LEARNING HYBRID BNs FROM DATA FACTS & FICTION

Anca Hanea & Stefan Blauw

6th Annual Conference of the Australasian Bayesian Network Modelling Society

November 26 - 27, 2014
A False Report That One Pretends Is True

“These random variables can be either continuous or discrete. For simplicity, in this paper we shall only consider discrete ones.”

“Our solution will benefit the determination of the structure of Bayesian networks in domains that contain any number of continuous, ordinal discrete and/or categorical attributes.”

“We can now address complete data problems with reasonable confidence in problems of moderate size.”
Outline

• The type of BNs we work with: non-parametric BNs

• A new structure learning algorithm for NPBNs

• Comparisons

• Conclusions
What type of BNs we work with?
Hybrid Bayesian Networks

- Any continuous variables – possibility to include discrete variables
  - marginal distributions
  - a measure of bivariate dependence
  - some assumption about the bivariate dependence
Measure of Bivariate Dependence

A measure of bivariate dependence

\[ \rho(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} \]

\[ r(X,Y) = \rho(F_X(X), F_Y(Y)) \]
“[...] the skeptic might still wonder how the numbers that are still required are, in fact, obtained. In all the examples described previously, they are made up. Naturally, nobody actually makes this statement. What one really says is that they are elicited from an expert who subjectively assesses them. This statement sounds a lot better, but there is really nothing wrong with making up numbers. For one thing, experts are fairly good at it.”
Why the Rank Correlation?

- It always exists
- It does not depend on the marginal distributions (non-parametric measure of dependence)
- It measures monotone dependence, i.e. it assesses how well an arbitrary monotonic function describes the relationship between two variables
- It is a property of a copula
- We know how to elicit it from experts
“Shape” of Bivariate Dependence

An assumption about the bivariate dependence - copula

Density of Clayton copula with correlation 0.7.

Density of Frank copula with correlation 0.7.
Non-Parametric Bayesian Networks
Non-Parametric Bayesian Networks
Non - Parametric Bayesian Networks

- **nodes** – arbitrary continuous, invertible distribution functions
- **influences** – (conditional) copulae parametrised by the rank correlation
- **rank correlations** – any number between -1 and 1 is a valid choice
- **absence of arcs** – independent copula

Quantify with data and/or experts

The conditional rank correlations, the copula, the one-dimensional margins and the conditional independence statements implied by the graph **uniquely** determine the joint distribution

---

**ceb ra**
Centre of Excellence for Biosecurity Risk Analysis

**TU Delft**
Delft University of Technology
Discrete Nodes in NPBNs

The dependence structure is specified with respect to the uniform variates

- Discrete ordinal variables can be written as monotone transforms of uniform variables
- Adjusted rank correlation coefficient
  \[ \tilde{r} = f(C_r, \text{marginal distributions}) \]
What if Data is Not Available?

Structured Expert Judgment

Ask experts about:

• Marginal distribution of $X_6$

• The rank correlation between $X_4$ and $X_6$
Non - Parametric Bayesian Networks

Functional relationships can also be represented in a NPBN
Sampling & Inference in a NPBN

**Sampling**

- Use an arbitrary copula
- Use the normal copula

**Inference**

- Re-sample the NPBN for any new conditionalisation
- Use a hybrid approach - sample once and use a discretised version of the NPBN
- Assume the normal copula - use analytical expressions
NPBN - Applications
Learning the Structure of a NPBN

Assumption: normal copula
Validation: possible
How should we build a NPBN?
• Start with an empty graph
• Assume/Test the normal copula
• Add arcs based on highest rank correlations (arbitrary directions)
• Stop when you represent “enough” dependence [1]

• The data is continuous, so let it be!
• Use the normal copula, and the highest rank correlation to add edges
• Direct the edges using smart rules
• Use independence tests for continuous variables
• Use expert opinion [4]

• No, no, no
• Discretise and use a well-established algorithm for discrete
• …and then deal with all the problems of learning a structure on many variables from a discretised small data set [2]
The new structure learning algorithm for NPBNs
Right! I now have enough material to improve the improvements, design a new algorithm and compare it with established ones on known data sets! [5]
Learning the Structure of a NPBN

- **Validation phase**: the copula assumptions of the NPBN is validated
- **Initial phase**: an empty graph (nodes only) is filled with N edges
  - use a Kernel-based independence test [6] to determine pairs of independent variables
  - add edges based on the highest rank correlation
  - make sure you never add an edge between independent variables
- **Main phase**: each edge is directed and becomes an arc
  - Use similar rules as the IC and PC algorithms
  - Use the Kernel based conditional independence test
  - Extra rules for avoiding colliders
  - A tiny amount of randomness at the end
- **Final phase**: validate that the NPBN represents “enough” dependence
Comparisons
Abalone Dataset

- The Abalone dataset contains measurements from a study conducted to predict the age of abalones from their physical characteristics.
- There are 4177 joint samples and no missing values.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Node name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>LENG</td>
<td>Longest shell measurement [mm]</td>
</tr>
<tr>
<td>Diameter</td>
<td>DIAM</td>
<td>Perpendicular to length [mm]</td>
</tr>
<tr>
<td>Height</td>
<td>HEIG</td>
<td>With meat in shell [mm]</td>
</tr>
<tr>
<td>Whole weight</td>
<td>WHWE</td>
<td>Whole abalone [g]</td>
</tr>
<tr>
<td>Shucked weight</td>
<td>SHWE</td>
<td>Weight of meat [g]</td>
</tr>
<tr>
<td>Viscera weight</td>
<td>VIWE</td>
<td>Gut weight (after bleeding) [g]</td>
</tr>
<tr>
<td>Shell weight</td>
<td>SLWE</td>
<td>After being dried [g]</td>
</tr>
<tr>
<td>Number of rings</td>
<td>RING</td>
<td>+1.5 gives the age in years</td>
</tr>
</tbody>
</table>
Abalone Dataset

Output of the new algorithm

Output of PC algorithm with the RM test [7]
Housing Dataset

- This dataset considers the value of houses in downtown Boston
- It contains 13 continuous variables and 1 binary (dummy) variable
- 3 of the 13 continuous variables can be considered to be discrete, hence 10 variables will be considered in building a NPBN
- 506 joint samples are available and there are no missing values

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRIME</td>
<td>Crime rate per town</td>
</tr>
<tr>
<td>INDUSTRY</td>
<td>Proportion of non-retail business acres per town</td>
</tr>
<tr>
<td>NOX</td>
<td>Nitric oxides concentration</td>
</tr>
<tr>
<td>ROOMS</td>
<td>Average number of rooms per dwelling</td>
</tr>
<tr>
<td>AGE</td>
<td>Proportion of owner-occupied units built prior to 1940</td>
</tr>
<tr>
<td>DISTANCE</td>
<td>Weighted distances to five Boston employment centres</td>
</tr>
<tr>
<td>TAX</td>
<td>Full-value property-tax rate per $10,000</td>
</tr>
<tr>
<td>B</td>
<td>$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town</td>
</tr>
<tr>
<td>LOWER STATUS</td>
<td>% Lower status of the population</td>
</tr>
<tr>
<td>VALUE</td>
<td>Median value of owner-occupied homes in $1000’s</td>
</tr>
<tr>
<td>ZN</td>
<td>Proportion of residential land zoned for lots over 25,000 sq.ft.</td>
</tr>
<tr>
<td>RIVER</td>
<td>Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)</td>
</tr>
<tr>
<td>TEACHER</td>
<td>Pupil-teacher ratio by town</td>
</tr>
<tr>
<td>RAD</td>
<td>Index of accessibility to radial highways</td>
</tr>
</tbody>
</table>
Housing Dataset

Output of the new algorithm (with the KCI test)

Output of PC algorithm with the KCI test
Housing Dataset

Output of PC algorithm with the $C_{\text{I}_{\text{perm}}}$ test [6]

Output of PC algorithm with the KCI test
Housing Dataset

Output of PC algorithm with the $C_{\text{perm}}$ test

Output of PC algorithm with the RM test
Some conclusions

“Our evaluation on both real and artificial data sets shows that [our algorithm] 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”

"We applied the method on both simulated and real world data, and the results suggested that the method outperforms existing techniques in both accuracy and speed"
Conclusions

• The proposed algorithm still needs to be tested … much more

• The algorithm adds 1.5 times more edges (on average) than the other algorithms – thinning phase

• Directing the arcs rarely produces contradictions

• No arcs remain undirected (because of the last step) – but is this a good idea?

• Starting with a pre-prepared DAG might be a better idea
Conclusions

- Expert opinion is crucial – but how do we do it in a structured, reproducible, reliable way?

- One should refrain from using BN modelling software as a “black box”

- *In most applications, no standard solutions exist and insight in the different steps of the process is key to obtaining useful and reliable results*
Thank You
anca.maria.hanea@gmail.com
References


