# Likelihood-based inference for probabilistic graphical models: Some preliminary results

Marco E. G. V. Cattaneo
Department of Statistics, LMU Munich, Germany
cattaneo@stat.uni-muenchen.de

**Abstract**

A method for calculating some profile likelihood inferences in probabilistic graphical models is presented and applied to the problem of classification. It can also be interpreted as a method for obtaining inferences from hierarchical networks, a kind of imprecise probabilistic graphical models.

## 1 Introduction

The main result of the present paper is a method for calculating profile likelihood functions for an important class of probabilistic inferences, when the probabilities of a Bayesian network are learned from data. This result can also be interpreted as a method for obtaining inferences from hierarchical networks, which are networks with imprecisely known probabilities.

Likelihood-based inference is briefly outlined in the next section, while Section 3 contains the main result, stated in Theorem 1 (whose proof will be given in an extended version of the paper). Section 4 presents an application of this result to the problem of classification.

## 2 Likelihood

Let $P_\theta$ be a parametric probabilistic model for some discrete random variables $X, Y, \ldots$, where $\theta \in \Theta$ is the parameter (vector) and $\Theta$ is the parameter space. The observation of a realization $X = x$ induces the (normalized) likelihood function $lik$ on $\Theta$ defined by

$$lik(\theta) = \frac{P_\theta(X = x)}{\sup_{\theta' \in \Theta} P_{\theta'}(X = x)}$$

for all $\theta \in \Theta$. Moreover, when $X = x$ is observed, the model $P_\theta$ is updated into the conditional model $P_\theta(\,\cdot \,|\, X = x)$.

The inference about the value $g(\theta)$ of a function $g : \Theta \to \mathcal{G}$ (where $\mathcal{G}$ can be any set) can be based on the (normalized) profile likelihood function $lik_g$ on $\mathcal{G}$ defined by

$$lik_g(\gamma) = \sup_{\theta \in \Theta \,:\, g(\theta) = \gamma} lik(\theta)$$

for all $\gamma \in \mathcal{G}$, where $\sup \varnothing$ is interpreted as 0. In particular, if there is a unique $\gamma \in \mathcal{G}$ such that $lik_g(\gamma) = 1$, then $\gamma$ is the maximum likelihood estimate of $g(\theta)$. More generally, the likelihood-based confidence region for $g(\theta)$ with cutoff point $\beta \in [0, 1[$ is the set

$$\{\gamma \in \mathcal{G} : lik_g(\gamma) > \beta\},$$

whose confidence level can often be approximated thanks to the result of (Wilks, 1938). These are the usual likelihood-based point estimates and set estimates; more general ways of basing inferences and decisions directly on the likelihood function are discussed in (Cattaneo, 2007).

**Example 1.** Let $X_1, \ldots, X_{10}$ be 10 categorical variables taking values in the set $\{a, b, c\}$. The assumption that $X_1, \ldots, X_{10}$ are independent and identically distributed leads to a parametric probabilistic model $P_\theta$ with $P_\theta(X_i = \omega) = \theta_\omega$, where the parameter $\theta = (\theta_a, \theta_b, \theta_c)$ is an element of the standard 2-dimensional simplex

$$\Theta = \left\{(\theta_a, \theta_b, \theta_c) \in [0, 1]^3 : \theta_a + \theta_b + \theta_c = 1)\right\}.$$

Assume that the realizations of $X_1, \ldots, X_9$ are observed, and the values $a$, $b$, and $c$ appear 2, 4, and 3 times, respectively. This observation

induces the (normalized) multinomial likelihood function $lik$ on $\Theta$ defined by

$$lik(\theta) = \tfrac{14348907}{1024}\, \theta_a^2\, \theta_b^4\, \theta_c^3$$

for all $\theta \in \Theta$. The inference about the probability that the realization of $X_{10}$ will be either $a$ or $b$ can be based on the profile likelihood function $lik_g$ on $[0,1]$, with $g : \Theta \to [0,1]$ defined by

$$g(\theta) = P_\theta\left(X_{10} \in \{a,b\}\right) = \theta_a + \theta_b$$

for all $\theta \in \Theta$ (note that conditioning $P_\theta$ on the observed realizations of $X_1, \ldots, X_9$ has no influence on the probability distribution of $X_{10}$). Since the (normalized) profile likelihood function $lik_g$ on $[0,1]$ satisfies

$$lik_g(\gamma) = \tfrac{19683}{64}\, \gamma^6\, (1-\gamma)^3$$

for all $\gamma \in [0,1]$, the maximum likelihood estimate of $P_\theta\left(X_{10} \in \{a,b\}\right)$ is $\widehat{\gamma} = \frac{2}{3}$, while for instance $[0.35, 0.90]$ is an approximate 95% confidence interval for $P_\theta\left(X_{10} \in \{a,b\}\right)$, since it corresponds approximately to the likelihood-based confidence region with cutoff point $\beta = 0.15$.

### 2.1 Hierarchical model

The parametric probabilistic model and the likelihood function can be considered as the two levels of a hierarchical model, in which the likelihood function describes the relative plausibility of the parameter values. As noted above, when $X = x$ is observed, the set $\{P_\theta : \theta \in \Theta\}$ is updated by conditioning each element $P_\theta$ on the observed event: this corresponds to the updating of the imprecise Bayesian model studied in (Walley, 1991). Hence, the hierarchical model generalizes the imprecise Bayesian model, in the sense that the second level (that is, the likelihood function) describes additional information about the relative plausibility of the elements of $\{P_\theta : \theta \in \Theta\}$.

This additional information allows fundamental advantages of the hierarchical model over the imprecise Bayesian model, such as the possibility of starting without prior information and the increased robustness of the conclusions: see for example (Cattaneo, 2009). Moreover, since the membership functions of fuzzy sets are often interpreted as likelihood functions (the extension principle of possibility theory corresponds then to the use of profile likelihood functions), the hierarchical model can handle fuzzy data and possibilistic information or variables in a unified and well-founded way: see for instance (Cattaneo, 2008).

## 3 Networks

When $X$ is a categorical variable, let $\Omega_X$ denote the set of all possible realizations of $X$. Moreover, let $\mathcal{F}_X$ denote the set of all possible real functions on $\Omega_X$, and let $\mathcal{S}_X$ denote the set of all possible probability distributions on $\Omega_X$. Hence, $\mathcal{S}_X \subset \mathcal{F}_X$, and $\mathcal{S}_X$ can be identified with the standard simplex of dimension $|\Omega_X| - 1$, where $|\Omega_X|$ denotes the cardinality of $\Omega_X$. Moreover, let $0_X$ denote the function on $\Omega_X$ with constant value 0 (therefore, $0_X \in \mathcal{F}_X$, but $0_X \notin \mathcal{S}_X$). Finally, if $f \in \mathcal{F}_X$ with $f(x) \geqslant 0$ for all $x \in \Omega_X$, then let $\langle f \rangle$ denote the probability distribution on $\Omega_X$ proportional to $f$ when $f \neq 0_X$, and the uniform probability distribution on $\Omega_X$ when $f = 0_X$. That is, $\langle f \rangle \in \mathcal{S}_X$, and for all $x \in \Omega_X$,

$$\langle f \rangle(x) = \begin{cases} \frac{f(x)}{\sum_{x' \in \Omega_X} f(x')} & \text{if } \sum_{x' \in \Omega_X} f(x') > 0, \\ \frac{1}{|\Omega_X|} & \text{if } \sum_{x' \in \Omega_X} f(x') = 0. \end{cases}$$

Let $X_1, \ldots, X_k$ be $k$ categorical variables such that $|\Omega_{X_i}| \geqslant 2$ for all $i \in \{1, \ldots, k\}$. Assumptions about conditional independencies among the variables $X_1, \ldots, X_k$ can be encoded in a directed acyclic graph $G$ with nodes $X_1, \ldots, X_k$: see for example (Jensen and Nielsen, 2007). Let $\Pi_i$ denote the joint variable composed of all parents of $X_i$ according to $G$, where $\Pi_i$ is assumed to be constant when $X_i$ has no parents. The other component of a Bayesian network, besides the graph $G$, are the probability distributions of $X_i$ conditional on $\Pi_i = \pi_i$, for each $i \in \{1, \ldots, k\}$ and each $\pi_i \in \Omega_{\Pi_i}$. Altogether, these conditional probability distributions can be described by the parameter $\theta \in \Theta_G$, where

$$\Theta_G = \bigtimes_{i=1}^{k} \bigtimes_{\pi_i \in \Omega_{\Pi_i}} \mathcal{S}_{X_i}$$

is a Cartesian product of the simplexes $\mathcal{S}_{X_i}$. For each $\theta \in \Theta_G$, let $\theta_{X_i|\pi_i}$ denote the corresponding probability distribution of $X_i$ conditional on $\Pi_i = \pi_i$ (hence, $\theta_{X_i|\pi_i} \in \mathcal{S}_{X_i}$). The Bayesian network described by the graph $G$ and the parameter $\theta \in \Theta_G$ determines the joint probability distribution $P_\theta$ on $\Omega_{X_1} \times \cdots \times \Omega_{X_k}$ defined by

$$P_\theta(x_1, \ldots, x_k) = \prod_{i=1}^{k} \theta_{X_i|\pi_i}(x_i)$$

for all $(x_1, \ldots, x_k) \in \Omega_{X_1} \times \cdots \times \Omega_{X_k}$, where $\pi_i$ are the corresponding realizations of the joint variables $\Pi_i$.

To allow uncertainty about the involved probability values, Bayesian networks have been generalized to credal networks, which can be described by a directed acyclic graph $G$ and a set $\Theta \subseteq \Theta_G$ of parameters: see for instance (Antonucci and Zaffalon, 2008). A credal network determines an imprecise Bayesian model $\{P_\theta : \theta \in \Theta\}$, instead of a single probability distribution $P_\theta$. That is, a credal network corresponds mathematically to a set of Bayesian networks with the same graph $G$. A credal network is said to be separately specified if

$$\Theta = \bigtimes_{i=1}^{k} \bigtimes_{\pi_i \in \Omega_{\Pi_i}} \Theta_{X_i|\pi_i} \tag{1}$$

is the Cartesian product of the sets $\Theta_{X_i|\pi_i}$, with $\Theta_{X_i|\pi_i} \subseteq \mathcal{S}_{X_i}$ for all $i \in \{1, \ldots, k\}$ and all $\pi_i \in \Omega_{\Pi_i}$. That is, a separately specified credal network consists of all Bayesian networks with graph $G$ and probability distributions of $X_i$ conditional on $\Pi_i = \pi_i$ freely selected from the sets $\Theta_{X_i|\pi_i}$ (note that only the so-called strong extension of a separately specified credal network is considered in the present paper).

To allow additional information about the relative plausibility of the involved probability values, credal networks have been generalized to hierarchical networks, which can be described by a directed acyclic graph $G$, a set $\Theta \subseteq \Theta_G$ of parameters, and a (normalized) likelihood function $lik$ on $\Theta$: see for example (Cattaneo, 2009). A hierarchical network determines a hierarchical model with as first level the imprecise Bayesian model $\{P_\theta : \theta \in \Theta\}$, and as second level the likelihood function $lik$ on $\Theta$, describing the relative plausibility of the elements of $\{P_\theta : \theta \in \Theta\}$. That is, a hierarchical network corresponds mathematically to a set of Bayesian networks with the same graph $G$ but in general with different degrees of plausibility. A hierarchical network is said to be separately specified if $\Theta$ satisfies (1), and $lik$ is the product of the local likelihood functions $lik_{X_i|\pi_i}$ on $\Theta_{X_i|\pi_i}$, in the sense that

$$lik(\theta) = \prod_{i=1}^{k} \prod_{\pi_i \in \Omega_{\Pi_i}} lik_{X_i|\pi_i}(\theta_{X_i|\pi_i})$$

for all $\theta \in \Theta$, with $\Theta_{X_i|\pi_i} \subseteq \mathcal{S}_{X_i}$ and $lik_{X_i|\pi_i} : \Theta_{X_i|\pi_i} \to [0, 1]$ for all $i \in \{1, \ldots, k\}$ and all $\pi_i \in \Omega_{\Pi_i}$. That is, the separately specified hierarchical networks generalize the separately specified credal networks by adding information about the relative plausibility of the elements of the sets $\Theta_{X_i|\pi_i}$.

### 3.1 Learning probabilities from data

Learning networks from data is a fundamental problem. In the present paper, only the simplest case is considered: the directed acyclic graph $G$ is assumed known, and the dataset is complete. That is, the dataset consists of $n$ realizations of the joint variable $X = (X_1, \ldots, X_k)$. For each $i \in \{1, \ldots, k\}$ and each $\pi_i \in \Omega_{\Pi_i}$, let $n_{X_i|\pi_i}$ denote the function on $\Omega_{X_i}$ assigning to each $x_i \in \Omega_{X_i}$ the number of realizations of the joint variable $X$ such that $X_i = x_i$ and $\Pi_i = \pi_i$. Hence, $n_{X_i|\pi_i} \in \mathcal{F}_{X_i}$, and for all $i \in \{1, \ldots, k\}$,

$$\sum_{\pi_i \in \Omega_{\Pi_i}} \sum_{x_i \in \Omega_{X_i}} n_{X_i|\pi_i}(x_i) = n. \tag{2}$$

When the $n$ realizations of the joint variable $X$ are considered independent and identically distributed according to the joint probability distribution $P_\theta$ with $\theta \in \Theta_G$, they induce the (normalized) likelihood function $lik$ on $\Theta_G$ defined by

$$lik(\theta) = \prod_{i=1}^{k} \prod_{\pi_i \in \Omega_{\Pi_i}} \prod_{x_i \in \Omega_{X_i}} \frac{\big(\theta_{X_i|\pi_i}(x_i)\big)^{n_{X_i|\pi_i}(x_i)}}{\big(\langle n_{X_i|\pi_i} \rangle(x_i)\big)^{n_{X_i|\pi_i}(x_i)}}$$

for all $\theta \in \Theta_G$, where $0^0$ is interpreted as 1. The denominators of the fractions normalize the likelihood function, since $lik$ is maximized by the parameter $\widehat{\theta} \in \Theta_G$ such that $\widehat{\theta}_{X_i|\pi_i} = \langle n_{X_i|\pi_i} \rangle$ for all $i \in \{1, \ldots, k\}$ and all $\pi_i \in \Omega_{\Pi_i}$. However, $\widehat{\theta}$ is the unique parameter maximizing $lik$ only if $n_{X_i|\pi_i} \neq 0_{X_i}$ for all $i \in \{1, \ldots, k\}$ and all $\pi_i \in \Omega_{\Pi_i}$ (that is, only if all possible realizations $\Pi_i = \pi_i$ appear at least once in the dataset).

Hence, the likelihood function $lik$ on $\Theta_G$ factorizes in the local likelihood functions $lik_{X_i|\pi_i}$ on $\mathcal{S}_{X_i}$ defined by

$$lik_{X_i|\pi_i}(\theta_{X_i|\pi_i}) = \prod_{x_i \in \Omega_{X_i}} \frac{\big(\theta_{X_i|\pi_i}(x_i)\big)^{n_{X_i|\pi_i}(x_i)}}{\big(\langle n_{X_i|\pi_i} \rangle(x_i)\big)^{n_{X_i|\pi_i}(x_i)}}$$

for all $\theta_{X_i|\pi_i} \in \mathcal{S}_{X_i}$. Estimates of the conditional probability distributions $\theta_{X_i|\pi_i}$ can easily be based on the multinomial likelihood functions $lik_{X_i|\pi_i}$: if $n_{X_i|\pi_i} \neq 0$, then $\widehat{\theta}_{X_i|\pi_i} = \langle n_{X_i|\pi_i} \rangle$ is the maximum likelihood estimate; alternatively, $lik_{X_i|\pi_i}$ can be combined with a prior probability distribution on $\mathcal{S}_{X_i}$ (usually a Dirichlet distribution) to obtain a Bayesian estimate of $\theta_{X_i|\pi_i}$. However, the Bayesian network corresponding to the estimated conditional probability distributions does not contain any information about the uncertainty of those estimates, and consequently it does not contain any information about the uncertainty of the resulting probabilistic inferences.

To include some information about the uncertainty of the conditional probability distributions and of the resulting probabilistic inferences, set estimates $\Theta_{X_i|\pi_i} \subseteq \mathcal{S}_{X_i}$ of the conditional probability distributions $\theta_{X_i|\pi_i}$ can be based on the local likelihood functions $lik_{X_i|\pi_i}$ (instead of point estimates $\theta_{X_i|\pi_i} \in \mathcal{S}_{X_i}$). The set estimates $\Theta_{X_i|\pi_i}$ determine a separately specified credal network, and inferences can then be based on the corresponding imprecise Bayesian model. The usual way of obtaining an imprecise probability distribution $\Theta_{X_i|\pi_i}$ from a multinomial likelihood function $lik_{X_i|\pi_i}$ is by combining it with a set of prior Dirichlet distributions, called imprecise Dirichlet model: see for example (Walley, 1996). But the confidence level of the set estimates $\Theta_{X_i|\pi_i}$ obtained from the imprecise Dirichlet model can be arbitrarily low (for sufficiently large $n$: compare with Example 2): see for instance Wilson's comment in the discussion of (Walley, 1996). To avoid this problem, the sets $\Theta_{X_i|\pi_i} \subseteq \mathcal{S}_{X_i}$ could be estimated as likelihood-based confidence regions for $\theta_{X_i|\pi_i}$, according to the multinomial likelihood functions $lik_{X_i|\pi_i}$, but in general the closure of the resulting set estimates $\Theta_{X_i|\pi_i}$ would be convex with infinitely many extreme points when $|\Omega_{X_i}| \geqslant 3$, and this would lead to computational difficulties.

Instead of reducing them to likelihood-based confidence regions $\Theta_{X_i|\pi_i}$, it is better to maintain the whole likelihood functions $lik_{X_i|\pi_i}$ as descriptions of the uncertainty about the conditional probability distributions $\theta_{X_i|\pi_i}$. The likelihood function $lik$ on $\Theta_G$ describes then the uncertainty about the whole Bayesian network (given the graph $G$), and corresponds to a separately specified hierarchical network with $\Theta = \Theta_G$. Learning hierarchical networks from data is straightforward (when the graph $G$ is assumed known), and no estimates or prior distributions are necessary, but in general the calculation of profile likelihood functions (on which inferences and decisions are based) is not so simple. However, the following theorem shows that for particular classes of functions $g$ on $\Theta$, obtaining the profile likelihood function $lik_g$ is straightforward too.

**Theorem 1.** *For each $i \in \{1, \ldots, k\}$ and each $\pi_i \in \Omega_{\Pi_i}$, let $d_{X_i|\pi_i}, q_{X_i|\pi_i} \in \mathcal{F}_{X_i}$ with $d_{X_i|\pi_i}(x_i) \geqslant 0$ for all $x_i \in \Omega_{X_i}$. Moreover, let $lik : \Theta_G \to [0,1]$ and $g : \Theta_G \to [0, +\infty]$ be defined by*

$$lik(\theta) = \prod_{i=1}^{k} \prod_{\pi_i \in \Omega_{\Pi_i}} \prod_{x_i \in \Omega_{X_i}} \frac{\big(\theta_{X_i|\pi_i}(x_i)\big)^{d_{X_i|\pi_i}(x_i)}}{\big(\langle d_{X_i|\pi_i} \rangle(x_i)\big)^{d_{X_i|\pi_i}(x_i)}}$$

*and*

$$g(\theta) = \prod_{i=1}^{k} \prod_{\pi_i \in \Omega_{\Pi_i}} \prod_{x_i \in \Omega_{X_i}} \big(\theta_{X_i|\pi_i}(x_i)\big)^{q_{X_i|\pi_i}(x_i)},$$

*respectively, for all $\theta \in \Theta_G$, where $0^0$ is interpreted as 1, and $0^x$ is interpreted as $+\infty$ for all*

*negative $x$ (but $g(\theta)$ is undefined when both 0 and $+\infty$ appear in the same product). Finally, let $\underline{\alpha}$ and $\overline{\alpha}$ be the infimum and the supremum, respectively, of the set*

$$\left\{\alpha \in \mathbb{R} : (d_{X_i|\pi_i} + \alpha\, q_{X_i|\pi_i})(x_i) \geqslant 0 \ \forall\, i, \pi_i, x_i\right\}$$

*(linear combinations of functions are to be interpreted pointwise).*

- *If $\underline{\alpha} = \overline{\alpha} = 0$, then $lik_g(\gamma) = 1$ for all $\gamma \in [0, +\infty]$.*
- *Otherwise, define $\underline{\theta}, \theta[\alpha], \overline{\theta} \in \Theta_G$ as follows:*

$$\underline{\theta}_{X_i|\pi_i} = \begin{cases} \langle d_{X_i|\pi_i} + \underline{\alpha}\, q_{X_i|\pi_i}\rangle & \text{if } \underline{\alpha} \neq -\infty \\ \quad \text{and } d_{X_i|\pi_i} + \underline{\alpha}\, q_{X_i|\pi_i} \neq 0_{X_i}, & \\ \langle q_{X_i|\pi_i}\rangle & \text{if } \underline{\alpha} \neq -\infty \\ \quad \text{and } d_{X_i|\pi_i} + \underline{\alpha}\, q_{X_i|\pi_i} = 0_{X_i}, & \\ \langle -q_{X_i|\pi_i}\rangle & \text{if } \underline{\alpha} = -\infty \\ & \text{and } q_{X_i|\pi_i} \neq 0_{X_i}, \\ \langle d_{X_i|\pi_i}\rangle & \text{if } \underline{\alpha} = -\infty \\ & \text{and } q_{X_i|\pi_i} = 0_{X_i}, \end{cases}$$

$$\theta[\alpha]_{X_i|\pi_i} = \langle d_{X_i|\pi_i} + \alpha\, q_{X_i|\pi_i}\rangle,$$

$$\overline{\theta}_{X_i|\pi_i} = \begin{cases} \langle d_{X_i|\pi_i} + \overline{\alpha}\, q_{X_i|\pi_i}\rangle & \text{if } \overline{\alpha} \neq +\infty \\ \quad \text{and } d_{X_i|\pi_i} + \overline{\alpha}\, q_{X_i|\pi_i} \neq 0_{X_i}, & \\ \langle -q_{X_i|\pi_i}\rangle & \text{if } \overline{\alpha} \neq +\infty \\ \quad \text{and } d_{X_i|\pi_i} + \overline{\alpha}\, q_{X_i|\pi_i} = 0_{X_i}, & \\ \langle q_{X_i|\pi_i}\rangle & \text{if } \overline{\alpha} = +\infty \\ & \text{and } q_{X_i|\pi_i} \neq 0_{X_i}, \\ \langle d_{X_i|\pi_i}\rangle & \text{if } \overline{\alpha} = +\infty \\ & \text{and } q_{X_i|\pi_i} = 0_{X_i}, \end{cases}$$

*respectively, for all $\alpha \in ]\underline{\alpha}, \overline{\alpha}[$, all $i \in \{1, \ldots, k\}$, and all $\pi_i \in \Omega_{\Pi_i}$.*

*Then $lik_g(g(\underline{\theta})) = lik(\underline{\theta})$ and $lik_g(g(\overline{\theta})) = lik(\overline{\theta})$.*

*If $g(\underline{\theta}) > 0$, then for all $\gamma \in [0, g(\underline{\theta})[$,*

$$lik_g(\gamma) = \begin{cases} \left(\frac{\gamma}{g(\underline{\theta})}\right)^{-\underline{\alpha}} lik(\underline{\theta}) & \text{if } \underline{\alpha} \neq -\infty, \\ 0 & \text{if } \underline{\alpha} = -\infty. \end{cases}$$

*If $g(\underline{\theta}) < g(\overline{\theta})$, then the graph of the restriction of $lik_g$ to $]g(\underline{\theta}), g(\overline{\theta})[$ is the set*

$$\{(g(\theta[\alpha]), lik(\theta[\alpha])) : \alpha \in ]\underline{\alpha}, \overline{\alpha}[\,\},$$

*and $g(\theta[\alpha])$ is a continuous, strictly increasing function of $\alpha \in ]\underline{\alpha}, \overline{\alpha}[$.*

*If $g(\overline{\theta}) < +\infty$, then for all $\gamma \in ]g(\overline{\theta}), +\infty]$,*

$$lik_g(\gamma) = \begin{cases} \left(\frac{\gamma}{g(\overline{\theta})}\right)^{-\overline{\alpha}} lik(\overline{\theta}) & \text{if } \overline{\alpha} \neq +\infty, \\ 0 & \text{if } \overline{\alpha} = +\infty. \end{cases}$$

The likelihood function $lik$ of Theorem 1 generalizes the likelihood function on $\Theta_G$ induced by a complete dataset, for which the functions $d_{X_i|\pi_i} = n_{X_i|\pi_i}$ can take only integer values and must satisfy conditions such as (2). The function $g$ of Theorem 1 can for example describe the probability of a particular realization $(x_1, \ldots, x_k) \in \Omega_{X_1} \times \cdots \times \Omega_{X_k}$ of the joint variable $X = (X_1, \ldots, X_k)$; that is,

$$g(\theta) = P_\theta(X_1 = x_1, \ldots, X_k = x_k)$$

for all $\theta \in \Theta_G$. In this case, $q_{X_i|\pi_i} = n_{X_i|\pi_i}$ for the particular dataset consisting of the single realization $(x_1, \ldots, x_k)$ of the joint variable $X$. If the functions $n'_{X_i|\pi_i}$ describe a second dataset consisting of the single realization $(x'_1, x_2, \ldots, x_k)$ of the joint variable $X$, then the function $g$ with $q_{X_i|\pi_i} = n_{X_i|\pi_i} - n'_{X_i|\pi_i}$ satisfies

$$g(\theta) = \frac{P_\theta(X_1 = x_1 \,|\, X_2 = x_2, \ldots, X_k = x_k)}{P_\theta(X_1 = x'_1 \,|\, X_2 = x_2, \ldots, X_k = x_k)}$$

for all $\theta \in \Theta_G$ such that the right-hand side is well-defined. That is, $g$ describes the probability ratio of the possible realizations $x_1$ and $x'_1$ of $X_1$ conditional on the realizations of $X_2, \ldots, X_k$. In the next section, Theorem 1 with this kind of function $g$ is used in the problem of classification: the goal is to determine the realization of $X_1$ given the realizations of $X_2, \ldots, X_k$.

The formulation of Theorem 1 is rather complex, because several special cases must be considered, but the central part of the theorem is pretty simple: it is the parametric expression for the graph of the profile likelihood function $lik_g$ restricted to the interval $]g(\underline{\theta}), g(\overline{\theta})[$. For example, in the problem of classification studied in the next section, it suffices to consider this central part, since $]g(\underline{\theta}), g(\overline{\theta})[ = ]0, +\infty[$.

The idea behind the parametric expression of the graph of $lik_g$ is the following: if $\theta[\alpha]$ maximizes $(g(\theta))^\alpha\, lik(\theta)$ over all $\theta \in \Theta_G$ for some $\alpha \in \mathbb{R}$, then $\theta[\alpha]$ maximizes $lik(\theta)$ over all $\theta \in \Theta_G$ such that $g(\theta) = g(\theta[\alpha])$, and therefore $lik_g(g(\theta[\alpha])) = lik(\theta[\alpha])$. For the particular classes of functions $lik$ and $g$ considered in Theorem 1, finding the parameter $\theta[\alpha]$ maximizing $(g(\theta))^\alpha\, lik(\theta)$ over all $\theta \in \Theta_G$ is extremely simple. For more general classes of functions $lik$ and $g$, the above idea can be combined with approximation algorithms for maximizing $(g(\theta))^\alpha\, lik(\theta)$, such as the EM algorithm of (Dempster et al., 1977), but this goes beyond the scope of the present paper.

## 4 Naive classifiers

Let $C, F_1, \ldots, F_{k-1}$ be $k$ categorical variables. The variables $F_1, \ldots, F_{k-1}$ describe $k-1$ features of an object, while $C$ is the variable of interest: it describes the object's class. Having observed $m$ features of an object, say $F_1 = f_1, \ldots, F_m = f_m$, with $m \in \{0, \ldots, k-1\}$, the goal is to classify it; that is, to predict the realization of $C$. The problem is particularly simple if the features $F_1, \ldots, F_{k-1}$ are assumed to be conditionally independent given the class $C$. This assumption can be encoded in the directed acyclic graph $G_N$ with nodes $C, F_1, \ldots, F_{k-1}$ such that $C$ has no parents and is the only parent of $F_1, \ldots, F_{k-1}$.

The Bayesian network described by the graph $G_N$ and a parameter $\theta \in \Theta_{G_N}$ is called naive Bayes classifier (NBC): such classifiers were proposed in (Duda and Hart, 1973). For each pair of different classes $a, b \in \Omega_C$, let $g_{a,b} : \Theta_{G_N} \to [0, +\infty]$ be defined by

$$g_{a,b}(\theta) = \frac{P_\theta(C = a, F_1 = f_1, \ldots, F_m = f_m)}{P_\theta(C = b, F_1 = f_1, \ldots, F_m = f_m)}$$

for all $\theta \in \Theta_G$, where $\frac{x}{0}$ is interpreted as $+\infty$ for all positive $x$, and as 1 when $x = 0$. A strict partial preference order $\succ$ on $\Omega_C$ is obtained by considering the values $g_{a,b}(\theta)$, for the parameter $\theta$ of the Bayesian network and all pairs of different classes $a, b \in \Omega_C$: if $g_{a,b}(\theta) > 1$, then $a \succ b$ (that is, $a$ is preferred to $b$), while if $g_{a,b}(\theta) < 1$, then $b \succ a$; finally, if $g_{a,b}(\theta) = 1$, then there is no preference between $a$ and $b$. The NBC returns as prediction of $C$ the maximal elements of $\Omega_C$ according to $\succ$ (that is, the $c \in \Omega_C$ such that there is no $c' \in \Omega_C$ with $c' \succ c$). Usually the prediction consists of a single class, but sometimes it can consist of several classes (with no preference among them). The parameter $\theta$ of the Bayesian network can be estimated from training data (for example by maximum likelihood estimation: see Subsection 3.1), but the resulting NBC does not contain any information about the uncertainty of the estimate $\theta$ and of the inferred values $g_{a,b}(\theta)$.

The credal network described by the graph $G_N$ and a set $\Theta \subseteq \Theta_{G_N}$ of parameters is called naive credal classifier (NCC): such classifiers were proposed in (Zaffalon, 2002). A strict partial preference order $\succ$ on $\Omega_C$ (called credal dominance) is obtained by considering the values $g_{a,b}(\theta)$, for all parameters $\theta \in \Theta$ and all pairs of different classes $a, b \in \Omega_C$: there is a preference between $a$ and $b$ only if either $g_{a,b}(\theta) > 1$ for all $\theta \in \Theta$ (in which case $a \succ b$), or $g_{a,b}(\theta) < 1$ for all $\theta \in \Theta$ (in which case $b \succ a$). The NCC returns as prediction of $C$ the maximal elements of $\Omega_C$ according to $\succ$; hence, the prediction often consists of more than one class. The set $\Theta$ of parameters can be estimated from training data (for example on the basis of the imprecise Dirichlet model: see Subsection 3.1): the resulting NCC contains some information about the uncertainty of the inferred values $g_{a,b}(\theta)$, and the number of classes returned as prediction of $C$ depends on the amount of uncertainty (the more uncertainty, the more classes).

The hierarchical network described by the graph $G_N$ and a (normalized) likelihood function $lik$ on $\Theta_{G_N}$ can be called naive hierarchical classifier (NHC). For each $\beta \in [0, 1[$, a strict partial preference order $\succ_\beta$ on $\Omega_C$ is obtained by considering the profile likelihood functions $lik_{g_{a,b}}$ on $[0, +\infty]$, for all pairs of different classes $a, b \in \Omega_C$: there is a preference between $a$ and $b$ only if either $lik_{g_{a,b}}(\gamma) \leqslant \beta$ for all $\gamma \in [0, 1]$ (in which case $a \succ_\beta b$), or $lik_{g_{a,b}}(\gamma) \leqslant \beta$ for all $\gamma \in [1, +\infty]$ (in which case $b \succ_\beta a$). The NHC returns as prediction of $C$ with cutoff point $\beta$

the maximal elements of $\Omega_C$ according to $\succ_\beta$. Hence, the prediction can consist of one or more classes, and the number of classes increases as $\beta$ decreases, in the sense that additional classes can be included in the prediction as $\beta$ decreases. The likelihood function $lik$ on $\Theta_{G_N}$ can be induced by training data, and when $lik$ satisfies the condition of Theorem 1, the profile likelihood functions $lik_{g_{a,b}}$ are easily obtained. In order to satisfy that condition, it is not necessary for the training dataset to be complete (the features of the objects in the dataset need not be observed), but when it is complete, Theorem 1 implies the following simple result.

**Corollary 1.** *Let $a, b \in \Omega_C$ be two different classes, and for both $c \in \{a, b\}$ and each $i \in \{1, \ldots, m\}$, let $n_c$ and $n_{c,i}$ be the numbers of objects in the complete training dataset with $C = c$, and with $C = c$ and $F_i = f_i$, respectively. Moreover, define*

$$\underline{\alpha} = -\min\{n_a, n_{a,1}, \ldots, n_{a,m}\}$$

*and*

$$\overline{\alpha} = \min\{n_b, n_{b,1}, \ldots, n_{b,m}\}.$$

- *If $\underline{\alpha} = \overline{\alpha} = 0$, then $lik_{g_{a,b}}(\gamma) = 1$ for all $\gamma \in [0, +\infty]$.*
- *Otherwise, let $x_{a,b}, y_a, y_b : [\underline{\alpha}, \overline{\alpha}] \to [0, +\infty]$ be defined by*

$$x_{a,b}(\alpha) = \frac{n_a + \alpha}{n_b - \alpha} \prod_{i=1}^{m} \left( \frac{n_{a,i} + \alpha}{n_a + \alpha} \frac{n_b - \alpha}{n_{b,i} - \alpha} \right),$$

$$y_a(\alpha) = \frac{(n_a + \alpha)^{n_a}}{{n_a}^{n_a}} \prod_{i=1}^{m} \frac{{n_a}^{n_a} (n_{a,i} + \alpha)^{n_{a,i}}}{{n_{a,i}}^{n_{a,i}} (n_a + \alpha)^{n_a}},$$

$$y_b(\alpha) = \frac{(n_b - \alpha)^{n_b}}{{n_b}^{n_b}} \prod_{i=1}^{m} \frac{{n_b}^{n_b} (n_{b,i} - \alpha)^{n_{b,i}}}{{n_{b,i}}^{n_{b,i}} (n_b - \alpha)^{n_b}},$$

*respectively, for all $\alpha \in [\underline{\alpha}, \overline{\alpha}]$, where $0^0$ is interpreted as 1, and $\frac{x}{0}$ is interpreted as $+\infty$ for all positive $x$, and as 1 when $x = 0$.*

*Then $x_{a,b}$ is an increasing bijection, and the graph of $lik_{g_{a,b}}$ is the set*

$$\{(x_{a,b}(\alpha),\, y_a(\alpha)\, y_b(\alpha)) : \alpha \in [\underline{\alpha}, \overline{\alpha}]\}.$$

If the NHC is learned from training data, then for sufficiently large $\beta \in [0, 1[$, the predictions with cutoff point $\beta$ correspond to the ones returned by the NBC based on maximum likelihood estimation (if this is well-defined). But as $\beta$ decreases, more and more classes are included in the predictions with cutoff point $\beta$; and for sufficiently small $\beta$, the predictions are vacuous, in the sense that they consist of all possible classes. Hence, the NHC learned from training data can be interpreted as a description of the uncertainty about the NBC based on maximum likelihood estimation: when the cutoff point $\beta \in ]0, 1[$ is fixed, the numbers of classes in the predictions depend on the amount of uncertainty (the more uncertainty, the more classes). In particular, if $c$ is the prediction of $C$ returned by the NBC, then $\beta_c = \max_{c' \in C \setminus \{c\}} lik_{g_{c,c'}}(1)$ is the minimum value of $\beta \in ]0, 1[$ such that the prediction of $C$ with cutoff point $\beta$ returned by the NHC is $c$ as well. Therefore, $\beta_c$ is an index of the uncertainty about the prediction $c$: the larger $\beta_c$, the more uncertainty; in fact, $\beta_c$ is the likelihood ratio test statistic for the set of all parameters $\theta \in \Theta_{G_N}$ such that the corresponding NBC does not return $c$ as prediction of $C$: see for instance (Wilks, 1938).

The strict partial preference order $\succ_\beta$ for the NHC with likelihood function $lik$ on $\Theta_{G_N}$ corresponds to credal dominance for the NCC with as set $\Theta$ of parameters the likelihood-based confidence region $\{\theta \in \Theta_{G_N} : lik(\theta) > \beta\}$. When the NCC is learned from training data, the set $\Theta$ of parameters is usually estimated on the basis of the imprecise Dirichlet model: this model depends on a hyperparameter $s \in ]0, +\infty[$, and the behavior of the resulting predictions as $s$ varies from 0 to $+\infty$ is similar to the behavior as $\beta$ varies from 1 to 0 of the predictions with cutoff point $\beta$ returned by the NHC learned from the same training data. Besides the theoretical advantages of not needing prior distributions and of having the whole information encoded in the model (whereas to each $s \in ]0, +\infty[$ corresponds a different NCC), the main practical advantage of the NHC over the NCC when they are learned from training data is that, unlike the hyperparameter $s$, the cutoff point $\beta$ has a

frequentist interpretation in terms of (approximate) confidence levels, thanks to the result of (Wilks, 1938), as shown in the next example. A much more thorough comparison of these naive classifiers will be presented in (Antonucci et al., 2011).

**Example 2.** The simplest nontrivial classification problem corresponds to the case with $\Omega_C = \{a, b\}$ and $m = 0$. Assume that $P(C = a) = \frac{1}{2}$; in this case, the vacuous prediction of $C$ can be considered as the theoretically correct classification, since there is no reason for preferring either of the two possible classes to the other. Consider the NHC learned from a complete training dataset consisting of $n$ objects, and consider the NCC learned from the same training data on the basis of the imprecise Dirichlet model with the standard choice $s = 2$ for the hyperparameter. The probability that the prediction of $C$ with cutoff point $\beta = 0.15$ returned by the NHC is vacuous is approximately 94.3% when $n = 100$ and 94.6% when $n = 1000$, while the probability that the prediction of $C$ returned by the NCC is vacuous is approximately 23.6% when $n = 100$ and 7.6% when $n = 1000$. Hence, in this perfectly symmetric situation the probability that the NCC returns the vacuous prediction (that is, the theoretically correct classification) decreases as the number of objects in the training dataset increases.

## 5 Conclusion

When the likelihood function for the probabilities of a Bayesian network factorizes in multinomial likelihood functions, Theorem 1 gives a method for calculating profile likelihood functions for a particular class of probabilistic inferences. In the future, this method will be generalized to non-factorizing likelihood functions and more general classes of probabilistic inferences, by combining it with approximation algorithms (such as the EM algorithm) and exploiting the algebraic structure of the likelihood functions. Another interesting research topic is the combination of these methods with the learning of the graph of the Bayesian network.

**Acknowledgments**

The author wishes to thank Alessandro Antonucci for stimulating discussions on imprecise probabilistic graphical models, and the anonymous referees for their helpful comments.

## References

Alessandro Antonucci and Marco Zaffalon. 2008. Decision-theoretic specification of credal networks: A unified language for uncertain modeling with sets of Bayesian networks. *Int. J. Approx. Reasoning*, 49(2):345–361.

Alessandro Antonucci, Marco E. G. V. Cattaneo, and Giorgio Corani. 2011. The naive hierarchical classifier. *In preparation.*

Marco E. G. V. Cattaneo. 2007. *Statistical Decisions Based Directly on the Likelihood Function.* Ph.D. thesis, ETH Zurich.

Marco E. G. V. Cattaneo. 2008. Fuzzy probabilities based on the likelihood function. In *Soft Methods for Handling Variability and Imprecision*, pages 43–50. Springer.

Marco E. G. V. Cattaneo. 2009. A generalization of credal networks. In *ISIPTA '09*, pages 79–88. SIPTA.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., Ser. B*, 39(1):1–38.

Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis.* Wiley.

Finn V. Jensen and Thomas D. Nielsen. 2007. *Bayesian Networks and Decision Graphs.* Springer, second edition.

Peter Walley. 1991. *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall.

Peter Walley. 1996. Inferences from multinomial data: Learning about a bag of marbles. *J. R. Stat. Soc., Ser. B*, 58(1):3–57.

Samuel S. Wilks. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 9(1):60–62.

Marco Zaffalon. 2002. The naive credal classifier. *J. Stat. Plann. Inference*, 105(1):5–21.