TKK Dissertations in Information and Computer Science
Espoo 2008 TKK-ICS-D5

ADVANCES IN MINING BINARY DATA: ITEMSETS AS SUMMARIES

Nikolaj Tatti

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

TKK Dissertations in Information and Computer Science
Espoo 2008 TKK-ICS-D5

ADVANCES IN MINING BINARY DATA: ITEMSETS AS SUMMARIES

Nikolaj Tatti

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium TU1 at Helsinki University of Technology (Espoo, Finland) on the 6th of June, 2008, at 12 noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 5400
FI-02015 TKK
FINLAND
URL: http://ics.tkk.fi
Tel. +358 9 451 1
Fax +358 9 451 3369
E-mail: series@ics.tkk.fi

ISBN 978-951-22-9375-9 (Print)
ISBN 978-951-22-9376-6 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)
URL: http://lib.tkk.fi/Diss/2008/isbn9789512293766/

Multiprint
Espoo 2008

Tatti N. (2008): **Advances in Mining Binary Data: Itemsets as Summaries.** Doctoral Thesis, Helsinki University of Technology, TKK Dissertations in Information and Computer Science, Report D5, Espoo, Finland.

Keywords: data mining, frequent itemset, boolean queries, safe projections, itemset ranking.

Abstract

Mining frequent itemsets is one of the most popular topics in data mining. Itemsets are local patterns, representing frequently cooccurring sets of variables. This thesis studies the use of itemsets to give information about the whole dataset.

We show how to use itemsets for answering queries, that is, finding out the number of transactions satisfying some given formula. While this is a simple procedure given the original data, the task transforms into a computationally infeasible problem if we seek the solution using the itemsets. By making some assumptions of the structure of the itemsets and applying techniques from the theory of Markov Random Fields we are able to reduce the computational burden of query answering.

We can also use the known itemsets to predict the unknown itemsets. The difference between the prediction and the actual value can be used for ranking itemsets. In fact, this method can be seen as generalisation for ranking itemsets based on their deviation from the independence model, an approach commonly used in the data mining literature.

The next contribution is to use itemsets to define a distance between the datasets. We achieve this by computing the difference between the frequencies of the itemsets. We take into account the fact that the itemset frequencies may be correlated and by removing the correlation we show that our distance transforms into Euclidean distance between the frequencies of parity formulae.

The last contribution concerns calculating the effective dimension of binary data. We apply fractal dimension, a known concept that works well with real-valued data. Applying fractal dimension dimension directly is problematic because of the unique nature of binary data. We propose a solution to this problem by introducing a new concept called normalised correlation dimension. We study our approach theoretically and empirically by comparing it against other methods.

Tiivistelmä

Kattavien joukkojen louhinta on yksi suosituimmista tiedon louhinnan teemoista. Kattavat joukot ovat paikallisia hahmoja: ne edustavat usein esiintyviä muuttujakombinaatioita. kattavien joukkojen käyttöä koko tietokantaa kuvaaviin tarkoituksiin.

Kattavia joukkoja voidaan käyttää Boolen kyselyihin vastaamiseen, ts. annetun Boolen kaavan toteuttavien tietuiden lukumäärän arviointiin. Tehtävästä tulee kuitenkin laskennallisesti vaativa, jos käytössä ovat vain kattavat joukot. Väitöskirjassa osoitetaan, että tietyin oletuksin ongelman ratkaisemista voidaan helpottaa käyttäen hyväksi tekniikoita, jotka perustuvat Markov-kenttiin.

Väitöskirjassa tutkitaan myös miten kattavia joukkoja voidaan käyttää tuntemattomien joukkojen frekvenssin ennustamiseen. Varsinaisen datasta lasketun frekvenssin ja ennusteen välistä erotusta voidaan käyttää kattavan joukon merkitsevyyden mittana. Tämä lähestymistapa on itseasiassa tiedon louhinnassa usein toistuvan tärkeysmitan yleistys, jossa kattavan joukon tärkeys on sen poikkeama riippumattomuusoletuksesta.

Väitöskirjan seuraava tutkimusaihe on kattavien joukkojen käyttö tietokantojen välisen etäisyyden määrittelemiseen. Etäisyys määritellään kattavien joukkojen frekvenssien erotuksena. Kattavien joukkojen frekvenssien välillä saattaa olla korrelaatiota ja eliminoimalla tämä korrelaatio työssä osoitetaan, että etäisyys vastaa tiettyjen pariteettikyselyiden välistä euklidista etäisyyttä.

Väistökirjan viimeinen teema on binääritietokannan efektiivisen dimension määritteleminen. Työssä sovelletaan fraktaalidimensiota, joka on suosittu menetelmä ja soveltuu hyvin jatkuvalle datalle. Tämän lähestymistavan soveltaminen diskreettiin dataan ei kuitenkaan ole suoraviivaista. Työssä ehdotetaan ratkaisuksi normalisoitua korrelaatiodimensiota. Lähestymistapoja tarkastellaan sekä teoreettisesti että empiirisesti vertailemalla sitä muihin tunnettuihin menetelmiin.

Contents

List of Figures

List of Notations

$\mathbb{Q}$	rational numbers
$\mathbb{R}$	real numbers
x^T	transpose of a vector x
x_i	ith element of a vector x
Ω	sample space, usually $\{0,1\}^K$
ω	binary vector
K	dimension of sample space, number of attributes
a_i	ith attribute
A	set of all attributes, $A = \{a_1, \ldots, a_K\}$
B, C, Q	itemsets
$\mathcal{F}, \mathcal{G}, \mathcal{I}, \mathcal{C}, \mathcal{A}$	families of itemsets
D	binary data set
$\|D\|$	number of elements in D
$\cup$	set union
$\cap$	set intersection
$\vee$	disjunctive operator
$\wedge$	conjunctive operator
$\oplus$	exclusive-or (XOR) operator
S_F	indicator function of a Boolean formula F, returns 1 if and only if F is satisfied
$S_{\mathcal{F}}$	indicator function for a family $\mathcal{F}$ of itemsets
p, q	distributions
p^*, p^{ME}	distribution derived using Maximum Entropy principle
$\mathrm{E}_p[X]$	mean of X with respect to p
$\mathrm{Std}_p[X]$	standard variation of X with respect to p
$\mathrm{Var}_p[X]$	variance of X with respect to p
$\mathrm{Cov}_p[X]$	covariance matrix of X with respect to p

$\mathbb{P}(X)$	probability of an event X
$p\,(B = \omega)$	probability $p\,(b_1 = \omega_1, \ldots, b_L = \omega_L)$
$p\,(B = 1)$	probability $p\,(b_1 = 1, \ldots, b_L = 1)$
$\mathcal{E}(p)$	entropy of distribution p
$\mathrm{KL}(p;\, q)$	Kullback-Leibler divergence between p and q
$\mathcal{P}(S, \theta)$	space of distributions satisfying $\mathrm{E}[S] = \theta$
$d_{CM}(D_1, D_2;\, \mathcal{F})$	constrained minimum (CM) distance between D_1 and D_2 based on itemsets $\mathcal{F}$
$\mathrm{fi}(Q; \mathcal{F}, \theta)$	frequency interval of Q derived from the frequencies θ of itemsets $\mathcal{F}$
$\mathrm{cd}_{\mathrm{A}}(D;\, \alpha_1, \alpha_2)$	correlation dimension of D with α_1 and α_2 tail cuts
$\mathrm{cd}_{\mathrm{R}}(D;\, r_1, r_2)$	correlation dimension of D calculated using radii r_1 and r_2
Z_D	distance between two random points from D
$\mathrm{acd}(D)$	approximative correlation dimension of D, $\mathrm{acd}(D) = \mathrm{E}[Z_D]\,/\mathrm{Std}[Z_D]$
$\mathrm{ind}(D)$	a data set having equal margins as D but independent attributes
$\mathrm{ind}(L, s)$	a data set having L independent attributes with margins equal to s
$\mathrm{ncd}_{\mathrm{A}}(D;\, \alpha_1, \alpha_2)$	normalised correlation dimension of D with α_1 and α_2 tail cuts

List of Publications

This thesis consists of an introduction and the following papers:

I Nikolaj Tatti. Distances between data sets based on summary statistics. Journal of Machine Learning Research, 8:131–154, Jan 2007.

II Nikolaj Tatti. Computational complexity of queries based on itemsets. Information Processing Letters, pages 183–187, June 2006.

III Nikolaj Tatti. Safe projections of binary data sets. Acta Informatica, 42(8-9):617–638, April 2006.

IV Nikolaj Tatti. Maximum Entropy Based Significance of Itemsets. Accepted for publication in Knowledge and Information Systems (KAIS).

V Nikolaj Tatti, Taneli Mielikäinen, Aristides Gionis, and Heikki Mannila. What is the dimension of your binary data. In Proceedings of Sixth IEEE International Conference on Data Mining (ICDM 2006), pages 603–612, 2006.

We will use this numbering throughout the thesis.

Preface

I've been lucky in many ways. One of the greatest privileges I had over past years is that I have been able to do the work that I wanted to do. In my case that work is being an academic researcher. Over the years I have learned to fully appreciate this fact. To me being a researcher isn't about the money, nor it is the fame. In the end it is a simple joy of solving puzzles, seeing order in chaos. Curiosity, if you will.

Having said that, it is obvious that I am most grateful to people and institutions that have helped me during my journey. I did my work at the Laboratory of Computer and Information Science (CIS) at the Helsinki University of Technology.[1] I am also affiliated with the Basic Research Unit of the Helsinki Institute for Information Technology (HIIT BRU). I was funded by ComMIT graduate school and the Academy of Finland. I am also grateful for a personal grant given by the Foundation of Technology (TES).

I would like to thank my advisor Heikki Mannila whose vision, something that I deeply admire, has provided me with a direction and guidance. Jaakko Hollmén has always had time for my problems and has always cheered me with his upbeat spirit. I am also grateful for the help I received from Jouni K. Seppänen whose calm and down-to-earth attitude I greatly respect. In addition, I would like to thank Aris Gionis and Taneli Mielikäinen with whom I had the pleasure to collaborate. Thanks are also due to Dr. Szymon Jaroszewicz from the National Institute of Telecommunications and Prof. Bart Goethals from University of Antwerp who have reviewed the manuscript of the thesis and provided excellent and insightful comments.

[1] currently part of the Department of Information and Computer Science (ICS).

In addition, I would like to thank Antti Rasinen, Hannes Heikinheimo, Antti Ukkonen, Kai Puolamäki, Heli Hiisilä, Janne Toivola, and Mikko Korpela. I would also like to thank Gemma Garriga who has helped me greatly by reading my thesis and providing comments that significantly improved the readability of the work. My deepest thanks are due to my close friend Anne Patrikainen who originally inspired me to begin working at CIS.

В заключение я хочу поблагодарить мою семью, маму, папу, Валеру и Аниту. Вы для меня важнее всех во всем свете.

Otaniemi, April 2008

Nikolaj Tatti / Николай Татти

Chapter 1

Introduction

> There's a war out there, old friend. A world war. And it's not about who's got the most bullets. It's about who controls the information. What we see and hear, how we work, what we think... it's all about the information!
>
> *Cosmo, Sneakers*

On data mining. It is appropriate to begin this work by discussing goals and origin of data mining. One of the possible definitions of the field is the following.

> Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner.
>
> [HMS01]

The need for data mining comes from the gap between the traditional statistics and computer science: The focus in the traditional statistics is testing hypotheses, the computation involved is usually ignored. On the other hand, in computer science we are focused on fast computing but statistical analysis of the underlying data does not play any major role. However, there is a growing need for doing both simultaneously. Modern, constantly improving technology has enabled us to store huge data sets, both in academia and in industry [HAK+02, KRS02]. These databases are large enough so that manual exploration is infeasible. Hence, we need to either summarise data or explore data automatically for useful statistically relevant information.

Soda	SodaLight	Diapers	Beer
0	1	1	1
1	0	1	1
0	1	0	0
1	0	0	1
1	1	0	0

Table 1.1: A toy example of market basket data with 5 transactions and 4 items.

Data mining, however, is more than simply statistics for large data sets; namely, the actual mining process is involved. In traditional hypothesis test we already know what we are testing whereas in a typical data mining scenario we do not know what are our hypotheses. The novelty of data mining is to find possibly interesting information from data — statistics only provides means for testing whether our acquired information is statistically significant.

The aforementioned definition of the field allows us to divide the data mining techniques into two categories: global methods and local methods. The goal of global methods is to summarise the data as a whole. Prime examples of such techniques are decision trees [Qui93], graphical modelling [Jor99], clustering tasks [DHS00], principal component analysis [Hay98], and graph analysis [BRRT05]. On the other hand, local methods concentrate finding patterns in data. Thus, the goal is not to cover all data but only interesting parts of it. Classic examples of such methods are association rules [AIS93, AMS+96] and episodes [MTV97].

The division of the field into local and global methods is not crisp. For instance, clustering can be viewed as an attempt to model the whole data. However, a single cluster can also be considered as a pattern explaining only a portion of data. Similarly, a collection of local patterns can be viewed as a summary for the complete data set. A large portion of this thesis concentrates on using itemsets, local patterns for binary data, as a summary of the original binary data.

On mining of binary data. Perhaps the most classical example of binary data is market basket data (see e.g. [BSVW99]). A toy example of a such data set is given in Table 1.1. Such data consists of transactions represented by binary vectors. An element of a transaction describes whether the customer bought the corresponding product. For instance, in Table 1.1 transactions 2, 4, and 5 contain Soda.

Association rules are one of the most classical methods for analysing binary data [AIS93, AMS+96]. These are statements of type 'If a customer buys product X, then with a high probability he also buys product Y'. For instance, we may

Itemset	# of customers
Soda	3
SodaLight	3
Diapers	2
Beer	3
Soda, SodaLight	1
Diapers, Beer	2

Table 1.2: A toy example of a collection of itemsets as a surrogate for the data set given in Table 1.1.

conclude from our toy data that customers buying diapers also buy beer.[1] A more fundamental patterns for the analysis are itemsets. Itemsets are sets of products. For instance, in our toy data these are 'Soda', 'SodaLight', 'Soda, SodaLight', and so on. To each itemset we assign a number which we call support of frequency. This is the number of transactions in which every product included in the itemset occurs. For instance, in our toy data we have 3 customers buying Soda and 3 customers buying SodaLight but only one customer buying both. Thus the support for 'Soda' is 3, but the support for 'Soda, SodaLight' is only 1.

In addition to market basket analysis, binary data occur in a wide range of scenarios, such as bag-of-words representations of text documents [BFS03], fossil occurrence at paleontogical sites [For05], geographical co-occurrence [HFEM07] of mammals, traffic accident reports [GWBV03], click-stream data [KBF+00], co-authorship and citations in scientific papers [BMS02, GKM03].

Topic of the thesis. A large portion of the thesis focuses on using itemsets as a surrogate for the original data set. In other words, imagine that instead of a binary data set given in Table 1.1 you are provided with a collection of itemsets presented in Table 1.2. The idea behind this scenario is that the itemsets capture the essential information from the original data set. Note that in our toy example we are not able to fully construct the original data set from the itemsets, that is, there are other data sets that produce equivalent itemsets. This is an important observation since this enables using this scenario in privacy-preserving data mining.

Once we have replaced the original data with a family of itemsets, many simple tasks become difficult. For instance, it is easy to deduce from Table 1.1 that we have one customer buying Beer and SodaLight. On other hand, deducing the

[1] This particular example is actually an infamous urban legend in the data mining community, http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html

number of customers buying Beer and SodaLight using Table 1.2 is surprisingly difficult. In fact, we will show that in general solving such a query problem takes an exponential amount of time. However, by imposing some restrictions we are able to ease the computational burden.

We can use the family of itemsets to estimate the frequencies of unknown itemsets. For instance, we can estimate that the number of customers buying Beer and SodaLight is $3 \times 3/5 = 9/5$. By computing the difference between the estimate and the actual support we receive a number telling how surprising is the itemset. This number can be used for ranking itemsets, that is, the more we are surprised by the support of the itemset, the more important the itemset is.

We can also use itemsets for computing the difference between two entire data sets. Assume that we have market basket data from two different months. We can compute some selected family of itemsets from the data sets and compute the difference by comparing the supports of the itemsets. Defining the distance between data sets allows us treat data sets, highly complex objects, as units and enables us to use traditional data mining tools.

Our last topic of the thesis deals with defining an intrinsic dimension for binary data. Traditionally, the dimension is understood to be the number of columns in the data set, usually a very high number. On the other hand, binary data is usually sparse and contains structure and hence is less complex than the number of columns would suggest. One possible way of defining intrinsic dimension is to use fractal dimension which is often used with real-valued data sets. We study how the fractal dimension can be applied to binary data.

Contributions of the thesis. This thesis is based on the publications listed in List of Publications. The content of the papers is discussed in greater detail at the beginning of Chapters 6, 7, and 8.

In Publication II we show that certain query problems, namely predicting itemsets from a set of known itemsets, are infeasible. Problems of the same type are well-studied [Cal04, Cal03], but we focus more on downward closed families of itemsets. We use a construction similar to [Coo90] to show that even the restriction of being downward closed does not prevent the problem from being infeasible.

The query theme continues in Publication III where we provide a novel optimisation scheme for solving a classic query problem [Hai65, Cal03, BSH04] by applying theorems from Markov Random Field theory [CDLS99].

In Publication IV we introduce a method for ranking itemsets by comparing the observed frequency against a Maximum Entropy estimate [PMS03]. Our work can be seen as an extension of the approach in [BMS97] in which the observed value is compared against the independence assumption.

In Publication I we study the idea of computing the distance between data sets via itemset frequencies. Similar approaches has been suggested for example in [HSM03]. However, we show that our distance possesses many theoretical properties, some of them being unique among alternative distances.

In Publication V we apply correlation dimension, a well-known concept [Bar88, Ott97], to binary data sets. Our investigation leads to a novel idea, called normalised correlation dimension, that takes into account the unique nature of binary data.

Contributions of the author. The author of the thesis is the sole author of Publications I, II, III, and IV.

In Publication V the author collaborated with the co-authors of the paper. The theoretical analysis of the paper is a joint work of the author and H. Mannila. The author is responsible for introducing the concept of the normalised correlation dimension. The author implemented all the experiments which were designed jointly with H. Mannila, T. Mielikäinen and A. Gionis. The paper was written jointly with the co-authors.

The structure of the thesis. The purpose of the subsequent chapters is to provide the needed background mathematics used in the articles so that a reader with a reasonable knowledge in statistics, calculus and algebra will able to follow the articles. The chapters also summarise the articles, emphasising heavily the theoretical side. We also review the research related to our ideas.

The first four chapters focus solely on background mathematics. They provide a sound base for the three remaining chapters which describe the key theorems in the articles. The style of the thesis is a traditional definition-theorem-example approach. The dependencies of theorems presented in introduction is given in Figure 1.1.

In Chapter 2 we introduce the notation and basic concepts related to binary data and itemsets. In Chapter 3 we introduce basic theory of Linear Programming. In Chapter 4 we study Markov Random Fields and in Chapter 5 we introduce Kullback-Leibler divergence and Maximum Entropy principle. In Chapter 6 we discuss the problem of predicting itemset frequencies from a known set of itemsets. We also introduce a rank measure for itemsets that uses information available from the sub-itemsets. In Chapter 7 we introduce the idea of using itemsets for computing a distance between two binary data sets. Finally, in Chapter 8 we use concepts from fractal theory for defining an effective dimension of a binary data set. Proofs for some theorems are provided in Appendix.

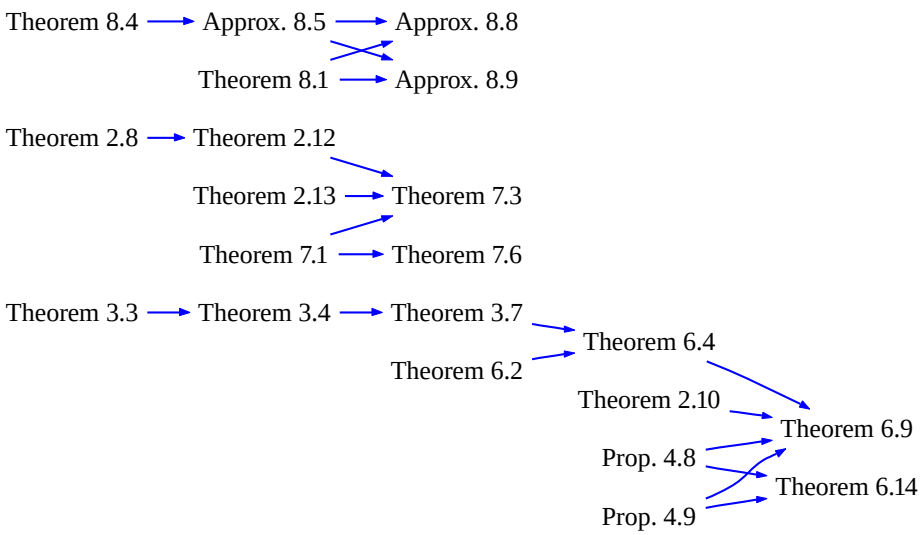


Figure 1.1: Dependency graph of theorems presented in the thesis.

Chapter 2

Binary data

> The world isn't run by weapons anymore, or energy, or money. It's run by little ones and zeroes, little bits of data. It's all just electrons.
>
> *Cosmo, Sneakers*

Extracting patterns from binary data is an active subfield of data mining. The most popular patterns are itemsets, sets of columns, that have unusually high concentration of 1s. Originally, itemsets were used as intermediate result for finding association rules [AMS+96, AIS93]. Nowadays, they are considered to be interesting patterns on their own and they have other applications in addition to association rules [MT96]. One major factor for the popularity of itemsets is their anti-monotonicity property which allows using level-wise mining algorithms, for example, Apriori [AMS+96].

The main theme of the thesis is to use itemsets as a surrogate for the original binary data. We can think that the mined itemsets act like a summary of the original data.

2.1 Binary data and Itemsets

We begin by defining basic concepts related to mining binary data. A *binary data set* is a finite multiset of a binary vectors of length K. In other words, it is a collection of elements from a space $\Omega = \{0,1\}^K$. A single element of a data set is called a *transaction*. The space Ω is called *sample space* and K is the *dimension* of Ω. The number of binary vectors in a data set D is denoted by $|D|$. It is customary to visualise binary data as a binary matrix of size $|D| \times K$

binary data set
transaction
sample space
dimension

	attributes		
	a_1	a_2	a_3
transactions	1	0	0
	0	1	0
	0	0	1
	1	1	0
	0	1	1
	0	1	1

Table 2.1: An example of a binary data set represented as a binary matrix. The data contains $K = 3$ attributes, namely a_1, a_2, and a_3, and 6 transactions.

(see Table 2.1 for such an example). In this matrix the rows are the transactions. Note that for our purposes the order of the transactions is irrelevant.

attribute

Let $\omega \in D$ be a randomly selected binary vector from D. We define an *attribute* a_i to be the boolean random variable representing the ith component of ω. If the data is represented as a matrix (see Table 2.1), then the attributes represent the columns of the matrix. We set $A = \{a_1, \ldots, a_K\}$ to be the collection of all attributes.

empirical distribution

In the context of this work it is more convenient to talk about distributions rather than data sets. We can represent a data set D by an *empirical distribution* p_D defined on a sample space Ω by setting

$$p_D\left(a_1 = \omega_1, \ldots, a_K = \omega_K\right) = \frac{\text{number of elements in } D \text{ equal to } \omega}{|D|},$$

where $\omega \in \Omega$ is a binary vector of length K, ω_i is the ith element of ω, and a_i is the corresponding attribute, a boolean random variable representing the ith dimension of the data set.

Example 2.1. Consider the data set given in Table 2.1. The data set has 3 attributes $A = \{a_1, a_2, a_3\}$ and 6 transactions. The corresponding empirical distribution is

$$\begin{aligned}
p_D\left(a_1 = 0, a_2 = 0, a_3 = 0\right) &= 0, & p_D\left(a_1 = 0, a_2 = 0, a_3 = 1\right) &= 1/6,\\
p_D\left(a_1 = 0, a_2 = 1, a_3 = 0\right) &= 1/6, & p_D\left(a_1 = 0, a_2 = 1, a_3 = 1\right) &= 1/3,\\
p_D\left(a_1 = 1, a_2 = 0, a_3 = 0\right) &= 1/6, & p_D\left(a_1 = 1, a_2 = 0, a_3 = 1\right) &= 0,\\
p_D\left(a_1 = 1, a_2 = 1, a_3 = 0\right) &= 1/6, & p_D\left(a_1 = 1, a_2 = 1, a_3 = 1\right) &= 0.
\end{aligned}$$

□

We use some convenient abbreviations. Given a distribution p defined on Ω, a set of attributes $B = \{b_1, \ldots, b_L\} \subseteq A$,[1] and a binary vector ω of length L, we use $p(B = \omega)$ for $p(b_1 = \omega_1, \ldots, b_L = \omega_L)$. By writing $p(B = 1)$ we mean $p(B = \omega)$, where ω is a vector containing only 1s.

Example 2.2. Consider the distribution p_D given in Example 2.1. Then, for instance, we have

$$p_D(a_1 a_2 = 1) = 1/6, \quad p_D(a_1 = 1) = 1/3,$$
$$p_D(a_2 a_3 = 1) = 1/3, \quad p_D(a_3 = 1) = 1/2.$$

□

One of the most useful properties of binary data is that we can apply Boolean logic. Assume that we are given a Boolean formula F defined on (a subset of) the attributes A. Let $S_F : \Omega \to \{0, 1\}$ be an *indicator function*, *indicator function*

$$S_F(\omega) = \begin{cases} 1 & \omega \text{ satisfies } F, \\ 0 & \text{otherwise.} \end{cases}$$

Given a distribution p, the *frequency* θ_F of F is then the probability of S_F being 1, that is, it is the mean $\theta_F = \mathrm{E}_p[S_F]$. *frequency*

Example 2.3. Consider the data set given in Table 2.1. Consider the formulae $F_1 = a_1$, $F_2 = a_2$, $F_3 = a_3$, $F_4 = a_1 \wedge a_2$, and $F_5 = a_2 \wedge a_3$. For example, the indicator function of F_5 is

$$S_{F_5}(\omega) = \begin{cases} 1 & \omega_2 = \omega_3 = 1, \\ 0 & \text{otherwise.} \end{cases}$$

There are 2 transactions that satisfy F_5, namely transactions 5 and 6, hence the frequency for F_5 is $\theta_{F_5} = \frac{1}{3}$. Similarly, the frequencies for the rest of the formulae are $\theta_{F_1} = \frac{1}{3}$, $\theta_{F_2} = \frac{4}{6}$, $\theta_{F_3} = \frac{1}{2}$, and $\theta_{F_4} = \frac{1}{6}$. □

Given a formula F, a frequency θ, we say that a distribution p *satisfies the frequency* θ if $\mathrm{E}_p[S_F] = \theta$. We can easily extend these definitions for multiple Boolean formulae. If we have a family $\mathcal{F} = \{F_1, \ldots, F_N\}$ of Boolean formulae, then the frequencies $\theta_{\mathcal{F}} = (\theta_{F_1}, \ldots, \theta_{F_N})$ is a vector containing N elements. Similarly, the indicator function is $S_{\mathcal{F}} : \Omega \to \{0, 1\}^N$, defined as $S_{\mathcal{F}}(\omega) = (S_{F_1}(\omega), \ldots, S_{F_N}(\omega))$. *satisfies the frequency*

[1] We will rather use $\{b_1, \ldots, b_L\}$ instead of cumbersome $\{a_{i_1}, \ldots, a_{i_L}\}$.

Example 2.4. We continue Example 2.3. Consider a family of formulae

$$\mathcal{F} = \{a_1, a_2, a_3, a_1 \wedge a_2, a_2 \wedge a_3\}.$$

Example 2.3 tells us that the corresponding frequencies are $\theta_{\mathcal{F}} = \frac{1}{6}(2, 4, 3, 1, 2)$. We can easily see that the empirical distribution p_D given in Example 2.1 satisfies the frequencies since $\mathrm{E}_{p_D}[S_{\mathcal{F}}] = \theta_{\mathcal{F}}$. The distribution p_D is not the only distribution satisfying $\theta_{\mathcal{F}}$. In fact, there are infinite number of such distributions. For example, a distribution q defined as

$$\begin{array}{llll} q(a_1 = 0, a_2 = 0, a_3 = 0) = 1/6, & & q(a_1 = 0, a_2 = 0, a_3 = 1) = 1/6, \\ q(a_1 = 0, a_2 = 1, a_3 = 0) = 1/6, & & q(a_1 = 0, a_2 = 1, a_3 = 1) = 1/6, \\ q(a_1 = 1, a_2 = 0, a_3 = 0) = 1/6, & & q(a_1 = 1, a_2 = 0, a_3 = 1) = 0, \\ q(a_1 = 1, a_2 = 1, a_3 = 0) = 0, & & q(a_1 = 1, a_2 = 1, a_3 = 1) = 1/6. \end{array}$$

also satisfies the frequencies. □

itemsets

Our special interest lies in conjunctive formulae, that is, formulae having the form $b_1 \wedge \cdots \wedge b_L$. Such formulae are called *itemsets* and they are usually represented by a subset of attributes $B = \{b_1, \ldots, b_L\}$. Often the condensed notation is used $B = b_1 \cdots b_L$. Note that if B is an itemset, then the frequency θ_B can be expressed in a form $p(B = 1)$.

Itemsets possess many useful properties, the most important one is the anti-monotonic property:

Proposition 2.5. *Assume that we are given two itemsets U and V such that $U \subseteq V$, then the frequencies obey $\theta_U \geq \theta_V$.*

σ-frequent
downward closed
antimonotonic

One of the largest research areas in binary data mining is retrieving σ-frequent itemsets. In other words, given a data set (or a distribution), the frequency θ_B of an itemset B is *σ-frequent* if $\theta_B \geq \sigma$. A family $\mathcal{F}$ of itemsets is said to be *downward closed* or *antimonotonic* if each subset of each member of $\mathcal{F}$ is also included in $\mathcal{F}$. Proposition 2.5 says that a family containing all σ-frequent itemsets is downward closed.

Example 2.6. The $\frac{1}{3}$-frequent itemsets in Example 2.3 are $\emptyset$, a_1, a_2, a_3, and a_2a_3. Note that this family is downward closed. □

Our main interest is to study the properties of downward closed families of itemsets. The fundamental property of such families as that we are able to express the frequency of a Boolean formula as a linear combination of frequencies of itemsets. This is illustrated in the following example.

Example 2.7. Let $B = a_1 \vee a_2$ be a disjunctive formula of two attributes. Let $\mathcal{F}$ be a family of itemsets $\mathcal{F} = \{a_1, a_2, a_1a_2\}$. We can express the indicator function S_B as a linear combination of indicator functions of itemsets:

$$S_B = S_{a_1} + S_{a_2} - S_{a_1a_2}.$$

Given a distribution p we can express the frequency of H as

$$\theta_B = \mathrm{E}_p[S_B] = \mathrm{E}_p[S_{a_1}] + \mathrm{E}_p[S_{a_2}] - \mathrm{E}_p[S_{a_1a_2}] = \theta_{a_1} + \theta_{a_2} - \theta_{a_1a_2}.$$

□

The following theorem states the property of downward families of itemsets used in the previous example.

Theorem 2.8 (Proposition 1 in [MT96])**.** *Let $\mathcal{F}$ be a downward closed family of itemsets along with the frequencies $\theta_{\mathcal{F}}$. Let p be a distribution satisfying $\theta_{\mathcal{F}}$. Let B be a Boolean formula and let S_B be its indicator function. If we assume that B depends only on variables that are contained in some member of $\mathcal{F}$, then there is a set of constants $\{u_C\}$ not depending of $\theta_{\mathcal{F}}$ such that*

$$\theta_B = \mathrm{E}_p[S_B] = \sum_{C \subseteq B} u_C \theta_C,$$

where C ranges over all subsets of B.

More generally, if $\mathcal{B} = \{B_1, \ldots, B_M\}$ is a collection of Boolean formulae such that a formula B_i depends only on variables containing in some member of $\mathcal{F}$, then there is a matrix U of size $|\mathcal{B}| \times |\mathcal{F}|$ not depending of $\theta_{\mathcal{F}}$ such that

$$\theta_{\mathcal{B}} = \mathrm{E}_p[S_{\mathcal{B}}] = U\theta_{\mathcal{F}}.$$

The immediate corollary of Theorem 2.8 states that certain marginal distributions obtained from a distribution satisfying the frequencies are unique. We illustrate this with the following toy example.

Example 2.9. Consider two attributes a and b and let their frequencies be $\theta_a = 0.5$ and $\theta_b = 0.6$. Assume two distributions p and q being

$$\begin{aligned} p\,(a = 1, b = 1) &= 0.5, \quad p\,(a = 0, b = 1) = 0.1, \\ p\,(a = 1, b = 0) &= 0.0, \quad p\,(a = 0, b = 0) = 0.4 \end{aligned}$$

and

$$\begin{aligned} q\,(a = 1, b = 1) &= 0.3, \quad q\,(a = 0, b = 1) = 0.3, \\ q\,(a = 1, b = 0) &= 0.2, \quad q\,(a = 0, b = 0) = 0.2. \end{aligned}$$

Although p and q are different distributions, they both satisfy the itemsets a and b. This implies that

$$p(a=1) = q(a=1), \quad p(a=0) = q(a=0),$$
$$p(b=1) = q(b=1), \quad p(b=0) = q(b=0).$$

In other words, p and q are equal when they are marginalised to a (or to b). □

We generalise the preceeding example in the following corollary of Theorem 2.8.

Corollary 2.10. *Let $\mathcal{F}$ be a downward closed family of itemsets and let $\theta_{\mathcal{F}}$ be the corresponding frequencies. Let $B = b_1 \cdots b_L \in \mathcal{F}$ be an itemset from $\mathcal{F}$. If p and q satisfy the frequencies $\theta_{\mathcal{F}}$, then $p(B = \omega) = q(B = \omega)$ for any ω. In other words, the distribution obtained by ignoring the attributes outside B from a distribution p is unique.*

This corollary combined with the theory of Markov Random Fields (Chapter 4) will play a crucial role in Chapter 6.

parity formula

In addition to itemsets, there are also other families of Boolean functions that satisfy Theorem 2.8. A *parity formula* $B = b_1 \oplus \cdots \oplus b_L$, where $\oplus$ is the XOR-operator, returns 1 if and only if and odd number of the variables b_i are equal to 1. We can express parity functions as a linear combination of conjunctive functions and visa versa.

Example 2.11. Let us continue Example 2.3. Consider the following parity functions $H_1 = a_1$, $H_2 = a_2$, $H_3 = a_3$, $H_4 = a_1 \oplus a_2$, and $H_3 = a_2 \oplus a_3$ and let their frequencies be $\theta_{\mathcal{H}} = \frac{1}{6}(2, 4, 3, 4, 3)$. We know that

$$S_{a \oplus b} = S_a + S_b - 2S_{ab}.$$

This implies that

$$\theta_{a \oplus b} = \mathrm{E}[S_{a \oplus b}] = \mathrm{E}[S_a] + \mathrm{E}[S_b] - 2\mathrm{E}[S_{ab}] = \theta_a + \theta_b - 2\theta_{ab}.$$

We can restate this connection by using vector notation. Let U be

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -2 & 0 \\ 0 & 1 & 1 & 0 & -2 \end{bmatrix}.$$

Recall that the frequencies in Example 2.3 were $\theta_{\mathcal{F}} = \frac{1}{6}(2, 4, 3, 1, 2)$. By multiplying $\theta_{\mathcal{F}}$ with U we obtain a vector

$$U\theta_{\mathcal{F}} = \frac{1}{6}(2, 4, 3, 4, 3)$$

that corresponds to the parity frequencies $\theta_{\mathcal{H}}$. It is important to note that U is invertible, that is, we can transform parity frequencies into itemset frequencies by a linear transformation. □

The following proposition generalises the preceeding example.

Proposition 2.12. *Let $\mathcal{F}$ be a downward closed family of itemsets. Define*

$$\mathcal{H} = \{b_1 \oplus \cdots \oplus b_L;\ B = \{b_1, \ldots, b_L\} \in \mathcal{F}\}$$

to be the set of corresponding parity formulae. Then there is an invertible matrix U such that $\theta_{\mathcal{H}} = U\theta_{\mathcal{F}}$.

The proof of the theorem is provided in Appendix A.1.

The theorem tells us that once we know the frequencies for itemsets, then we can deduce by linear transformation to parity formulae without making any additional queries from the data set. We can also deduce itemsets from parity formulae. In other words, parity formulae and itemsets contain essentially the same information. This idea turns out to be important in Chapter 7 in which we study the distances between data sets.

The reason why we are interested in parity formulae is that the frequencies are very regular when computed from the uniform distribution. This regularity enables us to ease the computational burden when solving the distances discussed in Chapter 7.

Proposition 2.13 (Lemma 8 in Publication I)**.** *Let B_1 and B_2 be two parity formulae such that $B_1 \neq B_2$. Let p be the uniform distribution defined on Ω. Then, $\mathrm{E}_p[S_{B_1}] = \frac{1}{2}$ and $\mathrm{E}_p[S_{B_1} S_{B_2}] = \frac{1}{4}$.*

Chapter 3

Linear Programming for Predicting Itemset Frequencies

Our need for understanding Linear Programming (LP) stems from Chapter 6. In that chapter we study the problem of deducing the frequency of an itemset from a known set of itemsets. We provide a sneak peak of this scenario in the following example.

Example 3.1. Assume that we are given two attributes, namely a and b. Assume also that we have the frequency for a equal to 0.5 and, similarly, 0.6 for b. We wish to find a distribution p having the highest possible frequency $p(ab = 1)$ for the itemset ab while at the same time having $p(a = 1) = 0.5$ and $p(b = 1) = 0.6$. In this particular case the distribution is equal to

$$p(a = 1, b = 1) = 0.5, \quad p(a = 0, b = 1) = 0.1,$$
$$p(a = 1, b = 0) = 0.0, \quad p(a = 0, b = 0) = 0.4.$$

Hence, the maximum frequency for ab is 0.5. □

We study the problems similar to the one given in Example 3.1 in Chapter 6. It turns out that these problems can be solved with LP.

Linear programming is perhaps the most classical constrained optimisation problem. The modern theory of LP was developed during the decades 1930–1950, however, the roots go as far as the 18th century. In those decades, the need for solving optimisation problems sprang from the industrial and military management, especially in the United States of America. World War II and the Great Depression had influences on these developments, as well as rapid development of computers. Perhaps the most important important event, a breakthrough, is the invention of Simplex, an algorithm for solving linear program, by George

Dantzig in 1947. Another famous event was the conference at the University of Chicago in 1949, arranged by Tjalling Koopmans under the sponsorship of the Cowles Commission for Research in Economics. At the conference the power of linear programming was demonstrated through different military and industrial applications. The articles of the conference are available in [Koo51]. A reader interested in the history of the development of Linear Programming Theory is advised to chapters 1–2 in [Dan63] and the introduction in [Koo51].

In this chapter we represent rudimentary theory of Linear Programming. In Section 3.1 we LP and analyse its solutions. In Section 3.2 we review some of widely known solving algorithms.

3.1 Theory

In this section we define the linear program and using geometrical intuition explain how the solution of the problem is found.

In the standard form of linear program, we are given a vector $c \in \mathbb{R}^N$, a $M \times N$ matrix A, and a vector $b \in \mathbb{R}^M$. Linear program involves in finding real vector $x \in \mathbb{R}^N$ such that the following optimisation problem is solved:

$$\begin{aligned} \min\, & c^T x \\ & Ax = b \\ & x \geq 0 \end{aligned} \tag{3.1}$$

feasible set
standard form

In other words, we are asked to minimise $c^T x$ under certain constraining conditions. A set containing the vectors x satisfying the constraints is called - *feasible set*. The problem given in Eq. 3.1 is known as LP in *standard form*. There are alternative ways of stating the same problem but they can be polynomically reduced into the standard form [PS98, Section 2.1].

Example 3.2. Let us consider the following linear program:

$$\begin{aligned} \min\, & c_1 x_1 + c_2 x_2 + c_3 x_3 \\ & x_1 + x_2 = 1 \\ & x_1 + 2x_3 = 1 \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Here we have denoted x_i as the ith component of a vector $x \in \mathbb{R}^3$ and c_i as the ith component of a vector $c \in \mathbb{R}^3$. The feasible set of this program is a segment lying in $\mathbb{R}^3$ and having $(1, 0, 0)$ and $\left(0, 1, \frac{1}{2}\right)$ as end points. The program is in standard form. By combining the conditions we can rewrite the problem in a simpler form

$$\min \left(c_1 - c_2 - \frac{1}{2} c_3 \right) x_1 + 2$$
$$1 \geq x_1 \geq 0.$$

We see that the solution is attained, depending on the sign of $c_1 - c_2 - \frac{1}{2}c_3$, either at $(1, 0, 0)$ or $\left(0, 1, \frac{1}{2}\right)$, the end points of the feasible set. If $c_1 - c_2 - \frac{1}{2}c_3 = 0$, then c is orthogonal with the feasible set and any point has the minimal value. □

From now on, we will assume that the feasible set is not empty and that there exists a finite solution.

We have seen from Example 3.2 that the optimal solution was always a corner point (a vertex) of the feasible set. In general, the feasible set is a polytope lying in $\mathbb{R}^N$. Scaling the vector c does not change the outcome of LP, hence we can assume that the length of c is 1. Let x be a vector in the feasible set. Then $c^T x$ is the length of the orthogonal projection of x into c (See Figure 3.1).

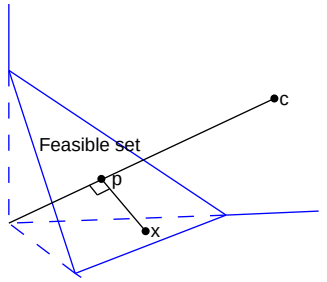


Figure 3.1: A geometrical interpretation of Linear Programming. The feasible set in this case is a triangle. If we assume that the vector c has length 1 and that x is a point from the feasible set, then $c^T x$ is the length of vector p, the orthogonal projection of x into c.

This geometrical interpretation reveals us the following theorem:

Theorem 3.3 (Theorems 2.4 and 2.6 in [PS98])**.** *There exists a vertex x of the feasible set of given LP such that x results the optimal value of LP.*

There exists a clever algebraic way of expressing the vertices of the feasible set. Consider the $M \times N$ matrix A in Eq. 3.1. We may safely assume that $N \geq M$ and that the rank of A is M. Otherwise, we can reduce the number of constraints, so that the condition holds.

Assume now that we are given a set $U = \{u_i\} \subseteq \{1, \ldots, N\}$ of M integers. Let A_U be a submatrix of A containing only the columns corresponding to U. Assume that A_U is invertible. Let $x_{u_i} = A_U^{-1} b_i$, and 0 for the rest entries. Such x is called *basic solution*. The vector x satisfies $Ax = b$ in Eq. 3.1, but it is not guaranteed that x has only positive elements. However, if this is the case, then x is called *basic feasible solution* (BFS).

basic solution

basic feasible solution

The following theorem shows that the vertices of the feasible set and BFS's are equivalent concepts.

Theorem 3.4 (Theorems 2.4 and 2.6 in [PS98])**.** *A basic feasible solution x is a vertex of the feasible set. In the other direction, if x is a vertex, then there is U such that $x_{u_i} = A_U^{-1} b_i$, and 0 for the rest entries. Consequently, there exists a basic feasible solution producing the optimal value for LP.*

Example 3.5. Consider the standard form of LP and assume that we have $N = 3$, $M = 1$, $A = [1, 2, 3]$, and $b = 1$. The feasible set is a triangle having the vertices $(1, 0, 0)$, $\left(0, \frac{1}{2}, 0\right)$, and $\left(0, 0, \frac{1}{3}\right)$. Let $U = \{1\}$, $V = \{2\}$, and $W = \{3\}$. We have $A_U = 1$, $A_V = 2$, $A_W = 3$. Hence the basic solutions are $x = (1, 0, 0)$, $y = \left(0, \frac{1}{2}, 0\right)$, and $z = \left(0, 0, \frac{1}{3}\right)$. These solutions are all feasible and they correspond to the vertices of the feasible area. □

Example 3.6. Let us consider Example 3.2. The condition variables in this case are

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Let $U = \{1, 2\}$, $V = \{2, 3\}$, and $W = \{1, 3\}$. The submatrices are

$$A_U = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad A_V = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad A_W = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}.$$

The corresponding basic solutions are $x = (1, 0, 0)$, $y = \left(0, 1, \frac{1}{2}\right)$, and $z = (1, 0, 0)$. We see that these solutions are feasible and they correspond to the vertices of the feasible set. □

We have interest in investigating the complexity problems involved with linear programming. Hence, Theorem 3.4 has the following important corollary:

Theorem 3.7 (Lemma 2.1 in [PS98])**.** *There is a solution x for LP having only M non-zero elements. Also, if A and b contains rational elements expressed with L bits, then we can express a non-zero element of x in $f(M, L)$ bits, where f is a function of polynomial growth.*

In complexity theory we are often required to provide a polynomial-size certificate. Theorem 3.7 allows us to use the optimal solution x as a certificate (if we needed to) because we can express it in polynomial space.

3.2 Algorithms Solving Linear Programs

While the theory of Linear Programming is straightforward and simple, solving LP in practice is a complex task.

Simplex

The most known algorithm for solving LP is Simplex [Dan51]. Roughly put, the algorithm solves the problem by finding the best basic feasible solution using a hill-climbing approach. The algorithm is easy to implement and it works fairly well in practice. However, there is a major drawback: The number of basic feasible solutions (vertices of the feasible set) can be exponential. We can construct a problem such that finding the optimal solution using Simplex requires an exponential number of steps [PS98, Section 8.6].

Ellipsoid Algorithm

The question whether LP can be solved in polynomial time remained open until Ellipsoid Algorithm was introduced in [Kha79]. The algorithm does solve the problem in polynomial time but it is highly complex and cumbersome. Hence, while this algorithm has important theoretical value, it is not used in practice.

Primal-Dual Path-Following Algorithm

A modern polynomial-time algorithm used for solving LP is called Primal-Dual Path-Following Algorithm [BSS93, Section 9.5]. The idea is that we remove the condition $x \geq 0$ from Eq. 3.1 and instead of minimising $c^T x$ we minimise $c^T x - \mu \sum_j \log x_j$, where μ is some small constant. We can show that by letting μ approach 0, the solution of this modified problem approaches the optimal solution of the original LP problem.

Chapter 4

Markov Random Field Theory for Optimising Prediction of Itemset Frequencies

In Chapter 2 we considered distributions defined over a sample sample Ω, the set of all binary vectors of length K. Although, these distributions have finite number of elements, they are too large to work with. However, we are able to reduce the number of free parameters by using the techniques from Markov Random Field (MRF) theory.

Our interest in MRFs is somewhat unorthodox. We are mainly interested in decomposable distributions. Such distributions can be expressed efficiently — the MRF theory states that we need only the cliques of a certain graph. In Chapter 6 we use these decomposable distributions, and especially Proposition 4.9, to drastically ease the computational burden of one of our main tasks considered in this work. In order to justify our need for MRF theory even further we consider the following example.

Example 4.1. Let us consider Example 3.1. We have two attributes a and b with the corresponding the frequencies 0.5 and 0.6, respectively. The maximum value for the frequency of ab is 0.5. Now consider adding a third attribute c with a frequency 0.3. We wish to find a distribution p maximising the frequency for ab and having $p(a = 1) = 0.5$, $p(b = 1) = 0.6$, $p(c = 1) = 0.3$. It turns out that the maximum frequency of ab is again 0.5. In fact, the frequency of c *does not play any role* in maximising the frequency for ab. To see this, note that we can expand any distribution satisfying itemsets a and b into a full joint distribution satisfying itemsets a, b, and c. □

The reason that we were able to prune out the attribute c in Example 4.1

is that we did not have any constraints on itemsets ac and bc. We will see in Chapter 6 that MRF Theory provides a neat framework for identifying the attributes that can be pruned.

Frameworks for graph-based modelling in order to represent efficiently multivariate distributions were developed in the late 1980's [Pea88, LS90]. The causality between the attributes is expressed by a directed (acyclic) graph. Such concepts had led into research area called graphical modelling. Markov Random Fields (MRF) can be considered as undirected version of Bayesian networks.

4.1 Theory

In this section we introduce the basic concepts of Markov Random Fields. We explain how to obtain junction tree from the dependency graph and using the junction tree obtain a decomposition for certain distributions. For more detailed description on Markov Random Fields see e.g., [CDLS99].

A major part of Markov Random Field theory, and the part in which we are interested, is the way of expressing distributions effectively. To illustrate the situation, let us provide a simple example:

Example 4.2. Consider K binary variables a_i, $i = 1, \ldots, K$. Without any assumptions, to express a joint distribution p of these variables, we need to store 2^K elements. On other hand, if we assume that the variables are independent, then we can express the distribution p as

$$p(a_1, \ldots, a_K) = p(a_1) \cdots p(a_K). \tag{4.1}$$

To express such a distribution, we need to store only K elements. □

Generally speaking, consider that we have K binary random variables a_i In this case, a joint distribution contains 2^K elements. However, if we make some assumptions (very similar to the independence assumption in Example 4.2), MRF theory allows us to express the distributions more succinctly.

Our interest is to study decomposable distributions. We have already seen one group of such distributions in Example 4.2 but the independence model is very strict. We will consider more general case by applying MRF concept.

Let us consider an undirected graph $G = (V, E)$ containing K nodes $V = \{v_i\}$; a node v_i represents the random variable a_i. We say that nodes v_i and v_j are *connected nodes* connected if there is an edge (v_i, v_j). The edges of the graph represent the dependencies between the variables a_i. Roughly put, an edge (v_i, v_j) tells us that we have some dependency between a_i and a_j and hence these variables should not be split in different components. Note that if the graph has no edges, then

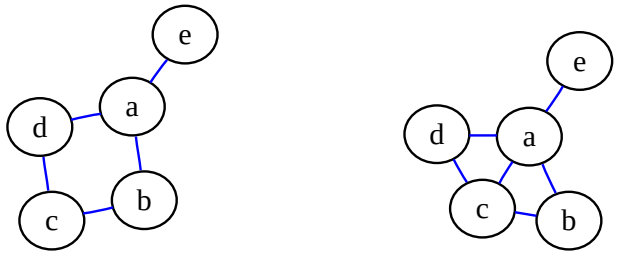
this property is equal to the independence assumption demonstrated in Example 4.2. Our goal is to decompose p into components similarly to decomposition demonstrated in Eq. 4.1.

Our next goal is to make sure that the dependency graph is regular enough. In order to this, we need to introduce some concepts from graph theory. A *cycle* is a set of nodes $\{v_1, \ldots, v_N\} \subseteq V$ such that v_j and v_{j+1} are connected, and v_1 and v_N are connected. Any possible additional edges between the nodes $\{v_1, \ldots, v_N\}$ are called *chord* edges. A cycle without chord edges is called *chordless*. A graph is called *triangulated* if there are no chordless cycles.

cycle

chord
chordless
triangulated

Example 4.3. Consider the graph given in Figure 4.1(a). There is a chordless cycle a–b–c–d. This cycle can be removed, for example, by adding a chord edge a–c. We see that the resulting graph (Figure 4.1(b)) contains no chordless cycles and it is therefore triangulated. Note that this is not the only possible triangulation, adding a chord b–d results also a triangulated graph.



(a) A non-triangulated graph (b) A triangulated graph

Figure 4.1: An example of a non-triangulated graph and a graph resulting from a triangulation process. In the left graph there is a chordless cycle a–b–c–d. This cycle is removed in the right graph by adding a chord a–c.

□

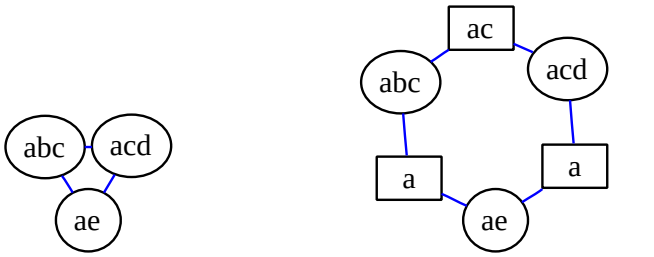
We will see that a triangulated graph possesses useful properties but let us consider how we can triangulate the graph G. The idea is to find the chordless cycles and add the missing edges until no chordless cycle can be found. A simple algorithm, called ELIMINATION Algorithm, iteratively picks a node v_i, connects its immediate neighbours, and delete the node [CDLS99, Section 4.4.1]. The graph G with the edges added during the elimination process is guaranteed to be triangulated.

Elimination Algorithm

Let us assume that G is triangulated. Consider an undirected graph H having a node C_i for each clique of G. Two nodes C_i and C_j are connected in H if they

clique graph share a node in G. The graph H is called the *clique graph* of G. Let us fix $C_i \in V(H)$ and $C_j \in V(H)$ and assume that they are connected by an edge $S \in E(H)$. We associate a set of mutual nodes (lying in $V(G)$) $C_j \cap C_j$ to the *separator* edge S. This set is called a *separator*.

Example 4.4. Let us continue Example 4.3. The cliques of the graph in Figure 4.1(b) are abc, acd, and ae. The separators are $abc \cap acd = ac$, $abc \cap ae = a$, and $acd \cap ae = a$. The resulting clique graph H (given in Figure 4.2) is a triangle. In the figure, the oval nodes are the cliques and the square nodes are the separators.



(a) A clique graph (b) A clique graph with separators

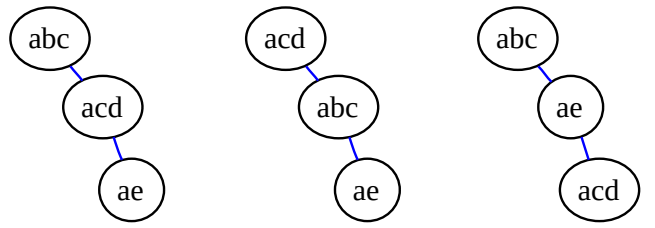
Figure 4.2: A clique graph of the graph G given in Figure 4.1(b). The oval nodes are the cliques of G and the square nodes are the separators, that is, intersections of the immediate cliques.

□

Assume for simplicity that the clique graph H is connected. Consider a spanning tree T of H. Select two cliques C_i and C_j sharing a mutual node v. We *running intersection property* say that C_i and C_j have *running intersection property* if the intermediate cliques connecting C_i and C_j in T contain also v. If this property holds for any pair of cliques C_i and C_j sharing a mutual node, then we say that the tree T has the *junction tree* running intersection property. Such a tree is called *junction tree*. There can be several junction trees and not all spanning trees are junction trees.

Example 4.5. We continue Example 4.4. There are three possible spanning trees (illustrated in Figure 4.3) of the clique graph given in Figure 4.2. However, only two of these are junction trees. The tree in Figure 4.3(c) does not satisfy the running intersection property because the node c is not included in ae, a clique lying on a path between the cliques abc and acd.

□



(a) A junction tree (b) A junction tree (c) A spanning tree

Figure 4.3: Spanning trees of the clique graph given in Figure 4.2. The tree in Figure 4.3(c) does not satisfy the running intersection property since the node c is not included in ae. The left and the centre tree are junction trees.

The following states that junction trees always exist, if G is triangulated.

Theorem 4.6 (Theorems 4.4 and 4.6 in [CDLS99])**.** *If the graph G is triangulated and connected, then there exists a junction tree T, that is, a spanning tree of the clique graph H satisfying the running intersection property.*

Let T be a junction tree. Let $C = \{C_1, \ldots, C_N\}$ be the set of cliques (the nodes of the clique graph H) and let $\{S_1, \ldots, S_{N-1}\}$ be the separators of the junction tree. We say that a distribution p is decomposable with respect to T if p has a form

decomposable distribution

$$p(a_1, \ldots, a_K) = \prod_{i=1}^{N} p(C_i) \Big/ \prod_{j=1}^{N-1} p(S_j).$$

There are several reasons why we are interested in decomposable distributions. The first reason is that storing such distributions requires less space as the following example demonstrates.

Example 4.7. We continue Examples 4.3–4.5. Assume that we have 5 binary variables and that the dependency graph is given in Figure 4.1. The number of elements in a general joint distribution is $2^5 = 32$. However, if p is decomposable with respect to tree given in Figure 4.3(a), then p has the form

$$\frac{p(a,b,c)\, p(a,c,d)\, p(a,e)}{p(a,c)\, p(a)}.$$

Since we can obtain the separator components $p(ac)$ and $p(a)$ from the clique components, we need to store only the clique components. The total number of elements in the components is $2^3 + 2^3 + 2^2 = 20$. □

The second reason is seen in a way the decomposition takes into account the dependency edges of the graph G. If we have a fully connected set of nodes W in G, then there is a clique that contains W. The following theorem follows.

Proposition 4.8. *Let G be a graph and let T be its junction tree. Assume that W is a set of fully connected nodes in G. Then there is a component C in T such that $W \subseteq C$. Consequently, if the distribution is decomposed using T, there there is a component in the decomposition containing W.*

Final (and the most important) reason is in the way we can compose a joint distribution from the components

Proposition 4.9. *Let T be a junction tree. Let $C = \{C_1, \ldots, C_N\}$ be the corresponding set of cliques and let $\{S_1, \ldots, S_{N-1}\}$ be the separators of T. Assume that for each cliques C_i we have a distribution p_i defined on C_i. Assume also that for two connected (in T) cliques C_i and C_j the components p_i and p_j are equivalent when marginalised to the separator. Then, there is a decomposable distribution p with respect to T such that marginalising p to C_i produces p_i. Moreover, p is equal to*

$$p(a_1, \ldots, a_K) = \prod_{i=1}^{N} p_i(C_i) \Big/ \prod_{j=1}^{N-1} p_j(S_j).$$

The proof of the theorem is provided in Appendix A.2.

Example 4.10. We continue Example 4.7. Consider the following three distributions

$$\begin{aligned}
p_1(abc = (0,0,0)) &= 0/8, & p_1(abc = (0,0,1)) &= 3/8,\\
p_1(abc = (0,1,0)) &= 1/8, & p_1(abc = (0,1,1)) &= 1/8,\\
p_1(abc = (1,0,0)) &= 0/8, & p_1(abc = (1,0,1)) &= 2/8,\\
p_1(abc = (1,1,0)) &= 0/8, & p_1(abc = (1,1,1)) &= 1/8,
\end{aligned}$$

$$\begin{aligned}
p_2(acd = (0,0,0)) &= 1/8, & p_2(acd = (0,0,1)) &= 0/8,\\
p_2(acd = (0,1,0)) &= 2/8, & p_2(acd = (0,1,1)) &= 2/8,\\
p_2(acd = (1,0,0)) &= 0/8, & p_2(acd = (1,0,1)) &= 0/8,\\
p_2(acd = (1,1,0)) &= 2/8, & p_2(acd = (1,1,1)) &= 1/8,
\end{aligned}$$

and

$$\begin{aligned}
p_3(ae = (0,0)) &= 3/8, & p_3(ae = (0,1)) &= 2/8,\\
p_3(ae = (1,0)) &= 2/8, & p_3(ae = (1,1)) &= 1/8.
\end{aligned}$$

The distributions p_3 and p_2 match at the separator a

$$\begin{aligned}
p_3(a = 0) = p_2(a = 0) = 5/8,\\
p_3(a = 1) = p_2(a = 1) = 3/8.
\end{aligned}$$

Also, the distributions p_1 and p_2 match at the separator ac

$$\begin{aligned}
p_1(ac = (0,\,0)) &= p_2(ac = (0,\,0)) = 1/8,\\
p_1(ac = (0,\,1)) &= p_2(ac = (0,\,1)) = 4/8,\\
p_1(ac = (1,\,0)) &= p_2(ac = (1,\,0)) = 0/8,\\
p_1(ac = (1,\,1)) &= p_2(ac = (1,\,1)) = 3/8.
\end{aligned}$$

Theorem 4.9 now states that there is a distribution p having the marginals

$$p(a,b,c) = p_1(a,b,c)\,, \quad p(a,c,d) = p_2(a,c,d)\,, \quad p(a,e) = p_3(a,e)\,.$$

□

In other words, as long as the component distributions match at the separators, we can join them to create a joint distribution. The total number of variables in the components is usually drastically smaller than the number of variables in a full joint distribution. These savings help us to reduce the computational burden of the problems introduced in Chapter 6.

Chapter 5

Kullback-Leibler Divergence for Ranking Itemsets

In this chapter we briefly review the theory related to information entropy and Kullback-Leibler divergence. The idea of information entropy in the context of communication theory was introduced introduced by Shannon in [Sha48], although the concept of thermodynamical entropy already existed in physics. Kullback and Leibler introduced Kullback-Leibler divergence in [KL51].

A concept that we are particularly interested in is the principle of maximum entropy which is discussed in Section 5.2. The idea was adopted from physics by Jaynes in [Jay57].

Concepts introduced in this chapter are used in Section 6.5. In that section we consider a ranking measure of itemsets by comparing the prediction made by Maximum Entropy against the actual value obtained from the data set. The difference is measured using Kullback-Leibler divergence. To motivate this chapter we provide the following sneak-peak example.

Example 5.1. We continue Example 3.1. Assume that we have two attributes a and b with the corresponding the frequencies 0.5 and 0.6, respectively. What value for the frequency of ab we expect to have? One approach is to consider the independence model, that is, the distribution p defined as

$$p(a=1,b=1)=0.5\times 0.6,\quad p(a=0,b=1)=(1-0.5)\times 0.6,$$
$$p(a=1,b=0)=0.5\times(1-0.6),\quad p(a=0,b=0)=(1-0.5)\times(1-0.6).$$

This is, in fact, the Maximum Entropy distribution satisfying the itemsets a and b. Now consider seeing the actual frequency of ab and assume that it is equal to

0.5. The empirical distribution is equal in this case

$$q(a=1,b=1)=0.5, \quad q(a=0,b=1)=0.1,$$
$$q(a=1,b=0)=0.0, \quad q(a=0,b=0)=0.4.$$

Our measure for the significance of ab is the difference between q and p, which is measured using Kullback-Leibler divergence. In this particular case, we have $\mathrm{KL}(q;\, p) = 0.42$. □

5.1 Definitions

In this section we will define Kullback-Leibler divergence, an asymmetric distance between two distributions, and the related quantity called entropy.

The distributions in this chapter are defined on the sample space Ω, a set of binary vectors of length K. However, we should point out that the concepts of this chapter work directly with any other finite space. The finiteness of Ω enables us to define a distribution as a function $p : \Omega \to [0, 1]$ mapping a point $\omega \in \Omega$ to a number between 0 and 1 such that $\sum_{\omega\in\Omega} p(\omega) = 1$. This naïve approach is adequate for our purposes but it should be kept in mind that the concepts introduced in this section can be expanded to arbitrary distributions.

entropy

We define the *entropy* of a distribution p to be

$$\mathcal{E}(p) = -\sum_{\omega\in\Omega} p(\omega)\log p(\omega).$$

Here we use the natural logarithm and we use the convention $0 \cdot \log 0 = 0$.

Example 5.2. Consider the distribution q given in Example 5.1. The entropy is equal to

$$\mathcal{E}(q) = -0.5\log(0.5) - 0.1\log(0.1) - 0\log(0) - 0.4\log(0.4) = 0.94.$$

□

Theorem 5.3 (Lemma 2.2.1 and Theorem 2.6.4 in [CT91])**.** *The entropy $\mathcal{E}(p)$ is a finite positive number. Among all the distributions defined on Ω, the uniform distribution has the largest entropy.*

absolutely continuous

Kullback-Leibler divergence

Assume that we are given two distributions p and q. We say that q is *absolutely continuous* with respect to p if $p(\omega) = 0$ implies that $q(\omega) = 0$. Given p and q such that q is absolutely continuous with respect to p we define the *Kullback-Leibler divergence* to be

$$\mathrm{KL}(q;\, p) = \sum_{\omega\in\Omega} q(\omega)\log\frac{q(\omega)}{p(\omega)}.$$

If q is not absolutely continuous with respect to p, then the divergence $\mathrm{KL}(q;\, p)$ is defined to be infinite.

Example 5.4. The divergence between q and p given in Example 5.1 is

$$\mathrm{KL}(q;\, p) = 0.5 \log\left(\frac{0.5}{0.3}\right) + 0.1 \log\left(\frac{0.1}{0.3}\right) + 0 \log\left(\frac{0}{0.2}\right) + 0.4 \log\left(\frac{0.4}{0.2}\right) = 0.42.$$

□

Theorem 5.5 (Theorem 3.1 in [Kul68])**.** *The divergence* $\mathrm{KL}(q;\, p)$ *is a finite positive number if* q *is absolutely continuous with respect to* p*, and infinite otherwise. We have that* $\mathrm{KL}(q;\, p) = 0$ *if and only if* p *equals to* q*.*

Assume that p is the uniform distribution, then

$$\mathrm{KL}(q;\, p) = \sum_{\omega \in \Omega} q(\omega) \log \frac{p(\omega)}{|\Omega|^{-1}} = -\mathcal{E}(q) + \log |\Omega| \,.$$

Hence, we can see the entropy $\mathcal{E}(q)$ as a measure of closeness (in a Kullback-Leibler sense) of q to the uniform distribution: The higher the entropy, the closer is q to the uniform distribution.

The following theorem describes a possible interpretation of the values produced by the divergence.

Theorem 5.6 (Section 5.6 in [Kul68])**.** *Let* $p(\omega; \theta)$ *be a family of distributions parametrised by a vector* $\theta \in \mathbb{R}^K$*. Given* θ_0*, let* D *be a collection of* n *independent points sampled from a distribution* $p(\omega; \theta_0)$*. Let* θ_n *be an estimate of* θ_0 *from* D*. Then, under some broad regularity conditions*[1]*, we have that*

$$2n \sum_{\omega \in \Omega} p(\omega; \theta_n) \log \frac{p(\omega; \theta_n)}{p(\omega; \theta_0)}$$

converges weakly into a χ^2 *distribution with* K *degrees of freedom, as* n *goes into infinity.*

The theorem is stated but not proven in [Kul68]. We provide a proof for the case of a finite sample space Appendix A.3. The finite case is sufficient for our purposes.

According to the theorem we can use the quantity $2n\mathrm{KL}(\theta_n;\, \theta_0)$ as a statistical test by comparing the P-value $\mathbb{P}\big(\chi^2(K) < 2n\mathrm{KL}(\theta_n;\, \theta_0)\big)$ to the selected risk threshold.

5.2 Maximum Entropy

Maximum Entropy Our next topic, *Maximum Entropy*, is closely related to the Kullback-Leibler divergence. Assume that we are given a function $S : \Omega \to \mathbb{R}^K$ mapping a sample $\omega \in \Omega$ into a vector of length K. Assume also that we are given a vector $\theta \in \mathbb{R}^K$.

satisfies We say that a distribution p *satisfies* θ if

$$\theta = \mathrm{E}_p[S] = \sum_{\omega \in \Omega} p(\omega) S(\omega),$$

that is, the mean of S taken with respect to p is equals to θ. We denote the set of all distributions satisfying θ by $\mathcal{P}(S, \theta)$, that is,

$$\mathcal{P}(S, \theta) = \{p; \mathrm{E}_p[S] = \theta\}.$$

Assume that the set $\mathcal{P}(S, \theta)$ is not empty. We denote the distribution from $\mathcal{P}(S, \theta)$ having the maximal entropy by p^*. Note that p^* depends on S and θ but we have ignored these variables from the notation for the sake of clarity. The *exponential form* distribution p^* can be expressed with an *exponential form*.

Theorem 5.7 (Theorem 3.1 in [Csi75])**.** *Let $S : \Omega \to \mathbb{R}^K$ be a function and let θ be a vector of length K. Assume that $\mathcal{P}(S, \theta)$ is not empty and let p^* be the distribution maximising the entropy. There exists a vector $r \in \mathbb{R}^K$, a real number r_0, and a set $Z \subseteq \Omega$ such that*

$$p^*(\omega) = \begin{cases} \exp\left(r_0 + r^T S(\omega)\right) & , \text{ if } \omega \notin Z \\ 0 & , \text{ if } \omega \in Z \end{cases}.$$

Moreover, for every $q \in \mathcal{P}(S, \theta)$ and $\omega \in Z$ we have $q(\omega) = 0$.

Example 5.8. Assume our sample space is the set of all binary vectors of length K, that is, $\Omega = \{0, 1\}^K$. Let $S(\omega) = \omega$. The mean $\theta = \mathrm{E}[S(\omega)]$ is now a vector containing the margins of the individual attributes, that is, $\theta_i = p(a_i = 1)$. Theorem 5.7 now states that p^* has the form

$$p^*(\omega) = \exp\left(r_0 + r^T S(\omega)\right) = \exp(r_0) \prod_{i=1}^{K} \exp(r_i \omega_i).$$

Define $p_i(a_i = 1) = \theta_i$ and $p_i(a_i = 0) = 1 - \theta_i$. It turns out that p_i is proportional to $\exp(r_i \omega_i)$. Thus we must have

$$p^*(\omega) = p_1(\omega_1) p_2(\omega_2) \cdots p_K(\omega_K).$$

This is the distribution related to the independence model. For example, the distribution p in Example 5.1 obeys the independence model. □

Iterative Scaling Algorithm

In practice, the distribution p^* is solved using the ITERATIVE SCALING Algorithm. The algorithm was introduced in its modern form in [DR72]. However, the basic idea was introduced originally in [DS40], where the authors considered the problem of solving cell probabilities in the contingency tables. The algorithm applies Theorem 5.7 in the following way: Instead of exploring the space of distributions satisfying θ, the algorithm searches the distribution of an exponential form that satisfies θ. Theorem 5.7 guarantees that once such distribution is found, it will have the maximal entropy. The algorithm works in an iterative fashion: Assume that p is the current distribution, and let $\theta_{old} = \mathrm{E}_p[S]$. The algorithm picks a new distribution q such that $\theta_{new} = \mathrm{E}_q[S]$ is closer to θ than θ_{old} was [DR72, Theorem 1]. The step is repeated with q replacing p. Under certain conditions the algorithm can be speeded up considerably by decomposing the distributions [JP95].

[1] *The exact conditions are stated in Appendix A.3.*

Chapter 6

Predicting Itemset Frequencies

A large portion of the literature related to mining binary data deals with finding frequent itemsets or condensing them into smaller space. Itemsets can be viewed as a summary of the relevant information of the data set. We will focus on a scenario where the original data is replaced by a family of itemsets. Such an scenario is interesting theoretically and computationally, but it can also occur in practice in privacy-preserving data mining where the researcher has no access to the original data, instead he is given a family of itemsets. Namely, we consider a specific problem of finding the frequency of an unknown itemset from a set of known itemsets. This classic problem can be reduced to a linear program. Unfortunately, the solution is intractable since the program contains an exponential amount of variables. However, we can greatly reduce the number of variables by applying ideas from Markov Field Theory.

We begin by defining the query problem and reducing it to a linear program in Section 6.1. We point out in Section 6.2 that the query problem is intractable, unless $\mathbf{P} = \mathbf{NP}$. In Sections 6.3 and 6.4 we discuss how can we reduce the number of variables in the linear program. In Section 6.5 we define a framework for ranking itemsets based on the information available from sub-itemsets.

Contribution of the papers. This chapter is based on Publications II, III, and IV.

In Publication II we show that the query problems we are considering are intractable: Finding the possible frequencies of an itemset from a family of known itemsets is **NP**-hard. In the paper, the problems Consistent and MaxQuery can be seen as special case of FreqSAT, an **NP**-complete problem introduced in [Cal04, Cal03]. In FreqSAT, the constraining family of itemsets need not to be downward closed and we are also allowed to have inequality constraints. The proof for the **NP**-hardness of FreqSAT given in [Cal03] is actually a valid proof

for Consistent although this is not explicitly mentioned. In a more general scenario we are allowed to have conditional first-order logic sentences as constraints and queries [Luk01]. In PSat, a famous **NP**-complete problem, we are given a CNF-formula, a frequency for each clause, and we are asked to find a distribution satisfying these frequencies [GKP88]. The construction in the paper resembles the technique used in [Coo90], in which it is used to prove that the interference of Belief Networks is **NP**-hard.

The query problem can be reduced to linear program [Hai65]. The reduction, however, contains an exponential number of variables with respect to the number of attributes. To remedy this problem the attributes outside the query are ignored [BSH04, PMS03]. In Publication III we show that this may change the outcome. We develop a novel idea of safe sets, a set of attributes that are guaranteed to produce the right outcome for the query. We present a polynomial algorithm for finding the minimal safe sets. We also present a heuristic for finding restricted safe sets, that is, sets with a limited number of attributes. In our experiments, using restricted safe sets improves 10% of the queries in which the restricted safe set is larger than the actual query.

In Publication IV we study the idea of ranking itemsets based on their deviation from the prediction. Namely, we are given a set of known itemsets and a query itemset. We use the Maximum Entropy principle to predict the contingency table and compare it using Kullback-Leibler divergence against the actual contingency table obtained from the data set. Our prediction method is equivalent to the approach used in [PMS03].

Many measures has been suggested for ranking itemsets [AY98, Omi03, AIS93, GCB07, DP01], association rules [PS91, BMUT97, AIS93, JS02], and other related patterns [BMS97, HHM+07]. In many of these works the comparison is based on the independence model [BMS97, AY98, DP01, GCB07, PS91, BMUT97]. In some approaches the frequency is compared to a more flexible model. For instance, in [JS04, JS05] the frequency of an itemset is compared to an estimate obtained from the Bayes network. In [Meo00] the authors introduce the concept of dependence value, a difference between the actual frequency of an itemset and its Maximum Entropy estimate.

A special case of our framework results in a measure that ranks itemsets based on their deviation from the independence model. On the other hand, our proposal allows us to use richer models, such as, discrete Gaussian model. Our technique resembles greatly the active interestingness in [JS02] in which Kullback-Leibler divergence along with the Maximum Entropy principle is used for ranking association rules.

In a related work [HHM+07] the authors seek tree patterns with low entropy. In this approach interesting patterns would be those trees that have strong correlations compared to our approach in which interesting patterns would be the

contingency tables that cannot be modelled well by trees.

6.1 Definition of the Problem

Given a data set, solving a frequency of an itemset is a straightforward task, a single data scan. A more complicated scenario is when we are given, instead of the data set, a family of itemsets $\mathcal{F}$ along with their frequencies and we are asked to deduce the frequency of some unknown itemset, say, Q from $\mathcal{F}$. In this section we will discuss this particular problem and show how the task can be solved using Linear Programming.

We should point out immediately that generally deducing the frequency of the itemset Q from the family $\mathcal{F}$ of itemsets does not yield a unique solution. One can easily form data sets having different frequencies for Q but same frequencies for $\mathcal{F}$. Hence our interest is to find *all* possible frequencies of Q that can be produced by data sets having some specific frequencies for $\mathcal{F}$.

To continue our analysis let us rephrase the problem using the distributions. We will point out later that this transformation has no particular effect on the outcome. Given a family of itemsets $\mathcal{F}$ along their frequencies θ and a (query) itemset Q we define a *frequency interval* $\mathrm{fi}(Q;\mathcal{F},\theta)$ to be a set

frequency interval

$$\mathrm{fi}(Q;\mathcal{F},\theta) = \{p\,(Q=1)\,;p \text{ is a distribution satisfying } \theta\}$$

of possible frequencies of Q produced by distributions satisfying θ. Our goal in this section is to provide a method for solving this interval.

Example 6.1. Assume that we have two attributes a_1 and a_2 and that our family of itemsets consists of one-order itemsets $\mathcal{F} = \{a_1, a_2\}$. The corresponding frequencies are set to be $\theta = (\theta_{a_1}, \theta_{a_2})$, where $\theta_{a_1} = 0.6$ and $\theta_{a_2} = 0.7$. Let us calculate the frequency interval $\mathrm{fi}(Q;\mathcal{F},\theta)$ for $Q = a_1a_2$.

We know that we must have

$$p\,(Q=1) \leq \min\{p\,(a_1=1)\,, p\,(a_2=1)\} = 0.6$$

and

$$p\,(Q=1) \geq p\,(a_1=1) + p\,(a_2=1) - 1 = 0.3.$$

Let us define distributions p_1 and p_2 as

$$\begin{aligned} p_1(a_1=0,a_2=0) &= 0, & p_1(a_1=1,a_2=0) &= 0.3\\ p_1(a_1=0,a_2=1) &= 0.4, & p_1(a_1=1,a_2=1) &= 0.3 \end{aligned}$$

and

$$\begin{aligned} p_2(a_1=0,a_2=0) &= 0.3, & p_2(a_1=1,a_2=0) &= 0\\ p_2(a_1=0,a_2=1) &= 0.1, & p_2(a_1=1,a_2=1) &= 0.6 \end{aligned}$$

We see that p_1 and p_2 are genuine distributions. They also satisfy the frequencies θ:

$$\begin{aligned} p_1(a_1 = 1) &= 0.3 + 0.3 = 0.6 = \theta_{a_1} \\ p_1(a_2 = 1) &= 0.4 + 0.3 = 0.7 = \theta_{a_2} \\ p_2(a_1 = 1) &= 0 + 0.6 = 0.6 = \theta_{a_1} \\ p_2(a_2 = 1) &= 0.1 + 0.6 = 0.7 = \theta_{a_2} \end{aligned}$$

We have $p_1(Q = 1) = 0.3$ and $p_2(Q = 1) = 0.6$. Hence we know that $\max(\text{fi}(Q; \mathcal{F}, \theta)) = 0.6$ and $\min(\text{fi}(Q; \mathcal{F}, \theta)) = 0.3$. The discussion below shows us that we have $\text{fi}(Q; \mathcal{F}, \theta) = [0.3, 0.6]$. □

Let us next analyse the frequency interval. Assume that two distributions p_0 and p_1 satisfy the given frequencies θ and that $p_0(Q = 1) = \eta_0$ and $p_1(Q = 1) = \eta_1$. Let a be a real number, $0 \leq a \leq 1$. Then a distribution $p_a = (1 - a)p_0 + ap_1$ satisfies the frequencies θ and $p_a(Q = 1) = (1 - a)\eta_0 + a\eta_1$. We conclude that $\text{fi}(Q; \mathcal{F}, \theta)$ is truly an interval and hence to solve this set we need to solve the extrema points.

To solve the right side of the interval $\text{fi}(Q; \mathcal{F}, \theta)$ we consider the following optimisation problem:

$$\begin{aligned} &\max p(Q = 1) \\ &\quad p(F_i = 1) = \theta_i, \text{ for } i \in \{1, \ldots, N\}. \end{aligned}$$

This problem resembles greatly linear program and we can transform this problem into a linear form. Note that the distribution p is defined on a sample space $\Omega = \{0, 1\}^K$ of binary vectors of length K. Let $\omega \in \Omega$ be a binary vector and let p_ω be the corresponding probability of p producing ω. Let Q be an itemset and let S_Q be the corresponding indicator function. We can formulate the optimisation problem in the following form

$$\begin{aligned} \max \sum_{\omega \in \Omega} & S_Q(\omega) p_\omega \\ \sum_{\omega \in \Omega} & S_{F_i}(\omega) p_\omega = \theta_i, \text{ for } i \in \{1, \ldots, N\} \\ \sum_{\omega \in \Omega} & p_\omega = 1 \\ & p_\omega \geq 0, \text{ for } \omega \in \Omega. \end{aligned}$$

Clearly, this is a linear program of a standard form. The left side of the interval $\text{fi}(Q; S, \theta)$ can be solved similarly. The following theorem summarises the previous discussion:

Theorem 6.2 (Theorem 1 in [BSH04]). *Given a family $\mathcal{F}$ of itemsets along their frequencies θ, a frequency interval $\mathrm{fi}(Q;S,\theta)$ for a query itemset is an interval whose boundaries can be solved using Linear Programming.*

Example 6.3. Let us reformulate the setup given in Example 6.1 as a linear program. In order to do that let p_{yz} represent the probability $p(a_1 = y, a_2 = z)$. The following linear program solves the right side of the frequency interval:

$$\begin{aligned} &\max p_{11} \\ p_{10} + p_{11} &= 0.6 \\ p_{01} + p_{11} &= 0.7 \\ \sum_{y,z\in\{0,1\}} p_{yz} &= 1 \\ p_{yz} &\geq 0, \text{ for } y,z \in \{0,1\}. \end{aligned}$$

The min-version of the program results in the left side of $\mathrm{fi}(Q;\mathcal{F},\theta)$. □

Let us now return to our original setup and consider what are the possible frequencies for query produced by *data sets*. The main difference here is that data sets must have finite number of elements. Thus, the empirical distribution formed from a finite data set has rational probabilities. This means that the possible frequencies for a query itemset Q should be rational. Given a distribution having only rational probabilities, we can easily form a data set having the distribution as empirical distribution. Theorem 3.7 guarantees that given a rational frequency $\eta \in \mathrm{fi}(Q;S,\theta)$ there is a rational distribution producing η as a frequency for Q. Theorem 3.7 also guarantees that the boundaries of $\mathrm{fi}(Q;S,\theta)$ are rational. We summarise this in the following theorem:

Theorem 6.4 (Lemma 1 in [BSH04]). *Given a family $\mathcal{F}$ of itemsets along their frequencies θ, possible frequencies for a query itemset Q produced by data sets satisfying the frequencies θ are $\mathrm{fi}(Q;S,\theta) \cap \mathbb{Q}$. Also, boundaries of the interval $\mathrm{fi}(Q;S,\theta)$ are rational and hence there are data sets producing these extrema frequencies.*

Example 6.5. We see that the boundaries in Example 6.1 are rational. For instance, a data set satisfying the frequencies θ and producing a frequency 0.6 for $Q = a_1a_2$ is

$$D = \left\{ \begin{array}{l} (0,0),(0,0),(0,0),(0,1),(1,1), \\ (1,1),(1,1),(1,1),(1,1),(1,1) \end{array} \right\}.$$

□

6.2 Complexity of Querying Itemsets

The major drawback in the query problem is that the number of variables in the linear program is $|\Omega| = 2^K$, where K is the number of attributes. In this section we will demonstrate that solving the query problem is **NP**-complete.

Recall that the downward closed family of itemsets is the one in which the subsets of a member itemset are also members. Consider the following problems:

Consistent
- Consistent: Given a set of downward closed family of itemsets $\mathcal{F}$ and a set of rational frequencies θ, decide if there is a data set that produces θ for $\mathcal{F}$.

MaxQuery
- MaxQuery: Given a set of downward closed family of itemsets $\mathcal{F}$, a set of rational consistent frequencies θ, and a query Q, find the maximal frequency of Q that a data set satisfying θ may achieve.

EntrQuery
- EntrQuery: Given a set of downward closed family of itemsets $\mathcal{F}$, a set of rational consistent frequencies θ, and a query Q, calculate the frequency $p^*(Q = 1)$, where p^* has the highest entropy among distributions satisfying θ.

Solving MaxQuery is equal to solving the right side of $\mathrm{fi}(Q; \mathcal{F}, \theta)$. EntrQuery is relevant because empirical tests indicate that this method leads to a good approximation of the frequency of Q [PMS03].

The following theorem states the complexity results of these problems.

Theorem 6.6 (Theorems 4, 6, and 7 in Publication II)**.** Consistent *and the decision version of* MaxQuery *are* ***NP****-complete. The decision version of* EntrQuery *is* ***PP****-complete.*

6.3 Safe sets

As we have pointed out in Section 6.2, the evaluation time of the query time is exponential with respect to the number of attributes. Hence, we can speed up the algorithm if we can reduce the attributes: Assume that we are given a query itemset Q and a family of itemset $\mathcal{F}$ along with the frequencies θ. Define $\mathcal{F}_Q$ to contain only the itemsets from $\mathcal{F}$ that are subsets of Q. Let θ_Q be the corresponding frequencies. Instead of computing $\mathrm{fi}(Q; \mathcal{F}, \theta)$, we *project* out the variables outside Q and compute $\mathrm{fi}(Q; \mathcal{F}_Q, \theta_Q)$. In doing this, we reduce the number of attributes from K to $|Q|$. The downside is that the frequency interval may change.

project

Example 6.7. Assume that we have three attributes a, b, and c. Let $\mathcal{F}$ be $\{a, b, c, ab, ac\}$, and $\theta = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$. Let $Q = bc$ be the query itemset. Then

$\mathcal{F}_Q = \{b, c\}$. In this case $\mathrm{fi}(Q; \mathcal{F}_Q, \theta_Q) = \left[0, \frac{1}{2}\right]$. However, it follows from θ that $a = b$ and $a = c$, hence we must have $b = c$. This implies that $\mathrm{fi}(Q; \mathcal{F}, \theta) = \frac{1}{2}$. □

Our goal is to study which attributes we can remove and which attributes we must keep. We begin by giving some definitions. Let $A = \{a_1, \ldots, a_K\}$ be the set of all attributes and B a subset of A. Let q be a distribution defined on B. We say that p, a distribution defined on A, is an *extension* of q, if the marginalisation of p on B is equal to q, that is, $p(B = \omega) = q(B = \omega)$ for any binary vector ω. Given a family of itemsets $\mathcal{F}$ and frequencies θ, we say that B is θ-*safe* if a distribution q (defined on B) satisfying θ_B can be extended into p satisfying θ. If B is θ-safe for all θ, we say that B is *safe*.

extension

θ-safe

safe

It is easy to see that if B is a safe set and a query itemset Q is a subset of B, then we can remove the attributes outside B without changing the outcome.

The rest of this section is devoted to the analysis of the safe sets. Namely, we will apply Markov Random Field Theory in order to characterise safe sets. Let G be a graph of K nodes, each node corresponding to an attribute. For each itemset X in the family $\mathcal{F}$, we connect the nodes in G corresponding to X, that is X is a clique in G. We call G a *dependency graph* of $\mathcal{F}$. If, in addition, we connect all the nodes from B, we obtain a dependency graph of $\mathcal{F} \cup \{B\}$. The following theorem provides a neat way of characterising safe sets using dependency graphs.

dependency graph

Example 6.8. Assume that we have 6 attributes. Let

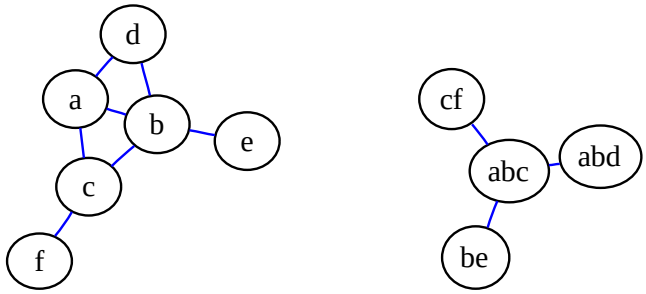
$$\mathcal{F} = \{a, b, c, d, e, f, ab, bc, ac, ad, bd, be, cf\}$$

be a family of itemsets. Let $B = abc$. The dependency graph of $\mathcal{F} \cup B$ is given in Figure 6.1(a). We are interested in finding out whether B is a safe set. To prove this we need to show that the distribution q defined on B can be extended into distribution p defined on all attributes.

Consider the junction tree of the dependency graph (given in Figure 6.1(b)). Let p_1 be a distribution defined on abd, p_2 a distribution defined on be, and p_3 a distribution defined on cf. The separator between abd and abc is ab. Note that ab is a member of $\mathcal{F}$. Hence, Theorem 2.10 implies that p_1 and q are equal at the separator ab. We have also p_2 and q being equal at the separator b and p_3 and q being equal at the separator c. Now we can apply Proposition 4.9 to combine q, p_1, p_2, and p_3 into a joint distribution. □

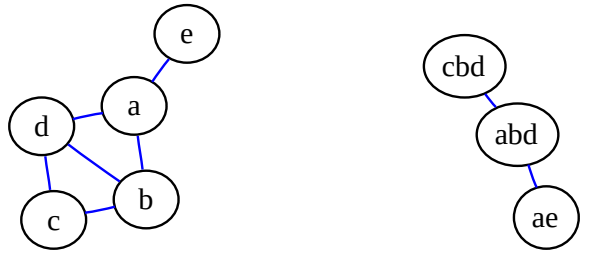
The following theorem shows that the constuction done the previous example holds also in general case.

Theorem 6.9 (Theorems 1–2 in Publication III)**.** *Let $\mathcal{F}$ be a downward closed family of itemsets. Let $B \notin \mathcal{F}$ be a subset of attributes. Let G be a dependency graph of $\mathcal{F} \cup \{B\}$. Then B is safe if and only if there is a junction tree T of G such that B is a node of T and all the separators of B are in $\mathcal{F}$.*



(a) The dependency graph of $\mathcal{F} \cup \{B\}$ (b) A junction tree

Figure 6.1: Graphs related to Example 6.8.



(a) The dependency graph of $\mathcal{F} \cup \{B\}$ (b) A junction tree

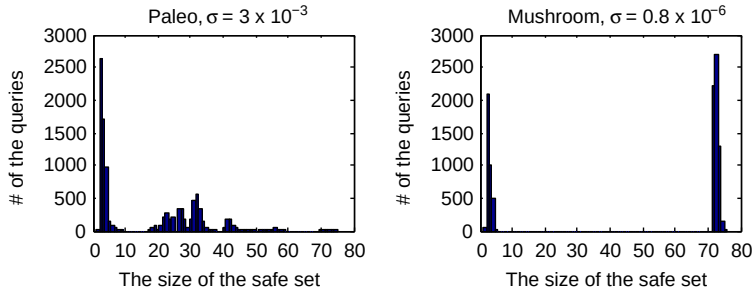
Figure 6.2: Graphs related to Example 6.10.

Example 6.10. Assume that we have 5 attributes. Let

$$\mathcal{F} = \{a, b, c, d, e, ab, bc, cd, ad, ae\}$$

be a family of itemsets. Let $B = abd$. The dependency graph of $\mathcal{F} \cup B$ is given in Figure 6.2(a) and its junction tree is given in Figure 6.2(b). The separators of B are a and bd. The itemset bd is not included in $\mathcal{F}$, hence B is not a safe set. However, if we augment $\mathcal{F}$ with bd, then B becomes a safe set. □

Given a family of itemsets $\mathcal{F}$ and a query itemset Q, our goal is to find a safe set B that contains Q. The algorithm for finding such a set is described in Algorithm 1 in Publication III. The algorithm starts by setting $B = Q$ and augments B with attributes until B is safe. The addition order of the attributes is selected such that when B becomes safe, it is guaranteed that B will be also



(a) *Paleo* (b) *Mushroom*

Figure 6.3: Distributions of the sizes of safe sets for random queries containing 2–4 attributes. The left histogram is obtained from the *Paleo* data set and the right histogram is obtained from the *Mushroom* data set.

the minimal safe set. The exact details of the algorithm are outside the scope of this introduction.

Theorem 6.11 (Theorem 6 in Publication III)**.** *Let $\mathcal{F}$ be a downward closed family of itemsets and let Q be a query itemset. The minimal safe set B containing Q is unique. There exists a polynomial algorithm for finding B.*

Example 6.12. In this example we consider the sizes of safe sets for random queries. We used two data sets: *Paleo*[1], a data set containing information of species fossils found in specific paleontological sites in Europe [For05], and *Mushroom*, a data set available from the FIMI repository[2].

From these data sets we extracted a family of itemsets using a modified version of APRIORI (see Publication III for more details). Using these families as surrogates for the data sets we calculated the minimal safe sets for 10000 random queries having 2–4 attributes. The results given in Figure 6.3 show that even though the queries were relatively simple they may produce large safe sets, that is, we need a large amount of of additional attributes to guarantee that the prediction boundaries are correct.

□

[1]NOW public release 030717 available from [For05], preprocessed as in [FGJM06].
[2]http://fimi.cs.helsinki.fi

6.4 Optimising Linear Program via Markov Random Fields

In the previous section we have used MRF Theory to remove the attributes from the query problem. In this section we demonstrate that we can use similar ideas to reduce the complexity of the query problem even further.

Before presenting the main theorem of this section we will demonstrate the technique with an example.

Example 6.13. Assume that we have K attributes $A = \{a_1, \ldots, a_K\}$. The family $\mathcal{F}$ contains $2K - 1$ itemsets,

$$\mathcal{F} = \{a_i; i = 1, \ldots, K\} \cup \{a_i a_{i+1}; i = 1, \ldots, K-1\}.$$

Let query be $Q = a_1 a_K$. We see that the minimal safe set is A itself. Hence, we have 2^K variables in the linear program. However, we can reformulate the program in the following way: Let G be a dependency graph of $\mathcal{F} \cup \{Q\}$. The graph G is a cycle (see Figure 6.4(a)). Triangulate the graph by connecting a_1 with the rest of the attributes (Figure 6.4(b)). A junction tree T has $K-2$ nodes (cliques) $C_i = a_1 a_{i+1} a_{i+2}$, where $i = 1, \ldots, K-2$ (Figure 6.4(c)). Consecutive nodes C_i and C_{i+1} are connected. Let S_j be the separators of T. Note that $S_j = a_1 a_{j+2}$.

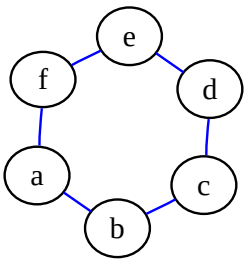
Let θ be the frequencies for $\mathcal{F}$. Let p_i be a distribution defined on a clique C_i. Let $\mathcal{F}_i = \mathcal{F}_{C_i}$ be the subset of $\mathcal{F}$ containing only itemsets that are subsets of C_i. Consider the following linear program:

$$\begin{aligned} &\max p_{K-2}(Q = 1) \\ &p_i(X = 1) = \theta_X, \quad X \in \mathcal{F}_i, \quad i = 1, \ldots, K-2, \\ &p_j = p_{j+1} \text{ at } S_j, \quad j = 1, \ldots, K-3. \end{aligned} \tag{6.1}$$

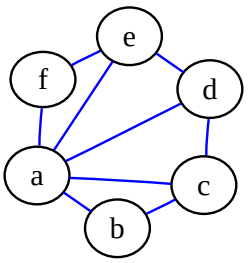
The first set of conditions says that p_i must satisfy the related frequencies. The second set of conditions forces p_i to be consistent with respect to each other. Let p be a distribution (defined on A) maximising the frequency of Q. Clearly p can be decomposed into p_i such that the conditions in Eq. 6.1 hold. Assume now that p_i solves Eq. 6.1. We can apply Theorem 4.9 and compose p_i into p such that p satisfies θ.

This implies that we can solve $\mathrm{fi}(Q; \mathcal{F}, \theta)$ by solving the linear program in Eq. 6.1. The number of variables in the program is $(K-2) \times 8$ which is drastically smaller than 2^K. □

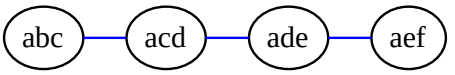
The following theorem summarises the previous discussion.



(a) The dependency graph of $\mathcal{F} \cup \{Q\}$



(b) A triangulated dependency graph



(c) The junction tree

Figure 6.4: Graphs related to Example 6.13.

Theorem 6.14 (Theorem 9 in Publication III)**.** *Let $\mathcal{F}$ be the family of itemsets along with the corresponding frequencies θ. Let Q be the query itemset. Let G be a dependency graph for $\mathcal{F} \cup Q$ and let T be its junction tree. Denote the nodes of T by C_i and let $\mathcal{F}_i = \mathcal{F}_{C_i}$. Assume that $Q \subseteq C_1$. Let p_i be a distribution defined on C_i. Then the frequency interval $\mathrm{fi}(Q; \mathcal{F}, \theta)$ can be solved with a linear program (and its min-version)*

$$\begin{aligned} &\max p_1(Q = 1) \\ &p_i \text{ satisfies } \mathcal{F}_i, \quad \text{for each } C_i \\ p_i = p_j &\text{ at the separator}, \quad \text{for each neighbours (in } T) \ C_i, C_j. \end{aligned}$$

The number of variables in the program is $\sum_i 2^{|C_i|}$.

The method works well when the queries are relatively small but it fails if the query contains all the attributes in which case we asking the probability of having a transaction with no 0s. For this particular case we can use an alternative approach decribed in [DF00]. In this approach we are able to solve the queries without using linear programming. The limitation for this approach is that the query must contain all the attributes.

6.5 Entropy Based Ranking of Itemsets

So far we have been interested in finding the frequency interval of an itemset given some known itemsets. A similar approach can be used in ranking itemsets:

We predict the frequency of an itemsets from some set of known itemsets and compare the actual value with the prediction. The more the actual value deviate from the prediction, the more the itemset is an interesting one. We will use Maximum Entropy as our estimation method and Kullback-Leibler divergence for comparison between the actual value and estimate.

Assume that we have a family of itemsets $\mathcal{F}$ and an itemset $Q \notin \mathcal{F}$. Assume for simplicity that there are no attributes in the data set outside Q. If there are, then these attributes are projected out. Let θ be the frequencies for $\mathcal{F}$. Let p^* be distribution having the highest entropy and satisfying the frequencies θ. Let q be the empirical distribution obtained from the data set D,

$$q(\omega) = \frac{\text{transactions equal to } \omega \text{ in } D}{|D|}.$$

We define the rank of an itemset to be

$$r(Q; \mathcal{F}, D) = \mathrm{KL}(q; p^*) = \sum_{\omega \in \Omega} q(\omega) \log \frac{q(\omega)}{p^*(\omega)}.$$

Example 6.15. Assume that we have $Q = ab$ and $\mathcal{F} = \{a, b\}$. Let data set D be

$$D = \left\{ \begin{array}{l} (0,0), (0,0), (0,0), (0,1), (1,1), \\ (1,1), (1,1), (1,1), (1,1), (1,1) \end{array} \right\}.$$

The frequency for a is 0.6 and the frequency for b is 0.7. We know that in this case the maximum entropy distribution p^* is equal to the independence model. Hence, we have

$$\begin{aligned} p^*(a=0, b=0) &= 0.4 \times 0.3, & p^*(a=0, b=1) &= 0.4 \times 0.7, \\ p^*(a=1, b=0) &= 0.6 \times 0.3, & p^*(a=1, b=1) &= 0.6 \times 0.7. \end{aligned}$$

On the other hand, the empirical distribution is equal to

$$\begin{aligned} q(a=0, b=0) &= 0.3, & q(a=0, b=1) &= 0.1, \\ q(a=1, b=0) &= 0.0, & q(a=1, b=1) &= 0.6. \end{aligned}$$

The rank of Q is equal to $r(Q; \mathcal{F}, D) = 0.3859$. □

Let us briefly discuss the evaluation of our rank measure. Distributions q and p^* have $2^{|Q|}$ entries. In addition, Theorem 6.6 points out that solving p^* is a **PP**-complete problem. Hence, we cannot solve this rank for large itemsets. Nevertheless, the rank is doable for itemsets of smaller size.

The following theorem, which follows from Theorem 5.6, explains the asymptotical behaviour of the measure.

Theorem 6.16 (Theorem 5 in Publication IV)**.** *Let $\mathcal{F}$ be a family of itemsets and $Q \notin \mathcal{F}$. Let D be a data set with N points sampled from p^*. If Q is non-derivable, then the quantity $2Nr(Q; \mathcal{F}, D)$ approaches to a χ^2 distribution with $2^{|Q|} - 1 - |F|$ degrees of freedom as N approaches the infinity.*

Theorem suggests that the ranks should be normalised — instead of using the raw values we should compare the P-values.

Example 6.17. We continue Example 6.15. The number of degrees is

$$2^{|Q|} - 1 - |F| = 2^2 - 1 - 2 = 1.$$

The P-value in our case is

$$\mathbb{P}\big(\chi^2 \leq 2 \times 10 \times 0.3859\big) = 0.9945.$$

Such a high P-value tells us that the actual empirical frequency of the itemset ab is statistically significant different than the prediction based on the independence model. □

Example 6.18. Consider a synthetic data set D with 100 independent columns and 1000 transactions. From this data set we select a certain set of queries (see Publication IV for more details) and calculate 3 different rank measures:

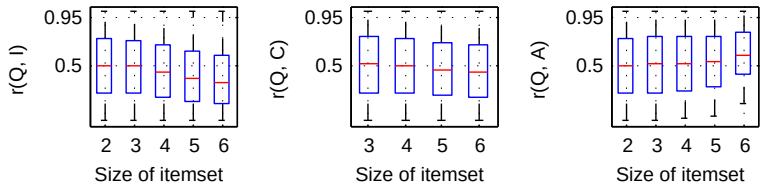
1. Measure $r(Q; \mathcal{I})$, where $\mathcal{I}$ is the set of itemsets of size 1. In this case, the Maximum Entropy distribution p^* is equal to the independence model.
2. Measure $r(Q; \mathcal{C})$, where $\mathcal{C}$ is the set of itemsets of size 1, 2. In this case, p^* is equal to the discrete Gaussian model.
3. Measure $r(Q; \mathcal{A})$, where $\mathcal{A}$ is the set of all proper sub-itemsets of Q.

The normalised ranks are given in Figure 6.5. We see that the ranks for $r(Q; \mathcal{I})$ are relatively small. This is a natural result since the 0-hypothesis of Theorem 6.16 holds for this particular data set. We also note that the ranks for richer models tends to be higher than for the independence model. In other words, the measure overfits the data and the prediction is misguided by the noise in the frequencies of the itemsets with the higher number of attributes.

□

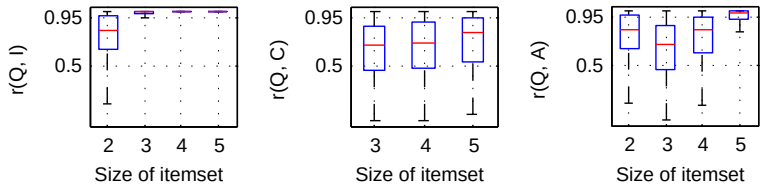
Example 6.19. We repeat Example 6.18 using *Paleo*[3], a dataset containing information of species fossils found in specific paleontological sites in Europe [For05]. The normalised ranks are given in Figure 6.6.

[3]NOW public release 030717 available from [For05], preprocessed as in [FGJM06].



(a) Independence model (b) Gaussian model (c) "All" model

Figure 6.5: Ranks for queries from synthetic data set. Each box represents queries with particular number of attributes.



(a) Independence model (b) Gaussian model (c) "All" model

Figure 6.6: Ranks for queries from the *Paleo* data set. Each box represents queries with particular number of attributes.

Here we see that the ranks for $r(Q; \mathcal{I})$ are high. The attributes in the data set is known to be highly correlated so the independence model produces poor estimates. The Gaussian model produces more accurate prediction and, hence, smaller ranks. We also note that the "All" model overfits and produces higher ranks than the Gaussian model.

□

Chapter 7

Distances between Binary Data Sets

The notion of similarity plays a crucial part in data mining. Once a distance between two objects is established a large number of data mining algorithms can be applied and the highly complex objects can be studied as units.

In this chapter we discuss the distances between binary data sets. Instead of defining the distance directly we apply the theme of Chapter 6 and use itemsets as a surrogate for the actual data. We begin by defining the distance in Section 7.1 via geometrical notions. We point out in Section 7.2 that our distance is the only one among Mahanalobis distances that satisfies certain broad assumptions.

The contribution of the paper. This chapter is based on Publication I. In the paper we define a computable distance between two binary distances with solid statistical properties. We approach the problem by first calculating the frequencies of some given itemsets and compare the frequencies. Since we expect the frequencies to correlate we provide a proper normalisation. We provide 3 different definitions for the distance: Firstly, we use geometrical notions to define the distance. Secondly, we show that the distance is the unique Mahalanobis distance satisfying specific axioms. Thirdly, we show that the distance is the unique Mahalanobis distance that generalises the L_2 distance between two empirical distributions. We show that the distance can be solved in cubic time with the respect to the number of itemsets and in linear time if the itemsets form a downward closed family. Our experiments with real-world data show that our distance produces interesting results that agree with our expectations. An alternative approach to define a data set distance is to use some natural distance between single data points and apply some known set distance. Some data set distances defined in this way can be evaluated in cubic time with respect to the

number of transactions [EM97]. However, this is too slow for us since we may have a vast amount of data points. We can also approach the problem by considering distances for distributions (see [Bas89] for a nice review). From these distances the CM distance resembles the statistical tests involved with Minimum Discrimination Theorem [Kul68, Csi75]: From each data set a Maximum Entropy distribution is calculated and they are compared using Kullback- Leibler divergence. However, the major drawback is that solving the distributions is an **NP**-hard problem [Coo90].

7.1 Constrained Minimum Distance

In our approach we do not compute the distance directly between data sets. Rather, we compare the frequencies of some given family of itemsets. Such an approach provides us a flexible family of distances, since we can choose which itemsets we are interested in. A problem with the itemset frequencies is that they correlate: If the frequency of an attribute a is small, we expect that the frequency of an itemset ab is also small. Thus, we should seek a distance that decorrelates the frequencies.

Assume that we are given two binary data sets D_1 and D_2, both having K attributes. Also assume that we are given a family of itemsets $\mathcal{F}$. We let $\mathcal{P}(\mathcal{F}, \theta)$ to be the set of distributions satisfying the frequencies θ, that is, $p \in \mathcal{P}(\mathcal{F}, \theta)$ if and only if $p(F_i = 1) = \theta_i$, for all $F_i \in \mathcal{F}$. The set $\mathcal{P}(\mathcal{F}, \theta)$ can be seen as a polytope in $\mathbb{R}^{2^K}$. We use the notation $\mathcal{P}(\mathcal{F}, D)$ if θ is calculated from a data set D.

One approach for defining the distance between D_1 and D_2 is to use the sets $\mathcal{P}(\mathcal{F}, D_1)$ and $\mathcal{P}(\mathcal{F}, D_2)$. However, as we have seen in Section 6.2, these sets are difficult to compute. Thus, another approach is needed. Recall that $S_{\mathcal{F}}$ is an indicator function for $\mathcal{F}$, that is, the ith component of $S_{\mathcal{F}}(\omega)$ is equal to 1 if ω satisfies $F_i \in \mathcal{F}$, and 0 otherwise. We define the set $\mathcal{C}(\mathcal{F}, \theta)$ to be

$$\mathcal{C}(\mathcal{F}, \theta) = \left\{ x \in \mathbb{R}^{2^K}; \sum_{\omega} x_\omega = 1, \sum_{\omega} S_{\mathcal{F}}(\omega) x_\omega = \theta \right\},$$

that is, $\mathcal{C}(\mathcal{F}, \theta)$ is similar to $\mathcal{P}(\mathcal{F}, \theta)$ except that we are allowed to have negative elements. See Figure 7.1 for illustration.

We see immediately that $\mathcal{C}(\mathcal{F}, \theta)$ is an affine space (a linear subspace shifted by vector) and that $\mathcal{P}(\mathcal{F}, \theta)$ is a subset of $\mathcal{C}(\mathcal{F}, \theta)$. In addition, the spaces $\mathcal{C}(\mathcal{F}, D_1)$ and $\mathcal{C}(\mathcal{F}, D_2)$ are parallel. This fact enables us to define the *constrained minimum (CM) distance* as

constrained minimum (CM) distance

$$d_{CM}(D_1, D_2; \mathcal{F}) = \sqrt{2^K} \times \begin{array}{l} \text{the shortest distance between} \\ \mathcal{C}(\mathcal{F}, D_1) \text{ and } \mathcal{C}(\mathcal{F}, D_2). \end{array}$$

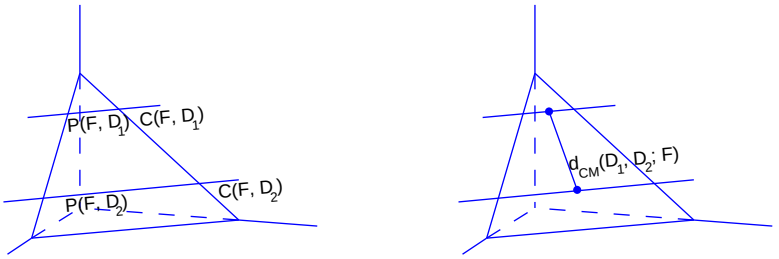


Figure 7.1: Illustration of the CM distance. The triangle represents the set of all possible distributions. The sets $\mathcal{C}(\mathcal{F}, D_i)$ are lines and the sets $\mathcal{P}(\mathcal{F}, D_i)$ are the segments containing the joint points from the set of all distributions and $\mathcal{C}(\mathcal{F}, D_i)$. The CM distance is proportional to the shortest distance between the spaces $\mathcal{C}(\mathcal{F}, D_1)$ and $\mathcal{C}(\mathcal{F}, D_2)$.

See Figure 7.1 for illustration.

It turns out that we can compute $d_{CM}(D_1, D_2; \mathcal{F})$ in polynomial time.

Theorem 7.1 (Theorem 1 in Publication I)**.** *Assume two data sets D_1 and D_2, and let $\mathcal{F}$ be a family of itemsets. Let θ and η be the frequencies calculated from D_1 and D_2, respectively. Let q be the uniform distribution and define a covariance matrix C as*

$$C_{ij} = q(F_i = 1, F_j = 1) - q(F_i = 1)q(F_j = 1), \quad F_i, F_j \in \mathcal{F}.$$

We have

$$d_{CM}(D_1, D_2; \mathcal{F})^2 = (\theta - \eta)^T C^{-1} (\theta - \eta).$$

Theorem 7.1 suggests that the CM distance is an L_2 distance of the decorrelated itemset frequencies. We demonstrate Theorem 7.1 with the following example.

Example 7.2. Assume that we have D_1 and D_2 defined as

$$D_1 = \left\{ \begin{array}{l} (0,0),(0,0),(0,0),(0,1),(0,1), \\ (0,1),(1,0),(1,1),(1,1),(1,1) \end{array} \right\}$$

and

$$D_2 = \left\{ \begin{array}{l} (0,0),(0,0),(0,1),(0,1),(0,1), \\ (1,0),(1,0),(1,0),(1,1),(1,1) \end{array} \right\}.$$

Let $\mathcal{F} = \{a_1, a_2, a_1 a_2\}$. The frequencies for D_1 are equal to $\theta = (0.4, 0.6, 0.3)$ and the frequencies for D_2 are equal to $\eta = (0.5, 0.5, 0.2)$. The covariance matrix

C is equal to

$$C = \begin{bmatrix} \frac{1}{4} & 0 & \frac{1}{8} \\ 0 & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{3}{16} \end{bmatrix}.$$

Theorem 7.1 implies that the distance is

$$d_{CM}(D_1, D_2; \mathcal{F})^2 = (\theta - \eta)^T C^{-1} (\theta - \eta) = 0.24.$$

Let p_1 be the empirical distribution for D_1,

$$p_1(a_1 = 0, a_2 = 0) = 0.3, \quad p_1(a_1 = 0, a_2 = 1) = 0.3,$$
$$p_1(a_1 = 1, a_2 = 0) = 0.1, \quad p_1(a_1 = 1, a_2 = 1) = 0.3.$$

and let p_2 be the empirical distribution for D_2,

$$p_2(a_1 = 0, a_2 = 0) = 0.2, \quad p_2(a_1 = 0, a_2 = 1) = 0.3,$$
$$p_2(a_1 = 1, a_2 = 0) = 0.3, \quad p_2(a_1 = 1, a_2 = 1) = 0.2.$$

Note that $\mathcal{C}(\mathcal{F}, D_1) = \{p_1\}$ and $\mathcal{C}(\mathcal{F}, D_2) = \{p_2\}$. Hence, by the definition of the CM distance, we have

$$\begin{aligned} d_{CM}(D_1, D_2; \mathcal{F})^2 &= 2^K \|p_1 - p_2\|_2^2 \\ &= 4\left[(0.3 - 0.2)^2 + (0.3 - 0.3)^2 + (0.1 - 0.3)^2 + (0.3 - 0.2)^2\right] \\ &= 0.24. \end{aligned}$$

□

We can also express CM distance in a neat form using parity functions. The following theorem is a corollary of Theorems 2.12, 2.13, and 7.1.

Theorem 7.3 (Section 3.1 in Publication I[1])**.** *Assume two data sets D_1 and D_2, and let $\mathcal{F}$ be a downward closed family of itemsets. Let $\mathcal{H}$ be the corresponding set of parity functions. Let α and β be the frequencies for H calculated from D_1 and D_2, respectively. Then*

$$d_{CM}(D_1, D_2; \mathcal{F}) = 2\|\alpha - \beta\|_2.$$

Example 7.4. We continue Example 7.2. Recall that the parity function results 1 if and only if an odd number of attributes are active. The parity frequencies for

[1] In Eq. 4 in Publication I there should be 2 instead of $\sqrt{2}$.

D_1 are $\alpha = (0.4, 0.6, 0.4)$ and the parity frequencies for D_2 are $\beta = (0.5, 0.5, 0.6)$. Theorem 7.3 implies that

$$\begin{aligned} d_{CM}(D_1, D_2; \mathcal{F})^2 &= 4 \|\alpha - \beta\|_2^2 \\ &= 4 \left[(0.4 - 0.5)^2 + (0.6 - 0.5)^2 + (0.4 - 0.6)^2 \right] = 0.24. \end{aligned}$$

□

Example 7.5. We consider the following 3 data set families: *Bible*, a collection of 73 books from the Bible[2], *Addresses*, a collection of 55 inaugural addresses given by the presidents of the U.S.[3], and *Abstract*, was composed of abstracts describing NSF awards from 1990–1999[4] (see Publication I for more details).

We calculated the distance matrices for each data set collection using the following 3 itemset families: *ind*, the collection of itemsets containing only one attribute, *cov*, the family of itemsets containing 1–2 attributes, and $freq$ a collection of $10K$ most frequent itemsets, where K is the dimension of the dataset.

From the results given in Figure 7.2 we see temporal behaviour in the data sets *Abstract* and *Addresses*. In *Bible* we note two clusters which are the Old and New Testaments.

□

7.2 Alternative Definition

In this section we will give an alternative definition for the CM distance. We have pointed out in Example 7.2 that if we know the frequencies of all itemsets, then the CM distance is basically an L_2 distance between the empirical distributions. It turns out that this property is almost sufficient to characterise the CM distance.

We say that a distance $d(x, y)$ is a *Mahanalobis distance* if it can be expressed as *Mahanalobis distance*

$$d(x, y) = (x - y)^T C (x - y),$$

where C is a symmetric invertible matrix not depending on x or y. Theorem 7.1 shows that the CM distance is a Mahanalobis distance.

Assume that we have a Mahanalobis distance between data sets having the form

$$d(D_1, D_2; \mathcal{F}) = (\theta - \eta)^T C (\theta - \eta),$$

where θ and η are the frequencies of $\mathcal{F}$ calculated from D_1 and D_2, respectively. The matrix C does not depend of the data sets but may depend of $\mathcal{F}$.

[2]The books were taken from http://www.gutenberg.org/etext/8300 in 20. July 2005

[3]The addresses were taken from http://www.bartleby.com/124/ in 17. August 2005

[4]The data set was taken from http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.data.html in 13. January, 2006

We impose the following assumptions on the distance.

(A1) Adding extra attributes, but not changing $\mathcal{F}$, does not change the distance.

(A2) Let $\mathcal{F} \subseteq \mathcal{G}$ be two families of itemsets. Then,

$$d(D_1, D_2; \mathcal{F}) \leq d(D_1, D_2; \mathcal{G}) .$$

(A3) Let $\mathcal{A}$ be the family of all itemsets. Let p_i be the empirical distribution of D_i. Then,

$$d(D_1, D_2; \mathcal{A}) = \|p_1 - p_2\|_2 .$$

Assumption A1 can be justified by noting that we haven't changed anything essential. We have added some extra attributes but they are ignored in $\mathcal{F}$. Assumption A2 states that additional information can only increase the difference between the data sets. Assumption A3 is motivated by Theorem 2.10 which states that we can deduce the empirical distribution if know the frequencies of all itemsets. Hence, we are able to use some distance between the distributions. In this case, we use the L_2 distance.

The following theorem states that the distance satisfying the aforementioned assumptions is essentially the CM distance.

Theorem 7.6 (Theorem 9 in Publication I)**.** *Let $d(D_1, D_2; \mathcal{F})$ be a Mahanalobis distance satisfying Assumptions A1–A3. Let $\mathcal{F}$ be a downward closed family of itemsets. Then,*

$$d(D_1, D_2; \mathcal{F}) = \alpha d_{CM}(D_1, D_2; \mathcal{F}) ,$$

where α is a constant not depending on D_1, D_2, or $\mathcal{F}$.

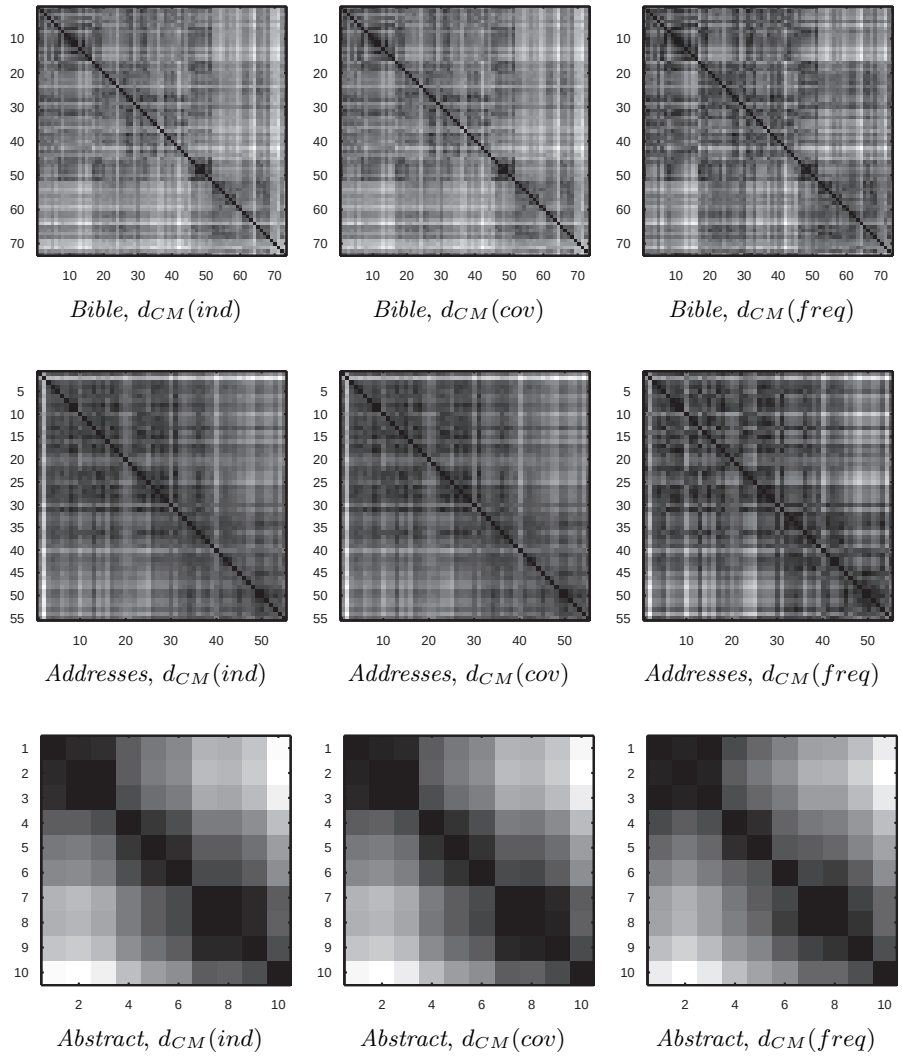


Figure 7.2: Distance matrices for *Bible*, *Addresses*, and *Abstract*. Dark values indicate small distances. In the first column the feature set *ind* contains the independent means, in the second feature set *cov* the pairwise correlation is added, and in the third column the feature set $freq$ consists of $10K$ most frequent itemsets, where K is the number of attributes. Darker colours indicate smaller distances.

Chapter 8

Fractal Dimension of Binary Data

When asked about the dimensionality of a data set one's first answer would be that the dimension of a data set is equal to the number of columns. This, however, is too simplistic approach: Imagine a curve in a plane. The number of columns needed to represent the curve is 2. However, a more natural dimension for the curve is 1. Our goal is to define and measure this intrinsic dimension. The curve example points out that this dimension should take into account the structure of data.

We use fractal dimension, a popular and successful notion, to determine the intrinsic dimension of a binary data set. In Section 8.1 we will introduce the correlation dimension and provide some analysis. The problem with the fractal dimensions is that they are designed for continuous data and have some undesired properties. We remedy these problems in Section 8.2 by defining the normalised correlation dimension.

The contribution of the paper. This chapter is based on Publication V. In the paper we study the idea of using fractal dimension with binary data. We study the behaviour of the correlation dimension, one of the many fractal dimensions. However, the dimension has some undesired properties that are directly related to binary data: For instance, the dimension depends on the sparsity of data and the dimension is not a linear function of the number of attributes. We overcome these problems by defining the normalised correlation dimension. The idea is to compare the correlation dimension of the original data set against the correlation dimension against the correlation dimension of the data set having equal margins but independent attributes. We provide approximations for both dimensions and show empirically that these estimates yield good results. We also compare the dimension against PCA.

There has been a significant amount of work in defining the concept of dimensionality in datasets. Even though most of the methods can be adapted to the case of binary data, they are not specifically tailored for it. For instance, many methods assume real-valued numbers and they compute vectors/components that have negative or continuous values. Such methods include, PCA, SVD, and non-negative matrix factorisation (NMF) [Jol02, LS01]. Other methods such as multinomial PCA (mPCA) [BP03], and latent Dirichlet allocation (LDA) [BNJ03] assume specific probabilistic models of generating the data and the task is to discover latent components in the data rather than reasoning about the intrinsic dimensionality of the data. Methods for exact and approximate decompositions of binary matrices in Boolean semiring have also been proposed [GGM04, MMG+06, MPR95], but similarly to mPCA and LDA, they focus on finding components instead of the intrinsic dimensionality. In addition, many different notions of complexity of binary datasets have been proposed and used in various contexts, for instance VC-dimension [AB97], discrepancy [Cha00], Kolmogorov complexity [LV97] and entropy-based concepts [CT91, POP04]. Finally, methods such as multidimensional scaling (MDS) [Kru64] and Isomap [TdSL00] focus on embedding the data (not necessarily binary) in low-dimensional spaces with small distortion, mainly for visualisation purposes. A key difference with many above approaches is that fractal dimension does not provide a mapping to a lower-dimensional space, and hence traditional applications, such as feature reduction, are not (directly) possible. However, fractal dimension has been used in many applications related to database and data mining, such as, making nearest neighbour computations more efficient [PKF00], speeding up feature selection methods [TJTWF00], outlier detection [PKGF03], and performing clustering tasks based on the local dimensionality of the data points [GHPT05].

8.1 Correlation Dimension

There are infinite number of ways in defining the fractal dimension; see, e.g., [Bar88, Ott97] for a survey. The standard definitions involve usually partitioning the data into infinitesimal pieces and study how the data is distributed with respect to the partition. This cannot be done with finite data but the definitions can be modified to fit our purposes.

correlation dimension

Given a data set D with K columns, let $0 \leq r_1 < r_2 \leq K$. Let Z_D be the distance between two randomly picked points from D. The *correlation dimension* $\mathrm{cd}_{\mathrm{R}}(D; r_1, r_2)$ for a binary data set D and radii r_1 and r_2 is the fraction

$$\mathrm{cd}_{\mathrm{R}}(D; r_1, r_2) = \frac{\log \mathbb{P}(Z_D < r_2) - \log \mathbb{P}(Z_D < r_1)}{\log r_2 - \log r_1},$$

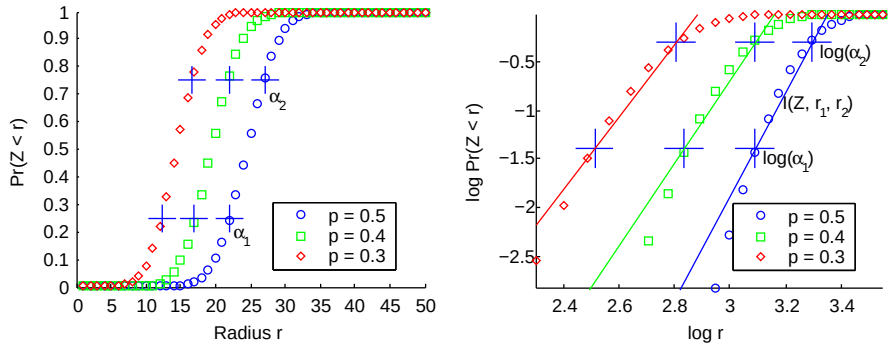


Figure 8.1: Examples of $\mathrm{cd}_{\mathrm{A}}(D)$ for different data sets. Plots represent three different data sets, each of them having 50 independent columns. The probability of a variable being 1 is p (indicated in the legend). The left figure is a regular plot of $\mathbb{P}(Z_D < r)$. The right figure is a log-log plot of $\mathbb{P}(Z_D < r)$. The crosses indicate the end points r_1 and r_2 that were determined by using $\alpha_1 = 1/4$ and $\alpha_2 = 3/4$. The slopes of the straight lines in the log-log plot are $\mathrm{cd}_{\mathrm{A}}(D; 1/4, 3/4)$. Note that the lines are gentler for smaller p.

that is, the correlation dimension is the slope of a line fitted into a log-log plot of the cumulative distribution function of Z_D[1]. For more details about the correlation dimension see e.g., [CY92].

The correlation dimension cd_{R} assumes that the radii r_1 are r_2 are given. The drawback with this approach is that the radii cannot be constant but should depend on the data set: For instance, $r_1 = 25$ and $r_2 = 75$ may be reasonable for a data set with 100 attributes but are absurd for a data set with only 20 columns. To remedy this problem we infer the radii from the distribution of Z_D: Assume that we are given α_1 and α_2 such that $0 \leq \alpha_1 < \alpha_2 \leq 1$. We define $\mathrm{cd}_{\mathrm{A}}(D; \alpha_1, \alpha_2)$ to be $\mathrm{cd}_{\mathrm{R}}(D; r_1, r_2)$, where the radii r_i are set such that $\alpha_i = \mathbb{P}(Z_D < r_i)$. For example,

$$\mathrm{cd}_{\mathrm{A}}(D; 1/4, 3/4) = \frac{\log 3/4 - \log 1/4}{\log r_2 - \log r_1},$$

where r_1 is the lower quartile point and r_2 is the upper quartile point. See Figure 8.1 for an illustration.

A direct analysis of the correlation dimension is difficult. To overcome these problems we define a much simpler quantity that can be used to approximate the

[1]The definition given in Publication V is somewhat more complex. However, the above definition is adequate for our purposes.

correlation dimension.

Given a binary data set D, let Z_D be the distance between two randomly picked points from D. Given a real number $0 < \alpha < 1/2$ we define the *approximative correlation dimension* to be

approximative correlation dimension

$$\operatorname{acd}(D) = \frac{\mathrm{E}[Z_D]}{\mathrm{Std}[Z_D]},$$

where $\mathrm{E}[Z_D]$ is the average distance and $\mathrm{Std}[Z_D]$ is the standard variation of Z_D. We will see in Theorem 8.4 that $\operatorname{acd}(D)$ is asymptotically proportional to the correlation dimension $\operatorname{cd}_{\mathrm{A}}(D; \alpha, 1-\alpha)$.

The following theorem describes $\operatorname{acd}(D)$ when D has independent attributes.

Theorem 8.1 (Proposition 2 in Publication V)**.** *Assume that the data set D has K independent variables, and that the probability of the variable i being 1 is p_i for each i, and let $q_i = 2p_i(1-p_i)$. We have*

$$\operatorname{acd}(D) = \frac{\sum_i q_i}{\sqrt{\sum_i q_i(1-q_i)}},$$

In particular, if all probabilities p_i are equal to p, then for $q = 2p(1-p)$ we have

$$\operatorname{acd}(D) = \sqrt{\frac{Kq}{1-q}}.$$

Corollary 8.2 (Corollary 3 in Publication V)**.** *Assume the data set D has independent columns. The correlation dimension* $\operatorname{acd}(D)$ *is maximised if the variables have frequency $\frac{1}{2}$.*

Given a data set D with K columns, we denote by $\operatorname{ind}(D)$ a random binary data set having K independent variables such that the probability of ith variable being 1 is equal to the probability of ith column of a random transaction sampled from D being 1. Alternatively, $\operatorname{ind}(D)$ can be considered as a data set obtained by permuting each column of D independently. We call $\operatorname{ind}(D)$ a *permuted data set*. By permuting we keep the margins of the individual attributes but destroy any inter-column dependence.

permuted data set

Theorem 8.3 (Discussion after Conjecture 6 in Publication V)**.** *Assume the marginal probability of all original variables are less than* 0.5*, and that all pairs of original variables are positively correlated. Then*

$$\operatorname{acd}(D) \leq \operatorname{acd}(\operatorname{ind}(D)),$$

i.e., the approximative correlation dimension of the original data is not larger than the approximative correlation dimension of the data with each column permuted randomly.

The proof of the theorem is given in Appendix A.4.

The following theorem points out that $\mathrm{acd}(D)$ is asymptotically proportional to the correlation dimension.

Theorem 8.4. *Assume that we have a sequence of independent binary variables X_i. Let D_K be the data set containing K first variables. Assume that $\mathrm{Std}[Z_{D_K}]$ goes to infinity as K approaches infinity. We have*

$$\lim_{K\to\infty} \frac{\mathrm{cd}_{\mathrm{A}}(D_K;\, \alpha, 1-\alpha)}{C(a)\mathrm{acd}(D_K)} = 1,$$

where $C(\alpha)$ is a constant depending only on α.

The proof of the theorem is given in Appendix A.5.

Theorem 8.4 justifies the following approximation.

Approximation 8.5. *Assume that the data set D has K independent variables. Assuming that K is large enough, we have*

$$\mathrm{cd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha) \approx C(\alpha)\mathrm{acd}(D)\,,$$

where $C(\alpha)$ is a constant depending only on α.

The assumption of independence in the statement of Theorem 8.4 is needed for estimating Z_D with a normal distribution. There are alternative versions of central limit theorems that allow non-independency, such as, central limit theorem for m-dependent variables [Ber73]. Hence, we can partly justify Approximation 8.5 for non-independent case.

Example 8.6. We study how accurate is Approximation 8.5 with synthetic data sets. We generated 100 data sets with independent columns and random margins (see Publication V for more details). The results given in Figure 8.2 show that $\mathrm{acd}(D)$ yields a good approximation of the correlation dimension.

□

8.2 Normalised Correlation Dimension

The scale of the correlation dimension is not very intuitive: the dimension of a dataset with K independent variables is not K, although this would be the most natural value. In fact, Theorem 8.1 implies that the correlation dimension is proportional to $\sqrt{K}$ for large K. The correlation dimension gives much smaller values and hence we need some kind of normalisation. Informally, we define the normalised correlation dimension of a dataset D to be the number of variables

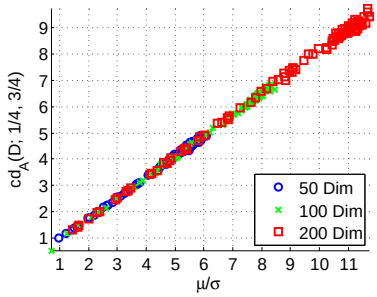


Figure 8.2: Correlation dimension $\mathrm{cd}_{\mathrm{A}}(D;\, 1/4)$ as a function of $\mathrm{acd}(D)$ for data with independent columns (see Proposition 8.5). The y-axis is $\mathrm{cd}_{\mathrm{A}}(D;\, 1/4)$ and the x-axis is $\mathrm{acd}(D) = \mu/\sigma$, where $\mu = \mathrm{E}[Z_D]$ and $\sigma^2 = \mathrm{Var}[Z_D]$. The slope of the line is about $C(1/4) = 0.815$.

that a dataset with independent variables must have in order to have the same correlation dimension as D has.

More formally, let $\mathrm{ind}(H, p)$ be a dataset with H independent variables, each of which is equal to 1 with probability p. From Proposition 8.1 we have an approximation for $\mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(H, p)\,;\, \alpha, 1-\alpha)$: setting $q = 2p(1-p)$ we have

$$\mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(H, p)\,;\, \alpha, 1-\alpha) \approx C(\alpha)\sqrt{\frac{Hq}{1-q}}. \tag{8.1}$$

If the dataset would have the same marginal frequency, say s, for each variable, the normalised correlation dimension of a dataset D could be defined to be the number H such that

$$\mathrm{cd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha) = \mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(H, s)\,;\, \alpha, 1-\alpha)\,.$$

The problem with this way of normalising the dimension is that it takes as the point of comparison a dataset where all the variables have the same marginal frequency. This is very far from being true in real data. We overcome this problem by first finding a value s such that

$$\mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(K, s)\,;\, \alpha, 1-\alpha) = \mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(D)\,;\, \alpha, 1-\alpha)\,,$$

that is, a summary of the marginal frequencies of the columns of D: s is the frequency that variables of an independent dataset should have in order that it has the same correlation dimension as D has when the columns of D have been randomised. We define the *normalised correlation dimension*, denoted by

normalised correlation dimension

$\mathrm{ncd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha)$, to be an integer H such that

$$\mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(H,s)\,;\, \alpha, 1-\alpha) = \mathrm{cd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha)\,.$$

The process is illustrated in Figure 8.3.

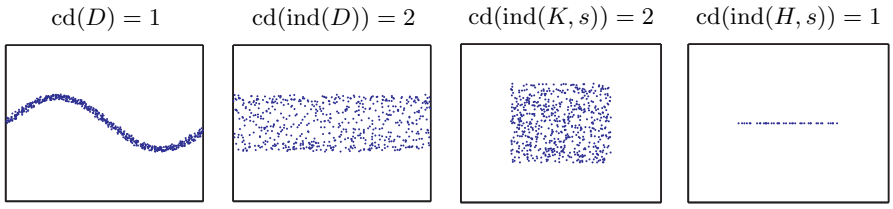


Figure 8.3: An illustration of computing normalised correlation dimension. The original data D is permuted, thus obtaining $\mathrm{ind}(D)$. The margins of $\mathrm{ind}(D)$ are forced to be equal such that the resulting dataset $\mathrm{ind}(K, s)$ has the same correlation dimension. The dataset $\mathrm{ind}(H, s)$ is computed such that $\mathrm{cd}(\mathrm{ind}(H, s)) = \mathrm{cd}(D)$. H is the normalised correlation dimension.

Example 8.7. We examine the normalised correlation dimension using the toy data sets. We generated 100 data sets with independent columns and random margins and calculated $\mathrm{ncd}_{\mathrm{A}}(D;\, 1/4)$ for each data set. Figure 8.4(a) shows that the $\mathrm{ncd}_{\mathrm{A}}(D)$ is concentrated around the number of attributes, as expected, since the attributes are independent. On the other hand, Figure 8.4(b) shows that the sparsity of the data set does not change the normalised correlation dimension. □

The estimate in Eq. 8.1 implies the following approximation.

Approximation 8.8. *Given a data set D with K columns, the normalised dimension $\mathrm{ncd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha)$ can be approximated by*

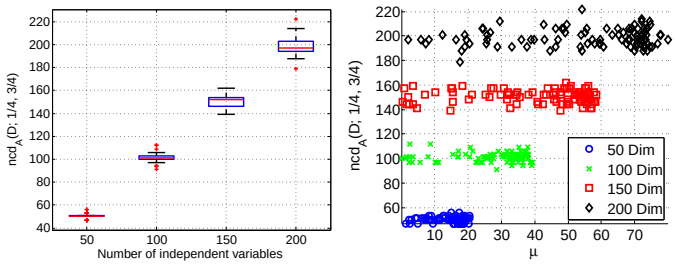
$$\mathrm{ncd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha) \approx \left(\frac{\mathrm{cd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha)}{\mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(D)\,;\, \alpha, 1-\alpha)} \right)^2 K.$$

We can estimate even further by approximating the correlation dimensions $\mathrm{cd}_{\mathrm{A}}(D)$ and $\mathrm{cd}_{\mathrm{A}}(\mathrm{ind}(D))$. This gives us

Approximation 8.9. *Given a data set D with K columns, the normalised dimension $\mathrm{ncd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha)$ can be approximated by*

$$\mathrm{ncd}_{\mathrm{A}}(D;\, \alpha, 1-\alpha) \approx \frac{\mathrm{Var}\left[Z_{\mathrm{ind}(D)}\right]}{\mathrm{Var}[Z_D]} K = \frac{\sum_i C\left(Z_D\right)_{ii}}{\sum_{i,j} C\left(Z_D\right)_{ij}} K,$$

where $C\left(Z\right)$ is the covariance matrix $C\left(Z\right)_{ij} = \mathrm{E}[Z_i Z_j] - \mathrm{E}[Z_i]\,\mathrm{E}[Z_j]$.

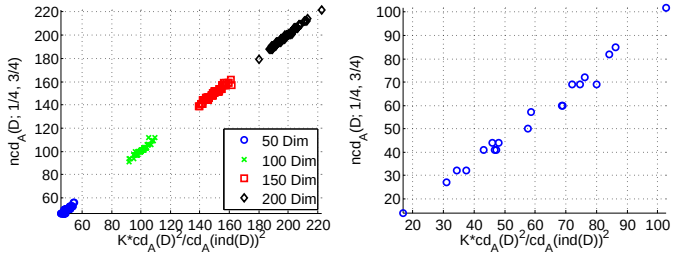


(a) Box plots of $\mathrm{ncd}_\mathrm{A}(D)$. (b) $\mathrm{ncd}_\mathrm{A}(D)$ as a function of μ.

Figure 8.4: Normalised correlation dimension for data having K independent dimensions for $K \in \{50, 100, 150, 200\}$. In Figure 8.4(a) the normalised correlation dimension $\mathrm{ncd}_\mathrm{A}(D)$ is concentrated around the number of attributes. In Figure 8.4(b) $\mathrm{ncd}_\mathrm{A}(D)$ is plotted as a function of μ, the average distance between two random points. The x-axis is $\mu = \mathrm{E}[Z_D]$ and the y-axis is $\mathrm{ncd}_\mathrm{A}(D; 1/4)$.

Note that the estimate in Approximation 8.9 does not depend on α.

Example 8.10. We tested Approximation 8.8 with synthetic data set having independent columns and *20 Newsgroups*[2], a collection of approximately 20 000 newsgroup documents across 20 different newsgroups [Lan95]. Figure 8.5 shows that Approximation 8.8 yields a good estimate for the selected data sets. □



(a) Synthetic data set (b) *20 Newsgroups*

Figure 8.5: Normalised correlation dimension as a function of $K\mathrm{cd}_\mathrm{A}(D)^2 / \mathrm{cd}_\mathrm{A}(\mathrm{ind}(D))^2$. Each point represents one data set. Figure 8.5(a) contains data sets with independent columns and Figure 8.5(b) contains data sets from the *20 Newsgroups* collection.

[2] http://people.csail.mit.edu/jrennie/20Newsgroups/

Appendix A

Proofs for the Theorems

A.1 Proof of Proposition 2.12

The existence of U follows directly from Theorem 2.8. To prove the invertibility of U let $B = \{b_1, \ldots, b_L\}$ be a set of items. We need to show that there is a set of constraints u_C such that

$$S_B(\omega) = \sum_{C \subseteq B} u_C S_{\oplus C}(\omega).$$

We claim that $u_C = (-1)^{|C|-1} 2^{1-L}$. We first note that

$$\sum_{C \subseteq B} u_C S_{\oplus C}(1) = \sum_{C \subseteq B} 2^{1-L} \left[|C| \text{ is odd}\right] = 1 = S_B(1).$$

Now, let ω be a binary vector of length L having some elements as 0. Let $O = \{b_i; \omega_i\}$ be the subset of attributes. We can rewrite the sum as

$$\begin{aligned}
\sum_{C \subseteq B} (-1)^{|C|-1} 2^{1-L} S_{\oplus C}(\omega) &= -2^{1-L} \sum_{X \subseteq O} \sum_{Y \subseteq B-O} (-1)^{|X|+|Y|} S_{\oplus X \oplus Y}(\omega) \\
&= -2^{1-L} \sum_{X \subseteq O} \sum_{Y \subseteq B-O} (-1)^{|X|+|Y|} S_{\oplus X}(\omega) \\
&= -2^{1-L} \sum_{X \subseteq O} (-1)^{|X|} S_{\oplus X}(\omega) \sum_{Y \subseteq B-O} (-1)^{|Y|} \\
&= -2^{1-L} \sum_{X \subseteq O} (-1)^{|X|} S_{\oplus X}(\omega) 0. \\
&= 0 = S_B(\omega)
\end{aligned}$$

This completes the proof.

A.2 Proof of Proposition 4.9

Let

$$p\left(a_1, \ldots, a_K\right) = \prod_{i=1}^{N} p_i\left(C_i\right) / \prod_{j=1}^{N-1} p_j\left(S_j\right).$$

Fix i and set C_i to be the root in T and define P_j to be the separator between the clique C_j and its parent (with respect to the root). Set $P_i = \emptyset$. The distribution p can now be written as

$$p\left(a_1, \ldots, a_K\right) = \prod_{i=1}^{N} p_i\left(C_i; P_i\right). \tag{A.1}$$

Let C_k be a leaf node. There is a node $v \in C_k$ such that $v \notin P_k$. Otherwise, p_k can be removed from the product. Note that v is not included in any other p_j since otherwise the running intersection property is violated. Hence, we can marginalise v out and still have the form of Eq. A.1. We repeat the marginalisation until we are left with p_i. This proves that the marginal distribution of p to C_i is equal to p_i.

A.3 Proof of Theorem 5.6

Before we state the regularity conditions, let us introduce some notation:

- By $\mathrm{E}_\theta[\cdot]$ we denote the mean taking with respect to $p(\omega;\theta)$.
- The partial derivatives $\partial p(\omega;\theta)/\partial\theta_i$ and $\partial^2 p(\omega;\theta)/\partial\theta_i\partial\theta_j$ are shortened into $p_i(\omega;\theta)$ and $p_{ij}(\omega;\theta)$, respectively.
- The partial derivatives $\partial \log p(\omega;\theta)/\partial\theta_i$ and $\partial^2 \log p(\omega;\theta)/\partial\theta_i\partial\theta_j$ are shortened into $l_i(\omega;\theta)$ and $l_{ij}(\omega;\theta)$, respectively.
- A vector (depending on ω and θ) $l(\omega;\theta) = [l_1(\omega;\theta), \ldots, l_K(\omega;\theta)]^T$ is called the score vector.
- A $K \times K$ matrix $I_\theta = \mathrm{E}_\theta\left[l(\omega;\theta)l(\omega;\theta)^T\right]$ is called Fisher's information matrix.

The regularity conditions are:

1. θ_0 is an inner point of Θ, that is, there is an open K-dimensional ball B around θ_0 such that $B \subseteq \Theta$.
2. The family $p(\omega;\theta)$ is homogeneous in B, that is, if $\alpha \in B$ and $p(\omega;\alpha) = 0$, then $p(\omega;\beta) = 0$ for all $\beta \in B$.

3. Derivatives l_i and l_{ij} exist and are continuous (with respect to θ) for each $\omega \in \Omega$ and each $\theta \in B$.

4. θ_n is an efficient asymptotic normal estimate of θ_0, that is, $\theta_n \rightsquigarrow \theta_0$ and $\sqrt{n}\,(\theta_n - \theta_0) \rightsquigarrow N(0, I_{\theta_0}^{-1})$, where $I_{\theta_0}^{-1}$ is the inverse of Fisher's information matrix. Note that we assume that I_{θ_0} is invertible.

Remark A.1. *The definition of weak convergence (or convergence in law) is given in [vdV98, Section 2.1]. We denote the weak convergence by $X_n \rightsquigarrow X$.*

Remark A.2. *We have assumed for simplicity that the sample space Ω is finite. The theorem also holds for general case under some additional regularity conditions.*

We need the following lemmae for proving the theorem:

Lemma A.3. $\mathrm{E}_\theta[l_i(\omega;\theta)] = 0$.

Proof.

$$\mathrm{E}_\theta[l_i(\omega;\theta)] = \sum_{\omega\in\Omega} p(\omega;\theta)\frac{p_i(\omega;\theta)}{p(\omega;\theta)} = \sum_{\omega\in\Omega} p_i(\omega;\theta) = \frac{\partial}{\partial\theta_i}\sum_{\omega\in\Omega} p(\omega;\theta) = \frac{\partial}{\partial\theta_i}1 = 0.$$

□

Lemma A.4. $\mathrm{E}_\theta[l_{ij}(\omega;\theta)] = -\mathrm{E}_\theta[l_i(\omega;\theta)l_j(\omega;\theta)]$.

Proof.

$$\begin{aligned}
\mathrm{E}_\theta[l_{ij}(\omega;\theta)] &= \sum_{\omega\in\Omega} p(\omega;\theta)\frac{\partial}{\partial\theta_j}\frac{p_i(\omega;\theta)}{p(\omega;\theta)} \\
&= \sum_{\omega\in\Omega} p(\omega;\theta)\left[\frac{p_{ij}(\omega;\theta)}{p(\omega;\theta)} - \frac{p_i(\omega;\theta)p_j(\omega;\theta)}{p(\omega;\theta)^2}\right] \\
&= \sum_{\omega\in\Omega} p_{ij}(\omega;\theta) - \sum_{\omega\in\Omega} p(\omega;\theta)\frac{p_i(\omega;\theta)p_j(\omega;\theta)}{p(\omega;\theta)^2} \\
&= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{\omega\in\Omega} p(\omega;\theta) - \sum_{\omega\in\Omega} p(\omega;\theta)l_i(\omega;\theta)l_j(\omega;\theta) \\
&= 0 - \mathrm{E}_\theta[l_i(\omega;\theta)l_j(\omega;\theta)]\,.
\end{aligned}$$

□

Lemma A.5 (Lemma 17.1 in [vdV98])**.** *Let Z be a random vector of length K distributed as $N(0, C)$, where C is invertible. Then $Z^T C^{-1} Z$ is distributed as χ^2 with K degrees of freedom.*

Proof of Theorem 5.6. Let $\alpha, \beta \in B$ be two vectors. Since l_i and l_{ij} are continuous, we can use Multidimensional Taylor's Theorem to obtain

$$\log p(\omega;\alpha) - \log p(\omega;\beta) = (\alpha-\beta)^T l(\omega;\beta) + \frac{1}{2}(\alpha-\beta)^T H(\omega;\gamma)(\alpha-\beta)^T,$$

where $\gamma \in B$ is a vector lying on a segment between α and β, and $H(\omega;\gamma)$ is a Hessian matrix $H_{ij}(\omega;\gamma) = l_{ij}(\omega;\gamma)$. By taking the mean and applying Lemma A.3 we obtain

$$-\mathrm{KL}(\beta;\,\alpha) = \mathrm{E}_\beta[\log p(\omega;\alpha) - \log p(\omega;\beta)] = \frac{1}{2}(\alpha-\beta)^T \mathrm{E}_\beta[H(\omega;\gamma)](\alpha-\beta)^T.$$

Assume for time being that $\theta_n \in B$. Let $\alpha = \theta_0$, $\beta = \theta_n$ and denote the resulting γ by η_n. If θ_n is outside B, set $\eta_n = 0$. Since $\theta_n \rightsquigarrow \theta_0$, we know from Theorem 2.7 in [vdV98] that $\eta_n \rightsquigarrow \theta_0$.

Define $g : \mathbb{R}^K \times \mathbb{R}^K \times \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ to be

$$g(a,b,c,d) = \begin{cases} -a^T \mathrm{E}_b[H(\omega;c)]\,a & \text{, if } b \in B \\ -\frac{2}{d}\mathrm{KL}(b;\,\theta_0) & \text{, if } b \notin B \end{cases}.$$

Note that

$$2n\mathrm{KL}(\theta_n;\,\theta_0) = g(\sqrt{n}\,(\theta_n-\theta_0), \theta_n, \eta_n, 1/n).$$

Let $\sqrt{n}\,(\theta_n-\theta_0) \rightsquigarrow Z$, a random variable distributed as $N\left(0, I_{\theta_0}^{-1}\right)$. Since g is continuous at $(Z,\theta_0,\theta_0,0)$ we can apply Continuous Map Theorem [vdV98, Theorem 2.3] to obtain

$$2n\mathrm{KL}(\theta_n;\,\theta_0) \rightsquigarrow g(Z,\theta_0,\theta_0,0) = -Z^T \mathrm{E}_{\theta_0}[H(\omega;\theta_0)]\,Z.$$

An application of Lemma A.4 leads us to

$$2n\mathrm{KL}(\theta_n;\,\theta_0) \rightsquigarrow Z^T I_{\theta_0} Z.$$

Since $I_{\theta_0}^{-1}$ is the covariance matrix of Z, we can apply Lemma A.5 to obtain the desired result. □

A.4 Proof of Theorem 8.3

We begin by noting that $\mathrm{E}[Z_D] = \mathrm{E}\left[Z_{\mathrm{ind}(D)}\right]$. Let $C(D)$ be the covariance matrix of the distance vector between two random points, that is,

$$C(D)_{ij} = \mathrm{E}[Z_i Z_j] - \mathrm{E}[Z_i]\,\mathrm{E}[Z_j],$$

where Z_i is the indicator variable having value 1 if two randomly chosen elements from D disagree at ith dimension. Note that $\mathrm{Var}[Z_D] = \sum_{ij} C(D)$. Let $U =$

$C(D)$ and $V = C(\mathrm{ind}(D))$. Note that V is a diagonal matrix having the diagonal equal to the diagonal of U. Hence, to prove the theorem we need to show that U contain only positive entries.

Fix j and j and abbreviate

$$x = \mathbb{P}(a_i = 1)\,, \quad y = \mathbb{P}(a_j = 1)\,, \quad z = \mathbb{P}(a_i = 1, a_j = 1)\,.$$

The entry U_{ij} can be written as

$$\begin{aligned} U_{ij} &= 2z(1 - x - y + z) + 2(x - z)(y - z) - 4xy(1 - x)(1 - y) \\ &= 4z^2 + (2 - 4x - 4y)z + 2xy - 4xy(1 - x)(1 - y). \end{aligned}$$

The value of z minimising U_{ij} is

$$z = \frac{4x + 4y - 2}{8} = \frac{1}{2}(x + y) - \frac{1}{4}.$$

Since $x, y \leq \frac{1}{2}$, we have

$$z - xy = \frac{1}{2}(x + y) - \frac{1}{4} - xy = -\left(\frac{1}{2} - x\right)\left(\frac{1}{2} - y\right) \leq 0.$$

But we have assumed that $z \geq xy$, hence U_{ij} obtains its minimum value when $z = xy$, that is, a_i and a_j are independent. In this case $U_{ij} = 0$ and we have proved the statement.

A.5 Proof of Theorem 8.4

We need the following lemma.

Lemma A.6. *Assume sequences x_n, a_n, and b_n such that $x_n \to \infty$, $a_n \to a$ and $b_n \to b$. Then*

$$\left(\frac{x_n + a_n}{x_n + b_n}\right)^{x_n} \underset{n\to\infty}{\to} \exp(a - b)\,.$$

Proof of Theorem 8.4. Recall that

$$\mathrm{cd}_{\mathrm{A}}(D;\, \alpha, 1 - \alpha) = \frac{\log(1 - \alpha) - \log \alpha}{\log r_2 - \log r_1},$$

where r_1 and r_2 are such that $\alpha = \mathbb{P}(Z_D < r_1)$ and $1 - \alpha = \mathbb{P}(Z_D < r_2)$. The numerator is $\log((1-\alpha)/\alpha)$. Assume that K is large enough Let $r_1(K)$ and $r_2(K)$ be the corresponding radii for the dataset D_K.

We next study the denominator $\log r_2 - \log r_1$. We have to analyse the distribution of the random variable Z_D, the L_1 distance between two randomly chosen points from D. For simplicity, we denote Z_{D_K} by Z_K in the sequel. Let Z_i be the indicator variable having value 1 if two randomly chosen elements from X_i disagree; then $Z_K = \sum_{i=1}^{K} Z_i$. Let

$$\mu_K = \mathrm{E}[Z_K]\,, \quad \mu_i = \mathrm{E}[Z_i]\,, \quad \sigma_K^2 = \mathrm{E}\Big[(Z_K - \mu_K)^2\Big]\,.$$

Our goal is to estimate Z_K with the normal distribution. In order to do this we define

$$Y_{K,i} = \frac{Z_i - \mu_i}{\sigma_K}.$$

The sufficient condition for Lindeberg-Feller central limit theorem [vdV98, Theorem 2.27] is that for a fixed $\epsilon > 0$ there is L such that $|Y_{K,i}| \leq \epsilon$, whenever $K > L$. But this is true since $\sigma_K \to \infty$ and $|Z_i - \mu_i| \leq 1$. The central limit theorem implies that

$$\frac{Z_K - \mu_K}{\sigma_K} = \sum_{i=1}^{K} Y_{K,i} \rightsquigarrow N(0,1).$$

This implies that

$$\frac{r_1(K) - \mu_K}{\sigma_K} \to -c \text{ and } \frac{r_2(K) - \mu_K}{\sigma_K} \to c,$$

where c is the inverse of the cumulative distribution function of the normal distribution with parameters 0 and 1, that is, $c = \Phi^{-1}(\alpha) = \sqrt{2}\,\mathrm{erf}^{-1}(2\alpha - 1)$.

Define $e_2(K) = r_2(K) - (\mu_K + c\sigma_K)$ and $e_1(K) = r_1(K) - (\mu_K - c\sigma_K)$. Also, let $z_K = \mu_K/\sigma_K$. Consider the ratio

$$\begin{aligned}\left(\frac{r_2(K)}{r_1(K)}\right)^{z_K} &= \left(\frac{\mu_K + c\sigma_K + e_2(K)}{\mu_K - c\sigma_K + e_1(K)}\right)^{z_K} \\ &= \left(\frac{z_K + c + e_2(K)/\sigma_K}{z_K - c + e_1(K)/\sigma_K}\right)^{z_K}.\end{aligned}$$

Note that $e_1(K)/\sigma_K \to 0$ and $e_2(K)/\sigma_K \to 0$. A straightforward calculation shows that $\mu_K \geq \sigma_K^2$. This implies that

$$z_K = \frac{\mu_K}{\sigma_K} \geq \frac{\sigma_K^2}{\sigma_K} = \sigma_K \to \infty$$

We can now apply the lemma to obtain

$$\left(\frac{r_2(K)}{r_1(K)}\right)^{z_K} \to \exp(2c)\,.$$

Hence, we must have

$$\frac{\mu_K}{\sigma_K}\left(\log r_2(K) - \log r_1(K)\right) = z_K \log \frac{r_2(K)}{r_1(K)} \to 2c.$$

By setting

$$C(\alpha) = \frac{\log((1-\alpha)/\alpha)}{2c} = \frac{\log((1-\alpha)/\alpha)}{2\sqrt{2}\,\mathrm{erf}^{-1}(2\alpha - 1)}$$

we have the desired result. □

Bibliography

[AB97] Martin Anthony and Norman Biggs. *Computational Learning Theory: An Introduction*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 1997.

[AIS93] Rakesh Agrawal, Tomazc Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[AMS+96] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press/The MIT Press, 1996.

[AY98] Charu C. Aggarwal and Philip S. Yu. A new framework for itemset generation. In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 18–24. ACM Press, 1998.

[Bar88] Michael Barnsley. *Fractals Everywhere*. Academic Press, 1988.

[Bas89] Michéle Baseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, 1989.

[Ber73] Kenneth N. Berk. A central limit theorem for m-dependent random variables with unbounded m. *The Annals of Probability*, 1(2):352–354, Apr. 1973.

[BFS03] Pierre Baldi, Paolo Frasconi, and Padhraic Smyth. *Modeling the Internet and the Web*. John Wiley & Sons, 2003.

[BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 265–276. ACM Press, May 1997.

[BMS02] Ella Bingham, Heikki Mannila, and Jouni K. Seppänen. Topics in 0–1 data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 450–455, New York, NY, USA, 2002. ACM.

[BMUT97] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, May 1997.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BP03] Wray Buntine and Sami Perttu. Is multinomial PCA multi-faceted clustering or dimensionality reduction? In C.M. Bishop and B.J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pages 300–307, 2003.

[BRRT05] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, 5(1):231–297, 2005.

[BSH04] Artur Bykowski, Jouni K. Seppänen, and Jaakko Hollmén. Model-independent bounding of the supports of Boolean formulae in binary data. In Pier Luca Lanzi and Rosa Meo, editors, *Database Support for Data Mining Applications: Discovering Knowledge with Inductive Queries*, LNCS 2682, pages 234–249. Springer Verlag, 2004.

[BSS93] Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, second edition, 1993.

[BSVW99] Tom Brijs, Gilbert Swinnen, Koen Vanhoof, and Geert Wets. Using association rules for product assortment decisions: A case study. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, CA, USA*, pages 254–260. ACM, 1999.

[Cal03] Toon Calders. *Axiomatization and Deduction Rules for the Frequency of Itemsets*. PhD thesis, University of Antwerp, Belgium, 2003.

[Cal04] Toon Calders. Computational complexity of itemset frequency satisfiability. In *Proceedings of the 23nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database System*, 2004.

[CDLS99] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and Davig J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag, 1999.

[Cha00] Bernhard Chazelle. *The Discrepancy Method*. Cambridge University Press, 2000.

[Coo90] Gregory Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, Mar. 1990.

[Csi75] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, Feb. 1975.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[CY92] Sangit Chatterjee and Mustafa R. Yilmaz. Chaos, fractals and statistics. *Statistical Science*, 7(1):49–68, Feb. 1992.

[Dan51] George B. Dantzig. Programming of interdependent activities, II, mathematical model. In Koopmans [Koo51], pages 19–32.

[Dan63] George B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.

[DF00] Adrian Dobra and Stephen E. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 97(22):11885–11892, Oct. 2000.

[DHS00] Richard Duda, Peter Hart, and David Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition, 2000.

[DP01] William DuMouchel and Daryl Pregibon. Empirical bayes screening for multi-item associations. In *Knowledge Discovery and Data Mining*, pages 67–76, 2001.

[DR72] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[DS40] W. Edwards Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, Dec 1940.

[EM97] Thomas Eiter and Heikki Mannila. Distance measures for point sets and their computation. *Acta Informatica*, 34(2):109–133, 1997.

[FGJM06] Mikael Fortelius, Aristides Gionis, Jukka Jernvall, and Heikki Mannila. Spectral ordering and biochronology of european fossil mammals. paleobiology. *Paleobiology*, 32(2):206–214, 2006.

[For05] Mikael Fortelius. Neogene of the old world database of fossil mammals (NOW). University of Helsinki, http://www.helsinki.fi/science/now/, 2005.

[GCB07] Arianna Gallo, Nello Cristianini, and Tijl De Bie. Mini: Mining informative non-redundant itemsets. In *11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 438–445, 2007.

[GGM04] Floris Geerts, Bart Goethals, and Taneli Mielikäinen. Tiling databases. In Einoshin Suzuki and Setsuo Arikawa, editors, *Discovery Science*, volume 3245 of *Lecture Notes in Computer Science*, pages 278–289. Springer, 2004.

[GHPT05] Aristides Gionis, Alexander Hinneburg, Spiros Papadimitriou, and Panayiotis Tsaparas. Dimension induced clustering. In Robert Grossman, Roberto Bayardo, and Kristin P. Bennett, editors, *KDD*, pages 51–60. ACM, 2005.

[GKM03] Aristides Gionis, Teija Kujala, and Heikki Mannila. Fragments of order. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–136, New York, NY, USA, 2003. ACM.

[GKP88] George Georgakopoulos, Dimitris Kavvadias, and Christos H. Papadimitriou. Probabilistic satisfiability. *Journal of Complexity*, 4(1):1–11, March 1988.

[GWBV03] Karolien Geurts, Geert Wets, Tom Brijs, and Koen Vanhoof. Profiling high frequency accident locations using association rules. In *Proceedings of the 82nd Annual Transportation Research Board, Washington DC. (USA), January 12-16*, 2003.

[Hai65] Theodore Hailperin. Best possible inequalities for the probability of a logical function of events. *The American Mathematical Monthly*, 72(4):343–359, Apr. 1965.

[HAK+02] Jiawei Han, Russ B. Altman, Vipin Kumar, Heikki Mannila, and Daryl Pregibon. Emerging scientific applications in data mining. *Communications of the ACM*, 45(8):54–58, 2002.

[Hay98] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2nd edition, 1998.

[HFEM07] Hannes Heikinheimo, Mikael Fortelius, Jussi Eronen, and Heikki Mannila. Biogeography of european land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6):1053–1064, 2007.

[HHM+07] Hannes Heikinheimo, Eino Hinkkanen, Heikki Mannila, Taneli Mielikäinen, and Jouni K. Seppänen. Finding low-entropy sets and trees from binary data. In *Knowledge Discovery and Data Mining*, 2007.

[HMS01] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 2001.

[HSM03] Jaakko Hollmén, Jouni K Seppänen, and Heikki Mannila. Mixture models and frequent sets: combining global and local methods for 0-1 data. In *Proceedings of the SIAM Conference on Data Mining (2003)*, 2003.

[Jay57] Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, May 1957.

[Jol02] Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2nd edition, 2002.

[Jor99] Michael I. Jordan, editor. *Learning in graphical models*. MIT Press, 1999.

[JP95] Radim Jiroušek and Stanislav Přeušil. On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics and Data Analysis*, 19:177–189, 1995.

[JS02] Szymon Jaroszewicz and Dan A. Simovici. Pruning redundant association rules using maximum entropy principle. In *Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference, PAKDD'02*, pages 135–147, May 2002.

[JS04] Szymon Jaroszewicz and Dan A. Simovici. Interestingness of frequent itemsets using bayesian networks as background knowledge. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 178–186, New York, NY, USA, 2004. ACM.

[JS05] Szymon Jaroszewicz and Tobias Scheffer. Fast discovery of unexpected patterns in data, relative to a bayesian network. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 118–127, New York, NY, USA, 2005. ACM.

[KBF+00] Ron Kohavi, Carla Brodley, Brian Frasca, Llew Mason, and Zijian Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations*, 2(2):86–98, 2000.

[Kha79] Leonid G. Khachian. A polynomial algorithm for linear programming. *Doklady Akad. Nauk USSR*, 244(5):1093–1096, 1979. Translated in *Soviet Math. Dokldady*, 20, 191–194.

[KL51] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mar. 1951.

[Koo51] Tjalling C. Koopmans, editor. *Activity Analysis of Production and Allocation*. John Wiley & Sons, 1951.

[KRS02] Ron Kohavi, Neal J. Rothleder, and Evangelos Simoudis. Emerging trends in business analytics. *Communications of the ACM*, 45(8):45–48, 2002.

[Kru64] Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of t to a nonmetric hypothesis. *Psychometrica*, 29:1–26, 1964.

[Kul68] Solomon Kullback. *Information Theory and Statistics*. Dover Publications, Inc., 1968.

[Lan95] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[LS90] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. pages 415–448, 1990.

[LS01] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562, 2001.

[Luk01] Thomas Lukasiewicz. Probabilistic logic programming with conditional constraints. *ACM Transactions on Computational Logic (TOCL)*, 2(3):289–339, July 2001.

[LV97] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer-Verlag, 3rd edition, 1997.

[Meo00] Rosa Meo. Theory of dependence values. *ACM Trans. Database Syst.*, 25(3):380–406, 2000.

[MMG+06] Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila. The discrete basis problem. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Knowledge Discovery in Databases: PKDD 2006 – 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, Croatia, September 18–22, 2006, Proceedings*, Lecture Notes in Computer Science. Springer, 2006.

[MPR95] Sylvia D. Monson, Norman J. Pullman, and Rolf Rees. A survey of clique and biclique coverings and factorizations of $(0, 1)$-matrices. *Bulletin of the ICA*, 14:17–86, 1995.

[MT96] Heikki Mannila and Hannu Toivonen. Multiple uses of frequent sets and condensed representations (extended abstract). In *Knowledge Discovery and Data Mining*, pages 189–194, 1996.

[MTV97] Heikki Mannila, Hannu Toivonen, and Aino Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289, 1997.

[Omi03] Edward R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, 2003.

[Ott97] Edward Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 1997.

[Pea88] Judea Pearl. *Probabilistic Inference in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[PKF00] Bernd-Uwe Pagel, Flip Korn, and Christos Faloutsos. Deflating the dimensionality curse using multiple fractal dimensions. In *ICDE*, pages 2589–2598. IEEE Computer Society, 2000.

[PKGF03] Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B. Gibbons, and Christos Faloutsos. LOCI: fast outlier detection using the local correlation integral. In Umeshwar Dayal, Krithi Ramamritham, and T. M. Vijayaraman, editors, *ICDE*, pages 315–326. IEEE Computer Society, 2003.

[PMS03] Dmitry Pavlov, Heikki Mannila, and Padhraic Smyth. Beyond independence: Probabilistic models for query approximation on binary transaction data. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1409–1421, 2003.

[POP04] Paolo Palmerini, Salvatore Orlando, and Raffaele Perego. Statistical properties of transactional databases. In Hisham Haddad, Andrea Omicini, Roger L. Wainwright, and Lorie M. Liebrock, editors, *SAC*, pages 515–519. ACM, 2004.

[PS91] Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[PS98] Christos Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization Algorithms and Complexity*. Dover, 2nd edition, 1998.

[Qui93] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, Oct. 1948.

[TdSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[TJTWF00] Caetano Traina Jr., Agma J. M. Traina, Leejay Wu, and Christos Faloutsos. Fast feature selection using fractal dimension. In Karin Becker, Adriano Augusto de Souza, Damires Yluska de Souza Fernandes, and Daniela Coelho Freire Batista, editors, *SBBD*, pages 158–171. CEFET-PB, 2000.

[vdV98] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Index