Learning Complex Uncertain State Changes via Asymmetric Hidden Markov Models: an Industrial Case

M. Bueno, A. Hommersom, P. Lucas, S. Verwer, A. Linard

mbueno@cs.ru.nl

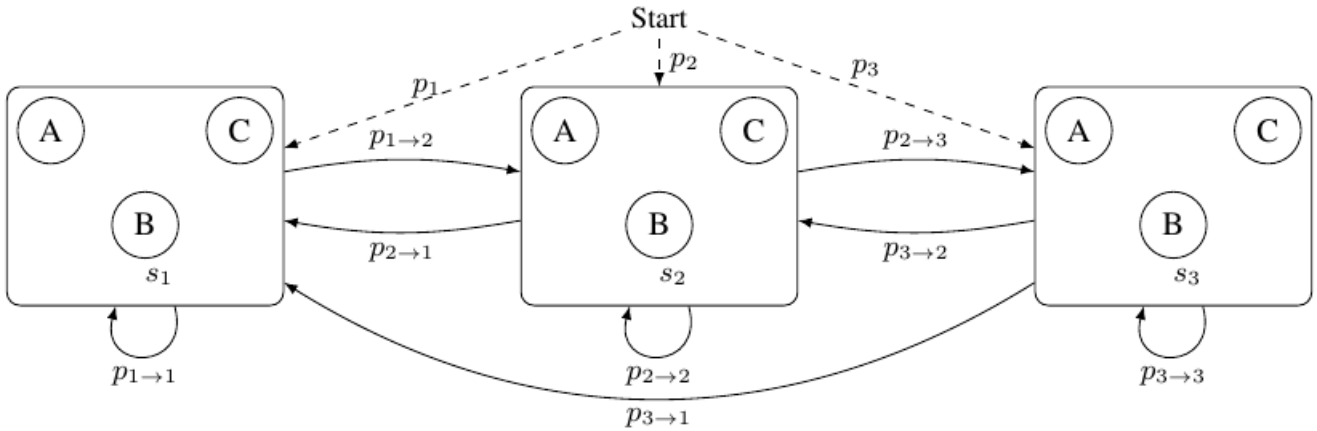
Radboud University Nijmegen, the Netherlands
Open University of the Netherlands
Leiden University, the Netherlands
Delft University of Technology, the Netherlands

PGM 2016
Lugano, Switzerland

Introduction

Hidden Markov Models (HMMs) are intensively used in sequential and temporal problems.

- Setting: Multivariate observations, discrete time.
- Most common HMMs: use the "naive" structure.
- This might require large state spaces for fitting the data.
- *Issues*: typically, scarce data – large state space leads to *overfitting*.
- *Issues*: large data – very little or no practical problem insight.



Introduction

How to deal with this?

- Additional structure: space of **observables** (e.g. trees), **multiple hidden chains** (e.g. Factorial HMMs), **autoregressions**, etc.
- Step further: capturing **asymmetric or contextual independencies** – each hidden state induces a particular structure over observables.

Research to capture asymmetries in HMMs:

- Chow-Liu trees (generative) [Kirshner et al., 2004]
- Multinets (discriminative) [Bilmes, 2000].

Contributions

In this paper, we propose:

- **Asymmetric Hidden Markov Models** to represent state-specific Bayesian-network distributions on observables.
  - Generalize emissions to general BNs.
  - Thus: We extend Chow-Liu trees (generative)
  - (Multinets do **not** capture these interactions.)

We also empirically investigate:

- Importance of modeling structure in HMMs.
- Relationship between state space dimension and amount of data.

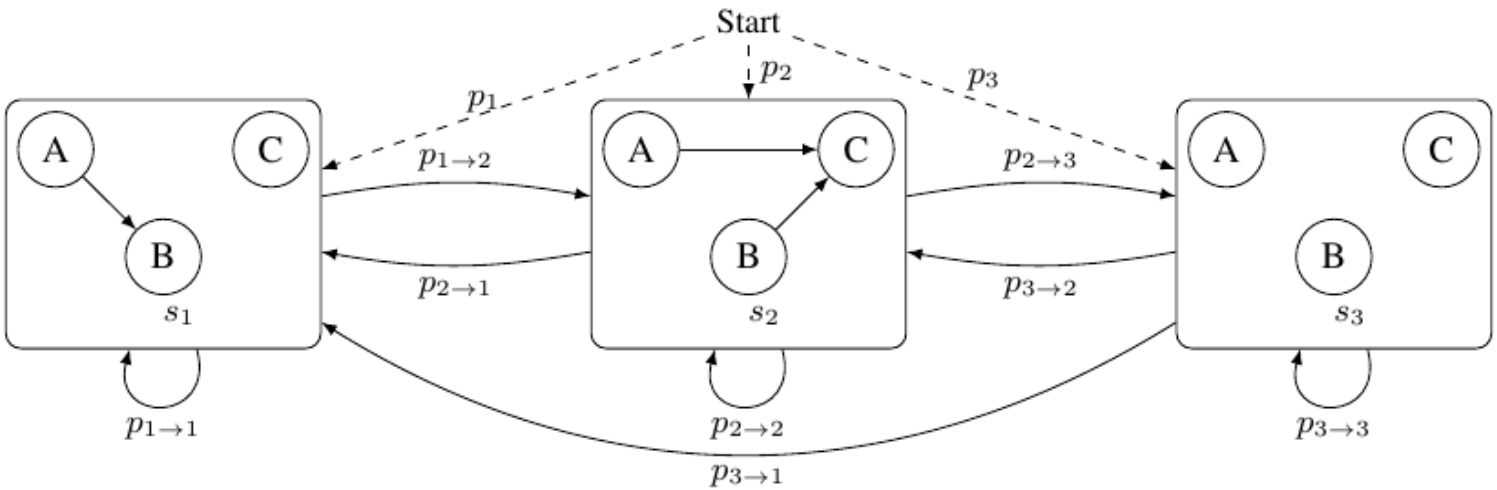
Outline

Model specification

- Observable features $= \mathbf{X}$; State space $= S$
- In HMM-As, we associate each state in S to a BN over $\mathbf{X}$.
- The emissions distributions are defined as:

$$\begin{aligned} P(\mathbf{X} \mid S) &= P_S(\mathbf{X}) \\ &= \prod_{i=1}^{n} P_S(X_i \mid \mathrm{Pa}_S(X_i)) \end{aligned}$$

Example 1



Learning asymmetric HMMs

Learning setting: missing data and *unknown structure*.

- Our procedure is based on structural expectation-maximization.

E-step: compute expected sufficient statistics

- We compute via forward-backward calculations.
- It has the same complexity as indep. HMM: $O(nmTk^3)$, $k = \#\text{states}$

M-step: as opposed to HMM-Is, cannot be solved analytically.

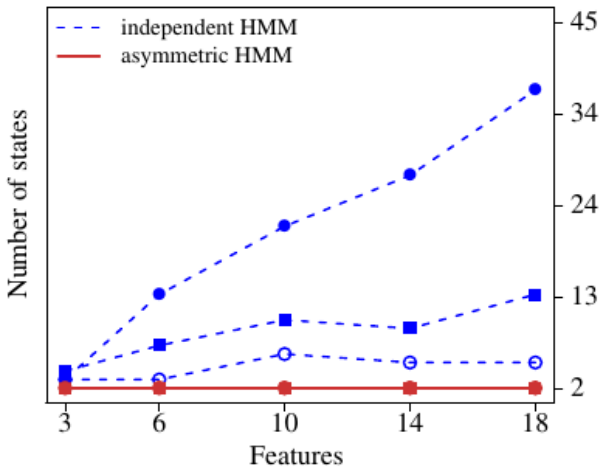
- Structure learning is performed *per state* using expected score:

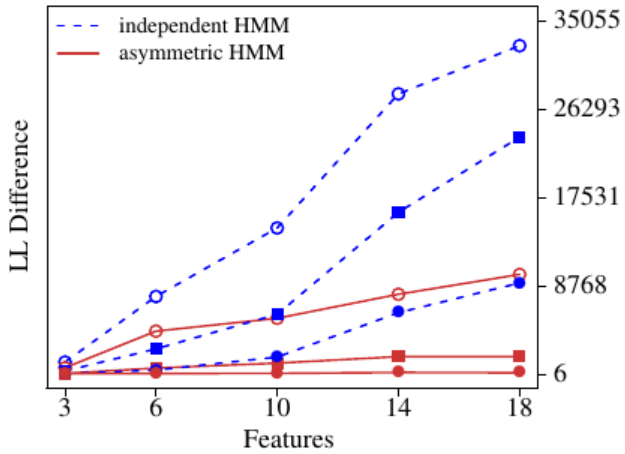
$$E[\log P(\mathcal{D}, S \mid \lambda)] + \text{Pen} = \sum_{i=1}^{k} \Bigg[\underbrace{\sum_{t=0}^{T} \sum_{l=1}^{m} \log P(s_i^{(t)} \mid \mathcal{D}_l) P(\mathcal{D}_l \mid \lambda)}_{\text{Score for each state } i=1,\ldots,k} \Bigg] + \text{Pen}$$

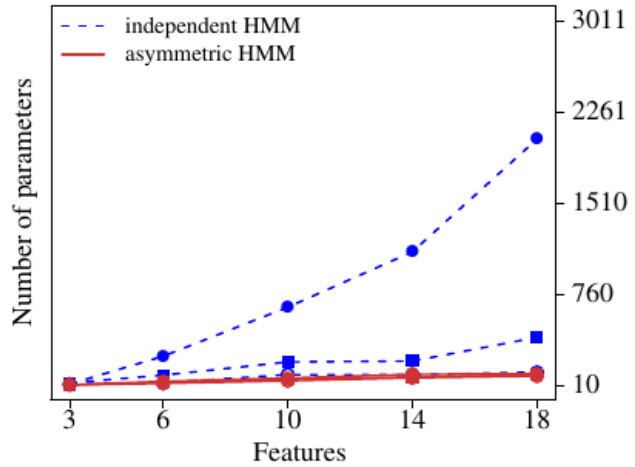
- In practice: reasonable solutions in reasonable time using heuristics.

Simulations

True HMM-A models: #Hidden states = 2, max node degree = 3
Datasets: 50 seqs. (○), 200 seqs. (■), 1,000 seqs. (●)
Evaluation: CV; #Hidden states = $[2, \ldots, k]$ up to overfitting.





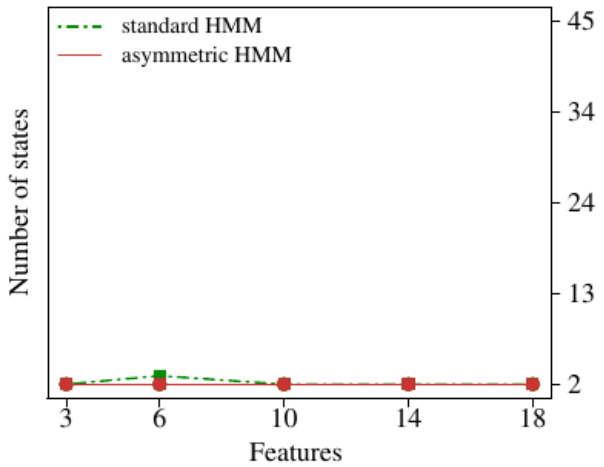


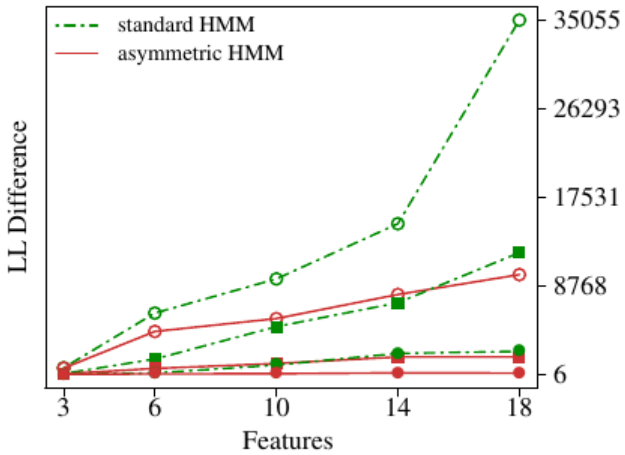
Asymmetric HMMs versus Independent HMMs

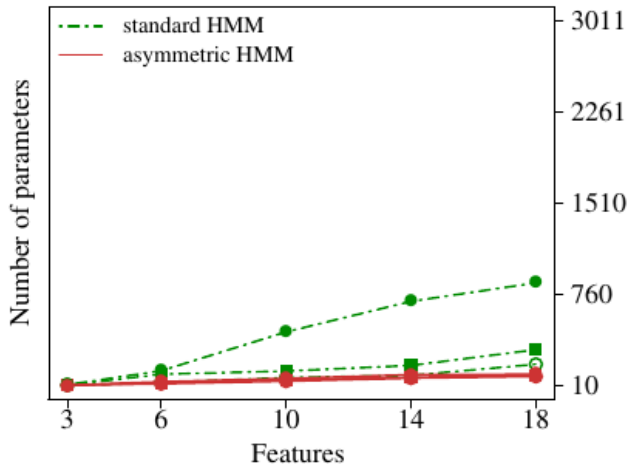
- State spaces of HMMs grow until overfitting is reached.
- Independent HMMs: reaches overfitting with less model quality.
- Asymmetric HMMs: give better fit even when learned with less data.
- Much lower number of parameters in asymmetric HMMs.

Simulations

True HMM-A models: #Hidden states = 2, max node degree = 3
Datasets: 50 seqs. (○), 200 seqs. (■), 1,000 seqs. (●)
Evaluation: CV; #Hidden states = $[2, \ldots, k]$ up to overfitting.





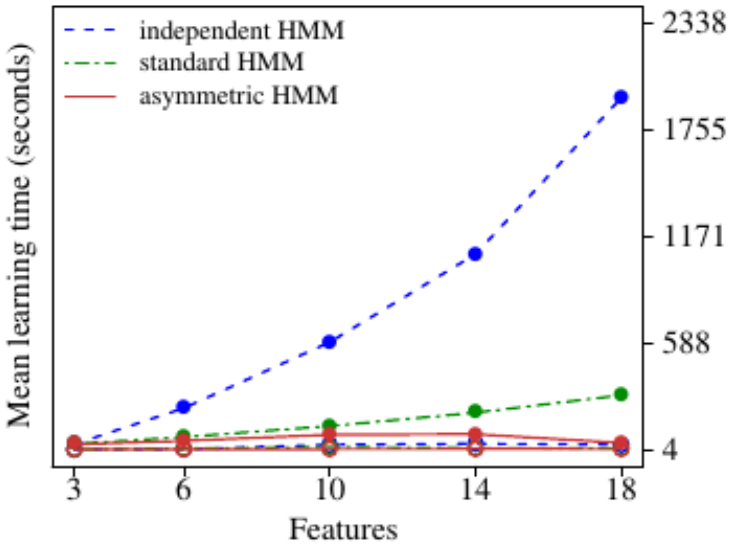


Asymmetric HMMs versus Standard HMMs

- Standard HMMs: overfit with (very) dense structures and less model quality.
- Asymmetric HMMs: give better fit even when learned with less data.
- Still lower number of parameters in asymmetric HMMs.

Simulations - Running time

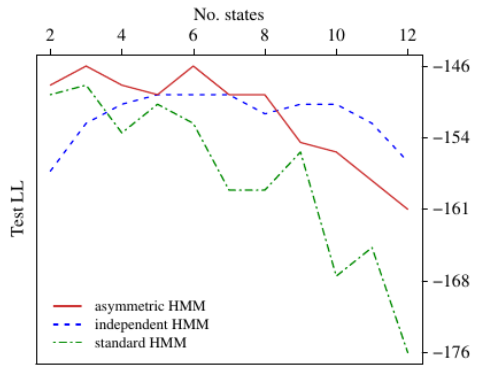
True HMM-A models: #Hidden states = 2, max node degree = 3
Datasets: 50 seqs. (∘), 200 seqs. (■), 1,000 seqs. (•)
Evaluation: CV; #Hidden states = $[2, \ldots, k]$ up to overfitting.

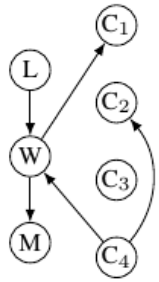


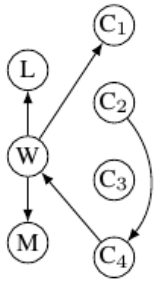
- Empirically: Structural EM was faster than common EM on most cases.
- This is due to the large state space of independent HMMs.

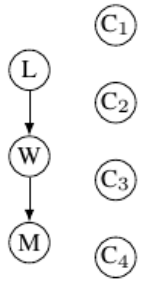
Learning from large-scale printers data

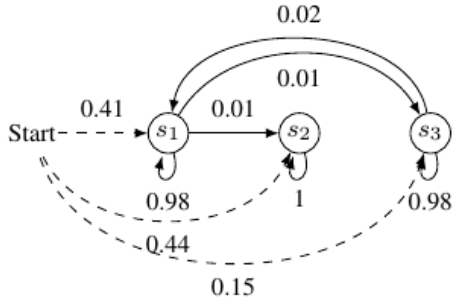
- Discrete sequence data from two large-scale industrial printers.
- **Datasets**: $\mathcal{R}_1$: 27 sequences; $\mathcal{R}_2$: 52 sequences ($l = 15$)
- **Observables**: interval duration (L), total workload (W), maintenance frequency (M), color-related features ($C_{1,2,3,4}$).

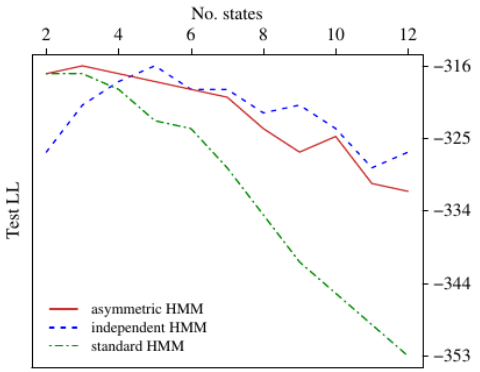


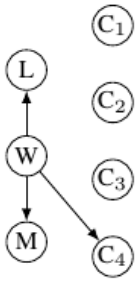


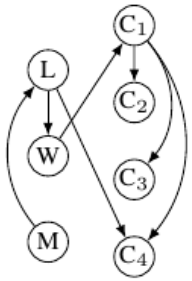


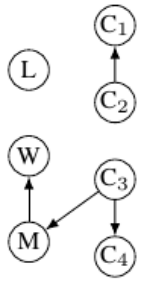


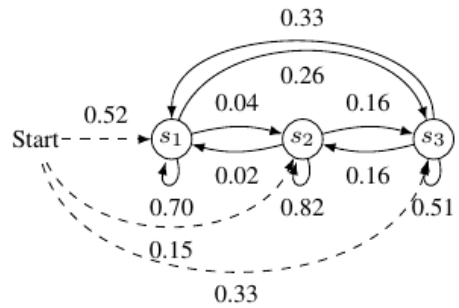












Conclusions

Conclusions:

- We proposed asymmetric HMMs to represent BN distributions on observables space.
- Experiments showed that data with underlying structure is better captured with structured models + asymmetries modeling.
- Asymmetric HMMs are more compact than non-asymmetric HMMs (independent and standard HMMs).
- Asymmetric HMMs were able to better deal with overfitting.

Future work:

- Comparisons with other HMMs that capture asymmetries.
- Possibly extend representation, e.g. modeling autoregressions.
- Other applications.


Bilmes, J. (2000).
Dynamic Bayesian multinets.
In *Proc. of the Sixteenth conference on Uncertainty in Artificial Intelligence*, pages 38–45.


Kirshner, S., Smyth, P., and Robertson, A. (2004).
Conditional Chow-Liu tree structures for modeling discrete-valued vector time series.
In *Proc. of the 20th UAI*, pages 317–324.

Relation to non-asymmetric HMMs

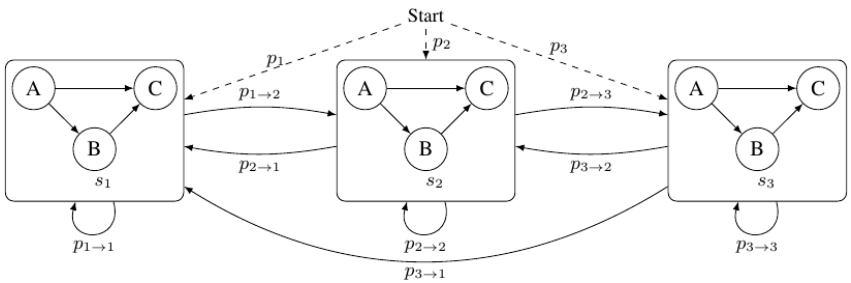
Non-asymmetric HMMs: special cases of HMM-As.

$$P(\mathbf{X} \mid S) = \prod_{i=1}^{n} P_S(X_i \mid \mathrm{Pa}_S(X_i)) \text{ where } \mathrm{Pa}_S \text{ is fixed}$$

Can non-asymmetric models represent HMM-As distributions (over observables)?

Standard HMM
(any structure over observables)

- No additional states needed.
- Equivalence: Merging arcs from states into a single structure.



Independent HMM
(naive structure: Pa $= \{S\}$)

- *Additional states might be needed!* Strong independence assumptions.
- Equivalence: "Determinizing" HMM-As emissions (one option).