Parameter Learning for Bayesian Networks with Strict Qualitative Influences

Ad Feelders and Robert van Straalen

Utrecht University, Department of Information and Computing Sciences,
P.O. Box 80089, 3508TB Utrecht, The Netherlands,
ad@cs.uu.nl, Robert.vanStraalen@phil.uu.nl

Abstract. We propose a new method for learning the parameters of a Bayesian network with qualitative influences. The proposed method aims to remove unwanted (context-specific) independencies that are created by the order-constrained maximum likelihood (OCML) estimator. This is achieved by averaging the OCML estimator with the fitted probabilities of a first-order logistic regression model. We show experimentally that the new learning algorithm does not perform worse than OCML, and resolves a large part of the independencies.

1 Introduction

In recent work, Wittig and Jameson [8], Altendorf et al. [1] and Feelders and van der Gaag [4] have shown that the use of qualitative influences can improve the probability estimates in Bayesian networks, in case relatively few observations are available. Apart from improvement of the parameter estimates, in [8] and [4] it was argued that networks with probability estimates that reflect the qualitative knowledge specified by the domain experts are less likely to exhibit counterintuitive reasoning behaviour and are therefore more likely to be accepted.

Feelders and van der Gaag [4] provide a simple algorithm, based on the isotonic regression, to compute the order-constrained maximum likelihood (OCML) estimates for networks of binary variables. A disadvantage of the OCML estimates is that in case order reversals are present in the unconstrained estimates, these are resolved by setting blocks of violating estimates equal to their weighted average. This results in unwanted (context-specific) independencies in the network.

We present a new estimator that aims at enforcing strict inequalities between parameters, thereby avoiding the creation of unwanted independencies in the network. This is achieved by combining the OCML with a first-order logistic regression model.

The paper is organized as follows. In section 2 we introduce the necessary notation, and introduce some important concepts that are used throughout the paper. In section 3 we discuss parameter learning with qualitative influences, and explain the shortcoming of the order-constrained maximum likelihood estimator.Subsequently, in section 4, we discuss our compound estimator that aims

to remove this shortcoming. This new parameter learning method is evaluated experimentally in section 5. We end with conclusions.

2 Preliminaries

A *Bayesian network* is a concise representation of a joint probability distribution over a collection of stochastic variables $\mathbf{V} = (V_1, \ldots, V_m)$; in the sequel, we assume all variables to be binary, taking the value 0 or 1. The network consists of an acyclic directed graph in which each node corresponds to a variable and the arcs capture the dependence structure of the distribution; the network further includes a number of conditional probabilities, or *parameters*, $P(V_i \mid \mathbf{V}_{\mathrm{pa}(i)})$ for each variable V_i given its parents $\mathbf{V}_{\mathrm{pa}(i)}$ in the graph. The graphical structure and associated probabilities together represent a unique joint probability distribution over the variables involved, which is factorised according to

$$\Pr(\mathbf{V}) = \prod_{i=1}^{m} P(V_i \mid \mathbf{V}_{\mathrm{pa}(i)})$$

In estimating the parameters from data, we only have to consider one node and its parents at a time. To simplify notation, we will do so in the sequel. Let $\mathbf{X} = (X_1, \ldots, X_k)$ be the parents of a variable Y, and let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k = \{0,1\}^k$ consist of vectors $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ of values for the k variables in $\mathbf{X}$, that is, $\mathcal{X}$ is the set of all *parent configurations* of Y. Slightly abusing terminology, we sometimes say that X_i *occurs* or *is present* if it has the value one. We write $\mathbf{X}_a$ for the sub-vector of $\mathbf{X}$ containing the variables X_j for $j \in a$, where a is a subset of $K = \{1, \ldots, k\}$. We further write $\mathbf{X}_{-a}$ for $\mathbf{X}_{K \setminus a}$, and $\mathbf{X}_{-i}$ for $\mathbf{X}_{K \setminus \{i\}}$, where $i \in K$. Furthermore, we write for example $n(y, \mathbf{x})$ to denote the number of observations in the data with $Y = y$ and $\mathbf{X} = \mathbf{x}$.

A *qualitative influence* [7] between two variables expresses how observing a value for the one variable affects the probability distribution for the other variable. A *positive* influence of X_i on Y along an arc $X_i \to Y$ means that the occurrence of X_i *does not decrease* the probability that Y occurs, regardless of any other direct influences on Y, that is

$$P(y = 1 | x_i = 1, \mathbf{x}_{-i}) \geq P(y = 1 | x_i = 0, \mathbf{x}_{-i}), \tag{1}$$

where $\mathbf{x}_{-i}$ is any configuration of the parents of Y other than X_i. Since the inequality in (1) is not strict, the technically correct, if somewhat awkward, verbal description is *does not decrease* rather than *increases*. In this paper, the distinction between the two is crucial however. Similarly, there is a *negative* influence of X_i on Y along an arc $X_i \to Y$ if the occurrence of X_i *does not increase* the probability that Y occurs. From now on we assume, without loss of generality, that all qualitative influences are positive.

Finally, we say a positive qualitative influence is *strict* if the occurrence of X_i *increases* the probability that Y occurs, regardless of any other direct influences on Y, that is

$$P(y = 1 | x_i = 1, \mathbf{x}_{-i}) > P(y = 1 | x_i = 0, \mathbf{x}_{-i}). \tag{2}$$

3 Parameter learning with qualitative influences

As we saw in the previous section, qualitative influences correspond to certain constraints between the parameters in the Conditional Probability Table (CPT) of Y. In earlier work, several methods have been proposed to exploit these constraints in estimating the parameters from data. In [4] it was shown how the order-constrained maximum likelihood (OCML) estimates, using the non-strict inequalities in (1), can be computed using the isotonic regression. The problem with these estimates is that they create unwanted (context-specific) independencies. This is illustrated by the following example.

Consider a node Y with two parents, X_1 and X_2, both with a positive influence on Y. Suppose we observe the data given in the left part of Table 1.

Table 1. Data for example. In the left table, each cell gives $n(y = 1, \mathbf{x})/n(\mathbf{x})$. The middle table gives the unconstrained ML estimates of $P(Y = 1|X_1, X_2)$. The table on the right gives the order-constrained ML estimates.

X_1/X_2	0	1
0	4/10	1/3
1	6/10	18/20

X_1/X_2	0	1
0	0.40	0.33
1	0.60	0.90

X_1/X_2	0	1
0	0.38	0.38
1	0.60	0.90

The first row of the unconstrained estimates contains an order reversal, since it is decreasing rather than increasing. The OCML estimator resolves this order violation by taking the weighted average of the two violating cells, and assigning this value to the both of them. Hence the value

$$\frac{4+1}{10+3} = \frac{5}{13} \approx 0.38$$

in the first row of the rightmost table in Table 1. The result is a context-specific independence: according to the OCML estimates, Y is independent of X_2 for $X_1 = 0$. This is probably not what the expert intended when she specified a positive influence of X_2 on Y. This raises the question if it wouldn't be better to enforce the strict inequalities given in equation (2) rather than the non-strict inequalities given in (1). The problem is that for the strict inequalities, the OCML estimates do not exist in case one of the order constraints is violated. This can be intuitively appreciated by looking at the solution obtained in the example above. In the strict version the likelihood would keep increasing as we make the difference between the violating estimates $\hat{P}(Y = 1|0, 0)$ and $\hat{P}(Y = 1|0, 1)$ smaller and smaller.

In order to enforce strict inequalities, Altendorf et al. [1] specified a minimum margin between two "contiguous" parameters, that is, (1) was replaced by

$$P(y = 1|x_i = 1, \mathbf{x}_{-i}) \geq P(y = 1|x_i = 0, \mathbf{x}_{-i}) + \varepsilon_{\mathbf{x}_{-i}}, \tag{3}$$

where $\varepsilon_{\mathbf{x}_{-i}}$ is the minimum difference required, which in general may depend on the values at which the remaining parents are held constant. The problem with this approach is the selection of appropriate values for the $\varepsilon_{\mathbf{x}_{-i}}$. One could try to elicit these margins from domain experts. Experience shows however that experts have a hard time in reliably providing such numerical information. If they had been very good at it, we might just as well have asked them for the CPTs straight away. Another possibility is to specify some fixed value for the margin, that is applied to every pair of contiguous parameters. This option is chosen in [1]. The problem is, however, that the choice of value for ε becomes arbitrary.

Therefore, we propose a different, data-driven, approach to obtain strict inequalities between contiguous parameters. This approach is presented in the next section.

4 Learning with strict qualitative influences

The basic idea of our approach is to try to remove unwanted independencies by combining the OCML estimates with estimates produced by a logistic regression model. A similar idea in the different setting of numeric isotonic regression with just one predictor variable was proposed by Wright [9]. For the logistic regression model, we assume the log-odds is linear in the parent variables, that is

$$\log\left\{\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}\right\} = \beta_0 + \sum_{i=1}^{k} \beta_i X_i \tag{4}$$

The important property of this model is that $\beta_i > 0$ corresponds to a strictly positive influence of X_i on Y. The proposed compound estimator $\hat{P}^*$ is given by

$$\hat{P}^*(Y|\mathbf{X}) = \gamma \hat{P}_{\text{LR}}(Y|\mathbf{X}) + (1-\gamma)\hat{P}_{\text{OCML}}(Y|\mathbf{X}),$$

where $0 < \gamma < 1$, and $\hat{P}_{\text{LR}}(Y|\mathbf{X})$ are the fitted probabilities of a first-order logistic regression model. By taking the weighted average with the fitted logistic regression probabilities, we enforce strict inequalities as long as the signs of the logistic regression coefficients are positive. There are still two loose ends in this proposal

1. What if one or more of the logistic regression coefficient estimates are negative instead of positive?
2. How do we choose the value of γ, the weight of the logistic regression model?

The first issue is addressed as follows. We start by estimating the full model, i.e. including all parents of Y as predictors. Then we check if there are any parents having a negative coefficient. If so, these parents are removed from the model, and the model is re-estimated with the remaining parents. If necessary, this procedure is repeated until only parents with positive coefficients remain. Note that the removal of X_i from the model (i.e. setting $\beta_i = 0$) results in

$$\hat{P}_{\text{LR}}(y=1|x_i=1,\mathbf{x}_{-i}) = \hat{P}_{\text{LR}}(y=1|x_i=0,\mathbf{x}_{-i}), \tag{5}$$

for every configuration $\mathbf{x}_{-i}$. Hence, if the OCML estimates satisfy

$$\hat{P}_{\text{OCML}}(y = 1|x_i = 1, \mathbf{x}_{-i}) = \hat{P}_{\text{OCML}}(y = 1|x_i = 0, \mathbf{x}_{-i}), \tag{6}$$

for one or more configurations $\mathbf{x}_{-i}$, these equalities will *not* be resolved by the compound estimator.

With respect to the second issue, we should keep in mind that the primary reason to combine the OCML estimates with the logistic regression model is to obtain data-driven margins between contiguous parameters. We do not want its weight to become too big, unless the LR model actually gives a good fit of the data. Therefore, we chose to let γ depend on the p-value of the observed deviance of the fitted logistic regression model, under the null hypothesis that the logistic regression model is the correct model specification.

To illustrate the basic idea of our compound estimator, we continue the example of section 3. Recall that the OCML estimates created an independence between X_2 and Y for $X_1 = 0$. To remove this independence, we combine the OCML estimates with those obtained by fitting a logistic regression model with Y as the response, and X_1 and X_2 as the predictors. This results in positive coefficients for both X_1 ($\hat{\beta}_1 = 1.47$) and X_2 ($\hat{\beta}_2 = 1.09$). Note that X_2 still has a positive coefficient, because the order reversal in the first row of the observed relative frequencies is more than compensated by the increasing second row. If the second row were also decreasing, X_2 would have had a negative coefficient, and consequently the problem could not be fixed by the compound estimator. The fitted probabilities of the logistic regression model are given in Table 2.

Table 2. Fitted probabilities of the logistic regression model estimated on the data in Table 1 on the left. On the right, the fitted probabilities of the compound estimator.

X_1/X_2	0	1
0	0.32	0.59
1	0.68	0.86

X_1/X_2	0	1
0	0.37	0.43
1	0.62	0.89

To compute the compound estimator, we still have to determine the weight of the LR model. Here we take γ equal to the p-value; in the experimental section, we also consider two other possibilities. We compute the p-value by using the fact that the deviance has approximately a $\chi^2_{n-k'-1}$ distribution under the null hypothesis, where k' is the number of remaining variables in the logistic regression model. Hence, we look up the area of the tail of this distribution to the right of the observed deviance, and find a p-value of approximately 0.26. Hence the compound estimator becomes

$$\hat{P}^*(Y = 1|X_1, X_2) = 0.26 \cdot \hat{P}_{\text{LR}}(Y = 1|X_1, X_2) + 0.74 \cdot \hat{P}_{\text{OCML}}(Y = 1|X_1, X_2),$$

The compound estimates are given in Table 2 on the right.We have achieved our goal: the unwanted equality in the first row has been resolved, and the margin between the two contiguous probabilities has been determined by the data.

5 Experiments

To evaluate the proposed parameter learning method, we performed experiments on both artificial data and real data. We are interested in two aspects of performance:

1. How good are the estimates produced by our compound estimator?
2. How many of the unwanted independencies in the OCML estimates are fixed, in the sense that our compound estimator turns them into strict inequalities?

Our objective is to remove as many unwanted independencies as possible, while retaining high quality estimates.

5.1 Artificial data

In order to test our approach under various circumstances, we set up several environments to measure the performance of the proposed compound estimator. Parameters in these test environments are

1. The number of parent nodes: 2, 3 and 4.
2. The size of the data sample: 20, 50 and 100 for 2 or 3 parents; 50, 100 and 200 for 4 parents.
3. The margins between the conditional probabilities $P(Y = 1|\mathbf{X})$ for contiguous values of $\mathbf{X}$ of the underlying true distribution: small and large.

To elaborate on the third point: we suspect that with small margins, reversed signs of the coefficients in the LR model will occur more often, leading to a lower fraction of resolved independencies. An example of small and large margins for the case of two parents is given in Figure 1.

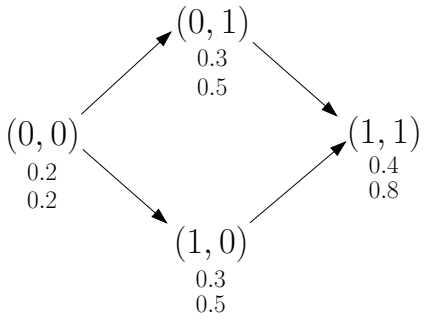

Fig. 1. Small and large margins between $P(Y = 1|\mathbf{X})$ for contiguous values of $\mathbf{X}$. The first number below each parent configuration $\mathbf{x}$ denotes $P(Y = 1|\mathbf{x})$ corresponding to the small margin; the second number corresponds to the large margin. Arrows between the parent configurations denote direct precedence in the order ("contiguous" configurations)

We considered three definitions of the LR weight γ in the experiments:

1. γ_1 = p-value: Just the p-value as discussed in section 4.
2. $\gamma_2 = \min(1, \frac{2^{|\mathcal{X}|}}{n} \cdot \text{p-value})$.
3. $\gamma_3 = \min(0.1, 10 \cdot \text{p-value})$: The p-value is multiplied by 10, to avoid γ being too small. Also it has an upper threshold of 0.1 to avoid too large values. Here an extra factor is included, which decreases as the sample size n (relative to the number of parameters) increases. It has a maximum of 1 and goes to 0 as n goes to infinity. This definition is inspired by the view that with more data, the ML (and OCML) estimates become more reliable, requiring a smaller weight for the LR model.

It is clear that none of these three weighting methods has any deep theoretical justification, so the experiments will have to make clear which one works best.

In our experiments we fitted five different models: $\hat{P}_{ML}$ (the unconstrained maximum likelihood estimates), $\hat{P}_{OCML}$ (the order-constrained maximum likelihood estimates), and $\hat{P}_i^*$, $i = 1, 2, 3$ (the compound estimates, using γ_i). To determine the quality of the estimates, we computed the Kullback-Leibler (KL) divergence between the true and the fitted distributions. We applied the Laplace correction to avoid possible infinity values. We performed a 1000 replications of the experiment, and averaged the Kullback-Leibler divergence over these 1000 replications. The results are given in Tables 3, 4, and 5, for 2,3, and 4 parents respectively.

Table 3. Results for artificial data, 2 parent nodes.

ntrain	KL_{ML}	KL_{OCML}	KL_{C1}	KL_{C2}	KL_{C3}	fixratio
distribution with large margins						
20	0.0598	0.0436	0.0692	0.0448	0.0443	63%
50	0.0319	0.0277	0.0278	0.0277	0.0274	89%
100	0.0185	0.0172	0.017	0.0172	0.017	99%
distribution with small margins						
20	0.0576	0.0342	0.0539	0.0353	0.0345	31%
50	0.0305	0.0197	0.0207	0.0197	0.0195	43%
100	0.0184	0.0129	0.013	0.0129	0.0128	49%

There are several general observations clearly shown in the tables. First of all, the ML model has the worst performance. Second, the compound estimator using γ_1 does not perform well: the LR model becomes too dominant and spoils the estimates. Third, the OCML model and the models using γ_2 and γ_3 are close, but rather consistently the compound estimator using γ_3 performs the best by a small margin. Finally, as the sample size increases, the differences in performance become smaller. Summarizing, we can say that with the right weight (γ_3) for the LR model, the compound estimator works well under all circumstances we have studied in the experiments.

The next question is, what fraction of the unwanted independencies created by OCML are resolved by the compound estimator. This fraction is referred to

Table 4. Results for artificial data, 3 parent nodes.

ntrain	KL_{ML}	KL_{OCML}	KL_{C1}	KL_{C2}	KL_{C3}	fixratio
distribution with large margins						
20	0.0718	0.0325	0.0719	0.0381	0.0339	58%
50	0.0514	0.0286	0.0299	0.0287	0.0282	78%
100	0.0329	0.0223	0.0219	0.0223	0.0212	91%
distribution with small margins						
20	0.0702	0.0324	0.0868	0.037	0.0332	42%
50	0.0519	0.0231	0.025	0.0233	0.0228	48%
100	0.0325	0.0163	0.0163	0.0163	0.0157	62%

Table 5. Results for artificial data, 4 parent nodes.

ntrain	KL_{ML}	KL_{OCML}	KL_{C1}	KL_{C2}	KL_{C3}	fixratio
distribution with large margins						
50	0.0716	0.0314	0.0348	0.0313	0.0302	84%
100	0.0522	0.0265	0.0256	0.0263	0.0247	93%
200	0.0336	0.0213	0.0206	0.0212	0.0196	99%
distribution with small margins						
50	0.0732	0.0312	0.038	0.0304	0.0289	59%
100	0.0516	0.0215	0.0212	0.0213	0.02	71%
200	0.0321	0.0143	0.0138	0.0143	0.0132	85%

in the following as the *fixratio*. As we expected, the fixratio is higher for large margins, and it also increases with the sample size. Table 6 shows the average fixratios.

Table 6. *Average fixratios.*

distribution	average fixratio
large margins	84%
small margins	54%
overall	69%

5.2 Real data

We also performed experiments on five real datasets, four of which have been taken from the UCI repository [2]. All non-binary variables were made binary by making approximately equal-frequency bins of 0 and 1-values. We used the dependent variable and a selection from the attributes to construct a small Bayesian network fragment. To determine the sign of the influences, we used the

information attached to the datasets, previous articles [1] and [3], and commonsense. We did not look at the data itself to determine the signs. The data sets used are:

- *Windsor housing data*: used as an example by Koop [5]. It has 546 examples of sale prices of houses dependent on different variables. We take the sale price as the child variable, and lot size and presence or absence of a basement and air conditioning as parent nodes with positive influences.
- *Wisconsin Breast Cancer Database* [6]: 699 instances of data about the class (benign or malignant) of breast tumors. We used the class as the child variable and clump thickness, single epithelial cell size and bland chromatin as parent nodes, all with positive influences.
- *Pima Indians diabetes*: 768 cases of medical data on Pima Indians. The class variable denotes whether the person has diabetes or not. We used the body mass index, diabetes pedigree function and age of the person as parent nodes, all with positive influences.
- *Haberman's Survival Data*: 306 cases from a study on the survival of patients who had undergone surgery for breast cancer. Survival of the patient is the child variable. The patient's age, year of operation and number of nodes detected are the parent nodes, again all with positive influences.
- *CPU Performance Data*: 209 cases of characteristics and performance of various types of computer CPU models. Performance is the dependent variable. Machine cycle time, minimum main memory and maximum main memory are the parent nodes. Machine cycle time has a negative influence, the other two parents have a positive influence.

Like with the artificial data, we performed a thousand replications of the experiment. Table 7 shows the results. For all estimators, we computed the average log-loss and its standard error on the data not used for training as follows:

$$\mathcal{L}_{\text{test}} = \frac{-\sum_{i=1}^{\text{ntest}} \log \hat{P}(y_i|\boldsymbol{x}_i)}{\text{ntest}},$$

where ntest is the number of observations in the test set. The results are fairly in line with those obtained on the artificial data. Model C3 (the compound model using γ_3) seems to perform slightly better than the other estimators again. Table 8 shows the average fixratios per dataset. Table 9 shows the average fixratios per sample size. The overall fixratio of the compound estimator is about 68%, and increases with the sample size.

6 Conclusions

We have proposed a new method for learning the parameters of a Bayesian network with qualitative influences. This method aims at avoiding unwanted independencies created by the order-constrained maximum likelihood (OCML) estimator. To obtain data-driven margins between contiguous parameters, the

Table 7. Average log-loss for respectively the Windsor housing, Wisconsin breast cancer, Pima Indians, Haberman survival, and CPU performance datasets for different sizes of the training sample.

ntrain	$\mathcal{L}_{ML}$	$\mathcal{L}_{OCML}$	$\mathcal{L}_{C1}$	$\mathcal{L}_{C2}$	$\mathcal{L}_{C3}$	fixratio
20	0.5888	0.5667	0.7624	0.5689	0.5645	70%
	(± 0.043)	(± 0.026)	(± 0.859)	(± 0.032)	(± 0.027)	
50	0.5539	0.5428	0.5634	0.5427	0.5418	89%
	(± 0.027)	(± 0.02)	(± 0.123)	(± 0.02)	(± 0.02)	
100	0.5354	0.5293	0.5305	0.5291	0.5284	99%
	(± 0.018)	(± 0.015)	(± 0.022)	(± 0.015)	(± 0.015)	
20	0.2999	0.2944	1.1285	0.2677	0.2853	5%
	(± 0.03)	(± 0.023)	(± 1.13)	(± 0.031)	(± 0.024)	
50	0.2415	0.2396	0.4461	0.2357	0.237	82%
	(± 0.017)	(± 0.015)	(± 0.435)	(± 0.015)	(± 0.015)	
100	0.2189	0.2182	0.2458	0.2175	0.2174	91%
	(± 0.011)	(± 0.011)	(± 0.156)	(± 0.01)	(± 0.01)	
20	0.6526	0.6164	0.6901	0.6198	0.6151	54%
	(± 0.042)	(± 0.024)	(± 0.332)	(± 0.03)	(± 0.025)	
50	0.6223	0.6024	0.6085	0.6023	0.6011	72%
	(± 0.027)	(± 0.018)	(± 0.067)	(± 0.018)	(± 0.018)	
100	0.6008	0.5916	0.5909	0.5914	0.5904	79%
	(± 0.018)	(± 0.013)	(± 0.014)	(± 0.013)	(± 0.013)	
20	0.6157	0.5872	0.7017	0.5883	0.5843	41%
	(± 0.043)	(± 0.028)	(± 0.572)	(± 0.034)	(± 0.03)	
50	0.5837	0.5648	0.5735	0.5647	0.5635	47%
	(± 0.027)	(± 0.019)	(± 0.057)	(± 0.02)	(± 0.02)	
100	0.5657	0.5539	0.5547	0.5538	0.5531	56%
	(± 0.026)	(± 0.021)	(± 0.026)	(± 0.021)	(± 0.021)	
20	0.499	0.4883	0.7738	0.4877	0.4849	56%
	(± 0.036)	(± 0.025)	(± 0.9)	(± 0.032)	(± 0.026)	
50	0.4728	0.4673	0.5248	0.4665	0.466	84%
	(± 0.028)	(± 0.024)	(± 0.249)	(± 0.026)	(± 0.025)	
100	0.4605	0.4579	0.4603	0.4572	0.4568	97%
	(± 0.038)	(± 0.038)	(± 0.061)	(± 0.038)	(± 0.038)	

Table 8. *Average fixratios per dataset.*

dataset	average fixratio
Windsor housing	86%
Wisconsin breast cancer	60%
Pima indians diabetes	68%
Haberman survival	48%
CPU performance	79%
Overall	68%

Table 9. *Average fixratios per sample size.*

sample size	average fixratio
20	45%
50	75%
100	84%
Overall	68%

OCML estimates are combined with the fitted probabilities of a first-order logistic regression model. We have shown that the new compound estimator performs well, and is able to remove a large fraction of the independencies created by the OCML estimator.

References

1. E.A. Altendorf, A.C. Restificar, and T.G. Dietterich. Learning from sparse data by exploiting monotonicity constraints. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 18–25. AUAI Press, 2005.
2. C.L. Blake and C.J. Merz. UCI repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/mlrepository.html], 1998.
3. A. Feelders and M. Pardoel. Pruning for monotone classification trees. In M.R. Berthold, H-J. Lenz, E. Bradley, R. Kruse, and C. Borgelt, editors, *Advances in Intelligent Data Analysis V*, Lecture Notes in Computer Science 2810, pages 1–12. Springer, 2003.
4. A. Feelders and L. van der Gaag. Learning Bayesian network parameters with prior knowledge about context-specific qualitative influences. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 193–200. AUAI Press, 2005.
5. Gary Koop. *Analysis of Economic Data*. John Wiley and Sons, 2000.
6. Olvi L. Mangasarian, W. Nick Street, and William H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Technical Report MP-TR-1994-10, 1994.
7. M.P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44:257–303, 1990.
8. F. Wittig and A. Jameson. Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks. In C. Boutilier and M. Goldszmidt, editors, *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 644–652. Morgan Kaufmann, 2000.
9. F.T. Wright. Estimating strictly increasing regression functions. *Journal of the American Statistical Association*, 73(363):636–639, 1978.