Available online at www.sciencedirect.com
ScienceDirect
International Journal of Approximate Reasoning
48 (2008) 659–666
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
www.elsevier.com/locate/ijar

Bayesian networks with a logistic regression model for the conditional probabilities

Frank Rijmen *

Department of Clinical Epidemiology and Biostatistics, VU Medical Centre, Amsterdam, The Netherlands

Received 26 March 2007; received in revised form 9 January 2008; accepted 9 January 2008
Available online 20 January 2008

Abstract

Logistic regression techniques can be used to restrict the conditional probabilities of a Bayesian network for discrete variables. More specifically, each variable of the network can be modeled through a logistic regression model, in which the parents of the variable define the covariates. When all main effects and interactions between the parent variables are incorporated as covariates, the conditional probabilities are estimated without restrictions, as in a traditional Bayesian network. By incorporating interaction terms up to a specific order only, the number of parameters can be drastically reduced. Furthermore, ordered logistic regression can be used when the categories of a variable are ordered, resulting in even more parsimonious models. Parameters are estimated by a modified junction tree algorithm. The approach is illustrated with the Alarm network.

Keywords: Bayesian networks; Logistic regression; Generalized linear models; Restricted conditional probabilities

0. Introduction

In a probabilistic graphical model, random variables are represented by nodes, and the (absence of) edges between nodes represent conditional (in)dependence relations. Apart from offering an appealing way to represent models visually, efficient computational schemes can be constructed by working on the graph associated with a probabilistic model [7].

Recently, research has been focused on structural learning. That is, how can we identify a set of conditional dependence relations that is both parsimonious and provides an adequate fit to a given dataset? Several procedures have been proposed in the literature (for reviews, see [2,8,12]). In this paper on the other hand, we focus on learning the parameters of an inferred (or a priori given) network structure. We consider Bayesian networks for discrete variables, where dependence relations are encoded through directed edges. More specifically, we show how the number of effective parameters of the network can be reduced by adopting a logistic regression framework for modelling the conditional dependence relations.

* Address: Rosedale Road, Princeton, NJ 08541, United States.
E-mail address: frijmen@ets.org

0888-613X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2008.01.001

In a Bayesian network, the probability distribution of a set of random variables $\mathbf{X} = (X_1, \ldots, X_M)'$ can be recursively factorized as

$$\Pr(\mathbf{X}) = \prod_{m=1}^{M} \Pr(X_m | \mathrm{pa}(X_m)), \tag{1}$$

where $\mathrm{pa}(X_m)$ is the set of random variables that are parents of X_m in the directed acyclic graph that is associated with $\Pr(\mathbf{X})$. Learning the parameters of a Bayesian network for discrete variables hence comes down to learning the parameters that govern the conditional probability tables $\Pr(X_m|\mathrm{pa}(X_m))$. Usually, these conditional probability tables are not restricted beyond the obvious restriction that $\sum_{j=1}^{J_m} \Pr(X_m = j|\mathrm{pa}(X_m)) = 1$, where J_m is the number of distinct values X_m can take. In some model families, equality restrictions between conditional probability tables are encountered as well. For example, in hidden Markov type of models, a default assumption is that the conditional probability tables do not change over time. Regardless of the latter type of restrictions, each additional parent adds a dimension to the conditional probability table, so that the number of parameters increases exponentially with the number of parents when these conditional probabilities are not further restricted. Consequently, for small to moderately sized data sets, parameters can only be reliably estimated for fairly simple network structures. In the Bayesian networks field, this problem is most often tackled by incorporating "prior information," leading to either penalized maximum likelihood estimation or a fully Bayesian approach. When prior information is available through substantive knowledge or previous studies (rather than the prior 'knowledge' that extreme probabilities are unlikely), this is quite a reasonable approach.

In this paper, an alternative approach to tackle the estimation problem is proposed. More specifically, the number of parameters is controlled by modelling the conditional probabilities as a function of a limited set of parameters using logistic regression.

1. Modelling the conditional probabilities with multinomial logistic regression

Let y_i, $i = 1, \ldots, n$ denote a set of independent realizations of a categorical outcome variable Y, and $\mathbf{z}_i$ the corresponding vector of realizations of p covariates. Then, a multinomial logistic regression model can be specified as follows (e.g. [4]):

- y_i is a realization from a multinomial distribution

$$\Pr(Y_i = j) = \pi_{ij} \quad \text{with} \quad \sum_j \pi_{ij} = 1 \tag{2}$$

- The parameter vector $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ-1})'$ (π_{iJ} is redundant since $\sum_j \pi_{ij} = 1$) is related to the *linear predictor* $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iJ-1})'$ via the multinomial link function:

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \eta_{ij}. \tag{3}$$

- $\boldsymbol{\eta}_i = \mathbf{Z}_i\boldsymbol{\beta}$, where $\mathbf{Z}_i$ is the so-called design matrix of size $J - 1$ by p constructed from $\mathbf{z}_i$; and $\boldsymbol{\beta}$ is a p-dimensional parameter vector.

The multinomial logistic regression model can be integrated into a Bayesian network by modelling each conditional probability table $\Pr(X_m|\mathrm{pa}(X_m))$ of a particular Bayesian network with a multinomial logistic regression model, where X_m is the outcome variable and the design matrix $\mathbf{Z}_{mi}$ is constructed from $pa(X_m)$.

A Bayesian network without restrictions on the conditional probability tables is obtained by constructing $\mathbf{Z}_{mi}$ from $\mathrm{pa}(X_m)$ as follows. For each possible configuration s on $\mathrm{pa}(X_m), s = 1, \ldots, S = \prod_{k:X_k \in \mathrm{pa}(X_m)} J_k$, a dummy variable is defined. For each case i, the covariate vector $\mathbf{z}_{im} = (z_{im1}, \ldots, z_{imS})'$ is defined as an indicator vector with $z_{ims} = 1$ if configuration s is observed, and $z_{ims} = 0$ otherwise. The $(J_m - 1)$ by $(J_m - 1) \times S$ design matrix $\mathbf{Z}_{im}$ is constructed from $\mathbf{z}_{im}$ as

$$\mathbf{Z}_{im} = \begin{bmatrix} \mathbf{z}'_{im} & & & \\ & \mathbf{z}'_{im} & & \\ & & \ddots & \\ & & & \mathbf{z}'_{im} \end{bmatrix},$$

where all entries that are not displayed equal zero. For example, suppose that X_m can take three possible values ($J_m = 3$) and has two parents, each of them having four possible categories. Then, the expression for the linear predictor for each case i is instantiated by

$$\boldsymbol{\eta}_{im} = \begin{pmatrix} \eta_{im1} \\ \eta_{im2} \end{pmatrix} = \mathbf{Z}_{im}\boldsymbol{\beta}_m = \begin{bmatrix} z_{im1} & \cdots & z_{im16} & & & \\ & & & z_{im1} & \cdots & z_{im16} \end{bmatrix} \begin{bmatrix} \beta_{m11} \\ \vdots \\ \beta_{m116} \\ \beta_{m21} \\ \vdots \\ \beta_{m216} \end{bmatrix}. \tag{4}$$

From the example, it is easily verified that this is no more then expressing the probability parameters of the Bayesian network on a different scale. There are as many logistic regression parameters as there are free probabilities parameters ($4 \times 4 \times 2 = 32$). For any of the 16 possible configurations on $\mathrm{pa}(X_m)$, each time two different logistic regression parameters are selected by pre-multiplying $\boldsymbol{\beta}_m$ with $\mathbf{Z}_{im}$, one for each response category j, $j = 1, \ldots, J_m - 1$. The corresponding conditional probabilities are obtained by applying the inverse of the link function to the linear predictor:

$$\pi_{mj} = \frac{\exp(\eta_{mj})}{1 + \sum_{k=1}^{J_m-1} \exp(\eta_{mk})} = \frac{\exp(\beta_{mjs})}{1 + \sum_{k=1}^{J_m-1} \exp(\beta_{mks})}. \tag{5}$$

A model that does not impose any restriction on the conditional probabilities is called a *saturated* model.

2. Restricting the conditional probabilities of a Bayesian network

In the previous section, it was explained how a conditional probability table $\Pr(X_m|\mathrm{pa}(X_m))$ of a traditional Bayesian network can be modelled with a saturated multinomial logistic regression model that incorporates as many covariates as there are free probabilities in the conditional probability table. That a distinct set of parameters is defined for each configuration on $\mathrm{pa}(X_m)$ actually means that the model incorporates the highest order interaction between the categorical covariates. In the example discussed above: the 'effect' of the first categorical covariate (with 4 categories) was allowed to differ across the 4 categories of the second covariate, resulting in 16 logistic regression parameters to model each of the two nonredundant category probabilities of the outcome variable.

A natural way to reduce the number of parameters of the model is to include only main effects of the covariates and interaction terms up to a specific order. This is illustrated by skipping the interaction between the two covariates in the example. Then, $\mathbf{z}_{im}$ has the following structure for each case:

$$\mathbf{z}_{im} = \begin{pmatrix} 1 & \mathbf{z}'_{im1} & \mathbf{z}'_{im2} \end{pmatrix}',$$

where $\mathbf{z}_{iml}$, $l = 1, 2$ is an indicator vector of length 3 ($J_l - 1$) with $z_{imlj} = 1$ if category j is observed for covariate l, and $z_{imlj} = 0$ otherwise. The length of each $\mathbf{z}_{iml}$ is only $J_l - 1$ because the last category is coded as a vector of zeros. $\mathbf{Z}_{im}$ is constructed from $\mathbf{z}_{im}$ as before:

$$\mathbf{Z}_{im} = \begin{bmatrix} 1 & \mathbf{z}'_{im1} & \mathbf{z}'_{im2} & & & \\ & & & 1 & \mathbf{z}'_{im1} & \mathbf{z}'_{im2} \end{bmatrix}.$$

Hence, omitting the interaction between the two covariates reduces the number of parameters from 32 to 14.

For ordered outcome variables, the cumulative logistic link function is often used instead of the multinomial link function:

$$\log\left(\frac{\pi_{i1}+\ldots+\pi_{ij}}{\pi_{ij+1}+\ldots+\pi_{iJ_m}}\right)=\eta_{ij},\quad j=1,\ldots,J_m-1. \tag{6}$$

The cumulative link function can be motivated from a category boundaries approach [3]. In this approach, it is assumed that the categorical outcome variable Y results from categorizing an underlying latent variable Y^*

$$Y_i=j\Longleftrightarrow\gamma_{j-1}<Y^*\leqslant\gamma_j\quad\text{where}\quad-\infty=\gamma_0<\cdots<\gamma_J=\infty. \tag{7}$$

The underlying latent variable is further modelled as a linear function of the covariates

$$Y_i^*=\mathbf{z}_i'\boldsymbol{\alpha}+\varepsilon_i. \tag{8}$$

When a logistic distribution function is assumed for ε_i, the cumulative logistic link function is obtained. Note that the cumulative logistic function is a valid choice for a link function in its own and does not have to be grounded in an underlying continuous variable approach.

Motivated by the category boundaries approach, a common assumption is that the covariates have the same weight over the categories of the outcome variable. For the illustrative example, this means that the model only incorporating the main effects of the covariates is further simplified:

$$\boldsymbol{\eta}_{im}=\mathbf{Z}_{im}\boldsymbol{\beta}_m=\begin{bmatrix}1 & & \mathbf{z}_{im1}' & \mathbf{z}_{im2}'\\ & 1 & \mathbf{z}_{im1}' & \mathbf{z}_{im2}'\end{bmatrix}\begin{bmatrix}\gamma_1\\ \gamma_2\\ -\boldsymbol{\alpha}\end{bmatrix}. \tag{9}$$

The model incorporates eight parameters: two category boundary parameters, three parameters to code for the main effect of the first covariate and three for the main effect of the second covariate. Hence, the difference between response categories solely relies on a shift in the linear predictor. The number of parameters is only one fourth of the number of parameters of a model with no restrictions on the conditional probabilities.

3. Estimation

When there are no missing data or unobserved variables, standard procedures for generalized linear models (of which the logistic regression model is a special case) such as iteratively weighted least squares can be used to obtain maximum likelihood parameter estimates and corresponding asymptotic standard errors. These procedures operate on the frequency tables of $\{X_m\}\cup pa(X_m)$. Detailed descriptions can be found in many textbooks on generalized linear models (e.g., [4]). In case of missing observations and/or unobserved variables, maximum likelihood estimates can be obtained using the EM-algorithm. Lauritzen [9] described how, in the E-step, the tables of *expected* frequencies of $\{X_m\}\cup pa(X_m)$ can be calculated efficiently by local computations on the junction tree [7]. The M-step is again any standard procedure for generalized linear models, operating on these tables of expected frequencies. Maximum likelihood estimates are consistent when the missing mechanism is ignorable, i.e. when data are missing at random or completely at random in the terminology of [10].

In a logistic regression model, the existence of unique and finite parameter estimates depends on the pattern of the datapoints. For example, in a saturated model, whenever a response category is not observed for a particular combination of parent variables, the corresponding parameter will have a nonunique solution at minus infinity [1]. To avoid finite parameter estimates, one can restrict the parameters to fall within a certain range, or add a small amount to each cell of the frequency table, tantamount to the use of a prior in a Bayesian framework.

A set of Matlab functions that implement the EM-algorithm for Bayesian networks with (ordered or multinomial) logistic regression models for the conditional probability tables can be obtained from the author.

4. Simulation: Alarm network

The original Alarm network [6] is a toy example of a Bayesian network that was crafted by hand without imposing specific restrictions on the conditional probability tables (such as, only main effects of the parent

variables, or restrictions across categories of the child variable). Because the variables in the Alarm network are either binary or ordered (as apparent from the use of terms such as "low", "normal", and "high" as category labels), it is interesting to assess the performance of a Bayesian network with its conditional probability tables restricted according to ordered logistic regression models.

Data were simulated for three different models. In the first model, data were generated from the original Alarm network [6], without restrictions on the conditional probabilities. In a second model, the probabilities for all conditional probability tables were restricted according to an ordered logistic regression model containing category boundary parameters and main effects of the parents. The third model contained unrestricted conditional probabilities for half of the tables as in Model 1 (every even table in alphabetical order); the probabilities for the other half of the tables (every odd table) were obtained from the same ordered logistic regression models as in Model 2. To stay close to the original Alarm network, the generating parameters for the restricted networks were obtained by generating 25,000 cases from the original alarm network, and then estimating them under the restricted model.

Under each of the three models, two datasets were generated for sample sizes of 50, 100, 200, and 500 cases. One dataset was used for parameter estimation (training set), the other for cross-validation (test set). Furthermore, incomplete datasets were created by declaring 20% of the observations in the complete datasets as missing.

Three models were estimated on each training set: the model without restrictions on the conditional probability tables, the model in which all conditional probability tables were restricted according to the aforementioned ordered logistic regression model, and the 'true' model with half of its conditional probability tables unrestricted, and the other half restricted according to the logistic regression model. To avoid infinite parameter estimates, a small amount (0.1% of the sample size N) was added to all frequencies. The cross-validated deviances ($-2 \times$ loglikelihood) divided by sample size are depicted in Figs. 1–3. Fig. 1 displays the results for the case where the true network contained no restrictions on the conditional probability tables. For all sample sizes, the model with no restrictions on the conditional probabilities showed the lowest cross-validated deviance. Hence, the true model was always the preferred one. The model with all conditional probabilities restricted to an ordered logistic regression model containing only main effects performed the worst for all sample sizes, and the model with the conditional probabilities restricted in half of the tables and unrestricted in the other half had always a cross-validated deviance in between the deviances of the other two models.

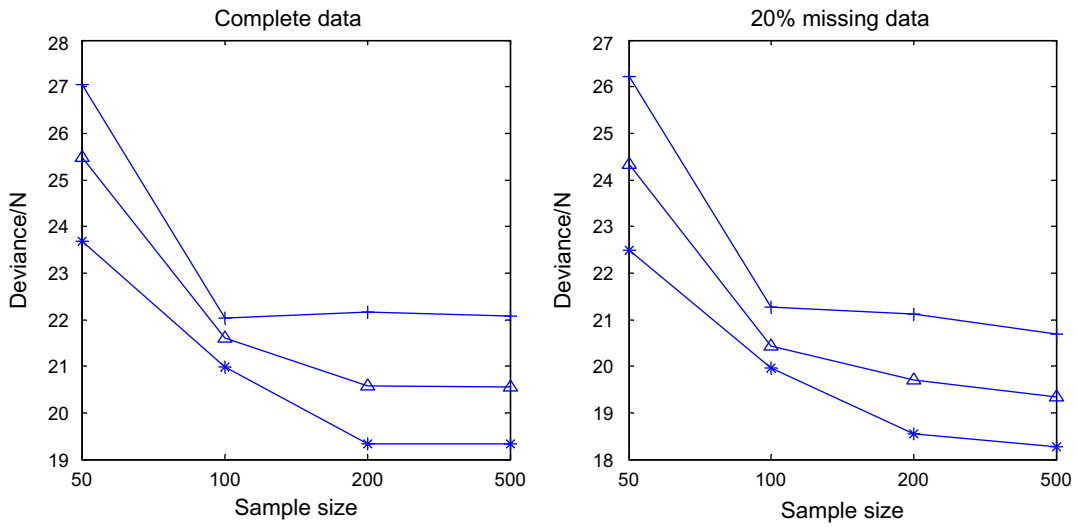

Fig. 1. Cross-validated deviances for the model without restrictions on the conditional probabilities (*; true model), all conditional probabilities restricted according to a logistic regression model with main effects of the parents (+), and 50% of the conditional probability tables without restrictions and 50% restricted according to a logistic regression model with main effects of the parents (Δ). Deviances are divided by sample size.

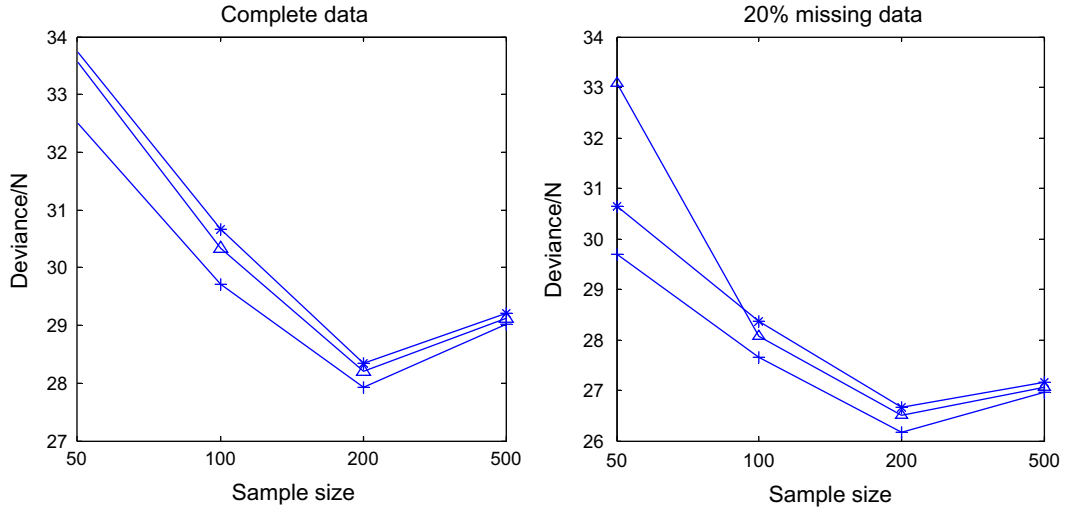

Fig. 2. Cross-validated deviances for the model without restrictions on the conditional probabilities (*), all conditional probabilities restricted according to a logistic regression model with main effects of the parents (+; true model), and 50% of the conditional probability tables without restrictions and 50% restricted according to a logistic regression model with main effects of the parents (Δ). Deviances are divided by sample size.

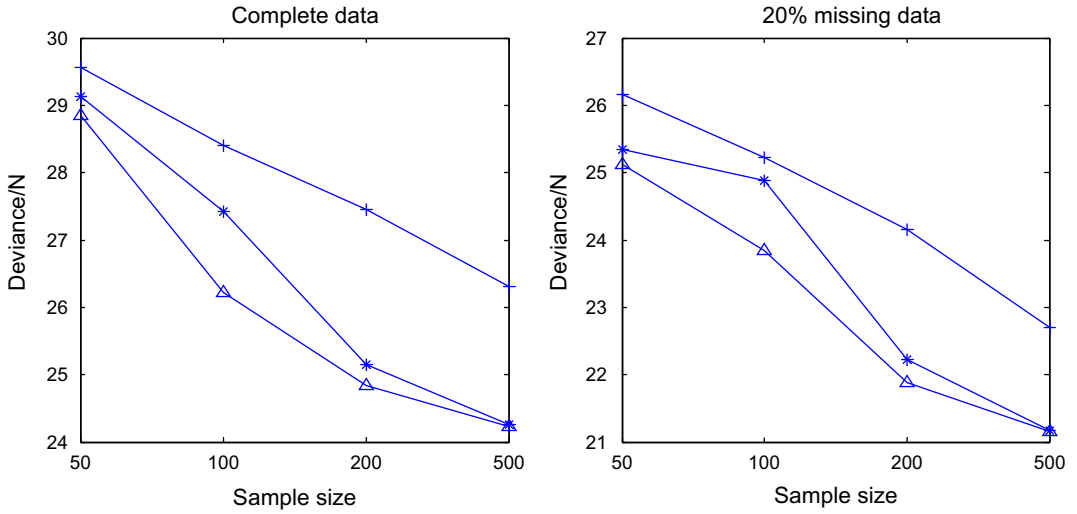

Fig. 3. Cross-validated deviances for the model without restrictions on the conditional probabilities (*), all conditional probabilities restricted according to a logistic regression model with main effects of the parents (+), and 50% of the conditional probability tables without restrictions and 50% restricted according to a logistic regression model with main effects of the parents (Δ; true model). Deviances are divided by sample size.

On the other hand, when the true network was the network in which the conditional probabilities were restricted for all tables, that model outperformed the other two models, see Fig. 2. In addition, the differences between the three models became smaller for a sample size of 500. Apparently, that is the sample size for the Alarm network where the cost of estimating superfluous parameters becomes smaller.

For the third scenario, where the true network had half of its conditional probability tables restricted according to a main effects ordered logistic regression model, and the other half unrestricted, again the true model was always the preferred one, see Fig. 3. The model that performs second best for all sample sizes is

the model with none of its conditional probabilities restricted. Apparently, the cost of an overparameterized model was smaller than the cost of a more parsimonious but incorrectly specified model. Furthermore, for a sample size of 500, the model with no restrictions performed almost as well as the true model, in which the probabilities for half of the conditional probability tables were restricted. As was also observed in Fig. 2, 500 seems to be the sample size for the Alarm network where the cost of estimating superfluous parameters becomes smaller.

In sum, the correct model was always selected: for all sample sizes, both for the complete datasets and for the datasets in which 20% of the data was missing, and irrespective of which model was chosen as the true network.

5. Discussion

In this paper, we described how the conditional probability tables of a Bayesian network can be modelled as logistic regression models, in which the parent variables define the covariates. When all main effects and interactions between the parent variables are incorporated as covariates, the conditional probabilities are estimated without restrictions, as in a traditional Bayesian network. By incorporating interaction terms up to a specific order only, the number of parameters can be drastically reduced. When the categories of a variable are ordered, ordered logistic regression can be used.

Restricting the conditional probabilities not only results in a more parsimonious model, but also enhances the interpretability of a model. A conditional probability table in which all probabilities are estimated freely corresponds to a logistic regression model that contains all interaction terms between covariates. However, interaction terms of a higher order are often quite hard to interpret. By including interaction terms only up to a specific order, this problem is remedied.

When all variables (observed and hidden) are binary, and only main effects of parents are included, a sigmoid belief network is obtained [11]. Hence, the estimation procedure based on generalized linear models outlined above can be used to estimate the parameters of sigmoid belief networks. However, sigmoid belief networks are often characterized by densely connected (hidden) nodes, in which case exact inference becomes intractable and one has to rely on approximate methods [13].

Modeling the conditional probability tables with logistic regression can also be seen as a generalization of the use of 'default tables' [5] to avoid an exponentially growing number of parameters. In a default table, some conditional probabilities are restricted to have the same value. In a logistic regression framework, this is obtained by having identical rows in the design matrix.

In the research reported in this paper, the restrictions on the conditional probability tables were pre-specified for the three investigated models, and subsequently the three models were compared to each other using cross-validation. An obvious next step would be to incorporate into the learning scheme an automatic procedure for selecting, for each table, the proper order of interactions between the parent variables. That is, the learning scheme would alternate repeatedly between estimation and cross-validation. One could either start with a model with no restrictions at all and gradually remove interactions (starting with the highest order interactions), or start with a main-effects model and gradually add interaction terms. The implementation of a flexible automatic procedure would require the development of a routine for constructing design matrices for any particular interaction order that can handle any number of parents which each can have any number of categories.

When there are no missing data, the order of interactions can be selected independently for each conditional probability table because the factors of the likelihood corresponding to each table vary independently. In the presence of missing data however, the expected frequencies (on which estimation is based) in one table may depend on other tables, and hence the selection of a specific interaction order in one table may depend as well on which interaction terms are selected in other tables. Further research is needed to develop adequate heuristics for this case.

Friedman and Goldszmidt [5] used default tables within a structure learning procedure, trading restrictions on the conditional probability tables for a richer set of conditional dependence relations that can be inferred from a given dataset. The use of logistic regression could be incorporated in a structure learning procedure in an analogous way.

References

[1] A. Albert, J.A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, Biometrika 71 (1984) 1–10.
[2] W.L. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE Transactions on Knowledge and Data Engineering 8 (1996) 195–210.
[3] A. Edwards, L. Thurstone, An internal consistency check for scale values determined by the method of successive intervals, Psychometrika 17 (1952) 169–180.
[4] L. Fahrmeir, G. Tutz, Multivariate Statistical Modelling Based on Generalized Linear Models, second ed., Springer-Verlag, New York, 2000.
[5] N. Friedman, M. Goldszmidt, Learning Bayesian networks with local structure, in: M.I. Jordan (Ed.), Learning in Graphical Models, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
[6] E.H. Herskovits, Computer-based probabilistic-network construction. PhD Thesis. Stanford University, 1991.
[7] F.V. Jensen, S.L. Lauritzen, K.G. Olesen, Bayesian updating in causal probabilistic networks by local computation, Computational Statistics Quarterly 4 (1990) 269–282.
[8] M.I. Jordan (Ed.), Learning in Graphical Models, Kluwer Academic Publishers., Dordrecht, The Netherlands, 1998.
[9] S.L. Lauritzen, The EM algorithm for graphical association models with missing data, Computational Statistics and Data Analysis 19 (1995) 191–201.
[10] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 1987.
[11] R.M. Neal, Connectionist learning of belief networks, Artificial Intelligence 56 (1992) 71–113.
[12] R.E. Neapolitan, Learning Bayesian Networks, Pearson Prentice Hall, 2004.
[13] L.K. Saul, T.S. Jaakkola, M.I. Jordan, Mean field theory for sigmoid belief networks, Journal of Artificial Intelligence Research 4 (1996) 61–76.