rc}}, \quad (3.16)$$

and

$$V(n_{ij}) = \frac{\alpha_{rc}\beta_{rc}}{(\alpha_{rc} + \beta_{rc})^2} \left(1 + \frac{n_{..} - 1}{\alpha_{rc} + \beta_{rc} + 1}\right). \quad (3.17)$$

Substituting the sample mean and variance, $\bar{n}_{ij}$ and s^2 , for $E(n_{ij})$ and $V(n_{ij})$ gives the MM estimates for α_{rc} and β_{rc} . Similarly we can obtain estimates for the other prior parameters $\alpha_r, \beta_r, \alpha_c, \beta_c$.

The posterior estimate (or Empirical Bayes estimate) for p_{ij} is then given by

$$\hat{p}_{ij} = \frac{n_{ij} + \hat{\alpha}_{rc}}{n_{..} + \hat{\alpha}_{rc} + \hat{\beta}_{rc}}. \quad (3.18)$$

3.3.5 Bayesian Nonparametrics

We can use a Bayesian Nonparametric approach to modelling our data. We implement a Gibbs sampler in a mixture of Dirichlet process (MDP) model (for a complete description of the general sampling scheme see MacEachern & Muller (1998) and MacEachern & Muller (2000)). Escobar & West (1995) gave a mixture of normals formulation while Quintana (2002) presented the approach in a Multinomial context. Here we will give the basic formulation of the MDP model in the context of a Poisson likelihood, according to the general scheme given in MacEachern & Muller (2000).

The basic MDP approach is as follow: The data has likelihood $y_i \sim Po(\theta_i)$ with prior $\theta_i \sim G$, where G is a random distribution generated by a Dirichlet process with base measure MG_0 and total mass parameter M such that $G \sim DP(G_0, M)$. Here

we choose the conjugate base measure $G_0 = \text{Gamma}(\alpha, \beta)$ for ease of computation.

We will now define notations:

Let $\theta = \{\theta_1, \dots, \theta_n\}$ be the vector of likelihood parameters.

Let $\phi = \{\phi_1, \dots, \phi_k\}$ denote the set of distinct θ_i 's.

Let $s = \{s_1, \dots, s_n\}$ denote the vector of configuration indicators defined by $s_i = j$ iff $\theta_i = \phi_j$, $i = 1, \dots, n$.

Let $n = \{n_1, \dots, n_k\}$ denote the vector where n_j is the number of $s_i = j$. Gibbs sampling is implemented by iterative sampling from the full conditionals, and the full sampling scheme is as follow:

- To resample (θ_i, s_i) , we assign $s_i = j$ for existing cluster j with probability $n_j p(y_i)$, where n_j denotes the size of cluster j with θ_i removed from consideration and $p(y_i)$ is the posterior predictive distribution for y_i , again with θ_i removed. We assign $s_i = k + 1$ for new cluster $k + 1$ with probability $M \int p_\theta(y_i) G_0(\theta) d\theta$.

In our Poisson-Gamma context, $p(y_i) \sim NB(\alpha, \beta) = \frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha)\Gamma(y_i+1)} (\frac{\beta}{\beta+1})^\alpha (\frac{1}{\beta+1})^{y_i}$.

- To resample ϕ_j , we simply sample from the posterior for ϕ_j over all data points with $s_i = j$.

In our present context, this amounts to sampling ϕ_j from $\phi_j \sim \text{Gamma}(\alpha + n_j, \beta + \sum y_i)$ where the sum is over all y_i such that $s_i = j$.

- To resample M , we need to define an auxiliary function on M with an additional latent variable x such that $p(x|k, M) = Be(M + 1, n)$.
- To resample all other hyperparameters, we simply use the standard Bayes formulation and resample from the relevant conjugate posterior.

The MDP model is a popular one in the Bayesian Nonparametrics literature. Bayesian Nonparametrics has the advantage of not requiring a specific model formu-

lation. We can think of this as modelling the data as a countably infinite component mixture, as opposed to some of the previous methods in Pharmacovigilance which uses mixtures of two component priors. Clearly we obtain a much richer model in this case. There are also various other aspects to this model. We already have clustering in the data and we can devise a test for independence in the contingency table and we can test for homogeneity between rows and columns.

We are left with the task of specifying the parameters in the Dirichlet distribution prior. The simplest way is to specify a reasonably flat prior. We use the following scheme (for drug profiles): Calculate the mean of each ADE across all drugs and set the prior Dirichlet parameter α_i for drug i as

$$\alpha_i = \lambda_i \mu_i + \delta_i,$$

where μ_i is the mean for each ADE and suitable values of λ_i and δ_i can be chosen. Initially we will simply use the same value for λ_i and δ_i for all ADEs. We can try a range of values in an empirical study.

An obvious advantage of this approach over the DuMouchel approach is that we can now set prior parameters according to the importance we attach to a particular ADE, by changing values of λ_i and δ_i (small values giving more emphasis to the data). This information can be easily obtained from lists giving the severity of each ADE.

3.4 Application

We apply the various methods to a real data set consisting of 172 drugs and 2245 ADEs. Note that here the data is selected in such a way because of availability. In ideal circumstances it would be more representative to choose a data-set which has more balanced numbers of drugs and ADEs. We apply each of the methods (the OR,

the Gamma-Poisson model, the GPS, the Beta-Binomial model and the BCPNN) to the data. For each method we calculate the list of signals which comprise the drug-ADE combinations that return the highest numbers in terms of the signals generated. The top 40 signals for each method are generated and shown in Table A.1 to A.5 in Appendix A. As we would expect, the high variation of the OR means that the highest signals all comprise of low counts (typically of the order 1 or 2). Moreover the high variation is further confirmed by the excessively high numbers that make up for the signals. Perhaps surprisingly, the Gamma-Poisson model and the Beta-Binomial model both return as unstable signals as the OR. A closer inspection reveals that these signals generated by the Gamma-Poisson model and the Beta-Binomial model are in fact nearly identical to the ones generated by the OR (the actual drug-ADE combinations are the same). This means that there has been very little smoothing applied. However, if we examine the way by which the smoothing is actually performed in both of these methods, we can easily see that this is to be expected. We can examine the difference when it comes to the GPS and the BCPNN, in both of which the basic model is modified to allow smoothing to be applied in different ways. Note that for the BCPNN we use the uniform prior in this case. We can see from both the counts of the combinations and the signals themselves that these methods produce much more stable signals. In particular we notice that the counts almost entirely consist of higher numbers, which incidentally are much higher for the BCPNN than the GPS (this is partly because we are using the uniform prior for the BCPNN). From the formulation of the GPS and the BCPNN the results are also to be expected since the smoothing is applied to the baseline in both cases which would significantly reduce the problem with excessively low baseline value faced by the OR, Gamma-Poisson model and the Beta-Binomial model. On the other hand, we must consider that it is not necessarily better in all cases to apply smoothing to such an extent, since we are now faced with the

alternative problem that all the low counts are being smoothed out and the signals would consist entirely of higher numbered counts. We will discuss this in more depth in the next section.

3.5 Discussion

In this section we discuss the advantages and disadvantages of the methods described above, as well as a general summary of the whole approach. There is some literature which gives comparisons of the different methods, such as van Puijenbroek et al. (2002), Evans (2002), Meza (2003) and Fedorov et al. (2003).

The odds ratio clearly suffers from the problem of sampling variation at low counts⁶. In many cases this leads to unstable baseline information, resulting in the fact that the highest odds ratios tend to be very small counts (often ones and twos). This is clearly unsatisfactory. Attempts to remedy this problem have been to either ignore the low counts or to use the Empirical Bayes based smoothing methods⁷. Ignoring the low counts has a major drawback, which is that the unknown rare ADEs tend to be precisely these low counts. Of the Empirical Bayes methods, the Gamma-Poisson model and the Beta-Binomial model with Empirical Bayes prior do not apply very much smoothing and are very close to the odds ratio in practice while the GPS and the Beta-Binomial model with uniform prior apply a lot more smoothing. However, what can we make of this smoothing? First of all, it has to be said that at large counts the smoothed estimators approach the odds ratio, so it is really the small counts which are affected. As mentioned, this is convenient since these tend to be what we are interested in anyway. The smoothing is essentially

⁶DuMouchel (1999) accounted that “Measures based on raw ratios of observed to expected seem too ready to select cells with very small counts”.

⁷DuMouchel (1999) accounted that “One way to reduce the sampling variation in RR would be to require a minimum value of N , such as at least 5 observations in a cell. But one needs a rationale for choosing the cutoff value”.

controlled by a few prior parameters. While this works well for some applications, in some areas of contingency table analysis, spatial analysis (Cressie (1993)) and in particular spatial disease mapping (Group (2000), Marshall (1991*b*), Manton et al. (1989), Marshall (1991*a*) and Hills & Alexander (1989)), these tend to involve much smaller datasets where one would be interested in better estimators or in tests for the entire data set such as the test for independence in the entire table. Empirical Bayes estimators would be suitable since, as mentioned, they have smaller overall risks than MLEs. However, in the present context, this is not particularly useful when producing single cell estimates in very large contingency tables.

In practice (as shown in section 3.4), there is very little difference among all these methods and very careful follow-up work is required to validate the signals. van Puijenbroek et al. (2002) accounted that “The heterogeneity of the data collected in databases of spontaneous reporting systems and the variety of biases influencing data (such as under-reporting) are likely to have more influence on the potential for signal detection than the small behavioural differences among the measures detected in this study”. Because of this, one of the main priorities of any signal detection process is to reduce the number of false positive signals. There is an inherent danger in producing single cell estimates, which is that the data is not particularly reliable, resulting inevitably in many false positive signals. We can thus venture an opinion that the deficiencies of the present methods arise not so much due to the methods themselves, but more as a result of the general approach. Along the same lines we can say that significant improvement will not be obtained by refinement of the methods described in this chapter, but rather from alternative approaches which we will subsequently describe.

We now take a moment to discuss the issue of validation. There are various ways to perform validation, the most common being to check the resulting signals with pharmaceutical experts. However, in the course of this thesis it has proven

difficult to obtain such validation, mainly due to the fact that it is often sensitive to obtain information pertaining to the side effects of certain drugs. In an alternative approach, one can ask for a data set with known signals to be able to check the efficacy of detection. This involves some significant resources, however, and have been difficult to obtain.

3.6 Other Issues

Many problems with the current methodology have been raised in certain sources, although the literature is very restricted with regard to the evaluation of the methods and suggestions for possible improvements. Most of the points being made seem to be concerned with the fact that there is very little knowledge of the performance assessment and comparison of the current methods. In particular, there is very little known about the behaviour of these methods when data issues complicate matters. We list some of the suggested issues below:

- There is no established literature about the false signal generation rate (as mentioned, this poses a major drain on resources) of each method. The performance of either of the methods described in section 3.3 in terms of generating false positive signals is not known. It is suspected and many practices feel that both methods do tend to generate a large number of false positives. One of the difficulties in this aspect is the precise meaning of a false positive signal. In real world data sets it is often the case that very severe ADEs need to be considered regardless of the signals generated and in this case there is no such thing as a false positive signal. That said, however, there is a genuine need to carry out well designed simulations to study the effects of false signal generation. The analyses currently being performed on real data sets can only be verified through previously analysed (independently) ADEs.

- There is no real effort made to assess the effects of other factors and variables. Although stratification is possible in these methods, it is currently not sufficient to account for each patient profile. Additionally the effects of higher order interactions are not well catered for and little is understood about this case. Whilst higher order associations (in particular drug-drug interactions) can be analysed using the current methods, the analyses are rather rudimentary and at best extend to higher order counts. More needs to be studied in terms of the functionality of these methods for the aforementioned scenarios.

Chapter 4

Alternative Approaches

As we have discussed in section 3.5, there are many problems associated with the existing methods of signal detection in Pharmacovigilance. In this chapter we introduce a new approach to detecting signals. We discuss the motivation behind the new approach, and the steps needed to arrive at practical conclusions.

4.1 Motivation

Underlying the new approach is the need to give the data some structure. At the outset, we have a categorical contingency table of raw counts. The present methods all involve detecting anomalously high counts within the raw table. However, given that this is a table of drugs against ADEs, it would be natural to ask whether some of the ADEs and drugs are closely related to each other. The obvious way to obtain such information would be through expert medical knowledge. However, this is by no means an easy task when one has tens of thousands of such drugs and ADEs. Such a task would comprise a large project in itself and may take years for its completion. The standard medical dictionaries only class ADEs and drugs according to the convenience of classifying terms and not according to their characteristics. This is not particularly useful for the purposes of research. Given

the lack of such an appropriate dictionary, we switch focus to an analytical approach to derive associations.

4.2 Improvements

How can we improve on the past methods if we have sets of strongly associated drugs and ADEs? Firstly, it must be noted that it is interesting in itself to have a classification of drugs and ADEs, and to compare the classes derived analytically to those from an external source (expert knowledge or even chemical information). For instance, one can break down a drug chemically and study it in terms of reactions it would cause due to its components. A comparison of this with its real world effects would be extremely useful in areas such as drug development. Moreover a classification would also allow many other tasks to be performed which are related to the concepts of similarity among drugs and ADEs. We will describe these later in this chapter.

In terms of anomaly detection, the first logical step is to analyze the reduced contingency table, with classes of drugs and ADEs instead of the original individual drugs and ADEs. This would effectively pool the raw counts and reduce the problems associated with sample variation at low counts. Additionally this would allow us to analyse the pooled classes of drugs and ADEs in blocks and we can look for anomalies within these blocks. Secondly, we can still analyze the individual drugs and ADEs, but use the pooled information in the baseline instead. Again this would reduce sample variation, which is most severe in the baseline information and also we would be able to look for individual drug or ADE against the pooled baseline. And lastly, we can derive an alternative representation to that of a categorical contingency table, a spatial map of drug and ADE combinations for instance, with which we can apply more sophisticated models.

In effect, the improvements all center on the fact that we are no longer treating each cell in the original contingency table as a single entity, but we seek to gain knowledge of the inherent structure present in data. The aim is then to use this structural knowledge to incorporate into more sophisticated methods of anomaly detection than the smoothing methods. This structural information will be represented by similarities (distances) among drugs and ADEs and with these similarities we can pick out groups which are close together for the purposes of analysis. We are then no longer restricted to performing anomaly detection on single entities (which will always suffer from the effects of sampling variation and poor quality data, regardless of the amount of smoothing or method or estimation), but we can apply a much more global perspective.

4.3 Profiling and Expert Knowledge

The analytical approach, which we will describe in depth in the next chapter, mainly focuses on how a drug reacts to different ADEs, and vice versa. In this way, we can think of the counts of ADEs arising according to a drug as a profile of that drug. Given only a raw contingency table, this is the simplest and most intuitive approach to deriving associations among drugs and ADEs. The expert knowledge can come from two separate sources. One can simply ask a medical expert to classify the drugs and ADEs according to similarities in practice and obtain a distance matrix this way. However, as mentioned, this is time consuming and infeasible for large numbers of drugs and ADEs (or, in the cases of new drugs and ADEs, one may know very little about them). Alternatively, but only in the case of drugs, one can obtain the chemical structures. This, on the other hand, has the disadvantage of not necessarily giving the right information (one can never be sure what the overall effect of a combination of chemical components is).

4.4 Outline of Remainder of Thesis

Below we give a flow diagram of the general outline of the work in the remainder of this thesis.

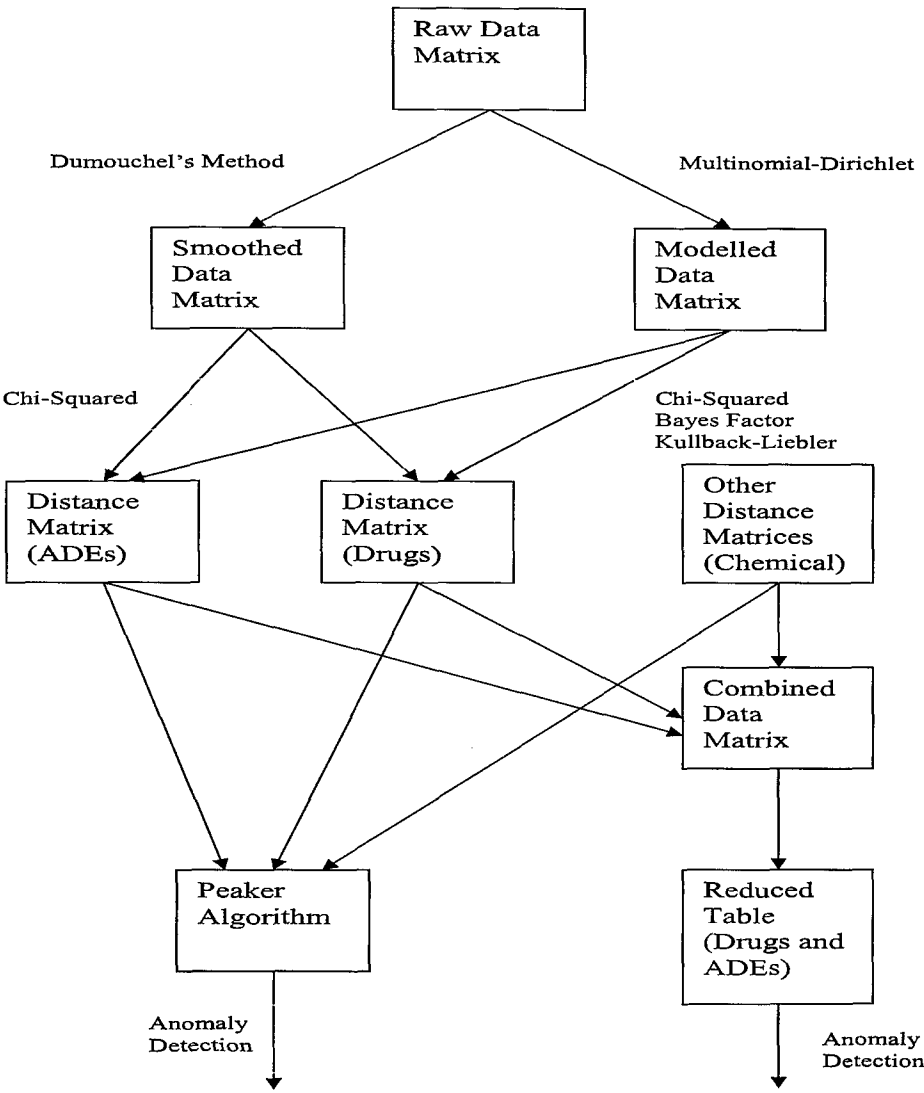


Figure 4.1: Flow Diagram of Remainder of Thesis

To start with, we have the table with the raw cell counts. We can use a variety of methods (such as DuMouchel's GPS method) to smooth the raw counts so as to obtain a table of smoothed counts. From the smoothed data distances are then produced between drugs and ADEs. We will describe in detail in the next chapter

how to calculate these distances and in particular, the suitability of a variety of distance metrics. With the distance matrices in place, we are ready then to perform anomaly detection with the aid of data mining tools. In particular, we will introduce a new data mining tool which will allow us to detect peaks of drugs and ADEs given the distances between them. Additionally we have the idea of using external information, the already mentioned chemical information which we can organise into a distance matrix among drugs and other similar types of distances matrices. We will then incorporate these external distances alongside of the existing distance matrices we obtained using analytical methods. The combination of these disparate distances presents quite a challenge and we will devote further analysis to the most suitable method of achieving this.

Chapter 5

Profiling

In this chapter we describe the ideas of profiling a drug or an ADE such that we can measure the similarities among drugs and ADEs. We will describe the basic ideas of the profiling approach and discuss the various considerations we need to take into account. We will then describe and compare a variety of distance metrics for the purposes of measuring the similarities in our particular type of data. Keeping in mind the properties of our data we need to carefully choose a suitable distance metric. We apply these distances to both simulated and real data sets to compare the behaviour of these metrics. In particular we will be concerned with the behaviour of the various distances under the condition of sparsely populated data in high-dimensional space. In general, such sparsely populated data sets require specifically designed distances and in certain instances fractional distances can be helpful for such purposes. We will discuss this further later in the chapter.

5.1 Basics

As discussed in section 4.3, the basic idea is that pairs of drugs can have inherent similarities and that we can estimate these similarities in a variety of ways (in terms of chemical similarities or, if that is not available, profiles of ADEs for each

drug). However, there are problems associated with generating distances between raw counts. The data itself is unreliable and contains a larger number of false reports; this can be very misleading in a low count profile. Together with the large sampling variation at low counts, one tends to obtain some rather arbitrary profiles and consequently distances. The following section discusses this in more detail.

5.2 Uncertainty

If we are working with the raw counts, problems arise due to the differing sample sizes in the profiles. This is a problem which is not fully addressed in the present literature. The standard method of getting around this difficulty is by some form of standardising. Even after careful standardization, however, it is possible for anomalous cases to arise. Consider the example in Table 5.1, how can we obtain an accurate measure of the distance between these three vectors?

Vector 1	Vector 2	Vector 3
0	50	3
0	50	2
1	50	300
0	100	5

Table 5.1: Three Vectors

Any way of standardising would give vector 1 as closer to vector 3 than vector 2. However, there is no real evidence to suggest that this should be true, since there is essentially very little information contained in vector 1. Testing procedures of the equality of these vectors would advocate that vector 1 has too small a sample size. However, as discussed, the low count profiles tend to be the ones we are interested in and we need some way of generating a distance.

In order to fully eradicate all anomalous cases, we need some way of smoothing

the data so that sampling variation is reduced (in particular for low sample sizes). It is with this in mind that the approaches in the following sections are adapted.

5.3 DuMouchel's Approach

DuMouchel advocated clustering the ADEs according to their drug profiles, so as to obtain a reduced set of ADEs to work with instead of the MedDRA hierarchy.

To obtain a set of distances for clustering, he first smoothed the table using GPS, and then calculated distances according to these smoothed odds ratios, in the same way as the profiling approach described already. We will discuss this method further in the subsequent sections.

5.4 Distance Metric

A metric $d(\cdot)$ has to satisfy the following properties:

1. Identity $d(x, x) = 0$.
2. Symmetry $d(x, y) = d(y, x)$.
3. Triangular Inequality $d(x, y) + d(y, z) \geq d(x, z)$.

We now give some of the most commonly used distance metrics that can be applied, as well as one which has not appeared in the literature.

5.4.1 Euclidean Distance

One of the simplest L_2 metrics is the Euclidean distance, defined by

$$d(X, Y) = \sqrt{\sum_i (x_i - y_i)^2}. \quad (5.1)$$

In our Multinomial-Dirichlet formulation, we can calculate the Euclidean distance between posterior estimates of p_{ij} , so that we have

$$d(P_1, P_2) = \sqrt{\sum_i (\hat{p}_{1i} - \hat{p}_{2i})^2}. \quad (5.2)$$

The Euclidean distance obeys all metric properties.

5.4.2 χ^2 Distance

In our problem, we use the posterior estimates of p_{ij} , $\hat{p}_{ij}$, through which we can calculate two forms of the χ^2 distance.

The first is the well-known χ^2 test statistic for independence. For two Multinomial vectors, P_1 and P_2 , the test has the null hypothesis $H_0 : P_1 = P_2$ and uses the statistic

$$T = \sum_i \sum_j \frac{(\hat{n}_{ij} - o_{ij})^2}{o_{ij}}, \quad (5.3)$$

where $\hat{n}_{ij} = N\hat{p}_{ij}$ and $o_{ij} = N\hat{p}_i\hat{p}_{.j}$. The second is the χ^2 distance, defined by

$$d(P_1, P_2) = \sum_j \frac{(\hat{n}_{1j} - \hat{n}_{2j})^2}{\hat{n}_{1j} + \hat{n}_{2j}}. \quad (5.4)$$

The two distances are equivalent, so we will only use the χ^2 test statistic in the subsequent sections.

The χ^2 distance obeys all metric properties.

5.4.3 Jensen-Shannon Divergence

We use a symmetric form of the Kullback Liebler divergence, known as the Jensen-Shannon Divergence (Lamberti & Majtey (2003)), to calculate the divergence be-

tween the two distributions, $f(\theta)$ and $g(\theta)$,

$$d(f(\theta), g(\theta)) = \frac{1}{2} \left(\int f \ln \frac{f}{g} \delta\theta + \int g \ln \frac{g}{f} \delta\theta \right). \quad (5.5)$$

In our particular application, where f and g are the posterior distributions of two Multinomial probability vectors, so that $f \sim \text{Dir}(\mu_i + n_{1i})$ and $g \sim \text{Dir}(\mu_i + n_{2i})$, the Jensen-Shannon divergence takes the form

$$\begin{aligned} d(f, g) = & \frac{1}{2} \left(\ln \Gamma\left(\sum_i \alpha_i\right) - \sum_i \ln \Gamma(\alpha_i) + \sum_i (\alpha_i - 1) (\Psi(\alpha_i) - \Psi(\sum_i \alpha_i)) \right. \\ & \left. - \ln \Gamma\left(\sum_i \alpha_i\right) + \sum_i \ln \Gamma(\alpha_i) - \sum_i (\alpha_i - 1) (\Psi(\alpha_i) - \Psi(\sum_i \alpha_i)) \right) \\ & - \frac{1}{2} \left(\ln \Gamma\left(\sum_i \beta_i\right) - \sum_i \ln \Gamma(\beta_i) + \sum_i (\beta_i - 1) (\Psi(\beta_i) - \Psi(\sum_i \beta_i)) \right. \\ & \left. - \ln \Gamma\left(\sum_i \beta_i\right) + \sum_i \ln \Gamma(\beta_i) - \sum_i (\beta_i - 1) (\Psi(\beta_i) - \Psi(\sum_i \beta_i)) \right), \end{aligned} \quad (5.6)$$

where $\alpha_i = \mu_i + n_{1i}$ and $\beta_i = \mu_i + n_{2i}$.

The Jensen-Shannon divergence obeys all metric properties.

5.4.4 Bayes Factor

The Bayes Factor does not appear in the literature as a distance measure¹. This is strange since it is an intuitive measure of the similarity of two attribute vectors.

We define our distance between two Multinomial probability vectors P_1 and P_2 as the Bayes Factor when testing the null hypothesis $H_0 : P_1 = P_2$. Let $D(x, y)$ be the distance between x and y . The Bayes Factor is calculated as follows.

We take independent Dirichlet prior distributions for the P_i 's under the alterna-

¹For literature on the Bayes Factor see Kass & Raftery (1995), Kass (1993) and Gunel & Dickey (1974).

tive hypothesis H_1 , so that

$$P_i|H_1 \sim Dir(\alpha_i),$$

then

$$P_i|H_1, D \sim Dir(\alpha_i + n_i).$$

We can then calculate the Bayes Factor from the definition

$$BF = \frac{\int \pi(D|\theta)\pi(\theta|H_0)d\theta}{\int \pi(D|\theta)\pi(\theta|H_1)d\theta}, \quad (5.7)$$

eventually arriving at the analytical form

$$\begin{aligned} BF = \sum \{ & (n_{1j} + \alpha_{1j} - 1)(\Psi(n_{1j} + \alpha_{1j}) - \Psi(\sum(n_{1j} \\ & + \alpha_{1j}))(n_{2j} + \alpha_{2j} - 1)(\Psi(n_{2j} + \alpha_{2j}) - \Psi(\sum(n_{2j} \\ & + \alpha_{2j}))(n_{1j} + \alpha_{1j} - 1)(\Psi(n_{2j} + \alpha_{2j}) - \Psi(\sum(n_{2j} \\ & + \alpha_{2j}))(n_{2j} + \alpha_{2j} - 1)(\Psi(n_{1j} + \alpha_{1j}) - \Psi(\sum(n_{1j} \\ & + \alpha_{1j}))) \}. \end{aligned} \quad (5.8)$$

Clearly the Bayes Factor does obey the Symmetry property. The Bayes Factor does not obey the Identity property, but this can be remedied easily by setting $D(P_1, P_1) = 0$.

5.4.5 Comparisons of Distances

First of all we would like to discuss some points concerning distance metrics. There are a wide variety of methods used in the data mining context associated with prob-

lems such as clustering and nearest neighbour search². In most instances suitable distance metrics are key to the performance of these methods. This can be particularly difficult when we are concerned with certain problems of the data such as sparseness and high dimensionality. It can be quite difficult to generate meaningful distances under these conditions (we will discuss more on this later). For the remainder of this section we will keep these considerations in mind when comparing the various distances that we can utilise.

As discussed, an alternative notion of generating a distance between two Multinomial vectors is that of testing the equality of two Multinomial vectors. The most commonly used statistic is the χ^2 test statistic, given in section 5.4.2. The χ^2 test has been extensively studied, such as in Cressie & Read (1984), Read (1984) and Tate & Hye (1973). This has been well studied in terms of power calculations and other criterion, with various minimum expectations suggested (Fisher et al. (1994) recommended 5, Kendall (1952) suggested 20). Improvements have been sought in using the general family of power divergence statistics, of the form

$$\frac{2}{\lambda(\lambda + 1)} \sum_i n_i [(\frac{n_i}{E_i})^\lambda - 1], \quad (5.9)$$

with a smaller value for λ suggested for small sample inference. The expectations we have in the present context are much smaller than would be involved in traditional χ^2 tests, so that no modification is sufficient to provide a reliable test statistic.

In the same way the Bayes Factor is derived from testing equality of two Multinomial vectors. It is severely inflated by a flat prior and is unreliable as a test statistic. Conversely, since it is a probabilistic statement in favour of the alternative hypothesis, it takes into account the uncertainty present in the data.

The Jensen-Shannon Divergence is a symmetrized and smoothed extension of

²Mostly in specialised fields, for examples see Cheng et al. (2004), Huang (1998) and Pekalska et al. (2001).

the Kullback-Liebler Divergence (Menendez et al. (1997), Morales et al. (1994) and Lamberti & Majtey (2003)), which also belongs to a wider class of information-theoretic based divergence measures. It is a natural measure of the divergence of two probability distributions. However, as with all information-theoretic measures, it gives the same distance to large objects that differ by a small amount as small objects that differ by the same amount. This stems from the fact that all information-theoretic measures are based on the idea of the amount of information required to tranform one object to another. Clearly this deficiency can be distorting in the present context.

The suitability of a statistic in a testing procedure is very different from its suitability as a distance metric. We are no longer concerned with the asymptotic property of a test, but rather the monotonic nature of a distance. To this aim we devise the following criterion, which is that a distance should obey the following ordering when applied to two profiles. If we have simulated Multinomial vectors under two Multinomial distributions, the intuitive ordering of distances should be as shown in Table 5.2 (from 1 – 6, with 1 being the smallest in orders).

	M1(high)	M1(low)	M2(high)	M2(low)
M1(high)	1	3	6	5
M1(low)	3	2	5	4
M2(high)	6	5	1	3
M2(low)	5	4	3	2

Table 5.2: Ordering of Distances (M1(high) denotes vectors of high counts generated under Multinomial distribution 1, *etc.*)

5.5 Multinomial Formulation

5.5.1 Method

We have demonstrated some of the problems associated with the approach outlined in section 5.1. One improvement we can make is to take into account variations in the profiles and to smooth over these variations.

As noted in section 5.3, DuMouchel uses a smoothing approach to produce a distance. He produces a measure for each cell using the Gamma-Poisson Shrinker and defines profiles of drugs and ADEs using these measures. This does indeed smooth over some of the variation in the profiles.

Our approach uses the raw counts to produce our distance metric, rather than a derived measure which seems to be rather *ad hoc*. On the one hand, we approach the problem from a statistically intuitive way. On the other, we believe that we can much better outline the requirements of our distance metric and make improvements where necessary.

We assume that each profile follows a Multinomial distribution, so that

$$p(x_1, \dots, x_n | p_1, \dots, p_n) = \binom{N}{x_1 \dots x_n} p_1^{x_1} \dots p_n^{x_n}, \quad (5.10)$$

where $x_1 \dots x_n$ and $p_1 \dots p_n$ denote cell counts and parameters in a single profile, and N is the sum of the profile.

We can estimate the distribution through maximum likelihood, and we would obtain the maximum likelihood estimates

$$p_i = \frac{x_i}{N}. \quad (5.11)$$

Clearly the estimates are affected by sample size, and at low counts are unreliable.

One way to get around this is to take a Bayesian approach to the problem, by

assuming a prior for the Multinomial likelihood. We can use the conjugate prior to the Multinomial distribution, which is the Dirichlet distribution, such that

$$p(p_1, \dots, p_n | \mu_1, \dots, \mu_n) = \frac{\Gamma(\sum \mu_i)}{\prod \Gamma(\mu_i)} \prod p_i^{\mu_i-1}. \quad (5.12)$$

The posterior is also a Dirichlet distribution, of the form

$$p(p_1, \dots, p_n | \gamma_1, \dots, \gamma_n, x_1, \dots, x_n) = \frac{\Gamma(\sum (\mu_i + x_i))}{\prod \Gamma(\mu_i + x_i)} \prod p_i^{\mu_i+x_i-1}. \quad (5.13)$$

In the absence of prior knowledge, we can simply use a uniform prior. However, in some cases we would have some prior knowledge, and the prior parameters can be set according to our prior knowledge and in such a way that strong information in the data is needed to significantly change the shape of the distribution in the posterior. This would then be an intuitive way of smoothing over the data to obtain a posterior distribution.

The above approach already has one major advantage over using a derived measure; the ability to incorporate prior belief into our distance metric. Another intuitive way of specifying the prior is to give greater weight to the more severe ADEs, and this can be achieved by setting a weaker prior parameter for the ADE so that the posterior is affected more sharply by the information in the data.

Once we have the posterior distributions for the profiles, we can then obtain a distance metric based on the posteriors. The Kullbeck-Liebler distance is one natural metric, as it measures the discrepancy between two distributions. Alternatively, since we have the distributions of the profiles, we can construct a distribution of the distance between two profiles. This would have the advantage of modelling the uncertainty of distances between profiles.

5.5.2 Prior Specification

We will now discuss how we should specify the prior in our Multinomial distribution.

Although we have discussed the use of a uniform prior, this is not particularly applicable in this case. Clearly we need to attach different prior parameters to different parameters in the likelihood to reflect the differing baseline information. One approach we can take is to assign the prior parameters as the means of respective ADEs across all drugs. This would have the effect that the shape of the posterior is changed according to a combination of differing baselines and observed counts. Alternatively, we can assign a uniform prior, but allocate differing weights to the counts for each ADE. This would have the effect that the shape of the posterior would change at the same scale of magnitude. Again these weights can be decided according to baseline information.

The above approaches to prior specification depend on a combination of subjective assessment by the analyst and information in the data. An alternative is to use an entirely data-dependent method such as Empirical Bayes. We prefer, in this instance, to use more subjective priors, since we can specify clearly what we require.

5.5.3 Distance

As already mentioned, one distance metric we can use is the Kullback-Liebler distance between distributions, which is of the form

$$KLD(q|p) = \int \log\left(\frac{p}{q}\right)p. \quad (5.14)$$

For two Dirichlet densities p and q such that

$$p(p_1, \dots, p_n | \mu_1, \dots, \mu_n) = \frac{\Gamma(\sum \mu_i)}{\sum \Gamma(\mu_i)} \prod p_i^{\mu_i - 1},$$

$$q(p_1, \dots, p_n | \lambda_1, \dots, \lambda_n) = \frac{\Gamma(\sum \lambda_i)}{\sum \Gamma(\lambda_i)} \prod p_i^{\lambda_i - 1},$$

the Kullback-Liebler distance is then

$$KLD(q|p) = \log\left(\frac{\Gamma(\sum \mu_i) \sum \Gamma(\lambda_i)}{\sum \Gamma(\mu_i) \Gamma(\sum \lambda_i)}\right) + \sum ((\mu_i - \lambda_i)(\psi(\mu_i) - \psi(\sum \mu_i))). \quad (5.15)$$

Note that the Kullback-Liebler distance as given above is of the form q given p and is not the same as $KLD(p|q)$. Thus the Kullback-Liebler distance is not symmetric. The Kullback-Liebler distance also does not obey the Triangular Inequality, so that only one of the three properties of a metric is satisfied ($KLD(p|p) = 0$). While non-symmetry is a problem, the disobedience to the Triangular Inequality is particularly undesirable when we need to look for clusters in high dimensional space. This renders the Kullback-Liebler unsuitable to our purposes. To illustrate the above, we have plotted 4 distribution functions of the $Beta(\alpha, \beta)$ distribution, which is a 2-parameter Dirichlet distribution.

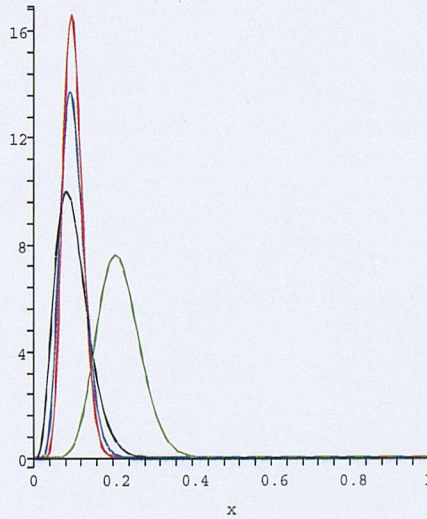


Figure 5.1: Plots of Four Beta Distributions

In Figure 5.1 the distribution plotted in black is $Beta(5, 45)$, which we will take as

our prior distribution. The distributions plotted in green and blue are $Beta(10, 90)$ and $Beta(15, 135)$ respectively and that plotted in red is $Beta(13, 47)$. Clearly we can see that there is a large difference between $Beta(13, 47)$ and the other three distributions while there is a small difference between those three distributions. This is to be expected, simply by a calculation of the mean and variance of the distributions. As such we would expect this to be reflected in our distance metric between the distributions. Table 5.3 shows the Kullback-Liebler distances between these distributions.

	$Beta(5, 45)$	$Beta(10, 90)$	$Beta(15, 135)$	$Beta(13, 47)$
$Beta(5, 45)$	0	0.16	0.47	3.86
$Beta(10, 90)$	0.10	0	0.05	3.58
$Beta(15, 135)$	0.22	0.04	0	3.57
$Beta(13, 47)$	2.55	4.83	7.27	0

Table 5.3: Kullback-Liebler Distances Between Four Distributions

As we can see from the last row, these distances are inflated and do not obey the Triangular Inequality. We can note at this point that the Kullback-Liebler distance is generally used in the literature to give an idea of the discrepancy of two distributions and, as we have argued, is clearly not suitable if we then proceed to work implicitly with these distances.

As we can see from Figure 5.1, an alternative, natural distance to use would be the overlapping regions under two $Beta$ distributions. However, to extend this to high dimensions would be difficult to evaluate and almost certainly computer intensive. One way to do this would be to sample points under the two distributions and thus sample the percentage of points that fall under both distributions, which we can then use as the distance metric. However, while this is applicable to distributions of two parameters such as the Beta or Normal distributions, in higher dimensions there is no straightforward analogy of such overlapping regions (indeed we can no

longer visualize exactly what overlapping regions mean).

Another approach to take would be to either work with the posterior Dirichlet parameters or a scaled version. If we scale them by summing them to unity, then we are simply using the posterior estimates of the likelihood parameters. Again in this case we can use forms of the Euclidean or χ^2 distances between the posterior parameters or estimators to evaluate a distance.

5.5.4 Application

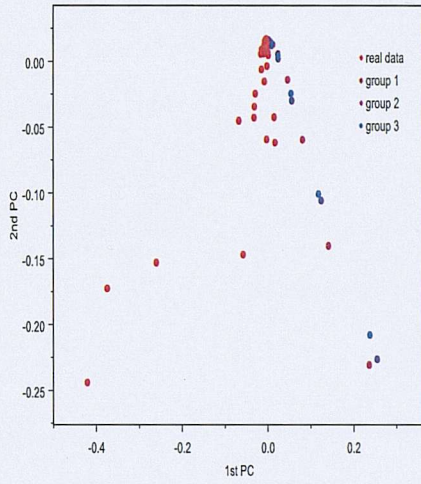
Here we describe the simulated dataset of section 5.6. We will have a prior based on the mean. Subsequently we will use a χ^2 distance between the posterior estimates of individual likelihood parameters.

If we use the raw means for the prior, we will find that some of the means are very small, which is then equivalent to using a weakly informative prior. We need to use a stronger prior so that strong data is needed to change the posterior, therefore we will make the following initial modifications to the prior, shown in Table 5.4.

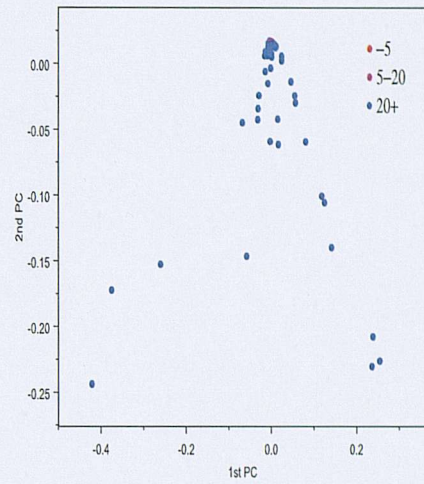
Mean	Prior Parameter
< 1	$mean + 1$
1 – 2	$mean + 2$
2 – 10	$mean + 3$
10+	$mean + 5$

Table 5.4: Prior Specification Based on the Mean

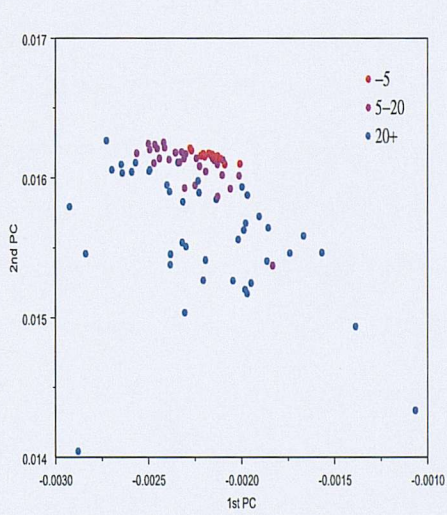
We can see from Figure 5.2 that we have now a fan shaped effect in our distances. From Figure 5.2(b) and 5.2(c), we can see that the profiles with low counts are now clustered together and the counts increase in an outward direction from the lowest counts. In effect, low counts provide no real information and thus do not change the posterior parameters very much. This has achieved one of aims of our distance metric, in that uncertainty is incorporated. However, we now have a clear separation



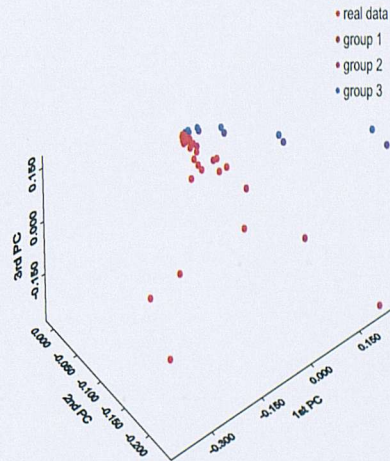
(a) All Points



(b) Points of Differing Counts



(c) Points of Differing Counts (Zoomed)



(d) 3D Plot of All Points

Figure 5.2: Plots of 1st v 2nd Principal Component of Distance Matrix Obtained by Multinomial Approach

between profiles with low counts and profiles with high counts. This is clearly not entirely desirable. Whilst we would like to have this separation to a degree, we also need our distance metric to take into account the profile distributions. This above analysis shows an opposite effect to the basic profiling method discussed earlier. We have now a condensed region of low counts, which are now close in distance, and the distances tend to increase with increasing count values. Looking at Figure 5.2(d), we can see a very clear trend of the simulated data points from low counts near the central condensed region to high counts. We need to take this into consideration when applying the distances to real data.

5.6 Application and Discussion

In this section we apply all the above methods to a set of simulated data. The data comprises of 70 drugs and 2245 ADEs. The reason for simulating such a data set is that we wish to investigate the behaviour of the distance metrics when applied to a data set closely mirroring that of real world data. We also wish to differentiate easily amongst the distances. hence we will only simulate 70 drugs so as to be able to easily interpret the results. We will apply our methods to the drug profiles. The 70 drug profiles are simulated from 3 distinct Multinomial distributions (this is done for convenience due to the fact that it is difficult to simulate an adequate data set from Poisson distributions that closely resemble real world data), of sizes 20, 30 and 20. Within each group of simulated data, we change the sample sizes from very low (50 counts in 2245 ADEs) to high (100,000 counts in 2245 ADEs).

The methods we will use are:

1. DuMouchel's method, with χ^2 distance
2. Multinomial-Dirichlet formulation, with distances:

- Kullback-Liebler distance
- χ^2 distance
- Bayes Factor

We use hierarchical cluster analysis to demonstrate whether the distances obey the ordering in Table 5.2. We can see from Figure B.1 to B.8 (Appendix B) that DuMouchel’s method performs the best with complete-link clustering, also note the similarities between the B.3 and B.5 and between B.4 and B.6 as these two distances perform similarly for the particular data we analyse. In fact, all the methods apart from the Bayes Factor perform much better with complete-link than single-link, whereas for the Bayes Factor this trend is reversed. The reason is that for all other distances apart from the Bayes Factor, two profiles with low counts are closer in distance than two profiles with large counts from the same group. Thus single-link clustering would perform poorly as low profiles from all groups would cluster first, followed by higher profiles, leading to arbitrary clustering. For the Bayes Factor this is not so, thereby leading to better single-link clustering. It is difficult, however, to infer more useful information regarding the performances of the distances from the clustering alone. We will also mention that the performance of the Bayes factor in terms of separating the simulated Multinomial distributions is superior to the other distance metrics.

The results show that the only distance capable of fulfilling the criteria of Table 5.2 is the Bayes Factor. One would intuitively expect the method by DuMouchel of comparing smoothed odds ratios not to follow the above ordering. In particular, at low counts, one will simply have many of the smoothed odds ratios being the same for two different profiles due to the fact that they are dominated by the prior smoothing parameters. This leads to low counts being very close to each other, regardless of the distance used (any L_2 norms, for instance). The Euclidean distance clearly suffers

from the same problem of having low counts close to one another, due to the lack of normalization. The Jensen-Shannon divergence has the problem of giving two distributions as being closer to one another if there is a large proportional change in small parameter values, as opposed to a small proportional change in large parameter values. The χ^2 distance is not appropriate for small expected values of cell counts, and distorts the distances for such values to be smaller than the weight of evidence suggests. The Bayes Factor, on the other hand, does obey the ordering in Table 5.2, since it is a probabilistic statement in favour of the alternative hypothesis of inequality of Multinomial vectors. It should take into account the sample sizes of the vectors (in other words, the uncertainty of making such a probabilistic statement).

Finally, it should be remarked that, of the ordering in Table 5.2, the focus of the previous analysis has been on the ordering numbers 1 and 2. This is not negligence, but all other orderings are rather obvious and one would expect any distance to demonstrate such behaviour. Of the ordering numbers 1 and 2, we can of course force any distance to follow this order by normalizing with a size factor. Indeed, it has been suggested that this should be performed on information-theoretic divergence measures due to their deficiencies as discussed in section 5.4.5. The simplest such device is to normalize each distance by the sum of counts in the two profiles,

$$D_{new}(Pr_1, Pr_2) = \frac{D(Pr_1, Pr_2)}{N(Pr_1) + N(Pr_2)}, \quad (5.16)$$

where $N(Pr_1)$ is the total count in profile 1. However, such normalization destroys other inherent orders in the distance metric. For instance, if we have the following vectors, normalization can easily give vectors 3 and 4 being closer than vectors 1 and 2, an undesirable situation.

In the next analysis we have simulated a dataset of 20 drug profiles under three different Multinomial distributions. We then insert this simulated dataset into a

Vector 1	Vector 2	Vector 3	Vector 4
0	0	50	3
0	0	50	2
10	20	50	300
0	0	100	5

Table 5.5: Four Vectors

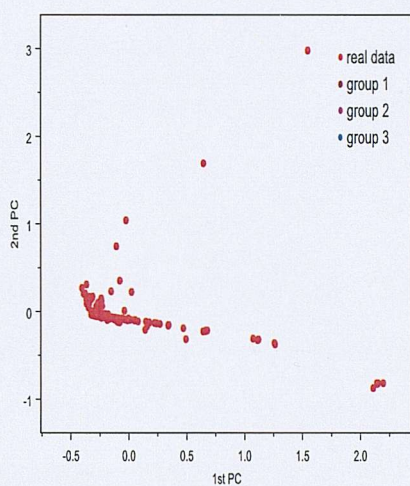
real dataset of 152 drug profiles. In each profile there are 2245 ADEs. The aim is to see how well the profiling method separates the simulated groups from the real noisy data. In the simulated dataset, the marginal counts of each profile range from low to high (50 to 10,000), which will give us an idea of the performance of the profiling method for differing counts.

We calculate our distance metric using the Euclidean distance, after standardizing by summing each profile to unity. We then apply Multidimensional Scaling (MDS) to the resulting distances and plot the first two components against each other for the purposes of visual inspection.

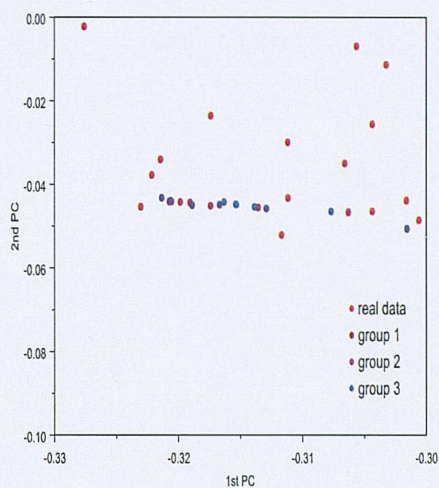
In Figure 5.3(b), we can see the separation of the different groups. It can be seen that the groups have a high degree of overlap, which is not desirable since we would expect more distinct groups. The distinct groups which do arise are profiles with high counts. The simulated profiles with lower counts, on the other hand, are not picked out at all well.

It is also interesting to note that the simulated profiles cluster together, regardless of which groups they belong to. This can be seen clearly from both Figure 5.3(a) and 5.3(b). We believe this is due to the fact that the simulated profiles have a much more regular structure, whereas real datasets tend to be much more noisy and irregular.

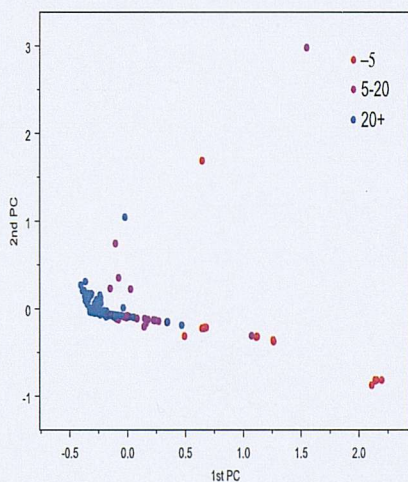
From Figure 5.3(a), we can see the general structure of the profiles. The elongated shape is a little worrying, since we should expect a more spherical shape, although we can see that the vast majority of the profiles are clustered in a small



(a) All Points



(b) Points of Simulated Data (Zoomed)



(c) Points of Differing Count Values

Figure 5.3: Plots of 1st v 2nd Principal Component of Distance Matrix Obtained by Basic Profiling Method

condensed region. This can be explained by Figure 5.3(c). We can see that the profiles represented by red dots are profiles that contain very few counts (those to the right only contain 1 or 2 counts). As such, they are essentially outliers. Indeed, if we were to plot all the components in high dimensional space, then these profiles with low counts would be much further away from each other. However, since these counts are so low, we really can not trust to this distance. Rather, we would ideally like these profiles to be closer together to reflect the fact that we can make no clear decision.

It is worth noting at this point that the two dimensional MDS plot is simply a tool to enable us to visualize the distances. In reality, we would be working with the distances alone. Clusters which are close together in these plots are certainly not always close together in higher dimensions.

As a result of the above analysis, we can gather that the basic profiling method has some undesirable properties. Firstly, it does not separate profiles well when counts are relatively low. Secondly, profiles with very low counts can be very far away from one another. This is clearly undesirable since we have insufficient information to draw such a conclusion. And finally, we have no indication of the uncertainty in the data.

5.6.1 Conclusion

In this chapter we have seen that the basic profiling method does not perform very well when profiles contain low counts. It is an interesting question to ask what we can do with these low counts. Some of the previous approaches have simply eradicated profiles with low counts. However, as we have argued, this is a dubious practice at best. What we need is some way of taking into account the uncertainty in the profiles and construct our distance metric to reflect this uncertainty.

Finally we would like to discuss some traditional approaches in data mining.

There are a wide variety of methods used in the data mining context associated with problems such as clustering and nearest neighbour search. These are based on using some distance metric between points in high dimensional space. The problems associated with the curse of dimensionality are well known in this context. For instance, Aggarwal et al. (2001) and Beyer et al. (1999) gave accounts of data mining in high dimensional space. Some of the proposed solutions consist of data mining in subspaces and using modified distance metrics (Beyer et al. (1999)). However, we have in our problem a different formulation to the traditional approaches, which are mainly concerned with large data sets containing many observations with a large number of well defined and measurable attributes. One then proceeds to define distance metrics between attribute values and perform clustering. We have no such luxury, in that our attributes contain differing counts and therefore differing degrees of uncertainty. Our primary concern then is to define distance metrics which can take uncertainty into account. In this context there has been little work in the present literature, in particular there has been little discussion of a suitable distance metric to use.

Chapter 6

Data Mining Software

This chapter is a rather stand-alone chapter from the rest of the thesis. However data mining software plays an essential role in analysing the real data sets, therefore we feel it is apt to discuss the central issues in data mining software before going further.

With the relatively recent explosion of data mining applications, so has a need for data mining software arisen. Corresponding to this there has been an explosion in the number of data mining packages marketed in recent years, to the extent that there are at present a bewildering amount of software available for both the commercial and academic users. These include some major packages which have been maintained over many years and are consequently very mature, some more recent packages as well as some very application specific packages.

For many users, choosing a suitable package is not an easy choice. There has been very little in the literature to help facilitate such a decision (some useful comparisons can be found in Davi et al. (2003)), which can perhaps be explained by the fact that this is rather difficult to achieve and it is much more common to tailor a package to a user's specific requirements. In this chapter we will review some of the packages available and compare them according to a set of criteria we have devised. In

particular, we will be looking at the packages SPSS Clementine, Statistica, KXEN, Ghostminer and XLminer. While this is rather a small selection, we have made this choice so that there is a wide representation of the types of packages available. We will mention more on this in the remainder of this chapter.

It must be stated here that the goal of this chapter is not to look in detail at the technical aspects of each package in terms of the types of analyses available, but rather to give a feel for the general issues that arise when using each package and to hopefully relate to the common users. To this end we will use some or all of the following list of criteria in the following sections:

1. General features (including pricing, learning curve, user interface, output).
2. Toolset.
3. Advanced features (including customisation, exportability, database extraction).
4. Balance between basic and expert users.
5. Scalability.

The rest of this chapter is organised in the following way. Section 6.1 gives a review of the capabilities of present-day software. Section 6.2 compares the pros and cons of the various packages. Some relevant examples are given in section 6.3.

6.1 Features of a Modern Data Mining Package

In this section, we will describe some of the aspects of a modern data mining package.

6.1.1 User Interface

There are generally three types of user interfaces for a standard package, a menu based approach, a stream based approach and scripting. The menu approach is generally for the more basic users. The stream based approach is slightly more advanced in that more care must be taken in deciding on the connection of nodes. Streams have two key advantages over the menu options. They allow complicated streams to be saved as projects and there is a clear sense of separating parts of a larger project into smaller manageable sub-projects. At the most advanced level, one can script for user customisation. Figure 6.1 shows the workspace for the Statistica package.

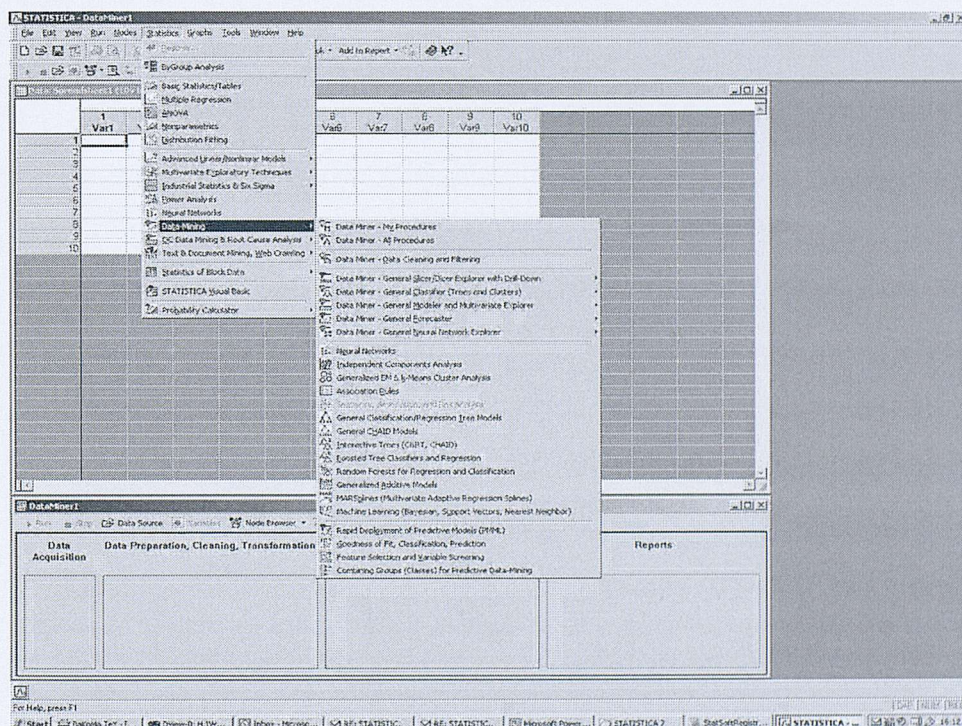


Figure 6.1: Workspace for the Statistica Package

We see the typical elements of a workspace, a drop-down menu of options and a spreadsheet for manual data input. At the bottom of Figure 6.1 we can see the stream interface. Figure 6.2 shows in more detail the stream workspace of the Statis-

tica package. Statistica takes the approach of forcing the user to create a stream of four separate parts, in order to maintain a clear and manageable appearance. This is not true of other packages such as Clementine. On the other hand, complicated streams can be confusing when links point back and forth and it can be visually clearer to keep streams pointing in a single direction.

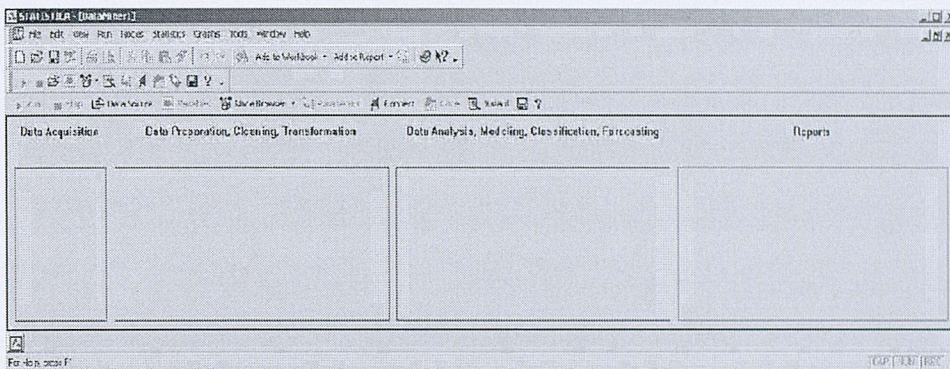


Figure 6.2: Stream Workspace for the Statistica Package

6.1.2 Toolset

In this section we list some of the common data mining tools available today. It must be noted that a comprehensive coverage of tools does not necessarily make for a good package. As is often the case, packages that can perform a specific set of tasks well may be preferable to the more general packages in terms of speed and accuracy. Moreover, it can be important to limit applications for the novice users who can be confused by a vast array of analytical tools and options. Choosing an appropriate method of analysis can often be difficult without the necessary statistical training and often the best option is to fully automate the modelling process to provide quick and easy solutions.

Table 6.1 summarises the types of analyses available in each of the packages.

	Statistica	Ghostminer	Clementine	KXEN	XL
Data Cleaning/Manipulation	y	y	y	y	y
Missing Data Imputation	y	y	y		y
Generalised Linear Models	y		y	y	y
Multivariate Analysis	y		y	y	y
Principal Components Analysis	y		y		y
Factor Analysis	y		y		
Canonical Analysis	y		y		
Multidimensional Scaling	y		y		
Nearest Neighbour	y	y	y	y	y
Correspondence Analysis	y		y		
Discriminant Analysis	y		y		y
Classification	y	y	y	y	y
CART	y	y	y	y	y
FSM		y			
kNN	y	y	y	y	y
SSV Tree		y			
Committee	y	y	y		
Neural Networks	y		y		y
Multilayer Perceptrons	y		y		y
Kohonen			y		
Radial Basis Functions	y		y		y
Time Series Analysis	y		y	y	
Cluster Analysis	y	y	y	y	y
Association Rules	y		y	y	y
CHAID Models	y				
Random Forest Models	y				
MAR Splines	y				
Support Vector Machines	y		y	y	

Table 6.1: Summary of Available Analyses

6.1.3 Visualisation and Graphics

Visualisation can often be the most effective way to start any analysis. Today's data mining packages often provide powerful graphical features to enable fully integrated support. In particular, many packages provide flexible graphical features which are fully customisable along with additional user-defined graphs. Graphs can also be dynamically linked to the data (even external remote data bases) to enable automatic redrawing and updating with real time data changes.

Table 6.2 summarises the types of visualisation and graphical tools available in each of the packages.

	Statistica	Ghostminer	Clementine	KXEN	XL
Visualisation	y	y	y	y	y
Histograms / Bar Charts	y	y	y	y	y
Pie Charts	y	y	y	y	y
Box-Whisker Plots	y		y		
Scatter Plots	y		y		
Probability Plots	y				
2D Graphs	y				
Scatter Plots	y	y	y		y
Range Plots	y				
Quantile Plots	y		y		
Normal Probability Plots	y				
Line Plots	y		y		
Sequential/Stacked Plots	y				
3D Graphs					
Scatter Plots	y		y		
Bivariate Histograms	y		y		
Range Plots	y		y		
Surface Plots	y		y		
Contour Plots	y		y		
Ternary Plots	y				
Categorized Plots	y				
Smoothing Plots	y				
Tessellation Plots	y				
Matrix Plots	y				
Lift Charts				y	
Correlation Plots			y		
Misclassification Plots	y	y		y	

Table 6.2: Summary of Visualisation Tools

6.1.4 Customisation

Ultimately packages cannot provide the exact features that an advanced user is entirely satisfied with. The user can customise the modules to suit individual preferences. Most packages provide easy to edit scripting features with each pre-defined module. Figure 6.3 shows this feature in Statistica, which uses industry standard Visual Basic as its scripting language.

Of course, to customise code requires a certain degree of advanced understanding,

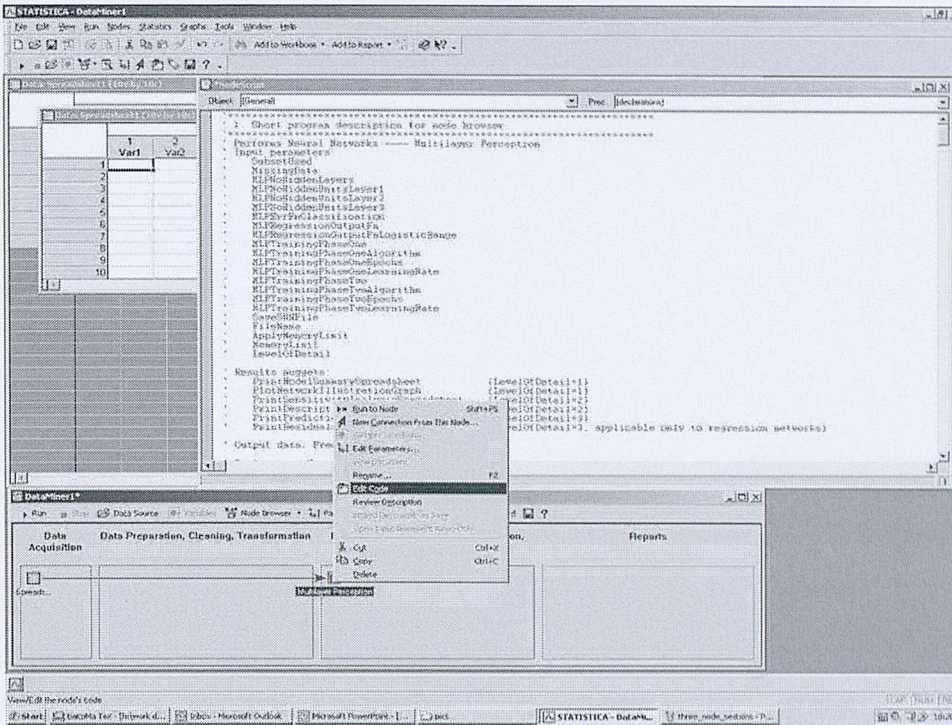


Figure 6.3: Customisation in Statistica

but this is generally quite straightforward to learn. In particular, a nice feature of most data mining packages is that all menu based tools and options can be converted into script and this greatly facilitates the scripting procedure since command line scripts can simply consist of portions of existing scripts.

6.1.5 Database Extraction

The scalability of modern day packages means that they can handle massive amounts of data. In most instances it is impossible to store the data on a local machine. Remote data warehouses are used to store the data source and any package must have the facility to access this data source. Mostly this is quite straightforward with in-built functionalities which allow easy access of compatible data sources.

6.1.6 Deployment and Exporting

Deployment refers to the use of an existing predictive model on a fresh set of data. Most packages have such a feature available. Moreover, models can be readily exported and deployed in other integrated applications. This can be done by either exporting the model itself to compatible applications or by converting the model to script and exporting the script. It is very important for a modern day package to be able to export its modules across to other packages. To this end a programming language, Predictive Model Markup Language (PMML), has been devised to enable universal compatibility among applications when defining statistical and data mining models. This means that any particular package can convert its modules into PMML and subsequently run in another application. Most packages provide such a facility for several other languages such as Java, C++ and variants of Visual Basic.

Figure 6.4 shows the exporting option in Statistica.

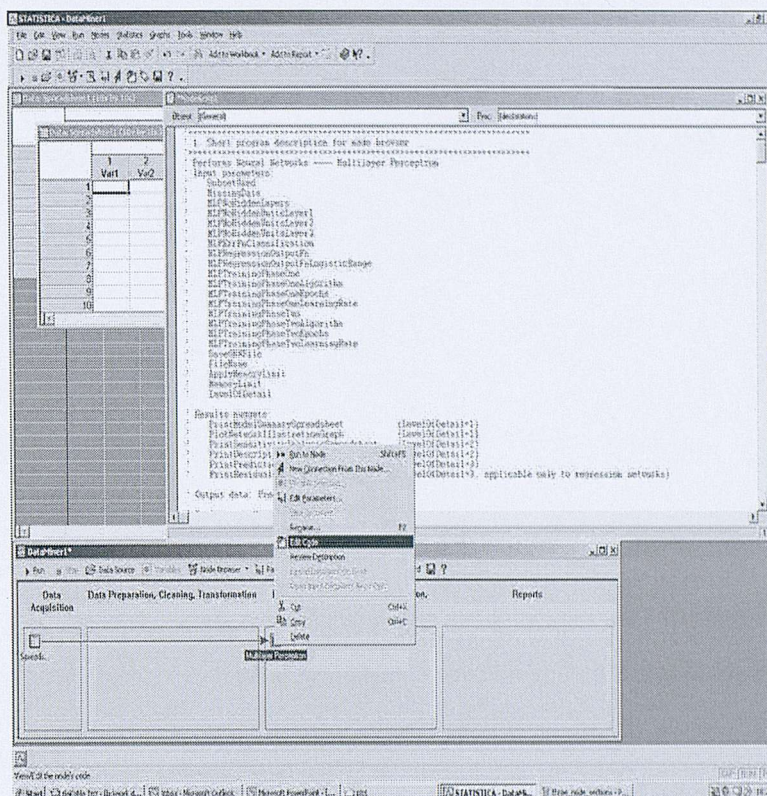


Figure 6.4: Model and Script Exporting in Statistica

6.2 Comparison of Software

6.2.1 Advantages and Disadvantages of Data Mining Packages

Statistica

Statistica is a very complete package in terms of both the analyses that it can perform and the graphical tools available. As one can see from section 6.1, the data mining package contains a very extensive set of methods including some very specialised analyses (possibly more so than any other existing package today). It is also one of the very few packages that incorporates rich sets of both parametric statistics and data mining tasks. Each individual method contains a powerful set of

user definable features which cover most aspects of the method. Additionally it is highly flexible both in terms of user definable features and easy to use customisation of existing features.

The user interface is easy to use and intuitive, with most aspects adjustable. On the other hand, there is a deliberate effort to simplify some of the features, especially for the stream nodes, in an attempt to avoid confusing users with overly complicated analysis options. These options can of course be customised through scripting.

Scalability is impressive in Statistica, having the ability to handle large datasets and databases across different platforms. Indeed, it is one of the leading applications in terms of scalability.

One of the nice features of Statistica is that it caters well for both the basic and advanced users, with a relatively straightforward learning curve but also very powerful customisable features for the advanced users. Indeed, one of the most important aspects of a universal data mining package is its ability to cater for users from all walks of life.

The only problem which has arisen is the propensity for Statistica to crash occasionally for rather arbitrary reasons.

Statistica provides an academic license and is affordable for personal use. Statistica is possibly the most complete package available today. Clearly a great deal of care has been taken over this package and most aspects have been fine tuned.

KXEN

KXEN is a specialised package that implements the Support Vector Machine. It performs a limited set of methods which includes mainly classification and regression, but in this capacity it is extremely fast and accurate. The philosophy of KXEN is to deliver fast business solutions on vast amounts of data. It dispenses with providing the user with a variety of analytical tools in favour of one approach. The aim is to

provide a non-expert user with a fast and accurate model on large data sets. We say non-expert because KXEN does not provide a set of tools to choose from and the process of data analysis is fully automated once a clean data set is presented to the internal Consistent Coder (for data preparation). Because of the full automation of KXEN, no sophisticated user interface is employed for the front-end. Figure 6.5 and 6.6 show two of the KXEN user interfaces. The user is asked initially to choose a model and input data, after which the whole analysis is carried out by simply clicking the Analyse button shown in Figure 6.6.

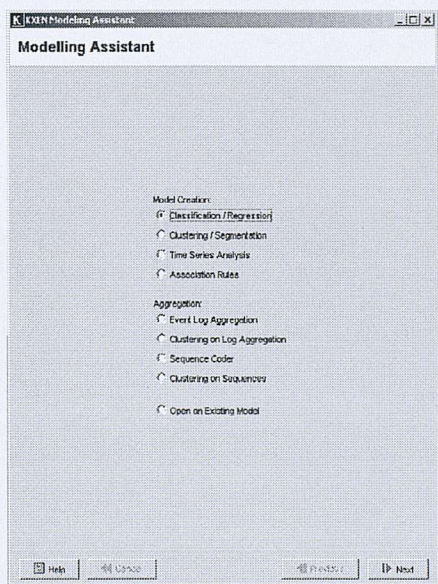


Figure 6.5: KXEN User Interface

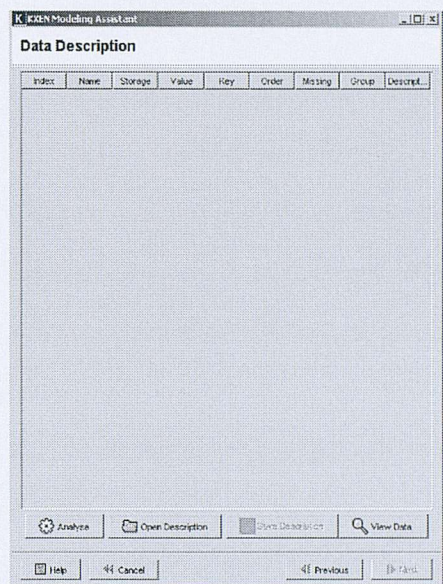


Figure 6.6: KXEN User Interface

The output is again entirely automated and rather limited, and user defined graphics cannot be performed. For instance, in Regression/Classification, the output is based on the performance indicators KI and KR which correspond to model *consistency* and *robustness*. Figure 6.7 and 6.8 show examples of model output that are available.

In essence, KXEN is a nice tool that accomplishes well what it is designed to do, but it is rather limited. Its main use is to give a fast overview (consisting mainly of variable selection, event associations, data sufficiency and exploratory modelling) of

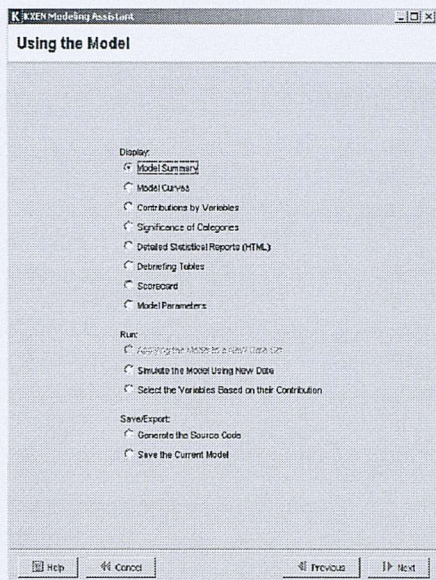


Figure 6.7: KXEN Model Output Options

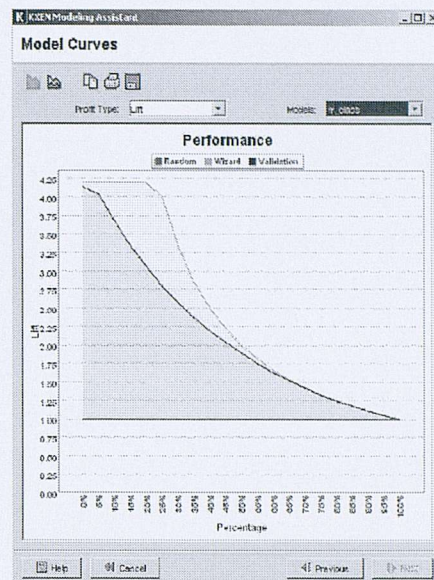


Figure 6.8: KXEN Lift Curves

a large data set to the business user. It cannot perform several other key aspects of data mining such as data cleaning/manipulation, data visualisation, exploratory data analysis and other in-depth statistical analyses.

Clementine

Clementine is a very mature package. This means that it is easily integrable within a work environment and has many third party vendors for complimentary features. Clementine contains an extensive set of data type and quality nodes and performs very powerful data checking and manipulation operations as well as missing data imputation. Additionally, powerful optimal parameter search routines are built into most of the data mining algorithms to enable fast computation. It contains an useful variety of algorithms which can be customised through its scripting language CLEM. On the downside, it has a limited range of graphical features. Its price is also prohibitive for personal and academic use.

Ghostminer

Ghostminer is a package that concentrates extensively on classification methods. As such it has a substantial range of such algorithms. It is reasonably fast computationally for such purposes. However, this is to the exclusion of most other data mining techniques and it is therefore very restricted in the analyses it can perform. Given the rather restricted nature of its algorithms, it seems not to have any obvious advantages over other packages when it comes down to both the computational aspect and also the level of customisability. In terms of the graphical packages and data manipulation Ghostminer is also not outperforming its competitors.

XLminer

XLminer is a limited package in terms of the methods and features it contains. Because it is an Excel add-on, it contains many of the limitations of Excel, in particular scalability and very limited graphical tools. On the other hand it is very cheap and fits within the Excel package which most people already have. Whilst it cannot be expected to perform in-depth analysis, it contains a useful set of methods for the basic users to try their hands at performing some data mining tasks. In fact, it contains a very useful set of data mining applications which include most of the common types of analyses. More features are being developed to add to the existing package which should enhance its capabilities somewhat and these include database extraction, data manipulation tools and graphical enhancements. At its price, XLminer provides a nice package for the non-serious user.

6.3 Examples

In this section we will examine the performance of some packages when applied to an example data set. The data set we used is the ADULT data set, which is available

through the online UCI Data Repository. The data set contains 48842 cases with 14 attributes which are split into 6 continuous and 8 nominal attributes. There are 7% missing values in the data.

In Statistica, we simply used a classification tree. The analysis took 3 minutes to complete.

In KXEN, the analysis took a few seconds. The model output gives $KR = 0.986$ and $KI = 0.809$. We show two of the most important output graphics in Figure 6.9 and 6.10.

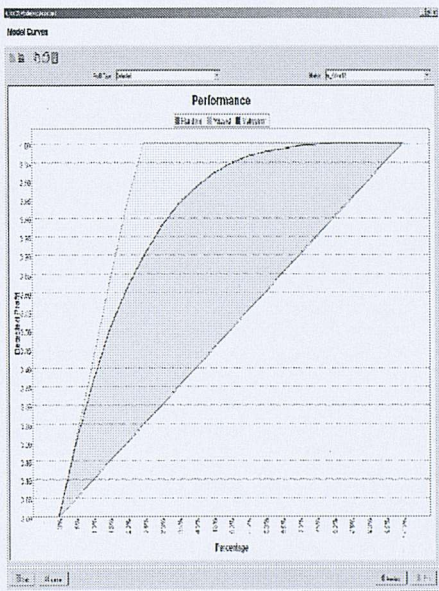


Figure 6.9: Model Curves in KXEN for Adult Data Set

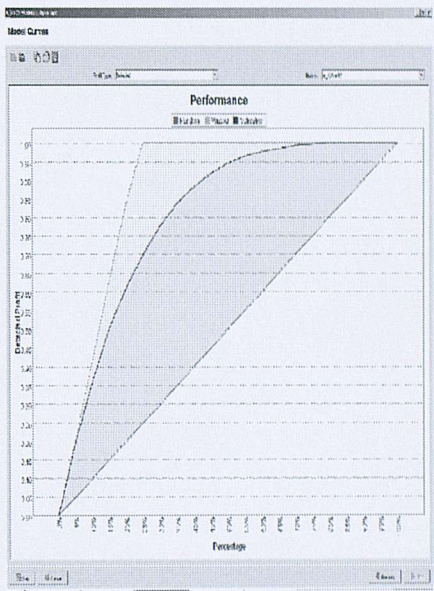


Figure 6.10: Variable Contribution in KXEN for Adult Data Set

Chapter 7

Peaker

In this chapter we discuss a pattern discovery tool called Peaker. Peaker is used to identify modes (or dense regions) in a data set.

Mode-hunting (bump-hunting) has attracted little attention over the years. The problem, in its basic formulation, is to identify a set of modes (areas of high probability density) in the data. This of course is closely linked to the problem of density estimation. However, as opposed to estimation of the underlying density, detection and verification¹ of a bump is more of a significance testing problem. The simplest way of detecting a bump is by computing the maxima of the density estimates. In this sense obtaining a good density estimate is critical to the detection of significant bumps (for literature in this context see Good & Gaskins (1980)). In many contexts, the problem we are faced with can be related to that of outlier detection. Outlier detection is well established in statistics and there are many methods available which have sound underlying statistical properties. Moreover, in computational sciences there have also been many advances. We will note here, however, that one of differences is that we are more concerned with the properties and behaviour of any patterns that we find rather than the process of looking for them. In particu-

¹These are known as pattern detection and pattern verification which we will describe later, discussions on these can be found in Hand & Bolton (2004).

lar, we look for the underlying mechanism that gives rise to a pattern rather than stopping at the detection and verification of an outlier. That said, there are many similarities and it is useful to keep this mind since many techniques have already been developed which can be useful as alternative approaches to the problem we try to solve.

In the data mining context bump-hunting is more concerned with looking for localized modes in the sample space rather than the exact estimation of the underlying density. The existing methods approach the problem by looking for subregions with higher densities than the whole sample space (see Friedman & Fisher (1999) and Barnett (1976)).

Peaker takes a different approach. It looks for data points (which we call peaks) at the location of which there is a higher estimated probability density than at neighbouring data points. In this respect it only looks for approximations of sample modes (points which are close to modes). Peaker has the advantage of being simple to compute, intuitively clear and easy to extend to higher-dimensional space². However, we are faced with a new set of problems. In the first instance, we need to identify modes by other means to computing maxima³. Secondly, verification of a mode takes on a different perspective. The work in this chapter follows the earlier results presented in Adams et al. (2001) and Bolton et al. (2004), and follows closely the work presented in Zhang & Hand (2005) and Hand & Zhang (2005).

7.1 Method

The basic Peaker algorithm is as follows: Given a data set $(x_i, i = 1, \dots, n)$ of some dimension d , we calculate the probability density estimates $\hat{f}(x_i)$ at each point x_i .

²Although it can be computationally expensive over a large data set with many observations.

³We have no direct means of evaluating a maximum, since we are not evaluating the density surface.

For each point x_i , we say that it is a peak with $M(x_i) = m$ iff $\hat{f}(x_i) > \hat{f}(x_j), \forall x_j \in N_m(x_i)$ but $\hat{f}(x_i) < \hat{f}(x^{(m+1)})$, where $N_m(x_i)$ is the set of the m nearest neighbours to x_i (defined by some distance function $d(x_i, x_j)$) and $x^{(m+1)}$ is the $(m+1)$ th nearest neighbour to x_i . Initially we call all points x_p , for which $M(x_p) > 0$, peaks. In the above, the distance function we will typically use is the Euclidean distance and the probability density estimates will be obtained via kernel methods.

We will now also give an algorithmic outline of Peaker:

1. Begin with the data set $(x_i, i = 1, \dots, n)$.
2. Calculate $\hat{f}(x_i)$ for all x_i and store resultant values in vector **FX**[n].
3. For each point x_i :
 - (a) Calculate distances from x_i to all other points and store in vector **DIST**[$n-1$].
 - (b) Sort **DIST**[$n-1$] in ascending order and store the resulting indices (of data points) in vector **ORDER**[$n-1$].
 - (c) Look through **ORDER**[$n-1$] and return $M(x_i)$.

Note that in the above algorithm, we would already obtain the inter-point distances in step (2), which we could store for later usage in step (3a). However, this is not possible as sample size becomes larger, therefore we have to calculate the distances again.

Peaker essentially estimates the pdf at each data point and performs nearest neighbour search on all data points, identifying those points which have a higher pdf estimate than their surrounding points (we will call these local peaks). Clearly peaks arise as a result of a local region of anomalously high concentration of data points. A local mode of the underlying true density will thus tend to lead to a

peak. We can thus regard peaks as approximations to local modes in the underlying true density. We say approximations since peaks a) derive from density estimates rather than the true underlying density, and b) are restricted to the locations of actual data points, which are unlikely to be in exactly the same position as the local pdf modes. Note that peaks can also arise, by random variation, even when there is no underlying local pdf mode. The essence of verification, discussed below, is to determine when peaks do reflect real underlying modes, or when they are merely chance occurrences. This random occurrence of peaks which do not reflect underlying modes can occur a) because of a chance grouping of data points, or b) because of an interaction between the estimated pdfs and the locations of the data points. We illustrate this last situation later in the chapter.

One attractive property of Peaker is its computational efficiency in high-dimensional spaces which follows from restricting the search to the actual data points only. This avoids the potentially huge search required (for example, in scan statistics) when moving a window over a high-dimensional space.

As a last note we would like to remark that Peaker does not serve the same function as a clustering tool. As mentioned, Peaker is well suited to detecting small local patterns⁴. Figure 7.1 shows one instance of a problem that cluster analysis would not be expected to perform well on, but Peaker is designed to detect.

7.2 Density Estimation and Bandwidth Selection

In section 7.1, we mentioned the use of kernel density estimation. In the paper by Adams et al. (2001), it was suggested that a simplified method could be used to

⁴On the other hand, clustering is generally used to find overall partitions of the data and in most instances will search for large obvious clusters rather than small local patterns.

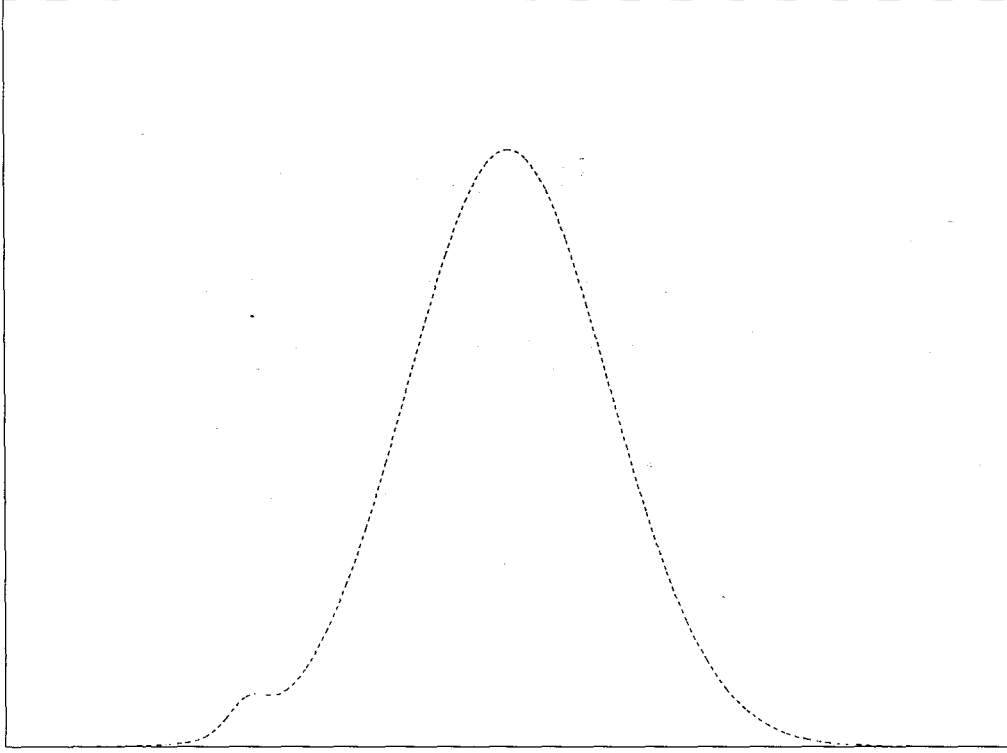


Figure 7.1: Example of a Small Local Peak

obtain the pdf estimates. This was given by initially calculating

$$\hat{g}(x) = \frac{1}{K} \sum_{y \in N_K(x)} d(x, y), \quad (7.1)$$

for some nearest neighborhood size K , and setting $\hat{f}(x)$ as inversely proportional to $\hat{g}(x)$. This is akin to using a uniform kernel with K being the bandwidth. The reason cited for its use was the speed of computation. We feel that it would be more appropriate to use a more generalized kernel as the penalty to computation is not severe (from the algorithmic outline in section 7.1, we can see that we do not store the distance ordering in step (1) and therefore would need to do an additional sort for the uniform kernel).

Throughout the rest of the chapter we will assume the use of a multivariate

Gaussian kernel, of the form

$$\hat{f}(x) = \frac{1}{nh_1 \dots h_d} \sum_{i=1}^n \prod_{j=1}^d K\left(\frac{x_j - x_{ij}}{h_j}\right), \quad (7.2)$$

where $K(y) = \exp(-\frac{1}{2}y^2)$. As we will see, this has certain advantages. Given the kernel function, we now need to choose appropriate values of the bandwidth h . There has been much literature on the selection of an optimal bandwidth, with different approaches given in Silverman (1986), Taylor (1989), Hall et al. (1991), Sheath & Jones (1991), Wand & Jones (1994), Stone (1984) and Scott & Terrell (1987)⁵. Sain et al. (1994) gave multivariate extensions of these results. These methods are all based on minimizing data-driven approximations to the integrated squared error (ISE) and the mean integrated squared error (MISE), given by

$$ISE(h) = \int_{R_d} (\hat{f}(x) - f(x))^2 dx, \quad (7.3)$$

$$MISE(h) = \int_{R_d} E(\hat{f}(x) - f(x))^2 dx. \quad (7.4)$$

Of course, such data-driven optimization is computationally expensive and can sometimes give only local minima. Sain et al. (1994) showed that under the constraint $h_0 = h_1 = \dots = h_d$, one can obtain an exact expression for the minimizer of the asymptotic mean integrated squared error (AMISE). For instance, under the assumption of the data being generated from a single Gaussian $N(0, \sigma \mathbf{I})$, we have $h_0 = [\frac{4\sigma}{n(d+2)}]^{\frac{1}{d+4}}$. This bypasses the optimization and we will make use of this result when appropriate.

Of course, while it is desirable from a theoretical point of view to use the optimal bandwidth, we must never forget that Peaker is a practical tool and as such it may

⁵For categorical data see Bowman (1980), Fienberg & Holland (1973), Brown & Rundell (1985) and Simonoff (1995).

be more appropriate in some cases to use a smaller bandwidth and thus better investigate local peaks, since the optimum bandwidth is a global optimum and it may be better to focus on detecting possible peaks. The Gaussian kernel also has the property that the number of modes is non-decreasing with increasing h . It may therefore be useful to vary the bandwidth. We will investigate this in a simulated study later.

7.3 Identification of Significant Peaks

This section describes the criteria by which we can detect peaks and the methods with which we can verify the significance of a peak once detected⁶. Moreover we should consider whether we are interested in detecting sharp peaks or peaks that occupy a large proportion of the density space (which we will refer to simply as large peaks). We can say that large peaks represent the more obvious structures within a data set, which would be picked up by standard modeling strategies, and that we are mainly concerned with only the small sharp peaks.

All the above considerations (as well as the choices of distance metrics and density estimation) lead to different variants of the Peaker algorithm. As such, there is no single best algorithm which can answer all the questions posed. Peaker is an exploratory tool, and any of its different variants might be useful in detecting interesting and useful local structures.

Before we move on, we must make an important distinction. Any data point can be one of the following:

1. A mode (a maximum) - one which is a maximum in the density surface. To decide whether a point is a mode one needs to evaluate the entire density

⁶To determine whether a detected peak represents a true anomaly of the underlying mechanism or whether it is a chance occurrence.

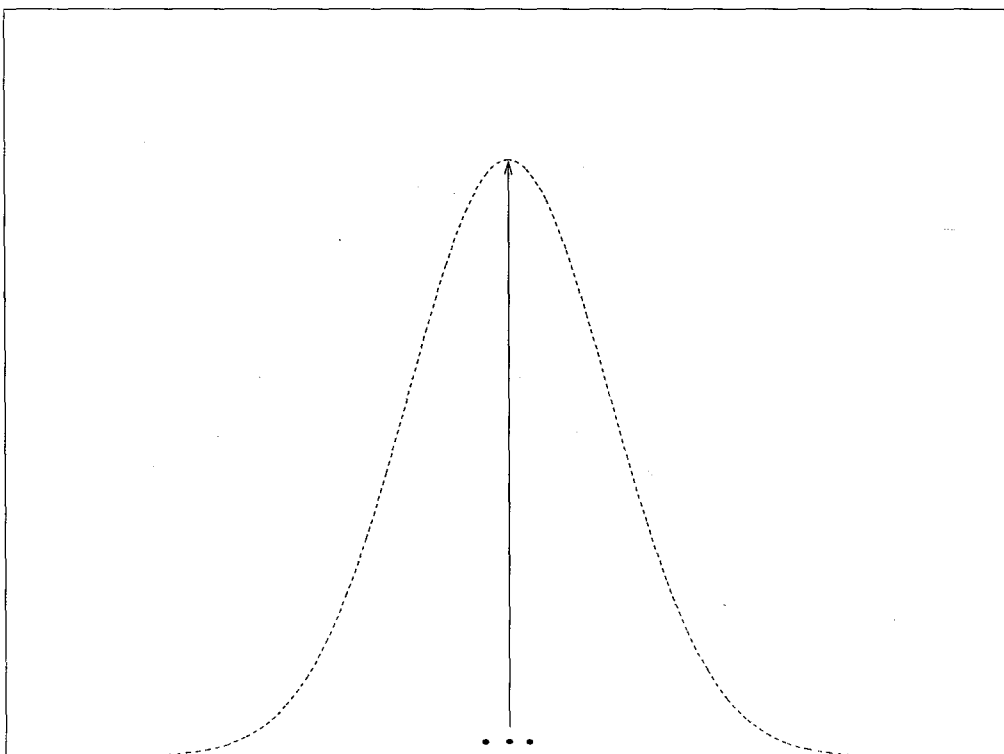


Figure 7.2: A Maximum

surface around the point⁷. An example is shown in Figure 7.2.

2. A maximum peak - a data point with higher pdf estimate than its nearest neighbours and which is also an approximation to a maximum. This is of course the essence of Peaker, since we only make use of density estimates at the data points and not the entire surface and so can only look for approximations. An example is shown in Figure 7.3.
3. A non-maximum peak - This is simply a point which has a larger pdf estimate than some neighbourhood of data points. Such a point does not need to approximate a maximum and arises due to a chance configuration of the neighbouring points (in one dimension, for instance, the nearest points with lower pdf estimates lying to one side of the peak). An example is shown in Figure 7.4.

⁷In high-dimensional space this is very difficult.

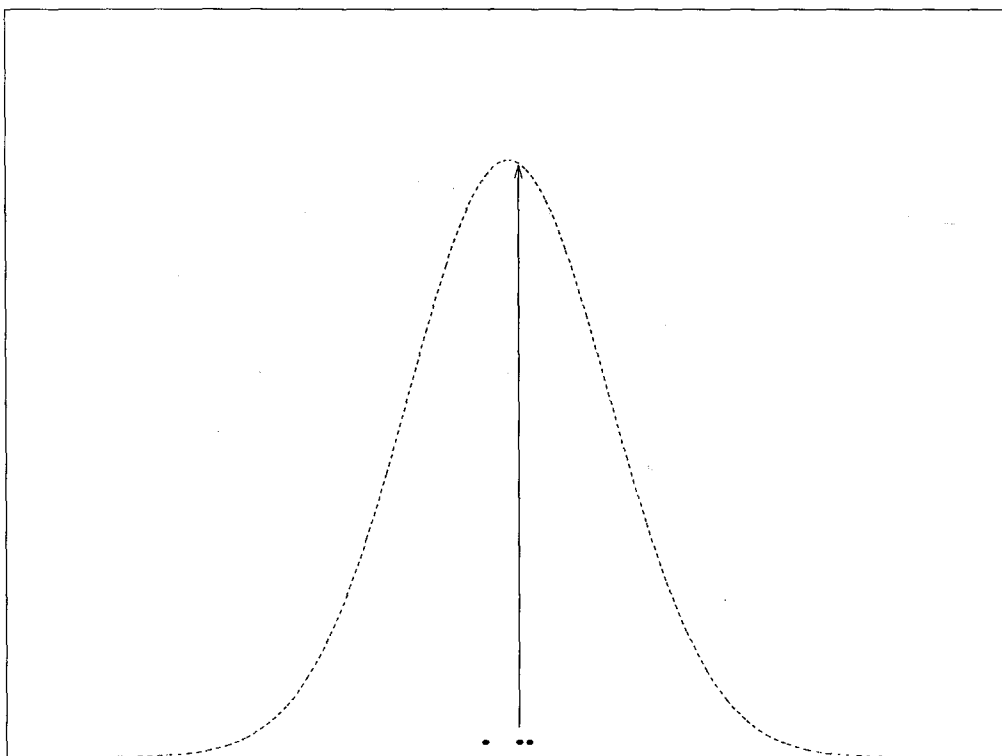


Figure 7.3: A Maximum Peak

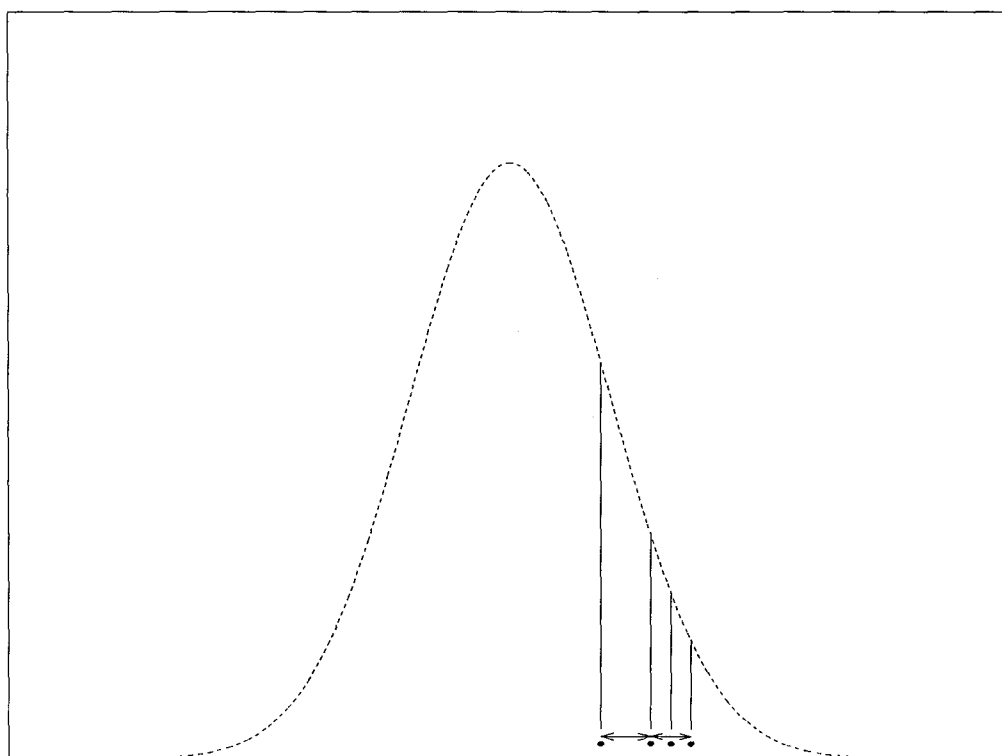


Figure 7.4: A Non-Maximum Peak

4. A non-peak - A data point which has lower pdf estimate than its nearest neighbour.

Intuitively one would like to detect maxima or approximations to maxima. This is rather difficult since it involves evaluation of the density surface. We will describe a method in the next section by which we can achieve this probabilistically.

7.3.1 Setting M

We have already mentioned the types of false peaks which can arise in the previous section. To make this more precise, we restrict our discussion to the univariate case⁸.

Let a set of points $(x_1, \dots, x_n)$ be generated from some underlying probability density f , with *cdf* F . For any three consecutive points x_{i-1}, x_i, x_{i+1} , with true *pdf* $f(x_{i-1}) > f(x_i) > f(x_{i+1})$, then x_i is a peak with $M(x_i) > 0$ iff $d(x_{i-1}, x_i) > d(x_i, x_{i+1})$. In this instance x_i is a false peak since it is not a local maxima. We can say then that x_i has some probability P_i of being a false peak where $P_i = P(d(x_{i-1}, x_i) > d(x_i, x_{i+1}))$. Over the whole sample we can average over P_i to give an overall probability R , which is the average probability that some point is a false peak. We can extend this further. For a set of consecutive points $x_{i-1}, x_i, \dots, x_{i+m}$ with $f(x_{i-1}) > f(x_i) > \dots > f(x_{i+m})$, x_i is a false peak with $M(x_i) = m$ and probability $R_m = R^m(1 - R)$. This gives us a way of setting a value M initially such that we retain only the set of peaks $x_i, M(x_i) > m$ which are significant in the sense of being actual bumps in the data.

We will now show that, as $n \rightarrow \infty$, $d((x_{i-1}, x_i) - d(x_i, x_{i+1})) \rightarrow 0$ by showing that $E(d(x_{i-1}, x_i) - d(x_i, x_{i+1})) \rightarrow 0$. To do this, we use order statistics and again assume the univariate case.

Given a set of points $(x_1, \dots, x_n)$ generated from some probability density f with

⁸Analogies to the multivariate case can also be made, but there are theoretical difficulties involved.

cdf F on $\mathfrak{R}$, the set of order statistics $y_1, \dots, y_n$ is obtained by arranging the x s in ascending order on $\mathfrak{R}$. The density function of the k th order statistic, Y_k , is given by

$$f_{Y_k}(y_k) = \frac{n!}{(k-1)!(n-k)!} F(y_k)^{k-1} f(y_k) (1 - F(y_k))^{n-k} \quad -\infty < y_k < \infty. \quad (7.5)$$

The joint density for Y_k and Y_{k-1} is given by

$$f_{Y_{k-1}, Y_k}(y_{k-1}, y_k) = \frac{n!}{(k-2)!(n-k)!} F(y_{k-1})^{k-2} f(y_{k-1}) f(y_k) (1 - F(y_k))^{n-k} \\ -\infty < y_{k-1} < y_k < \infty. \quad (7.6)$$

We can then go on to obtain the density function for $D_k = Y_k - Y_{k-1}$ ($2 \leq k \leq n$), given by

$$f_{D_k}(d_k) = \frac{n!}{(k-2)!(n-k)!} \int_{-\infty}^{\infty} F(x)^{k-2} f(x) f(x + d_k) (1 - F(d_k))^{n-k} dx, \quad (7.7) \\ 0 < d_k < \infty.$$

The joint density function for (D_k, D_{k+1}) is given by

$$f_{D_k, D_{k+1}}(d_k, d_{k+1}) = n! \int_{-\infty}^{\infty} \frac{F(x)^{k-2} (1 - F(x + d_k + d_{k+1}))^{n-k-1}}{(k-2)! (n-k-1)!} \\ f(x) f(x + d_k) f(x + d_k + d_{k+1}) dx, \quad (7.8) \\ 0 < d_k < \infty, 0 < d_{k+1} < \infty.$$

Unfortunately, as one can see, none of the above expressions are simple for most f . We can, however, use the following (Pyke (1965)): For $0 < u < v < 1$, suppose $s = F^{-1}(u)$ and $t = F^{-1}(v)$ are uniquely defined, then if $\frac{i}{n} \rightarrow u$ and $\frac{j}{n} \rightarrow v$

$$\lim_{n \rightarrow \infty} F_{nD_i, nD_j}(x, y) = (1 - \exp(-f(s)x))(1 - \exp(-f(t)y)). \quad (7.9)$$

The density function for (D_i, D_j) is then given by

$$f_{D_i, D_j}(d_i, d_j) = f(s)f(t)n^2 \exp(-f(s)nd_i) \exp(-f(t)nd_j), \quad (7.10)$$

$$0 < d_i < \infty, 0 < d_j < \infty.$$

We can then obtain the density for $Z_i = D_i - D_{i-1}$, given by

$$f_{Z_i}(z_i) = \frac{nf(t)f(s)}{f(s) + f(t)} \exp(nf(s)z_i), \quad d_i < d_{i-1}$$

$$f_{Z_i}(z_i) = \frac{nf(t)f(s)}{f(s) + f(t)} \exp(-nf(s)z_i), \quad d_i > d_{i-1}, \quad (7.11)$$

and Z_i has expectation zero.

Going back to the original problem, we can thus show that $E(d(x_{i-1}, x_i) - d(x_i, x_{i+1})) = 0$ as $n \rightarrow \infty$.

There are three points we need to consider with the above. Firstly, we have assumed that we know the true underlying distribution of the data. This is of course not true in practice and we only have the kernel density estimates. But we feel justified in using the above as an approximation since we only require some *ad hoc* way of setting M , especially so since we will be using optimal bandwidth selection in the kernel estimation and can be reasonably confident of obtaining smooth estimates. Secondly we have shown the above for the case of univariate data only. The idea of order statistics in higher-dimensions is much more complicated and there is no natural analogy of the univariate ordering. Barnett (1976) reviewed order statistics in multivariate data and the various proposed ideas of ordering, including M-ordering (marginal ordering) which orders the data in one of the dimensions and P-ordering (aggregate ordering) which derives a single value for each multivariate data point and orders the data accordingly. With these types of ordering it is not difficult to see that we can readily extend the ideas of this section. The last point is that we have

derived only a concrete method of setting M as $n \rightarrow \infty$. We can try to empirically evaluate R , with one approximation being the ratio $R_s = \frac{\text{No. of peaks}}{n}$. Here we have to be careful to eliminate the points which cannot be peaks⁹ from all data points. To do this we can simulate different sample sizes for some dimension and look for convergence towards some value $R_s = r$. This is the value at which the true R is $\frac{1}{2}$, from which we can derive the proportion of points which can never be peaks. Fortunately, we show below that, at reasonably low dimensions, we do not need a large number of data points for the limiting behavior to be true.

Figure 7.5 shows that R_s converges to some value at dimension 1 to 5, for data size below 30,000 and is in fact close to the convergent value for much smaller sample sizes (around 10,000). Therefore we can reasonably assume $R = \frac{1}{2}$ unless for low sample sizes.

7.3.2 Sharpness of a Peak

Detecting peaks, as potential modes of the underlying pdf, is one thing, but we need to be reasonably confident that they are genuine, and not merely due to random fluctuation. We can explore this using ideas of significance tests, and of the probability mass of a peak. We now describe some methods of verifying the significance of detected peaks.

A simple approach is to calculate the ratio of the pdf estimate at a peak to the average of the pdf estimates at the points in a surrounding neighbourhood. This is a measure of sharpness of the peak (which we call T_{sharp}). The general form is given by

$$T_{sharp} = \frac{N_s \hat{f}(x)}{\sum_{i=1}^{N_s} \hat{f}(x_i)}, \quad (7.12)$$

where N_s is the number of nearest neighbours from which we take the average. For

⁹In the univariate case the two end points are points which cannot be peaks.

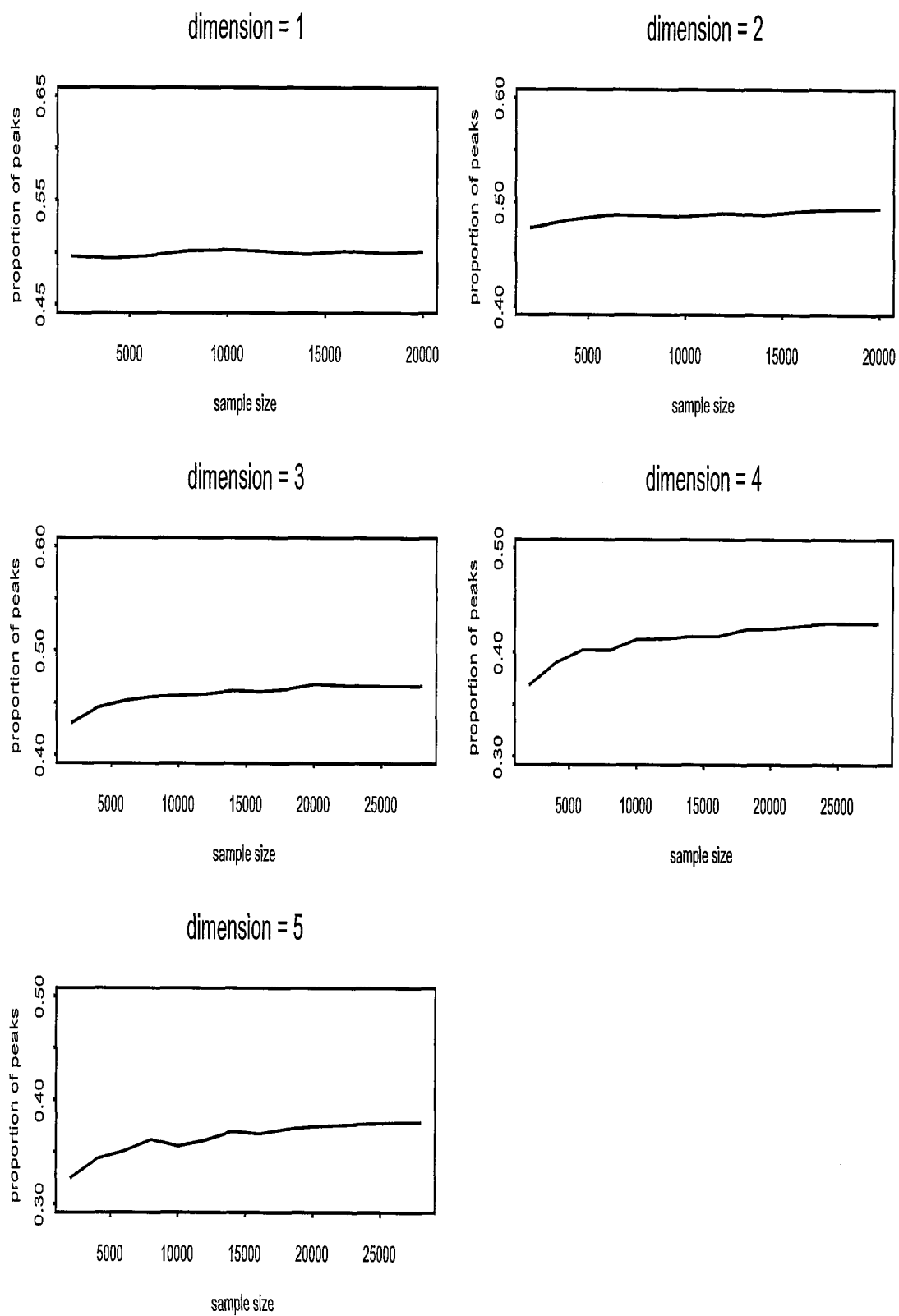


Figure 7.5: Simulation Results for Convergence

simplicity we can set $N_s = \frac{m}{2}$. Of course this is again rather arbitrary and moreover prone to distortion by specific arrangements of peaks. For instance, the sharpness of one peak can be very much affected by the shape of the neighbouring peaks (if a small sharp peak is close to a large flat peak, it is likely to have a smaller value of T_{sharp}).

There are more sophisticated methods of measuring the sharpness of peaks. Bolton et al. (2004) described two such tests¹⁰. The idea is to assume a null hypothesis in which the points are locally uniformly distributed in a d -dimensional hyper-sphere of radius R . Denoting the distance of each point to the centre as X , then the transformed variate $Y = (\frac{X}{R})^d$ is distributed as $U(0, 1)$. The two tests are based on this transformed variate Y .

The first test uses the test statistic

$$T_{lepto1} = \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(Y_i), \quad (7.13)$$

where Φ^{-1} is the inverse of the cumulative normal distribution. T_{lepto1} is distributed as a $N(0, \frac{1}{n})$ distribution. This test is based on the premise that a local dense region will have an abundance of short distances and a large negative value of T_{lepto1} .

The second test uses the r th order statistic $Y_{(r)} = (\frac{X_{(r)}}{R})^d$ and $Y_{(r)}$ follows a $Beta(r, n - r + 1)$. Here $Y_{(r)}$ will be smaller if there is a region of high density.

7.3.3 Probability Mass Contained in Peaks

An alternative to testing for the sharpness of a peak, is to derive a measure of the probability mass contained in a peak. This section follows the approaches contained in Silverman (1981), Good & Gaskins (1971), Good (1971) and Minnotte (1997), which are used to test for modes in the one-dimensional case. We assume the use of

¹⁰They refer to these tests as tests for leptokurtosis.

Gaussian kernel estimates¹¹. In Silverman (1981), to test the null hypothesis that the underlying density g has k modes, the test proposed uses the test statistic

$$h_{crit} = \inf(h; \hat{f}(x, h) \text{ has at most } k\text{-modes}). \quad (7.14)$$

Here h_{crit} is the smallest value at which the data remains k -modal. With the property of the Gaussian kernel, one can perform a binary search to calculate h_{crit} . To obtain the significance level $P(h_{crit} > h_0)$, Silverman (1981) proposed sampling for g from the distribution g if g is known or sampling from a bootstrap density g_0 obtained from the data if g is unknown.

Minnotte (1997) extended the above to investigate each mode in turn. For a single mode v_i , he defined $h_{test,i}$ as the smallest h at which the mode remains a single mode, and proposed the test statistic

$$M_i = \int_{u_{i-1}}^{u_{i+1}} [\hat{f}(x) - \max(\hat{f}(u_{i-1}), \hat{f}(u_{i+1}))] dx, \quad (7.15)$$

where u_{i-1} and u_{i+1} are the two anti-modes either side of the mode u_i . In effect M_i is the area of probability mass above the higher of the two surrounding anti-modes. To obtain a p -value for M_i , the approach of Silverman (1981) is followed and points are sampled from a representative density of the region with no modes (possibly a uniform density).

Higher-dimensional extensions to the above ideas are difficult to implement since it is computationally expensive to either a) look for anti-modes in high-dimensional space or b) obtain accurate significance levels by re-sampling.

For our case, we simply calculate the ratio between an estimate of the density in the local region, and the density that this region would have if the local region is

¹¹The Gaussian kernel has the property that the number of modes found in the data is non-decreasing with decreasing bandwidth h , which facilitates the testing procedures.

part of a larger region in which the density is uniform. Such a measure would have the generic form

$$T_{density} = \frac{\sum_{N_s} \hat{f}(x_i)/N_s}{\sum_M \hat{f}(x_j)/M}, \quad (7.16)$$

where $\sum_{N_s} \hat{f}(x_i)/N_s$ is an estimate of the density contained in the local region, defined as the region containing the nearest N_s points, and $\sum_M \hat{f}(x_j)/M$ is an estimate of the density over the larger region, containing the nearest M points. There is, of course, some arbitrariness about the choices of M and N_s .

7.3.4 Testing for Large Peaks

We base our approach on that of Minnotte (1997). In our case, we do not calculate h_{test} , since this is computationally expensive. Moreover, the use of h_{test} is based on the desire to maximize the power of the test, since the use of the smallest bandwidth minimizes the anti-modes and therefore maximizes the probability mass in the critical region. This is therefore a trade-off between power and computational complexity. We also require the anti-modes. This is quite difficult to find in high dimensions, particularly since we can not even make use of maximization. In this case, we make use of our version of the anti-mode, the anti-peak. An anti-peak is simply the opposite of a peak, so that u_i is an anti-peak with $A(u_i) = A$ iff $\hat{f}(u_i) < \hat{f}(u_j), \forall u_j \in N_a(u_i)$ but $\hat{f}(u_i) > \hat{f}(u^{(a+1)})$. In keeping with the last section, we define the nearest anti-peak to a peak as the closest anti-peak with some $A(x) > a$ or the largest $A(x) \leq a$. However, in this case, we do not wish to obtain the p -value by resampling from a representative local distribution with no modes, which would be very expensive computationally. Instead we wish to obtain a more general statistic which would be monotonically increasing with decreasing p -value. In the approach of Minnotte (1997), the p -value is the percentage of s samples gen-

erated from the local representative distribution which has more probability mass over the next higher anti-mode. Therefore it is clear that one obtains a low p -value if the probability mass constitutes a high proportion of all local probability mass.

To test the significance of a peak x_i , we use the test statistic

$$T = \frac{M \sum_A \hat{f}(x_i)}{A \sum_M \hat{f}(x_j)}, \quad (7.17)$$

where A is the number of points to the nearest anti-peak and M the number of points to the next nearest neighbour with higher estimated pdf. T has to be scaled carefully with respect to the proportion of the sample space occupied by the A nearest neighbours and M nearest neighbours.

7.3.5 Size of the Local Neighbourhood

Much of the discussion in the previous sections is based on rather arbitrary selections of the neighbourhood size that we use to calculate various measures. This includes the neighbourhood size N_s used in T_{sharp} and $T_{density}$ and the size of the uniform hyper-sphere used in the two tests by Bolton et al. (2004). The problem in all these cases is that these measures work well when the null model of local uniformity holds. For the various tests and ranking measures described so far, they only work well when the assumptions in the region of the test hold. When this is not the case, there are various ways for these measures to be distorted. For instance, if we have two identical peaks, one against a uniform background of higher pdf estimates than the other, then the measure T_{sharp} would return very different values for the two peaks. Similarly for the two tests contained in Bolton et al. (2004), it is imperative that the local hyper-sphere holds only one peak. In Minnotte (1997), such problems are resolved by only testing for uniformity of a mode in the local region bounded by the two nearest anti-modes. In our case, the neighbourhood M extends all the way

to the next peak.

In the case of Peaker we cannot evaluate the anti-modes since we only work with the data points themselves. The most intuitive analogy to the anti-mode is the anti-peak (*i.e.*, a point which has a lower pdf estimate than its nearest points). We can make use of this in deciding on the neighbourhood sizes in that we can restrict any testing procedure to within the neighbourhoods bounded by the anti-peaks. Clearly there are also problems associated with this, since anti-peaks may be false and so make the neighbourhood sizes again rather arbitrary. Additionally, the meaning of a bounded region by anti-peaks is unclear, since there is no clear analogy to the high-dimensional case. In one dimension it is clear where the valleys on either side of a mode are, but in higher dimensions it is unclear which particular valleys we should extend the local region to.

Having said this, we must not forget the exploratory nature of Peaker and as such it is important to experiment with differing neighbourhood sizes to look for otherwise undetected peaks. Therefore we are quite content in using rather arbitrary parameters in the Peaker algorithm knowing that one needs to explore these parameters to obtain peaks of different properties.

7.4 Simulated Study

The first example demonstrates the use of Peaker as a high-dimensional exploratory tool. The data set consists of 5000 data points in 10-dimensional space drawn from a mixture of background uniform and three Gaussian distributions. More data points in higher dimensional space can also be simulated. However, we simulate the above data set to reduce computational time while still giving a representative insight into the behaviour of Peaker. The Gaussian distributions have covariance matrices $0.025\mathbf{I}$, $0.025\mathbf{I}$, $0.1\mathbf{I}$ with mixing probabilities 0.0025, 0.0025, 0.02 respectively. The

background uniform is drawn in the hyper-box $(-2, 2), \dots, (-2, 2)$. For visualization we have reduced the data set down to two dimensions by plotting the points on the first two principal components.

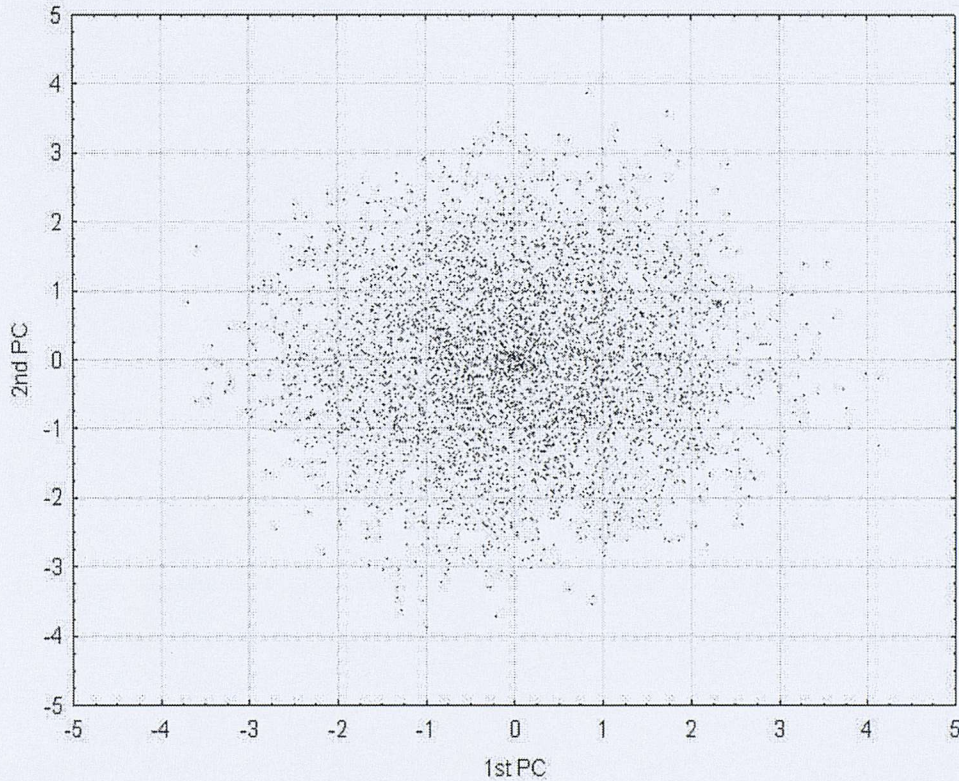


Figure 7.6: First Two Principal Components of 10-dimensional Data With Three Modes

The three Gaussian means are at $(0, 0)$, $(2.32, 0.81)$, $(-2.32, -0.81)$ in Figure 7.6. The peak at (0.0) is detectable visually¹². It is not nearly as clear if we simply plot any two random dimensions, but the other two peaks are not at all clear by visualization (since the mixing probabilities are so low). It is interesting to note that here many peaks are thrown up for $M = 200$ say, but the three Gaussians are clearly detected when one uses the measure T_{sharp} (in that they give much higher values of T_{sharp} than other peaks). Of course Peaker is well designed for detecting

¹²However, this is distorted by the fact that we are plotting the first two principal components, which pushes the points towards the center.

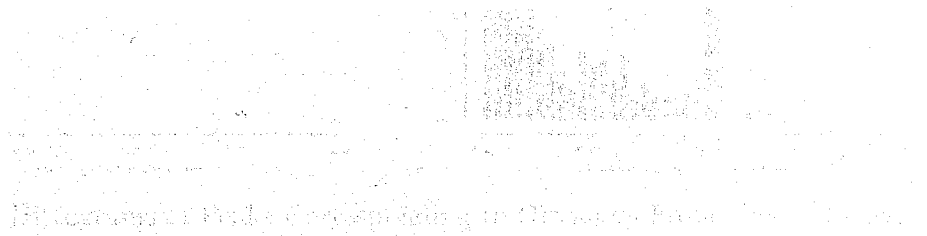
such peaks otherwise undetectable by eye (since at high dimensions points are sparse and a small but dense region would be easily detected by Peaker).

In the next example, we simulated from a mixture distribution of two independent multivariate Gaussian distributions (MVN_L, MVN_S). We change the variables of the simulated data as follows:

1. dimensionality, $d = 2, 5$.
2. sample size, $N = 1000, 5000$.
3. covariance matrix of MVN_L , with mean at $\mathbf{0}$, $\sigma_L = 0.25\mathbf{I}$, $\mathbf{I}$.
4. mixing probabilities for MVN_S , $p_{mix} = 0.025$, $p = 0.1$.
5. distance of the mean of MVN_S to the mean of MVN_L , $dist = \sigma_L, 0.5\sigma_L$.

For each combination of variables, 10 repeat samples are generated.

To investigate the quality of the peaks obtained, we set $M = 15$, and the distance of each peak to the nearest true Gaussian mean (in the mixture distribution of two Gaussian distributions) is calculated. We then plot histograms of the peaks which fall under each distance interval.



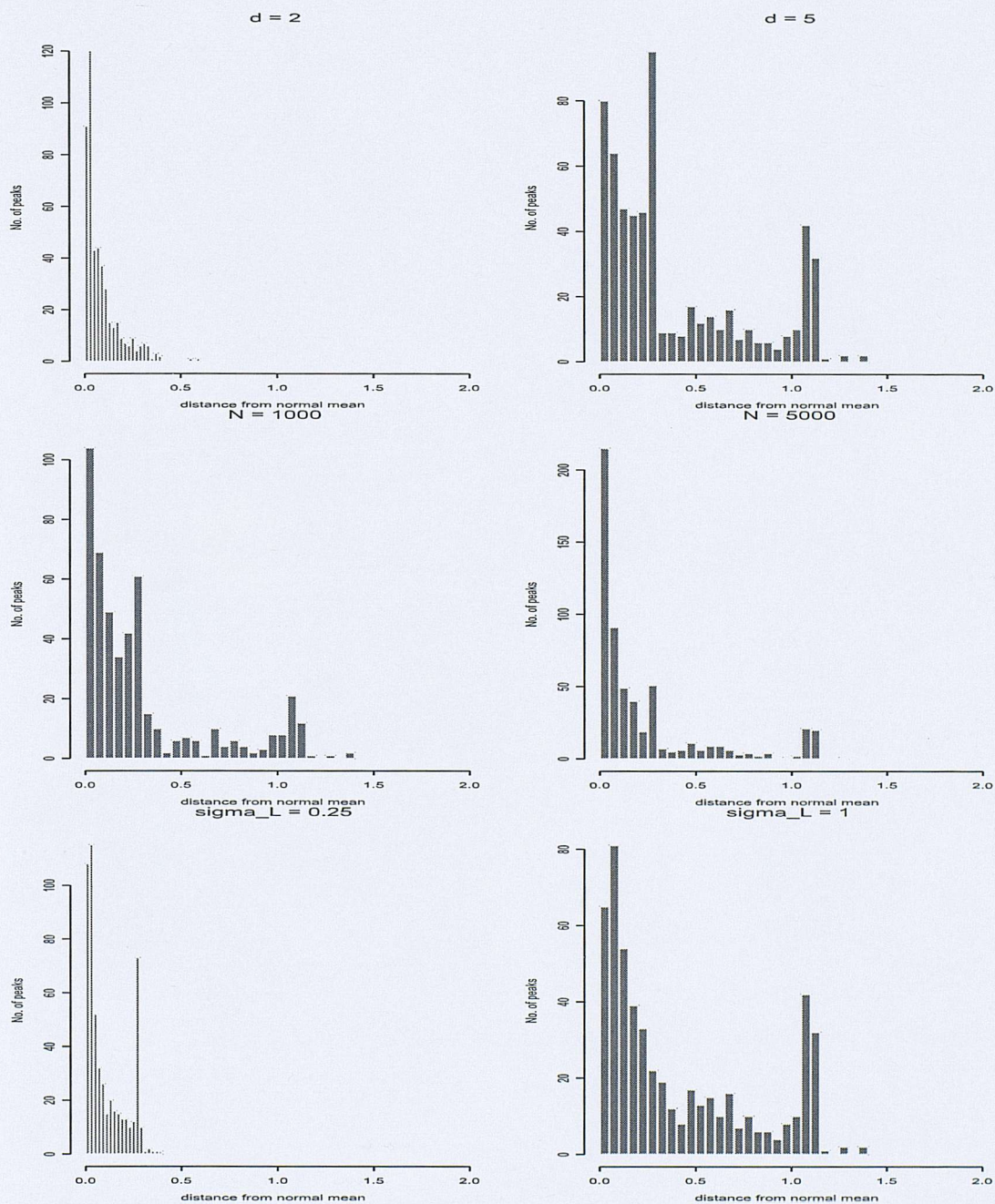


Figure 7.7: Histograms of Peaks Corresponding to Distances From True Gaussian Means

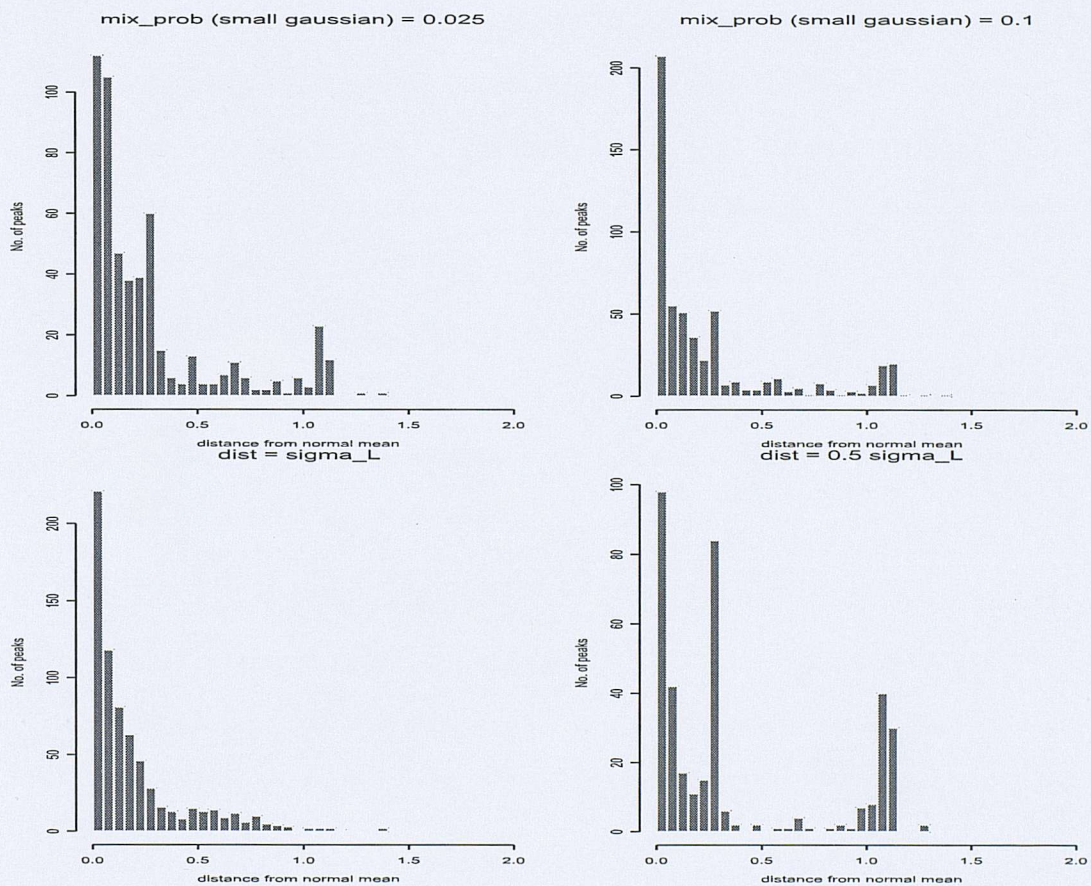


Figure 7.8: Histograms of Peaks Corresponding to Distances From True Gaussian Means (Continued)

We can see from Figure 7.7 and Figure 7.8 that we obtain peaks closer to the true Gaussian means with larger N , smaller d , smaller σ_L , larger p_{mix} and larger $dist$. These would be the expected results. If we look closer at the plots, we can make the following points:

1. Dimensionality is clearly the most important factor in determining how close each detected peak is to the true mean. It is perhaps a surprise that sample size does not affect the performance of Peaker as much, but this can be explained by the fact that an increase in dimensionality makes the data points far sparser than a decrease in sample size. In higher dimensions the tendency is that the difference of the nearest neighbour to the furthest neighbour becomes smaller and more difficult to distinguish. In this regard it is not surprising that increasing the dimensionality makes the results worse. Moreover, it is often the case that the nearest data point approximating a true peak can be quite far away from the true peak.
2. In terms of the two mixing probabilities, we can see that the performance of Peaker is quite similar. This suggests that Peaker is quite robust in detecting small local peaks with regards to the size and sharpness of the peak. This is a desirable feature of Peaker which sets it apart from other unsupervised pattern detection tools.
3. In terms of the distance between the two Gaussian means, we can see quite clearly a significant decrease in performance as the two means become closer. At $dist = 0.5\sigma_L$, there is a large number of peaks at around distance 0.25 and 1.1 from the true mean, which occurs when Peaker only detects one merged peak of the two true Gaussian means. However, this can be attributed more to the density estimation process than properties of Peaker. This effect can perhaps be reduced by using a smaller bandwidth.

4. Following on from the last point, the peaks at around distance 0.25 and 1.1 from the true mean only occur for $d = 5$. This suggests that 'Peaker can still detect distinct close peaks given a more densely populated space. Of course, this again is a property of the density estimation process since a large data set will give better estimates.

Chapter 8

Anomaly Detection Under the New Approach

We compare the signals obtained under a new¹ approach to that obtained using the traditional methods². Under the new approach, we use distance metrics obtained using the chemical properties of drugs and analytic properties of drugs and ADEs. We use Peaker to detect anomalously high concentrations of side effects under the effect of similarities among drugs.

8.1 Detecting Groups of Similar Drugs

We use a real data set to pick out interesting peaks amongst the drugs. The data set consists of 931 drugs. Chemical fingerprint data of the presence or absence of different molecular structures are available and the distance metric is the Jaccard coefficient. We run Peaker to detect dense regions of drugs which may have similar properties. We show the peaks obtained in reduced 2-dimensional space.

We indicate various peaks in Figure 8.1 which represent some interesting pat-

¹We will describe the new approach in this section.

²Traditional methods denote the odds ratio, the GPS *etc.*

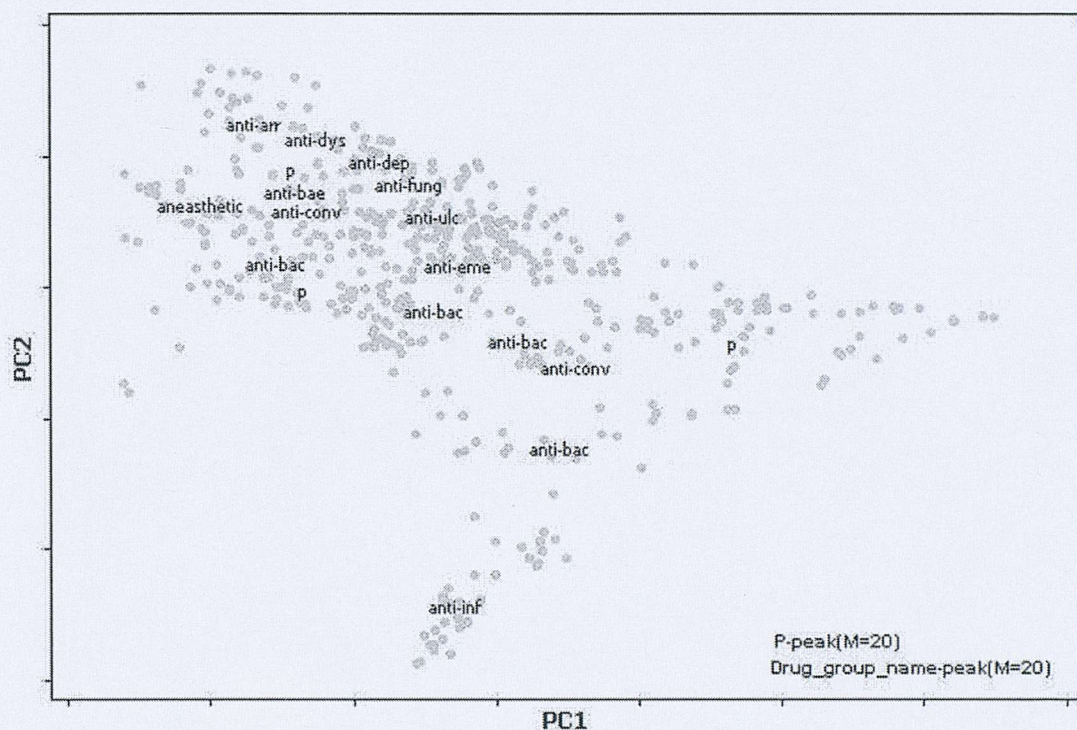


Figure 8.1: First Two Principal Components Drugs Data Set Showing Peaks of the Common Drugs

terns. All these peaks contain drugs almost exclusively of individual types. With a smooth bandwidth and $M = 20$, we discover mostly peaks of the more common drugs (including anti-inflammatory, antineoplastic, anti-bacterial, anti-convulsant, anesthetic, anti-depressant and analgesic drugs) as shown in Figure 8.1. It is surprising that we have managed to find patterns of such purity (in terms of homogeneous drugs).

Figure 8.2 shows the peaks found using a smaller bandwidth and $M = 5$. Some of the peaks found with $M = 5$ are not immediately obvious in similarity and it is only with further investigation that they are revealed as interestingly similar drugs. Certain drug peaks found have very low incidence rate in the data set (sometimes only 3 or 4 drugs). Overall there are not very many false peaks and most of the peaks found consist of interestingly similar groups of drugs, but this may be due to the closeness of chemical similarity to general behaviour of a drug. The general

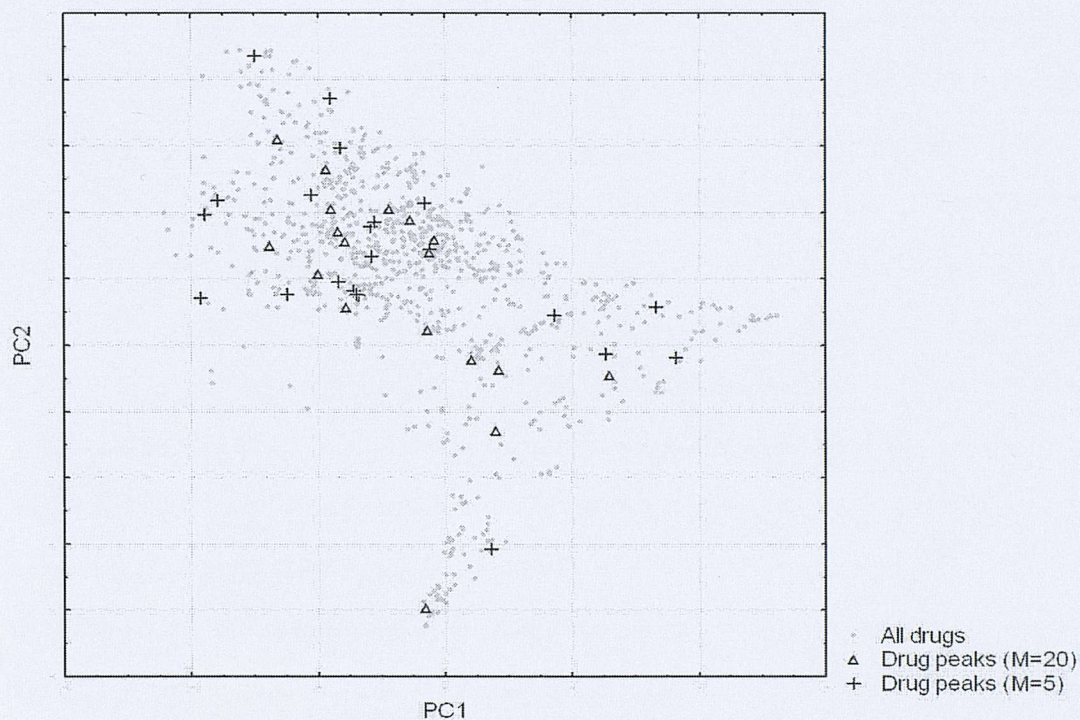


Figure 8.2: First Two Principal Components Drugs Data Set Showing All Peaks

trend seems to be that sharper peaks tend to show more homogeneous drug clusters. Indeed it does seem that Peaker can be applied to search for the rarer drugs that constitute interesting patterns or associations.

8.2 Comparison of Signals Using Simulated Data

8.2.1 Simulation

In this section we use a simulated example to compare the differences in signals obtained using the groups of similar drugs or ADEs.

The data is simulated using a Multinomial background model for the drugs and ADEs. We simulate 300 drugs each with a profile of 200 ADEs. We add particular signals into the background Multinomial model and compare the tendencies of detecting these signals. We keep track of all the artificial signals and detect them using

the Gamma-Poisson Shrinker and the new approach. The 300 drugs are separated into groups of 10 each such that each group comprises of drugs of similar background Multinomial rates. Subsequently we can compare signals when there are inherent similarities among the drugs. We also insert 20 artificial signals at various points of the data set. Of course, this is a rather exaggerated situation when compared to real data, since in real data we would not have similar background Multinomial rates even when there are similarities in behaviour between drugs, but we can have a reasonable idea of the comparisons from this type of simulation. From section 8.1, we have seen that meaningful and interesting groups of drugs can be picked up using their chemical properties which constitute the real world data situation. Lastly, for the simulation we have deliberately made the drug profiles so that they are non-sparse. We do this so that the distance metric used is no longer as important an issue which simplifies our task significantly.

We first use GPS to analyse the signals. Here we only use a two parameter model to smooth the signals for the sake of simplicity. Table 8.1 shows the 20 simulated signals and their strengths ordered by the GPS measure. We see that to a reasonable extent the signals are well picked up. The top signals comprise mostly of the true anomalies inserted. Moreover, keeping in mind that the inserted signals are themselves generated, so that some are not nearly as strong as others, we do see a good representative set of signals which are picked up by GPS.

We run Peaker on the simulated drugs to look for drug groupings. We show the drugs plotted on the reduced 2-dimensional graph. We simply use the Euclidean distance between profiles of drugs and ADEs. The interesting point to note here is that most of the simulated drug signals are somewhat separated from the main clusters of drugs. This is to be expected since the simulated signals are different to the other drugs of the same simulated group. Indeed Peaker confirms this result.

Drug index	Order of strength of signal
31742	1
234	2
27124	3
49109	6
47801	7
44699	8
100	13
16245	15
37853	33
51612	42
55123	163
38968	194
41670	211
53117	1195
59144	1288
1021	1485
57138	1502
43984	8147
42771	10265
28136	19510

Table 8.1: Strength of Simulated Signals With GPS

Here we simply use the quick method of estimating density (as described in chapter 7) which estimates the density at each point as a weighted sum of the inverse of distances to the neighbouring points. Depending on the weights used in the density estimate, Peaker detects around 30 peaks with $M \leq 9$ and these comprise the 30 groups of simulated drugs. It is interesting to note that none of the 20 simulated drug signals are at each one of these peaks, although some do appear at neighbouring points to these peaks. Indeed, the majority of the simulated drug signals only show as peaks with $M = 1$.

The above analysis shows that a possible way to make use of the groups of drugs is to merge similar groups together. Since the simulated signals will be likely to be in individual groups by themselves, this should enable us to perform signal detection of the simulated signals against a background of merged drugs which can give us

better background rates and smooth out the random variation. To merge the drugs, we can potentially use Peaker, but this will lead to some overlap as well as having groups which can comprise of drugs which should not be in the group (without due care) since the drugs on the boundary of the neighbourhood of a peak may not necessarily belong to the same group as other drugs in the same neighbourhood. On the other hand, it is not ideal to use some clustering technique, since many of the signals are likely to be included in the large clusters if these signals are not strong enough. Indeed this is the case when K-means clustering is used. Here only 25 clusters are produced and most of the simulated signals are incorporated into the larger clusters.

We will perform two separate analyses using Peaker. In the first instance, we will use Peaker to pick out the groups of similar drugs, with due care to avoid overlap. We will restrict each group that constitutes a peak to the 7 drugs which are closest to the peak, thus avoiding some of the problem posed when disparate drugs are included in a group. This gives us 196 drugs into 28 distinct groups, we will treat the other drugs as individual groups. In fact, we have reduced our original 300 drugs into 132 groups and we hope that the combined drugs will give us better baselines. After grouping by Peaker, we see that, given our criteria of keeping only the 7 closest drugs to a peak in a group, in fact none of the simulated drug signals are included in the 28 groups.

Table 8.2 shows the 20 simulated signals and their strengths ordered by the GPS measure after reduction of the drugs into groups by Peaker. Bearing in mind that the number of drugs has been reduced so that there are less signals overall, there is a vast improvement in terms of detection against all other background signals (although it is somewhat better in terms of ordering even given the reduced number of drugs). But the important point to keep in mind is that we have been able to vastly reduce the number of all background signals whilst still being able to detect the artificial

true signals in a somewhat better way (of course we would have expected this since none of the simulated signals are included in the large drug groups, as mentioned in the above paragraph). This in itself is enough to justify using the new approach since we have far less false positives to deal with.

Drug index	Order of strength of signal
31742	1
234	2
27124	3
49109	5
47801	6
44699	7
100	9
16245	10
37853	20
51612	24
55123	53
38968	68
41670	88
53117	392
59144	439
1021	519
57138	528
43984	1283
42771	2247
28136	5302

Table 8.2: Strength of Simulated Signals After Reducing Size of Table With Peaker

For the second analysis, we will include all drugs in groups as detected by Peaker and try to look for signals (using GPS) within each of the groups. Table 8.3 shows the 20 simulated signals and their strengths ordered within each group of drugs as picked up by Peaker. Here we see a dramatic change in terms of the strengths of the signals. Most of the simulated signals are now very well picked to the extent that 14 of them now occupy positions 1 or 2 in the within-group ordering. In a way, this is not as surprising as one may assume, since within each group the baseline

estimates are much better than can be obtained from all drugs. As such, it is much easier to pick up the true artificial signals.

Drug index	Order of strength of signal
31742	2
234	2
27124	1
49109	1
47801	10
44699	1
100	8
16245	1
37853	1
51612	108
55123	52
38968	5
41670	1
53117	1
59144	1
1021	1
57138	1
43984	22
42771	2
28136	1

Table 8.3: Strength of Simulated Signals Within Each Group Picked by Peaker

8.2.2 Discussion

The simulated examples in section 8.2.1 show the value of making use of the similarities among drugs and ADEs when attempting to detect signals. We have used Peaker initially to look for meaningful groups of drugs and subsequently performed analyses on the groups rather than the whole data set. We have seen that reducing the size of the overall data set by merging drugs within each group enables us to eradicate many false positive signals while not losing the true signals, whilst analysing each group individually enables us to much better detect the true signals.

A combination of these two approaches can also be utilised for potentially better combined results.

Although the results do seem encouraging, it must be said that there are several points to consider. Firstly this simulation has been made to generate a non-sparse data set, which is contradictory to the real world scenario. With non-sparse data sets, it is much easier to employ meaningful distance metrics between profiles of drugs and ADEs and consequently to obtain meaningful groups. With sparse data sets, however, the choice of distance metric is critical and it is difficult to devise overall satisfactory metrics. Secondly we have simulated the data so that there are inherent similarities between drugs in terms of response to ADEs. This is by no means the case if we were to employ properties (such as chemical) in the real world scenario. However, it is to be hoped that such properties can be reasonable informative in terms of drug or ADE responses.

Along with the distance matrices obtained analytically in chapter 5, we have other distance matrices obtained from chemical and biological information. It is therefore desirable to incorporate all the information into our distance. We can attempt to combine these distances. With the distances obtained we can subsequently perform pattern search using Peaker.

8.3 Combining Distances

There are two separate cases in terms of combining distances:

1. Given two distances obtained under two distance measures on the same variables. This situation arises when we have two distances (*e.g.*, analytical and chemical) calculated between $drug_1$ and $drug_2$, and we need to calculate an overall distance. Given $d_1(x_{[1]}, y_{[1]})$ and $d_2(x_{[2]}, y_{[2]})$ (where $x_{[1]}$ and $x_{[2]}$ denote different attribute vectors measured on the variable x), the question is how we

can combine d_1 and d_2 to obtain

$$d(x, y | d_1, d_2).$$

One way is to combine the attribute vectors so that the distance is given by

$$d([x_{[1]}, x_{[2]}], [y_{[1]}, y_{[2]}]),$$

where $[x_{[1]}, x_{[2]}]$ denotes a combined vector of $x_{[1]}$ and $x_{[2]}$.

But in this case we have to use a single distance measure on the combined attribute vector, which may well not be appropriate.

Alternatively we can put weights on the separate distances, such that

$$d = \alpha d_1 + \beta d_2,$$

and alter the weights α and β in the pattern search process to look for anomalies under different weights. This approach, while in principle sensible, does put a lot of strain on the pattern search and requires careful interpretation of the anomalies found.

2. The second issue concerns combining distances obtained under the same distance measure on different variables. This situation arises when we have distances between two drugs and between two ADEs, and we require a distance between two (drug, ADE) combinations. If we have $d(x_1, x_2)$, $d(y_1, y_2)$, we look for a combined distance

$$d((x_1, y_1), (x_2, y_2)).$$

This is in principle easier, since we can now simply combine the attribute

vectors. The only problem we need to take care with is the scaling of the combined vector so that we give equal weights to the two original vectors.

8.4 Anomaly Detection

We have three sets of distances on different variables that we can perform anomaly detection on. One between drugs, one between ADEs and one between combinations of drug and ADE. The next step is to use Peaker to detect anomalies in the data, with the final objective being to pick out unusually high counts of a (drug, ADE) combination.

This particular approach to anomaly detection is based on three ideas:

1. We do not want to assume independence in the contingency table as we would like to look for signals around a more localised setting, whereby we look for drugs which behave anomalously among similar drugs (similarly for ADEs and (drug, ADE) combinations).
2. We minimize the effect of smoothing (with global smoothing parameters) and only base signals on localised behaviour. Of course, in analytically obtaining a distance matrix between drugs or ADEs, we have to use a degree of smoothing. Moreover, in applying Peaker, we have smoothing in the kernel-estimates. But these are very different from actually applying global smoothing parameters to any particular cell-count. In particular, we try to lessen the effect of having a few global smoothing parameters to smooth the measure of association and control for all the variation in the data.
3. We take each drug and use the ADE counts to detect high anomalous counts at each drug. As a baseline at each drug we will use the smoothed counts estimated from the independence model. Aggregating the two counts will

show us the anomalously high side-effect regions.

There are some problems associated with this approach. In the first instance, it may seem that to accurately detect anomalies, we need to obtain accurate distances in the first place. If we obtain only approximate distances, then it may seem the case that localised anomalies are meaningless. However, in terms of applying Peaker and thus calculating kernel estimates of a particular data point (in this case a drug for instance), we are only giving more weight to localised points around the data point. This, regardless of the degree of meaningfulness of the distance measure (assuming that the distances do actually have some significance and not completely unrelated to the proximity in behaviour of two drugs say), is surely better than assuming complete independence and thereby assigning equal weight to all points in the space. The second point is that without the smoothing of the cell estimates, we would detect anomalies which have small cell counts (akin to using the odds ratio). In this instance, we can use the same argument as before, which is that we need to look at all anomalies regardless of size and sharpness of the anomaly (indeed we can hope to detect the rare events better in this approach since we are detecting them in regions where they are more likely to occur).

8.5 Example

We provide an illustrative example using a real data set in the section. The data comes from a real data set provided by GlaxoSmithKline which contains 931 drugs crossed with 245 ADEs. Also provided with the contingency table is a set of chemical information comprising of the chemical fingerprints of the drugs. Jaccard coefficients are taken of the chemical information as a measure of the dissimilarities between the drugs and a distance matrix is provided. To illustrate our approach using the profiling method we analyse the data in the following steps:

- Using the chemical data, find all subsets of drugs which display some inherent similarities. Note that this example uses the same set of drugs as in section 8.1 so that we can use the peaks that we have already detected in that section.
- Using the profiling methods, find all subsets of drugs and ADEs which display some inherent similarities. We use the Bayes Factor only as the distance measure in this example and we look for instances whereby the groups are sensibly separated by some known similarities within the groups.
- For each group of drugs, we take the ADEs as a set of counts from which we would look for peaks, taking into account the similarities between drugs. To do this, we look for high counts of ADEs for drugs that comprise each peak. A baseline for each such drug-ADE combination can be generated using the GPS and we compare this baseline with the ADE counts. In particular, we have the following three stage process:
 1. Given the groups of drugs detected using Peaker, use the ADEs as a profile for each drug.
 2. Compare the counts of the ADEs to the baseline obtained using the GPS.
 3. Look for anomalously high counts of ADEs such that within each group it is much higher than the baseline.

Note that in the above approach it is not the case that we will end up with the drugs which provide high counts of ADEs as signals (drugs that produce nausea for instance). This is because drugs with high counts of a particular ADE will not necessarily constitute peaks, whereas we are specifically looking for high counts of a particular drug as compared to other drugs within the peaks detected.

Table 8.4 to 8.6 show the signals we have generated using the approach outlined above. The odds ratio remains variable of course, this is reflected by the high

number of low counts that constitute the signals for the odds ratio. For the combined distance approach, shown in Table 8.6, we notice an interesting feature of the signals, which is that they are somewhat more stable than signals generated by the odds ratio but do not have such high counts as the smoothing method (GPS) produced in Table 8.5. In this sense we can say that these signals benefit from the smoothing being applied to neighbourhoods of similar drugs and ADEs. Moreover, we notice that some of the signals themselves derive from each other in the sense that they are neighbouring peaks within regions of similarities. In a wider sense, we can say that deriving signals in this way has allowed us to detect groups of signals and to be able to analyse these within the groups and between the groups. This is clearly a much more sensible way to calculate the signals than with an independence assumption.

We now look more closely at Table 8.6. Note first of all that the column signal denotes the GPS baseline measure which is given to the drug-ADE combination. The drug-ADE combination itself is detected by running Peaker on each of the side-effects against the drugs. We have merely given some of the more obvious peaks in Table 8.6 and there is no particular ordering in the table. We will go through the first few signals to give an idea of the peaks detected. The first signal comprises a high count in an anti-bacterial drug (as detected by Peaker in section 8.1). The signal is detected by comparing the count of the ADE with the baseline GPS of the ADE. We see that the GPS measure is by no means high compared to the highest GPS measures in Table 8.5. This is partly due to the fact that the signals are aggregated and partly due to the fact that the drugs have been grouped. The second signal in the table is one that comprises of the group anti-inflammatory drugs. Again here we note the GPS measure is very low. The third signal is that of anti-bacterial drugs again, but this time in a different peak to the first.

More importantly, the new approach has allowed us to extract signals (combined or otherwise) which we were not able to before. In particular, the combined signals

in general are quite reflective of the behaviour that drugs in each group tend to exhibit individually.

Drug index	ADE index	Count	Odds Ratio
13	156	3	10275
475	54	1	10264
115	44	1	9231
115	23	1	9224
235	512	2	9212
532	23	2	8132
251	143	1	8131
545	232	3	8112
565	113	2	8110
567	62	2	8065
123	15	3	8064
877	34	2	7764
231	54	4	7157
876	15	5	6446
126	54	3	6245
543	52	2	5544
23	20	2	5043
89	27	4	4634
165	58	1	4233
171	87	2	4133

Table 8.4: Highest 20 Odds Ratios

Drug index	ADE index	Count	GPS Signal
263	226	32	201
275	54	54	154
165	72	23	151
753	41	56	143
841	12	223	141
62	72	12	130
2	17	11	123
72	156	65	112
74	113	42	112
85	237	32	105
73	164	13	103
157	186	56	99
853	219	12	96
425	12	32	92
523	19	8	91
682	16	13	91
25	202	56	89
166	178	28	89
245	18	16	88
467	62	27	87

Table 8.5: Highest 20 GPS Signals

Drug index	ADE index	Count	GPS Signal
163	192	29	35
345	42	45	12
62	92	8	72
35	103	29	28
247	104	95	51
922	182	4	46
36	52	67	24
328	94	49	17
467	62	27	87
103	125	59	8
57	204	29	48
764	205	93	26
92	20	190	34
823	49	5	13
62	72	12	130
49	98	59	11
204	184	20	52
694	127	24	28
195	18	75	29
954	202	12	41

Table 8.6: 20 Signals (Peaks) Detected Using New Approach (Using GPS as the Baseline Measure)

8.6 Conclusion

In this chapter we have applied the methods arrived at in the previous chapters to a set of real data. The main result shows that the new method is indeed able to derive different sets of signals to the older methods and that there is real scope to expand on this new approach. On the other hand, it is possible that the older methods can be employed alongside the new approach to achieve useful results. In a sense we are relying on the older methods to detect the actual signals, but within a more structured data set. This is superior in many ways and we have illustrated the new signals which we are able to detect. It also is a great help that we are able to use chemical information in this context since this greatly enhances the way that the profiling can be done. Bearing in mind the fact that it is very difficult to be able to profile appropriately using just the inherent drug and ADE counts, chemical properties is another way to be able to remedy this problem.

Chapter 9

Conclusion and Future Research

We have presented in this thesis the methods and ideas behind the important area of Pharmacovigilance. In this particular aspect, we have carried out analyses with regard to the detection and verification of anomalous signals which arise due to the adverse reactions that arise as a consequence of some drugs. It is imperative that such reactions can be detected quickly and effectively dealt with. Whether this can be achieved using the current methodology is a point of debate and we hope to have illustrated the many shortcomings that exist in the field today.

In a sense, the remarks made in chapter 1 are particularly apt when considering the remainder of the thesis. We encounter the difficulties of data quality more than anything else in this area. When the data present difficulties, it is sometimes rather pointless to delve into deep and sophisticated methods to solve such problems. The point is that one must first gain a thorough understanding of the data and the inherent behaviour before one can apply automated procedures. Perhaps this is precisely the reason why so many methods perform poorly and in practice are similar when used in this context (chapter 3 contains a detailed comparison so we will not explain further here). In defense of these methods, it might be said that nothing really can be achieved without the basis of using ratio measures as the central ideas

of such methods, since in most instances companies are interested in concrete results rather than a willingness to explore.

Furthermore, we illustrate the alternatives which are possible. The central theme here is very simple, to utilise all information about the data before making a concrete decision on the type of analyses that we should perform. The extra information in our data comes in the shape of both chemical information and the inherent structures of the data. The idea is that, by using this information, we can establish the basis for measuring similarities among drugs and ADEs. These similarities will then further allow analyses based on the groups within the data. This is a significant improvement on the independence structure of the older models, which is severely restrictive on the types of analyses that can be performed. To obtain such similarities, we introduce the important concept of a suitable distance metric for sparse and high dimensional data. In particular we stress the importance attached to the ideas of distortive combinations of count data. Establishing that the Bayes Factor is a viable measure under such conditions, it becomes clear that this would allow us to establish a sensible and workable profiling system.

To explore the data structures we have obtained, we introduce an important data exploratory tool called Peaker which helps to facilitate our analyses. Peaker is specifically designed to detect anomalously high concentrations of data points in high dimensional space. More to the point, it is designed to detect small localised regions of such high concentrations. In particular, Peaker only evaluates the density at each data point and in so doing bypasses the problem of high dimensions in which it would be extremely difficult to estimate the density surface. Being exploratory, however, also comes with difficulties when an accurate evaluation is needed in order to assess the significance of the detected peaks. In this regard we have introduced several methods of verification which can be applied. We emphasise exploratory here because that is precisely the philosophy we must take when applying Peaker.

We cannot expect everything to be answered by a simple press of a button, but rather we need to tweak the various parameters to obtain an in-depth knowledge of the structures within the data.

9.1 Data Mining

It has often been said that data mining is a discipline which by itself has no underlying methodologies. This, however, in no way detracts from its importance as an area that contributes fully to the many facets of science. In Hand et al. (2001), data mining is described as the “search for anomalous local patterns within large databases, which are both interesting and useful to the user”. The point is indeed to look for patterns which are of interest. This is a very apt description in the sense that within the real world environment, it is much more interesting and productive to look for unexpected patterns rather than to partition the data in some way. Large partitions of the data generally reveal only obvious facts about the data which are well known already. On the other hand, it is precisely in the case of small local anomalies that one might discover interesting patterns that reveal insight into particular aspects of the data. Hand et al. (2001) furthermore gives the following representation of data

$$data = structural\ part + patterns + random\ part.$$

Note that this is in contrast to the otherwise held view in traditional statistics that only the structural and random parts are considered, or in classical language, structure and noise. We note that the addition of the patterns component has certainly changed the analysis somewhat since we would now be concerned with the detection and evaluation of these patterns as well as the structure and noise analysed in traditional statistics. Moreover the methods needed to analyse these

patterns need to be developed which can cope with the task of looking for patterns in often large and high dimensional data sets. This then, is perhaps one of the fundamental ideas of data mining, that of searching for patterns in large data sets. The methods developed in this thesis is one particular instrument in aid of searching for patterns, there are many more. In particular, methods such as that described in Friedman & Fisher (1999) and various clustering procedures spring to mind. These methods are developed to combat the many issues arising from the data which have perhaps never been the focus of traditional statistical methods (which are generally concerned with statistical soundness rather than practical usefulness). Moreover, in the age of computer analysis, we are able to perform tasks on data sets which were not possible before. This in turn has led to a case of “if we are willing to explore then we can find some means to analyse even the very massive of data sets”.

The explosion in interest in data mining in recent years is a testament to the continuing growth of this area and the reliance many sectors now place upon such techniques. The underlying principle of most data mining tasks is the search for structures within large datasets. What sets this apart from many other more traditional techniques such as cluster analysis and classification is the interest in more localised structures which are otherwise difficult to detect when they are embedded in huge datasets. This is applicable in various areas of society and is indeed one of the most sought after advances in statistics today, driven by the even expanding databases which companies are now collecting. These applications include areas such as fraud detection (typically in large bank transaction databases), market analysis (an example would be customer behaviour in large supermarket transaction databases). Of course, another aspect of this wide-ranging capacity of data mining is the number of subject areas that it is applicable, including the physical sciences, the social sciences (such as psychology) and finance. The appreciation is that as data becomes more widely available it is also imperative to develop techniques which can

handle large quantities of data. Of course, with large datasets there is a tendency to develop methods which are rather *ad hoc* and perhaps do not possess the solid statistical foundations of the classical techniques. Indeed there is a general lack for unified methods and a formal approach to areas such as pattern discovery (in contrast to statistical modelling say) which is attributable to the area still being relative young and unexplored. This, however, is by no means disadvantageous if we can treat the problem with the right approach since any exploratory insight into the data can only be helpful towards the final goal. We have already mentioned one tool that we have developed in this thesis, which is called Peaker. The spirit of using this tool is precisely an exploratory approach. However, we hope to have demonstrated that one can fine-tune the many aspects of such a tool to obtain desired results.

The particular aspect of data mining that we are concerned in this thesis is the area of pattern discovery, more specifically that of unsupervised pattern discovery. In Hand & Bolton (2004), unsupervised pattern discovery is described as “characterizing and finding sets of values of the recorded variables which correspond to real features of the data generating process (they are not merely chance relationships) which are too small or local ... to form part of a model constructed to describe the data, and which are, in some sense, interesting or valuable”. A pattern itself is defined as “a local structure that generates data with an anomalously high density compared with that expected under the (global) baseline model”. Pattern discovery moreover is split into the two parts of pattern detection and pattern verification. In this thesis we have given examples of both in chapter 7. Pattern detection is the search for useful and interesting patterns while pattern verification is the testing thereafter carried out to ensure the veracity of the detected patterns. The detection of patterns usually involves searching for anomalously high regions of density in the data. In this sense anomalously high is relative to the baseline model.

Alongside with data mining techniques there is also the important contribution

that software packages have made to the area of data mining. The general consensus is that today the computational power of most packages allows the user means to apply sophisticated techniques to large data sets, even without any particularly in-depth knowledge about the tasks they are performing. We saw the power of such tools in chapter 6. In particular we illustrated the many facets to a data mining package that is required to compete in today's market. All these various aspects are crucial to the overall composition of a package.

9.2 Summary of Thesis

We made a case for the inefficiencies of many of the current methods being deployed in this area and suggested schematically ideas which could make improvements based on the use of inherent similarities among drugs and ADEs. At this point it would be as well to summarise the findings presented in the thesis so far and we will then discuss whether any improvements have been made.

- In chapter 1 we present the background to the area of Pharmacovigilance and the ideas of signal detection within this area. We stress to the reader the scale of the problem associated with adverse reactions to drugs and discuss at length the various difficulties that can arise in the analysis phase due to the inherent data issues and political issues. As a result of some of these difficulties we impress upon the reader the need for data mining techniques to be applied and the suitability of such techniques in this particular case.
- Chapter 2 gives an in-depth summary of the data and issues concerning such data. In particular, we stress many of the deficiencies of the data in a spontaneous reporting database. This is even more valid for some databases which are less well organised than others. We then introduce an alternative type of

database which has not yet been properly explored for such analyses and give the advantages and disadvantages of utilising such a database.

- Chapter 3 contains a review of the current state of technology in Pharmacovigilance. In particular the methods presented here are utilised for the purposes of signal detection. We present and compare four different methods with illustrative examples. In the final analysis we comment on the similarities of these methods in practice. The shortcomings of such methods are discussed and these as a whole are attributable to the general approach to the problem rather than any theoretical shortfalls. We introduce an idea by which we could hope to overcome such shortcomings.
- Chapter 4 introduces the central theme of the thesis, which is an alternative approach based on the profiling approach. We describe the motivation behind the new approach as well as introduce the ideas and concepts in a generic sense. Also an outline for the remainder of the thesis is given.
- Chapter 5 describes the profiling approach in detail. We also outline and discuss in depth the important concept of distance and the role it plays in our subsequent analyses. In particular we discuss the distances when under the conditions of sparse and high dimensional data. We compare the merits of four distance metrics under such conditions. Tied to all of this is of course the ideas of profiling and we illustrate using the distances metrics to compare the similarities of drugs and ADEs. The examples at the end of the chapter illustrate that the Bayes Factor would be likely to perform well with such data.
- Chapter 6 is a rather stand-alone chapter separate from the rest of the thesis. It does, however, discuss one of the central issues in data mining today, namely the role of the software package. Here we use five packages to demonstrate the capabilities that software has evolved to this day. Comparisons are drawn

among the various packages as an illustration of the many advantages and disadvantages of designing a software package in a specific way.

- Chapter 7 contains a tool called Peaker which is used to explore high dimensional data. Tied into the chapter we also introduce the ideas of pattern detection and verification. The underlying method to Peaker is fairly straightforward, however it is by no means simple to discover its properties. We introduce many ideas based on the verification of a peak once it has been detected, which include both fairly *ad hoc* devices and more sophisticated tests. In particular, we impress upon the reader that as an exploratory tool it is best to utilise Peaker to its full capacity in order to achieve the best results.
- In chapter 8 we utilise all the methods used in previous chapters in order to perform the anomalous signal detection using the alternative approach. We describe the way by which we combine the previously developed methods. We first look for simple groups within the drugs which reveal that many similar and moreover interesting groups can be detected. This further verifies the validity of taking such an approach. The final illustrative example shows the advantages of using the new approach. We are now able to perform many more types of analyses with the groups than the current methods. We can also illustrate by the signals that these are indeed better and more revealing than those detected with the older methods.

9.3 Conclusion

So what conclusions can we draw from the research carried out in this thesis? From a general point of view, we have advocated an approach of analysing Pharmacovig-

ilance data¹ which differs in principle to the older methods employed in this area. The main difference is that we dispense with the older approach that drugs and ADEs are independent entities and can only be analysed under the assumption of independence. Rather we attempt to look for structures that exist within the data and aim to use these structures to analyse the data in a more thorough way. The problem with taking such an approach, which has traditionally been considered impossible, is that it is rather difficult to obtain such structures due to both the poor quality of the data and the high dimensionality of the data. More concretely, we have mentioned the traditionally held view that drugs should not be grouped according to their ADE profiles and vice versa. In a sense this is true, since this can be highly distorting to the final analysis. However, we have demonstrated that with the right approach and a willingness to explore, useful information can indeed be obtained from such data. In particular, we have illustrated that taking the new approach is certainly not disadvantageous in obtaining signals, but moreover can reveal aspects of the data which are not picked up by traditional methods. The signals detected are more stable when using the new approach and new signals are detected when analysing the groups both as a whole and within which would not be otherwise detectable under the older approaches.

In chapter 8 we devised an alternative approach to detecting signals in Pharmacovigilance data. In particular we used Peaker to look for signals within groups of drugs such that we can compare with the baseline signals generated by the GPS. The signals we generated showed that there is a substantial difference between the old and new signals. In particular, we have detected some signals which would not have been detectable under the older approach. In the main we have looked for signals which to some degree are aggregated between the peaks and the baseline. This has given us signals which are both robust and useful. The outline of such an

¹This is not restricted to Pharmacovigilance only, but is rather a general approach to many data mining tasks.

approach is that we first take the peaks found amongst the drugs and then match the ADEs to these drugs to discover the hidden possibilities.

In the main we have derived an alternative approach to analysing signals. This is carried out in several steps, all of which are integral to the final analysis. Firstly we need to devise a sensible way of measuring similarities among drugs and ADEs. In more detail we need a suitable distance for this purpose. We show that the Bayes Factor is the most useful distance out of a selection of candidates. Using the Bayes Factor we can derive suitable similarities in high dimensional sparse data which is precisely the type of data that we are faced with. Secondly we need a useful way of exploring the similarities that we have derived. This is solved by using the Peaker algorithm which is designed to detect patterns in high dimensional space (in particular localised patterns). We utilise this to look for signals in groups of similar drugs and we have illustrated the results thereby arrived at.

9.4 Future Work

The three main areas of research that can be furthered are as follows:

- Further research can be performed on the anomaly detection. In particular, further investigation into a suitable distance metric to utilise when comparing and combining drug and ADE profiles is desirable. This is by no means an easy task since intrinsically the data quality is a main obstacle in assessing the performance of such metric. We hope to explore this further in conjunction with any improvements that can be made possible in the quality of the data.
- The data mining techniques that can be applied to detect anomalies can also be further investigated. We have introduced one tool for such purposes, but undoubtedly there are many other means through which we can identify signals. In particular, alternative methods to explore high dimensional spaces can

be applied which will perhaps reveal different insight into the data. Spatial modelling is another possibility, in which case it is probable that we would be able to obtain a more sophisticated representation of the data through which we would be able to utilise more complicated models.

- Alternative databases can be explored alongside the SRDs. We have mentioned one such database (the GPRD) but have yet to be able to make good use of the type of data it holds. This is mainly because of the lack of insight into appropriate ways to make sense of such data. The GPRD is precisely opposite to the SRDs since it contains longitudinal data rather than the cross-sectional data contained in SRDs. Means to combining the two are particularly problematic since the GPRD is arguably of even poorer quality. The advantage, on the other hand, is the sheer volume of information that the GPRD contains and it would certainly be a major step forward if it becomes possible to utilise this information.

Appendix A

Tables of Signals

Results are generated using each of the methods OR, the Gamma-Poisson model, the Beta-Binomial model, the GPS and the BCPNN. The data used is a set of 172 drugs and 2245 ADEs. The results are obtained by programs written in C. Detailed discussions of the results can be found in chapter 3.

Drug index	ADE index	Count	Odds Ratio
99	24	1	14137.9
130	379	1	10603.4
62	890	1	9896.5
44	861	1	8996.82
75	1075	2	8482.71
26	542	1	8482.71
26	1936	1	8482.71
26	1705	1	8482.71
53	1371	1	8247.08
87	1090	2	7068.93
78	644	1	7068.93
111	648	1	5821.47
31	469	2	5498.06
97	1171	1	4948.25
111	257	4	4657.18
16	1802	1	4638.98
16	1786	4	4638.98
80	1261	1	4567.62
59	1018	2	4498.41
26	225	2	4241.36
110	1249	9	4241.36
18	1297	1	3906.51
28	373	3	3806.35
44	2180	2	3598.73
147	161	1	3534.46
61	258	2	3492.88
77	1090	2	3298.83
48	1927	1	3298.83
21	926	1	3298.83
153	588	1	3298.83
13	1503	1	3298.83
81	1670	2	3141.75
83	855	3	3029.54
44	2064	2	2998.94
42	311	1	2939.55
72	1552	1	2910.74
34	294	2	2827.57
26	939	1	2827.57
27	352	3	2635.16
17	1158	1	2604.34

Table A.1: Highest 40 Odds Ratios

Drug index	ADE index	Count	GPS signals
158	2063	20	158.615
135	2175	228	147.185
144	2012	37	144.92
169	1181	16	138.522
135	317	158	136.952
135	716	257	132.183
135	707	57	124.293
100	589	21	118.15
42	95	12	112.125
5	6	24	100.138
93	1167	12	99.7102
110	1249	9	99.3357
139	560	9	97.7911
110	180	9	93.6065
5	809	19	91.7269
135	360	190	91.1696
169	1723	10	89.492
125	1156	13	89.4524
143	313	23	88.4922
132	1338	28	87.416
143	1822	11	87.3194
143	358	25	86.6622
122	421	11	84.3258
92	2050	9	83.4615
135	123	19	83.1669
135	1376	21	82.6935
132	621	24	82.1242
39	1219	12	81.412
93	1164	10	80.3663
56	2050	9	79.0055
140	1219	12	77.5629
142	1340	9	76.4447
102	1670	7	75.3204
71	671	8	74.2088
122	1368	11	72.8894
171	1672	40	71.7762
151	1155	38	71.7431
151	1154	29	71.0506
71	1265	10	69.8434
56	371	7	67.6004

Table A.2: Highest 40 GPS Signals

Drug index	ADE index	Count	Gamma-Poisson measures
99	24	1	13356.8
130	379	1	10069.5
62	890	1	8872.82
75	1075	2	7849.48
53	1371	1	7732.57
44	861	1	7121.67
26	542	1	6717.16
26	1936	1	6717.16
26	1705	1	6717.16
78	644	1	6665.46
87	1090	2	6531.22
31	469	2	5035.71
111	648	1	4618.5
97	1171	1	4302.48
16	1786	4	4225.42
16	1802	1	4190.01
111	257	4	4153.91
59	1018	2	4005.55
110	1249	9	3884.34
26	225	2	3853.79
18	1297	1	3779.65
80	1261	1	3626.96
28	373	3	3519.2
147	161	1	3347.01
44	2180	2	3319.15
21	926	1	3301.52
13	1503	1	3301.52
48	1927	1	3227.68
153	588	1	3136.35
77	1090	2	3072.23
81	1670	2	2968.64
61	258	2	2959.98
72	1552	1	2920.48
44	2064	2	2794.71
26	939	1	2681.62
34	294	2	2635.97
83	855	3	2632.01
27	352	3	2463.09
42	311	1	2336.88
139	560	9	2302.72

Table A.3: Highest 40 Gamma-Poisson Measures

Drug index	ADE index	Count	Beta-Binomial measures
99	24	1	14012.7
130	379	1	10489
62	890	1	9705.81
44	861	1	8663.96
75	1075	2	8429.46
26	542	1	8169.04
26	1936	1	8169.04
26	1705	1	8169.04
53	1371	1	8137.73
87	1090	2	7020.41
78	644	1	7018.85
111	648	1	5606.76
31	469	2	5473.37
97	1171	1	4891.82
111	257	4	4622
16	1786	4	4616.53
16	1802	1	4551.4
59	1018	2	4472.07
80	1261	1	4399.36
110	1249	9	4235.27
26	225	2	4200.92
18	1297	1	3868.25
28	373	3	3789.63
44	2180	2	3571.13
147	161	1	3489.46
61	258	2	3427.66
21	926	1	3285.26
13	1503	1	3285.26
48	1927	1	3283.34
153	588	1	3278.65
77	1090	2	3277.59
81	1670	2	3127.64
83	855	3	2991.79
44	2064	2	2979.73
72	1552	1	2899.63
42	311	1	2831.45
34	294	2	2809.51
26	939	1	2791.79
27	352	3	2627.02
124	1076	1	2575.42

Table A.4: Highest 40 Beta-Binomial Measures (Method of Moments)

Drug index	ADE index	Count	BCPNN signals
135	2175	228	92.8163
135	716	257	90.123
135	317	158	76.7478
135	360	190	63.601
171	276	121	44.981
135	707	57	42.1396
144	2012	37	32.4907
117	645	396	31.4952
171	1360	54	31.404
117	1036	69	30.8728
134	652	59	28.6499
171	1672	40	27.7622
171	1316	45	27.3953
151	1155	38	26.8419
171	1227	39	25.9428
171	1485	78	25.7037
155	761	99	25.5929
89	1273	66	25.1328
170	340	35	24.8149
5	8	195	24.4783
68	1040	105	24.1953
115	2185	34	24.1169
115	643	39	23.7953
115	1368	56	23.7201
5	512	275	23.6285
132	1338	28	23.5009
151	1154	29	22.5322
5	270	33	22.152
143	358	25	21.6299
5	6	24	21.6137
90	1505	30	21.3815
145	363	258	21.1906
132	621	24	20.7302
89	1563	45	20.4809
143	313	23	20.454
167	1506	42	20.2408
158	2063	20	20.2287
89	276	84	20.1911
100	589	21	20.1795
171	87	82	19.8618

Table A.5: Highest 40 BCPNN Signals (Uniform Prior)

Appendix B

Cluster Analysis Results

Figures are generated using single link and complete cluster analysis on each of the distances DuMouchel's method, Kullback-Liebler, χ^2 and the Bayes Factor. The results are obtained by programs written in C. Detailed discussions of the results can be found in chapter 5.

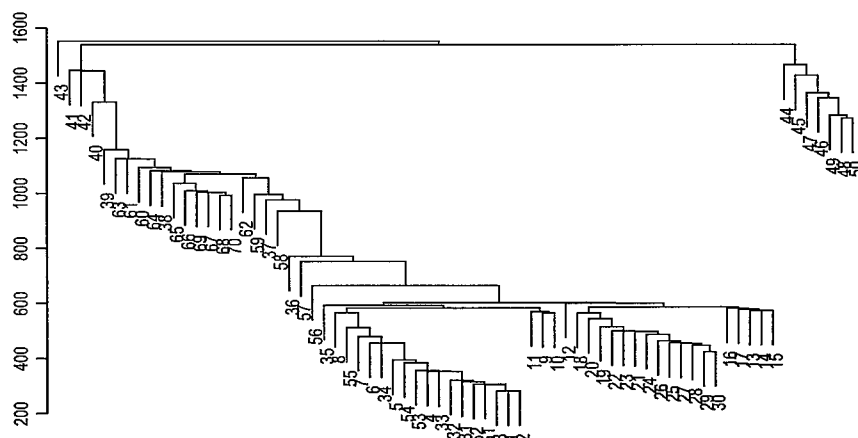


Figure B.1: Cluster Analysis Plot of DuMouchel's Method (Single Linkage)

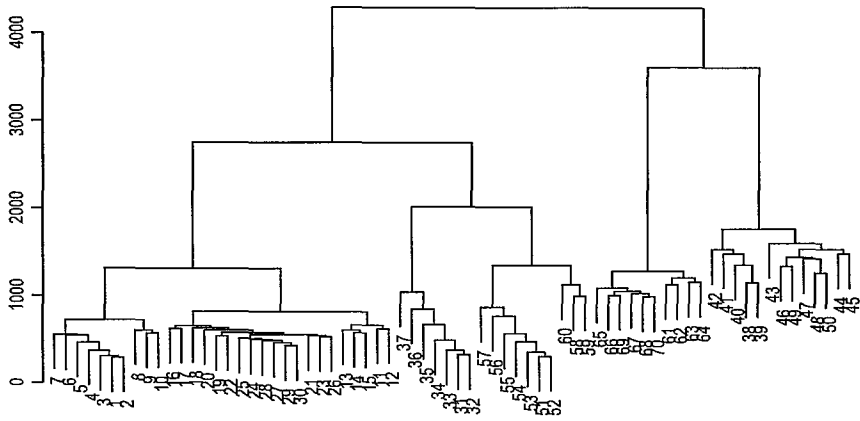


Figure B.2: Cluster Analysis Plot of DuMouchel's Method (Complete Linkage)

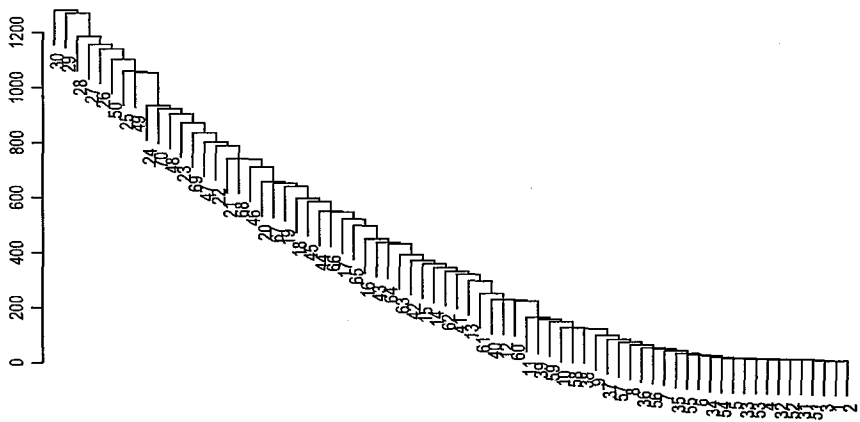


Figure B.3: Cluster Analysis Plot of Kullback-Liebler Distance (Single Linkage)

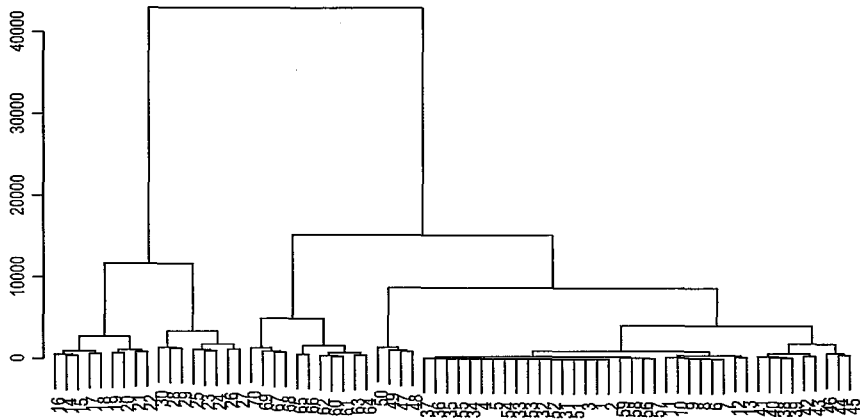


Figure B.4: Cluster Analysis Plot of Kullback-Liebler Distance (Complete Linkage)

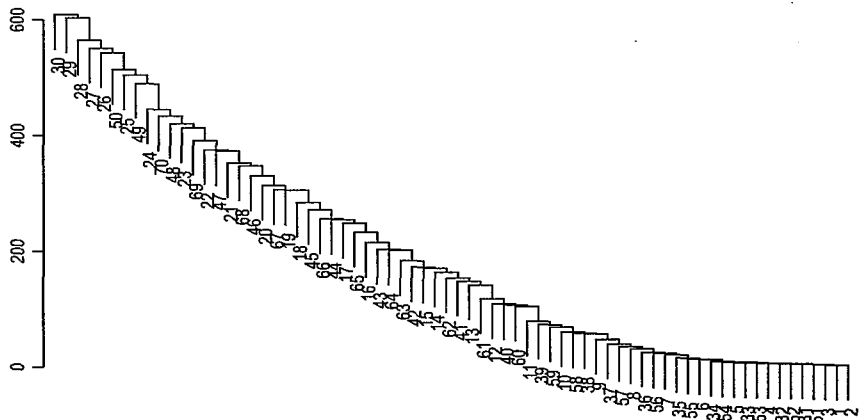


Figure B.5: Cluster Analysis Plots of χ^2 Distance (Single Linkage)

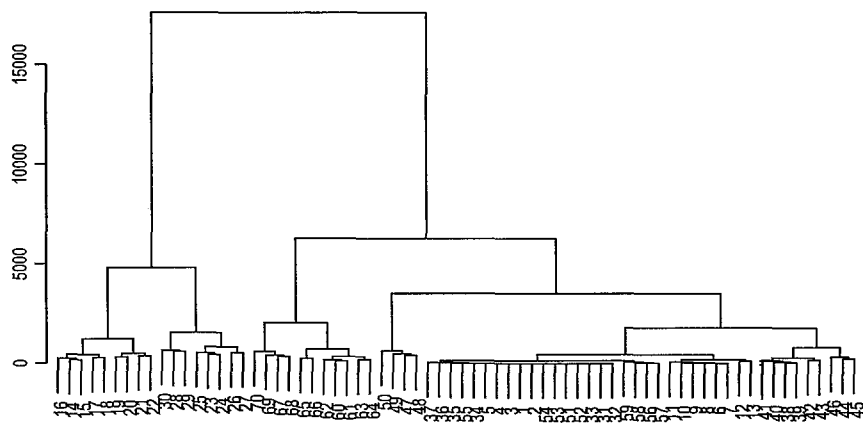


Figure B.6: Cluster Analysis Plots of χ^2 Distance (Complete Linkage)

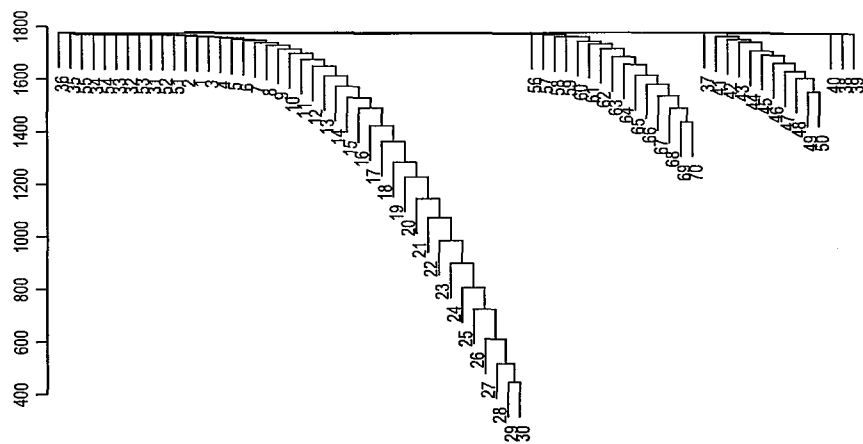


Figure B.7: Cluster Analysis Plots of Bayes Factor Distance (Single Linkage)

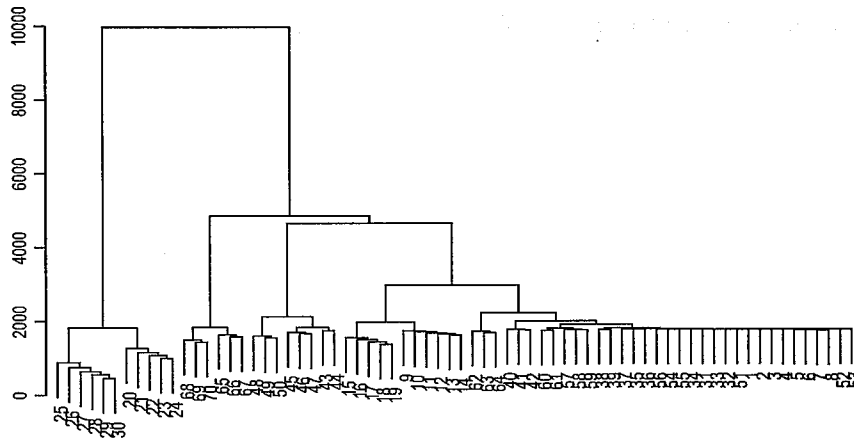
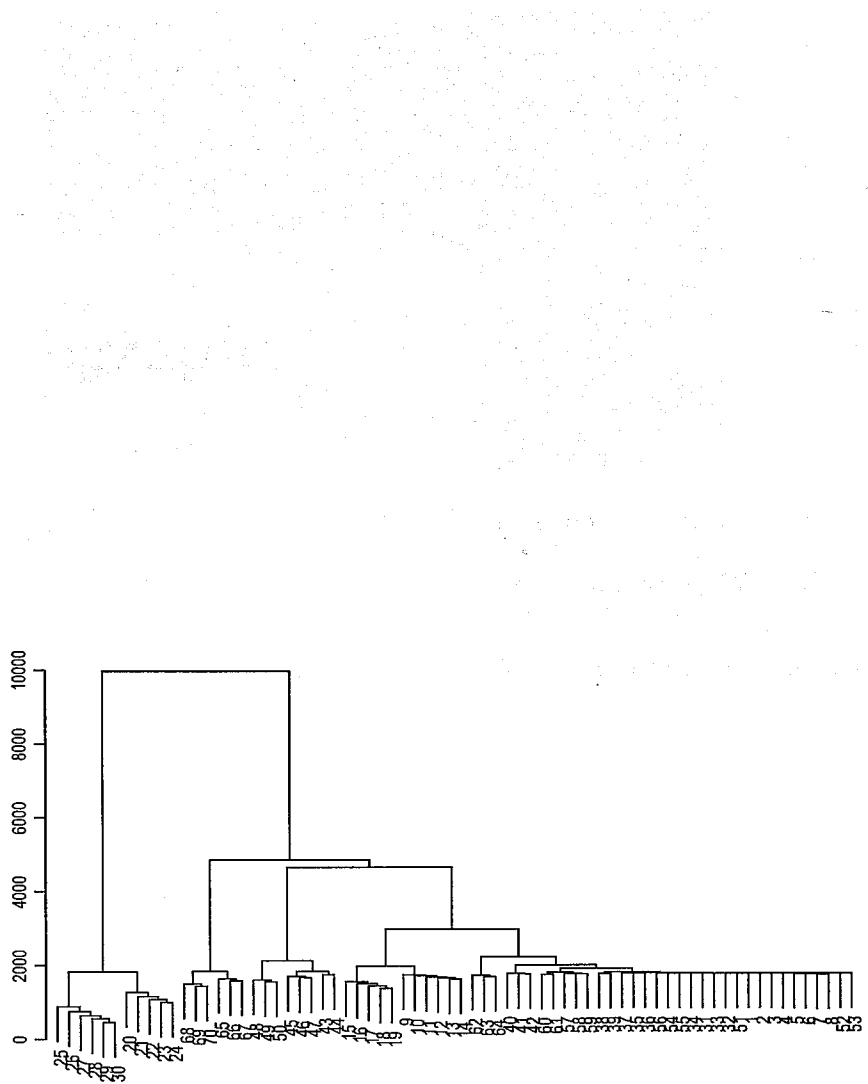


Figure B.8: Cluster Analysis Plots of Bayes Factor Distance (Complete Linkage)

Bibliography

- Adams, N., Hand, D. & Till, R. (2001), 'Mining for classes and patterns in behavioural data', *Journal of the Operational Research Society* **52**, 1017–1024.
- Aggarwal, C., Hinneburg, A. & Keim, D. (2001), 'On the surprising behavior of distance metrics in high dimensional space', *Database Theory - ICDT 2001, Proceedings Lecture Notes in Computer Science 3512* pp. 420–434.
- Agrawal, R., Imielinski, T. & Swami, A. (1993), 'Mining association rules between sets of items in large databases', *In Proceedings of the ACM SIGMOD Conference on Management of Data* pp. 207–216.
- Agrawal, R. & Sikrant, R. (1994), 'Fast algorithms for mining association rules', *In Proceedings of the 20th International Conference on Very Large Databases* pp. 487–499.
- Agresti, A. (1990), *Categorical Data Analysis*, Wiley, New York.
- Amery, W. (1999a), 'Assessment of signals generated from spontaneously reported adverse events', *Pharmacoepidemiology and Drug Safety* **8**, 301–304.
- Amery, W. (1999b), 'Signal generation from spontaneous adverse event reports', *Pharmacoepidemiology and Drug Safety* **8**, 147–150.
- Barnett, V. (1976), 'The ordering of multivariate data', *Journal of the Royal Statistical Association. Series A* **139**, 318–354.

- Bate, A., Lindquist, M. & Edwards, I. (1998), 'A bayesian neural network method for adverse drug reaction signal generation', *Eur J Clin Pharmacol* **54**, 315–321.
- Bate, A., Lindquist, M., Edwards, I. & Orre, R. (2002), 'A data mining approach for signal detection and analysis.', *Drug Safety* **26**, 393–397.
- Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. (1999), 'When is "nearest neighbour" meaningful?', *Proceedings of the International Conference on Database Theory (ICDT)* pp. 217–235.
- Bolton, R., Hand, D. & Crowder, M. (2004), 'Significance tests for unsupervised pattern discovery in large continuous multivariate data sets', *Computational Statistics and Data Analysis* **46**, 57–79.
- Bortnichak, E., Wise, R., Salive, M. & Tilson, H. (2001), 'Proactive safety surveillance', *Pharmacoepidemiology and Drug Safety* **10**, 191–196.
- Bowman, A. (1980), 'A note on consistency of the kernel method for the analysis of categorical data', *Biometrika* **67**, 682–684.
- Brown, E. (2002), 'Effects of coding dictionary on signal generation', *Drug Safety* **25**, 445–452.
- Brown, P. & Rundell, P. (1985), 'Kernel estimates for categorical data', *Technometrics* **27**, 293–299.
- Chen, S. & Liu, X. (2004), 'The contribution of data mining to information science', *Journal of Information Science* **30**(6), 550–558.
- Cheng, V., Li, C., Kwok, J. & Li, C. (2004), 'Dissimilarity learning for nominal data', *Pattern Recognition* **37**, 1471–1477.
- Cressie, N. (1993), *Statistics for Spatial Data*, Wiley, NY.

- Cressie, N. & Read, T. (1984), 'Multinomial goodness-of-fit tests', *Journal of the Royal Statistical Society, Series B* **46**, 440–464.
- Davi, A., Deichmann, J., Eshqhi, A., Sayek, S., Teebaqy, N. & Topi, H. (2003), 'A review of software packages for data mining', *American Statistician* **57**(4), 290–309.
- Deely, J. & Lindley, D. (1981), 'Bayes empirical bayes', *Journal of the American Statistical Association* **76**, 833–841.
- DuMouchel, W. (1999), 'Bayesian data mining in large frequency table with application to fda spontaneous reporting system (with discussion)', *The American Statistician* **53**(3), 177–207.
- DuMouchel, W. & Pregibon, D. (2001), 'Empirical bayes screening for multi-item association', *In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining* pp. 67–76.
- Efron, B. & Morris, C. (1973), 'Stein's estimation rule and its competitors: An empirical bayes approach', *Journal of the American Statistical Association* **68**(341), 117–130.
- Escobar, M. & West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association* **90**, 577–588.
- Evans, S. (2000), 'Pharmacovigilance, a science or fielding emergencies?', *Statistics in Medicine* **19**, 3199–3209.
- Evans, S. (2002), *Pharmacovigilance*, John Wiley and Sons, Ltd.
- Fedorov, V., Lin, X. & Patwardhan, R. (2003), 'Comparisons of different filters in detection of weak signals for large size contingency tables', *GSK BDS Technical Report*.

- Fienberg, S. & Holland, P. (1973), ‘Simultaneous estimation of multinomial cell probabilities’, *Journal of the American Statistical Association* **68**, 683–691.
- Fisher, N., Mammen, E. & Marron, J. (1994), ‘Testing for multimodality’, *Computational Statistics and Data Analysis* **18**, 499–512.
- Friedman, J. H. & Fisher, N. I. (1999), ‘Bump hunting in high-dimensional data’, *Statistics and Computing* **9**, 123–143.
- Good, I. (1971), ‘The utility of a distribution’, *Biometrika* **58**, 255–277.
- Good, I. & Gaskins, R. (1971), ‘Nonparametric roughness penalties for probability densities’, *Biometrika* **58**, 255–277.
- Good, I. J. & Gaskins, R. A. (1980), ‘Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data’, *Journal of the American Statistical Association* **75**, 42–56.
- Group, D. M. C. (2000), ‘Disease mapping models: An empirical evaluation’, *Statistics in Medicine* **19**, 2217–2241.
- Gruber, M. (1998), *Improving Efficiency by Shrinkage*, Marcel Dekker, Inc.
- Gunel, E. & Dickey, J. (1974), ‘Bayes factors for independence in contingency tables’, *Biometrika* **61**, 545–557.
- Hall, P., Sheather, S., Jones, M. & Marron, J. (1991), ‘On optimal data-based bandwidth selection in kernel density estimation’, *Biometrika* **78**, 263–269.
- Hand, D. (1998), ‘Data mining: statistics and more?’, *The American Statistician* **52**, 112–118.
- Hand, D., Adams, N. & Bolton, R. (2002), *Pattern Detection and Discovery*, Springer.

- Hand, D., Blunt, G., Kelly, M. & Adams, N. (2000), 'Data mining for fun and profit', *Statistical Science* **15**, 111–131.
- Hand, D. & Bolton, R. (2004), 'Pattern discovery and detection: A unified statistical methodology', *Journal of Applied Statistics* **31**, 885–924.
- Hand, D., Mannila, H. & Smyth, P. (2001), *Principles of Data Mining*, MIT Press.
- Hand, D. & Zhang, Z. (2005), 'Spotting the difference: detecting local structures in large data sets', *Expert Update* **8**, 11–18.
- Hills, M. & Alexander, F. (1989), 'Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: A review', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **152**(3), 353–384.
- Hipp, J., Guntzer, U. & Nakhaeizadeh, G. (2000), 'Algorithms for association rule mining a general survey and comparison', *SIGKDD Explorations* **2**, 58–64.
- Huang, Z. (1998), 'Extensions to the k-means algorithm for clustering large data sets with categorical values', *Data Mining and Knowledge Discovery* **2**, 283–304.
- James, W. & Stein, C. (1961), 'Estimation with quadratic loss', *Proceedings of the Fourth Berkeley Symposium on Mathematics and Statistics* **1**, 361–379.
- Jones, K. (2001), 'The role of data mining technology in the identification of signals of possible adverse drug reactions: Value and limitations', *Current Therapeutic Research* **62**, 664–672.
- Kass, R. (1993), 'Bayes factor in practice', *Statistician* **42**, 551–560.
- Kass, R. & Raftery, A. (1995), 'Bayes factor', *Journal of the American Statistical Association* **90**, 773–795.

- Kendall, M. (1952), *The Advanced Theory of Statistics*, Vol. 1, 5 edn, Griffin, London.
- Lamberti, P. & Majtey, A. (2003), 'Non-logarithmic jensen-shannon divergence', *Physica A* **329**, 81–90.
- Lilienfeld, D. E. (2004), 'A challenge to the data miner', *Pharmacoepidemiology and Drug Safety* **13**, 881–884.
- Lindquist, M., Edwards, I., Bate, A., Fucik, H., Nunes, A. & Stahl, M. (1999), 'From association to alert – a revised approach to international signal analysis', *Pharmacoepidemiology and Drug Safety* **8**, s15–s25.
- MacEachern, S. & Muller, P. (1998), 'Estimating mixture of dirichlet process models', *Journal of Computational and Graphical Statistics* **7**, 223–238.
- MacEachern, S. N. & Muller, P. (2000), Efficient mcmc schemes for robust model extensions using encompassing dirichlet process mixture models, *in* D. Insua & F. Ruggeri, eds, 'Robust Bayesian Analysis', Springer-Verlag, New York, pp. 295–315.
- Manton, K., Woodbury, M., Stallard, E., Riggan, W., Creason, J. & Pellom, A. (1989), 'Empirical bayes procedures for stabilizing maps of u.s. cancer mortality rates', *Journal of the American Statistical Association* **84**(407), 637–650.
- Marshall, R. (1991a), 'Mapping disease and mortality rates using empirical bayes estimators', *Applied Statistics* **40**(2), 283–294.
- Marshall, R. (1991b), 'A review of methods for the statistical analysis of spatial patterns of disease', *Journal of the Royal Statistical Society. Series A* **154**(3), 421–441.

- Menendez, M., Pardo, J., Pardo, L. & Pardo, M. (1997), ‘The jensen-shannon divergence’, *J. Franklin Inst.* **334B**, 307–318.
- Meza, J. (2003), ‘Empirical bayes estimation smoothing of relative risks in disease mapping’, *Journal of Statistical Planning and Inference* **112**, 43–62.
- Minnotte, M. (1997), ‘Nonparametric testing of existing modes’, *The Annals of Statistics* **25**, 1646–1660.
- Morales, D., Pardo, L., Saalicro, M. & Menendez, M. (1994), ‘Asymtotic properties of divergence statistics in astratified sampling and its applications to test statistical hypotheses’, *Journal of Statistical Planning and Inference* **38**, 201–222.
- Morris, C. (1983), ‘Parametric empirical bayes inference: Theory and application’, *Journal of the American Statistical Association* **78**(381), 47–55.
- O’Neill, R. & Szarfman, A. (2001), ‘Some us food and drug administration perspectives on data mining for pediatric safety assessment’, *Current Therapeutic Research* **62**(9), 381–392.
- Orre, R., Lansner, A., Bate, A. & Lindquist, M. (2000), ‘Bayesian neural networks with confidence estimation applied to data mining’, *Computational Statistics and Data Analysis* **34**, 473–493.
- Pekalska, E., Paclik, P. & Duin, R. (2001), ‘A generalized kernel approach to dissimilarity based classification’, *Journal of Machine Learning Research* **2**, 175–211.
- Pyke, R. (1965), ‘Spacings’, *Journal of the Royal Statistical Society. Series B* **27**, 395–449.
- Quintana, A. (2002), Nonparametric bayesian analysis for assessing homogeneity in $k \times l$ contingency tables with fixed right margin totals, Technical report, Pontificia Universidad Catolica de Chile.

- Read, T. (1984), 'Small-sample comparisons for the power divergence goodness-of-fit statistics', *Journal of the American Statistical Association* **79**, 929–935.
- Sain, S., Baggerly, K. & Scott, D. (1994), 'Cross-validation of multivariate densities', *Journal of the American Statistical Association* **89**, 807–817.
- Scott, D. & Terrell, G. (1987), 'Biased and unbiased cross-validation in density estimation', *Journal of the American Statistical Association* **82**, 1131–1146.
- Sheath, S. & Jones, M. (1991), 'A reliable data-based bandwidth selection method for kernel density estimation', *Journal of the Royal Statistical Society, Series B* **53**, 683–690.
- Silverman, B. (1981), 'Using kernel density estimates to investigate multimodality', *Journal of the Royal Statistical Association. Series B* **43**, 97–99.
- Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Simonoff, J. (1995), 'Smoothing categorical data', *Journal of Statistical Planning and Inference* **47**, 41–69.
- Stone, C. (1984), 'An asymptotically optimal window selection rule for kernel density estimates', *The Annals of Statistics* **12**, 1285–1297.
- Szarfman, A., Machado, S. & O'Neill, R. (2002), 'Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fdas spontaneous reports database', *Drug Safety* **6**, 381–392.
- Tate, M. & Hye, L. (1973), 'Inaccuracy of the χ^2 test of goodness of fit when expected frequencies are small', *Journal of the American Statistical Association* **68**, 836–841.

- Taylor, C. (1989), 'Bootstrap choice of the smoothing parameter in kernel density estimation', *Biometrika* **76**, 705–712.
- Trontell, A. (2001), 'How the us food and drug administration defines and detects adverse drug events', *Current Therapeutic Research* **62**(9), 641–649.
- van der Heijden, P., van Puijenbroek, E., van Buuren, S. & van der Hofstede, J. (2002), 'On the assessment of adverse drug reactions from spontaneous reporting systems, the influence of under-reporting on odds ratio', *Statistics in Medicine* **21**, 2027–2044.
- van Puijenbroek, E., Bate, A., Leufkens, G., Lindquist, M., Orre, R. & Egberts, A. (2002), 'A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions', *Pharmacoepidemiology and Drug Safety* **11**, 3–10.
- Waller, P. C., Coulson, R. A. & Wood, S. M. (1996), 'Regulatory pharmacovigilance in the united kingdom: Current principles and practice', *Pharmacoepidemiol Drug Safety* **5**, 363–375.
- Wand, M. & Jones, M. (1994), 'Multivariate plug-in bandwidth selection', *Computational Statistics* **9**, 97–116.
- Wong, C. K. (1999), 'Pharmacovigilance resources in the united kingdom', *Pharmaceutical Journal* **263**, 285–288.
- Zhang, Z. & Hand, D. (2005), 'Detecting groups of anomalously similar objects in large data sets', In *Advances in Intelligent Data Analysis VI, 6th International Symposium on Intelligent Data Analysis* pp. 509–519.