UniS Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

HMM tutorial 5

by Dr Philip Jackson

- Implementing B-W formulae
  - Review of α, β, γ and ξ
  - Forward & backward procedures
  - Accummulate & update
- Emission or output pdfs
  - Multivariate Gaussian
  - Alternative functions
  - Gaussian mixture pdf
  - Re-estimation with mixtures
- Practical issues

http://www.ee.surrey.ac.uk/Personal/P.Jackson/ce.adsp/

Implementing B-W re-estimation

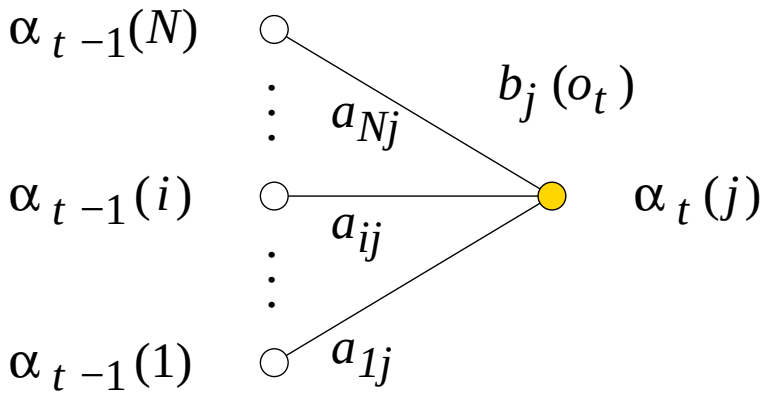
Overview

- **Forward**: compute likelihoods $\alpha_t(i)$
- **Backward**: compute likelihoods $\beta_t(i)$
- **Parallel**:
  - compute occupation $\gamma_t(i)$
  - compute transition $\xi_t(i,j)$
  - accumulate $\underline{\pi}_i$, $\underline{a}_{ij}$ and $\bar{a}_i$
  - accumulate $\underline{\boldsymbol{\mu}}_i$, $\underline{\boldsymbol{\Sigma}}_i$ and $\bar{b}_i$
- *Repeat* for all files in the training set, $r \in \{1, 2, \ldots, R\}$.

Forward likelihood

The forward likelihood is a joint probability,

$$\alpha_t(i) = P\left(x_t = i, \mathbf{o}_1^t | \lambda\right). \tag{1}$$

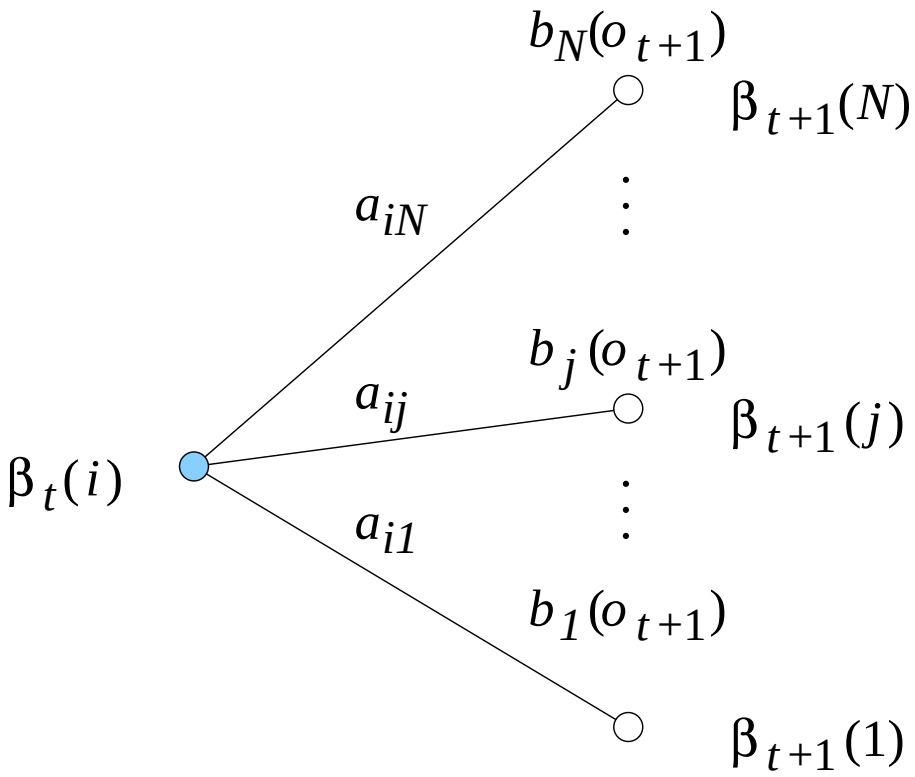


Trellis fragment depicting calculation of a forward likelihood.

Backward likelihood

The backward likelihood is a conditional probability,

$$\beta_t(j) = P\left(\mathbf{o}_{t+1}^T | x_t = j, \lambda\right). \tag{2}$$

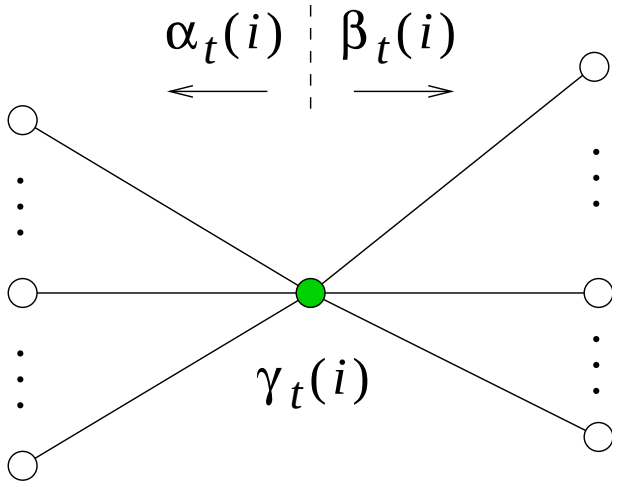


Trellis fragment depicting calculation of a backward likelihood.

Occupation likelihood

The occupation likelihood is a conditional probability,

$$\gamma_t(i) = P\left(x_t = i|\mathcal{O}, \lambda\right). \tag{3}$$

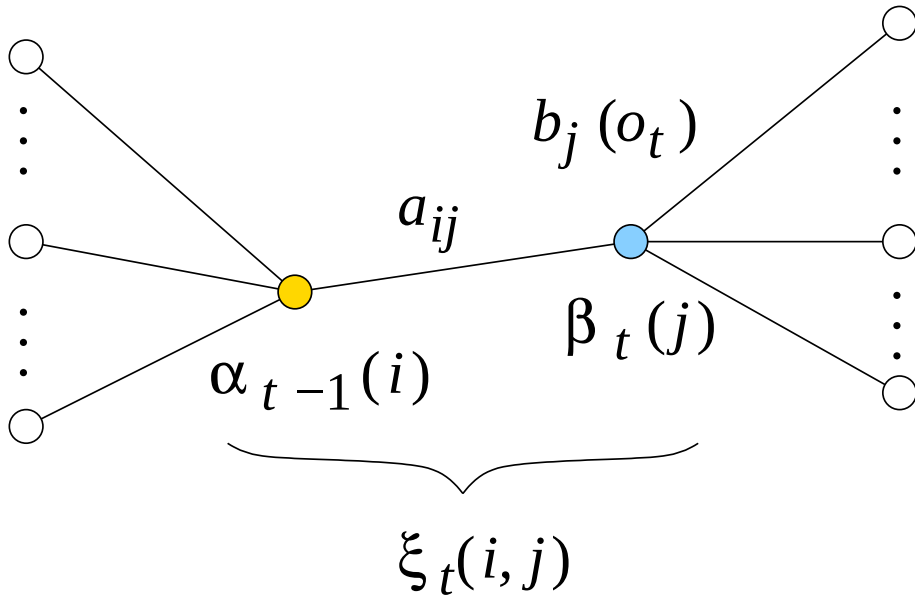


Trellis fragment depicting calculation of occupation likelihood.

Transition likelihood

The transition likelihood is the joint probability of states i and j, conditional on the observations,

$$\xi_t(i,j) = P\left(x_{t-1} = i, x_t = j | \mathcal{O}, \lambda\right). \tag{4}$$



Trellis fragment depicting calculation of a transition likelihood.

Forward procedure (Problem 1)

1. Initially,
   $\alpha_1(i) = \pi_i \, b_i(\boldsymbol{o}_1)$, for $1 \leq i \leq N$;

2. For $t = 2, 3, \ldots, T$,
   $\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) \, a_{ij}\right] b_j(\boldsymbol{o}_t)$, for $1 \leq j \leq N$;

3. Finally, for the rth file,
   $P^r = P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

Backward procedure (Problem 1)

1. Initially,
   $\beta_T(i) = 1$, for $1 \leq i \leq N$;

2. For $t = T-1, T-2, \ldots, 1$,
   $\beta_t(i) = \sum_{j=1}^{N} a_{ij} \, b_j(\boldsymbol{o}_{t+1}) \, \beta_{t+1}(j)$, for $1 \leq i \leq N$.

3. Alternatively, we have $P^r = \sum_{i=1}^{N} \pi_i \, b_i(\boldsymbol{o}_1) \, \beta_1(i)$.

Parallel procedure

Occupation likelihoods:

$$\gamma_t(j) = \frac{\alpha_t(j)\,\beta_t(j)}{P^r}$$

Transition likelihoods:

$$\xi_t(i,j) = \frac{\alpha_{t-1}(i)\;\; a_{ij}\, b_j(\boldsymbol{o}_t)\;\; \beta_t(j)}{P^r}$$

Transition accumulators:

$$\begin{aligned}
\underline{\pi}_i &\mapsto \underline{\pi}_i + \gamma_1(i) \\
\underline{a}_{ij} &\mapsto \underline{a}_{ij} + \sum_{t=2}^{T} \xi_t(i,j) \\
\bar{a}_i &\mapsto \bar{a}_i + \sum_{t=2}^{T} \gamma_{t-1}(i)
\end{aligned}$$

Parallel (continued)

Emission accumulators:

$$\underline{\boldsymbol{\mu}}_i \;\mapsto\; \underline{\boldsymbol{\mu}}_i + \sum_{t=1}^{T} \gamma_t(i)\boldsymbol{o}_t$$

$$\underline{\boldsymbol{\Sigma}}_i \;\mapsto\; \underline{\boldsymbol{\Sigma}}_i + \sum_{t=1}^{T} \gamma_t(i)(\boldsymbol{o}_t - \boldsymbol{\mu}_i)(\boldsymbol{o}_t - \boldsymbol{\mu}_i)'$$

$$\bar{b}_i \;\mapsto\; \bar{b}_i + \sum_{t=1}^{T} \gamma_t(i)$$

Repeat

For all files in the training set:

1. recompute the forward and backward likelihoods;
2. recompute the occupation and transition likelihoods;
3. increment the transition and emission accumulators.

Update (Problem 3)

Finally, we update the models:

$$\hat{\pi}_i = \frac{1}{R}\sum_{r=1}^{R} \underline{\pi}_i$$

$$\hat{a}_{ij} = \frac{\sum_{r=1}^{R} \underline{a}_{ij}}{\sum_{r=1}^{R} \bar{a}_i}$$

$$\hat{b}_i \begin{cases} \hat{\boldsymbol{\mu}}_i = \dfrac{\sum_{r=1}^{R} \underline{\boldsymbol{\mu}}_i}{\sum_{r=1}^{R} \bar{b}_i} \\ \\ \hat{\boldsymbol{\Sigma}}_i = \dfrac{\sum_{r=1}^{R} \underline{\boldsymbol{\Sigma}}_i}{\sum_{r=1}^{R} \bar{b}_i} \end{cases}$$

where $\hat{}$ denotes new values of the model parameters, $\lambda = \{\pi, A, B\}$.

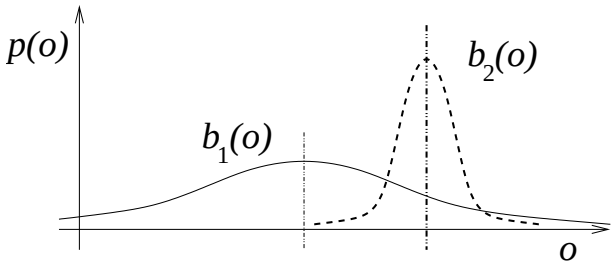
Summary of the training procedure

1. initialise accumulators for all HMMs' parameters
2. read the next training utterance
3. join HMMs in sequence to make composite HMM
4. calculate forward & backward probabilities
5. increment accumulators using forwd/backwd probs
6. repeat from step 2 until all utterances processed
7. use accumulators to update parameters for all HMMs

Parametric emission pdfs

Univariate Gaussian

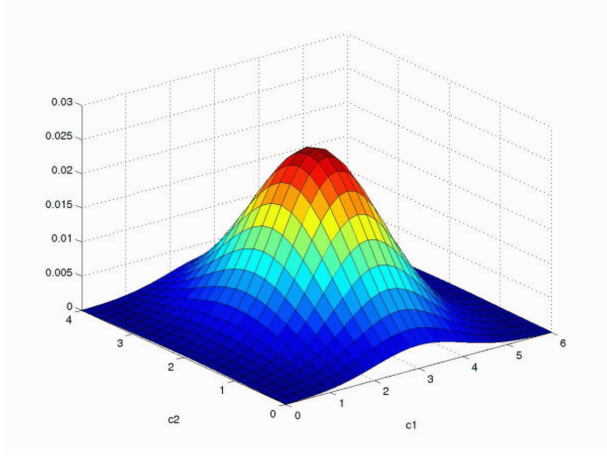
$$b_i(o_t) = \frac{1}{\sqrt{2\pi\Sigma_i}} \exp\left[-\frac{(o_t - \mu_i)^2}{2\Sigma_i}\right].$$



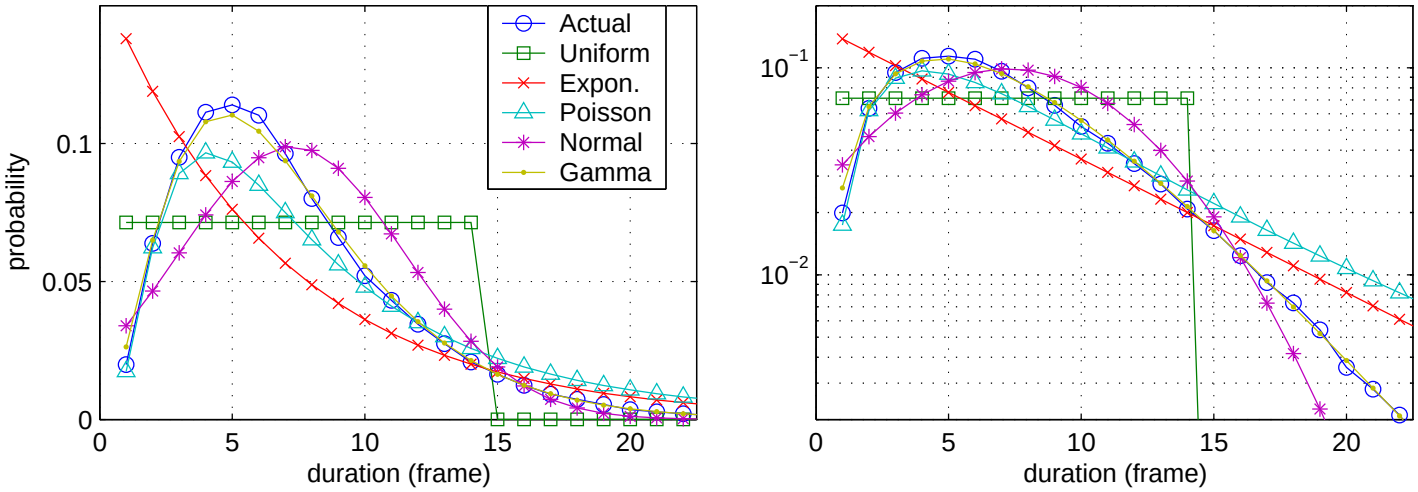
Multivariate Gaussian

$$b_i(\boldsymbol{o}_t) = \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}_i|}} \exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)'\right],$$

where the dimensionality of the observation space is K.



Other distributions



Various functions on (left) linear and (right) logarithmic scales.

- Exponential
- Poisson
- Gamma
- Log-normal
- Gaussian mixture
- Ricean
- Rayleigh
- Beta
- student-t
- Cauchy

Gaussian mixture pdfs

Univariate Gaussian mixture

$$b_i(o_t) = \sum_{m=1}^{M} c_{im}\,\mathcal{N}(o_t; \mu_{im}, \Sigma_{im})$$

$$= \sum_{m=1}^{M} \frac{c_{im}}{\sqrt{2\pi\Sigma_{im}}} \exp\left[-\frac{(o_t - \mu_{im})^2}{2\Sigma_{im}}\right], \quad (5)$$

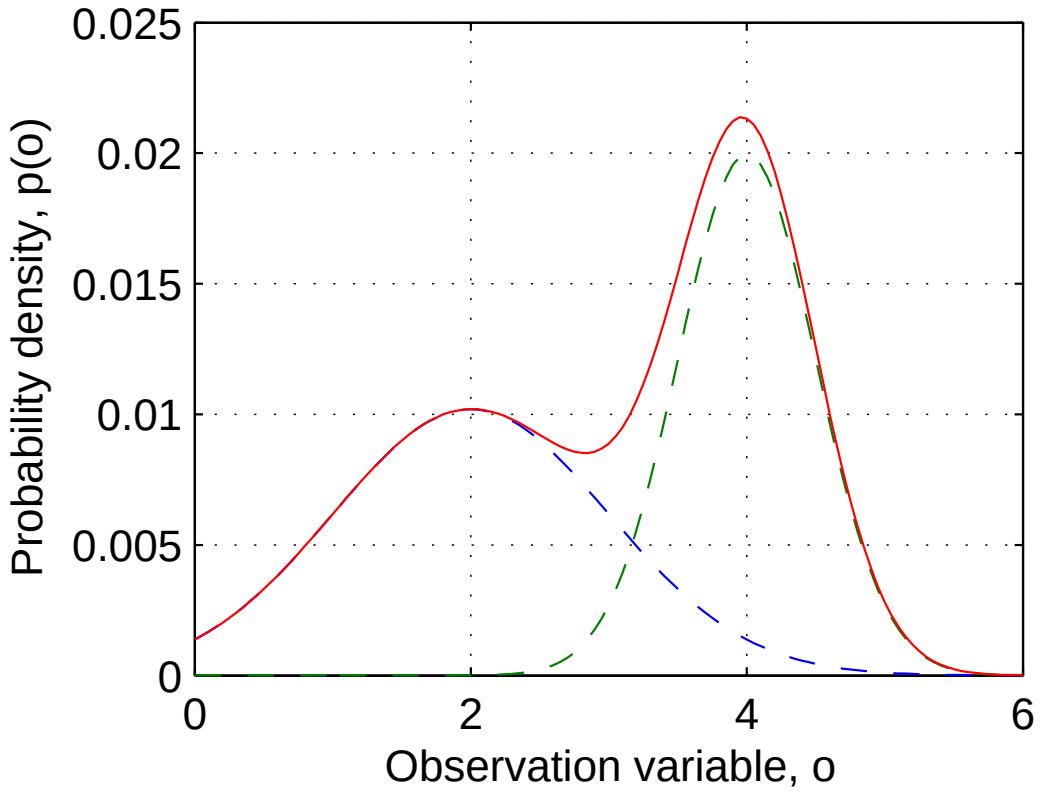
where M is the number of mixture components in the Gaussian mixture (i.e. M-mix), and $\sum_{m=1}^{M} c_{im} = 1$.

Multivariate Gaussian mixture

$$b_i(\boldsymbol{o}_t) = \sum_{m=1}^{M} c_{im}\,\mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}), \quad (6)$$

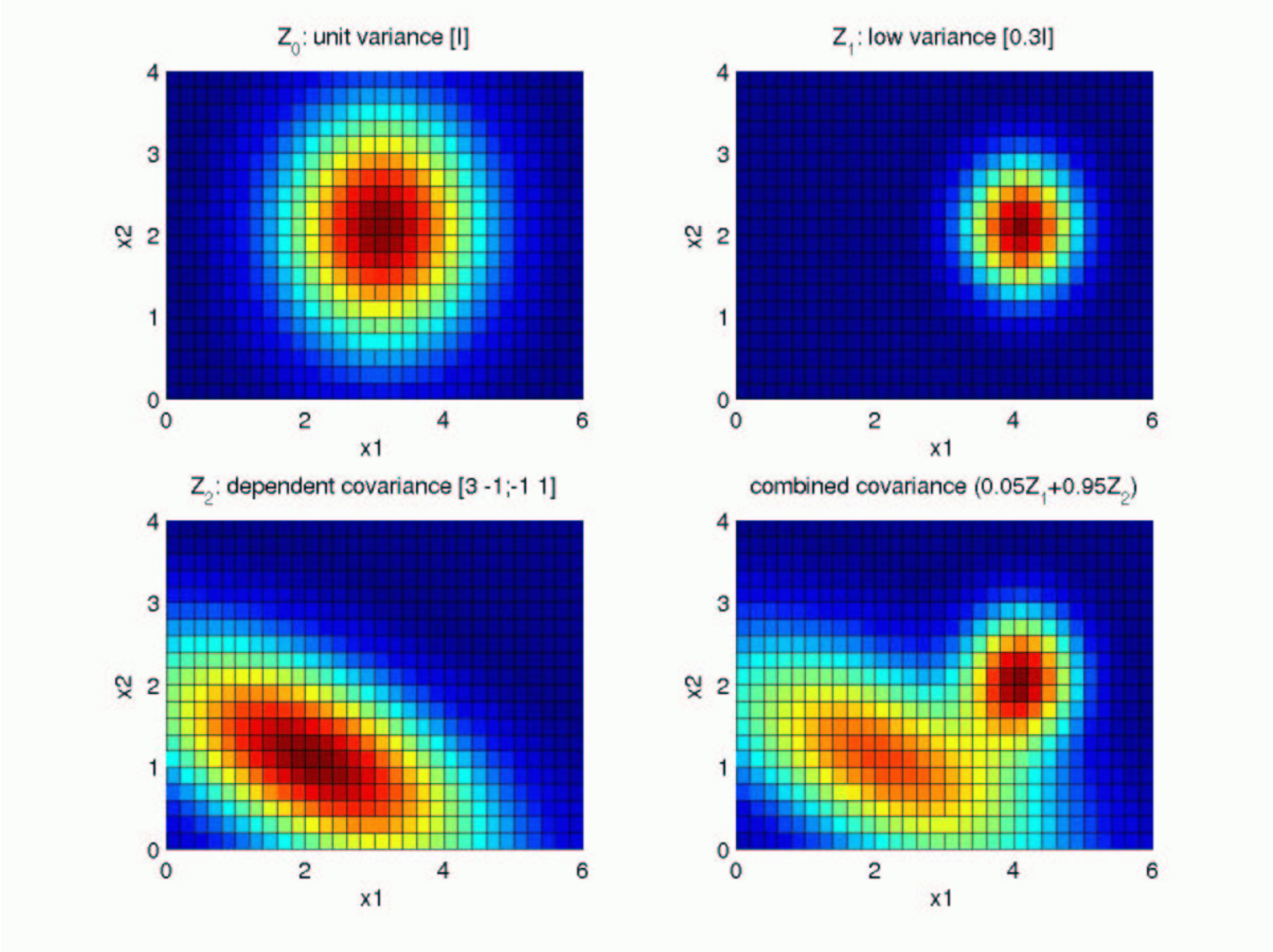
where $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \Sigma)$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ evaluated at $\boldsymbol{o}$.

Univariate Gaussian mixture



Example of a univariate Gaussian mixture pdf.

Multivariate Gaussian mixture



Examples of multivariate Gaussian pdfs
and a multivariate Gaussian mixture.

Viterbi training with mixtures

Mean: $$\widehat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} q_t(j,m)\boldsymbol{o}_t}{\sum_{t=1}^{T} q_t(j,m)},$$

Covariance: $$\widehat{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} q_t(j,m)(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})'}{\sum_{t=1}^{T} q_t(j,m)},$$

Weights: $$\widehat{c}_{jm} = \frac{\sum_{t=1}^{T} q_t(j,m)}{\sum_{t=1}^{T} q_t(j)},$$

where $q_t(j,m)$ and $q_t(j)$ are occupation indicators:

$$q_t(j) = \begin{cases} 1 & \text{for } i = x_t; \\ 0 & \text{otherwise,} \end{cases}$$

$$q_t(j,m) = \begin{cases} 1 & \text{for } i = x_t,\ m = \widehat{m}; \\ 0 & \text{otherwise,} \end{cases}$$

and the occupied mixture is

$$\widehat{m} = \arg\max_m c_{jm}\mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right).$$

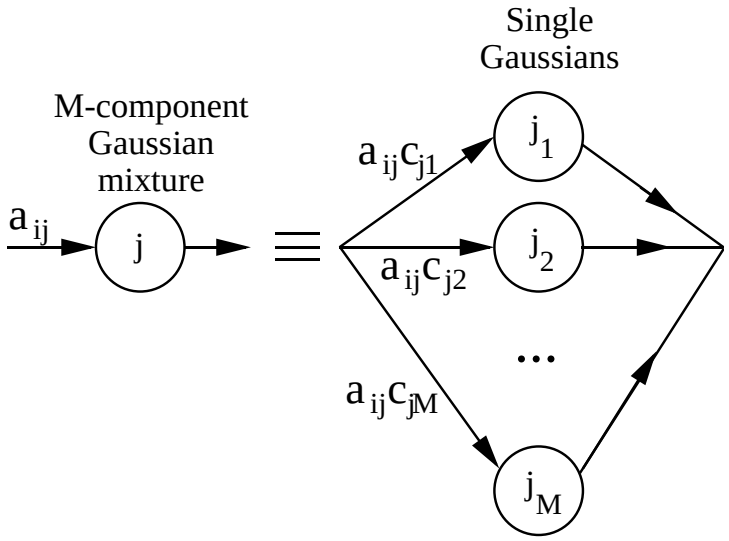
Baum-Welch training with mixtures

Mixture-occupation likelihood

$$\gamma_t(j,m) = \frac{\alpha_t^{\emptyset}(j)\ c_{jm}\,\mathcal{N}\left(\boldsymbol{o}_t;\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_{jm}\right)\ \beta_t(j)}{P(\mathcal{O}|\lambda)}, \qquad (7)$$

where

$$\gamma_t(j) = \sum_{m=1}^{M} \gamma_t(j,m), \quad \text{and} \quad \alpha_t^{\emptyset}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)a_{ij}\right].$$



Representing a mixture of Gaussians.*

B-W re-estimation of mixture parameters

Mean:

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)\boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(j,m)}, \tag{8}$$

Covariance:

$$\hat{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})'}{\sum_{t=1}^{T} \gamma_t(j,m)}, \tag{9}$$

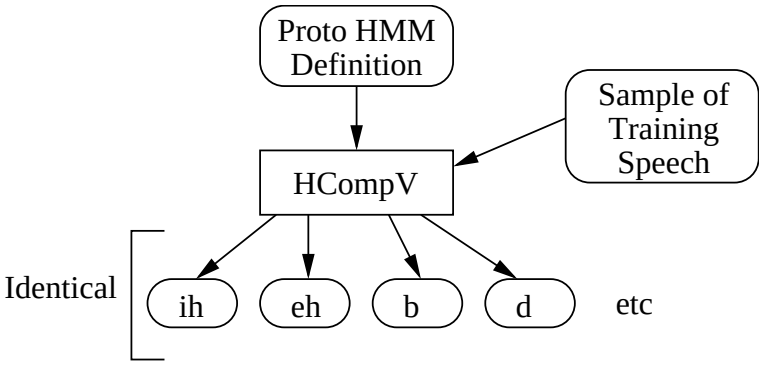
Mixture weights:

$$\hat{c}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)}{\sum_{t=1}^{T} \gamma_t(j)}. \tag{10}$$

Practical issues

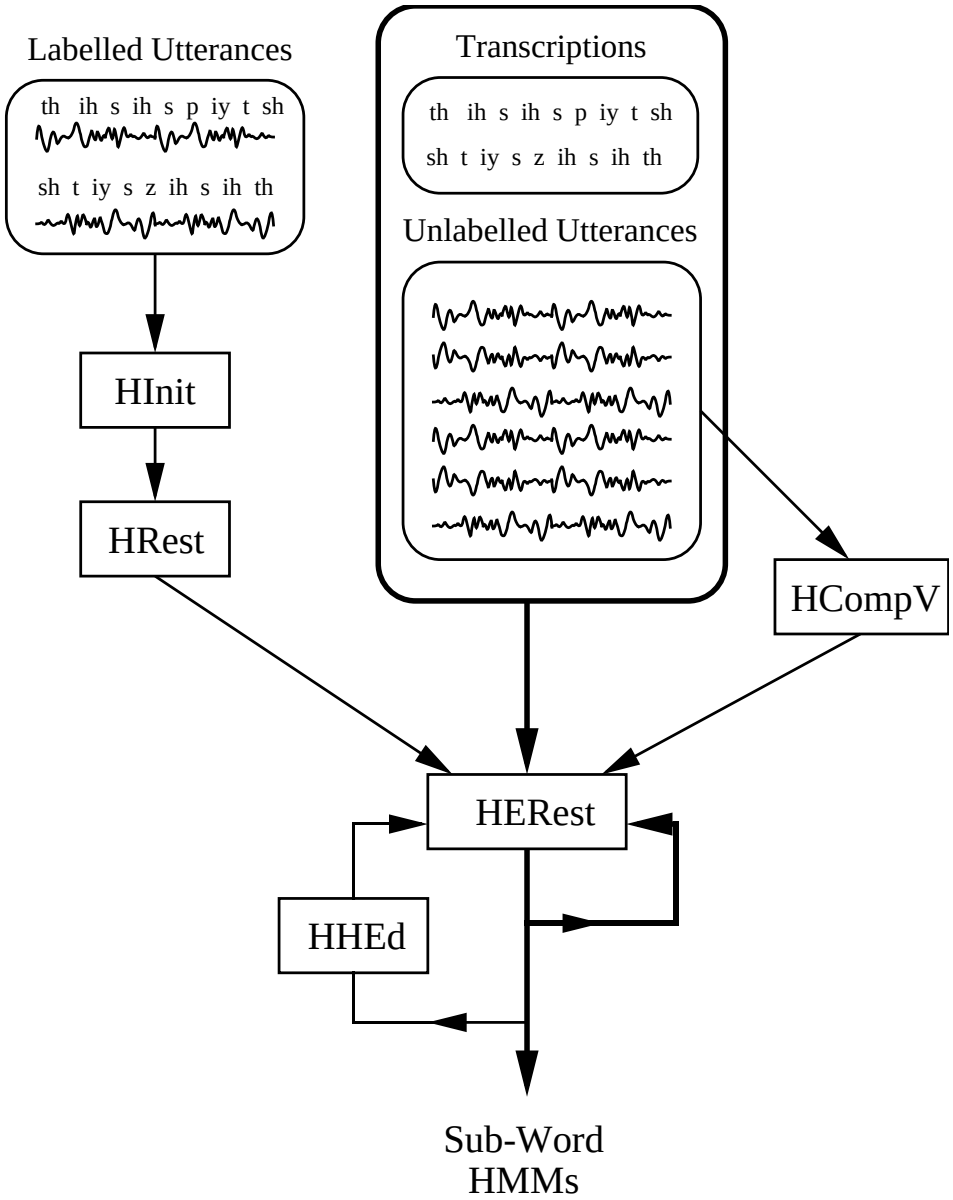
Model initialisation

1. Random
2. Flat start
3. Least squares
4. Viterbi
5. Baum-Welch (supervised)
6. Baum-Welch (unsupervised)



Flat start.*

Re-estimation and embedded re-estimation



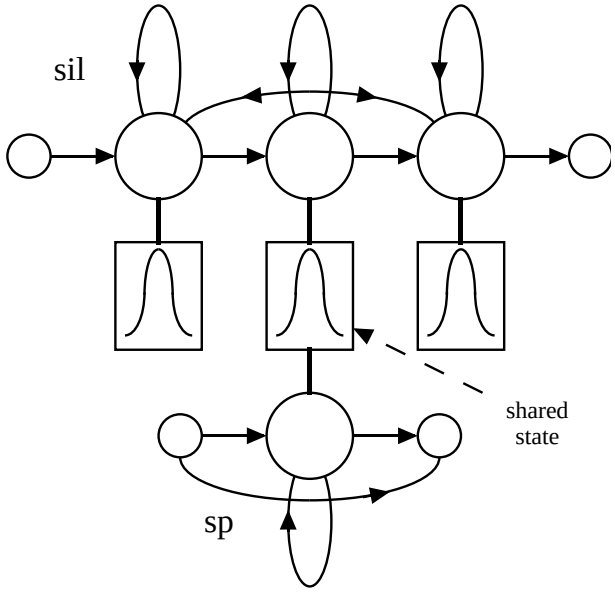
Training sub-word HMMs.*

Number of parameters

- Parsimony
  - Occam's razor
- Amount of training data

Regularisation

- variance floor
- parameter tying
  - context sensitvity
  - agglomerative clustering
  - phonetic decision tree



Linking the silence and short pause models.*

Context-sensitive models

- **Problem:** coarticulation causes phones to vary

- **Solution:** train separate models for each context

  – monophone:

  $$\text{sil - t - e - n - sil}$$

  – triphone:

  $$\text{sil}_{t} - {}_{sil}\text{t}_{e} - {}_{t}\text{e}_{n} - {}_{e}\text{n}_{sil} - {}_{n}\text{sil}$$

Tutorial 5 summary

- Review of likelihoods α_t, β_t, γ_t and ξ_t
- Implementation of B-W formulae:
  - forward, backward, accummulate & update
- Output probability distribution functions
  - Multivariate Gaussian mixture, etc.
- Practical issues
  - Annotations, initialisation & tying
  - Context-sensitive models