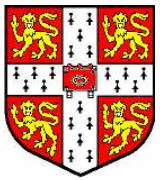
Linear Gaussian models for speech recognition

Antti-Veikko Rosti

November 2002



Cambridge University Engineering Department

SVR Speech Seminar

Overview

- State-space models
- Linear Gaussian models
- Bayesian networks
- Factor analysed HMMs
- Segment models
- Conclusions

State-space models

Generative model of a state-space model:

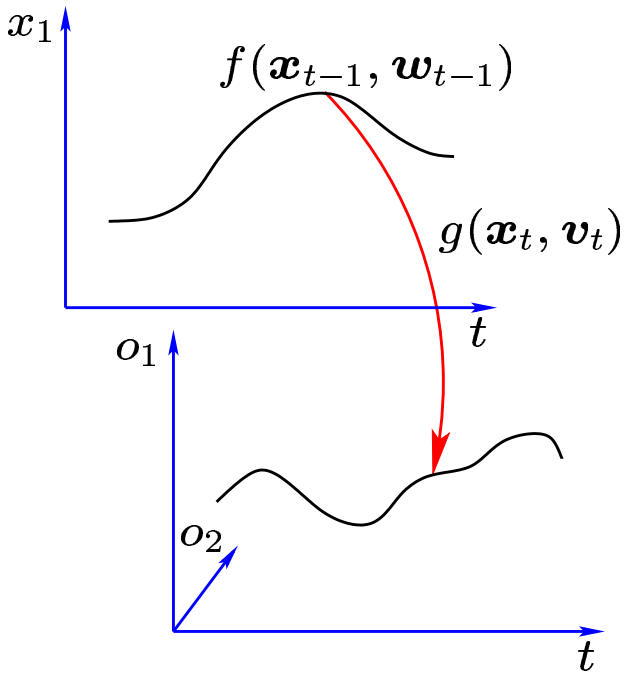
$$\begin{aligned} \boldsymbol{x}_{t+1} &= f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \boldsymbol{w}_t) \\ \boldsymbol{o}_t &= g(\boldsymbol{x}_t, \boldsymbol{v}_t) \end{aligned}$$

- k-dimensional state vectors $\boldsymbol{x}_t$;
- p-dimensional observation vectors $\boldsymbol{o}_t$;
- state evolution noise $\boldsymbol{w}_t$;
- observation noise $\boldsymbol{v}_t$.

Model for speech production

- articulator positions in state vectors;
- acoustic realisations in observation vectors;
- mappings $f(\cdot)$ and $g(\cdot)$ non-linear!

Linear Gaussian models



State-space models with

- linear mappings $f(\cdot)$ and $g(\cdot)$;
- Gaussian distributed noise $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$;
- usually Markov in state-space; i.e., $\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{w}_t)$;
- not a good model for speech production!

Interpretation:

- state vector components are covariance model factors;
- observations are linear combinations of state vectors and noise.

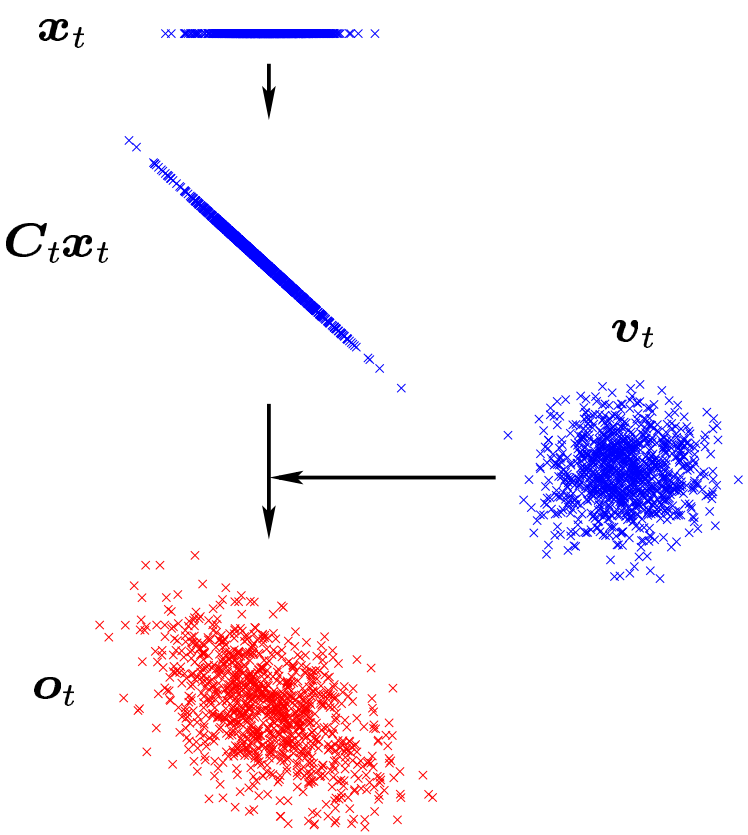
What about the mappings...

Observation process

Factor analysis (FA)

$$\begin{aligned} \boldsymbol{o}_t &= \boldsymbol{C}_t \boldsymbol{x}_t + \boldsymbol{v}_t \\ \boldsymbol{v}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}_t^{(o)}, \boldsymbol{\Sigma}_t^{(o)}\right) \end{aligned}$$

- $\boldsymbol{x}_t$ are $\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right)$ distributed factors;
- $\boldsymbol{C}_t$ are matrices of factor loadings;
- $\boldsymbol{v}_t$ are errors with diagonal $\boldsymbol{\Sigma}_t^{(o)}$;
- observation covariance $\boldsymbol{C}_t \boldsymbol{C}_t' + \boldsymbol{\Sigma}_t^{(o)}$;
- fewer parameters than full covariance if $k < (p-1)/2$.



Extensions: SFA, IFA, ...

Linear discriminant analysis

Generative model:

$$\begin{aligned} \boldsymbol{o}_t &= \boldsymbol{C}_t \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{v}_t \end{bmatrix} \\ \boldsymbol{v}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}_t^{(o)}, \boldsymbol{\Sigma}_t^{(o)}\right) \end{aligned}$$

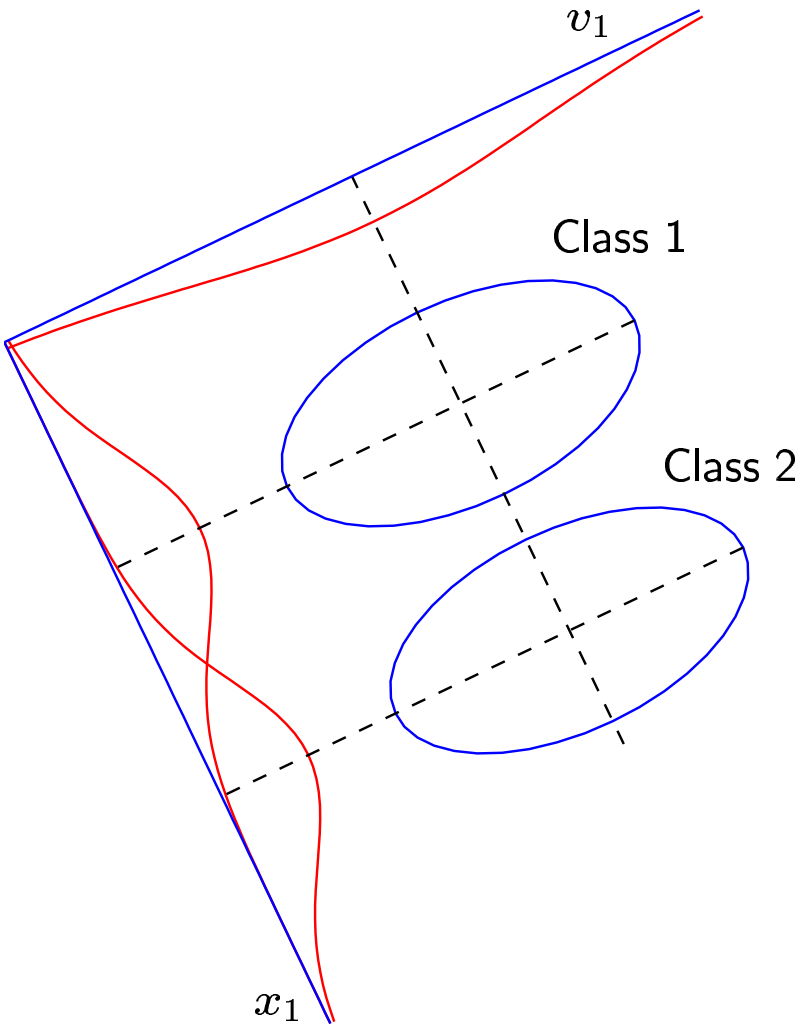
Projection scheme which

- minimises within class variance;
- maximises between class variance.

Example:

- two classes of 2-d observations;
- project down to optimal axis, x_1.

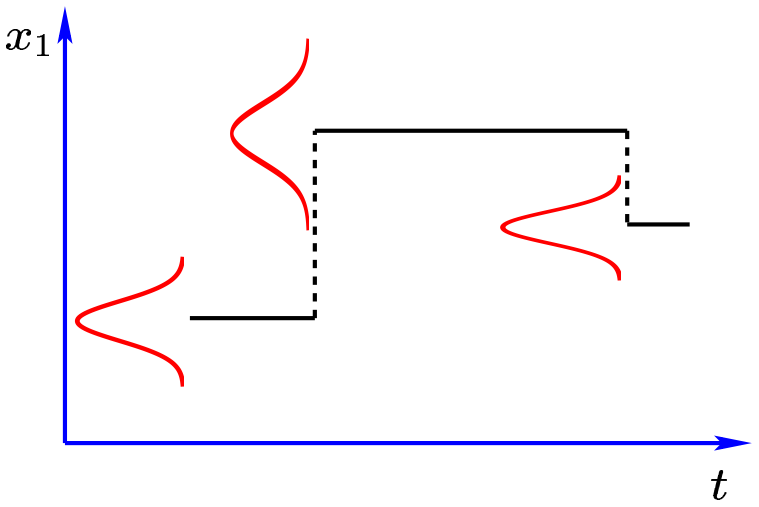
Extensions: HLDA, MLDA, HDA, ...



State evolution process

Hidden Markov model

$$\begin{aligned} \boldsymbol{x}_t &\sim \mathcal{M}^{hmm} \\ \mathcal{M}^{hmm} &= \{a_{ij}, \boldsymbol{\mu}_j^{(x)}, \boldsymbol{\Sigma}_j^{(x)}\} \end{aligned}$$



- hidden discrete states $q_t = j$, $j \in [1, N_s]$;
- diagonal covariances $\boldsymbol{\Sigma}_j^{(x)}$.

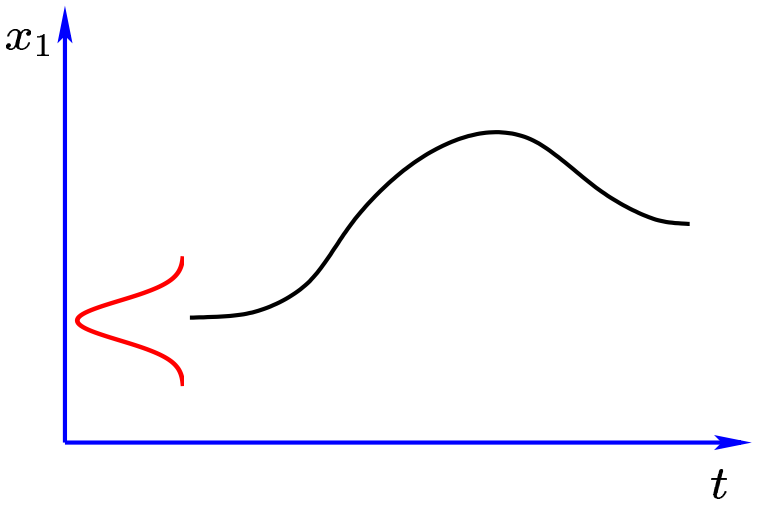
Properties:

- piecewise constant trajectory;
- state conditional independence assumption.

Linear first-order Gauss-Markov

Generative model:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{w}_t \\ \boldsymbol{w}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}_t^{(x)}, \boldsymbol{\Sigma}_t^{(x)}\right) \end{aligned}$$



- transition matrix $\boldsymbol{A}_t$;
- diagonal covariances $\boldsymbol{\Sigma}^{(x)}$;
- initial distribution: $\boldsymbol{x}_1 = \mathcal{N}\left(\boldsymbol{\mu}_t^{(i)}, \boldsymbol{\Sigma}_t^{(i)}\right)$.

Properties:

- continuous trajectory;
- models temporal correlation explicitly.

Bayesian networks

Graphical models to illustrate independence assumptions

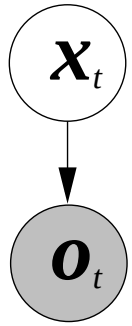
- round node: continuous variable;
- square node: discrete variable;
- shaded node: observable;
- no arrow: conditional independence.

Examples:

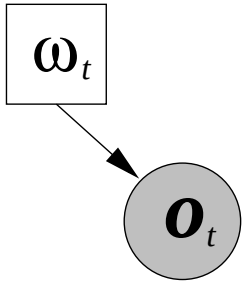
1. Factor Analysis:
   $p(\boldsymbol{o}_t|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{C}_t\boldsymbol{x}_t + \boldsymbol{\mu}_t^{(o)}, \boldsymbol{\Sigma}_t^{(o)}\right)$

2. Mixture of Gaussians:
   $p(\boldsymbol{o}_t|\omega_t = n) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right)$

1



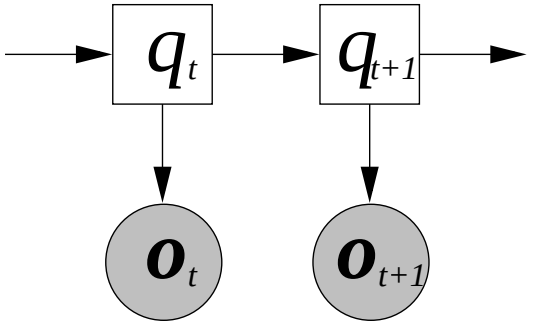
2



Dynamic Bayesian networks

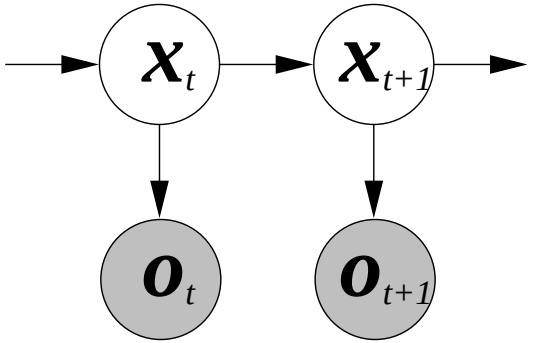
Hidden Markov model

- hidden discrete Markov sequence:
  $P(q_{t+1}|q_1,\ldots,q_t) = P(q_{t+1}|q_t)$
- state conditional observation distribution:
  $p(\boldsymbol{o}_t|q_t = j) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$



Linear dynamical system

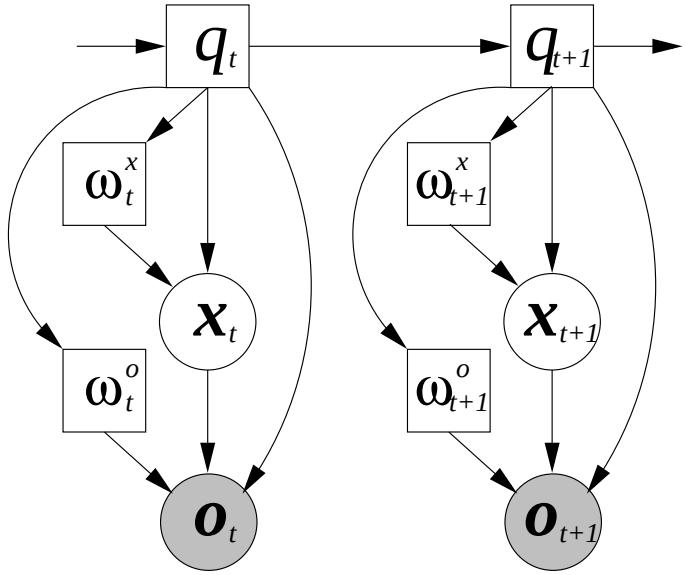
- hidden Gauss-Markov sequence:
  $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_{t+1}; \boldsymbol{A}_t\boldsymbol{x}_t + \boldsymbol{\mu}_t^{(x)}, \boldsymbol{\Sigma}_t^{(x)}\right)$
- factor analysis observation process:
  $p(\boldsymbol{o}_t|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{C}_t\boldsymbol{x}_t + \boldsymbol{\mu}_t^{(o)}, \boldsymbol{\Sigma}_t^{(o)}\right)$



Factor analysed HMM

Generative model:

$$\begin{aligned} \boldsymbol{x}_t &= \mathcal{M}^{hmm} \\ \boldsymbol{o}_t &= \boldsymbol{C}_t \boldsymbol{x}_t + \boldsymbol{v}_t \end{aligned}$$



- mixture of Gaussians HMM: $\mathcal{M}^{hmm} = \{a_{ij}, c_{jn}^{(x)}, \boldsymbol{\mu}_{jn}^{(x)}, \boldsymbol{\Sigma}_{jn}^{(x)}\}$;
- mixture of Gaussians noise: $\boldsymbol{v}_t \sim \sum_m c_{jm}^{(o)} \mathcal{N}(\boldsymbol{\mu}_{jm}^{(o)}, \boldsymbol{\Sigma}_{jm}^{(o)})$.

Effectively $M^{(o)}M^{(x)}$ component full covariance matrix system:

- mean vectors $\boldsymbol{\mu}_{jmn} = \boldsymbol{C}_j \boldsymbol{\mu}_{jn}^{(x)} + \boldsymbol{\mu}_{jm}^{(o)}$;
- covariance matrices $\boldsymbol{\Sigma}_{jmn} = \boldsymbol{C}_j \boldsymbol{\Sigma}_{jn}^{(x)} \boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jm}^{(o)}$.

Properties of FAHMM

Number of free parameters:

- order of $2(M^{(x)}-1)k + pk + 2M^{(o)}p$ per state;
- arbitrary tying of model parameters.

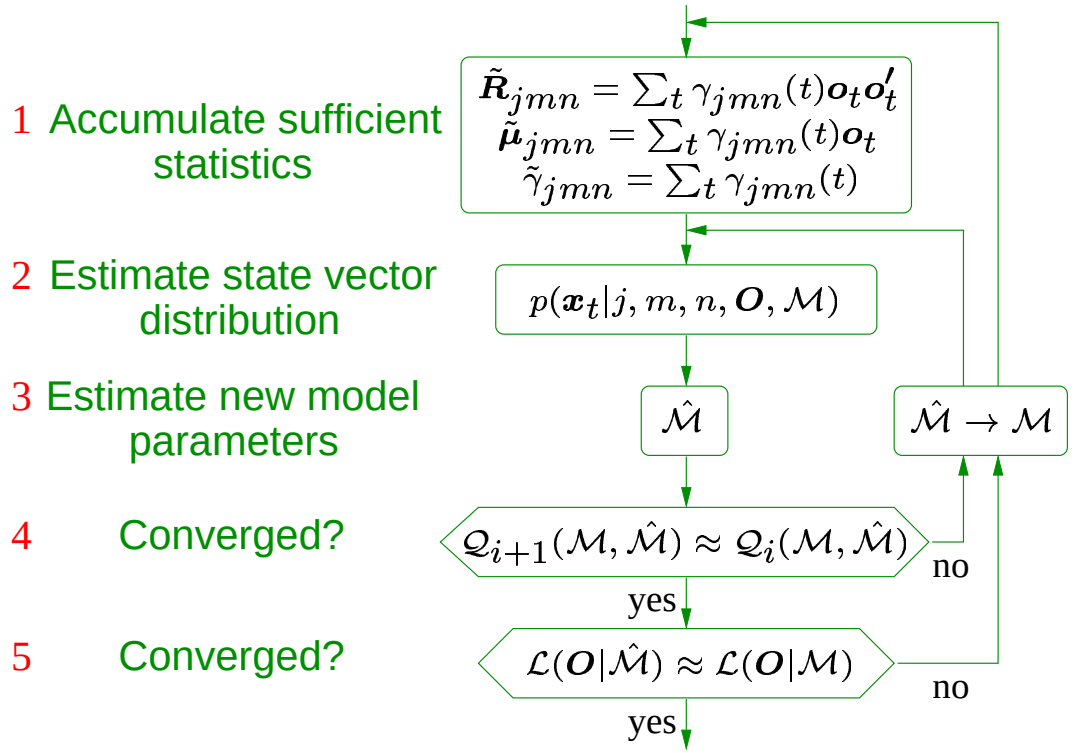
Related standard systems include:

- $M^{(x)} = 0 \quad \mapsto \quad$ hidden Markov model;
- $M^{(x)} = 1 \quad \mapsto \quad$ shared factor analysis;
- $M^{(o)} = 1 \quad \mapsto \quad$ dynamic version of independent factor analysis;
- $k = p$ and $\boldsymbol{v}_t = \mathbf{0} \quad \mapsto \quad$ semi-tied covariance matrix HMM (STC);
- $k > p$ and $\boldsymbol{v}_t = \mathbf{0} \quad \mapsto \quad$ extended MLLT (restricted).

Training FAHMM

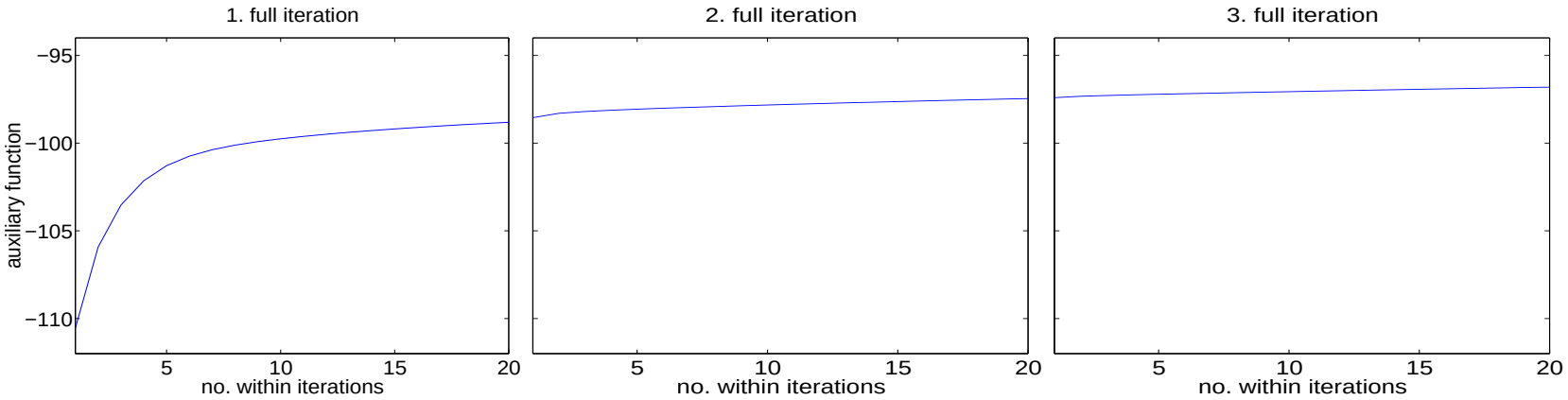
Expectation maximisation algorithm:

- auxiliary function $\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) = E\{ \log p(\boldsymbol{O}, \Omega^o, \boldsymbol{X}, \Omega^x, Q|\hat{\mathcal{M}}) \big| \boldsymbol{O}, \mathcal{M} \}$;
- forward-backward algorithm gives $\gamma_{jmn}(t) = P(j, m, n|\boldsymbol{O}, \mathcal{M})$



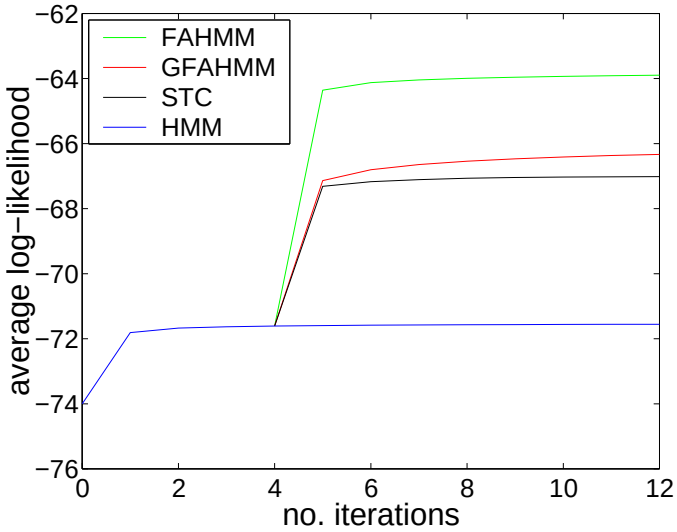
Training FAHMM continued

Multiple within iterations increase log-likelihood faster:



Log-likelihoods vs. full iterations:

- standard HMM;
- global full C STC;
- global C FAHMM ($k = 39$);
- untied FAHMM ($k = 13$).



Resource Management experiments

- 3990 sentence training data set (`train+dev_aug`);
- 39 dimensional feature vectors (MFCC+E+D+A);
- decision tree clustered cross-word triphone models;
- 1200 sentence evaluation data set (`feb89+oct89+feb91+sep92`);
- word-pair grammar;
- best HMM performance 3.99% (6 comps);
- global full transform STC;
- global observation matrix FAHMM with $k = 39$.

no. comps	1	5
HMM	7.79%	4.09%
STC	7.06%	3.83%
FAHMM	6.52%	3.68%

Switchboard Minitrain experiments

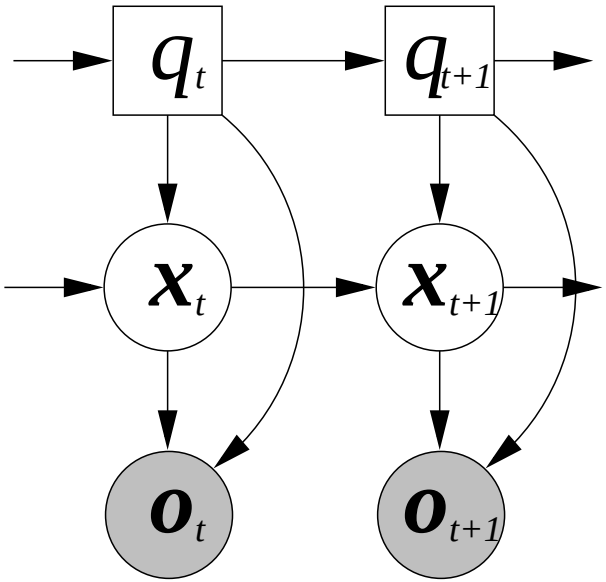
- 18 hour training data set;
- 39 dimensional feature vectors (MF-PLP+A+D);
- VTLN + mean and variance side based normalisation;
- decision tree clustered cross-word triphone models;
- 10 conversation sides of SWBD-2 and 10 of CHE evaluation set;
- untied FAHMM with $k = 13$;
- best baseline performance with 12 components;
- baseline performance exceeded with $M^{(o)} = 2$ and $M^{(x)} = 4$ (nfp 741);
- best performance with $M^{(o)} = 2$ and $M^{(x)} = 6$.

no. comps	WER	nfp
HMM	51.0%	936
FAHMM	50.7%	793

Switching linear dynamical system

Generative model:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{w}_t \\ \boldsymbol{o}_t &= \boldsymbol{C}_t \boldsymbol{x}_t + \boldsymbol{v}_t \end{aligned}$$



- HMM indexes standard LDS parameters: $\boldsymbol{\mu}_j^{(i)}, \boldsymbol{\Sigma}_j^{(i)}, \boldsymbol{A}_j, \boldsymbol{\mu}_j^{(x)}, \boldsymbol{\Sigma}_j^{(x)}, \boldsymbol{C}_j, \boldsymbol{\mu}_j^{(o)}, \boldsymbol{\Sigma}_j^{(o)}$;
- continuous state vector evolution.

Intractable inference!

Given the segmentation

- Kalman filtering for likelihood computation;
- Kalman smoothing for parameter estimation.

Stochastic segment model vs. SLDS

Stochastic segment model

- first segment model with standard LDS as the trajectory model;
- segments assumed independent given the discrete state; i.e., state vectors initialised on the segment boundaries;
- training alternates between segmentation and parameter estimation.

Switching LDS

- relaxes independent segment assumption by propagating state vector posteriors over the segment boundaries;
- exact inference results in exponential growth of paths (N_s^T).

Both model unimodal distributions!

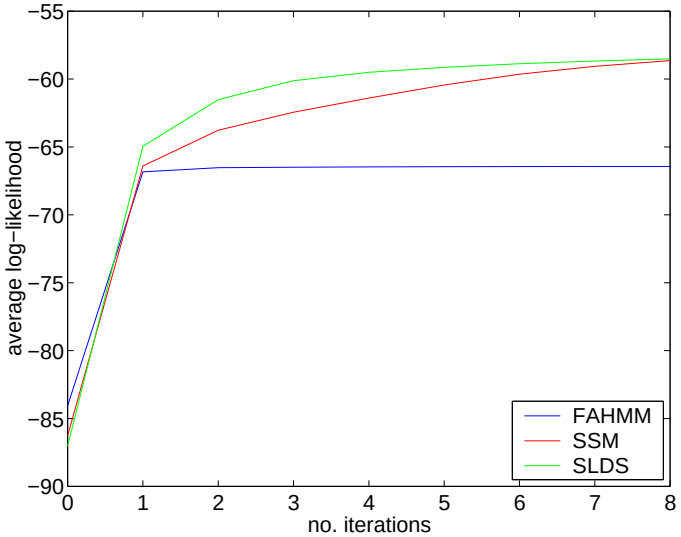
Resource Management experiments

Aim to compare discrete vs. continuous state evolution processes

- decision tree clustered cross-word triphone system at model level;
- single state per model in SSM and SLDS, $k = 13$;
- three state FAHMM, $k = 13$, all parameters except $\boldsymbol{\mu}_j^{(x)}$ tied at model level;
- model aligned training data using alignment FAHMM.

Forced alignment training

- flat start tied three state FAHMM;
- flat start single state SSM;
- flat start single state SLDS.



Resource Management results

20-best re-scoring

- hypotheses and alignments using alignment FAHMM;
- hypotheses re-scored using FAHMM, SSM and SLDS.

Test data word error rates:

FAHMM	SSM	SLDS
11.00 %	10.69 %	13.45 %

Subset of training data word error rates:

FAHMM	SSM	SLDS
4.30 %	3.96 %	4.30 %

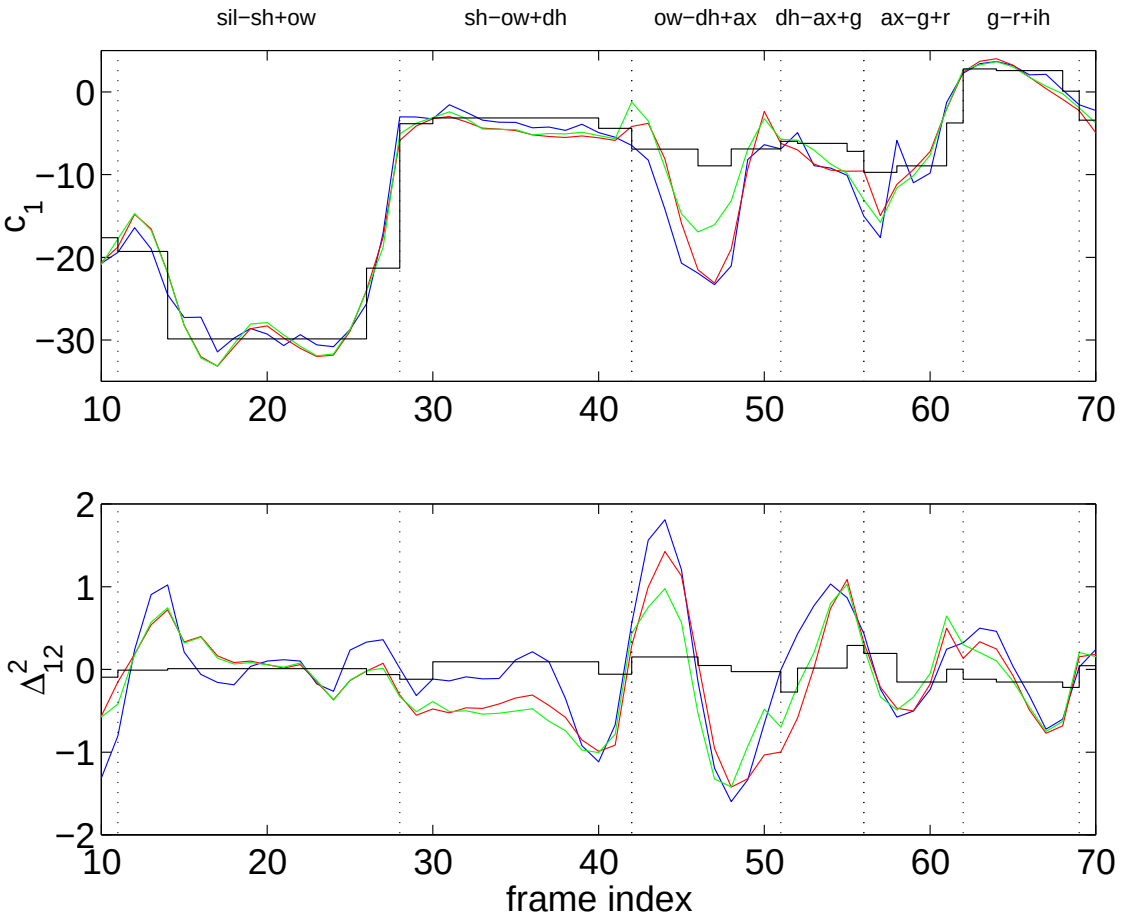
SSM outperforms SLDS!

True and estimated trajectories

- estimates given by $\hat{\boldsymbol{o}}_t = \boldsymbol{C}_t \hat{\boldsymbol{x}}_t + \boldsymbol{\mu}_t^{(o)}$, where $\hat{\boldsymbol{x}}_t = E\{\boldsymbol{x}_t | \boldsymbol{O}\}$.

Legend:

- True
- SSM
- SLDS
- FAHMM



Conclusions

- general framework of linear Gaussian models developed;
- developed and evaluated factor analysed HMMs;
- initial experiments with SSM and SLDS.

Future work

- develop decoding algorithms for SSM and SLDS;
- investigate multi-modal distributions for SSM and SLDS.

Any suggestions ...