# What to expect from a set of itemsets?

Thomas Delacroix, Philippe Lenca, Stéphane Lallich

# What to expect from a set of itemsets?

T. Delacroix[1,*], P. Lenca[b], S. Lallich[c]

[a] *Université Paris Saclay, Polytech Paris Saclay, Orsay, France*
[b] *IMT Atlantique, Lab-STICC, F-29238 Brest, France*
[c] *Université de Lyon 2, Laboratoire ERIC, Lyon, France*

**Abstract**

Dealing with redundancy is one of the main challenges in frequency based data mining and itemset mining in particular. To tackle this issue in the most objective possible way, we introduce the theoretical bases of a new probabilistic concept: Mutual constrained independence (MCI). Thanks to this notion, we describe a MCI model for the frequencies of all itemsets which is the least binding in terms of model hypotheses defined by the knowledge of the frequencies of some of the itemsets. We provide a method for computing MCI models based on algebraic geometry.

We establish the link between MCI models and a class of MaxEnt models which has already known to be used in pattern mining. As such, our research presents further insight on the nature of such models and an entirely novel approach for computing them.

*Keywords:* pattern mining, itemset mining, interestingness, redundancy, maximum entropy model, independence model

## 1. Introduction

Research in frequency based itemset mining and, more generally, frequency based pattern mining has identified the reduction of the huge quantities of patterns extracted through mining as a key objective [1, 21, 49]. End users prefer quality over quantity and data miners should provide for adequate tools to tackle this issue. One manner of considering this issue relies on the idea that some knowledge may be sufficient to infer other elements of knowledge. Elements of knowledge which can be inferred may then be considered as redundant information with low interest to the user. In order to build a pattern mining process based on this idea, three questions must be answered:

*Corresponding author
*Email addresses:* thomas.delacroix-sadighiyan@universite-paris-saclay.fr (T. Delacroix), philippe.lenca@imt-atlantique.fr (P. Lenca), stephane.lallich@univ-lyon2.fr (S. Lallich)

1. What information can be inferred about a pattern or a set of patterns from information about other patterns in the data?
2. How do we measure the redundancy within information about patterns given the notion of inference defined above?
3. How do we extract information about patterns with low redundancy?

We focus here specifically on an answer to the first of these three questions. As such, our work provides for novel theoretical approaches and algorithmic methods with applications in pattern mining, as well as information theory. Conversely, it is not the purpose of this article to present a complete method for the mining of non-redundant itemsets. Note that redundancy is not the only way to consider the interestingness of patterns in frequency based pattern mining and there is a wide variety of alternative approaches, notably those based on interestingness measures, for quantifying the interestingness of individual patterns [16, 30, 29, 48, 28].

In itemset mining, the issue of redundancy was first addressed using exact approaches: given the knowledge of the frequency of some itemsets it might be possible, in some cases, to determine the exact frequency of other itemsets. This gave rise to a number of concepts such as non-derivable itemsets [6, 7], closed itemsets [50] or minimal generators [44]. However, the number of itemsets needed to provide a full exact description of the frequencies of the itemsets in a dataset can still be much too important to render such exact approaches entirely satisfactory [49].

Furthermore, they do not solve the issue of redundancy completely. Indeed, even though we may not infer the frequency of a particular itemset exactly from the known frequencies of other itemsets, this does not mean we have no knowledge about the frequency of this itemset at all. Certain values could be much more surprising than others and the additional knowledge of the unsurprising frequency of an itemset, based on the prior knowledge of the frequencies of other itemsets, can be considered redundant with this prior knowledge. In order to tackle the issue of redundancy from this perspective, it is necessary to determine how the prior knowledge of the frequencies of certain itemsets can provide information about the frequencies of other itemsets. Various approaches have been suggested to address this point based on randomization methods [17, 22], maximum entropy (MaxEnt) models [24, 46, 47, 12, 33, 10] or constrained independence models [14].

We suggest a novel approach towards this issue based on a notion of mutual constrained independence (MCI). This notion generalizes to all sets of itemsets the notion of constrained independence introduced in [14], which applied only to a very specific type of set of itemsets (the set of all proper itemsets of an itemset). We present the theoretical and mathematical foundations of MCI, together with a method for computing MCI models. Furthermore, by exhibiting the relationship between MCI models and MaxEnt models, we show how our computation method can help decrease computation times for MaxEnt models in itemset mining by several orders of magnitude.

For readers familiar with the random-worlds framework [2, 19, 20], the ratio-

nale underlying the construction of the MCI model, in the context of itemsets, is closely related to the one behind the random-worlds framework, in the more general context of formulas in first-order logic. In fact, the MCI model may be seen as a specific instance of the random-worlds framework when considering simplified constraints. As such, the main contribution of this article relies in the algorithmic methods presented for the computation of MCI models as well as the mathematical characterizations of the models on which these methods are grounded. However, the mathematical and algorithmic tools presented in this paper may only be applied within a similarly simplified context. Hence, as the random-worlds framework is not necessary for the construction of the MCI model and may also raise theoretical issues, the MCI approach is presented here independently of this broader framework.

## 2. Preliminaries

### *2.1. Notations*

Before we lay down the theoretical foundations to mutual constrained independence, it is important to settle on a certain number of notations. Indeed, standard notations can differ greatly between the computer science and data mining literature, on the one hand, and the probability and statistics literature, on the other hand. This is particularly the case when considering itemsets.

In an itemset mining context, focus is set on itemsets (i.e. subsets of a set $\mathcal{A} = \{a_1, ..., a_m\}$) on a dataset of transactions (i.e. a list of subsets of $\mathcal{A}$) and $f_X$ typically designates the frequency of transactions within the considered dataset that contain an itemset $X$. Hence, if $X = \{a\}$ and $Y = \{b\}$, $f_{X \cup Y} = f_{\{a,b\}}$ would designate the frequency of transactions in which both $a$ and $b$ are present. Note also that brackets are generally dropped in the itemset mining literature so that $\{a, b\}$ is simply written $ab$. Conversely, in a statistics context, focus is set on events. In this context, it would be more common to have $X$ designate the event associated to the presence of $a$ and $Y$ the event associated to $b$, so that $f_{X \cap Y}$ would typically designate the frequency referred to as $f_{X \cup Y}$ in the itemset literature. Similarly, standard notations in an itemset mining context allow for $f_\emptyset = 1$ to be true while standard notations in a statistics context allow for $f_\emptyset = 0$.

In order to avoid any confusion between these two conflicting notations, we will use a third notation based on logical symbols and Boolean algebras which should represent common ground for both computer scientists and mathematicians. In the following, we will consider:

- a set $\mathcal{A} = \{a_1, ..., a_m\}$ of $m$ items;
- a free Boolean algebra $\mathcal{B}$, generated by $\mathcal{A}$, with operators $\wedge$, $\vee$ and $\neg$, as well as bottom and top elements $\bot$ and $\top$;
- $d = 2^m - 1$;

- a set $\mathcal{I} = (I_i)_{0 \leq i \leq d} \subset \mathcal{B}$ corresponding to all itemsets ordered by natural lexicographic order as illustrated in Table 1;
- a set $\Omega = (\omega_i)_{0 \leq i \leq d} \subset \mathcal{B}$ corresponding to the atoms of $\mathcal{B}$ ordered by natural lexicographic order as illustrated in Table 2.

In tables 1 and 2, we give the correspondence between the standard notations for itemsets and generalized itemsets of size $m$ (see [7]) in the itemset mining literature and our own notations when considering 3 items. Note that the elements in $\mathcal{I}$ can easily be expressed as disjunctions of elements in $\Omega$ as illustrated in Table 3.

| Itemset | $X \in \mathcal{I}$ | | |
|---|---|---|---|
| $\emptyset$ | $I_0$ | $=$ | $\top$ |
| $a_3$ | $I_1$ | $=$ | $a_3$ |
| $a_2$ | $I_2$ | $=$ | $a_2$ |
| $a_2a_3$ | $I_3$ | $=$ | $a_2$ |
| $a_1$ | $I_4$ | $=$ | $a_1$ |
| $a_1a_3$ | $I_5$ | $=$ | $a_1 \wedge a_3$ |
| $a_1a_2$ | $I_6$ | $=$ | $a_1 \wedge a_2$ |
| $a_1a_2a_3$ | $I_7$ | $=$ | $a_1 \wedge a_2 \wedge a_3$ |

Table 1: Correspondence between itemsets and elements in $\mathcal{I}$ for $m = 3$.

| Generalized itemset | $X \in \Omega$ |
|---|---|
| $\overline{a_1}\overline{a_2}\overline{a_3}$ | $\omega_0 = \neg a_1 \wedge \neg a_2 \wedge \neg a_3$ |
| $\overline{a_1}\overline{a_2}a_3$ | $\omega_1 = \neg a_1 \wedge \neg a_2 \wedge a_3$ |
| $\overline{a_1}a_2\overline{a_3}$ | $\omega_2 = \neg a_1 \wedge a_2 \wedge \neg a_3$ |
| $\overline{a_1}a_2a_3$ | $\omega_3 = \neg a_1 \wedge a_2 \wedge a_3$ |
| $a_1\overline{a_2}\overline{a_3}$ | $\omega_4 = a_1 \wedge \neg a_2 \wedge \neg a_3$ |
| $a_1\overline{a_2}a_3$ | $\omega_5 = a_1 \wedge \neg a_2 \wedge a_3$ |
| $a_1a_2\overline{a_3}$ | $\omega_6 = a_1 \wedge a_2 \wedge \neg a_3$ |
| $a_1a_2a_3$ | $\omega_7 = a_1 \wedge a_2 \wedge a_3$ |

Table 2: Correspondence between generalized itemsets and elements in $\Omega$ for $m = 3$.

### *2.2. Measures on $\mathcal{B}$ and transfer matrix*

As there is a natural isomorphism between $(\Omega, \mathcal{B})$ and the measurable space $(\{0,1\}^m, \mathcal{P}(\{0,1\}^m))$ where $\mathcal{P}(\{0,1\}^m)$ is the powerset of $\{0,1\}^m$, we can consider measures on $\mathcal{B}$. If $\mathbf{p}$ is a measure on $\mathcal{B}$, we will 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 $X \in \mathcal{B} \setminus \Omega$.

Note that any measure on $\mathcal{B}$ is defined naturally by its values on the atoms $\Omega$ of $\mathcal{B}$ so that $(p_i)_{0 \leq i \leq d}$ defines $\mathbf{p}$ entirely on $\mathcal{B}$. Furthermore, a measure can also be entirely defined by its values on the elements of $\mathcal{I}$ (i.e. the itemsets). Hence, these two families of patterns can be seen as bases for representing probability

$$
\begin{array}{lcl}
I_0 & = & \omega_0 \vee \omega_1 \vee \omega_2 \vee \omega_3 \vee \omega_4 \vee \omega_5 \vee \omega_6 \vee \omega_7 \\
I_1 & = & \omega_1 \vee \omega_3 \vee \omega_5 \vee \omega_7 \\
I_2 & = & \omega_2 \vee \omega_3 \vee \omega_6 \vee \omega_7 \\
I_3 & = & \omega_3 \vee \omega_7 \\
I_4 & = & \omega_4 \vee \omega_5 \vee \omega_6 \vee \omega_7 \\
I_5 & = & \omega_5 \vee \omega_7 \\
I_6 & = & \omega_6 \vee \omega_7 \\
I_7 & = & \omega_7
\end{array}
$$

Table 3: Elements in $\mathcal{I}$ as disjunctions of elements in $\Omega$ for $m = 3$.

measures on $\mathcal{B}$. We will make this explicit by defining a transfer matrix that allows to switch easily from one representation to the other and which will be used extensively in the rest of this article.

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

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

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:

$$
T^{-1}_{k,l} = \begin{cases} (-1)^{(\mathbf{l}-\mathbf{k})\cdot\mathbf{d}} & \text{if} \quad (\mathbf{d} - \mathbf{l}) \cdot \mathbf{k} = 0 \\ 0 & \text{if} \quad (\mathbf{d} - \mathbf{l}) \cdot \mathbf{k} \neq 0 \end{cases} \tag{2}
$$

Equation (1) is obtained quite directly from the definition of $T$. Indeed, $(\omega_l \implies I_k)$, 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 (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 (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. Similarly, from $(\mathbf{d}-\mathbf{j})\cdot\mathbf{k}=0$, 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 obtain:

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

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 2.

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

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

## 3. Rationale of the MCI approach

### *3.1. Problem statement*

Our aim is to define a means to determine what to expect from the frequencies of all itemsets given the knowledge of the frequencies of some itemsets. In other words, we need to determine a natural probability measure on $\mathcal{B}$ given constraints for its values on a number of 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.

*3.2. Objective hypotheses*

Before we provide an answer to the problem statement described above, let us consider the wider issue of formulating a hypothesis about a mathematical object based on partial knowledge of this object.

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[1]. 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 one of the defining characteristics of intelligence.

However, it would be quite limiting to consider that the scope of the knowledge of a system can only be reached through exact 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 necessarily 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. This last distribution represents the best of our knowledge about this variable. Note that this general approach is similar to the one presented in the random-worlds

[1] 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 focus here on this second aspect.

framework [20, 19, 2]. However, as we show in section 6.3, our specific focus on itemsets rather than the complete set of formulas within a logical structure allows us to better address the issue of redundancy and leads us directly to the algebraic approach described in section 7.

Now, several approaches can be used to determine an objective hypothetical value for this variable from this distribution. 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 we focus on here, 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). As we will show in section 6.2, each of these options correspond to a different approach towards the definition of MaxEnt models.

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 costly computation of the underlying uniform distribution.

### *3.3. Application to the problem statement*

Let us now try to understand how the approaches described in section 3.2 can apply to the problem statement defined in section 3.1. 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 3.2, 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. As we will 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 will 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 5, is central to the notion of mutual constrained independence described in this article.

## 4. 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 $\mu$ 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:

$$\mu = \frac{\sum_{\mathbf{h} \in \mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}}} \frac{1}{\mathbf{h}!}\mathbf{h}}{\sum_{\mathbf{h} \in \mathcal{M}_{\overline{\mathcal{K}},\mathbf{g}_{|\overline{\mathcal{K}}}}} \frac{1}{\mathbf{h}!}} \ . \tag{3}$$

By linearity, $\mu$ 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 show in the third of the following three examples.

### *4.1. 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 $\mu_i$ 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 absolute 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, 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 column vectors 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 specific case is considered in less general terms in [14] and other similar approaches have been discussed in [35, 46]. 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 2.2, 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, presented in [6] in the context of non-derivable itemsets, 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:

$$\mu(h_{I_d}) = \frac{\sum_{k=l}^{u} \frac{k}{\prod_{i=0}^{d} h_i(k)!}}{\sum_{k=l}^{u} \frac{1}{\prod_{i=0}^{d} h_i(k)!}} \tag{4}$$

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) becomes:

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

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 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: 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 $\mathcal{K} = \mathcal{A}$.

### *4.2. Computing $\mu$*

Another one of the issues with the finite approach is the difficulty in computing the value of $\mu$. Indeed, if we set aside some trivial cases such as the one corresponding to the independence model for which the formula simplifies easily, computing $\mu$ directly from equation 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 both $m \geq 3$ and $n \times m \geq 10^3$. Therefore, other means for computing $\mu$ must be envisaged.

One alternative approach is to use randomization methods in order to determine an approximate value for $\mu$. Such methods have been considered in itemset mining for a similar yet distinct problem (see [22]) 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 [22], they cannot be reasonably computed if both $m \geq 3$ 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.

Another means to approximate $\mu$ is through the mutual constrained independence models which we will define in the following section. Indeed, we will show in section 5 that $\frac{1}{n}\mu$ converges towards a distribution and this limit may be used to approximate $\mu$. Moreover, we will show that this value is arguably a more relevant theoretical choice than the measure $\mu$ which is tied to the number of transactions.

## 5. Asymptotic approach

### *5.1. Mutual constrained independence (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 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:*

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

*imply that* $\frac{1}{g_{\top}^{(k)}} \mu^{(k)} \underset{k\to+\infty}{\longrightarrow} \mathbf{p}$*, where* $\mu^{(k)}$ *is the measure defined by* $\left(\overline{\mathcal{K}}, \mathbf{g}_{|\overline{\mathcal{K}}}^{(k)}\right)$ *as in section 4.*

### *5.2. Model justification*

Assuming the validity of Theorem 1 (the proof of which is provided in section 5.3), 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 3.1.

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 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 $\mathbf{p}$ 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 $n$ because the number $m$ of items must still be taken into account.

As we will make explicit in section 6.2, the link between $\mathbf{p}$ and MaxEnt models further justifies the use of this asymptotic approach.

### *5.3. Proof of the convergence theorem*

Our proof of Theorem 1 is a constructive one which allows to characterize $\mathbf{p}$. 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 $\mathbf{p}$.

*Preliminary step 1: Reduced transfer matrix.* Recall that the aim is to define the probability measure $\mathbf{p}$ 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 $\Omega$, 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 $\mathcal{C} = (\mathcal{K}, \mathbf{f}_{|\mathcal{K}})$ 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 $\mathbf{g}$, we see that $\mathbf{g}$ satisfies $\mathcal{C}$ if and only if $T_{\mathcal{K}} X_{\mathbf{g}} = K$.

Table 5 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 $a_1 \wedge a_2 \wedge a_3$.

| $X \in \mathcal{K}$ | $f_X$ |
|---|---|
| $\top$ | 1 |
| $a_3$ | $1/2$ |
| $a_2$ | $1/3$ |
| $a_1 \wedge a_2 \wedge a_3$ | $1/5$ |

$$\longleftrightarrow \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} g_0 \\ \vdots \\ g_7 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/5 \end{bmatrix}$$

Table 5: A constraint system and its corresponding matrix equation.

As we will make explicit with Theorem 2, 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 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 5.

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

Figure 2: The basis $\mathcal{B}_{\mathcal{K}}$ of $\mathrm{Ker}(T_{\mathcal{K}})$ with $T_{\mathcal{K}}$ as in Table 5.

*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 3.2. This comes down to the same problem as distinguishing between derivable and non-derivable itemsets [6]. For this purpose, we introduce the notions of derivable constraint system and largest derivable constraint system.

**Definition 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 6: 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 7: An incomplete LDCS

In Table 6 and Table 7, we give examples of constraint systems and their corresponding largest derivable constraint systems. In Table 6, the LDCS is complete. This means that there is only one probability measure on $\mathcal{B}$ which satisfies the constraints. In Table 7, 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 demonstrated further in the proof to Theorem 2, the limit in Theorem 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 1.** *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:*

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

*Proof.* Suppose $X$ and $Y$ are two such vectors. Then $Y - X \in \mathrm{Ker}(T_{\mathcal{K}^*})$ 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 $\varphi$ 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: $\nabla\varphi(Z) \in \mathrm{Im}(T_{\mathcal{K}^*}^T)$. 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, $\mathrm{Im}(T_{\mathcal{K}^*}^T) = \mathrm{Ker}(T_{\mathcal{K}^*})^{\perp}$. Furthermore, $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathrm{Im}(T_{\mathcal{K}^*}^T)$, so the condition becomes $\underline{\ln}(Z) \in \mathrm{Ker}(T_{\mathcal{K}^*})^{\perp}$. By the strict concavity of $\varphi$, we conclude on the uniqueness of such an optimum thus obtaining the desired result. □

*Strong version of Theorem 1 and proof.* Lemma 1 is central in the proof we provide to Theorem 1. As stated previously, this proof is constructive, leading to the following stronger result.

**Theorem 2** (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;*
- $\mathcal{C}^* = (\mathcal{K}^*, \mathbf{f}^*_{|\mathcal{K}^*})$ *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 $\operatorname{Ker}\left(T_{\mathcal{K}^*}\right)$ 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 $\mathbf{g}^{(k)}_{|\overline{\mathcal{K}}}$. If we look at a dataset in $\mathcal{D}_k$, each line of the dataset corresponds uniquely to an element $\omega_i$ of $\Omega$. 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 $\omega_i$ 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 demonstrate that both ${}^{|\mathcal{D}_{k,Y^+}|}\!/_{N_Y!|\mathcal{D}_k|} \underset{k\to+\infty}{\longrightarrow} \prod_{y_i=1} x_i$ and ${}^{|\mathcal{D}_{k,Y^-}|}\!/_{N_Y!|\mathcal{D}_k|} \underset{k\to+\infty}{\longrightarrow} \prod_{y_i=-1} x_i$ hold. To prove this point, we consider a probability with uniform distribution on the finite set of matrices $\mathcal{D}_k$. We note this probability $\mathrm{Prob}_k$. Let $[L_j = \omega_i]$ denote the set of matrices of $\mathcal{D}_k$ for which the $j$-th row corresponds to $\omega_i$ and $[|\omega_i| = l]$ the set of matrices of $\mathcal{D}_k$ for which exactly $l$ rows correspond to $\omega_i$. 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. Moreover, 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 $\sigma$ 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 $\omega_i$ 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 us consider $i \neq j$ such that $y_i = y_j = 1$ and the define a sequence $(H_k)$ by $H_k = \mathrm{Prob}_k\left([L_1 = \omega_i] \cap [L_2 = \omega_j]\right) - \mathrm{Prob}_k\left(L_1 = \omega_i\right)\mathrm{Prob}_k\left(L_2 = \omega_j\right)$. Our aim is to show that the $(H_k)$ converges towards 0 when $k$ goes to infinity. 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 $\mathrm{Prob}_k\left(L_2 = \omega_j \mid L_1 = \omega_i\right) = \mathrm{Prob}_k\left(L_2 = \omega_j \mid |\omega_i| \geq 1\right) = \mathrm{Prob}_k\left(L_2 = \omega_j\right)\frac{\mathrm{Prob}_k(|\omega_i|\geq 1 \mid L_2=\omega_j)}{\mathrm{Prob}_k(|\omega_i|\geq 1)} = \mathrm{Prob}_k\left(L_2 = \omega_j\right)\frac{1-\mathrm{Prob}_k(|\omega_i|=0 \mid L_2=\omega_j)}{1-\mathrm{Prob}_k(|\omega_i|=0)}$. Hence, $H_k = \mathrm{Prob}_k\left(L_1 = \omega_i\right)\mathrm{Prob}_k\left(L_2 = \omega_j\right)\left[\frac{1-\mathrm{Prob}_k(|\omega_i|=0 \mid L_2=\omega_j)}{1-\mathrm{Prob}_k(|\omega_i|=0)} - 1\right]$.

But both $\mathrm{Prob}_k\left(|\omega_i| = 0\right) \underset{k\to+\infty}{\longrightarrow} 0$ and $\mathrm{Prob}_k\left(|\omega_i| = 0 \mid L_2 = \omega_j\right) \underset{k\to+\infty}{\longrightarrow} 0$. 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 $(x_{k,i})_{k\geq 1}$ is strictly positive for large enough $k$.

Now, the results of the two previous paragraphs can be combined and we get $\prod_{y_i=1} x_i = \prod_{y_i=-1} x_i$. Hence, $\sum_{y_i=1}\ln(x_i) - \sum_{y_i=-1}\ln(x_i) = 0$, 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 $\mathrm{Ker}(T_{\mathcal{K}^*})$,

this gives $\underline{\ln}(X) \in \mathrm{Ker}\,(T_{\mathcal{K}^*})^{\perp}$.

We conclude from lemma 1 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.

## 6. Mutual constrained independence

### *6.1. Formal definitions*

In section 3, 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 5, 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 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* $\Omega$*. 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 1 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 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:*

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

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

*6.2. Relation to MaxEnt models*

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

**Theorem 3.** *Consider notations as in section 6.1. 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*

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

*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 1. 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 2, it coincides therefore with this optimum. □

This result implies that the MCI models which we have constructed are in fact 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 $\mathcal{K}$. As such, our general approach towards the definition of MCI models brings further insight to the maximum entropy principle. Indeed, there are two main and complementary views towards the rationale of MaxEnt models: the standard approach based on Shannon's original interpretation of entropy as an information measure, in which MaxEnt models are seen as the models that add the least information to the system [41, 9]; and E. T. Jaynes approach presented in his famous article "On The Rationale of Maximum-Entropy Methods" [25, 26], in which MaxEnt models are seen as models that have the maximum likelihood given the asymptotic behavior of a uniform distribution on the set of all possible models (similarly to our own asymptotic approach). In our own approach, the MaxEnt models considered are seen as an average model given the asymptotic behavior of a uniform distribution on the set of all

possible models. Note that this view is not limited to the context of itemsets and mutual constrained independence. In fact, we can also derive directly from Jayne's entropy concentration theorem [25] that both the model with highest likelihood and the average model asymptotically converge towards the same MaxEnt model.

Through the expression of MCI models as MaxEnt models given by Theorem 3, we can see that models of this precise nature have already been considered for pattern mining [35, 46, 32, 33, 49] based either on Shannon's or Jayne's approach towards MaxEnt models (or both). As discussed above, the MCI approach brings further justification for the use of maximum entropy methods in pattern mining. But more importantly, the MCI characterization of these models allows to envisage new methods for computing them, which we present in section 7.

### *6.3. Relation to the random-worlds framework*

The MCI approach is closely related to the random-worlds framework [20, 19, 2] as well as methods for probabilistic propositional reasoning [38, 18] which can also be embedded in the random-worlds framework [19]. However, in the MCI approach, the constraints are purposefully limited to constraints on the frequencies of itemsets, while the random-worlds framework considers a wider range of logical formulas.

While such a limitation on the scope of the constraints may be due to an initial focus on applications in itemset mining, it is in fact necessary in order to obtain the mathematical characterization of the MCI model in Definition 3 and apply the algorithmic method presented in section 7 on which it is grounded. More precisely, both the mathematical characterization of the MCI model and the algorithmic method for computing it are based on the notion of the transfer matrix $T$ and its inverse $T^{-1}$, which are necessary in order to express the constraints on the model. As this matrix represents the decomposition of elements of $\mathcal{I}$ (i.e. itemsets) as disjunctions of elements in $\Omega$ (i.e. the set of atoms of $\mathcal{B}$), constraints must be limited to itemsets. Note that this would still work similarly if we considered any other family of patterns $\mathcal{F} \subset \mathcal{B}$ such that the corresponding transfer matrix is invertible, instead of the set of itemsets $\mathcal{I}$, but the method described would not hold when considering a wider range of logical formulas as the transfer matrix would no longer be invertible.

More generally, the invertibility of the transfer matrix can be related to the notion of redundancy. Indeed, if the transfer matrix associated to a set of patterns $\mathcal{F}$ is not injective this implies that a description of a probability measure on $\mathcal{B}$ given by its values on the elements of $\mathcal{F}$ contains redundant information. On the other hand, the surjectivity of the matrix ensures that such a description entirely defines the probability measure. Therefore, if the aim is to reduce redundancy in information, it makes sense to consider a family of patterns for which the transfer matrix is invertible. Moreover, the set of itemsets is arguably the simplest family of patterns satisfying this condition [13].

## 7. Algebraic geometry for computing MCI models

As MaxEnt models are normally presented as the solution to an optimization problem, they are traditionally computed using numerical methods. This is the case for all the examples from the itemset mining literature that we have found in which MaxEnt models are explicitly computed [35, 46, 32, 33]. However, the models described in [14], of which MCI models are a generalization, are presented as the solution of an algebraic problem and computed as such. This led us to research the possibility of computing MCI models using algebraic methods. In the following section, we present a method for computing such models based on algorithms from real algebraic geometry. As we will show, our method may be used to reduce computation times for MaxEnt models by several orders of magnitude in itemset mining contexts.

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 [43]. However, the algorithm he suggests remains an analytical one.

### *7.1. Algebraic geometry for polynomial system solving*

As we 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 [43, 3, 5].

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 article with the terminology defined in [3].

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 [3]) 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 $\mathcal{K}^*$ 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 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 (when performing a single computation), 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 $\mathbb{Q}(f_1, ..., f_d)[\mathbf{X}]$ which requires 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 multiple times very efficiently for any MCI model corresponding to a given $\mathcal{K}^*$.

*7.2. 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 3. The vector $X$ is characterized as the unique solution to a linear and loglinear problem (Theorem 2). 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:

$$\begin{aligned}
L_0 &= X_0 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 - 1 \\
L_1 &= X_1 + X_3 + X_5 + X_7 - f_1 \\
L_2 &= X_2 + X_3 + X_6 + X_7 - f_2 \\
L_3 &= X_3 + X_7 - f_3 \\
L_4 &= X_4 + X_5 + X_6 + X_7 - f_4 \\
L_5 &= X_5 + X_7 - f_5 \\
L_6 &= X_6 + X_7 - f_6 \\
L_7 &= X_7 - f_7
\end{aligned}$$

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 1: Computing $\mathcal{P}_L$

*Loglinear part.* Secondly, let $Y = \begin{bmatrix} y_0 \\ \vdots \\ y_d \end{bmatrix} \in \operatorname{Ker}\left(T_{\mathcal{K}^*}\right) \cap \mathbb{Z}^{d+1}$. 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 $\underline{\ln}(X) \in \operatorname{Ker}\left(T_{\mathcal{K}^*}\right)^{\perp}$ implies that $M_Y(X) = 0$. Our aim now is to pick a family of vectors in $\operatorname{Ker}\left(T_{\mathcal{K}^*}\right) \cap \mathbb{Z}^{d+1}$ 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 $\mathrm{Ker}\,(T_{\mathcal{K}^*})$ (see section 5.3). However, this does not always result in a zero-dimensional polynomial system. Indeed, consider $M_j$ be the polynomial defined by the $j$-th column of $T^{-1}$, for each $j \in [\![1, d]\!]$. This gives, for example, when $m = 3$:

$$\begin{array}{rcl} M_1 & = & X_1 - X_0 \\ M_2 & = & X_2 - X_0 \\ M_3 & = & X_0X_3 - X_1X_2 \\ M_4 & = & X_4 - X_0 \\ M_5 & = & X_0X_5 - X_1X_4 \\ M_6 & = & X_0X_6 - X_2X_4 \\ M_7 & = & X_1X_2X_4X_7 - X_0X_3X_5X_6 \end{array}$$

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 $X \in \mathcal{Z}_M \cap \left(\mathbb{R}^*\right)^{d+1}$ because $X \in \mathcal{Z}_M$ implies $|X| \in \mathcal{Z}_M$ where $|X| = \begin{bmatrix} |x_0| \\ \vdots \\ |x_d| \end{bmatrix} \in \left(\mathbb{R}_+\right)^{d+1}$.

Indeed, $\forall M_Y \in \mathcal{P}_M,\ M_Y(X) = 0 \iff \prod_{\substack{i=0\\ y_i>0}}^{d} x_i^{y_i} - \prod_{\substack{i=0\\ y_i<0}}^{d} x_i^{-y_i} = 0 \iff$

$\prod_{\substack{i=0\\ y_i>0}}^{d} x_i^{y_i} = \prod_{\substack{i=0\\ y_i<0}}^{d} x_i^{-y_i} \implies \left|\prod_{\substack{i=0\\ y_i>0}}^{d} x_i^{y_i}\right| = \left|\prod_{\substack{i=0\\ y_i<0}}^{d} x_i^{-y_i}\right| \iff \prod_{\substack{i=0\\ y_i>0}}^{d} |x_i|^{y_i} = \prod_{\substack{i=0\\ y_i<0}}^{d} |x_i|^{-y_i}$

$\iff \prod_{\substack{i=0\\ y_i>0}}^{d} |x_i|^{y_i} - \prod_{\substack{i=0\\ y_i<0}}^{d} |x_i|^{-y_i} = 0 \iff M_Y(|X|) = 0.$

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 2.** *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 2 relies on the other following lemma from linear algebra.

**Lemma 3.** *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 3.* 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 $s$. □

*Proof of lemma 2.* 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 3, as $\dim(\mathrm{Ker}(T_{\mathcal{K}^*})) = s > s'$, there exists a vector $Y' \in \mathrm{Ker}(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1} \setminus \{0\}$ 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. □

Through lemma 2, 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 2 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 2. 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:

1. initialize $J \leftarrow \emptyset$;
2. initialize $\mathcal{Y}_L \leftarrow ()$;
3. for $j$ from $0$ to $d-1$:
4. if $j \notin J$:
5. add $j$ to $J$;
6. for $j'$ from $j+1$ to $d$:
7. if $j' \notin J$:
8. $Y = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$, $y_j = 1$, $y_{j'} = -1$;
9. if $Y \in \mathrm{Ker}\,(T_{\mathcal{K}^*})$:
10. add $j'$ to $J$;
11. add $Y$ to $\mathcal{Y}_L$;

Algorithm 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 $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$, 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 $\mathrm{Ker}\,(T_{\mathcal{K}^*}) \cap \mathbb{Z}^{d+1}$;
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 3: Computing $\mathcal{Y}_{NL}$

The issue of choosing $Y'$ in step 5 of the previous algorithm can be resolved as follows:

1. consider the matrix $B$ such that each row corresponds to an element from $\mathcal{B}_{\mathcal{K}^*}$;
2. reorder the columns of $B$ so that the first $s'$ columns correspond to the columns with indices in $S'$;
3. reduce $B$ to its reduced row echelon form;
4. set $Y'$ to the last row of $B$;
5. rearrange the columns of $Y'$ back to the original order of indices;

Algorithm 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 2. We reduce the size of $\mathcal{Y}$ using a greedy algorithm:

1. 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 $\mathcal{T}' = \emptyset$:
2. remove $Y$ from $\mathcal{Y}$;

Algorithm 5: Reducing $\mathcal{Y}_{NL}$

By combining all these algorithms, we obtain an algorithm for computing $\mathcal{P}_M$:

1. compute $\mathcal{Y}_L$ via algorithm 2;
2. compute $\mathcal{Y}_{NL}$ via algorithm 3 and algorithm 4;
3. reduce $\mathcal{Y}_{NL}$ via algorithm 5;
4. initialize $\mathcal{P}_{M_L} \leftarrow \emptyset$;
5. for $Y \in \mathcal{Y}_L$:
6. add $M_Y$ to $\mathcal{P}_{M_L}$;
7. initialize $\mathcal{P}_{M_{NL}} \leftarrow \emptyset$;
8. for $Y \in \mathcal{Y}_{NL}$:
9. add $M_Y$ to $\mathcal{P}_{M_{NL}}$;
10. $\mathcal{P}_M \leftarrow \mathcal{P}_{M_L} \sqcup \mathcal{P}_{M_{NL}}$;

Algorithm 6: Computing $\mathcal{P}_M$

Finally, we can determine $\mathcal{P}$ through algorithm 1 and algorithm 6:

1. compute $\mathcal{P}_L$ via algorithm 1;
2. compute $\mathcal{P}_M$ via algorithm 6;
3. $\mathcal{P} \leftarrow \mathcal{P}_L \sqcup \mathcal{P}_M$;

Algorithm 7: Computing $\mathcal{P}$

Note that the reduction of the problem to a zero-dimensional polynomial system is critical in order to use an algorithm for determining a univariate representation of the system. To the best of our knowledge, the solution provided above is the first one which allows to unlock this possibility.

### *7.3. 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 7.1 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 section 5.3), 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 $D$;[2]
2. determine $\mathcal{P}$ from $\left(\mathcal{K}^*, \mathbf{f}_{|\mathcal{K}^*}\right)$ via algorithm 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 7.1);
5. find $X \in \mathcal{Z}$ such that $x_i > 0$, $\forall i \in [\![0, d]\!] \setminus D$;

Algorithm 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 $D = \emptyset$ (as all cases in which $D \neq \emptyset$ can be obtained by continuity from cases in which $D = \emptyset$).

### *7.4. 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: 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 $i \in [\![1, 5]\!]$.

[2] This is a simple problem in linear programming which can be solved through the use of a simplex algorithm for example.

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 [33].

### *7.5. Speed-ups for step 4*

As stated previously, the bottleneck of algorithm 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 2 (noted $\mathcal{P}_{M_L}$ in algorithm 6) and the polynomials added to $\mathcal{P}$ in step 3 of algorithm 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 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 [33] 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.

*7.6. Algebraic solutions for all cases when $m \leq 4$*

In section 7.1, 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 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 [42], 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 7.4. The number of such cases corresponds to sequence A323819 in [42], 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 | $9.223 \times 10^{18}$ | $1.281 \times 10^{16}$ | $1.281 \times 10^{16}$ |
| 7 | $1.701 \times 10^{38}$ | $3.376 \times 10^{34}$ | $3.376 \times 10^{34}$ |

Table 8: Sequences for the number of cases to compute.

The number of cases to compute is therefore reasonable enough for us to en-

visage 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 $f_1, ..., f_d$).

### *7.7. Solutions for $m = 3$*

We list below the computed algebraic expressions corresponding to representatives for each of the 30 different equivalence classes described in section 7.6 when $m = 3$. 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.

| $\{f_7\}$ | $\{f_6, f_7\}$ | $\{f_5, f_6\}$ $Q$ |
|---|---|---|
| $x_0 = \frac{1-f_7}{7}$ | $x_0 = \frac{1-f_6}{6}$ | $x_7 = t$ |
| $x_1 = x_0$ | $x_1 = x_0$ | $x_0 = \frac{1-f_5-f_6+x_7}{5}$ |
| $x_2 = x_0$ | $x_2 = x_0$ | $x_1 = x_0$ |
| $x_3 = x_0$ | $x_3 = x_0$ | $x_2 = x_0$ |
| $x_4 = x_0$ | $x_4 = x_0$ | $x_3 = x_0$ |
| $x_5 = x_0$ | $x_5 = x_0$ | $x_4 = x_0$ |
| $x_6 = x_0$ | $x_6 = f_6 - f_7$ | $x_5 = f_5 - x_7$ |
| $x_7 = f_7$ | $x_7 = f_7$ | $x_6 = f_6 - x_7$ |

$Q = 4T^2 - (1 + 4(f_5 + f_6))T + 5f_5f_6$

| $\{f_5, f_6, f_7\}$ | $\{f_4, f_7\}$ | $\{f_4, f_6, f_7\}$ |
|---|---|---|
| $x_0 = \frac{1-f_5-f_6+f_7}{5}$ | $x_0 = \frac{1-f_4}{4}$ | $x_0 = \frac{1-f_4}{4}$ |
| $x_1 = x_0$ | $x_1 = x_0$ | $x_1 = x_0$ |
| $x_2 = x_0$ | $x_2 = x_0$ | $x_2 = x_0$ |
| $x_3 = x_0$ | $x_3 = x_0$ | $x_3 = x_0$ |
| $x_4 = x_0$ | $x_4 = \frac{f_4-f_7}{3}$ | $x_4 = \frac{f_4-f_6}{2}$ |
| $x_5 = f_5 - f_7$ | $x_5 = x_4$ | $x_5 = x_4$ |
| $x_6 = f_6 - f_7$ | $x_6 = x_4$ | $x_6 = f_6 - f_7$ |
| $x_7 = f_7$ | $x_7 = f_7$ | $x_7 = f_7$ |

| $\{f_4, f_5, f_6\}$ | $\{f_4, f_5, f_6, f_7\}$ | $\{f_3, f_5, f_6\}$ $Q$ |
|---|---|---|
| $x_7 = \frac{f_5 f_6}{f_4}$ | $x_0 = \frac{1-f_4}{4}$ | $x_7 = t$ |
| $x_0 = \frac{1-f_4}{4}$ | $x_1 = x_0$ | $x_0 = \frac{1-f_3-f_5-f_6+2x_7}{4}$ |
| $x_1 = x_0$ | $x_2 = x_0$ | $x_1 = x_0$ |
| $x_2 = x_0$ | $x_3 = x_0$ | $x_2 = x_0$ |
| $x_3 = x_0$ | $x_4 = f_4 - f_5 - f_6 + f_7$ | $x_3 = f_3 - x_7$ |
| $x_4 = f_4 - f_5 - f_6 + x_7$ | $x_5 = f_5 - f_7$ | $x_4 = x_0$ |
| $x_5 = f_5 - x_7$ | $x_6 = f_6 - f_7$ | $x_5 = f_5 - x_7$ |
| $x_6 = f_6 - x_7$ | $x_7 = f_7$ | $x_6 = f_6 - x_7$ |

$Q = 20T^3 + 4(1 - 5(f_3 + f_5 + f_6))T^2 + ((1 - (f_3 + f_5 + f_6))^2 + 16(f_3 f_5 + f_3 f_6 + f_5 f_6))T - 16 f_3 f_5 f_6$

| $\{f_3, f_5, f_6, f_7\}$ | $\{f_3, f_4, f_7\}$ | $\{f_3, f_4, f_6\}$ $Q$ |
|---|---|---|
| $x_0 = \frac{1-f_3-f_5-f_6+2f_7}{4}$ | $x_0 = \frac{1-f_3-f_4+f_7}{3}$ | $x_7 = t$ |
| $x_1 = x_0$ | $x_1 = x_0$ | $x_0 = \frac{1-f_3-f_4+x_7}{3}$ |
| $x_2 = x_0$ | $x_2 = x_0$ | $x_1 = x_0$ |
| $x_3 = f_3 - f_7$ | $x_3 = f_3 - f_7$ | $x_2 = x_0$ |
| $x_4 = x_0$ | $x_4 = \frac{f_4-f_7}{3}$ | $x_3 = f_3 - x_7$ |
| $x_5 = f_5 - f_7$ | $x_5 = x_4$ | $x_4 = \frac{f_4-f_6}{2}$ |
| $x_6 = f_6 - f_7$ | $x_6 = x_4$ | $x_5 = x_4$ |
| $x_7 = f_7$ | $x_7 = f_7$ | $x_6 = f_6 - x_7$ |

$Q = 2T^2 + (f_4 - 2f_3 - 3f_6 - 1)T + 3f_3 f_6$

| $\{f_3, f_4, f_6, f_7\}$ | $\{f_3, f_4, f_5, f_6\}$ $Q$ | $\{f_3, f_4, f_5, f_6, f_7\}$ |
|---|---|---|
| $x_0 = \frac{1-f_3-f_4+f_7}{3}$ | $x_7 = t$ | $x_0 = \frac{1-f_3-f_4+f_7}{3}$ |
| $x_1 = x_0$ | $x_0 = \frac{1-f_3-f_4+x_7}{3}$ | $x_1 = x_0$ |
| $x_2 = x_0$ | $x_1 = x_0$ | $x_2 = x_0$ |
| $x_3 = f_3 - f_7$ | $x_2 = x_0$ | $x_3 = f_3 - f_7$ |
| $x_4 = \frac{f_4-f_6}{2}$ | $x_3 = f_3 - x_7$ | $x_4 = f_4 - f_5 - f_6 + f_7$ |
| $x_5 = x_4$ | $x_4 = f_4 - f_5 - f_6 + x_7$ | $x_5 = f_5 - f_7$ |
| $x_6 = f_6 - f_7$ | $x_5 = f_5 - x_7$ | $x_6 = f_6 - f_7$ |
| $x_7 = f_7$ | $x_6 = f_6 - x_7$ | $x_7 = f_7$ |

$Q = 4T^3 + (1 - 4(f_3 + f_5 + f_6))T^2 + (3(f_3 f_5 + f_3 f_6 + f_5 f_6) + (1 - f3 - f4)(f4 - f5 - f6))T - 3f_3 f_5 f_6$

| $\{f_2, f_4, f_7\}$ $Q$ | $\{f_2, f_4, f_6, f_7\}$ | $\{f_2, f_4, f_5, f_7\}$ |
|---|---|---|
| $x_6 = t$ | $x_0 = \frac{1-f_2-f_4+f_6}{2}$ | $x_6 = \frac{(f_4-f_5)(f_2-f_7)}{1-f_5}$ |
| $x_0 = \frac{1-f_2-f_4+f_7+x_6}{2}$ | $x_1 = x_0$ | $x_0 = \frac{1-f_2-f_4+f_7+x_6}{2}$ |
| $x_1 = x_0$ | $x_2 = \frac{f_2-f_6}{2}$ | $x_1 = x_0$ |
| $x_2 = \frac{f_2-f_7-x_6}{2}$ | $x_3 = x_2$ | $x_2 = \frac{f_2-f_7-x_6}{2}$ |
| $x_3 = x_2$ | $x_4 = \frac{f_4-f_6}{2}$ | $x_3 = x_2$ |
| $x_4 = \frac{f_4-f_7-x_6}{2}$ | $x_5 = x_4$ | $x_4 = f_4 - f_5 - x_6$ |
| $x_5 = x_4$ | $x_6 = f_6 - f_7$ | $x_5 = f_5 - f_7$ |
| $x_7 = f_7$ | $x_7 = f_7$ | $x_7 = f_7$ |

$Q = T^2 + (2 - f_2 - f_4)T - (f4 - f7)(f2 - f7)$

| $\{f_2, f_4, f_5, f_6\}$ | $\{f_2, f_4, f_5, f_6, f_7\}$ | $\{f_2, f_3, f_4, f_5\}$ $Q$ |
|---|---|---|
| $x_7 = \frac{f_5 f_6}{f_4}$ | $x_0 = \frac{1-f_2-f_4+f_6}{2}$ | $x_6 = t$ |
| $x_0 = \frac{1-f_2-f_4+f_6}{2}$ | $x_1 = x_0$ | $x_7 = x_6 - 1 - f_2 + 2f_3 - f_4 + 2f_5 + \frac{2(f_2-f_3)(f_4-f_5)}{x_6}$ |
| $x_1 = x_0$ | $x_2 = \frac{f_2-f_6}{2}$ | $x_0 = \frac{1-f_2-f_4+x_6+x_7}{2}$ |
| $x_2 = \frac{f_2-f_6}{2}$ | $x_3 = x_2$ | $x_1 = x_0$ |
| $x_3 = x_3$ | $x_4 = f_4 - f_5 - f_6 + f_7$ | $x_2 = f_2 - f_3 - x_6$ |
| $x_4 = f_4 - f_5 - f_6 + x_7$ | $x_5 = f_5 - f_7$ | $x_3 = f_3 - x_7$ |
| $x_5 = f_5 - x_7$ | $x_6 = f_6 - f_7$ | $x_4 = f_4 - f_5 - x_6$ |
| $x_6 = f_6 - x_7$ | $x_7 = f_7$ | $x_5 = f_5 - x_7$ |

$Q = T^3 + (1 - (1 + f_2 - f_3)(1 + f_4 - f_5) - (1 - f_3)(1 - f_5))T^2 + (f_2 - f_3)(f_4 - f_5)(f_2 - 2f_3 + f_4 - 2f_5 + 3)T - 2(f_2 - f_3)^2(f_4 - f_5)^2$

| $\{f_2, f_3, f_4, f_5, f_7\}$ $Q_1$ | $\{f_2, f_3, f_4, f_5, f_6\}$ $Q_2$ | $\{f_2, f_3, f_4, f_5, f_6, f_7\}$ |
|---|---|---|
| $x_6 = t$ | $x_7 = t$ | $x_0 = \frac{1-f_2-f_4+f_6}{2}$ |
| $x_0 = \frac{1-f_2-f_4+f_7+x_6}{2}$ | $x_0 = \frac{1-f_2-f_4+f_6}{2}$ | $x_1 = x_0$ |
| $x_1 = x_0$ | $x_1 = x_0$ | $x_2 = f_2 - f_3 - f_6 + f_7$ |
| $x_2 = f_2 - f_3 - x_6$ | $x_2 = f_2 - f_3 - f_6 + x_7$ | $x_3 = f_3 - f_7$ |
| $x_3 = f_3 - f_7$ | $x_3 = f_3 - x_7$ | $x_4 = f_4 - f_5 - f_6 + f_7$ |
| $x_4 = f_4 - f_5 - x_6$ | $x_4 = f_4 - f_5 - f_6 + x_7$ | $x_5 = f_5 - f_7$ |
| $x_5 = f_5 - f_7$ | $x_5 = f_5 - x_7$ | $x_6 = f_6 - f_7$ |
| $x_7 = f_7$ | $x_6 = f_6 - x_7$ | $x_7 = f_7$ |

$Q_1 = T^2 - (1 + f_2 - 2f_3 + f_4 - 2f_5 + f_7)T + 2(f_2 - f_3)(f_4 - f_5)$

$Q_2 = 2T^3 + (f_2 - 2f_3 + f_4 - 2f_5 - 3f_6)T^2 + (f_2f_4 - f_2f_5 - f_2f_6 - f_3f_4 + 2f_3f_5 + 2f_3f_6 - f_4f_6 + 2f_5f_6 + f_6^2)T - f_3f_5f_6$

| $\{f_1, f_2, f_4, f_7\}$ $Q$ | $\{f_1, f_2, f_4, f_6, f_7\}$ | $\{f_1, f_2, f_4, f_5, f_6\}$ |
|---|---|---|
| $x_6 = t$ | $x_5 = \frac{(f_1-f_7)(f_4-f_6)}{1-f_6}$ | $x_7 = \frac{f_5f_6}{f_4}$ |
| $x_5 = \frac{(f_1-f_7)(f_4-f_7-x_6)}{1-f_7-x_6}$ | $x_3 = \frac{(f_2-f_6)(f_1-f_7-x_5)}{1-f_4}$ | $x_3 = \frac{(f_1-f_5)(f_2-f_6)}{1-f_4}$ |
| $x_3 = \frac{(f_1-f_7-x_5)(f_2-f_7-x_6)}{1-f_4}$ | $x_0 = 1 - f_1 - f_2 - f_4 + f_6 + f_7 + x_3 + x_5$ | $x_0 = 1 - f_1 - f_2 - f_4 + f_5 + f_6 + x_3$ |
| $x_0 = 1 - f_1 - f_2 - f_4 + 2f_7 + x_3 + x_5 + x_6$ | $x_1 = f_1 - f_7 - x_3 - x_5$ | $x_1 = f_1 - f_5 - x_3$ |
| $x_1 = f_1 - f_7 - x_3 - x_5$ | $x_2 = f_2 - f_6 - x_3$ | $x_2 = f_2 - f_6 - x_3$ |
| $x_2 = f_2 - f_7 - x_3 - x_6$ | $x_4 = f_4 - f_6 - x_5$ | $x_4 = f_4 - f_5 - f_6 + x_7$ |
| $x_4 = f_4 - f_7 - x_5 - x_6$ | $x_6 = f_6 - f_7$ | $x_5 = f_5 - x_7$ |
| $x_7 = f_7$ | $x_7 = f_7$ | $x_6 = f_6 - x_7$ |

$Q = (1 - f_1)T^2 - (1 - 2f_7 - f_1f_2 - f_1f_4 + 2f_1f_7 + f_2f_4)T + (f_2 - f_7)(f_4 - f_7)(1 - f_1)$

| $\{f_1, f_2, f_4, f_5, f_6, f_7\}$ | $\{f_1, f_2, f_3, f_4, f_5, f_6\}$ $Q$ | $\{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$ |
|---|---|---|
| $x_3 = \frac{(f_1-f_5)(f_2-f_6)}{1-f_4}$ | $x_7 = t$ | $x_0 = 1 - f_1 - f_2 + f_3 - f_4 + f_5 + f_6 - f_7$ |
| $x_0 = 1 - f_1 - f_2 - f_4 + f_5 + f_6 + x_3$ | $x_0 = 1 - f_1 - f_2 + f_3 - f_4 + f_5 + f_6 - x_7$ | $x_1 = f_1 - f_3 - f_5 + f_7$ |
| $x_1 = f_1 - f_5 - x_3$ | $x_1 = f_1 - f_3 - f_5 + x_7$ | $x_2 = f_2 - f_3 - f_6 + f_7$ |
| $x_2 = f_2 - f_6 - x_3$ | $x_2 = f_2 - f_3 - f_6 + x_7$ | $x_3 = f_3 - f_7$ |
| $x_4 = f_4 - f_5 - f_6 + f_7$ | $x_3 = f_3 - x_7$ | $x_4 = f_4 - f_5 - f_6 + f_7$ |
| $x_5 = f_5 - f_7$ | $x_4 = f_4 - f_5 - f_6 + x_7$ | $x_5 = f_5 - f_7$ |
| $x_6 = f_6 - f_7$ | $x_5 = f_5 - x_7$ | $x_6 = f_6 - f_7$ |
| $x_7 = f_7$ | $x_6 = f_6 - x_7$ | $x_7 = f_7$ |

$Q = T^3 - (f_3 + f_5 + f_6 - f_1f_2 - f_1f_4 - f_2f_4 + f_1f_6 + f_2f_5 + f_3f_4)T^2 + (f_3f_5 + f_3f_6 + f_5f_6 + f_1f_2f_4 - f_1f_2f_5 - f_1f_2f_6 - f_1f_3f_4 - f_1f_4f_6 + f_1f_6^2 - f_2f_3f_4 - f_2f_4f_5 + f_2f_5^2 + f_3^2f_4 + 2f_3f_5f_6)T + f_3f_5f_6(f_1 + f_2 - f_3 + f_4 - f_5 - f_6 - 1)$

### *7.8. 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 expression for MCI models in a generic case, 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. Note that this can be reasonably expected as the problem of computing the MCI model (with rational constraints) in general is an NP-hard problem [45][3] and any algorithm which computes MCI models in general is affected by this difficulty (including the algebraic approach). However, the main difference between the algebraic approach versus any other algorithm based on optimization methods is that this high complexity will only affect the computation of the univariate representation in a generic case for a given set of itemsets but it will not affect the computation of any specific instance corresponding to this set of itemsets based on this univariate representation (which breaks down to determining the roots of a rational polynomial which has polynomial complexity [34]).

[3]The proof of NP-hardness in [45] is relative to the computation of models which are a subclass of MCI models.

In order to check this empirically, we chose 20 different samples of $m = 3$ items among the 70 items of the plants database [36], 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 7.7 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 [11][4]. 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 Python3 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 7.6, 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 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 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

[4]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 [24, 39, 46, 47, 33]

help us determine an explicit algebraic formula for such models and provide an interesting alternative to Chow-Liu tree models [8].

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.

## 8. Conclusion

Given the knowledge of the frequencies of a set of itemsets, what frequencies can one reasonably expect for the remaining itemsets? In the course of this paper, we have presented a solution to this question based on a new notion of mutual constrained independence: the MCI model. As we have made explicit, the theoretical basis to our answer is very solid and it has been designed to be the most objective possible answer to this question. We present mathematical proofs for the existence and characterization of MCI models, as well as a complete method for computing them explicitly.

Furthermore, we have shown that our solution to this question coincides exactly with other solutions to the same question which can be obtained when considering objectivity in a different light: MaxEnt models. As such, the MCI approach sheds a new light on the maximum entropy principle and provides a new means to compute such models. We show that our computation method can provide these models at a much greater speed (by a factor over 100) when considering models with a small number of items in comparison with standard numerical methods.

As this paper focuses on theoretical and computational aspects of MCI models, we have not presented an explicit application of the MCI approach. However, the link we have established between MCI models and a specific class of MaxEnt models already used in itemset mining allows to assert that our approach has applications in a wide variety of domains such as text mining, bioinformatics or sociology to list just a few [37, 15, 31] . Furthermore, we believe the notion of mutual constrained independence can also find simple and straightforward applications in statistical hypothesis testing. Indeed, as mutual constrained independence is a natural generalization of independence, we can easily define a generalization of the $\chi^2$ test of independence to a $\chi^2$ test of mutual constrained independence (based on a $\chi^2$ test of goodness of fit for the corresponding MCI

model). For example, one could test a hypothesis that the interactions between multiple variables are simply the result of their pairwise interactions using a test of mutual constrained independence. These tests could suitably be used in a wide variety of domains, ranging from the study of comorbidities in medical research to research on intersectionality, which are two current hot topics in which the difficulty of defining adequate quantitative methods has been clearly pointed out [27, 23, 40, 4].

## References


[1] C. C. Aggarwal. *An Introduction to Frequent Pattern Mining*, pages 1–17. Springer International Publishing, 2014. ISBN 978-3-319-07821-2. doi: 10.1007/978-3-319-07821-2_1. URL https://doi.org/10.1007/978-3-319-07821-2_1.

[2] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. *Artificial intelligence*, 87(1-2):75–143, 1996.

[3] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in Real Algebraic Geometry*. Springer, 2006.

[4] G. R. Bauer and A. I. Scheim. Advancing quantitative intersectionality research methods: Intracategorical and intercategorical approaches to shared and differential constructs. *Social Science & Medicine*, 226:260–262, 2019.

[5] J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

[6] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 74–86. Springer, 2002.

[7] T. Calders and B. Goethals. Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, 14(1):171–206, 2007.

[8] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968. ISSN 0018-9448. doi: 10.1109/TIT.1968.1054142.

[9] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. ISBN 9781118585771. URL https://books.google.fr/books?id=VWq5GG6ycxMC.

[10] S. Dalleiger and J. Vreeken. The relaxed maximum entropy distribution and its application to pattern discovery. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2020.

[11] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, pages 1470–1480, 1972.

[12] T. 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.

[13] T. Delacroix. *Meaningful objective frequency-based interesting pattern mining*. PhD thesis, 2021.

[14] T. Delacroix, A. Boubekki, P. Lenca, and S. Lallich. Constrained independence for detecting interesting patterns. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2015.

[15] P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, and H. B. Le. A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4):e1207, 2017.

[16] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3), 2006. doi: 10.1145/1132960.1132963. URL http://doi.acm.org/10.1145/1132960.1132963.

[17] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(3):14–es, 2007.

[18] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):220–232, 1993.

[19] A. J. Grove, J. Y. Halpern, and D. Koller. Random worlds and maximum entropy. *Journal of Artificial Intelligence Research*, 2:33–88, 1994.

[20] J. Y. Halpern. An analysis of first-order logics of probability. *Artificial intelligence*, 46(3):311–350, 1990.

[21] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, 15 (1):55–86, 2007. ISSN 1573-756X. doi: 10.1007/s10618-006-0059-1. URL https://doi.org/10.1007/s10618-006-0059-1.

[22] S. Hanhijärvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila. Tell me something I don't know: randomization strategies for iterative data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 379–388. ACM, 2009. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557065. URL http://doi.acm.org/10.1145/1557019.1557065.

[23] A. Hassaine, G. Salimi-Khorshidi, D. Canoy, and K. Rahimi. Untangling the complexity of multimorbidity with machine learning. *Mechanisms of ageing and development*, 190:111325, 2020.

[24] S. Jaroszewicz and D. A. Simovici. Pruning redundant association rules using maximum entropy principle. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 135–147. Springer, 2002.

[25] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982. ISSN 0018-9219. doi: 10.1109/PROC.1982.12425.

[26] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[27] M. C. Johnston, M. Crilly, C. Black, G. J. Prescott, and S. W. Mercer. Defining and measuring multimorbidity: a systematic review of systematic reviews. *European journal of public health*, 29(1):182–189, 2019.

[28] S. O. Kuznetsov and T. Makhalova. On interestingness measures of formal concepts. *Information Sciences*, 442:202–219, 2018.

[29] Y. Le Bras, P. Lenca, and S. 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.

[30] P. Lenca, P. Meyer, B. Vaillant, and S. 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.

[31] J. M. Luna, P. Fournier-Viger, and S. Ventura. Frequent itemset mining: A 25 years review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1329, 2019.

[32] M. Mampaey, N. Tatti, and J. 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.

[33] M. Mampaey, J. Vreeken, and N. Tatti. Summarizing data succinctly with the most informative itemsets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):16, 2012. ISSN 1556-4681. doi: 10.1145/2382577.2382580. URL http://doi.acm.org/10.1145/2382577.2382580.

[34] J. M. McNamee and V. Pan. *Numerical Methods for Roots of Polynomials-Part II*. Newnes, 2013.

[35] R. Meo. Theory of dependence values. *ACM Transactions on Database Systems (TODS)*, 25(3):380–406, 2000. ISSN 0362-5915. doi: 10.1145/363951.363956. URL http://doi.acm.org/10.1145/363951.363956.

[36] National Plant Data Center. The plants database, 2008. URL https://archive.ics.uci.edu/ml/datasets/Plants.

[37] S. Naulaerts, P. Meysman, W. Bittremieux, T. N. Vu, W. Vanden Berghe, B. Goethals, and K. Laukens. A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 16(2):216–231, 2015.

[38] N. J. Nilsson. Probabilistic logic. *Artificial intelligence*, 28(1):71–87, 1986.

[39] D. N. Pavlov, H. Mannila, and P. Smyth. Beyond independence: Probabilistic models for query approximation on binary transaction data. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1409–1421, 2003. ISSN 1041-4347. doi: 10.1109/TKDE.2003.1245281.

[40] N. A. Scott and J. Siltanen. Intersectionality and quantitative methods: assessing regression from a feminist perspective. *International Journal of Social Research Methodology*, 20(4):373–385, 2017.

[41] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://ieeexplore.ieee.org/document/6773024.

[42] N. J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences*. published electronically at https://oeis.org, 2019.

[43] B. Sturmfels. *Solving systems of polynomial equations*. Number 97. American Mathematical Soc., 2002.

[44] L. Szathmary, A. Napoli, and S. O. Kuznetsov. Zart: A multifunctional itemset mining algorithm. Research report, 2006. URL https://hal.inria.fr/inria-00001271.

[45] N. Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*, 98(5):183–187, 2006. ISSN 0020-0190. doi: https://doi.org/10.1016/j.ipl.2006.02.003. URL http://www.sciencedirect.com/science/article/pii/S0020019006000408.

[46] N. Tatti. Maximum entropy based significance of itemsets. *Knowledge and Information Systems*, 17(1):57–77, 2008. ISSN 0219-3116. doi: 10.1007/s10115-008-0128-4. URL https://doi.org/10.1007/s10115-008-0128-4.

[47] N. Tatti and M. Mampaey. Using background knowledge to rank itemsets. *Data Mining and Knowledge Discovery*, 21(2):293–309, 2010.

[48] C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, 28(4):1004–1045, 2014.

[49] J. Vreeken and N. Tatti. *Interesting Patterns*, pages 105–134. Springer International Publishing, 2014. ISBN 978-3-319-07821-2. doi: 10.1007/978-3-319-07821-2_5. URL https://doi.org/10.1007/978-3-319-07821-2_5.

[50] M. 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.