Disclaimer

- The material provided in this document is not my original work and is a summary of some one else’s work(s).
- A simple Google search of the title of the document will direct you to the original source of the material.
- I do not guarantee the accuracy, completeness, timeliness, validity, non-omission, merchantability or fitness of the contents of this document for any particular purpose.
-

Theory and Implementation of Hidden Markov Models

Outline

- Introduction
- Discrete Time Markov Processes
- Extensions to Hidden Markov Models
- The Three Basic Problems of HMMs
- Types of HMMs
- Continuous Observation Densities in HMMs
- Autoregressive HMMs
- Variants on HMM Structures
- Inclusion of Explicit State Duration Density in HMMs
- Optimization Criterion– ML, MMI, and MDI
- Comparisons of HMMs
- Implementation Issues for HMMs
- Improving the Effectiveness of Model Estimates
- Model Clustering and Splitting
- HMM System for Isolated Word Recognition

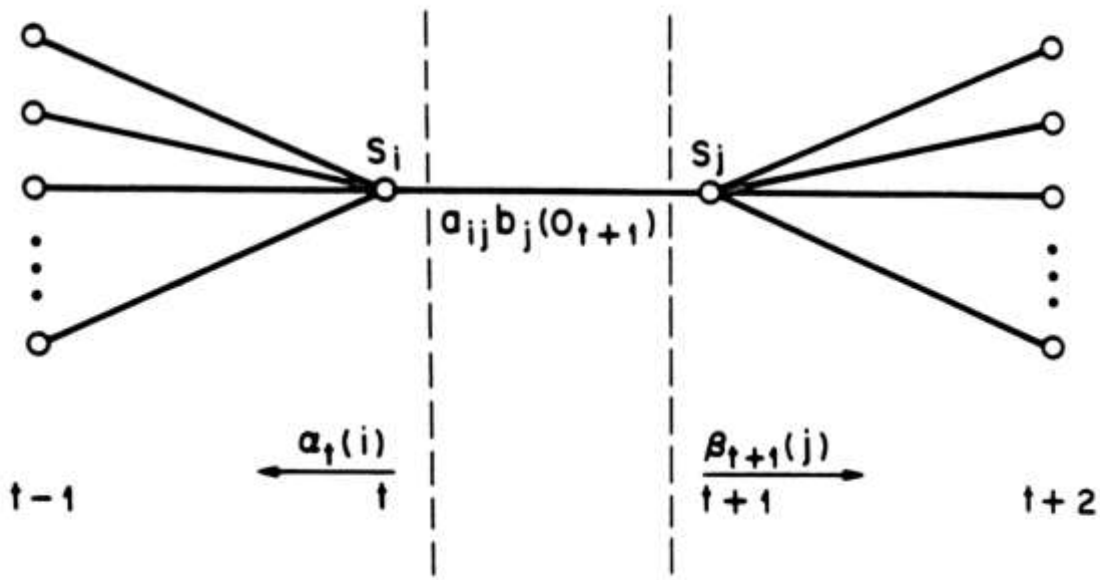
Solution to Problem 3

- The third problem of HMMs is to determine a method to adjust the model parameters $\lambda=(A,B,\pi)$ to maximize the probability of the observation sequence given the model
- There is no known way to analytically solve for the model which maximizes the probability of the observation sequence
- We choose $\lambda=(A,B,\pi)$ such that $P(O|\lambda)$ is locally maximized using an iterative procedure such as the Baum-Welch method

Solution to Problem 3

- First define $\varepsilon_t(i,j)$, the probability of being in state i at time t, and state j at time t + 1, given the model and the observation sequence

$$\xi_t(i,j) = P(q_t = i,\ q_{t+1} = j|\mathbf{O}, \lambda).$$



Solution to Problem 3

- First define εt(i,j), the probability of being in state i at time t, and state j at time t + 1, given the model and the observation sequence

$$\xi_t(i,j) = P(q_t = i,\ q_{t+1} = j | \mathbf{O}, \lambda).$$

$$\xi_t(i,j) = \frac{P(q_t = i,\ q_{t+1} = j,\ \mathbf{O} \mid \lambda)}{P(\mathbf{O} \mid \lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)}{P(\mathbf{O}|\lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i)\, a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)}.$$

Solution to Problem 3

- The probability of being in state i at time t, given the observation sequence and the model

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j).$$

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from state } i \text{ in } \mathbf{O}$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{expected number of transitions from state } i \text{ to state } j \text{ in } \mathbf{O}.$$

Solution to Problem 3

- Using the above formulas (and the concept of counting event occurrences) we can give a method for re-estimation of the parameters of an HMM

$$\bar{\pi}_j = \text{expected frequency (number of times) in state } i \text{ at time } (t = 1) = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } i \text{ to state } j}{\text{expected number of transitions from state } i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } \mathbf{v}_k}{\text{expected number of times in state } j}$$

$$= \frac{\sum_{\substack{t=1 \\ \text{s.t. } \mathbf{o}_t = \mathbf{v}_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}.$$

Solution to Problem 3

- We define the current model as $\lambda=(A,B,\pi)$ and use that to compute the right-hand sides of the above equations
- We define the re-estimated model as $\bar{\lambda}=(\bar{A},\bar{B},\bar{\pi})$ as determined from the left hand side of the above equations
- It has been proven by Baum and his colleagues that either
  - The initial model λ defines a critical point of the likelihood function $\lambda=\bar{\lambda}$
  - Model $\bar{\lambda}$ is more likely than model λ in the sense that $P(O|\bar{\lambda}) > P(O|\lambda)$

Solution to Problem 3

- Based on the above procedure, if we iteratively use $\bar{\lambda}$ in place of λ and repeat the re-estimation calculation, we then can improve the probability of 0 being observed from the model until some limiting point is reached
- The final result of this re-estimation procedure is called a maximum likelihood estimate of the HMM

Solution to Problem 3

- The re-estimation formulas can be derived directly by maximizing Baum's auxiliary function

$$Q(\lambda', \lambda) = \sum_{\mathbf{q}} P(\mathbf{O}, \mathbf{q}|\lambda') \log P(\mathbf{O}, \mathbf{q}|\lambda)$$

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \Rightarrow P(\mathbf{O}|\lambda) \geq P(\mathbf{O}|\lambda')$$

Derivation of Re-estimation formulas from Q function

$$P(\mathbf{O}, \mathbf{q}|\lambda) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\mathbf{o}_t)$$

$$\log\ P(\mathbf{O}, \mathbf{q}|\lambda) = \log\ \pi_{q_0} + \sum_{t=1}^{T} \log\ a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log\ b_{q_t}(\mathbf{o}_t)$$

$$Q(\lambda', \lambda) = Q_\pi(\lambda', \boldsymbol{\pi}) + \sum_{i=1}^{N} Q_{a_i}(\lambda', \mathbf{a}_i) + \sum_{i=1}^{N} Q_{b_i}(\lambda', \mathbf{b}_i)$$

Derivation of Re-estimation formulas from Q function

- We can maximize Q by maximizing the individual terms separately subject to the stochastic constraints

$$\sum_{j=1}^{N} \pi_j = 1$$

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad \forall j$$

$$\sum_{k=1}^{K} b_i(k) = 1, \qquad \forall i.$$

Derivation of Re-estimation formulas from Q function

- Because the individual auxiliary functions all have the form

$$\sum_{j=1}^{N} w_j \log y_j$$

- It attains a global minimum at the single point

$$y_j = \frac{w_j}{\sum_{i=1}^{N} w_i}, \qquad j = 1, 2, \ldots, N$$

Derivation of Re-estimation formulas from Q function

- The maximization leads to the model re-estimate $\bar{\lambda}=(\bar{A},\bar{B},\bar{\pi})$ where

$$\bar{\pi}_i = \frac{P(\mathbf{O}, q_0 = i|\lambda)}{P(\mathbf{O}|\lambda)}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} = i, q_t = j|\lambda)}{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} = i|\lambda)}$$

$$\bar{b}_i(k) = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_t = i|\lambda)\ \delta(\mathbf{o}_t, \mathbf{v}_k)}{\sum_{t=1}^{T} P(\mathbf{O},\ q_t = i|\lambda)}$$

$$\begin{aligned} \delta(\mathbf{o}_t, \mathbf{v}_k) &= 1 \quad \text{if } \mathbf{o}_t = \mathbf{v}_k \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Derivation of Re-estimation formulas from Q function

- Using the forward backward variables

$$P(\mathbf{O}, q_t = i|\lambda) = \alpha_t(i)\beta_t(i)$$

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) = \sum_{i=1}^{N} \alpha_T(i)$$

$$P(\mathbf{O}, q_{t-1} = i,\ q_t = j|\lambda) = \alpha_{t-1}(i)\, a_{ij} b_j(\mathbf{o}_t)\beta_t(j)$$

- Giving

$$\bar{\pi}_i = \frac{\alpha_0(i)\beta_0(i)}{\sum_{j=1}^{N} \alpha_T(j)} = \gamma_0(i)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i)\, a_{ij} b_j(\mathbf{o}_t)\beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)} = \frac{\sum_{t=1}^{T} \xi_{t-1}(i,j)}{\sum_{t=1}^{T} \gamma_{t-1}(i)}$$

$$\bar{b}_i(k) = \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)\delta(\mathbf{o}_t, \mathbf{v}_k)}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)} = \frac{\sum_{\substack{t=1 \\ \text{s.t. } \mathbf{o}_t = \mathbf{v}_k}}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$

Notes on the Re-Estimation Procedure

- The re-estimation formulas can readily be interpreted as an implementation of the EM algorithm of statistics
  - E (expectation) step is the calculation of the auxiliary function $Q(\bar{\lambda}, \lambda)$
  - M (modification) step is the maximization over $\bar{\lambda}$
- The stochastic constraints of the HMM parameters, are automatically incorporated at each iteration

$$\sum_{i=1}^{N} \bar{\pi}_i = 1 \qquad \sum_{k=1}^{M} \bar{b}_j(k) = 1,$$

$$\sum_{j=1}^{N} \bar{a}_{ij} = 1,$$

Notes on the Re-Estimation Procedure

- By looking at the parameter estimation problem as a constrained optimization of $P(O|\lambda)$, $P(O|\lambda)$ can be maximized if the following conditions are met

$$\pi_i = \frac{\pi_i \dfrac{\partial P}{\partial \pi_i}}{\displaystyle\sum_{k=1}^{N} \pi_k \frac{\partial P}{\partial \pi_k}} \qquad a_{ij} = \frac{a_{ij} \dfrac{\partial P}{\partial a_{ij}}}{\displaystyle\sum_{k=1}^{N} a_{ik} \frac{\partial P}{\partial a_{ik}}}$$

$$b_j(k) = \frac{b_j(k) \dfrac{\partial P}{\partial b_j(k)}}{\displaystyle\sum_{\ell=1}^{M} b_j(\ell) \frac{\partial P}{\partial b_j(\ell)}}.$$

Types of HMMs

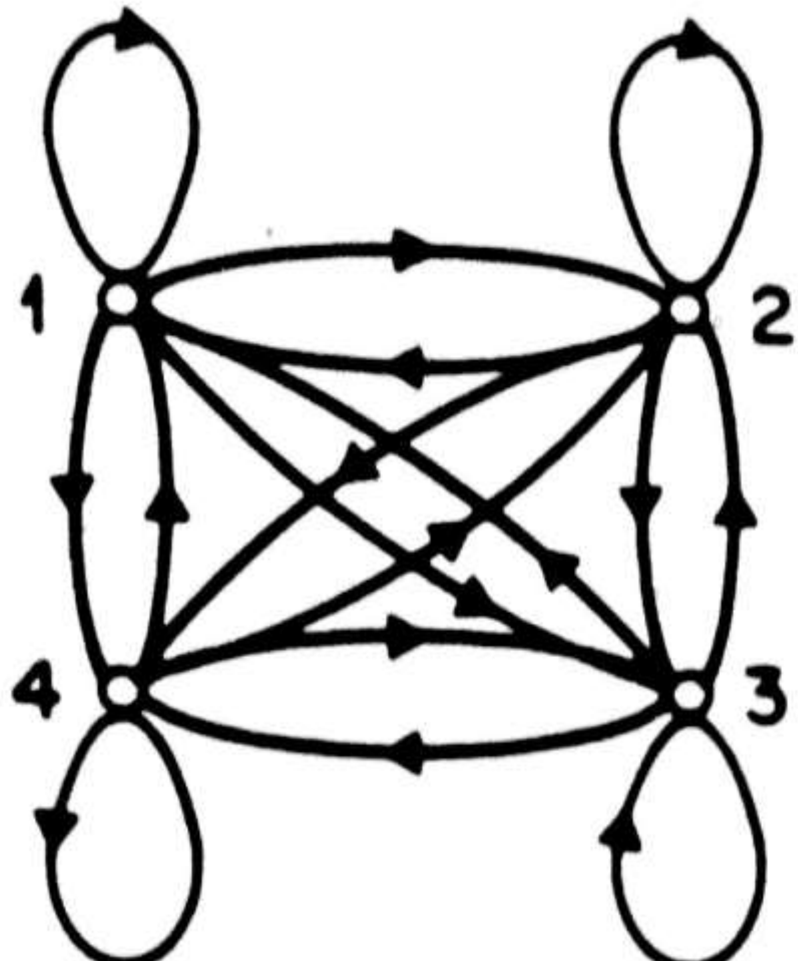
- Ergodic

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}.$$

Types of HMMs

- Ergodic

$A =$



Types of HMMs

- Left to Right

$$a_{ij} = 0, \qquad j < i$$

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases}$$

$$a_{ij} = 0, \quad j > i + \Delta i$$

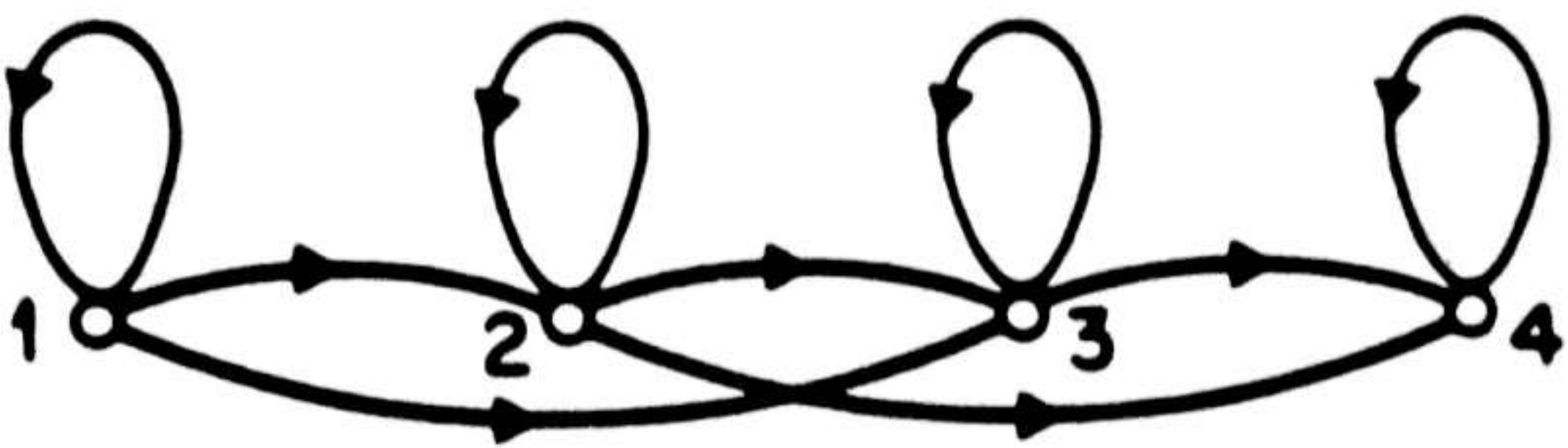
$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix}$$

Types of HMMs

- Left to Right

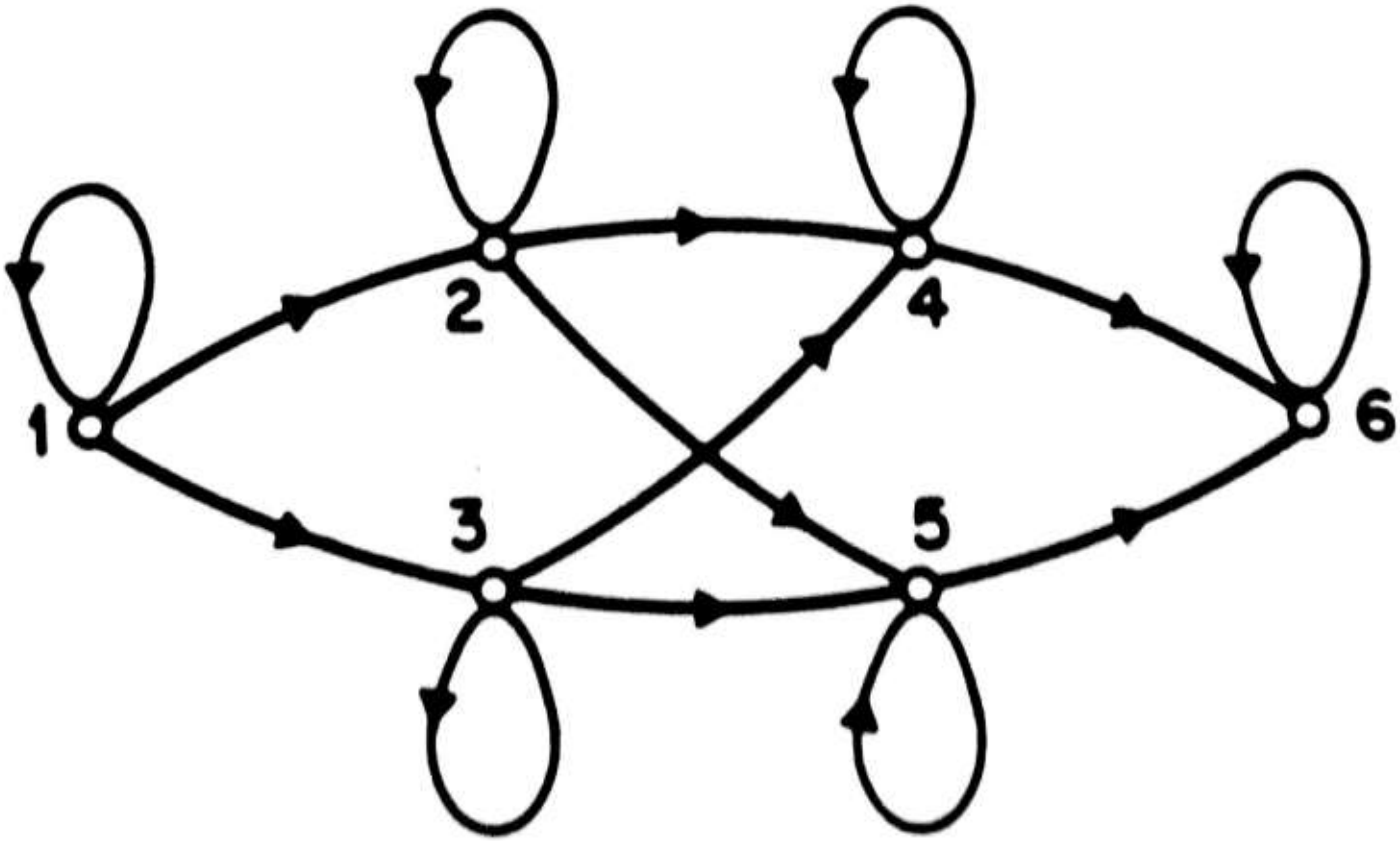
$$a_{ij} = 0, \qquad j < i$$

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases}$$



Types of HMMs

- Parallel Path Left to Right



Continuous Observation Densities in HMMs

- In order to use a continuous observation density, some restrictions have to be placed on the form of the model pdf to insure that the parameters of the pdf can be re-estimated in a consistent way
- The most general representation of the pdf, for which a re-estimation procedure has been formulated is a finite mixture of the form

$$b_j(\mathbf{o}) = \sum_{k=1}^{M} c_{jk} \mathcal{N}(\mathbf{o}, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}), \qquad 1 \leq j \leq N$$

Continuous Observation Densities in HMMs

- In order to use a continuous observation density, some restrictions have to be placed on the form of the model pdf to insure that the parameters of the pdf can be re-estimated in a consistent way
- The mo… e pdf, for whi… s been formula… m

$$\sum_{k=1}^{M} c_{jk} = 1, \qquad 1 \le j \le N$$

$$c_{jk} \ge 0, \; 1 \le j \le N, \; 1 \le k \le M$$

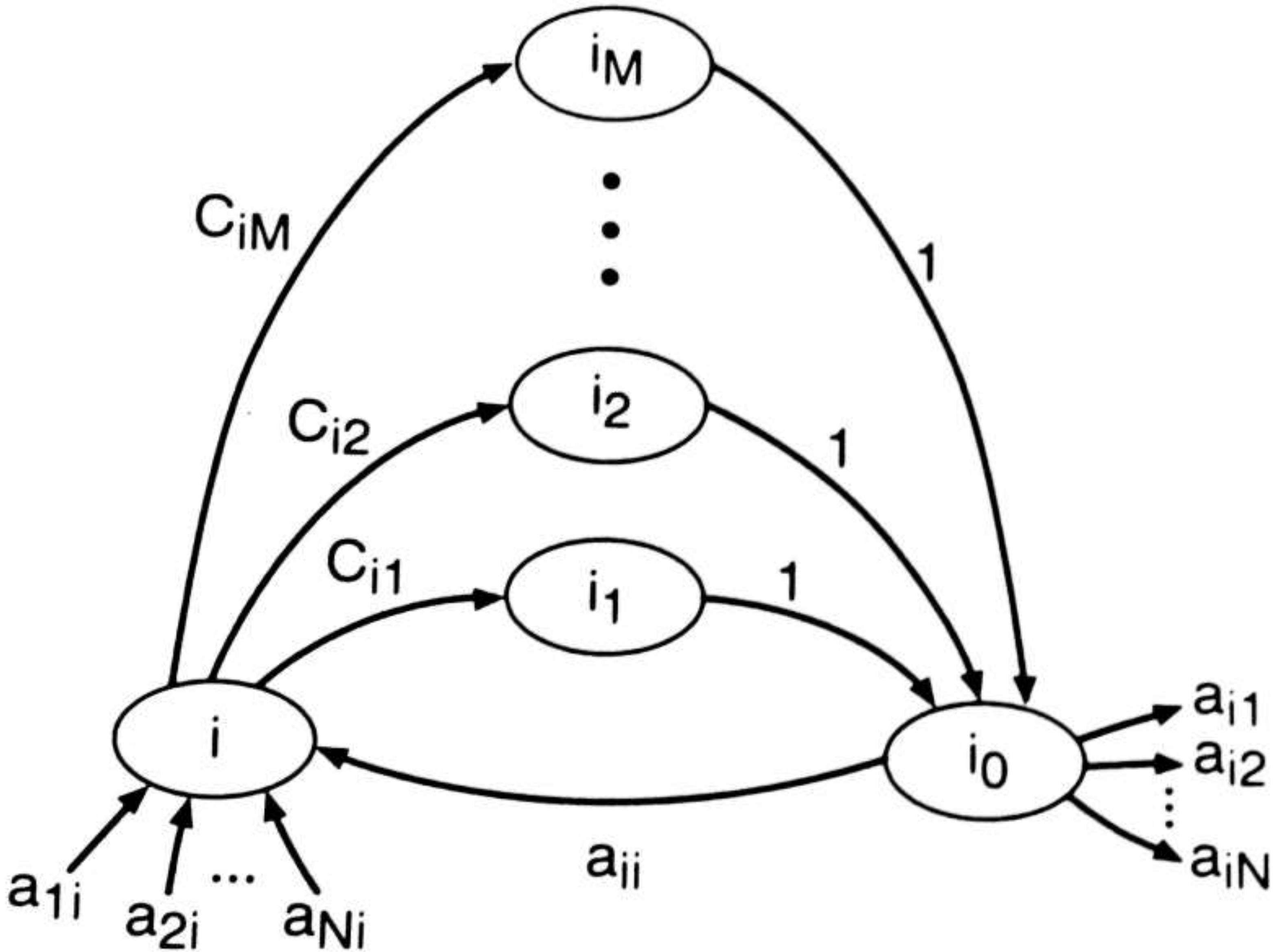
so that

$$\int_{-\infty}^{\infty} b_j(\mathbf{x})\, d\mathbf{x} = 1, \qquad 1 \le j \le N.$$

Continuous Observation Densities in HMMs

- The above pdf can be used to approximate, arbitrarily closely, any finite, continuous density function
- HMM states with mixture density is equivalent to a multistate single mixture density
- The mixture weights are interpreted as transition probabilities to sub-states

Continuous Observation Densities in HMMs



Continuous Observation Densities in HMMs

- It can be shown that the re-estimation formulas for the coefficients of the mixture density, i.e. c_{jk} and μ_{jk} and U_{jk} are of the form

$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j, k)}$$

$$\bar{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k) \cdot \mathbf{O}_t}{\sum_{t=1}^{T} \gamma_t(j, k)}$$

$$\bar{\boldsymbol{U}}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k) \cdot (\mathbf{O}_t - \boldsymbol{\mu}_{jk})(\mathbf{O}_t - \boldsymbol{\mu}_{jk})'}{\sum_{t=1}^{T} \gamma_t(j, k)}$$

Continuous Observation Densities in HMMs

- It can be shown that the re-estimation formulas for the coefficients of the mixture density, i.e. c_{jk} and μ_{jk} and U_{jk} are of the form

$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j, k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j, k)}$$

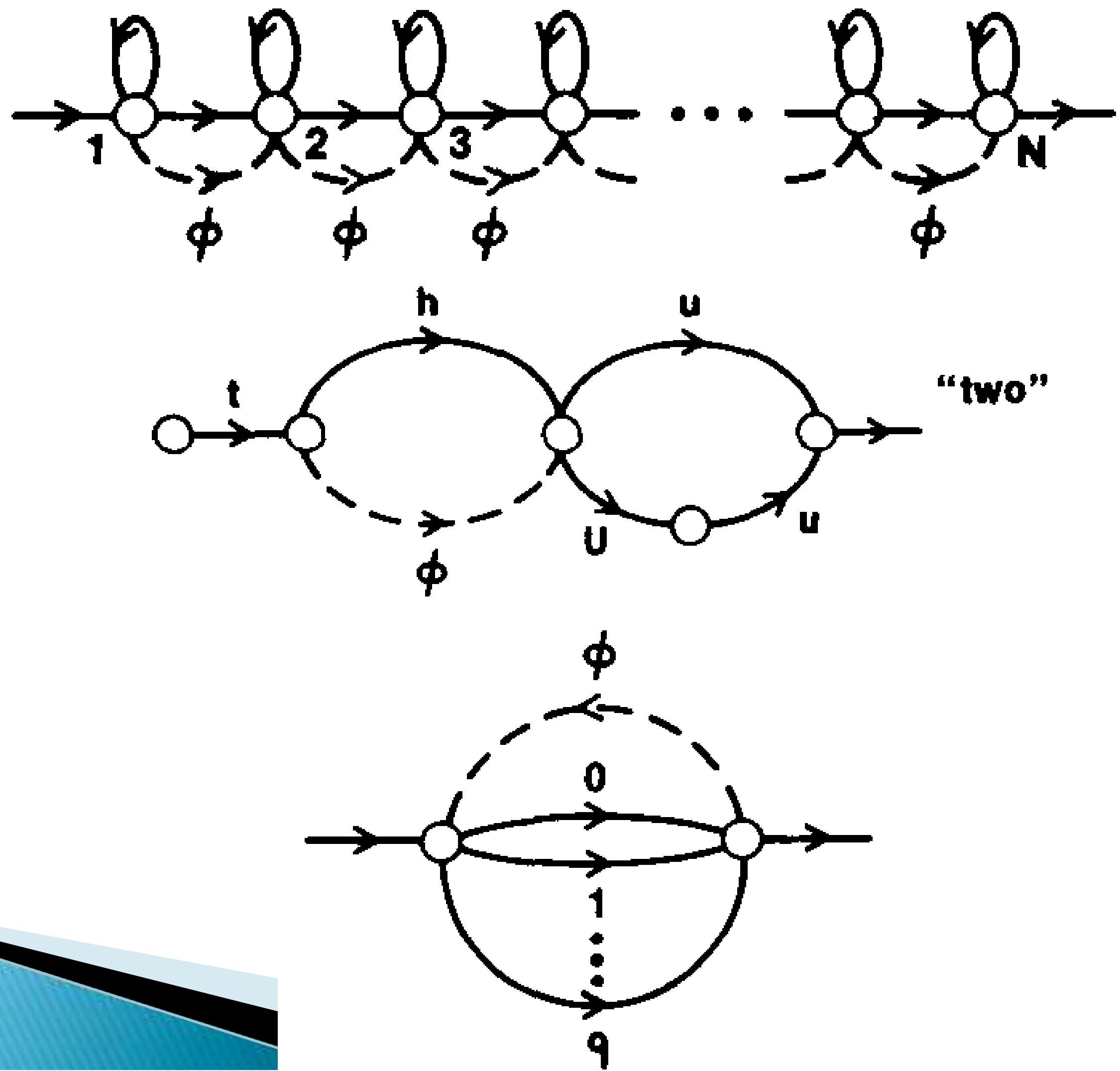
$$\gamma_t(j, k) = \left[\frac{\alpha_t(j)\, \beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\, \beta_t(j)} \right] \left[\frac{c_{jk}\, \mathfrak{N}(\mathbf{O}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{U}_{jk})}{\sum_{m=1}^{M} c_{jm}\, \mathfrak{N}(\mathbf{O}_t, \boldsymbol{\mu}_{jm}, \boldsymbol{U}_{jm})} \right].$$

$$\bar{\boldsymbol{U}}_{jk} = \frac{\sum_{t=1} \gamma_t(j, k) \cdot (\mathbf{O}_t - \boldsymbol{\mu}_{jk})(\mathbf{O}_t - \boldsymbol{\mu}_{jk})'}{\sum_{t=1}^{T} \gamma_t(j, k)}$$

Autoregressive HMMs

Variants on HMM Structures

- Null Transitions



Variants on HMM Structures

- Parameter Tying
- An equivalence relation is set up between HMM parameters in different states
- In this manner the number of independent parameters in the model is reduced and the parameter estimation becomes somewhat simpler
- Parameter tying is used in cases where the observation density is known to be the same in 2 or more states

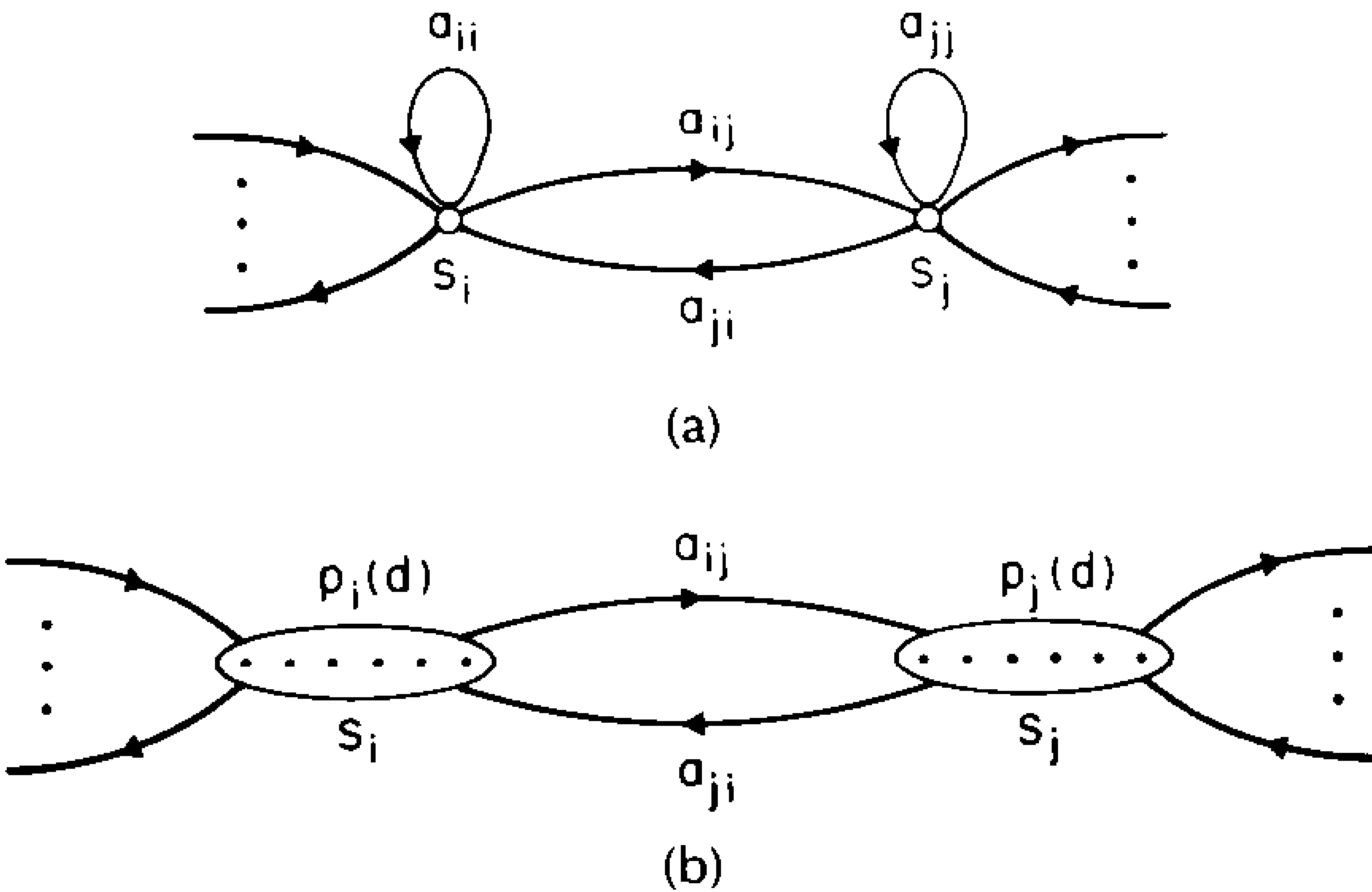
Inclusion of Explicit State Duration Density in HMMs

- The inherent duration probability density $p_i(d)$ associated with state i with self transition coefficient $\boldsymbol{a_{ii}}$ was of the form

$$p_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$$
$$= \text{probability of } d \text{ consecutive observations in state } S_i.$$

- For most physical signals, this exponential state duration density is inappropriate
- We would prefer to explicitly model duration density in some analytic form

Inclusion of Explicit State Duration Density in HMMs



Inclusion of Explicit State Duration Density in HMMs

- Based on the simple model of Fig above, the sequence of events of the variable duration HMM is as follows
  - An initial state, $q_1 = i$, is chosen according to the initial state distribution π_i
  - A duration d_1 is chosen according to the state duration density $p_{q1}(d_1)$
  - Observations $O_1, O_2 \dots O_{d1}$ are chosen according to the joint observation density, $b_{q1}(O_1, O_2 \dots O_{d1})$
  - The next state, $q_2 = i$ is chosen according to the state transition probabilities, a_{q1q2} with the constraint that $a_{q1q1} = 0$

Inclusion of Explicit State Duration Density in HMMs

- Using the above formulation, several changes must be made to the re-estimation formulas to allow calculation of P(0Iλ) and for re-estimation of all model parameters
- We assume that the first state begins at t = 1 and the last state ends at t = T
- The forward variable then becomes

$$\alpha_t(i) = P(O_1\, O_2 \cdots O_t,\, S_i \text{ ends at } t|\lambda).$$

- We assume that a total of r states have been visited during the first t observations
- We denote the states as $q_l, q_2, \ldots, q_r$ with durations associated with each state of d_l, $d_2, \ldots, d_r$

Inclusion of Explicit State Duration Density in HMMs

- Constraints

$$q_r = i \qquad \sum_{s=1}^{r} d_s = t.$$

- The forward variable becomes

$$\alpha_t(i) = \sum_q \sum_d \pi_{q_1} \cdot p_{q_1}(d_1) \cdot P(O_1\, O_2 \cdots O_{d_1}|q_1)$$

$$\cdot\, a_{q_1 q_2} p_{q_2}(d_2)\, P(O_{d_1+1} \cdots O_{d_1+d_2}|q_2) \cdots$$

$$\cdot\, a_{q_{r-1} q_r} p_{q_r}(d_r)\, P(O_{d_1+d_2+\cdots+d_{r-1}+1} \cdots O_t|q_r)$$

- Where the sum is over all states q and all possible state durations *d*

Inclusion of Explicit State Duration Density in HMMs

- By induction

$$\alpha_t(j) = \sum_{i=1}^{N} \sum_{d=1}^{D} \alpha_{t-d}(i)\, a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\mathbf{O}_s)$$

- To initialize the computation

$$\alpha_1(i) = \pi_i p_i(1) \cdot b_i(\mathbf{O}_1)$$

$$\alpha_2(i) = \pi_i p_i(2) \prod_{s=1}^{2} b_i(\mathbf{O}_s) + \sum_{\substack{j=1 \\ j \neq i}}^{N} \alpha_1(j)\, a_{ji} p_i(1)\, b_i(\mathbf{O}_2)$$

$$\alpha_3(i) = \pi_i p_i(3) \prod_{s=1}^{3} b_i(\mathbf{O}_s) + \sum_{d=1}^{2} \sum_{\substack{j=1 \\ j \neq i}}^{N} \alpha_{3-d}(j)\, a_{ji} p_i(d) \cdot \prod_{s=4-d}^{3} b_i(\mathbf{O}_s)$$

Inclusion of Explicit State Duration Density in HMMs

- By induction

$$\alpha_t(j) = \sum_{i=1}^{N} \sum_{d=1}^{D} \alpha_{t-d}(i)\, a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\mathbf{O}_s)$$

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Inclusion of Explicit State Duration Density in HMMs

- To derive re-estimation formula, we define three more forward-backward variables

$$\alpha_t^*(i) = P(O_1\, O_2 \cdots O_t,\, S_i \text{ begins at } t+1|\lambda)$$

$$\beta_t(i) = P(O_{t+1} \cdots O_T|S_i \text{ ends at } t,\, \lambda)$$

$$\beta_t^*(i) = P(O_{t+1} \cdots O_T|S_i \text{ begins at } t+1,\, \lambda).$$

Inclusion of Explicit State Duration Density in HMMs

$$\alpha_t^*(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij}$$

$$\alpha_t(i) = \sum_{d=1}^{D} \alpha_{t-d}^*(i)\, p_i(d) \prod_{s=t-d+1}^{t} b_i(\mathbf{O}_s)$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \beta_t^*(j)$$

$$\beta_t^*(i) = \sum_{d=1}^{D} \beta_{t+d}(i)\, p_i(d) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s).$$

Inclusion of Explicit State Duration Density in HMMs

- Based on the above relationships and definitions, the re-estimation formulas for the variable duration HMM are

$$\bar{\pi}_i = \frac{\pi_i \beta_0^*(i)}{P(O|\lambda)} \qquad \bar{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha_t(i)\, a_{ij} \beta_t^*(j)}{\sum_{j=1}^{N} \sum_{t=1}^{T} \alpha_t(i)\, a_{ij} \beta_t^*(j)}$$

$$\bar{b}_i(k) = \frac{\sum_{\substack{t=1 \\ \text{s.t. } O_t = k}}^{T} \left[\sum_{\tau<t} \alpha_\tau^*(i) \cdot \beta_\tau^*(i) - \sum_{\tau<t} \alpha_\tau(i)\, \beta_\tau(i) \right]}{\sum_{k=1}^{M} \sum_{\substack{t=1 \\ \text{s.t. } O_t = v_k}}^{T} \left[\sum_{\tau<t} \alpha_\tau^*(i) \cdot \beta_\tau^*(i) - \sum_{\tau<t} \alpha_\tau(i)\, \beta_\tau(i) \right]}$$

$$\bar{p}_i(d) = \frac{\sum_{t=1}^{T} \alpha_t^*(i)\, p_i(d)\, \beta_{t+d}(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s)}{\sum_{d=1}^{D} \sum_{t=1}^{T} \alpha_t^*(i)\, p_i(d)\, \beta_{t+d}(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{O}_s)}.$$

Optimization Criterion– ML, MMI, and MDI

Comparisons of HMMs

- A distance measure $D(\lambda_1,\lambda_2)$ can be defined between two Markov models, λ_1 and λ_2 as

$$D(\lambda_1, \lambda_2) = \frac{1}{T}[\log P(O^{(2)}|\lambda_1) - \log P(O^{(2)}|\lambda_2)]$$

- A natural expression of this measure is the symmetrized version

$$D_s(\lambda_1, \lambda_2) = \frac{D(\lambda_1\lambda_2) + D(\lambda_2, \lambda_1)}{2}.$$