# On Identifying Tree-Structured Perfect Maps

Christian Borgelt

Dept. of Knowledge Processing and Language Engineering
School of Computer Science
Otto-von-Guericke-University of Magdeburg
Universitätsplatz 2, 39106 Magdeburg, Germany
borgelt@iws.cs.uni-magdeburg.de

**Abstract.** It is well known that tree-structured perfect maps can be uniquely identified by computing a maximum weight spanning tree with mutual information providing the edge weights. In this paper I generalize the edge evaluation measure by stating the conditions such a measure has to satisfy in order to be able to identify tree-structured perfect maps. In addition, I show that not only mutual information, but also the well-known $\chi^2$ measure satisfies these conditions.

## 1 Introduction

At the core of the theory of graphical models [10, 6, 9, 2, 1], that is, of Bayes networks and Markov networks, is the notion of a so-called *conditional independence graph* or *independence map* for a given multidimensional probability distribution. It allows us to determine the conditional independence statements obtaining in the probability distribution by applying a simple graph theoretic criterion, which is based on node separation.

The exact form of this criterion depends on whether the graph is directed (Bayes network) or undirected (Markov network). If it is undirected, so-called *u-separation* is defined as follows: Let $X$, $Y$, and $Z$ be three disjoint sets of nodes of a graph $G$. Then $Z$ $u$-separates $X$ and $Y$ iff all paths from a node in $X$ to a node in $Y$ contain a node in $Z$. If the graph is directed, the slightly more complicated notion of *d-separation* is used [10]: Here $Z$ $d$-separates $X$ and $Y$ iff there is no path, i.e., no sequence of consecutive edges (of any directionality) along which the following two conditions hold:

1. every node, at which edges of the path converge (i.e., both edges are directed towards the node), either is in $Z$ or has a descendant in $Z$,
2. every other node is not in $Z$.

As already stated above, a graph $G$ is called a *conditional independence graph* or an *independence map* [10] iff all conditional independences that can be read from it using these criteria actually hold in the associated probability distribution $p$, or formally

$$\langle X \mid Z \mid Y \rangle_G \quad \Rightarrow \quad X \perp\!\!\!\perp_p Y \mid Z,$$

where $\langle X \mid Z \mid Y \rangle_G$ denotes that $Z$ separates $X$ and $Y$ in the graph $G$ and $X \perp\!\!\!\perp_p Y \mid Z$ means that

$$\begin{aligned}
&\forall x \in \operatorname{dom}(X) : \forall y \in \operatorname{dom}(Y) : \forall z \in \operatorname{dom}(Z) : \\
&\quad P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z).
\end{aligned}$$

However, there may be additional conditional independences holding in the distribution that are not captured by the graph.

The dual concept is a so-called *conditional dependence graph* or *dependence map* [10], which captures all conditional dependences holding in the distribution, or—stated the other way round—all conditional independences obtaining in the distribution are represented in the graph. Formally, we have

$$X \perp\!\!\!\perp_p Y \mid Z \quad \Rightarrow \quad \langle X \mid Z \mid Y \rangle_G.$$

The graph may represent additional conditional independences that do not hold in the distribution, but wherever it indicates a conditional dependence, this dependence holds in the distribution.

If a graph is both an independence map as well as a dependence map, that is, if it captures exactly the conditional independence statements holding in the distribution, no more and no less, it is called a *perfect map* [10].

In this paper I consider tree-structured perfect maps and examine how they can be determined by constructing a maximum weight spanning tree for given edge weights. While it is well-known that tree-structured perfect maps can be uniquely determined if mutual information is used to compute the edge weights, I state here more generally what conditions the edge evaluation measure has to satisfy for this task. Furthermore, I show that not only mutual information, but also the well-known $\chi^2$ measure satisfies these conditions.

## 2 Identifying Tree-Structured Perfect Maps

The best-known greedy approach to induce a graphical model—and at the same time the oldest—is *optimum weight spanning tree construction* and was first suggested in [3]. All possible (undirected) edges over the set $U = \{A_1, \ldots, A_n\}$ of attributes used to describe the multidimensional domain under consideration are evaluated with an evaluation measure (in [3] *mutual information* was used). Then an optimum weight spanning tree is constructed with either the (well-known) Kruskal algorithm [7] or the (somewhat less well-known) Prim algorithm [12] (or any other greedy algorithm for this task).

I am interested in this approach here, because if the probability distribution, for which a graphical model is desired, has a perfect map that is a tree, optimum weight spanning tree construction is guaranteed to find this perfect map, provided the evaluation measure used has a certain property.

However, before I state the corresponding theorem, I should introduce the notion of *symmetry* (although it is, of course, canonical): An evaluation measure $m : U \times U \to \mathrm{I\!R}$ is called *symmetric* iff $\forall A, B : m(A, B) = m(B, A)$.

**Theorem 1.** *Let $m$ be a symmetric evaluation measure satisfying*

$$\forall A, B, C : \quad m(C, AB) \;\geq\; m(C, B),$$

*with equality obtaining only if the attributes $A$ and $C$ are conditionally independent given $B$. ($AB$ is a pseudo-attribute with values in* $\mathrm{dom}(A) \times \mathrm{dom}(B)$*.) Let $G$ be a singly connected (or tree-structured) undirected perfect map of a probability distribution $p$ over a set $U$ of attributes. Then constructing a maximum weight spanning tree for the attributes in $U$ with $m$ (computed from $p$) providing the edge weights uniquely identifies $G$.*

In order to prove this theorem, it is convenient to prove first the following lemma, by which an important property of the measure $m$ is established:

**Lemma 1.** *Let $m$ be a symmetric evaluation measure satisfying*

$$\forall A, B, C : \quad m(C, AB) \;\geq\; m(C, B)$$

*with equality obtaining only if the attributes $C$ and $A$ are conditionally independent given $B$. Furthermore, let $p$ be the probability distribution from which $m$ is computed. If $A$, $B$, and $C$ are three attributes satisfying $A \perp\!\!\!\perp_p C \mid B$, but neither $A \perp\!\!\!\perp_p B \mid C$ nor $C \perp\!\!\!\perp_p B \mid A$, then*

$$m(A, C) \;<\; \min\{m(A, B), m(B, C)\}.$$

*Proof.* From the facts that $m$ is symmetric and $A \perp\!\!\!\perp_p C \mid B$ we know that

$$m(C, AB) = m(C, B) \qquad \text{and} \qquad m(A, CB) = m(A, B).$$

Since it is $A \not\!\perp\!\!\!\perp_p B \mid C$ and $C \not\!\perp\!\!\!\perp_p B \mid A$, we have

$$m(C, AB) > m(C, A) \qquad \text{and} \qquad m(A, CB) > m(A, C).$$

Consequently, $m(C, A) < m(C, B)$ and $m(C, A) < m(A, B)$. □

*Proof.* (of Theorem 1)
Let $C$ and $A$ be two arbitrary attributes in $U$ that are not adjacent in $G$. Since the graph $G$ is singly connected there is a unique path connecting $C$ and $A$ in $G$. I show that any edge connecting two consecutive nodes on this path has a higher weight than the edge $(C, A)$.

Let $B$ be the successor of $C$ on the path connecting $C$ and $A$ in $G$. Then it is $C \perp\!\!\!\perp_p A \mid B$, but neither $C \perp\!\!\!\perp_p B \mid A$ nor $A \perp\!\!\!\perp_p B \mid C$, because $G$ is a perfect map. Consequently, it is $m(C, A) < m(C, B)$ and $m(C, A) < m(B, A)$. If $B$ is the predecessor of $A$ on the path, we already have that all edges on the path have a higher weight than the edge $(C, A)$. Otherwise we have that the edge $(C, B)$ has a higher weight than the edge $(C, A)$. For the remaining path, i.e., the path that connects $B$ and $A$, the above argument is applied recursively.

Therefore any edge between two consecutive nodes on the path connecting any two attributes $C$ and $A$ has a higher weight than the edge $(C, A)$. From this it is immediately clear, for example by considering how the Kruskal algorithm [7] works, that constructing the optimum weight spanning tree with $m$ providing the edge weights uniquely identifies $G$. □

In the next section I show in Theorems 3 and 4 that at least mutual information and the $\chi^2$ measure have the property presupposed in this theorem.

It is clear that the above theorem holds also for directed trees, since any undirected conditional independence graph that is a tree can be turned into an equivalent directed tree by choosing an arbitrary root node and (recursively) directing the edges away from this node. However, with an additional requirement, it can also be extended to polytrees.

**Theorem 2.** *Let $m$ be a symmetric evaluation measure satisfying*

$$\forall A, B, C: \quad m(C, AB) \;\geq\; m(C, B)$$

*with equality obtaining only if the attributes $A$ and $C$ are conditionally independent given $B$ and*

$$\forall A, C: \quad m(C, A) \;\geq\; 0$$

*with equality obtaining only if the attributes $A$ and $C$ are (marginally) independent. Let $\boldsymbol{G}$ be a singly connected directed perfect map of a probability distribution $p$ over a set $U$ of attributes. Then constructing a maximum weight spanning tree for the attributes in $U$ with $m$ (computed from $p$) providing the edge weights uniquely identifies the so-called* skeleton *of $\boldsymbol{G}$, i.e., the undirected graph that results if all edge directions are discarded.*

*Proof.* Let $C$ and $A$ be two arbitrary attributes in $U$ that are not adjacent in $\boldsymbol{G}$. Since the graph $\boldsymbol{G}$ is singly connected, there is a unique path connecting $C$ and $A$. Suppose first that this path does not contain a node with converging edges (from its predecessor and its successor on the path). In this case the proof of Theorem 1 can be transferred, because, according to $d$-separation, we have $C \perp\!\!\!\perp_p A \mid B$, but neither $C \perp\!\!\!\perp_p B \mid A$ nor $A \perp\!\!\!\perp_p B \mid C$ (because $\boldsymbol{G}$ is a perfect map). Therefore the value of $m$ must be less for the edge $(C, A)$ than for any pair of consecutive nodes on the path connecting $C$ and $A$.

Suppose next that the path connecting $C$ and $A$ in $\boldsymbol{G}$ contains at least one node with converging edges (from its predecessor and its successor on the path). According to the $d$-separation criterion (see Section 1 for the definition), $C$ and $A$ must be marginally independent and hence it is $m(C, A) = 0$. However, no pair $(B_i, B_j)$ of consecutive nodes on the path is marginally independent (since $\boldsymbol{G}$ is a perfect map) and thus $m(B_i, B_j) > 0$.

Therefore any edge between two nodes on a path connecting two nonadjacent nodes in the perfect map $\boldsymbol{G}$ has a higher weight than the edge connecting them directly. From this it is immediately clear, for example by considering how the Kruskal algorithm [7] works, that constructing the maximum weight spanning tree with $m$ providing the edge weights uniquely identifies the skeleton of $\boldsymbol{G}$. □

Note that the above theorem is an extension of a theorem shown in [14, 10], where it was proven with mutual information providing the edge weights. Note also that the edges of the skeleton found with the above approach may be directed with an algorithm presented in [14, 10], although the result may not be unique, because often the direction of some edges can be chosen arbitrarily.

## 3 Edge Evaluation Measures

The theorems in the preceding section are formulated in a general way with an evaluation measure $m$ that has to satisfy certain properties. In this section I show that at least *mutual information* [8,3], which is also known under the names of *cross entropy* or *information gain* [13], and the $\chi^2$ measure satisfy these conditions, so both can be used to identify tree-structured perfect maps.

### 3.1 Notation

In the following I will use the following notation: Let $A$, $B$, and $C$ be three attributes with domains $\text{dom}(A) = \{a_1, \ldots, a_{n_A}\}$, $\text{dom}(B) = \{b_1, \ldots, b_{n_B}\}$, and $\text{dom}(C) = \{c_1, \ldots, c_{n_C}\}$, respectively. Furthermore, let $P$ be a strictly positive probability measure defined on the joint domain of $A$, $B$, and $C$. In order to make the formulae easier to read, I introduce the following abbreviations:

$$\begin{array}{rlrl} p_{i..} &= P(C = c_i), & p_{ij.} &= P(C = c_i, A = a_j), \\ p_{.j.} &= P(A = a_j), & p_{i.k} &= P(C = c_i, B = b_k), \\ p_{..k} &= P(B = b_k), & p_{.jk} &= P(A = a_j, B = b_k), \quad \text{and} \\ & & p_{ijk} &= P(C = c_i, A = a_j, B = b_k), \end{array}$$

i.e., the index $i$ always refers to the attribute $C$, the index $j$ always refers to the attribute $A$, and the index $k$ always refers to the attribute $B$. If a formula refers only to two attributes $C$ and $A$, the third index $k$ is dropped.

### 3.2 Mutual Information

The *mutual information* of two attributes $C$ and $A$ w.r.t. $P$ can be defined in different ways. In the first place, it can be defined as a pointwise comparison of the actual joint distribution, as it is described by $p_{ij}$, to a hypothetical independent distribution, as it can be computed by $p_{i.}p_{.j}$. That is,

$$I_{\text{mut}}(C, A) \;=\; \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} p_{ij} \log_2 \frac{p_{ij}}{p_{i.}p_{.j}}.$$

Alternatively, one may draw on the notion of the *Shannon entropy* $H$ of a probability distribution [15], which leads to

$$\begin{aligned} I_{\text{mut}}(C, A) &= H(C) + H(A) - H(CA) \\ &= -\sum_{i=1}^{n_C} p_i \log_2 p_i - \sum_{j=1}^{n_A} p_j \log_2 p_j + \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} p_{ij} \log_2 p_{ij}, \end{aligned}$$

which can be interpreted intuitively as measuring the reduction of the expected number of yes/no questions one has to ask in order to determine the obtaining value combination, or the reduction of the expected binary code length for transmitting the value tuple [1]. Obviously, the two definitions are equivalent.

The following theorem shows that mutual information satisfies the prerequisites of Theorem 1. Although the property of mutual information stated in it is well-known and the proof is merely a technical task, I provide a full proof (derived from a proof in [11] that mutual information is always nonnegative), because it is rarely spelled out clearly and thus is difficult to find.

**Theorem 3.** *Let $A$, $B$, and $C$ be three attributes with finite domains and let their joint probability distribution be strictly positive, i.e., let $\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) : P(A = a, B = b, C = c) > 0$. Then*

$$I_{\mathrm{mut}}(C, AB) \;\geq\; I_{\mathrm{mut}}(C, B),$$

*with equality obtaining only if the attributes $C$ and $A$ are conditionally independent given $B$.*

*Proof.* Since it makes the proof much simpler, I show that

$$I_{\mathrm{mut}}(C, B) - I_{\mathrm{mut}}(C, AB) \leq 0,$$

from which the original statement follows trivially.

$$\begin{aligned}
&I_{\mathrm{mut}}(C, B) - I_{\mathrm{mut}}(C, AB)\\
&= H(C) + H(B) - H(CB) - (H(C) + H(AB) - H(CAB))\\
&= -H(CB) - H(AB) + H(CAB) + H(B)\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B} p_{i.k}\log_2 p_{i.k} + \sum_{j=1}^{n_A}\sum_{k=1}^{n_B} p_{.jk}\log_2 p_{.jk}\\
&\quad - \sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\sum_{k=1}^{n_B} p_{ijk}\log_2 p_{ijk} - \sum_{k=1}^{n_B} p_{..k}\log_2 p_{..k}\\
&= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\sum_{k=1}^{n_B} p_{ijk}\log_2 \frac{p_{i.k}p_{.jk}}{p_{ijk}p_{..k}}\\
&= \frac{1}{\ln 2}\sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\sum_{k=1}^{n_B} p_{ijk}\ln \frac{p_{i.k}p_{.jk}}{p_{ijk}p_{..k}}\\
&\leq \frac{1}{\ln 2}\sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\sum_{k=1}^{n_B} p_{ijk}\left(\frac{p_{i.k}p_{.jk}}{p_{ijk}p_{..k}} - 1\right)\\
&= \frac{1}{\ln 2}\left[\sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\sum_{k=1}^{n_B} \frac{p_{i.k}p_{.jk}}{p_{..k}} - \underbrace{\sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\sum_{k=1}^{n_B} p_{ijk}}_{=1}\right]\\
&= \frac{1}{\ln 2}\left[\left(\sum_{k=1}^{n_B}\frac{1}{p_{..k}}\sum_{i=1}^{n_C}\sum_{j=1}^{n_A} p_{i.k}p_{.jk}\right) - 1\right]
\end{aligned}$$

$$= \frac{1}{\ln 2}\left[\left(\sum_{k=1}^{n_B}\frac{1}{p_{..k}}\underbrace{\left(\sum_{i=1}^{n_C}p_{i.k}\right)}_{=p_{..k}}\underbrace{\left(\sum_{j=1}^{n_A}p_{.jk}\right)}_{=p_{..k}}\right)-1\right]$$

$$= \frac{1}{\ln 2}\left(\underbrace{\left(\sum_{k=1}^{n_B}\frac{p_{..k}^2}{p_{..k}}\right)}_{=1}-1\right)$$

$$= \frac{1}{\ln 2}(1-1) \;=\; 0,$$

where the inequality follows from the fact that

$$\ln x \leq x - 1,$$

with equality obtaining only for $x = 1$. (This can most easily be seen from the graph of $\ln x$.) As a consequence, $I_{\text{gain}}(C, AB) = I_{\text{gain}}(C, B)$ only if

$$\forall i, j, k : \frac{p_{i.k}p_{.jk}}{p_{ijk}p_{..k}} = 1 \quad \Leftrightarrow \quad \forall i, j, k : p_{ij|k} = p_{i.|k}p_{.j|k},$$

where $p_{ij|k} = P(C = c_i, A = a_j \mid B = b_k)$ and $p_{i.|k}$ and $p_{.j|k}$ likewise. That is, $I_{\text{gain}}(C, AB) = I_{\text{gain}}(C, B)$ only holds if the attributes $C$ and $A$ are conditionally independent given attribute $B$. □

Note that with the above theorem it is easily established that mutual information is always nonnegative and zero only for independent attributes: Assume that attribute $B$ has only one value. In this case it is $I_{\text{gain}}(C, B) = 0$, since the joint distribution on the values of the two attributes clearly coincides with the distribution on the values of $C$. In addition, the combination of the attributes $A$ and $B$ is obviously indistinguishable from $A$ alone and thus we get $I_{\text{gain}}(C, AB) = I_{\text{gain}}(C, A)$. Consequently, we have as a corollary:

**Corollary 1.** *Let $C$ and $A$ be two attributes with finite domains and let their joint probability distribution be strictly positive, i.e. $\forall c \in \text{dom}(C)\colon \forall a \in \text{dom}(A)\colon P(C = c, A = a) > 0$. Then*

$$I_{\text{gain}}(C, A) \;\geq\; 0,$$

*with equality obtaining only if $C$ and $A$ are (marginally) independent.*

Therefore mutual information also satisfies the prerequisites of Theorem 2.

### 3.3 $\chi^2$ Measure

As mentioned above, one way to define mutual information relies on a pointwise comparison of the actual joint distribution, as it is described by $p_{ij}$, to a hypothetical independent distribution, as it can be computed by $p_i p_j$. The

$\chi^2$ *measure*, which is well known in statistics, does the same, but instead of the pointwise quotient (as mutual information does) it computes the pointwise squared difference of the two distributions. It is usually defined as

$$\begin{aligned}\chi^2(C,A) &= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} \frac{(E_{ij}-N_{ij})^2}{E_{ij}} \qquad\qquad \text{where } E_{ij} = \frac{N_{i.}\,N_{.j}}{N_{..}}\\ &= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} \frac{N_{..}^2\left(\frac{N_{i.}}{N_{..}}\frac{N_{.j}}{N_{..}} - \frac{N_{ij}}{N_{..}}\right)^2}{N_{..}\,\frac{N_{i.}}{N_{..}}\frac{N_{.j}}{N_{..}}}\\ &= N_{..}\sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\frac{(p_{i.}\,p_{.j}-p_{ij})^2}{p_{i.}\,p_{.j}},\end{aligned}$$

where the $N$'s are counters for the occurrence of certain value combinations in a sample. From these counters the probabilities are estimated by simple maximum likelihood estimation (i.e. as relative frequencies).

With the above transformation it is obvious that the numerator of the fraction is the squared difference of the actual joint distribution and the hypothetical independent distribution. The denominator serves to weight these pointwise differences. In order to render this measure independent of the number of sample cases, the factor $N_{..}$ (the size of the sample) is often discarded.

For the $\chi^2$ measure we have a direct analog of Theorem 3. That is, the $\chi^2$ measure also satisfies the prerequisites of Theorem 1 and may thus also be used to identify tree-structured perfect maps. The proof is also mainly a technical task, although it is slightly more complicated than the proof of Theorem 3.

**Theorem 4.** *Let $A$, $B$, and $C$ be three attributes with finite domains and let their joint probability distribution be strictly positive, i.e. let $\forall a \in \text{dom}(A) : \forall b \in \text{dom}(B) : \forall c \in \text{dom}(C) : P(A=a, B=b, C=c) > 0$. Then*

$$\chi^2(C,AB) \;\geq\; \chi^2(C,B),$$

*with equality obtaining only if the attributes $C$ and $A$ are conditionally independent given $B$.*

*Proof.* Since it makes the proof much simpler, I show

$$\frac{1}{N_{..}}\left(\chi^2(C,AB)-\chi^2(C,B)\right) \;\geq\; 0,$$

from which the original statement follows trivially.

$$\begin{aligned}&\frac{1}{N_{..}}\left(\chi^2(C,AB)-\chi^2(C,B)\right)\\ &= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A}\sum_{k=1}^{n_B}\frac{(p_{ijk}-p_{i..}p_{.jk})^2}{p_{i..}p_{.jk}} - \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\frac{(p_{i.k}-p_{i..}p_{..k})^2}{p_{i..}p_{..k}}\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\left(\sum_{j=1}^{n_A}\frac{p_{ijk}^2 - 2p_{ijk}p_{i..}p_{.jk} + p_{i..}^2 p_{.jk}^2}{p_{i..}p_{.jk}} - \frac{p_{i.k}^2 - 2p_{i.k}p_{i..}p_{..k} + p_{i..}^2 p_{..k}^2}{p_{i..}p_{..k}}\right)\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\left(\sum_{j=1}^{n_A}\left(\frac{p_{ijk}^2}{p_{i..}p_{.jk}} - 2p_{ijk} + p_{i..}p_{.jk}\right) - \frac{p_{i.k}^2}{p_{i..}p_{..k}} + 2p_{i.k} - p_{i..}p_{..k}\right)\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\left(\sum_{j=1}^{n_A}\frac{p_{ijk}^2}{p_{i..}p_{.jk}} - 2p_{i.k} + p_{i..}p_{..k} - \frac{p_{i.k}^2}{p_{i..}p_{..k}} + 2p_{i.k} - p_{i..}p_{..k}\right)\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\frac{1}{p_{i..}p_{..k}}\left(p_{..k}\sum_{j=1}^{n_A}\frac{p_{ijk}^2}{p_{.jk}} - p_{i.k}\sum_{j=1}^{n_A}p_{ijk}\right)\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\frac{1}{p_{i..}p_{..k}}\left[\left(\sum_{j_1=1}^{n_A}p_{.j_1k}\right)\left(\sum_{j_2=1}^{n_A}\frac{p_{ij_2k}^2}{p_{.j_2k}}\right) - \left(\sum_{j_1=1}^{n_A}p_{ij_1k}\right)\left(\sum_{j_2=1}^{n_A}p_{ij_2k}\right)\right]\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\frac{1}{p_{i..}p_{..k}}\left(\sum_{j_1=1}^{n_A}\sum_{j_2=1}^{n_A}\frac{p_{.j_1k}p_{ij_2k}^2}{p_{.j_2k}} - \sum_{j_1=1}^{n_A}\sum_{j_2=1}^{n_A}p_{ij_1k}p_{ij_2k}\right)\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\frac{1}{p_{i..}p_{..k}}\left(\sum_{j_1=1}^{n_A}\sum_{j_2=1}^{n_A}\frac{p_{.j_1k}^2 p_{ij_2k}^2 - p_{ij_1k}p_{ij_2k}p_{.j_1k}p_{.j_2k}}{p_{.j_1k}p_{.j_2k}}\right)\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\frac{1}{2p_{i..}p_{..k}}\sum_{j_1=1}^{n_A}\sum_{j_2=1}^{n_A}\frac{(p_{.j_1k}p_{ij_2k} - p_{ij_1k}p_{.j_2k})^2}{p_{.j_1k}p_{.j_2k}}\\
&= \sum_{i=1}^{n_C}\sum_{k=1}^{n_B}\sum_{j_1=1}^{n_A}\sum_{j_2=1}^{n_A}\frac{(p_{.j_1k}p_{ij_2k} - p_{ij_1k}p_{.j_2k})^2}{2p_{i..}p_{..k}p_{.j_1k}p_{.j_2k}} \quad \geq \quad 0,
\end{aligned}$$

where the semi-last step follows by duplicating the term in parentheses and then interchanging the indices $j_1$ and $j_2$ in the second instance (which is possible, because they have the same range). From the result it is immediately clear that $\chi^2(C, AB) \geq \chi^2(C, B)$: Since each term of the sum is a square divided by a product of (positive) probabilities, each term and thus the sum must be non-negative. It also follows that the sum can be zero only if all of its terms are zero, which requires their numerators to be zero:

$$\begin{aligned}
\forall i, j_1, j_2, k : p_{.j_1k}p_{ij_2k} - p_{ij_1k}p_{.j_2k} = 0 &\Leftrightarrow \forall i, j_1, j_2, k : \frac{p_{ij_2k}}{p_{.j_2k}} = \frac{p_{ij_1k}}{p_{.j_1k}}\\
&\Leftrightarrow \forall i, j_1, j_2, k : p_{i|j_2k} = p_{i|j_1k},
\end{aligned}$$

where $p_{i|j_\alpha k} = P(C = c_i \mid A = a_{j_\alpha}, B = b_k)$ with $\alpha \in \{1, 2\}$. As a consequence we have that $\chi^2(C, AB) = \chi^2(C, B)$ only holds if the attributes $C$ and $A$ are conditionally independent given attribute $B$. □

Note that no corollary is needed in this case, because from the definition of the $\chi^2$ measure it is already obvious that $\chi^2(C, A) \geq 0$. Therefore the $\chi^2$ measure also satisfies the prerequisites of Theorem 2 and thus may also be used to identify the skeleton of a polytree.

## 4 Conclusions

In this paper I provided a general statement of the conditions an edge evaluation measure has to satisfy in order to be able to identify a tree-structured perfect map or the skeleton of a polytree that is a perfect map. This generalizes the well-known fact that applying maximum weight spanning tree construction with mutual information providing the edge weights solves these tasks. In addition I showed that not only mutual information, but also the well-known $\chi^2$ measure satisfies these conditions, so that it may be used for the same task. However, for mutual information also a stronger statement holds, namely that if there is no tree-structured perfect map, constructing a maximum weight spanning tree yields the best tree-structured approximation w.r.t. the Kullback-Leibler information divergence [8] between the original distribution and the distribution represented by the tree [3, 10]. This result even generalizes to the construction of tree-augmented naive Bayes classifiers, where the star-like structure of such a classifier is augmented by edges that form a tree [5, 4]. To find out whether a similar result can be obtained for the $\chi^2$ measure, for example, w.r.t. the difference between the original distribution and the approximation as it can be measured by an adapted $\chi^2$ measure itself, remains as future work.

## References

1. C. Borgelt and R. Kruse. *Graphical Models — Methods for Data Analysis and Mining.* J. Wiley & Sons, Chichester, United Kingdom 2002
2. E. Castillo, J.M. Gutierrez, and A.S. Hadi. *Expert Systems and Probabilistic Network Models.* Springer-Verlag, New York, NY, USA 1997
3. C.K. Chow and C.N. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Trans. on Information Theory* 14(3):462–467. IEEE Press, Piscataway, NJ, USA 1968
4. N. Friedman and M. Goldszmidt. Building Classifiers using Bayesian Networks. *Proc. 13th Nat. Conf. on Artificial Intelligence (AAAI'96, Portland, OR, USA)*, 1277–1284. AAAI Press, Menlo Park, CA, USA 1996
5. D. Geiger. An entropy-based learning algorithm of Bayesian conditional trees. *Proc. 8th Conf. on Uncertainty in Artificial Intelligence (UAI'92, Stanford, CA, USA)*, 92–97. Morgan Kaufmann, San Mateo, CA, USA 1992
6. F.V. Jensen. *An Introduction to Bayesian Networks.* UCL Press, London, United Kingdom 1996
7. J.B. Kruskal. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. American Mathematical Society* 7(1):48–50. American Mathematical Society, Providence, RI, USA 1956

8. S. Kullback and R.A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics* 22:79–86. Institute of Mathematical Statistics, Hayward, CA, USA 1951
9. S.L. Lauritzen. *Graphical Models.* Oxford University Press, Oxford, United Kingdom 1996
10. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, USA 1988 (2nd edition 1992)
11. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C — The Art of Scientific Computing (2nd edition).* Cambridge University Press, Cambridge, United Kingdom 1992
12. R.C. Prim. Shortest Connection Networks and Some Generalizations. *The Bell System Technical Journal* 36:1389-1401. Bell Laboratories, Murray Hill, NJ, USA 1957
13. J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, USA 1993
14. G. Rebane and J. Pearl. The Recovery of Causal Polytrees from Statistical Data. *Proc. 3rd Workshop on Uncertainty in Artificial Intelligence (Seattle, WA, USA)*, 222–228. USA 1987.
15. C.E. Shannon. The Mathematical Theory of Communication. *The Bell System Technical Journal* 27:379–423. Bell Laboratories, Murray Hill, NJ, USA 1948