Pp. 95–102 of Proc. 12th Belgian-Dutch Conference on Machine Learning, (Benelearn-02), Utrecht, 2002

Supervised Learning of Bayesian Network Parameters Made Easy

Hannes Wettig⋆, Peter Grünwald◦, Teemu Roos⋆, Petri Myllymäki⋆, and Henry Tirri⋆

⋆ **Complex Systems Computation Group**
Helsinki Inst. for Inf. Tech. (HIIT)
University of Helsinki
& Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
{*Firstname*}.{*Lastname*}*@hiit.fi*

◦ **CWI**
P.O. Box 94079
NL-1090 GB Amsterdam, The Netherlands.
Peter.Grunwald@cwi.nl

Abstract

Bayesian network models are widely used for supervised prediction tasks such as classification. Usually the parameters of such models are determined using 'unsupervised' methods such as maximization of the joint likelihood. In many cases, the reason is that it is not clear how to find the parameters maximizing the supervised (conditional) likelihood. We show how the supervised learning problem can be solved efficiently for a large class of Bayesian network models, including the Naive Bayes (NB) and tree-augmented NB (TAN) classifiers. We do this by showing that under a certain general condition on the network structure, the supervised learning problem is exactly equivalent to logistic regression. Hitherto this was known only for Naive Bayes models. Since logistic regression models have a concave log-likelihood surface, the global maximum can be easily found by local optimization methods.

1 Introduction

In recent years it has been recognized that for supervised prediction tasks such as classification, we should use a supervised learning algorithm such as supervised (conditional) likelihood maximization (Friedman et al., 1997; Greiner et al., 1997; Ng and Jordan, 2001; Kontkanen et al., 2001; Greiner and Zhou, 2002). Nevertheless, in most related applications the model parameters are still determined using unsupervised methods such as maximization of the unsupervised (joint) likelihood or (ordinary, unsupervised) Bayesian methods. One of the main reasons for this discrepancy is the difficulty in finding the global maximum of the supervised likelihood. In this paper, we show that this problem can be solved for Bayesian network models, as long as they satisfy a particular additional condition. The condition is satisfied for many existing Bayesian-network based classifiers including the Naive Bayes (NB), TAN (tree-augmented NB) and 'diagnostic' classifiers (Kontkanen et al., 2001).

We find the maximum supervised likelihood parameters by parametrizing our models in a different manner; roughly speaking, the parameters in our parametrization correspond to logarithms of parameters in the standard Bayesian network parametrization. In this way, each conditional Bayesian network model is mapped to a logistic regression model. However, in some cases the parameters of this logistic regression model are not allowed to vary freely. In other words, the Bayesian network model corresponds to a subset of a logistic regression model rather than a 'full' logistic regression model.

We provide a general condition on the network structure under which, as we prove, the Bayesian network model is mapped to a full logistic regression model with freely varying parameters. The supervised log-likelihood for logistic regression models is a concave function of the parameters. Since, under our condition, these parameters are allowed to vary freely over $\mathbb{R}^k$ for some k, our condition implies that in the new parametrization the supervised log-likelihood becomes a concave function of parameters in a convex set. This implies that we can find the global maximum supervised likelihood parameters by simple local optimization techniques such as hill climbing.

We remark that viewing Bayesian network classifiers as logistic regression models is not new; it was used earlier in papers such as (Heckerman and Meek, 1997a; Ng and Jordan, 2001; Greiner and Zhou, 2002). Also, the concavity of the log-likelihood surface for logistic regression is known. Our main contribution is to sup-

ply the condition under which Bayesian network models correspond to logistic regression with *completely freely varying parameters*. Only in this latter case can we guarantee that there are no local maxima in the likelihood surface. As a direct consequence of our result, we show for the first time that the supervised likelihood of, for instance, the tree-augmented Naive Bayes (TAN) model has no local maxima.

This paper is organized as follows. In Section 2 we introduce Bayesian networks and an alternative so-called L-parametrization. In Section 3 we show that the L-parametrization allows us to consider Bayesian network classifiers as logistic regression models. Based on earlier results in logistic regression, we conclude that in the L-parametrization the supervised log-likelihood is a concave function. In Section 4 we present our main result, giving conditions under which the two parametrizations correspond to exactly the same conditional distributions and the L-parametrization preserves all the independence assumptions encoded by the network structure. Conclusions are summarized in Section 5.

2 Bayesian Networks

We assume that the reader is familiar with the basics of the theory of Bayesian networks see, e.g., (Pearl, 1988).

Consider a random vector $X = (X_0, X_1, \ldots, X_{M'})$, where each X_i takes values in $\{1, \ldots, n_i\}$. Let $\mathcal{B}$ be a Bayesian network structure over X, which factorizes $P(X)$ into

$$P(X) = \prod_{i=0}^{M'} P(X_i \mid Pa_i), \tag{1}$$

where $Pa_i \subseteq \{X_0, \ldots, X_{M'}\}$ is the parent set of variable X_i in $\mathcal{B}$.

We are interested in predicting some class variable X_m for some $m \in \{0, \ldots, M'\}$ conditioned on all X_i, $i \neq m$. Without loss of generality we may assume that $m = 0$ (i.e., X_0 is the class variable) and that the children of X_0 in $\mathcal{B}$ are $\{X_1, \ldots, X_M\}$ for some $M \leq M'$. For instance, in the so-called Naive Bayes model (leftmost picture in Figure 1), we have $M = M'$ and the children of the class variable X_0 are independent given the value of X_0. The Bayesian network model corresponding to $\mathcal{B}$ is the set of all distributions satisfying the conditional independencies encoded in $\mathcal{B}$. It is usually parametrized by vectors $\Theta^{\mathcal{B}}$ with components of the form $\theta^{\mathcal{B}}_{x_i|pa_i}$ defined by

$$\theta^{\mathcal{B}}_{x_i|pa_i} := P(X_i = x_i \mid Pa_i = pa_i), \tag{2}$$

where pa_i is any configuration (set of values) for the parents Pa_i of X_i. Whenever we want to emphasize that each pa_i is determined by the complete data vector $x = (x_0, \ldots, x_{M'})$, we write $pa_i(x)$ to denote the configuration of Pa_i in $\mathcal{B}$ given by the vector x. For a given data vector $x = (x_0, x_1, \ldots, x_{M'})$, we sometimes need to consider a modified vector where x_0 is replaced by x'_0 and the other entries remain the same. We then write $pa_i(x'_0, x)$ for the same configuration given by $(x'_0, x_1, \ldots, x_{M'})$.

We let $\mathcal{M}^{\mathcal{B}}$ be the set of *conditional* distributions $P(X_0 \mid X_1, \ldots, X_{M'}, \Theta^{\mathcal{B}})$ corresponding to distributions $P(X_0, \ldots, X_{M'} \mid \Theta^{\mathcal{B}})$ satisfying the conditional independencies encoded in $\mathcal{B}$. The conditional distributions in $\mathcal{M}^{\mathcal{B}}$ can be written as

$$P(x_0 \mid x_1, \ldots, x_{M'}, \Theta^{\mathcal{B}}) = \frac{\theta^{\mathcal{B}}_{x_0|pa_0(x)} \prod_{i=1}^{M'} \theta^{\mathcal{B}}_{x_i|pa_i(x)}}{\sum_{x'_0=1}^{n_0} \theta^{\mathcal{B}}_{x'_0|pa_0(x)} \prod_{i=1}^{M'} \theta^{\mathcal{B}}_{x_i|pa_i(x'_0,x)}}, \tag{3}$$

extended to N outcomes by independence.

Given a complete data-matrix $D = (x^1, \ldots, x^N)$, the *supervised log-likelihood*, $S^{\mathcal{B}}(D; \Theta^{\mathcal{B}})$, of parameters $\Theta^{\mathcal{B}}$ is given by

$$S^{\mathcal{B}}(D; \Theta^{\mathcal{B}}) := \sum_{j=1}^{N} S^{\mathcal{B}}(x^j; \Theta^{\mathcal{B}}), \tag{4}$$

where

$$S^{\mathcal{B}}(x; \Theta^{\mathcal{B}}) := \log P(x_0 \mid x_1, \ldots, x_{M'}, \Theta^{\mathcal{B}}). \tag{5}$$

Note that in (3), and hence also in (4), all $\theta^{\mathcal{B}}_{x_i|pa_i}$ with $i > M$ (standing for nodes that are neither the class variable nor any of its children) cancel out, since for these terms we have $pa_i(x) \equiv pa_i(x'_0, x)$ for all x'_0. Thus the only relevant parameters for determining the conditional likelihood are of the form $\theta^{\mathcal{B}}_{x_i|pa_i}$ with $i \in \{0..M\}$, $x_i \in \{1..n_i\}$ and pa_i any configuration of values of Pa_i. We order these parameters

lexicographically and define $\mathbf{\Theta}^{\mathcal{B}}$ to be the set of vectors constructed this way, with $\theta^{\mathcal{B}}_{x_i|pa_i} > 0$ and $\sum_{x_i=1}^{n_i} \theta^{\mathcal{B}}_{x_i|pa_i} = 1$ for all $i \in \{0, \ldots, M\}$, x_i and all values (configurations) of pa_i. Note that we require all parameters to be strictly positive.

The model $\mathcal{M}^{\mathcal{B}}$ does not contain any notion of the 'unsupervised' distributions: Probabilities such as $P(X_i \mid Pa_i)$, where $M < i \leq M'$, are undefined, and neither are we interested in them. Our task is prediction of X_0 *given* $X_1, \ldots, X_{M'}$. Heckerman and Meek call models such as $\mathcal{M}^{\mathcal{B}}$ *Bayesian regression/classification* (BRC) models (Heckerman and Meek, 1997a; Heckerman and Meek, 1997b).

For an arbitrary supervised Bayesian network model $\mathcal{M}^{\mathcal{B}}$, we now define the so-called L-model, another set of conditional distributions $P(X_0 \mid X_1, \ldots, X_{M'})$. This model, which we denote $\mathcal{M}^L$, is parametrized by vectors Θ^L in some set $\mathbf{\Theta^L}$ that closely resembles $\mathbf{\Theta}^{\mathcal{B}}$. Each different $\mathcal{M}^{\mathcal{B}}$ will give rise to a corresponding $\mathcal{M}^L$, although we do not necessarily have $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$. For each component $\theta^{\mathcal{B}}_{x_i|pa_i}$ of each vector $\Theta^{\mathcal{B}} \in \mathbf{\Theta}^{\mathcal{B}}$, there is a corresponding component $\theta^L_{x_i|pa_i}$ of the vectors $\Theta^L \in \mathbf{\Theta^L}$. The components $\theta^L_{x_i|pa_i}$ take values in the range $(-\infty, \infty)$ rather than $(0, 1)$. Each vector $\Theta^L \in \mathbf{\Theta^L}$ defines the following conditional distribution:

$$P(x_0 \mid x_1, \ldots, x_{M'}, \Theta^L) := \frac{\exp \theta^L_{x_0|pa_0(x)} \prod_{i=1}^{M} \exp \theta^L_{x_i|pa_i(x)}}{\sum_{x'_0=1}^{n_0} \exp \theta^L_{x'_0|pa_0(x)} \prod_{i=1}^{M} \exp \theta^L_{x_i|pa_i(x'_0,x)}}. \quad (6)$$

The model $\mathcal{M}^L$ is the set of conditional distributions $P(X_0 \mid X_1, \ldots, X_{M'}, \Theta^L)$ indexed by $\Theta^L \in \mathbf{\Theta^L}$, extended to N outcomes by independence. Given a data-matrix D, let $S^L(D; \Theta^L)$ be the supervised log-likelihood of parameters Θ^L, defined analogously to (4) with (6) in place of (3).

Theorem 1. $\mathcal{M}^{\mathcal{B}} \subseteq \mathcal{M}^L$.

Proof. The theorem is immediate from doing the log-parameter transformation, i.e., setting $\theta^L_{x_i|pa_i} = \log \theta^{\mathcal{B}}_{x_i|pa_i}$ for all i, x_i and pa_i. ∎

In words, all the conditional distributions that can be represented by parameters $\Theta^B \in \mathbf{\Theta}^{\mathcal{B}}$ can also be represented by parameters $\Theta^L \in \mathbf{\Theta^L}$. However, whereas in the usual parametrization we require $\sum_{x_i=1}^{n_i} \theta^{\mathcal{B}}_{x_i|pa_i} = 1$ for each $i \in \{0, \ldots, M'\}$ and pa_i, there is no corresponding condition for the parameters of the L-model. Consequently, the converse of Theorem 1, i.e., $\mathcal{M}^L \subseteq \mathcal{M}^{\mathcal{B}}$, is true only under some additional conditions on the network structure. We return to this topic in Section 4 but first we take a closer look at the L-model.

3 The L-model Viewed as Logistic Regression

Although the L-model is closely related to and in some cases formally identical to Bayesian network classifiers, it can also be interpreted in terms of *logistic regression*. We can think of the conditional model $\mathcal{M}^L$ as a predictor that combines the information of the attributes using the so-called *softmax* rule (Bishop, 1995; Heckerman and Meek, 1997b; Ng and Jordan, 2001). Figure 1 gives an interpretation of this, depicting a Naive Bayes model and the corresponding L-model in their Bayesian network guises.

In terms of logistic regression, the L-model has one (binary) regressor variable for each configuration of each of the parent sets Pa_i, and one (binary) output variable for each possible value of the class variable. Having established that the L-model is a logistic regression model, we may use a well-known fact that holds for logistic regression models in general. Namely, the supervised log-likelihood in the L-parametrization is a concave function of the parameters:

Theorem 2. *(Santner and Duffy, 1989) The parameter set $\mathbf{\Theta^L}$ is convex, and the supervised log-likelihood $S^L(D; \Theta^L)$ is concave, though not strictly concave.*

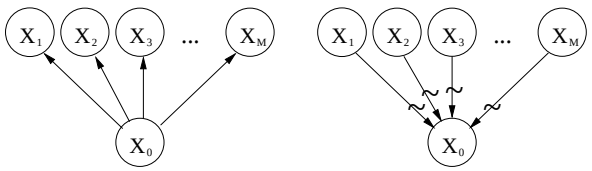

Figure 1: Standard Naive Bayes structure (left) and the corresponding L-model (right). The arcs of the network have been reversed and the resulting product distribution has been replaced by softmax (denoted by tildes).

Proof. The first part is obvious since each parameter can take values in $(-\infty, \infty)$. Concavity of $S^L(D; \Theta^L)$ is a direct consequence of the fact that $\mathcal{M}^L$ is a logistic regression model; see, e.g., (Santner and Duffy, 1989, p. 234). For an example where the supervised log-likelihood is not strictly concave, see (Wettig et al., 2002). ∎

From the theorem we directly obtain the following corollary.

Corollary 1. *There are no local (non-global) maxima in the likelihood surface of an L-model.*

The conditions under which a global maximum exists are discussed in, e.g., (Santner and Duffy, 1989) and references therein. A possible solution in cases where no maximum exists is to assign a prior on the model parameters and maximize the 'supervised posterior' (Grünwald et al., 2002; Wettig et al., 2002) instead of the likelihood.

The global supervised maximum likelihood parameters obtained from training data can be used for prediction of future data. In addition, as discussed in (Heckerman and Meek, 1997a) they can be used to perform model selection among several competing model structures using, e.g., the BIC (Schwarz, 1978) or (approximate) MDL (Rissanen, 1978) criteria. In (Heckerman and Meek, 1997a) it is stated that for general supervised Bayesian network models $\mathcal{M}^{\mathcal{B}}$, "although it may be difficult to determine a global maximum, gradient-based methods [...] can be used to locate local maxima". What is more, Theorem 2 shows that when dealing with L-models even a *global* maximum can be found if it exists.

In the original parametrization, the log-likelihood surface is not necessarily *concave*, as the following example shows.

Example 1. Consider a Bayesian network where the class variable X_0 has only one child, X_1, and both variables take values in $\{1, 2\}$. Let the training data be given by

$$D = \big((1,1), (1,2), (2,1), (2,2)\big).$$

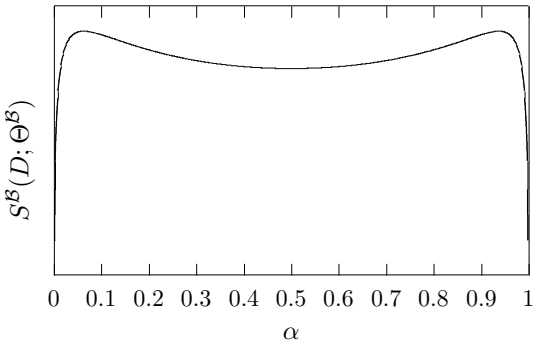

Figure 2: The supervised log-likelihood of Example 1 peaks twice in the original parametrization along a line defined by $\alpha \in (0, 1)$.

Set the parameters $\Theta^{\mathcal{B}}$ as follows:

$$\theta^{\mathcal{B}}_{x_0} = \begin{cases} 0.1 & \text{if} \quad x_0 = 1, \\ 0.9 & \text{if} \quad x_0 = 2, \end{cases}$$

$$\theta^{\mathcal{B}}_{x_1|x_0} = \begin{cases} 0.5 & \text{if} \quad x_0 = 1, x_1 = 1, \\ 0.5 & \text{if} \quad x_0 = 1, x_1 = 2, \\ \alpha & \text{if} \quad x_0 = 2, x_1 = 1, \\ 1 - \alpha & \text{if} \quad x_0 = 2, x_1 = 2. \end{cases}$$

Figure 2 shows the supervised log-likelihood given data D as a function of α. The figure shows a bimodal curve that clearly violates concavity. ◇

If the network structure $\mathcal{B}$ is such that the two models are equivalent, $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$, we can find the global maximum of the supervised likelihood by reparametrizing $\mathcal{M}^{\mathcal{B}}$ in the L-parameterization, and using some local optimization method. Because the log-transformation is continuous, it follows (with some calculus) that in this case all local maxima of the supervised likelihood are global maxima also in the original parametrization $\mathbf{\Theta}^{\mathcal{B}}$.

4 Main Result

Theorems 1 and 2 suggest that in order to find parameters maximizing the supervised likelihood for any Bayesian network model, we could use the L-model where optimization is easy. However, the resulting parameters may violate the 'sum-up-to-one constraint', i.e., we may have $\sum_{x_i=1}^{n_i} \exp \theta^L_{x_i|pa_i} \neq 1$ for some $i \in \{0, \ldots, M'\}$ and pa_i. This *could* correspond to

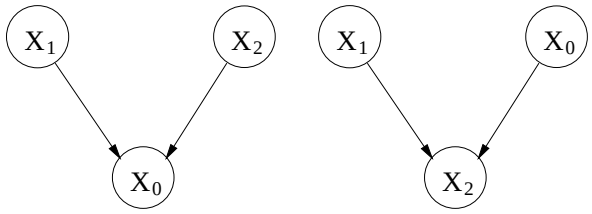

Figure 3: A simple Bayesian network (the class variable is denoted by X_0) satisfying Condition 1 (left); and a network that does not satisfy the condition (right).

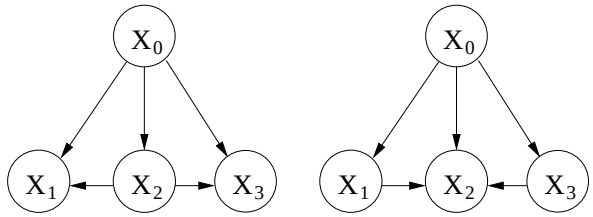

Figure 4: A tree-augmented Naive Bayes (TAN) model satisfying Condition 1 (left); and a network that is not TAN (right). Note that even though in both cases the class variable X_0 is a moral node, the network on the right does not satisfy Condition 1.

a violation of the independence assumptions of the model structure $\mathcal{B}$ as demonstrated by Example 2 below. However, for some structures $\mathcal{B}$, we can show that even if Θ^L is such that $\sum_{x_i=1}^{n_i} \exp \theta^L_{x_i|pa_i} \neq 1$ for some $i \in \{0, \ldots, M'\}$ and pa_i, the conditional distribution indexed by Θ^L is still in $\mathcal{M}^{\mathcal{B}}$.

Our main result is that the independence assumptions encoded by $\mathcal{B}$ are guaranteed to be preserved in the L-model if $\mathcal{B}$ satisfies the following condition:

Condition 1 For all $j = 1..M$, there exists $X_i \in Pa_j \cap \{X_0, \ldots, X_M\}$ such that $Pa_j \subseteq Pa_i \cup \{X_i\}$.

Remark. Condition 1 holds for $\mathcal{B}$ (as can be seen by induction) if and only if any parent set Pa_j of a child X_j of the class X_0 is 'conditionally fully connected', i.e., fully connected modulo arcs (between parents of X_0) that have no effect on the conditional $P(X_0 \mid Pa_j \setminus \{X_0\})$. A necessary but not sufficient condition is that the class X_0 must be a 'moral node', i.e., it cannot have a common child with a node it is not directly connected with; see Figure 4. $\diamond$

Example 2. Consider the Bayesian networks depicted in Figure 3. The leftmost network, $\mathcal{B}_1$, satisfies Condition 1, unlike the rightmost network, $\mathcal{B}_2$. Let $\mathcal{M}_2^{\mathcal{B}}$ denote the 'usual' model corresponding to $\mathcal{B}_2$ indexed by parameters in $\mathbf{\Theta}^{\mathcal{B}}$, and let $\mathcal{M}_2^L$ denote the corresponding L-model. The parameters used to define $\mathcal{M}_2^L$ are $\theta^L_{x_0}$ and $\theta^L_{x_2|x_0,x_1}$ where each x_i varies in $\{1, \ldots, n_i\}$. Consider the distributions in $\mathcal{M}_2^L$ indexed by a Θ^L such that, for some x_0 and x_1, $\sum_{x_2=1}^{n_2} \exp \theta^L_{x_2|x_0,x_1} \neq 1$. Some of these distributions violate the independence assumptions of $\mathcal{B}_2$, and therefore, are not in $\mathcal{M}_2^{\mathcal{B}}$. For instance, suppose that for $x_0 = x_1 = 1$, summing over the values of X_2 gives something *more* than one, whereas for $x_0 = 2, x_1 = 1$, summing over the values of X_2 gives something *less* than one. This would correspond to the situation where given $X_1 = 1$ it is more probable to have $X_0 = 1$ than $X_0 = 2$, i.e., X_1 and X_0 would be dependent *without* X_2 being given. It follows that in this case $\mathcal{M}_2^L$ contains some conditional distributions that do not satisfy the independence assumptions encoded by $\mathcal{B}_2$. This can not happen in the leftmost network $\mathcal{B}_1$ since the structure allows all conditional distributions of X_0 given X_1 and X_2. $\diamond$

As examples of network structures that satisfy Condition 1, we mention the Naive Bayes (NB) and the tree-augmented Naive Bayes (TAN) models (Friedman et al., 1997). The latter is a generalization of the former in which the children of the class variable are allowed to form tree-structures; see Figure 4.

Proposition 1. *Condition 1 is satisfied by the Naive Bayes and the tree-augmented Naive Bayes structures.*

Proof. For Naive Bayes, we have $Pa_j \subseteq \{X_0\}$ for all $j \in \{1, \ldots, M\}$. For TAN models, all children of the class variable have either one or two parents. For children with only one parent (the class variable), we can use the same argument as in the NB case. For any child X_j with two parents, let X_i be the parent that is not the class variable. Because X_i is also a child of the class variable, we have $Pa_j \subseteq Pa_i \cup \{X_i\}$. ∎

Condition 1 is also automatically satisfied if

X_0 only has incoming arcs[1] ('diagnostic' classifiers, see (Kontkanen et al., 2001)). For Bayesian network structures for which the condition does not hold, we can always add some arrows to arrive at a structure $\mathcal{B}'$ for which the condition does hold (for instance, add an arrow from X_1 to X_3 in the rightmost network in Figure 4). Therefore, the model $\mathcal{M}^{\mathcal{B}}$ is always a submodel of a larger model $\mathcal{M}^{\mathcal{B}'}$ for which the condition holds. For these reasons, we regard Condition 1 as relatively mild.

We are now ready to present our main result:

Theorem 3. *If $\mathcal{B}$ satisfies Condition 1, then $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^{L}$.*

Together with Corollary 1, this shows that if Condition 1 holds for $\mathcal{B}$, then the supervised likelihood surface of $\mathcal{M}^{\mathcal{B}}$ has no local (non-global) maxima. Proposition 1 now implies that, for example, the supervised likelihood surface for the TAN classifiers has no local (non-global) maxima. Therefore, this maximum can be found by local optimization techniques.

Proof. In the following, we will often speak of the parent configuration pa_0 of X_0. In case X_0 has no parents (i.e., $M = M'$), Pa_0 is the empty set and $pa_0(x)$ is constant with respect to $x = (x_0, \ldots, x_{M'})$.

We introduce some more notation. For $j \in \{1, \ldots, M\}$, let m_j be the maximum number in $\{0, \ldots, M\}$ such that $X_{m_j} \in Pa_j$, $Pa_j \subseteq Pa_{m_j} \cup \{X_{m_j}\}$. Such an m_j exists by Condition 1. Condition 1 implies that pa_j is completely determined by the pair (x_{m_j}, pa_{m_j}). We can therefore introduce functions Q_j mapping (x_{m_j}, pa_{m_j}) to the corresponding pa_j. Hence, for all $x = (x_0, \ldots, x_{M'})$ and $j \in \{1, \ldots, M\}$ we have

$$pa_j = Q_j(x_{m_j}, pa_{m_j}). \tag{7}$$

We introduce, for all $i \in \{0, \ldots, M\}$ and for each configuration pa_i of Pa_i, a constant $c_{i|pa_i}$ and define, for any $\Theta^L \in \mathbf{\Theta^L}$,

$$\theta^{(c)}_{x_i|pa_i} := \theta^{L}_{x_i|pa_i} + c_{i|pa_i} - \sum_{j:m_j=i} c_{j|Q_j(x_i,pa_i)}. \tag{8}$$

[1]It is easy to see that in that case the maximum supervised likelihood parameters may even be determined analytically.

The parameters $\theta^{(c)}_{x_i|pa_i}$ constructed this way are combined to a vector $\Theta^{(c)}$ which is clearly a member of $\mathbf{\Theta^L}$.

After introducing this additional notation, we proceed to the first stage of the proof.

Stage 1 In this stage of the proof, we show that no matter how we choose the constants $c_{i|pa_i}$, for all Θ^L and corresponding $\Theta^{(c)}$ we have $S^L(D; \Theta^{(c)}) = S^L(D; \Theta^L)$.

We first show that, for all possible vectors x and the corresponding parent configurations, no matter how the $c_{i|pa_i}$ are chosen, it holds that

$$\sum_{i=0}^{M} \theta^{(c)}_{x_i|pa_i} = \sum_{i=0}^{M} \theta^{L}_{x_i|pa_i} + c_{0|pa_0}. \tag{9}$$

To derive (9) we substitute all terms of $\sum_{i=0}^{M} \theta^{(c)}_{x_i|pa_i}$ by their definition (8). Clearly, for all $j \in \{1, \ldots, M\}$, there is exactly one term of the form $c_{j|pa_j}$ that appears in the sum with a positive sign. Since for each $j \in \{1, \ldots, M\}$ there exists exactly one $i \in \{0, \ldots, M\}$ with $p_j = i$, it must be the case that for all $j \in \{1, \ldots, M\}$, a term of the form $c_{j|Q_j(x_i,pa_i)}$ appears exactly once in the sum with a negative sign. By (7) we have $c_{j|Q_j(x_i,pa_i)} = c_{j|pa_j}$. Therefore all terms $c_{j|pa_j}$ that appear once with a positive sign also appear once with a negative sign. It follows that, except for $c_{0|pa_0}$, all terms $c_{j|pa_j}$ cancel. This establishes (9). By plugging in (9) into (6), it follows that $S^L(D; \Theta^{(c)}) = S^L(D; \Theta^L)$ for all D. This concludes Stage 1 of the proof.

Stage 2 Set, for all x_i and pa_i,

$$\theta^{\mathcal{B}}_{x_i|pa_i} = \exp \theta^{(c)}_{x_i|pa_i}. \tag{10}$$

In this stage we show that we can determine the constants $c_{i|pa_i}$ such that for all $i \in \{0, \ldots, M\}$ and pa_i, the 'sum up to one' constraint is satisfied, i.e., we have

$$\sum_{x_i=1}^{n_i} \theta^{\mathcal{B}}_{x_i|pa_i} = 1. \tag{11}$$

We will achieve this by sequentially determining values for $c_{i|pa_i}$ in a particular order. We now need some terminology: we say 'c_i *is determined*' if for all configurations pa_i of Pa_i, we

have already determined $c_{i|pa_i}$. We say 'c_i *is undetermined*' if we have determined $c_{i|pa_i}$ for *no* configuration pa_i of Pa_i. We say 'c_i is *ready to be determined*' if c_i is undetermined and at the same time all c_j with $p_j = i$ have been determined.

We first note that as long as some c_i with $i \in \{0, \ldots, M\}$ are undetermined, there must exist $c_{i'}$ that are ready to be determined. To see this, first take any $i \in \{0, \ldots, M\}$ with c_i undetermined. Either c_i itself is ready to be determined (in which case we are done), or there exists $j \in \{1, \ldots M\}$ with $p_j = i$ (and hence $X_i \in Pa_j$) such that c_j is undetermined. If c_j is ready to be determined, we are done. Otherwise, there must exist some k with $X_j \in Pa_k$ such that c_k is undetermined. We can now repeat the argument, and move forward in the Bayesian network structure $\mathcal{B}$ restricted to $\{X_0, \ldots, X_M\}$ until we find a c_l that is ready to be determined. Because $\mathcal{B}$ is acyclic, we must find such a c_l (within $M + 1$ steps).

We now describe an algorithm that sequentially assigns values to $c_{i|pa_i}$ such that (11) will be satisfied. We start with all c_i undetermined and repeat the following steps:

WHILE there exists $i \in \{0, \ldots, M\}$ such that c_i is undetermined
DO

1. Pick the largest i such that c_i is ready to be determined.
2. Set, for all configurations pa_i of Pa_i, $c_{i|pa_i}$ such that $\sum_{x_i=1}^{n_i} \theta^{\mathcal{B}}_{x_i|pa_i} = 1$ holds.

DONE

The algorithm will loop $M + 1$ times and then halt. Step 2 does not affect the values of $c_{j|pa_j}$ for any j, pa_j such that $c_{j|pa_j}$ has already been determined. Therefore, after the algorithm halts, (11) holds. This concludes Stage 2 of the proof.

Let $\Theta^L \in \mathbf{\Theta^L}$. Each choice of constants $c_{i|pa_i}$ determines a corresponding vector $\Theta^{(c)}$ with components given by (8). This in turn determines a corresponding vector $\Theta^{\mathcal{B}}$ with components given by (10). In Stage 2 we showed that we can take the $c_{i|pa_i}$ such that (11) holds. This is the choice of $c_{i|pa_i}$ which we adopt. With this particular choice, $\Theta^{\mathcal{B}}$ indexes a distribution in $\mathcal{M}^{\mathcal{B}}$. By applying the log-transformation to the components of $\Theta^{\mathcal{B}}$ we find that for any D of any length, $S^{\mathcal{B}}(D; \Theta^{\mathcal{B}}) = S^L(D; \Theta^{(c)})$, where $S^{\mathcal{B}}(D; \Theta^{\mathcal{B}})$ denotes the supervised log-likelihood of $\Theta^{\mathcal{B}}$ as given by summing the logarithm of (3). The result of Stage 1 now implies that $\Theta^{\mathcal{B}}$ indexes the same conditional distribution as Θ^L. Since $\Theta^L \in \mathbf{\Theta^L}$ was chosen arbitrarily, this shows that $\mathcal{M}^L \subseteq \mathcal{M}^{\mathcal{B}}$. Together with Theorem 1 this concludes the proof. ∎

5 Concluding Remarks

We showed that by using the parameter transformation described above, one can effectively find the parameters maximizing the supervised (conditional) likelihood of NB, TAN and many other Bayesian network models. For an arbitrary Bayesian network, this transformation may yield a slightly more powerful model class, i.e., remove some of the independence assumptions of the network structure. We also gave a condition under which the transformation does not change the class of models considered. Test runs for the Naive Bayes case in (Wettig et al., 2002) have shown that maximizing the supervised likelihood in contrast to the usual practice of maximizing the unsupervised (joint) likelihood is feasible and yields greatly improved classification. In the future we intend to study more complicated models as well as use the L-parametrization for model selection.

Acknowledgments. This research has been supported by the National Technology Agency, and the Academy of Finland. The authors thank Wray Buntine for many useful comments.

References

C.M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

N. Friedman, D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning*, 29:131–163.

R. Greiner and W. Zhou. 2002. Structural extension to logistic regression: Discriminant parameter learning of belief net classifiers. In *Proceedings of the 18th Annual National Conference on Artificial Intelligence (AAAI-02)*, pages 167–173, Edmonton.

R. Greiner, A. Grove, and D. Schuurmans. 1997. Learning Bayesian nets that perform well. In *Pro-*

ceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97), pages 198–207. Morgan Kaufmann Publishers, San Francisco, CA.

P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig. 2002. Supervised posterior distributions. Presented at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain.

D. Heckerman and C. Meek. 1997a. Embedded bayesian network classifiers. Technical Report MSR-TR-97-06, Microsoft Research.

D. Heckerman and C. Meek. 1997b. Models and selection criteria for regression and classification. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 223–228. Morgan Kaufmann Publishers, San Francisco, CA.

P. Kontkanen, P. Myllymäki, and H. Tirri. 2001. Classifier learning with supervised marginal likelihood. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 277-284. Morgan Kaufmann Publishers, San Francisco, CA.

A.Y. Ng and M.I. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, 14:605–610.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA.

J. Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14:445–471.

T. Santner and D. Duffy. 1989. *The Statistical Analysis of Discrete Data*. Springer Verlag, New York.

G. Schwarz. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

H. Wettig, P. Grünwald, T. Roos, P. Myllymäki, and H. Tirri. 2002. On supervised learning of Bayesian network parameters. Technical Report HIIT-2002–1, Helsinki Institute for Information Technology (HIIT). Available at http://cosco.hiit.fi/Articles/hiit2002-1.ps.