A Dynamic Bayesian Network Based Structural Learning towards Automated Handwritten Digit Recognition

Olivier Pauplin and Jianmin Jiang

Digital Media & Systems Research Institute
University of Bradford, United Kingdom

Email: o.pauplin@bradford.ac.uk

HAIS 2010, 23-25 June

San Sebastian, Spain

Outline

- Background on probabilities
- Introduction to Static and Dynamic Bayesian Networks
- Machine Learning with DBNs
  - Parameter learning
  - Structure learning
- Models for handwritten digit recognition
- Results

Background

- A, B: random variables
- Prior probability of A: $P(A)$
- Joint probability of A and B: $P(A,B)$
- Posterior probability (or conditional probability): $P(A|B)$

(conditional probability that event A occurs given that event B has occurred)

- Bayes' rule: $P(A,B) = P(A|B).P(B) = P(B|A).P(A)$

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

- Extension to n random variables:

$$\begin{aligned} P(X_1,\ldots,X_n) &= P(X_n \mid X_{n-1},\ldots,X_1).P(X_{n-1},\ldots,X_1) \\ &= P(X_1).\prod_{i=2}^{n} P(X_i \mid X_{i-1},\ldots,X_1) \end{aligned}$$

BNs and DBNs

Bayesian networks (BNs) allow:

- Efficient representation of uncertain knowledge

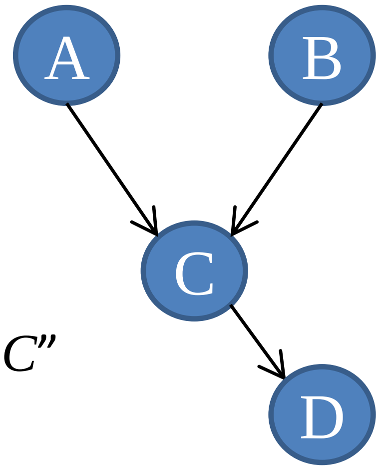
  BNs represent the dependencies among variables and give a concise specification of any full joint probability distribution.

- Learning from experience

A simple BN:

Nodes of the graph = random variables

Arrows between nodes link "parents" of X_i to X_i
(In this example, A and B are the parents of C)



- An arrow between A and C means: "A has a direct influence on C"

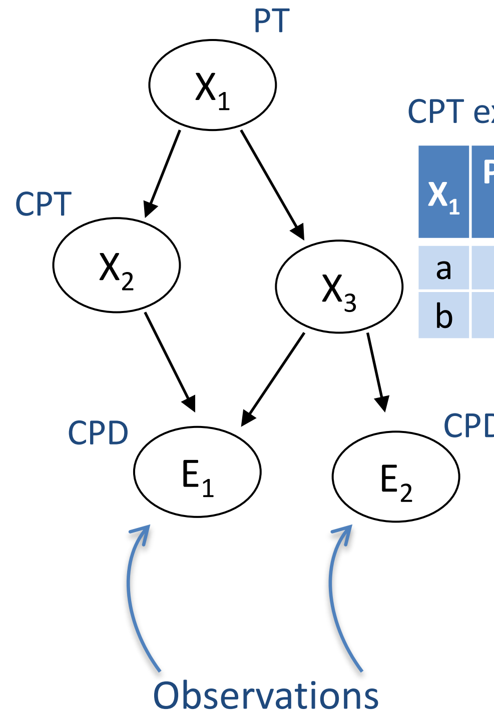
- The effect on a node of its parents is quantified by:

- The graph has no directed cycles (DAG: Directed Acyclic Graph)

- Full joint probability of a BN: $P(X_1,\ldots,X_n)=\prod_{i=1}^{n} P(X_i \mid parents(X_i))$

- The full specification of a BN requires:
  → A topology (nodes, arrows);
  → For each node, a conditional probability table (discrete node) or a conditional probability distribution (continuous node) $P(Xi \mid parents(Xi))$ that quantifies the effects of the parents on the node.

  The parameters (CPDs and CPTs) can be obtained from data analysis.

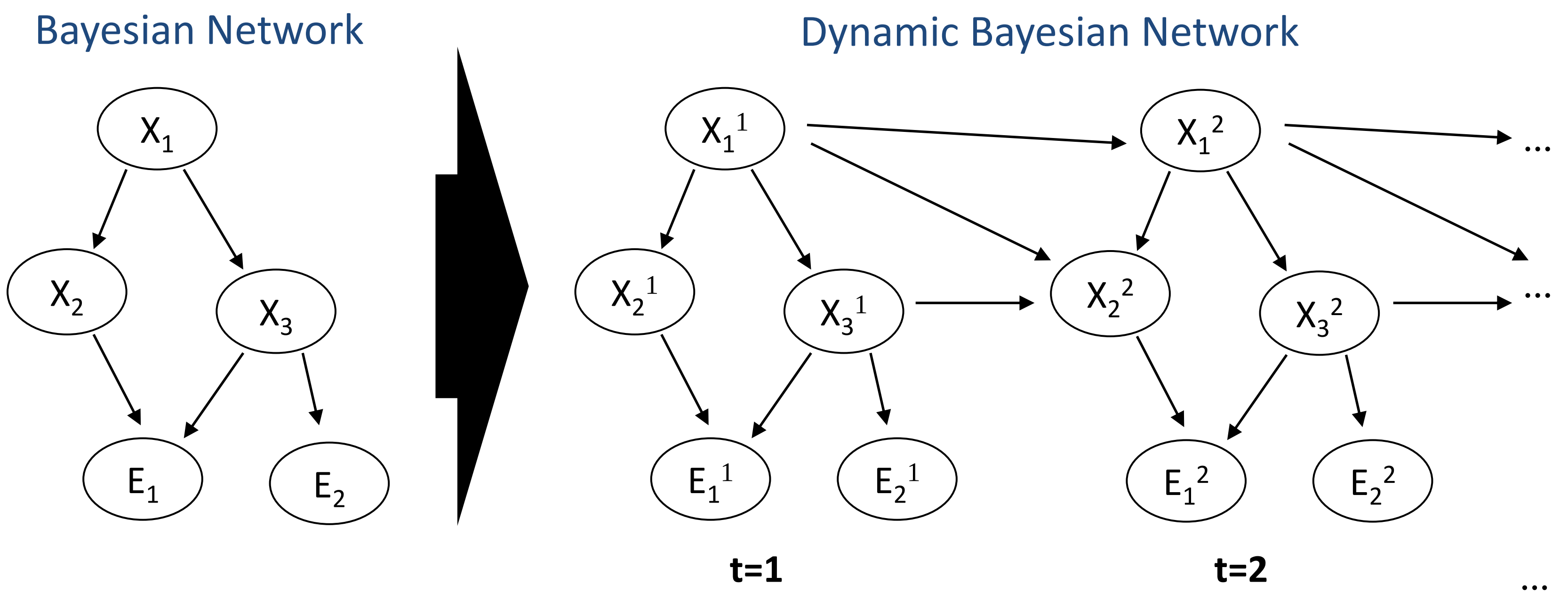


CPT example:

X_1	$P(X_3) = c$	$P(X_3) = d$
a	0.3	0.7
b	0.1	0.9

Dynamic Bayesian Network (DBN):

A temporal extension of Bayesian Networks



- Stationarity: DBNs are time-invariant (parameters are the same for all *t*)
- Markov property: The current state depends on only a finite history of previous states (usually only the previous state)

➔ 2 time slices are enough to describe the whole DBN

Learning with DBNs

Parameter learning

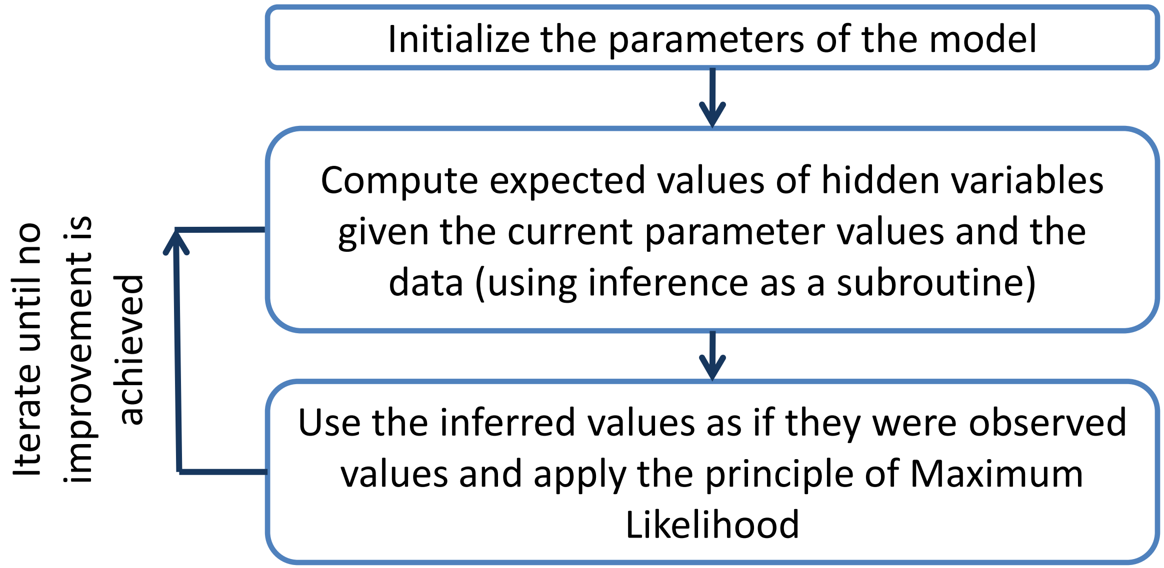
- Principle of Maximum Likelihood (ML)

$\Theta=\{\theta 1,\ldots, \theta m\}$ set of parameters

$D=\{d1,\ldots,dn\}$ data (observations)

$$\max[\,P(D\,|\,\Theta)] \quad ?$$

- In case of incomplete data: Expectation-Maximization algorithm:



Structure learning

- In the general case, with hidden nodes: very computationally intensive
- To overcome that problem: links between consecutive time slices are learnt with the following restrictions:

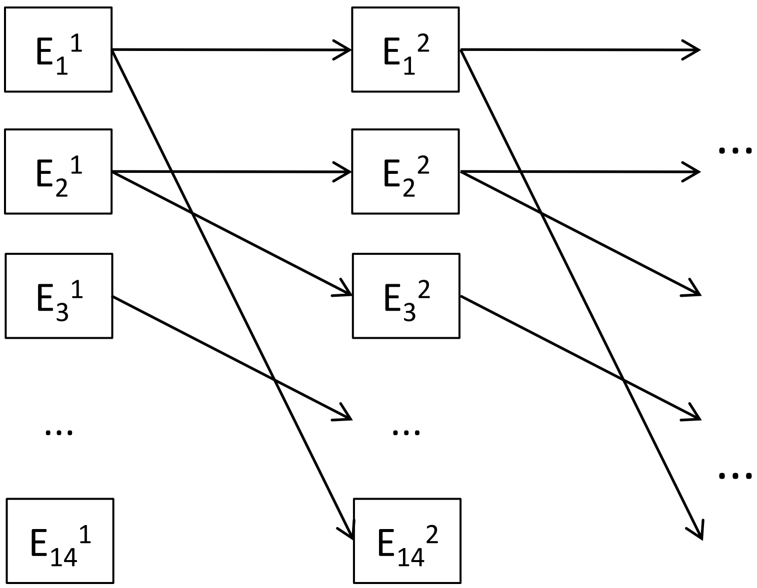
  1- All nodes must be observed (no hidden node)

  2- All nodes must be discrete (data is binarised beforehand)

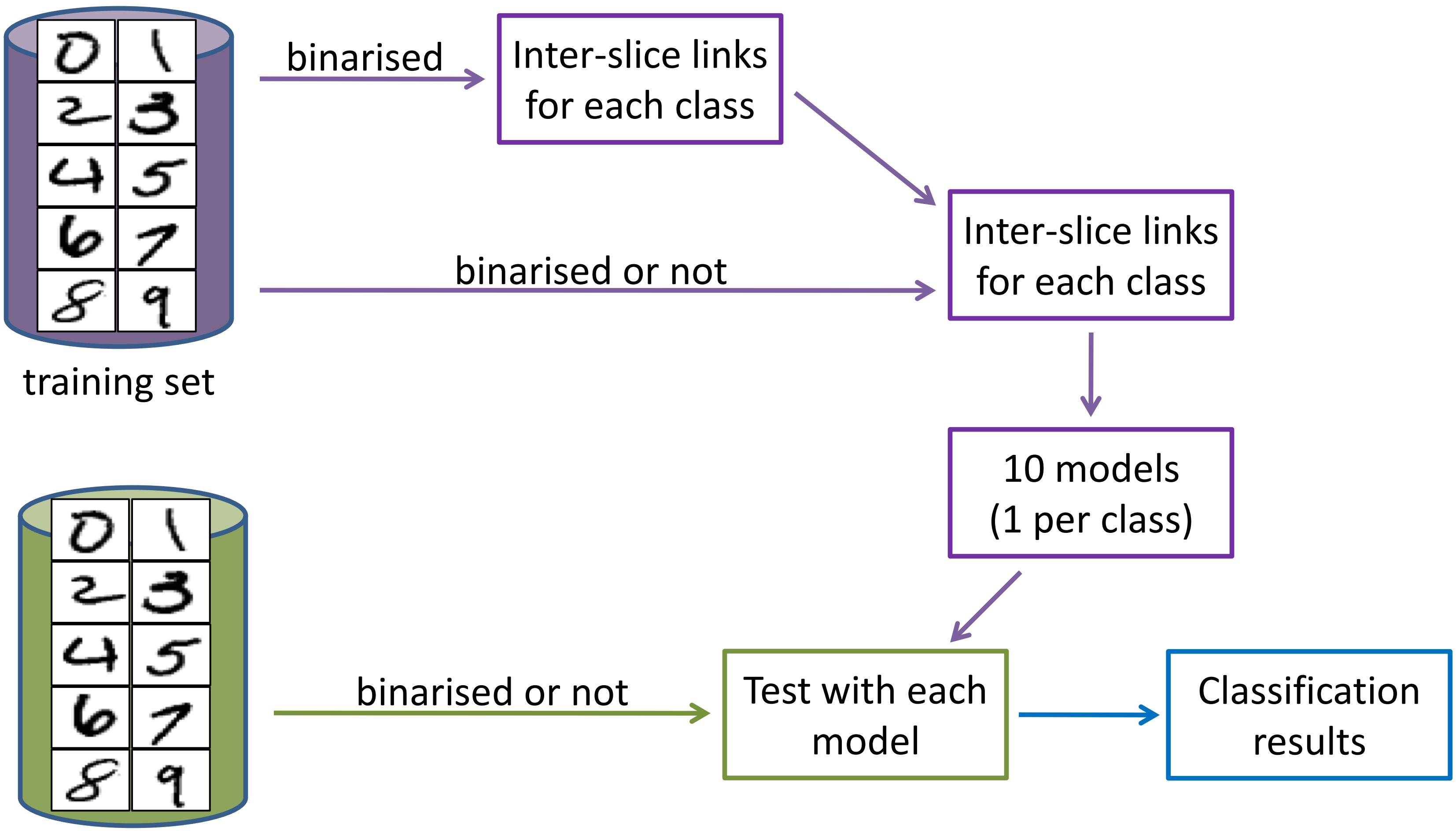
Links maximise the Bayesian Information Criterion (BIC score):

$$\mathrm{BIC} = \log[\mathrm{P}\,(D|G,\hat{\Theta})] - \frac{\log[\,Ns\,]}{2} \times Np$$

- D : data
- $\hat{\Theta}$: set of parameters maximising the likelihood of D
- Ns : number of data sample in D
- Np : dimension of graph G (number of free parameters)

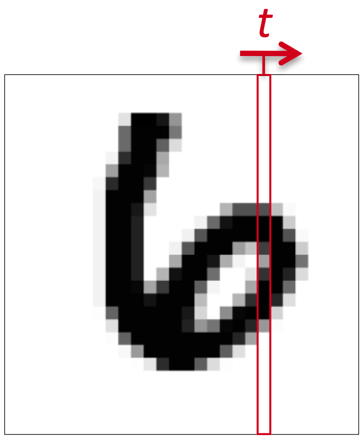


Models for handwritten digit recognition

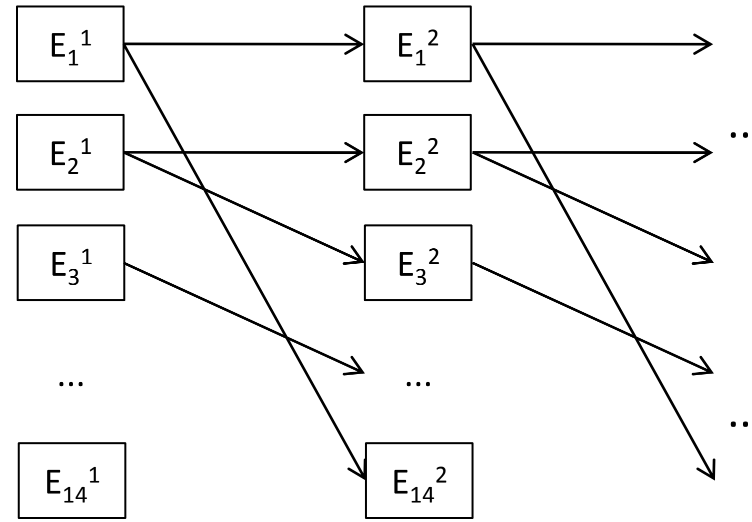


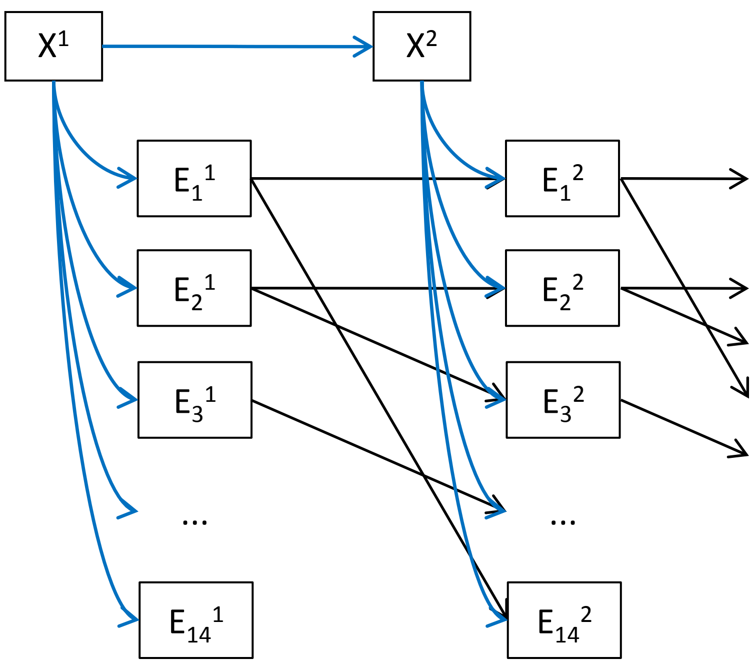
test set

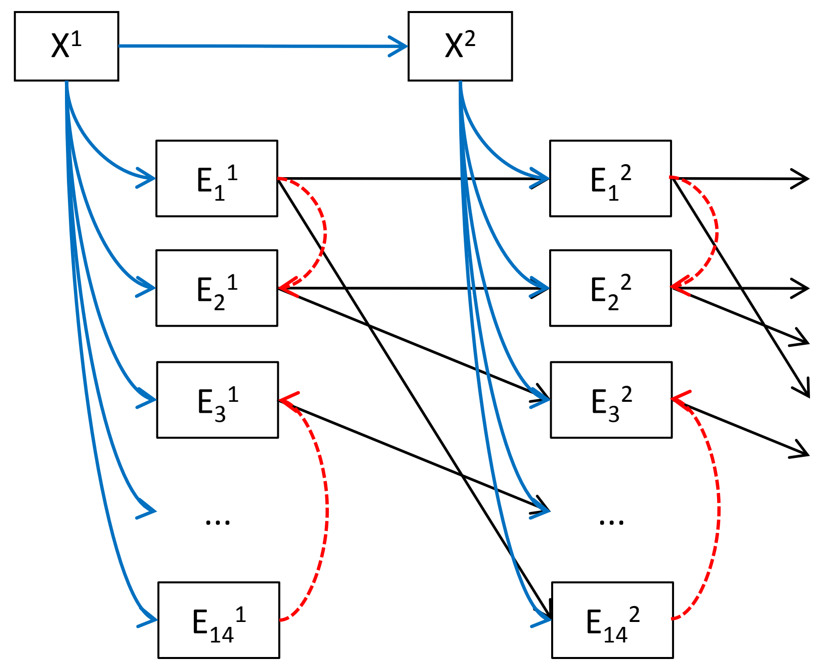
Kinds of models tested:

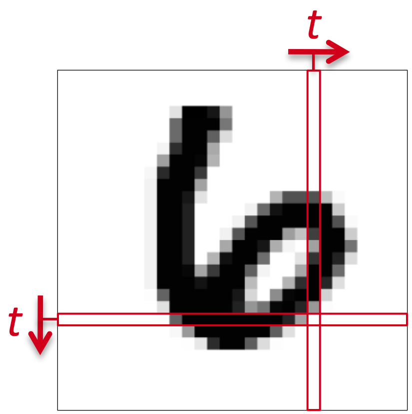


Observations are columns of pixels

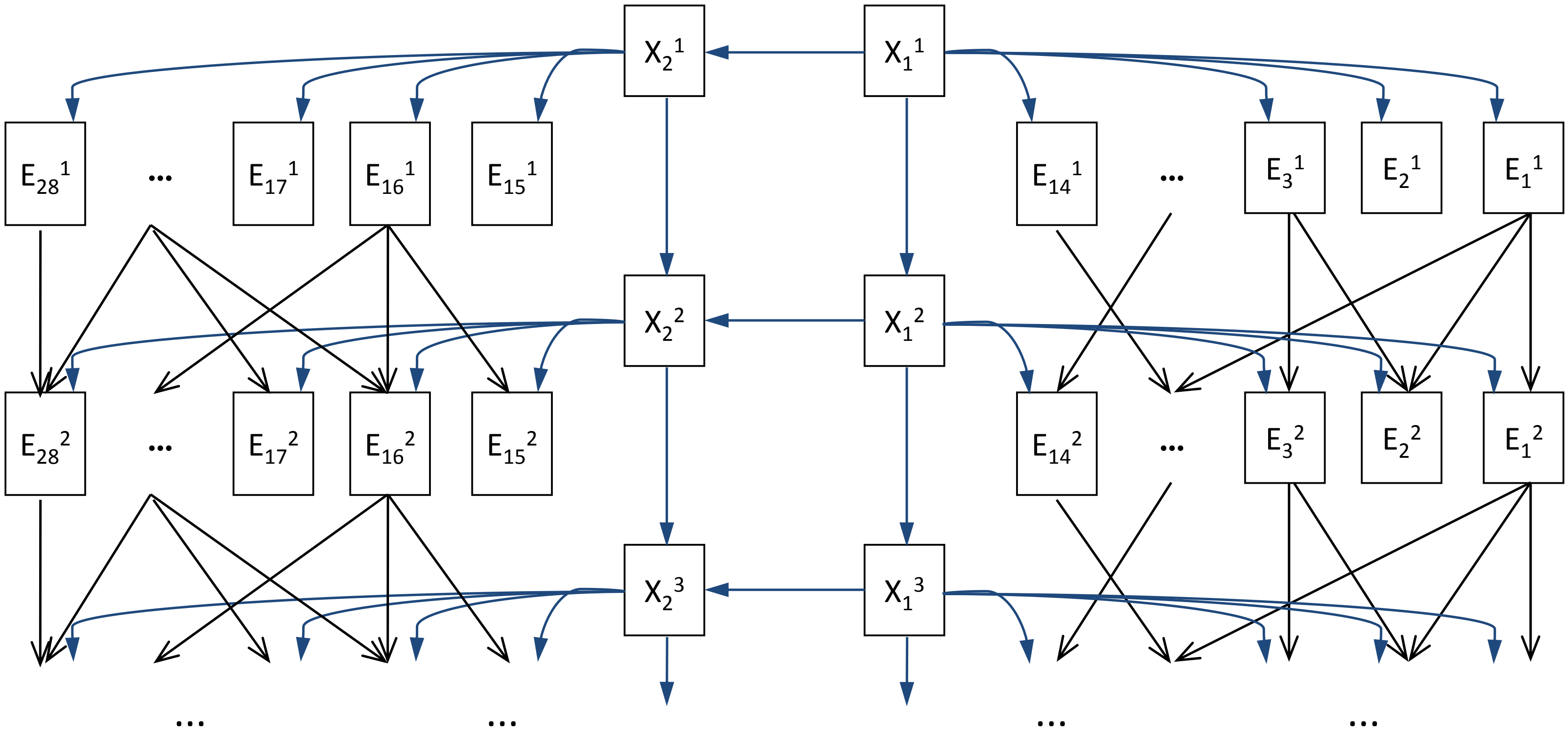




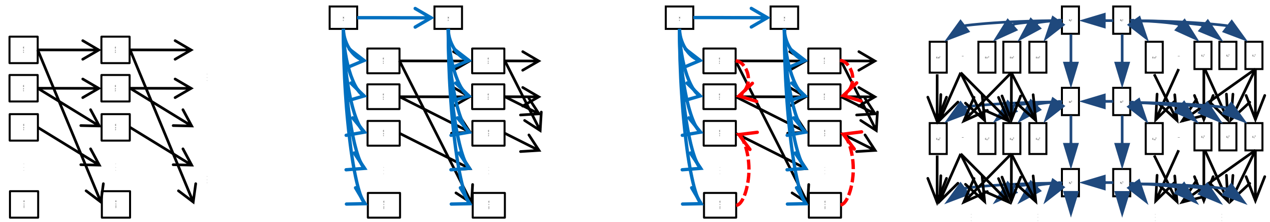




Observations are lines and columns of pixels



Results



	Inter-slice links learnt from the data			
	Observations = columns of pixels			Columns+lines
	No hidden nodes	1 hidden node per time slice	1 hidden node per *t* + learnt intra-slice links	2 hidden nodes per *t*
Discrete evidence nodes	67.7	69.6	71.2	74.8
Gaussian evidence nodes	81.0	90.2	90.6	93.3

Thank you for your attention

Any questions?