In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI)*, Pittsburgh, PA, July 2005

Distribution-Free Learning of Bayesian Network Structure in Continuous Domains

Dimitris Margaritis
Dept. of Computer Science
Iowa State University
Ames, IA 50011
dmarg@cs.iastate.edu

Abstract

In this paper we present a method for learning the structure of Bayesian networks (BNs) without making any assumptions on the probability distribution of the domain. This is mainly useful for continuous domains, where there is little guidance and many choices for the parametric distribution families to be used for the local conditional probabilities of the Bayesian network, and only a few have been examined analytically. We therefore focus on BN structure learning in continuous domains. We address the problem by developing a conditional independence test for continuous variables, which can be readily used by any existing independence-based BN structure learning algorithm. Our test is non-parametric, making no assumptions on the distribution of the domain. We also provide an effective and computationally efficient method for calculating it from data. We demonstrate the learning of the structure of graphical models in continuous domains from real-world data, to our knowledge for the first time using independence-based methods and without distributional assumptions. We also experimentally show that our test compares favorably with existing statistical approaches which use prediscretization, and verify desirable properties such as statistical consistency.

Introduction and Motivation

One of the first problems that a researcher who is interested in learning a graphical model from data is faced with, is making a choice on the kind of probability distributions she will use. Such distributions are used to model local interactions among subsets of variables in the model. For example, in a Bayesian network (**BN**), a local probability distribution function (**PDF**) needs to be defined between every variable and its parents. The choice is easier in discrete domains, where every variable can take only a fixed number of values; the standard choice for discrete PDFs is the multinomial, which is usually sufficient for modeling complex interactions among variables, and whose parameters are straightforward to estimate from data.

In continuous or mixed continuous-discrete domains however the problem is considerably harder, prompting the use of simplifying assumptions. The common assumption is for the local PDFs between parents and children to be linear relations with additive Gaussian errors (Geiger & Heckerman 1994; Spirtes, Glymour, & Scheines 1993). However, there are many real-life situations where this assumption fails (*e.g.*, stock market prices, biometric variables, weather status, *etc.*), where the interactions are far from linear. In these cases, such an assumption can lead to inaccurate networks that are a poor fit to the data and the underlying probability distribution, and can produce incorrect structures.

Our work addresses the learning of the structure of graphical models without making any such assumptions about the distributions of the local PDFs. Although we focus on Bayesian network structure learning, application to Markov network structure learning is straightforward. To learn the structure of a Bayesian network, there exist two general classes of algorithms. The first, called the *score-based* approach, employs a search in the space of all possible legal structures guided by a heuristic function, usually penalized log-likelihood (Lam & Bacchus 1994). The search procedure maximizes this score, usually by hill-climbing. Other search techniques have also been used (Heckerman 1995).

The second class of Bayesian network structure learning algorithms use the fact that the structure of a BN implies that a set of conditional independence statements hold in the domain it is modeling. They exploit this property by conducting a number of statistical conditional independence (**CI**) tests on the data and use their results to make inferences about the structure. Assuming no errors in these tests, the idea is to constrain, if possible, the set of possible structures that satisfy the conditional independencies that are found in the data to a singleton, and infer that structure as the only possible one. For this reason these algorithms are called *constraint-based* or *independence-based*.

Our work uses the second class of algorithms to learn BN structure in continuous domains. The crucial observation is that *the independence-based algorithms interact with the data only through the conditional independence tests done during their operation*. Therefore, the use of a CI test between continuous variables, and in particular one that does not assume anything about the distribution of the variables involved in the test, would be sufficient for learning the structure in such domains in a distribution-independent fashion. CI tests that do not assume any particular family of distributions are called *non-parametric*. Although such tests exist for discrete variables—the χ^2 (chi-square) test of inde-

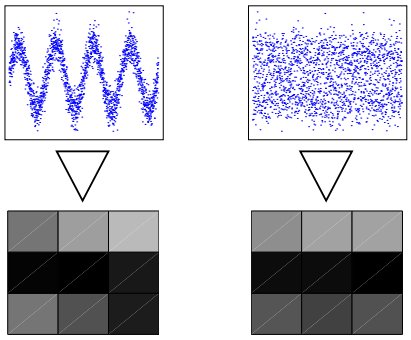

Figure 1: Two very different data sets have similar 3×3 histograms. **Left:** X, Y strongly dependent. **Right:** X, Y independent by construction.

pendence is perhaps the most common one—for continuous variables the standard approach is to discretize the continuous variables and perform a discrete test. One problem with this method is that the discretization has to be done with care; for example, Fig. 1 depicts two very different situations where X and Y are dependent (left) and independent (right), that produce two very similar histograms after a 3×3 discretization. The (unconditional) multi-resolution test of Margaritis & Thrun (2001), outlined in the next section, addresses cases such as this. Independence-based algorithms however require *conditional* independence tests. In this paper we extend the above-mentioned unconditional test to a conditional version that carefully discretizes the Z axis, performs an (unconditional) independence test on the data in each resulting bin, and combines the results into a single probabilistic measure of conditional independence.

Outline of Unconditional Test

In this section we outline the unconditional multi-resolution test of independence between two continuous variables X and Y of Margaritis & Thrun (2001). A central idea is the comparison of two competing statistical models, $M_{\mathcal{I}}$ (the *independent* model) and $M_{\neg\mathcal{I}}$ (the *dependent* model), according to the *data likelihood* of a data set consisting of (X, Y) pairs. For a given fixed resolution, the test uses a discretized version of the data set at that resolution (resolution is the size of the histogram or "grid" placed over the data *e.g.*, 3×3 in Fig. 1). The dependent model $M_{\neg\mathcal{I}}$ corresponds to a joint multinomial distribution while the independent model $M_{\mathcal{I}}$ to two marginal multinomials along the X- and Y-axes. Margaritis & Thrun calculate the data likelihoods of each model analytically:

$$\Pr(\mathbf{D} \mid M_{\neg\mathcal{I}}) = \frac{\Gamma(\gamma)}{\Gamma(\gamma + N)} \prod_{i=1}^{N} \prod_{j=1}^{M} \frac{\Gamma(\gamma_{ij} + c_{ij})}{\Gamma(\gamma_{ij})}$$

and

$$\Pr(\mathbf{D} \mid M_{\mathcal{I}}) = \left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{i=1}^{N} \frac{\Gamma(\alpha_i + c_{i+})}{\Gamma(\alpha_i)} \right) \times \left(\frac{\Gamma(\beta)}{\Gamma(\beta + N)} \prod_{j=1}^{M} \frac{\Gamma(\beta_j + c_{+j})}{\Gamma(\beta_j)} \right),$$

where $\mathbf{D}$ denotes the data set and c_{ij} the counts of the $N \times M$ histogram. This closed form is due to the use of a Dirichlet conjugate prior distribution over the parameters of the multinomials (which are typically unknown). The terms γ_{ij}, α_i and β_j are the parameters of the Dirichlet (hyperparameters). Using Bayes' theorem, the posterior probability of $M_{\mathcal{I}}$ is given by

$$\Pr(M_{\mathcal{I}} \mid \mathbf{D}) = 1 \Big/ \left(1 + \frac{1 - \mathcal{P}}{\mathcal{P}} \cdot \frac{\Pr(\mathbf{D} \mid M_{\neg\mathcal{I}})}{\Pr(\mathbf{D} \mid M_{\mathcal{I}})} \right)$$

and $\Pr(M_{\neg\mathcal{I}} \mid \mathbf{D}) = 1 - \Pr(M_{\mathcal{I}} \mid \mathbf{D})$. $\mathcal{P} \stackrel{\text{def}}{=} \Pr(\mathcal{M}_{\mathcal{I}})$ is the prior probability of the independent model. A common choice for $\mathcal{P}$ in situations where there is no prior information is guided by the principle of indifference, which assigns equal prior probability to either case *i.e.*, $\mathcal{P} = 0.5$.

The above equations hold for a given resolution and a given placement of the histogram boundaries. Margaritis & Thrun average (integrate) in a Bayesian fashion over all possible positions that the histogram boundaries may take. Since this is prohibitively expensive for anything but very small data sets, the practical implementation employs a maximum-value approximation to the Bayesian integral. They also use this fixed-resolution test to develop a multi-resolution test by discretizing the data set at a number of gradually increasing resolutions and combining the results—see Margaritis & Thrun (2001) for details.

To conclude independence or dependence, the ratio $\Pr(M_{\mathcal{I}} \mid \mathbf{D}) / \Pr(M_{\mathcal{I}})$ is compared with unity:

$$\frac{\Pr(M_{\mathcal{I}} \mid \mathbf{D})}{\Pr(M_{\mathcal{I}})} = \begin{cases} \geq 1 \Rightarrow X, Y \text{ independent} \\ < 1 \Rightarrow X, Y \text{ dependent.} \end{cases}$$

The comparison of the ratio to a number greater (smaller) than unity can be used to reflect a bias of a researcher toward dependence (independence).

An extension to testing for independence between two *sets* of variables $\mathbf{X}$ and $\mathbf{Y}$, rather than single variables X and Y, is straightforward.

Probabilistic Measure of Conditional Independence

The unconditional test described in the previous section takes a data set of continuous-valued (X, Y) pairs and returns the estimated posterior probability of the model representing independence vs. the one representing dependence. In this paper we extend this to a conditional version that tests for conditional independence of X and Y given Z, in a domain that contains continuous-valued triples (X, Y, Z). This is denoted as $X \perp\!\!\!\perp Y \mid Z$ and defined at "all values of Z except a zero-support subset" (Lauritzen 1996).

We make the standard assumption that the data points are drawn independently and from the same distribution, namely the domain PDF. This is frequently called the *i.i.d.* assumption, standing for independently and identically drawn data, and is a basic assumption in statistics—for example the analytical forms of traditional distributions (multinomial, Poisson, hypergeometric etc.) depend on it.

Our procedure for testing for conditional independence of X and Y given Z can be summarized in the following three steps:

1. Subdivide the Z axis into m bins $b_1, b_2, \ldots, b_m$, resulting in a partition of the data set $\mathbf{D}$ of size N into $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_m$ disjoint data sets.
2. Measure conditional independence in each bin by performing an unconditional independence test of X and Y, using the test described in the previous section (obtaining a posterior probability of independence from each bin).
3. Combine the results from each bin into a single number.

This immediately introduces the following three questions: (a) What discretization of the Z axis is appropriate (step 1)? (b) When and why is conducting an *unconditional* test in a bin sufficient to measure *conditional* independence in that bin (step 2)? (c) How do we combine the (unconditional) test results using the result from each bin (step 3)?

Testing for Conditional Independence In Each Bin

For clarity of exposition, we start with question (b). Let us assume we are given a subset of the data $\mathbf{D}_i$ that fall within a particular bin b_i, produced by some discretization of the Z-axis. To determine whether $(X \perp\!\!\!\perp Y \mid Z)$ in this bin, it is sufficient to test $(X \perp\!\!\!\perp Y)$ if we know that $(\{X, Y\} \perp\!\!\!\perp Z)$ (in this bin), as the following theorem states.

Theorem 1. *If* $(\{X, Y\} \perp\!\!\!\perp Z)$*, then* $(X \perp\!\!\!\perp Y \mid Z)$ *if and only if* $(X \perp\!\!\!\perp Y)$.

Proof. We are using the following theorems that hold for any probability distribution, due to **?** (**?**): Decomposition, Contraction, and Weak Union.

(Proof of $\Rightarrow$): We want to prove that
$(\{X, Y\} \perp\!\!\!\perp Z)$ and $(X \perp\!\!\!\perp Y \mid Z) \Rightarrow (X \perp\!\!\!\perp Y)$.

- From Decomposition and $(\{X, Y\} \perp\!\!\!\perp Z)$ we get $(X \perp\!\!\!\perp Z)$.
- From Contraction and $(X \perp\!\!\!\perp Z)$ and $(X \perp\!\!\!\perp Y \mid Z)$ we get $(X \perp\!\!\!\perp \{Y, Z\})$.
- From Decomposition and $(X \perp\!\!\!\perp \{Y, Z\})$ we get $(X \perp\!\!\!\perp Y)$.

(Proof of $\Leftarrow$): We want to prove that
$(\{X, Y\} \perp\!\!\!\perp Z)$ and $(X \perp\!\!\!\perp Y) \Rightarrow (X \perp\!\!\!\perp Y \mid Z)$.

- From Weak Union and $(\{X, Y\} \perp\!\!\!\perp Z)$ we get $(X \perp\!\!\!\perp Z \mid Y)$.
- From Contraction and $(X \perp\!\!\!\perp Z \mid Y)$ and $(X \perp\!\!\!\perp Y)$ we get $(X \perp\!\!\!\perp \{Z, Y\})$.
- From Weak Union and $(X \perp\!\!\!\perp \{Z, Y\})$ we get $(X \perp\!\!\!\perp Y \mid Z)$.

We can use the above theorem as follows. Suppose that we conduct an unconditional test $(\{X, Y\} \perp\!\!\!\perp Z)$ in bin b_i, using the unconditional test of Margaritis & Thrun (2001) (extended to handle sets), obtaining $\Pr(\{X, Y\} \perp\!\!\!\perp Z \mid \mathbf{D}_i)$. Denoting $(\{X, Y\} \perp\!\!\!\perp Z) \stackrel{\text{def}}{=} U$, $(X \perp\!\!\!\perp Y) \stackrel{\text{def}}{=} I$, and $(X \perp\!\!\!\perp Y \mid Z) \stackrel{\text{def}}{=} CI$, consider two extreme cases:

- Given *certain independence* of $\{X, Y\}$ and Z, *i.e.*, $\Pr(U \mid \mathbf{D}_i) = 1$, conditional independence of X and Y given Z is the same as unconditional, according to the theorem. Therefore, $\Pr(CI \mid U, \mathbf{D}_i) = \Pr(I \mid \mathbf{D}_i)$.
- Given *certain dependence* of $\{X, Y\}$ and Z, *i.e.*, $\Pr(U \mid \mathbf{D}_i) = 0$, nothing can be said about the conditional independence of X and Y given Z without actually conducting a conditional test; this is because the distribution of $\{X, Y\}$ is certain to change with Z within the bin, making the theorem inapplicable. Therefore, in this case, the posterior is taken equal to the prior probability: $\Pr(CI \mid \neg U, \mathbf{D}_i) = \mathcal{P} = 0.5$.

In the general case, by the theorem of total probability, $\Pr(CI \mid \mathbf{D}_i)$ is:

$$\begin{aligned} \Pr(CI \mid \mathbf{D}_i) = {} & \Pr(CI \mid U, \mathbf{D}_i) \Pr(U \mid \mathbf{D}_i) + \\ & \Pr(CI \mid \neg U, \mathbf{D}_i) \Pr(\neg U \mid \mathbf{D}_i) = \\ & \Pr(I \mid \mathbf{D}_i) \Pr(U \mid \mathbf{D}_i) + \mathcal{P}(1 - \Pr(U \mid \mathbf{D}_i)) \end{aligned} \tag{1}$$

The test of Margaritis & Thrun (2001) can be used for both $\Pr(U \mid \mathbf{D}_i)$ and $\Pr(I \mid \mathbf{D}_i)$ since they are unconditional. We therefore now have a way of estimating the posterior probability of conditional independence from the results of two unconditional independence tests, in each bin of a given discretization of the Z-axis.

Discretization and Probabilistic Measure

We now address questions (c) and (a). Even though we can use Eq. (1) directly to calculate the posterior probability of conditional independence on the entire data set $\mathbf{D}$ of size N, this will be useless in most cases, because the distribution of $\{X, Y\}$ will likely vary along the Z-axis, resulting in $\Pr(U \mid \mathbf{D}) \approx 0$ and therefore $\Pr(CI \mid \mathbf{D}) \approx \mathcal{P} = 0.5$ according to Eq. (1). To improve the chance that the PDF of $\{X, Y\}$ will be uniform along Z, we divide the Z axis into subregions, resulting into a partition of $\mathbf{D}$ into subsets.

Let us assume that we are given a discretization resulting in m subsets $\mathbf{D}_1, \ldots, \mathbf{D}_m$, and we have calculated the independence of X with Y in each bin b_i, *i.e.*, $\Pr(CI \mid \mathbf{D}_i) \stackrel{\text{def}}{=} \mathcal{I}_i$ and the independence of $\{X, Y\}$ with Z in b_i, *i.e.*, $\Pr(U \mid \mathbf{D}_i) \stackrel{\text{def}}{=} \mathcal{U}_i$. These estimates are independent for different values i and j due to the i.i.d. assumption, which implies that the (X, Y, Z) values of the i-th data point do not depend on the values any of the remaining $(N - 1)$ data points. Therefore, $\mathcal{I}_i$ is independent of $\mathcal{I}_j$ (and $\mathcal{U}_i$ of $\mathcal{U}_j$) for $i \neq j$, because they are functions of the independent data sets $\mathbf{D}_i$ and $\mathbf{D}_j$.

We therefore measure conditional independence by the product of the corresponding estimates $\mathcal{I}_i$ over all bins. Due to the practical consideration that the fewer data points an interval contains, the more unreliable the test of independence becomes, we scale each term in the product by the fraction of data points in the corresponding bin, resulting in the following expression:

$$\boxed{\mathcal{I} = \prod_{i=1}^{m} \Pr(CI \mid \mathbf{D}_i)^{c_i/N} = \prod_{i=1}^{m} \mathcal{I}_i^{c_i/N}.} \tag{2}$$

where c_i is the number of data points in bin b_i. Each term in the product can be calculated using Eq. (1). $\mathcal{I}$ ranges from 0 to 1. We use it as a probabilistic measure of conditional independence for a given discretization. To conclude dependence or independence, we compare $\mathcal{I}$ with its value when we have no data *i.e.*,

$$\prod_{i=1}^{m} \Pr(CI)^{c_i/N} = \prod_{i=1}^{m} \mathcal{P}^{c_i/N} = \mathcal{P}.$$

This addresses question (c) (combination of results). For question (a), the determining factor on the appropriateness

$(\mathcal{I}, \mathcal{U}) =$ RECURSIVE-MEDIAN$(X, Y, Z, \mathbf{D})$:

1. **if** $|\mathbf{D}| \leq 1$,
2. **return** $(0.5, 0.5)$
3. $\mathcal{U} = \Pr(\{X, Y\} \perp\!\!\!\perp Z \mid \mathbf{D})$
4. $\mathcal{I} = \Pr(X \perp\!\!\!\perp Y \mid \mathbf{D}) \times \mathcal{U} + \mathcal{P} \times (1 - \mathcal{U})$
5. $z^* = median(\mathbf{D}, Z)$
6. $\mathbf{D}_1 = \{$points j of $\mathbf{D}$ such that $z_j \leq z^*\}$
7. $\mathbf{D}_2 = \{$points j of $\mathbf{D}$ such that $z_j > z^*\}$
8. $(\mathcal{I}_1, \mathcal{U}_1) =$ RECURSIVE-MEDIAN$(X, Y, Z, \mathbf{D}_1)$
9. $(\mathcal{I}_2, \mathcal{U}_2) =$ RECURSIVE-MEDIAN$(X, Y, Z, \mathbf{D}_2)$
10. $f_1 = (z^* - z_{min})/(z_{max} - z_{min})$ /* $f_1 \approx 0.5$ */
11. $f_2 = (z_{max} - z^*)/(z_{max} - z_{min})$ /* $f_2 \approx 0.5$ */
12. $\mathcal{I}' = \exp(f_1 \ln \mathcal{I}_1 + f_2 \ln \mathcal{I}_2)$
13. $\mathcal{U}' = \exp(f_1 \ln \mathcal{U}_1 + f_2 \ln \mathcal{U}_2)$
14. **if** $\mathcal{U} > \mathcal{U}'$,
15. **return** $(\mathcal{I}, \mathcal{U})$
16. **else**
17. **return** $(\mathcal{I}', \mathcal{U}')$

Figure 2: The recursive-median algorithm.

of a discretization is a measure of how uniform each bin is with respect to $\{X, Y\}$ and Z, allowing the use of Theorem 1. We measure uniformness in a fashion similar to $\mathcal{I}$, calling the result $\mathcal{U}$:

$$\mathcal{U} = \prod_{i=1}^{m} \Pr(U \mid \mathbf{D}_i)^{c_i/N} = \prod_{i=1}^{m} \mathcal{U}_i^{c_i/N}. \qquad (3)$$

As with $\mathcal{I}$, $\mathcal{U}$ also ranges from 0 to 1. The insight here is that $\mathcal{U}$ is theoretically constant and equal to 1 everywhere along the Z-axis at the limit where the length of each interval tends to 0. (This is a consequence of $\mathcal{U}$ monotonically tending to 1 within an interval as its length is made shorter. The proof of this is simple once we observe that independence of $\{X, Y\}$ with Z within an interval is logically implied by independence within two intervals that partition it.) We can therefore use $\mathcal{U}$ to compare two discretizations by their resulting $\mathcal{U}$ values, preferring the one with the greater $\mathcal{U}$ value (*i.e.*, the one closer to 1).

Recursive-median algorithm

We can now put all the above together in the recursive-median algorithm shown in Fig. 2. The algorithm takes as input the names of variables X, Y and Z and a data set $\mathbf{D}$. It begins by calculating the measure of posterior probability of independence $\mathcal{I}$ (using Eq. (1)) and $\mathcal{U}$ using a single interval along the Z-axis that contains the entire data set. It then splits the data set along the Z-axis at the median, producing in two non-overlapping intervals containing the same number of points (plus or minus one) and recursively calls itself on each of the two subsets. When only one point remains, the recursion reaches its base case; in this case 0.5 (the prior) is returned both for $\mathcal{I}$ and $\mathcal{U}$, since both the independent and the dependent model are supported by the single data point equally well. Upon return from the two recursive calls, the results $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{U}_1$, $\mathcal{U}_2$ are combined into $\mathcal{I}'$ and $\mathcal{U}'$ respectively using Eqs. (2) and (3), respectively. The greater of $\mathcal{U}$ (representing the estimated uniformness of the interval before splitting) and $\mathcal{U}'$ (representing the estimated uniformness after splitting) is returned, along with the corresponding conditional independence estimate ($\mathcal{I}$ or $\mathcal{I}'$). At the end of the run on the entire data set, the value of $\mathcal{U}$ returned can be discarded or used as a measure of confidence in the main result ($\mathcal{I}$ value), if desired. We conclude conditional independence if and only if $\mathcal{I}/\mathcal{P} \geq 1$.

Related Work

Work specific to non-parametric conditional independence testing in continuous domains is scarce. The most relevant work is in the area of discretization. Virtually all work in this area has focused in the past on PDF estimation (*e.g.*, Scott (1992)). However, it is arguable whether a method well-suited for PDF estimation can also perform well in independence testing (those tested here did not, as will be seen in the Experiments section below). Discretization methods can be categorized as supervised or unsupervised. Since here there exists no concept of "class," supervised methods are not applicable. Unfortunately, these form the bulk of the research in the area (*e.g.*, see survey in Dougherty, Kohavi, & Sahami (1995)). A small number of unsupervised methods exist:

- **Sturges's method** is widely recommended in introductory statistics texts (Scott 1992). It dictates that the the optimal number of equal-width histogram bins is $k = 1 + \log_2 N$, where N is the number of points.
- **Scott's method** (Scott 1979) dictates that the optimal bin width is $\hat{h} = 3.5\hat{\sigma}N^{-1/(2+d)}$ for multiple dimensions, where $\hat{\sigma}$ is the sample standard deviation and d is the number of dimensions.
- **Freedman and Diaconis's method** (Freedman & Diaconis 1981) is an improvement over Scott's method intended to be more robust to outliers. It sets $\hat{h} = 2(\text{IQ})N^{-1/(2+d)}$, where IQ is the interquartile range (the distance between the points splitting the data into 25%–75% and 75%–25% subsets).

We compare our approach with these methods in the next section.

Experiments

Real-World BN Structure Learning Experiments

We evaluated the suitability of the recursive-median algorithm for learning the structure of a graphical model on the BOSTON-HOUSING and ABALONE real-world data sets:

Data set	No. of data points	No. of variables
HOUSING	506	14
ABALONE	4177	8

HOUSING is a relatively small data set, making non-parametric testing for independence especially difficult. We applied the PC algorithm (Spirtes, Glymour, & Scheines 1993) using the recursive-median test for learning the structure of a graphical model in this domain. The output of PC is not always a Bayesian network; in domains where its assumptions fail, the resulting structure may contain bidirected

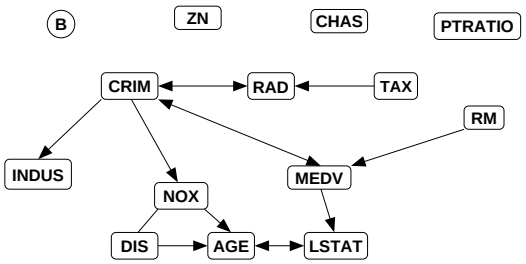

Figure 3: Output of the PC algorithm for the HOUSING domain. The model contains bidirected and undirected edges (see text).

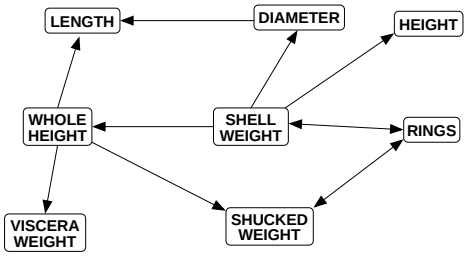

Figure 4: Output of the PC algorithm for the ABALONE domain. The model contains bidirected edges (see text).

(indicating the existence of a latent variable) as well as undirected (undirectable) edges. The resulting graphical model is shown in Fig. 3. This structure implies a set of conditional independencies; for example that INDUS (proportion of non-retail business acres per town) is independent of other variables given CRIM (per capita crime rate by town). We note that the resulting network agrees with the original study of the data that concluded that there is no direct link between NOX (a pollution indicator) and MEDV (the median value of owner-occupied homes)—their dependence in Fig. 3 is through a number of intermediate relationships.

The ABALONE data set contains measurements from a study conducted to predict the age of abalones from their physical characteristics. The domain contains 9 variables in total, one of which is categorical (SEX), 7 continuous, and one ordinal discrete (the number of rings, which denotes the age). Since the test for continuous variables assumes an ordering of the values, we did not use the categorical variable. The network structure resulting for the ABALONE domain is shown in Fig. 4. The structure indicates a number of conditional independencies, including that HEIGHT is independent of other variables given the SHELL WEIGHT, and LENGTH is independent of other variables given WHOLE WEIGHT and DIAMETER.

Comparison with Prediscretization Methods

We compared the recursive-median algorithm with methods that prediscretized the data using Sturges's, Scott's, and Freedman and Diaconis's method, described in the Related Work section. After discretization, we conducted a χ^2 (chi-square) test of independence for each bin of the conditioning variable—combining over all Z intervals in the standard way—to produce a significance (a p-value) in the null hypothesis (conditional independence). To evaluate the performance of these tests, artificial data sets were necessary. This

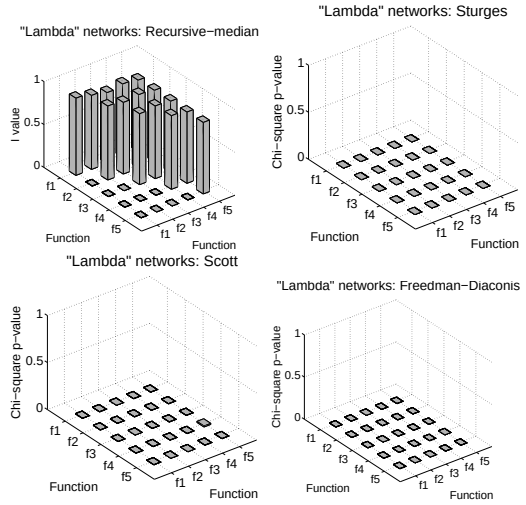

Figure 5: 15 Lambda-network results from recursive-median, Sturges, Scott, and Freedman-Diaconis methods. Only recursive-median (correctly) indicates conditional independence in all cases.

is because the ground truth (whether the variables in the data set are in fact conditionally independent or not) cannot be known with certainty except for artificially generated data sets. We therefore used data sets generated from the following 3-variable BNs:

- **"Lambda" networks:** These are structured as $X \leftarrow Z \rightarrow Y$, and therefore $(X \perp\!\!\!\perp Y \mid Z)$ by construction.
- **"V" networks:** These are structured as $X \rightarrow Z \leftarrow Y$, and therefore $(X \not\perp\!\!\!\perp Y \mid Z)$ by construction.

We used all pairs of the functions shown in the table on the right for the BN edges, combined additively in the case of the V-networks, and added Gaussian noise. 20,000 data points were used in each experiment. The matrices of CI test results for all function combinations are summarized in Figs. 5 and 6. For clarity we do not show results below the matrix diagonal since it is symmetric. As can be seen, while all prediscretization methods were able to capture conditional dependence (Fig. 6), only the recursive-median algorithm was able to detect conditional independence ($\mathcal{I} > 0.5$) in all 15 Lambda-networks (Fig. 5).

Functions	
f_1	x
f_2	$2\sin(x)$
f_3	$\ln \lvert x \rvert$
f_4	$1/(x/5+1)$
f_5	e^x

We also measured statistical consistency, which is defined as a property of an estimator to have mean square error tend to 0 as the data set size tends to infinity. For this purpose we used the MEANDER data set, shown in the top left part of Fig. 7. It resembles a spiral, and is challenging because the joint PDF of X and Y given Z changes dramatically with Z, making Theorem 1 inapplicable without subdivision of the Z axis. The data set was generated using the following nonlinear equations:

$$\begin{aligned} Z &\sim 0.75 \times N(0, 1/5) + 0.25 \times N(1, 1/3) \\ X &\sim Z/10 + (1/2)\sin(2\pi Z) + 0.1 \times N(0, 1) \\ Y &\sim Z/5 + \sin(2\pi Z + 0.35) + 0.1 \times N(0, 1) \end{aligned}$$

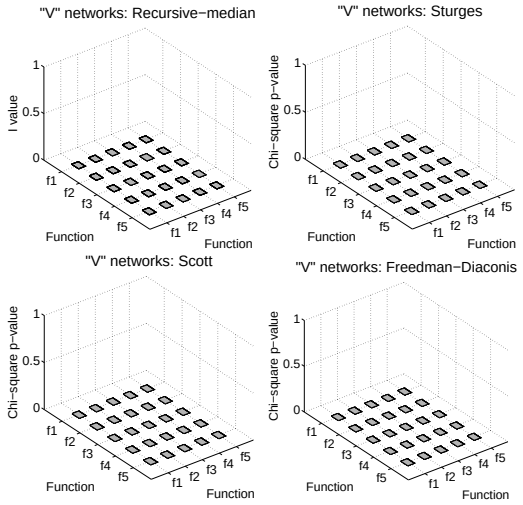

Figure 6: **Top row:** 15 V-network results from recursive-median, Sturges, Scott, and Freedman-Diaconis methods. All methods (correctly) indicate conditional dependence.

where $N(\mu, \sigma)$ denotes a normal (Gaussian) distribution with mean μ and standard deviation σ. According to the these functional descriptions, X and Y are independent given Z. However, as can be seen in the top right part of Fig. 7, they are unconditionally dependent. We generated 20 data sets of 25,000 examples each, and then ran all algorithms on subsets of each, of sizes ranging from 1,000 to all 25,000 data points. For each data set size, we averaged over the 20 data sets of that size to obtain a 95% confidence interval of our estimates of $\mathcal{I}$ and $\mathcal{U}$ using the recursive-median algorithm. We plot $\mathcal{I}$ and $\mathcal{U}$ values for recursive-median, and also the p-values for the prediscretization methods in Fig. 7 (bottom). As can be seen, all three prediscretization methods returned a very low p-value (too close to 0 to be seen clearly in the graph), strongly indicating conditional dependence (incorrectly), with little or no improvement for larger data sets. The recursive-median returns values of $\mathcal{I}$ much greater than 0.5 and tending to 1 with increasing sample size, indicating statistical consistency of both $\mathcal{I}$ and $\mathcal{U}$.

Conclusion

To the best of our knowledge, this is the first time the learning of structure of graphical models in continuous domains is demonstrated, using independence-based methods and without making use of any distributional assumptions. This was made possible by the use of the probabilistic nonparametric conditional independence test presented in this paper, and an effective algorithm (recursive-median) for estimating it from a data set. Our evaluation on both real and artificial data sets shows that can be used to learn a graphical model consistent with previous studies of the application domain, and that it performs well against alternative methods drawn from the statistical literature.

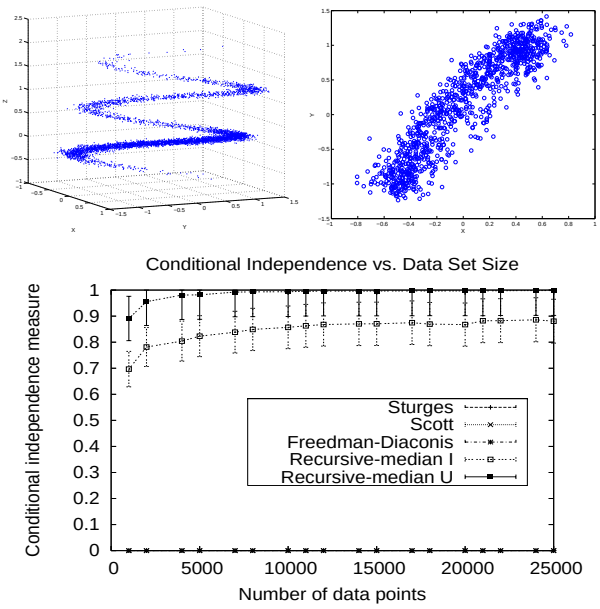

Figure 7: **Top left:** Three-dimensional plot of the MEANDER data set. $(X \perp\!\!\!\perp Y \mid Z)$ by construction (see text). **Top right:** Projection of data along Z axis (XY plane). $(X \not\perp\!\!\!\perp Y)$ (unconditionally). **Bottom**: $\mathcal{I}$ and $\mathcal{U}$ values for the recursive-median method, and p-values of the χ^2 test for Sturges's, Scott's and Freedman and Diaconis's methods. $\mathcal{I}$ (correctly) indicates conditional independence. All p-values are close to zero, (mistakenly) indicating conditional dependence.

References

Dougherty, J.; Kohavi, R.; and Sahami, M. 1995. Supervised and unsupervised discretization of continuous features. In Prieditis, A., and Russel, S., eds., *Proceedings of the Twelfth International Conference on Machine Learning*, 194–202. Morgan Kaufmann Publishers, San Francisco, CA.

Freedman, D., and Diaconis, P. 1981. On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57:453–476.

Geiger, D., and Heckerman, D. 1994. Learning gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research, Advanced Technology Division.

Heckerman, D. 1995. A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division.

Lam, W., and Bacchus, F. 1994. Learning Bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence* 10:269–293.

Lauritzen, S. L. 1996. *Graphical models*. New York: Oxford University Press.

Margaritis, D., and Thrun, S. 2001. A Bayesian multiresolution independence test for continuous variables. In *Uncertainty in Artificial Intelligence (UAI)*, 346–353. Morgan Kaufmann.

Scott, D. W. 1979. On optimal and data-based histograms. *Biometrika* 66:605–610.

Scott, D. W. 1992. *Multivariate Density Estimation*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc.

Spirtes, P.; Glymour, G.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag.