A pot-pourri of Bayesian network learning methods

Robert Cowell

Faculty of Actuarial Science and Insurance
Cass Business School

LMS Durham Symposium
Mathematical Aspects of Graphical models
1^{st} July 2008

Topics

- Parameter estimation for discrete probability tables.
- Structural learning of discrete Bayesian networks.
- Learning structure from large datasets *exploiting pairwise marginals*
- Testing (conditional) independence of continuous random variables for learning structure of continuous Bayesian networks.

Topics

- Parameter estimation for discrete probability tables.
- Structural learning of discrete Bayesian networks.
- Learning structure from large datasets *exploiting pairwise marginals*
- Testing (conditional) independence of continuous random variables for learning structure of continuous Bayesian networks.

Pretty much a non-Bayesian talk, more engineering.

Parameter estimation for discrete BNs

Basic to learning the structure of a discrete Bayesian network is some method for estimating marginal or conditional probability tables, for use in (a) score based methods or (b) conditional independence test methods.

- Usual to have a complete dataset
- Here relax to incomplete data, assuming Missing At Random (MAR)
- Specifically advocate use method in AISTATS '99 paper, based on maximum entropy.

Estimate joint distribution of two binary variables X and Y from incomplete dataset.
Introduce random variables: $X^* \in (x_0^* \equiv ?, x_1^* \equiv x_1, x_2^* \equiv x_2)$ and $Y^* \in (y_0^* \equiv ?, y_1^* \equiv y_1, y_2^* \equiv y_2)$.
In terms of these variables, the dataset is *complete*.

	$Y = ?$	$Y = y_1$	$Y = y_2$
$X = ?$	n_{00}	n_{01}	n_{02}
$X = x_1$	n_{10}	n_{11}	n_{12}
$X = x_2$	n_{20}	n_{21}	n_{22}

Set $p_{ij,kl} := P(x_i, y_j, x_k^*, y_l^*)$.
Under MAR assumption, we have non-linear constraints:

$$
\begin{aligned}
p_{ij,00} &= P(x_i, y_j)P(x_0^*, y_0^*) \\
p_{ij,i0} &= P(y_j \mid x_i)P(x_i^*, y_0^*) \\
p_{ij,0j} &= P(x_i \mid y_j)P(x_0^*, y_j^*).
\end{aligned}
$$

Maximize entropy the joint distribution $P(X, Y, X^*, Y^*)$, subject to constraints, by iteration. Then marginalize to get the desired estimate of $P(X, Y)$.
Equivalent to EM algorithm.

Local EM Estimation

In a BN, I estimate the conditional table of a node given parents using only the data on the family (hence *local EM*): Estimate the joint, marginalize to parents, then condition.

- Use previous iteration scheme with two or more variables if data incomplete.
- Fast compared to full EM, and usually quite accurate.
- Estimates can be used as starting point for full EM estimation in BN.

Equivalence of scoring and conditional independence tests (UAI 2001).

In learning BN structure:

- Assume complete data (discrete).
- Assume node ordering.
- No latent variables.

Then: Incremental structure learning based on

- conditional independence tests using cross-entropy, and
- score based search using maximum likelihood

are equivalent.

Nested models $g \subset g$‘ differing in parent set in one node X_i: $pa(X_i : g') \supset pa(X_i : g)$. Log-likelihood difference of models:

$$\log \frac{L(\hat{p}_{g'})}{L(\hat{p}_g)} = \sum_{x_i, pa(x_i:g')} n(x_i, pa(X_i : g')) \log \frac{n(x_i, pa(x_i : g'))/n(pa(x_i : g'))}{n(x_i, pa(x_i : g))/n(pa(x_i : g))}$$

Equal to conditional independence (conditional cross-entropy) test-statistic after scaling:

$$\frac{1}{N} \log \frac{L(\hat{p}_{g'})}{L(\hat{p}_g)} = \sum_{x_i, pa(x_i:g')} \hat{p}(x_i, pa(x_i : g')) \log \frac{\hat{p}(x_i \mid pa(x_i : g'))}{\hat{p}(x_i \mid pa(x_i : g))},$$

So I use the conditional independence testing approach.

Learning structure from large datasets

- Main computational bottle-neck is traversing dataset to estimate joint probabilities for CI tests (uses up to 98% of processing time).
- Look at ways to reduce this using **bivariate** probability tables.
- Finding Chow-Liu tree requires using only marginal and pairwise tables.

Learning structure from large datasets

- Main computational bottle-neck is traversing dataset to estimate joint probabilities for CI tests (uses up to 98% of processing time).
- Look at ways to reduce this using **bivariate** probability tables.
- Finding Chow-Liu tree requires using only marginal and pairwise tables.

Question Does anyone know of an example of a connected Bayesian network whose induced Chow-Liu tree is not a subgraph of the network?

A simple lemma

Lemma

Let A, B and C denote three mutually disjoint sets of discrete random variables, having joint probability distribution $P(A, B, C)$. Let I_{XY} denote the Kullback–Leibler divergence

$$I_{XY} = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

where X and Y are disjoint sets of discrete random variables. Then, if A and C are conditionally independent given B:

$$I_{AC} \leq I_{AB} + I_{BC}.$$

A simple lemma

Lemma

Let A, B and C denote three mutually disjoint sets of discrete random variables, having joint probability distribution $P(A, B, C)$. Let I_{XY} denote the Kullback–Leibler divergence

$$I_{XY} = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

where X and Y are disjoint sets of discrete random variables. Then, if A and C are conditionally independent given B:

$$I_{AC} \leq I_{AB} + I_{BC}.$$

Note: Only requires bivariate tables for singleton sets.

A simple lemma

Lemma

Let A, B and C denote three mutually disjoint sets of discrete random variables, having joint probability distribution $P(A, B, C)$. Let I_{XY} denote the Kullback–Leibler divergence

$$I_{XY} = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

where X and Y are disjoint sets of discrete random variables. Then, if A and C are conditionally independent given B:

$$I_{AC} \leq I_{AB} + I_{BC}.$$

Note: Only requires bivariate tables for singleton sets.
In fact: (Pearl, 8.15) $I_{AC} \leq \min(I_{AB}, I_{BC})$.

More lemmas

Under assumptions of previous lemma:

Lemma

$$I_{AC} \leq \min(I_{AB}, I_{BC}) \textit{ and } I_{AB} + I_{BC} \leq I_{AC} - I_B$$

where $I_B = \sum_b p(b) \log p(b)$.

More lemmas

Under assumptions of previous lemma:

Lemma

$$I_{AC} \leq \min(I_{AB}, I_{BC}) \textit{ and } I_{AB} + I_{BC} \leq I_{AC} - I_B$$

where $I_B = \sum_b p(b) \log p(b)$.

Lemma

$$\check{I}_{AC} \leq \check{I}_{AB} + \check{I}_{BC}$$

where

$$\check{I}_{XY} = \sum_{x,y} p(x)p(y) \log \frac{p(x)p(y)}{p(x,y)}$$

More lemmas

Under assumptions of previous lemma:

Lemma

$$I_{AC} \leq \min(I_{AB}, I_{BC}) \textit{ and } I_{AB} + I_{BC} \leq I_{AC} - I_B$$

where $I_B = \sum_b p(b) \log p(b)$.

Lemma

$$\check{I}_{AC} \leq \check{I}_{AB} + \check{I}_{BC}$$

where

$$\check{I}_{XY} = \sum_{x,y} p(x)p(y) \log \frac{p(x)p(y)}{p(x,y)}$$

Conjecture $\check{I}_{AC} \leq \min(\check{I}_{AB}, \check{I}_{BC})$.

IPF estimation

Suppose during incremental BN building X has parent set Y, and we wish to check if $X \perp\!\!\!\perp Z \mid Y$ for singleton Z.
Could estimate $P(X, Y, Z)$ from data, but this could have problems:

- Slow to do this many times.
- Table counts could get quite sparse, so estimate of $P(X, Y, Z)$ unreliable

Approximate solution: use IPF (iterative proportional fitting) using bivariate tables to estimate $P(X, Y, Z)$.

IPF estimation

Suppose during incremental BN building X has parent set Y, and we wish to check if $X \perp\!\!\!\perp Z \mid Y$ for singleton Z.
Could estimate $P(X, Y, Z)$ from data, but this could have problems:

- Slow to do this many times.
- Table counts could get quite sparse, so estimate of $P(X, Y, Z)$ unreliable

Approximate solution: use IPF (iterative proportional fitting) using bivariate tables to estimate $P(X, Y, Z)$.

Lemma

If A, B and C are discrete random variables, and $A \perp\!\!\!\perp B \mid C$, then IPF using bivariate marginals gives the correct joint table $P(A, B, C)$ (starting from uniform joint table).

Advantages of IPF estimation

- Bivariate marginals usually estimated and cached during the initial BN model search stage.
- For 'typical' problems, could expect that estimates should be reasonable: in terms of log-linear models, higher-order interactions have smaller influence on joint distribution 'typically' in natural world.
- Avoids bottleneck problem, no need to traverse the dataset again.
- Fast - can also combine with local-EM for estimates using incomplete data.

Advantages of IPF estimation

- Bivariate marginals usually estimated and cached during the initial BN model search stage.
- For 'typical' problems, could expect that estimates should be reasonable: in terms of log-linear models, higher-order interactions have smaller influence on joint distribution 'typically' in natural world.
- Avoids bottleneck problem, no need to traverse the dataset again.
- Fast - can also combine with local-EM for estimates using incomplete data.

This seems to work moderately well (at least on a few small examples) though I don't have extensive results to back this up.

Learning continuous BN's from data

Interested in not making distributional assumptions:
non-parametric test.
Assume X, Y and Z are continuous. Have a complete dataset, size N.

Learning continuous BN's from data

Interested in not making distributional assumptions: non-parametric test.
Assume X, Y and Z are continuous. Have a complete dataset, size N. For test of $X \perp\!\!\!\perp Y$ use

$$P(x_1 < X < x_2 \text{ and } y_1 < Y < y_2) = P(x_1 < X < x_2)P(y_1 < Y < y_2).$$

Estimate various probabilities using frequency counts of data. Use Pearson goodness-of-fit statistic to assess independence

Learning continuous BN's from data

Interested in not making distributional assumptions: non-parametric test.
Assume X, Y and Z are continuous. Have a complete dataset, size N. For test of $X \perp\!\!\!\perp Y$ use

$$P(x_1 < X < x_2 \text{ and } y_1 < Y < y_2) = P(x_1 < X < x_2)P(y_1 < Y < y_2).$$

For test of $X \perp\!\!\!\perp Y \mid Z$ use

$$P(x_1 < X < x_2, y_1 < Y < y_2, z_1 < Z < z_2)$$

$$\approx \frac{P(x_1 < X < x_2, z_1 < Z < z_2)P(y_1 < Y < y_2, z_1 < Z < z_2)}{P(z_1 < Z < z_2)}$$

Estimate various probabilities using frequency counts of data. Use Pearson goodness-of-fit statistic to assess independence /conditional independence.

Independence test

For test of $X \perp\!\!\!\perp Y$, standard Pearson goodness-of-fit test would partition (X, Y) region with rectangular grid.

- How many grid lines?
- Where to place the grid-lines?
- For smaller dataset, grid may be too coarse to be useful.
  - Even more so when extending to conditional independence test.

Suggested solution: Median partitioning

Consider test of $X \perp\!\!\!\perp Y$. Instead of making a rectangular grid, make a *random* plane partitioning using the following recursive method:

1. Randomly choose X or Y.

Suggested solution: Median partitioning

Consider test of $X \perp\!\!\!\perp Y$. Instead of making a rectangular grid, make a *random* plane partitioning using the following recursive method:

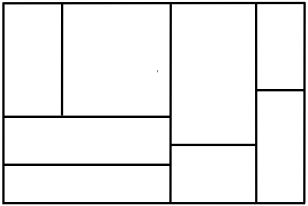
1. Randomly choose X or Y.
2. Partition dataset into two parts, according to median of chosen variable.

Suggested solution: Median partitioning

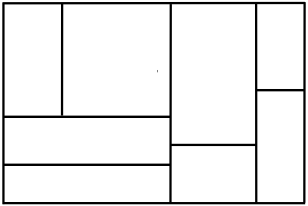
Consider test of $X \perp\!\!\!\perp Y$. Instead of making a rectangular grid, make a *random* plane partitioning using the following recursive method:

1. Randomly choose X or Y.
2. Partition dataset into two parts, according to median of chosen variable.
3. Repeat Steps 1 and 2 on the data subsets, recursively, until datasets are smaller than some number of observations (25 say).

Illustration



Illustration



For each rectangular region i, can:

- Find observed counts o_i
- Estimate expected counts e_i based on independence assumption.
- Find correlation c_i between X-Y values.

These can be cumulated over all regions to give:

- Goodness of fit statistic: $\chi^2 = \sum_{i=1}^{I}(o_i - e_i)^2/e_i$.
  - What distribution does it follow? Probably χ^2 (Wilks' theorem) but what degree of freedom? Fractal dimension?
- Estimate of I_{XY}: $(1/N)\sum_{i=1}^{I} o_i \log(o_i/e_i)$
  - Note: If independence holds, then $\chi^2 \approx 2NI_{XY}$.
- Root-mean-square correlation $\sqrt{\sum_{i=1}^{I} c_i^2/I}$.
  - Could also look at $\sum_{i=1}^{I} \mathrm{abs}(c_i)/I$

These can be cumulated over all regions to give:

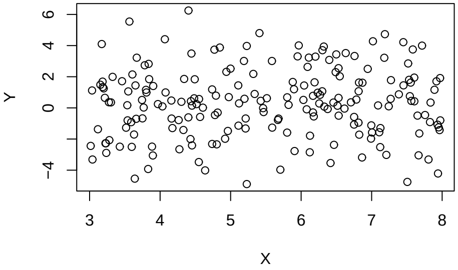
- Goodness of fit statistic: $\chi^2 = \sum_{i=1}^{I}(o_i - e_i)^2/e_i$.
    - What distribution does it follow? Probably χ^2 (Wilks' theorem) but what degree of freedom? Fractal dimension?
- Estimate of I_{XY}: $(1/N)\sum_{i=1}^{I} o_i \log(o_i/e_i)$
    - Note: If independence holds, then $\chi^2 \approx 2NI_{XY}$.
- Root-mean-square correlation $\sqrt{\sum_{i=1}^{I} c_i^2/I}$.
    - Could also look at $\sum_{i=1}^{I} \text{abs}(c_i)/I$

By repeating the partitioning process, distributions of these three quantities may be built up.

Example 1: Independent data

Data size $N = 200$. Correlation in data 0.0585099.

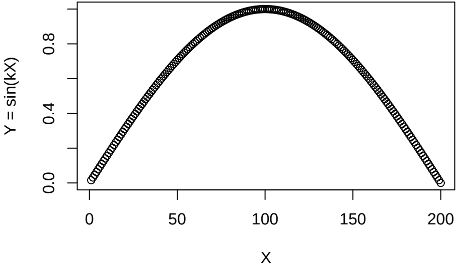
Number cells	χ^2	I_{XY}	$\chi^2/(2NI_{XY})$	RMS-corr
16	4.16627	0.0098185	1.06082	0.277886
16	7.21718	0.0181548	0.99384	0.244964
16	5.81515	0.0143282	1.01463	0.227799
16	4.20133	0.0101659	1.03319	0.313294
16	6.88750	0.0161347	1.06719	0.255480



Example 2

Data size $N = 201$. Correlation in data 3.0088e-06.

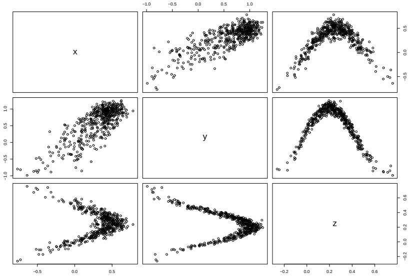
Number cells	χ^2	I_{XY}	$\chi^2/(2NI_{XY})$	RMS-corr
16	30.8015	0.074389	1.03000	0.790612
16	72.1086	0.151718	1.18229	0.995518
16	50.4350	0.100801	1.24463	0.995501
16	138.447	0.290167	1.18689	0.995514
16	134.818	0.244971	1.36901	0.911523

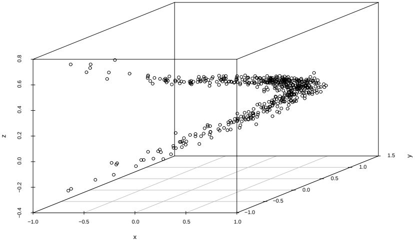


Example 3: $X \perp\!\!\!\perp Y \mid Z$

Data size $N = 20,000$. $X - Y$ Correlation in data 0.8078973.

Number cells	χ^2	$I_{XY \mid Z}$	$\chi^2/(2NI_{XY \mid Z})$	RMS-corr
1024	515.526	0.0125959	1.02320	0.240660
1024	442.617	0.0108970	1.01545	0.229239
1024	476.859	0.0117177	1.01739	0.221301
1024	535.646	0.0129877	1.03107	0.233340
1024	492.561	0.0119363	1.03165	0.234522





References

- R. G. Cowell (1999), Parameter estimation from incomplete data for Bayesian networks, In Artificial Intelligence and Statistics 99: Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics. (D. Heckerman and J. Whittaker, eds.) pp. 193-196. Morgan Kaufmann, San Francisco
- R. G. Cowell (2001), Conditions Under Which Conditional Independence and Scoring Methods Lead to Identical Selection of Bayesian Network Models, In Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001) (J. Breese and D. Koller, eds.)pp 91-97.
- J. Pearl. Probabilistic Reasoning in Intelligent Systems, Morgan-Kaufmann, 1988.

Postscript

Following the talk, F. Matus gave me a class of Bayesian networks for which an induced Chow-Liu tree is not a subgraph of the skeleton of the Bayesian network, of which this is an example.
Let A and B be independent binary random variables with states $\{0,1\}$ and uniform distribution $p(0) = p(1) = 0.5$ on each variable.
Let X and Y also be binary random variables with states $\{0,1\}$.
Make A and B parents of X, with conditional probability table given by $p(x|a,b) = 1$ if $x = a + b$ mod 2, and zero otherwise.
Also make A and B parents of Y with the same logical dependence.
Then all variables in the four node network are pairwise independent, and hence have zero cross-entropy, except for the pair (X, Y) which are logically dependent and therefor have non-zero cross entropy: however there is no direct edge between X and Y.