Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Adaptive Thresholding in Structure Learning of a Bayesian Network

Boaz Lerner, Michal Afek, Rafi Bojmel
Ben-Gurion University of the Negev, Israel
boaz@bgu.ac.il; {caspimic,rafibojmel}@gmail.com

Abstract

Thresholding a measure in conditional independence (CI) tests using a fixed value enables learning and removing edges as part of learning a Bayesian network structure. However, the learned structure is sensitive to the threshold that is commonly selected: 1) arbitrarily; 2) irrespective of characteristics of the domain; and 3) fixed for all CI tests. We analyze the impact on mutual information – a CI measure – of factors, such as sample size, degree of variable dependence, and variables' cardinalities. Following, we suggest to adaptively threshold individual tests based on the factors. We show that adaptive thresholds better distinguish between pairs of dependent variables and pairs of independent variables and enable learning structures more accurately and quickly than when using fixed thresholds.

1 Introduction

Constraint-based (CB) structure-learning algorithms of Bayesian networks (BN) use conditional independence (CI) tests to threshold information-based measures, such as mutual information (MI), or statistics, such as X^2 or G. CB algorithms are popular, but arbitrariness in CI testing increases the chances of learning redundant edges, removing true edges, and orienting edges wrongly or not at all, and thus learning an erroneous structure [10, 14].

Arbitrariness is a result of two main sources: 1) inaccurate estimation of the CI measure, especially using small samples and large cardinalities of the conditioning set and the involved variables (henceforth "cardinality"), and 2) uneducated selection of the threshold on the measure.

As the degrees of freedom (df) for χ^2 and the cardinality for conditional MI (CMI) increase, X^2 statistic and CMI increase, respectively, because both sum increasing numbers of non-negative terms. As the sample size increases, the statistic increases and the bias between the estimated $\widehat{\mathrm{MI}}$ and real MI decreases [32]. Compared to X^2, CMI has no means, such as df, to adapt to increase in cardinality, and both measures have no means to adapt to the sample size. So far, the analysis of cardinality (or df) and sample size (henceforth "factors"), as sources of arbitrariness in estimating CI measures, was limited [10, 12, 14, 15, 31, 33] yielding no practical ways to tackle this arbitrariness.

Regarding the second main source of arbitrariness – threshold selection – a wrong threshold influences structure learning [2, 12, 14, 20]. Commonly, a threshold is selected according to three main methods: as a user default [10, 29, 31, 33], based on limited trial and error experimentation [5], or automatically as the threshold, from a range of candidates, that max(in)imizes a performance measure [9, 14, 20, 23, 35]. The first method provides simplicity and low runtime, but cannot guarantee satisfactory performance for most problems. The second method overcomes exhaustive experimentation using a limited number of candidates, but it is manual and incomplete. The third method may find a global optimal threshold but may also suffer from a long runtime.

A threshold is usually not only selected arbitrarily but also irrespectively of test factors. While the critical value increases with df, it is indifferent to sample size. The CMI threshold is indifferent to all factors. To our best knowledge, no idea for setting a CMI threshold that depends on the factors has ever been offered. Also, no proposal has ever been given to adapt a threshold to factors of each CMI test.

We suggest adaptive thresholding for individual CI tests based on test factors, and apply this concept in CMI-based CI testing. Statistical tests are well established and consistent but compared to CMI tests can only reject independence, and not confirm dependence as CMI, and lose accuracy due to multiple comparisons. Section 2 reviews common tests and motivates adaptive thresholding. Section 3 suggests adaptive thresholds, which depend on test factors. Section 4 shows that CMI-based adaptive thresholds: 1) follow MI sensitivity to the factors; 2) distinguish pairs of dependent from pairs of independent variables better than fixed thresholds; 3) enable learning structures that are significantly more accurate than if using fixed thresholds; 4) facilitate complexity and expedite learning; and 5) allow learning structures that fit and classify data more accurately. Conclusions are in Section 5.

2 Background

BN is a graphical model that describes independence relations between variables. Its structure is a directed acyclic graph (DAG) composed of nodes representing variables and directed edges that connect nodes and probabilistically quantify their connection [21, 28, 31].

Learning a BN structure is NP-hard [12], and we focus on the CB approach to structure learning [9, 10, 20, 28, 31, 35]. First, CB methods (starting from a complete graph) test pairs of variables to determine whether edges should be

removed. If a statistical test is used, the null hypothesis is that the two variables are (conditionally) independent. If the *p*-value, corresponding to the probability of obtaining a value as extreme as the one observed, given that the null is true, is less than the significance level, α, the null is rejected [25]. If a CMI test is used [13], two (conditionally) independent variables are assumed (for an infinite sample size) to share a zero CMI [30], hence we test them (for a finite sample size) against a small threshold [10]. Failing in the test (CMI) or failing to reject the null (χ^2) indicates node (conditional) independence that yields edge removal.

By performing tests for conditioning sets of different sizes, we form an undirected graph [10, 29, 31, 35] that is directed using orientation rules [17, 26, 27]. The result is a complete partial DAG (CPDAG) that shares the same oriented edges with the true DAG and uses undirected edges to represent statistically indistinguishable connections in the true graph [11, 31]. We concentrate on ways to improve CI testing in domains with discrete variables and complete data.

2.1 Common CI Tests

Common ways of CI testing are by thresholding CMI or a statistic that measures statistical independence between variables (in Pearson's chi-square or likelihood ratio G-test).

Mutual information MI between variables X and Y measures the amount of information shared between these variables. It also measures how much uncertainty about Y decreases when X is observed (and vice versa) [13]

$$MI(X;Y) = \sum_{x\epsilon X, y\epsilon Y} P(x,y)\log\{P(x,y)/P(x)P(y)\}. \quad (1)$$

MI is the KL divergence between $P(x,y)$ and $P(x)P(y)$ [13], measuring how much the joint is different from the marginals' product, or how much the variables can be considered not independent. CMI between X and Y measures information flow between X and Y given a conditioning set $\boldsymbol{S}$

$$\mathrm{CMI}(X;Y|\boldsymbol{S}) = \sum_{x\epsilon X, y\epsilon Y, s\epsilon \boldsymbol{S}} P(x,y,s)\log\frac{P(x,y|s)}{P(x|s)P(y|s)}. \quad (2)$$

By definition, MI(X;Y) and CMI(X;Y|**S**) are non-negative. MI(X;Y) = 0 (CMI(X;Y|**S**) = 0) iff X and Y are independent (given $\boldsymbol{S}$). The true MI is unknown, and the estimated $\widehat{\mathrm{MI}}$ is larger than MI [32], and thus for independent variables larger than 0. Practically, $\widehat{\mathrm{MI}}$ is compared to a small threshold, ε, to distinguish pairs of dependent and pairs of independent variables [2, 6, 9, 10]. If $\widehat{MI}(X;Y) < \varepsilon$, X and Y are regarded as independent and the edge connecting them is removed. The test for CI using CMI is similar.

Pearson's chi-square and G test Statistical tests compare the null hypothesis that two variables are independent with the alternative hypothesis. If the null is rejected (cannot be rejected), the edge is learned (removed). A statistic, which is asymptotically chi-square distributed, is calculated and compared to a critical value. If it is greater (smaller) than the critical value, the null is rejected (cannot be rejected) [1, 31]. In Pearson's chi-square test, the statistic X^2_{st} is

$$X^2_{\mathrm{st}} = \sum_{x\epsilon X, y\epsilon Y} (O_{xy} - E_{xy})^2/E_{xy} \sim \chi^2_{\mathrm{d.f}=(|X|-1)(|Y|-1)}, \quad (3)$$

where O_{xy} (E_{xy}) is the number of records (expected to be if the null was correct) for which $X = x$, $Y = y$, and $|X|$ and $|Y|$ are the corresponding cardinalities. If the null is correct, $P(x,y) = P(x)\cdot P(y), \forall x\epsilon X, y\epsilon Y$, and we expect that $E_{xy}/N = (E_x/N)\cdot(E_y/N), \forall x\epsilon X, y\epsilon Y$ and $E_{xy} = E_x \cdot E_y/N$ for E_x and E_y, which are the numbers of records in which $X = x$ and $Y = y$, respectively, and N is the total number of records. If X^2_{st} is larger than a critical value for a significance value α, $X^2_{st} > X^2_{\mathrm{d.f}=(|X|-1)(|Y|-1),\alpha}$, then we reject the null.

Instead, based on maximum likelihood, if the statistic G_{st}

$$G_{\mathrm{st}} = 2\cdot \sum_{x\epsilon X, y\epsilon Y} O_{xy}\, ln(O_{xy}/E_{xy}) \sim \chi^2_{\mathrm{d.f}=(|X|-1)*(|Y|-1)} \quad (4)$$

is larger than the previous critical value $G_{\mathrm{st}} > X^2_{d.f=(|X|-1)(|Y|-1),\alpha}$, then we reject the null.

2.2 Why Adaptive Thresholding?

Too high a threshold applied in CI testing leads to many unjustified edge removals and thus a sparse network that represents the domain partially. Too low a threshold yields erroneous learning of too many edges and thus a dense network that overfits the data and requires high-order tests, which decrease reliability and increase complexity and run-time. Clearly, both networks are wrong. Not only a threshold should not be set arbitrarily, but it also should be sensitive to the factors (cardinalities and sample size) that affect the measure. In this study, we concentrate on, and analyze, the sensitivity of the MI measure to the factors.

MI sensitivity to sample size If the sample was unlimited, $\widehat{\mathrm{MI}}$ would estimate MI accurately, $\widehat{\mathrm{MI}}$ of a pair of independent variables would be 0, and a threshold would be redundant. But, the sample size is finite and affects MI [32, 33]. As the sample size decreases, the bias between $\widehat{\mathrm{MI}}$ and MI increases, and $\widehat{\mathrm{MI}}$ for independent variables increases from 0. Then, it is more complicated to find a threshold to distinguish between dependent and independent pairs. Moreover, the sample size changes with the domain.

MI sensitivity to cardinality Asymptotically for independent variables, MI(X;Y) = 0, and there is no dependence on variable cardinality. However, cardinality affects MI [18, 32, 34]. Since $\mathrm{MI}(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, we can write $\mathrm{MI}(X;Y) \le \min\{H(X), H(Y)\}$. For totally dependent variables (i.e., have the same values), $\mathrm{MI}(X;X) = H(X)$. As variable cardinality increases, variable entropy increases (due to increased uncertainty) as does its MI with other variables (e.g., for a uniform distribution, the entropy is log of cardinality [13]). Also statistical tests (3) (4) depend on the cardinality (df). Naturally, the cardinality changes with the CI test and so should do the threshold.

3 Adaptive Thresholds

Focusing on MI, our adaptive thresholds change their values among CI tests and, similar to MI, increase with the cardinality and decrease with the sample size.

3.1 MI Threshold Candidates

C_1 threshold A relation between MI and $\widehat{\mathrm{MI}}$ is [32]

$$\widehat{\mathrm{MI}}(X,Y) = \mathrm{MI}(X,Y) + \textstyle\sum_{m=1}^{\infty} C_m, \quad (5)$$

where $\sum_{m=1}^{\infty} C_m$ is a non-converging series of coefficients. $C_1 = (1/2N\ln 2)(|X|-1)(|Y|-1)$ approximates $\widehat{\mathrm{MI}} - \mathrm{MI}$ [32] when the sample $N \gg |X|\cdot|Y|$. Thus, to separate

between independent variables (MI=0) and dependent variables (MI>0), we suggest an adaptive threshold,

$$\varepsilon \equiv C_1 = (1/2N \ln 2)(|X|-1)(|Y|-1)\,. \quad (6)$$

S_1 threshold Since G_{st} and $\widehat{\text{MI}}$ are related [7]

$$G_{st} = 2 \cdot N \cdot \widehat{\text{MI}}(X;Y)\,, \quad (7)$$

we suggest a second adaptive threshold, which is (see (4))

$$\varepsilon \equiv S_1 = (1/2N) \cdot X^2_{\text{d.f}=(|X|-1)(|Y|-1),\alpha}\,. \quad (8)$$

S_2 threshold Since (7) is true when $\widehat{\text{MI}}$ is calculated using the natural logarithm, whereas MI in (1) is calculated using base 2, we propose S_2 by changing bases for S_1, i.e.,

$$\varepsilon \equiv S_2 = S_1 \log_2 e = (\log_2 e\,/2N) \cdot X^2_{\text{d.f}=(|X|-1)(|Y|-1),\alpha} \quad (9)$$

Between the thresholds All thresholds decrease with N, as desired. S_2 is larger than S_1 by $1/ln2$ and thus increases with cardinality faster than S_1. C_1's numerator equals the expected value of a chi-square distributed variable, which is the number of variable's df [1], whereas S_2's numerator is the chi-square critical value. Thus, the latter is higher than the former. Both S_1's numerator and denominator are larger than C_1's numerator and denominator, and S_1 is larger than C_1 if $(|X|-1)(|Y|-1)/X^2_{\text{d.f}=(|X|-1)(|Y|-1),\alpha} < 0.693$. Usually ($\alpha$=0.05), this happens for df ≤ 35, which is common.

3.2 Threshold Extension to CMI

Extending the thresholds from MI to CMI, we account for the increase in df due to the cardinalities of all variables Z_i in set $\boldsymbol{S}$. Thus, when conditionally testing two variables, the adaptive thresholds are corrected by $\prod_{\forall Z_i \in \boldsymbol{S}} |Z_i|$ [31, 33],

$$\begin{aligned} C_1 &= (1/2N \ln 2)(|X|-1)(|Y|-1) \prod\nolimits_{\forall Z_i \in \boldsymbol{S}} |Z_i| \\ S_1 &= (1/2N) \cdot X^2_{\text{d.f}=(|X|-1)(|Y|-1)\prod_{\forall Z_i \in S}|Z_i|,\alpha} \\ S_2 &= (1/2N \ln 2) \cdot X^2_{\text{d.f}=(|X|-1)(|Y|-1)\prod_{\forall Z_i \in S}|Z_i|,\alpha} \end{aligned} \quad (10)$$

3.3 Complexity

Complexities of the CMI, X^2_{st}, and G_{st} measures are $O(c^3 n)$ for n variables and maximal variable cardinality c. The complexity of a fixed threshold is $O(1)$ and that of an adaptive threshold (and $X^2_{\text{d.f},\alpha}$ for thresholding X^2_{st}) is $O(n)$; hence, adaptive thresholding does not add to complexity.

4 Empirical Evaluation

In six experiments, we compared adaptive to fixed thresholds. We used ANOVA with a 0.05 significance level to evaluate the significance of the results. ANOVA examines [3] whether the effect of a factor (e.g., sample size, variable cardinality) on the dependent variable (e.g., CMI value) is significant. The null hypothesis is that all factors perform similarly and the observed differences are merely random; thus, the effect is insignificant. ANOVA divides the total variability into that caused by each of the factors and the residual (error) variability. If the factor variability is considerably larger than the error variability, the null is rejected, and we conclude that some factor has a significant effect on the dependent variable. Then, we find which factor makes the effect using a post hoc test (Tukey-HSD) [16].

4.1 Experiment 1

To learn about the suitability of the adaptive thresholds to MI-based CI testing, we sampled datasets of 500, 1,000, and 5,000 records from the Alarm, Insurance, and Child networks [4]. For every dataset and pair of variables, we plotted $\widehat{\text{MI}}$ against thresholds' values, as computed by (6), (8), and (9). We plotted $\widehat{\text{MI}}$ for increasing products of cardinality values ($|X| \cdot |Y|$). For each product, we presented thresholds and $\widehat{\text{MI}}$ for all pairs of variables with the same product along with the average $\widehat{\text{MI}}$. Figure 1, for Alarm and 500 records, validates the analysis in Section 3.1, by which S_2 is larger than S_1, which is larger than C_1. The analysis for all networks and dataset sizes reveals that $\widehat{\text{MI}}$ increases with the decrease in the dataset size, but it is yet not clear if the $\widehat{\text{MI}}$ increase with cardinality is not due to high dependence between tested variables with high cardinality.

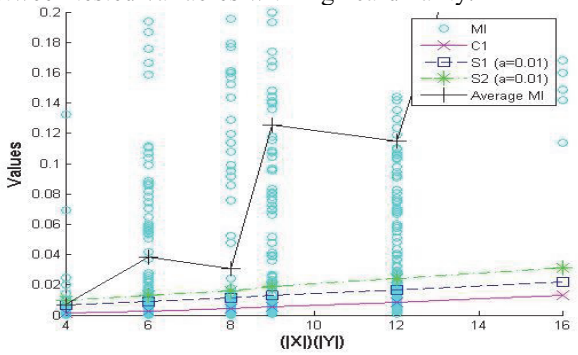
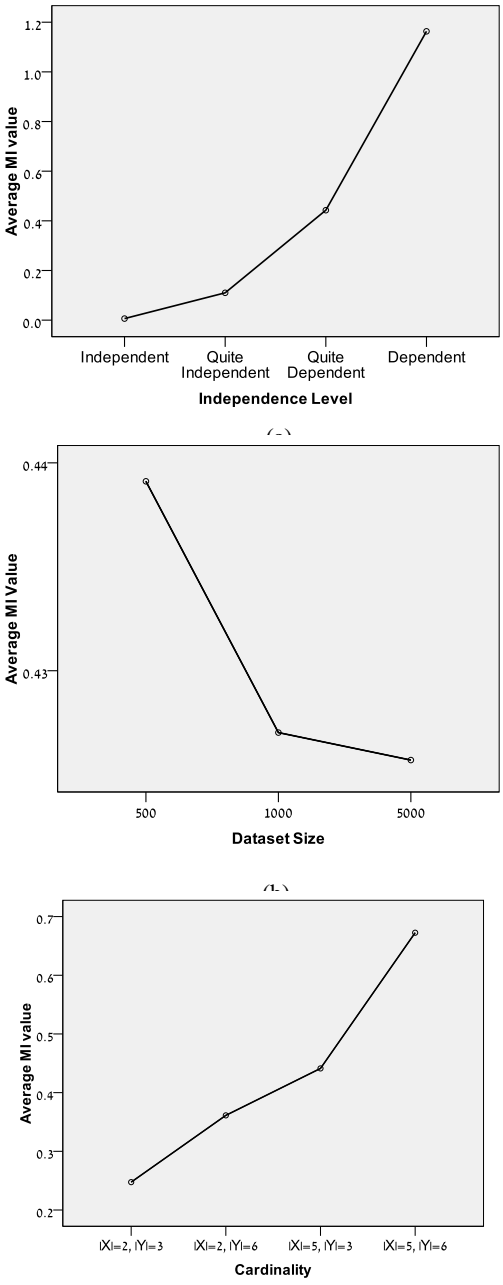


Figure 1. MI and thresholds' values (Alarm, 500 records).

4.2 Experiment 2

To understand the joint effect on $\widehat{\text{MI}}$ of cardinality, sample size, and the independence level between the variables, we generated 48 synthetic datasets. They describe, for two variables connected by a directed edge, all combinations of: 1) $|X_1|$=2, 5; 2) $|X_2|$=3, 6; 3) 500, 1,000, and 5,000 records; and 4) four variable independence levels: independent, "quite" independent, "quite" dependent, and dependent.

To create *independent* variables, each variable was generated based on a different sequence of random numbers drawn from U(0,1). For example, if $|X_1|$=2, and the first sampled random number from the first sequence was lower than 0.5, then X_1's value in the first record was 1, otherwise it was 2. If $|X_2|$=3, and the first sampled random number from the second sequence was lower than 0.33, then X_2's value in the first record was 1, otherwise, if it was lower than 0.66, then X_2's value in the first record was 2, otherwise 3. In practice, N random numbers were sampled to represent each variable; hence, the variables are fully independent [24]. To create "*quite" independent* variables, 66.6% of the variables' values were generated based on different sequences of random numbers (as for independent variables) and 33.3% of the values based on the same sequence of random numbers for both X_1 and X_2. To create *"quite" dependent* variables, 33.3% of the variables' values were generated based on different sequences and 66.6% of the variables' values based on the same sequence for both X_1 and X_2. To create *dependent* variables, the variables were sampled based on the same sequence of random numbers, i.e., each random number in the sequence simultaneously determines a specific value for both X_1 and X_2.



(c)

Figure 2. Average $\widehat{MI}$ values as a function of (a) independence level, (b) dataset size, and (c) cardinality.

Each of the 48 datasets was randomly replicated 30 times. For every dataset replication, $\widehat{MI}$ values between the two variables were calculated. The effects of the factors on $\widehat{MI}$ were analyzed using a three-way ANOVA test (the three examined factors were independence level, dataset size, and cardinality). It was followed by a Tukey-HSD post hoc analysis [16], when significance was found by ANOVA [3].

Independence level As expected, $\widehat{MI}$ increases with the degree of dependence (Figure 2(a)) and this degree was found significant to $\widehat{MI}$ in statistical analysis (p-value=0.00).

Dataset Size Figure 2(b) shows that $\widehat{MI}$ decreases with the dataset size and the size was also found significant.

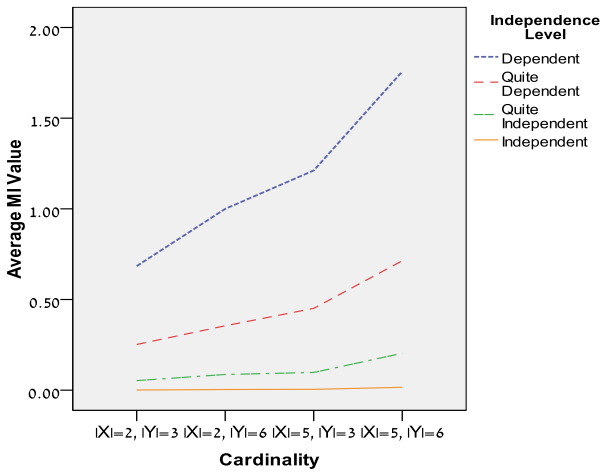


Figure 3. Average $\widehat{MI}$ values as a function of cardinality and independence level.

Cardinality $\widehat{MI}$ increases with cardinality (Figure 2(c)). This factor was also found significant statistically.

Figure 3 shows results for the interaction between cardinality and independence level, which is the most interesting one, because we expect that when the variables are dependent, the effect of the cardinality will be stronger than when the variables are independent. Figure 3 shows that $\widehat{MI}$ increases faster with cardinality when the variables are more dependent. This interaction was found significant (p-value=0.00), as those between cardinality and dataset size and independence level and dataset size. Overall, the results of this experiment encourage the use of thresholds that are adaptive to the sample size and cardinality.

4.3 Experiment 3

We evaluated the adaptive thresholds in distinguishing pairs of independent variables from pairs of dependent variables using the Alarm, Insurance, Child, Mildew, Asia, and Hail-Finder networks [4]. Samples for each network included 500, 1,000, and 5,000 records. For every pair of variables, we calculated $\widehat{MI}$ and each of the adaptive thresholds. If the $\widehat{MI}$ value was higher than a tested threshold, we decided that, based on data, the variables are dependent; otherwise, we decided they are independent. By examining d-separation between all variable pairs (for conditioning sets of 0 order) using the true network, we could test these decisions for all variable pairs and compute the accuracy of the decisions for each adaptive threshold. We repeated this procedure with two common, fixed thresholds, 10^{-2} and 10^{-3} [2, 10, 33, 35], and computed an average accuracy (weighted by the number of pairs tested for each network) for each threshold. The results show that S_1 (α = 0.01, 0.05) and S_2 (α = 0.01) are the best thresholds and the fixed thresholds (and C1) are the worst. ANOVA showed that threshold type significantly affects the percentage of correct decisions (p-value=0.00). Table 1 presents a Tukey-HSD post hoc analysis by which S_1 (α = 0.01, 0.05) and S_2 (α = 0.01) make a subgroup of thresholds that distinguish between pairs of independent variables and pairs of dependent variables most accurately, and 10^{-3} is the least accurate threshold.

Threshold	T	Subset				
		1	2	3	4	5
10^{-3}	10110	.5934				
C1	10110		.6436			
10^{-2}	10110		.6624	.6624		
S1(0.1)	10110			.6693	.6693	
S2 (0.01)	10110			.6816	.6816	.6816
S1(0.05)	10110				.6843	.6843
S1(0.01)	10110					.6920

Table 1. Homogenous subsets of the accuracy of the thresholds in distinguishing pairs of independent from pairs of dependent variables. The number of observations (T) is the number of tested pairs in 6 networks (e.g., 666 for Alarm) and 3 dataset sizes.

4.4 Experiment 4

We checked the effects on $\widehat{\text{CMI}}$ of the number of variables in the conditioning set and their cardinalities.

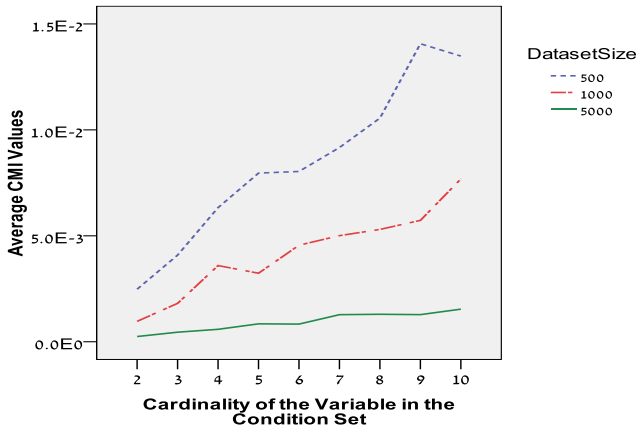
In *Experiment A*, we generated a serial connection, $X_1 \rightarrow X_2 \rightarrow X_3$ with cardinalities $|X_1|=|X_3|=2$, and $|X_2|=2,\ldots,10$, yielding nine BN configurations. For each configuration, we randomly generated the BN parameters from a Dirichlet distribution (α = 0.9). Datasets that contained 500, 1,000, and 5,000 records were sampled from each BN. For each of 30 data replications and each combination of a configuration and dataset size, we calculated $\widehat{\text{CMI}}(X_1, X_3|X_2)$.

In *Experiment B*, we generated a serial connection that contains k nodes, k = 3,...,6, $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_{k-1} \rightarrow X_k$. All cardinalities were fixed, $|X_i|= 2\ \forall i = 1 \ldots k$. Parameters were randomly generated from a Dirichlet distribution (α = 0.9). Datasets that contained 500, 1,000, and 5,000 records were sampled for each BN. We computed $\widehat{\text{CMI}}(X_1, X_k|X_2, \ldots, X_{k-1})$ for each of 30 data replications and each combination of the conditioning set size and dataset size.

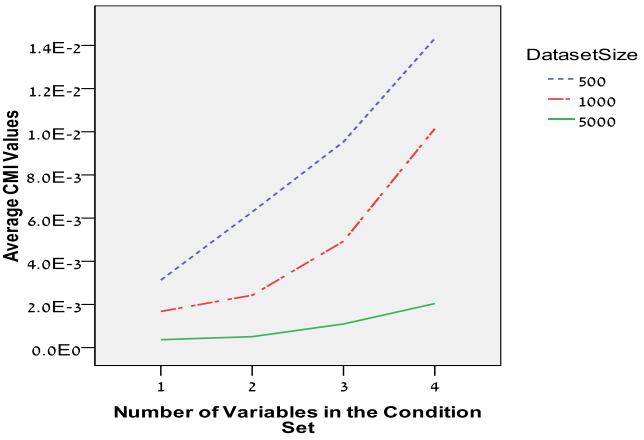
The statistical analysis for *Experiment A* shows that the cardinality of the variable in the conditioning set (X_2) has a significant effect on $\widehat{\text{CMI}}$ (p-value=0.00). Figure 4(a) demonstrates that $\widehat{\text{CMI}}$ increases with X_2's cardinality and this increase goes faster for smaller dataset sizes. The analysis for *Experiment B* shows that the cardinality of the conditioning set has a significant effect on $\widehat{\text{CMI}}$ (p-value=0.00). Figure 4(b) shows that $\widehat{\text{CMI}}$ increases with cardinality and this increase goes faster for smaller dataset sizes. These results reinforce threshold augmentation based on the cardinalities of the conditioning set (as in (10)).

4.5 Experiment 5

We evaluated structure learning using adaptive thresholds. Eight networks were examined: Alarm, Insurance, Child, Mildew, Hail-Finder, Asia, BN_10 (random modification of Alarm), and BN_126 (a coding graph) (the last 2 from the UAI 2008 competition). For each network, we sampled 500, 1,000, and 5,000 records. We used ten random replications for each sample size. For every dataset, the BN structure was learned (up to a CI-test order of 6) using the PC algorithm [31]. Learning was performed with adaptive thresholds: C_1, $S_1(0.01)$, $S_1(0.05)$, $S_1(0.1)$, and $S_2(0.01)$ and



(a)



(b)

Figure 4. Effect on $\widehat{\text{CMI}}$ in (a) *Experiment A* and (b) *Experiment B*.

with fixed thresholds: 10^{-2} and $5 \cdot 10^{-3}$ [2, 10, 35]. To evaluate learning, the learned CPDAGs were compared with the true networks (after being transferred to CPDAGs) to derive the structural hamming distance (SHD) [33], which is structural error. In addition, we calculated the number of CI tests performed and run-time, as measures of complexity.

SHD Figure 5 shows that the advantage of adaptive over fixed thresholds decreases with the dataset size. This is because $\widehat{\text{CMI}}$ reduces as a power of the size [32] and the fixed thresholds are smaller than the adaptive. Yet, even for 5,000 records, two adaptive thresholds yield the best SHD and another one is comparable to the fixed thresholds. Figure 6 shows this advantage in learning curves for Alarm.

Complexity Figures 7 and 8 show that adaptive thresholds reduce the run-time and number of high-order tests (order >2), respectively (averaged over all datasets and networks).

4.6 Experiment 6

We learned BNs for classification using 21 UCI databases [19]. We discretized continuous variables and omitted records with missing values. We performed parameter learning using a Dirichlet prior with zero hyper-parameters [21]. We computed the classification accuracy and BDeu score [8, 22] over a cross-validation experiment.

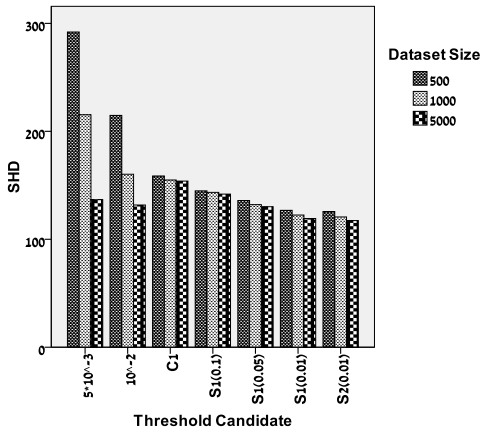


Figure 5. SHD averaged over all datasets of all eight networks.

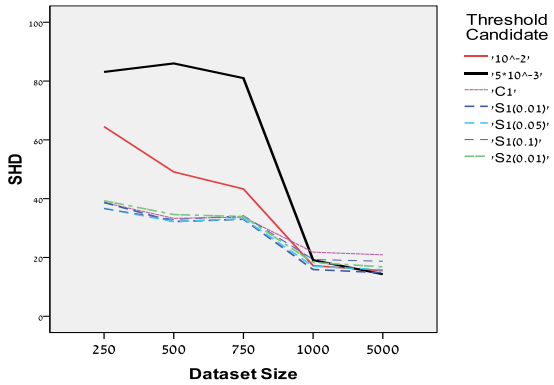


Figure 6. Learning curves for the Alarm network.

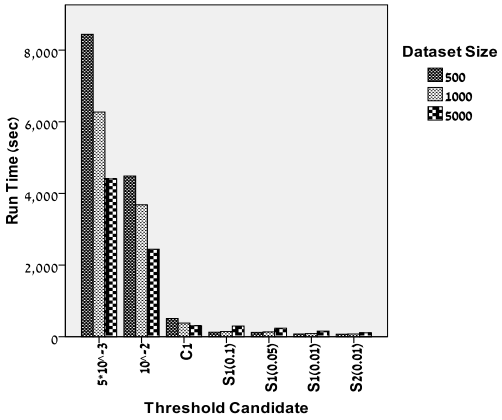


Figure 7. Run-times averaged over datasets and networks.

Figure 9 shows advantage in accuracy (averaged over all databases) for adaptive thresholds. They were significantly more accurate than fixed thresholds. Figure 10 shows BDeu is higher for adaptive thresholds than for fixed thresholds.

5 Conclusion and Future Work

Arbitrariness in estimating a CI measure and selecting a threshold undermines and extends learning a BN structure.

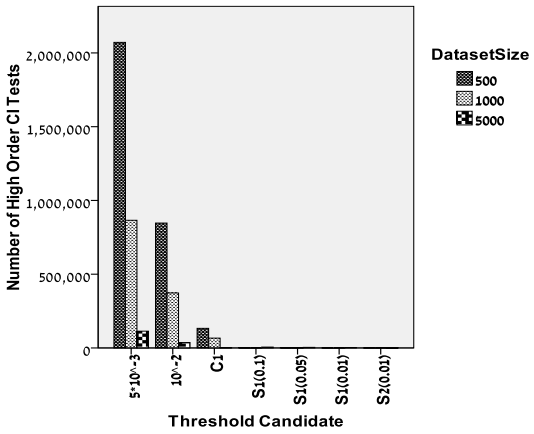


Figure 8. Numbers of high-order tests for different thresholds.

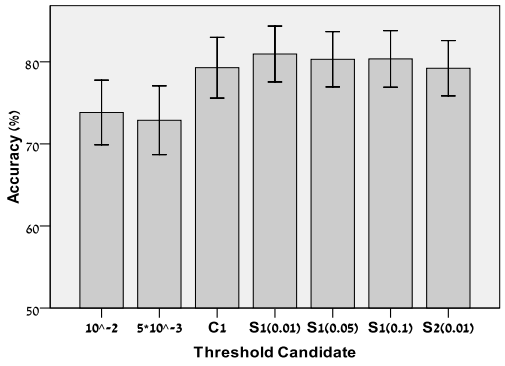


Figure 9. Classification accuracies averaged over 21 databases.

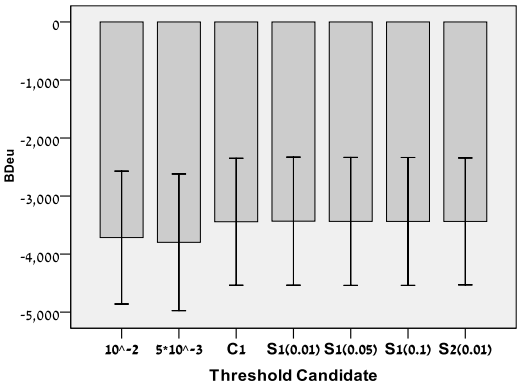


Figure 10. BDeu values averaged over 21 databases.

For MI, we explored arbitrariness with respect to factors, such as the degree of dependence, sample size, and cardinalities of involved variables. We suggested three adaptive thresholds and showed their advantages over fixed thresholds in 1) separating pairs of dependent variables from pairs of independent variables; 2) learning an accurate BN structure; 3) learning using smaller sample sizes; 4) learning faster and with smaller complexity; and 5) learning structures that fit and classify the data more accurately. Our current efforts focus on extending the study for PC to other algorithms, expanding the evaluation to non-fixed thresholds and to statistical measures, and studying additional methods of adaptive thresholding.

References

[1] Agresti, A. (2002). *Categorical Data Analysis.* Wiley Series in Probability and Statistics.

[2] Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010a). Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, 11, 171–234.

[3] Armstrong, R. A., Slade, S. V., & Eperjesi, F. (2002). An Introduction to Analysis of Variance (ANOVA) with Special Reference to Data from Clinical Experiments in Optometry. *Ophthalmic and Physiological Optics*, 20, 235-241.

[4] Bayesian Network Repository, http://www.cs.huji.ac.il/site/labs/compbio/Repository/

[5] Belyavin, H., & Lerner, B. (2011). Machine Learning in Predicting and Explaining Failure Using Class-Imbalance FAB Data. In *The 21st International Conference on Production Research.* Stuttgart.

[6] Besson, P. (2010). Bayesian Networks and Information Theory for Audio-Visual Perception Modeling. *Biological Cybernetics ,* 103, 213-226.

[7] Brilinger, D. (2005). Some Data Analyses Using Mutual Information. *Brazilian Journal of Probability and Statistics ,* 18, 163-182.

[8] Buntine, W. (1991). Theory Refinement of Bayesian Networks. *7th Conf. on UAI*, pp. 52-60.

[9] Cheng, J., & Greiner, R. (1999). Comparing Bayesian Network Classifiers. *15th Conf. on UAI*, pp. 101-107.

[10] Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian Network from Data: An Information-Theory Based Approach. *Artificial Intelligence ,* 137, 43-90.

[11] Chickering, D. M. (2002). Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3, 507-554.

[12] Cooper, G. F., & Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Network from Data. *Machine Learning ,* 9, 309-347.

[13] Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory* (2nd ed.). Wiley Series in Telecommunications and Signal Processing.

[14] Dash, D., & Druzdzel, R. (1999). A Hybrid Anytime Algorithm for the Construction of Causal Models From Sparse Data. *15th Conf. on UAI*, pp. 142-149.

[15] de Campos, L. M. (2006). A Scoring Function for Learning Bayesian Networks Based on Mutual Information and Conditional Independence Tests. *Journal of Machine Learning Research,* 7, 2149-2187.

[16] Demŝar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research ,* 7, 1-30.

[17] Dor, D. & Tarsi, M. (1992). A Simple Algorithm to Construct a Consistent Extension of a Partially Oriented Graph. TR185, Computer Science Department, UCLA.

[18] Filho, J., & Wainer, J. (2008). HPB: A Model for Handling BN Nodes with High Cardinality Parents. *Journal of Machine Learning Research,* 9, 2141-2170.

[19] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository, University of California, Irvine, http://archive.ics.uci.edu/ml.

[20] Fung, R. M. & Crawford, S. L. (1990). Constructor: A System for the Induction of Probabilistic Models. *8th National Conf. on AI,* pp. 762-769.

[21] Heckerman, D. (1995). A Tutorial on Learning with Bayesian Networks. TR-95-06. MS Research.

[22] Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20, 197-243.

[23] Kalisch, M. & Buhlman, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC Algorithm. *Journal of Machine Learning Research*, 8, 613-636.

[24] Law, A., & Kelton, D. (2000). *Simulation Modling and Analysis.* McGraw Hill.

[25] Lehmann, E. L. and Romano, J. P. (2010). *Testing Statistical Hypothesis*. Springer, New York, NY.

[26] Meek, C. (1995). Causal Inference and Causal Explanation with Background Knowledge. *11th Conf. on UAI*, pp. 403-410.

[27] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (2nd edition). Morgan-Kaufmann, San-Francisco.

[28] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge.

[29] Ramsey, J., Spirtes, P., & Zhang, J. (2006). Adjacency-Faithfulness and Conservative Causal Inference. *22nd Conf. on UAI*, pp. 401-408.

[30] Rebane, G., & Pearl, J. (1987). The Recovery of Causal Poly-trees from Statistical Data. *Workshop on UAI*, pp. 222–228.

[31] Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction and Search* (2nd ed.). MIT Press.

[32] Treves, A., & Panzeri, S. (1995). The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Computation ,* 7, 399-407.

[33] Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Journal of Machine Learning Research ,* 65, 31–78.

[34] Vinh, N. X., Epps, J., & Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison:Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research ,* 11, 2837-2854.

[35] Yehezkel, R., & Lerner, B. (2009). Bayesian Network Structure Learning by Recursive Autonomy Identification. *Journal of Machine Learning Research ,* 10, 1527-1570.