A statistically efficient and scalable method for log-linear analysis of high-dimensional data

François Petitjean, Lloyd Allison and Geoffrey I. Webb
Faculty of Information Technology
Monash University, Melbourne, Australia
Email: firstname.lastname@monash.edu

***Abstract*—Log-linear analysis is the primary statistical approach to discovering conditional dependencies between the variables of a dataset. A good log-linear analysis method requires both high precision and statistical efficiency. High precision means that the risk of false discoveries should be kept very low. Statistical efficiency means that the method should discover actual associations with as few samples as possible. Classical approaches to log-linear analysis make use of χ^2 tests to control this balance between quality and complexity. We present an information-theoretic approach to log-linear analysis. We show that our approach 1) requires significantly fewer samples to discover the true associations than statistical approaches – statistical efficiency – 2) controls for the risk of false discoveries as well as statistical approaches – high precision – and 3) can perform the discovery on datasets with hundreds of variables on a standard desktop computer – computational efficiency.**

***Keywords*-Association discovery; Data modeling; High-dimensional data; Log-linear Analysis; Graphical models, Statistical inference; Information theory**

I. Introduction

Association discovery is a fundamental data mining task. Association discovery is generally divided into two main approaches: finding associations between values or items [1], [2], [3], or between variables [4], [5], [6].

Log-linear analysis is the well established statistical technique for finding associations between discrete variables in data [7]. In contrast, data mining research into association discovery has focused primarily on finding associations between variable-values or items [1]. Each of these approaches, finding associations between variables or between variable-values has distinct contexts in which it is more useful. In some cases, the focus is on which settings of which variables are associated with which specific outcomes. In other cases, for example, if one wishes to model a complex multivariate distribution, one needs to know which variables interact in which ways.

The general objective of log-linear analysis (LLA) is to select a log-linear model that satisfactorily explains the observed frequencies of a given dataset. General approaches to LLA are exponential with respect to the number of variables, because the model must be evaluated over all its possible outcomes. There are 2^M outcomes for M binary variables, which makes LLA infeasible for large M. This is why, in the general case, finding a statistically significant log-linear model for a dataset is limited to a dozen variables at most. However, with our Chordalysis-χ^2 technique [6] we recently demonstrated that for the class of *multiplicative* or *decomposable* log-linear models, a statistically standard χ^2-based evaluation can be performed, which allows LLA to scale to high-dimensional data[1].

We propose a new evaluation metric for decomposable models that is based on the information-theoretic Minimum Message Length (MML) principle [8]. We demonstrate that our method discovers existing correlations with significantly fewer samples than statistical methods based on χ^2 goodness-of-fit tests, while consistently controlling for the risk of false discoveries. In addition, we show that, by melding advanced data mining techniques with results from graph theory, we can perform LLA on datasets with hundreds of variables on a standard desktop computer.[2]

This paper is organized as follows. In Section II, we formalize the problem. In Section III, we present our solution Chordalysis-Mml, which enables the discovery of statistically sound multi-way correlations between variables for hundred-dimensional datasets. In section IV, we place this work in the context of related research. In Section V, we conduct experiments that demonstrate 1) the quality and performance of our approach compared to the state of the art and 2) the relevance of our approach on real-world high-dimensional datasets. Finally, we conclude this work and describe future research in Section VI.

II. Problem statement

A. Log-linear models and log-linear analysis

Let $\mathcal{D}$ be a dataset of N samples over a set of M discrete variables $\mathcal{V} = \{V_1, \cdots, V_M\}$. Every variable V takes values in $\text{Dom}(V)$. $\mathcal{D}$ is drawn from a probability distribution $p_{\mathcal{V}}$ over $\mathcal{V}$, giving rise to maximum likelihood estimates $\hat{p}_{\mathcal{V}}$:

$$\begin{array}{rcccc} \hat{p}_{\mathcal{V}} & : & \text{Dom}(V_1) \times \cdots \times \text{Dom}(V_M) & \rightarrow & [0,1] \subset \mathbb{R} \\ & & \mathbf{x} = \big(\mathbf{x}(1), \cdots, \mathbf{x}(M)\big) & \mapsto & \text{O}_{\mathbf{x}} \,/N \end{array}$$

[1]Throughout the paper, we use the term "high-dimensional data" to designate datasets with a high number of variables, regardless of the number of samples.

[2]The open source Chordalysis-Mml software is available at https://sourceforge.net/p/chordalysis.

where $O_{\mathbf{x}}$ designates the observed frequencies for a vector of values $\mathbf{x}$ in $\mathcal{D}$.

Log-linear models use a first-degree polynomial function to model the logarithm of the frequencies that can be observed in a contingency table. The expected frequency $m_{\mathbf{x}}$ of $\mathbf{x}$ under the model has the form:

$$\begin{aligned}\log(m_{\mathbf{x}}) &= u + \sum_{i=1}^{M} u_i(\mathbf{x}(i)) + \sum_{1 \leqslant i < j \leqslant M} u_{i,j}\left(\mathbf{x}(i), \mathbf{x}(j)\right) \\ &+ \cdots + u_{1,\cdots,M}(\mathbf{x}) \end{aligned} \quad (1)$$

The u functions represent the interactions between variables that are used to model the observed frequencies. For example, a model that includes the term $u_{i,j}$ considers the interaction between the i^{th} and the j^{th} variable. Similarly, a model that includes the term $u_{i,j,k}$ will consider the three-way interaction between the i^{th}, the j^{th} and the k^{th} variable, while a model that does not include $u_{i,j,k}$ (having set it to zero) will not consider the corresponding three-way interaction. The *saturated model* is the one including all terms.

Log-linear analysis (LLA) is the general name given to methods that seek to select a statistically significant log-linear model from data. This corresponds to determining which $u_{\ldots}$ terms have to be part of the model. LLA classically uses hypothesis testing to decide if the current reference model (the *null* hypothesis) has to be replaced by a candidate model that is a variation of the reference model (the tested hypothesis).

The objective of LLA is different from the one of probabilistic inference. LLA seeks an *explanatory* model from which conclusions about conditional dependencies/indepencies can be drawn. Probabilistic inference seeks a *predictive* model. Note that an explanatory model may also predict well and conversely; we emphasize here that the focus and use of the two families of methods differ, and have thus led to different methods. In practice, LLA starts from an initial model (the *null* model), and iteratively performs transformations to this model, as long as there is enough evidence to add (or remove) a term to (from) the model. LLA is thus much more conservative than probabilistic inference: terms will be added or removed from the model, if and only if there is enough evidence to be confident that no false discovery will be committed. In medicine for example, this conservative behaviour is critical, because it makes it possible to confidently decide that drugs have an impact on diseases. This is mainly why LLA has such widespread use in medicine, but also in statistics and in social sciences, and why it is integrated into all the classical statistical packages (SPSS, SAS, R, and so on). To summarize, LLA methods must have the following two characteristics:

1. precision – very low rate of false discoveries: minimize the probability of accepting incorrect hypotheses;

2. statistical efficiency: infer the correct hypotheses with as little data as possible.

B. LLA for high-dimensional data

Methods for LLA methods do not scale up beyond a dozen variables for the general class of log-linear models. This is because evaluating the replacement of the current reference model is exponential in the number of variables. This assessment implies iteration over all possible combinations of values. While this is feasible for ten binary variables ($\#operations \geqslant 2^{10} = 1024$), it becomes infeasible when the number of variables is high (*e.g.*, for 100 variables, $\#operations \geqslant 2^{100} > 10^{30}$). Furthermore, these operations have to be performed not just once, but for every potential model that is considered – a number that itself grows exponentially with the dimensionality of the data.

To scale LLA to high-dimensional data, researchers have focused on a subclass of log-linear models that are *decomposable* (or *multiplicative*). We first introduce decomposable models, and then explain why these models are necessary to scale up LLA to high-dimensional data.

Definition 1: [9] A log-linear model is *graphical* if, whenever the model contains all two-factor terms generated by a higher-order interaction, the model also contains the higher-order interaction.

Property 1: Being completely determined by its two-factor terms, *graphical* models can be represented by an undirected graph, where the vertices represent the variables and the edges represent the two-factor terms.

Note that graphical log-linear models are equivalent to Markov networks.

Definition 2: A graphical log-linear model is *decomposable* iff the corresponding graph is *chordal*, *i.e.*, iff the graph does not admit chord-less cycles of length strictly greater than three.

There are three main reasons to focus on decomposable models:

1. Closed-form MLE: Decomposable models are the only log-linear models that have closed-form maximum likelihood estimates (MLEs) [7]. This dramatically improves the computational efficiency of their evaluation.

2. Scalability of the evaluation: Several evaluation metrics can be rewritten for decomposable models, so that they scale up to high-dimensional data: Kullbach-Leibler [10], an MDL score [11] and χ^2 goodness-of-fit tests [6].

3. Usefulness: This sub-class is not only practical but also a useful class of models. This is ensured by the fact that, for any non-decomposable log-linear model μ, there always exists a log-linear model that is decomposable and that subsumes μ and hence can exactly model any distribution modeled by μ. The proof comes from the fact that any minimal triangulation [12] will provide such a subsuming model.

C. Limitations of current approaches

As mentioned above, evaluation frameworks that scale up to high-dimensional data *via* decomposable models include KL divergence in [13], [10] (KL-2001), an MDL

score in [11] (AH-2004, for Altmueller-Haralick) and χ^2 goodness-of-fit tests [6] (CHORDALYSIS-χ^2). KL-2001 usually overfits the true structure, because it optimizes the entropy of the model without taking into account the complexity of the model. AH-2004 often favors overfitting models, because it uses an compression-inefficient encoding scheme: it over-estimates the number of parameters of the model, and uses an inefficient encoding for both the graph structure and the data. Of the three existing techniques, only the standard method in statistics – namely χ^2 goodness-of-fit testing – has a very low-rate of false discoveries, and thus qualifies as a sound LLA framework. These conclusions will be confirmed in the experiments.

However, the low-rate of false discoveries of χ^2 based methods comes at a price: they require many more samples to accept correct hypotheses. Moreover, χ^2-based methods have two functional drawbacks: 1) they rely on the existence of the Maximum Likelihood Estimates (MLE), while particular configurations of zeros in the observed frequencies lead to nonexistent MLE [14], [7], and 2) the confidence on the tests rely on the expected frequencies being greater than 5 [15], which prevents testing of high-order interactions for most datasets.

III. CHORDALYSIS-MML

Our method for log-linear analysis, CHORDALYSIS-MML, is presented in this section. We start by briefly outlining our approach, and then describe in detail its main features.

Our approach includes the following three features:

1. An information-theoretic score: We propose an information-theoretic scoring method for decomposable models. We prove that the length of a lossless encoding scheme can be expressed in terms of the graph structure associated with the model.

2. Efficiently scoring edges: During the forward selection, the hypotheses/models compared always differ by one edge only. We prove that the evaluation of any new edge depends upon the scoring of four local sub-structures of the graph only.

3. Efficiently scoring sub-structure: We show that the scoring of sub-structures of the graph exhibits a number of overlapping computations at different levels. Once every sub-score is correctly memoized, we show that the scoring relies completely on the computation of marginal frequencies for different combinations of values.

We will show that the conjunction of these features allow us to overcome the limitations of state-of-the-art methods, *i.e.*, they make CHORDALYSIS-MML:

1. control for false discoveries as well as χ^2 tests
2. require far fewer samples to accept true hypotheses
3. not reliant on MLE existing, contrary to χ^2 methods
4. not require the expected frequencies to be greater than 5, contrary to χ^2-based methods
5. require the computation of a very limited number of marginal frequencies, and thus is scalable to datasets with hundreds of variables on a standard computer.

A. An MML score for decomposable models

We start by briefly presenting the theoretical framework that is used for developing our scoring scheme, as well as formalizing decomposable models, we present introduce our score for decomposable models.

Overview of Minimum Message Length (MML): The MML criterion provides an information-theoretic objective for problems of inference where the goal is to find the best *explanation* (or *theory*, *hypothesis*, *model*) for a set of observed data [16]. MML relies on quantifying the amount of information *required* to convey losslessly the observed data in an *explanation message*. The best hypothesis is the one that can convey the entire data set in the shortest possible explanation message.

MML [16], like its close cousin MDL (Minimum Description Length) [17], is a practical version of Kolmogorov Complexity [18]. All three embrace the motive: *Induction by Compression*. MML allows priors to be stated more naturally [8]. Thus, it is particularly suited for measuring the quality of an explanation, which aligns perfectly with the explanatory needs of LLA (see Section II-A).

More formally, for some observed data $\mathcal{D}$ and a hypothesis $\mathcal{H}$ that offers an explanation of $\mathcal{D}$, Bayes's theorem gives: $p(\mathcal{H},\mathcal{D}) = p(\mathcal{H}) \times p(\mathcal{D}|\mathcal{H}) = p(\mathcal{D}) \times p(\mathcal{H}|\mathcal{D})$ where $p(\mathcal{H})$ is the prior probability of hypothesis $\mathcal{H}$, $p(\mathcal{D})$ is the *prior* probability of data $\mathcal{D}$, $p(\mathcal{H}|\mathcal{D})$ is the *posterior* probability of $\mathcal{H}$ given $\mathcal{D}$, and $p(\mathcal{D}|\mathcal{H})$ is the likelihood. Using Shannon's theory of communication, the amount of information for an explanation of the $\mathcal{D}$ with $\mathcal{H}$ is:

$$I(\mathcal{H},\mathcal{D}) = I(\mathcal{H}) + I(\mathcal{D}|\mathcal{H}) = I(\mathcal{D}) + I(\mathcal{H}|\mathcal{D}) \qquad (2)$$

where $I(x) = -\log(p(x))$ gives the optimal code length to convey some event x whose probability is $p(x)$. This immediately gives an objective means to compare competing hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ on the same data $\mathcal{D}$:

$$I(\mathcal{H}_1|\mathcal{D}) - I(\mathcal{H}_2|\mathcal{D}) = I(\mathcal{H}_1) + I(\mathcal{D}|\mathcal{H}_1) - I(\mathcal{H}_2) - I(\mathcal{D}|\mathcal{H}_2) \qquad (3)$$

A concrete realization of this framework comes from describing it as a communication process between an imaginary transmitter (Alice) and receiver (Bob) connected over a Shannon channel. Alice's objective is to send the observed data $\mathcal{D}$ using an explanation message in a form such that Bob can receive and decode the data $\mathcal{D}$ precisely as Alice sees it. If Alice can find the best hypothesis on the data, Bob will receive a decodable explanation message most economically: The best inference about the data is the hypothesis that minimizes the total message length.

Alice sends the explanation message of $\mathcal{D}$ in two parts. In the first part, she transmits the hypothesis, $\mathcal{H}$, she could find on the data $\mathcal{D}$, taking $I(\mathcal{H})$ bits. In the second, she transmits the details of the observed data $\mathcal{D}$ not explained by $\mathcal{H}$, taking $I(\mathcal{D}|\mathcal{H})$ bits (*i.e.*, the deviations from $\mathcal{H}$).

In our case, the hypothesis $\mathcal{H}$ on the data is a parameterized decomposable model $\mathcal{M}$. To transmit the explanation message, Alice thus need to transmit, first the structure

of the graph $\mathcal{G}$ associated to $\mathcal{M}$, then the parameters $\mathcal{P}$ of the models, and finally the data $\mathcal{D}$ given $\mathcal{M}$. Later in this section, we will describe the lossless encoding of these three parts of the explanation message.

Decomposable models: Let us introduce some further notation for decomposable models, which we use to construct our encoding. For a decomposable model μ, its probability p_μ can be evaluated using marginal probabilities of a small number of terms. Let $\mathcal{G} = (\mathcal{V}, E)$ be the graph associated to the model. The expression of p_μ is directly linked to $\mathcal{G}$, and relies on the maximal *cliques*, $\mathcal{C}$, and minimal *separators*, $\mathcal{S}$, of the graph $\mathcal{G}$. The probability p_μ of a vector $\mathbf{x}$ under a decomposable model $\mathcal{M}$ is expressed by:

$$p_\mu(\mathbf{x}) = \frac{\prod_{C \in \mathcal{C}} p_C(\mathbf{x})}{\prod_{S \in \mathcal{S}} p_S(\mathbf{x})} \tag{4}$$

with p_A representing the marginal probability of $\hat{p}_\mathcal{V}$ over a set of variables A. The definition and efficient computation of $\mathcal{C}$ and $\mathcal{S}$ are presented in [6] (App. 1). Note that their computation requires only $O(|\mathcal{V}| + |E|)$ operations for chordal graphs.

1) Encoding the graph: First to be transmitted is the structure of the graph $\mathcal{G}$ that is associated to our model $\mathcal{M}$. To this end, it is sufficient to send the edges E of the graph: we first state the number of edges ($|E|$), and then the particular combination of $|E|$ edges that the graph exhibits. Note that the variables do not need to be transmitted because it is common to all possible models, and thus will not play any role in the model selection.

Given that the number of possible edges in a graph is $\#maxE = \frac{|\mathcal{V}| \cdot |\mathcal{V}-1|}{2}$, we have:

$$I(\mathcal{G}) = \log\left(1 + \#maxE\right) + \log \binom{\#maxE}{|E|} \tag{5}$$

2) Encoding the parameters: Once the structure of the graph has been sent, the decoder can reconstruct the graph, as well as the maximal cliques. We now have to transmit the parameters $\mathcal{P}$ of model p_μ. It follows from Eqn. 4 that:

$$I(p_\mu) \quad = \quad \sum_{C \in \mathcal{C}} I(p_C) - \sum_{S \in \mathcal{S}} I(p_S) \tag{6}$$

We thus have to evaluate the length of the message for the transmission of the marginal probabilities $p_A, \forall A \in \mathcal{C} \cup \mathcal{S}$. Let us use $\mathrm{O}^A_\mathbf{x}$ to designate the observed frequencies for the configuration $\mathbf{x}$ with respect to the set of variables A. For any $A \in \mathcal{C} \cup \mathcal{S}$, the MLE for p_A is: $\hat{p}_A(\mathbf{x}) = \frac{\mathrm{O}^A_\mathbf{x}}{N}$. This means that any marginal probability over $A \in \mathcal{C} \cup \mathcal{S}$ can be transmitted by sending N and the associated frequencies $\mathrm{O}^A_\mathbf{x}$. N need not be transmitted because it is common to any model of $\mathcal{D}$. In addition, any frequency $\mathrm{O}^A_\mathbf{x}$ can be transmitted in $\log(N+1)$ bits, because $\mathrm{O}^A_\mathbf{x} \in [\![0, N]\!]$. Note that the last frequency does not have to be transmitted, given that it will be equal to the difference between N and the sum of the transmitted frequencies (which aligns with the number of degrees of freedom). The message length of the parameters $\mathcal{P}$ given the graph structure $\mathcal{G}$ is then:

$$I(\mathcal{P}|\mathcal{G}) = \log(N+1) \cdot \left(\sum_{C \in \mathcal{C}} \#\mathrm{Param}(C) - \sum_{S \in \mathcal{S}} \#\mathrm{Param}(S) \right) \tag{7}$$

with $\#\mathrm{Param}(A) = -1 + \prod_{V \in A} |\,\mathrm{Dom}(A)|$.
Note that $\mathcal{G}$ and $\mathcal{P}$ constitute the model $\mathcal{M}$.

3) Encoding the Data: The classical way to transmit the data $\mathcal{D}$ is to send every instance $\mathbf{x}$ of $\mathcal{D}$ in $-\log\left(p_\mu(x)\right)$ bits. This leads to:

$$I(\mathcal{D}|\mathcal{M}) = N \cdot H(\mathcal{M}) = \sum_{C \in \mathcal{C}} H(C) - \sum_{S \in \mathcal{S}} H(S) \tag{8}$$

with $H(\mathcal{M}) = \sum_x p_\mu(x) \log(p_\mu(x))$ the entropy of p_μ.

However, in our case, we transmitted the parametrized model using the exact frequencies observed in $\mathcal{D}$. We will show how this information can be used to shorten the part of the explanation message that is about the data $\mathcal{D}$.

Consider an example to motivate our intuition. Imagine that we have a model of a tossed coin. This model has only one parameter: the probability of a head $p(H)$ (the probability of a tail is $1 - p(H)$). Let us assume that we tossed the coin 100 times and observed 60 heads. MLE gives $p(H) = 0.6$. Transmitting the data will then take:

$$\begin{aligned} I(\mathcal{D}|\mathcal{M}) \quad &= \quad 100\log(100) - 60\log(60) - 40\log(40) \\ &\approx \quad 97.1 \text{ bits} \end{aligned} \tag{9}$$

However, if the receiver knows that 60 heads have been observed, the only information that is required is the way these heads appeared in $\mathcal{D}$. This can be done by stating what combination of 60 heads among 100 tosses has been observed, which is exactly $\binom{100}{60}$. Sending this combination will then take:

$$\begin{aligned} I(\mathcal{D}|\mathcal{M}) = \log \binom{100}{60} \quad &= \quad \log(100!) - \log(60!) - \log(40!) \\ &\approx \quad 93.5 \text{ bits} \end{aligned} \tag{10}$$

The quantity $I(\mathcal{D}|\mathcal{M})$ from Eqn. 10 will always be smaller than the one from Eqn. 9 [19] and is thus a shorter way to send the data.

We will now demonstrate how to find the length of sending the data given the model, for decomposable models.

Stating the position of the observed frequencies for every maximal clique would state the data. For the model over three variables with two edges $A - B$ and $B - C$, stating the positions of all combinations of A and B would indeed make it possible to reconstitute $\mathcal{D}$ for these variables, and similarly for B and C. However, such a transmission is redundant, because the combinations for B would have been stated twice. This is because B is present (and hence stated) in two different cliques. To avoid "overcounting" the combinations involving B, we have to correct the length of the message by the length of stating the combinations for B.

The difficulty is to know what the message should be corrected by, if the combinations are stated for every

maximal clique. This corresponds to knowing what information would be sent several times if we were to send the combinations for the maximal cliques. Similarly to B being the intersection between AB and BC, it turns out that the set of all the pairwise intersections between the maximal cliques is the *minimal separators*.

We now introduce the length of stating the data $\mathcal{D}$ given that the model $\mathcal{M}$ has been stated using the frequencies, as in Eqn. 7. Stating the combinations for a subset of variables $A \subseteq \mathcal{V}$ given the frequencies O^A is given by [19]:

$$I_A(\mathcal{D}|\,\mathrm{O}^A) \quad = \quad \log(N!) - \sum_{\mathbf{x}\in \mathrm{Dom}(A)} \log(O^A_{\mathbf{x}}!) \qquad (11)$$

This directly gives the length of stating $\mathcal{D}$ given $\mathcal{M}$:

$$\begin{aligned} I(\mathcal{D}|\mathcal{M}) \quad &= \quad |\mathcal{C}|\log\left(N!\right) - \sum_{C\in\mathcal{C}} \sum_{\mathbf{x}\in\mathrm{Dom}(C)} \log\left(O^C_{\mathbf{x}}!\right) \\ &- \quad |\mathcal{S}|\log\left(N!\right) + \sum_{S\in\mathcal{S}} \sum_{\mathbf{x}\in\mathrm{Dom}(S)} \log\left(O^S_{\mathbf{x}}!\right) \quad (12) \end{aligned}$$

B. Efficient scoring edges

The classical search strategy for LLA in high-dimension is *forward selection*, which yield a hill-climbing algorithm [9]. Forward selection starts with a simple model (usually all variables independent) and iteratively adds terms (accepting more complex hypotheses), so long as there is sufficient evidence to accept new hypotheses. Note that hill-climbing strategies are currently the only ones that are compatible with LLA, because statistical goodness-of-fit tests (χ^2) require the compared models to be nested.

For forward selection, the generation of candidate alternatives to a current model relies on the addition of edges, because graphical models are completely defined by their edges (or two-factor terms). In order to ensure that the candidate graphs remain decomposable, it is necessary to consider only edges that result in a chordal graph. Such edges are called 2-pairs [20]. Every time a new model is chosen to replace the previous best one (*i.e.*, at each iteration of the forward selection process), we build a set of eligible interactions associated with the new model. This can be efficiently done using the clique-graph [13]. One new candidate model will then be constructed for every eligible interaction/edge.

A reference model $\mathcal{M}^\star$ will thus be repeatedly compared to a candidate model $\mathcal{M}^c$ that differs by only one edge. Below, we show that the score that we use to decide if $\mathcal{M}^c$ should replace $\mathcal{M}^\star$, $I(\mathcal{M}^c|\mathcal{D}) - I(\mathcal{M}^\star|\mathcal{D})$, is computed from a very limited number of marginal frequencies only.

1) Length difference for the graph structure: Let us denote $\mathcal{G}^\star = (\mathcal{V}, E^\star)$ and $\mathcal{G}^c = (\mathcal{V}, E^c)$ for the respective graphs of $\mathcal{M}^\star$ and $\mathcal{M}^c$. Eqn. 5 directly gives:

$$I(\mathcal{G}^c) - I(\mathcal{G}^\star) = \log\left(\#maxE - |E^\star|\right) - \log\left(|E^c|\right) \quad (13)$$

This information can be computed in $O(1)$ operations.

2) Length difference for the parameters: Let us denote $\mathcal{C}^\star$ (resp. $\mathcal{S}^\star$) and $\mathcal{C}^c$ (resp. $\mathcal{S}^c$) the sets of maximal cliques (resp. minimal separators) of models $\mathcal{M}^\star$ and $\mathcal{M}^c$. Given Eqn. 7, we have $I(\mathcal{P}^c|\mathcal{G}^c) - I(\mathcal{P}^\star|\mathcal{G}^\star) =$

$$\begin{aligned} &\log(N+1)\cdot\Bigg(\sum_{C^c\in\mathcal{C}^c}\#\mathrm{Param}(C^c) - \sum_{C^\star\in\mathcal{C}^\star}\#\mathrm{Param}(C^\star) \\ &+\sum_{S^\star\in\mathcal{C}^\star}\#\mathrm{Param}(S^\star) - \sum_{S^c\in\mathcal{S}^c}\#\mathrm{Param}(S^c)\Bigg) \qquad (14) \end{aligned}$$

This formula can be extremely simplified. When $\mathcal{M}^c$ replaces $\mathcal{M}^\star$ in the forward selection procedure, they differ by one edge only. The associated graphs thus have close structures and many of the terms expressed in Eqn. 14 cancel out.

Consider the example illustrated in Fig. 1. Because of the similar cliques and separator between $\mathcal{M}^\star$ and $\mathcal{M}^c$, Eqn. 14 will be simplified to:

$$\begin{aligned} &I(\mathcal{P}^c|\mathcal{G}^c) - I(\mathcal{P}^\star|\mathcal{G}^\star) = \log(N+1)\cdot\big(\#\mathrm{Param}(\{1234\}) \\ &+\#\mathrm{Param}(\{23\}) - \#\mathrm{Param}(\{123\}) - \#\mathrm{Param}(\{234\})\big) \end{aligned}$$

This is a direct consequence of graph-theoretical results on chordal graphs: if two decomposable models differ only in one edge $\{a, b\}$, then the maximal cliques and minimal separators differ only in a local sub-structure of the graph, namely around the minimal separator of a and b [13]. Using [13] (Theorem 4.2 and Corollary 4.1), we can formulate the following theorem:

Theorem 1: If two decomposable models $\mathcal{M}^c \subset \mathcal{M}^\star$ differ only in one edge $\{a, b\}$, and let S_{ab} be the minimal separator of $\{a, b\}$, then we have $I(\mathcal{P}^c|\mathcal{G}^c) - I(\mathcal{P}^\star|\mathcal{G}^\star) =$

$$\begin{aligned} &\log(N+1)\cdot\big(\#\mathrm{Param}(S_{ab}\cup\{a,b\}) + \#\mathrm{Param}(S_{ab}) \\ &\quad -\#\mathrm{Param}(S_{ab}\cup\{a\}) - \#\mathrm{Param}(S_{ab}\cup\{b\})\big) \quad (15) \end{aligned}$$

Note that in the example in Fig.1, $S_{14} = \{2, 3\}$.

3) Length difference for the Data: Similarly, and given Eqn. 12, we can formulate the following theorem about assessing the difference in the message length to state $\mathcal{D}$:

Theorem 2: If two decomposable models $\mathcal{M}^c \subset \mathcal{M}^\star$ differ only in one edge $\{a, b\}$, and let S_{ab} be the minimal separator of $\{a, b\}$, then we have $I(\mathcal{D}|\mathcal{M}^c) - I(\mathcal{D}|\mathcal{M}^\star) =$

$$\begin{aligned} &= \sum_{\mathbf{x}\in\mathrm{Dom}(S_{ab}\cup\{a\})} \log\left(O^{S_{ab}\cup\{a\}}_{\mathbf{x}}!\right) + \sum_{\mathbf{x}\in\mathrm{Dom}(S_{ab}\cup\{b\})} \log\left(O^{S_{ab}\cup\{b\}}_{\mathbf{x}}!\right) \\ &- \sum_{\mathbf{x}\in\mathrm{Dom}(S_{ab}\cup\{a,b\})} \log\left(O^{S_{ab}\cup\{a,b\}}_{\mathbf{x}}!\right) - \sum_{\mathbf{x}\in\mathrm{Dom}(S_{ab})} \log\left(O^{S_{ab}}_{\mathbf{x}}!\right) \quad (16) \end{aligned}$$

In summary, we have shown that assessing the replacement of $\mathcal{M}^\star$ by $\mathcal{M}^c$ depends upon sub-scores that are associated to four different cliques only. This extremely reduced expression of message length dramatically improves

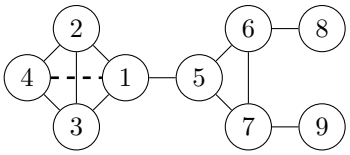


Figure 1. Illustrative example of two decomposable models with nine variables: $\mathcal{M}^\star$ is depicted with strong lines and adding the edge $\{1, 4\}$ (dashed line) results in the model $\mathcal{M}^c$.

the scalability of our approach: the evaluation step only depends on a local graph sub-structure of the models.

C. Efficiently computing sub-scores

As a result of Section III-B, and using Eqn. 3, we can rewrite the function that we are trying to minimize:

$$\begin{aligned} & I(\mathcal{M}^c|\mathcal{D}) - I(\mathcal{M}^\star|\mathcal{D}) \\ = \; & I(\mathcal{G}^c) - I(\mathcal{G}^\star) + I_\mathcal{D}(S_{ab} \cup \{a\}) + I_\mathcal{D}(S_{ab} \cup \{b\}) \\ & - I_\mathcal{D}(S_{ab} \cup \{a,b\}) - I_\mathcal{D}(S_{ab}) \end{aligned} \quad (17)$$

where

$$I_\mathcal{D}(A) = -\log(N+1) \cdot \#\text{Param}(A) + \sum_{\mathbf{x} \in \text{Dom}(A)} \log\left(O_\mathbf{x}^A!\right) \quad (18)$$

Given that we have seen in Eqn. 13 that $I(\mathcal{G}^c) - I(\mathcal{G}^\star)$ is a constant time operation, the scalability of our approach relies on the ability to compute efficiently sub-scores $I_\mathcal{D}(A)$ for different $A \subseteq \mathcal{V}$.

Computing every sub-score once. Let us consider again the example illustrated in Fig. 1. Assessing the replacement of $\mathcal{M}^\star$ by $\mathcal{M}^c$ requires computation of $I_\mathcal{D}(\{123\})$, $I_\mathcal{D}(\{234\})$, $I_\mathcal{D}(\{23\})$ and $I_\mathcal{D}(\{1234\})$. Among these four sub-scores, the first two are sub-scores associated to maximal cliques of $\mathcal{M}^\star$. As a consequence, they were previously computed when $\mathcal{M}^\star$ was a candidate for replacing the former reference model. The sub-score $I_\mathcal{D}(\{23\})$ has also been computed in the process of selecting either $\{123\}$ or $\{234\}$. Clearly, the forward selection procedure exhibits many overlapping sub-problems. As a result, we make Chordalysis-Mml memoize these partial solutions. The replacement of $\mathcal{M}^\star$ by $\mathcal{M}^c$ is then reduced to a function of only one new term, namely $I_\mathcal{D}(\{1234\})$. This compares to the direct calculation of $I(\mathcal{M}^c|\mathcal{D}) - I(\mathcal{M}^\star|\mathcal{D})$ that would require the computation of 20 different sub-scores.

Computing every logarithm and log factorial once. Another case of overlapping sub-problems can be found at a lower level. The evaluation of $I_\mathcal{D}(A)$ for a set of variables $A \subseteq \mathcal{V}$ repeatedly computes logarithms and factorials. It is well-known that the logarithm of a factorial can be computed with $\log(n!) = \sum_{k=1}^{n} \log(k)$. There are however two main issues with this computation: 1) the log(.) function is computationally expensive and 2) for every $I_\mathcal{D}(A)$, the log function is going to be called n times. However, we can observe that $\forall A \subseteq \mathcal{V}, \forall \mathbf{x} \in A, \mathrm{O}_\mathbf{x}^A \leqslant N$. We thus make Chordalysis-Mml pre-compute all log-factorials and store them in an array of size $N+1$ (with $O(1)$ access). Similarly, we pre-compute and store all logarithms up to $\max(N+1, \#maxE)$.

Computing marginal frequencies. The scalability of Chordalysis-Mml now mainly relies on the ability to efficiently compute marginal frequencies $(\mathrm{O}_\mathbf{x}^A, \forall \mathbf{x} \in A$ and for different $A \subseteq \mathcal{V})$. The evaluation of $\mathcal{M}^\star$ vs. $\mathcal{M}^c$, most of the time, will require computing a single sub-score (for a single $A \subseteq \mathcal{V}$). Association discovery between values has carefully studied the efficient computation of marginal frequencies. Chordalysis-Mml uses the Tidsets vertical description of $\mathcal{D}$ with a bitset representation, because it ensures efficient computation of marginal frequencies through the CPU primitive `AND`.

IV. Related research

Researchers have investigated the learning of graphical log-linear models, also named Markov networks or Markov random fields from high-dimensional data.

A first approach consists of building log-linear models on subsets of variables – for which the classical LLA scales up – and then to combine these sub-models [21], [22]. However, because they make strong assumptions about the independence between the variables, they often inaccurately discover associations between variables (see for example Section 5.2 in [22]), and thus do not align with the required high-precision of LLA.

A second approach uses ℓ_1-regularizers. This makes the search possible in high dimensional spaces, by biasing the search towards models for which many parameters are zero. Different configurations have been studied: performing a logistic regression for every variable independently [23], focusing on a reduced subset of features [24] or finding a set of variables that best divides the graph [25]. Because these methods aim mostly at being predictive (as opposite to explanatory), and because they focus on local substructures, they often result in false discoveries (see for example the precision trend depicted in [23] – Section 6) and thus cannot be considered as LLA methods.

A third approach evaluates the trade-off quality/complexity of the models in order to ensure that only associations, for which there is enough evidence, are included in the model. This family is mostly represented by the LLA methods that have been introduced in statistics, and that use of χ^2 goodness-of-fit tests to assess this trade-off on the full model [9]. However, these approaches have progressively lost favor in the last decade due to their exponential complexity with the number of variables, which limited them to datasets with at most a dozen variables. In parallel, researchers have investigated decomposable models: efficient algorithms to compute their maximal cliques and minimal separators [20], to triangulate graphs [12] or to perform hill-climbing search [13]. To the best of our knowledge, the first attempt at performing the full LLA, with the only condition for the graph to be chordal, is from [13], after [10] showed that the Kullback-Leibler divergence can be computed using marginal frequencies that align with the maximal cliques and minimal separators of the graph. The Kullback-Leibler divergence however does not take into account any complexity of the model (other than the regularization). As a result, KL-based methods usually exhibit numerous false discoveries, thus being incompatible with LLA.

To address this issue, two different methods have been proposed. On the one hand, [11] developed an scoring metric for decomposable models that penalizes more complex models (AH-2004). On the other hand, [6] showed with CHORDALYSIS-χ^2 that the statistically established χ^2-framework can be fully applied, while scaling up to datasets with hundreds of variables. Our experiments in the next section show that the CHORDALYSIS-χ^2 is already superior to AH-2004: it has a lower rate of false discovery while being slightly more statistically efficient.

In experimental evaluation in the next section, we show that our CHORDALYSIS-MML method outperforms the state-of-the-art techniques: no false discovery has ever been observed, while CHORDALYSIS-MML is shown to require fewer samples than the state-of-the-art methods to discover true associations.

V. EXPERIMENTS

A. Datasets from known models

Assessing the quality of LLA requires having knowledge about the multi-way interactions that take place in data. Therefore we start by evaluating the discovery with data that is sampled from known distributions (sets of interactions and associated probability tables). We can then compare the discovered interactions to the true structure from which the data was sampled.

The structure of a graphical model is completely determined by its pairwise interactions. We can thus assess the recovery of the structure in terms of the edges of the graph. Each possible edge in the graph can be present or absent in the true model and each true edge can also have been discovered or not. This corresponds to the standard scheme true/false positive/negative, to which are associated the usual precision $\mathcal{P}$, recall $\mathcal{R}$ and F-measure $\mathcal{F} = 2 \cdot \frac{\mathcal{P} \cdot \mathcal{R}}{\mathcal{P}+\mathcal{R}}$. Two main criteria are used to assess LLA methods: false discovery rate (FDR), $(1-\mathcal{P})$, and statistical efficiency. We thus report the FDR and the F-measure (to quantify the recovery of the graph structure with increasing number of samples), as well as the execution time.

We compare CHORDALYSIS-MML to all state-of-the-art methods: KL-2001 [13], AH-2004 [11], and CHORDALYSIS-χ^2 [6]. Note, we implemented all four methods as modifications to the original CHORDALYSIS-χ^2 software that we previously released [6], thus making the execution times truly comparable.

Data structures: We designed three different models from which the data was sampled. This allows comparison of the behavior of all the methods in terms of false discovery, statistical efficiency and runtime, while controlling what the *true* model is. We designed:

- $\mathcal{D}_1$ to have three independent variables only, in order to verify that no method is finding a correlation when none have to be discovered (Fig. 2-(a)).
- $\mathcal{D}_2$ to have three 3-way and three 2-way interactions, in order to test the ability of the different methods to recover conditional dependencies on distributions over 9 variables (Fig. 2-(b)).
- $\mathcal{D}_3$ to finely compare all the methods on a complex distribution over 150 variables. $\mathcal{D}_3$ comprises 150 variables and includes 24 5-way (in three interlaced groups of eight 5-ways), three 4-way, two 3-way and three 2-way interactions as well as 55 independent variables (Fig. 2-(c)).

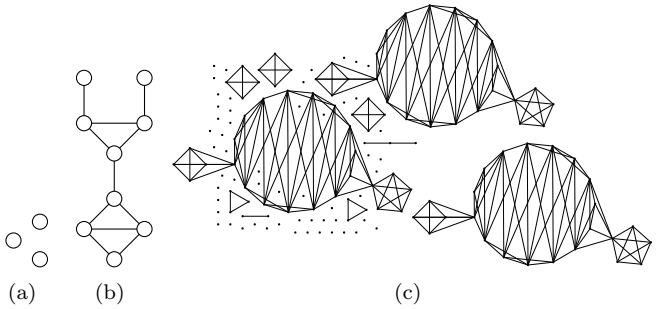


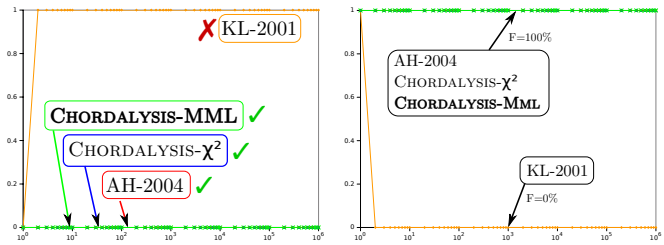
Figure 2. Data structures that are randomly sampled. (a) $\mathcal{D}_1$. (b) $\mathcal{D}_2$. (c) $\mathcal{D}_3$.

Configuration note: To handle high-dimensional data with KL-2001, it is necessary to set the maximum clique size (treewidth) to 5 for $\mathcal{D}_3$. Note that this is actually "helping" the method. Following recommendations in statistics regarding the p-value threshold [26], we set $p = 0.001$ for CHORDALYSIS-χ^2.

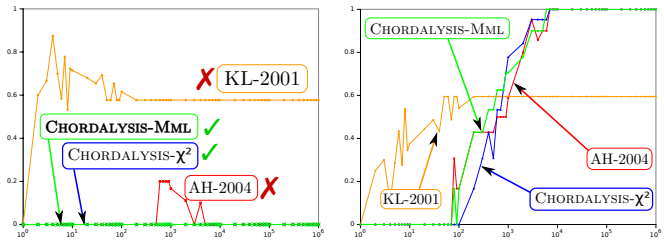
AH-2004's encoding: We will see that AH-2004 exhibits a very different behavior to CHORDALYSIS-MML. This is due to three main elements of the encoding that AH-2004 uses. First, it over-estimates the number of parameters of the model by undercounting for the overlap between different cliques (as opposed to Sec. III-A1). Second, it uses an inefficient encoding of the data given the model $N \cdot H(\mathcal{M})$ (see Sec. III-A3). Third, it encodes the graph by favoring complex structure and largely over-estimates the lengths of edges (far more than one bit per edge).

1) Sanity check – Results for $\mathcal{D}_1$ and $\mathcal{D}_2$: Dataset $\mathcal{D}_1$ was used to test if the methods include correlations in the model when there are actually no correlations at all between the variables. The results are depicted on the left side of Fig. 3(a). CHORDALYSIS-χ^2, AH-2004 and the proposed CHORDALYSIS-MML behave consistently and never discover any correlation (0% of false discoveries and 100% of the structure recovered). KL-2001, on the other hand, quickly retrieves the model with all possible edges (saturated model), leading to 100% of false discoveries. This "disqualification" of the KL-based method is consistent with classical results (see [6] for example), because KL-based methods optimize the entropy of the model without taking into account the complexity of the model.

Dataset $\mathcal{D}_2$ was used to test the discovery of simple conditional dependencies/independencies. The results are depicted on the right side of Fig. 3(b). Similarly to the previous results, KL-2001 rapidly (and incorrectly) converges to the saturated model, while CHORDALYSIS-χ^2 and CHORDALYSIS-MML behave consistently: 1) they never discover any nonexistent correlation (0% of false



(a) $\mathcal{D}_1$: FDR (left) and Graph recovery ($\mathcal{F}$ – right) vs. the number of samples



(b) $\mathcal{D}_2$: FDR (left) and Graph recovery ($\mathcal{F}$ – right) vs. the number of samples

Figure 3. Results of the experiments on $\mathcal{D}_1$ and $\mathcal{D}_2$.

discoveries) and 2) they recover the complete graph structure (with similar statistical efficiency).[3] However, on this slightly more complicated structure, AH-2004 exhibits a number of false discoveries. Due to the inefficient encoding elements described above, AH-2004 overfits when only a few thousand samples are available as evidence of the correlations in the graph.

Note that CHORDALYSIS-χ^2, AH-2004 and CHORDALYSIS-MML selected the model in less than 100 ms, regardless of the number of samples.

2) High-dimensional experiment – Results for $\mathcal{D}_3$: $\mathcal{D}_3$ allows us to perform full comparison of the methods. It is a high-dimensional dataset with 150 variables that is drawn from a complex distribution with multi-way correlations between the variables (Fig. 2-(c)). This makes the discovery difficult, because there is very little evidence of high-order correlation from lower-order ones in data. This graph structure (and associated dataset) are thus a good test bed to compare the methods at a large scale.

Fig. 4(a) illustrates the false discovery rate on this dataset. KL-2001 is confirmed as not suitable for LLA by retrieving incorrect correlations more than half of the time. The overfitting behavior of AH-2004 is also confirmed: it exhibits about 10% of false discoveries, which makes it unsuitable for most uses of LLA. LLA is indeed often used for example in medicine to decide upon the conditional dependencies/independencies of different medical conditions and treatments. Such a high false discovery rate is clearly incompatible with the validation, for instance, of the influence of a drug on a disease. The stability and consistency of CHORDALYSIS-χ^2 and CHORDALYSIS-MML are confirmed with 0% of false discovery, regardless of the number of samples that is used.

Fig. 4(b) illustrates the statistical efficiency on $\mathcal{D}_3$ and confirms that CHORDALYSIS-MML outperforms the state-of-the-art methods:

1. CHORDALYSIS-MML is the only method that recovers the full structure of the graph (CHORDALYSIS-χ^2 misses 10 edges while AH-2004 misses 20 edges and includes 21 incorrect edges).

2. CHORDALYSIS-MML is slightly more statistically efficient: it requires fewer samples to reach the same recovery-rate of the graph structure. For example, with 100,000 samples CHORDALYSIS-MML recovers 66% of the graph while CHORDALYSIS-χ^2 recovers 49% only. In addition, CHORDALYSIS-MML requires only 150,000 samples to recover 90% of the graph structure; CHORDALYSIS-χ^2 requires more than half a million samples to reach the same level of quality.

In addition, note that CHORDALYSIS-MML is very stable: the quality of the recovery does not significantly oscillate. This is a strong evidence of the quality and stability of our approach, because once CHORDALYSIS-MML decides upon the presence of a correlation, increasing the amount of evidence does not challenge the decision. This supports the statistical power of the decision taken by CHORDALYSIS-MML being sound and consistent.

Finally, we consider the computational efficiency. Fig. 4(c) shows that both Chordalysis approaches are faster than AH-2004. This is mainly due to the fact that the latter does not have any edge-optimized scoring, and performs additional operations on the clique graph (*e.g.*, spanning tree, topological sort) that are not used in CHORDALYSIS-MML (note that we use the memoization strategies from Sec. III-C for all four methods). The computation time for this high-dimensional dataset is stable at around 10 min while slowly increasing with the number of samples. The slight difference between CHORDALYSIS-χ^2 and CHORDALYSIS-MML is due to the statistical efficiency of CHORDALYSIS-MML: as it discovers more of the true correlations with fewer samples, it actually explores more of the search space, which takes more time.

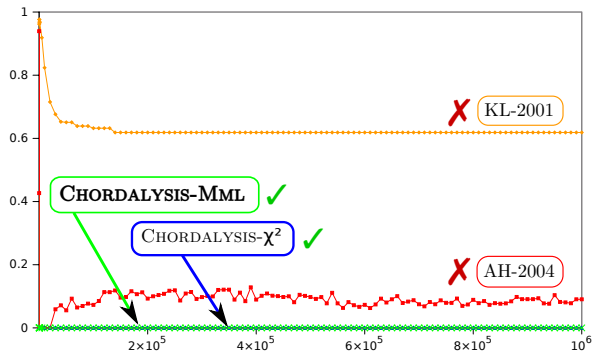
B. Results on a real dataset

To demonstrate real-world performance we compare the results of CHORDALYSIS-χ^2 and of CHORDALYSIS-MML to a dataset from an epidemiological study of the elderly (EPESE) [27].[4] It is important to note that only qualitative analysis is possible on real data, because there is no "ground truth". This is why we have conducted the previous experiments on datasets for which we can control the structure from which the data has been generated.

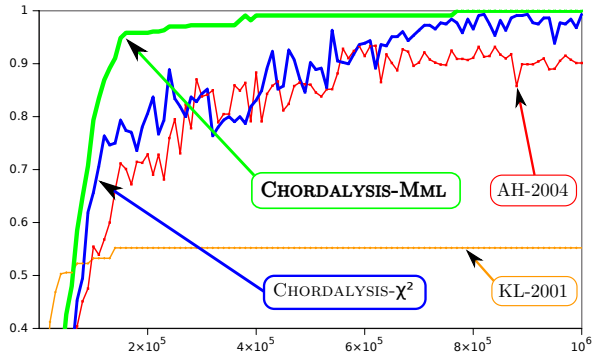
The resulting model (selected in less than 4 s) is shown in Fig. 5. Expert assessment of this dataset is provided

[3]Note that KL-2001 has a higher $\mathcal{F}$-measure with fewer samples, because an overfitting model will also find actual correlations, since it will eventually include all of them.

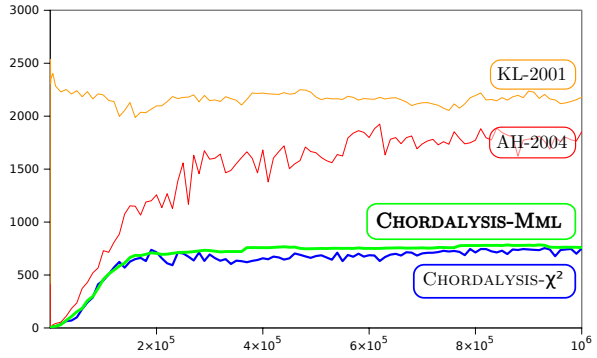
[4]We have shown in the previous section that the results of KL-2001 and AH-2004 do not meet the standards of log-linear analysis. However, for completeness, the reader can find the corresponding results at http://www.tiny-clues.eu/Research/ICDM2014-MML/.



(a) $\mathcal{D}_3$: FDR vs. the number of samples



(b) $\mathcal{D}_3$: **Graph recovery** ($\mathcal{F}$) vs. the no. of samples



(c) $\mathcal{D}_3$ – Execution time in s.

Figure 4. Results of the experiments on $\mathcal{D}_3$.

in [28]. Many of the multi-way relationships retrieved by both methods have supporting evidence, such as the links between blood pressure variables and the use of medication to treat high-blood pressure, and the correlations between pain in the chest or shortness of breath and heart attack. Due to limited space, we now focus on the main differences between the two results. The correlations that are retrieved by CHORDALYSIS-MML are generally more sensible. For example, CHORDALYSIS-MML linked `Insulin` to `Diabetes` rather than to the ability to walk a mile (`WalkMile`); taking medication for high blood pressure is linked to the fact of having a high blood pressure; `Smoking` is linked to the fact of having ever smoked (`EverSmoked`) rather than to being married (`Married`); having been married is linked to being married rather than to being retired, etc.

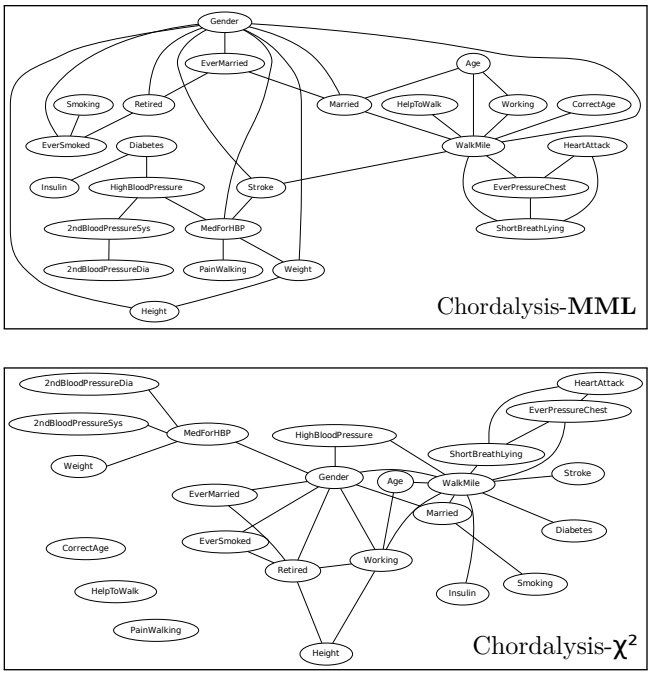


Figure 5. Models selected for the EPESE dataset.

CHORDALYSIS-χ^2 also detected no correlation between the need for help to walk (`HelpToWalk`) and any other variable. In contrast, CHORDALYSIS-MML quite logically linked it with the ability for the patient to walk for a mile by him- or her-self. More subtly, CHORDALYSIS-χ^2 detected no correlation including `CorrectAge` while CHORDALYSIS-MML has associated it to the ability for a patient to walk a mile by him- or her-self. Knowing that correctness of the declared age is indicative of the patient's mental health, this correlation could be explained by the loss of mobility that is often observed with the late stages of dementia. More surprisingly, CHORDALYSIS-MML linked `PainWalking` with taking medication for high blood pressure. This might however be explained by high blood pressure being commonly treated with diuretics, which often decrease the body's levels of potassium, leading specifically to a high-probability of leg cramps.

VI. CONCLUSIONS

In this paper, we proposed an information-theoretic approach to log-linear analysis (LLA), that is based on the Minimum Message Length principle. Our experiments have shown that our method never committed any false discovery and requires fewer samples to reach the same quality as state-of-the-art methods. Moreover, our theorems for decomposable models, melded with advanced data mining techniques and results from graph theory, allow our method to scale up to datasets with more than a hundred variables on a standard desktop computer. Our contributions to association discovery between variables include:

1) A new scoring for decomposable models that results in the best LLA method so far.

2) Because our statistic applies to any set of frequencies, it makes LLA possible where classical χ^2 tests cannot be used, *i.e.*, when the MLEs do not exist or when co-occurrence matrices exhibit small frequencies. This is a major theoretical property that extends the frontiers of the applicability of LLA.

3) Proof that our statistic can be expressed in terms of the graph structure of the model.

4) Proof that our statistic for comparing two decomposable models that differ by a single edge can be calculated using a function of four marginal entropies only.

5) Efficient techniques for computing our statistic using the above proofs and techniques developed for itemset mining, as well as memoization of marginal entropies, ensuring that marginal entropies are only computed once.

One limitation of our approach is that the search only considers additions to a model that involve adding a single edge that results in a decomposable graph. It may thus fail finding associations between variables where the addition of an edge would result in a non-chordal graph. It would be valuable to explore techniques that can either step through graphs that are not decomposable, or can consider steps that involve addition of multiple edges, specifically, an edge of interest and the additional edges required to triangulate the resulting graph. This is a difficult problem because there can be many ways to triangulate a single graph and there is no obvious efficient way to select one from the many [12]. Additionally, contrary to classical approaches in statistics that rather assess the modification of two nested models, Chordalysis-Mml can assess models in isolation. This property opens the way to randomized search (*e.g.*, simulated annealing), which should improve the quality of models selected by LLA procedures.

Association discovery is a fundamental data mining task. We believe that we have significantly improved the discovery of trustworthy associations between variables in high-dimensional data, and hope that this will prove to be a powerful addition to the data mining toolbox.

VII. Acknowledgments

This research has been supported by the Australian Research Council under grant DP120100553. The authors would like to thank Arun Konagurthu, Jilles Vreeken and Ann E. Nicholson for their fruitful comments on this work and for their review of the manuscript.

References

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Int. Conf. on Very Large Data Bases*, 1994, pp. 487–499.

[2] G. I. Webb, "Discovering significant patterns," *Machine Learning*, vol. 68, no. 1, pp. 1–33, 2007.

[3] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in *Int. Conf. on Knowledge Discovery and Data Mining*, 1997, pp. 283–286.

[4] H. Heikinheimo, J. K. Seppänen, E. Hinkkanen, H. Mannila, and T. Mielikäinen, "Finding low-entropy sets and trees from binary data," in *Int. Conf. on Knowledge Discovery and Data Mining*, 2007, pp. 350–359.

[5] M. Mampaey and J. Vreeken, "Summarizing categorical data by clustering attributes," *Data Mining and Knowledge Discovery*, vol. 26, no. 1, pp. 130–173, 2013.

[6] F. Petitjean, G. I. Webb, and A. E. Nicholson, "Scaling log-linear analysis to high-dimensional data," in *IEEE Int. Conf. on Data Mining*, 2013, pp. 597–606.

[7] S. J. Haberman, *The analysis of frequency data*. University of Chicago Press, 1974.

[8] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, ser. Information Science and Statistics. Springer-Verlag, 2005.

[9] R. Christensen, *Log-Linear Models and Logistic Regression Second Edition*. Springer, 1997.

[10] F. Malvestuto, "Approximating discrete probability distributions with decomposable models," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, no. 5, pp. 1287–1294, 1991.

[11] S. Altmueller and R. M. Haralick, "Approximating high dimensional probability distributions," in *IEEE Int. Conf. on Pattern Recognition*, 2004, pp. 299–302.

[12] P. Heggernes, "Minimal triangulations of graphs: A survey," *Discrete Mathematics*, vol. 306, no. 3, pp. 297–317, 2006.

[13] A. Deshpande, M. Garofalakis, and M. I. Jordan, "Efficient stepwise selection in decomposable models," in *Uncertainty in Artificial Intel.*, 2001, pp. 128–135.

[14] S. E. Fienberg and A. Rinaldo, "Maximum likelihood estimation in log-linear models," *Annals of Statistics*, vol. 40, no. 2, pp. 996–1023, 2012.

[15] J. H. McDonald, *Handbook of Biological Statistics*. Sparky House Pub., 2009, ch. Tests for nominal variables, pp. 80–83.

[16] C. S. Wallace and D. M. Boulton, "An information measure for classification," *The Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.

[17] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[18] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problemy Peredachi Informatsii*, vol. 1, no. 1, pp. 3–11, 1965.

[19] D. Boulton and C. Wallace, "The information content of a multistate distribution," *Journal of Theoretical Biology*, vol. 23, no. 2, pp. 269–278, 1969.

[20] A. Berry, A. Sigayret, and C. Sinoquet, "Maximal subtriangulation in pre-processing phylogenetic data," *Soft Computing*, vol. 10, no. 5, pp. 461–468, 2006.

[21] X. Wu, D. Barbará, and Y. Ye, "Screening and interpreting multi-item associations based on log-linear modeling," in *Int. Conf. on Knowledge Disc. and Data Mining*, 2003, pp. 276–285.

[22] C. Dahinden, M. Kalisch, and P. Bühlmann, "Decomposition and model selection for large contingency tables," *Biometrical Journal*, vol. 52, no. 2, pp. 233–252, 2010.

[23] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty, "High-dimensional graphical model selection using ℓ_1-regularized logistic regression," in *Advances in Neural Information Processing Systems*, 2007, pp. 1465–1472.

[24] S.-I. Lee, V. Ganapathi, and D. Koller, "Efficient Structure Learning of Markov Networks using ℓ_1-Regularization," in *Advances in neural information processing systems*, 2006, pp. 817–824.

[25] V. Gogate, W. A. Webb, and P. Domingos, "Learning Efficient Markov Networks," in *Advances in Neural Information Processing Systems*, 2010, pp. 748–756.

[26] V. E. Johnson, "Revised standards for statistical evidence," *Proc. of the National Academy of Sciences of the USA*, vol. 110, no. 48, pp. 19 313–19 317, 2013.

[27] J. Taylor, R. Wallace, A. Ostfeld, and D. Blazer, "Established Populations for Epidemiologic Studies of the Elderly, 1981-1993," 1998, http://dx.doi.org/10.3886/ICPSR09915.

[28] J. Flores, A. E. Nicholson, A. Brunskill, K. B. Korb, and S. Mascaro, "Incorporating expert knowledge when learning Bayesian network structure," *Artificial Intelligence in Medicine*, vol. 53, no. 3, pp. 181–204, 2011.