Scaling log-linear analysis to high-dimensional data

François Petitjean, Geoffrey I. Webb and Ann E. Nicholson
Faculty of Information Technology
Monash University, Melbourne, Australia
Email: {francois.petitjean,geoff.webb,ann.nicholson}@monash.edu

***Abstract*—Association discovery is a fundamental data mining task. The primary statistical approach to association discovery between variables is log-linear analysis. Classical approaches to log-linear analysis do not scale beyond about ten variables. We develop an efficient approach to log-linear analysis that scales to hundreds of variables by melding the classical statistical machinery of log-linear analysis with advanced data mining techniques from association discovery and graphical modeling.**

***Keywords*-Association Discovery; Data Modeling; High-dimensional Data; Log-linear Analysis**

I. Introduction

Log-linear analysis is a well established statistical technique for finding associations between variables in data [1]. In contrast, data mining research into association discovery has focused primarily on finding associations between variable-values or items [2]. Each of these approaches, finding associations between variables or find associations between variable-values, has distinct contexts in which it is more useful. Sometimes the focus will be 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.

Classical approaches to log-linear analysis are exponential with respect to the number of variables, as they calculate the frequency for all combinations of values. For M binary variables, this requires 2^M operations. This is not feasible for high M. In practice, classical techniques are limited to a dozen variables at most.

A number of researchers have investigated approaches that build log-linear models on subsets of the variables and combine them [3], [4], [5], [6]. However, these are unable to recover complex high-order interactions.

In this paper, we demonstrate that, by focusing on a powerful sub-class of log-linear models – namely, decomposable models – and taking advantage of theory developed in data mining for association discovery, log-linear analysis can be scaled to high-dimensional data, with no further restrictions. In particular, we prove that for decomposable models, χ^2 tests can be computed in a significantly reduced number of marginal contingency tables.

This paper is organized as follows. In Section II, we formalize the problem. In Section III, we present our solution *Chordalysis*, which enables the discovery of statistically sound multi-way interactions between variables for hundred-dimensional datasets. In section IV, we present related research. In Section V, we conduct experiments that demonstrate 1) the relevance of our approach on real-world high-dimensional datasets; 2) its scalability; and 3) the importance of statistical control for association discovery between variables. Finally, we conclude and describe future research in Section VI.

II. Problem statement

A. Definitions: log-linear models and analysis

Log-linear models: 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 $\hat{p}_\mathcal{V}$:

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

where $\text{O}_\mathbf{x}$ 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. With M variables, the model for a vector of values $\mathbf{x}$, the expected frequency $m_\mathbf{x}$ has the form:

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

The $u_{\dots}$ functions represent the interactions between variables that are used to model the observed frequencies. Terms are removed from the model by setting them to zero. For example, a model that includes the term $u_{i,j}$ will consider the interaction between the i^{th} and the j^{th}

To be published in Proceedings of the IEEE International Conference on Data Mining, 2013. Copyright (c) IEEE 2013.

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 doesn't include $u_{i,j,k}$ (having set it to zero) will not consider the corresponding three-way interaction. A model that includes all terms is called a *saturated model*.

Notation: Log-linear models are often represented using the highest-order interactions that they include. For example, `[ABC][CD]` represents a model among four variables, which includes a three-way interaction $u_{A,B,C}$, as well as a two-way interaction $u_{C,D}$. Moreover, this notation implies that all the lower-order interactions are also included, *i.e.*, $u_{A,B}$, $u_{A,C}$, $u_{B,C}$, as well as u_A, u_B, u_C, u_D and u. Note that the log-linear models that are completely determined by these highest-order interactions are called *hierarchical*.

Log-linear analysis: Log-linear analysis is the general name given to methods that seek to select a consistent log-linear model from data. This corresponds to determining which $u_{...}$ terms can be removed from the general form, without loosing too much "predictive power". The saturated model will always fit $\mathcal{D}$ perfectly. Log-linear analysis methods add or remove terms to an initial model, for as long as the quality/complexity trade-off improves. Classical methods address this trade-off with χ^2 goodness-of-fit tests.

B. Forward selection

Log-linear analysis methods are divided into two families: *backward elimination* and *forward selection*. Both methods iterate so long as the quality/complexity improves. Backward elimination starts with a complex model (usually saturated) and iteratively removes terms. Forward selection starts with a simple model (usually independence between the variables) and iteratively adds terms. The saturated model for 100 variables includes $\sum_{k=0}^{100}\binom{100}{k} = 2^{100} > 10^{30}$ different terms. As it is infeasible to iterate through such a set, and without loss of generality with regard to our theoretical results, we focus on *forward selection*.

Forward selection methods proceed as follows:

1) Start with an initial model $\mathcal{M}^\star$.
2) Consider the set of terms that can be added to $\mathcal{M}^\star$.
3) Select from that set of terms the one that produces the candidate model $\mathcal{M}^c$ with the best statistical significance to replace $\mathcal{M}^\star$.
    - if the associated statistical significance achieves a predetermined minimum level, replace $\mathcal{M}^\star$ by $\mathcal{M}^c$ and loop to step 2.
    - else the procedure terminates and returns $\mathcal{M}^\star$.

The structure of the resulting model (*i.e.*, the interactions that are considered) represents the multi-way associations – between the variables – that are statistically significant in the dataset.

This paper shows how to scale up this procedure to datasets with hundreds of variables.

C. Obstacles for high-dimensional datasets

Existing approaches to forward selection log-linear analysis do not scale up to more than about ten variables, because they include several sub-processes that are exponential with respect to the number of variables:

1. Evaluation: preferring one model over another is usually assessed through χ^2 goodness-of-fit tests [7, pp. 94–98]. These tests require the comparison of observed frequencies to the ones predicted by the model for all possible combinations of values, a number that is exponential with the number of variables. While it can be feasible to perform an exponential number of operations for ten binary variables ($\#\textit{operations} \geqslant 2^{10} = 1024$), it becomes infeasible when the number of variables is high (*e.g.*, for 100 variables, $\#\textit{operations} \geqslant 2^{100} > 10^{30}$). 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. Thus, the number of operations prevents log-linear analysis from being scaled up to high-dimensional data.

2. Fitting the model – MLE: Let us assume that a model is considered to represent a better quality/complexity compromise by the log-linear analysis process (*i.e.*, that the previous problem of evaluation is solved). To assess the model, its parameters first have to be estimated. Classically, this procedure is performed *via* a Newton-Raphson procedure [7, pp. 346–347]. Considering high-order interactions will exponentially increase the size of the considered contingency tables. As a consequence, the probability that contingency tables contain zeros will increase exponentially (since the size of the dataset remains constant). Zeros in the contingency tables often lead to non-existing maximum likelihood estimates (MLEs) [1], making it impossible to fit, and thus to evaluate the model. See the work by [8] for an up-to-date overview of this problem.

III. Chordalysis

Our method for association discovery between variables is presented in this section. By focusing on a particular sub-class of log-linear models – *decomposable* or *multiplicative* models – the full log-linear analysis paradigm can be extended to high-dimensional data. We start by briefly presenting decomposable models, outline our solution, then describe in detail the four features of the method.

A. Decomposable models

Definition 1: [7] 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, a *graphical* model can be represented by an undirected graph, where the vertices represent the variables and the edges represent the two-factor terms included in the model.

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

Decomposable models are the log-linear models that have closed-form maximum likelihood estimates (MLEs) [1]. Note that not only do decomposable models have closed-form MLE, but that all the log-linear models that have closed-form MLE are decomposable. This subclass is not only practical but also a sound 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 general proof of this existence comes from the fact that the saturated model, which subsumes any other log-linear model, is decomposable. In practice, for a graphical model that is not decomposable, any minimal triangulation of its graph [9] will provide such a decomposable model.

The model's probability p_μ can be evaluated using marginal probabilities of a small number of terms. In contrast, non-decomposable models can only be fitted to the data if the dataset is low-dimensional [8] (no more than a dozen variables). The expression of p_μ is linked to the graph representation of the model, and relies on the maximal *cliques*, $\mathcal{C}$, and minimal *separators*, $\mathcal{S}$, of the graph. The probability p_μ of a vector $\mathbf{x}$ under a decomposable model $\mathcal{M}$ is then expressed by:

$$p_\mu(\mathbf{x}) \quad = \quad \frac{\displaystyle\prod_{C\in\mathcal{C}} p_C(\mathbf{x})}{\displaystyle\prod_{S\in\mathcal{S}} p_S(\mathbf{x})} \qquad (1)$$

with p_A representing the marginal probability of $\hat{p}_\mathcal{V}$ over a set of variables A. Note that $\mathcal{S}$ is a multi-set of sets of variables, since a minimal separator can separate several maximal cliques.

The definition and efficient computation of $\mathcal{C}$ and $\mathcal{S}$ are presented in App. A. Note that their computation requires only $O(|\mathcal{V}|+|E|)$ operations for chordal graphs.

B. Sketch of our solution

As we show below, decomposable models are closely linked to *chordal* graphs, which gives the name to our method *Chordalysis*: a scalable method for association discovery between variables. Chordalysis is a forward selection approach among decomposable models. The obstacles described above are addressed with the following four main features.

1. Searching among decomposable models: Decomposable models are intrinsically linked to *chordal* graphs. We show how to take advantage of recent discoveries in this field to perform the forward selection procedure among decomposable models.

2. Rewriting the G^2 statistic: Taking advantage of the closed-form MLEs, we show that a χ^2 goodness-of-fit test can be expressed without requiring an exponential number of operations in terms of the number of variables. In fact, we prove that the likelihood ratio statistic (G^2) for decomposable models can be expressed in terms of the graph structure associated with the model.

3. Efficient marginal entropies computation: We show that the computation of the rewritten G^2 statistic relies completely on the computation of marginal frequencies for different combinations of values. We can thus take advantage of the intersection-optimized data structures that have been developed for itemset mining (*e.g.*, Tidsets [10]). In practice, forward selection procedures can be seen as a breadth-first search. Our solution combines a data structure based on a partial lattice, and memoization of intermediate solutions.

4. Layered critical values: When statistical tests are used repeatedly, to control the familywise error rate, the significance threshold must be corrected. We propose a correction scheme based on [11], that assigns different critical values to different areas of the search space.

The conjunction of these features makes it possible to scale log-linear analysis to hundreds of variables on a standard desktop computer.

C. Searching among decomposable models

As described above, at each iteration, forward selection methods generate a set of more complex models, by adding a single term per candidate model that is not contained in the current best model. Decomposable models are completely defined by their two-way interactions (or two-factor terms). As a result, the generation of candidate alternatives to a current model involves the addition of a two-factor term. In order to ensure that the candidate graphs remain decomposable, it is necessary to consider only terms that result in a decomposable model, *i.e.*, for which its graph is chordal.

We need to determine whether adding an edge $\{a,b\}$ to a chordal graph $\mathcal{G} = (\mathcal{V}, E)$ will result in a further chordal graph. If a and b belong to different connected components of $\mathcal{G}$ (*i.e.*, if there is no path from a to b), then the graph $(\mathcal{V}, E \cup \{\{a,b\}\})$ will always be chordal. However, if a and b belong to the same connected component (*i.e.*, there is a path from a to b), the problem involves the concept of a 2-pair [12].

Definition 3: Given a chordal graph $G_1 = (\mathcal{V}, E)$, a pair $\{a, b\}$ of non-adjacent vertices is called a 2-pair iff the graph $G_2 = (\mathcal{V}, E \cup \{\{a, b\}\})$ is chordal.

Recall that a chordal graph contains no chordless cycle longer than three (triangles). The intuition behind the above definition is then that the only cycles that will be closed by adding the edge $\{a, b\}$ are cycles of length three, which keeps the graph chordal. Because the cycles created by joining a to b all are of length three, any chordless paths between a and b are of length two, hence the name 2-pair.

In this way, every time a new model is chosen to replace the previous best one (*i.e.*, at each iteration of the forward selection method), we build a set of eligible interactions associated with the new model. Therefore, we select all the pairs of variables for which all the chordless paths from the one to the other are of length two. Then, one new candidate model will be constructed for every eligible interaction.

D. Rewriting the G^2 statistic

In this section, we show that using the likelihood ratio test statistic (G^2), we can exactly express the fit of a model to $\mathcal{D}$, using marginal probabilities only. Let us recall that the G^2 statistic for a model $\mathcal{M}$ with df number of degrees of freedom approximates a $\chi^2(df)$ distribution for large samples. First, we develop and simplify the G^2 statistic for the evaluation of one model. Then, we express, develop and simplify the G^2 statistic for the replacement of $\mathcal{M}^\star$ by $\mathcal{M}^c$.

1) G^2 of one model: We will use $\mathrm{O}^A_{\mathbf{x}}$ (resp. $\mathrm{E}^A_{\mathbf{x}}$) to designate the observed (resp. expected from a model $\mathcal{M}$) frequencies for the configuration $\mathbf{x}$ with respect to the set of variables A, and $H(.)$ to denote the entropy.

$$\begin{aligned}\mathrm{G}^2(\mathcal{M}) &= 2 \cdot \sum_{\mathbf{x} \in \mathrm{Dom}(\mathcal{V})} \mathrm{O}^{\mathcal{V}}_{\mathbf{x}} \cdot \ln\left(\frac{\mathrm{O}^{\mathcal{V}}_{\mathbf{x}}}{\mathrm{E}^{\mathcal{V}}_{\mathbf{x}}}\right) \\ &= 2 \cdot N \left(\sum_{\mathbf{x} \in \mathrm{Dom}(\mathcal{V})} \hat{p}_{\mathcal{V}}(\mathbf{x}) \ln \hat{p}_{\mathcal{V}}(\mathbf{x}) - \sum_{\mathbf{x} \in \mathrm{Dom}(\mathcal{V})} \hat{p}_{\mathcal{V}}(\mathbf{x}) \ln p_{\mu}(\mathbf{x}) \right)\end{aligned} \tag{2}$$

In addition, the first term within the brackets in the last equation corresponds to $-\mathrm{H}(\mathcal{V})$. Moreover, [13] showed that, for decomposable models, the second term can be simplified in the following way:

$$-\sum_{\mathbf{x} \in \mathrm{Dom}(\mathcal{V})} \hat{p}_{\mathcal{V}}(\mathbf{x}) \ln p_{\mu}(\mathbf{x}) = \sum_{C \in \mathcal{C}} \mathrm{H}(C) - \sum_{S \in \mathcal{S}} \mathrm{H}(S) \tag{3}$$

Replacing in Eqn. 2, we now have:

$$\mathrm{G}^2(\mathcal{M}) = 2 \cdot N \left(\sum_{C \in \mathcal{C}} \mathrm{H}(C) - \sum_{S \in \mathcal{S}} \mathrm{H}(S) - \mathrm{H}(\mathcal{V}) \right) \tag{4}$$

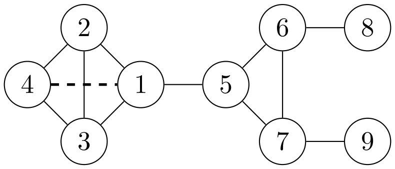


Figure 1. Illustrative example of two decomposable models with nine variables: $\mathcal{M}^\star = [123][234][15][567][68][79]$ is depicted with strong lines and adding the edge $\{1, 4\}$ (dashed line) results in the model $\mathcal{M}^c = [1234][15][567][68][79]$.

2) $\mathrm{G}^2(\mathcal{M}^\star$ vs. $\mathcal{M}^c)$: Given a reference model $\mathcal{M}^\star$ and a (strictly larger) candidate model $\mathcal{M}^c$, MLEs for log-linear models satisfy[7, pp. 96–97]:

$$\mathrm{G}^2(\mathcal{M}^\star \text{ vs. } \mathcal{M}^c) = \mathrm{G}^2(\mathcal{M}^\star) - \mathrm{G}^2(\mathcal{M}^c) \tag{5}$$

Then, given Eqn. 4, it is simple to show that to test $\mathcal{M}^\star$ against $\mathcal{M}^c$, where $\mathcal{M}^c$ is strictly larger than $\mathcal{M}^\star$, the entropy of the dataset $\mathrm{H}(\mathcal{V})$ disappears:

$$\begin{aligned}\mathrm{G}^2(\mathcal{M}^\star \text{ vs. } \mathcal{M}^c) &= 2 \cdot N \left(\sum_{C \in \mathcal{C}^\star} \mathrm{H}(C) - \sum_{S \in \mathcal{S}^\star} \mathrm{H}(S) \right. \\ &\qquad \left. - \sum_{C \in \mathcal{C}^c} \mathrm{H}(C) + \sum_{S \in \mathcal{S}^c} \mathrm{H}(S) \right)\end{aligned} \tag{6}$$

Also, when $\mathcal{M}^c$ replaces $\mathcal{M}^\star$ in the forward selection procedure, they differ by one edge only. Thus, the associated graphs will have very close structures and many entropies expressed in Eqn. 6 cancel each other out.

Consider the example illustrated in Fig. 1: let $\mathcal{M}^\star = [123][234][15][567][68][79]$ and the additional interaction under consideration be $\{1, 4\}$. The candidate model is thus $\mathcal{M}^c = [1234][15][567][68][79]$. Computing Eqn. 6, because of the similar cliques and separator between $\mathcal{M}^\star$ and $\mathcal{M}^c$, many terms cancel out, leading to $\mathrm{G}^2(\mathcal{M}^\star \text{ vs. } \mathcal{M}^c) = 2 \cdot N\big(\mathrm{H}(\{123\}) + \mathrm{H}(\{234\}) - \mathrm{H}(\{1234\}) - \mathrm{H}(\{23\})\big)$.

This result is a 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[14]. Using [14, 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:

$$\begin{aligned}\mathrm{G}^2(\mathcal{M}^c \text{ vs. } \mathcal{M}^\star) &= 2 \cdot N\big(\mathrm{H}(S_{ab} \cup \{a\}) + \\ &\quad \mathrm{H}(S_{ab} \cup \{b\}) - \mathrm{H}(S_{ab} \cup \{a, b\}) - \mathrm{H}(S_{ab})\big)\end{aligned} \tag{7}$$

Note that in the previous example, we had $S_{14} = \{23\}$.

Then, assessing the statistical significance of the replacement of $\mathcal{M}^\star$ by $\mathcal{M}^c$ is a function of only four different entropies. This extremely reduced expression of the G^2 statistic improves dramatically the scalability of our approach. The evaluation step is no longer exponential in terms of the number of variables, but only depends on a local graph sub-structure of the models.

E. Efficient marginal entropies computation

Evaluating the replacement of one model by another relies on the computation of marginal entropies (Eqn. 7). Chordalysis includes three main optimizations of this computation. First, we ensure that every marginal entropy is only computed once. Second, we show that all the possible marginal entropies are sums of a limited number of partial entropies. Third, we propose a data structure, based on a partial lattice, that enables the fast computation of all the required marginal frequencies.

1) Computing every marginal entropy once: Let us consider again the example illustrated in Fig. 1. We have seen that the simplified G^2 statistic requires computation of $H(\{123\})$, $H(\{234\})$, $H(\{23\})$ and $H(\{1234\})$. Among these four entropies, the first two are entropies of 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 entropy $H(\{23\})$ has also been computed in the process of selecting either $\{123\}$ or $\{234\}$.

Clearly, the forward selection procedure exhibits many overlapping sub-problems. We propose to memoize these partial solutions. The calculation of the G^2 statistic is then reduced to a function of only one new term, namely $H(\{1234\})$. This compares to the direct calculation of Eqn. 6 that included 20 different entropies.

2) Computing every logarithm once: Another case of overlapping sub-problems can be found at a lower level. The marginal entropy for a set of variables $A \subseteq \mathcal{V}$ can be expressed in terms of the observed frequencies:

$$H(A) = -\frac{1}{N}\sum_{\mathbf{x}\in A} \mathrm{O}_{\mathbf{x}}^{A} \cdot \left(\ln \mathrm{O}_{\mathbf{x}}^{A} - \ln N\right) \qquad (8)$$

which can be re-expressed as:

$$H(A) = -\frac{1}{N}\sum_{\mathbf{x}\in A} \mathit{partial\_entropy}(\mathrm{O}_{\mathbf{x}}^{A}) \qquad (9)$$

Given that the $\ln(.)$ function is computationally expensive and that $\forall A \subseteq \mathcal{V}, \forall \mathbf{x} \in A, \mathrm{O}_{\mathbf{x}}^{A} \leqslant N$, we pre-compute all values of the partial entropies in an array of size N (with $O(1)$ access).

3) Computing marginal frequencies: The scalability of Chordalysis 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. $\mathcal{M}^c$ will most of the time require computing a single marginal entropy. We showed that all the possible partial entropies can be pre-computed. The only missing elements are thus $\mathrm{O}_{\mathbf{x}}^{A}$ for different sets of variables A and for all the combinations of values for this set.

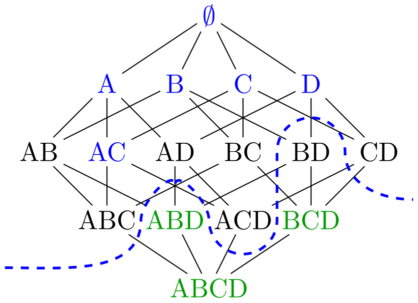


Figure 2. Lattice on four variables A, B, C, D. The decomposable model [ABC][ACD], which excludes the interaction between B and D, is illustrated as a cut on the lattice. The green nodes haven't been explored yet, and are thus not stored. The blue nodes need not be stored anymore.

Association discovery between values has carefully studied the efficient computation of marginal frequencies. The main methods rely on a vertical description of the data – for instance Tidsets [10] or Diffsets [15].

Chordalysis uses Tidsets to store the dataset. Computing marginal frequencies relies on the computation of intersections between Tidsets. To optimize this computation, a good data structure would minimize both the number of Tidsets whose intersections are calculated and their sizes. We propose a data structure based on a partial lattice, as illustrated in Fig. 2. Every node of the partial lattice stores one Tidset for every associated combination of values. For example, the node AB will store $(|\operatorname{Dom}(A)| \cdot |\operatorname{Dom}(B)|)$ Tidsets. We call this structure a *partial* lattice for two reasons: (1) nodes that have not been explored are not stored (*e.g.*, in Fig. 2, the nodes ABD, BCD and $ABCD$ are not part of the partial lattice); (2) if all the children of a node have been explored, the Tidsets of the node can be removed from the lattice – this is a consequence of the *forward* progression of the method. In Fig. 2, the first level need not be stored because the entire second level has been explored. Similarly, AC need not be stored, because its two children ABC and ACD have been explored. However, even though it has been explored, BD is kept in memory, because it can be used to explore ABD and BCD.

This partial lattice makes it possible to guarantee that every Tidset will be computed with a single intersection and the number of operations to compute this intersection is minimal. When a new node n is explored – in order to compute the associated marginal entropy – a Tidset has to be computed for every associated combination of values. Every Tidset can be computed from the corresponding Tidsets of any two parents of n. This ensures that only one intersection is required. Moreover, we choose the parents that have the smallest corresponding Tidsets. This ensures the number of operations for the single intersection is minimal.

F. Layered critical values

Chordalysis makes intensive use of statistical testing. It is well known that multiple testing can result in many false discoveries. Given the size of the search space, a large number of tests will be performed. As a result, modifications of the models may be accepted too often. This can be avoided by using *layered critical values* [11], a variant of the Bonferroni correction that increases the number of significant patterns discovered while still maintaining strict control over the risk of false discoveries. Given the p-value threshold α (usually $\alpha = 0.05$), the layered p-value α_L at step L with a search space of size S_L is defined as:

$$\alpha_L = \frac{\alpha}{2^L \cdot S_L} \qquad (10)$$

where L is the number of edges in the current best model $\mathcal{M}^\star$, and S_L is the number of chordal graphs that can be formed by adding a single edge to $\mathcal{M}^\star$.

G. Complexity

We compare here the complexity of Chordalysis against that of a standard log-linear modeling forward selection procedure among graphical models. In both cases, the worst number of steps corresponds to the construction of the saturated model, *i.e.*, $max\_\#steps = \frac{M\cdot(M-1)}{2}$ steps. At step s, both methods will select a $\mathcal{M}^c$ for which the corresponding graph includes s edges. At step s, both methods consider the replacement of $\mathcal{M}^\star$ by $(max\_\#steps - s + 1)$ $\mathcal{M}^c$. So far, both methods have a similar worst case. However, the evaluation of every $\mathcal{M}^c$ is drastically different. In the standard log-linear analysis case, evaluating any $\mathcal{M}^c$ requires $O(2^M)$ operations (assuming binary variables) irrespective of its structure. In Chordalysis, the evaluation depends on the structure of the graph and requires at most $O(2^{\sqrt{s}})$ operations.

Proof: At step s, all the $\mathcal{M}^c$ contain exactly s edges. The computation of the G^2 statistic is upper bounded by the evaluation of the biggest clique that can be created by adding the s^{th} edge to the graph. If all s edges are composing a single clique, the clique has at most $k = \frac{1+\sqrt{1+8s}}{2}$ variables.[1] Assuming binary variables, thus the evaluation requires at most $O(2^{\sqrt{s}})$ operations. □

Note also that it is very unlikely that all the edges will be in the same clique. In practice, while the standard approach will require $O(2^M)$ operations, the number of operations will be much smaller for Chordalysis.

IV. Related research

Researchers have investigated the learning of 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 log-linear analysis scales up – and then to combine these sub-models [3], [5].

A second approach consists of optimizing the negative log-likelihood of the different neighboring of the graph [6]. ℓ_1-regularized logistic regression is used to discover the local structure of every variable. In a similar fashion, [16] propose to focus on sets of variables that will best divides the graph.

A third approach focuses on a reduced subset of features [17]. This is achieved using ℓ_1-regularization which biases the search towards models for which many parameters are zero.

However, by focusing on local structures, these methods remain *ad hoc*, and cannot ensure that all the subtleties of interactions will be captured.

To the best of our knowledge, [13], [14] are the only cases in which researchers have tried to learn the structure of the model without focusing on local structures. The objective function is the Kullback-Leibler divergence, which is minimized when the observed frequencies are equal to the modeled frequencies.

The saturated model (containing the full-way interaction) always minimizes the KL-divergence, because it predicts exactly the observed frequencies. As a consequence, models built with this objective function will retrieve the saturated model if this model is explored. This is an important issue because the MLEs require a number of samples that is exponential with the number of variables. With 100 binary variables, the MLE will require a dataset with more than 10^{30} instances. Moreover, such a model is of no interest to the data analyst, because it doesn't give any information about the underlying dependencies that take place in data. To address this issue, one solution is to limit the number of variables that can interact in the model to a given parameter k [18], known as the *treewidth* of the graph. This constraint guarantees that the saturated model will not be explored. However, this constraint raises two other issues: (1) there is no method to determine k in advance and (2) even if the best k is chosen, the approach extremely overfits the data. We will show that KL approaches are far from being optimal. This constraint compares with the discovery of itemsets/associations involving no more than k items. There will rarely be a uniformly optimal k because the complexity of data is usually not homogeneous: if one 5-way interaction has to be discovered, and k is set to 5, then many other 5-way interactions will be retrieved. In this paper, we argue that ensuring statistical significance constitutes a sound way to discover associations between variables.

[1] A clique of k variables contains $\frac{k\cdot(k-1)}{2}$ edges. Solving $\frac{k\cdot(k-1)}{2} = s$ gives $k = \frac{1+\sqrt{1+8s}}{2}$.

V. Experiments

A. Datasets from known models

Assessing the quality of association discovery between variables requires having knowledge about the multi-way interactions that take place in data. Therefore we evaluate the discovery with data that is randomly sampled from a known distribution (set of interactions and associated probability tables). We can then compare the discovered interactions to the true structure from which the data was sampled.

As we pursue the class of decomposable models, the structure of the model is completely determined by its pairwise interactions. We can thus assess the recovery of the structure in terms of the edges in the associated graph. Each possible edge in the graph can be present or absent in the true model and each edge can also have been discovered or not. This corresponds to the standard scheme true/false positive/negative. To take into account both the precision and the recall of the discovery process, we evaluate the models using the F-measure, which corresponds to the harmonic mean between precision and recall:

$$\mathcal{F} = 2 \cdot \frac{\mathcal{P} \cdot \mathcal{R}}{\mathcal{P} + \mathcal{R}}$$

We compare Chordalysis to the state-of-the-art method based on the KL-divergence [13], [14]. The results are compared in terms of the quality of the recovery of the structure and in terms of the execution time and complexity.[2]

Data structures: We designed five models (Fig. 3) from which the data will be sampled. The first four models are used to investigate the behavior of the methods on particular types of distribution: three independent variables ($\mathcal{D}_1$), one triple interaction ($\mathcal{D}_2$), one triple interaction and two independent variables ($\mathcal{D}_3$) and a more complicated structure with one 4-way, one 3-way and three 2-way interactions ($\mathcal{D}_4$). These four models are intended to investigate the relative overfitting behavior of the KL-based method in contrast with the conservative behavior of Chordalysis. Model $\mathcal{D}_5$ investigates the scalability (quality and performance) on high-dimensional data. It 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.

1) $\mathcal{D}_1$ to $\mathcal{D}_4$: The results are given in Fig. 4, with the x-axis depicted in log-scale. For all these structures, once 100 samples are present in the dataset, the KL-based method terminates with the saturated model (the model with the full-way interaction). The saturated model includes all the possible interactions, and has thus a recall of 100 %. It is obvious that returning all the possible interactions – whatever the true model is – will lead in the general case to a very low precision. The poor quality of the set of discovered interactions is illustrated in the charts, with the exception of $\mathcal{D}_2$ for which the true model actually is saturated.

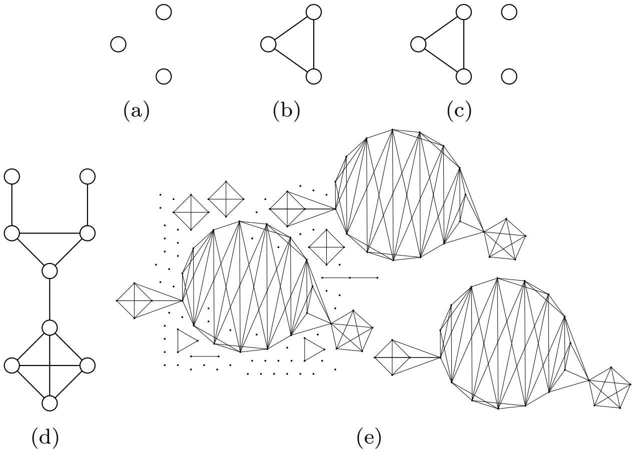


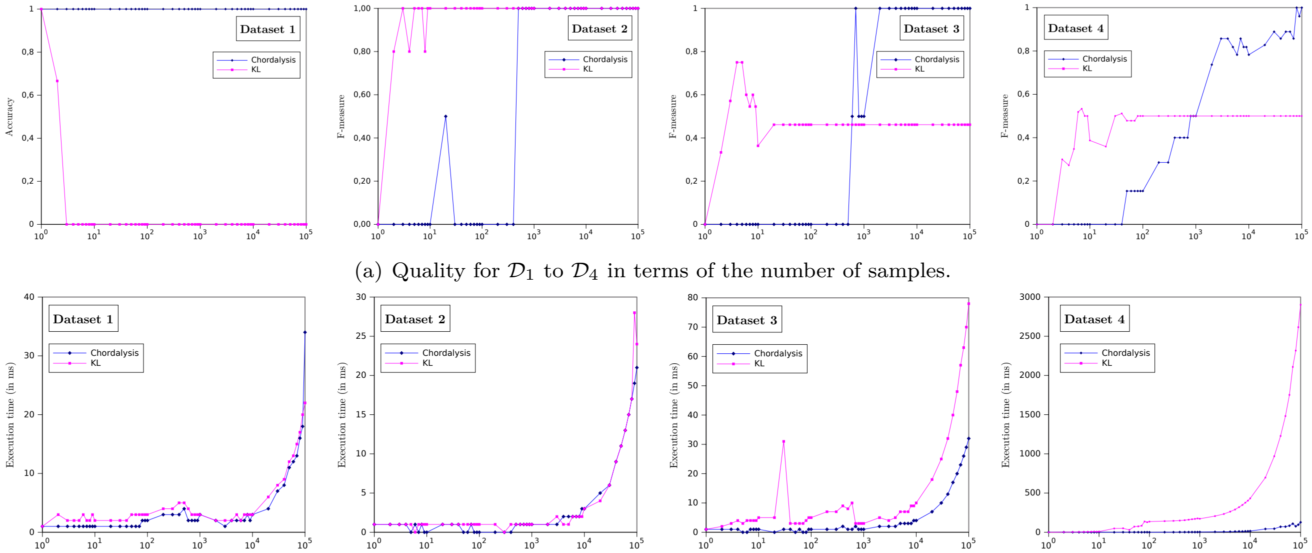
Figure 3. Data structures that are randomly sampled. (a) $\mathcal{D}_1$. (b) $\mathcal{D}_2$. (c) $\mathcal{D}_3$. (d) $\mathcal{D}_4$. (e) $\mathcal{D}_5$.

To the contrary, Chordalysis is much more consistent. The quality of the model increases with the number of samples and always ends up with $\mathcal{F} = 1.0$ indicating recovery of the exact model. It is also interesting to notice that the complexity of the interactions that can be recovered generally depends upon the number of available samples. This behavior comes from the statistical significance that is enforced by Chordalysis.

Note that the efficient data structure enables Chordalysis to recover the true model with less than 150 ms. The KL-based method requires twenty times longer to converge for $\mathcal{D}_4$, because it considers overly complex models, even for these low-dimensional datasets.

2) $\mathcal{D}_5$: $\mathcal{D}_5$ includes 150 variables. The complexity of the KL approach is exponential with the size of the maximal cliques that are retrieved (*i.e.*, the size of the multi-way interactions). This is a consequence of the stopping criteria: it stops when there is no more eligible pairs of vertices that would keep the graph decomposable. For the previous experiments, this happened when reaching the saturated model. However, for this 150-dimensional dataset the complexity makes the unconstrained approach infeasible and another stopping criteria has to be set. We limit the KL approach to consider at most k-way interactions, with $k \in [\![2, 7]\!]$. In this case, limiting k to 7 is consistent since we know that $\mathcal{D}_5$ include at most 5-way interactions. However, for real cases there may be no way to determine k_{max}, because the model from which the data is drawn is not known. This is an additional advantage of Chordalysis, which

[2] Note that both methods are implemented inside the exact same framework, thus, the execution times are truly comparable.



(a) Quality for $\mathcal{D}_1$ to $\mathcal{D}_4$ in terms of the number of samples.

(b) Execution time for $\mathcal{D}_1$ to $\mathcal{D}_4$ in terms of the number of samples.

Figure 4. Results of the experiments comparing Chordalysis to the KL-divergence approach on datasets $\mathcal{D}_1$ to $\mathcal{D}_4$.

doesn't need to set such a parameter.

Fig. 5(a) illustrates the results in terms of the quality of returned structure. The KL approach achieves better recall than Chordalysis with small quantities of data because it is less conservative. However, from $50,000$ samples, Chordalysis outperforms KL at all settings of k, and with $500,000$ samples, the model is retrieved with $\mathcal{F} = \mathbf{90}$ % ($\mathcal{P} = 83$ %, $\mathcal{R} = 98$ %). This compares to the best scores obtained with the KL method and $k = 4$ with only $\mathcal{F} = 59$ % ($\mathcal{P} = 45$ %, $\mathcal{R} = 88$ %).

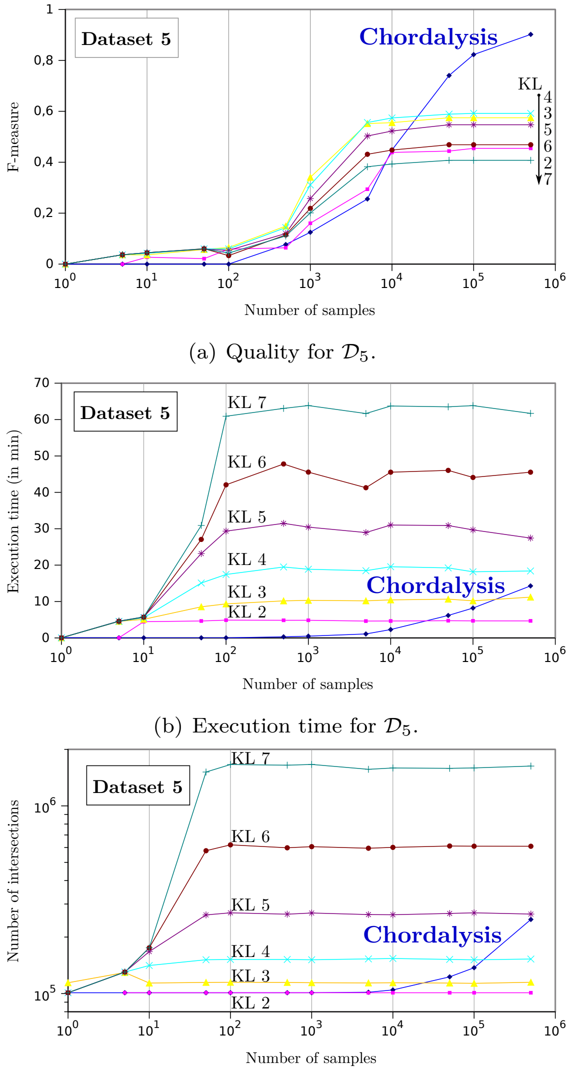
Let us explain why the quality for the KL-based method is ordered as $KL4 > KL3 > KL5 > KL6 > KL2 > KL7$. The KL-based method will retrieve as many k ways interactions as possible; even the 55 independent variables are included in k-way interactions. The recall of the edges of the graph progressively increases with the highest-order interaction k. From $k = 5$, the recall is greater than 99 %. As a consequence of this overfitting behavior, the precision follows an opposite trend. Starting at 57 % for $k = 2$, the precision drops down to 38 % for $k = 5$ and only 26 % for $k = 7$. Overall, whatever the configuration, the precision of the retrieved structure is too low to provide a consistent model.

Not only are the structure quality results of Chordalysis much better, but its execution is also much quicker (Fig. 5(b)). Chordalysis focuses on the statistically significant part of the search space. As a consequence, Chordalysis explores far fewer nodes of the lattice (multi-way interactions) than the KL approach. Fewer combinations of values are considered, and so fewer intersections are computed (Fig. 5(c)). With $500,000$ samples, Chordalysis explores almost as much as KL5. However, by focusing on the statistically significant part of the search space, Chordalysis obtains much better results. Fig. 5(b) shows that Chordalysis only requires half the time of KL5. This is another consequence of the statistical significance of Chordalysis: the associated decomposable graph is much simpler, which reduces the maintenance time for the data structure.

B. Results on a real dataset

To demonstrate real-world performance we apply Chordalysis to a 25 variable dataset from an epidemiological study of the elderly (EPESE) [19]. The resulting model (selected in less than 2 s) is shown in Fig. 6. Expert assessment of this dataset is provided in [20].

Many of the multi-way relationships retrieved by Chordalysis have supporting evidence. For example, the high-order interactions between age and gender and a third variable are explained by the fact that, generally, if the patient is older, then there has been a longer period over which they have had the opportunity to have smoked, been married or have retired (and hence not be working), while it is well-known that women get married earlier than men. Many identified interactions have a direct medical interpretation, including the relationships between diabetes and taking insulin, smoking (or having smoked) and having had cancer, etc.. Other identified interactions are less obvious. For example, the connection between taking high blood pressure medication and having pain walking might be explained by the fact that high blood pressure is often treated with diuretics, which may decrease the body's levels of the mineral potassium, leading specifically to possible leg cramps. An interesting example is the four-way relationship identified by our method between the age, whether the patient provided the correct age, and the two smoking variables. The correctness of the declared age is indicative of the patient's



(a) Quality for $\mathcal{D}_5$.

(b) Execution time for $\mathcal{D}_5$.

(c) Number of intersections performed for $\mathcal{D}_5$.

Figure 5. Results of the experiments comparing the KL-divergence approach to the proposed χ^2 based one on dataset $\mathcal{D}_5$. The KL-divergence approach has been limited to $k \in [\![2,7]\!]$.

mental health. Mental health decreases with age and it is more likely for someone to have smoked if he/she is older. The *a priori* surprising fact is the inclusion of smoking variables in a group of variables related to mental health. It turns out that a recent neurological study [21] established that smoking increases the risk of dementia and Alzheimer's disease.

VI. Conclusion and future research

While the data mining community has focused on associations between variable values, it is often useful to directly find associations between variables. Statisticians have well-developed statistically sound techniques for the latter task, but they have not scaled up to more than about ten variables at a time. We have melded the statistical machinery of log-linear analysis together with the computational methods of association rule discovery and network models, to develop a statistically sound approach to discovering associations between variables in high-dimensional data.

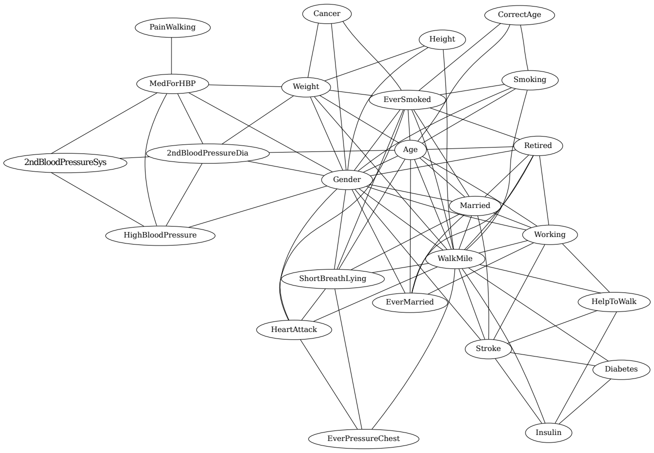


Figure 6. Decomposable model selected for the EPESE dataset.

Our contributions to log-linear analysis include:

- Proof that the likelihood ratio statistic (G^2) for decomposable models can be expressed in terms of the graph structure of the model.
- Proof that the G^2 statistic for comparing two decomposable models that differ by the inclusion of a single edge can be calculated using only a function of four marginal entropies. This dramatically reduces the amount of computation relative to previous approaches.
- Efficient techniques for computing G^2 using the above proofs and techniques developed for itemset mining.
- Efficient techniques for finding all candidate single edge additions to an arbitrary decomposable model that result in decomposable models. This is the basis of our efficient forward-selection search.
- A variant of the layered critical values technique to strictly control the familywise risk of Type 1 error.
- Memoization of marginal entropies, ensuring that every marginal entropy will only be computed once.

One limitation of our approach is that it only considers additions to a model that involve adding a single edge that results in a decomposable graph. This means that our approach may fail to find associations between variables where the addition of the association in a single step 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 single-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.

Our research builds upon a large body of work in the learning of graphical models. It remains an important direction for future research to compare the utility of the statistical techniques we have scaled up, to Bayesian and information theoretic approaches (*e.g.*, [22], [23]) to managing the trade-off between model complexity and goodness of fit.

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

References

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

[2] 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.

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

[4] S.-H. Kim and S. Lee, "Searching model structures based on marginal model structures," in *New Developments in Robotics Automation and Control*. InTech, 2008, pp. 355–376.

[5] 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.

[6] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

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

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

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

[10] 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.

[11] G. I. Webb, "Layered critical values: a powerful direct-adjustment approach to discovering significant patterns," *Machine Learning*, vol. 71, no. 2–3, pp. 307–323, 2008.

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

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

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

[15] M. J. Zaki and K. Gouda, "Fast vertical mining using diffsets," in *Int. Conf. on Knowledge Discovery and Data Mining*. ACM, 2003, pp. 326–335.

[16] V. Gogate, W. A. Webb, and P. Domingos, "Learning Efficient Markov Networks," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 748–756.

[17] 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 18*, 2006, pp. 817–824.

[18] A. Deshpande, M. Garofalakis, and R. Rastogi, "Independence is good: dependency-based histogram synopses for high-dimensional data," in *Int. Conf. on Management of Data*, 2001, pp. 199–210.

[19] J. O. Taylor, R. B. Wallace, A. M. Ostfeld, and D. G. Blazer, "Established Populations for Epidemiologic Studies of the Elderly, 1981-1993," Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 1998, http://dx.doi.org/10.3886/ICPSR09915.

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

[21] C. Reitz, T. den Heijer, C. van Duijn, A. Hofman, and M. Breteler, "Relation between smoking and risk of dementia and alzheimer disease: The Rotterdam study," *Neurology*, vol. 69, no. 10, pp. 998–1005, 2007.

[22] W. Lam and F. Bacchus, "Learning Bayesian belief networks: An approach based on the MDL principle," *Computational Intelligence*, vol. 10, pp. 269–293, 1993.

[23] C. S. Wallace, K. Korb, and H. Dai, "Causal discovery via MML," in *Int. Conf. on Machine Learning*, 1996, pp. 516–524.

[24] A. Berry and R. Pogorelcnik, "A simple algorithm to generate the minimal separators and the maximal cliques of a chordal graph," *Information Processing Letters*, vol. 111, no. 11, pp. 508–511, 2011.

Appendix

A. Maximal cliques and minimal separators

Let $\mathcal{G} = (\mathcal{V}, E)$ be the undirected graph, where $\mathcal{V}$ is the set of variables and E the set of edges in $\mathcal{G}$.

Definition 4: A set $C \subseteq \mathcal{V}$ is a *clique* of $\mathcal{G}$ iff all its vertices are pairwise adjacent.

Definition 5: A clique C is *maximal* iff there is no vertex $V \in \mathcal{V}, V \notin C$ such that $C \cup \{V\}$ is a clique.

Definition 6: A set $S \subseteq \mathcal{V}$ is a *separator* of $\mathcal{G}$ if $G = (\mathcal{V} - S, E)$ is unconnected.

Definition 7: A separator set S of $\mathcal{G}$ is *minimal* if no subset of S is a separator.

Chordal graphs (graphs corresponding to decomposable models) are an important family of graphs, for which many polynomial time algorithms are available for complex problems. This is due to the fact that a *perfect elimination ordering* (peo) on the vertices can be established with a Lexicographic Breadth First Search or by a Maximum Cardinality Search, *i.e.*, in $O(|\mathcal{V}| + |E|)$ operations. Moreover, for chordal graphs, the maximal cliques and the minimal separators can be found in a single pass [24].