Meaningful objective frequency-based interesting pattern mining

Thomas Delacroix

DOCTORAT BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

Thomas DELACROIX

Meaningful objective frequency-based interesting pattern mining

Extraction objective et signifiante de motifs intéressants sur la base de leur fréquence

Thèse présentée et soutenue à IMT Atlantique, le 21 mai 2021
Unité de recherche : UMR CNRS 6285 Lab-STICC
Thèse N° : 2021IMTA0251

Rapporteurs avant soutenance :

Jean-Paul Haton	Professeur émérite, Université de Lorraine
Gilbert Saporta	Professeur émérite, Conservatoire National des Arts et Métiers

Composition du Jury :

Président :	Jérôme Azé	Professeur, Université de Montpellier
Examinateurs :	Jean-Paul Haton	Professeur émérite, Université de Lorraine
	Gilbert Saporta	Professeur émérite, Conservatoire National des Arts et Métiers
	Pascale Kuntz	Professeure, Université de Nantes
	Franck Vermet	Maître de conférences, Université de Bretagne Occidentale
Dir. de thèse :	Philippe Lenca	Professeur, IMT Atlantique

Contents

CHAPTER 1

Foreword

1.1 The two cultures of statistical modeling

In 2001, Professor Leo Breiman described "*Two Cultures*" of statistical modeling: the data modeling culture, on the one side, and the algorithmic modeling culture, on the other side [B$^+$01]. According to Breiman, both cultures study the same object: data consisting of a vector $\mathbf{x}$ of input variables and a vector $\mathbf{y}$ of output variables. Nature operates in a certain manner to associate to each input variable $\mathbf{x}$ an output variable $\mathbf{y}$ but, to the data scientist, this is a black box. Both cultures also share the same two goals: prediction (what will be the outputs to future inputs?) and information (how does nature associate outputs to inputs?). However, their approaches differ. On one side of the spectrum, members of the data modeling culture which, he argued, was the mainstream culture of the statistics community at that time, would try to fill in the black box with a statistical model such as linear regression or logistic regression. On the other side, members of the algorithmic modeling culture would simply substitute nature's black box by another black box consisting of objects such as neural nets, forests or support vectors.

Professor Breiman was a strong proponent of the algorithmic modeling culture. According to this statistician, the data modeling culture belonged to the past as it could not answer the problems of the twenty first century and its huge data flow. Conversely, the algorithmic modeling culture was the way of the future and was supported by a younger generation of computer scientists, physicists and engineers. He argued that the gains in the accuracy of the predictions brought forth by the methods within this developing culture trumped the losses that this gain induced in terms of interpretability. This article gave rise to a certain number of critical comments, notably from the statisticians community, some of which were published along with the article [Cox01, Efr01]. However, the vision it defended also appealed to a certain number of actors in the data mining and machine learning communities who could identify to this algorithmic modeling culture. Such opinions became increasingly popular and trending in these communities and eventually mainstream, thus making room for even more radical visions.

1.2 The end of theory

In June 2008, WIRED magazine released an issue containing an editorial by Chris Anderson, its editor-in-chief at the time, entitled "*The end of theory : The data deluge makes the scientific method obsolete*" [And08]. In this article, Anderson wrote:

> *Petabytes allow us to say: "correlation is enough". We can stop looking for models. [...] Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?*

The point developed by Chris Anderson was that the vast amounts of data now available, a.k.a. Big Data, could be processed and mined to reveal information (in this case, scientific knowledge) without the need to understand where this information came from and if it could be explained. Once more, this article brought forth some strong opposition from members of the scientific community [Pig09, Man13, Maz15, LLLW16, Got16] but it also obtained a large adherence within the data mining and machine learning communities and industries. If a system works, why would you need to understand how or why it works? From a commercial and industrial perspective, this makes perfect sense: if you only need a system to work, and you trust this system to work, trying to understand how and why it works would simply represent an additional and superfluous cost. As cost efficiency is a main drive for commercial and industrial actors, meaningfulness is often absent in data mining and machine learning algorithms. The opinion that meaningfulness and interpretability could be sacrificed to the benefit of higher accuracy in predictions, or simply economic efficiency, has continued its expansion up until today both within the computer science community and more broadly, to businesses and society in general [O'N16, RS17]. More recently, on the 27th of March 2019 the Association for Computing Machinery awarded the 2018 ACM Alan Mathison Turing Award to Yoshua Bengio, Geoffrey Hinton and Yann LeCun the "*Fathers of the Deep Learning Revolution*" [ACM], a revolution which led to the proliferation of algorithmic black boxes as described by Leo Breiman [VBB+18b].

1.3 The right to an explanation

The general development of this opinion and particularly the consequences of its influence, ranging from the production of scientific research to the implementations of everyday algorithms, has also led many to oppose and reject it These include of course statisticians and members of the scientific community, but also men and women within the civil society thus inciting policy makers to react [RS17, O'N16, CGM14, ZBB+17, VBB+18a, EU216]. In her 2016 best-seller, *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy*, Cathy O'Neil advocates for transparency, fairness and accountability in algorithmic models [O'N16]. "*Opaque and invisible models are the rule,*" she says, "*and clear ones very much the exception*", but if a decision is made based on the conclusions of a black box algorithm, how can this decision be contested? The accountability of algorithmic models depends therefore on their explainability. Such accountability became mandatory in the European Union after the implementation of the General Data Protection Regulation in May 2018 [EU216] which also introduced a new "*right to an explanation*" of automated decision-making [EV18]. In 2018, Cédric Villani, Fields medalist and member of the French parliament, produced a six part report on artificial intelligence for the French government, one of which deals solely with the ethics of artificial intelligence [VBB+18a]. In this section, he defends that priority should be given to increasing the transparency of automated processes and the possibility to audit them, by massively funding research on explainability and aiming at "*opening the black boxes*" within machine learning algorithms. Explaining artificial intelligence is also a priority for the United

States military which started funding the "*Explainable Artificial Intelligence (XAI)*" program through its Defense Advanced Research Projects Agency in 2016 [Gun17]. This ongoing program funds thirteen research projects which aim to "*produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)*" and "*enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners*".

1.4 A culture shock

Coming from a background in fundamental mathematics, specifically algebraic geometry, I was myself very much surprised by the algorithmic model culture which I discovered in the field of machine learning and data mining. Indeed, the fundamental mathematics culture is one based on the notion of deductive reasoning, while the algorithmic model culture is based on inductive reasoning. In a sense, the scientific culture is located somewhere in between the two because it is a subtle mix between logical and empirical approaches [Pop59, LWC82, Cha13, Pot17]. That is why pure mathematics is generally not considered a science but rather an art ([Rus07, Har40, Dev00, Cel15]), a language ([Gal23, FP88, CC89, Sch71, O'H04, RSHF15]) or a philosophy ([Kör60, GV01, Jac01, Bro08]). Conversely, pure empiricism is no more than observation.

From my perspective as a mathematician, good mathematical modeling is an inextricable part of high quality scientific research and an indispensable means to achieve explainability, meaningfulness and accountability. I therefore needed to understand why the mainstream culture of an entire scientific community (i.e. the machine learning and data mining communities) would be adverse to mathematical modeling. Many aspects of this thesis may be better understood in light of this culture shock and the subsequent will to create a bridge between different cultures. In a sense, this can be seen as the underlying premise of this thesis as it is founded on the coming together of three different persons representing three different cultures: my supervisor, Philippe Lenca, who is a computer scientist, my co-adviser, Stéphane Lallich, who is a statistician, and myself, a mathematician.

How my mathematical knowledge and mathematical approach could benefit the fields of data mining, machine learning and artificial intelligence was a question I kept in mind during the entire duration of my doctoral research. But I also came to understand much more about my own culture, about mathematics and mathematical modeling; how and why they bring meaningfulness to scientific research. I also came to understand their limitations and the need for compromise in order to reach and satisfy the goals and expectations of the field in which I was working.

1.5 Objective frequency-based interesting pattern mining

The specific domain in which I conducted my research was frequency-based[1] pattern mining with a particular focus on rule and itemset mining as well as the search for objectively interesting patterns. I present a brief review of the history and state-of-the-art of these topics in chapter 2.

One of the first goals which we set up together with my supervisor was to try to understand if and how meaningfulness was a criteria for defining interestingness in current itemset and rule mining procedures. This exercise helped me establish a link between meaningfulness and mathematical modeling. I therefore undertook the task of characterizing mathematical modelings in frequency-based pattern mining while keeping in mind their propensity to be explained and carry meaning. The product of this research constitutes chapter 3 of this thesis. Its contents pertain to the fields of pattern mining, applied mathematics and the philosophy of science. The contributions of the thesis within chapter 3 include:

- The presentation of a novel framework for the qualitative analysis of modeling processes and their meaningfulness comprising a new formulation for the notions of model and modeling and the definition of new concepts for the characterization of modeling processes such as phenotypic and genotypic modeling, pragmatic modeling, and patchwork and holistic modeling.

[1] We use the terminology *frequency-based pattern mining* rather than the more common *frequent pattern mining* terminology because this last term is used ambiguously to refer both to the process of mining patterns that are frequent in the data (*frequent pattern mining*) or to the process of mining patterns based on their frequency in the data (what we call *frequency-based pattern mining*).

- An analysis of the impact in terms of meaningfulness of various modeling choices in frequency-based pattern mining based on these tools.

- The presentation of a Boolean lattice type structure fr representing the patterns to be mined and the demonstration that there are objective arguments for considering itemsets rather than other types of patterns (including association rules) in frequency-based pattern mining.

- A new approach towards reconciling the hypothetico-deductive model of the scientific method and frequency-based pattern mining, based on a notion of confidence in the empirical data.

While chapter 3 focuses on establishing a number of general recommendations for the definition of meaningful mathematical modelings in objective frequency-based interesting pattern mining, chapter 4 concentrates on the definition of specific mathematical models following some of these recommendations. The main objects in chapter 4 are mutual constrained independence (MCI) models which are particular cases of MaxEnt models. The definition of these models is the result of a two phase process which emerged during my doctoral research: first, a more specific model (easier to formalize and compute); second, a generalization of this previous model whose definition and computation rely on much more elaborate mathematical tools, notably based on modern algebraic geometry. The presentation of the mathematical modelings which lead to these models, together with their mathematical definitions and properties, as well as novel algorithms for computing them, make up for chapter 4 of this thesis. The contributions of the thesis within chapter 4 include:

- The mathematical proofs for the existence and the characterization of MCI models, as well as their relationship to MaxEnt models.

- The algebraic expressions and mathematical properties for MCI models in which the constraints are defined on all proper subitemsets of an itemset.

- The algebraic expressions for all MCI models in which the number of items m is equal to 4 or less.

- The presentation of an algorithm based on tools from algebraic geometry to determine the algebraic expression for any MCI model.

The last chapter of this doctoral thesis concentrates on mining algorithms which allow to extract interesting patterns, following the principles for mathematical modeling elaborated in chapter 3 and the mathematical tools developed in 4. As the limits of a direct implementation of the principles previously defined are made apparent, further paths in research are suggested. The contributions of the thesis within chapter 5 include:

- The reduction of objective frequency-based interesting pattern mining to a specific mathematical optimization problem.

- The presentation of an operational algorithm, based on a greedy strategy, for solving this problem.

1.6 Acknowledgments

This thesis is dedicated to my son Sacha Ali Karl Delacroix Sadighiyan who came into this world halfway through the completion of my doctoral research. Thank you for giving me the strength to go through with this task. I hope that you will be proud of me for this accomplishment as you grow up. I thank my wife Saena for her continuing support and love, as well as the inspiration she has provided to me during all these years. I thank my parents for their support without which I would not have managed to pursue this work. I thank my supervisor, Philippe Lenca, and co-adviser, Stéphane Lallich, for accompanying me and advising me all along the way. I thank Jean-Paul Haton and Gilbert Saporta for their meticulous work in reviewing my thesis. I also thank the further members of the jury Jérome Azé, Pascale Kuntz and Franck Vermet for assessing my research. I thank Alain Hillion and Cédric Wemmert for their participation in the Follow-up Committee for this thesis. I would like to thank all my former colleagues, particularly those at IMT Atlantique and the University of Strasbourg, who have helped me develop my understanding of the various fields of study that I have approached in this thesis. I would also like to thank the researchers that I have met and discussed with at all the seminars and conferences which I have attended during these years. I thank all the researchers whom I have not met and whose research has nourished my own. I thank all the friends that have followed me through this endeavor and who have helped my ideas mature through our discussions. Last, I thank

the open science community and particularly Alexandra Elbakyan for their invaluable contribution to scientific progress.

CHAPTER 2

Mining objectively interesting itemsets and rules - History and state-of-the-art

2.1 Early developments

For more than a century now, quite a few decades before the dawn of the Information Age, mathematicians and scientists alike have tried to define methods for exhibiting interesting associations and relations between nominal attributes based on observed frequencies in data. As such, Karl Pearson, who was already working "*on the correlation of characters not quantitatively measurable*" at the end of the nineteenth century by analyzing contingency tables [Pea00], can be seen as a pioneer. By developing the first chi-squared test in 1900 and later adapting it to be able to test for independence in contingency tables in 1904 [Sti02], he set the foundations to the first formal mathematical method for analyzing such associations. His method was specified and corrected by Ronald A. Fisher in the 1920s to obtain what is now taught as Pearson's chi-squared test for statistical independence [Fis22, Fis24, Jay03]. This method, as well as most early methods on nominal attribute association analysis, were only aimed at rejecting (or not) hypotheses individually defined by humans and which mostly involved only two different attributes.

The development of the computer and the beginning of the digital revolution allowed for much more possibilities. New approaches started to develop aiming both at considering relations and associations between much more than simply two attributes and at using automated systems to discover these associations. In 1966, a team of researchers at the University of Prague, Peter Hàjek, Ivan Havel and Metoděj Chytil, published a paper entitled "*The GUHA*

Method of Automatic Hypotheses Determination" [HHC66] which presented the foundations of a theory for extracting logical relations between nominal attributes of significant interest. This theory, deeply rooted in fundamental logics and statistical analysis, continued its development with a number of puplished scientific articles and a book "*Mechanizing Hypothesis Formation (mathematical foundations for general theory)*", published by Springer-Verlag in 1978, but its impact remained quite local, mostly at the University of Prague [HH12, Háj01, Hol98, HHR10]. It seems that the combined barriers represented by the language (a lot of publications on the GUHA were in Czech), the geopolitical situation at the time (Prague was located behind the Iron Curtain), cultural differences in terms of scientific culture (the literature on the GUHA method, even in English, can seem exceedingly cryptic for someone who is not accustomed with the formal theory of mathematical logic) and simply bad timing (the GUHA method developed at the early beginning of the digital revolution) did not allow for a large dissemination and global recognition of the historical precedence in the field of itemset and rule mining of the work conducted on the GUHA method in Prague.

In the 1980s, numerous studies for characterizing rule type patterns and extracting them from datasets started to develop [Cen87, CN89, Qui87, LGR81a, PS91]. While some notions and approaches stayed quite confined to a restricted circle (see, for example, statistical implicative analysis which continues its mostly separate development up to this day [GL92, GBPP98, GRMG13]), a few dominant trends and key notions started to emerge. These included a differentiation between rules that are always correct (exact rules) and rules that are almost always correct (strong rules), together with the idea that the strength (or interest) of a rule could be characterized by a "*rule-interest measure*" (the terminology later settling on "*interestingness measure*").

Enjoying now a much more favorable conjuncture and building on the principles for characterizing strong rules suggested by Gregory Piatetsky-Shapiro [PS91], the framework for mining frequent itemsets and association rules which was developed at the IBM Almaden Research Center in the 1990s became quickly widespread. This framework was first presented in Mining Association Rules between Sets of Items in Large Databases, a 1993 article by Rakesh Agrawal, Tomasz Imieliński and Arun Swami [AIS93] and quickly followed in 1994 by Fast Algortihms for Mining Association Rules by Rakesh Agrawal and Ramakrishnan Srikant [AS94a] which presented the now famous Apriori algorithm. By contrast with the GUHA method, which contained descriptions of objects equivalent to association rules in its 1966 version, the great simplicity of both the framework and the Apriori algorithm made it easily accessible to a large number. Furthermore, the presentation of the framework towards its direct application to market basket analysis, in both articles, undoubtedly appealed to many more. Through the general adherence it had gained, this framework was rapidly established as the reference for mining associations and relations between nominal attributes in data. The Apriori algorithm figured in the famous list of the "*Top 10 algorithms in Data Mining*" [WKRQ$^+$08] and the current most popular textbooks in data mining all confer an important part to frequent itemset mining and association rule mining [HPK11, ZMJM14, Agg15, HTF09, TSKK18]. As for the two articles figuring above, they now rank amongst the most cited articles in the field of data mining, or even within the more general field of computer science, with respectively 21,285 and 24,654 referenced citations by Google Scholar as of August 2019.

When conducting research in the field of itemset mining, a recollection of its genealogy as presented above is far from superfluous. The canons in terms of terminology and representations have been defined and structured by this history. It is quite notable that the presentations of modern itemset and rule mining have strongly inherited from their early ties to market basket analysis. Such examples as the common storytelling of the discovery of a relation between beer and diaper sales by a large retail store company, even though it is mainly mythology (see [Pow02] for a detailed explanation), helped to forge many of the representations of the people who study and work with itemset mining. Understanding the history allows one to perceive the reasons behind certain of the constraints and limitations of the mainstream framework and question their necessity. By acknowledging that not all choices by a given scientific community are motivated by scientific reasons, it is easier to depart from this mainstream framework when necessary.

2.2 Frequent itemsets and association rules

2.2.1 The models

2.2.1.1 Frequent itemsets

The canonical example of a dataset in itemset mining is a dataset consisting of information on the purchases of customers at a retail store which sells items such as beer, diapers, eggs, milk and bread. The information in the data is very basic. For each individual purchase, called a transaction, we know which set of items were bought from the store. We do not know which quantity of each item was purchased nor their price, neither do we possess any further information regarding the transactions (such as a customer ID or a time of purchase), we only know that it corresponds to a unique transaction. This gives rise to the following data representation: a table of transaction IDs together with the corresponding transactions. The transaction IDs (often generally shortened to TID) can be any unique identifier and, although they are usually numerical, their ordering is irrelevant to the model. The transactions themselves can be represented as sets of items also known as itemsets.

TID	Transaction
0	$\{a_1, a_3\}$
1	$\{a_1, a_2\}$
2	$\{a_1, a_3, a_4\}$
3	$\{a_3, a_4\}$
4	$\{a_2\}$
5	$\{a_2, a_4\}$
6	$\{a_3\}$
7	$\{a_1, a_3\}$
8	$\{a_2, a_3, a_4\}$
9	$\{a_1, a_3\}$

Table 2.1: A toy database with four items and ten transactions.

The corresponding mathematical model can be formalized as follows. Let $\mathcal{I} = \{a_1, ..., a_m\}$ be a set of m elements hereinafter referred to as items. Define an itemset X to be a subset of the set of items $\mathcal{I}$, i.e. $X \subset \mathcal{I}$. Define a database of transactions to be a set $\mathcal{T} = \{T_1, ..., T_n\}$ of n itemsets hereinafter referred to as transactions. The support of an itemset X in a database of transactions

$\mathcal{T}$ is the proportion $\text{supp}_{\mathcal{T}}(X)$ (or simply $\text{supp}(X)$ if there is no ambiguity) of transactions in $\mathcal{T}$ that contain X:

$$\text{supp}(X) = \frac{\text{card}\left(\{T \in \mathcal{T} \mid X \subset T\}\right)}{\text{card}\left(\mathcal{T}\right)}$$

The frequent itemset mining problem is defined as determining the set of itemsets $\mathcal{FI}_{\mathcal{T},\text{minsupp}}$ (or simply $\mathcal{FI}$ if there is no ambiguity) for which the support in the database $\mathcal{T}$ is greater than a given threshold minsupp.

Itemset	Support
$\{a_3\}$	0.7
$\{a_1\}$	0.5
$\{a_1, a_3\}$	0.4
$\{a_2\}$	0.4
$\{a_4\}$	0.4
$\{a_3, a_4\}$	0.3

Table 2.2: The frequent itemsets ordered by support for minsupp = 0.3 for the database in Table 2.1.

There is much to say about the adequacy and the meaningfulness of the modeling choices resulting in the definition of the transaction/itemset model and the frequent itemset problem. This will be addressed specifically in the next chapter.

2.2.1.2 Association rules

Historically, the mining of rule type patterns was the main focus of frequent pattern mining because rules offered greater interpretability than most other patterns. Itemsets, which are conjunction type patterns, were originally only designed for the purpose of defining association rules [AIS93, AS94a]. An association rule is a rule between two disjoint itemsets X and Y, noted $X \to Y$, which is characterized by two measures. The support of the rule, noted $\text{supp}(X \to Y)$, measures the observed frequency of the rule, that is the proportion of transactions that contain both itemsets. The confidence of the rule, noted $\text{conf}(X \to Y)$, measures the observed frequency of Y when X occurs. Formally this is expressed as follows.

Consider a set of items $\mathcal{I}$ and a database of transactions $\mathcal{T}$as defined previously. Let $X, Y \subset \mathcal{I}$. Then $X \to Y$ is an association rule if $X \cap Y = \emptyset$ and

$\mathrm{supp}(X \cup Y) \neq 0$. The antecedent and the consequent of the rule are X and Y respectively. The support and confidence of the rule are defined by

$$\mathrm{supp}(X \to Y) = \mathrm{supp}(X \cup Y)$$

and

$$\mathrm{conf}(X \to Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)}$$

The association rule mining problem is defined as determining the set of association rules $\mathcal{AR}_{\mathcal{T},\mathrm{minsupp},\mathrm{minconf}}$ (or simply $\mathcal{AR}$ if there is no ambiguity) for which the support in the database is greater than a given threshold minsupp and the confidence is greater than a given threshold minconf.

Association rule	Support	Confidence
$\{a_1\} \to \{a_3\}$	0.5	1
$\{a_3\} \to \{a_1\}$	0.5	0.71
$\{a_4\} \to \{a_3\}$	0.3	0.75

Table 2.3: Association rules for minsupp = 0.3 and minconf = 0.5 for the database in Table 2.1.

2.2.2 Algorithms

Solving both the frequent pattern mining problem and the association rule mining problem has been an important focus of research in this field and, as a result, quite a few algorithms for solving these problems have been suggested. As this would lead us astray from the scope of this thesis, we do not dwell much on the specifics of these different mining methods. We will briefly present the general principles behind the Apriori algorithm, as these are intrinsically tied to the elaboration and development of the frequent itemset and association rule models, and simply recall the subsequent development of alternative approaches.

2.2.2.1 The Apriori algorithm

The Apriori algorithm holds a special place in the history of frequent itemset mining because the development of the field itself can be partly attributed to the popular success of this algorithm. Furthermore, even though it was presented in the second of the two founding articles previously cited, its underlying principles were already present in the first article. As we will discuss in the following chapter, this raises issues regarding the rationale behind the definition of the frequent pattern and association rule models.

The Apriori algorithm relies on two aspects which are related to the representation and definition of these models. The first aspect on which it relies is the underlying lattice structure of the set of itemsets ordered by inclusion.

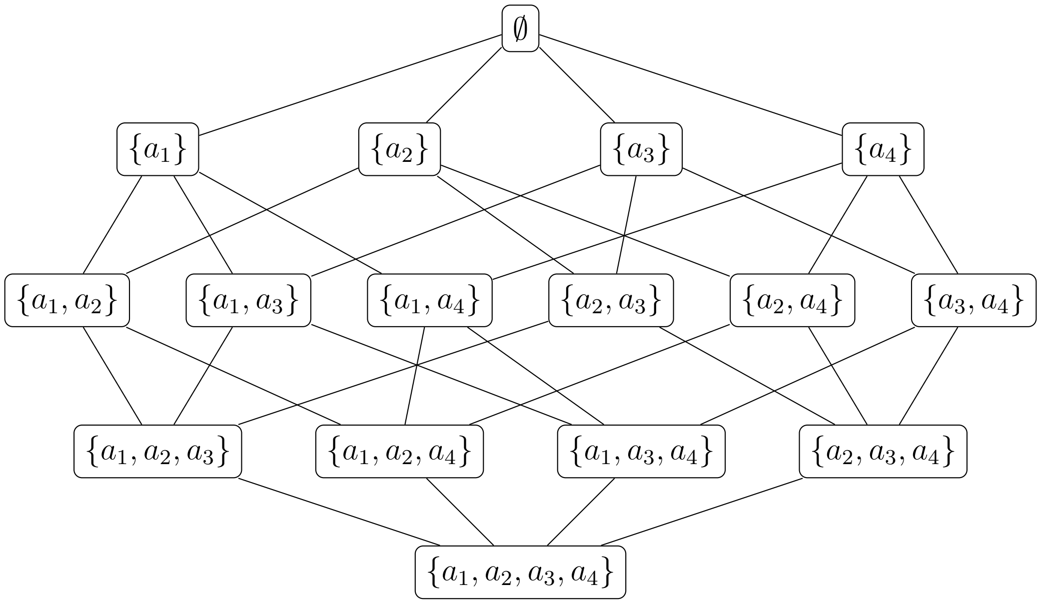

Figure 2.1: The itemset lattice for four items.

The second aspect is the monotonicity properties of the support and confidence measures which can be expressed as follows.

Proposition 2.2.1 (Support anti-monotone property). *Let X_1, X_2 be two itemsets such that $X_1 \subset X_2$, then* $\text{supp}(X_1) \geq \text{supp}(X_2)$.

This anti-monotone property leads directly to the following downward closure property also known as the Apriori principle.

Corollary 2.2.0.1 (Apriori principle). *The subsets of a frequent itemset are also frequent.*

And it also leads easily to the confidence monotone property.

Proposition 2.2.2 (Confidence monotone property). *Let $X_1 \to Y_1$ and $X_2 \to Y_2$ be two association rules such that $X_1 \subset X_2$ and $X_1 \cup Y_1 = X_2 \cup Y_2$, then* $\text{conf}(X_1 \to Y_1) \leq \text{conf}(X_2 \to Y_2)$.

These properties are essential in the process defined by the algorithm. Indeed, for discovering frequent itemsets, the Apriori algorithm follows a bottom-up, level wise approach. This means it scans the itemset lattice to find frequent itemsets starting from the bottom and moving upwards layer by layer. Of course, a complete scan is practically infeasible because the number of non-empty itemsets is equal to $2^m - 1$ where m is the number of items. To tackle this issue, the algorithm uses the Apriori principle to prune off entire branches of the lattice which it knows does not contain frequent itemsets: it will only scan the supersets of itemsets which are frequent by generating candidates (i.e. potential frequent itemsets) in a given layer from the frequent itemsets already discovered in the previous layer. The algorithm also makes use of the natural lexicographic ordering between itemsets within the lattice to avoid multiple scans of a single itemset. Similarly, a brute force approach towards association rule mining is technically infeasible because the number of potential association rules is equal to $3^m - 2^{m+1} + 1$. The Apriori algorithm uses then a top-down,

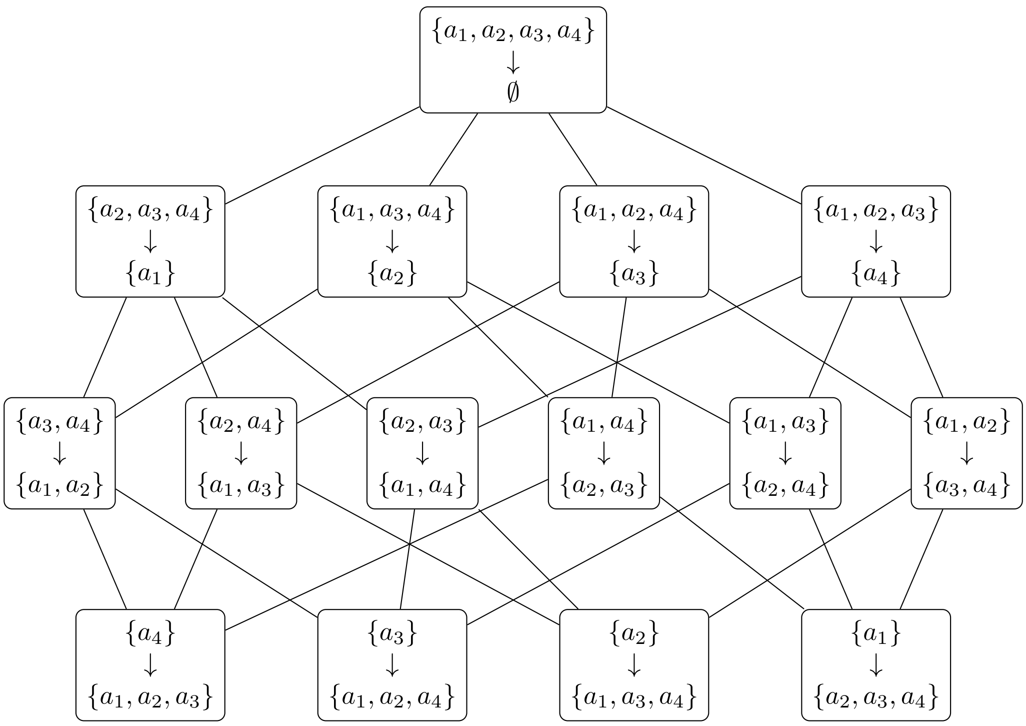

Figure 2.2: The association rule lattice (one lattice for each frequent itemset of size 2 or more).

level wise approach on each lattice of potential association rules corresponding to a frequent itemset of at least two items (as represented in Table 2.2) while using the confidence monotone property for pruning.[1]

2.2.2.2 Other mining methods

As interest grew for frequent pattern mining and the research that had been developed at the IBM Almaden Research Center became widespread in the data mining community, many researchers took upon them to solve the frequent itemset mining problem as efficiently as possible. The Apriori algorithm was indeed relatively easy to apprehend and implement but it was still generally quite slow and would not function for lower values of minimal support. Defining the most efficient algorithm became a computational challenge for researchers and the main focus of early research in itemset mining [ABH14, Goe03, HCXY07], reaching a peak with the workshops on Frequent Itemset Mining Implementations in the early 2000s [FIM03, FIM04].

The majority of the different algorithms which were proposed can be characterized by their differences to Apriori with respect to three different aspects. Firstly, many algorithms focused primarily on improving the cost of the support-counting process while retaining an Apriori-like structure. The Direct Hashing and Pruning algorithm accomplished this by trimming items from transactions [PCY95]; other algorithms, such as Apriori_LB [BJ98], used mathematical properties of the support function to avoid this process altogether; vertical algorithms, such as Monet [HKMT95], Partition [SON95], Eclat [ZPOL97, Zak00b] or VIPER [SHS+00], used the transposed of the sparse matrix representation of the transaction database to perform efficient support-counting; and projection-based algorithms, such as TreeProjection [AAP00, AAP01] or FP-Growth [HP00], used local projected databases when considering an itemset. Secondly, some algorithms explored other approaches for scanning the itemset lattice than the breadth-first (i.e. level wise) approach of Apriori. This included depth-first algorithms such as the TreeProjection algorithm in [AAP00], FP-Growth or dEclat [ZG03], or combined breadth-first and depth-first approaches such as the TreeProjection algorithm in [AAP01]. Lastly, some algorithms, such as FP-Growth, introduced specific data structures for a compressed representation of the databases. Pattern growth algorithms, following the example set by FP-Growth, are now considered to be the

[1] Note that this last process is not described in the most commonly cited version of the paper which is known for introducing the Apriori algorithm [AS94a] but in its expanded version [AS94b], which was published concomitantly.

state-of-the-art of frequent pattern mining algorithms [Agg14, HP14].

2.3 Interestingness

Having reached its peak a decade or so following their introduction, the enthusiasm for the original frequent pattern mining and association rule problems started to dwindle. The state-of-the-art algorithms were generally capable of mining large databases for frequent itemsets and association rules but the usefulness of the results they produced was put into question. Indeed, in many applications, the patterns extracted for high values of support and confidence were quite obvious. Conversely, when lowering the values for support and confidence, the number of patterns extracted grew exceedingly large. The end user was stuck between gathering a small amount of information which brought little insight about the data on the one hand, and dealing with an information overload on the other hand. Neither one of these represented interesting pieces of information to the user. The promise of the discovery of a small number of interpretable yet unexpected patterns, supported by the beers and diapers mythology, failed to deliver. We will refer to this general issue as the interestingness issue. From a chronological perspective, it is important to note that the interestingness issue did not appear in research specifically during the first decade of the twenty-first century: it had been addressed explicitly right after the development of the itemset and association rule models and even before that. However, this period does correspond to a much larger development of this topic in research.

The notion of interest, as it is commonly understood, is of course essentially subjective (though some of the research focuses on objective qualifications of interestingness while others deal with specifically subjective aspects). Therefore a large variety of research focusing on finding interesting patterns has been conducted within the field of frequent pattern mining (or more generally frequency-based pattern mining) some of them holding radically different views of what it means for patterns to be interesting, leading to entirely diverging branches in the field. For example, research that focuses on rare patterns suggest that such outliers are the most interesting patterns in many contexts, often using the extraction of gold nuggets from a mine as an analogy. This view point is entirely antithetical to the frequent pattern mining approach which suggests that the most frequent patterns are the most interesting ones. This led to the development of an entire branch of frequent pattern mining research which actually focuses on infrequent or negative frequent patterns [SON98, KR05, AWF07, SNV07, TSB09, SBK12, MSG14, ALZ14, Agg17]. This shows to tell that the notion of interestingness may have a great number of possible definitions and, though we will try to present the most significant approaches towards interestingness, we cannot address them all.

2.3.1 Interestingness measures

One of the first approaches towards the issue of interestingness was to consider alternative or additional interestingness measures to determine the patterns that were truly interesting. This approach, which focused much more on rules than itemsets, was both a prolongation of the association rule model and a renewal with previous methods.

2.3.1.1 A prolongation of the association rule model

Indeed, the association rule model identifies interestingness with high support and high confidence: two measures of interestingness. Hence, for someone who accepts the general idea behind the association rule model, the strategy for solving the interestingness issue is quite simple.

This depends on how the issue itself is perceived. Either the association rules mined are deemed uninteresting. In this case, one can believe it is simply because the measures of interestingness were not the right ones and they should be replaced by other measures. Or the rules are indeed considered interesting but not interesting enough, leading to an excessive number of patterns. In this case, using additional measures for interestingness might lead to the extraction of a fewer number of genuinely interesting rules. The Apriori algorithm already consisted of two successive steps each corresponding to a different interestingness measure: first, find the rules with high support, and second, find the rules with high confidence amongst the rules with high support. Adding a third step, based on a third measure, may seem like a natural progression, and such a process could be iterated as many times as needed until gaining satisfaction.

In both cases, the solutions are compatible with the general principles which justify the association rule model: interesting rules are those whose measure of interestingness is above a given threshold. As such these approaches are prolongations of the association rule model.

2.3.1.2 Finding the right interesting measure

In order to adopt these strategies, one must have alternative interestingness measures at ones disposal. A number of researchers therefore undertook the task of proposing a large number of interestingness measures, as well as different principles and strategies to choose an interesting measure wisely.

Subjective and objective interestingness measures. To accomplish such a task, one must first define what it means to be an interestingness measure. One of the main debates relative to this issue, which resulted in the development of two very separate branches of research, focused on considering either subjective interestingness measures or objective interestingness measures. The advocates for subjective interestingness measures suggested that the main reason explaining that patterns extracted through standard mining procedures were not considered interesting was because they conformed to the preconceptions on the data held by the users. For the proponents of subjective interestingness, if the beer and diapers association was a good example of an interesting pattern, it was because it went against our common subjective belief that beer and diapers should not be purchased simultaneously. Methods relying on interestingness measures which integrated the user's subjective belief system were proposed to extract rules which did not conform to those beliefs. In such methods, the user's belief system could be directly specified beforehand by the user [LHC97, LHML99, ST95, ST96] or learned from the user's feedback on the interestingness of proposed patterns [Sah99, AT01]. However, for the proponents of objective interestingness measures, one of the main aims of data mining is to relieve the user from having to analyze the data. For them, even though a mining process may be defined by user specifications, the process itself should not rely on the user but simply on the data. As the research in this thesis focuses on objective interestingness (although, as made explicit in the following chapter, we adopt a much stricter view on the definition of objectivity), we will focus slightly more on such objective measures.

Objective interestingness measures: a definition. The general consensus regarding the definition of an objective interestingness measure for a rule of type $X \to Y$ (up to a few exceptions as will be explicited further in this section) is that it is a function of four parameters n, n_X, n_Y and $n_{X \cup Y}$, corresponding respectively to the size of the dataset and the observed absolute frequencies in the dataset of X, Y and $X \cup Y$[2], which models the interestingness of the rule by returning a real value [OKO+04, TKS04, McG05, GH06, Vai06, LMVL08, BGK09, LB11, Hah15]. In other words, interestingness is quantified by a function with real values defined on the set of all possible contingency tables for two nominal variables X and Y, as there is a one-to-one correspondence between such contingency tables and the four parameters previously stated. In most cases, this definition can be restricted to three parameters corresponding to the relative frequencies f_X, f_Y and $f_{X \cup Y}$ of observing respectively X, Y and $X \cup Y$ in the data.

	X	$\neg X$
Y	$n_{X \cup Y}$	$n_Y - n_{X \cup Y}$
$\neg Y$	$n_X - n_{X \cup Y}$	$n + n_{X \cup Y}$ $-n_X - n_Y$

	X	$\neg X$
Y	$f_{X \cup Y}$	$f_Y - f_{X \cup Y}$
$\neg Y$	$f_X - f_{X \cup Y}$	$1 + f_{X \cup Y} -$ $f_X - f_Y$

Figure 2.3: Contingency tables for absolute and relative frequencies.

Note that, in the vast majority of research papers concerning objective interesting measures there is no distinction between observed frequencies (f_X, f_Y and $f_{X \cup Y}$) and probabilities (p_X, p_Y and $p_{X \cup Y}$) and the latter notation is often preferred together with the term probability (with a few rare exceptions such as [GSS12] which makes this quite explicit). This leads to modeling issues as we will discuss in the following chapter.

Regarding the specifics of the function or even which values should model low or high interestingness (such as 0, 1 or $+\infty$) there is no consensus. Therefore a great number of functions can be considered. In the search for the right objective interestingness measure, more than sixty different measures

[2]We use the itemset notation $X \cup Y$ here for continuity with the previous sections rather than the notation for conjunction of attributes $X \wedge Y$ which we prefer for the rest of the thesis. Both are equivalent in this context and both are used in the literature though the even more ambiguous XY notation is the most common.

were suggested in the literature ([OKO$^+$04] considers 39 measures, [TKS04] 21 measures, [GH06] 38 measures, [Vai06] 20 measures, [LMVL08] 20 measures, [BGK09] 29 measures, [LB11] 42 measures, [Hah15] 45 measures).

Objective interestingness measures for rules between itemsets? It must be remarked that most of these measures were not novel at all. In his PhD thesis [LB11], Yannick Le Bras presents a table of 42 measures in which he indicates the original scientific reference that he had found for each of these measures for rule interestingness. Out of the 42 measures which he presented, he managed to find the original references to 35 of them, only 9 of which were posterior to the 1993 paper on association rules. In this sense, the search for interestingness measures represented a renewal with previous research, because researchers were borrowing ideas from preexisting scientific work to tackle the interestingness issue. This is worth noting because it might explain why, even though all of the research papers which are referenced in this section quote association rule mining as a defining paradigm, the underlying itemset structure of the rules considered in association rule mining is not taken into account by the objective interestingness measures presented. Indeed, as these measures fall mostly into the category defined above (i.e. functions on the values of the contingency table for two nominal variables X and Y), there is no place for integrating aspects relative to the itemset structure of X or Y for the measurement of interestingness. As such, X and Y are not treated as patterns themselves but simply as items.

To be entirely precise, out of all the objective interestingness measures proposed in the scientific literature which we have scrutinized, only five of these utilized the itemset structure within the patterns to measure interestingness. Out of these five measures, one of them, the cross-support ratio, was initially not intended to be an interestingness measure but rather described the upper bound of another interestingness measure in [XTK03] (this "interestingness measure" was likely classified as such by mistake in [Hah15]). Three other measures only addressed the interestingness of itemsets and not rules: the all-confidence ([Omi03] or h-confidence in [XTK03]); the collective strength [AY98]; and lift as defined in [VT14]. Therefore, the only remaining measure on that list: improvement [BJ98], was the one and only objective interestingness measures for rules proposed, out of more than fifty, that utilized the itemset structures of the antecedents and consequents of the rules.

Properties of objective interestingness measures. In order to suggest the most adequate interestingness measures, researchers performed detailed analyses of the properties and performances of objective interestingness measures. Many different types of properties were considered, corresponding mostly to different views of what an objective interestingness measure should be.

Algorithmic properties. The study of algorithmic properties of objective interestingness measures corresponds to a very pragmatic view towards the definition of interestingness measures: if interestingness measures have nice algorithmic properties, it is much easier to define algorithms for discovering interesting patterns. This is actually an important factor in the definition of support and confidence as interestingness measures in the association rule model. Indeed, their monotonicity properties, described previously in section 2.2.2.1, are essential to association rule mining algorithms. Such algorithmic properties are related to the way interestingness measures behave with regards to the underlying mathematical structures of the search spaces (generally, one of the lattice structures described in section 2.2.2.1). Some research, such as [WHC01, LBLL09, LBLML09, LBLL10, LB11, LBLL12b] focused primarily on the study and generalization of such algorithmic properties in the context of itemset and association rule mining. An entire branch of frequent pattern mining which developed particularly well, constraint-based (or query-based) frequent pattern mining, also focuses nearly exclusively on such properties [NLHP98, LNHP99, BJ98, PHL01, BGMP03, ZYHP07, NZ14]. In fact, in constraint-based frequent pattern mining the issue of defining interestingness itself is left to the user. The database is queried for frequent patterns which satisfy a user-defined constraint formulated using a typical data query language syntax. As this branch of pattern mining has strong ties to the data management communities, it shares its traditions which explains why the focus is set on the algorithmic properties of interestingness measures. Indeed, database management algorithms rely strongly on the algorithmic properties of the functions defined within the query languages to provide fast responses.

Good modeling properties. In a defining article [PS91] for interestingness measures, Gregory Piatetsky-Shapiro suggested "*several intuitive principles that all rule-interest functions should satisfy*". These principles corresponded to three mathematical properties that an objective interestingness measure μ of parameters n, n_X, n_Y and $n_{X\wedge Y}$[3] should satisfy to be a good modeling of interestingness, namely that:

(P$_1$) $\mu = 0$ if $n_{X\wedge Y} = \frac{n_X n_Y}{n}$ (i.e. if X and Y are statistically independent, the rule is not interesting).

(P$_2$) μ increases with $n_{X\wedge Y}$ when all other parameters remain unchanged.

(P$_3$) μ decreases with n_X (or n_Y) when all other parameters remain unchanged.

Relying on the pragmatic algorithmic argument that the simplest function satisfying these principles would be more easily computed, this led him to suggest the following objective interestingness measure:

$$\mu_{\mathcal{PS}}(n, n_X, n_Y, n_{X\wedge Y}) = n_{X\wedge Y} - \frac{n_X n_Y}{n}$$

Following Piatetsky-Shapiro's lead, researchers suggested a growing number of properties that an objective interestingness should satisfy in order to be a good model for interestingness. A fourth principle, similar in its mathematical formulation, was added to the list in [MM95]. [TKS04] added five more properties to the list, which were presented as mathematical properties of the matrix operators that are interestingness measures, such as symmetry under variable permutation, row and column scaling invariance or antisymmetry under row or column permutation. These five properties were not, however, presented as necessary for a good mathematical model of interestingness but rather as potentially relevant depending on the specific context and the specific view towards interestingness which was adopted by the user. Furthermore, [TKS04] suggested that certain mathematical properties which relied on arbitrary choices (such as the choice of 0 in (P$_1$)) could be made less strict by accepting measures which would satisfy the property if conveniently normalized. The measures were classified with respect to the properties presented. Furthermore, this paper analyzed how these measures ranked a number of contingency tables, comparing this to rankings suggested by experts.

[3]We use the logical conjunction notation here, which corresponds to the original notation in [PS91], rather than the equivalent standard notation for itemsets, which is $n_{X\cup Y}$.

Expanding on previous work in French [Lal02, LMP+03, LMV+04, LT04, GCB+04] in which interestingness measures had been described using at least five additional properties, [LMVL08] selected eight properties of objective interestingness measures to be used in a multicriteria decision aid (MCDA) process to help a user choose the most adapted interestingness measures. The MCDA process described required the intervention of two different experts, an expert analyst (expert in MCDA and KDD) and an expert user (expert in the data). Out of the eight properties presented in the paper, five were attached to the expert analyst with the three remaining attached to the expert user. Though the majority of these criteria, six to be precise, were clearly mathematically defined, the definitions of the two remaining were more subjective. These were defined as the "*easiness to fix a threshold*" and the "*intelligibility*" of the interestingness measures and were both evaluated using a nominal score: easy or hard for the first criteria; a,b or c for the second criteria.

A total of fifteen properties of objective interestingness measures were described in a thorough survey on both subjective and objective interestingness by L. Geng and Howard J. Hamilton [GH06], two of which were presented as novel with the thirteen remaining associated to the papers referenced above. Later, the study of the robustness (i.e. the ability to tolerate noise in data) of the various interestingness measures was accomplished in [LBMLL10a, LBMLL10b, LB11], adding an extra criteria for the choice of the perfect interestingness measure.

Meaningfulness. We note that the meaningfulness of interestingness measures, though not explicitly defined as a property, was nevertheless identified as a possible criteria for choosing an objective interestingness measure. In [OKO+04], the authors identified five general factors for categorizing interestingness measures: Subject (Who evaluates?); Object (What is evaluated?); Unit (By how many objects?); Criterion (Based on what criterion?); and Theory (Based on what theory?). The last theoretical factor was explicitly linked to the meaningfulness of the measures and a list of 39 measures were classified into five different categories depending on what they represented: N (Number of instances included in the antecedent and/or consequent of a rule); P (Probability of the antecedent and/or consequent of a rule); S (Statistical variable based on P); I (Information of the antecedent and/or consequent of a rule); and D (Distance of a rule from the others based on rule attributes). However, this classification was quite loose. Indeed, many of the measures that were classified in the categories N, P or I did not actually represent a specific number of instances, probability or quantity of information but rather used such quantities in their definition. Furthermore, this classification was not put to use for the evaluation of interestingness measures proposed which relied rather on the subjective evaluation by an expert of both interestingness and understandability of a number of given rules.

The two criteria in [LMVL08], relative to the easiness to fix a threshold for interestingness and the intelligibility of the interestingness measures, can also be seen as related to the meaningfulness of the interestingness measure. However, both properties related more to the notion of interpretability than the notion of meaningfulness. As we will discuss in the following chapter, the first pertains to the concept of *giving meaning* whereas the second pertains to the idea of *finding meaning*. Another paper [BGK09], suggested a classification of interestingness measures based on three aspects: subject, scope and nature. These aspects, defined respectively as "*the notion measured by the index*", "*the entity concerned by the result of the measure*" and "*the descriptive or statistical feature of the index*", were explicitly presented as a means to "*grasp the meaning of rule interestingness measures*". However, the classification was based on a list of nine mathematical properties which were reformulated versions of some of the mathematical properties previously presented in the literature. The perspective they offered towards understanding the meaning of the interest measures was indeed closer to a "*grasp*" than to a full understanding of the meaning of the measures.

Measuring the interestingness of interestingness measures. Between the work presented by Gregory Piatetsky-Shapiro in 1991 [PS91] and the research conducted in the following two decades presented in the previous paragraphs, there is a quite notable evolution regarding the approach towards choosing an interestingness measure. On the one hand, Piatetsky-Shapiro defines a number of principles for interestingness measures and defines a single measure from these principles. On the other hand, a number of measures are analyzed, using semi-automated methods, to determine the most interesting interestingness measure for a specific user based on a number of different criteria. Interestingly, the robustness index [LBMLL10a, LBMLL10b, LB11], for example, can naturally be considered as a measure of the interestingness of interestingness measures! In a sense, data analysis is used to define the notion of interestingness. This meta-analysis of interestingness encourages to consider as many interestingness measures as possible. However, determining the best model for measuring rule interestingness within a long list of interestingness measures is not necessarily the most appropriate way to go. Setting aside the fact that a number of the objective interestingness measures considered in the literature have double or more counterparts after normalization [GSS12], such an approach does lead to theoretical modeling issues which will be explored in the following chapter.

2.3.2 Exact summarizations of itemsets

An entirely different approach towards the interestingness issue focuses on exact summarizations (also often referred to as lossless condensed representations) of the mined patterns. Though first introduced in 1996 [MT96], research on exact summarizations emerged mostly after the development of frequent itemset mining algorithms. This approach views the interestingness issue as an issue of information overload which is due to the high redundancy of the patterns extracted rather than them being uninteresting. That is the patterns extracted might be each individually interesting but many carry the same information and they are therefore collectively redundant. To deal with this issue, a straightforward approach is to consider a subset of non-redundant patterns which describes perfectfully the set of all individually interesting patterns. Such a subset of non-redundant patterns is called an exact summarization.

By contrast with the research on interestingness measures, the research conducted on exact summarization focused primarily on itemsets rather than rules. This is because such approaches rely strongly on the underlying mathematical structure of the search space which is particularly easy to formalize with lattices when dealing with itemsets. As a matter of fact, some of the approaches towards exact summarization, including minimal generators [BTP$^+$00] or closed itemsets [PBTL99], find their roots in formal concept analysis, a mathematical theoretical framework in lattice theory [Wil82, GW12, GSW05, GORS16].

2.3.2.1 Maximal frequent itemsets

Introduced in [AMS+96], the idea behind maximal frequent itemsets is no more than the Apriori principle (see corollary 2.2.0.1). If the subsets of a frequent itemset are also frequent and one knows that a given itemset is frequent then the information that a subset of this itemset is frequent represents redundant information with regards to the initial knowledge. Hence, the set of all frequent itemsets is entirely defined by the set of all maximal frequent itemsets, were a maximal frequent itemset is a frequent itemset that has no frequent supersets. Within the itemset lattice, the set of maximal frequent itemsets can be seen as a border which splits the lattice into two areas: frequent itemsets on one side and infrequent itemsets on the other. In a sense, the focus switches from a 2-dimensional area to a 1-dimensional border. Therefore, when using the set of maximal frequent itemsets as a condensed representation for the set of all frequent itemsets, the reduction in terms of number of itemsets can be theoretically significant if the dataset contains long frequent itemsets.

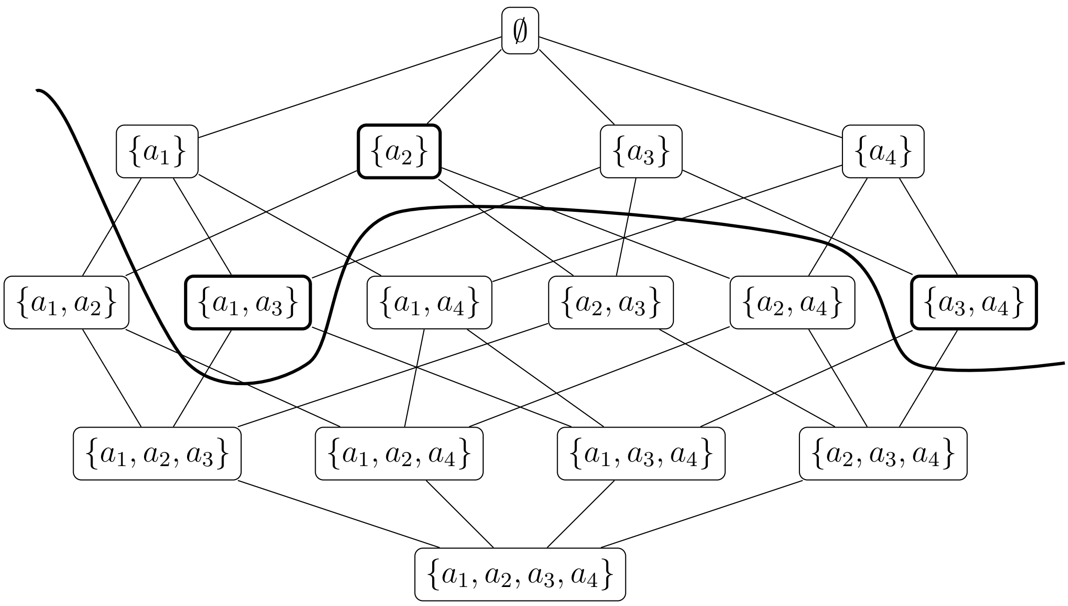

Figure 2.4: The border of maximal itemsets within the lattice of itemsets for the toy database in Table 2.1.

The theoretical gain was confirmed in practice with various datasets and researchers suggested different algorithms for specifically mining maximal frequent itemsets [AMS+96, LK98, AAP00, BCG01, GZ05]. However, although this approach leads to a lossless representation of the set of frequent itemsets, it represents a lossy representation of the set of frequent itemsets together with their frequencies. This represents an issue when mining for association rules as the frequencies of the subsets of a given frequent itemset are needed to determine the confidence of the rules whose scope is defined by that given itemset. This also represents a modeling issue which is whether itemsets themselves can be interesting or if it is the itemsets together with their frequencies which can be interesting.

2.3.2.2 Closed itemsets and minimal generators

Closed itemsets. Introduced in [PBTL99], closed itemsets were a response to the issues described above as they provide a lossless representation of the set of frequent itemsets together with their frequencies. The notion of closed itemsets relies on the scope of an itemset, which is the set of transactions (or tidset) that contains a given itemset. An itemset is said to be closed, if all of the scopes of its supersets are strictly contained in its scope.

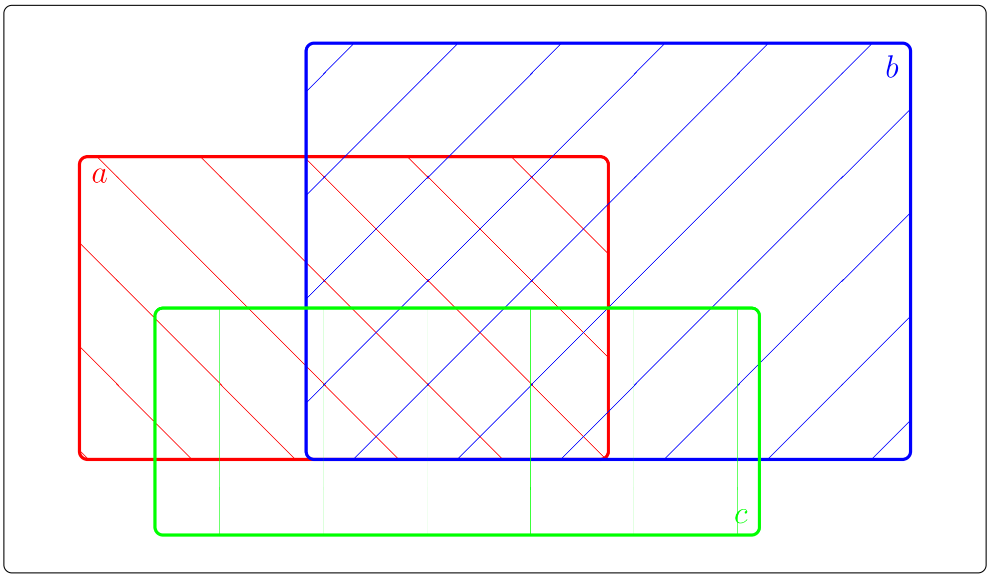

Figure 2.5: In this representation of tidsets using a Venn diagram all eight subitemsets of $\{a, b, c\}$ are closed.

As we are considering a discrete quantity of transactions, the notion of closed itemsets can also be formalized using the support measure: $X \subset \mathcal{I}$ is closed if and only if, $\forall X \subset Y \subset \mathcal{I}$, $\text{supp}(X) \neq \text{supp}(Y)$. Given the frequency of all closed itemsets, one can easily determine the frequency of any itemset: it is the frequency of its smallest closed superset (also called the itemset's

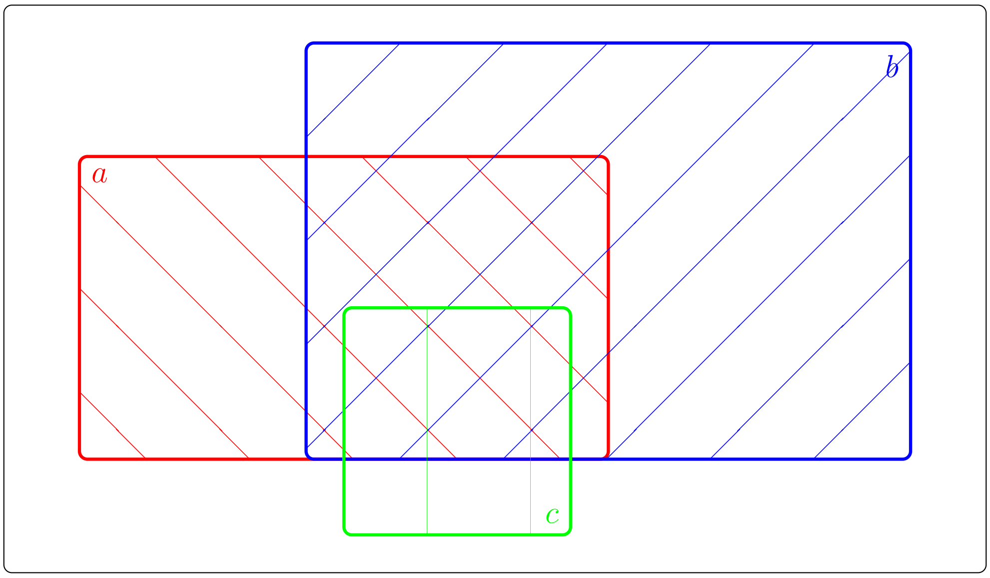

Figure 2.6: In this case, itemsets $\{a, c\}$ and $\{b, c\}$ are not closed as their scope is equal to the scope of $\{a, b, c\}$.

closure). This is also true when considering only frequent itemsets and closed frequent itemsets. Therefore, the set of closed frequent itemsets together with their frequencies allows for a lossless compression of the set of frequent itemsets together with their frequencies.

As a maximal frequent itemset is necessarily a closed itemset, the set of closed frequent itemsets contains the set of maximal itemsets. Therefore, the set of closed frequent itemsets is at least as big as the set of maximal frequent itemsets (and generally significantly larger). This is quite expected as the summarization of frequent sets using closed itemsets contains more information than the one using maximal itemsets. Although the reduction in terms of itemsets provided by closed itemsets is not as important as with maximal frequent itemsets, it can be significant nevertheless and numerous algorithms for specifically mining closed frequent itemsets have been suggested in the literature [PBTL99, ZH$^+$99, Zak00a, PHM$^+$00, WHP03, ZH05].

It is worth noting that the redundancy reduction obtained through closed itemsets for frequent itemset mining also transposes to association rule mining. Indeed, the confidence of an association rule is equal to the confidence of the rule between the closures of its antecedent and consequent [ZH$^+$99]. Therefore the set of association rules between closed itemsets, together with their frequencies and confidences, is an exact summarization of the set of association rules, together with their frequencies and confidences.

Minimal generators. Minimal generators [SNK06, VL09, GORS16] (also referred to in previous research as key patterns [BTP$^+$00] or free itemsets [BBR03]) are closely related to closed itemsets. An itemset is a minimal generator if all of the scopes of its subsets strictly contain its scope. While there is only one closed itemset for a given scope, there can be several minimal generators that share the same scope. Therefore, the number of minimal generators is at least as big as the number of closed itemsets and they do not provide a more condensed representation than closed itemsets for the set of frequent itemsets. However, they have been shown to be quite useful for defining exact summarizations of association rules [SNK06, VL09]. Furthermore, more elaborate representations based on minimal generators have been shown to provide condensed representations (such as the generalized disjunction-free generators representation [KG02a]) which can be more concise than the condensed representation given by frequent closed itemsets in some cases [KG02b].

2.3.2.3 Non-derivable itemsets

The notion of non-derivable itemsets was introduced by Toon Calders and Bart Goethals in [CG02]. This approach towards exact summarization relies principally on a generalization of the itemset model together with the centuries-old exclusion-inclusion principle from combinatorial mathematics. The generalized itemset model defined in the context of non-derivable itemsets considers both items x and their negations $\overline{x}$. The model defines a generalized itemset X as any subset of $\mathcal{I} \cup \overline{\mathcal{I}}$ (where $\mathcal{I} = \{x_1, ..., x_m\}$ is the set of items and $\overline{\mathcal{I}} = \{\overline{x_1}, ..., \overline{x_m}\}$ is the set of their negations) which does not contain both an item and its negation. The set of generalized itemsets $\mathcal{GI}$ can therefore be expressed as such:

$$\mathcal{GI} = \left\{ X \subset \mathcal{I} \cup \overline{\mathcal{I}} \;\middle|\; \forall i \in [\![1, m]\!],\ \neg\left(x_i \in X \wedge \overline{x_i} \in X\right) \right\}$$

A transaction in the database is said to contain a generalized itemset if it contains all of the items within the generalized itemset and none of those whose negation is in the generalized itemset.

For k given items corresponding to an itemset I (a regular itemset), there are 2^k generalized itemsets of size k. As the support of each of these generalized

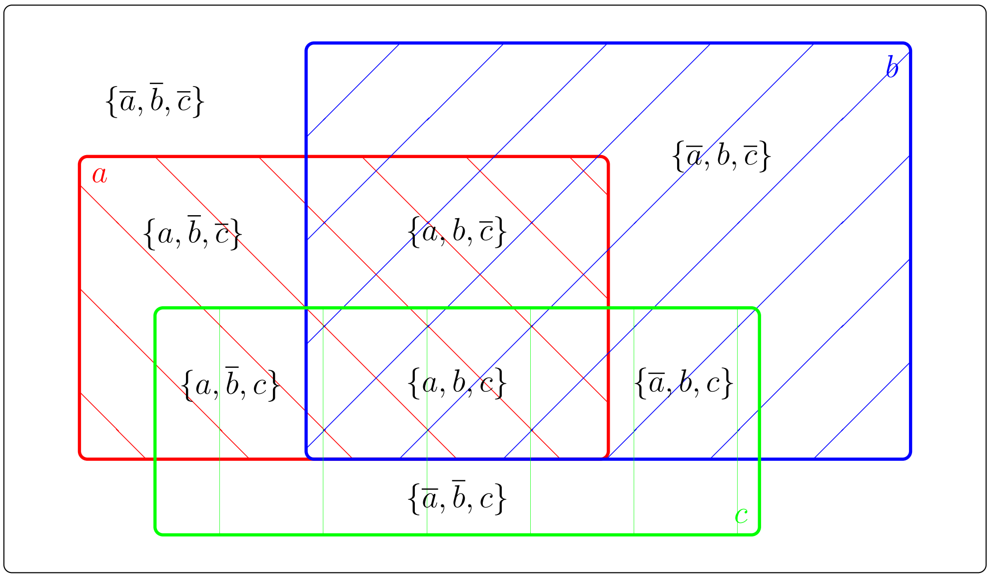

Figure 2.7: The eight generalized itemsets of size three based on itemset $\{a, b, c\}$ correspond to the eight disjoint areas in the Venn diagram.

itemsets cannot be less than zero, this leads to 2^k inequalities. For each subset $X \subset I \subset \mathcal{I}$, define $\delta_X(I)$ as:

$$\delta_X(I) = \sum_{X \subseteq J \subsetneq I} (-1)^{|I \setminus J|+1} \text{supp}(J)$$

Then, using the inclusion-exclusion principle, one can show that the inequalities described above are equivalent to the set of inequalities defined, for all $X \subset I \subset \mathcal{I}$, by:

$$\text{supp}\,(I) \leq \delta_X(I) \quad \text{if } |I \setminus X| \text{ odd}$$

and:

$$\text{supp}\,(I) \geq \delta_X(I) \quad \text{if } |I \setminus X| \text{ even}$$

Hence, $\text{supp}\,(I) \in [\text{LB}(I), \text{UB}(I)]$ where:

$$\text{LB}(I) = \max_{\substack{X \subseteq J \subset \mathcal{I} \\ |I \setminus X| \text{ even}}} (\delta_X\,(I))$$

and:

$$\text{UB}(I) = \min_{\substack{X \subseteq J \subset \mathcal{I} \\ |I \setminus X| \text{ odd}}} (\delta_X\,(I))$$

This defines an interval for the support of I based on the knowledge of the supports for all of its proper subsets. Furthermore, this is the best possible interval given this knowledge as one can construct possible datasets to fit each of the values within this interval [Cal04].

An itemset (i.e. a regular itemset) whose support can be inferred from the knowledge of the support on its subsets is called a derivable itemset. From what precedes, the set of derivable itemsets is exactly the set of itemsets I such that $\mathrm{LB}(I) = \mathrm{UB}(I)$. Itemsets for which $\mathrm{LB}(I) \neq \mathrm{UB}(I)$ are defined as non-derivable itemsets.

Derivable and non-derivable itemsets have interesting mathematical properties which can be used effectively in data mining processes. This includes a monotonicity property.

Proposition 2.3.1. *The supersets of a derivable itemset are also derivable.*

Which can be slightly generalized as follows.

Proposition 2.3.2. *If I is an itemset such that* $\mathrm{supp}\,(I) \in \{\mathrm{LB}(I), \mathrm{UB}(I)\}$ *then all supersets of I are derivable.*

Another important property gives a bound on the size of non-derivable itemsets, depending on the size n of the database (i.e. the number of transactions).

Proposition 2.3.3. *Let I be an itemset. If* $|I| > \log_2(n)+1$ *then I is derivable.*

This last property shows that, for a fixed n, the number of non-derivable itemsets is at most polynomial in the number of items m. Furthermore, the set of frequent non-derivable items together with their frequencies gives a lossless condensed representation of the set of frequent itemsets together with their frequencies. This representation has been shown to be theoretically more concise than other representations such as those based on minimal generators [Cal04, CG07]. With regards to the frequent closed itemsets representation, it is neither more nor less concise as this can vary depending on the dataset, and empirical studies have shown that they are relatively comparable [CG07, VT14].

2.3.3 Local and global models for mining informative and significant patterns

One of the key aspects of the interestingness issue in frequent pattern mining is redundancy: the idea that some patterns can carry redundant information, whether individually or collectively. As made explicit in the previous section, this aspect is central in pattern mining approaches which use exact summarizations. However, in the context of exact summarizations, redundancy of information is only addressed exactly. That is, a given piece of information is considered redundant with other elements of information only if it can be inferred entirely and certainly from these elements. This vision is quite restrictive because sometimes such elements of information can tell us mostly (albeit not exactly) what there is to know about this piece of information.

In order to consider how a piece of information can be mostly inferred from other elements of information, various frameworks have been envisaged rooted in either statistics or information theory. These frameworks allow for the definition of either local or global models for considering pattern interestingness and identifying redundancy either in individual patterns or collectively.

Note that the emphasis put on the difference between local and global models for classifying different pattern mining approaches corresponds to our own general understanding of the various approaches in the literature and the issues, in terms of model consistency, related to the use of local models, which we will address in detail in the following chapter. This classification is the central theme of one of our publications [DLL17].

2.3.3.1 Local data models for identifying local redundancy within individual patterns

Given a single pattern, we use the term local data model to refer to a data model based on the observation of its proper components in the data (i.e. based on partial descriptions of the pattern). If the local data model mostly predicts the pattern, then the pattern is locally redundant. As such, many objective interestingness measures rely on (implicit) local models to identify local redundancy.

The fundamental idea behind this approach is to define a redundancy score for each individual pattern, or at least to incorporate this aspect in a more general interestingness measure. A redundancy score can be used to rank patterns and define a set of non-redundant patterns to be presented to the user (either the top-k less redundant patterns or all those whose redundancy score is beyond a given threshold).

Local redundancy in rules. Most objective interestingness measures for rules do not take into account the mathematical structures of the antecedents and consequents of the rules which they consider. They rely solely on a simple rule structure between two attributes (an antecedent X and a consequent Y) whose parameters are entirely defined by the absolute frequencies in the 2×2 contingency table for these attributes (see section 2.3.1.2). Therefore, identifying redundancy within a rule is equivalent to identifying redundancy within the contingency table.

Statistical models. Throughout the past century and more, such contingency tables have been extensively studied in the field of statistics [Pea00, Fis22, Fis24, Jay03]. The likelihood of observing such a table in a given context may be modeled by using a number of various probability distributions [LGR81a, LGR81b] and different tests have been designed to measure the surprisingness of having observed one given table.

If there is no background knowledge about the data, the simplest model for the database is that it is a random sample of n independent identically distributed random variables $(\mathbf{x}, \mathbf{y})$ with values in $\{0, 1\}^2$. The probability distribution for this random variable can be entirely defined by three values $p_X = \text{Prob}(\mathbf{x} = 1)$, $p_Y = \text{Prob}(\mathbf{y} = 1)$ and $p_{X \wedge Y} = \text{Prob}(\mathbf{x} = \mathbf{y} = 1)$. Information on these values is provided by the contingency table of observed frequencies.

If there is nothing to say about the relationship between the antecedent and the consequent (i.e. there is no rule between the two), then the safest assumption to make is that $\mathbf{x}$ and $\mathbf{y}$ are independent random variables. In this case, $p_{X \wedge Y} = p_X p_Y$. The local independence model for a rule is the model defined by $p_X = f_X$, $p_Y = f_Y$ and the independence of $\mathbf{x}$ and $\mathbf{y}$.

Although generally not based on proper statistical tests of independence, many objective interestingness measures are designed to discriminate against rules whose antecedent and consequent appear to be independent and it is widely defended that this is a necessary property for defining an objective interestingness measure (see section 2.3.1.2). In table 2.4, we make explicit the manner in which some objective interestingness measures in the literature correspond to a naive comparison between a given statistic and its predicted value given the independence model.

Measure	Regular formula	Comparative formula
Added value	$\max\left(f_{X\mid Y} - f_X, f_{Y\mid X} - f_Y\right)$	$\max\left(f_{X\mid Y} - p_{X\mid Y}, f_{Y\mid X} - p_{Y\mid X}\right)$
Bayes factor	$\frac{f_{X\mid Y}}{f_{X\mid \neg Y}}$	$\frac{f_{X\mid Y}}{f_{X\mid \neg Y}} \div \frac{p_{X\mid Y}}{p_{X\mid \neg Y}}$
Centered confidence	$f_{Y\mid X} - f_Y$	$f_{Y\mid X} - p_{Y\mid X}$
Collective strength	$\frac{f_{X\wedge Y}+f_{\neg X\wedge\neg Y}}{f_X f_Y + f_{\neg X} f_{\neg Y}} \times \frac{f_{\neg X} f_Y + f_X f_{\neg Y}}{f_{\neg X\wedge Y}+f_{X\wedge\neg Y}}$	$\frac{f_{X\wedge Y}+f_{\neg X\wedge\neg Y}}{f_{\neg X\wedge Y}+f_{X\wedge\neg Y}} \div \frac{p_{X\wedge Y}+p_{\neg X\wedge\neg Y}}{p_{\neg X\wedge Y}+p_{X\wedge\neg Y}}$
Conviction	$\frac{f_X f_{\neg Y}}{f_{X\wedge\neg Y}}$	$\frac{p_{X\wedge\neg Y}}{f_{X\wedge\neg Y}}$
Interest	$\lvert f_{X\wedge Y} - f_X f_Y\rvert$	$\lvert f_{X\wedge Y} - p_{X\wedge Y}\rvert$
Lift	$\frac{f_{X\wedge Y}}{f_X f_Y}$	$\frac{f_{X\wedge Y}}{p_{X\wedge Y}}$
Loevinger	$1 - \frac{f_X f_{\neg Y}}{f_{X\wedge\neg Y}}$	$1 - \frac{p_{X\wedge\neg Y}}{f_{X\wedge\neg Y}}$
Piatetsky-Shapiro1	$f_{X\wedge Y} - f_X f_Y$	$f_{X\wedge Y} - p_{X\wedge Y}$
Piatetsky-Shapiro2	$n\left(f_{X\wedge Y} - f_X f_Y\right)$	$n\left(f_{X\wedge Y} - p_{X\wedge Y}\right)$
Relative risk	$\frac{f_{Y\mid X}}{f_{Y\mid \neg X}}$	$\frac{f_{Y\mid X}}{f_{Y\mid \neg X}} \div \frac{p_{Y\mid X}}{p_{Y\mid \neg X}}$
Relative specificity	$f_{\neg X\mid\neg Y} - f_{\neg X}$	$f_{\neg X\mid\neg Y} - p_{\neg X\mid\neg Y}$
Varying rates liaison	$1 - \frac{f_{X\wedge Y}}{f_X f_Y}$	$1 - \frac{f_{X\wedge Y}}{p_{X\wedge Y}}$

Table 2.4: Objective interestingness measures that naively compare a statistic to its predicted value for the standard independence model.

In addition to the standard independence model described above, other models have been proposed to represent the database, notably the Poisson model and the fixed column margins model. On the one hand, the Poisson model relies on the assumption that the number of transactions n is the result of the observation of a random variable, noted N, which follows a Poisson distribution of parameter n, rather than being a fixed parameter. Furthermore, conditionally to the realization of any event $[N = k]$, the dataset is seen as the random sample of k independent identically distributed random variables $(\mathbf{x}, \mathbf{y})$ following the distribution given by the local independence model. On the other hand, in the fixed column margins model, not only is n considered as being a fixed parameter, but so are n_X and n_Y. As such, the dataset is considered to be the random sampling of identically distributed dependent random variables that follow the distribution given by the local independence model (which is equivalent to considering a uniform distribution on all datasets satisfying the conditions given by the fixed parameters)[4]. In Table 2.5, we present the known probabilistic behaviors of various statistics (each of which are also seen as objective interestingness measures in the literature) with regards to the data model considered. Statistical tests based on these probabilistic behaviors (or some analog version such as considering $\neg Y$ rather than Y in statistical implicative analysis) can in turn be considered as objective interestingness measures for rules [LMVL08, LB11, GRMG13].

[4] In a sense, the Poisson model and the fixed column margins model diverge from the standard independence model in opposing directions. In all models, the empirical dataset is seen as one observation out of all the potential datasets which can be generated by the random process described by the model. The set of all potential datasets in the standard model has cardinality 4^n, it is strictly contained in the set of potential datasets in the Poisson model which has infinite cardinality, but it strictly contains the set of potential datasets in the fixed column margins model.

[5] We use the following notations for Table 2.5:

- S is the random variable associated to the statistic, $\sim$ indicates its distribution and $\rightsquigarrow$ its convergence in distribution;
- $\mathcal{B}(n, p)$ for the binomial distribution with parameters n and p;
- $\mathcal{P}(\lambda)$ for the Poisson distribution with parameter λ;
- $\mathcal{H}(n, n_1, n_2)$ for the hypergeometric distribution with parameters n, n_1 and n_2;
- $\mathcal{N}(\mu, \sigma^2)$ for the normal distribution with parameters μ and σ^2;
- $\chi^2(d)$ for the chi-square distribution with d degrees of freedom;
- $\mathcal{T}(d)$ for Student's t-distribution with d degrees of freedom.

Measure	Regular formula	Probabilistic behavior[5]	Model
Support	$n_{X\wedge Y}$	$S \sim \mathcal{B}(n, f_X f_Y)$	Standard
		$S \sim \mathcal{P}(n f_X f_Y)$	Poisson
		$S \sim \mathcal{H}(n, n_X, n_Y)$	Fixed column margins
Correlation coefficient	$\frac{f_{X\wedge Y} - f_X f_Y}{\sqrt{f_X f_Y f_{\neg X} f_{\neg Y}}}$	$S\sqrt{\frac{n-2}{1-S^2}} \rightsquigarrow \mathcal{T}(n-2)$	Standard
Normalized support	$\sqrt{n}\frac{f_{X\wedge Y} - f_X f_Y}{\sqrt{f_X f_Y (1 - f_X f_Y)}}$	$S \rightsquigarrow \mathcal{N}(0,1)$	Standard
	$\sqrt{n}\frac{f_{X\wedge Y} - f_X f_Y}{\sqrt{f_X f_Y}}$		Poisson
	$\sqrt{n-1}\frac{f_{X\wedge Y} - f_X f_Y}{\sqrt{f_X f_Y f_{\neg X} f_{\neg Y}}}$		Fixed column margins
χ^2 statistic	$n\left(\frac{f^2_{X\wedge Y}}{p_{X\wedge Y}} + \frac{f^2_{\neg X\wedge Y}}{p_{\neg X\wedge Y}} + \frac{f^2_{X\wedge\neg Y}}{p_{X\wedge\neg Y}} + \frac{f^2_{\neg X\wedge\neg Y}}{p_{\neg X\wedge\neg Y}} - 1\right)$	$S \rightsquigarrow \chi^2(1)$	Standard
		$S \rightsquigarrow \chi^2(2)$	Poisson
			Fixed column margins

Table 2.5: Statistics and their behavior.

Information theory models. Another framework which allows to identify local redundancy in rules is information theory. Information theory provides a complete modeling for information [Sha48, Kul59, Mac03, CT12, Jay03, Bor11] which comprises:

- a model for the amount of information gained from the observation of a random event: the *information content* of the event;
- a model for the average amount of information required to describe a single sample of a random variable: the *information entropy* of a probability mass function ;
- a model for the amount of information lost when a probability mass function **p** is used to approximate a probability mass function **f**: the

Kullback–Leibler divergence between distributions.

For a distribution $\mathbf{p}$ in the contingency table, the information entropy $H(\mathbf{p})$ is defined as:

$$H(\mathbf{p}) = -p_{X \wedge Y} \log p_{X \wedge Y} - p_{X \wedge \overline{Y}} \log p_{X \wedge \overline{Y}} - p_{\overline{X} \wedge Y} \log p_{\overline{X} \wedge Y} - p_{\overline{X} \wedge \overline{Y}} \log p_{\overline{X} \wedge \overline{Y}}$$

High entropy indicates high randomness while low entropy indicates high structure. Given partial knowledge on the values of this distribution, then there is a unique distribution which maximizes the entropy while matching this knowledge. This maximum entropy distribution corresponds to the idea that there is no more structure to the data than what is already described through the given partial knowledge. Typically, if p_X and p_Y are given by f_X and f_Y, the maximum entropy model corresponds exactly to the previous independence model.

The Kullback-Leibler divergence $D_{\mathrm{KL}}(\mathbf{f} \parallel \mathbf{p})$ between the empirical distribution $\mathbf{f}$ in the contingency table and the distribution given by the local independence model $\mathbf{p}$ is given by:

$$D_{\mathrm{KL}}(\mathbf{f} \parallel \mathbf{p}) =$$
$$f_{X \wedge Y} \log \frac{f_{X \wedge Y}}{f_X f_Y} + f_{X \wedge \neg Y} \log \frac{f_{X \wedge \neg Y}}{f_X f_{\neg Y}} + f_{\neg X \wedge Y} \log \frac{f_{\neg X \wedge Y}}{f_{\neg X} f_Y} + f_{X \wedge \neg Y} \log \frac{f_{\neg X \wedge \neg Y}}{f_{\neg X} f_{\neg Y}}$$

If the Kullback-Leibler divergence is small, there is little information lost when representing $\mathbf{f}$ by $\mathbf{p}$. Hence, the distribution $\mathbf{f}$ may be considered redundant with regards to its parts which describe $\mathbf{p}$.

A few objective interestingness measures identify redundancy based on information theory and a local maximum entropy model (i.e. the local independence model) as we make explicit in table 2.6.

Local redundancy in rules between itemsets. As stated in section 2.3.1.2, we have found but a single example of an objective interestingness measure for rules between itemsets that actually includes both rule and itemset

[6]In addition to the notations previsouly defined, we use the following notations for Table 2.6:

- $I_{\mathbf{p}}(E)$ for the information content of the observation of E given a distribution $\mathbf{p}$;
- $D_{\mathrm{KL}}(\mathbf{p} \parallel \mathbf{q}, E)$ for the Kullback-Leibler divergence between $\mathbf{p}$ and $\mathbf{q}$ restricted to E.

Measure	Regular formula	Information theory formula[6]
Information gain	$\log \frac{f_{X\wedge Y}}{f_X f_Y}$	$I_{\mathbf{p}}(X \wedge Y) - I_{\mathbf{f}}(X \wedge Y)$
J-Measure	$f_{X\wedge Y} \log \frac{f_{X\wedge Y}}{f_X f_Y} + f_{X\wedge \neg Y} \log \frac{f_{X\wedge\neg Y}}{f_X f_{\neg Y}}$	$D_{\mathrm{KL}}(\mathbf{f} \parallel \mathbf{p}, X)$
Normalized mutual information	$-\frac{f_{X\wedge Y} \log \frac{f_{X\wedge Y}}{f_X f_Y} + f_{X\wedge\neg Y} \log \frac{f_{X\wedge\neg Y}}{f_X f_{\neg Y}} + f_{\neg X\wedge Y} \log \frac{f_{\neg X\wedge Y}}{f_{\neg X} f_Y} + f_{X\wedge\neg Y} \log \frac{f_{\neg X\wedge\neg Y}}{f_{\neg X} f_{\neg Y}}}{f_X \log f_X + f_{\neg X} \log \neg X}$	$\frac{D_{\mathrm{KL}}(\mathbf{f} \Vert \mathbf{p})}{H((f_X, f_{\neg X}))}$
One way support	$f_{Y\vert X} \log \frac{f_{Y\vert X}}{f_Y}$	$D_{\mathrm{KL}}(\mathbf{f} \parallel \mathbf{p}, Y\vert X)$
Two way support	$f_{X\wedge Y} \log \frac{f_{X\wedge Y}}{f_X f_Y}$	$D_{\mathrm{KL}}(\mathbf{f} \parallel \mathbf{p}, X \wedge Y)$
Two way support variation	$f_{X\wedge Y} \log \frac{f_{X\wedge Y}}{f_X f_Y} + f_{X\wedge\neg Y} \log \frac{f_{X\wedge\neg Y}}{f_X f_{\neg Y}} + f_{\neg X\wedge Y} \log \frac{f_{\neg X\wedge Y}}{f_{\neg X} f_Y} + f_{X\wedge\neg Y} \log \frac{f_{\neg X\wedge\neg Y}}{f_{\neg X} f_{\neg Y}}$	$D_{\mathrm{KL}}(\mathbf{f} \parallel \mathbf{p})$

Table 2.6: Objective interestingness measures based on information theory that discriminate against locally redundant rules given the independence model.

structures, the *improvement* measure [BJ98]:

$$\mathrm{imp}(X \to Y) = \min_{X' \subsetneq X} (\mathrm{conf}(X \to Y) - \mathrm{conf}(X' \to Y))$$

This measure does discriminate against locally redundant rules as it compares the confidence of the rule to the confidence of any proper subrule with the same consequent. In terms of local data models, this can be seen as the difference between the observed frequency $f_{Y|X}$ and its predicted value $p_{Y|X} = \max_{X' \subseteq X} p_{Y|X'}$. However, this measure was constructed as an intuitive indicator for local redundancy and any justification for why a local data model would predict such a value would be quite ad hoc.

Local redundancy in itemsets. In the case of itemsets, the proper components of the pattern which usually define local data models are either its items or, more generally, its subsets (both together with their frequencies).

As in the case of rules, the statistical data models described in this section comply with the following classical structure for data models. Consider an itemset $X = \{x_1, ..., x_k\}$. The database is modeled as a random sample of n independent identically distributed random variables $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_k)$ with values in $\{0,1\}^k$. The probability distribution for this random variable can be entirely described by the $2^k - 1$ values $p_{X'}$ for all $\emptyset \subsetneq X' \subset X$.

To simplify notations in this section, we will use f_i and p_i rather than $f_{\{x_i\}}$ and $p_{\{x_i\}}$ for the frequencies and probabilities associated to a single item x_i.

The independence model The independence model is the simplest local statistical model for an itemset. It is based on the frequencies of the items that compose the item and the mutual independence of their associated random variables:

$$\forall i \in [\![1,k]\!],\ p_i = f_i \text{ and } \mathbf{x}_i \text{ are all mutually independent}$$

In this model, the value of the probability of any subset $X' \subset X$ is equal to the product of the frequencies of the items that compose X':

$$p_{X'} = \prod_{x_i \in X'} f_i$$

Different tests and measures for identifying redundancy using the independence model have been suggested in the literature. These can rely on a specific statistic. The most common statistic used is the frequency of the itemset f_X which is compared to $p_X = \prod_{1 \leq i \leq k} f_i$. The lift measure for itemsets ([VT14]) compares the two through a simple ratio:

$$\text{lift}(X) = \frac{f_X}{p_X}$$

and evaluation of p-values allow for stronger statistical testing ([VT14]).

Another statistic has been compared to its expected value given the independence model in the literature through the collective strength measure ([AY98]). This measure considers the ratio between the number of agreements to X (transactions that contain X or contain none of the items of X) and the number of violations to X (the remaining transactions) and divides by its expected value given the independence model:

$$\mathrm{cs}(X) = \frac{\frac{1-f_v}{f_v}}{\frac{1-p_v}{p_v}} \text{ where } f_v = f_{\overline{X} \wedge \bigvee_{1 \le i \le k} \{x_i\}} \text{ and } p_v = 1 - \left(\prod_{1 \le i \le k} p_i + \prod_{1 \le i \le k} (1 - p_i) \right)$$

However, testing a single statistic against its value given by the independence model can lead to misleading conclusions. Indeed, there are multiple different models which may be equal to the independence model on a given number of statistics. Hence, for one single statistic which is well predicted, it is fallacious to say that the hypotheses for the independence model led to this good prediction. Therefore, some authors have privileged tests that allow complete comparisons of the local independence data model to the local empirical distribution. These include the G-test [BMS97] or Pearson's χ^2 test [VT14]. Note that, in such a case, the size of the itemset must stay reasonably small because the computation complexity of such tests grows exponentially with the size of the itemset. In any case, as we will discuss in the next chapter, the significance of such tests fall extremely low for larger itemsets unless impossibly massive amounts of data are collected.

An important remark relative to the local independence model for a given itemset is that it is compatible with the global independence model for all itemsets. That means the probabilities given by any local model correspond with those given by the global model. This is quite a unique property which allows to consider multiple local models while maintaining global consistency between models. This does not, however, prevent other issues with the use of multiple local models as we will discuss in the next chapter.

MaxEnt models. Maximum entropy models (or MaxEnt models) are models which maximize the information entropy of the probability mass function of the underlying statistical model given a set of constraints. As we will discuss further in this thesis, the rationale for using MaxEnt models can be founded on three different approaches towards the meaning of entropy: Shannon's approach [Sha48], Jayne's approach [Jay82, Jay03], and the constrained independence approach which we have developed. In any case, MaxEnt models may be seen as a generalization of independence models. They are the least binding models, in terms of model hypotheses, given the set of constraints on which they are defined. As this thesis proposes a detailed study of such models, we will limit this section to a brief description of the classical representation of MaxEnt models in itemset mining.

Let $X = \{x_1, ..., x_k\}$ be the itemset for which a local MaxEnt model is defined. Consider $\Omega = \{0, 1\}^k$ the set of all possible transactions (which can also be seen as the set of all generalized itemsets of size k). These represent the 2^k different possible values which can be taken by the random variable $\mathbf{X}$. Hence, we can define the information entropy for the associated probability mass function as:

$$H = -\sum_{\omega \in \Omega} p_\omega \ln(p_\omega)$$

where p_ω is the probability that $\mathbf{X} = \omega$.

Now consider a set $\mathcal{C}$ of non empty subsets of X (or more generally a set of generalized itemsets based on X). We can define a MaxEnt model by considering the probability function that maximizes H while satisfying the constraints that $p_{X'} = f_{X'}$ for all $X' \in \mathcal{C}$.

If $\mathcal{C} = \emptyset$, then we obtain the model for k independent random coin tosses. If $\mathcal{C} = \{\{x_1\}, ..., \{x_k\}\}$, then we obtain the independence model. If $\mathcal{C} = \mathcal{P}(X) \setminus \{\emptyset\}$ (where $\mathcal{P}(X)$ represents the partition set of X), then we obtain the local empirical data model.

In addition to these simple cases, a few other specific cases are known to be solved using specific procedures.

The oldest example is the Chow-Liu tree model which describes the MaxEnt model were the constraints are given by the set of all itemsets of size both one and two [CL68]. A generalization of this method, known as k-width junction cherry trees, has been suggested to define models based on all the itemsets of size k or less [KS10, SK12] [7]. However, to our knowledge, the models provided have only been proven to maximize entropy among a certain class of probabilities defined by k-width junction trees [SK12]. Chow-Liu tree models, as well as k-width junction cherry trees have both been used in the context of itemset mining.

Our own constrained independence model describes, in its first version, a specific fast computable method for defining the MaxEnt model when the set of constraints contains all proper subsets of the itemset [DBLL15].

Our further work on mutual constrained independence model, presented in

[7] The itemset framework is, however, not mentioned in these articles.

chapter 4 of this thesis, allowed us to present explicit algebraic formulas for the local MaxEnt models for any itemset of size 4 or less, given any set of constraints. More generally, methods for computing the local MaxEnt models for any size of itemset and any set of constraints are known. We have contributed to this topic by defining a new general method for computing MaxEnt models based on algebraic geometry. However, all these methods are technically impracticable for large itemsets [Tat06].

Local MaxEnt models have been used to identify local redundancy within individual patterns using both specific [Meo00, Tat08, PMS03, DBLL15] and general cases of MaxEnt models [Tat08, PMS03]. Redundancy scores are defined similarly as in the case of the independence model: either by comparing the observed value of a given statistic to its predicted value; or by a complete comparison of the local MaxEnt model and the local empirical model.

Note that, regardless of the comparison tool utilized, a single MaxEnt model must always be considered in the end. If this is quite straightforward when considering a specific case of MaxEnt models, it is less so when considering more general cases of MaxEnt models. In this scenario, one general MaxEnt model must still be chosen. This has been done by considering the MaxEnt model defined by the frequencies of only frequent subsets of the itemset [PMS03]; by considering the optimal tree model [Tat08]; or by considering the optimal family model [Tat08] (i.e. the local MaxEnt model which generates the highest redundancy score for the pattern).

Another important remark is that, unlike the independence models, local MaxEnt models are not, a priori, consistent among each other [DBLL15] which raises some theoretical issues addressed in the next chapter.

Other models Various other local models have been tested to determine the redundancy of an individual itemset including: mixtures of independence models [PMS03]; an inclusion-exclusion model based on AD trees [PMS03]; and partition models [Web10, Web11].

2.3.3.2 Testing redundancy against global background knowledge models

In the previous section, we have described approaches that search for redundancy within individual patterns, that is the redundancy of a pattern with regards to its components. The redundancy of an individual pattern has also been considered relative to background knowledge which can be represented by a global data model.

In the context of itemset mining, the independence model can very well be considered as a global background model. In that case, it is equivalent to the local independence model for the itemset containing all items. As all local independence models are consistent with the global independence model (that is they are local projections of the global model) in the context of itemset mining, checking for the redundancy of an itemset with regards to the background knowledge represented by the global independence model or checking for the redundancy of an itemset relative to its local independence model is, in fine, quite equivalent. This is, however, not the case when mining rules between itemsets.

Furthermore, there is a distinctive theoretical nuance between both approaches as will be discussed in the next chapter. One important aspect to keep in mind is that defining a background model is akin to classical statistical modeling, not to data mining. A data mining layer can be added on top of this data modeling layer, but the two must not be mistaken for each other.

Different statistical models for background knowledge have been suggested using global MaxEnt models. These models rely mainly on constraints defined otherwise than by the frequencies of itemsets. In [TM10], the authors suggest four possible statistics for defining a MaxEnt model: the standard column and row margins; the lazarus counts; and transaction bounds. They test global MaxEnt models, based on these constraints or combinations of these constraints, as background knowledge models. As these models are MaxEnt models for the probability mass function defined on the space of possible transactions, whether they belong to the data modeling layer or to the data mining layer described above is ambiguous (see discussion in the next chapter).

This is not the case for the MaxEnt model defined on the stricter fixed row and column margins constraint, known as the Rasch model [Ras60], which is a MaxEnt model for the probability mass function defined on the space of possible datasets. This model has also been used as a global background model for identifying redundancy in itemset mining directly [KDB10] or indirectly, through the use of randomization methods based on Monte Carlo Markov chains and swap randomization [GMMT07, HOV$^+$09].

2.3.3.3 Iterative learning

One of the major criticisms which can be addressed towards itemset ranking based on individual interestingness measures or redundancy scores is that it is a forgetful process: itemsets are ranked individually regardless of any rankings previously determined. There is no learning process.

This contrasts with the vision that these methods aim at gaining knowledge, in other words learning, about the underlying mechanisms that generate the data. Learning, be it human or machine learning, is usually described as a never-ending incremental process[8]. We present here the main approaches in the literature towards implementing the incremental aspect of learning. The endless aspect of learning implies learning knowledge from infinite or dynamic databases and involves an entirely different branch of research which is beyond the scope of this thesis.

Consider the process of determining the redundancy of a pattern with regards to background knowledge as described in the previous section. If a pattern (for example an itemset together with its frequency) is non-redundant with that background knowledge, then it can be added to that background knowledge and the process may be repeated. Hence, the background knowledge of any given step represents the entire knowledge, previously known and acquired up to that step during the learning process[9].

Different variations to this principle have been suggested in the itemset literature.

In [WP06a, WP06b], Markov random fields (MRF) are used to iteratively construct global models for the data. The algorithm proposed is a level-by-level approach. First, the frequencies of all itemsets of size 1 are used to build the MRF model for predicting the frequencies of all itemsets of size 2. These predictions are compared to the corresponding empirical frequencies. If the prediction is too far away from the empirical value for a given itemset, the empirical frequency of that itemset is added as a constraint to the MRF model. Once all itemsets of size 2 have been scanned, the additional knowledge is used to recalculate the MRF model using a junction tree algorithm or an

[8] Note that this description also applies to the scientific process, as we will discuss in the following chapter.

[9] As we will show in the next chapter, it is important in such a process to dissociate between background knowledge based on the type of patterns that are being mined and background knowledge based on other aspects in the data.

approximation through an MCMC method. The process then goes on until a given itemset size is reached. Note that the model given by the junction tree algorithm, let alone its approximation obtained via an MCMC method, is not equal to the MaxEnt model based on the same constraints. Indeed as stated previously, junction tree algorithms are only known to converge towards the MaxEnt model in a specific class of probabilities.

In an article entitled "*Tell me something I don't know*" [HOV$^+$09], the authors use randomization methods applied to an iterative learning process. Using the p-value given by the randomization model for all itemsets, the algorithm suggested finds the most surprising itemset at each step of the iteration and adds it to the randomization model. Given the exponential number of itemsets, the method proposed only considers itemsets of size 2 or 3. Furthermore, the complexity of the randomization task being to great when adding exact constraints on the frequencies of itemsets in the randomized data, the constraints are softened (the frequencies of the constrained itemsets in a randomized dataset can be different from the empirical frequencies in the original dataset but the probability of obtaining a randomized dataset exponentially decreases with its distance from the exact constraints). It is also important to note that the fixed rows and column margins constraint is used as an additional constraint in all the randomization processes described in [HOV$^+$09].

In [LPP14] and [MVT12], the practical approaches are similar in the sense that the mining algorithms focus, at each step, on finding the itemset whose frequency diverges the most from the current data model (a randomization model in [LPP14] and a MaxEnt model in [MVT12]). However, the theoretical approach is different in the sense that the method is presented as a greedy heuristic in order to determine the most interesting set of patterns. We address this perspective in the next section.

There are a number of other possible variations to the general iterative learning process and we will present a few of these in chapter 5, some based on previous suggestions we made in [DLL17].

2.3.3.4 Global models defined by interesting and non-redundant sets of patterns

In the approaches which we have described previously, redundancy and, more generally, interestingness are always regarded as properties of a single pattern with respect to a certain amount of knowledge: the interestingness of a pattern with respect to its components, the interestingness of a pattern with respect to predefined background knowledge, or the interestingness of a pattern with respect to acquired background knowledge based on other patterns.

Another approach regards interestingness as a property of sets of patterns. In fact, the focus is not so much aimed towards patterns but towards the models they define. An interesting set of patterns is one that defines an interesting model.

Model evaluation. As the focus is set on models, the issue of evaluating a model becomes central with this approach. For defining a good model, most of the literature focuses on two aspects of the model: its ability to predict the data, on the one hand, and its simplicity, on the second hand. The first aspect is quite easy to define. Measures using likelihood or distances characterize this reasonably well. However, the best model to predict the data is the empirical data model itself, so relying simply on such an aspect would defeat the whole point of pattern mining.

One of the simplest approaches towards this issue is to fix the number of patterns which define the model [MVT12, LPP14]. In this approach, the model with the best prediction defined by k patterns is the best model. However, the value of k is difficult to determine and this results generally in a quite ad hoc choice.

This is not an issue if the measure for ranking models decreases with regards to the complexity of the models (for a given precision in prediction). This is the case while considering statistical testing, such as Pearson's χ^2 test of adequacy, or measures from information theory, such as the Bayesian information criterion (BIC) or the minimum description length (MDL), both present in the itemset mining literature [VVLS11, MVT12, VLV14]. We address the specific case of MDL in the next paragraph.

Pattern mining through compression. Minimum description length has been used in a number of research studies focusing on interesting pattern mining [VLV14, SK11, TV12], and interesting itemset mining in particular [TV08, VVLS11, MVT12]. In order to understand what the approach towards MDL corresponds to, we briefly recall its history and theoretical foundations.

MDL was introduced in the 1978 by Jorma Rissanen [Ris78]. It is founded on algorithmic information theory, which was invented in the 1960s by Andrey Kolmogorov, Ray Solomonoff and Gregory Chaitin [Kol63, Sol64a, Sol64b, Cha69]. Algorithmic information theory is itself founded on both Shannon's information theory and Turing's computability theory [10].

The main idea behind information theory is that, given the output of a program generated by a universal machine, there is a shortest possible program that generates the exact same output. The length of this program in bits is the Kolmogorov complexity of the output. Using this notion and relying on the principle of Occam's razor, a program that generates the output and whose length equals the Kolmogorov complexity of that output is a sounder explanation than any other longer program. Because Kolmogorov complexity is not computable, computable counterparts have been suggested, such as MDL.

To transpose to pattern mining, a pattern language and a model class must be chosen. The pattern language defines which patterns may be mined (for example, in the case of itemsets, the pattern language is the set of all itemsets). The model class defines how the language transposes to a data model (for example, MaxEnt models in [MVT12] or code tables in [VVLS11, SK11, SV12, TV12]). The MDL is computed as:

$$L(D, M) = L(M) + L(D|M)$$

where $L(M)$ is the length of the description of the model in the pattern language and $L(D|M)$ is the length of the description of the deviation of the data with regards to the model. The model which minimizes $L(D, M)$ is considered the best model. In other words, the strongest lossless compression of the data is considered to be the best explanation of the data. The set of patterns which define this compression is therefore considered to be the most interesting set of patterns.

Although this approach is one of the most solidly theoretically founded in pattern mining, it does come with certain limitations. We address some of these limitations in the following chapter and suggest an alternative general theoretical framework for pattern mining.

[10] As Gregory Chaitin, one of its founders, described, Algorithmic information theory is "*the result of putting Shannon's information theory and Turing's computability theory into a cocktail shaker and shaking vigorously.*" [Cal13].

A necessary resort to heuristics. When considering the issue of maximizing interestingness with regards to sets of patterns rather than simply patterns (regardless of the measure for model interestingness chosen among the ones previously cited), the search space for an optimal solution is doubly exponential rather than simply exponential. In the case of itemsets, for m items, there are 2^m itemsets and 2^{2^m} sets of itemsets. Considering only 8 items, the order of magnitude is close to the estimated number of atoms in the Universe and for any value above 5, an exhaustive search within this space is technically infeasible. There are no known results that allow to reduce the search space significantly and sufficiently. Some results show that this may not be the case altogether as the search for an optimal solution has been shown to be NP-hard in some cases [LPP14].

As such, the search of the optimal solution must resort to various heuristics and is limited to the search of locally optimal solutions. Greedy algorithms have been utilized to this effect for models based on randomization methods [LPP14] and MaxEnt methods [MVT12]. We suggest a few possible improvements to such greedy algorithms in chapter 5.

Note that such approaches only bring down the complexity of the method from a double to a single exponential which remains insufficient in many cases. In order to bring this down further still, the problem may be simplified by partitioning the set of items and searching for an optimal solution within the scope of each block of items. This is similar to the approach in [PNS+07], in which clustering is used to partition the items before mining for association rules. However, the models of interestingness considered for mining association rules within each cluster in [PNS+07] are local even within the scope of cluster. Such approaches are also discussed in chapter 5.

2.4 Conclusion

In this chapter, we have presented the history and state-of-the-art of frequent rule and itemset mining while focusing on the approaches that mine for objectively interesting patterns. Of course, there are many other topics covered in the scientific literature on rule and itemset mining and we could not address these all. Other notable areas of research include, for example, the issue of huge data [RU11, AISK14, GLCZ17, GLFV+19], uncertain data [ALWW09, TCCY12, LMT14, LGFV^{+}16] or mining in data streams [JA07, LJA14, AH17]. Furthermore, as our work focuses on theoretical aspects of rule and itemset mining, we have not mentioned the very wide range of applications. This include notably data clustering [FAT^{+}14, ZAV14, ZMM15, DBFVL18], data classification [GBMTCO10, NVHT13, ZN14, EGB^{+}17] and a large panel of application domains ranging from web mining [IV06, SSLL08, NB12, K^{+}12, SO15] to text mining [BEX02, ZYTW10, AZ12], biology [PCT^{+}03, MO04, CTTX05, MMB^{+}18, ZAZ19] to chemistry [BB02, DKWK05, HXH^{+}20], medicine [Kha11, CTH^{+}13, TPMD^{+}13] to sociology [AAR09, NCC^{+}12, FC13, MJM^{+}17] and many more.

In the following chapters, we will present our contributions to this field. In chapter 3, we analyze various theoretical issues in current approaches and the theoretical boundaries of the field. From this analysis, we suggest a number of recommendations for defining a general theoretical framework for pattern mining. In chapter 4, we present our novel approach towards MaxEnt models through mutual constrained independence (MCI). We show that MCI offers both further insight on the rationale behind MaxEnt models and novel techniques for computing such models. Theses techniques, based on tools from algebraic geometry, allow us to provide for direct exact solutions to a class of MaxEnt models which had not been previously solved. In chapter 5, we present algorithms for extracting objectively interesting patterns from data based on the principles which are defined in chapter 3 and the tools which are presented in chapter 4. We discuss the issues that arise from a direct implementation of these principles and give guidelines for future research that could allow to tackle these issues, while suggesting current solutions based on compromise.

CHAPTER 3

Meaningful mathematical modeling of the objective interestingness of patterns

3.1 Introduction

In the previous chapter, we have presented the history, development and state-of-the-art of frequency-based[1] itemset and rule mining, with a specific focus on the issue of extracting objectively interesting patterns. Each and every approach which we have described relies on mathematical models. The mathematical definitions of these models are usually quite explicitly laid out in the literature. However, the modeling processes themselves are not always made apparent. In fact, more often than not, they are brushed aside as if they were irrelevant or simply and totally ignored. As we have discussed in our foreword, this tendency to be oblivious to the mathematical modeling process can be seen as a cultural trend which goes far beyond the specific field of frequency-based pattern mining.

Mathematical modeling is often described as the act of establishing a correspondence between a system (usually corresponding to some aspect of the real world) and a mathematical model. When applied to a system, every mathematical model relies, explicitly or implicitly, on a mathematical modeling. The mathematical modeling is what provides meaning to the mathematical model. It is the link between mathematics and reality. Without this link, a

[1]Recall that we use the less ambiguous frequency-based itemset mining terminology or, more generally, frequency-based pattern mining rather than the more common frequent pattern mining terminology which is used ambiguously to refer both to the process of mining patterns that are frequent in the data or to the process of mining patterns based on their frequency in the data.

mathematical model is just an abstract construction and the results that we can obtain through the model are meaningless with regards to reality.

Therefore, questioning the meaningfulness of the answers provided by a method based on mathematical models implies a necessary inspection of the corresponding underlying mathematical modeling processes. If the modeling processes are implicit, they need to be made explicit in order to do so. As one of the main aims in this doctoral thesis is to provide both efficient and meaningful methods for extracting objectively interesting patterns, we undertook the task of identifying the various underlying modeling processes inherent to the different approaches in the literature. This led us to analyzing different mechanisms involved in these modeling processes and their impact, in terms of meaningfulness, on the general modeling process. We came to understand that there are certain number of recurrent issues within the research in the literature directly related to the modeling processes described or the absence of explicit descriptions of these modeling processes. Hence, we suggest a number of recommendations for a meaningful mathematical modeling of objective interestingness of patterns based on our analysis.

3.1.1 Modeling and mathematical modeling

3.1.1.1 Modeling as translation

Before we address any specific issue related to modeling processes, we must start by clearly defining the notion of mathematical modeling and, more generally, of modeling. The issue of defining the notions of models and modeling has been extensively debated by researchers studying the philosophy of science and definitions vary from one author to the other [Apo61, Sup61, Hes65, Min65, Sup69, Sta73, McM85, Gie88, RWLN89, BJ99, C+99, Gie99, BJ02, GS06, TJ06, Sup07, Wei07, G+09, BJ09, Con10, Cha13, Pot17]. We build upon some of these definitions to suggest a new description of models and modeling. The definition we propose here follows in the tradition of philosophers who have adopted a rather broad view towards the definitions of models and modeling. These include the definitions given by Leo Apostel in [Apo61]:

> *This will be our final and most general hint towards the definition of model: any subject using a system A that is neither directly nor indirectly interacting with a system B to obtain information about*

the system B, is using A as a model for B.

Marvin Minsky, whose definition of a model was given in [Min65]:

> *To an observer B, an object A^* is a model of an object A to the extent that B can use A^* to answer questions that interest him about A.*

as well as the definition by Jeff Rothenberg in [RWLN89]:

> *Modeling in its broadest sense is the cost-effective use of something in place of something else for some cognitive purpose.*

and the idea by Herbert Stachowiak in [Sta73] that:

> *all of cognition is cognition in models or by means of models, and in general, any human encounter with the world needs a "model" as the mediator [...].*[2]

We defend the idea that modeling is a translation. It is the act of establishing a correspondence between two languages each of which allow a representation of the world. The model corresponds to the representation of the world in the modeling language. The aim is to use the second language in order to answer questions about the world which cannot be satisfactorily answered in the first language as illustrated by Figure 3.1. Therefore the second language must allow for inference. It should also generally be more structured than the initial language, allowing for more acceptable forms of inference, in order to satisfy the aim of the modeling process. Our approach of modeling and models intentionally differs from that of a number of philosophers who see modeling exclusively as a representation of the real world (such as Giere [G^+09], Magnani [Mag12] or Portides [Por14]) or based on formal models (such as Suppes [Sup61, Sup64, Sup68, Sup69], van Fraassen [VF67, VF70, VF10] or da Costa and French [DCF90, DCF03]). Firstly, we support the views of Stachowiak and Podnieks that the world is always accessed through models [Sta73, Pod18]. Hence our starting point is not the world per se, but some representation of the world in an initial language (even though we might refer to this representation as *the world* for further simplicity in the other sections of

[2] Translation from German to English from [Pod18].

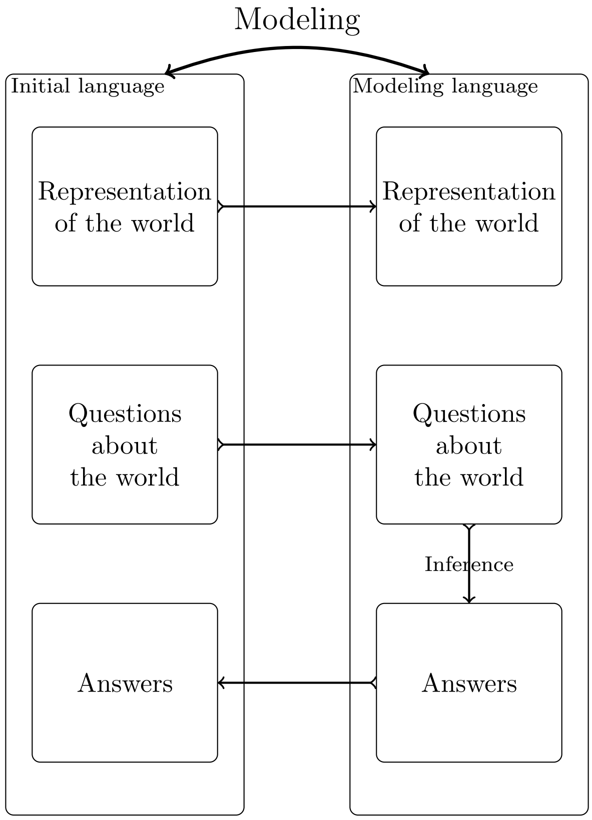

Figure 3.1: Modeling as a translation

this chapter). Secondly, our approach allows to consider mathematical modelings as well as a wide variety of other modelings, as long as we accept a broad definition of the notion of language. For example, we can easily include computer models if we consider programming languages as possible modeling languages[3]. Indeed, in such a case, we can model our initial representation through a computer program implemented in a given programming language together with a set of input variables. A question about the world can then be represented by a subroutine of the program and the answer to this question by the output of this subroutine. The inference step of the modeling process corresponds here to the execution of the subroutine on a computer. Note here that the ability to provide for answers in the modeling language depends on the systems that perform the inference step of the modeling process: computers in the previous example; mathematicians in the case of mathematical models; etc.

[3] Martin Thomson-Jones states that "*one outstanding issue*" with his own taxonomy of models is that it does not encompass such models [TJ06].

We must add that the use of the term *translation* can also be seen as a relevant choice for describing the idea, emphasized as an important aspect of models by authors such as Cartwright [CM84, C+99], Strevens [Str08], Wimsatt [Wim87] or Potochnik [Pot17], that models are necessarily idealized and therefore false representations of the reality they describe (see also [Wei07] on idealizations). Similarly, a translation can never perfectly transpose every aspect of an original text, a characteristic often referred to by the Italian expression *traduttore, traditore* meaning *translator, traitor*. As translations betray original texts, models lie about the world. The analogy may even be pushed slightly further as, in both contexts, authors argue that a good translation or a good model is necessarily unfaithful or a lie (see [Mou76] for translations and [CM84, Wim87, Str08] for models).

3.1.1.2 Mathematical modeling

Through this approach, mathematical modeling is simply modeling into mathematical language. A mathematical modeling process establishes a correspondence between the representation of the world in the initial language (usually a technical form of a natural language) and a mathematical model. It also establishes a correspondence between questions about the world in the initial language and mathematical problems. Most importantly, it allows to establish a correspondence between mathematical solutions to the mathematical problems and answers to the questions in the initial language.

The significant advantage of using a mathematical modeling compared to other types of modeling is that the mathematical language is exact and infallible. Indeed, once a mathematical problem, or a class of mathematical problems, has been solved mathematically (i.e. there is a mathematically correct proof of the solution to the problem) it has been solved definitely. Moreover, the solution is independent from the proof: even if the proof may be extremely tedious and complex, the formulation of the problem and its solution may be quite simple, allowing for it be used easily, over and over again, as many times as needed. Hence, provided that we can trust mathematicians to deliver correct mathematical solutions to mathematical problems, we can entirely trust the problem solving step in the modeling process. Hence, the only debatable step in a mathematical modeling process is the establishment of the mathematical modeling correspondence. This allows to circumscribe issues related to

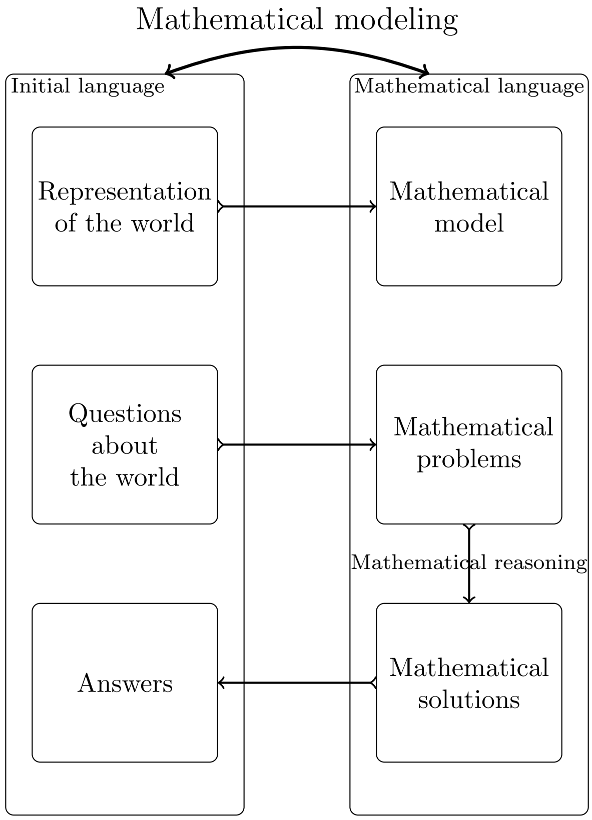

Figure 3.2: Mathematical modeling

meaningfulness to the correspondence established by the modeling: a general mathematical modeling process is meaningful if and only if the mathematical modeling correspondence is meaningful. In particular, if an answer given by a mathematical modeling is to be contested, then the only thing there is to contest is the adequacy of the mathematical modeling with regards to the simplifications, approximations and representation choices which were made.

Another argument in favor of the use of mathematical models is based on the idea that mathematics is the fundamental language of the Universe. This theory, defended by many mathematicians and scientists every since Galileo Galilei expressed it in Il Saggiatore [Gal23], implies that there is a mathematical modeling which exactly represents the world as it is. Hence, the only way to truly access the essence of reality is through mathematical models. In contrast with the idea that models lie about reality, this theory suggests that the truth about reality can only be described by a mathematical model. Note that this is not incompatible with the idea that models, as translations, move away from the initial representation of the world which is modeled. Indeed, as we have stated, this initial representation is already at a distance from the world per se. Hence moving away from this representation may in some cases lead us towards reality rather than away from reality.

3.1.1.3 Complex modeling processes

Although the description of modeling and models which we have given in this section allows for a broad and elegant approach towards these notions, we understand that it falls short when it comes to analyzing the complex structure of actual modeling processes if it is not refined. Indeed, more often than not, multiple modelings of various aspects of the world are used, expressed in multiple modeling languages, and combined with modelings of modelings, thus creating a complex structure for the general modeling process. This does not imply that Figure 3.1 is not valid for such cases: it stays valid if we consider that the modeling language is a complex aggregation of all the modeling languages used within the general process. However, it does not inform much about such complex modelings and we need to be able to describe these complex processes. Though are main focus in the rest of this chapter is mathematical modeling, we also describe a type of such complex modeling approaches which we name patchwork modelings in section 3.5.

Furthermore, in the pattern mining modeling processes that we address in this thesis, at least two modeling languages are used within the general modeling processes: the mathematical language; and a programming language. In some cases the modeling in the programming language will describe most of the mathematical objects from the mathematical model (this is usually the case for randomization methods), while in other cases the modeling in the programming language will simply be used to compute the mathematical solutions defined through the mathematical modeling but without any regards towards the other aspects of this mathematical modeling. However, in all of the cases we address, the computer modelings follow in sequence after the mathematical modelings in the general modeling processes: they model aspects of the mathematical modeling. Hence, the meaningfulness of the general modeling process lies foremost in the mathematical modeling. Of course, the computer modeling plays a distinctive role in the definition of the general modeling process and even of the mathematical modeling. In particular, the ability to infer an answer (in this case to reach an output) in the computer modeling sets the restrictions, in terms of complexity, for the usefulness of a mathematical modeling. However, the meaningfulness of the computer modeling is, in the cases we address, not a significant issue.

3.1.2 Chapter outline

Following the definition of the notion of mathematical modeling in this section, we address a number of aspects related to mathematical modeling in objective frequency-based interesting pattern mining. We define novel terminology for describing general features of mathematical modeling processes which we consider to be quite relevant in this context but also in most other contexts in which mathematical models are applied. The necessity for defining this new terminology stems from the fact that we have not found any preexisting terminology allowing to precisely describe the issues which we pinpoint in any of the literature which we have knowledge of, whether in the fields of computer science, mathematics, or the philosophy of science. As such, this chapter can be viewed as much as a contribution to the philosophy of science and mathematics as a contribution to the field of pattern mining.

In section 3.2, we compare modeling processes in which the data is the main subject of the modeling process to those in which it is only a deriving object and the main subject is the mechanism that generated such data. In section 3.3, we compare two general approaches towards the definition of a mathematical model: phenotypic and genotypic modeling. We show how different modelings in pattern mining relate to these approaches and discuss some issues of phenotypic modeling. In section 3.4, we address the issue of pragmatic modeling (i.e. modeling which is based foremost on pragmatic considerations). In section 3.5, we introduce the notions of patchwork and holistic mathematical modelings and address specific issues related to patchwork modeling in pattern mining. In section 3.6, we consider the modeling of patterns within the general modeling process and particularly the use of a mathematical model based on measure spaces and Boolean lattices. In section 3.7, we focus on the modeling of objectivity within the general modeling process. Finally, we conclude in section 3.8 on the definition of a mathematical modeling satisfying all the various recommendations previously defined in this chapter.

Note that the length of each of these sections may vary significantly. However, this only indicates how much detail we felt was necessary to present each of these aspects and does not reflect in any way the importance of the related recommendations.

3.2 The data: subject or object of the modeling process

One important aspect that allows to categorize most of the different modeling approaches in the literature is related to the position held by the modeling of the data within the general modeling process. We discern two main categories. On the one hand, we consider the modeling processes in which the data is regarded as a main subject of the modeling process, and on the other hand, those in which the data is regarded as an object deriving from a main subject of the modeling process. By subject, we mean that the data is modeled using a mathematical model which does not conceptually derive from any another mathematical model within the general modeling process. Conversely, we use the term object to signify that the data is modeled as the result of a mechanism which is itself described in the modeling process through a mathematical model. In the first case, the main subject of the modeling process is the data and, in the second case, it is the mechanism that generates the data.

Classical statistical approaches fall into the second category. Consider, for example, the modeling process described in section 2.3.3.1 in which the database is modeled as a random sample of n independent identically distributed random variables $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_k)$ with values in $\{0, 1\}^k$. In this case, the subject of the modeling process is the random process which generated the data. The data is only seen as a means to gain information regarding that random process. This is essential because the random process corresponds to a lower level of modeling than the database which has a much larger scope: the modeling process considers any other transaction on the same items to be a result of that single random process. This in turn provides a basis for justifying the generalization of the results obtained on one database to other databases.

Conversely, any modeling process that considers the database as a subject per se does not provide any basis for utilizing the results obtained on one database to other databases. Note that this is not necessarily an issue. If the aim of a particular data mining method is to provide the best compression for one very large video file or the best summary for one given book, it is not a necessary condition to consider the mechanisms that generate video files or books in the mathematical modeling for the data mining method. In both examples, there is no need to generalize the results.

In many other applications of pattern mining, generalizing is nevertheless necessary. Take classification, for example. In order to be able to justify why a classification method learned on a given database can be used to classify elements from another database, it is necessary that the modeling process for this classification method include descriptions not only for the initial database, but also for other potential databases, and for the relationship between them that justifies that knowledge about this one database is knowledge about all databases. Modeling the underlying mechanisms that generate the data through a random distribution is one way to go but there could be many other possible manners to accomplish this.

Note that, for a statistician, what we describe here may seem quite obvious. But recall from our foreword that the field of our study is at the junction of different scientific cultures and that what may be perceived as general knowledge in one culture may not be perceived as such in the other. The examples in the literature show that this is indeed not the case.

We present here two general cases of ambiguous mathematical modeling with regards to the position held by the modeling of the data within the general modeling process. In both cases, this leads to theoretical issues.

3.2.1 The data modeling process in the case of objective interestingness measures for rules

Consider the case of objective interestingness measures for rules as presented in section 2.3.1.2. In the vast majority of the literature dealing with such measures, there is no explicit reference to a random variable that generates the data. If this were the case, there would be a distinction between the probabilities p_X, p_Y and $p_{X \wedge Y}$ which define the probability distribution for the random variable defined on contingency tables and the frequencies f_X, f_Y and $f_{X \wedge Y}$ which correspond to the distribution in the empirical data. As we have stated previously, we have found only a few cases in the literature pertaining specifically to objective interestingness measures of rules were this distinction is made explicit, and none of them correspond to the most cited papers (see [GSS12], for example). In most cases, it is the probabilistic notation, however, which is preferred rather than one using frequencies. There are three possible explanations for this.

The first possibility is that this corresponds to the case were the data is modeled as a subject and the probabilities that are described are simply referring to the probability distribution of the empirical data. In this case, as explained in the beginning of this section, there is no theoretical ground for generalizing the results from this dataset to any other dataset.

The second possibility is the case in which the data is modeled implicitly as an object deriving from a distinguishable subject. Even if this implies that a same notation is used alternatively to describe the probability distribution for a random variable and the frequencies in the empirical distribution, this would make sense for a certain number of interestingness measures as we have described in section 2.3.3.1. This is not an issue if a single rule is considered. However, when considering multiple rules on itemsets, this could generate a conflict within the general modeling process, in the sense that a single aspect of reality can be modeled through two different and incompatible mathematical models. We will address such issues in detail in section 3.5.

A third option is that the data is modeled as an object deriving from an indistinguishable subject. In this case, the data is considered to be generated by a random variable. However, based on a frequentist approach towards probability, the distribution of this random variable is exactly equated with the empirical distribution in the data. In this sense, summarizing the dataset is equivalent to summarizing the mechanisms that generate such datasets. Regretfully, this is mostly an artifice for presenting the first case with a means to justify that the results of the mining process may be generalized from one dataset to other datasets. Indeed, while there is sufficient ground for justifying that the distribution of the random variable be equated to the distribution of the empirical in the modeling process when considering a few items as long as the number of transactions n is large enough, there is no way that the number of transactions considered will ever be large enough to justify such an approximation when considering a hundred items, let alone a hundred thousand items (see section 3.7 for more details).

In all these cases, the modeling process raises issues. Moreover, the fact that the modeling process for the data is not made explicit makes it complicated to address these issues.

3.2.2 The data modeling process when considering the fixed row and column margins constraint

The second example we present here is related to the fixed row and column margin model that is one of the few global models which is considered in itemset mining. In the itemset mining literature, this model is especially present in research papers focusing on swap-randomization methods [GMMT07, HOV+09, LPP14] but also commonly used when considering MaxEnt models [KDB10, TM10, DB11, MVT12, MV13]. We show here how considering this model may hinder the ability to meaningfully generalize the obtained results.

The swap-randomization methods in the literature mentioned above all share a common genealogy with a general problem from discrete tomography: describe the space of all $n \times m$ binary matrices with given row and column margins [HK08, HK12]. It has been demonstrated that this space can be represented via a graph, in which the vertices represent the matrices and the edges represent a swapping operation between two matrices, composed of a single connected component [Bru80]. This representation provides for a mean to utilize methods such as random walks on graphs to randomly generate a matrix satisfying the same conditions on the margins as a given initial matrix. Swap-randomization methods have been used to model data particularly in the domains of ecology [SMS98, CDHL05, SNB+14] and psychometrics [Pon01, CS05, Ver08].

When modeling data using a fixed row and column margins model, the idea is that these margins correspond to defining characteristics in the system that is described by the data. Hence, the data is modeled as a single random sample of a random variable $\mathbf{D} = \begin{pmatrix} \mathbf{d}_{1,1} & \cdots & \mathbf{d}_{1,m} \\ \vdots & \ddots & \vdots \\ \mathbf{d}_{n,1} & \cdots & \mathbf{d}_{n,m} \end{pmatrix}$ following a uniform distribution on all $n \times m$ matrices satisfying the fixed row and column margins constraint.

This makes sense for a certain number of applications in ecology, such as the case of the classical study of the distribution of bird species among islands in an archipelago [SMS98, CDHL05]. In this case, the datasets indicate the presence of a given species of bird (among m species) on a given island (among n islands). As specialists defend that the natural characteristics of each island only allow for a given number of species to cohabit while the natural characteristics of each species define their widespreadness, the model can be justified. Any divergence from the model indicates information about the specific distribution which is studied (i.e. how these species of birds are distributed within this archipelago) and the modeling process does not provide for a basis to justify the generalizing of the results obtained to any other system (at best a specialist with further knowledge about such systems might rely on these results to formulate general hypotheses).

This could also make sense in the case of evaluation datasets that indicate which questions (among m questions in an examination) were correctly answered by a given examinee (among n examinees). Indeed, both the general levels of the examinees (represented by the number of correct answers given by each examinee) and the general difficulty of the questions (represented by the number of correct answers at a given question) may be seen as defining characteristics of the system. This explains the popularity of fixed row and column margins model in psychometrics, whether based on swap-randomization [Pon01, CS05, Ver08] or the corresponding MaxEnt model: the Rasch model [Mas82, Kel84, KKB$^+$17]. However, it must be understood that the associated modeling processes do not provide any basis for generalizing the results obtained to a larger pool of potential examinees.

Similarly, in the classical itemset example of market basket analysis, it makes sense to consider that the size of a consumer's basket (i.e. the number of items in a transaction) is a defining characteristic of that consumer's purchasing habits. However, one of the main goals of market basket analysis is to gain knowledge about general consumer habits rather than those simply related to the pool of consumers within the dataset.

Let us examine more closely what it means in terms of data modeling when these fixed row and column margin models are combined with itemset mining models as in [GMMT07] or in [LPP14]. If the aim is to generalize the results obtained to other transactions there are necessarily theoretical issues with the general modeling process. Indeed, either we consider that the frequencies of an itemset I can inform us about the distribution of a random variable $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)$ which generated the data. In this case, we have two competing and incompatible models for the dataset: as a random sample of n independent identically distributed random variables $\mathbf{X}$, on the one hand, and as single sample of the random variable $\mathbf{D}$, on the other hand. Alternatively, if we only consider the data modeling process associated with the fixed row and column margins model, then there is little basis for justifying that the observed value of $\sum_{\substack{1\leq i\leq n\\ j\in I}} \mathbf{d}_{i,j}$ (corresponding to the frequency of itemset I) for a single sample gives any substantial information about the distribution defining the random variable $\mathbf{D}$. Once more, note that if the aim is simply data compression and not generalizing there are no issues with this modeling process.

In contrast, the model suggested in [TM10] does not suffer from these theoretical issues. Indeed, the authors of this paper suggest that rather than considering the fixed row and column margins constraint similar constraints could be defined with regards to the random variable $\mathbf{X}$. First, the fixed column constraint is easily replaced by the constraint based on the frequencies of an item:

$$p_i = f_i$$

where p_i is the probability that $\mathbf{x}_i = 1$ and f_i is the observed frequency of item a_i. Second, the fixed row constrain is replaced by a constraint based on the frequencies of transactions of size k:

$$\forall k \in [\![0, m]\!],\ p_{|\mathbf{X}|=k} = f_{|\mathbf{X}|=k}$$

where $|\mathbf{X}|$ is the number of ones in $\mathbf{X}$. Note that these are, in fact, two well defined linear constraints with regards to the elementary probabilities p_ω where $\omega \in \Omega = \{0,1\}^m$:

$$p_i = \sum_{\substack{\omega_i=1\\ \omega\in\Omega}} p_\omega \quad \text{and} \quad p_{|\mathbf{X}|=k} = \sum_{\substack{|\omega|=k\\ \omega\in\Omega}} p_\omega$$

Of course, it could be argued that if n and m are large enough the two approaches described in this section give similar results. However, if we are concerned by the meaningfulness of the process, theoretical rigor is necessary and a distinction between the two must be made. If a swap-randomization method, based on a fixed row and column margins constraint, is the easier option in terms of computability (which is rarely the case) and that the method is aimed at prediction, then it should be specified that the swap-randomization method is only used to approximate the probability distribution defined in the second approach and the appropriateness of the approximation should be justified. Otherwise, the use of fixed row and column margins models should be limited to exhaustive datasets, for describing other inherent properties of a given system in which they justifiably correspond to background knowledge (such as in the previous example in ecology), or for summarization.

3.2.3 Recommendation

We have shown in this section that, in order to provide for a meaningful justification of the utilization of a data mining method for predictive tasks, the mathematical model for the empirical dataset must derive from a more general mathematical model encompassing all potential datasets. The model must also allow for a justification to why the observation of one dataset can provide information about the general mathematical model for all potential datasets. These elements are, however, not compulsory for compression tasks.

In any case, it is important that the modeling process leading to the definition of the mathematical model for the data may be made explicit in order to provide for a meaningful explanation of the mining method and detect possible inconsistencies within the modeling process. This is not a superfluous task, as such inconsistencies are not rare in the literature.

3.3 Phenotypic modeling and genotypic modeling of interestingness

3.3.1 Phenotypic and genotypic modeling: a definition

In the previous section, we addressed a particular aspect of the general modeling process related to the modeling of data. In this section, we focus on another aspect of the modeling process related to the modeling of interestingness. To describe this aspect, we will borrow elements from the terminology of the field of biology which illustrate quite adequately this notion: phenotypes and genotypes. In biology, the phenotype of an organism is its observable characteristics while its genotype is its genetic makeup. The phenotype is an expression of the genotype within a certain environment. Before the discovery of the role of DNA in genetic transmission and the development of DNA sequencing technologies, the only way of describing an organism was through its phenotype. Nowadays, accessing the organism's genotype provides for deeper understanding of both the organism and its phenotype.

Similarly, in order to model an object or a concept, one can rely on its traits and characteristics (i.e. its phenotype) or on its core code from which these characteristics derive (i.e. its genotype). These correspond to two different modeling approaches, which we shall refer to as phenotypic modeling and genotypic modeling. While the first approach provides for a method to model an object whose intrinsic nature is unknown or unclear, the second approach generally provides for a more meaningful explanation of the properties of the object.

Indeed, if a mathematical model is defined so that some of its mathematical properties correspond to some of the characteristics of the modeled object, then it is still hard to justify why the other mathematical properties of the model should correspond to actual characteristics of the modeled object. This is an issue because there is no reason a priori that these other mathematical properties do not influence the way the model behaves in a given context.

Note that the concepts of genotypic and phenotypic modelings are close to the concepts of *white-box* and *black-box models* as described in [KK15]. However, this is not the standard use of the term black box model which corresponds usually to the inability to fully understand the mechanisms of the model itself and is an entirely different issue. This motivated our use of this novel terminology.

Phenotypic approaches, on the one hand, are commonly used to describe systems or concepts for which there is more consensus on their characteristics than on their intrinsic nature. This is often the case for notions in social sciences like economics where the use of indicators relies mostly on phenotypic modeling approaches. This can also be the case for the notion of interestingness as we will discuss further in this section.

On the other hand, genotypic approaches can be used for systems or concepts which have been clearly defined as resulting from some underlying mechanism. In this case, the modeled concept is represented by a mathematical object which is defined by other mathematical objects such as a quantity, a distance, a probability, or any other more complex mathematical object.

3.3.2 Phenotypic and genotypic approaches for measuring objective interestingness

3.3.2.1 Phenotypic approaches for defining objective interestingness measures

When it comes to objective interestingness measures for rules, an entire portion of the literature specifically focuses on phenotypic modeling approaches (see section 2.3.1.2 for a detailed review of this portion of the literature). Phenotypic modeling approaches have been used for defining individual measures [PS91, GCB+04, BGK09, GSS12] as well as choosing a measure among potential measures [LMP+03, TKS04, OKO+04, LMV+04, LT04, GH06, LMVL08, LB11].

Indeed, the three principles, presented by Gregory Piatetsky-Shapiro in [PS91] as intuitive principles that all interestingness measure for a rule should satisfy, are in fact mathematical properties which model what he perceives as specific traits of interestingness. He then suggests the use of a measure that he defines as the most easily computable function which satisfies these properties. In a sense, the traits he considers are not defining characteristics of interestingness because there is an infinite number of possible functions which satisfy these properties while exhibiting very different behaviors otherwise. However, he still relies on these properties to propose a mathematical model for interestingness. This type of modeling approach falls completely within the phenotypic modeling category which we have described.

In the later scientific contributions involving the definition of additional properties of objective interestingness measures and the classification of a number of such measures with regards to these properties [OKO+04, GH06, LMVL08, BGK09, LB11, GSS12][4], the modeling processes for interestingness are slightly different but still fall into the same general category of phenotypic modeling approaches. Indeed, in these papers the properties suggested are meant to model possible traits of interestingness. The idea is to choose a model for interestingness, i.e. an objective interestingness measure, based on which properties the measure satisfies, that is based on which traits are chosen to characterize interestingness. The exact list of traits which are intended to be utilized in order to characterize interestingness in a given modeling process

[4]See section 2.3.1.2 for a detailed review.

is not defined a priori, contrarily to what was the case in [PS91], but rather left to the user or an expert to specify. In any case, as in [PS91], the lists of traits chosen and their associated properties do not entirey characterize the measures that are picked to model interestingness. Indeed, the measures are only chosen from a relatively short list of measures defined explicitly in the literature, short with respect to the infinite number of substantially different potential measures which exist for any given combination of these supposedly characteristic properties. Therefore, such approaches fully register as phenotypic modeling approaches and carry the related issues in terms of meaningfulness.

It can be argued that, what these methods lose in terms of meaningfulness of the modeling process for interestingness, they gain in terms of flexibility towards the definition of interestingness. Indeed, for any combination of supposedly characteristic traits of interestingness[5], one can always either find an objective interestingness measure in the literature that fits the corresponding properties or suggest a novel objective interestingness measure which would fit. For example, in [BGK09], the authors suggest two new measures in order to satisfy two combinations of properties which they believe could correspond to some users' view towards interestingness and for which they did not find any corresponding measure in the literature. However, the fact that no prior existing objective measure corresponded to these two visions of interestingness could also suggest that not all possible ways of describing interestingness correspond to what is seen as objective interestingness by most researchers.

It is worth noting, at this point, that there can be a semantic confusion regarding the expression objective interestingness measure. The term *objective* can be seen as qualifying the measure, interestingness, or both. If the term only applies to the measure (see paragraph 2.3.1.2 for more details concerning subjective and objective measures), then the modeling process described above qualifies as modeling interestingness through an objective interestingness measure. However, if the term applies to interestingness, as in *mining for objectively interesting patterns*, then this process appears as highly subjective. If the essence of interestingness is fundamentally subjective, then this is inevitable. Nevertheless, as we will expose in section 3.7, a notion of objective interestingness can very well be defined, and it certainly cannot be described by using alternative opposing views. Hence, these phenotypic ap-

[5] We use the term *supposedly characteristic* as they are meant to characterize interestingness but fail to do so entirely as we have previously mentioned.

proaches carry inherent issues regarding the two main aspects of our goal to provide for meaningful modeling of objective interestingness.

3.3.2.2 Genotypic approaches for modeling interestingness

In the case of genotypic modeling, the modeled concept (in this case interestingness) is modeled as a mathematical object which is defined by other mathematical objects within the model. This is the case for the models described in sections 2.3.3.1, 2.3.3.2 and 2.3.3.3 in which interestingness of a pattern is defined as its statistical surprisingness or informativeness (in the sense of information theory) relative to some data model. This is also the case for the various approaches described in section 2.3.3.4 in which sets of patterns are considered interesting if the data model they define is a good model for the dataset and where the specifics of the modeling process depends on the choice of the measure for evaluating the different data models. Notably, this is the case for the data mining as summarization approach which uses the minimum description length principle. Indeed, in this approach, the interestingness of a given set of patterns is equated with the length (or more exactly the shortness) of the description of the data given by the data model defined by the given set of patterns with the correction corresponding to the error between the data model and the empirical data.

In all cases described above, the modeling of interestingness is associated to a clear definition of interestingness. Although one might question the adequacy of the specific definition chosen for interestingness, the modeling processes are clearly meaningful with regards to this choice.

3.3.3 When phenotypic modeling meets genotypic modeling: modeling information

We set aside the issue of modeling interestingness here to focus on the modeling of informativeness and, more generally, information through entropy. As we have mentioned previously, maximum entropy (MaxEnt) models are commonly used in frequency-based interesting pattern mining. A study of the rationale towards the use of entropy as a mathematical model for information is therefore essential for understanding the meaningfulness of a modeling process based on MaxEnt models. We will show here that there exists both phenotypic and genotypic modeling approaches of information which lead to the same mathematical model of entropy.

3.3.3.1 Phenotypic approaches for modeling entropy

Information entropy was first presented by Claude E. Shannon in 1948 [Sha48] as a function H defined for any probability distribution $\mathbf{p} = (p_i)_{1 \leq i \leq n}$ as below.

$$H(\mathbf{p}) = -\sum_{i=0}^{n} p_i \log p_i$$

In this founding paper, Shannon asks the following question:

> *Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?*

He then pursues immediately by asserting that:

> *If there is such a measure, say $H(p_1, p_2, \cdots, p_n)$, it is reasonable to require of it the following properties:*

followed by a list of three fundamental properties. The paper then continues by stating as a theorem that entropy is the only function satisfying these properties up to a multiplicative constant. However, this theorem is not presented as the main justification for the use of entropy as a model for information:

> *This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications. [...] The quantity H has a number of interesting properties which further substantiate it as a reasonable measure of choice or information.*

which is followed by six important properties of entropy.

Hence, Shannon establishes a distinction between axiomatic properties which uniquely define entropy and inferred properties. However, the rationale for the use of entropy as a model for information is based mostly on the properties of entropy as a whole, axiomatic and inferred. As such, this is clearly a phenotypic modeling approach. Later in the literature, some debate arose over whether the rationale for entropy relied on its definition through axiomatic properties or its properties as a whole [Csi76, SJ80]. A number of different axiomatizations of entropy were suggested [Fad56, CM60, AFN74, AD75, For75, SJ80]. While it was acknowledged that these axiomatizations supported the justification of entropy as the unique natural model for information (as it is clearly stated in W. Weaver's introductory notes to the 1949 republication of Shannon's paper [Sha49]), the rationale for the entropy model for information was mostly based on its properties as a whole [CK11, CT12]. In any case, both approaches are phenotypic in the sense that the modeling is based on the properties which are expected of a mathematical model for information. However, the uniqueness of such a measure obtained through the axiomatic approach suggests that a genotypic approach should yield the same model. Indeed, if this were not the case, it would imply either that some of the axiomatic properties chosen to define the entropy model do not correspond to characteristics of information or that information simply cannot be modeled mathematically. As we show in the next paragraph, genotypic approaches for modeling information do indeed yield the same entropy model.

3.3.3.2 Genotypic approaches for modeling information

A first genotypic approach towards entropy was described by Edwin T. Jaynes in [Jay82]. For Jaynes, the rationale behind the use of the entropy model and, more specifically, MaxEnt models before his work was based on a number of intuitive principles. However, he stated that:

> *While each of these intuitions doubtless expresses an element of truth, none seems explicit enough to lend itself to a "hard" quantitative demonstration of the kind we are accustomed to having in other areas of applied mathematics. Accordingly, many approaching this field are disconcerted by what they sense as a kind of vagueness, the underlying theory lacking solid content.*

The opposition that Jaynes placed in prior approaches towards entropy based on *intuitions* and an approach based on *a "hard" quantitative demonstration* corresponds here exactly to the opposition between phenotypic and genotypic modeling approaches which we have described in this section.

Consider a partition of N elements into n different categories. Let $N_1, ..., N_n$

be the number of elements in each category. There are

$$W = \frac{N!}{N_1!N_2!...N_n!}$$

partitions corresponding to the values $N_1, ..., N_n$. Hence, if we consider a random variable $\mathbf{X} = (X_1, ..., X_n)$ for the partition of N elements in n categories given by the uniform distribution on the set of all such partitions (which is isomorphic to $[\![1, n]\!]^N$), we have:

$$\text{Prob}\,(X_1 = N_1, ..., X_n = N_n) = \frac{W}{n^N}$$

More generally, if we limit the set of possible partitions by constraining them such that $(N_1, ..., N_n)$ belongs to a non empty subset S of $\mathcal{P}_{N,n}$ as defined below:

$$\emptyset \subsetneq S \subset \mathcal{P}_{N,m} = \left\{ (N_i)_{1\leq i\leq n} \in [\![0, N]\!]^n \;\middle|\; \sum_{i=0}^{n} N_i = N \right\}$$

Then:

$$\text{Prob}\,(X_1 = N_1, ..., X_n = N_n) = \frac{W}{|S|}$$

In both cases, the probabilities are proportional to W, so we can compare two probabilities associated to two cardinalities W and W' by comparing these two cardinalities.

Furthermore, consider a probability distribution $\mathbf{p} = (p_i)_{1\leq i\leq n}$ and a sequence $\left(\left(N_i^{(k)}\right)_{1\leq i\leq n}\right)_{k\in\mathbb{N}}$ such that:

$$N^{(k)} = \sum_{i=1}^{n} N_i^{(k)} \xrightarrow[k\to+\infty]{} +\infty$$

and:

$$\forall i \in [\![1, n]\!],\; \frac{N_i^{(k)}}{N^{(k)}} \xrightarrow[k\to+\infty]{} +\infty$$

Then, if $\left(W^{(k)}\right)_{k\in\mathbb{N}}$ is the corresponding sequence of cardinalities:

$$\frac{1}{N^{(k)}} \log W^{(k)} \xrightarrow[k\to+\infty]{} H(\mathbf{p})$$

which is obtained through Stirling's approximation.

Now, let $\left(S^{(k)}\right)_{k\in\mathbb{N}}$ be a sequence of non empty subsets of $\mathcal{P}_{N^{(k)},n}$ defined by:

$$S^{(k)} = \left\{ \left(N_i^{(k)}\right)_{1\leq i\leq n} \in \mathcal{P}_{N^{(k)},n} \;\middle|\; \mathbf{f}\left(\frac{N_1^{(k)}}{N^{(k)}}, ..., \frac{N_n^{(k)}}{N^{(k)}}\right) = \mathbf{v}^{(k)} \right\}$$

where $\mathbf{f} : [0,1]^n \to \mathbb{R}^m$ is a continuous function and $\mathbf{v}^{(k)} \xrightarrow[k\to+\infty]{} \mathbf{v}$ so that $\mathbf{f}(\mathbf{p}) = \mathbf{v}$ defines a valid constraint on $\mathbf{p}$. Then the probabilities defined previously go towards zero:

$$\frac{W^{(k)}}{|S^{(k)}|} \xrightarrow[k\to+\infty]{} 0$$

and the cardinalities alone go towards infinity

$$W^{(k)} \xrightarrow[k\to+\infty]{} 0$$

so that it is necessary to consider an asymptotic behavior such as described through entropy to compare the likelihood of two probability distributions $\mathbf{p}$ and $\mathbf{q}$ satisfying the same constraint given by $\mathbf{f}(\mathbf{p}) = \mathbf{f}(\mathbf{q}) = \mathbf{v}$. In particular,

$$\arg\max_{\mathbf{f}(\mathbf{p})=\mathbf{v}} H(\mathbf{p}) = \lim_{k\to+\infty} \left(\frac{1}{N^{(k)}} \arg\max_{\left(N_i^{(k)}\right)_{1\leq i\leq n} \in S^{(k)}} \frac{W(N_1^{(k)}, ..., N_n^{(k)})}{|S^{(k)}|} \right).$$

The previous expression means that maximizing entropy corresponds to maximizing the likelihood of the associated partition while considering a uniform distribution on all partitions of N elements into n categories in the limit case when N goes to infinity.[6] This allows for a genotypic approach towards the use of MaxEnt models based on an argument of maximum likelihood.

Jaynes' work also explores another aspect of this asymptotic behavior through his entropy concentration theorem. The theorem states that, in the case where $\mathbf{f}$ is a linear function of rank $m < n$ and $\mathbf{v}^{(k)} = \mathbf{v}$, then we have asymptotically

$$2N^{(k)} \left(\max_{\mathbf{f}(\mathbf{p})=\mathbf{v}} H(\mathbf{p}) - H\left(\frac{1}{N^{(k)}}\mathbf{X}^{(\mathbf{k})}\right)\right) \sim \chi^2(n-m-1).$$

[6] Note that the formulation presented here is a slight generalization of what is presented in [Jay82] which only considers a linear constraint (i.e. $\mathbf{f}$ is a linear function) which is needed for considering the number of degrees of freedom in the concentration theorem but not necessary at this point.

One of the corollaries to this theorem is that the average partition for a given N and $\mathbf{f}$, divided by N, converges towards the MaxEnt model when N goes to infinity:

$$\frac{\mathbb{E}\left(\mathbf{X}^{(k)}\right)}{N^{(k)}} \xrightarrow[k\to+\infty]{} \operatorname*{arg\,max}_{\mathbf{f}(\mathbf{p})=\mathbf{v}} H(\mathbf{p})$$

This last expression is equivalent to our own genotypic modeling towards MaxEnt models through mutual constrained independence. In this modeling approach, which we focus on in the following chapter of this thesis, we consider that the distribution which is the least binding in terms of model hypothesis other than defined by a set of linear constraints is precisely given by $\lim_{k\to+\infty} \frac{\mathbb{E}\left(\mathbf{X}^{(k)}\right)}{N^{(k)}}$. However, the formulation in chapter 4 of the results relative to the existence and convergence of this limit (Theorem 4.1.3) and its relationship to MaxEnt models (Theorem 4.1.4), as well as their proofs, differ significantly from the presentation above and provide deeper insight on the nature of this limit.

Through this analysis, we have shown that some phenotypic approaches can lead to the same mathematical models as genotypic approaches. This should be the case if the model is uniquely defined by the model properties defined by the phenotypic modeling. Nevertheless, the genotypic approach always adds further understanding and meaningfulness to the modeling process.

3.3.4 Recommendation

In this section, we have introduced two new concepts that describe two different approaches towards modeling: phenotypic and genotypic approaches. We have shown that the modeling processes of the latter type carry much more meaning than those of the prior type. Therefore, if meaningfulness is desired, they should be preferred when applicable. Although phenotypic approaches are the only way to proceed in certain cases, such as when dealing with highly complex notions, we do not believe this is the case for objective interestingness. Hence, we recommend that a genotypic approach towards the modeling of interestingness be adopted to provide for higher meaningfulness.

3.4 Pragmatic modeling

One important aspect when opting for a modeling process, is that the mathematical model which is used in the end must be computable, at least to a certain extent, in order to provide for answers to the questions initially formulated about the system which is modeled. In a sense, one must be certain that the modeling intention meets with the practical capacities. While there is no inherent issue with this idea, we would like to underline the fact that focusing primarily on pragmatic aspects for defining modeling processes leads to issues in terms of meaningfulness.

We use the term pragmatic modeling to describe modeling processes for which the question *what can we do?* seems to come before the question *what should we do?* when justifying their utilization. If pragmatic modeling can provide answers, these are not necessarily useful because they are answers to a question which was formulated in order to be able to provide for such answers. This is a widely known issue in statistics, often phrased simply as *the right answer to the wrong question.* A linear regression model, for example, is always the right answer to the question of finding the linear model which fits the data best but this question might be completely irrelevant and meaningless with regards to the data. John Tukey, a statistician who is considered to have been one of the pioneers in the development of exploratory data analysis and the establishment of principles for statistical practice [Tuk62, Tuk77, Tuk80], summarized this particular principle in [Tuk62] as such:

> *Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.*

3.4.1 Pragmatic modeling of interestingness

The frequent itemset mining problem and the association rule mining problem presented in [AIS93] (see sections 2.2.1.1 and 2.2.1.2 for more details) represent two typical cases of pragmatic modeling. In both cases, the seemingly vague question of finding interesting associations or interesting rules between associations in data are modeled by the very precise questions of finding all frequent items or all strong association rules. In both cases, the main motivation behind the modeling is the monotonicity properties of the support and confidence measures which provide a basis for the efficient mining of frequent itemsets and strong association rules. In other words, the Apriori algorithm is the main motivation for the definition of the itemset and association rule models. The meaningfulness and the adequacy of these modeling choices are secondary considerations at best.

With retrospect, the fact that pragmatic modeling and its consequences were not systematically and explicitly pinpointed as an issue here had a larger impact than one might imagine at first. Indeed, the popularity of frequent itemset and association rule mining led many researchers to focus on efficiently computing (see section 2.2.2) or perfectly summarizing (see section 2.3.2) what was simply the right answer to the wrong question in most applied cases.

Pragmatic modeling in the field of pattern mining is not limited to frequent itemset mining and association rule mining. As described in section 2.3.1.2, measures can be chosen to model interestingness based on the facility to compute them or their algorithmic properties with little regard to their meaningfulness.

3.4.2 Meaningfulness first, computability second

It is important to note that trying to define computable models is not in itself an issue. The issue only arises when meaningfulness comes after computability. Following John Tukey's principle, if a meaningful model is defined and that it is not directly computable, then it is better to try to approximate the model than to entirely redefine the problem statement. Examples of such an approach may be found in the scientific literature on interesting pattern mining specifically in papers related to the *mining through compression* paradigm (see section 2.3.3.4). Indeed, the proponents of this approach defend that the most interesting set of patterns in a dataset are those that can be used to provide the most concise lossless compression of the data, the size of such an optimal compression being modeled through Kolmogorov complexity. As such, the modeling of interestingness is highly meaningful. However, there is a serious computability issue with this modeling. As a matter of fact, it is a proven result in algorithmic information theory that Kolmogorov complexity is uncomputable: no program can compute the Kolmogorov complexity for all possible datasets [LV08]. Hence, Kolmogorov complexity is approximated using other notions such as the minimum description length [GMP05].

3.4.3 Recommendation

Similarly to the popular saying that states that *just because you can, doesn't mean you should*, just because a model is computable doesn't mean it carries any meaning. If meaningfulness is a criteria, a modeling process for objective interestingness should start by considering a truly meaningful approach towards modeling objective interestingness regardless of any computational aspects. This might lead to the case in which computing the solutions to the mathematical problems associated with the model is technically or theoretically infeasible. In such a case, means for approximating these solutions in a computable way should be envisaged.

3.5 Patchwork and holistic modeling processes

In this section we address the issue of meaningfully and consistently connecting the various parts that compose a general modeling process. As was the case for the issues described in the previous sections, we feel that there is no adequate preexisting terminology in order to describe this particular modeling issue. Therefore, we introduce two concepts which describe two opposing approaches towards mathematical modeling. On the one hand, we consider *holistic modeling* and, on the other hand, *patchwork modeling*. Holistic modeling describes a general modeling process in which a single mathematical model englobes every particular aspect of the world considered within the general modeling process. Patchwork modeling describes a general modeling process in which multiple mathematical models are used for modeling the different aspects of the world which is considered within the general modeling process. The use of the terminology holistic is based on the notion conveyed in holism that the whole is not equivalent to the sum of its parts. By contrast, a patchwork is constructed precisely as the sum of different parts.

Patchwork modelings can be decomposed into two steps. The first step consists in modeling the world by its different aspects. We call this step the projection modeling as, in a sense, it corresponds to the projection of the world onto each of the different aspects considered. Note that this step can be lossy (if not all aspects of the world are modeled) as well as redundant (if some elements of the world are described in several aspects of the projection modeling). The second step consists in the mathematical modelings for each of the local projections considered through the projection modeling. Together these two steps make up a general patchwork modeling process as illustrated in Figure 3.3.

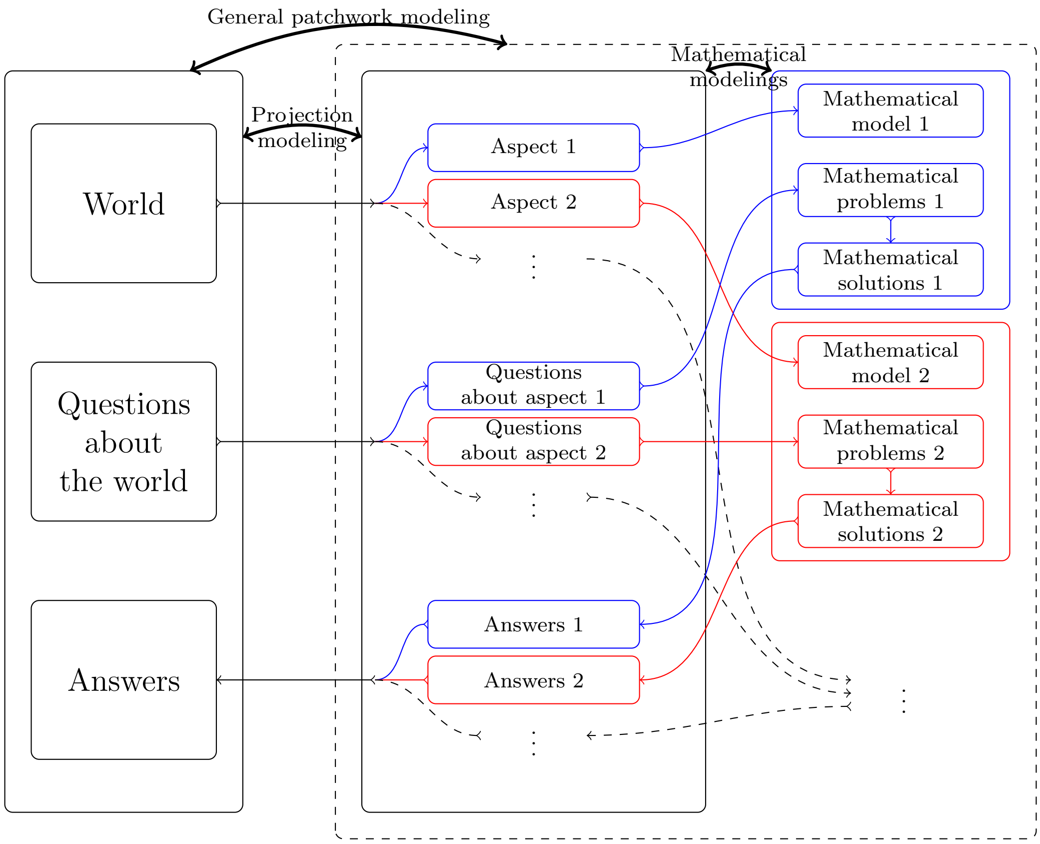

Figure 3.3: Patchwork modeling

If a truly meaningful explanation of a complex modeling process can only be provided for mathematical modeling processes, then holistic modeling is the only approach that allows for a meaningful explanation of the general modeling process. Indeed, in a patchwork modeling approach the general modeling process is not a mathematical modeling process. Even if each part of the world is modeled by mathematical models, the world itself is modeled by this patchwork of mathematical models, which is not a priori a mathematical model. Nevertheless, the fact that a patchwork model is not described through a mathematical modeling process does not necessarily imply that such a description cannot be given.

In some cases, a general mathematical layer is omitted simply because defining it feels unnecessarily tedious. We note this category of modeling processes PW0 (for patchwork type 0). However, in many cases, it is impossible to define a general mathematical modeling that is both meaningful and consistent with each local mathematical modeling within the general modeling process. We note PW1 the category of modeling processes for which a consistent yet meaningless general mathematical modeling may be defined, at least meaningless in the sense that it carries no further meaning than being simply the junction of the different models from the projection step in the modeling process. Conversely, we note PW2 the category in which consistency is an issue. PW2 corresponds to cases in which the different aspects in the projection are not entirely disjoint and one of the elements of the world, associated to two different aspects in the projection, is modeled by two different mathematical objects rendering a general mathematical model inconsistent. In other cases, this general mathematical modeling layer is absent because the current status of mathematical knowledge does not allow for its definition. We note PW3 this last category.

To each of these different categories correspond different modeling issues. For a modeling process in the PW0 category, this may simply correspond to a case of implicit holistic modeling and the general mathematical modeling layer is not described because its definition is obvious. While this may well be the case, we still recommend that this be made explicit because what may seem obvious at first is not necessarily obvious when formalized. Omitting the general mathematical modeling may in fact hide issues related to the other PW1, PW2 and PW3 categories described above. Modeling processes pertaining to the PW1 and PW2 categories present irresolvable issues towards the definition of a meaningful holistic mathematical modeling. We will exhibit how some of the modeling processes used in frequency-based interesting pattern mining belong to these categories.

The PW3 category corresponds to modeling processes for which we do not know yet how or even if it is possible to define a holistic mathematical model that is proven to be consistent with the local mathematical models within the general modeling process. In such a case, the modeling process cannot be meaningfully explained because we do not know exactly why it should be a good modeling process or not. Although we did not encounter modeling processes in the frequency-based itemset mining literature which belonged to this category, we believe it is still worth mentioning here because it is linked to the main debate surrounding the notion of meaningfulness in the field of artificial intelligence: the possibility to meaningfully explain deep learning algorithms. Indeed, the local mathematical models that compose deep learning algorithms such as artificial neurons are easily understood and explained. Furthermore, in the case of single layer neural networks, the relationship between the artificial neurons and the general network can be meaningfully explained because we have sufficient mathematical knowledge on the convergence of such models [Nov63]. However, there are no sufficiently general mathematical results known as of today which prove that multilayer neural networks converge towards mathematical solutions to the mathematical problems which model the real life questions that these algorithms aim at answering (even though it must be admitted that a number of intermediary theoretical results have been proven quite recently [APVZ14, LY17, ACGH18, AZLS18, BHL19, CJRR21]). This means that we can observe that deep learning algorithms work but we cannot really explain why.

Note that holistic modeling cannot be reduced to a vain pursuit of mathematical formalism. Well defined holistic mathematical modeling are often behind the great paradigm shifts in science and technology. Considering the World Wide Web as a single oriented graph with web pages as nodes and hyperlinks as vertices is what allowed Sergei Brin and Larry Page to define the PageRank algorithm [BP98]. Thanks to this holistic modeling they understood that the general structure of the web pointed preferentially towards certain pages and their algorithm revolutionized the world of search engines.

3.5.1 Patchwork modeling in interesting pattern mining

3.5.1.1 Type 1 patchwork modelings (PW1)

Phenotypic modeling approaches, as those we have described in section 3.3, may often be associated to the PW1 category. Indeed, a phenotypic modeling approach is an attempt to define a holistic mathematical modeling which matches with all the local mathematical models arising from the projection step and that correspond to characteristics of the world. As we have described previously, this process does not necessarily allow for the general modeling process to carry any larger meaning than the description of the individual characteristics of the subject of the modeling. In fact, the only case in which phenotypic modelings are not akin to PW1 modelings is when the characteristics fully and uniquely characterize the subject of the modeling as is the case with the modelings involving entropy described in section 3.3.3.

Other examples may include the use of multiple objective interestingness measures. For example, considering that interesting rules are defined by a support greater than 25%, a confidence greater than 95% and a lift larger than 1.5 corresponds to finding a rule $a \to b$ such that the vector $(f_{a\wedge b}, \frac{f_{a\wedge b}}{f_a}, \frac{f_{a\wedge b}}{f_a f_b})$ is greater than $(.25, .95, 1.5)$ for the partial product ordering and this carries no more meaning than the sum of the individual models given by $f_{a\wedge b} \geq .25$, $\frac{f_{a\wedge b}}{f_a} \geq .95$ and $\frac{f_{a\wedge b}}{f_a f_b} \geq 1.5$.

3.5.1.2 Type 2 patchwork modelings (PW2)

We have noted a high frequency of issues related to PW2 modelings when considering local models for redundancy (see section 2.3.3.1). To illustrate this we give two examples: one based on local independence models for rule mining; and one based on local MaxEnt models for itemset mining.

Consider two rules between itemsets $a_1 \to a_2$ and $a_1 \wedge a_2 \to a_3$ and a modeling for interestingness based on the local independence model between the antecedent and the consequent of rules. This means that we consider implicitly at least that the data corresponding to items a and b is generated by a random variable $\mathbf{X} = (X_1, X_2)$ such that $p_{a_1} = f_{a_1}$, $p_{a_2} = f_{a_2}$ and $p_{a_1\wedge a_2} = p_{a_1}p_{a_2} = f_{a_1}f_{a_2}$ for $a_1 \to a_2$; and that the data corresponding to items a_1, a_2 and a_3 is generated by another random variable $\mathbf{X}' = (X_1', X_2', X_3')$ such that $p_{a_1\wedge a_2} = f_{a_1\wedge a_2}$, $p_{a_3} = f_{a_3}$ and $p_{a_1\wedge a_2\wedge a_3} = p_{a_1\wedge a_2}p_{a_3} = f_{a_1\wedge a_2}f_{a_3}$ for $a_1 \wedge a_2 \to a_3$. If $f_{a_1\wedge a_2} \neq f_{a_1}f_{a_2}$, which is generally the case, these two models are not consistent. They cannot be seen as local projections of a global model in which the data is considered to be generated by a random variable $\mathbf{X} = (X_1, ..., X_m)$. Note that an approach based on mutual local independence models for itemsets does not, however, share this issue as such local models are local projections of the global mutual independence model.

We will now consider an example in which the use of local MaxEnt models also leads to a case of global inconsistency. Consider four items a_1, a_2, a_3, a_4 and the four local MaxEnt models for each of the itemsets of size 3, defined by the frequencies of all itemsets of size 2 or less. Then these models are not necessarily globally consistent. Indeed, not only are they generally not a projection of the global MaxEnt model for four items defined by all itemsets of size 2 or less (unless there is a specific pattern of independence), but they might not even be the projection of any possible probability distribution model. We give one specific counter-example supporting this last statement.

Consider an empirical distribution satisfying the conditions given by the table 3.1 below. Note that these are valid constraints for an empirical distribution as the dataset for which the empirical distribution is exactly defined by table 3.2 satisfy these conditions. Using the four local MaxEnt models for the

Itemset	Frequency
a_1	.330
a_2	.680
a_3	.558
a_4	.613
$a_1 \wedge a_2$	.222
$a_1 \wedge a_3$	.133
$a_1 \wedge a_4$	.157
$a_2 \wedge a_3$	.277
$a_2 \wedge a_4$	.360
$a_3 \wedge a_4$	.269

Table 3.1: A set of conditions on itemset frequencies

Minimal generators[7]	Absolute frequency
ω_0	5
ω_1	3
ω_2	5
ω_3	199
ω_4	9
ω_5	228
ω_6	195
ω_7	26
ω_8	15
ω_9	16
ω_{10}	42
ω_{11}	35
ω_{12}	69
ω_{13}	97
ω_{14}	47
ω_{15}	9

Table 3.2: A corresponding empirical distribution

[7]See section 3.6.2.5 for an explanation of the notation used in Table 3.2.

itemsets of size 3 defined by the frequencies of the itemsets of size 2 or less we obtain the probabilities in table 3.3.

Itemset	Probability
$a_1 \wedge a_2 \wedge a_3$	.0048564
$a_1 \wedge a_2 \wedge a_4$	.0083261
$a_1 \wedge a_3 \wedge a_4$	.0030280
$a_2 \wedge a_3 \wedge a_4$	.0054097

Table 3.3: Probabilities defined by the local MaxEnt models

However, we see that the following condition given by the inclusion-exclusion principle:

$$p_{a_1\wedge a_2\wedge a_3\wedge a_4} \leq p_{a_1} - p_{a_1\wedge a_2} - p_{a_1\wedge a_3} - p_{a_1\wedge a_4} + p_{a_1\wedge a_2\wedge a_3} + p_{a_1\wedge a_2\wedge a_4} + p_{a_1\wedge a_3\wedge a_4}$$

does not allow to define a globally consistent model as we have:

$$p_{a_1} - p_{a_1\wedge a_2} - p_{a_1\wedge a_3} - p_{a_1\wedge a_4} + p_{a_1\wedge a_2\wedge a_3} + p_{a_1\wedge a_2\wedge a_4} + p_{a_1\wedge a_3\wedge a_4} = -.0019895$$

which would imply a negative value for $p_{a_1\wedge a_2\wedge a_3\wedge a_4}$.

Note that such counter-examples correspond to skewed distributions. Indeed, in order to present such a case here, we randomly generated one million hypothetical empirical distributions $\mathbf{f} = (f_i)_{0\leq i\leq 15}$ by randomly picking 16 integers between 1 and 100 and dividing them by there sum. For each of these distributions, we determined the values for the probabilities of the itemsets of size 3 given by the local MaxEnt models defined by the frequencies of itemsets of size 2 or less. We then proceeded by checking whether the values of the frequencies of itemsets of size 2 or less together with the probabilities of the itemsets of size 3 defined by the local MaxEnt models where globally consistent using the necessary and sufficient conditions given by the inclusion-exclusion principle (see for example Figure 1 in [CG07]). Out of the million distributions that were generated, only forty were not globally consistent, each of which corresponded to skewed distributions. As such distributions correspond to structured data, they are the most relevant in interesting pattern mining. Therefore, this reveals one of the important limits to a naive use of local MaxEnt models for interesting pattern mining.

From the example which precedes, we can say that local MaxEnt models based on a global set of constraints do not necessarily provide for a globally consistent model. However, this does not mean that the result holds for every set of constraints. Indeed, the MaxEnt model given by the set of constraints defined by the frequencies of items is the mutual independence model and we know that, in this case, the local models are projections of the global model. Hence, there are sets of constraints that do not always allow for consistency between local models and sets of constraints that do. Nevertheless, and this is simply a conjecture, we believe that the latter are quite the exception and that only a negligible fraction of the 2^d constraints defined by sets of itemsets always allow for global consistency between local models as in the case of mutual independence.

3.5.2 Recommendation

In this section, we have introduced two new concepts to qualify two opposing modeling approaches: patchwork and holistic modeling. We have shown that patchwork modeling can lead to meaningless or inconsistent mathematical models. Even when this is not the case, meaningfulness arises when making explicit an implicit holistic model. If meaningfulness is a main criteria, holistic modelings should be preferred. This does not imply that local projections of a holistic model should never be considered, but only that they should be considered as such. In other words, the definition of any local model should stem from the prior definition of a global model in order to be able to justify the meaningfulness of the general modeling process.

3.6 Mathematical modeling of patterns

In this section, we focus on the structure of the mathematical model for describing the patterns which are extracted in the mining process. In the standard itemset model, the patterns are sets of items (a.k.a. itemsets) and the set of all possible patterns is simply described as the set of all itemsets with its natural lattice structure (see section 2.2.1.1). This model, still currently very much in use in the data mining community, is a heritage from the early presentation of itemset mining as a tool for market basket analysis by Agrawal et al. in [AIS93]. If it is quite simple to apprehend, it is also quite limited and limiting with regards to the possibilities that it allows in terms of mathematical modeling. In fact, it can be considered a rather poor modeling choice for studying the fundamental measure in itemset mining: frequency[8]. Indeed, some of the most basic properties of frequency, for example those based on the inclusion-exclusion principle, are much more naturally understood if a more general structure is considered, which is why generalized itemsets were introduced in this specific example (see section 2.3.2.3).

At this point, it is important to recall that, in this thesis, we are addressing specific issues in frequency-based pattern mining. This means that frequency is at the core of any mining process, which might not necessarily be the case in other domains of pattern mining such as pattern recognition. In fact, the patterns which are considered are usually objects together with their frequencies. If we sometimes use the term pattern to refer to an object regardless of its frequency, as in an itemset is a pattern, then it is generally a misnomer. Note that this can also be seen as part of Agrawal's legacy. Indeed, the phrasing of the frequent itemset mining problem suggests that the set of frequent itemsets alone, rather than the set of frequent itemsets together with their frequencies, is interesting. However, it is now widely acknowledged in the literature, whether explicitly or implicitly, that an object must always be considered together with its frequency.

3.6.1 Measure spaces and Boolean lattices

The natural mathematical structure for considering both itemsets and their frequencies is the measurable space generated by the set of all itemsets fitted with the empirical distribution in the dataset. For this reason and those we will present in section 3.6.2, this model is by far a more satisfying choice than the simple set model.

In order to consider frequency as a measure (in the sense of mathematical measure theory), terminology must first be explicitly specified as the standard usage in both itemset mining and measure theory may reflect opposite ideas. Indeed, in itemset mining, focus is set on itemsets. Hence, the frequency of a given itemset X is noted as defined by X, for example f_X. The frequency of a union of itemsets $X = X_1 \cup X_2$ can therefore be noted $f_{X_1 \cup X_2}$. By contrast, in measure theory the focus is set on measurable sets (i.e. the equivalent of events in probability theory), which can be equated to tidsets in the itemset context.

[8] We prefer the term *frequency* to the term *support*, which are synonyms, as the term frequency is more commonly used in the broader statistical literature.

Hence, if A, A_1 and A_2 are the measurable sets associated to X, X_1 and X_2 respectively and $X = X_1 \cup X_2$, then $A = A_1 \cap A_2$. Therefore, if the same notations are used for itemsets and their associated measurable sets (which is not uncommon), the notation $f_{X_1 \cap X_2}$ in the context of measure theory might very well be used to designate the same notion as the notation $f_{X_1 \cup X_2}$ in the context of itemset mining.

One approach which allows to remove all ambiguity is to consider the natural Boolean lattice structure of the measure space considered here. In this case, we can associate an itemset $X = \{x_1, ..., x_k\}$ to a propositional logic formula $X = x_1 \wedge ... \wedge x_k$ using the same notations for itemsets and their corresponding formulas, as well as items and their corresponding atomic formulas, without cause for concern. Using this representation, the notation for the frequency of the unioned itemset $X_1 \cup X_2$ becomes $f_{X_1 \wedge X_2}$. The Boolean lattice model corresponds to identifying the measurable sets in the measure space model by their corresponding propositions in the Boolean lattice. This additional modeling layer helps for removing ambiguities such as described above and is also generally quite suited to the context of frequency-based itemset mining as we show in section 3.6.2.

3.6.2 Benefits of the measure space and Boolean lattice models

We present here some aspects in which using the measure space and Boolean lattice models for describing the pattern space can be quite beneficial to the general modeling process. The first focuses on the measure space model, the next three focus on the Boolean lattice model, and the last addresses benefits of the combined approach. We also question the relevance of the rule mining paradigm in light of these explanations.

3.6.2.1 Modeling the dataset using a random variable

Previously in this thesis, we have established that the dataset can be adequately modeled as the result of the random sampling of n independent identically distributed random variables with values in $\{0, 1\}^m$. Modeling the patterns in the data as measurable sets from a measure space provides a context for transferring information on the patterns in order to define the probability distribution of the random variables because measure spaces and probability spaces have exactly the same structure (a probability space is simply a measure space such that the measure of the entire space equals one). Indeed, a distribution for the random variable can be defined by a certain number of constraints corresponding to observed patterns in the data together with a method for defining a unique probability distribution based on these constraints (using the MaxEnt principle or junction trees for example).

One may then wonder why we do not simply choose to consider two probability spaces rather than a measure space for the patterns to extract and a probability space for the random variable. This is because there is nothing probabilistic in the observation of the patterns in the data, probabilities are only used here to describe our ignorance about the general process that generated the data.

3.6.2.2 Pattern diversity

One of the major advantages of the Boolean lattice is that it allows to consider much more patterns than just itemsets, without the need to add an extra layer of modeling each and every time a new type of pattern needs to be considered. The Boolean lattice contains propositions which correspond to itemsets and negative itemsets. It also contains propositions corresponding to the 3^m generalized itemsets, but more generally still it contains exactly all of the 2^{2^m} disjunctions of the 2^m minimal generators of the Boolean lattice (which correspond to the generalized itemsets of size m). Of course this is too large to hope to ever consider all possible types of patterns. We will explain further in this section how we can choose to constrain the mining process to a smaller family of patterns. However, considering this huge space is necessary if we want to include the majority of the great diversity of patterns which we have encountered in the literature.

For example, the Boolean lattice contains all logical implications between itemsets which are the core patterns of statistical implicative analysis [GRMG13]. Another example is given by the four different constraints used as possible background knowledge in [TM10]: column margins, row margins, lazarus counts and transaction bounds. Indeed, each of these constraints can be associated to a specific set of elements of the Boolean lattice. While column margins correspond quite trivially to the set of items, which are the simplest propositions in the Boolean lattice, the other constraints are described using much more complex propositions. The row margins constraint is associated to the following set of the $m+1$ propositions:

$$\bigvee_{I \in \binom{[\![1,m]\!]}{k}} \left[\left(\bigwedge_{i \in I} x_i \right) \wedge \left(\bigwedge_{i \notin I} \neg x_i \right) \right]$$

such that $k \in [\![0, m]\!]$. Similarly, the lazarus counts constraint, where a lazarus count gives the number of zeros in between two ones in a transaction, is associated to the set of all propositions:

$$\bigvee_{1 \leq a \leq b \leq m} \left[\left(\bigwedge_{i<a} \neg x_i \right) \wedge x_a \wedge \left(\bigvee_{I \in \binom{]\!]a,b[\![}{k}} \left[\left(\bigwedge_{i \in I} x_i \right) \wedge \left(\bigwedge_{i \notin I} \neg x_i \right) \right] \right) \wedge x_b \wedge \left(\bigwedge_{b<i} \neg x_i \right) \right]$$

where $k \in [\![0, m-2]\!]$. And finally, the transaction bounds constraint, where bounds correspond to the first and last positions for a one in a transaction, is associated to the set of propositions:

$$\left(\bigwedge_{i<a} \neg x_i \right) \wedge x_a \wedge x_b \wedge \left(\bigwedge_{b<i} \neg x_i \right)$$

where $1 \leq a \leq b \leq m$ or $a = b = 0$.

3.6.2.3 Pattern complexity

One of the important issues when considering the interestingness of a pattern is its complexity. Indeed, if a single pattern allows to define a model that closely fits the data but the description of the pattern itself is impossible to apprehend then the pattern should not be considered interesting. While this issue is widely acknowledged in the field of pattern mining, it has been rarely addressed both theoretically and objectively, with the notable exclusion of the research conducted within the mining as compression paradigm. Indeed, the fact that a large portion of the literature focuses on rule mining is mainly motivated by the fact humans, among other intelligent systems, easily interpret rules. This is however a very subjective approach and, to define the notion of complexity objectively, one requires an objective description of the language in which the patterns are expressed.

The Boolean lattice model provides for a good basis for such a description. In fact, the first-order languages which derive from such structures have been extensively studied in formal language theory [Rau06]. Moreover, the study of the synonyms in such languages, that is logically equivalent propositions, is by itself an important field of research [HS06, MT12, Vin13]. Indeed, each individual element in the Boolean lattice corresponds to an infinite number of propositional logic formulas each of which have their own complexity. For a given propositional logic formula, the task of transforming it into a synonym of a given type to reduce its complexity is a well known problem in computer science often referred to as the *minimum equivalent expression problem* [Uma01, BU11].

Note that this is, in itself, a highly complex issue and beyond the scope of this thesis. However, I have conducted research in this specific area during the time of my doctoral thesis and published a paper analyzing the complexity of different conjunctive normal form encodings of cardinality constraints for Boolean satisfiability problems [Del18]. Such cardinality constraints are involved in particular in the definition of the row margins constraint as an element of the Boolean lattice. While the row margins constraint is easily apprehensible by a human, its corresponding proposition is too long (and irreducibly so in the language deriving from the Boolean lattice) for it to be effectively used by a computer program. To be more precise, the constraint is concisely expressed in second-order logic but has a lengthy expression in first-order logic. As the state-of-the-art of propositional satisfiability solvers operate in first-order logic (even though there have been some recent attempts at defining second-order SAT solvers [DKL14]), problems expressed in second-order logic are still generally transposed into first-order logic when considered by a computer program. In this case, one must rely on an extension of the language (i.e. the addition of new Boolean variables together with a characterizing propositional formula known as the *encoding*) in order to be able to reduce the complexity of the proposition. This example, among other aspects which will be addressed further in this chapter, supports the idea which we defend that no artificial intelligence may be developed to gain substantial objective and interesting information about the world unless it is capable of complexifying the native language which it uses to describe the world (by resorting to new variables and encodings in first-order logic or defining propositions in second-order logic). This is exactly what allows humans to consider that certain patterns, which are inherently more complex than other patterns as is the case with the row margins constraint, are still more interpretable than simpler patterns.

3.6.2.4 Type diversity

Up to this point, we have only considered that the variables in the data are binary. Though we will maintain this view in the rest of this thesis, we briefly present here possible generalizations to include other types of variables which can be easily integrated within the Boolean lattice model. The first type consists of categorical variables. A categorical variable has a natural representation using Boolean variables and a constraint. Indeed, if a is a variable which can take any one of p values $v_1, ..., v_p$, then we can consider p variables $a_1, ..., a_p$ together with the cardinality constraint that exactly one of these variables is true. This cardinality constraint can be given optimally by its naive conjunctive normal form:

$$\left(\bigvee_{1 \leq i \leq p} a_i\right) \wedge \left(\bigwedge_{1 \leq i < j \leq p} (\neg a_i \vee \neg a_j)\right)$$

for $p \leq 5$ or using a more elaborate encoding for larger values of p (see [Del18]). Similarly, an ordinal variable a taking values in $[\![0, p]\!]$ can be replaced by $p + 1$ binary variables $a_0, ..., a_p$ together with the CNF constraint that:

$$\left(\bigvee_{0 \leq i \leq p} a_i\right) \wedge \left(\bigwedge_{1 \leq i \leq p} (a_{i-1} \vee \neg a_i)\right)$$

On another level, fuzzy logic may be used as a generalization of Boolean logic to consider numeric variables which correspond to fuzzy data. Fuzzy approaches have been considered in itemset mining [DMSV03, DHP06] and, in the general modeling process which we describe, a fuzzy modeling layer may be considered without raising any theoretical issues.

3.6.2.5 Sound and complete families of patterns

As we have mentioned previously, the size of the Boolean lattice is extremely large and the mining process must therefore be constrained to certain types of propositions within the lattice. In order to determine which type of patterns to mine for, one possible criteria is that the family of patterns considered allows for a *complete* and *sound* description of the frequency measure. This means that any frequency measure is uniquely defined on the entire measure space by the frequencies of the patterns in the family (*completeness*) and that no subfamily of the family holds this property (*soundness*). We will consider three families of patterns which satisfy these conditions: minimal generators; itemsets; and implications between complementary itemsets. In order to prove this result, we introduce a few elements of notation and definitions.

Consider m items $a_1, ..., a_m$ and the associated Boolean lattice $\mathcal{B}$. We give the following definitions for completeness and soundness.

Definition 3.6.1 (Completeness)**.** Let $\mathcal{F}$ be an ordered subset of $\mathcal{B}$. Then $\mathcal{F}$ is a complete family of patterns if, and only if, for any two measures μ_1 and μ_2 defined on $\mathcal{B}$, $(\forall P \in \mathcal{F},\ \mu_1(P) = \mu_2(P)) \implies (\mu_1 = \mu_2)$.

Definition 3.6.2 (Soundness)**.** $\mathcal{F}$ is a sound family of patterns if, and only if, for every subset $\mathcal{F}' \subsetneq \mathcal{F}$, there exists two measures μ' and μ defined on $\mathcal{B}$ such that $\forall P \in \mathcal{F}',\ \mu'(P) = \mu(P)$ and $\exists P \in \mathcal{F},\ \mu'(P) \neq \mu(P)$.

We note Ω the set of minimal generators of $\mathcal{B}$ defined by:

$$\Omega = \left\{ \left(\left(\bigwedge_{i \in A} a_i \right) \wedge \left(\bigwedge_{i \notin A} \neg a_i \right) \right) \in \mathcal{B} \;\middle|\; A \subset [\![1, m]\!] \right\}$$

For conciseness, we note $d = 2^m - 1$. We consider the natural lexicographic order on Ω given by the following sequence of increasing bijections:

$$\begin{array}{ccccc} [\![0, d]\!] & \to & \{0,1\}^m & \to & \Omega \\ k & \mapsto & \mathbf{k} & \mapsto & \omega_k \end{array}$$

where $\mathbf{k}$ is the binary representation of k as a tuple in $\{0,1\}^m$ and $\omega_k = \left(\bigwedge_{i \in A_k} a_i \right) \wedge \left(\bigwedge_{i \notin A_k} \neg a_i \right)$ such that $A_k = \{i \in [\![1, m]\!] \mid \mathbf{k}_i = 1\}$. Which gives, for example:

$$\omega_0 = \bigwedge_{1 \le i \le m} \neg a_i \ , \ \ \omega_1 = \left(\bigwedge_{1 \le i < m} a_i \right) \wedge a_m \ \text{ and } \ \omega_d = \bigwedge_{1 \le i \le m} a_i$$

Furthermore, we note $\mathcal{I}$ the set of all itemsets of $\mathcal{B}$:

$$\mathcal{I} = \left\{ \left(\bigwedge_{i \in A} a_i \right) \in \mathcal{B} \;\middle|\; A \subset [\![1, m]\!] \right\}$$

Similarly, we consider the natural lexicographic order on $\mathcal{I}$ given by the following sequence of increasing bijections:

$$\begin{array}{ccccc} [\![0, d]\!] & \rightarrow & \{0,1\}^m & \rightarrow & \mathcal{I} \\ k & \mapsto & \mathbf{k} & \mapsto & I_k \end{array}$$

where $I_k = \bigwedge\limits_{i \in A_k} a_i$ and both $\mathbf{k}$ and A_k are defined as previously. Which gives here:

$$I_0 = \top \ , \ \ I_1 = a_m \ \text{ and } \ I_d = \bigwedge_{1 \leq i \leq m} a_i$$

Lastly, we note $\mathcal{R}$ the set of all implications between complementary itemsets defined by:

$$\mathcal{R} = \left\{ \left(\left(\bigwedge_{i \in A} a_i \right) \implies \left(\bigwedge_{i \notin A} a_i \right) \right) \in \mathcal{B} \;\middle|\; A \subset [\![1, m]\!] \right\}$$

Again, we consider the natural lexicographic order on $\mathcal{R}$ given by the following sequence of increasing bijections:

$$\begin{array}{ccccc} [\![0, d]\!] & \rightarrow & \{0,1\}^m & \rightarrow & \mathcal{R} \\ k & \mapsto & \mathbf{k} & \mapsto & R_k \end{array}$$

where $R_k = \left(\bigwedge\limits_{i \in A_k} a_i \right) \implies \left(\bigwedge\limits_{i \notin A_k} a_i \right)$ with the same notations as above. In which case, we have:

$$R_0 = \bigwedge_{1 \leq i \leq m} a_i \ , \ \ R_1 = \left(\bigwedge_{1 \leq i < m} a_i \right) \implies a_m \ \text{ and } \ R_d = \top$$

Notice that, for any pattern $P \in \mathcal{B}$, there is a unique minimal decomposition as a disjunction of elements in Ω and we can therefore associate this decomposition to a unique vector in $\{0,1\}^{d+1}$ representing P in the basis Ω (given an order on Ω). Hence, any family of patterns $\mathcal{F} = (P_i)_{0 \leq i \leq l}$ can be represented by a unique binary $(d+1)\times l$ matrix. The family is complete if and only if the matrix is surjective and sound if and only if it is injective. Hence, $\mathcal{F}$ is a complete and sound family of patterns if and only if its corresponding matrix representation in base Ω is invertible.

Proposition 3.6.1. *Ω, $\mathcal{I}$ and $\mathcal{R}$ are sound and complete families of patterns in $\mathcal{B}$.*

This result is trivial for Ω by definition as they are minimal disjoint generators of $\mathcal{B}$. It is less trivial but still common knowledge for $\mathcal{I}$ and can be demonstrated using the exclusion-inclusion principle for defining the matrix representation of $\mathcal{I}$ in Ω. We give the proof for $\mathcal{R}$ that is based on the following lemma (which incidentally also provides a proof for $\mathcal{I}$ without having to define the corresponding matrix).

Lemma 3.6.1. *Consider $\mathcal{S} = (S_k)_{0\leq k\leq d}$ a family of patterns in $\mathcal{B}$ defined by $S_0 = R_d$, $S_k = \neg R_k$ for all $k \in [\![1, d-1]\!]$ and $S_d = R_0$. Then, $\forall k \in [\![0, d]\!]$,*

$$\left(\bigvee_{k\leq i\leq d} \omega_i\right) = \left(\bigvee_{k\leq i\leq d} S_i\right) = \left(\bigvee_{k\leq i\leq d} I_i\right)$$

Proof of the lemma. For all $k \in [\![1, d-1]\!]$,

$$\begin{aligned}
S_k &= \neg\left(\left(\bigwedge_{i\in A_k} a_i\right) \implies \left(\bigwedge_{i\notin A_k} a_i\right)\right)\\
&= \neg\left(\neg\left(\bigwedge_{i\in A_k} a_i\right) \vee \left(\bigwedge_{i\notin A_k} a_i\right)\right).\\
&= \left(\bigwedge_{i\in A_k} a_i\right) \wedge \left(\bigvee_{i\notin A_k} \neg a_i\right)
\end{aligned}$$

$\omega_k = \left(\bigwedge_{i\in A_k} a_i\right) \wedge \left(\bigwedge_{i\notin A_k} \neg a_i\right)$ and $I_k = \left(\bigwedge_{i\in A_k} a_i\right)$, this gives:

$$\forall j \in [\![1, d-1]\!], \quad \omega_k \implies S_k \implies I_k$$

Moreover, $\omega_d = S_d = I_d = \bigwedge_{1\leq i\leq m} a_i$. Therefore,

$$\forall k \in [\![1, d]\!], \quad \left(\bigvee_{k\leq i\leq d} \omega_i\right) \implies \left(\bigvee_{k\leq i\leq d} S_i\right) \implies \left(\bigvee_{k\leq i\leq d} I_i\right)$$

Furthermore, as $\bigvee_{0 \leq i \leq d} \omega_i = S_0 = I_0 = \top$, the previous relationship is still true for all $j \in [\![0, d]\!]$.

Now consider the sequences of implications:

$$\omega_d \implies (\omega_{d-1} \vee \omega_d) \implies ... \implies \left(\bigvee_{k < i \leq d} \omega_i \right) \implies \left(\bigvee_{k \leq i \leq d} \omega_i \right) \implies ... \implies \top$$

and

$$I_d \implies (I_{d-1} \vee I_d) \implies ... \implies \left(\bigvee_{k < i \leq d} I_i \right) \implies \left(\bigvee_{k \leq i \leq d} I_i \right) \implies ... \implies \top$$

Both are strictly monotonic. This is trivial for the first sequence and we can see this for the second by noticing that $\omega_k \wedge \left(\bigvee_{k < i \leq d} I_i \right) = 0$ whereas $\omega_k \wedge \left(\bigvee_{k \leq i \leq d} I_i \right) = \omega_k$.

Finally, we see that both sequences have maximal length $d + 1$ (which corresponds to the number of possible interpretations in $\mathcal{B}$). Hence, they are equal, as well as the third sequence which lies in between the two. $\square$

Proof of the proposition. From the previous lemma, we can see that the matrix representations for $\mathcal{S}$ and $\mathcal{I}$ in Ω are invertible triangular and hence they are both complete and sound families of patterns. The definition of $\mathcal{S}$ from $\mathcal{R}$ implies that the latter is also complete and, given its cardinality, it is necessarily sound. This concludes the proof of the proposition. $\square$

Now that we have presented the notions of complete and sound families of patterns and given a few examples of such families, we address two questions: first the relationship of these families with regards to a random variable model for the data; and second the issue of deciding which family to choose for the mining process.

In a sense, complete and sound families allow for a full and minimal description of any measure on the measure space. In the theoretical case in which the dataset corresponds to a random sample of an infinite number of independent identically distributed random variables with values in $\{0, 1\}^m$ (which implies a dataset with infinitely many transactions) then a complete and sound family of patterns allows for a full description of the probability distribution for the random variables with but a single redundancy corresponding to the fact that probability of the entire space (corresponding to the truth value $\top$ in the Boolean lattice) is always equal to 1. An analogous notion of complete and sound families of patterns for defining the probability distribution can be given by considering, for all complete and sound family of patterns $\mathcal{F} = (P_i)_{0\leq i\leq d}$ (with regards to the definition of a measure), the subsets of size d, $\mathcal{F}_j = (P_i)_{\substack{0\leq i\leq d\\ i\neq j}}$, such that $\bigvee\limits_{i\neq j} P_i$ is not a tautology. In the case of itemsets, this corresponds necessarily to removing the empty itemset I_0, in the case of $\mathcal{R}$, to removing R_d, and in the case of Ω, to removing any one of its elements.

Now, considering the issue of picking one particular family of patterns for the mining process, what criteria can we base this decision on, knowing that they are all already minimal? First, notice that they are only minimal in the sense that they contain a minimal number of patterns, not in the sense of the complexity of the patterns themselves. For example, the description of any pattern in Ω contains m literals while an itemset can contain but a single literal. This may seem to be a large difference but it can be moderated by the fact that the total number of literals for every pattern in Ω is equal to $m2^m$ while it is equal to $\sum\limits_{i=0}^{m} \binom{m}{i} i = m2^{m-1}$ for itemsets. Even though the number is smaller in the case of itemsets, it is only twice as small as for minimal generators. As it is usually easier to represent any given pattern in $\mathcal{B}$ as a disjunction of minimal generators rather than as combination of itemsets, it is possible that a complete description of a measure is better given by its values on Ω than on $\mathcal{I}$. Note however, that the value $m2^{m-1}$ obtained for the total number of literals in $\mathcal{I}$ is the lowest possible bound for a complete and sound family of patterns in $\mathcal{B}$. Furthermore, complete descriptions are usually not what we are aiming for. Indeed, we do not have infinite datasets so a complete description of the data would not give a complete description of the random variable but rather suffer from data over-fitting. Hence considering itemsets, particularly with a level-wise approach makes sense because it allows to consider the least complex patterns first. Another approach is to consider that certain patterns are more interpretable than others, such as rule type patterns of which implications are the standard example in Boolean logic. This motivated our search for a complete and sound family of implications which lead us to the definition of $\mathcal{R}$.

We must also insist on the fact that it is not necessarily important to constrain mining processes to patterns within a sound family of patterns as long as the family of patterns mined during the process stays sound (which depends mostly on the size of m). For example, when mining for itemsets given the row margins constraint, we can consider that the we are mining for patterns in a family of $2^m + m + 1$ patterns, $m + 1$ of which are already known. While this family is complete, because it contains the complete family of itemsets, it is not sound. However, if m is very large, the mining process will never reach a point in which the itemsets mined, together with the $m + 1$ row margins constraint, do not make up for a sound family of patterns. Nevertheless, this issue may easily arise in the case of small values of m.

3.6.2.6 Rule mining

Note that throughout this section, we have hardly addressed the issue of rule mining and not, in any case, considered rules such as they are defined in association rule mining or, more generally, rules characterized by any objective interestingness measure. This is due to the fact that such rules do not correspond to any element in the Boolean lattice. To be more precise, they correspond to tuples of elements from the Boolean lattice together with their associated frequencies in the measure space.

Indeed, recall that in the standard itemset model an association rule $X \to Y$ can be defined for any two itemsets $X = \{x_1, ..., x_k\}$ and $Y = \{y_1, ..., y_l\}$ such that $X \cap Y = \emptyset$ and is characterized by its support and confidence measures which are equal to $f_{X \cup Y}$ and $\frac{f_{X \cup Y}}{f_X}$ respectively. Hence, in the measure space and Boolean lattice model, an association rule can be identified by a tuple (X, Z) such that $X, Z \in \mathcal{I} \subset \mathcal{B}$ and $Z \implies X$. In this representation, if $X = \bigwedge_{i \in A} a_i$ then $Z = \bigwedge_{i \in B} a_i$ with $A \subset B$ and $Y = \bigwedge_{i \in B \setminus A} a_i$ so that $Z = X \wedge Y$. This means that if we consider three distinct itemsets $X, Y, Z \in \mathcal{B}$ such that $Z \implies Y \implies X$ then the tuples (X, Y), (Y, Z) and (X, Z) correspond to three different association rules, but the information on the measure given by (X, Z) is contained in the information on the measure given by (X, Y) and (Y, Z).

More generally, a rule $X \to Y$ between any two patterns, characterized by one or several interestingness measures defined on the values for the frequencies in the contingency tables for X and Y, corresponds to a triple (X, Y, Z), if the rules are defined by the relative frequencies f_X, f_Y and $f_{X \wedge Y}$, or a quadruple

$(1, X, Y, Z)$, if the rules are defined by the absolute frequencies n, n_X, n_Y and $n_{X \wedge Y}$, where $Z = X \wedge Y$ in both cases. Hence, association rules, and any other type of rule defined on contingency tables even more so, correspond to naturally redundant families of patterns with regards to the definition of a measure.

This brings further insight on the issue discussed in section 3.5. Indeed, either considering rules leads to redundancy or, if there is no redundancy, it means the model is necessarily inconsistent. As meaningfulness cannot be justified in an inconsistent model, an additional modeling layer for objectively interesting rules would be at best superfluous. Therefore, one simple question is to ask whether rule mining should be included in a meaningful frequency-based objectively interesting pattern mining process. Based on Occam's razor, our simple answer is no. Focus has been set on rule mining mainly because rules are more interpretable by humans. This is a highly subjective criteria and, though it might be very useful in a number of contexts, it does not apply to the specific goal we have defined. In any case, if rule type patterns are required, a mining process can be defined to mine for patterns within a complete and sound family of implications in the Boolean lattice such as $\mathcal{R}$.

3.6.3 Recommendation

In this section, we have addressed the issue of the mathematical modeling of the patterns within the general mathematical modeling process used in objective frequency-based interesting pattern mining. From our analysis, we strongly recommend that the set of minable patterns be modeled by a measure space in which the measurable sets are identified with the elements of a Boolean lattice. Furthermore, we recommend that the mining process be restricted to a complete and sound family of patterns, such as itemsets which also constitute a reasonable option in other regards, and we disqualify the mining of rules defined as particular tuples of patterns[9].

[9]Note that these recommendations specifically aim at defining meaningful mathematical modeling in objective frequency-based interesting pattern mining and that we do not disqualify rule mining in pattern mining in general.

3.7 Modeling objectivity

In this section, we address one last aspect of mathematical modeling for objective frequency-based interesting pattern mining: the issue of including the notion of objectivity in the general modeling process. Though some may argue that interestingness is inherently subjective, we defend the idea that the aim of science is precisely to objectively provide for interesting information. Objectivity in science is obtained by conforming to the process known as the scientific method. Though there is much debate on the exact definition of the scientific method and to what extent it allows to reach scientific objectivity, we will set these issues aside in this thesis and focus mainly on one description of the scientific method: the hypothetico-deductive model. Hence, our question here is whether and how it is possible to include a modeling of scientific objectivity and, specifically, the hypothetico-deductive model within the general mathematical modeling which we use to give meaning to a pattern mining process.

The scientific method as described by the hypothetico-deductive model is seen as a dynamic process involving a number of different steps. While the number and the nature of these steps vary from author to author, most agree on a fundamental basis for the hypothetico-deductive model consisting in the four following steps as illustrated in Figure 3.4:

1. Empirical observation;

2. Hypotheses formulation;

3. Prediction;

4. Hypotheses evaluation.

Observation of the world leads to formulating hypotheses about the world. We use these hypotheses to predict aspects of the world. The hypotheses are evaluated by comparing the predictions to new empirical observations. If the evaluation is positive (the hypotheses have not been falsified), we can continue evaluating them. Otherwise, we try to formulate new hypotheses based on further empirical observations. The hypothetico-deductive model is usually understood as a never-ending dynamic process because hypotheses may never

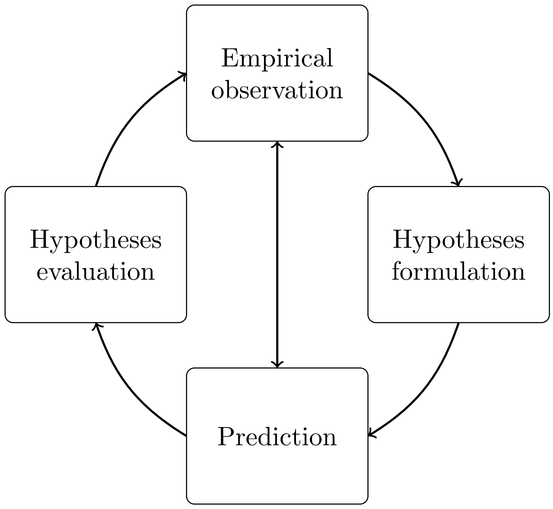

Figure 3.4: Hypothetico-deductive model for the scientific method

be entirely verified, they can only be falsified. This explains why Figure 3.4 is represented as a cycle.

Our aim is to include this representation of the scientific method into a mathematical modeling process for pattern mining. In a sense, this brings us back to the very beginning of itemset mining which can be seen to have originated in Prague in the 1960s with the GUHA method [HHC66, HH12]. Indeed, GUHA stands for General Unary Hypotheses Automaton and it was clearly intended as a tool for assisting researchers in the hypotheses formulation and evaluation steps of the hypothetico-deductive model. However, the GUHA method is essentially based on local models and, as such, suffers from a number of modeling issues which are described in this chapter (see section 3.5).

In the following, we will present all the issues regarding the modeling of the hypothetico-deductive model which we have identified, as well as corresponding solutions.

3.7.1 A static finite model for a dynamic never-ending process?

One of the first issues which arises when considering the modeling of the hypothetico-deductive model is that it is described as a dynamic never-ending process while the data which we consider is static and finite.

In the hypothetico-model, it is often argued that the empirical observations which are used to formulate hypotheses should not be the same as those which are used to evaluate those hypotheses. Indeed, if one formulates hypotheses on the basis of an observation and uses that same observation to evaluate these hypotheses, what is to say that the formulation of the hypotheses does not depend on the same aspects of the observation as the evaluation step? This issue is, in a sense, akin to the overfitting bias. As such, many of the standard methods that are used to prevent overfitting, particularly in supervised learning, could be envisaged in this case. Indeed, machine learning methods in supervised learning will often dissociate *training data* from *validation data* in a dataset, which corresponds in our case to dissociating the empirical observations on which the formulation step is based from those on which the evaluation step is based. However, this only allows to consider one cycle of the hypothetico-deductive model while the process involves many cycles.

One option is to conclude that this is simply the wrong type of data for modeling such a process and that we should use data streams (as humans do with their senses). We do acknowledge that this is a serious theoretical issue and that it is indeed impossible to consider a data mining modeling based on finite data for the hypothetico-deductive model strictly speaking. More precisely, it is impossible if the variables considered in the process (i.e. the items in itemset mining) have not been previously designated, which is the case if we consider the scientific method to produce science in general. Indeed, if the hypotheses formulated are rejected after evaluation, then the new hypotheses defined may characterize entirely new variables which had not necessarily been observed before. In fact, the formulation step involves both the definition of a set of variables and the definition of the hypotheses about these variables. As it is not possible to observe every possible variable simultaneously (no system is all-perceiving) and that we cannot know in advance which variables should be chosen (because choosing the variables is part of the scientific method itself), an objective method for designating the variables to observe between different observations must be elaborated if we hope to conceive a strong objective artificial intelligence (i.e. an artificial scientist). Though such an endeavor is eminently interesting it goes far beyond the scope of this thesis and we set it aside for now to concentrate on a more achievable goal.

This leaves us with the following question, is it possible to include the hypothetico-deductive model in a mathematical modeling based on a finite dataset when the variables considered in the process are previously designated? This last condition corresponds to the idea that we are constraining the scientific method to a certain aspect of the world which is defined and described by the variables in the data. To this question, we answer by the positive and we shall detail this throughout this section. Intuitively, this seems quite normal. Indeed, consider all the data that has been used in scientific research up to this day, it is a large yet finite amount of data and it is defined on a large yet finite number of variables. Now, consider this data as a single dataset defined on a given set of variables. It should be possible for a system based on a mathematical modeling to derive scientific knowledge from this dataset because we, as humans, have managed to do so.

Note also that, even though we acknowledge the idea that the production of science is a never-ending process, this does not imply that scientific knowledge may not be produced in a finite amount of time and, hence, with finite data (which is exactly what scientists do). In fact, this idea implies rather that scientific knowledge is produced with a certain degree of confidence[10]. Hypotheses are considered as scientific knowledge if we are sufficiently confident that they are correct. Further data may increase our confidence in the hypotheses (or falsify the hypotheses altogether), but we can never reach full confidence with a finite amount of data.

Throughout the rest of this section, we will consider the classical statistical model for the data in which the data was generated by a random sampling of n independent identically distributed random variables following an unknown distribution $\mathbf{p} = (p_i)_{0 \leq i \leq d}$. The aim is to formulate hypotheses about this distribution and evaluate them, in order to extract those in which we are confident (akin to scientific knowledge about the distribution). Our modeling represents a self-contained process, in the sense that no knowledge is given outside from the data. As such, the hypotheses which we will consider are necessarily data-driven data models. In order to be able to produce hypotheses in which we might be confident, we must first define some conditions on the data. We address this issue in section 3.7.2. We then specify the type of hypotheses which may be formulated in section 3.7.3 and present means to evaluate them in section 3.7.4. We then discuss the limits of our approach before concluding with some recommendations in section 3.7.5.

[10]We purposefully avoid the use of the term probability here.

3.7.2 Prerequisites to considering data-driven data models

Consider the following dataset in Table 3.4 which has been obtained by throwing two dice ten times and recording whenever one of the dice landed on a figure equal to five or six. As we have knowledge about how the world works,

a_1	a_2
0	0
1	0
1	0
0	0
0	1
0	0
0	0
1	1
0	0
0	0

Table 3.4: Example dataset

and particularly about how dice work, we will consider the distribution in which $p_{a_1} = 1/6$, $p_{a_2} = 1/6$ and the two tosses are considered independent, regardless of the data. Moreover, the hypothesis that the data follows such a distribution would not be rejected if we were to conduct a statistical test for this hypothesis.

However, if we have no knowledge about anything else than the dataset itself, let alone about dice, then we have no reason to consider this distribution at all. The distributions considered will be data-driven. If we test for independence between the two variables for example, we will consider the hypothesis based on the distribution in which $p_{a_1} = f_{a_1} = 0.3$, $p_{a_2} = f_{a_2} = 0.2$ and $p_{a_1 \wedge a_2} = p_{a_1} p_{a_2} = 0.06$ while considering that it has one degree of freedom. Similarly, this hypothesis would likely not be rejected by a statistical test. Implicitly, it means however that we accept the constraints given by the values f_{a_1} and f_{a_2} in the empirical distribution as reasonable and this is not necessarily justified. In fact, this entirely depends on the size of the dataset and we see, in this example that we are considering data driven hypotheses that are quite far from the distribution which was used to generate the data.

If we do not have enough data to be certain that the empirical distribution is close enough to the unknown theoretical distribution which we are trying to describe, we cannot justify the use of data-driven models. This means that we must be able to say that we do not have enough data to suggest any knowledge of scientific value concerning the theoretical distribution. In this respect, we depart significantly from the pattern mining as compression paradigm for which our approach has otherwise a number of similarities. Indeed, mining as compression approaches compress the empirical distribution regardless of the size of the dataset which defines the distribution.

It remains then to define what it means to be certain that the empirical distribution is close enough to the theoretical distribution. We suggest that we can equate this to the following condition: any hypothesis stating that the data was generated following a distribution which is not close to the empirical distribution would be rejected if a statistical test was to be performed. We present a formalization of this condition and discuss its properties in sections 3.7.2.1 and 3.7.2.2.

3.7.2.1 Confidence in the empirical distribution

The notion of confidence in the empirical distribution which we wish to define is based on the following idea. If we know that any potential hypothesis, in which we consider that the data was generated by a probability distribution that is far away from the empirical distribution, would be rejected by a statistical test, then we know that the only hypotheses which would not be rejected correspond to probability distributions which are close to the empirical distribution. Hence, we can say that we are confident that the empirical distribution is a good approximation of the theoretical distribution. In order to give a formal definition of this notion, we must start by specifying the set of potential probability distributions which can be considered and define a distance among these probability distributions. In the following, we consider an empirical distribution $\mathbf{f} = (f_i)_{0 \leq i \leq d}$ on $\{0, 1\}^m$, corresponding to a dataset D of n transactions (where $d = 2^m - 1$).

Potential distributions. Regarding the set of potential probability distributions, a first idea would be to consider all probability distributions that could have generated the dataset D. This is the set $\mathcal{S}_{\mathbf{f}}$ of probability distributions

$\mathbf{p} = (p_i)_{0\le i\le d}$ on $\{0,1\}^m$ such that $p_i > 0$ whenever $f_i > 0$:

$$\mathcal{S}_{\mathbf{f}} = \left\{\mathbf{p} = (p_i)_{0\le i\le d} \;\middle|\; \forall i \in [\![0,d]\!],\ f_i > 0 \implies p_i > 0\right\}$$

We may also consider including information provided by background knowledge (i.e. not inferred from the dataset D) about a more specific class of potential probability distributions $\mathcal{C} \subset \mathcal{S}_{\mathbf{f}}$. This could be the case if categorical or ordinal variables are considered as described in section 3.6.2.4. However, as no statistical test can be performed for a hypothesis that a probability is null for a given event, we must necessarily assume that the class of distributions $\mathcal{C}$ only contains probability distributions such that $p_i = 0$ only when $f_i = 0$. This means that $\mathcal{C}$ must be contained in the set $\mathcal{Z}_{\mathbf{f}}$ defined by:

$$\mathcal{Z}_{\mathbf{f}} = \left\{\mathbf{p} = (p_i)_{0\le i\le d} \;\middle|\; \forall i \in [\![0,d]\!],\ f_i > 0 \iff p_i > 0\right\}$$

Note that, if we may infer from the data that the theoretical distribution $\mathbf{p}$ satisfies $p_i > 0$ whenever $f_i > 0$, we may not infer the converse from the data and, therefore, this must correspond to background knowledge. In other words, if $f_i = 0$ for some $i \in [\![0,d]\!]$, we must be able to infer this from background knowledge if we wish to proceed rigorously with the process which we describe here.

Distance. Considering now a set of potential distributions $\mathcal{C}$, we wish to define a distance between the elements of this set. A simple choice would be to consider the Chebyshev distance defined, for two distributions $\mathbf{p}$ and $\mathbf{q}$, by:

$$\delta_{\mathrm{Cheb}}(\mathbf{p},\mathbf{q}) = \max_{0\le i\le d} |p_i - q_i|$$

However, this choice fails to take into account the fact that $\mathbf{p}$ and $\mathbf{q}$ are taken in $\mathcal{C}$. Hence, we prefer the normalized Chebyshev distance defined by:

$$\delta_{\mathbf{c}}(\mathbf{p},\mathbf{q}) = \max_{\substack{0\le i\le d\\ c_i\neq 0}} \frac{|p_i - q_i|}{c_i}$$

where $\mathbf{c}$ is the probability distribution with maximal entropy in $\mathcal{C}$. Indeed, $\mathbf{c}$ holds a centered position within $\mathcal{C}$ (see previous section 3.3.3 and further developments in chapter 4) so that the distance between two elements in $\mathcal{C}$

given by $\delta_{\mathbf{c}}$ expresses the skewness of the distribution of the elements within $\mathcal{C}$. Note that, in the case in which no background knowledge is assumed (which is the case that we will study more thoroughly), $\mathbf{c}$ is the uniform distribution given by $c_i = \frac{1}{2^m}$ for all $i \in [\![0, d]\!]$ so that $\delta_{\mathbf{c}}$ is equal to the standard Chebyshev distance multiplied by 2^m. In this case, we write δ rather than $\delta_{\mathbf{c}}$.

Statistical test. For all $\mathbf{p} \in \mathcal{C}$, let $H_{\mathbf{p}}$ be the hypothesis that the dataset D was generated by $\mathbf{p}$ and:

$$\chi^2_{\mathbf{p},\mathbf{f}} = n \sum_{\substack{0 \le i \le d \\ c_i \ne 0}} \frac{(f_i - p_i)^2}{p_i}$$

the corresponding χ^2 statistic.

The hypothesis $H_{\mathbf{p}}$ is rejected by the χ^2 test of goodness of fit for the threshold α if:

$$\chi^2_{\mathbf{p},\mathbf{f}} > \chi^2_{\alpha}(d_{\mathbf{c}})$$

where $d_{\mathbf{c}} + 1 = |\{i \in [\![0, d]\!] \mid c_i > 0\}|$ is the number of non-zero values c_i of $\mathbf{c}$ (which correspond also to the non-zero values f_i of $\mathbf{f}$), $\chi^2_{\alpha}(d_{\mathbf{c}})$ is the value such that $\mathrm{Prob}\left(Z > \chi^2_{\alpha}(d_{\mathbf{c}})\right) = \alpha$ if $Z \sim \chi^2(d_{\mathbf{c}})$ and α is a fixed probability threshold (typically .95 or .99).

Confidence. Rejecting the hypothesis means that it is considered unreasonable to assume that the data was generated by the corresponding distribution. Conversely, the hypothesis cannot be rejected if it is not highly unlikely that the data was generated by $\mathbf{p}$. As we want to consider the empirical distribution as an approximation for the theoretical distribution, we cannot accept a situation in which the dataset can reasonably be considered to be generated by a distribution which is too far away from the empirical distribution. This gives rise to the following definition which means that we are confident in the empirical distribution if any hypothesis which is not close to the distribution is rejected by a χ^2 test of goodness of fit.

Definition 3.7.1. We say that we are *confident in the empirical distribution* $\mathbf{f}$ *to the precision* ε *and to the degree given by the probability threshold* α if, for all distributions $\mathbf{p} \in \mathcal{C}$, we have:

$$\delta_{\mathbf{c}}(\mathbf{p}, \mathbf{f}) > \varepsilon \implies \chi^2_{\mathbf{p},\mathbf{f}} > \chi^2_\alpha(d_{\mathbf{c}})$$

Precise values of $\chi^2_\alpha(d_{\mathbf{c}})$ are known for low values of $d_{\mathbf{c}}$. However, if $d_{\mathbf{c}} = d$ (which is the case if no background knowledge is considered), d increases exponentially with m and we use the normal approximation for the χ^2 distribution given by $\frac{\chi^2(k)-k}{\sqrt{k}} \xrightarrow{d} \mathcal{N}(0,1)$ in order to approximate $\chi^2_\alpha(d)$ for values of m larger than 5.[11] This gives:

$$\chi^2_\alpha(d) \approx d + \mathcal{N}_\alpha\sqrt{2d}$$

with $\mathcal{N}_{.95} \approx 1.645$ and $\mathcal{N}_{.99} \approx 2.33$. Note that we obtain $\chi^2_\alpha(d) \sim 2^m$ so that the increase in m is asymptotically exponential.

3.7.2.2 How many transactions are needed?

We now try to determine bounds on the number of transactions n which allow us to be confident or not in the empirical distribution.

Background knowledge. For this section, we will consider only the specific case in which no background knowledge is assumed, that is $d_{\mathbf{c}} = d$, $c_i = \frac{1}{2^m}$ for all $i \in [\![0, d]\!]$ and $\delta_{\mathbf{c}} = \delta$ is the standard Chebyshev distance multiplied by 2^m.

Note that all of the results presented in the following paragraphs can easily be generalized when $\mathbf{c}$ describes a uniform distribution on its non null space (i.e. $c_i = \frac{1}{d_{\mathbf{c}}+1}$ for $d_{\mathbf{c}} + 1$ values of i in $[\![0, d]\!]$ and $c_i = 0$ otherwise). This corresponds to the case where the only background knowledge given is that some events are impossible. However, we do not cover more general types of background knowledge which would likely lead to much more complicated results.

Precision. We discuss, in this paragraph, the necessity to fix some bounds on the values of ε in order to preserve the meaningfulness of the process. For this, consider $\mathbf{e}$ such that $\mathbf{f} = \mathbf{p} + \mathbf{e}$, which represents the error between a

[11] There are, of course, much better approximations for $\chi^2_\alpha(d)$ (see, for example, [Beh18] on such approximations) but this simple approximation is sufficient for the purpose of our asymptotic analysis here.

potential theoretical distribution $\mathbf{p} \in \mathcal{C}$ and the empirical distribution $\mathbf{f}$. If $\|\cdot\|$ is the norm associated to the distance δ (i.e. $\|\mathbf{x}\| = 2^m \max_{0 \leq i \leq d} |x_i|$), then:

$$\delta(\mathbf{f}, \mathbf{p}) \geq \varepsilon \iff \|\mathbf{e}\| \geq \varepsilon$$

Now suppose that we allow all distributions $\mathbf{p}$ such that $|e_i| \geq f_i$ for some $i \in [\![0, d]\!]$, then we could have p_i equal to zero and which would mean that we are considering distributions outside of $\mathcal{C}$. Therefore, in order to ensure that we stay strictly within the interior of $\mathcal{C}$, we must necessarily have $\max_{0 \leq i \leq d} |e_i| < \min_{0 \leq i \leq d} f_i$ or, equivalently, $\|\mathbf{e}\| < 2^m \min_{0 \leq i \leq d} f_i$. Hence, it would not be meaningful to consider any precision greater than $2^m \min_{0 \leq i \leq d} f_i$. In other words, we must have:

$$\varepsilon < 2^m \min_{0 \leq i \leq d} f_i$$

Furthermore, it is also meaningless to consider a better precision for e_i than that given by the minimal quantity of information in the empirical distribution corresponding to a single transaction which is $\frac{1}{n}$. Hence, necessarily:

$$\frac{1}{n} \leq \frac{\varepsilon}{2^m} < \min_{0 \leq i \leq d} f_i$$

As the frequencies in the empirical distribution are necessarily whole fractions over n, this gives:

$$\frac{2}{n} \leq \min_{0 \leq i \leq d} f_i$$

Which implies that confidence in the precision of an empirical distribution cannot be estimated meaningfully if there are less than two transactions for each element in $\{0, 1\}^m$, setting a first lower bound on the number of transactions:

$$2^{m+1} \leq n$$

Conversely, if $n < 2^{m+1}$, a more restricted set of distributions $\mathcal{C}$ must necessarily be considered through some background knowledge (i.e. we must have $c_i = 0$ for some $i \in [\![0, d]\!]$) if we want to say that we are confident in the data.

Minimizing the χ^2 statistic. We now study the χ^2 statistic as a function on $(0,1)^{d+1}$ with values in $\mathbb{R}$:

$$\begin{array}{rccl} \chi^2_{\cdot,\mathbf{f}}: & (0,1)^{d+1} & \longrightarrow & \mathbb{R} \\ & \mathbf{p} & \longmapsto & \chi^2_{\mathbf{p},\mathbf{f}} \end{array}$$

in order to determine bounds for n. Indeed, if we can determine for which value $n_{\alpha,\mathbf{f}}$ we have:

$$\min_{\delta(\mathbf{p},\mathbf{f})\geq\varepsilon} \chi^2_{\mathbf{p},\mathbf{f}} = \chi^2_\alpha(d),$$

then we can say that we can be confident in the empirical distribution $\mathbf{f}$ to the precision ε if and only if $n \geq n_{\alpha,\mathbf{f}}$.

Determining $n_{\alpha,\mathbf{f}}$ is not an easy task unless the empirical distribution $\mathbf{f}$ is extremely simple (as is the case for the uniform distribution studied later on). However, we have determined lower and upper bounds for $n_{\alpha,\mathbf{f}}$, as presented in the following paragraphs.

In order to bound $n_{\alpha,\mathbf{f}}$, we start by noticing that, as:

$$\chi^2_{\mathbf{p},\mathbf{f}} = n\left(\sum_{i=0}^{d} \frac{f_i^2}{p_i}\right) - n$$

we have:

$$\frac{\partial}{\partial p_i}\chi^2_{\mathbf{p},\mathbf{f}} = -n\left(\frac{f_i}{p_i}\right)^2 \quad \forall i \in [\![0,d]\!]$$

and:

$$\frac{\partial^2}{\partial p_i \partial p_j}\chi^2_{\mathbf{p},\mathbf{f}} = \begin{cases} 0 & \text{if } i \neq j \\ \frac{2nf_i^2}{p_i^3} & \text{if } i = j \end{cases}$$

Hence, $\chi^2_{\cdot,\mathbf{f}}$ is strictly convex positive on $(0,1)^{d+1}$ with its unique minimum reached for $\mathbf{p} = \mathbf{f}$. Therefore:

$$\min_{\delta(\mathbf{p},\mathbf{f})\geq\varepsilon} \chi^2_{\mathbf{p},\mathbf{f}} = \min_{\delta(\mathbf{p},\mathbf{f})=\varepsilon} \chi^2_{\mathbf{p},\mathbf{f}}$$

and, more generally:

$$\min_{\delta(\mathbf{p},\mathbf{f})=\varepsilon'} \chi^2_{\mathbf{p},\mathbf{f}} \geq \min_{\delta(\mathbf{p},\mathbf{f})=\varepsilon} \chi^2_{\mathbf{p},\mathbf{f}} \quad \text{if } \varepsilon' \geq \varepsilon$$

Lower bounds for $n_{\alpha,\mathbf{f}}$. Now consider a distribution $\mathbf{p}$ and $0 < \eta_1 \leq \eta_2 < 1$ such that:

$$\forall i \in [\![0, d]\!], \quad \eta_1 f_i \leq |p_i - f_i| \leq \eta_2 f_i$$

Then:

$$f_i(1-\eta_2) \leq p_i \leq f_i(1-\eta_1)$$

and:

$$\frac{\eta_1^2}{1-\eta_2} f_i \leq \frac{(p_i - f_i)^2}{p_i} \leq \frac{\eta_2^2}{1-\eta_2} f_i$$

Hence, by summation:

$$\frac{\eta_1^2}{1-\eta_2} \leq \frac{1}{n}\chi^2_{\mathbf{p},\mathbf{f}} \leq \frac{\eta_2^2}{1-\eta_2}$$

Therefore, we can say that that the $H_{\mathbf{p}}$ hypothesis is rejected if $n \geq \frac{1-\eta_2}{\eta_1^2}\chi^2_\alpha(d)$ and that is not rejected if $n \leq \frac{1-\eta_2}{\eta_2^2}\chi^2_\alpha(d)$.

Moreover, $2^m \eta_1 \min\limits_{0\leq i\leq d} f_i \leq \|\mathbf{e}\| \leq 2^m \eta_2 \max\limits_{0\leq i\leq d} f_i$. Hence we can say that for any precision $\varepsilon \leq 2^m \eta_1 \min\limits_{0\leq i\leq d} f_i$, we must at least have $n \geq \frac{1-\eta_2}{\eta_1^2}\chi^2_\alpha(d)$ in order to be confident in $\mathbf{f}$ to the precision ε and to the degree α. As we have already established that we must have $\varepsilon < 2^m \min\limits_{0\leq i\leq d} f_i$, we can consider $\eta_1 = \frac{\varepsilon}{2^m \min\limits_{0\leq i\leq d} f_i}$ and $\eta_2 = \eta_1$. Note that this means that $\|\mathbf{e}\| = \frac{\max\limits_{0\leq i\leq d} f_i}{\min\limits_{0\leq i\leq d} f_i}\varepsilon \geq \varepsilon$ which gives us the condition $\delta(\mathbf{p},\mathbf{f}) \geq \varepsilon$. Hence, we have a first lower bound $a_{\alpha,\mathbf{f}}$ for $n_{\alpha,\mathbf{f}}$ which depends on the empirical distribution $\mathbf{f}$:

$$n_{\alpha,\mathbf{f}} \geq a_{\alpha,\mathbf{f}} \quad \text{where} \quad a_{\alpha,\mathbf{f}} = \frac{2^m \min\limits_{0\leq i\leq d} f_i \left(2^m \min\limits_{0\leq i\leq d} f_i - \varepsilon\right)}{\varepsilon^2}\chi^2_\alpha(d). \tag{3.1}$$

Furthermore, we have established that $\frac{2}{n} \leq \min\limits_{0\leq i\leq d} f_i$ is also a necessary condition for confidence in $\mathbf{f}$. Therefore:

$$\eta_1 = \frac{\varepsilon}{2^m \min\limits_{0\leq i\leq d} f_i} \leq \frac{n\varepsilon}{2^{m+1}}$$

and, as $x \mapsto \frac{1-x}{x^2}$ decreases on $(0,1)$:

$$\frac{1-\frac{n\varepsilon}{2^{m+1}}}{\left(\frac{n\varepsilon}{2^{m+1}}\right)^2} \leq \frac{1-\eta_1}{\eta_1^2}$$

Hence:

$$n \geq \frac{1 - \frac{n\varepsilon}{2^{m+1}}}{\left(\frac{n\varepsilon}{2^{m+1}}\right)^2} \chi^2_\alpha(d)$$

is also a necessary condition for confidence in $\mathbf{f}$ which is not dependent on the distribution. This last inequality may be reduced to the following polynomial inequality:

$$n^3 + \frac{2^{m+1}}{\varepsilon} \chi^2_\alpha(d) n - \left(\frac{2^{m+1}}{\varepsilon}\right)^2 \chi^2_\alpha(d) \geq 0$$

which is equivalent to:

$$g(x) = x^3 + px + q \geq 0$$

with:

$$p = \frac{2^{m+1}}{\varepsilon} \chi^2_\alpha(d) \quad \text{and} \quad q = -\left(\frac{2^{m+1}}{\varepsilon}\right)^2 \chi^2_\alpha(d)$$

This is a depressed cubic and, as we clearly have $4p^3 + 27q^2 \geq 0$, it is equivalent to $x \geq a_\alpha$ where a_α is the only real root of g given by:

$$a_\alpha = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} \tag{3.2}$$

The root a_α is a lower bound for $n_{\alpha,\mathbf{f}}$ which is easily computed for small values of m.

For higher values of m, we use the approximation of the χ^2 distribution by a normal distribution described above which gives $\chi^2_\alpha(d) \approx d + \mathcal{N}_\alpha \sqrt{2d}$. This also allows us to determine the asymptotic behavior of a_α when m increases which is given by:

$$a_\alpha \sim \gamma_a 2^m \quad \text{and} \quad \gamma_a = \frac{2^{1/3}}{\varepsilon^{2/3}} \left(\sqrt[3]{1 + \sqrt{1 + \frac{2\varepsilon}{27}}} + \sqrt[3]{1 - \sqrt{1 + \frac{2\varepsilon}{27}}} \right) \tag{3.3}$$

This last expression can be approximated for small values of ε by:

$$a_\alpha \sim \gamma_a 2^m \quad \text{and} \quad \gamma_a \approx \left(\frac{2}{\varepsilon}\right)^{2/3} \tag{3.4}$$

Upper bounds for $n_{\alpha,\mathbf{f}}$. We have determined a lower bound $a_{\alpha,\mathbf{f}}$ for $n_{\alpha,\mathbf{f}}$ which is given by the expression 3.1. We have also determined a larger lower bound a_α for $n_{\alpha,\mathbf{f}}$ whose expression, determined by the expression 3.2, does not depend on the distribution $\mathbf{f}$ and whose asymptotic behavior is approximately described by the expression 3.4. We will now give similar higher bounds for $n_{\alpha,\mathbf{f}}$.

Consider a distribution $\mathbf{p}$ such that $\delta(\mathbf{p},\mathbf{f}) = \varepsilon$. Then $\max\limits_{0\leq i\leq d} |p_i - f_i| = \frac{\varepsilon}{2^m}$ and:

$$\frac{1}{n}\chi^2 = \sum_{i=0}^{d} \frac{(p_i - f_i)^2}{p_i} \geq \frac{1}{\max\limits_{0\leq i\leq d} p_i} \left(\frac{\varepsilon}{2^m}\right)^2$$

Hence, if:

$$\frac{n}{\max\limits_{0\leq i\leq d} p_i} \left(\frac{\varepsilon}{2^m}\right)^2 \geq \chi^2_\alpha(d)$$

then $H_\mathbf{p}$ is rejected. As $\max\limits_{0\leq i\leq d} p_i \leq \frac{\varepsilon}{2^m} + \max\limits_{0\leq i\leq d} f_i \leq 1$. This gives an upper bound $b_{\alpha,\mathbf{f}}$ for $n_{\alpha,\mathbf{f}}$ which depends specifically on $\mathbf{f}$:

$$n_{\alpha,\mathbf{f}} \leq b_{\alpha,\mathbf{f}} \quad \text{where} \quad b_{\alpha,\mathbf{f}} = \left(\frac{2^m}{\varepsilon}\right)^2 \left(\frac{\varepsilon}{2^m} + \max_{0\leq i\leq d} f_i\right) \chi^2_\alpha(d)$$

and a more general upper bound b_α for $n_{\alpha,\mathbf{f}}$ which does not depend on $\mathbf{f}$:

$$n_{\alpha,\mathbf{f}} \leq b_\alpha \quad \text{where} \quad b_\alpha = \frac{2^m}{\varepsilon}\left(1 + \frac{2^m}{\varepsilon}\right) \chi^2_\alpha(d)$$

As $\chi^2_\alpha(d) \sim 2^m$, we have the following asymptotic behavior for b_α:

$$b_\alpha \sim \gamma_b 8^m \quad \text{and} \quad \gamma_b = \frac{1}{\varepsilon^2}$$

Note that both the lower and upper bounds, a_α and b_α, have asymptotic behaviors for large values of m which are constant in α.

The case of the uniform distribution. In the previous paragraphs, we have determined upper and lower bounds for $n_{\alpha,\mathbf{f}}$. In the general case, we have:

$$a_\alpha \leq n_{\alpha,\mathbf{f}} \leq b_\alpha$$

with the following asymptotic behaviors for the two bounds:

$$a_\alpha \sim \gamma_a 2^m \quad \text{and} \quad b_\alpha \sim \gamma_b 8^m$$

This gives quite a large interval for $n_{\alpha,\mathbf{f}}$ which can be reduced with the bounds $a_{\alpha,\mathbf{f}}$ and $b_{\alpha,\mathbf{f}}$ for a specific empirical distribution $\mathbf{f}$.

In the case of a uniform distribution defined by $f_i = \frac{1}{2^m}$ for all $i \in [\![0, d]\!]$, the specific bounds give the following interval for $n_{\alpha,\mathbf{f}}$:

$$\frac{1-\varepsilon}{\varepsilon^2}\chi^2_\alpha(d) \leq n_{\alpha,\mathbf{f}} \leq 2^m \frac{1+\varepsilon}{\varepsilon^2}\chi^2_\alpha(d)$$

However, in this specific case, the exact value for $n_{\alpha,\mathbf{f}}$ can be determined. Indeed, consider $\mathbf{p}$ such that $\delta(\mathbf{p}, \mathbf{f}) = \varepsilon$. We have $\max\limits_{0\leq i\leq d} |p_i - f_i| = \frac{\varepsilon}{2^m}$ so we can consider i_0 such that $|p_{i_0} - f_{i_0}| = \frac{\varepsilon}{2^m}$. Hence:

$$\frac{1}{n}\chi^2_{\mathbf{p},\mathbf{f}} = \sum_{i=0}^{d} \frac{f_i^2}{p_i} - 1 = \frac{1}{2^{2m}} \sum_{i=0}^{d} \frac{1}{p_i} - 1 = \frac{1}{2^{2m}} \left(\frac{1}{p_{i_0}} + \sum_{\substack{i=0 \\ i\neq i_0}}^{d} \frac{1}{p_i} \right) - 1$$

This is minimized for:

$$p_i = \frac{1 - p_{i_0}}{d} \quad \text{for all } i \neq i_0$$

Therefore, if $\mathbf{p}$ minimizes $\chi^2_{\mathbf{p},\mathbf{f}}$:

$$\frac{1}{n}\chi^2_{\mathbf{p},\mathbf{f}} = \frac{1}{2^{2m}} \left(\frac{1}{p_{i_0}} + \frac{d^2}{1 - p_{i_0}} \right) - 1 = \frac{(1 - 2^m p_{i_0})^2}{2^m p_{i_0} (2^m - 2^m p_{i_0})}$$

Furthermore, we have $|p_{i_0} - f_{i_0}| = \frac{\varepsilon}{2^m}$ so that:

$$2^m p_{i_0} = 1 - \varepsilon \quad \text{or} \quad 2^m p_{i_0} = 1 + \varepsilon$$

This gives:

$$\frac{1}{n}\chi^2_{\mathbf{p},\mathbf{f}} = \frac{\varepsilon^2}{(1-\varepsilon)(d+\varepsilon)} \quad \text{or} \quad \frac{1}{n}\chi^2_{\mathbf{p},\mathbf{f}} = \frac{\varepsilon^2}{(1+\varepsilon)(d-\varepsilon)}$$

As:

$$\frac{1}{(1-\varepsilon)(d+\varepsilon)} - \frac{1}{(1+\varepsilon)(d-\varepsilon)} = \frac{2\varepsilon(d-1)}{(1-\varepsilon^2)(d^2-\varepsilon^2)} > 0$$

we have:

$$\min_{\delta(\mathbf{p},\mathbf{f})=\varepsilon} \chi^2_{\mathbf{p},\mathbf{f}} = \frac{n\varepsilon^2}{(1+\varepsilon)(d-\varepsilon)}$$

Hence:

$$n_{\alpha,\mathbf{f}} = \frac{(1+\varepsilon)(d-\varepsilon)}{\varepsilon^2}\chi^2_\alpha(d)$$

Note that, in this case, $n_{\alpha,\mathbf{f}}$ is relatively close to the upper bound $b_{\alpha,\mathbf{f}}$ as:

$$b_{\alpha,\mathbf{f}} = \frac{(1+\varepsilon)(d+1)}{\varepsilon^2}\chi^2_\alpha(d)$$

and they have the same asymptotic behavior:

$$n_{\alpha,\mathbf{f}} \sim \gamma 4^m \quad \text{where} \quad \gamma = \frac{1+\varepsilon}{\varepsilon^2}$$

The limits of pure empirical science. Considering a precision $\varepsilon = 0.001$ and a degree of confidence $\alpha = 0.99$ as in Figure 3.5 and assuming the empirical distribution corresponds to the uniform case (which can also be seen as the theoretical distribution associated to m independent coin tosses), the size of a dataset which is necessary to be confident in the empirical distribution is $n = 34,057,279$ for $m = 2$, $n = 1,158,331,433,060$ for $m = 10$ and $n \approx 1.6 \times 10^{66}$ for $m = 100$. Considering much looser values of precision $\varepsilon = 0.05$

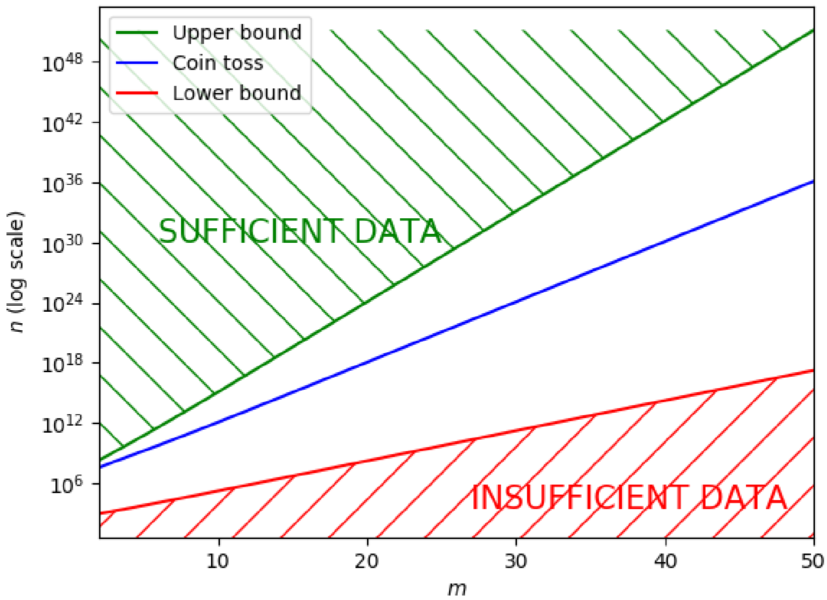

Figure 3.5: Upper and lower bounds for $n_{\alpha,\mathbf{f}}$ and exact values for the uniform distribution for $\alpha = 0.99$ and $\varepsilon = 0.001$.

and of degree of confidence $\alpha = 0.95$, we obtain $n = 9,863$ for $m = 2$, $n =$

$471,967,371$ for $m = 10$ and $n \approx 6.7 \times 10^{62}$ for $m = 100$. Hence, unless we consider extremely low values of m, we cannot be confident in the hypothesis that the data was generated by a distribution close to the empirical distribution if we do not have massive amounts of data. Even if we consider the loose values for ε and α, we would need a bit more than one petabyte of data to store any dataset with the minimum required size for $m = 20$ (note that the size of the dataset in bits is equal to $n \times m$). The necessary amounts of data required for larger values of m could not even be reached given the limitations of our physical world.

Moreover, as we have explained previously, such confidence is necessary for defining data-driven models. This leads to the following dilemma: either only consider small number of attributes which seriously limits the scope of scientific knowledge or accept that we cannot base all scientific knowledge purely on induction. The choice here is easily made. If we consider 100 coins being tossed, we want to be able to say that the mathematical model of 100 independent Bernoulli trials with parameters $(n, 1/2, 1/2)$ is a good model for the data without having to toss coins roughly 1.6×10^{66} times before we can consider any possible model. Scientific knowledge cannot only be inferred inductively.

Empirical distribution precision. In this section, we have mostly focused on determining the minimum number of transactions $n_{\alpha,\mathbf{f}}$ which is needed to be confident in a distribution $\mathbf{f}$ to the precision ε and the degree α. In many cases, the number of transactions will be a fixed parameter. Hence, it might be more interesting in these cases to determine the precision associated to a distribution $\mathbf{f}$, a number of transactions n and a degree α. This is simply the value $\varepsilon_{\alpha,\mathbf{f},n}$ defined as the infimum value for ε such that we are confident in the empirical distribution $\mathbf{f}$ to the precision ε and the degree α.

3.7.3 Formulating hypotheses

In this section, we will assume that we are considering a dataset for which we are confident in the empirical distribution to the precision ε and the threshold α. Our aim is now to determine which hypotheses can be formulated from this empirical distribution. We will distinguish between two types of hypotheses: global and local. A global hypothesis corresponds to a probability distribution

$\mathbf{p}$ on the Boolean lattice $\mathcal{B}$ described previously (see section 3.6) while a local hypothesis corresponds to the projection of such a distribution on a subset $\mathcal{L}$ of $\mathcal{B}$.

3.7.3.1 Global hypotheses

In order to qualify as truly objective, the formulation of hypotheses must rely on two scientific principles: *Newton's "Hypotheses non fingo"* and *Occam's razor*.

"*Hypotheses non fingo*", which translates to "*I do not feign hypotheses*", is associated to the notion that hypotheses must be entirely based on the data and no more[12]. Applying this principle, we determine that the only objective hypotheses which can be formulated on the probability distribution $\mathbf{p}$ is that it equates with the empirical distribution $\mathbf{f}$ at least partly. In other words, there is a subset $\mathcal{K}$ of the Boolean lattice $\mathcal{B}$ such that $\mathbf{p}_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$.

However, this is generally not sufficient in order to define a probability distribution on $\mathcal{B}$ and we must therefore rely on *Occam's razor* to determine $\mathbf{p}$. Indeed, this scientific principle supports that, if a choice must be made between various models which are equivalent in terms of results (in this context, possible probability distributions $\mathbf{p}$ such that $\mathbf{p}_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$), then the one based on the fewest assumptions should be considered. In the case of a constrained probability distribution as described above, *Occam's razor* translates to the *principle of maximum entropy* (see [Jay82, Jay03, CK11] as well as section 3.3.3 and chapter 4). This means that we consider the distribution corresponding to the MaxEnt model where entropy is maximized among all distributions $\mathbf{p}$ satisfying the constraint $\mathbf{p}_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$.

Note that there are two aspects in the hypotheses which we have defined: one approximate and one exact. Indeed, the first aspect of such a hypothesis is given by the equation $\mathbf{p}_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$ which is the approximation of the underlying distribution by the empirical distribution on $\mathcal{K}$. The empirical distribution could correspond exactly to the underlying distribution but this is highly unlikely and, in some cases, even impossible (if the underlying distribution takes irrational values on $\mathcal{K}$). The fact that we are confident in the empirical distribution ensures, however, that this approximation is close enough. The second

[12]This quote from Newton's Principia [New13] has been often mistranslated as "*I frame no hypotheses*" thus wrongly implying that Newton did not believe in formulating hypotheses at all [BC62].

aspect of the hypothesis consists in using the principle of maximum entropy to define $\mathbf{p}$ entirely and therefore implies that $\mathbf{p}$ can be naturally derived from its values on $\mathcal{K}$. In other words, the interactions between the attributes observed can be entirely described by those existing within $\mathcal{K}$ which is why we have described this property as the *mutual independence* of the attributes *constrained by* $\mathcal{K}$ (see chapter 4). This aspect of the hypothesis is exact in the sense that such statements are discrete and finite and at least one such statement is true about the underlying distribution. Hence, even though no such hypothesis can describe exactly an underlying distribution for which, for example, $p_a = \frac{\pi}{4}$, $p_b = \frac{1}{2}$ and $p_{a \wedge b} = \frac{\pi}{8}$, we can consider a hypothesis which describes exactly the fact that a and b are independent.

3.7.3.2 Local hypotheses

Given that the Boolean lattice $\mathcal{B}$ is generated by 2^m elements where m is the number of items considered, it may be difficult or even practically impossible to define a distribution $\mathbf{p}$ entirely on $\mathcal{B}$. Therefore, one might have to resort to local hypotheses rather than global hypotheses. Let $\mathcal{L} \subset \mathcal{B}$ such that $\mathcal{L}$ is still naturally isomorphic to a measurable space through the Boolean structure. Then we can define a local hypothesis on $\mathcal{L}$ for any subset $\mathcal{K} \subset \mathcal{L} \subset \mathcal{B}$ by considering a distribution $\mathbf{p}_{|\mathcal{L}}$ on $\mathcal{L}$ such that $\mathbf{p}_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$ and defined on the rest of $\mathcal{L}$ through the principle of maximum entropy.

Considering local hypotheses may lead to a number of various issues some of which we have already described. Indeed, when simultaneously considering multiple local hypotheses, one must be aware that they might be mutually inconsistent (see section 3.5 and particularly section 3.5.1.2). While this is not an issue if the hypotheses are considered to be concurrent, this is problematic if we want such hypotheses to jointly describe different local aspects of the single underlying probability distribution. Even when considering globally consistent local hypotheses, issues may arise in the evaluation step of the process as we will describe in section 3.7.4.

3.7.3.3 Selecting hypotheses

While we have identified what type of hypotheses can be formulated, we have not specified which of these hypotheses should be selected for evaluation. In the classical hypothetico-deductive model, individual hypotheses are formulated and evaluated through new empirical observations. However, as we are considering a static dataset, a slightly different approach must be adopted.

The formulation step in the hypothetico-deductive model is arguably the less transparent step in current scientific processes and is often associated to the notion of intuition. This simply shows that, although we believe hypotheses are formulated based on prior knowledge, we have little understanding of the exact processes which lead to their formulation. This explains why a hypothesis should normally not be tested based on the same empirical observations which led to its formulation: if we do not know exactly why a hypothesis was formulated we might end up validating a hypothesis based on the reasons which led to its definition (which comes down to the issue of overfitting as we have previously explained). As we are considering a single static dataset, we must adopt a different strategy in order to ensure that this does not occur. One solution is to define the hypotheses which we want to evaluate through this dataset without consideration for the dataset.

At first, this may seem to be in contradiction with our suggestion to consider hypotheses based on the empirical distribution as described in sections 3.7.3.1 and 3.7.3.2. However, considering a dataset for which we are confident in the empirical distribution allows to avoid this issue. Indeed, this ensures that the equality $\mathbf{p}_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$ for a given hypothesis is the only reasonable assumption (given the precision ε and the threshold α) and therefore not relevant towards the definition of the hypothesis. In a sense, considering a dataset for which we are confident reduces each hypothesis to its exact aspect (the mutual constrained independence of the attributes) as described in section 3.7.3.1. As this aspect of the hypothesis depends only on the choice of $\mathcal{K}$, we can select the hypotheses to evaluate based on $\mathcal{K}$.

Following the recommendations in section 3.6, we can first choose to restrict the selection of hypotheses by considering only subsets within a sound and complete family of patterns (such as the set of itemsets). Given such a restriction, we must still define which hypotheses to evaluate. An initial idea is to evaluate them all and consider the hypothesis which evaluates best. This is the same principle as the one used in strategies within the mining as compression paradigm. However, there is no theoretical guarantee in this case that overfitting is avoided. Another option is to define a total order on subsets of the family considered. Such an order should be associated to a notion of complexity so that the simplest set of patterns come first (this is not necessarily a trivial issue as described in section 3.6.2.3). Hypotheses may then be evaluated, following the order previously defined, until a hypothesis is tested positively. This last hypothesis is the least complex hypothesis which provides a reasonable explanation for the dataset and represents the scientific knowledge extracted from the dataset. This again is a direct transcription of Occam's razor.

3.7.4 Evaluating hypotheses

We consider, in this section, that the hypotheses which we mean to evaluate correspond to MaxEnt models as described in section 3.7.3. We will focus first on the issue of evaluating hypotheses based on global MaxEnt models (as described in section 3.7.3.1) before addressing some issues with the evaluation of local MaxEnt models (as described in section 3.7.3.2).

3.7.4.1 Evaluating global hypotheses

Compression scores. In the literature, a few criteria have been suggested by the proponents of the mining as compression paradigm (see section 2.3.3.4) for the evaluation of MaxEnt models defined by the frequencies of a set of itemsets. These include the Bayesian Information Criterion (BIC) or a Minimum Description Length (MDL) score [MVT12, VLV14]. As BIC may be seen as a simplified MDL score, we will concentrate on the latter in this paragraph.

Recall from section 2.3.3.4, that given the output of a program generated by a universal machine, there is a shortest possible program that generates the exact same output. The length of this program in bits is the Kolmogorov complexity of the output. Considering Kolmogorov complexity as a score for evaluating the best possible compression for a dataset would be solidly grounded in theory. However, Kolmogorov complexity has been shown to be uncomputable in general. Hence, computable substitutes, such as the ones mentioned above, have been considered instead. Note that, when considering such substitutes, a class of models must be defined a priori, together with a specific language to describe the models within this class. In the current context, the class of models considered is that of MaxEnt models defined by the frequencies of a set of itemsets.

In practice, two-part MDL scores are considered. They are computed as:

$$L(D, M) = L(M) + L(D|M)$$

where $L(M)$ is the length of the description of the model M in the specific language and $L(D|M)$ is the length of the description of the deviation of the data D with regards to the model. Such a score does not take into account the complexity of the task of generating the dataset from the model (i.e. the decompression task) and focuses instead on the complexity of the description of the model in the specific language. While [MVT12] justifies this simply by saying that the aim of the process is "*to summarize the data with a succinct set of itemsets, not model it with a distribution*", the authors of [VLV14] go further by stating that the complexity of the decompression algorithm is constant[13]. We entirely disagree with this statement as the complexity of the task of computing a MaxEnt model varies, in fact, significantly depending on the set of itemsets whose frequencies define the model. In practice, the variation is such that the search is limited to a subclass of MaxEnt models which can be reasonably computed (see [MVT12] and chapter 4 on these issues). This is an important theoretical issue because it undermines the theoretical foundations of the approach. Indeed, if the complexity of the task of generating the model from its description is irrelevant, what justifies the use of a general theoretical model in which it is not? In this respect, an approach based on statistical testing makes more sense as it disregards the issue of computing the model entirely.

Statistical testing. Statistical testing, which is the standard method for evaluating a hypothesis based on a data model, appears to be the most relevant option for evaluating a hypothesis in the current context. As any given hypothesis which we consider here defines a global model for the data, we should consider a statistical test which tests the model globally, such as a χ^2 test of adequacy. However, as the hypothesis considered corresponds to a description of the model given by the empirical frequencies of a set of itemsets

[13]The article states precisely while referring to the differences between two-part MDL and Kolmogorov complexity that "*One important difference is that $L(D, M)$ happily ignores the length of the decompression algorithm—which would be needed to reconstruct the data given the compressed representation of the model and data. The reason is simple: its length is constant, and hence does not influence the selection of the best model*".

or, in other words, the mutual independence of the items given a set of constraints, the number of degrees of freedom of the χ^2 test depends on the set of constraints (see proposition 5.1.1).

Now, considering that χ^2 tests are used to evaluate each individual hypothesis, we must still determine a means to discriminate between two concurrent hypotheses which both pass such a test. A seemingly simple solution would be to consider the χ^2 test as a score and look the hypothesis which scores best. However, such an approach conflicts with the principle of statistical testing itself, which should only be used to reject hypotheses. A more meaningful approach is to consider the simplest model which passes the test. Such an approach implies that a total order relation between all potential hypotheses, corresponding to a notion of relative complexity, be defined. As discussed previously in section 3.7.3.3, in order for the process to remain objective, the choice of such an order must be defined a priori. We set aside the issue of determining the most appropriate order for such hypotheses for the moment. Note, however, that such an order should be determined by objective factors (such as the number of itemsets considered or the total number of items within the itemsets considered) but must also necessarily rely on arbitrary factors (such as a lexicographic order between itemsets of same size) if we want to define a total order between hypotheses.

3.7.4.2 Evaluating local hypotheses

In section 3.7.3.2, we describe the possibility of defining local hypotheses rather than global ones. While such an approach is theoretically suboptimal, it can be necessary in practice if defining global MaxEnt models is technically infeasible.

The first issue to consider is the issue of global consistency of local, non-concurrent, hypotheses (see section 3.5.1.2). This is a highly complex topic in itself, which can be related to the study of junction trees (see, for example, [CL68, WP06a, KS10, SK12]), and a thorough analysis of this issue is well beyond the scope of this thesis. However, one simple method for dealing with this issue is to consider a partition of the set of items into blocks of items, so that any local model considered is defined on one of the local Boolean lattices $\mathcal{L}$ generated by the items within a given block. The issue of defining such a partition is quite tricky itself, at least if we wish to preserve meaningfulness throughout the process, but could eventually be justified by the knowledge (or simple assumption) that the itemsets in different blocks do not truly interact with each other (either because the itemsets within separate blocks are independent, or because they are incompatible, or a combination of both depending on the blocks considered). We discuss the possibility of such an approach in section 5.3.

The second issue to consider is one related to multiple testing. Indeed, when performing tests on a large number of local models, each corresponding to a different projection of a global model, false positives are known to appear [LTP06, Han11, KIA+17]. Once again, the study of this issue is far beyond the scope of this thesis and we do not provide any solution which allows to tackle it. However, we do believe that considering partition models as described above would, without solving the issue altogether, allow to apprehend it in simpler terms.

3.7.5 Recommendation

Throughout this section, we have defended the idea that we can use the hypothetico-deductive model within a general mathematical modeling process in order to meaningfully model the objectivity in objective frequency-based interesting pattern mining. The model we have proposed allows to define and evaluate hypotheses based on this principle.

Because we are dealing with static and finite data, we have argued that such a model must include a means of determining whether or not the number of transactions n in the data is sufficient to meaningfully and objectively define data-driven hypotheses. This led us to the definition of the notion of confidence in the empirical dataset. More generally, we recommend that such an indicator be used in pattern mining processes, to evaluate the confidence which can be expected in the scienticity of the information extracted in such processes.

Furthermore, we recommend that the hypotheses considered should be formulated as MaxEnt models associated to constraints defined by the empirical frequencies of a set of patterns in the dataset.

Finally, we support the idea that the patterns which are to be extracted in a meaningful objective frequency-based interesting pattern mining should correspond to the simplest possible hypothesis that passes a statistical test.

We fully acknowledge that a rigorous implementation of the model which we have described is practically quite difficult, either because we do not have sufficient data or because we do not have enough resources to entirely compute the hypotheses considered. However, it is important to understand the limits of our model in order to determine why and how compromises with such a rigorous approach can be made. In fact, these limits also provide information about some of the theoretical boundaries that surround empirical science. Hoping to be able to extract objective scientific information about the mutual interactions of a hundred items, based solely on the frequencies of their associations in a finite dataset is, as we have shown, completely unreasonable. This gives food for thought as a hundred items is most generally considered small data in the current data mining community. It also allows us to reflect on the manner one should go about building an artificial scientist because it means such an artificial intelligence should only consider datasets on a large number of items if it has developed some form of prior knowledge about the relationships between these items.

3.8 Conclusion

The meaningfulness of any process based on mathematical models, as well as the meaningfulness of the output of such a process, stems directly from the meaningfulness of the mathematical modeling process on which it relies. This is why it is important to make explicit such modeling processes and understand how choices in terms of modeling affect and determine the meaningfulness of the general process.

While focusing on the specific case of mathematical modeling for objective frequency-based interesting pattern mining processes, we have exhibited general issues related to mathematical modelings and, as such, much of the research presented in this chapter can be seen as a contribution in the field of the philosophy of science. This includes our definitions of the notions of model and modeling, the notions of phenotypic and genotypic modeling, the notion of pragmatic modeling, the notions of patchwork and holistic modeling, and our approach towards the modeling of the scientific method.

Concomitantly, we have determined a number of principles for meaningful mathematical modeling in the specific case of objective frequency-based interesting pattern mining processes which are summarized at the end of each section of this chapter. The definitions of our own pattern mining processes (see chapter 5), as well as the mutual constrained independence models they rely on (see chapter 4), are deeply rooted in these principles.

CHAPTER 4

Mutual constrained independence models

In the previous chapter, we have listed a number of recommendations for a meaningful modeling of the objective interestingness of itemsets. As we have presented, MaxEnt models are an essential tool for such a modeling because they provide for an objective answer to the issue of defining a probability measure given a set of constraints on that measure (see section 3.7.3). This statement can be supported by a number of various approaches towards the definition of MaxEnt models, the two most notable ones being Claude E. Shannon's original presentation of information entropy and E. T. Jayne's approach (see section 3.3.3 for more detail). In the course of this doctoral thesis, we have added a third approach towards the definition of MaxEnt models in the context of itemset mining: mutual constrained independence (MCI). We will start this chapter by defining this notion and proving the mathematical properties on which it is founded. We will then present some of the properties related to MCI models and exhibit their link to MaxEnt models. Finally, we will provide novel methods for computing these models.

4.1 Theoretical foundations of MCI

4.1.1 Preliminaries

4.1.1.1 Notations

Let $\mathcal{A} = \{a_1, ..., a_m\}$ be a set of m items. We will consider the following notations from section 3.6:

- $\mathcal{B}$ is the Boolean lattice associated to $\mathcal{A}$;

- $\top$ is the top element of $\mathcal{B}$ corresponding to the empty itemset;
- $d = 2^m - 1$;
- $\Omega = (\omega_i)_{0 \leq i \leq d} \subset \mathcal{B}$ is the set of minimal generators of $\mathcal{B}$ ordered by the natural lexicographic order as described in section 3.6.2.5;
- $\mathcal{I} = (I_i)_{0 \leq i \leq d} \subset \mathcal{B}$ is the set of all itemsets ordered by the natural lexicographic order described in the section mentioned above;
- for any probability measure $\mathbf{p}$ defined on $\mathcal{B}$, we write p_i in place of $\mathbf{p}(\omega_i)$, for all $i \in [\![0, d]\!]$, and p_X in place of $\mathbf{p}(X)$, for all $X \in \mathcal{B} \setminus \Omega$.

4.1.1.2 Transfer matrix

Before we make explicit how we aim to objectively define a probability measure on $\mathcal{B}$ from given constraints, we introduce a mathematical object which is quite useful for defining such probability measures: the transfer matrix from Ω to $\mathcal{I}$. Recall that any measure on $\mathcal{B}$ is defined naturally by its values on the minimal generators Ω of $\mathcal{B}$ (which correspond to the generalized itemsets of size m). Furthermore, a measure can also be entirely defined by its values on the itemsets $\mathcal{I}$, as we have described in section 3.6. As such, these families of patterns can be seen as bases for representing probability measures.

Switching from one representation to the other is easily accomplished using a transfer matrix. Indeed, consider the binary matrix T of size $2^m \times 2^m$ such that $T_{k,l} = 1$ if and only if $(\omega_l \implies I_k)$. It results from the properties of a measure that, for any measure $\mathbf{g}$ on $\mathcal{B}$, we have the following equality:

$$TX_{\mathbf{g}} = \begin{bmatrix} g_{I_0} \\ \vdots \\ g_{I_d} \end{bmatrix} \quad \text{where } X_{\mathbf{g}} = \begin{bmatrix} g_0 \\ \vdots \\ g_d \end{bmatrix}$$

The values for the coordinates $T_{k,l}$ of the matrix T can be computed directly from the indices k and l. To do this, we note that k and l can both naturally be represented by binary vectors $\mathbf{k} = (k_1, ..., k_m)$ and $\mathbf{l} = (l_1, ..., l_m)$ to which we associate them. The coordinates of the matrix T are then given by the

Generalized itemset	$X \in \Omega$
$\overline{a_1 a_2 a_3}$	$\omega_0 = \neg a_1 \wedge \neg a_2 \wedge \neg a_3$
$\overline{a_1 a_2} a_3$	$\omega_1 = \neg a_1 \wedge \neg a_2 \wedge a_3$
$\overline{a_1} a_2 \overline{a_3}$	$\omega_2 = \neg a_1 \wedge a_2 \wedge \neg a_3$
$\overline{a_1} a_2 a_3$	$\omega_3 = \neg a_1 \wedge a_2 \wedge a_3$
$a_1 \overline{a_2 a_3}$	$\omega_4 = a_1 \wedge \neg a_2 \wedge \neg a_3$
$a_1 \overline{a_2} a_3$	$\omega_5 = a_1 \wedge \neg a_2 \wedge a_3$
$a_1 a_2 \overline{a_3}$	$\omega_6 = a_1 \wedge a_2 \wedge \neg a_3$
$a_1 a_2 a_3$	$\omega_7 = a_1 \wedge a_2 \wedge a_3$

Itemset	$X \in \mathcal{I}$
$\emptyset$	$I_0 = \top$
a_3	$I_1 = a_3 = \omega_1 \vee \omega_3 \vee \omega_5 \vee \omega_7$
a_2	$I_2 = a_2 = \omega_2 \vee \omega_3 \vee \omega_6 \vee \omega_7$
$a_2 a_3$	$I_3 = a_2 \wedge a_3 = \omega_3 \vee \omega_7$
a_1	$I_4 = a_1 = \omega_4 \vee \omega_5 \vee \omega_6 \vee \omega_7$
$a_1 a_3$	$I_5 = a_1 \wedge a_3 = \omega_5 \vee \omega_7$
$a_1 a_2$	$I_6 = a_1 \wedge a_2 = \omega_6 \vee \omega_7$
$a_1 a_2 a_3$	$I_7 = a_1 \wedge a_2 \wedge a_3 = \omega_7$

Table 4.1: Correspondence between elements in $\mathcal{I}$ and Ω for $m = 3$.

following equation (where $\cdot$ is the dot product).

$$T_{k,l} = \begin{cases} 1 & \text{if} \quad (\mathbf{d} - \mathbf{l}) \cdot \mathbf{k} = 0 \\ 0 & \text{if} \quad (\mathbf{d} - \mathbf{l}) \cdot \mathbf{k} \neq 0 \end{cases} \tag{4.1}$$

Furthermore, we can see that T is invertible and that the value for the coordinates $T^{-1}_{k,l}$ of its inverse are given by the following equation.

$$T^{-1}_{k,l} = \begin{cases} (-1)^{(\mathbf{l}-\mathbf{k})\cdot\mathbf{d}} & \text{if} \quad (\mathbf{d} - \mathbf{l}) \cdot \mathbf{k} = 0 \\ 0 & \text{if} \quad (\mathbf{d} - \mathbf{l}) \cdot \mathbf{k} \neq 0 \end{cases} \tag{4.2}$$

Equation (4.1) is obtained quite directly from the definition of T. Indeed, we see that $(\omega_l \implies I_k)$, if and only if, $(\forall i \in [\![1, m]\!],\ k_i = 1 \implies l_i = 1)$, which is equivalent to the equation $(\mathbf{d} - \mathbf{l}) \cdot \mathbf{k} = 0$. Equation (4.2) can then be verified by multiplying both matrices. Indeed, let M be the matrix obtained by multiplying T with the matrix whose coordinates are defined by (4.2). The coordinates of M are given by:

$$M_{i,j} = \sum_{\substack{k=0 \\ (\mathbf{d}-\mathbf{k})\cdot\mathbf{i}=0 \\ (\mathbf{d}-\mathbf{j})\cdot\mathbf{k}=0}}^{d} (-1)^{(\mathbf{j}-\mathbf{k})\cdot\mathbf{d}} \ .$$

From $(\mathbf{d}-\mathbf{k})\cdot\mathbf{i} = 0$, we get that if $\mathbf{i}$ has a coordinate equal to 1, then the coordinate with the same index in $\mathbf{k}$ is also equal to 1. From $(\mathbf{d}-\mathbf{j})\cdot\mathbf{k} = 0$, we see that if $\mathbf{j}$ has a coordinate equal to 0, then the coordinate with the same index in $\mathbf{k}$ is also equal to 0. Hence, if $i > j$, $M_{i,j} = 0$. Furthermore, if $i = j$, then necessarily $k = i$ from which we get $M_{i,j} = 1$. Finally, if $i < j$, then by grouping all the values of k for which $\mathbf{k}$ has the same number $r = (\mathbf{j}-\mathbf{k})\cdot\mathbf{d}$, we get:

$$M_{i,j} = \sum_{r=0}^{(\mathbf{j}-\mathbf{i})\cdot\mathbf{d}} \binom{(\mathbf{j}-\mathbf{i})\cdot\mathbf{d}}{r} (-1)^r = 0 \ .$$

Hence, M is equal to the identity matrix which proves the result. Notice also that T^{-1} has all its coordinates in $\{-1, 0, 1\}$ which will be used in the proof of theorem 4.1.3.

$$T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, T^{-1} = \begin{bmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 4.1: The transfer matrix T and its inverse T^{-1} for $m = 3$.

4.1.1.3 Problem statement

As we have described previously, our aim here is to define a probability measure on $\mathcal{B}$ given a number of constraints on itemsets. We formalize this aim through the following problem statement.

Let $\mathcal{K} \subset \mathcal{I}$ be a set of itemsets and $\mathbf{f}_{|\mathcal{K}}$ be the restriction to $\mathcal{K}$ of a probability measure on $\mathcal{B}$ which corresponds to an empirical distribution in a dataset of transactions. In the following, we will refer to such a set of itemsets $\mathcal{K}$ as a constrained set, $\mathbf{f}_{|\mathcal{K}}$ as a constraint function and $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$ as a constraint system on $\mathcal{B}$. We say that a probability measure $\mathbf{p}$ on $\mathcal{B}$ satisfies the constraints given by the constraint system $\mathcal{C}$, if its restriction to $\mathcal{K}$ is equal to $\mathbf{f}_{|\mathcal{K}}$ (i.e. $\forall X \in \mathcal{K}$, $p_X = f_X$). We consider the problem of objectively hypothesizing the values of a probability measure on $\mathcal{B}$ which satisfies a constraint system $\mathcal{C}$.

In other words, we aim to define a probability measure $\mathbf{p}$ on $\mathcal{B}$ as a hypothesis for the value of $\mathbf{f}$, as naturally and objectively as possible, based on the sole knowledge that is given about $\mathbf{f}$ by the constraint system $\mathcal{C}$. Note that this problem statement is not a purely mathematical problem as the notion of objectivity is not a mathematical one per se. We must therefore model this notion in order to transform this into a purely mathematical problem.

4.1.1.4 Formulating objective hypotheses

Before we provide an answer to the problem statement described above, let us consider the wider issue of formulating hypotheses on mathematical objects based on partial knowledge of these objects. In section 3.7, we aimed at presenting a mathematical modeling for the discovery of scientific knowledge about the world from data. The modeling which is suggested in this previous section relies on the formulation of hypotheses based on knowledge acquired from the data in which we may be confident. However, in this specific step of hypothesis formulation, there is no intention to discover new knowledge from the data: the hypothesis must be formulated or, in other words, inferred based on knowledge already acquired from the data. More generally, for an intelligent system, we can differentiate between its ability to acquire knowledge from the world and its ability to reason based on its knowledge of the world. We will focus here on this second aspect.

Consider an intelligent system whose representation of the world is given by a mathematical model. The system has knowledge about the world stored in its memory from which it can directly infer further knowledge about the world using methods from mathematical reasoning. For example, if the system knows that $a = 768$, $b = 453$ and $c = a \times b$ as well as basic arithmetic, it will be able to answer that $c = 347,904$ to the question "What is c equal to?". This answer is part of the scope of the knowledge of the system even though it is not necessarily part of the knowledge stored in its memory. In a sense, the fact that the scope of the knowledge of the system reaches beyond the knowledge stored in its memory is a defining characteristic of intelligence.

However, it would be quite limiting to consider that the scope of the knowledge of a system can only be reached through mathematical reasoning. Indeed, one might be interested in an exact numerical value as an answer to a question when pure mathematical reasoning may only provide an interval. For example, even if one does not know the exact age of the last person one has met, and cannot derive it exactly from one's knowledge, one can generally still provide an answer if asked to guess that person's age. In every day life, such an answer is called an educated guess and is based on the person's prior knowledge about people and ages and the world in general (even though the mechanisms that lead to its formulation are essentially a black box). Similarly, we can formalize the notion of an educated guess in the case of an intelligent system whose knowledge of the world is a mathematical model. As we do not include any form of black box in our formulation process, we will use the term objective hypothesis rather than educated guess.

In order to formulate such objective hypotheses, we rely on the principle of indifference (also referred to as the principle of insufficient reason). This principle states that, when confronted to a model in which different possibilities arise and no information allows to differentiate between any of them, then each possibility should be considered as equally likely. The system should therefore consider every possible interpretation of the world as equally likely thus defining a uniform distribution on the set of possible interpretations of the world (that is, if such a probability measure is definable on this set, which is always the case if the set is finite but not, a priori, the case if the set is infinite). In the case in which a value must be provided for a variable, such a uniform distribution induces a distribution on the set of possible values for this variable. We can then use this last distribution to define an objective hypothetical value for this variable.

In such a case, several approaches can be used to determine this hypothetical value. A first approach, based on information theory, considers the value which adds the least information to the system. A second approach considers the value with the highest likelihood (which is not necessarily possible if the distribution is not discrete). A third option, which is the one that we shall study in detail in this chapter, is to consider the expected value for this variable (which is possible only if the variable is numerical and the expected value is well defined). These three approaches correspond to the three approaches towards the definition of MaxEnt models described previously (Shannon's, Jayne's and our own respectively) and we shall show how they all relate to each other in the specific context studied here.

Note that we have not discussed the practical manner in which an intelligent system may compute such hypotheses. This is of course a consideration of the utmost importance, notably because the theoretical scope of the knowledge of an intelligent system, which corresponds to the notion we have described above, is not a priori equal to the practical scope of its knowledge, which comprises only the conclusions the system might reach within the limits of its resources. Therefore, a process resulting in the formulation of hypotheses must be defined and its complexity must be taken into consideration. In particular, a naive process which would consist in an exhaustive review of all the different interpretations of the world would be practically infeasible in general. Hence, more elaborate mathematical tools are necessary to compute hypotheses while bypassing the computation of the underlying uniform distribution.

4.1.1.5 Application to the problem statement

Let us now try to understand how the approaches described in section 4.1.1.4 can apply to the problem statement defined in section 4.1.1.3. The world is represented here as a dataset of transactions on items whose empirical distribution is described by a probability measure $\mathbf{f}$ on $\mathcal{B}$. However, we only have partial knowledge about the world. The knowledge we have is represented by the restriction $\mathbf{f}_{|\mathcal{K}}$ of $\mathbf{f}$ to a set of itemsets $\mathcal{K} \subset \mathcal{I}$. In other words, our knowledge of the world is defined by the constrained system $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$. Our aim is to define a probability measure $\mathbf{p}$ on $\mathcal{B}$ which can be seen as an objective hypothesis about $\mathbf{f}$ based on the partial knowledge defined by $\mathcal{C}$. Given our representation of the world, any interpretation of the world corresponds to a dataset whose empirical distribution $\mathbf{h}$ satisfies the constraints given by the constraint system $\mathcal{C}$. As described in section 4.1.1.4, we would like to define $\mathbf{p}$ as the expected value for $\mathbf{h}$ given a uniform distribution on the set of all these datasets. However, this raises an issue as this set is infinite and there is no natural way to define a uniform distribution on it.

One first approach is to consider only datasets of a given size (i.e. the number of transactions n can therefore be seen as an additional constraint).

We call this approach the finite approach and discuss this in section 4.1.2. As we show, this approach poses both theoretical and practical issues. Another approach is to consider the limit, when n goes towards infinity, of the solutions obtained when considering datasets of size n. As we show, this limit is well defined and, in contrast with the finite approach, it does not suffer from the same theoretical issues and is more easily computed. This asymptotic approach, presented in section 4.1.3, is central to the notion of mutual constrained independence described in this article.

4.1.2 Finite approach

Consider a set of itemsets $\mathcal{K} \subset \mathcal{I}$ and define $\overline{\mathcal{K}} = \mathcal{K} \cup \{\top\}$. Let $\mathbf{g}_{|\overline{\mathcal{K}}}$ be the restriction to $\overline{\mathcal{K}}$ of a measure on $\mathcal{B}$ with integer values so that $\mathbf{g}_{|\overline{\mathcal{K}}}$ can be seen as corresponding to a dataset with n transactions where $n = g_\top$. Then, the set $\mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}}$ defined below as the set of all measures on $\mathcal{B}$ with integer values which are equal to $\mathbf{g}_{|\overline{\mathcal{K}}}$ for all itemsets in $\overline{\mathcal{K}}$ is finite:

$$\mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}} = \left\{\mathbf{h} = (h_i)_{0 \leq i \leq d} \in \mathbb{N}^{d+1} \;\middle|\; \forall X \in \overline{\mathcal{K}},\ h_X = g_X\right\}$$

Furthermore, for each measure $\mathbf{h} \in \mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}}$, there are exactly $\frac{n!}{\mathbf{h}!}$ distinct datasets which can be associated to $\mathbf{h}$ where $\mathbf{h}! = \prod_{i=0}^{d} h_i!$. Hence, we can define the expected measure μ when considering a uniform distribution on all possible datasets corresponding to a measure in $\mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}}$ by:

$$\mu = \frac{\sum\limits_{\mathbf{h} \in \mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}}} \frac{1}{\mathbf{h}!}\mathbf{h}}{\sum\limits_{\mathbf{h} \in \mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}}} \frac{1}{\mathbf{h}!}} . \tag{4.3}$$

By linearity, μ is of course a measure on $\mathcal{B}$ such that, $\forall X \in \overline{\mathcal{K}}$, $\mu_X = g_X$. In particular, $\mu_\top = n$. This measure is entirely defined by $\left(\overline{\mathcal{K}}, \mathbf{g}_{|\overline{\mathcal{K}}}\right)$. Note that $\left(\overline{\mathcal{K}}, \mathbf{g}_{|\overline{\mathcal{K}}}\right)$ is not a constraint system per se because $\mathbf{g}$ is not a probability measure (excluding the trivial case for which $n = 1$). We can naturally bring this problem down to probability measures and constraint systems by considering the constraint system $\mathcal{C}_n = \left(\mathcal{K}, \frac{1}{n}\mathbf{g}_{|\overline{\mathcal{K}}}\right)$ and noticing that $\frac{1}{n}\mu$ is a probability measure satisfying $\mathcal{C}_n$. However, this constraint system does not, in general, uniquely define $\frac{1}{n}\mu$ as we will show in the third of the following three examples.

4.1.2.1 Particular constrained sets

Empty set. The first specific case which we consider is the case in which $\mathcal{K} = \emptyset$ and, therefore, $\overline{\mathcal{K}} = \{\top\}$. This case is quite trivial and can be seen as the case in which there is only a constraint on the number of transactions. By symmetry, we see that all μ_i are equal. As there sum is equal to n, we get $\mu_i = \frac{n}{2^m}$ for all $i \in [\![0, d]\!]$. Hence, $\frac{1}{n}\mu$ corresponds to the theoretical probability distribution for m random independent coin tosses.

Independence model. In this case, $\mathcal{K} = \mathcal{A} = \{a_1, ..., a_m\}$. This corresponds to the case in which the frequencies $n_{a_1}, ..., n_{a_m}$ corresponding to each item, as well as the total number of transactions n, are fixed constraints. Considering the natural representation of a dataset of n transactions on these m items as a binary matrix, we see that the constraints correspond to the column margins. As each constraint corresponds to an individual column, we see that the set of all $n \times m$ binary matrices satisfying the constraints has a natural one-to-one correspondence with the Cartesian product of the m sets of vector columns of size n corresponding to each individual constraint. Therefore, in this case, $\frac{1}{n}\mu$ corresponds to the distribution given by the independence model.

All proper subitemsets. The last specific case we consider here is the case in which $\mathcal{K}$ contains all the proper subitemsets of a given itemset. This case was presented in [DBLL15]. Without any loss for generality, we may limit our study to the case in which the itemset considered is I_d (recall that $I_d = \bigwedge_{i=0}^{m} a_i$) and hence $\mathcal{K} = \mathcal{I} \setminus \{I_d\}$.

We suppose that we are considering measures $\mathbf{h}$ on $\mathcal{B}$ constrained so that, for all $i \in [\![0, d-1]\!]$, $h_{I_i} = n_i$, where the integers n_i correspond to some empirical dataset (note that $n_0 = n$ necessarily). Then, for all $j \in [\![0, d]\!]$, h_j is determined entirely by the values n_i together with one variable k such that $h_{I_d} = k$. More precisely, considering the transfer matrix T and its inverse as defined in section 4.1.1.2, we have:

$$\begin{bmatrix} h_0(k) \\ \vdots \\ h_d(k) \end{bmatrix} = T^{-1} \begin{bmatrix} n_0 \\ \vdots \\ n_{d-1} \\ k \end{bmatrix}.$$

Furthermore, we know that the possible values for h_{I_d} correspond exactly to an interval $[\![l, u]\!]$ whose bounds are entirely defined by the constraints n_i. This result, which is presented by Calders and Goethals in their work on non-derivable itemsets [CG02], can be rephrased using the transfer matrix. Indeed, recall that $k \in [\![l, u]\!]$ is equivalent to $h_i \geq 0$ for all $i \in [\![0, d]\!]$. Hence, if we write the previous equation as:

$$\begin{bmatrix} h_0(k) \\ \vdots \\ h_d(k) \end{bmatrix} = T^{-1} \begin{bmatrix} n_0 \\ \vdots \\ n_{d-1} \\ 0 \end{bmatrix} + T^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ k \end{bmatrix} = T^{-1} \begin{bmatrix} n_0 \\ \vdots \\ n_{d-1} \\ 0 \end{bmatrix} + k \begin{bmatrix} (-1)^{(\mathbf{d}-\mathbf{0})\cdot\mathbf{d}} \\ (-1)^{(\mathbf{d}-\mathbf{1})\cdot\mathbf{d}} \\ \vdots \\ (-1)^{(\mathbf{d}-\mathbf{d})\cdot\mathbf{d}} \end{bmatrix},$$

we can say that:

$$l = \max_{\substack{i \in [\![0,d]\!] \\ (\mathbf{d}-\mathbf{i})\cdot\mathbf{d} \text{ even}}} (-c_i) \qquad \text{and} \qquad u = \min_{\substack{i \in [\![0,d]\!] \\ (\mathbf{d}-\mathbf{i})\cdot\mathbf{d} \text{ odd}}} (c_i)$$

where:

$$\begin{bmatrix} c_0 \\ \vdots \\ c_d \end{bmatrix} = T^{-1} \begin{bmatrix} n_0 \\ \vdots \\ n_{d-1} \\ 0 \end{bmatrix}.$$

We can therefore express $\mu(h_{I_d})$ through the following formula:

$$\mu(h_{I_d}) = \frac{\sum\limits_{k=l}^{u} \frac{k}{\prod\limits_{i=0}^{d} h_i(k)!}}{\sum\limits_{k=l}^{u} \frac{1}{\prod\limits_{i=0}^{d} h_i(k)!}} \tag{4.4}$$

which can be computed directly using T^{-1}. The value obtained allows in turn to determine $\mathbf{h}$ entirely.

For the case in which $m = 3$, equation (4.4) becomes:

$$\mu(h_{I_7}) = \frac{\displaystyle\sum_{k=l}^{u} \frac{k}{\begin{array}{c}(n_0 - n_1 - n_2 + n_3 - n_4 + n_5 + n_6 - k)!(n_1 - n_3 - n_5 + k)!\\(n_2 - n_3 - n_6 + k)!(n_3 - k)!(n_4 - n_5 - n_6 + k)!(n_5 - k)!(n_6 - k)!k!\end{array}}}{\displaystyle\sum_{k=l}^{u} \frac{1}{\begin{array}{c}(n_0 - n_1 - n_2 + n_3 - n_4 + n_5 + n_6 - k)!(n_1 - n_3 - n_5 + k)!\\(n_2 - n_3 - n_6 + k)!(n_3 - k)!(n_4 - n_5 - n_6 + k)!(n_5 - k)!(n_6 - k)!k!\end{array}}}$$

where $l = \max(0, -n_1 + n_3 + n_5, -n_2 + n_3 + n_6, -n_4 + n_5 + n_6)$ and $u = \min(n_0 - n_1 - n_2 + n_3 - n_4 + n_5 + n_6, n_3, n_5, n_6)$.

This last formula allows us to check that $\frac{1}{n}\mu$ is not, in general, uniquely defined by $\mathcal{C}_n = \left(\mathcal{K}, \frac{1}{n}\mathbf{g}_{|\overline{\mathcal{K}}}\right)$. Indeed, the two set of values for n_i presented in table 4.2 correspond to a same constraint system yet do not yield the same value for $\frac{\mu(h_{I_7})}{n}$.

	Case 1	Case 2
n_0	12	24
n_1	7	14
n_2	8	16
n_3	4	8
n_4	9	18
n_5	5	10
n_6	6	12
$\frac{\mu(h_{I_7})}{n}$	0.241	0.237

Table 4.2: Finite constraints corresponding to a same constraint system.

This remark is important because it shows that the finite approach does not allow to define a hypothetical value for a probability distribution in general : the number of transactions must be defined. As such it does not provide for a generalization of the independence model (which can be defined regardless of the number of transactions) even though we do obtain the same model as the independence model when considering $\mathcal{K} = \mathcal{A}$.

4.1.2.2 Computing μ

Another one of the issues with the finite approach is the difficulty in computing the value of μ. Indeed, if we set aside some trivial cases such as the one corresponding to the independence model for which the formula simplifies easily, computing μ directly from equation 4.3 becomes practically infeasible as soon as n or m are too large. This is due to the combinatorial nature of this formula which contains many factorials. In fact, even in the particular case that we have described previously in which all proper subitemsets of an itemset are known (which can be considered an easy case because the number of liberties for $\mathbf{h}$ is equal to one), the formula cannot be reasonably computed if $m \geq 3$ and $n \times m \geq 10^3$. Therefore, other means for computing μ must be envisaged.

One alternative approach is to use randomization methods in order to determine an approximate value for μ. Such methods have been considered in itemset mining for a similar yet distinct problem (see [HOV+09]) in which the randomization method simulates a uniform distribution on all datasets of a given size that share the same row and column margins as a given dataset as well as constraints on the values of some itemsets. Such methods can be slightly more scalable than a direct computation but the gain is still limited and, given the results on complexity in [HOV+09], they cannot be reasonably computed if $m \geq 3$ and $n \times m \geq 10^6$. Furthermore, there is no reason to believe that removing the constraints on the row and column margins would help in this respect and more likely the opposite as the methods suggested are based on methods for randomly generating matrices based on their row and column margins. In any case, the size of the datasets that may be considered for such methods remain quite small compared to the Big Data considered in data mining processes and much too small with regards to the aim of discovering scientific knowledge in the data as presented in section 3.7.

Another means to approximate μ is through the MaxEnt model defined by the corresponding constraint system. Indeed, as we will show in section 4.1.3, $\frac{1}{n}\mu$ converges towards the distribution given by this MaxEnt model and this limit may be used to approximate μ. As a matter of fact, it is the observation of this convergence on specific examples that led us first to the invention of the notion mutual constrained independence. Moreover, we will show that this value is arguably a more relevant theoretical choice than the measure μ which is tied to the number of transactions.

4.1.3 Asymptotic approach

4.1.3.1 MCI convergence theorem

The main principle behind the asymptotic approach is that, when considering finite constraints all corresponding to a same constraint system (or at least corresponding to a converging sequence of constraint systems), the sequence of probability distributions resulting from finite approaches converges towards a limit. This is formalized through the following mathematical result.

Theorem 4.1.1 (MCI convergence theorem). *Given a constraint system $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$ on $\mathcal{B}$, there exists a unique probability measure $\mathbf{p}$ such that, for any sequence of functions $\left(\mathbf{g}^{(k)}_{|\overline{\mathcal{K}}}\right)_{k\in\mathbb{N}}$, the three following conditions:*

- *$\forall k \in \mathbb{N}$, $\mathbf{g}^{(k)}_{|\overline{\mathcal{K}}}$ is the restriction to $\overline{\mathcal{K}}$ of a measure on $\mathcal{B}$ with integer values;*
- *$g^{(k)}_{\top} \underset{k\to+\infty}{\longrightarrow} +\infty$;*
- *$\frac{1}{g^{(k)}_{\top}}\mathbf{g}^{(k)}_{|\mathcal{K}} \underset{k\to+\infty}{\longrightarrow} \mathbf{f}_{|\mathcal{K}}$;*

imply that $\frac{1}{g^{(k)}_{\top}}\mu^{(k)} \underset{k\to+\infty}{\longrightarrow} \mathbf{p}$, where $\mu^{(k)}$ is the measure defined by $\left(\overline{\mathcal{K}}, \mathbf{g}^{(k)}_{|\overline{\mathcal{K}}}\right)$ as in section 4.1.2.

4.1.3.2 Model justification

Assuming the validity of theorem 4.1.1 (the proof of which is provided in section 4.1.3.3), we can consider $\mathbf{p}$ to represent the objective hypothesis regarding $\mathbf{f}$ given the knowledge provided by the constraint system $\mathcal{C}$ as described by the problem statement in section 4.1.1.3.

In comparison to the answer provided by the finite approach, this answer is more satisfying theoretically in several respects. Indeed, in many cases the transactions observed in a dataset are but a sample of a much larger, potentially infinite pool of transactions. This is notably the case if the aim is to use the observed dataset to extrapolate about other unobserved datasets and, in particular, if the data is seen as being generated by a random variable which we aim to describe. In such a case, a hypothesis on the distribution of this random variable is better defined through this asymptotic behavior. Note also that, as $\mathbf{p}$ is defined uniquely by the constraint system, this approach provides for a true generalization of the notion of independence as we will formalize with the definition of mutual constrained independence. On a practical note, as **p** is not determined by any given number of transactions, the complexity for computing this probability measure is not determined by the number of transactions in a dataset. This allows to consider truly big data, at least in terms of the number of transactions as the number m of items must still be taken into account.

As we will make explicit in section 4.1.5, the link between **p** and MaxEnt models further justifies the use of this asymptotic approach.

4.1.3.3 Proof of the convergence theorem

Our proof of theorem 4.1.1 is a constructive one which allows to characterize **p**. Hence, we will at the same time give the proof to a stronger version of this theorem. We start by setting up some notions which will be useful for the characterization of **p**.

Preliminary step 1: Reduced transfer matrix. Recall that the aim is to define the probability measure **p** from a constraint system $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$ where $\mathcal{K} \subset \mathcal{I}$ is a set of itemsets. In the following, we will use the matrix T to transfer this question around $\mathcal{I}$ towards Ω, where it is more easily answered. We will then bring the problem back to $\mathcal{I}$. For this purpose, we introduce the notion of reduced transfer matrix and constraint vector.

Consider a constraint system $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$ on $\mathcal{B}$. We define the **reduced transfer matrix** $T_{\mathcal{K}}$ to be the submatrix of T composed of the lines of T corresponding to the elements in $\mathcal{K}$ and the **constraint vector** K to be the column vector with coordinates equal to f_{I_k} for all $I_k \in \mathcal{K}$. Now, for any probability measure **g**, we see that **g** satisfies $\mathcal{C}$ if and only if $T_{\mathcal{K}} X_{\mathbf{g}} = K$.

Table 4.3 gives an example of a constraint system and its corresponding matrix equation for $m = 3$. The constraints are given here on three itemsets: a_2, a_3 and $a_1 \wedge a_2 \wedge a_3$.

$$\begin{array}{c|c} X \in \mathcal{K} & f_X \\ \hline \top & 1 \\ a_3 & 1/2 \\ a_2 & 1/3 \\ a_1 \wedge a_2 \wedge a_3 & 1/5 \end{array} \longleftrightarrow \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} g_0 \\ \vdots \\ g_7 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/5 \end{bmatrix}$$

Table 4.3: A constraint system and its corresponding matrix equation.

As we will make explicit with theorem 4.1.3, the kernel of the reduced transfer matrix plays a significant role in obtaining the solution $\mathbf{p}$ to our problem. We can notice here that we can obtain a basis $\mathcal{B}_{\mathcal{K}}$ of $\mathrm{Ker}(T_{\mathcal{K}})$ by considering the columns of T^{-1} which correspond to the lines removed from the matrix T. Figure 4.2 gives the basis $\mathcal{B}_{\mathcal{K}}$ defined by the columns of T^{-1} for the constraint system given as an example in Table 4.3.

$$\mathcal{B}_{\mathcal{K}} = \left(\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right)$$

Figure 4.2: The basis $\mathcal{B}_{\mathcal{K}}$ of $\mathrm{Ker}(T_{\mathcal{K}})$ with $T_{\mathcal{K}}$ as in Table 4.3.

Preliminary step 2: Largest derivable constraint system. In order to prevent issues related to boundary conditions, we distinguish between the information that can be obtained directly through mathematical properties from the rest, as described in section 4.1.1.4. This comes down to the same problem as distinguishing between derivable and non-derivable itemsets [CG02]. For this purpose, we introduce the notions of derivable constraint system and largest derivable constraint system.

Definition 4.1.1. Let $\mathcal{C} = (\mathcal{K}, \mathbf{f}_{|\mathcal{K}})$ be a constraint system on $\mathcal{B}$. A **derivable constraint system** of $\mathcal{K}$ is a constraint system $\mathcal{C}' = (\mathcal{K}', \mathbf{f}'_{|\mathcal{K}})$ such that the probability measures on $\mathcal{B}$ that satisfy $\mathcal{C}$ are exactly those that satisfy $\mathcal{C}'$.

Notice that, if $\mathcal{C}'$ is a derivable constraint system of $\mathcal{C}$, then we can define the union constraint system $\mathcal{C}'' = (\mathcal{K}'', \mathbf{f}''_{|\mathcal{K}''})$ by $\mathcal{K}'' = \mathcal{K} \cup \mathcal{K}'$, $\mathbf{f}''_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$ and $\mathbf{f}''_{|\mathcal{K}'} = \mathbf{f}'_{|\mathcal{K}'}$. Furthermore, $\mathcal{C}''$ is a derivable constraint system of $\mathcal{C}$. Therefore, we can define a **largest derivable constraint system** (**LDCS**) $\mathcal{C}^* = (\mathcal{K}^*, \mathbf{f}^*_{|\mathcal{K}^*})$ of $\mathcal{C}$ by considering the union of $\mathcal{C}$ with all its derivable constraint systems. We say that the LDCS is **complete** if $\mathcal{C}^* = \mathcal{I}$ and **incomplete** otherwise.

In terms of linear equations, a probability measure $\mathbf{g}$ satisfying $\mathcal{C}$ corresponds to a vector $X_{\mathbf{g}}$ of $[0,1]^{2^m}$ such that $T_{\overline{\mathcal{K}}} X_{\mathbf{g}} = K$. The set of all probability measures satisfying $\mathcal{C}$ is therefore the convex polytope of $\mathbb{R}^{2^m}$ defined as the intersection of the hypercube $[0,1]^{2^m}$ and the affine space of equations $T_{\overline{\mathcal{K}}}X = K$. A constraint system $\mathcal{C}'$ is a derivable constraint system of $\mathcal{C}$ if and only if the polytope defined as the intersection of the hypercube $[0,1]^{2^m}$ and the affine space of equations $T_{\overline{\mathcal{K}'}}X = K'$ is the same as the one for $\mathcal{C}$. Hence, the largest derivable constraint system $\mathcal{C}^*$ corresponds to the smallest affine space such that the intersection with the polytope of probability measures gives the same convex polytope as for $\mathcal{C}$.

$I \in \mathcal{K}$	f_I
a_3	$1/2$
a_2	$1/2$
$a_2 \wedge a_3$	$1/6$
a_1	$1/2$
$a_1 \wedge a_3$	$1/6$
$a_1 \wedge a_2$	$1/6$

$\longrightarrow$

$I \in \mathcal{K}^*$	f_I^*
$\top$	1
a_3	$1/2$
a_2	$1/2$
$a_2 \wedge a_3$	$1/6$
a_1	$1/2$
$a_1 \wedge a_3$	$1/6$
$a_1 \wedge a_2$	$1/6$
$a_1 \wedge a_2 \wedge a_3$	0

Table 4.4: A complete LDCS

$I \in \mathcal{K}$	f_I
$\top$	1
a_3	$1/2$
a_2	$1/2$
a_1	$1/3$
$a_1 \wedge a_2 \wedge a_3$	$1/3$

$\longrightarrow$

$I \in \mathcal{K}^*$	f_I^*
$\top$	1
a_3	$1/2$
a_2	$1/2$
a_1	$1/3$
$a_1 \wedge a_3$	$1/3$
$a_1 \wedge a_2$	$1/3$
$a_1 \wedge a_2 \wedge a_3$	$1/3$

Table 4.5: An incomplete LDCS

In Table 4.4 and Table 4.5, we give examples of constraint systems and their corresponding largest derivable constraint systems. In Table 4.4, the LDCS is complete. This means that there is only one probability measure on $\mathcal{B}$ which satisfies the constraints. In Table 4.5, the LDCS is incomplete. There is therefore an infinite number of probability measures on $\mathcal{B}$ which satisfy these constraints.

Preliminary step 3: Equations. As we will demonstrate in the proof to theorem 4.1.3, the limit in 4.1.1 is obtained as the solution to two easily defined equations which we present in this section. The variable in these equations is a vector $X = \begin{bmatrix} x_0 \\ \vdots \\ x_d \end{bmatrix}$ in $[0,1]^{d+1}$ The solution corresponds to the vector $\begin{bmatrix} p_0 \\ \vdots \\ p_d \end{bmatrix}$, allowing to define the probability measure $\mathbf{p}$. We will also consider the vector $\underline{\ln}(X) = \begin{bmatrix} \underline{\ln}(x_0) \\ \vdots \\ \underline{\ln}(x_d) \end{bmatrix}$, where $\underline{\ln} : [0,+\infty) \to \mathbb{R}[\infty];\ x \mapsto \ln(x)$ if $x \neq 0$ and $-\infty$ if $x = 0$.

Lemma 4.1.2. *Consider $\mathcal{C}$, $\mathcal{C}^*$, $T_{\mathcal{K}^*}$ and K^*, with notations as above. Then, there exists at most one vector $X = \begin{bmatrix} x_0 \\ \vdots \\ x_d \end{bmatrix}$ in $[0,1]^{d+1}$ such that:*

$$T_{\mathcal{K}^*}X = K^* \quad \text{ and } \quad \underline{\ln}\,(X) \in \operatorname{Ker}\,(T_{\mathcal{K}^*})^{\perp}$$

Proof. Suppose X and Y are two such vectors. Then $Y - X \in \operatorname{Ker}(T_{\mathcal{K}^*})$ and $(Y-X)^T\underline{\ln}(X) = (Y-X)^T\underline{\ln}(Y) = 0$. Therefore, $Y^T\underline{\ln}(X) = X^T\underline{\ln}(X)$ and $X^T\underline{\ln}(Y) = Y^T\underline{\ln}(Y)$. As $X^T\underline{\ln}(X) \in \mathbb{R}$, we get $Y^T\underline{\ln}(X) \in \mathbb{R}$. Therefore $y_i = 0$ when $x_i = 0$. By symmetry, we get $x_i = 0 \iff y_i = 0$. We will therefore limit ourselves to the case where $y_i \neq 0$ for all i as the other indices may be dropped for our current purposes.

Define the function $\varphi_Y : (0,1]^{d+1} \to \mathbb{R};\ Z \mapsto Z^T\underline{\ln}(Z)$. We will consider the problem of minimizing φ under the constraint that $T_{\mathcal{K}^*}X = K^*$. Via the method of Lagrange multipliers we have the following necessary condition for a local optimum: $\nabla\varphi(Z) \in \operatorname{Im}(T_{\mathcal{K}^*}^T)$. Now, on the one hand, $\nabla\varphi(Z) = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \underline{\ln}(Z)$ and, on the other hand, as we are in finite dimension, $\operatorname{Im}(T_{\mathcal{K}^*}^T) = \operatorname{Ker}(T_{\mathcal{K}^*})^{\perp}$. Furthermore, $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \operatorname{Im}(T_{\mathcal{K}^*}^T)$, so the condition becomes $\underline{\ln}(Z) \in \operatorname{Ker}(T_{\mathcal{K}^*})^{\perp}$. By the strict concavity of φ, we conclude on the uniqueness of such an optimum thus obtaining the desired result. □

Strong version of Theorem 4.1.1 and proof. Lemma 4.1.2 is central in the proof we provide to Theorem 4.1.1. As stated previously, this proof is constructive, leading to the following stronger result.

Theorem 4.1.3 (MCI convergence theorem, strong version)**.** *Let $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$ be a constraint system on $\mathcal{B}$ and $\left(\mathbf{g}^{(k)}_{|\overline{\mathcal{K}}}\right)_{k\in\mathbb{N}}$ be a sequence of functions satisfying the three following conditions:*

- $\forall k \in \mathbb{N}$, $\mathbf{g}^{(k)}_{|\overline{\mathcal{K}}}$ *is the restriction to $\overline{\mathcal{K}}$ of a measure on $\mathcal{B}$ with integer values;*
- $g^{(k)}_{\top} \underset{k\to+\infty}{\longrightarrow} +\infty$*;*
- $\frac{1}{g^{(k)}_{\top}} \mathbf{g}^{(k)}_{|\mathcal{K}} \underset{k\to+\infty}{\longrightarrow} \mathbf{f}_{|\mathcal{K}}$.

Consider:

- $X_k = \frac{1}{g^{(k)}_{\top}} \begin{bmatrix} \mu^{(k)}_0 \\ \vdots \\ \mu^{(k)}_d \end{bmatrix}$ *where $\mu^{(k)}$ is the average finite measure defined by $\left(\overline{\mathcal{K}}, \mathbf{g}^{(k)}_{|\overline{\mathcal{K}}}\right)$ as in section 4.1.2;*
- $\mathcal{C}^* = (\mathcal{K}^*, \mathbf{f}^*_{|\mathcal{K}^*})$ *the largest derivable constraint system of $\mathcal{C}$;*
- *and $T_{\mathcal{K}^*}$ the reduced transfer matrix as defined above.*

Then $(X_k)_{k\in\mathbb{N}}$ converges towards the unique vector $X \in [0,1]^{d+1}$ such that:

$$T_{\mathcal{K}^*} X = K^* \quad \text{ and } \quad \underline{\ln}\,(X) \in \operatorname{Ker}\left(T_{\mathcal{K}^*}\right)^{\perp}$$

Proof. As $(X_k)_{k\in\mathbb{N}}$ is a sequence of vectors of $[0,1]^{d+1}$, which is a compact space, it is sufficient to show that all convergent subsequences of $(X_k)_{k\in\mathbb{N}}$ converge towards the same limit. Rather than considering a subsequence, we will consider, with no loss of generality, that $(X_k)_{k\in\mathbb{N}}$ converges towards a limit and show that this limit is uniquely defined by $\mathcal{C}$.

Let X be the limit of $(X_k)_{k\in\mathbb{N}}$. We know that, for all $k \in \mathbb{N}$, $T_{\mathcal{K}^*} X_k = K^*_k$, and that $K_k \underset{k\to+\infty}{\longrightarrow} K$. Hence, by continuity, $T_{\mathcal{K}^*} X = K^*$, which is the first of the two equations needed. Obtaining the second one is slightly more complex and is detailed in the following.

Let Y be a vector from the basis $\mathcal{B}_{\mathcal{K}^*}$ of $\operatorname{Ker}(T_{\mathcal{K}^*})$ as defined previously. We know that the coordinates of Y are in $\{-1, 0, 1\}$ and that $\sum_{i=0}^{d} y_i = 0$, so we can set $N_Y = \sum_{y_i=1} y_i = -\sum_{y_i=-1} y_i$. Let $n = g^{(k)}_{\top}$ and consider k so that $n \geq N_Y$.

We consider the space $\mathcal{D}_k$ of all datasets of size $n \times m$ satisfying the constraints given by $\mathbf{g}^{(k)}_{|\mathcal{K}}$. If we look at a dataset in $\mathcal{D}_k$, each line of the dataset corresponds uniquely to an element ω_i of Ω. Consider the subsets $\mathcal{D}_{k,Y_+}$ (resp. $\mathcal{D}_{k,Y^-}$) of $\mathcal{D}_k$ of all matrices for which each of the N_Y first lines correspond to one of the ω_i such that $y_i = 1$ (resp. $y_i = -1$). Then $|\mathcal{D}_{k,Y^+}| = |\mathcal{D}_{k,Y^-}|$. Notice here that $x_i \neq 0$ if $y_i \neq 0$. Indeed, if $x_i = 0$, this means that the convex polytope of the vectors Z which correspond to probability measures satisfying $\mathcal{K}^*$ is contained in the affine space defined by the equation $z_i = 0$. As the direction of this affine space is $\mathrm{Ker}(T_{\mathcal{K}^*})$, then for any vector Y from a basis of $\mathrm{Ker}(T_{\mathcal{K}^*})$, $y_i = 0$.

Furthermore, we will show that we have $|\mathcal{D}_{k,Y^+}|/N_Y!|\mathcal{D}_k| \underset{k\to+\infty}{\longrightarrow} \prod_{y_i=1} x_i$ and $|\mathcal{D}_{k,Y^-}|/N_Y!|\mathcal{D}_k| \underset{k\to+\infty}{\longrightarrow} \prod_{y_i=-1} x_i$. To prove this point, we will consider a probability with uniform distribution on the finite set of matrices $\mathcal{D}_k$. We note this probability Prob_k. Let $[L_j = \omega_i]$ denote the set of matrices of $\mathcal{D}_k$ for which the j-th row corresponds to ω_i and $[|\omega_i| = l]$ the set of matrices of $\mathcal{D}_k$ for which exactly l rows correspond to ω_i. We can hereafter express our previous quantities as probabilities. For the first of the two fractions, this gives: $|\mathcal{D}_{k,Y^+}|/N_Y!|\mathcal{D}_k| = \mathrm{Prob}_k\left(\bigcap_{j=1}^{N_Y}\left[L_j = \omega_{\sigma(j)}\right]\right)$ where $\sigma : [\![1, N_Y]\!] \to \{i \in \mathbb{N} \mid y_i = 1\}$ is any bijection. Note that we only need to consider one of the two cases as the following demonstration is easily transposed to the other case. Furthermore, by the definition of X_k, for all $j \in [\![1, n]\!]$ and $i \in [\![0, d]\!]$, we have $\mathrm{Prob}_k\left(L_j = \omega_i\right) = x_{k,i}$ (where $x_{k,i}$ is the i-th coordinate of X_k). In addition, as $X_k \underset{k\to+\infty}{\longrightarrow} X$, we have $\mathrm{Prob}_k\left(L_j = \omega_i\right) \underset{k\to+\infty}{\longrightarrow} x_i$. Hence, to prove our point, it is sufficient to show that $\mathrm{Prob}_k\left(\bigcap_{j=1}^{N_Y}\left[L_j = \omega_{\sigma(j)}\right]\right) - \prod_{j=1}^{N_Y}\mathrm{Prob}_k\left(L_j = \omega_{\sigma(j)}\right) \underset{k\to+\infty}{\longrightarrow} 0$ for any bijection σ defined as previously. As this is obvious for $N_Y = 1$, let us consider that $N_Y \geq 2$. The convergence towards 0 corresponds to the following intuitive idea. If N_Y is fixed while we consider larger and larger datasets (i.e. larger n), the events that any given one of the N_Y first rows corresponds to any given ω_i become gradually independent because the incidence that the value of one single row has on another single row becomes gradually negligible. We show this is true for two rows and the rest follows easily by iteration. Let $i \neq j$ such that $y_i = y_j = 1$ and $H_k = \mathrm{Prob}_k\left([L_1 = \omega_i] \cap [L_2 = \omega_j]\right) - \mathrm{Prob}_k\left(L_1 = \omega_i\right)\mathrm{Prob}_k\left(L_2 = \omega_j\right)$. We see that $H_k = \mathrm{Prob}_k\left(L_1 = \omega_i\right)\left(\mathrm{Prob}_k\left(L_2 = \omega_j \mid L_1 = \omega_i\right) - \mathrm{Prob}_k\left(L_2 = \omega_j\right)\right)$.

But we also have $\text{Prob}_k\left(L_2=\omega_j \mid L_1=\omega_i\right) = \text{Prob}_k\left(L_2=\omega_j \mid |\omega_i|\geq 1\right) = \text{Prob}_k\left(L_2=\omega_j\right)\frac{\text{Prob}_k(|\omega_i|\geq 1 \mid L_2=\omega_j)}{\text{Prob}_k(|\omega_i|\geq 1)} = \text{Prob}_k\left(L_2=\omega_j\right)\frac{1-\text{Prob}_k(|\omega_i|=0 \mid L_2=\omega_j)}{1-\text{Prob}_k(|\omega_i|=0)}$. Hence, $H_k = \text{Prob}_k\left(L_1=\omega_i\right)\text{Prob}_k\left(L_2=\omega_j\right)\left[\frac{1-\text{Prob}_k(|\omega_i|=0 \mid L_2=\omega_j)}{1-\text{Prob}_k(|\omega_i|=0)}-1\right]$. And both $\text{Prob}_k\left(|\omega_i|=0\right) \underset{k\to+\infty}{\longrightarrow} 0$ and $\text{Prob}_k\left(|\omega_i|=0 \mid L_2=\omega_j\right) \underset{k\to+\infty}{\longrightarrow} 0$. Therefore, $H_k \underset{k\to+\infty}{\longrightarrow} 0$, quod erat demonstrandum. Note that the previous demonstration is only valid because, if $y_i=1$, both $x_i \neq 0$ and the sequence $\left(x_{k,i}\right)_{k\geq 1}$ is strictly positive for large enough k.

Now, the results of the two previous paragraphs can be combined and we get $\prod\limits_{y_i=1} x_i = \prod\limits_{y_i=-1} x_i$. Hence, $\sum\limits_{y_i=1}\ln(x_i) - \sum\limits_{y_i=-1}\ln(x_i) = 0$, which can also be written $Y^T\underline{\ln}\left(X\right)=0$. As this is true for all Y from the basis $\mathcal{B}_{\mathcal{K}^*}$ of $\text{Ker}(T_{\mathcal{K}^*})$, this gives $\underline{\ln}\left(X\right) \in \text{Ker}\left(T_{\mathcal{K}^*}\right)^{\perp}$.

We conclude from lemma 4.1.2 that X is uniquely defined by $\mathcal{K}$ which ends the proof. □

Note that this result is not limited to the case in which $\mathbf{f}_{|\mathcal{K}}$ is necessarily the restriction of a probability measure corresponding to an empirical distribution. Indeed, the density of the rationals in the reals, together with the continuity of the functions defining the equations, ensure that it still holds if $\mathbf{f}_{|\mathcal{K}}$ is the restriction of any probability measure on $\mathcal{B}$. More precisely, such a condition on $\mathbf{f}_{|\mathcal{K}}$ is only necessary for defining constraint systems in the finite approach and can be omitted when defining the asymptotic constraint system here.

4.1.4 Definition of MCI

In section 4.1.1, we have presented an approach for formulating an objective hypothesis on the values of a probability measure for the distribution of items given constraints on the values of this measure for certain itemsets. In section 4.1.3, we have shown that this approach leads to a solution which we can characterize mathematically as the unique solution to a system of equations. Conversely, this characterization may be seen as a property of distributions of items which indicates how the items relate to each other: tied by a certain number of interrelations and entirely free otherwise. Because this characterization corresponds to the intuitive notion of independence under constraint and because it generalizes the mathematical notion of mutual independence, we have named this property **mutual constrained independence**. We give its formal definition below.

Definition 4.1.2 (Mutual constrained independence)**.** Consider a probability measure $\mathbf{p}$ on $\mathcal{B}$ and a set of itemsets $\mathcal{K} \subset \mathcal{I}$. Let $X = \begin{bmatrix} p_0 \\ \vdots \\ p_d \end{bmatrix}$ be the vector representation of $\mathbf{p}$ in the basis Ω. We say that the items $a_1, ..., a_m$ are mutually constrainedly independent in $\mathcal{B}$ with regards to the constraints defined by $\mathcal{K}$, if and only if $\underline{\ln}(X) \in \mathrm{Ker}(T_{\mathcal{K}^*})^{\perp}$. *(See notations preceding lemma 4.1.2 for the definition of* $\underline{\ln}$*.)*

Note that this definition is not restricted to the context of itemsets and to applications in data mining. It applies more generally to the field of probabilities, as any finite family of events $A_1, ..., A_m$ of a probability space can naturally be associated to a set of items $a_1, ..., a_m$. It is a straight forward generalization of the notion of mutual independence. Indeed, the mutual independence of m items corresponds to the mutual constrained independence of these items with regards to $\mathcal{K} = \{a_1, ..., a_m\}$. It is therefore quite natural to consider statistical tests for mutual constrained independence similarly as the well known tests of independence performed by statisticians. This implies that one might define a statistical MCI model from a dataset in the same fashion as one defines an independence model.

Definition 4.1.3 (MCI model)**.** Let $\mathbf{f}$ be a probability measure on $\mathcal{B}$ defined as the empirical distribution of a dataset of transactions on items and $\mathcal{K} \subset \mathcal{I}$ be a set of itemsets. The MCI model for the data defined by $\mathcal{K}$ is the probability measure $\mathbf{p}$ defined by its vector representation $X = \begin{bmatrix} p_0 \\ \vdots \\ p_d \end{bmatrix}$, such that:

$$T_{\mathcal{K}^*} X = K^* \quad \text{and} \quad \underline{\ln}(X) \in \mathrm{Ker}(T_{\mathcal{K}^*})^{\perp}$$

where K^* is the vector representation of $\mathbf{f}$ reduced to $\mathcal{K}^*$.

4.1.5 MCI and maximum entropy

As stated in the introduction, the notion we have defined is related to MaxEnt models. This is made explicit in the following theorem.

Theorem 4.1.4. *Consider notations as in section 4.1.4. Then $a_1, ..., a_p$ are mutually constrainedly independent in $\mathcal{B}$ with regards to the constraints defined by $\mathcal{K}$ if and only if*

$$X = \underset{\substack{T_{\mathcal{K}^*} Z = K^* \\ Z \in [0,1]^{2^p}}}{\arg\max} H(Z)$$

where H is the information entropy function and K^* is the reduction of X to $\mathcal{K}^*$.

Proof. The proof to this theorem is already contained in the proof to lemma 4.1.2. Indeed, we have shown the unicity of X (which corresponds to the solution to the mutual constrained independence problem) by showing that, if X exists, it is the minimum of a function which is none other than the opposite of the entropy function. As we have shown its existence in Theorem 4.1.3, it coincides therefore with this optimum. □

This result implies that MCI models are MaxEnt models, where the maximization of the information entropy is constrained by the values of the empirical frequencies of the itemsets within the constrained set $\mathcal{K}$. Such MaxEnt models have already been considered and employed successfully in the field of data mining (see section 2.3.3). As such, MCI models cannot be considered to be entirely novel models. However, the MCI approach towards the definition of these models is new and this novel approach brings forth a number of new perspectives. First, it brings further understanding to the maximum entropy principle as we have described in section 3.3.3 which helps to strengthen the use of MaxEnt models in mathematical modelings. Second, the MCI characterization of these models allows to envisage new properties of these models and, most notably, new methods for computing them, which we present in section 4.2.

4.2 MCI Models: properties and computation

In this section, we discuss some properties of MCI models and propose a novel approach for computing such models. We start by considering the specific case in which we consider all the proper subsets of $\mathcal{A}$ as the constrained set. We then propose a method based on algebraic geometry allowing to compute any MCI model. Finally, we use this method to provide exact algebraic formulas for all MCI models when $m \leq 4$.

4.2.1 $\mathcal{K} = \mathcal{I} \setminus \{I_d\}$

The case of computing an MCI model in which $\mathcal{K}$ is the set of all proper subsets of $\mathcal{I}$ is relatively easy to consider because the number of degrees of liberty defined by the linear constraints is equal to one. Chronologically, it is the first case which was studied in the course of this thesis and is the focus of one of our published papers [DBLL15]. The corresponding MaxEnt models have also been considered previously in the itemset literature but strictly from the perspective of an optimization problem [Meo00, Tat08].

4.2.1.1 Algebraic expression of the model

Consider a constraint system $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$ such that $\mathcal{K} = \mathcal{I} \setminus \{I_d\}$. Let $\mathbf{p}$ be the corresponding MCI model and let $X = \begin{bmatrix} p_0 \\ \vdots \\ p_d \end{bmatrix}$ be its vector representation in the basis Ω. Then:

$$TX = \begin{bmatrix} f_{I_0} \\ \vdots \\ f_{I_{d-1}} \\ p_d \end{bmatrix}$$

As in section 4.1.2.1, we can express all the coordinates of X as of function of p_d, like so:

$$X = C + p_d V$$

where:

$$C = T^{-1} \begin{bmatrix} f_0 \\ \vdots \\ f_{d-1} \\ 0 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} (-1)^{(\mathbf{d}-\mathbf{0})\cdot\mathbf{d}} \\ (-1)^{(\mathbf{d}-\mathbf{1})\cdot\mathbf{d}} \\ \vdots \\ (-1)^{(\mathbf{d}-\mathbf{d})\cdot\mathbf{d}} \end{bmatrix}$$

Furthermore, we know that $p_d \in [l, u]$ where:

$$l = \max_{\substack{i \in [\![0,d]\!] \\ v_i = 1}} (-c_i) \qquad \text{and} \qquad u = \min_{\substack{i \in [\![0,d]\!] \\ v_i = -1}} (c_i)$$

There are two possibilities:

- either $l = u$, which implies $p_d = l$ and the model is determined immediately (this corresponds to the case in which $\mathcal{K}^* = \mathcal{I}$);

- or $l < u$, in which case we use the MCI characterization to determine $\mathbf{p}$ (in this case, $\mathcal{K}^* = \mathcal{K}$).

In the second case, the MCI characterization translates directly to the following condition:

$$\prod_{\substack{i=0 \\ v_i=1}}^{d} p_i = \prod_{\substack{i=0 \\ v_i=-1}}^{d} p_i$$

Hence, p_d is a real root of the polynomial Q defined as follows:

$$Q(x) = \prod_{\substack{i=0 \\ v_i=1}}^{d} (c_i + x) \quad - \prod_{\substack{i=0 \\ v_i=-1}}^{d} (c_i - x)$$

Furthermore, this is necessarily the only real root of Q which lies in the interval $[l, u]$ as this would otherwise imply that the MCI characterization does not uniquely define the MCI model. Hence, this root may be computed from the expression of Q via numerical methods.

Note that Q is a monic polynomial with degree exactly equal to $\frac{d-1}{2}$. Indeed, if we factor each product and group by degree, we get:

$$Q(x) = \left(x^{\frac{d+1}{2}} + \left(\sum_{\substack{i=0 \\ v_i=1}}^{d} c_i \right) x^{\frac{d-1}{2}} + R_+(x) \right) - (-1)^{\frac{d+1}{2}} \left(x^{\frac{d+1}{2}} - \left(\sum_{\substack{i=0 \\ v_i=-1}}^{d} c_i \right) x^{\frac{d-1}{2}} + R_-(x) \right)$$

where R_+ and R_- have degree $\frac{d-3}{2}$ or less. As $d = 2^m - 1$, $\frac{d+1}{2} = 2^{m-1}$ is even for all $m \geq 2$. Hence:

$$Q(x) = \left(\sum_{i=0}^{d} c_i \right) x^{\frac{d-1}{2}} + R(x)$$

where $R = R_+ - R_-$ has degree $\frac{d-3}{2}$ or less. Finally, as:

$$\sum_{i=0}^{d} c_i = \sum_{i=0}^{d} (p_i - p_d v_i) = \sum_{i=0}^{d} p_i - p_d \sum_{i=0}^{d} v_i = 1 - p_d \times 0 = 1$$

we find that:

$$Q(x) = x^{\frac{d-1}{2}} + R(x)$$

For $m = 3$, we can see that p_7 is the only real root within the interval $[l, u]$ of the polynomial Q_3 defined by:

$$Q_3(x) = x^3 + \alpha x + \beta x + \gamma$$

where:

$$\begin{array}{lcl}
\alpha & = & f_{I_4} f_{I_2} + f_{I_4} f_{I_1} + f_{I_2} f_{I_1} - (f_{I_4} + 1) f_{I_3} - (f_{I_2} + 1) f_{I_5} - (f_{I_1} + 1) f_{I_6} \\
\beta & = & f_{I_4} f_{I_3} (f_{I_3} - f_{I_2} - f_{I_1}) + f_{I_2} f_{I_5} (f_{I_5} - f_{I_4} - f_{I_1}) + f_{I_1} f_{I_6} (f_{I_6} - f_{I_4} - f_{I_2}) \\
 & & +2 f_{I_6} f_{I_5} f_{I_3} + f_{I_4} f_{I_2} f_{I_1} + f_{I_6} f_{I_5} + f_{I_6} f_{I_3} + f_{I_5} f_{I_3} \\
\gamma & = & f_{I_6} f_{I_5} f_{I_3} (f_{I_4} + f_{I_2} + f_{I_1} - f_{I_6} - f_{I_5} - f_{I_3} - 1) \\
l & = & \max(0, -f_{I_1} + f_{I_3} + f_{I_5}, -f_{I_2} + f_{I_3} + f_{I_6}, -f_{I_4} + f_{I_5} + f_{I_6}) \\
u & = & \min(1 - f_{I_1} - f_{I_2} + f_{I_3} - f_{I_4} + f_{I_5} + f_{I_6}, f_{I_3}, f_{I_5}, f_{I_6})
\end{array}$$

4.2.1.2 Distance to the MCI model

Consider a probability measure $\mathbf{f}$ that defines a constraint system $\mathcal{C} = \left(\mathcal{K}, \mathbf{f}_{|\mathcal{K}}\right)$ such that $\mathcal{K} = \mathcal{I} \setminus \{I_d\}$ and let $\mathbf{p}$ be the associated MCI model. An interesting question is to determine how these two probability measures compare and one of the simplest ways to compare them is to subtract one to the other. As $f_I = p_I$ for all $I \neq I_d$, $\mathbf{f} - \mathbf{p}$ is equal to zero for all items except I_d. We will note $\delta_{\mathcal{A}}$ this difference:

$$\delta(\mathcal{A}) = f_d - p_d$$

Furthermore, for any itemset I in $\mathcal{I}$, we can always consider the localization on I of the problem defined previously (by transposing the entire problem from the set of items $\mathcal{A}$ to the corresponding subset of items $\mathcal{A}_I$), so that we can define a function Δ for all $I \in \mathcal{I}$ by:

$$\Delta(I) = \delta(\mathcal{A}_I)$$

The function Δ may be seen as an objective interestingness measure allowing to determine for local redundancy in itemsets (see section 2.3.3.1). Indeed, it compares the value for the frequency of an itemset to the value given by the local MCI model where the constraints are set by all the proper subitemsets of that itemset. Although we have expressed some theoretical reservations with regards to the use of such measures in objective itemset mining in Chapter 3, such measures do present some practical advantages due to their algorithmic properties and they have been successfully put to use in data mining processes. We exhibit here one such property for Δ which we originally presented in [DBLL15].

Proposition 4.2.1. *For all itemsets $I \in \mathcal{I}$,*

$$|\Delta(I)| \leq \frac{1}{2^{|I|}}$$

where $|I|$ is the number of items in I. Furthermore, there is a dataset for which this upper bound is reached.

Proof. Without loss of generality, we will show that $|\delta(\mathcal{A})| \leq \frac{1}{d+1}$. This proposition is related to a nonlinear optimization problem. It can be solved by using the Karush-Kuhn-Tucker theorem (see, for example, [Rus11], p.116, theorem 3.25).

Let $\mathcal{V}_+ = \{i \in [\![0, d]\!] \mid v_i = 1\}$ and $\mathcal{V}_- = \{i \in [\![0, d]\!] \mid v_i = 1\}$ where v_i are the coordinates of the vector V as defined in section 4.2.1.1 and let $\mathcal{V}_+^* = \mathcal{V}_+ \setminus \{d\}$ and $\mathcal{V}_-^* = \mathcal{V}_- \setminus \{d-1\}$. Consider the functions f, h_0, h_1, ..., h_{d+1} defined for all $z = (x_0, ..., x_d, y_0, ..., y_d) \in \mathbb{R}_+^{2d+2}$ by :

- $f(z) = x_d - y_d$;
- $h_k(z) = x_{d-1} + x_k - (y_{d-1} + y_k), \forall k \in \mathcal{V}_+^*$;
- $h_k(z) = x_d + x_k - (y_d + y_k), \forall k \in \mathcal{V}_-$;
- $h_d(z) = \prod_{i \in \mathcal{V}_+} x_i - \prod_{i \in \mathcal{V}_-} x_i$;
- $h_{d+1}(z) = \left(\sum_{i=0}^{d} x_i\right) - 1$;
- $h_{d+2}(z) = \left(\sum_{i=0}^{d} y_i\right) - 1$.

We will consider the nonlinear optimization problem of minimizing $f(z)$ on $\mathbb{R}_+^{2d+2}$ subject to $h_k(z) = 0$ for all $k \in [\![0, d+2]\!]$.

First, see that for $\tilde{z}$ such that :

- $\tilde{x}_k = \frac{1}{d+1}$ for all $k \in [\![0, d]\!]$;
- $\tilde{y}_k = \frac{2}{d+1}$ for all $k \in \mathcal{V}_+$;
- and $\tilde{y}_k = 0$ for all $k \in \mathcal{V}_-$;

the conditions are satisfied and $f(\tilde{z}) = -\frac{1}{d+1}$.

Let $\hat{z} = (\hat{x}_0, ..., \hat{x}_d, \hat{y}_0, ..., \hat{y}_d)$ be a minimum to the optimization problem (as all the coefficients of z are positive, the conditions given by h_{d+1} and h_{d+2} show that z belongs to the compact set $[0,1]^{2d+2}$ and, hence, that such a minimum exists). Notice that the condition $(h_k(\hat{z}) = 0,\ \forall k \in [\![0, d-1]\!])$ gives $f(\hat{z}) = \hat{x}_k - \hat{y}_k$ for all $k \in \mathcal{V}_+$ and $f(\hat{z}) = \hat{y}_k - \hat{x}_k$ for all $k \in \mathcal{V}_-$. Now, suppose $\hat{x}_k = 0$ for some $k \in [\![0, d]\!]$. The condition $h_d(\hat{z}) = 0$ tells us that there is a couple $(k^+, k^-) \in \mathcal{V}_+ \times \mathcal{V}_-$ such that $x_{k^+} = x_{k^-} = 0$. This gives $y_{k^+} + y_{k^-} = 0$ and, therefore, $y_{k^+} = y_{k^-} = 0$. Hence, $f(\hat{z}) = 0$, which is impossible as $f(\hat{z}) \leq -\frac{1}{d+1}$. Therefore, $\hat{x}_k$ is strictly positive for all k. Now, suppose $\hat{y}_k = 0$ for some $k \in \mathcal{V}_+$. Then $f(\hat{z}) = \hat{x}_k - \hat{y}_k = \hat{x}_k > 0$, which is impossible as previously. Hence, $\hat{y}_k$ is strictly positive for all $k \in \mathcal{V}^+$.

Let us now apply the Karush-Kuhn-Tucker theorem to the optimization problem. The theorem gives the existence of a $z = (x_0, ..., x_d, y_0, ..., y_d) \in \mathbb{R}_+^{2d+2}$ and $d+2$ Lagrange multipliers $\hat{\lambda}_k \in \mathbb{R}$ for $k \in [\![0, d+2]\!]$, such that $z \cdot \hat{z} = 0$ and $z = \nabla f(\hat{z}) + \sum_{k=0}^{d+2} \hat{\lambda}_k \nabla h_k(\hat{z})$. As all coordinates are positive, the orthogonality condition implies that $x_k = 0$ for all k and that $y_k = 0$ for all $k \in \mathcal{V}_+$. This translates as follows:

$$\begin{array}{llcll}
x_k = 0 & \forall k \in \mathcal{V}_+^* & \Longrightarrow & \hat{\lambda}_k + \hat{\lambda}_d \frac{M}{\hat{x}_k} + \hat{\lambda}_{d+1} = 0 & \forall k \in \mathcal{V}_+^* \quad (1) \\
x_k = 0 & \forall k \in \mathcal{V}_-^* & \Longrightarrow & \hat{\lambda}_k - \hat{\lambda}_d \frac{M}{\hat{x}_k} + \hat{\lambda}_{d+1} = 0 & \forall k \in \mathcal{V}_-^* \quad (2) \\
x_d = 0 & & \Longrightarrow & 1 + \sum_{i \in \mathcal{V}_-} \hat{\lambda}_i + \hat{\lambda}_d \frac{M}{\hat{x}_d} + \hat{\lambda}_{d+1} = 0 & (3) \\
x_{d-1} = 0 & & \Longrightarrow & \sum_{i \in \mathcal{V}_+^*} \hat{\lambda}_i + \hat{\lambda}_{d-1} - \hat{\lambda}_d \frac{M}{\hat{x}_{d-1}} + \hat{\lambda}_{d+1} = 0 & (4) \\
y_k = 0 & \forall k \in \mathcal{V}_+^* & \Longrightarrow & -\hat{\lambda}_k + \hat{\lambda}_{d+2} = 0 & \forall k \in \mathcal{V}_+^* \quad (5) \\
y_d = 0 & & \Longrightarrow & -1 - \sum_{i \in \mathcal{V}_-} \hat{\lambda}_i + \hat{\lambda}_{d+2} = 0 & (6)
\end{array}$$

where $M = \prod_{i \in \mathcal{V}_+} \hat{x}_i = \prod_{i \in \mathcal{V}_-} \hat{x}_i$. By combinations, we see that:

$$\begin{array}{lcl}
(1) + (5) & \Longrightarrow & \hat{\lambda}_d \frac{M}{\hat{x}_k} + \hat{\lambda}_{d+1} + \hat{\lambda}_{d+2} = 0 \quad \forall k \in \mathcal{V}_+^* \\
(3) + (6) & \Longrightarrow & \hat{\lambda}_d \frac{M}{\hat{x}_d} + \hat{\lambda}_{d+1} + \hat{\lambda}_{d+2} = 0
\end{array}$$

which gives

$$\hat{\lambda}_d M \left(\frac{1}{\hat{x}_k} - \frac{1}{\hat{x}_{k'}}\right) = 0 \quad \forall\,(k, k') \in (\mathcal{V}_+)^2$$

As $M > 0$, this implies that

$$\left(\hat{\lambda}_d = 0\right) \quad \text{or} \quad \left(\hat{x}_k = \hat{x}_{k'} \quad \forall\,(k, k') \in (\mathcal{V}_+)^2\right)$$

If $\hat{\lambda}_d = 0$, then (1) and (2) give $\hat{\lambda}_k = -\hat{\lambda}_{d+1}$ for all $k \in \mathcal{V}_+^* \cup \mathcal{V}_-^*$. Hence, $\sum\limits_{i \in \mathcal{V}_-} \hat{\lambda}_i = \sum\limits_{i \in \mathcal{V}_-^*} \hat{\lambda}_i + \hat{\lambda}_{d-1} = \sum\limits_{i \in \mathcal{V}_+^*} \hat{\lambda}_i + \hat{\lambda}_{d-1}$. Therefore, (3) $-$ (4) gives $1 = 0$ and, by contradiction, we now know that $\hat{x}_k = \hat{x}_{k'}$ for all $(k, k') \in (\mathcal{V}_+)^2$ or, equivalently, $\hat{x}_k = \hat{x}_d$ for all $k \in \mathcal{V}_+$.

Now, suppose that $y_k = 0$ for some $k \in \mathcal{V}_-$. If $k \neq d - 1$, this implies:

$$-\hat{\lambda}_k + \hat{\lambda}_{d+2} = 0 \qquad (7)$$

If $k = d - 1$, it implies

$$-\sum_{i \in \mathcal{V}_+^*} -\hat{\lambda}_{d-1} + \hat{\lambda}_{d+2} = 0 \qquad (8)$$

In both cases, by combining (2) + (7) or (4) + (8), we get

$$-\hat{\lambda}_d \frac{M}{\hat{x}_k} + \hat{\lambda}_{d+1} + \hat{\lambda}_{d+2} = 0$$

Hence, $\hat{x}_k = -\hat{x}_d < 0$. Again, this is a contradiction and we can conclude that $y_k > 0$ for all $k \in \mathcal{V}_-$, which in turn implies that $\hat{y}_k = 0$ for all $k \in \mathcal{V}_-$.

We have thus shown that both $\hat{x}_k = \hat{x}_d$ for all $k \in \mathcal{V}_+$ and $\hat{y}_k = 0$ for all $k \in \mathcal{V}_-$. Therefore, as $f(\hat{z}) = \hat{x}_k - \hat{y}_k$ for all $k \in \mathcal{V}_+$ and $f(\hat{z}) = \hat{y}_k - \hat{x}_k$ for all $k \in \mathcal{V}_-$, we have $\hat{y}_k = \hat{y}_d$ for all $k \in \mathcal{V}_+$ and $\hat{x}_k = \hat{y}_d - \hat{x}_d$ for all $k \in \mathcal{V}_-$. Moreover, from $h_{d+2}(\hat{z}) = 0$, we get $\frac{d+1}{2}\hat{y}_d = 1$ and from $h_d(\hat{z}) = 0$, $\hat{x}_d^{\frac{d+1}{2}} - (\hat{y}_d - \hat{x}_d)^{\frac{d+1}{2}}$. This solves easily to

$$(\hat{x}_d, \hat{y}_d) = \left(\frac{1}{d+1}, \frac{2}{d+1}\right)$$

This implies that $\hat{z} = \tilde{z}$ and, therefore, $\min f(z) = -\frac{1}{d+1}$. Moreover, the symmetries of the problem induce similarly that $\max f(z) = \frac{1}{d+1}$. Hence,

$$\max |f(z)| = \frac{1}{d+1}$$

This concludes the proof of the proposition as the set of vectors z that satisfy the constraints given by the optimization problem identifies to the set of tuples $(\mathbf{p}, \mathbf{f})$ where $\mathbf{f}$ is a probability measure on $\mathcal{B}$ and $\mathbf{p}$ is the corresponding MCI model where the constraints are taken on all proper subitemsets of I_d. $\square$

This result echoes with the shrinking property described in [CG07] which states that the width of the interval $[l, u]$ (see section 4.2.1.1) shrinks exponentially with m. More precisely, the proposition in [CG07] allows to say that, for two probability measures $\mathbf{f}$ and $\mathbf{g}$ on $\mathcal{B}$ which share the same values for the frequencies of all proper subitemsets of I_d,

$$|f_d - g_d| \leq \frac{1}{2^{m-1}}$$

Similarly, our own proposition shows that, if $\mathbf{p}$ is the MCI model associated to $\mathbf{f}$ and defined by the proper subitemsets of I_d, then

$$|f_d - p_d| \leq \frac{1}{2^m}$$

Such propositions may prove helpful if Δ is used as an objective interestingness measure in a mining process so that only itemsets I satisfying $\Delta \geq \eta$ are considered interesting, for some threshold η. For example, one of the main issues with level wise algorithms is the possible explosion in the number of candidate itemsets from one layer to the next and it is often decided to stop the algorithm at a fixed layer. These proposition can tell us at which layer we can stop, without having to define any candidates on a further layer. Compared to the shrinking property in [CG07], our own proposition allows to stop one layer earlier, which given the explosion of candidates between two layers can represent an important gain. As we have also shown that the bound which we have obtained is the best possible bound, proposition 4.2.1 may also be used to determine the algorithmic complexity of such mining algorithms.

4.2.1.3 Particularity of the MCI approach.

As we have stated previously, we do acknowledge that mathematically equivalent models have been considered in the literature before. Rosa Meo has even presented an interestingness measure in [Meo00] which is mathematically equivalent to Δ. This article, in which she calls p_d the "*maximum independence value*" and Δ the "*dependence value*", aims at defining and computing Δ. However, it is quite interesting to see that, though the focus of her article is on the same mathematical objects as this current section, none of the results which are presented in this section are presented in [Meo00]. In particular, she did not reduce the equations of the optimization problem to its algebraic solution, even for $m = 3$. This is also true for other related articles which we have found within the literature such as [Tat08]. We believe this is linked to the idea that, as the notions in these articles are defined with respect to a problem of entropy maximization, the tools that are considered to compute them come from standard optimization theory.

In comparison, the approach which has led us to the results presented here is entirely different because, when we first proved them, we had not yet established the link between MCI models and MaxEnt models. As such, we can see the MCI approach as a means to envisage MaxEnt models from a different mathematical perspective. In the specific case discussed in this section, it is quite apparent that the model can be computed by considering a polynomial from the MCI point of view. Interestingly enough from an epistemological perspective, the fact that we had not yet made the connection between our models and MaxEnt models allowed us to follow through. Hence, when we considered more general MCI models, we focused on developing a generalized approach for computing them also based on polynomials. We present this generalization in the next section.

4.2.2 Algebraic geometry for computing MCI models

In the particular case for which $\mathcal{K}^* = \mathcal{I} \setminus \{I_d\}$, we have shown that we can algebraically reduce the equations defining the MCI model $\mathbf{p}$ and determine a univariate real polynomial Q together with $d+1$ real affine functions A_i (for each $i \in [\![0, d]\!]$) such that, there exists a unique $t \in \mathbb{R}$ satisfying:

1. $Q(t) = 0$;

2. $\forall i \in [\![0, d]\!]$, $A_i(t) \geq 0$;

and solving this equation gives $\mathbf{p}$ as $p_i = A_i(t)$ for all $i \in [\![0, d]\!]$. Hence, we can easily determine $\mathbf{p}$ by computing each root of Q successively until we find t such that $A_i(t) \geq 0$ for all $i \in [\![0, d]\!]$ which gives us the solution. Note that we can compute the roots of a real univariate polynomial up to any desired precision with a number of algorithms from optimization theory which guarantee successful termination. Hence, we have defined a general process which allows to determine the MCI model with guaranteed success (provided sufficient resources) when $\mathcal{K}^* = \mathcal{I} \setminus \{I_d\}$. Furthermore, the reduction phase only needs to be computed once for any m so that computing an MCI model may be performed very efficiently for different values of K^* corresponding to a common value of $\mathcal{K}^*$ (which, in this case, is entirely determined by m).

Our goal now is to show that this principle can be generalized to any $\mathcal{K}^*$. This offers an interesting alternative to the more classical option of computing the model through constrained optimization algorithms based only on the equations in definition 4.1.3. Indeed, multivariate optimization algorithms do not allow to solve such problems in generic cases and thus must necessarily be performed over and over again for each specific case.

Note that this is not the first attempt to describe such models through algebraic geometry. In fact, Berd Sturmfels uses a similar description for a more general class of maximum likelihood models in [Stu02]. However, the algorithm he suggests remains an analytical one.

4.2.2.1 Algebraic geometry for polynomial system solving

As we will show, the equations defining the MCI model can easily be transposed into a multivariate polynomial system. Solving a multivariate polynomial system is a difficult task in general which has been mostly addressed within the field of algebraic geometry and a number of algorithms for solving real polynomial systems are now known to exist[Stu02, BPR06, BCR13].

We present here the main result on which our approach is based. However, we do not include a detailed presentation of the mathematical background in algebraic geometry which is necessary to fully grasp the concepts which we cover in this section. We refer the reader to the aforementioned literature for further insight on this topic. Furthermore, to avoid any ambiguity, we have conformed the terminology in algebraic geometry used in this thesis with the terminology defined in [BPR06].

The following notations will be used within this section. For any field $\mathbb{F}$, let $\mathbb{F}[\mathbf{X}] = \mathbb{F}[X_0, ..., X_d]$ be the ring of polynomials in $d+1$ variables $X_0, ..., X_d$ with coefficients in $\mathbb{F}$. The fields we will consider here all satisfy $\mathbb{Q} \subset \mathbb{F} \subset \mathbb{C}$. To maintain consistency with previous notations, X will be used to refer to an element of $\mathbb{F}^{d+1}$ with coordinates equal to $x_0, ..., x_d$. The term polynomial system will refer to a finite subset of $\mathbb{F}[\mathbf{X}]$ and we will generally note such a system $\mathcal{P}$. Solving a system $\mathcal{P}$ in $\mathbb{C}$ means determining the set of zeros of $\mathcal{P}$ in $\mathbb{C}^{d+1}$, which is the set:

$$\mathcal{Z}_{\mathcal{P}} = \left\{ X \in \mathbb{C}^{d+1} \;\middle|\; \bigwedge_{P \in \mathcal{P}} P(X) = 0 \right\}$$

and we will generally note $\mathcal{Z}$ for $\mathcal{Z}_{\mathcal{P}}$ unless there is some cause for ambiguity. The dimension of a polynomial system will refer to the dimension of its set of zeros in $\mathbb{C}$. Hence, a polynomial system is zero-dimensional if its set of zeros in $\mathbb{C}$ is finite.

Our approach is based on the fact that, given a zero-dimensional polynomial system $\mathcal{P} \subset \mathbb{F}[\mathbf{X}]$, there are algebraic algorithms (see, for example, algorithm 12.12 p.468 in [BPR06]) which allow us to determine (given sufficient computational resources) $d+3$ univariate polynomials $Q, B, A_0, ..., A_d$ with coefficients in $\mathbb{F}$ such that Q and B are coprime and

$$\mathcal{Z} = \left\{ \left(\frac{A_0}{B}\right), ..., \left(\frac{A_d}{B}\right) \in \mathbb{C}^{d+1} \;\middle|\; t \in \mathbb{C} \wedge Q(t) = 0 \right\}$$

In this case, $(Q, B, A_0, ..., A_d)$ is called a univariate representation of $\mathcal{P}$. This implies that, if we manage to express the equations defining an MCI model as a zero-dimensional polynomial system $\mathcal{P}$, we could break down the problem of determining the MCI model into two steps:

- determining a univariate representation of $\mathcal{Z}$;
- determining the MCI model from this univariate representation.

If the first step is performed, then the second step follows quite easily. In fact, we will show that the first step of the process may be performed only once for any $\mathcal{K}^*$ (as was the case for each $\mathcal{K}^*$ such that $\mathcal{K}^* = \mathcal{I} \setminus \{I_d\}$) which then allows for a very fast computation of MCI models in common cases of $\mathcal{K}^*$. Hence, the main focus here is on accomplishing the first step. However, computing a univariate representation raises two important issues.

Firstly, the polynomial system $\mathcal{P}$ which we consider must be zero-dimensional and, as we will show, this is not entirely straightforward. Secondly, algebraic algorithms do not tolerate approximate values well. In particular, floating point representations may not be used in the algorithms which we consider here. Instead, the coefficients of the polynomials considered in the algorithms, as well as the operations performed on these coefficients, must be considered within a formal calculus structure. While this is not technically infeasible, it may require significant computational resources both in time and memory. In order to accomplish this, two main options can be considered. The first option is to represent $\mathcal{P}$ as a system of polynomials in $\mathbb{Q}[\mathbf{X}]$ (which is technically the case if the constraints given by K are defined by an empirical dataset) and perform operations in a formal representation of $\mathbb{Q}[\mathbf{X}]$. This is the easier option of the two to code and is also generally faster to compute, but it only allows to determine a univariate representation corresponding to a particular constraint system defined by $\left(\mathcal{K}^*, \mathbf{f}_{|\mathcal{K}^*}\right)$. The other option is to consider that the polynomials in $\mathcal{P}$ belong to $\mathbb{Q}(f_1, ..., f_d)[\mathbf{X}]$ and we require a formal representation of $\mathbb{Q}(f_1, ..., f_d)$. While the latter option implies more elaborate programming, and calculations in $\mathbb{Q}(f_1, ..., f_d)$ may, in this case, represent the computational bottleneck of the general process, it does allow us to determine a definite univariate representation which can be used for any MCI model corresponding to a given $\mathcal{K}^*$.

4.2.2.2 A zero-dimensional polynomial system

Let $\left(\mathcal{K}^*, \mathbf{f}_{|\mathcal{K}^*}\right)$ be a constrained system and X the vector associated to the MCI model as in definition 4.1.3. The vector X is characterized as the unique solution to a linear and loglinear problem (theorem 4.1.3). We will show how we can transpose the equations of this characterization into a roughly equivalent zero-dimensional polynomial system in $\mathbb{R}[\mathbf{X}]$.

Linear part. Firstly, let us define, polynomials L_i for all $i \in [\![0, d]\!]$ such that:

$$L_i = \left(\sum_{j=0}^{d} t_{i,j} X_j \right) - f_i$$

in which $t_{i,j}$ are the coordinates of the matrix T. For example, when $m = 3$, this gives:

$$\begin{array}{rcl}
L_0 & = & X_0 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 - 1 \\
L_1 & = & X_1 + X_3 + X_5 + X_7 - f_1 \\
L_2 & = & X_2 + X_3 + X_6 + X_7 - f_2 \\
L_3 & = & X_3 + X_7 - f_3 \\
L_4 & = & X_4 + X_5 + X_6 + X_7 - f_4 \\
L_5 & = & X_5 + X_7 - f_5 \\
L_6 & = & X_6 + X_7 - f_6 \\
L_7 & = & X_7 - f_7
\end{array}$$

The linear equation $T_{\mathcal{K}^*}X = K^*$ is then equivalent to the polynomial system $\mathcal{P}_L = (L_j)_{j\in J}$ where $J = \{j \in [\![0, d]\!] \mid I_j \in \mathcal{K}^*\}$. We note $r = |J|$ the number of polynomials in $\mathcal{P}_L$ and we can easily notice that the dimension of $\mathcal{P}_L$ is equal to $s = 2^m - r$ (because it is equal to the dimension of its set of zeros $\mathcal{Z}_L$ as a vector space). The algorithm for computing $\mathcal{P}_L$ is here entirely straightforward:

```
1.  P_L ← ∅;
2.  for j in J:
3.      add L_j to P_L;
```

Algorithm 4.1: Computing $\mathcal{P}_L$

Loglinear part. Secondly, let $Y = \begin{bmatrix} y_0 \\ \vdots \\ y_d \end{bmatrix} \in \operatorname{Ker}(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$. Then, we can define the following polynomial:

$$M_Y = \prod_{\substack{i=0 \\ y_i>0}}^{d} X_i^{y_i} - \prod_{\substack{i=0 \\ y_i<0}}^{d} X_i^{-y_i} \in \mathbb{R}[\mathbf{X}]$$

and the equation $\underline{\ln}(X) \in \operatorname{Ker}(T_{\mathcal{K}^*})^{\perp}$ implies that $M_Y(X) = 0$. Our aim now is to pick a family of vectors in $\operatorname{Ker}(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$ which defines a polynomial system $\mathcal{P}_M$ that can be concatenated with $\mathcal{P}_L$ to obtain a polynomial system $\mathcal{P}$ which allows to characterize the MCI model X.

The first idea which comes to mind is to consider the basis $\mathcal{B}_{\mathcal{K}^*}$ of $\operatorname{Ker}(T_{\mathcal{K}^*})$

(see section 4.1.3.3). However, this does not always result in a zero-dimensional polynomial system. Indeed, let M_j be the polynomial defined by the j-th column of T^{-1}, for each $j \in [\![1, d]\!]$. For example, when $m = 3$, this gives:

$$\begin{array}{rcl} M_1 & = & X_1 - X_0 \\ M_2 & = & X_2 - X_0 \\ M_3 & = & X_0X_3 - X_1X_2 \\ M_4 & = & X_4 - X_0 \\ M_5 & = & X_0X_5 - X_1X_4 \\ M_6 & = & X_0X_6 - X_2X_4 \\ M_7 & = & X_1X_2X_4X_7 - X_0X_3X_5X_6 \end{array}$$

Now, suppose that we define $\mathcal{P}_M$ from these polynomials. Then $\mathcal{P}_M = (M_j)_{j\in\overline{J}}$ where $\overline{J} = \{j \in [\![0, d]\!] \mid I_j \notin \mathcal{K}^*\}$ and $\mathcal{P} = (L_j)_{j\in J} \sqcup (M_j)_{j\in\overline{J}}$. Considering the case in which $m = 3$ and $\mathcal{K}^* = \{\top\}$, we get:

$$\mathcal{P} = \begin{cases} X_0 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 - 1 & (L_0) \\ X_1 - X_0 & (M_1) \\ X_2 - X_0 & (M_2) \\ X_0X_3 - X_1X_2 & (M_3) \\ X_4 - X_0 & (M_4) \\ X_0X_5 - X_1X_4 & (M_5) \\ X_0X_6 - X_2X_4 & (M_6) \\ X_1X_2X_4X_7 - X_0X_3X_5X_6 & (M_7) \end{cases}$$

We can see that $\mathcal{P}$ is at least 3-dimensional. Indeed, consider $\mathcal{Z}'$ as below:

$$\mathcal{Z}' = \left\{X \in \mathbb{R}^{d+1} \;\middle|\; x_0 = x_1 = x_2 = x_4 = 0\right\}$$

Then, we get the following intersection between the set $\mathcal{Z}$ of zeros of $\mathcal{P}$ and $\mathcal{Z}'$:

$$\mathcal{Z} \cap \mathcal{Z}' = \left\{X \in \mathbb{R}^{d+1} \;\middle|\; \wedge \begin{array}{l} x_0 = x_1 = x_2 = x_4 = 0 \\ x_3 + x_5 + x_6 + x_7 - 1 = 0 \end{array}\right\}$$

which is a 3-dimensional linear space. Hence, in this case, $\mathcal{P}$ is at least 3-dimensional.

The issue in the example given here is that the dimension of $\mathcal{P}_M$ is at least equal to 4 (as $\mathcal{Z}' \subset \mathcal{Z}_M$) while we could expect it to be equal to 1. Indeed, the dimension of $\mathrm{Ker}\,(T_{\mathcal{K}^*})^{\perp}$ is equal to $r = 2^m - s$ so that the set of all $X \in \left(\mathbb{R}_+^*\right)^{d+1}$ satisfying $\underline{\ln}\,(X) \in \mathrm{Ker}\,(T_{\mathcal{K}^*})^{\perp}$ is a smooth r-manifold. Hence, $\mathcal{Z}_M$ is locally of dimension r around all $X \in \mathcal{Z}_M \cap \left(\mathbb{R}_+^*\right)^{d+1}$. This property extends to all $X \in \mathcal{Z}_M \cap (\mathbb{R}^*)^{d+1}$ because $X \in \mathcal{Z}_M$ implies $|X| \in \mathcal{Z}_M$ where $|X| = \begin{bmatrix} |x_0| \\ \vdots \\ |x_d| \end{bmatrix} \in (\mathbb{R}_+)^{d+1}$. Indeed, for any $M_Y \in \mathcal{P}_M$, then $M_Y(X) = 0 \iff \prod_{\substack{i=0\\y_i>0}}^{d} x_i^{y_i} - \prod_{\substack{i=0\\y_i<0}}^{d} x_i^{-y_i} = 0 \iff \prod_{\substack{i=0\\y_i>0}}^{d} x_i^{y_i} = \prod_{\substack{i=0\\y_i<0}}^{d} x_i^{-y_i} \implies \left|\prod_{\substack{i=0\\y_i>0}}^{d} x_i^{y_i}\right| = \left|\prod_{\substack{i=0\\y_i<0}}^{d} x_i^{-y_i}\right| \iff \prod_{\substack{i=0\\y_i>0}}^{d} |x_i|^{y_i} = \prod_{\substack{i=0\\y_i<0}}^{d} |x_i|^{-y_i} \iff \prod_{\substack{i=0\\y_i>0}}^{d} |x_i|^{y_i} - \prod_{\substack{i=0\\y_i<0}}^{d} |x_i|^{-y_i} = 0 \iff M_Y\,(|X|) = 0$. Therefore, if the dimension of $\mathcal{Z}_M$ is greater than r, this is necessarily due to its behavior within $\mathbb{R}^{d+1} \cap \left(\bigcup_{i=0}^{d} \mathcal{H}_i\right)$ where $\mathcal{H}_i$ is the hyperplane defined by $X_i = 0$.

In other words, if $\mathcal{P}_M$ is determined by a generating family of vectors of $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$, its dimension should be equal to r, unless there is a subset $S' \subset [\![0, d]\!]$ with cardinality $s' = |S'| < s$ defining a linear space $\mathcal{Z}' = \left\{X \in \mathbb{R}^{d+1} \;\middle|\; \forall i \in S',\ x_i = 0\right\}$ of dimension $r' = 2^m - s' > r$ such that $\mathcal{Z}' \subset \mathcal{Z}_M$. Hence, in order to show that there is a family of generating vectors of $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$ such that the associated polynomial system $\mathcal{P}_M$ has dimension r, we must show the following lemma.

Lemma 4.2.1. *There is a family of generating vectors of* $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$ *which defines a polynomial system* $\mathcal{P}_M$ *such that:*

$$\{S' \subset [\![0, d]\!] \mid (s' < s) \wedge (\mathcal{Z}' \subset \mathcal{Z}_M)\} = \emptyset$$

The proof of lemma 4.2.1 relies on the other following lemma from linear algebra.

Lemma 4.2.2. *Let* $\mathcal{V}$ *be a vector space of* $\mathbb{R}^{d+1}$ *such that:*

$$\exists S \subset [\![0, d]\!],\ \forall X \in \mathcal{V} \setminus \{0\},\ S_+(X) \cap S \neq \emptyset \ \text{ and } \ S_-(X) \cap S \neq \emptyset$$

where $S_+(X) = \{i \in [\![0, d]\!] \mid x_i > 0\}$ *and* $S_-(X) = \{i \in [\![0, d]\!] \mid x_i < 0\}$. *Then:*

$$\dim(\mathcal{V}) \leq s$$

where $s = |S|$.

Proof of lemma 4.2.2. Consider $S \subset [\![0, d]\!]$ such that,

$$\forall X \in \mathcal{V} \setminus \{0\},\ S_+(X) \cap S \neq \emptyset \ \text{ and } \ S_-(X) \cap S \neq \emptyset$$

Let $X, X' \in \mathcal{V} \setminus \{0\}$ such that $x_i = x'_i$ for all $i \in S$. Then, $Y = X - X' \in \mathcal{V}$ and $y_i = 0$ for all $i \in S$. Hence, $S_+(Y) \cap S = S_-(Y) \cap S = \emptyset$. Thus, $Y = 0$. Therefore, $X = X'$ and the dimension of $\mathcal{V}$ is at most s. $\square$

Proof of lemma 4.2.1. Let $\mathcal{Y}$ be a family of generating vectors of $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$ and $\mathcal{P}_M$ the corresponding polynomial system. Note $\mathcal{S}'$ the set defined by:

$$\mathcal{S}' = \{S' \subset [\![0, d]\!] \mid (s' < s) \wedge (\mathcal{Z}' \subset \mathcal{Z}_M)\}$$

and suppose $\mathcal{S}' \neq \emptyset$. Let $S' \in \mathcal{S}'$. Then, based on the converse of lemma 4.2.2, as $\dim(\mathrm{Ker}\,(T_{\mathcal{K}^*})) = s > s'$, there exists a vector $Y' \in \mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$ with $S_+(Y') \cap S' = \emptyset$ or $S_-(Y') \cap S' = \emptyset$. Note that this implies necessarily that $Y' \notin \mathcal{Y}$ as $\mathcal{Z}'$ cannot be contained in the set of zeros of $M_{Y'}$. Hence, if $\mathcal{Y}'$ is equal to the family $\mathcal{Y}$ augmented by Y' and $\mathcal{P}'_M$ is the corresponding polynomial system, then $\mathcal{Z}' \not\subset \mathcal{Z}'_M$ while $\mathcal{Z}'_M \subset \mathcal{Z}_M$ so that $\mathcal{S}'' = \{S'' \subset [\![0, d]\!] \mid (s'' < s) \wedge (\mathcal{Z}'' \subset \mathcal{Z}'_M)\}$ is strictly included in $\mathcal{S}'$.

If $\mathcal{S}'' = \emptyset$, we are done. Otherwise, we can repeat the process and define a strictly increasing sequence $\mathcal{Y} \subset \mathcal{Y}' \subset ... \subset \mathcal{Y}^{(k)}$ associated to a strictly decreasing sequence $\mathcal{Z}_M \supset \mathcal{Z}'_M \supset ... \supset \mathcal{Z}^{(k)}_M$ together with a strictly decreasing sequence $[\![0, d]\!] \supset \mathcal{S}' \supset \mathcal{S}'' \supset ... \supset \mathcal{S}^{(k-1)}$, until $\mathcal{S}^{(k-1)} = \emptyset$, which is bound to happen eventually as $[\![0, d]\!]$ is finite.

Hence, $\mathcal{Y}^{(k)}$ is a generating family of vectors of $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$ satisfying the desired property. $\square$

Through lemma 4.2.1, we see that we can consider a polynomial system $\mathcal{P}_M$ based on a generating family of vectors of $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$ which has dimension r and which defines a zero-dimensional $\mathcal{P}$ when concatenated with $\mathcal{P}_L$.

Computing $\mathcal{P}$. The proof to lemma 4.2.1 is a constructive one, which provides a baseline for an algorithm to determine $\mathcal{P}_M$ as desired: initialize $\mathcal{Y}$ to $\mathcal{B}_{\mathcal{K}^*}$ and incrementally add vectors to $\mathcal{Y}$ until $\mathcal{S}' = \emptyset$. However, a family of vectors $\mathcal{Y}$ obtained through such a process would not, a priori, have minimal cardinality. In the previous example, in which $m = 3$ and $\mathcal{K}^* = \{\top\}$, the cardinality of $\mathcal{Y}$ would be necessarily greater than 7, which is th cardinality of $\mathcal{B}_{\mathcal{K}^*}$, while the family $\mathcal{W}$ defined by:

$$\mathcal{W} = \left(\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} \right)$$

satisfies the conditions of lemma 4.2.1. Hence, we resort to a number of heuristics in order to obtain concise forms of $\mathcal{Y}$, leading to simpler polynomial systems to solve.

First, we can see that if Y is such that $y_i = 0$ for all $i \in [\![0, d]\!] \setminus \{j, j'\}$, $y_j = 1$ and $y_{j'} = -1$, for some $j, j' \in [\![0, d]\!]$ with $j \neq j'$, then M_Y is a linear function. Hence, if there is such a $Y \in \mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$, then we can consider M_Y within the linear part of the system, which can be solved first to reduce the general complexity of the problem. Therefore, we start by determining a subfamily of $\mathcal{Y}$, corresponding to such linear functions, which we note $\mathcal{Y}_L$. This can be accomplished through the following algorithm:

initialize $J \leftarrow \emptyset$;
initialize $\mathcal{Y}_L \leftarrow ()$;
for j from 0 to $d-1$:
    if $j \notin J$:
            add j to J;
        for j' from $j+1$ to d:
            if $j' \notin J$:
                $Y = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \ y_j = 1, \ y_{j'} = -1;$
                if $Y \in \mathrm{Ker}\,(T_{\mathcal{K}^*})$:
                    add j' to J;

add Y to $\mathcal{Y}_L$;

Algorithm 4.2: Computing $\mathcal{Y}_L$

Then, we need to add a family $\mathcal{Y}_{NL}$ to $\mathcal{Y}_L$, corresponding to the strictly non-linear part of $\mathcal{P}_M$, in order to define $\mathcal{Y}$. To do this, we can complete $\mathcal{Y}_L$ to form a basis of $\operatorname{Ker}(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$, initialize $\mathcal{Y}$ to be equal to this basis and then incrementally add vectors to $\mathcal{Y}$ until $\mathcal{S}' = \emptyset$ as described previously. Notice that, in this process, it suffices to consider the subset $\mathcal{T}' = \{S' \subset [\![0,d]\!] \mid (s' = s-1) \wedge (\mathcal{Z}' \subset \mathcal{Z}_M)\}$ of $\mathcal{S}'$ rather than $\mathcal{S}'$ because $\mathcal{T}' = \emptyset$ necessarily implies $\mathcal{S}' = \emptyset$. Furthermore, there is no need to iterate more than once over the elements of $\{S' \subset [\![0,d]\!] \mid s' = s-1\}$ because $\mathcal{T}'$ decreases when we add elements to $\mathcal{Y}$. Hence, the outline of the algorithm becomes as follows:

1. initialize $\mathcal{Y} \leftarrow \mathcal{Y}_L$;
2. complete $\mathcal{Y}$ to form a basis of $\operatorname{Ker}(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$;
3. for $S' \in \{S' \subset [\![0,d]\!] \mid s' = s-1\}$:
4.   if $\mathcal{Z}' \subset \mathcal{Z}_M$:
5.   choose Y' appropriately;
6.   add Y' to $\mathcal{Y}$;

Algorithm 4.3: Computing $\mathcal{Y}_{NL}$

The issue of choosing Y' in step 5 of the previous algorithm can be resolved as follows:

consider the matrix B such that each row corresponds to an element from $\mathcal{B}_{\mathcal{K}^*}$;
reorder the columns of B so that the first s' columns correspond to the columns with indices in S';
reduce B to its reduced row echelon form;
set Y' to the last row of B;
rearrange the columns of Y' back to the original order of indices;

Algorithm 4.4: Computing Y'

Furthermore, the cardinality of $\mathcal{Y}$ may eventually be reduced as it can contain a subfamily which satisfies the condition from lemma 4.2.1. We reduce the size of $\mathcal{Y}$ using a greedy algorithm:

while $\exists Y \in \mathcal{Y}_{NL}$ such that $\mathcal{Y} \setminus Y$ is a generating family

    of vectors of $\mathrm{Ker}(T_{\mathcal{K}^*})$ and $\mathcal{T}' = \emptyset$:
        remove Y from $\mathcal{Y}$;

Algorithm 4.5: Reducing $\mathcal{Y}_{NL}$

By combining all these algorithms, we obtain an algorithm for computing $\mathcal{P}_M$:

compute $\mathcal{Y}_L$ via algorithm 4.2;
compute $\mathcal{Y}_{NL}$ via algorithm 4.3 and algorithm 4.4;
reduce $\mathcal{Y}_{NL}$ via algorithm 4.5;
initialize $\mathcal{P}_{M_L} \leftarrow \emptyset$;
for $Y \in \mathcal{Y}_L$:
    add M_Y to $\mathcal{P}_{M_L}$;
initialize $\mathcal{P}_{M_{NL}} \leftarrow \emptyset$;
for $Y \in \mathcal{Y}_{NL}$:
    add M_Y to $\mathcal{P}_{M_{NL}}$;
$\mathcal{P}_M \leftarrow \mathcal{P}_{M_L} \sqcup \mathcal{P}_{M_{NL}}$;

Algorithm 4.6: Computing $\mathcal{P}_M$

Finally, we can determine $\mathcal{P}$ through algorithm 4.1 and algorithm 4.6:

compute $\mathcal{P}_L$ via algorithm 4.1;
compute $\mathcal{P}_M$ via algorithm 4.6;
$\mathcal{P} \leftarrow \mathcal{P}_L \sqcup \mathcal{P}_M$;

Algorithm 4.7: Computing $\mathcal{P}$

An implementation of this algorithm in Python 3 will be made freely available.

4.2.2.3 General structure of the algorithm

We have shown in the previous section that we can transpose the equations characterizing an MCI model into a zero-dimensional polynomial system. This system can be solved using algorithms from algebraic geometry as mentioned in section 4.2.2.1 and we can check each solution of the system (of which there is a finite number) until we find the one which corresponds to the characterization of the MCI model.

As any coordinate of the vector X defining the MCI model is equal to zero if and only if this can be derived directly from the constraints (in the sense of derivable itemsets, see sections 2.3.2.3 and 4.1.3.3), the MCI model corresponds to the only $X \in \mathcal{Z}$ such that $x_i = 0,\ \forall i \in D$ and $x_i > 0,\ \forall i \in [\![0, d]\!] \setminus D$, where D is the set of indices for which we can derive $x_i = 0$ directly. Hence, the general structure of the algorithm may be summarized as follows:

1. compute D;[1]
2. determine $\mathcal{P}$ from $\left(\mathcal{K}^*, \mathbf{f}_{|\mathcal{K}^*}\right)$ via algorithm 4.7;
3. add X_i to $\mathcal{P}$ for all $i \in D$;
4. solve $\mathcal{P}$ (i.e. determine a univariate representation of $\mathcal{Z}$ using an algorithm as mentioned in section 4.2.2.1);
5. find $X \in \mathcal{Z}$ such that $x_i > 0,\ \forall i \in [\![0, d]\!] \setminus D$;

Algorithm 4.8: Computing the MCI model

Note that this algorithm corresponds to the case in which the values in $\mathbf{f}_{|\mathcal{K}^*}$ are specified (otherwise D cannot be computed). By contrast, if the values in $\mathbf{f}_{|\mathcal{K}^*}$ are seen as formal variables, we can only perform steps 2 and 4 and, eventually, step 5 if it may be solved formally (or at least reduced) under the assumption that $D \neq \emptyset$ (as all cases in which $D \neq \emptyset$ can be obtained by continuity from cases in which $D = \emptyset$).

4.2.2.4 Speed-up for independence cases

The computational complexity of this algorithm is quite difficult to characterize because the computational complexity for determining a univariate representation of $\mathcal{Z}$ is itself quite difficult to characterize (unless a Gröbner basis for $\mathcal{P}$ is provided but this is not the case here). Obviously, the computational complexity increases at least exponentially with m as the number of variables considered is equal to $d+1 = 2^m$. But given m, the complexity varies also enormously with the structure of $\mathcal{K}^*$. Cases such as $\mathcal{K}^* = \{\top\}$ or $\mathcal{K}^* = \mathcal{I} \setminus \{I_d\}$ are extremely easy cases to compute while cases corresponding to standard (unconstrained) mutual independence between items or itemsets appear to be the most difficult ones. Hopefully, such cases may be identified and divided into cases corresponding to strictly smaller values of m which prove to be easier to compute.

Consider for example that $m = 5$ and $\mathcal{K}^* = \{\top, a_1 \wedge a_2, a_3 \wedge a_4, a_4 \wedge a_5, a_3 \wedge a_5\}$. None of the constraints on a_1 and a_2 are linked in any way to the constraints on a_3, a_4 and a_5. Hence, we can consider two MCI models:

[1] This is a simple problem in linear programming which can be solved through the use of a simplex algorithm for example.

the probability distribution $\mathbf{p}_1$ over the Boolean lattice $\mathcal{B}_1$ associated to $\mathcal{A}_1 = \{a_1, a_2\}$, defined by $\left(\mathcal{K}_1^*, \mathbf{f}_{|\mathcal{K}_1^*}\right)$ where $\mathcal{K}_1^* = \{\top, a_1 \wedge a_2\}$, on the one hand; and the probability distribution $\mathbf{p}_2$ over the Boolean lattice $\mathcal{B}_2$ associated to $\mathcal{A}_2 = \{a_3, a_4, a_5\}$, defined by $\left(\mathcal{K}_2^*, \mathbf{f}_{|\mathcal{K}_2^*}\right)$ where $\mathcal{K}_2^* = \{\top, a_3 \wedge a_4, a_4 \wedge a_5, a_3 \wedge a_5\}$, on the other hand. The MCI model $\mathbf{p}$ is then obtained by the independence of these two models via:

$$\mathbf{p}\left(a_1^*, a_2^*, a_3^*, a_4^*, a_5^*\right) = \mathbf{p}_1\left(a_1^*, a_2^*\right)\mathbf{p}_2\left(a_3^*, a_4^*, a_5^*\right)$$

where $a_i^* \in \{a_i, \overline{a_i}\}$ for all $i \in [\![1, 5]\!]$.

More generally, we can define the undirected graph $G = (V, E)$ of the mutual constraints between items by:

- $V = \{a_1, ..., a_m\}$;

- $\{a_i, a_j\} \in E$ if and only if $\exists I \in \mathcal{K}^*$ such that $I \implies (a_i \wedge a_j)$.

Let n_c be the number of connected components of G and $V_1, ..., V_{n_c}$ the set of items associated to each component. Then, each set of items V_i corresponds to an MCI model $\mathbf{p}_i$ over the Boolean lattice associated to V_i, defined by $\left(\mathcal{K}_i^*, \mathbf{f}_{|\mathcal{K}_i^*}\right)$ where:

$$\mathcal{K}_i^* = \left\{ I \in \mathcal{K}^* \;\middle|\; \bigwedge_{a_j \in V_i} a_j \implies I \right\}$$

and the MCI model $\mathbf{p}$ is entirely defined by:

$$\mathbf{p}\left(\bigwedge_{j=1}^{m} a_j^*\right) = \prod_{i=1}^{n_c} \mathbf{p}_i\left(\bigwedge_{a_j \in V_i} a_j^*\right)$$

If G has only one connected component, then there is no gain, but the cost of computing G and its connected components is highly negligible in comparison to the gain that occurs when G has at least two components. This is true when the MCI model is computed through algorithms in algebraic geometry, but it is also true if they are seen as MaxEnt models and computed through algorithms in optimization theory and a similar process is described in [MVT12].

4.2.2.5 Speed-ups for step 4

As stated previously, the bottleneck of algorithm 4.8 in terms of computational complexity resides in its step 4, in which a univariate representation of $\mathcal{Z}$ is computed. In order to speed this step up, we can use substitutions to reduce significantly the number of variables considered before solving the polynomial system. These speed-ups were essential to compute the algebraic forms of all MCI models for $m = 3$ and $m = 4$.

The first trick is to reduce the linear part of $\mathcal{P}$ separately and perform substitutions in the nonlinear part of $\mathcal{P}$ based on this reduction. The linear part of $\mathcal{P}$ comprises the polynomials in $\mathcal{P}_L$, as well as the polynomials in $\mathcal{P}_M$ which correspond to the family of vectors $\mathcal{Y}_L$ as determined by algorithm 4.2 (noted $\mathcal{P}_{M_L}$ in algorithm 4.6) and the polynomials added to $\mathcal{P}$ in step 3 of algorithm 4.8 (we will note these $\mathcal{P}_D$). Each of these polynomials corresponds naturally to a vector with coordinates in $(X_0, ..., X_d, 1)$ so that we can see the linear part of $\mathcal{P}$ as a matrix with $d + 2$ columns and as many row as polynomials in the sets mentioned above. We can then consider its reduced row echelon form and obtain a set of free variables from which the remaining pivot variables are entirely determined. The pivot variables are then substituted in the remaining polynomials of $\mathcal{P}$ (noted $\mathcal{P}_{M_{NL}}$ in algorithm 4.6) by their expressions as affine functions of the free variables. In this manner, a new zero-dimensional polynomial system is obtained whose variables are the free variables determined previously. The reduction in terms of number of variables is quite substantial. For $m = 3$, this brings down the number of variables down from 8 to 1, 2 or 3 depending on $\mathcal{K}^*$. For $m = 4$, this brings down the number of variables down from 16 to 7 or less. Note that the part of this reduction which is based on the elements of $\mathcal{P}_{M_L}$ is mostly equivalent to the reduction based on blocks described in [MVT12] for the computation of MaxEnt models.

Now that we have obtained this reduced polynomial system, the second trick is to find any variable for which at least one polynomial in the system has degree exactly 1. Indeed, if a polynomial P has degree 1 in a variable, say X_0, then $P(X_0, ..., X_d) = A(X_1, ..., X_d)X_0 + B(X_1, ..., X_d)$ and, therefore:

$$P(X_0, ..., X_d) = 0 \iff A(X_1, ..., X_d)X_0 = -B(X_1, ..., X_d)$$

(Note that we write $X_0, ..., X_d$ for simplicity even though we are now considering a set of variables which is strictly contained in $\{X_0, ..., X_d\}$.)

Furthermore, as we have $P(x_0, ..., x_d) = 0$ and $x_0 \neq 0$ when considering the MCI model (because the variables equal to zero have already been set aside in the reduction described above), then either $A(x_1, ..., x_d) = B(x_1, ..., x_d) = 0$ or $A(x_1, ..., x_d)B(x_1, ..., x_d) \neq 0$. Each of these cases can be associated to a zero-dimensional polynomial system which is easier to solve than the current one. On one side, if $A(x_1, ..., x_d) = B(x_1, ..., x_d) = 0$, we can consider the polynomial system in which P has been replaced by A and B. And, on the other side, if $A(x_1, ..., x_d)B(x_1, ..., x_d) \neq 0$, we can consider that $X_0 = -\frac{B}{A}$ (where A and B can be reduced so that they contain no common factors because x_0 does not correspond to a root of A or B) and thus substitute X_0 by $-\frac{B}{A}$ in all the polynomials of the system and multiply each of these by A as many times as necessary to obtain a polynomial (which corresponds to the degree of X_0 in the polynomial). In this case, the new polynomial system has one polynomial less (the polynomial P initially considered) and one variable less (X_0 in this example). Note that, in all the cases which we have computed for $m = 3$ and $m = 4$, when such a reduction was possible, the solution of the system associated to the MCI model always corresponded to the reduced polynomial system in which a variable was substituted by a rational expression $-\frac{B}{A}$. Hence, though we have not proved this generally, for all the cases which we have computed, such a reduction corresponds to decreasing the number of variables in the polynomial system by one.

This process may be repeated until the system may no longer be reduced in this manner. However, note that, if at one point in the process there is more than one variable which may be considered, the choice of the variable may influence how much the system may be reduced. In practice, the gain provided by reducing the number of variables is such that we explore all possible choices until we have found one which gives an optimal reduction in terms of number of variables.

4.2.2.6 Algebraic solutions for all cases when $m \leq 4$

In section 4.2.2.1, we explained that the computations for determining a univariate representation may be performed in $\mathbb{Q}$, based on specific rational values for $f_1, ..., f_d$, or in $\mathbb{Q}(f_1, ..., f_d)$, based on formal values for $f_1, ..., f_d$. In the case in which formal values are employed, the univariate representation obtained for a given $\mathcal{K}^*$ corresponds to a formal and simplified algebraic representation of the MCI model (for this given $\mathcal{K}^*$). This representation can be stored allowing for a fast and precise computation of the corresponding MCI models given any specific values for $\mathbf{f}_{|\mathcal{K}^*}$.

In the course of this doctoral research, we have computed such formal univariate representations for a sufficient number of cases of $\mathcal{K}^*$ such that $m \leq 4$, allowing for a fast computation of all MCI models in which $m \leq 4$ or consisting of independent groups of items satisfying this condition. The number of different cases of $\mathcal{K}^*$ for a given m is equal to 2^{2^m-1} which is the number of subsets of $\mathcal{I}$ that contain $\top$. However, it is sufficient to consider only a fraction of these cases because if a set $\mathcal{K}_1^*$ may be obtained from a set $\mathcal{K}_2^*$ by a simple permutation of the items defining the itemsets, then a formal univariate representation associated to $\mathcal{K}_1^*$ may be obtained from the formal univariate representation computed for $\mathcal{K}_2^*$. Hence, we need only consider a single representative for each equivalence class defined by the set of permutations on items which brings down the number of cases to compute significantly enough. This corresponds to sequence A000612 in [Slo19], which is described as the number of non-isomorphic sets of nonempty subsets of an n-set. The number of cases to compute can be brought down slightly further still by computing only the cases which do not correspond to independence cases using the principles described in section 4.2.2.4. The number of such cases corresponds to sequence A323819 in [Slo19], which is described as the number of non-isomorphic connected set-systems covering n vertices.

m	2^{2^m-1}	A000612	A323819
2	8	6	3
3	128	40	30
4	32,768	1,992	1,912
5	2,147,483,648	18,666,624	18,662,590
6	9.223×10^{18}	1.281×10^{16}	1.281×10^{16}
7	1.701×10^{38}	3.376×10^{34}	3.376×10^{34}

Table 4.6: Sequences for the number of cases to compute.

The number of cases to compute is therefore reasonable enough for us to envisage computing all the cases for $m \leq 4$ on a personal computer. Given more computational power, computing the cases for $m = 5$ may also be considered. However, though the gain in terms of number of cases to compute is asymptotically a factor $m!$, this is not sufficient to envisage an exhaustive computation of all cases for any value of m beyond $m = 5$.

Setting aside the question of computing a large number of cases, the issue with performing computations in the field $\mathbb{Q}(f_1, ..., f_d)$ (or, more precisely, in the polynomial space $\mathbb{Q}(f_1, ..., f_d)[X_0, ..., X_d]$) resides in the augmented cost of basic operations and simplifications of expressions which must be performed both a great many times and with expressions that are potentially quite long. However, in order to curtail the size of the expressions considered, the coefficients of the polynomials can always be reduced to an irreducible rational fraction (based on the continuity of the solution with regards to the variables $f_1, ..., f_d$). This means that we can also consider operations on polynomials with coefficients in $\mathbb{Q}[f_1, ..., f_d]$ that are setwise coprime, which is the option we have adopted in our implementation.

Last, once a formal univariate representation is computed it may possibly be reduced. Indeed, it may appear, in some cases, that one or several of the roots of the polynomial Q of a univariate representation $(Q, B, A_0, ..., A_d)$ can be ignored: either because they lead to solutions which can be formally identified as not satisfying the conditions of the MCI model (necessarily leading to negative or non real values for $x_0, ..., x_d$); or because they lead to solutions which necessarily correspond to a situation of derivability (where one of the values for $x_0, ..., x_d$ at least is equal to zero which can be ignored because of the continuity of the MCI model with regards to $f_1, ..., f_d$).

The code with which we have obtained the formal univariate representations described in this section will be made freely available.

Solutions for $m = 3$. We list below the computed algebraic expressions corresponding to representatives for each of the 30 different equivalence classes described above. For each case, we give the subset of $\{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$ corresponding to the fixed frequencies. If solving the system includes computing the roots of a polynomial Q with coefficients in $\mathbb{Z}[f_1, ..., f_7]$, we indicate this in the upper right corner and give the corresponding polynomial below. We then list the algebraic expressions for each x_i based on the values $f_1, ..., f_7$

as well as the previously computed values of x_i and a root t of Q. The MCI model is obtained by considering a root t of Q such that all x_i are positive.

$\{f_7\}$ $x_0 = \frac{1-f_7}{7}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = x_0$ $x_5 = x_0$ $x_6 = x_0$ $x_7 = f_7$	$\{f_6, f_7\}$ $x_0 = \frac{1-f_6}{6}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = x_0$ $x_5 = x_0$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$\{f_5, f_6\}$ Q $x_7 = t$ $x_0 = \frac{1-f_5-f_6+x_7}{5}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = x_0$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$

$Q = 4T^2 - (1 + 4(f_5 + f_6))T + 5f_5 f_6$

$\{f_5, f_6, f_7\}$ $x_0 = \frac{1-f_5-f_6+f_7}{5}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = x_0$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$\{f_4, f_7\}$ $x_0 = \frac{1-f_4}{4}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = \frac{f_4-f_7}{3}$ $x_5 = x_4$ $x_6 = x_4$ $x_7 = f_7$	$\{f_4, f_6, f_7\}$ $x_0 = \frac{1-f_4}{4}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = \frac{f_4-f_6}{2}$ $x_5 = x_4$ $x_6 = f_6 - f_7$ $x_7 = f_7$

$\{f_4, f_5, f_6\}$ $x_7 = \frac{f_5 f_6}{f_4}$ $x_0 = \frac{1-f_4}{4}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = f_4 - f_5 - f_6 + x_7$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$	$\{f_4, f_5, f_6, f_7\}$ $x_0 = \frac{1-f_4}{4}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = x_0$ $x_4 = f_4 - f_5 - f_6 + f_7$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$\{f_3, f_5, f_6\}$ Q $x_7 = t$ $x_0 = \frac{1-f_3-f_5-f_6+2x_7}{4}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = f_3 - x_7$ $x_4 = x_0$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$

$Q = 20T^3 + 4(1 - 5(f_3 + f_5 + f_6))T^2 + ((1 - (f_3 + f_5 + f_6))^2 + 16(f_3 f_5 + f_3 f_6 + f_5 f_6))T - 16f_3 f_5 f_6$

$\{f_3, f_5, f_6, f_7\}$ $x_0 = \frac{1-f_3-f_5-f_6+2f_7}{4}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = f_3 - f_7$ $x_4 = x_0$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$\{f_3, f_4, f_7\}$ $x_0 = \frac{1-f_3-f_4+f_7}{3}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = f_3 - f_7$ $x_4 = \frac{f_4-f_7}{3}$ $x_5 = x_4$ $x_6 = x_4$ $x_7 = f_7$	$\{f_3, f_4, f_6\}$ Q $x_7 = t$ $x_0 = \frac{1-f_3-f_4+x_7}{3}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = f_3 - x_7$ $x_4 = \frac{f_4-f_6}{2}$ $x_5 = x_4$ $x_6 = f_6 - x_7$

$Q = 2T^2 + (f_4 - 2f_3 - 3f_6 - 1)T + 3f_3 f_6$

$\{f_3, f_4, f_6, f_7\}$ $x_0 = \frac{1-f_3-f_4+f_7}{3}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = f_3 - f_7$ $x_4 = \frac{f_4-f_6}{2}$ $x_5 = x_4$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$\{f_3, f_4, f_5, f_6\}$ Q $x_7 = t$ $x_0 = \frac{1-f_3-f_4+x_7}{3}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = f_3 - x_7$ $x_4 = f_4 - f_5 - f_6 + x_7$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$	$\{f_3, f_4, f_5, f_6, f_7\}$ $x_0 = \frac{1-f_3-f_4+f_7}{3}$ $x_1 = x_0$ $x_2 = x_0$ $x_3 = f_3 - f_7$ $x_4 = f_4 - f_5 - f_6 + f_7$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$

$Q = 4T^3 + (1 - 4(f_3 + f_5 + f_6))T^2 + (3(f_3 f_5 + f_3 f_6 + f_5 f_6) + (1 - f3 - f4)(f4 - f5 - f6))T - 3f_3 f_5 f_6$

$\{f_2, f_4, f_7\}$ Q $x_6 = t$ $x_0 = \frac{1-f_2-f_4+f_7+x_6}{2}$ $x_1 = x_0$ $x_2 = \frac{f_2-f_7-x_6}{2}$ $x_3 = x_2$ $x_4 = \frac{f_4-f_7-x_6}{2}$ $x_5 = x_4$ $x_7 = f_7$	$\{f_2, f_4, f_6, f_7\}$ $x_0 = \frac{1-f_2-f_4+f_6}{2}$ $x_1 = x_0$ $x_2 = \frac{f_2-f_6}{2}$ $x_3 = x_2$ $x_4 = \frac{f_4-f_6}{2}$ $x_5 = x_4$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$\{f_2, f_4, f_5, f_7\}$ $x_6 = \frac{(f_4-f_5)(f_2-f_7)}{1-f_5}$ $x_0 = \frac{1-f_2-f_4+f_7+x_6}{2}$ $x_1 = x_0$ $x_2 = \frac{f_2-f_7-x_6}{2}$ $x_3 = x_2$ $x_4 = f_4 - f_5 - x_6$ $x_5 = f_5 - f_7$ $x_7 = f_7$

$Q = T^2 + (2 - f_2 - f_4)T - (f4 - f7)(f2 - f7)$

$\{f_2, f_4, f_5, f_6\}$	$\{f_2, f_4, f_5, f_6, f_7\}$	$\{f_2, f_3, f_4, f_5\}$ Q
$x_7 = \frac{f_5 f_6}{f_4}$ $x_0 = \frac{1-f_2-f_4+f_6}{2}$ $x_1 = x_0$ $x_2 = \frac{f_2-f_6}{2}$ $x_3 = x_3$ $x_4 = f_4 - f_5 - f_6 + x_7$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$	$x_0 = \frac{1-f_2-f_4+f_6}{2}$ $x_1 = x_0$ $x_2 = \frac{f_2-f_6}{2}$ $x_3 = x_2$ $x_4 = f_4 - f_5 - f_6 + f_7$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$x_6 = t$ $x_7 = x_6 - 1 - f_2 + 2f_3 - f_4 + 2f_5 + \frac{2(f_2-f_3)(f_4-f_5)}{x_6}$ $x_0 = \frac{1-f_2-f_4+x_6+x_7}{2}$ $x_1 = x_0$ $x_2 = f_2 - f_3 - x_6$ $x_3 = f_3 - x_7$ $x_4 = f_4 - f_5 - x_6$ $x_5 = f_5 - x_7$

$Q = T^3 + (1-(1+f_2-f_3)(1+f_4-f_5)-(1-f_3)(1-f_5))T^2 + (f_2-f_3)(f_4-f_5)(f_2-2f_3+f_4-2f_5+3)T - 2(f_2-f_3)^2(f_4-f_5)^2$

$\{f_2, f_3, f_4, f_5, f_7\}$ Q_1	$\{f_2, f_3, f_4, f_5, f_6\}$ Q_2	$\{f_2, f_3, f_4, f_5, f_6, f_7\}$
$x_6 = t$ $x_0 = \frac{1-f_2-f_4+f_7+x_6}{2}$ $x_1 = x_0$ $x_2 = f_2 - f_3 - x_6$ $x_3 = f_3 - f_7$ $x_4 = f_4 - f_5 - x_6$ $x_5 = f_5 - f_7$ $x_7 = f_7$	$x_7 = t$ $x_0 = \frac{1-f_2-f_4+f_6}{2}$ $x_1 = x_0$ $x_2 = f_2 - f_3 - f_6 + x_7$ $x_3 = f_3 - x_7$ $x_4 = f_4 - f_5 - f_6 + x_7$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$	$x_0 = \frac{1-f_2-f_4+f_6}{2}$ $x_1 = x_0$ $x_2 = f_2 - f_3 - f_6 + f_7$ $x_3 = f_3 - f_7$ $x_4 = f_4 - f_5 - f_6 + f_7$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$

$Q_1 = T^2 - (1 + f_2 - 2f_3 + f_4 - 2f_5 + f_7)T + 2(f_2 - f_3)(f_4 - f_5)$

$Q_2 = 2T^3 + (f_2 - 2f_3 + f_4 - 2f_5 - 3f_6)T^2 + (f_2f_4 - f_2f_5 - f_2f_6 - f_3f_4 + 2f_3f_5 + 2f_3f_6 - f_4f_6 + 2f_5f_6 + f_6^2)T - f_3f_5f_6$

$\{f_1, f_2, f_4, f_7\}$ Q	$\{f_1, f_2, f_4, f_6, f_7\}$	$\{f_1, f_2, f_4, f_5, f_6\}$
$x_6 = t$ $x_5 = \frac{(f_1-f_7)(f_4-f_7-x_6)}{1-f_7-x_6}$ $x_3 = \frac{(f_1-f_7-x_5)(f_2-f_7-x_6)}{1-f_4}$ $x_0 = 1-f_1-f_2-f_4+2f_7+x_3+x_5+x_6$ $x_1 = f_1 - f_7 - x_3 - x_5$ $x_2 = f_2 - f_7 - x_3 - x_6$ $x_4 = f_4 - f_7 - x_5 - x_6$ $x_7 = f_7$	$x_5 = \frac{(f_1-f_7)(f_4-f_6)}{1-f_6}$ $x_3 = \frac{(f_2-f_6)(f_1-f_7-x_5)}{1-f_4}$ $x_0 = 1-f_1-f_2-f_4+f_6+f_7+x_3+x_5$ $x_1 = f_1 - f_7 - x_3 - x_5$ $x_2 = f_2 - f_6 - x_3$ $x_4 = f_4 - f_6 - x_5$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$x_7 = \frac{f_5 f_6}{f_4}$ $x_3 = \frac{(f_1-f_5)(f_2-f_6)}{1-f_4}$ $x_0 = 1 - f_1 - f_2 - f_4 + f_5 + f_6 + x_3$ $x_1 = f_1 - f_5 - x_3$ $x_2 = f_2 - f_6 - x_3$ $x_4 = f_4 - f_5 - f_6 + x_7$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$

$Q = (1 - f_1)T^2 - (1 - 2f_7 - f_1f_2 - f_1f_4 + 2f_1f_7 + f_2f_4)T + (f_2 - f_7)(f_4 - f_7)(1 - f_1)$

$\{f_1, f_2, f_4, f_5, f_6, f_7\}$	$\{f_1, f_2, f_3, f_4, f_5, f_6\}$ Q	$\{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$
$x_3 = \frac{(f_1-f_5)(f_2-f_6)}{1-f_4}$ $x_0 = 1 - f_1 - f_2 - f_4 + f_5 + f_6 + x_3$ $x_1 = f_1 - f_5 - x_3$ $x_2 = f_2 - f_6 - x_3$ $x_4 = f_4 - f_5 - f_6 + f_7$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$	$x_7 = t$ $x_0 = 1-f_1-f_2+f_3-f_4+f_5+f_6-x_7$ $x_1 = f_1 - f_3 - f_5 + x_7$ $x_2 = f_2 - f_3 - f_6 + x_7$ $x_3 = f_3 - x_7$ $x_4 = f_4 - f_5 - f_6 + x_7$ $x_5 = f_5 - x_7$ $x_6 = f_6 - x_7$	$x_0 = 1-f_1-f_2+f_3-f_4+f_5+f_6-f_7$ $x_1 = f_1 - f_3 - f_5 + f_7$ $x_2 = f_2 - f_3 - f_6 + f_7$ $x_3 = f_3 - f_7$ $x_4 = f_4 - f_5 - f_6 + f_7$ $x_5 = f_5 - f_7$ $x_6 = f_6 - f_7$ $x_7 = f_7$

$Q = T^3 - (f_3 + f_5 + f_6 - f_1f_2 - f_1f_4 - f_2f_4 + f_1f_6 + f_2f_5 + f_3f_4)T^2 + (f_3f_5 + f_3f_6 + f_5f_6 + f_1f_2f_4 - f_1f_2f_5 - f_1f_2f_6 - f_1f_3f_4 - f_1f_4f_6 + f_1f_6^2 - f_2f_3f_4 - f_2f_4f_5 + f_2f_5^2 + f_3^2f_4 + 2f_3f_5f_6)T + f_3f_5f_6(f_1 + f_2 - f_3 + f_4 - f_5 - f_6 - 1)$

4.2.2.7 Pros and cons of the algebraic method

As stated previously, one of the important advantages of the algebraic method is that it allows us to determine reduced algebraic expressions for MCI models in generic cases, from which we can then compute specific MCI models very efficiently. When the corresponding generic cases have been computed, the increase in computation speed is quite astounding in comparison to standard methods for computing MaxEnt models.

In order to check this, we chose 20 different samples of $m = 3$ items among the 70 items of the plants database [Nat08], each corresponding to an empirical distribution $\mathbf{f}$ such that no single f_i could be derived from the other f_j for which $j \neq i$ (i.e. $D = \emptyset$). For each of these distributions, we considered the computation of 30 different MCI models, each of which corresponded to one of the pre-computed cases in section 4.2.2.6 above. Each of these computations were performed 100 times using the pre-computed algebraic expressions and 100 times using an implementation of the Iterative scaling procedure by Darroch and Ratcliff for computing MaxEnt models [DR72][2]. In order to make comparisons in terms of execution as meaningful as possible, the computations were performed on the same computer (Intel Core i7-8550U CPU 1.80GHz $\times$ 8, 7.7 GiB RAM) and both were based on a Python 3 implementation. The total execution time using the algebraic expressions was approximately equal to 2.14 seconds, while it took approximately 6 minutes and 21 seconds for the purely numerical method. Hence, the method based on the algebraic expressions was about 150 times faster here. Note that a more detailed observation of the execution times in the process described above allowed us to ensure that the gain in time was not concentrated on any distribution or constrained set in particular (though there was some variations between constraint sets).

Even though the gain in terms of execution time obtained here is quite impressive, it must be put into perspective. Such a gain can only be obtained if we consider specific cases corresponding to previously computed generic cases, the computation of which is itself quite time consuming. As mentioned in section 4.2.2.6, we have managed to compute all generic cases corresponding to $m \leq 4$ but we also acknowledge that doing so is intractable for any value of $m \geq 6$.

Nevertheless, the inability to compute the exhaustive list of all generic cases for larger values of m does not necessarily represent a serious limitation to the interest of the MCI approach, for both practical and theoretical applications. Regarding practical applications, it must be noted that it is, in general and regardless of the method employed, practically infeasible to consider a full description of a probability measure on $\mathcal{B}$, for even a limited number of itemsets, because such a description requires the definition of $2^m - 1$ individual values a priori. In itemset mining, global models (i.e. probability distributions where $\mathcal{B}$ is defined by all items) are not considered directly in practical applications. Instead, they are replaced by numerous small local models (where $\mathcal{B}$ is defined by a small subset of itemsets). If we are considering a large number of local

[2]This specific algorithm was chosen based on the fact that it has been commonly used for computing such MaxEnt models in the context of itemset mining [PMS03, TM10, MVT12]

MCI models, each of which are defined around 3 or 4 items, the algebraic method becomes highly relevant. Furthermore, the explicit computation of reduced algebraic expressions for MCI models can be useful from a theoretical perspective, as it may bring insight on the structure of these models. Notably, we have hope that the explicit computation of reduced algebraic expressions for MCI models based on the frequencies of all itemsets of size 1 and 2 for low values of m can help us determine an explicit algebraic formula for such models and provide an interesting alternative to Chow-Liu tree models [CL68].

Lastly, the previous remarks only apply to the approach in which we try to compute an MCI model using the algebraic method in a generic case before considering a specific case (that is we perform computations in $\mathbb{Q}(f_1, ..., f_d)$ before substituting the f_i by their values). If we compute the MCI model using the algebraic method in a specific case (that is we perform computations directly in $\mathbb{Q}$), the computation time is individually much lower than computing the generic case. Though we speculate that, for the computation of a specific individual case, the numerical method is faster still than the algebraic method, we have yet to perform comparisons between these two approaches. As the algebraic method on a specific case performs better when the values for the numerators and denominators of the f_i are small (which can notably be the case if the number of transactions is not too large), it is possible that the algebraic approach (eventually combined with an approximation scheme) may outperform the numerical method in a number of cases.

4.3 Conclusion

In this chapter, we have presented our own approach towards MaxEnt models in the context of itemsets: mutual constrained independence models. We have demonstrated how this perspective sheds further light upon the rationale of such models and described a new approach towards their computation, based on tools for algebraic geometry. This approach has allowed us to determine exact algebraic expressions for all the MCI models when $m \leq 4$, as well as for the MCI models defined by the frequencies of all proper subsets of an itemset for any m. These expressions allow for an increase in the computation speed of the corresponding MCI models by several orders of magnitude in comparison to standard methods for computing MaxEnt models. We are hopeful that further research based on this approach might help determine algebraic expressions for a wider range of models and we would like to investigate the issue of determining algebraic expressions for MCI alternatives to Chow-Liu tree models. Note also that, although we have defined the constraints of MCI models as constraints on the frequencies of itemsets, the results we have shown can easily be generalized to a much wider range of constraints. In particular, they still hold if we replace itemsets by any sound and complete family of patterns in $\mathcal{B}$ (see section 3.6.2.5) as this would only modify the expression of the transfer matrix T (see section 4.1.1.2).

MCI models correspond exactly to the type of models that define the objective data-driven hypotheses which we have described in chapter 3 (see section 3.7.3 in particular) and, as such, represent a central notion in our general framework for objective frequency-based interesting pattern mining. In the next chapter, we present the corresponding pattern mining algorithms.

CHAPTER 5

Extracting objectively interesting patterns from data

In the current and final chapter of this doctoral thesis, we present pattern mining algorithms based on the principles for a meaningful mathematical modeling of objective interestingness in patterns described in chapter 3, as well as the MCI models presented in chapter 4.

Recall from chapter 4 that we can include the scientific method within the mathematical modeling for the pattern mining process so that the pattern mining is seen as a process in which a scientifically valid hypothesis about the data is discovered. The hypotheses that we will consider in this chapter are mutual constrained independence hypotheses each of which is associated to a MCI model. Hence, each of the hypotheses considered corresponds to a set of itemsets with their associated frequencies. As such, the processes which we describe in this chapter fall within the field of itemset mining. However, as we have previously noted in section 4.3 regarding MCI models, such hypotheses could still be defined if we considered other types of patterns based on logical expressions. In particular, the contents of this chapter could easily be adapted if we replaced the set of itemsets $\mathcal{I}$ with any other sound and complete family of patterns (see section 3.6.2.5).

5.1 Testing the MCI hypothesis

5.1.1 Definition of the MCI hypothesis

Unless specified explicitly, we consider the same notations here as in chapter 4. Let $\mathcal{D}$ be a binary dataset corresponding to n observations of the presence or absence of m items in a statistical population (i.e. a dataset of n transactions on these items) and $\mathcal{K} \subset \mathcal{I}$ be a set of itemsets. We note $\mathbf{f}$ the probability measure on $\mathcal{B}$ defined by the empirical distribution in the dataset $\mathcal{D}$.

Definition 5.1.1. The mutual constrained independence hypothesis for the dataset $\mathcal{D}$ defined by $\mathcal{K}$ is the hypothesis that the dataset corresponds to n independent identically distributed samples of a random variable whose distribution $\mathbf{p}$ is given by the MCI model for the data defined by $\mathcal{K}$ (see definition 4.1.3).

For a detailed explanation of the rationale behind the definition of such a hypothesis, we refer to the two previous chapters.

5.1.2 Statistical testing of the MCI hypothesis

5.1.2.1 χ^2 statistic

Following the classical approach used for statistical tests of independence, we suggest the use of a χ^2 test for testing mutual constrained independence. The χ^2 statistic for the dataset is defined as:

$$\chi^2_{\mathbf{p},\mathbf{f}} = n \sum_{\substack{0 \le i \le d \\ i \notin D}} \frac{(f_i - p_i)^2}{p_i}$$

where D is the set of indices for which $p_i = 0$ as described in section 4.2.2.3[1]. The distribution of the $\chi^2_{\mathbf{p},\mathbf{f}}$ statistic converges towards a χ^2 distribution as described in the following proposition.

[1]Recall that, in the MCI model, we have $p_i \implies f_i$. Hence the sum only excludes indices for which both p_i and f_i are equal to 0.

Proposition 5.1.1. *The distribution of the* $\chi^2_{\mathbf{p},\mathbf{f}}$ *statistic for the dataset given the MCI hypothesis asymptotically converges towards a* χ^2 *distribution with:*

$$d + 1 - \#D - \#\mathcal{P}_{L|D}$$

degrees of freedom, where $\#D$ *is the number of indices for which* $p_i = 0$ *and* $\#\mathcal{P}_{L|D}$ *is the rank of the space generated by the polynomials in* $\mathcal{P}_L$ *when setting* $X_i = 0$ *for all* $i \in D$ *(see section 4.2.2.2).*

Proof. The proposition is a straight forward application of the standard result by Pearson, while taking into account the number of degrees of freedom (see, for example, [BP14]). The only important aspect here is to be careful when counting the number of degrees of freedom, in order to take into account the possible border effects when $D \neq \emptyset$ so that $d + 1$ is reduced by $\#D + \#\mathcal{P}_{L|D}$ rather than simply $\#\mathcal{P}_L$. □

Note that, because we cannot address every issue raised in this thesis, we do not provide a description of the rate at which the distribution of the $\chi^2_{\mathbf{p},\mathbf{f}}$ statistic converges towards the χ^2 distribution. Moreover, the issue of the error in the approximation of the distribution of the $\chi^2_{\mathbf{p},\mathbf{f}}$ statistic by a χ^2 distribution has been mostly set aside by researchers in the case of Pearson's test of independence as it is only significant when considering extremely small frequencies, [SRDCS19]. However, in this case, this issue could be more significant because of the possibly high dimension of the probability spaces considered and we believe it should be addressed eventually.

5.1.2.2 χ^2 test

We now define the χ^2 test of mutual constrained independence given $\mathcal{K}$ following standard methodology for χ^2 tests.

Definition 5.1.2. Let α be a predefined probability threshold. We note $\mathcal{H}$ the MCI hypothesis for a dataset $\mathcal{D}$ defined by a set of itemsets $\mathcal{K}$. We say that the hypothesis $\mathcal{H}$ is rejected by the χ^2 test of mutual constrained independence for the threshold α if:

$$\chi^2_{\mathbf{p},\mathbf{f}} > \chi^2_\alpha(d')$$

where $d' = d+1-\#D-\#\mathcal{P}_{L|D}$ is the number of degrees of freedom determined above and $\chi^2_\alpha(d')$ is the value such that $\text{Prob}\,(Z > \chi^2_\alpha(d')) = \alpha$ if $Z \sim \chi^2(d')$.

As for most statistical tests, it is important to note that a single MCI hypothesis corresponds in fact to three distinct statements, not all of which are tested equally by the statistical test. First, the hypothesis contains the statement that the we can model the data as n independent identically distributed samples of a random variable. Second, the hypothesis says that $\mathbf{p}_{|\mathcal{K}} = \mathbf{f}_{|\mathcal{K}}$. And third, the hypothesis states that the items $a_1, ..., a_m$ are mutually constrainedly independent in $\mathcal{B}$ with regards to the constraints defined by $\mathcal{K}$ for the measure $\mathbf{p}$ (see definition 4.1.2). The first statement is not tested per se by the statistical test because the test is only meaningful if such an assumption is made. The second and the third statements of the hypothesis, however, are both simultaneously tested by the test because they define the probability measure $\mathbf{p}$ which is compared to the empirical distribution $\mathbf{f}$. This is not necessarily an issue, if ones aim is really to test the hypothesis as a whole. But it is an issue, if one is more interested in the third statement of the hypothesis (i.e. the mutual constrained independence of the items relative to $\mathcal{K}$) than the second, which is quite often the case when researchers test for independence in current studies. Indeed, if the hypothesis $\mathcal{H}$ is rejected, this does not automatically imply that we would reject any alternative hypothesis based on a probability measure $\mathbf{q}$ such that the items $a_1, ..., a_m$ are mutually constrainedly independent in $\mathcal{B}$ with regards to the constraints defined by $\mathcal{K}$ for this measure $\mathbf{q}$ but for which $\mathbf{q}_{\mathcal{K}} \neq \mathbf{f}_{\mathcal{K}}$.[2] One way to tackle this issue is to use the notion of confidence in the empirical distribution which we have defined in section 3.7.2.1. Indeed, if we are confident in the empirical distribution, then we consider the second statement to be true whether the hypothesis is rejected or not. Hence, in such a case, we can consider that rejecting $\mathcal{H}$ does indeed imply a rejection of the more general hypothesis that the items are mutually constrainedly independent relative to $\mathcal{K}$.

In the following sections, we will use this statistical test as a tool for pattern mining. However, we also believe this tool could be put to use to assess single predefined hypotheses, as in standard statistical testing. In order to facilitate

[2] To prove this, it is sufficient to show an example in which $n\chi^2_{\mathbf{p},\mathbf{f}} > n\chi^2_{\mathbf{q},\mathbf{f}}$. Such an example can be obtained easily by twitching the parameters of a MCI model. For example, consider two items a_1 and a_2, an empirical distribution such that $\mathbf{f} = (0.4, 0.2, 0.3, 0.1)$ and the two following probability measures $\mathbf{p}$ and $\mathbf{q}$ both defined by the independence of a_1 and a_2 together with the constraints that $p_{a_1} = f_{a_1} = 0.4$ and $p_{a_2} = f_{a_2} = 0.3$, on the one hand, and $q_{a_1} = 0.401$ and $q_{a_2} = 0.302$, on the other hand. Then $n\chi^2_{\mathbf{p},\mathbf{f}} \approx 0.007937$ while $n\chi^2_{\mathbf{q},\mathbf{f}} \approx 0.007922$.

such a use, we are currently implementing a Python module for MCI testing in low dimension, based on the algorithms described in chapter 4.

5.2 Discovering a valid global MCI hypothesis

In this section, we will only consider global MCI hypotheses which are defined on the entire Boolean lattice $\mathcal{B}$ (see section 3.7.3.1 for more details on global hypotheses). Local MCI hypotheses will be discussed briefly in section 5.3.

We consider a binary dataset $\mathcal{D}$ as previously and we will assume that the number of transactions n in the dataset is sufficiently large to be confident in the empirical distribution. Furthermore, we will start by assuming that we have no prior knowledge about the data. Note that, together, these two assumptions imply that $f_i \neq 0$ for all $i \in [\![0, d]\!]$ (see section 3.7.2.2). We will address the issue of datasets for which some values of f_i are null in section 5.2.4.

5.2.1 Valid MCI hypotheses

As we have expressed previously, our aim is to discover scientific information about the data in the form of a MCI hypothesis. If a hypothesis qualifies as scientific information about the data, we will say it is a *valid hypothesis*. Of course, a valid hypothesis should not be rejected if tested. However, passing a test is not sufficient to qualify a hypothesis as scientific information. In fact, there could be a large number of hypotheses that would not be rejected if they were tested and not all of these hypotheses should be considered valid hypotheses. For a hypothesis to be considered valid, we must be able to justify that we have a reason to test this hypothesis. This idea is related to the notion of the burden of proof which can be illustrated by Bertrand Russell's famous *celestial teapot* analogy [SK97]: even if the claim that there is a china teapot orbiting the sun between the Earth and Mars cannot be disproved, we should still ignore it because there are no grounds for such a claim.

In the classical hypothetico-deductive model for the scientific method (see section 3.7), hypotheses are formulated based on prior observations and reasoning. If a hypothesis is rejected, then a new hypothesis may be formulated, based on the accumulated knowledge resulting from the prior knowledge, as well as the observations which led to the rejection of the initial hypothesis.

In our case however, as we have discussed in section 3.7.1, the fact that we are considering finite static data implies that we cannot formulate a hypothesis based on prior empirical observations. However, we can rely on Occam's razor and define the simplest hypothesis possible. If this hypothesis is not rejected given the data, then the process ends there. Otherwise, we consider the simplest remaining hypothesis and we continue the process until we have determined the simplest possible hypothesis which is not rejected. In order to accomplish this, we must therefore define a total order relation between all possible hypotheses, based on a notion of simplicity and regardless of the dataset considered, as mentioned in section 3.7.4.1.

5.2.2 Ordering $\mathcal{P}(\mathcal{I})$

Defining an order relation on all possible hypotheses regardless of the data considered comes down to defining an order relation on the powerset $\mathcal{P}(\mathcal{I})$. As we have mentioned above, the order defined should reflect a notion of simplicity (or, inversely, a notion of complexity) because we need to be able to compare two hypotheses in terms of simplicity to choose a hypothesis based on Occam's razor. This is not a trivial task and is linked to the issue of the complexity of Boolean expressions discussed in section 3.6.2.3. While we do not aim at providing a perfect solution to this problem, we suggest some criteria which may be taken into account and present an order relation defined accordingly. Note that considering a set $\mathcal{K}$ and the set $\mathcal{K} \cup \{\top\}$ are equivalent in terms of MCI hypotheses (because $\mathcal{K}^* = \mathcal{K} \cup \{\top\}$ if $f_i \neq 0$ for all $i \in [\![0, d]\!]$). In fact, the order relation which we are considering need only be defined on $\{\mathcal{K}^* \mid \mathcal{K} \subset \mathcal{I}\}$ which is equal to $\{\mathcal{K} \cup \{\top\} \mid \mathcal{K} \subset \mathcal{I}\}$ here. Hence, all sets $\mathcal{K}$ considered in the following will be such that $\top \in \mathcal{K}$.

5.2.2.1 A possible order relation

First, it seems legitimate to require that $\mathcal{K} \leq \mathcal{K} \cup \{I\}$ for any $\mathcal{K} \subset \mathcal{I}$ and $I \in \mathcal{I}$. In other words, adding more itemsets to the set of itemsets with constrained frequencies complexifies the hypothesis. Second, it also seems reasonable to require that $\mathcal{K} \cup \{I\} \leq \mathcal{K}' \cup \{J\}$ for all $\mathcal{K}, \mathcal{K}' \subset \mathcal{I}$, $I \in \mathcal{I} \setminus \mathcal{K}$ and $J \in \mathcal{I} \setminus \mathcal{K}'$, such that $\mathcal{K} \leq \mathcal{K}'$ and $|I| \leq |J|$. Indeed, we can reasonably consider that an itemset I is less complex than an itemset J, if I has less items than J, and that this relationship should pass on to sets of itemsets. However, these two requirements are not enough to define a unique ordering on $\mathcal{P}(\mathcal{I})$ and we need to add other criteria.

One approach is to build on the idea that standard mutual independence is considered a rather simple case. If we require that the set corresponding to the independence model is the simplest possible one (given the two previous requirements) then the set $\{\top, a_1, ..., a_m\}$, which contains all the single items, is considered less complex than $\{\top, I\}$ for any $I \in \mathcal{I}$ such that $|I| \geq 2$. Hence, a single itemset of size 2 or more is considered more complex than any number of itemsets of size 1. We can then generalize this idea so that any single itemset of a given size is considered more complex than any number of itemsets with strictly smaller size.

Nevertheless, such a criteria is still not sufficient to define a total order relation on $\mathcal{P}(\mathcal{I})$ because it does not allow to compare two sets of same size, each of which contain itemsets of a same given size. However, if we consider that all itemsets of a given size are equally complex because items are interchangeable in terms of complexity, then we might consider that all sets of a given number of itemsets of a given size are equally complex[3]. Hence, we can choose any total order relation between such sets and one of the simplest options is a double lexicographic order (itemsets of same size are ordered by lexicographic order and sets of itemsets of same size are ordered by lexicographic order based on the initial lexicographic order).

Formally, we can express the order described above as follows. For any $\mathcal{K} \subset \mathcal{I}$, we note:

$$\mathcal{K} = \left\{ \bigwedge_{i \in K_1} a_i, ..., \bigwedge_{i \in K_r} a_i \right\}$$

and associate, to each $\mathcal{K}$, sets C_j defined for all $j \in [\![0, m]\!]$ by:

$$C_j = \{K_i \mid i \in [\![1, r]\!] \wedge |K_i| = j\}$$

Now, we can say that:

$$\mathcal{K}^{(1)} \leq \mathcal{K}^{(2)}$$

[3]As we will show further on, this is also debatable, but we will still use this as an assumption here to define a total order on $\mathcal{P}(\mathcal{I})$.

if and only if:

$$\exists i \in [\![0,m]\!],\ \left[\left|C_i^{(1)}\right| < \left|C_i^{(2)}\right| \wedge \forall j \in [\![i+1,m]\!],\ \left|C_j^{(1)}\right| = \left|C_j^{(2)}\right|\right]$$

$$\bigvee$$

$$\left[\forall i \in [\![0,m]\!],\ \left|C_i^{(1)}\right| = \left|C_i^{(2)}\right|\right] \wedge \left[\exists i \in [\![0,m]\!],\ \left[C_i^{(1)} \prec C_i^{(2)} \wedge \forall j \in [\![i+1,m]\!],\ C_j^{(1)} = C_j^{(2)}\right]\right]$$

where $C_i^{(1)} \prec C_i^{(2)}$ is given by the double lexicographic order described above.

5.2.2.2 Further discussions on the definition of an order relation

One of the issues with this order relation is that sets of itemsets corresponding to a same equivalence class do not all appear subsequently in this order (where the equivalence relation is defined by the set of permutations on items as in section 4.2.2.6). For example, when considering three items a_1, a_2 and a_3, the order defined above gives:

$$\{\top, a_1, a_2, a_1 \wedge a_2\} < \{\top, a_1, a_2, a_1 \wedge a_3\} < \{\top, a_1, a_3, a_1 \wedge a_3\}$$

where $\{\top, a_1, a_2, a_1 \wedge a_2\}$ and $\{\top, a_1, a_3, a_1 \wedge a_3\}$ belong to a same equivalence class and $\{\top, a_1, a_2, a_1 \wedge a_3\}$ belongs to a different one. This is because we have considered that the complexity of sets of a given number of itemsets of a given size is constant so that we can choose to order such sets in any given way. One could consider instead that complexity is only constant on the equivalence classes described above, so that sets in any given equivalence class should appear subsequently for the total order which we define.

Such an approach could bring the notion of complexity which we are trying to model closer to a notion of Kolmogorov complexity because the computational complexity for obtaining the MCI model from the algebraic reductions presented in section 4.2.2 is the same for all the sets $\mathcal{K}$ in any given equivalence class. Note that, if the order relation is defined based solely on the computational complexity described above, the condition that any single itemset of a given size be considered more complex than any number of itemsets of strictly smaller size would no longer hold. Indeed, the case $\mathcal{K} = \mathcal{I} \setminus \{I_d\}$, which we have studied in section 4.2.1, corresponds to a relatively easy case in terms of computational complexity, while it necessarily holds a median position in the order relation if the aforementioned condition is true. Though we acknowledge this general issue, we do not address it any further in this thesis, leaving the task of quantifying the complexity of each equivalence class for further research.

Another approach, which allows to avoid the issue of forcibly and arbitrarily defining an order relation between sets of itemsets within a same class, is to only consider hypotheses which are defined by sets that are invariant by permutation of items. For example, when considering three items, we would only consider the following sets, ordered as such:

$$\begin{gathered}
\{\top\} \\
< \\
\{\top, a_1, a_2, a_3\} \\
< \\
\{\top, a_1 \wedge a_2, a_1 \wedge a_3, a_2 \wedge a_3\} \\
< \\
\{\top, a_1, a_2, a_3, a_1 \wedge a_2, a_1 \wedge a_3, a_2 \wedge a_3\} \\
< \\
\{\top, a_1 \wedge a_2 \wedge a_3\} \\
< \\
\{\top, a_1, a_2, a_3, a_1 \wedge a_2 \wedge a_3\} \\
< \\
\{\top, a_1 \wedge a_2, a_1 \wedge a_3, a_2 \wedge a_3, a_1 \wedge a_2 \wedge a_3\} \\
< \\
\{\top, a_1, a_2, a_3, a_1 \wedge a_2, a_1 \wedge a_3, a_2 \wedge a_3, a_1 \wedge a_2 \wedge a_3\}
\end{gathered}$$

This approach also has the advantage of only considering 2^m potential hypotheses rather than 2^{2^m-1}. However, it has the disadvantage of disregarding potentially interesting hypotheses.

5.2.3 Search algorithms

In the following, assume we are using the order relation described in section 5.2.2.1. Our aim now is to find the smallest possible $\mathcal{K}$ for this order relation, such that the associated hypothesis is not rejected by a χ^2 test of mutual constrained independence.

5.2.3.1 Comprehensive search

A naive approach would be to test all potential hypotheses, in increasing order, until a valid hypothesis is determined. This is practically infeasible for all datasets for which $m \geq 7$ (and still extremely difficult for $m = 6$) because the number of hypotheses to test is equal to 2^{2^m-1}. However, without any firm mathematical result on the behavior of the χ^2 statistic with respect to $\mathcal{K}$ (which we will note $\chi^2_{\mathcal{K}}$ in the following), it is impossible to suggest an alternative which would be both more efficient computationally and guaranteed to succeed (even if only asymptotically almost surely).

This is why limiting the search to a smaller number of hypotheses, as described in the last example in section 5.2.2.2 in which the search is limited to 2^m hypotheses, could be considered an option. We can even reduce the size of the search space further by another logarithmic factor if we start by $\{\top\}$ (which is the complete layer of itemsets of size 0) and incrementally complexify the hypothesis by adding the next complete layer of itemsets of a given size at each iteration. In such a case, the size of the search space is reduced to $m+1$, which is a tremendous gain in terms of computational complexity, but we must not forget that this also corresponds to a tremendously simplified version of the problem.

5.2.3.2 Greedy algorithms

Alternatively, we could rely on various heuristics to obtain an eventually sub-optimal solution. Greedy algorithms offer a simple and quite satisfying option. In the following, we consider two different greedy approaches, which can be combined together: greedy-up and greedy-down. In the algorithms presented below, we will use the following indicator:

$$\xi_{\mathcal{K}} = \frac{\chi^2_{\mathcal{K}}}{\chi^2_{\alpha}(2^m - 1 - |\mathcal{K}|)}$$

which is greater than one if and only if the hypothesis associated to $\mathcal{K}$ is rejected for the threshold α.

The greedy-up approach is detailed in the following algorithm:

```
1.  initialize  K ← ∅;
2.  while  (ξ_K ≥ 1)  and  (∃I ∈ I \ K, ξ_{K∪{I}} < ξ_K):
```

$\quad\quad J \leftarrow \underset{I \in \mathcal{I} \setminus \mathcal{K}}{\arg\min} \left(\xi_{\mathcal{K} \cup \{I\}}\right)$

$\quad\quad \mathcal{K} \leftarrow \mathcal{K} \cup \{J\}$

if $\left(\xi_{\mathcal{K}} < 1\right)$:

$\quad\quad$ output $\mathcal{K}$

else :

$\quad\quad$ output $\mathcal{I}$

Algorithm 5.1: Greedy-up

The algorithm consists in building a strictly increasing sequence of sets of itemsets ($\mathcal{K}$), corresponding to a strictly decreasing sequence ($\xi_{\mathcal{K}}$). It terminates as soon as we have reached a $\mathcal{K}$ for which the associated hypothesis is no longer rejected or if we have reached a local minimum for $\xi_{\mathcal{K}}$ for which the associated hypothesis is still rejected. In the first case, the associated hypothesis is considered a valid hypothesis and, in the second case, we consider that no valid hypothesis can be formulated about the data.

In the greedy-down approach, we start from a constrained set of itemsets $\mathcal{K}_0$, associated to a hypothesis $\mathcal{H}_0$ which is not rejected and gradually simplify it. Note that we can only apply such a process if we already have such a $\mathcal{K}_0$ at our disposal.

initialize $\mathcal{K} \leftarrow \mathcal{K}_0$;

while $\left(\exists I \in \mathcal{K},\ \xi_{\mathcal{K} \setminus \{I\}} < 1\right)$:

$\quad\quad J \leftarrow \underset{I \in \mathcal{K},\ \xi_{\mathcal{K} \setminus \{I\}} < 1}{\arg\min} \ \xi_{\mathcal{K} \setminus \{I\}}$

$\quad\quad \mathcal{K} \leftarrow \mathcal{K} \setminus \{J\}$

output $\mathcal{K}$

Algorithm 5.2: Greedy-down

A greedy-up algorithm can be combined with a greedy-down one: first, $\mathcal{K}$ is complexified until we find a hypothesis which is not rejected, and then this hypothesis is simplified as much as possible. In this case, we can consider the resulting hypothesis as valid.

In the worst case scenario, $\xi_{\mathcal{K}}$ is computed for $\frac{2^m(2^m-1)}{2}$ different values of $\mathcal{K}$ for the greedy-up algorithm and $\frac{k_0(k_0-1)}{2}$ values for the greedy-down algorithm (where $k_0 = |\mathcal{K}_0|$). Hence, the computational complexity of the combined greedy-up, greedy-down algorithm is in $O(2^{2m})$ (under the simplified assumption that the complexity for computing the MCI model is constant which is not necessarily true as we have previously mentioned).

While this is an improvement in comparison to testing all potential hypotheses, of which there are 2^{2^m-1}, it can still be considered too much in many cases. In order to reduce this further to a more reasonable quantity, one may consider limiting the size of the itemsets considered to a certain value k. In this case, the number of hypotheses to test is polynomial in m of degree $2k$.

5.2.3.3 Efficiency of the greedy algorithms

In order to assess the efficiency of the greedy approach, we performed some tests to check how often the algorithms described above reached the same solution as the naive approach. As the naive approach can only be performed for a small number of items, we fixed the number of items to $m = 5$. Furthermore, in order to be confident that the hypotheses extracted corresponded indeed to meaningful knowledge about the data, we only considered artificially generated data.

The datasets considered were generated as follows. For each dataset, five values $p_1, ..., p_5$ were randomly and independently picked between 0.1 and 0.9. We then generated a million random binary vectors in $\{0, 1\}^5$ each of which corresponded to five independent Bernoulli trials with parameters $p_1, ..., p_5$. This procedure was repeated 1000 times, thus generating 1000 different datasets. The probability threshold for rejecting hypotheses was set to $\alpha = 99.9\%$.

Out of the 1000 datasets generated, the optimal constrained set for the hypothesis corresponded to:

- $\mathcal{K}_{\text{ind}} = \{\top, a_1, a_2, a_3, a_4, a_5\}$ (i.e. the independence model) in 962 cases;
- a subset of $\mathcal{K}_{\text{ind}}$ within the permutation class of $\mathcal{K}_{\text{ind}} \setminus \{a_5\}$ in 32 cases (corresponding to cases in which the values of the randomly generated probabilities for the missing item were very close to $\frac{1}{2}$);
- a superset of $\mathcal{K}_{\text{ind}}$ within the permutation class of $\mathcal{K}_{\text{ind}} \cup \{a_1 \wedge a_2\}$ in 3 cases;
- a superset of $\mathcal{K}_{\text{ind}}$ within the permutation class of $\mathcal{K}_{\text{ind}} \cup \{a_1 \wedge a_2 \wedge a_3\}$ in the 3 remaining cases.

The greedy-up algorithm led to the optimal solution in 432 cases. The greedy-up algorithm followed by the greedy-down algorithm led to the optimal solution in 998 cases, leaving but 2 cases in which the optimal solution was not reached. These are extremely encouraging results, not only because the greedy-up, greedy-down approach manages to reach the optimal solution in 99.8% of the cases, but also because the general method allows to determine the exact nature of the generating model in 96.2% of the cases observed and a very close model in 100% of these cases.

Naturally, the tests which have been conducted concern a particular type of datasets but we hope to be able to confirm the success of this method when applied on artificial datasets generated through more elaborate models.

5.2.3.4 Comparison with compression approaches

It has also been suggested in the literature that the sets of itemsets for a given database can be ranked by interestingness based on the compression scores of their associated MCI models (see section 3.7.4.1). In order to compare this approach with our own, we also used compression scores (specifically MDL and BIC as defined in [MVT12]) to rank the sets of itemsets for each of the 1000 artificial datasets mentioned in section 5.2.3.3 above.

In every single case, the random coin toss model associated to $\{\top\}$ was ranked best. Even when removing this specific model, the best model after this one was never associated to the independence model, despite the fact that the data was generated based on this model. Hence, for this particular example at least, our approach seems to be much more adequate. Further tests will be conducted on artificial datasets generated through more elaborate models in order to assess the efficiency of our approach in comparison to approaches based on compression scores on a wider scale.

5.2.4 Dataset with locally null frequencies

In the previous sections, we have only considered datasets $\mathcal{D}$ for which $f_i \neq 0$ for all $i \in [\![0, d]\!]$. However in most datasets, there is a non-empty set D such that $f_i = 0$ for all $i \in D$.

5.2.4.1 Theoretical issues

One of the important issues with such datasets is that, we can never be confident in the corresponding empirical distribution, unless we have background knowledge that states that $p_i = 0$ for all $i \in D$. However, if we consider all potential hypotheses associated to MCI models, some of these will correspond to probability measures $\mathbf{p}$ such that $p_i \neq 0$ for some $i \in D$. Hence, if we consider all possible MCI models, we are ignoring the background knowledge which is actually necessary to consider the process of extracting scientific knowledge from the dataset.

The second important issue related to such datasets is actually determining the corresponding background knowledge. Up to now, we have considered this to be an independent issue which must be settled before we examine the data. Indeed, it seems difficult to justify that we use the empirical distribution to determine background knowledge which is necessary to assess the confidence which we have in the empirical distribution. However, there are also theoretical grounds that justify that we do use the dataset to define the set $D = \{i \in [\![0, d]\!] \mid p_i = 0\}$.

Indeed, one can argue that it is in fact a stronger assumption to presume the possibility of the existence of a given eventuality compared to the assumption that this eventuality is impossible. This is related to the philosophical notion of the burden of proof which we have already evoked in section 5.2.1. If an eventuality has never been observed, then we should have some kind of reason to believe that it could be observed if we want to consider it as a possibility. With this perspective, an absence of knowledge leading to the possible existence of any given eventuality should correspond to the belief in the impossibility of this eventuality. Hence, if we have no prior knowledge whatsoever about a dataset of transactions on items, then it is the fact that we observe the transactions in the dataset that allows us to say that these particular transactions can exist. By contrast, we have no basis to assume that the transactions which are not present in the data can exist. Hence, only hypotheses corresponding to probability measures such that $p_i = 0$ whenever $f_i = 0$ should be considered. If we follow this theoretical argument, then D should be in fact defined from the data as $D = \{i \in [\![0, d]\!] \mid f_i = 0\}$ and this knowledge should be not be ignored during the pattern mining process.

Nevertheless, note that this is a delicate stance to uphold. Indeed, it mostly never occurs that we can claim to have a complete absence of knowledge which would lead to the possible existence of any transaction before its observation in the data. In fact, if we choose to build a dataset of transactions between items, this already means we have some reason to believe that some transactions are possible (because, if this is not the case, the dataset itself would be an impossible object) but this does not necessarily imply that we have reason to believe that all transactions are possible. In the classical market basket example, we have reason to believe that some transactions are possible even if we have not observed a single transaction because we know how retail stores work. However, this does not imply that we have reason to believe that a single transaction containing all the items in the store is possible and, in fact, we also know that such a transaction is impossible because of our knowledge of how retail stores work. Such a nuanced position, in between the two extremes defined by $D = \emptyset$, on the one hand, and $D = \{i \in [\![0, d]\!] \mid f_i = 0\}$, on the other hand, cannot be meaningfully justified without an additional modeling layer, which would include a description of the knowledge we had on items prior to considering a dataset of transactions on these items. As addressing this issue goes way beyond the scope of this thesis, we leave it as open problem for further research.

5.2.4.2 Practical implications

Rather than complexifying the issue of finding a valid MCI hypothesis, considering a dataset for which $D \neq \emptyset$ can only make the process more easily computable. In fact, as we will show, this may even allow us to consider cases with a larger number of items m than what is technically feasible if $D = \emptyset$.

In the following, we will note t the number of different transactions existing within the dataset, so that $t = 2^m - |D|$. Furthermore, note that $t \leq n$. In practice, the number of transactions n in a dataset is rarely exponential in the number of items m. Hence, for most datasets considered in itemset mining, $t \ll 2^m$. The practical implications of this statement can be viewed with regards to three aspects.

Firstly, if we consider that D is determined from the data, then we must add the constraints that $p_i = 0$ for all in $i \in D$ when computing the probability measure $\mathbf{p}$ for an MCI model, regardless of the set of itemsets $\mathcal{K}$ defining this particular MCI model. If $D \neq \emptyset$, this can only reduce the complexity of computing $\mathbf{p}$ because the number of variables p_i to compute is equal to t rather than 2^m.

Secondly, the size of the search space is also reduced as D increases. Indeed, the search space corresponds to $\{K^* \mid K \subset \mathcal{I}\}$, which has cardinality 2^{2^m-1} if $D = \emptyset$, but decreases when D increases, because increasing D increases the number of frequencies of itemsets which can be directly derived from the frequencies of other itemsets. The manner in which the search space decreases is, however, not a trivial issue and we leave this to further research.

Thirdly, the number of transactions n which is necessary to be confident in the empirical distribution is at least equal to $O(2^m)$ and at most equal to $O(8^m)$ with a reasonable approximation at $O(4^m)$ (see section 3.7.2.2) if we consider that $D = \emptyset$. However, the demonstrations presented in section 3.7.2.2 still hold if we consider that the number of degrees of freedom is equal to $t - 1$ rather than simply $2^m - 1$ and the number of necessary transactions to be confident in the empirical distribution becomes at most equal to $O(t^3)$. This can seriously bring down the total number of transactions which is necessary to be confident in the empirical distribution if the number t of existing transactions in the data is reasonable.

5.3 Using local MCI hypotheses

In sections 3.5 and 3.7, we addressed a number of theoretical issues pertaining to the use of local models and advised mostly against their use if the meaningfulness of the modeling for the pattern mining process is a main concern. Oppositely, it is practically infeasible to compute multiple global MCI models and perform the corresponding χ^2 tests of mutual constrained independence in order to discover a valid global MCI hypothesis for datasets with values of m as low as 20 (at least if $D \neq \emptyset$). For this reason, some form of compromise should be made if we want to consider such datasets and replacing global MCI hypotheses with local MCI hypotheses is one way to go. As this comes to the cost of meaningfulness, which is a central issue in this doctoral thesis, we will not delve much on such an approach. We present in this section some of the theoretical issues related to an approach based on local MCI hypotheses, and layout the bases for future work on this topic.

In the following, we will assume that we must limit the number of different items which can be considered to define a MCI model to a fixed value $l_{\max}$ for technical reasons. Furthermore, we will consider a dataset $\mathcal{D}$ of n transactions on m items, with $m > l_{\max}$, so that we cannot define any global MCI models based on the empirical distribution defined by $\mathcal{D}$. The local MCI models which we will consider instead will all be defined as probability measures $\mathbf{p}'$ over a Boolean lattice $\mathcal{B}'$ associated to a subset $\mathcal{A}'$ of the set of items $\mathcal{A} = \{a_1, ..., a_m\}$. We will note l' the number of items in $\mathcal{A}'$ and we necessarily have $l' \leq l_{\max}$.

5.3.1 Theoretical issues

For each of the subsets $\mathcal{A}' \subset \mathcal{A}$ which we consider, we can use the methods described in section 5.2 to determine a valid local hypothesis[4] $\mathcal{H}'$ if such a hypothesis exists. Hence, we can obtain a set of valid local hypotheses, noted $\mathbb{H}$, which is uniquely determined by the dataset $\mathcal{D}$ and the set of subsets $\mathcal{A}'$ which we have considered, which we will note $\mathbb{A}$. As our aim is to extract some piece of scientifically valid information about $\mathcal{D}$ generally, we must construct a valid global hypothesis $\mathcal{H}$ which is consistent with the hypotheses $\mathcal{H}'$ in $\mathbb{H}$. This raises two issues: the consistency of $\mathcal{H}$ and the validity of $\mathcal{H}$.

5.3.1.1 The issue of overlapping and global consistency

The issue of global consistency, already mentioned in sections 3.5.1.2 and 3.7.4.2, is a complex issue which we will not settle in general in this section. We will limit the discussion here to how inconsistency might arise and how it can be simply avoided.

If $\mathbb{A}$ is not a partition of $\mathcal{A}$, that is, if two elements of $\mathcal{A}', \mathcal{A}'' \in \mathbb{A}$ overlap, then it is possible that an MCI model $\mathbf{p}'$ defined over $\mathcal{B}'$ and an MCI model $\mathbf{p}''$ defined over $\mathcal{B}''$ are not globally consistent, in the sense that there is no probability measure $\mathbf{p}$ on $\mathcal{B}$ such that the restriction of $\mathbf{p}$ to $\mathcal{B}'$ is equal to $\mathbf{p}'$ and the restriction of $\mathbf{p}$ to $\mathcal{B}''$ is equal to $\mathbf{p}''$ (see section 3.5.1.2 for an example of this type of inconsistency). This issue does not arise when considering non-overlapping elements of $\mathbb{A}$ (i.e. $\mathcal{A}'$ and $\mathcal{A}''$ such that $\mathcal{A}' \cap \mathcal{A}'' = \emptyset$), as we can at least consider a joint independent model as in section 4.2.2.4. As each valid local hypothesis $\mathcal{H}'$ corresponds to a local MCI model, if we consider that $\mathbb{A}$ is a non-overlapping covering of $\mathcal{A}$ (in other words a partition of $\mathcal{A}$), then we avoid the issue of inconsistency altogether.

Note, however, that it is not obvious that we could build an example of inconsistency between valid local MCI hypotheses as easily as we have con-

[4]We use the term valid local hypothesis to refer to a local hypothesis which would be valid if we were to consider the corresponding local context as a separate global context.

structed an example of inconsistency between local MCI models. Indeed, we would expect that one at least of the two local MCI hypotheses corresponding to two local MCI models would be rejected if these two models were globally inconsistent, because this is asymptotically the case and we are considering datasets for which we are confident in the empirical distribution. Nevertheless, without further mathematical knowledge with regards to this specific aspect, we will restrict ourselves to cases in which $\mathbb{A}$ is a partition of $\mathcal{A}$.

5.3.1.2 The issue of multiple testing and global validity

Global validity is another important issue which can be related to the effects of simultaneously testing a large number of hypotheses. The larger $\mathbb{A}$ is, the higher the likelihood that some of the elements in $\mathbb{H}$ correspond actually to false positives. This aspect refrains us from being able to assert that the hypotheses in $\mathbb{H}$ are globally valid. While this issue has been studied in other contexts, it has not been addressed with regards to MCI hypotheses (or equivalent hypotheses based on MaxEnt models) other than those associated with independence (see, for example,[LTP06, Han11, KIA$^+$17]). We set this topic aside here, leaving it to be investigated in further research.

Note that, even if we are certain that all the local hypotheses in $\mathbb{H}$ are considered valid, this does not imply that any global hypothesis $\mathcal{H}$, consistent with the local MCI hypotheses $\mathcal{H}'$ in $\mathbb{H}$, can also be considered valid. In particular, even a global MCI hypothesis which is consistent with a great number of valid local MCI hypotheses can be rejected.[5] Hence, we cannot define a valid global hypothesis $\mathcal{H}$ from $\mathbb{H}$ alone.

5.3.2 Thoughts on the partition model

As we have mentioned above, global inconsistency is no longer an issue if $\mathbb{A}$ is a partition of $\mathcal{A}$, regardless of the partition $\mathbb{A}$ chosen. Hence, considering a pattern mining approach based on local MCI models associated to such a partition seems to be an interesting perspective. However, meaningfully determining an appropriate partition is not a trivial task.

Indeed, if we would randomly partition $\mathcal{A}$ in blocks $\mathcal{A}'$ of size less than

[5]The fact that we can construct a probability measure such that all items are pairwise independent which, at the same time, does not correspond to the mutual independence model is sufficient to prove this point.

$l_{\max}$, the hypotheses $\mathcal{H}'$ associated to each block would be globally consistent. However, we would not know how we should join the local probability measures $\mathbf{p}'$ together to make a global probability measure $\mathbf{p}$. Furthermore, the information within each partition would not necessarily be the most interesting because items would not be grouped on the basis that they are strongly connected through a complex pattern. Hence, the most interesting patterns could be hidden within the definition of the junction model.

5.3.2.1 A simple junction model

Ideally, the partition of $\mathcal{A}$ should be determined from the data, concomitantly with a means to construct a junction model $\mathbf{p}$ from the local models $\mathbf{p}'$, and in such a way that the local models in each block of the partition contain the most complex aspects of the structure of $\mathbf{p}$. Intuitively, if the complexity of the general model is concentrated in the local models, then the junction model should be quite simple.

Given two models $\mathbf{p}'$ and $\mathbf{p}''$, defined for two disjoint blocks $\mathcal{A}'$ and $\mathcal{A}''$, we can consider two simple junction models:

- the joint independence model;
- the joint incompatibility model.

On the one hand, the joint independence model corresponds to the model defined in section 4.2.2.4. On the other hand, the joint incompatibility model is defined by $p_{a' \wedge a''} = 0$ for all $a' \in \mathcal{A}'$ and $a'' \in \mathcal{A}''$.

More generally, we can define the joint mutual independence model for any number of models corresponding to mutually disjoint blocks by analogy with the definition of mutual independence. Similarly, we can also define a joint mutual incompatibility model. Lastly, we can define a combined junction model so that local models $\mathbf{p}'$ can be joined in groups of joint mutually independent models, each of which can then be joined through a joint incompatibility model.

5.3.2.2 Determining the right partition

Partitioning the set of items $\mathcal{A}$ is effectively equivalent to variable clustering. Such clustering methods have been considered in the context of association rule mining [PNS+07] and similar methods could be envisaged in the present context. Note however that, for the sake of meaningfulness, the choice of a given clustering method should be consistent with a chosen junction model.

If a joint incompatibility model is considered, the clustering should reflect the notion that the frequency of any itemset spanning over multiple clusters is negligible. A very simple option would be to consider a graph such that: (1) vertices are identified with items; (2) there is an edge between two vertices if and only if the frequency of the itemset defined by the two corresponding items is above a given support threshold (under which the frequency of an itemset is considered negligible). We can then identify each cluster of items to a connected component of the graph. The Apriori principle ensures that the frequency of any itemset spanning over multiple clusters is lower than the given support threshold and can therefore be considered negligible. Note that the value of the support threshold should be the lowest possible value that guarantees that the size of each cluster is less than l_m in order to achieve computability while limiting the cost to meaningfulness.

Similarly, if a joint independence model is considered, the clustering method should ensure that the joint independence model between the local empirical distributions for each cluster may be considered a decent approximation of the global empirical distribution. This is a more complex issue than in the case of the joint incompatibility model and we leave it for further investigation.

Furthermore, in order to determine an adequate partition for a combined junction model, a clustering method adapted to the joint incompatibility model can easily be combined with a clustering method adapted to the joint independence model in a two phase process, with the initial clustering phase corresponding to the joint incompatibility model between the clusters and the second clustering phase corresponding to local joint independence models between subclusters within each initial cluster. In this manner, we can hope to reduce the complexity of determining an adequate clustering for a joint independence model by setting a cap on the size of the initial clusters.

Alternately, the choice of the partition could also rely on background knowledge about the items and their interactions, which brings us once again to the necessity of combining induction with other forms of inference.

5.4 Conclusion

In this chapter, we presented a novel approach for mining objectively interesting patterns in a dataset of transactions on items. The principle of this approach may be summarized as follows.

The patterns considered are sets of itemsets, together with their corresponding frequencies in the empirical distribution. Each pattern is associated to an objective hypothesis stating that the data is the result of the sampling of n independent identically distributed random variables whose distribution is given by the MCI model defined by the pattern. Hypotheses are then individually tested, following an increasing order of complexity, using a χ^2 test of mutual constrained independence. The process terminates as soon as it reaches a hypothesis which is not rejected by the test. Following Occam's razor, this particular hypothesis gives the objectively interesting pattern which we extract from the data.

Note that this approach relies on a total ordering of the hypotheses based on their complexity. Defining such an ordering is not necessarily trivial and we presented a possible solution to this problem.

Furthermore, the search space which contains all objective hypotheses has cardinality 2^{2^m-1}. This is much too large to consider an exhaustive search beyond $m = 5$ and we must resort to heuristics instead to determine a (possibly suboptimal) solution. Considering artificial datasets with $m = 5$ (thus allowing for an exhaustive search), we experimented with greedy approaches. We showed that a combined approach of a greedy-up algorithm, followed by a greedy-down algorithm, can produce very satisfying results, reaching the optimal solution in the vast majority of cases. These first experiments on artificial datasets are also encouraging with regards to the aim of the approach as the generating models for the artificial data were correctly identified in a large majority of cases. From this perspective, our approach surpasses by far the existing pattern mining approaches using MaxEnt models for summarization which systematically failed to determine these generating models.

Nevertheless, it is important to note that the approach described above is limited to small values of m (roughly less than 20), even when considering a greedy search algorithm. In order to consider possible applications for datasets with larger values of m, it must be combined with other approaches. We presented the contours of an approach based on the partitioning of the set of items in small independent or incompatible clusters which we intended to develop in the near future.

More generally, much of the work presented in this chapter calls for further investigation and we intend to pursue our research accordingly (see section 6.1 for more details).

CHAPTER 6

Epilogue

We conclude this doctoral thesis by presenting some specific aspects which we will be working on in the close future, followed by some more general considerations. For a concise summary of the specific contributions of this thesis to current research, we refer the reader to section 1.5.

6.1 Follow-up research

The research we have conducted during this doctoral thesis has opened a wide scope of potential developments and we cannot list them all here. Instead, we concentrate solely on the specific elements of research which we have already engaged in, or for which we have at least sketched up some ideas.

Regarding the research presented in chapter 3, our main development concerns:

- An expanded presentation of the concepts developed around models and modeling, including a much wider range of examples from various fields.

Regarding the research presented in chapter 4, we will concentrate on the following three elements:

- The identification of a specific algebraic characterization of MCI models defined by all itemsets of size 1 and 2 for any given m (i.e. an MCI alternative to the Chow-Liu tree model).

- An empirical study of the performances of the algebraic method for computing MCI models with pre-computed generic cases versus standard methods from numerical analysis, including an analysis of the variations when considering different constrained sets and different distributions.

- An empirical study of the performances of the algebraic method for computing MCI models without pre-computed generic cases versus standard methods from numerical analysis.

Regarding the research presented in chapter 5, we will concentrate on the following five elements:

- An evaluation of the ability of the greedy-up, greedy-down algorithm to correctly identify generating models with higher levels of complexity and a comparison of the results obtained with a wider range of pattern mining methods.
- The experimentation of the method on real data (with low values of m) and a comparison with other pattern mining methods on this aspect.
- The experimentation of the method on real data with higher values of m after reducing this value using feature selection and a comparison with other pattern mining methods on this particular aspect.
- The development of a clustering method adapted to the combined junction model as described in section 5.3.2.2, the experimentation of the corresponding search algorithm on data with large values of m and a comparison with other pattern mining methods on this specific aspect.

Finally, a number of elements have been coded and obtained through computation in this thesis. We are working on making some of these elements both freely available and intelligible, in the form of a Python 3 module including implementations for:

- The algorithms for computing MCI models based on the algebraic approach while performing computations in $\mathbb{Q}$.
- The algorithms for computing reduced algebraic expressions for generic MCI models while performing computations in $\mathbb{Q}[f_1, ..., f_d]$.
- An algorithm for computing any MCI model when $m \leq 4$ based on a database containing the corresponding pre-computed algebraic expressions.

- The iterative scaling algorithm for computing MCI models.
- The χ^2 test of mutual constrained independence.
- The greedy-up, greedy-down algorithm for discovering a valid MCI hypothesis.

6.2 The artificial scientist

During the twentieth century, the idea that the scientific method could not be logically formalized or, in other words, mathematically modeled became commonly accepted within the philosophy of science community [GRZ19]. This view was widely influenced by the positions defended by highly prominent philosophers such as Karl Popper and Hans Reichenbach. Indeed, both argued that, in the production of science, there is a clear distinction between the formulation of hypotheses and theories, on the one hand, and their evaluation, on the other hand, or, in Reichenbach's words, the context of discovery and the context of justification; and they both agreed that the initial creative aspect of science was an entirely subjective matter which could not be logically described [Rei38, Pop59].[1] In terms of artificial intelligence, this philosophical stance has important implications: if the process for discovering hypotheses cannot be mathematically modeled, then it cannot be implemented within the source code of an artificial intelligence, hence leading to the impossibility of conceiving an artificial scientist. In fact, the impossibility of developing a computer program that discovers hypotheses and theories is exactly the idea defended by another highly influential philosopher of science of the twentieth century, Carl Hempel, in *Thoughts on the Limitations of Discovery by Computer* [Hem85].

[1]This position can be easily summarized by Popper's own words in *The Logic of Scientific Discovery* [Pop59]:

> *I said above that the work of the scientist consists in putting forward and testing theories.*
> *The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it. The question how it happens that a new idea occurs to a man — whether it is a musical theme, a dramatic conflict, or a scientific theory — may be of great interest to empirical psychology; but it is irrelevant to the logical analysis of scientific knowledge.*

As we have moved well into the twentieth century, it is time to consider such positions outdated. The complexity of the task of elaborating a meaningful and objective approach towards automatic discovery in science should not lead us to discard the process as inherently subjective. In fact, throughout the entire course of this thesis, we have worked towards the goal of conceiving a meaningful and objective approach for producing scientific information.

In chapter 3, we focused on the modeling that underlies such an approach. In chapter 4, we studied some of the mathematical tools on which it relies. In chapter 5, we concentrated on its algorithmic aspects. Of course, the algorithms which we have presented in this last chapter represent but a very small step towards the definition of a general artificial scientist. For one part, the world as seen by our algorithms is extremely simple as it corresponds to no more than the empirical distribution of transactions on a limited number of items. And for another part, the scope of the potential the hypotheses that they can produce is also quite limited. Nevertheless, the results we have obtained through the implementation of our greedy-up, greedy-down algorithm (see sections 5.2.3.2 and 5.2.3.3) seem to show that, at least within this limited context, our approach performs very well.

In order to move forward towards the aim of developing an artificial scientist, the big challenge now is scaling up. Not so much in terms of the number of possible attributes considered, but in terms of the complexity of the representation of the observable world and the complexity of the hypotheses and theories that describe it. In [Hem85], Carl Hempel argued that a computer program may only discover hypotheses within the limits of the vocabulary of its language, while "*the formulation of powerful explanatory principles, and especially theories, normally involves the introduction of a novel conceptual and terminological apparatus*". Similarly, our study of the practical limitations on the number of attributes which can be considered in our approach (see section 3.7.2.2) and our discussion on the issue of pattern complexity (see section 3.6.2.3) both lead us to say that an artificial scientist cannot consider more than a limited number of attributes unless it is capable of complexifying its native vocabulary. However, this does not lead us to conclude on the impossibility of the artificial scientist: it simply tells us in which direction we must pursue our research.

CHAPTER 7

Extraction objective et signifiante de motifs intéressants sur la base de leur fréquence

Pour permettre son accès au plus grand nombre, cette thèse a été rédigée intégralement en anglais. Le texte qui suit présente un résumé succinct de son contenu en français. Lorsque la traduction de certains termes pourrait générer une ambiguïté, la terminologie anglosaxonne est maintenue. Ainsi, *pattern mining* est traduit ici par *extraction de motifs* mais le terme *itemset* est utilisé tel quel. Par souci de concision, la formulation précise des définitions, propriétés et algorithmes présentés dans la thèse n'est pas fournie ici. Elle n'est disponible, pour l'instant, qu'en version anglaise.

Problématique

Considérons un jeu de données binaires correspondant à la présence ou à l'absence d'un certain nombre d'attributs dans une population statistique. Supposons par ailleurs que les seules informations que l'on puisse obtenir en interrogeant ce jeu de données se rapportent à la fréquence dans les données de motifs correspondant à des conjonctions d'attributs. Pour utiliser la terminologie en extraction d'itemsets, considérons donc un jeu de données de transactions sur des items.

Comment pourrions-nous procéder pour extraire une information intéressante de ce jeu de données, en toute objectivité, et tout en explicitant clairement la signification du processus d'extraction ?

Contexte de recherche

Ce travail de thèse s'inscrit tout d'abord dans la continuité de travaux de recherche en fouille de données et en extraction de motifs et, plus particulièrement, de travaux sur l'extraction d'itemsets et de règles objectivement intéressants.

En effet, malgré l'engouement général qui a suivi la publication des travaux de Rakesh Agrawal et son équipe sur l'extraction d'itemsets fréquents [AIS93, AS94a], il a été rapidement établi que les itemsets fréquents (et les règles d'associations qui en découlaient) ne présentaient, en tant que tels, qu'un intérêt limité. Les chercheur·e·s se sont donc penché·e·s sur la question de l'intérêt de ces motifs et, plus particulièrement, de leur intérêt objectif. Parmi les travaux de recherche, on discerne différentes approches générales: les mesures cherchant à quantifier l'intérêt objectif d'un motif [TKS04, GH06, LMVL08, LBLL12a]; les représentations condensées exactes d'une classe de motifs [AIS93, PBTL99, CG02]; les méthodes identifiant l'intérêt objectif à l'étonnement statistique [HOV+09, LPP14] ou à l'informativité (au sens de la théorie de l'information) [MTV11, VVLS11, MVT12].

Toutefois, si la question de l'intérêt objectif des motifs extraits est centrale dans un nombre important de publications scientifiques dans ce domaine, la question de la signification du processus d'extraction a souvent été mise de côté comme si celle-ci était triviale voire secondaire. C'est pourtant une question fondamentale qu'il convient de poser car elle revient à expliquer en quoi on peut affirmer que les informations extraites des données par un tel processus présentent un intérêt objectif. Alors qu'un nombre croissant de voix s'élèvent aujourd'hui pour demander « l'ouverture de la boîte noire » et exiger une réelle transparence dans les prises de décisions assistées par des processus automatisés, il est essentiel que les chercheur·e·s qui développent ces méthodes soient pleinement capables d'en expliquer la signification [O'N16, RS17, VBB+18b].

C'est pourquoi, nous avons d'abord cherché à prendre un peu de recul pour analyser les mécanismes de modélisation mathématique qui permettent d'expliquer clairement la signification d'un processus d'extraction de motifs dits objectivement intéressants. Notre travail s'intègre ainsi également dans une recherche académique en philosophie des sciences et ceci à deux titres:

d'une part, sur la question de l'étude des modèles et des modélisations dans les processus scientifiques et technologiques [Min65, RWLN89, Sup07, Bok11, Pot17]; et d'autre part, sur la question de la création de connaissances objectives et, en particulier, sur la question de la formalisation et de l'automatisation de ce processus [Pop59, Hem85, GRZ19].

Modélisations mathématiques et signification dans les processus d'extraction de motifs

Nous soutenons l'idée que la signification d'un processus reposant sur des modèles mathématiques, ainsi que la signification de ce qui en aboutit en sortie, découle directement de la signification du processus de modélisation mathématique sur lequel il s'appuie. C'est pourquoi il est important de rendre explicite ce processus de modélisation et de comprendre la manière dont les choix en termes de modélisation impactent et déterminent la signification du processus d'extraction de motifs dans son ensemble.

En cherchant à caractériser les processus de modélisation mathématique dans le cadre spécifique de l'extraction de motifs objectivement intéressants, nous avons identifié des problématiques générales en modélisation mathématique qui dépassent le cadre de l'extraction de motifs. À ce titre, une partie de la recherche que nous présentons dans cette thèse peut être considérée comme une contribution en philosophie des sciences et nous définissons les contours d'un nouveau cadre pour l'étude et l'analyse qualitative des modélisations dans un processus scientifique. Nous présentons de nouvelles formulations pour définir et représenter les notions de modèle et de modélisation. En nous appuyons sur ces représentations, nous introduisons un certain nombre de concepts visant à caractériser les processus de modélisation, tels que: les notions de modélisation phénotypique et de modélisation génotypique; la notion de modélisation pragmatique; ou encore les notions de modélisation en patchwork et de modélisation holistique.

Nous utilisons ce cadre sur différents exemples directement issus du domaine de l'extraction de motifs afin d'expliciter comment les choix de modélisation (implicites ou explicites) induisent des différences fondamentales (et même parfois inconciliables) en termes de signification entre les différentes approches étudiées en extraction de motifs. Notre démarche est volontairement sceptique et nous cherchons à questionner tous les choix qui peuvent être faits en termes de modélisation mathématique dans un contexte d'extraction de motifs objectivement intéressants. C'est ainsi que nous interrogeons le choix de l'itemset en tant que motif fondamental dans un tel contexte et nous apportons en réponse un certain nombre d'éléments objectifs qui permettent de justifier la pertinence toute particulière de ce choix.

Par ailleurs, partant de l'idée que la recherche de connaissances objectives est une entreprise scientifique, nous examinons la possibilité d'intégrer une modélisation mathématique de la méthode scientifique, et plus particulièrement sa description via le modèle hypothético-déductif, dans la modélisation associée à un processus d'extraction de motifs. Nous établissons ainsi le lien entre notre problématique initiale et la question bien plus générale de la formalisation mathématique de la recherche scientifique en vue de son automatisation. Au premier abord, le contour particulier de notre problématique, partant d'un jeu de données statique et fini, semble peu adapté pour considérer le processus dynamique et sans fin qui est décrit par le modèle hypothético-déductif.

Toutefois, nous montrons que l'introduction de nouveaux outils de modélisation permet de contourner cette difficulté. Nous introduisons ainsi la notion de confiance en la distribution empirique, sur laquelle nous nous appuyons pour décrire l'extraction de motifs objectivement intéressants comme un processus dans lequel des hypothèses sont successivement formulées puis évaluées. Notons que l'une des particularités de notre approche est qu'elle va permettre à un processus automatisé d'extraction de motifs de répondre qu'il y a insuffisamment de données pour conclure, ce qui est totalement attendu dans un processus de recherche scientifique, mais généralement absent dans les processus d'extraction de motifs. Enfin, notre analyse sur la complexité de notre démarche nous amène à conclure que, en dehors de cadres très restreints, un système intelligent doit nécessairement être capable de complexifier le langage qu'il utilise pour décrire le monde si l'on veut qu'il en fasse une description scientifique et objective.

Nous proposons enfin, pour chacun des points d'analyse que nous considérons, un certain nombre de principes et de recommandations pour l'élaboration de modélisations mathématiques signifiantes dans le contexte spécifique de l'extraction de motifs objectivement intéressants sur la base de leur fréquence. Nous notons que les approches qui identifient l'extraction de motifs à de la compression de données [MTV11, MVT12, VVLS11] sont en concordance avec une majorité des recommandations que nous établissons. Une différence subsiste cependant sur la signification de ce qui est recherché. Là où ces méthodes cherchent à déterminer une compression optimale des données, nous cherchons plutôt à exhiber le mécanisme sous-jacent qui permet de générer les données. Si ces deux objectifs peuvent se rejoindre dans certains contextes, ce n'est pas nécessairement le cas a priori.

Indépendance contrainte mutuelle

En s'appuyant sur les principes que nous avons élaborés pour la modélisation mathématique des processus d'extraction de motifs objectivement intéressants, nous avons construit une notion d'indépendance contrainte mutuelle (abrégé MCI pour *Mutual Constrained Independence*). Cette notion permet de définir les différentes hypothèses objectives qui constituent les motifs objectivement intéressants (ou du moins potentiellement intéressants) dans notre approche : les modèles MCI.

Comme nous le démontrons les modèles MCI que nous définissons sont mathématiquement équivalents à des modèles de maximum d'entropie (MaxEnt models) [Jay03]. De tels modèles ont d'ailleurs déjà été considérés en tant que tels dans la littérature en extraction de motifs objectivement intéressants, particulièrement chez celles et ceux qui identifient l'extraction de motifs objectivement intéressants dans les données à la compression optimale de ces données [MTV11, MVT12]. Toutefois, l'approche MCI diffère significativement de l'approche entropique par sa construction. En effet, les modèles MCI sont construits comme la solution asymptotique d'un problème de recherche du modèle moyen d'un ensemble de modèles satisfaisant une contrainte particulière, ce qui permet d'entrevoir le principe de maximum d'entropie sous un angle nouveau.

Cette approche particulière de tels modèles de maximum d'entropie nous a permis d'exhiber de nouvelles propriétés de ces modèles ainsi que de nouvelles méthodes pour les calculer.

Nous présentons, dans un premier temps, le cas simple dans lequel le modèle MCI est défini par des contraintes sur l'ensemble des sous-ensembles propres d'un itemset. Nous déterminons une expression algébrique exacte pour les modèles MCI de cette classe et analysons les propriétés de la distance entre des données empiriques et le modèle correspondant.

En nous appuyant sur des algorithmes de géométrie algébrique réelle, nous généralisons cette approche au calcul de tout type de modèle MCI. Les différentes étapes de notre algorithme sont présentées en détails et nous montrons que celui-ci permet d'une part le calcul direct d'un modèle MCI défini par des données empiriques et d'autre part le calcul d'expressions algébriques réduites pour des classes de modèles MCI. L'utilisation de telles expressions algébriques calculées préalablement permet une réduction significative (d'un facteur 150 pour notre cas d'étude) du temps de calcul d'un modèle MCI, en comparaison avec la méthode de Darroch et Ratcliff [DR72] qui est une méthode de calcul numérique standard utilisée pour le calcul des modèles de maximum d'entropie équivalents en extraction de motifs [PMS03, TM10, MVT12]. Cet avantage doit être tempéré par l'impossibilité pratique de calculer les expressions algébriques réduites pour l'ensemble des classes possibles de modèles MCI. Toutefois, nous montrons que cette méthode peut être utilisée pour calculer des expressions algébriques réduites exactes pour tout type de modèle MCI défini sur un faible nombre d'items, ce qui lui permet tout de même de jouer un rôle important dans l'accélération de processus d'extraction de motifs objectivement intéressants.

Algorithmes d'extraction de motifs

En nous appuyant sur la recherche que nous avons menée sur les modèles MCI, nous montrons qu'il est possible de définir un test d'indépendance contrainte mutuelle. Nous utilisons alors ce test afin d'évaluer successivement des hypothèses d'indépendance contrainte mutuelle dans un algorithme d'extraction d'information scientifique sur les données. L'algorithme évalue d'abord si la taille de l'échantillon est suffisante afin d'avoir confiance en la distribution empirique puis, le cas échéant, évalue successivement les différentes hypothèses d'indépendance mutuelle contrainte par ordre de complexité croissante jusqu'à obtenir une hypothèse non réfutée qui constitue, selon le principe du rasoir d'Ockham, le meilleur modèle explicatif pour les données dans l'état des connaissances disponibles. Cet algorithme peut être assimilé à un processus d'extraction de motifs objectivement intéressants car l'hypothèse d'indépendance mutuelle contrainte retenue est définie par un ensemble d'itemsets. Il faut noter, que la définition d'un ordre sur la complexité d'une hypothèse d'indépendance mutuelle n'est pas nécessairement évidente et nous engageons donc une discussion sur différentes façons de définir un tel ordre.

Comme indiqué ci-dessus, l'algorithme a pour but la résolution d'un problème de recherche d'optimum: on cherche l'hypothèse la moins complexe qui n'est pas réfutée par un test. Or la taille de l'espace de recherche pour cette solution est doublement exponentiel en le nombre d'items et donc rapidement bien trop grand pour qu'une recherche exhaustive soit envisagée. Nous avons donc étudié différentes approches reposant sur une heuristique gloutonne pour remplacer la recherche exhaustive. En particulier, nous montrons la très grande efficacité de l'algorithme consistant à ajouter des contraintes sur des itemsets de manière gloutonne jusqu'à obtenir une hypothèse non rejetée, puis à en retrancher de manière gloutonne tant que l'hypothèse associée n'est pas rejetée.

De par son élaboration, l'algorithme que nous avons présenté constitue un exemple fonctionnel de l'automatisation du modèle hypothético-déductif de la méthode scientifique. Certes, le contexte dans lequel il peut s'appliquer reste extrêmement restreint, mais il faut noter que l'idée même que l'on puisse envisager d'automatiser ce processus, et en particulier la phase d'élaboration des hypothèses à tester, est elle-même extrêmement récente [GRZ19]. En ce sens, l'approche que nous avons développée peut être vue comme un point de départ pour des recherches futures sur la problématique de l'automatisation de la création de savoir scientifique.

Par ailleurs, les limites de l'étendue des applications pratiques de notre algorithme sont liées aux exigences extrêmement fortes que nous avons fixées sur la signification et l'objectivité du processus d'extraction. Nous discutons de ces limites et nous proposons un certain nombre d'adaptations qui permettent d'envisager l'utilisation des outils que nous avons développés dans des contextes moins restreints, au prix d'un compromis raisonnable sur ces exigences.

Bibliography

[AAP00] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*, pages 108–118, 2000.

[AAP01] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. *Journal of parallel and Distributed Computing*, 61(3):350–371, 2001.

[AAR09] Muhaimenul Adnan, Reda Alhajj, and Jon Rokne. Identifying social communities by frequent pattern mining. In *2009 13th International Conference Information Visualisation*, pages 413–418. IEEE, 2009.

[ABH14] Charu C. Aggarwal, Mansurul A. Bhuiyan, and Mohammad Al Hasan. *Frequent Pattern Mining Algorithms: A Survey*, pages 19–64. Springer International Publishing, 2014.

[ACGH18] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[ACM] Fathers of the deep learning revolution receive 2018 ACM A.M. Turing award. Bengio, Hinton and LeCun ushered in major breakthroughs in artificial intelligence.

[AD75] János Aczél and Zoltán Daróczy. On measures of information and their characterizations. *New York*, 122, 1975.

[AFN74] János Aczél, Bruno Forte, and Che Tat Ng. Why the Shannon and Hartley entropies are 'natural'. *Advances in applied probability*, 6(1):131–146, 1974.

[Agg14] Charu C. Aggarwal. *An Introduction to Frequent Pattern Mining*, pages 1–17. Springer International Publishing, 2014.

[Agg15] Charu C. Aggarwal. *Data mining: the textbook*. Springer, 2015.

[Agg17] Charu C. Aggarwal. *Outlier Analysis - Second Edition*. Springer, 2017.

[AH17] Daichi Amagata and Takahiro Hara. Mining top-k co-occurrence patterns across multiple streams. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2249–2262, 2017.

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, 1993.

[AISK14] David C. Anastasiu, Jeremy Iverson, Shaden Smith, and George Karypis. Big data frequent pattern mining. In *Frequent pattern mining*, pages 225–259. Springer, 2014.

[ALWW09] Charu C. Aggarwal, Yan Li, Jianyong Wang, and Jing Wang. Frequent pattern mining with uncertain data. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38, 2009.

[ALZ14] Luiza Antonie, Jundong Li, and Osmar Zaiane. *Negative Association Rules*, pages 135–145. Springer International Publishing, 2014.

[AMS+96] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A. Inkeri Verkamo, et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1):307–328, 1996.

[And08] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[Apo61] Leo Apostel. Towards the formal study of models in the non-formal sciences. In *The concept and the role of the model in mathematics and natural and social sciences*, pages 1–37. Springer, 1961.

[APVZ14] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International conference on machine learning*, pages 1908–1916, 2014.

[AS94a] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[AS94b] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. *IBM RJ*, 9839, June 1994.

[AT01] Gediminas Adomavicius and Alexander Tuzhilin. Expert-driven validation of rule-based user models in personalization applications. *Data Mining and Knowledge Discovery*, 5(1-2):33–58, 2001.

[AWF07] Mehdi Adda, Lei Wu, and Yi Feng. Rare itemset mining. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 73–80. IEEE, 2007.

[AY98] Charu C. Aggarwal and Philip S. Yu. A new framework for itemset generation. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '98)*, pages 18–24. ACM, 1998.

[AZ12] Charu C. Aggarwal and Cheng X. Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[AZLS18] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[B+01] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[BB02] Christian Borgelt and Michael R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 51–58. IEEE, 2002.

[BBR03] Jean-François Boulicaut, Artur Bykowski, and Christophe Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7(1):5–22, 2003.

[BC62] I. Bernard Cohen. The first English version of Newton's hypotheses non fingo. *Isis*, 53(3):379–388, 1962.

[BCG01] Douglas Burdick, Manuel Calimlim, and Johannes Gehrke. Mafia: A maximal frequent itemset algorithm for transactional databases. In *Proceedings of the 17th International Conference on Data Engineering*, volume 1, pages 443–452, 2001.

[BCR13] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

[Beh18] Eric Beh. Exploring how to simply approximate the p-value of a chi-squared statistic. *Austrian Journal of Statistics*, 47(3):63–75, 2018.

[BEX02] Florian Beil, Martin Ester, and Xiaowei Xu. Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 436–442, 2002.

[BGK09] Julien Blanchard, Fabrice Guillet, and Pascale Kuntz. Semantics-based classification of rule interestingness measures. In *Post-mining of association rules: techniques for effective knowledge extraction*, pages 56–79. IGI Global, 2009.

[BGMP03] Francesco Bonchi, Fosca Giannotti, Alessio Mazzanti, and Dino Pedreschi. Exante: Anticipated data reduction in constrained pattern mining. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 59–70. Springer, 2003.

[BHL19] Peter L. Bartlett, David P. Helmbold, and Philip M. Long. Gradient descent with identity initialization efficiently learns positive-definite linear transformations by deep residual networks. *Neural computation*, 31(3):477–502, 2019.

[BJ98] Roberto J. Bayardo Jr. Efficiently mining long patterns from databases. In *ACM Sigmod Record*, volume 27, pages 85–93. ACM, 1998.

[BJ99] Daniela M. Bailer-Jones. Tracing the development of models in the philosophy of science. In *Model-based reasoning in scientific discovery*, pages 23–40. Springer, 1999.

[BJ02] Daniela M. Bailer-Jones. Scientists' thoughts on scientific models. *Perspectives on Science*, 10(3):275–301, 2002.

[BJ09] Daniela M. Bailer-Jones. *Scientific models in philosophy of science*. University of Pittsburgh Pre, 2009.

[BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD '97)*, pages 265–276. ACM, 1997.

[Bok11] Alisa Bokulich. How scientific models can explain. *Synthese*, 180(1):33–45, 2011.

[Bor11] Monica Borda. *Fundamentals in information theory and coding*. Springer Science & Business Media, 2011.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. 1998.

[BP14] Aniello Buonocore and Enrica Pirozzi. On the pearson-fisher chi-squared theorem. *Applied Mathematical Sciences*, 8(134):6733–6744, 2014.

[BPR06] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry.* Springer, 2006.

[Bro08] J.R. Brown. *Philosophy of Mathematics: A Contemporary Introduction to the World of Proofs and Pictures.* Philosophical issues in science. Routledge, 2008.

[Bru80] Richard A. Brualdi. Matrices of zeros and ones with fixed row and column sum vectors. *Linear algebra and its applications*, 33:159–231, 1980.

[BTP+00] Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme, and Lotfi Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2:66–75, 2000.

[BU11] David Buchfuhrer and Christopher Umans. The complexity of boolean formula minimization. *Journal of Computer and System Sciences*, 77(1):142–153, 2011.

[C+99] Nancy Cartwright et al. *The dappled world: A study of the boundaries of science.* Cambridge University Press, 1999.

[Cal04] Toon Calders. Deducing bounds on the support of itemsets. In *Database Support for Data Mining Applications*, pages 214–233. Springer, 2004.

[Cal13] Cristian S. Calude. *Information and randomness: an algorithmic perspective.* Springer Science & Business Media, 2013.

[CC89] Jean-Pierre Changeux and Alain Connes. *Matière à pensée.* Editions Odile Jacob, 1989. English translation under the title Conversations on Mind, Matter and Mathematics, Princeton University Press, 1995, https://books.google.fr/books?id=hd89DwAAQBAJ.

[CDHL05] Yuguo Chen, Persi Diaconis, Susan P. Holmes, and Jun S. Liu. Sequential monte carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–120, 2005.

[Cel15] Carlo Cellucci. Mathematical beauty, understanding, and discovery. *Foundations of Science*, 20(4):339–355, 2015.

[Cen87] Jadzia Cendrowska. PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27(4):349–370, 1987.

[CG02] Toon Calders and Bart Goethals. Mining all non-derivable frequent itemsets. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 74–86. Springer, 2002.

[CG07] Toon Calders and Bart Goethals. Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, 14(1):171–206, 2007.

[CGM14] Kate Crawford, Mary L. Gray, and Kate Miltner. Big data— critiquing big data: Politics, ethics, epistemology— special section introduction. *International Journal of Communication*, 8:10, 2014.

[Cha69] Gregory J. Chaitin. On the simplicity and speed of programs for computing infinite sets of natural numbers. *Journal of the ACM (JACM)*, 16(3):407–422, 1969.

[Cha13] A.F. Chalmers. *What Is This Thing Called Science?* Hackett Publishing Company, Incorporated, 2013.

[CJRR21] Patrick Cheridito, Arnulf Jentzen, Adrian Riekert, and Florian Rossmannek. A proof of convergence for gradient descent in the training of artificial neural networks for constant target functions. *arXiv preprint arXiv:2102.09924*, 2021.

[CK11] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

[CL68] C. Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[CM60] T. W. Chaundy and J. B. McLeod. On a functional equation. *Edinburgh Mathematical Notes*, 43:7–8, 1960.

[CM84] Nancy Cartwright and Ernan McMullin. How the laws of physics lie, 1984.

[CN89] Peter Clark and Tim Niblett. The CN2 induction algorithm. *Machine learning*, 3(4):261–283, 1989.

[Con10] Gabriele Contessa. Scientific models and fictional objects. *Synthese*, 172(2):215, 2010.

[Cox01] David R. Cox. [statistical modeling: The two cultures]: Comment. *Statistical Science*, 16(3):216–218, 2001.

[CS05] Yuguo Chen and Dylan Small. Exact tests for the rasch model via sequential importance sampling. *Psychometrika*, 70(1):11–30, 2005.

[Csi76] Imre Csiszár. Review of "on measures of information and their characterizationś" by (Aczél, J., and Daróczy, Z.; 1975). *IEEE Transactions on Information Theory*, 22(6):765–766, 1976.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[CTH+13] Pei-Shan Chien, Yu-Fang Tseng, Yao-Chin Hsu, Yu-Kai Lai, and Shih-Feng Weng. Frequency and pattern of chinese herbal medicine prescriptions for urticaria in taiwan during 2009: analysis of the national health insurance database. *BMC complementary and alternative medicine*, 13(1):209, 2013.

[CTTX05] Gao Cong, Kian-Lee Tan, Anthony K. H. Tung, and Xin Xu. Mining top-k covering rule groups for gene expression data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 670–681, 2005.

[DB11] Tijl De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, 23(3):407–446, 2011.

[DBFVL18] Youcef Djenouri, Asma Belhadi, Philippe Fournier-Viger, and Jerry Chun-Wei Lin. Fast and effective cluster-based information retrieval using frequent closed itemsets. *Information Sciences*, 453:154–167, 2018.

[DBLL15] Thomas Delacroix, Ahcène Boubekki, Philippe Lenca, and Stéphane Lallich. Constrained independence for detecting interesting patterns. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2015.

[DCF90] Newton C. A. Da Costa and Steven French. The model-theoretic approach in the philosophy of science. *Philosophy of Science*, 57(2):248–265, 1990.

[DCF03] Newton C. A. Da Costa and Steven French. *Science and partial truth: A unitary approach to models and scientific reasoning.* Oxford University Press on Demand, 2003.

[Del18] Thomas Delacroix. Choisir un encodage cnf de contraintes de cardinalité performant pour sat. In *Conférence Nationale d'Intelligence Artificielle Année 2018*, page 61, 2018.

[Dev00] Keith Devlin. *Do Mathematicians have different Brains?*, chapter 5. Basic Books New York, 2000.

[DHP06] Didier Dubois, Eyke Hüllermeier, and Henri Prade. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2):167–192, 2006.

[DKL14] Cristina David, Daniel Kroening, and Matt Lewis. Second-order propositional satisfiability. *arXiv preprint arXiv:1409.4925*, 2014.

[DKWK05] Mukund Deshpande, Michihiro Kuramochi, Nikil Wale, and George Karypis. Frequent substructure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering*, 17(8):1036–1050, 2005.

[DLL17] Thomas Delacroix, Philippe Lenca, and Stéphane Lallich. Du local au global: un nouveau défi pour l'analyse statistique implicative. In *ASI 2017 : 9e Colloque International d'Analyse Statistique Implicative*, pages 103–116, 2017.

[DMSV03] Miguel Delgado, Nicolás Marín, Daniel Sánchez, and M-A Vila. Fuzzy association rules: general model and applications. *IEEE transactions on Fuzzy Systems*, 11(2):214–225, 2003.

[DR72] John N Darroch and Douglas Ratcliff. Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, pages 1470–1480, 1972.

[Efr01] Brad Efron. [statistical modeling: The two cultures]: Comment. *Statistical Science*, 16(3):218–219, 2001.

[EGB+17] Elias Egho, Dominique Gay, Marc Boullé, Nicolas Voisine, and Fabrice Clérot. A user parameter-free approach for mining robust sequential classification rules. *Knowledge and Information Systems*, 52(1):53–81, 2017.

[EU216] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), 2016.

[EV18] Lilian Edwards and Michael Veale. Enslaving the algorithm: From a "right to an explanatio" to a "right to better decisions"? *IEEE Security & Privacy*, 16(3):46–54, 2018.

[Fad56] Dmitrii Konstantinovich Faddeev. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Matematicheskikh Nauk*, 11(1):227–231, 1956.

[FAT+14] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279, 2014.

[FC13] Zahra Farzanyar and Nick Cercone. Efficient mining of frequent itemsets in social network data based on mapreduce framework. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1183–1188, 2013.

[FIM03] FIMI. Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations (FIMI '03). Melbourne, Florida, USA. volume 90, 2003.

[FIM04] FIMI. Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations (FIMI '04). Brighton, United Kingdom. volume 126, 2004.

[Fis22] Ronald A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[Fis24] Ronald A. Fisher. The conditions under which χ2 measures the discrepancey between observation and hypothesis. *Journal of the Royal Statistical Society*, 87(3):442–450, 1924.

[For75] B. Forte. Why Shannon's entropy. In *Symposia Mathematica*, volume 15, pages 137–152. Academic Press New York, 1975.

[FP88] Alan Ford and F. David Peat. The role of language in science. *Foundations of physics*, 18(12):1233–1242, 1988.

[G+09] Ronald N. Giere et al. *Why scientific models should not be regarded as works of fiction.* na, 2009.

[Gal23] Galileo Galilei. *Il saggiatore.* 1623. in English: The Assayer, Translation from the Italian by Stillman Drake in The Controversy on the Comets of

1618. Philadelphia: University of Pennsylvania Press. https://www.degruyter.com/view/product/483880.

[GBMTCO10] Milton García-Borroto, José Francisco Martínez-Trinidad, and Jesús Ariel Carrasco-Ochoa. A new emerging pattern mining algorithm and its application in supervised classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 150–157. Springer, 2010.

[GBPP98] R. Gras, H. Briand, P. Peter, and J. Philippe. Implicative statistical analysis. In Chikio Hayashi, Keiji Yajima, Hans-Hermann Bock, Noboru Ohsumi, Yutaka Tanaka, and Yasumasa Baba, editors, *Data Science, Classification, and Related Methods*, pages 412–419. Springer Japan, 1998.

[GCB+04] Régis Gras, Raphaël Couturier, Julien Blanchard, Henri Briand, Pascale Kuntz, and Philippe Peter. Quelques critères pour une mesure de qualité de règles d'association. *Revue des nouvelles technologies de l'information RNTI E-1*, pages 3–30, 2004.

[GH06] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3), 2006.

[Gie88] Ronald N. Giere. Explaining science: a cognitive approach. university of chicago press. *Chicago I. L. Gould P.*, pages 139–51, 1988.

[Gie99] Ronald N. Giere. Using models to represent reality. In *Model-based reasoning in scientific discovery*, pages 41–57. Springer, 1999.

[GL92] Régis Gras and Annie Larher. L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématiques et sciences humaines*, 120:5–31, 1992.

[GLCZ17] Wensheng Gan, Jerry Chun-Wei Lin, Han-Chieh Chao, and Justin Zhan. Data mining in distributed environment: a survey.

Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6):e1216, 2017.

[GLFV+19] Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S Yu. A survey of parallel sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(3):1–34, 2019.

[GMMT07] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(3):14–es, 2007.

[GMP05] Peter D. Grünwald, In Jae Myung, and Mark A. Pitt. *Advances in minimum description length: Theory and applications*. MIT press, 2005.

[Goe03] Bart Goethals. Survey on frequent pattern mining. *Univ. of Helsinki*, 19:840–852, 2003.

[GORS16] Bernhard Ganter, Sergei Obiedkov, Sebastian Rudolph, and Gerd Stumme. *Conceptual exploration*. Springer, 2016.

[Got16] Don Gotterbarn. The creation of facts in the cloud: a fiction in the making. *ACM SIGCAS Computers and Society*, 45(3):60–67, 2016.

[GRMG13] Régis Gras, Jean-Claude Régnier, Claudia Marinica, and Fabrice Guillet. *L'analyse statistique implicative Méthode exploratoire et confirmatoire à la recherche de causalités*. 2013.

[GRZ19] Clark Glymour, Joseph D. Ramsey, and Kun Zhang. The evaluation of discovery: Models, simulation and search through "big data". *Open Philosophy*, 2(1):39–48, 2019.

[GS06] Peter Godfrey-Smith. The strategy of model-based science. *Biology and philosophy*, 21(5):725–740, 2006.

[GSS12] Salvatore Greco, Roman Słowiński, and Izabela Szczech. Properties of rule interestingness measures and alternative approaches to normalization of measures. *Information Sciences*, 216:1–16, 2012.

[GSW05] Bernhard Ganter, Gerd Stumme, and Rudolf Wille. *Formal concept analysis: foundations and applications*, volume 3626. springer, 2005.

[Gun17] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017.

[GV01] A.L. George and D. Velleman. *Philosophies of Mathematics*. Wiley, 2001.

[GW12] Bernhard Ganter and Rudolf Wille. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[GZ05] Karam Gouda and Mohammed J. Zaki. Genmax: An efficient algorithm for mining maximal frequent itemsets. *Data Mining and Knowledge Discovery*, 11(3):223–242, 2005.

[Hah15] Michael Hahsler. A probabilistic comparison of commonly used interest measures for association rules. Research report, 2015.

[Háj01] Petr Hájek. The GUHA method and mining association rules. In *Proceedings CIMA'2001 International ICSC Congress on Computational Intelligence: Methods and Applications June 19 - 22, 2001 (Bangor, Wales, United Kingdom)*, pages 533–539. ICSC, 2001.

[Han11] Sami Hanhijärvi. Multiple hypothesis testing in pattern discovery. In *International Conference on Discovery Science*, pages 122–134. Springer, 2011.

[Har40] Godfrey Harold Hardy. *A mathematician's apology*. Cambridge University Press, 1940. Reprinted with foreword by C. P. Snow in 1967, https://books.google.fr/books?id=beImvXUGD-MC.

[HCXY07] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, 15(1):55–86, 2007.

[Hem85] Carl G Hempel. Thoughts on the limitations of discovery by computer. *Logic of discovery and diagnosis in medicine*, pages 115–122, 1985.

[Hes65] Mary B. Hesse. Models and analogies in science. 1965.

[HH12] Petr Hájek and Tomáš Havránek. *Mechanizing hypothesis formation: Mathematical foundations for a general theory.* Springer Science & Business Media, 2012.

[HHC66] Petr Hájek, Ivan Havel, and Michal Chytil. The GUHA method of automatic hypotheses determination. *Computing*, 1(4):293–308, 1966.

[HHR10] Petr Hájek, Martin Holeňa, and Jan Rauch. The GUHA method and its meaning for data mining. *Journal of Computer and System Sciences*, 76(1):34–48, 2010. Special Issue on Intelligent Data Analysis.

[HK08] Gabor T. Herman and Attila Kuba. *Advances in discrete tomography and its applications.* Springer Science & Business Media, 2008.

[HK12] Gabor T. Herman and Attila Kuba. *Discrete tomography: Foundations, algorithms, and applications.* Springer Science & Business Media, 2012.

[HKMT95] Marcel Holsheimer, Martin L. Kersten, Heikki Mannila, and Hannu Toivonen. A perspective on databases and data mining. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD 95)*, volume 95, pages 150–155, 1995.

[Hol98] Martin Holeňa. Fuzzy hypotheses for GUHA implications. *Fuzzy Sets and Systems*, 98(1):101–125, 1998.

[HOV+09] Sami Hanhijärvi, Markus Ojala, Niko Vuokko, Kai Puolamäki, Nikolaj Tatti, and Heikki Mannila. Tell me something I don't know: randomization strategies for iterative data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 379–388. ACM, 2009.

[HP00] Jiawei Han and Jian Pei. Mining frequent patterns by pattern-growth: Methodology and implications. *SIGKDD Explor. Newsl.*, 2(2):14–20, 2000.

[HP14] Jiawei Han and Jian Pei. *Pattern-Growth Methods*, pages 65–81. Springer International Publishing, 2014.

[HPK11] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[HS06] Gary D. Hachtel and Fabio Somenzi. *Logic synthesis and verification algorithms*. Springer Science & Business Media, 2006.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.

[HXH+20] Kexin Huang, Cao Xiao, Trong Hoang, Lucas Glass, and Jimeng Sun. Caster: Predicting drug interactions with chemical substructure representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 702–709, 2020.

[IV06] Renáta Iváncsy and István Vajk. Frequent pattern mining in web log data. *Acta Polytechnica Hungarica*, 3(1):77–90, 2006.

[JA07] Ruoming Jin and Gagan Agrawal. Frequent pattern mining in data streams. In *Data Streams*, pages 61–84. Springer, 2007.

[Jac01] D. Jacquette. *Philosophy of Logic: An Anthology*. Blackwell Philosophy Anthologies. Wiley, 2001.

[Jay82] Edwin T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[Jay03] Edwin T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[K+12] Ashok Kumar et al. Web log mining using k-apriori algorithm. *International Journal of Computer Applications*, 41(11), 2012.

[KDB10] Kleanthis-Nikolaos Kontonasios and Tijl De Bie. An information-theoretic approach to finding informative noisy tiles in binary databases. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 153–164. SIAM, 2010.

[Kel84] Hendrikus Kelderman. Loglinear rasch model tests. *Psychometrika*, 49(2):223–245, 1984.

[KG02a] Marzena Kryszkiewicz and Marcin Gajek. Concise representation of frequent patterns based on generalized disjunction-free generators. In Ming-Syan Chen, Philip S. Yu, and Bing Liu, editors, *Pacific-Asia Conference on Knowledge Discovery and Data Mining, Advances in Knowledge Discovery and Data Mining. PAKDD 2002*, pages 159–171. Springer, 2002.

[KG02b] Marzena Kryszkiewicz and Marcin Gajek. Why to apply generalized disjunction-free generators representation of frequent patterns? In Mohand-Saïd Hacid, Zbigniew W. Raś, Djamel A. Zighed, and Yves Kodratoff, editors, *International Symposium on Methodologies for Intelligent Systems, Foundations of Intelligent Systems. ISMIS 2002*, pages 383–392. Springer, 2002.

[Kha11] Hnin Wint Khaing. Data mining based fragmentation and prediction of medical data. In *2011 3rd International Conference on Computer Research and Development*, volume 2, pages 480–485. IEEE, 2011.

[KIA+17] Junpei Komiyama, Masakazu Ishihata, Hiroki Arimura, Takashi Nishibayashi, and Shin-ichi Minato. Statistical emerging pattern mining with multiple testing correction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 897–906, 2017.

[KK15] Lev V. Kalmykov and Vyacheslav L. Kalmykov. A white-box model of s-shaped and double s-shaped single-species population growth. *PeerJ*, 3:e948, 2015.

[KKB+17] Seock-Ho Kim, Minho Kwak, Meina Bian, Zachary Feldberg, Travis Henry, Juyeon Lee, Ibrahim Burak Olmez, Yawei Shen, Yanyan Tan, Victoria Tanaka, et al. A taxonomy of item response models in psychometrika. In *The Annual Meeting of the Psychometric Society*, pages 13–23. Springer, 2017.

[Kol63] Andrei N. Kolmogorov. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 369–376, 1963.

[Kör60] S. Körner. *The Philosophy of Mathematics: An Introductory Essay*. Harper torchbooks: Science library. Hutchinson, 1960.

[KR05] Yun Sing Koh and Nathan Rountree. Finding sporadic rules using apriori-inverse. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 97–106. Springer, 2005.

[KS10] Edith Kovács and Tamás Szántai. On the approximation of a discrete multivariate probability distribution using the new concept of t-cherry junction tree. In Kurt Marti, Yuri Ermoliev, and Marek Makowski, editors, *Coping with Uncertainty: Robust Solutions*, pages 39–56. Springer, 2010.

[Kul59] Solomon Kullback. *Information theory and statistics*. John Wiley & Sons, 1959.

[Lal02] Stéphane Lallich. Mesure et validation en extraction des connaissances à partir des données. *Habilitationa Diriger des Recherches–Université Lyon*, 2, 2002.

[LB11] Yannick Le Bras. *Contribution à l'étude des mesures de l'intérêt des règles d'association et à leurs propriétés algorihmiques*. PhD thesis, 2011.

[LBLL09] Yannick Le Bras, Philippe Lenca, and Stéphane Lallich. On optimal rule mining: A framework and a necessary and sufficient condition of antimonotonicity. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 705–712. Springer, 2009.

[LBLL10] Yannick Le Bras, Philippe Lenca, and Stéphane Lallich. Mining interesting rules without support requirement: a general universal existential upward closure property. In *Data mining*, pages 75–98. Springer, 2010.

[LBLL12a] Yannick Le Bras, Philippe Lenca, and Stéphane Lallich. Formal framework for the study of algorithmic properties of objective interestingness measures. In *Data Mining: Foundations and Intelligent Paradigms*, pages 77–98. Springer, 2012.

[LBLL12b] Yannick Le Bras, Philippe Lenca, and Stéphane Lallich. Optimonotone measures for optimal rule discovery. *Computational Intelligence*, 28(4):475–504, 2012.

[LBLML09] Yannick Le Bras, Philippe Lenca, Sorin Moga, and Stéphane Lallich. All-monotony: A generalization of the all-confidence antimonotony. In *2009 International Conference on Machine Learning and Applications*, pages 759–764. IEEE, 2009.

[LBMLL10a] Yannick Le Bras, Patrick Meyer, Philippe Lenca, and Stéphane Lallich. Mesure de la robustesse de règles d'association. In *QDC 2010: 6th Workshop on Qualité des Données et des Connaissances (in conjunction with the 10th Workshop Extraction et gestion des connaissances)*, pages 27–38, 2010.

[LBMLL10b] Yannick Le Bras, Patrick Meyer, Philippe Lenca, and Stéphane Lallich. A robustness measure of association rules. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 227–242. Springer, 2010.

[LGFV+16] Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-Pei Hong, and Vincent S. Tseng. Efficient algorithms for mining high-utility itemsets in uncertain databases. *Knowledge-Based Systems*, 96:171–187, 2016.

[LGR81a] I. C. Lerman, R. Gras, and H. Rostam. Élaboration et évaluation d'un indice d'implication pour des données binaires. i. *Mathématiques et Sciences humaines*, 74:5–35, 1981.

[LGR81b] I.C . Lerman, R. Gras, and H. Rostam. Élaboration et évaluation d'un indice d'implication pour des données binaires. ii. *Mathématiques et sciences humaines*, 75:5–47, 1981.

[LHC97] Bing Liu, Wynne Hsu, and Shu Chen. Using general impressions to analyze discovered classification rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97). Newport Beach, CA. 31–36.*, pages 31–36. AAAI, 1997.

[LHML99] Bing Liu, Wynne Hsu, Lai-Fun Mun, and Hing-Yan Lee. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):817–832, 1999.

[LJA14] Victor E. Lee, Ruoming Jin, and Gagan Agrawal. Frequent pattern mining in data streams. In *Frequent Pattern Mining*, pages 199–224. Springer, 2014.

[LK98] Dao-I Lin and Zvi M. Kedem. Pincer-search: A new algorithm for discovering the maximum frequent set. In *International conference on Extending database technology*, pages 103–119. Springer, 1998.

[LLLW16] Jianzheng Liu, Jie Li, Weifeng Li, and Jiansheng Wu. Rethinking big data: A review on the data quality and usage issues. *ISPRS journal of photogrammetry and remote sensing*, 115:134–142, 2016.

[LMP+03] Philippe Lenca, Patrick Meyer, Philippe Picouet, Benoît Vaillant, and Stéphane Lallich. Critères d'évaluation des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information*, RNTI-1:123–134, 2003.

[LMT14] Carson Kai-Sang Leung, Richard Kyle MacKinnon, and Syed K. Tanbeer. Fast algorithms for frequent itemset mining from uncertain data. In *2014 IEEE International Conference on Data Mining*, pages 893–898. IEEE, 2014.

[LMV+04] Philippe Lenca, Patrick Meyer, Benoît Vaillant, Philippe Picouet, and Stéphane Lallich. Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données)*, pages 219–246, 2004.

[LMVL08] Philippe Lenca, Patrick Meyer, Benoît Vaillant, and Stéphane Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European journal of operational research*, 184(2):610–626, 2008.

[LNHP99] Laks V. S. Lakshmanan, Raymond Ng, Jiawei Han, and Alex Pang. Optimization of constrained frequent set queries with 2-variable constraints. In *Proceedings of the 1999 ACM-SIGMOD International Conference on the Management of Data*, volume 28, pages 157–168. ACM, 1999.

[LPP14] Jefrey Lijffijt, Panagiotis Papapetrou, and Kai Puolamäki. A statistical significance testing approach to mining the most informative set of patterns. *Data Mining and Knowledge Discovery*, 28(1):238–263, 2014.

[LT04] Stéphane Lallich and Olivier Teytaud. Évaluation et validation de l'intérêt des regles d'association. *Revue des nouvelles Technologies de l'Information*, 1(2):193–218, 2004.

[LTP06] Stéphane Lallich, Olivier Teytaud, and Elie Prudhomme. Statistical inference and data mining: false discoveries control. In *Compstat 2006-Proceedings in Computational Statistics*, pages 325–336. Springer, 2006.

[LV08] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 3rd edition, 2008.

[LWC82] Imre Lakatos, John Worrall, and Gregory Currie. Philosophical papers. volume i: The methodology of scientific research programmes; volume ii: Mathematics, science and epistemology. *Tijdschrift Voor Filosofie*, (4):744–745, 1982.

[LY17] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in neural information processing systems*, pages 597–607, 2017.

[Mac03] David J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[Mag12] Lorenzo Magnani. Scientific models are not fictions. In *Philosophy and cognitive science*, pages 1–38. Springer, 2012.

[Man13] Yuri I. Manin. Kolmogorov complexity as a hidden factor of scientific discourse: from newton's law to data mining. *arXiv preprint arXiv:1301.0081*, 2013.

[Mas82] Geoff N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

[Maz15] Fulvio Mazzocchi. Could big data be the end of theory in science? *EMBO reports*, 16(10):1250–1255, 2015.

[McG05] Ken McGarry. A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(1):39–61, 2005.

[McM85] Ernan McMullin. Galilean idealization. *Studies in History and Philosophy of Science Part A*, 16(3):247–273, 1985.

[Meo00] Rosa Meo. Theory of dependence values. *ACM Transactions on Database Systems (TODS)*, 25(3):380–406, 2000.

[Min65] Marvin Minsky. Matter, mind and models. In *Proc. International Federation of Information Processing Congress*, volume 1, pages 45–49, 1965.

[MJM+17] Seyed Ahmad Moosavi, Mehrdad Jalali, Negin Misaghian, Shahaboddin Shamshirband, and Mohammad Hossein Anisi. Community detection in social networks using user frequent pattern mining. *Knowledge and Information Systems*, 51(1):159–186, 2017.

[MM95] John A. Major and John J. Mangano. Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems*, 4(1):39–52, 1995.

[MMB+18] Aida Mrzic, Pieter Meysman, Wout Bittremieux, Pieter Moris, Boris Cule, Bart Goethals, and Kris Laukens. Grasping frequent subgraph mining for bioinformatics applications. *BioData mining*, 11(1):20, 2018.

[MO04] Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM transactions on computational biology and bioinformatics*, 1(1):24–45, 2004.

[Mou76] Georges Mounin. *Linguistique et traduction*, volume 60. Dessart et Mardaga, 1976.

[MSG14] Sajid Mahmood, Muhammad Shahbaz, and Aziz Guergachi. Negative and positive association rules mining from text using frequent and infrequent itemsets. *The Scientific World Journal*, 2014.

[MT96] Heikki Mannila and Hannu Toivonen. Multiple uses of frequent sets and condensed representations (extended abstract). In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 189–194. AAAI Press, 1996.

[MT12] Christoph Meinel and Thorsten Theobald. *Algorithms and Data Structures in VLSI Design: OBDD-foundations and applications*. Springer Science & Business Media, 2012.

[MTV11] Michael Mampaey, Nikolaj Tatti, and Jilles Vreeken. Tell me what i need to know: succinctly summarizing data with itemsets. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 573–581, 2011.

[MV13] Michael Mampaey and Jilles Vreeken. Summarizing categorical data by clustering attributes. *Data Mining and Knowledge Discovery*, 26(1):130–173, 2013.

[MVT12] Michael Mampaey, Jilles Vreeken, and Nikolaj Tatti. Summarizing data succinctly with the most informative itemsets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):16, 2012.

[Nat08] National Plant Data Center. The plants database, 2008.

[NB12] Victoria Nebot and Rafael Berlanga. Finding association rules in semantic web data. *Knowledge-Based Systems*, 25(1):51–62, 2012.

[NCC+12] Puteri N. E. Nohuddin, Frans Coenen, Rob Christley, Christian Setzkorn, Yogesh Patel, and Shane Williams. Finding "interesting" trends in social networks using frequent pattern mining and self organizing maps. *Knowledge-Based Systems*, 29:104–113, 2012.

[New13] Isaac Newton. *Philosophiae naturalis principia mathematica. Editio secunda auctior et emendatior.* Cornelius Crownfield, Cambridge, 1713.

[NLHP98] Raymond T Ng, Laks V. S. Lakshmanan, Jiawei Han, and Alex Pang. Exploratory mining and pruning optimizations of constrained associations rules. In *Proceedings of the 1998 ACM-SIG-MOD International Conference on the Management of Data*, volume 27, pages 13–24. ACM, 1998.

[Nov63] Albert B. Novikoff. On convergence proofs for perceptrons. Technical report, Stanford Research Inst. Menlo Park CA, 1963.

[NVHT13] Loan T. T. Nguyen, Bay Vo, Tzung-Pei Hong, and Hoang Chi Thanh. Car-miner: An efficient algorithm for mining class-association rules. *Expert Systems with Applications*, 40(6):2305–2311, 2013.

[NZ14] Siegfried Nijssen and Albrecht Zimmermann. *Constraint-Based Pattern Mining*, pages 147–163. Springer International Publishing, 2014.

[O'H04] Kay O'Halloran. *Mathematical discourse: Language, symbolism and visual images*. Bloomsbury Publishing, 2004.

[OKO+04] Miho Ohsaki, Shinya Kitaguchi, Kazuya Okamoto, Hideto Yokoi, and Takahira Yamaguchi. Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *European Conference on Principles of Data Mining and Knowledge Discovery in Databases: PKDD 2004*, pages 362–373. Springer, 2004.

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

[O'N16] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books Limited, 2016.

[PBTL99] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *International Conference on Database Theory*, pages 398–416. Springer, 1999.

[PCT+03] Feng Pan, Gao Cong, Anthony K. H. Tung, Jiong Yang, and Mohammed J. Zaki. Carpenter: Finding closed patterns in long biological datasets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–642, 2003.

[PCY95] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. An effective hash-based algorithm for mining association rules. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD '95)*, pages 175–186. ACM, 1995.

[Pea00] Karl Pearson. I. mathematical contributions to the theory of evolution.—vii. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195(262-273):1–47, 1900.

[PHL01] Jian Pei, Jiawei Han, and Laks V. S. Lakshmanan. Mining frequent itemsets with convertible constraints. In *Proceedings 17th International Conference on Data Engineering*, pages 433–442. IEEE, 2001.

[PHM+00] Jian Pei, Jiawei Han, Runying Mao, et al. Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, volume 4, pages 21–30, 2000.

[Pig09] Massimo Pigliucci. The end of theory in science? *EMBO reports*, 10(6):534–534, 2009.

[PMS03] D. N. 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.

[PNS+07] Marie Plasse, Ndeye Niang, Gilbert Saporta, Alexandre Villeminot, and Laurent Leblond. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics & Data Analysis*, 52(1):596–613, 2007.

[Pod18] Karlis Podnieks. Philosophy of modeling: Neglected pages of history. 2018.

[Pon01] Ivo Ponocny. Nonparametric goodness-of-fit tests for the rasch model. *Psychometrika*, 66(3):437–459, 2001.

[Pop59] Karl R. Popper. The logic of scientific discovery. 1959. Karl Popper rewrote his book in English from the German original printed in 1943 and titled Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft.

[Por14] Demetris Portides. How scientific models differ from works of fiction. In *Model-based reasoning in science and technology*, pages 75–87. Springer, 2014.

[Pot17] A. Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, 2017.

[Pow02] Daniel J. Power. Ask dan! *DSS News*, 3(23), 2002.

[PS91] Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, pages 229–238, 1991.

[Qui87] J. Ross Quinlan. Generating production rules from decision trees. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence August 23-28,1987 Milan, Italy*, volume 87, pages 304–307. IJCAI, 1987.

[Ras60] G. Rasch. *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press; Expanded ed edition (January 1, 1980). Copenhagen: Danish Institute for Education, 1960.

[Rau06] Wolfgang Rautenberg. *A concise introduction to mathematical logic.* Springer, 2006.

[Rei38] Hans Reichenbach. Experience and prediction: An analysis of the foundations and the structure of knowledge. 1938.

[Ris78] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[RS17] Gernot Rieder and Judith Simon. Big data: A new empiricism and its epistemic and socio-political consequences. In *Berechenbarkeit Der Welt?*, pages 85–105. Springer, 2017.

[RSHF15] Paul J. Riccomini, Gregory W. Smith, Elizabeth M. Hughes, and Karen M. Fries. The language of mathematics: The importance of teaching and learning mathematical vocabulary. *Reading & Writing Quarterly*, 31(3):235–252, 2015.

[RU11] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[Rus07] Bertrand Russell. The study of mathematics. *The New Quarterly; A Review of Science and Literature*, (1):31–44, November 1907. included in Russell's book Mysticism and Logic, and Other Essays published in 1919 by Longman.

[Rus11] Andrzej Ruszczynski. *Nonlinear optimization*. Princeton university press, 2011.

[RWLN89] Jeff Rothenberg, Lawrence E. Widman, Kenneth A. Loparo, and Norman R. Nielsen. The nature of modeling. *Artificial Intelligence, Simulation and Modeling*, 1989.

[Sah99] Sigal Sahar. Interestingness via what is not interesting. In *KDD*, volume 99, pages 332–336. Citeseer, 1999.

[SBK12] Idheba Mohamad Ali O. Swesi, Azuraliza Abu Bakar, and Anis Suhailis Abdul Kadir. Mining positive and negative association rules from interesting frequent and infrequent itemsets. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 650–655. IEEE, 2012.

[Sch71] R. L. E. Schwarzenberger. The language of geometry. *Mathematical Spectrum*, 1971.

[Sha48] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[Sha49] Claude Elwood Shannon. *The Mathematical Theory of Communication, by CE Shannon (and Recent Contributions to the Mathematical Theory of Communication), W. Weaver*. University of illinois Press, 1949.

[SHS+00] Pradeep Shenoy, Jayant R. Haritsa, S. Sudarshan, Gaurav Bhalotia, Mayank Bawa, and Devavrat Shah. Turbo-charging vertical mining of large databases. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD '00)*, volume 29, pages 22–33. ACM, 2000.

[SJ80] John Shore and Rodney Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980.

[SK97] John Slater and Peter Köllner. The collected papers of Bertrand Russell, volume 11: Last philosophical testament 1947-68. 1997.

[SK11] Arno Siebes and René Kersten. A structure function for transaction data. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 558–569. SIAM, 2011.

[SK12] Tamás Szántai and Edith Kovács. Hypergraphs as a mean of discovering the dependence structure of a discrete multivariate probability distribution. *Annals of Operations Research*, 193(1):71–90, 2012.

[Slo19] Neil James Alexander Sloane. *The On-Line Encyclopedia of Integer Sequences*. published electronically at https://oeis.org, 2019.

[SMS98] James G. Sanderson, Michael P. Moulton, and Ralph G. Selfridge. Null matrices and the analysis of species co-occurrences. *Oecologia*, 116(1-2):275–283, 1998.

[SNB+14] Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature communications*, 5(1):1–9, 2014.

[SNK06] Laszlo Szathmary, Amedeo Napoli, and Sergei O. Kuznetsov. Zart: A multifunctional itemset mining algorithm. Research report, 2006.

[SNV07] Laszlo Szathmary, Amedeo Napoli, and Petko Valtchev. Towards rare itemset mining. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 1, pages 305–312. IEEE, 2007.

[SO15] Parth Suthar and Bhavesh Oza. A survey of web usage mining techniques. *Int. J. Comput. Sci. Inf. Technol.(IJCSIT)*, 6(6), 2015.

[Sol64a] Ray J. Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

[Sol64b] Ray J. Solomonoff. A formal theory of inductive inference. part ii. *Information and control*, 7(2):224–254, 1964.

[SON95] Ashok Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. In *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland.*, pages 432–444, 1995.

[SON98] Ashok Savasere, Edward Omiecinski, and Shamkant Navathe. Mining for strong negative associations in a large database of customer transactions. In *Proceedings 14th International Conference on Data Engineering*, pages 494–502. IEEE, 1998.

[SRDCS19] Nicola Serra, Teresa Rea, Paola Di Carlo, and Consolato Sergi. Continuity correction of pearson's chi-square test in 2x2 contingency tables: A mini-review on recent development. *Epidemiology, Biostatistics and Public Health*, 16(2), 2019.

[SSLL08] Z. Su, W. Song, M. Lin, and J. Li. Web text clustering for personalized e-learning based on maximal frequent itemsets.

In *2008 International Conference on Computer Science and Software Engineering*, volume 6, pages 452–455, 2008.

[ST95] Abraham Silberschatz and Alexander Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proceedings of the Ist International Conference on Knowledge Discovery and Data Mining (KDD-95) Montreal, Canada.*, volume 95, pages 275–281, 1995.

[ST96] Abraham Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and data engineering*, 8(6):970–974, 1996.

[Sta73] Herbert Stachowiak. Allgemeine modelltheorie. 1973.

[Sti02] Stephen M. Stigler. *Statistics on the table: The history of statistical concepts and methods.* Harvard University Press, 2002.

[Str08] Michael Strevens. *Depth: An account of scientific explanation.* Harvard University Press, 2008.

[Stu02] Bernd Sturmfels. *Solving systems of polynomial equations.* Number 97. American Mathematical Soc., 2002.

[Sup61] Patrick Suppes. A comparison of the meaning and uses of models in mathematics and the empirical sciences. In *The concept and the role of the model in mathematics and natural and social sciences*, pages 163–177. Springer, 1961.

[Sup64] Patrick Suppes. What is a scientific theory? US Information Agency, Voice of America Forum, 1964.

[Sup68] Patrick Suppes. The desirability of formalization in science. *The Journal of Philosophy*, 65(20):651–664, 1968.

[Sup69] Patrick Suppes. Models of data. In *Studies in the Methodology and Foundations of Science*, pages 24–35. Springer, 1969.

[Sup07] Patrick Suppes. Statistical concepts in philosophy of science. *Synthese*, 154(3):485–496, 2007.

[SV12] Koen Smets and Jilles Vreeken. Slim: Directly mining descriptive patterns. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 236–247. SIAM, 2012.

[Tat06] Nikolaj Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*, 98(5):183–187, 2006.

[Tat08] Nikolaj Tatti. Maximum entropy based significance of itemsets. *Knowledge and Information Systems*, 17(1):57–77, 2008.

[TCCY12] Yongxin Tong, Lei Chen, Yurong Cheng, and Philip S. Yu. Mining frequent itemsets over uncertain databases. *arXiv preprint arXiv:1208.0292*, 2012.

[TJ06] Martin Thomson-Jones. Models and the semantic view (with 2012 note to the reader). *Philosophy of science*, 73(5):524–535, 2006.

[TKS04] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004. Knowledge Discovery and Data Mining (KDD 2002).

[TM10] Nikolaj Tatti and Michael Mampaey. Using background knowledge to rank itemsets. *Data Mining and Knowledge Discovery*, 21(2):293–309, 2010.

[TPMD+13] Emiliano Torre, David Picado-Muiño, Michael Denker, Christian Borgelt, and Sonja Grün. Statistical evaluation of synchronous spike patterns extracted by frequent item set mining. *Frontiers in computational neuroscience*, 7:132, 2013.

[TSB09] Luigi Troiano, Giacomo Scibelli, and Cosimo Birtolo. A fast algorithm for mining rare itemsets. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 1149–1155. IEEE, 2009.

[TSKK18] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining (2Nd Edition)*. Pearson, 2nd edition, 2018.

[Tuk62] John W. Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[Tuk77] John W. Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

[Tuk80] John W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980.

[TV08] Nikolaj Tatti and Jilles Vreeken. Finding good itemsets by packing data. In *2008 Eighth IEEE International Conference on Data Mining*, pages 588–597. IEEE, 2008.

[TV12] Nikolaj Tatti and Jilles Vreeken. The long and the short of it: summarising event sequences with serial episodes. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2012.

[Uma01] Christopher Umans. The minimum equivalent dnf problem and shortest implicants. *Journal of Computer and System Sciences*, 63(4):597–611, 2001.

[Vai06] Benoît Vaillant. *Mesurer la qualité des règles d'association: études formelles et expérimentales.* PhD thesis, Télécom Bretagne, 2006.

[VBB+18a] Cédric Villani, Yann Bonnet, Charly Berthet, François Levin, Marc Schoenauer, Anne Charlotte Cornut, and Bertrand Rondepierre. *Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne.* Conseil national du numérique, 2018.

[VBB+18b] Cédric Villani, Yann Bonnet, Charly Berthet, François Levin, Marc Schoenauer, Anne Charlotte Cornut, and Bertrand Rondepierre. *Partie 5 : Quelle éthique de l'IA ?*, pages 138–161. Conseil national du numérique, 2018.

[Ver08] Norman D. Verhelst. An efficient mcmc algorithm to sample binary matrices with fixed marginals. *Psychometrika*, 73(4):705, 2008.

[VF67] Bas C. Van Fraassen. Meaning relations among predicates. *Nous*, pages 161–179, 1967.

[VF70] Bas C. Van Fraassen. On the extension of beth's semantics of physical theories. *Philosophy of science*, 37(3):325–339, 1970.

[VF10] Bas C. Van Fraassen. Scientific representation: Paradoxes of perspective, 2010.

[Vin13] Shimon Peter Vingron. *Switching theory: Insight through predicate logic*. Springer Science & Business Media, 2013.

[VL09] Bay Vo and Bac Le. Fast algorithm for mining minimal generators of frequent closed itemsets and their applications. In *2009 International Conference on Computers & Industrial Engineering*, pages 1407–1411. IEEE, 2009.

[VLV14] Matthijs Van Leeuwen and Jilles Vreeken. Mining and using sets of patterns through compression. In *Frequent Pattern Mining*, pages 165–198. Springer, 2014.

[VT14] Jilles Vreeken and Nikolaj Tatti. *Interesting Patterns*, pages 105–134. Springer International Publishing, 2014.

[VVLS11] Jilles Vreeken, Matthijs Van Leeuwen, and Arno Siebes. Krimp: mining itemsets that compress. *Data Mining and Knowledge Discovery*, 23(1):169–214, 2011.

[Web10] Geoffrey I. Webb. Self-sufficient itemsets: An approach to screening potentially interesting associations between items. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):3:1–3:20, 2010.

[Web11] Geoffrey I. Webb. Filtered-top-k association discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):183–192, 2011.

[Wei07] Michael Weisberg. Three kinds of idealization. *The journal of Philosophy*, 104(12):639–659, 2007.

[WHC01] Ke Wang, Yu He, and David W Cheung. Mining confident rules without support requirement. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 89–96. ACM, 2001.

[WHP03] Jianyong Wang, Jiawei Han, and Jian Pei. Closet+: Searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 236–245. ACM, 2003.

[Wil82] Rudolf Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In Ivan Rival, editor, *Ordered Sets - Proceedings of the NATO Advanced Study Institute held at Banff Canada August 28 to September 12 1981*, pages 445–470. Springer Netherlands, 1982. part of the NATO Advanced Study Institutes Series (Series C — Mathematical and Physical Sciences) ASIC vol 83.

[Wim87] William C Wimsatt. False models as means to truer theories. *Neutral models in biology*, pages 23–55, 1987.

[WKRQ+08] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[WP06a] Chao Wang and Srinivasan Parthasarathy. Summarizing itemset patterns using probabilistic models. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pages 730–735. ACM, 2006.

[WP06b] Chao Wang and Srinivasan Parthasarathy. Summarizing itemset patterns using probabilistic models. Research report, 2006.

[XTK03] Hui Xiong, P-N Tan, and Vipin Kumar. Mining strong affinity association patterns in data sets with skewed support distribution. In *Third IEEE International Conference on Data Mining*, pages 387–394. IEEE, 2003.

[Zak00a] Mohammed J. Zaki. Generating non-redundant association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 34–43. ACM, 2000.

[Zak00b] Mohammed J. Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.

[ZAV14] Arthur Zimek, Ira Assent, and Jilles Vreeken. Frequent pattern mining algorithms for data clustering. In *Frequent pattern mining*, pages 403–423. Springer, 2014.

[ZAZ19] Mohammed J. Zaki, Fatimah Audah, and Nurul Fariza Zulkurnain. Improved BVBUC algorithm to discover closed itemsets in long biological datasets. In *Applied Mechanics and Materials*, volume 892, pages 157–167. Trans Tech Publ, 2019.

[ZBB+17] Matthew Zook, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, et al. Ten simple rules for responsible big data research. *PLOS Computational Biology*, 13:1–10, 03 2017.

[ZG03] Mohammed J. Zaki and Karam Gouda. Fast vertical mining using diffsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03)*, pages 326–335. ACM, 2003.

[ZH+99] Mohammed J. Zaki, Ching-Jui Hsiao, et al. Charm: An efficient algorithm for closed association rule mining. Technical report, Citeseer, 1999.

[ZH05] Mohammed J. Zaki and C.-J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE transactions on knowledge and data engineering*, 17(4):462–478, 2005.

[ZMJM14] Mohammed J. Zaki, Wagner Meira Jr, and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press, 2014.

[ZMM15] Bo Zhu, Alexandru Mara, and Alberto Mozo. Clus: parallel subspace clustering algorithm on spark. In *East European Conference on Advances in Databases and Information Systems*, pages 175–185. Springer, 2015.

[ZN14] Albrecht Zimmermann and Siegfried Nijssen. Supervised pattern mining and applications to classification. In *Frequent pattern mining*, pages 425–442. Springer, 2014.

[ZPOL97] MJ Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proceedings of the 3rd Intl. Conf. on Knowledge Discovery and Data Mining (KDD 97)*. AAAI, 1997.

[ZYHP07] Feida Zhu, Xifeng Yan, Jiawei Han, and S. Yu Philip. gprune: a constraint pushing framework for graph pattern mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 388–400. Springer, 2007.

[ZYTW10] Wen Zhang, Taketoshi Yoshida, Xijin Tang, and Qing Wang. Text clustering using frequent itemsets. *Knowledge-Based Systems*, 23(5):379–388, 2010.

Titre : Extraction objective et signifiante de motifs intéressants sur la base de leur fréquence

Mots clés : extraction de motifs, extraction d'itemsets, intérêt des motifs, signification des processus d'extraction, maximum d'entropie, indépendance contrainte mutuelle

Résumé :
L'objet de cette thèse est l'étude des processus d'extraction d'informations objectives et intéressantes dans une base de données portant sur la fréquence de cooccurrence de différents attributs dans une population statistique (telles qu'utilisées en itemset mining notamment).
On s'intéresse aux notions de l'objectivité et de la signification de ces processus d'extraction. On relie la question de la signification d'un processus à celle de la modélisation mathématique qui lui est sous-jacente, et on présente une étude détaillée des impacts, en terme de signification, des différents choix de modélisations que l'on peut opérer.
Notre analyse fait ressortir la pertinence de l'utilisation de modèles de maximum d'entropie dans ces processus d'extraction. On présente une nouvelle construction mathématique de ces modèles, autour d'une notion d'indépendance contrainte, spécifiquement adaptée au contexte des itemsets. En s'appuyant sur cette construction et sur des outils de géométrie algébrique, on présente une approche exacte pour le calcul des modèles de maximum d'entropie.
Enfin, en s'appuyant sur l'ensemble des recommandations initiales sur la modélisation des processus d'extraction ainsi que sur la notion d'indépendance contrainte, on présente un nouvel algorithme d'extraction.

Title : Meaningful objective frequency-based interesting pattern mining

Keywords : pattern mining, itemset mining, pattern interestingness, meaningfulness of mining processes, maximum entropy, mutual constrained independence,

Abstract :
In this thesis, we study objective interesting pattern mining processes on datasets such as used in itemset mining. We focus on the notions of objectivity and meaningfulness in mining processes.
We establish a link between the meaningfulness of a mining process and that of its corresponding mathematical modeling. We formulate a number of recommendations in terms of modeling choices for increasing both meaningfulness and objectivity. We also establish a link between the study of objective interesting pattern mining and the issue of the automation of scientific discovery.
Our theoretical analysis exhibits the adequacy of considering maximum entropy models in such mining processes. We then proceed with presenting a novel mathematical construction for such models, based on a notion of constrained independence, which is specifically adapted to the itemset context. Based on this construction and on tools from algebraic geometry, we present an exact method for computing maximum entropy models.
Last, based on our recommendations for the mathematical modeling of pattern mining processes and our notion of constrained independence, we present a new pattern mining algorithm.