Non-stationary dynamic Bayesian network learning

Christophe Gonzales, Séverine Dubuisson
Cristina Manfredotti

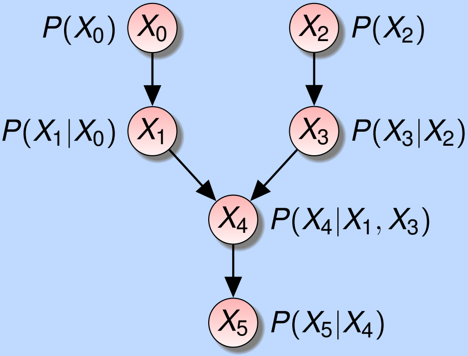
Sorbonne Universités, UPMC – LIP6, France

Outline

Bayesian network

Definition: Bayesian network [Pearl (1988)]

1. A directed acyclic graph (DAG):



joint distribution: $P(X_1, \ldots, X_5) = \prod_{i=1}^{5} P(X_i|\mathbf{Pa}(X_i))$

2. To each node X_i is assigned $P(X_i|\mathbf{Pa}(X_i))$

Bayesian network structure learning

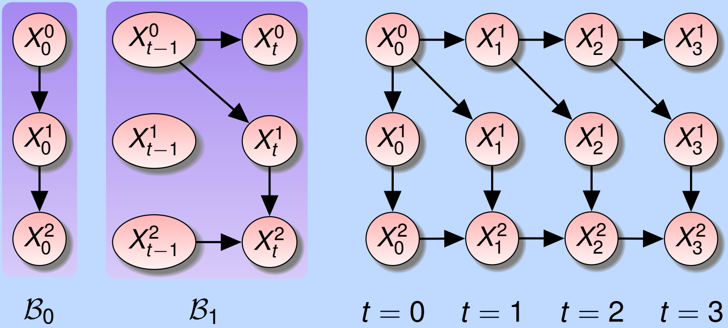
Structure Learning

- Database: **D**
- Problem: find structure $\mathcal{G}$ best fitting **D**

- 3 classes of algorithms:
  - search-based approaches: $\text{Argmax}_{\mathcal{G}}\ P(\mathcal{G}|\mathbf{D})$
    scoring (K2, BD, BDeu, BIC, AIC, *etc.*)
    [Cooper & Herskovits (92), Heckerman, Geiger & Chickering (95)]
  - constraint-based approaches:
    independence tests (χ^2, G^2, *etc.*)
    [Verma & Pearl (91), Spirtes, Glymour & Scheines (93)]
  - hybrid approaches [de Campos (06)]
- Key idea: start from $\mathcal{G}$ and search locally for a better structure $\mathcal{G}'$

Stationary dynamic Bayesian network learning

Dynamic Bayesian network (DBN) *[Dean & Kanazawa (89)]*



DBN structure learning:

Hypothesis: $\mathcal{B}_1$ indep. $\mathcal{B}_0$ given $\mathbf{D}$

$$\text{Argmax}_{\mathcal{G}_0,\mathcal{G}_1} P(\mathcal{G}_0,\mathcal{G}_1|\mathbf{D}) = \text{Argmax}_{\mathcal{G}_0,\mathcal{G}_1} P(\mathcal{G}_0|\mathbf{D})P(\mathcal{G}_1|\mathcal{G}_0,\mathbf{D})$$

$$= \left(\text{Argmax}_{\mathcal{G}_0} P(\mathcal{G}_0|\mathbf{D}),\ \text{Argmax}_{\mathcal{G}_1} P(\mathcal{G}_1|\mathbf{D})\right)$$

[Murphy (02)]

Non-stationary dynamic Bayesian network

Definition: non-stationary DBN

- Collection $\langle(\mathcal{B}_h, T_h)\rangle_{h=0}^{m}$
- T_h: transition time
- $\mathcal{B}_h$: Bayes net during epoch $\mathbf{E}_h = (T_{h-1}, T_h]$

Learning non-stationary DBNs

2 tasks:

1. Find the best set of transition times T_h
2. Find, within each epoch $\mathbf{E}_h = (T_{h-1}, T_h]$, the best BN $\mathcal{B}_h$

Detecting transitions T_h (1/2)

Robinson & Hartemink (2010)

- Multiple database records at each time step
- No transition detection criterion: optimization instead
- Determine $\underset{\{H,T_0,\ldots,T_H,\mathcal{B}_0,\ldots,\mathcal{B}_H\}}{\text{Argmax}} P(H, T_0, \ldots, T_H, \mathcal{B}_0, \ldots, \mathcal{B}_H | \mathbf{D})$
- Algorithm:

  start from a given set of transitions

  repeat local searches for:

      finding the best structure given transitions

      changing the dates of the transitions

      splitting/merging epochs

  until convergence

Caveat:

- All time slices need be observed

Detecting transitions T_h (2/2)

Nielsen & Nielsen (2008)

- A single database record at each time step
- Transitions detected in streaming mode
- Transition detection criterion:
  - Current BN fitting: $\log\left(\frac{P(X_i=x_i)}{P(X_i=x_i|X_j=x_j\forall j\neq i)}\right)$

    $\log \gg 0 \implies$ maybe not a good fit
  - Trend of the log toward high values $\implies$ transition (2nd Discrete Cosine Transform component)

Caveat:

- Transition identified a long time after it occurred
- The log formula is questionable:
  $P(X_i = x_i) = [0.8, 0.2]$ v.s. $P(X_i = x_i|X_j = x_j\forall j \neq i) = [0.7, 0.3]$

Constrained non-stationary DBN learning

- Structure not evolving: Grzegorczyk & Husmeier (2009)
- Structure evolving w.r.t. fixed transition proba:
  Robinson & Hartemink (2010)
- Parameter independence between $\mathcal{B}_h$ and $\mathcal{B}_{h+1}$:
  Robinson & Hartemink (2010)

2 A new learning algorithm

Detecting transition times

- Streaming mode
- Current BN $\mathcal{B}_h$ up to time $t - 1$
- Dataset $\mathbf{D}_t$ at time t

Algorithm:

1. if change in the set of X_i's $\Longrightarrow$ transition T_h
2. else if newly encountered values of X_i $\Longrightarrow$ transition T_h
3. else goodness-of-fit test:
   - perform χ^2 test on each $X_i \cup \mathbf{Pa}(X_i)$
   - if at least one test indicates a change $\Longrightarrow$ transition T_h

Learning a new BN fragment $\mathcal{B}_{h+1}$

Attractive features:

- Structure: limit the evolution
- Structure: take into account the strengths of the $\mathcal{B}_h$'s arcs
- Parameters: dependence w.r.t. $\mathcal{B}_h$

Learning $\mathcal{B}_{h+1}$: the key idea

$$
\begin{aligned}
P(\mathcal{G}_{h+1}|\mathbf{D}_t,\mathcal{B}_h) &\propto P(\mathcal{G}_{h+1},\mathbf{D}_t|\mathcal{B}_h) \\
&= \int_{\mathbf{\Theta}_{h+1}} P(\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1},\mathbf{D}_t|\mathcal{B}_h)d\mathbf{\Theta}_{h+1} \\
&= \int_{\mathbf{\Theta}_{h+1}} P(\mathbf{D}_t|\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1},\mathcal{B}_h)P(\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1}|\mathcal{B}_h)d\mathbf{\Theta}_{h+1} \\
&= \int_{\mathbf{\Theta}_{h+1}} P(\mathbf{D}_t|\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1})P(\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1}|\mathcal{B}_h)d\mathbf{\Theta}_{h+1} \\
&= \int_{\mathbf{\Theta}_{h+1}} P(\mathbf{D}_t|\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1})\pi(\mathbf{\Theta}_{h+1}|\mathcal{G}_{h+1},\mathcal{B}_h)P(\mathcal{G}_{h+1}|\mathcal{B}_h)d\mathbf{\Theta}_{h+1} \\
&= P(\mathcal{G}_{h+1}|\mathcal{B}_h)\int_{\mathbf{\Theta}_{h+1}} P(\mathbf{D}_t|\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1})\pi(\mathbf{\Theta}_{h+1}|\mathcal{G}_{h+1},\mathcal{B}_h)d\mathbf{\Theta}_{h+1}
\end{aligned}
$$

Graph transition distribution $P(\mathcal{G}_{h+1}|\mathcal{B}_h)$

- An atomic graph transformation = $(X_{(s)}, Y_{(s)}, A_s)$, $A_{(s)} \in \{$arc addition (add), arc deletion (del), arc reversal (rev)$\}$
- $\Delta(\mathcal{G}_h, \mathcal{G}_{h+1}) = \langle (X_{(s)}, Y_{(s)}, A_s) \rangle_{s=1}^{c}$
- Robinson & Hartemink (2010): $P(\mathcal{G}_{h+1}|\mathcal{B}_h) \propto e^{-\lambda|\Delta(\mathcal{G}_h,\mathcal{G}_{h+1})|} = e^{-\lambda c}$
  $\Longrightarrow$ strength of the arcs not taken into account

Generalization of the formula

$P(\mathcal{G}_{h+1}|\mathcal{B}_h) \propto \prod_{s=1}^{c} e^{f(X_{(s)}, Y_{(s)}, A_s)}$

$$f(\cdot,\cdot,\cdot) = \begin{cases} -\lambda_d I(X_{(s)}, Y_{(s)}|\mathbf{Pa}(Y_{(s)})\backslash\{X_{(s)}\}) & \text{if } A_s = \text{del} \\ -\lambda_a I(X_{(s)}, Y_{(s)}|\mathbf{Pa}(Y_{(s)})) & \text{if } A_s = \text{add} \\ \frac{1}{2}[f(X_{(s)}, Y_{(s)}, \text{del}) + f(Y_{(s)}, X_{(s)}, \text{add})] & \text{if } A_s = \text{rev} \end{cases}$$

$I(X, Y|\mathbf{Z})$: takes into account the strengh of the arc (X, Y)

Strength of the arcs

- Ebert-Uphoff (2007):

$$I(X, Y|\mathbf{Z}) = \sum_{X,\mathbf{Z}} P(X,\mathbf{Z}) \sum_{Y} P(Y|X,\mathbf{Z}) \log \frac{P(Y|X,\mathbf{Z})}{P(Y|\mathbf{Z})},$$

- Nicholson & Jitnah (1998): approximation

$$I(X, Y|\mathbf{Z}) \approx \sum_{X,\mathbf{Z}} P(X)P(\mathbf{Z}) \sum_{Y} P(Y|X,\mathbf{Z}) \log \frac{P(Y|X,\mathbf{Z})}{P(Y|\mathbf{Z})}.$$

$$P(\mathcal{G}_{h+1}|\mathbf{D}_t,\mathcal{B}_h) \propto P(\mathcal{G}_{h+1}|\mathcal{B}_h) \int_{\mathbf{\Theta}_{h+1}} P(\mathbf{D}_t|\mathcal{G}_{h+1},\mathbf{\Theta}_{h+1})\pi(\mathbf{\Theta}_{h+1}|\mathcal{G}_{h+1},\mathcal{B}_h)d\mathbf{\Theta}_{h+1}$$

- Geiger & Heckerman (97) : justification of Dirichlet priors

  $\Longrightarrow$ Bayesian Dirichlet (BD) score:

$$\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij}+\alpha_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}+\alpha_{ijk})}{\Gamma(\alpha_{ijk})}$$

- Problem: which Dirichlet hyperparameters?

Feature: allow dependence between CPTs of $\mathcal{B}_h$ and of $\mathcal{B}_{h+1}$

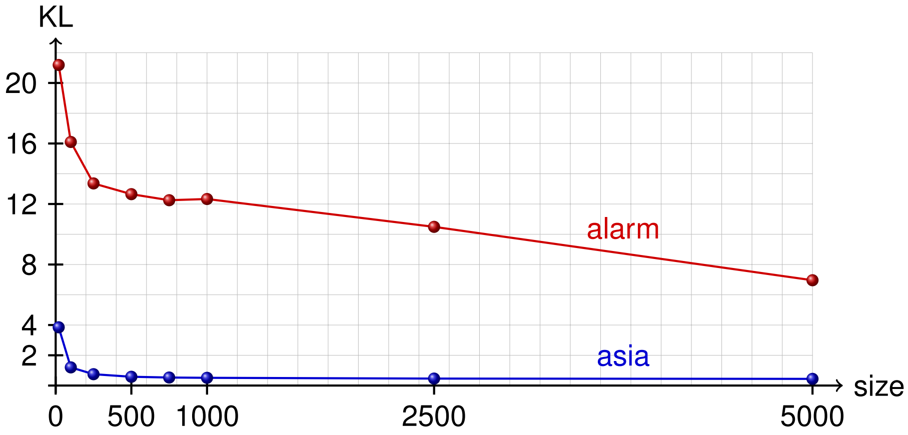
- $\hat{\mathcal{B}} = (\mathcal{G}_{h+1},\hat{\mathbf{\Theta}})$ = BN with minimal KL distance w.r.t. $\mathcal{B}_h$

  $\Longrightarrow$ hyperparameters = $N'\hat{\mathbf{\Theta}}$

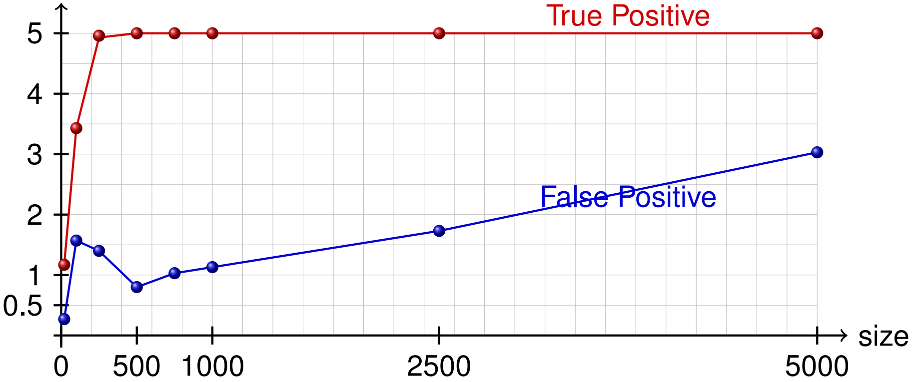
A flavor of experimentations

- DBN randomly generated from Alarm and Asia [Ide & Cozman (2002)]
- 5 epochs of 10 time slices

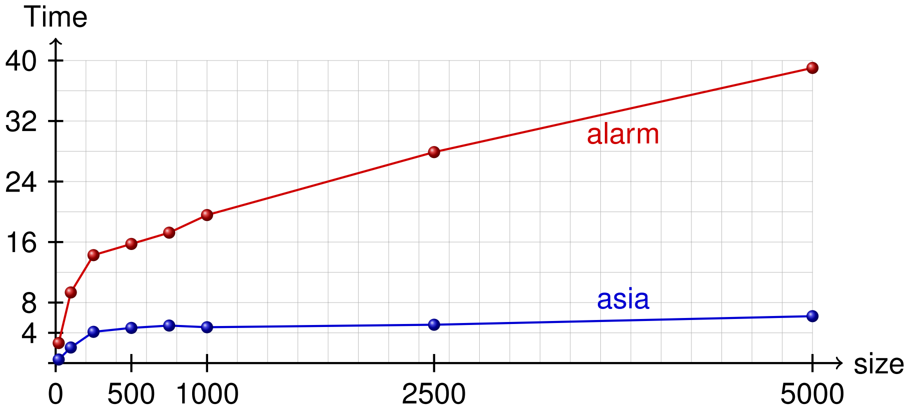
$\Longrightarrow$ grounded BNs $\approx$ (1850 nodes, 2500 arcs) and (400 nodes, 430 arcs)



A flavor of experimentations



A flavor of experimentations



grounded BNs ≈ (1850 nodes, 2500 arcs)
(400 nodes, 430 arcs)

Conclusions

- Framework mathematically sound
- Very flexible: take into account previous BNs:
  - Structure
  - Strength of the arcs
  - Parameters
- Scalable