

Introduction to Bayesian networks

- Probability theory
- Bayesian networks
 - Semantics
 - Learning
 - Example
- Dynamic Bayesian networks
- Gaussian networks
- Markov networks

Biological network = probability distribution

Measurements

Bayesian networks

"All models are wrong but some are useful"

11/14/2008 Jose M. Peña @ KI 1

Probability theory

- Biological network over genes, proteins, ... = probability distribution over genes, proteins, ...

GeneA → GeneB → GeneC

GeneA	GeneB	GeneC
0	0	0
1	1	1
0	1	1
1	1	0
0	1	0
1	0	1

- Deterministic relations ?
- Probabilistic relations ?
- Noise ?

11/14/2008 Jose M. Peña @ KI 2

Probability theory

- $p(X=x, Y=y) \geq 0$ for all x and y .
- $\sum_{x,y} p(X=x, Y=y) = 1$.
- Marginal:
 - $p(X=x) = \sum_y p(X=x, Y=y)$.
- Conditional:
 - $p(X=x | Y=y) = p(X=x, Y=y) / p(Y=y)$.
- Independence:
 - $p(X=x | Y=y) = p(X=x)$ for all x and y , or
 - $p(X=x, Y=y) = p(X=x)p(Y=y)$ for all x and y .

Conditional independence
 $p(X|YZ) = p(X|Z)$, or
 $p(X,Y|Z) = p(X|Z)p(Y|Z)$

11/14/2008 Jose M. Peña @ KI 3

Probability theory

$p(X,Y)$	Y=0	Y=1
X=0	0.3	0.3
X=1	0.2	0.2

	Y=0	Y=1
X=0	0.6	
X=1	0.4	

	Y=0	Y=1
$p(X Y)$		
X=0	0.3/0.5	0.3/0.5
X=1	0.2/0.5	0.2/0.5

11/14/2008 Jose M. Peña @ KI 4

Bayesian networks: Semantics

- $p(X,Y,Z) = p(X,Y)p(Z|X,Y) = p(X)p(Y|X)p(Z|X,Y)$.

- Drop the edge from X to Y, then
 - $p(X,Y,Z) = p(X)p(Y)p(Z|X,Y)$.
- In general: $p(X_1, \dots, X_n) = \prod_{X_i} p(X_i | \text{Parents}(X_i))$.

11/14/2008 Jose M. Peña @ KI 5

Bayesian networks: Semantics

- A Bayesian network (BN) over X_1, \dots, X_n consists of
 - a directed acyclic graph (DAG) over X_1, \dots, X_n and
 - probability distributions $p(X_i | \text{Parents}(X_i))$ for all i , and defines a probability distribution over X_1, \dots, X_n as $p(X_1, \dots, X_n) = \prod_{X_i} p(X_i | \text{Parents}(X_i))$.

SPRINKLER	T	F
RAIN	0.4	0.6
T	0.01	0.99

GRASS WET	T	F
SPRINKLER RAIN	1.0	0.0
F	0.8	0.2
T	0.9	0.1
T	0.99	0.01

RAIN	T	F
T	0.2	0.8
F	0.2	0.8

11/14/2008 6

Bayesian networks: Semantics

```

    graph TD
      X((X)) --> Z((Z))
      Y((Y)) --> Z((Z))
  
```

- $p(X)p(Y|X)p(Z|X,Y) = p(X,Y,Z) = p(X)p(Y)p(Z|X,Y)$, so $p(Y|X) = p(Y)$, so X is **independent** of Y !!!
 No numeric calculation required !!!

11/14/2008 Jose M. Peña @ KI 7

Bayesian networks: Semantics

- Which independencies are represented in a DAG ?

11/14/2008 Jose M. Peña @ KI 8

Bayesian networks: Semantics

```

    graph TD
      X((X)) --> Z((Z))
      Y((Y)) --> Z((Z))
  
```

- $p(X,Z) = \sum_y p(X)p(Y|X)p(Z|Y) \neq p(X)p(Z)$.
- $p(X,Z|Y) = p(X)p(Y|X)p(Z|Y)/p(Y) = p(X)p(X|Y)p(Z|Y)/p(Y) = p(X|Y)p(Z|Y)$.

11/14/2008 Jose M. Peña @ KI 9

Bayesian networks: Semantics

```

    graph TD
      U((U)) --> X((X))
      V((V)) --> Y((Y))
      X --> Z((Z))
      Y --> Z
      W((W)) --> Z
  
```

U ⊥ V ? U ⊥ V | ZW ?

11/14/2008 Jose M. Peña @ KI 10

Bayesian networks: Semantics

- Which independencies are represented in a DAG ? Use the **d-separation** criterion.
- X is **independent** of Y **given** Z when for every undirected path between X and Y there exists a node Z in the path such that
 - Z doesn't have two parents in the path and Z is in Z, or
 - Z has two parents in the path and neither Z nor any of its descendants is in Z.

11/14/2008 Jose M. Peña @ KI 11

Bayesian networks: Semantics

```

    graph TD
      U((U)) --> X((X))
      V((V)) --> Y((Y))
      X --> Z((Z))
      Y --> Z
  
```

$XU \perp YV ?$
 $X \perp Y | Z ?$
 $X \perp Y | UV ?$
 $U \perp Z | X ?$
 $U \perp Y | ZV ?$

- Two DAGs represent the same independencies iff they have the same **adjacencies** and **immoralities**.

11/14/2008 Jose M. Peña @ KI 12

Bayesian networks: Semantics

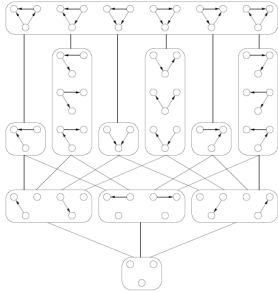


Figure 2: Hasse diagram of the space of Markov equivalence classes of Bayesian network structures over three variables. 13

Bayesian networks: Semantics

- The Markov boundary of X , $MB(X)$, is the minimal set such that $X \perp \text{Rest} \mid MB(X)$.

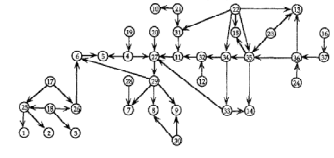


Figure 4: The ALARM belief network structure, containing 37 nodes and 46 arcs.

- $MB(X) = \text{Parents}(X) \cup \text{Children}(X) \cup \text{Spouses}(X)$.

Bayesian networks: Learning

- So far, we have a framework that
 - simplifies specifying probability distributions, and
 - enables us to reason quantitative and qualitatively.
- Is it learnable from data ?
- Learning a BN $B=(B_s, B_p)$ consists in
 - learning the best DAG B_s , and
 - learning the best probability distributions B_p for B_s .
- Access to a database $D = \{C_1, \dots, C_m\}$.

Bayesian networks: Learning

- Take any ordering X_1, \dots, X_n .
- For each i , if P_i is the smallest subset of X_1, \dots, X_{i-1} such that $X_i \perp \{X_1, \dots, X_{i-1}\} \setminus P_i \mid P_i$ then $\text{Parents}(X_i) = P_i$.
- The so-obtained DAG
 - only represents true independencies, and
 - no subgraph of it has this property.

Bayesian networks: Learning

The K2 algorithm

- Take any ordering X_1, \dots, X_n .
- For each i
 - $\text{Parents}(X_i) = \emptyset$.
 - Repeat
 - If there exists X_k in $\{X_1, \dots, X_{i-1}\} \setminus \text{Parents}(X_i)$ such that $X_i \perp X_k \mid \text{Parents}(X_i)$ then add it to $\text{Parents}(X_i)$.

Bayesian networks: Learning

Hill-climbing approach

- Start from the empty graph.
 - Repeat
 - Perform the edge addition/removal that improves the score the most*.
- * Check for cycles + store scores to avoid recomputing.

Bayesian networks: Learning

- $D = \{C_1, \dots, C_m\}$ is a database of m cases.
- $p(B_S|D) = p(B_S, D) / p(D)$
 $\alpha p(B_S, D) = p(B_S)p(D|B_S)$
 $= P(B_S) \int_{B_P} \left[\prod_{h=1}^m P(C_h | B_S, B_P) \right] f(B_P | B_S) dB_P$

11/14/2008 Jose M. Peña @ KI 19

Bayesian networks: Learning

The K2 score

Theorem 1. Let Z be a set of n discrete variables, where a variable x_i in Z has r_i possible value assignments: $(v_{i1}, \dots, v_{ir_i})$. Let D be a database of m cases, where each case contains a value assignment for each variable in Z . Let B_S denote a belief-network structure containing just the variables in Z . Each variable x_i in B_S has a set of parents, which we represent with a list of variables π_i . Let w_{ij} denote the j th unique instantiation of π_i relative to D . Suppose there are q_i such unique instantiations of π_i . Define N_{ijk} to be the number of cases in D in which variable x_i has the value v_{ik} and π_i is instantiated as w_{ij} . Let

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$

Suppose the following assumptions hold:

1. The variables in Z are discrete
2. Cases occur independently, given a belief-network model
3. There are no cases that have variables with missing values
4. Before observing D , we are indifferent regarding which numerical probabilities to assign to the belief network with structure B_S .

From these four assumptions, it follows that

$$P(B_S, D) = P(B_S) \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

Consistent, decomposable, non-equivalent

11/14/2008 20

Bayesian networks: Learning

The BDeu score

Consistent, decomposable, equivalent

THEOREM 4 (BDE METRIC) Given domain U , suppose that $p(\Theta_U | B_S^0, \xi)$ is Dirichlet with equivalent sample size N' for some complete network structure B_S^0 in U . Then, for any network structure B_S in U , Assumptions 1 through 3 and 5 through 7 imply

$$p(D, B_S^0 | \xi) = p(B_S^0 | \xi) \cdot \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}$$

- Use uninformative assignment $N'_{ijk} = N' / (r_i \cdot q_i)$
- Note that BDeu = K2 if $N'_{ijk} = 1$
- $E[p(X_i=k | Parents(X_i)=j)] = (N'_{ijk} + N_{ijk}) / (N'_{ij} + N_{ij})$.

11/14/2008 Jose M. Peña @ KI 21

Bayesian networks: Learning

The BIC/MDL score

Consistent, decomposable, equivalent

$$BIC(B_S, D) = \max_{B_P} p(D | B_S, B_P) - 0.5 \cdot \log m \cdot \sum_i q_i (r_i - 1)$$

where $\max_{B_P} p(D | B_S, B_P)$ is reached when

$$p(X_i=k | Parents(X_i)=j) = N_{ijk} / N_{ij}$$

Best parameter values for B_S !!!

11/14/2008 Jose M. Peña @ KI 22

Bayesian networks: Learning

Figure 2: Hasse diagram of the space of Markov equivalence classes of Bayesian network structures over three variables.

23

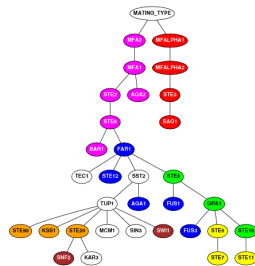
Bayesian networks: Example

- Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2002). Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models. In *Pacific Symposium on Biocomputing*, 437-449.
- **33 genes** involved in the budding yeast pheromone response.
- **320 samples** of the expression levels of the 33 genes under different conditions.
- Gene expression levels discretized into 4 states.

11/14/2008 Jose M. Peña @ KI 24

Bayesian networks: Example

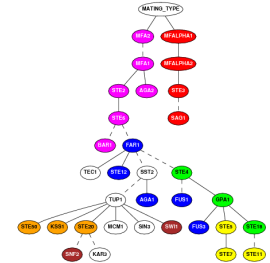
Group	Description
Magenta	Genes expressed only in MAT α cells
Red	Genes expressed only in MAT α cells
Blue	Genes with promoters bound by Ste12
Green	Genes coding for components of the heterotrimeric G-protein complex
Yellow	Genes coding for core components of the signaling cascade
Orange	Genes coding for auxiliary components of the signaling cascade
Brown	Genes coding for components of the SWI-SNF complex
White	Others



Left: Functional grouping of the genes. Right: Best local optimum.

Bayesian networks: Example

t	$k = 0.6$		$k = 0.8$		$k = 0.9$	
	FPS	FNS	FPS	FNS	FPS	FNS
1.00	0	30	0	25	0	22
0.95	0	22	0	15	0	12
0.90	0	17	0	11	0	10
0.85	0	12	0	8	0	7
0.80	0	11	0	6	0	3
0.75	0	8	0	2	0	1
0.70	0	5	0	1	0	1
0.65	0	2	0	1	0	1
0.60	0	1	0	0	0	0
0.55	0	1	0	0	0	0
0.50	0	0	0	0	0	0
0.45	0	0	0	0	0	0
0.40	0	0	0	0	0	0
0.35	0	0	0	0	0	0
0.30	1	0	0	0	0	0
0.25	6	0	0	0	0	0
0.20	9	0	4	0	2	0
0.15	11	0	7	0	6	0
0.10	17	0	11	0	10	0
0.05	25	0	18	0	14	0



Left: Trade-off between the number of FPs and FNs for undirected edges. Right: Dashed edges correspond to TPs at $t = 0.60$ (0 FPs, 0 FNs, 32 TPs) and solid edges at $t = 0.90$ (0 FPs, 11 FNs, 21 TPs).

Dynamic Bayesian networks

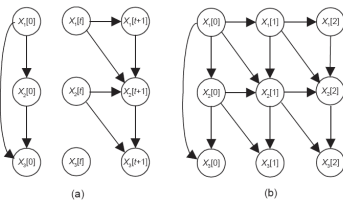


Figure 5.25. Prior and transition Bayesian networks are in (a). The resultant dynamic Bayesian network for $T = 2$ is in (b). Note that the probability distributions are not shown.

11/14/2008

Jose M. Peña @ KI

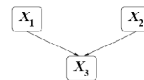
27

Gaussian networks

$$f(x | \theta, s^0) = f(x_1, \dots, x_n | \theta, s^0) = \prod_{i=1}^n f(x_i | \mu(s_i), \theta_i, s^0)$$

$$f(x_i | \mu(s_i), \theta_i, s^0) \sim \mathcal{N}(x_i; m_i + \sum_{X_k \in \text{Pa}(s_i)} b_{ik}(x_k - m_k), v_i)$$

• Model structure



• Model parameters and local probability density functions

$$\begin{aligned} \theta_1 &= (m_1, v_1) & f(x_1 | \theta_1, s^0) &\sim \mathcal{N}(x_1; m_1, v_1) \\ \theta_2 &= (m_2, b_{21}, v_2) & f(x_2 | \theta_2, s^0) &\sim \mathcal{N}(x_2; m_2, v_2) \\ \theta_3 &= (m_3, b_{31}, b_{32}, v_3) & f(x_3 | x_1, x_2, \theta_3, s^0) &\sim \mathcal{N}(x_3; m_3 + b_{31}(x_1 - m_1) + b_{32}(x_2 - m_2), v_3) \\ \theta_4 &= (b_{41}, b_{42}) \end{aligned}$$

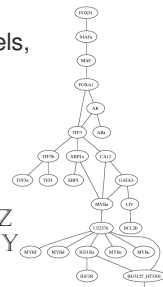
11/14/2008

Jose M. Peña @ KI

28

Markov networks: Semantics

- A.k.a Gaussian graphical models, covariance selection models, Markov random fields, ...
- Based on undirected graphs.
- $p(X_1, \dots, X_n) = (\prod_{C_i} q(C_i)) / Z$.
- u-separation criterion:
 X is **independent** of Y **given** Z if all the paths between X and Y are blocked by Z .



11/14/2008

Jose M. Peña @ KI

29

Markov networks: Learning

- Start from the empty graph.
- Repeat
 - If $X_i \perp X_k | \text{Rest}$ then remove the edge between X_i and X_k .
- Assuming $X = (X_1, \dots, X_n) \sim \mathcal{N}(\mu, \Sigma)$.
- $\mu = E[X]$ and $\Sigma = \text{cov}(X, X) = E[(X - \mu)(X - \mu)^T]$.
- $X_i \perp X_k | \text{Rest}$ if and only if $\Sigma^{-1}[i, k] = 0$.

11/14/2008

Jose M. Peña @ KI

30

Why Bayesian networks ?

- Solidly founded on probability theory.
- Can cope with noise and probabilistic relations.
- Graphical interface.
- Learnable from data and prior knowledge.
- Offer flexible reasoning.
- Accept both causal and acausal interpretation.
- Model both linear and non-linear interactions.
- **Too many samples** are required for accuracy.
- **Scalable** to thousands of genes ?
- Gaussian networks limited to **linear** interactions.
- Learning Markov networks is **not** so well studied.

11/14/2008

Jose M. Peña @ KI

31

More on Bayesian networks

- Continuous random variables, e.g. Gaussian networks.
- Learning via independence tests, e.g. PC algorithm.
- Learning in the space of equivalent BNs, e.g. GES, KES, PC algorithm.
- Model averaging, e.g. MCMC.
- Learning from interventional data.
- Causal Bayesian networks.

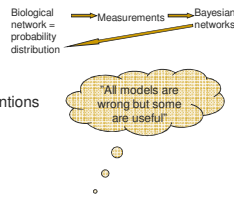
11/14/2008

Jose M. Peña @ KI

32

Causal Bayesian networks

- Causal Bayesian networks
- Interventions
- Learning from observations
- Learning from observations and interventions
- Example 1
- Example 2



11/14/2008

Jose M. Peña @ KI

33

Causal Bayesian networks

- Causal BN = BN + **causal** interpretation, i.e. Parents(X) are the **direct causes** of X.
- A causal BN is a BN and, thus,
 - the probability distribution factorizes accordingly,
 - d-separation applies,
 - it allows probabilistic inference, and
 - it is learnable from data.
- Causal BNs enables us to predict the effect of **interventions**, e.g. the effect of a drug. This is not possible with acausal BNs.

11/14/2008

Jose M. Peña @ KI

34

Interventions

- A BN can tell us how the distribution of X changes when **observing** Y, i.e. $p(X|Y=y)$.
- In addition to this, a causal BN can tell us how the distribution of X changes when **intervening** on Y, i.e. $p(X|\text{do}(Y=y))$.
- Check
 - $p(\text{Cancer}=\text{yes}|\text{WhiteTeeth}=\text{no})$
 - $p(\text{Cancer}=\text{yes}|\text{do}(\text{WhiteTeeth}=\text{no}))$



11/14/2008

Jose M. Peña @ KI

35

Interventions

- A causal BN enables us to predict the effect of an intervention on X_1 as
 - $p(X_2, \dots, X_n | \text{do}(X_1=x_1)) = \prod_{i \neq 1} p(X_i | \text{Parents}(X_i))$,
or as
 - delete the edges from Parents(X_1) to X_1 and, then, "observe" $X_1=x_1$.
- This is not possible with BNs.



11/14/2008

Jose M. Peña @ KI

36

Interventions

- Which nodes get affected by an intervention on X ? Use d-separation.
Answer: Only the descendants of X and, thus, $p(Y|do(X=x)) = p(Y)$ if Y is not one of them.
- Predicting the effect of an intervention that rewrites the causal BN
 - $p(X_2, \dots, X_n | do(Parents(X_1) = NewParents(X_1))) = p(X_1 | NewParents(X_1)) \prod_{i \neq 1} p(X_i | Parents(X_i))$
or as
 - rewire the causal BN accordingly.

11/14/2008 Jose M. Peña @ KI 37

Learning from observations

IC algorithm (≈PC algorithm)

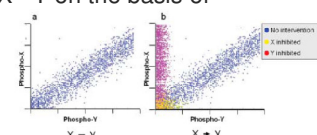
- Start from the complete graph. Try small sets first
- For each pair of nodes X_i and X_k
 - If $X_i \perp X_k | S_{jk}$ then remove the edge between X_i and X_k .
- For each induced subgraph $X_i - Y - X_k$
 - If Y is not in S_{jk} then $X_i \rightarrow Y \leftarrow X_k$. There exist rules for fulfilling this step
- Orient as many lines as possible. Output is not a BN but a class of equivalent BNs !!!

Alternative: Learn a BN first, then keep only compelled edges

11/14/2008 Jose M. Peña @ KI 38

Learning from observations and interventions

- Choose between the equivalent BNs $X \rightarrow Y$ and $X \leftarrow Y$ on the basis of


- $p(Y|do(X=x)) = p(Y|X=x) \neq p(Y)$
- $p(X|do(Y=y)) = p(X) \neq p(X|Y=y)$, so $X \rightarrow Y$.

11/14/2008 Jose M. Peña @ KI 39

Learning from observations and interventions

- As before

$$p(D, B_{ij}^k | \xi) = p(B_{ij}^k | \xi) \cdot \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}$$

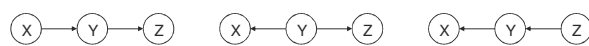
but now N_{ijk} is the number of cases in D in which X_i is **observed** in state k and its parents **are** in state j.

May be due to intervention As opposed to intervened or manipulated or perturbed !!!
- $E[p(X_i=k | Parents(X_i)=j)] = (N'_{ijk} + N_{ijk}) / (N'_{ij} + N_{ij})$.
- Similar for BIC/MDL.

11/14/2008 Jose M. Peña @ KI 40

Learning from observations and interventions

- Even with interventions ambiguity may remain.
- Two causal BNs are causally equivalent wrt $do(X=x)$ if they are acausally equivalent before and after $do(X=x)$.

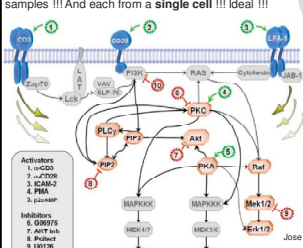

- E.g., the 2nd. and 3rd. are causally equivalent.
- In other words, two causal BNs are causally equivalent wrt $do(X=x)$ iff they are acausally equivalent and X has the same parents in both BNs.

11/14/2008 Jose M. Peña @ KI 41

Example 1

Sachs, K. et al.: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science 308 (2005) 523-529.

9 perturbations * 600 samples/perturbation = 5400 samples !!! And each from a single cell !!! Ideal !!!



Activators
1. cAMP
2. cAMP
3. cAMP
4. PKA
5. cAMP

Inhibitors
6. GTP/GDP
7. Akt
8. Phospho-ERK
9. MEK
10. L2726002

1. Perturbations

2. Multiparameter

3. Correlated phospho-measures per cell

4. Datasets of cells

5. Influence diagram of measured variables

6. Bayesian network analysis

Due to acyclicity? Dynamic BNs?

A Model inference result

Expected	15/17
Reported	17/17
Reversed	1
Missing	3

11/14/2008 Jose M. Peña @ KI

Example 1

11/14/2008 Jose M. Peña @ KI 43

Example 1

	A	B	C	Complete Dataset
Number of Edges	1710	1714	619	1457
Number of Nodes	119	114	116	217
Number of Interventions	NA	2	3	1
Number of Unintegrated Nodes	2	9	8	1
Number of Missing Edges	11	10	12	4

11/14/2008 Jose M. Peña @ KI 44

- Importance of interventions
- Importance of large datasets
- Importance of single cells
- Due to acyclicity? Dynamic BNs?
- May fit well the data, but doesn't reveal causality. Importance of interventions.

Example 2

Poumarat, I. and Wernisch, L.: Reconstruction of Gene Networks Using Bayesian Learning and Manipulation Experiments. *Bioinformatics* 20 (2004) 2934-2942.

Divide the equivalence class in many small subclasses

input: observational data D with N_s samples, limit on number of experiments, further experimental data as requested
 output: sequences of variables to manipulate

while limit not reached and variables not yet manipulated exist do
 learn TS-equivalence classes from D
 keep K classes with highest probability for each variable a not yet manipulated do
 $V_s \leftarrow$ expected loss $L_s(a, D)$ of manipulation a given current data D
 select a with L_s minimum
 output a
 $D_s \leftarrow N_s$ new samples after manipulating a
 $D \leftarrow D \cup D_s$ // update data

11/14/2008 Jose M. Peña @ KI

Bibliography

- Articles
 - Pe'er, D.: Bayesian Network Analysis of Signaling Networks: A Primer. *Science STKE* 281 (2005).
 - 1996 lecture by Judea Pearl (www.cs.ucla.edu/~judea).
 - <http://genomics.princeton.edu/~florian/docs/network-bib.pdf>
- Books
 - Pearl (1988, 2000), Castillo et al. (1997), Neapolitan (2003), Lauritzen (1999), Jensen (1996, 2000), ...
 - Articles, e.g. UAI.
- Software
 - www.hugin.dk
 - <http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html>

11/14/2008 Jose M. Peña @ KI 46