Non Parametric Bayesian Acoustic Model Discovery for Phoneme Classification

Tanuj Jain

26-03-2015

Problem Description

- Classification of Phonemes:
  - Unsupervised: Labels absent
  - Non – Parametric: Varying number of parameters
  - Language Independent

- Phoneme:
  - Basic unit of Language‘s Phonology
  - Number varies from language to language
  - Multiple phonemes combine to form meaningful entity
  - Example:

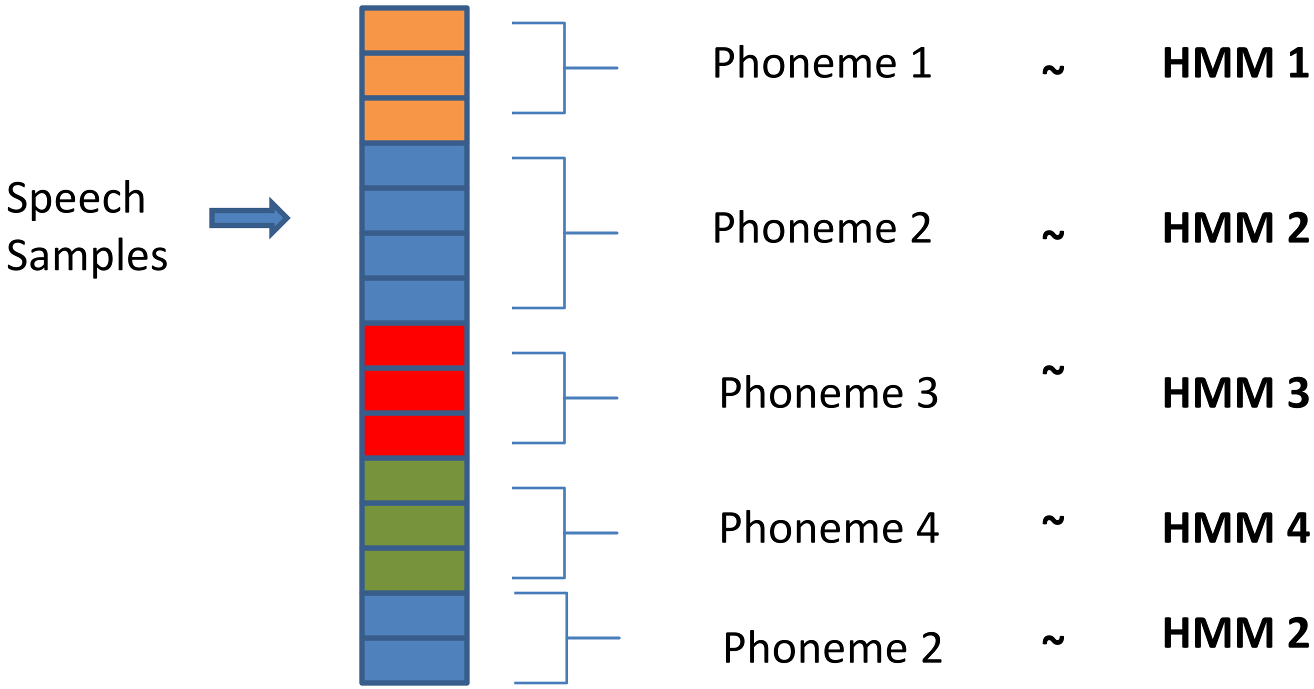
    table = /t/, /a/, /bl/

Agenda

- Modeling the Problem
  - Hidden Markov Model (HMM)
  - Basics of HMM
  - Our model
  - Baum Welch Algorithm
- Learning Approach: Gibbs Sampling
  - Basic Idea
  - Parameter sampling on a 2-D gaussian
  - Extension to estimation of parameters of our model
- Non-parametric case
- Results so far
- Future work

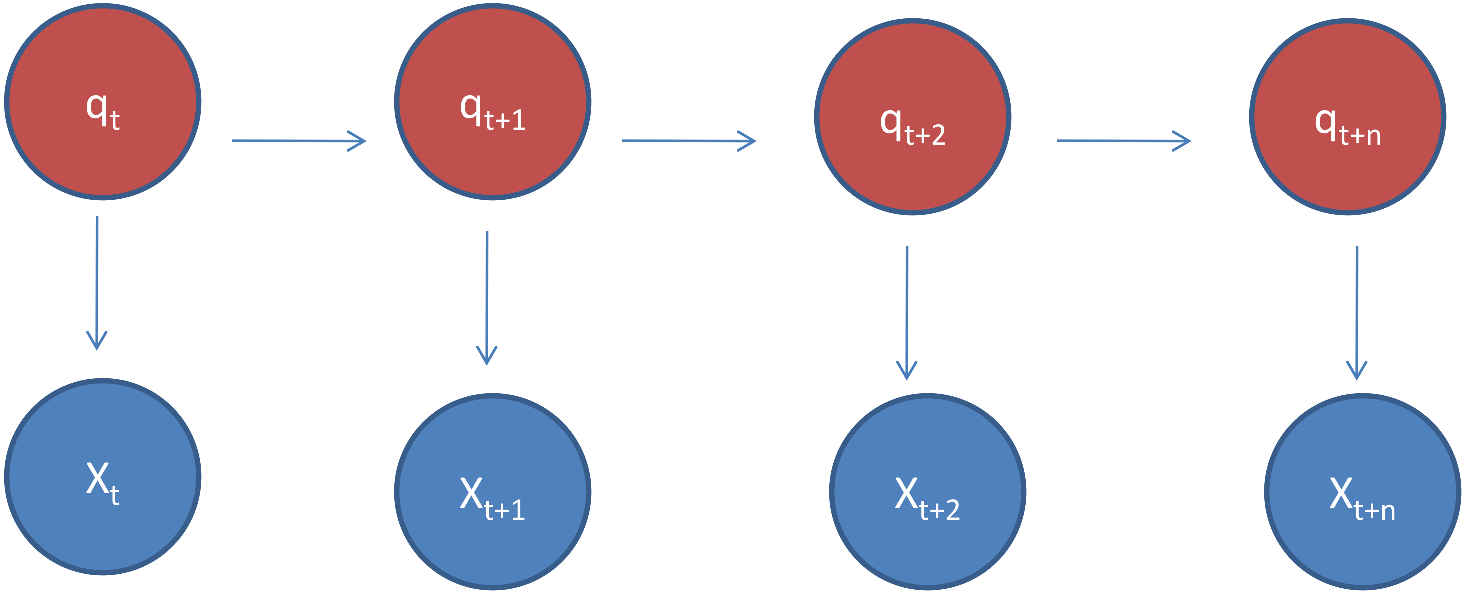
Modeling Phonemes

- Each Phoneme represented by an HMM



Why HMMs ?

1. Model Observations with temporal dependence



$$P(q_{t+1} = s_i \mid q_t = s_j, \ldots\ldots, q_0 = s_0) = P(q_{t+1} = s_i \mid q_t = s_j)$$

$$P(X_{t+1} \mid q_{t+1}, q_t) = P(X_{t+1} \mid q_{t+1})$$

Why HMMs ?

2. Given HMM $\boldsymbol{\theta} = \{\mathbf{A},\mathbf{B},\boldsymbol{\pi}\}$, Observations : $\boldsymbol{O} = \mathbf{O_1,O_2,.....,O_T}$

- Application:

  The Evaluation Problem: Find $P(\boldsymbol{O} | \boldsymbol{\theta}_i)$

- **The Learning Problem**:

  Adjust the Model Parameters $\boldsymbol{\theta} = \{\mathbf{A},\mathbf{B},\boldsymbol{\pi}\}$

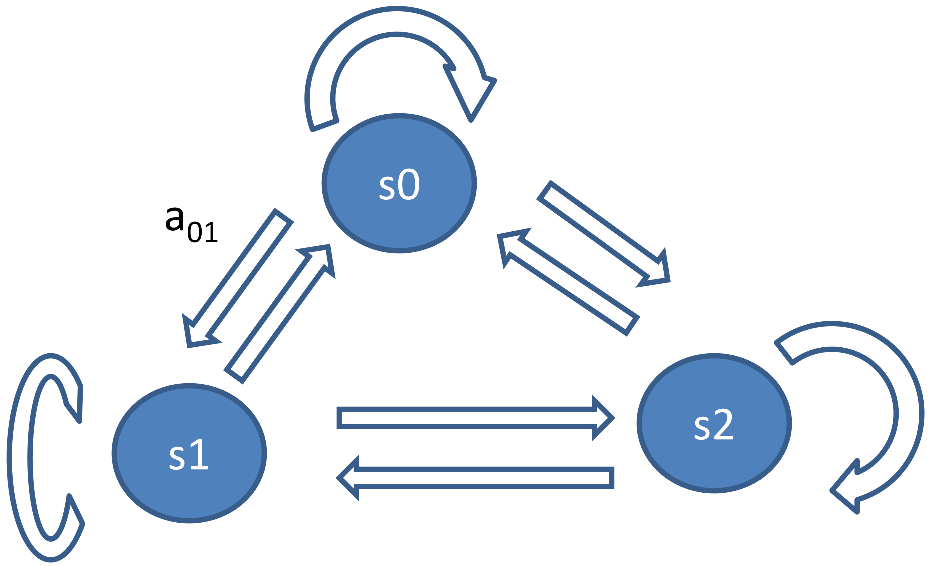
Hidden Markov Model: Markov Chain

- Markov Chain

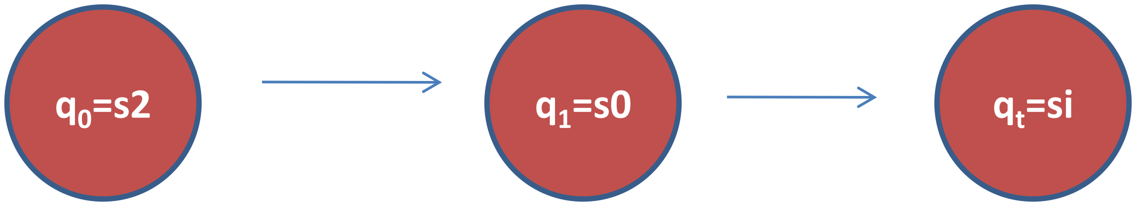
$\mathbf{\Omega}$ = {s0,s1,s2}

A={a_{ij}} = $P(q_{t+1} = s_i \mid q_t = s_j)$

π = {π_i} = $P(q_0 = s_i)$



- Sequence of Random Variables {q_t} on $\mathbf{\Omega}$ for t=0,1,2, ..



Hidden Markov Model: Markov Chain

- 1st Order property:

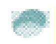
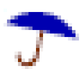
$$P(q_{t+1} = s_i \mid q_t = s_j, \ldots\ldots, q_0 = s_0) = P(q_{t+1} = s_i \mid q_t = s_j)$$

- Problems of type:
  - Given: **A, π**
  - To find: $P(q_t = s_i, q_{t+1} = s_j, q_{t+2} = s_k)$

Hidden Markov Model

- Bayes' Rule & 1st Order Assumption:

$$P(q_1, q_2, \ldots, q_n \mid X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid q_i) . \prod_{i=1}^{n} P(q_i \mid q_{i-1})$$

- Weather Example:
  - States:  ,  , 
  - Observations:
    - ➢ Umbrella 
    - ➢ No Umbrella 

  - Observation probability distribution

    B={b_i(v)} = $P(X_t = v \mid q_t = s_i)$

Weather Example : HMM

A =

Today's Weather	Tomorrow's Weather		
			
	0.8	0.05	0.15
	0.2	0.6	0.2
	0.2	0.3	0.5

B =

Weather	Umbrella Probability
	0.1
	0.8
	0.3

Weather Example : HMM

$$P(q_1 = \;, q_2 = \;, q_3 = \; \mid X_1 = \;, X_2 = \;, X_3 = \;) \;=$$

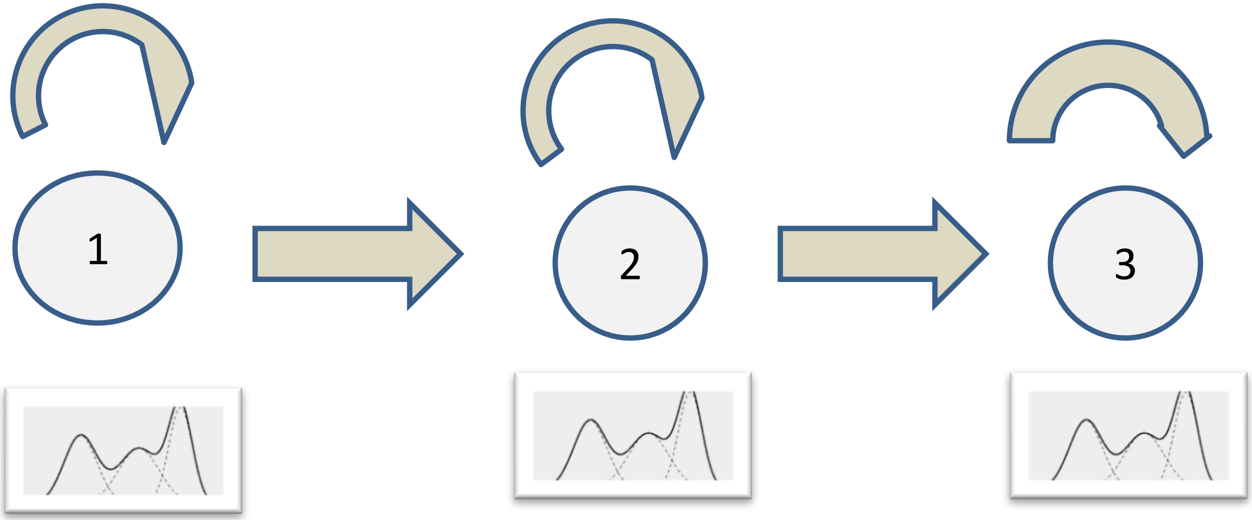
$$\Big\{ P(X_1 = \; \mid q_1 = \;)\; P(X_2 = \; \mid q_2 = \;)\; P(X_3 = \; \mid q_3 = \;) \Big\}$$

$$\times$$

$$\Big\{ P(q_1 = \;)\; P(q_2 = \; \mid q_1 = \;)\; P(q_3 = \; \mid q_2 = \;) \Big\}$$

Our Model

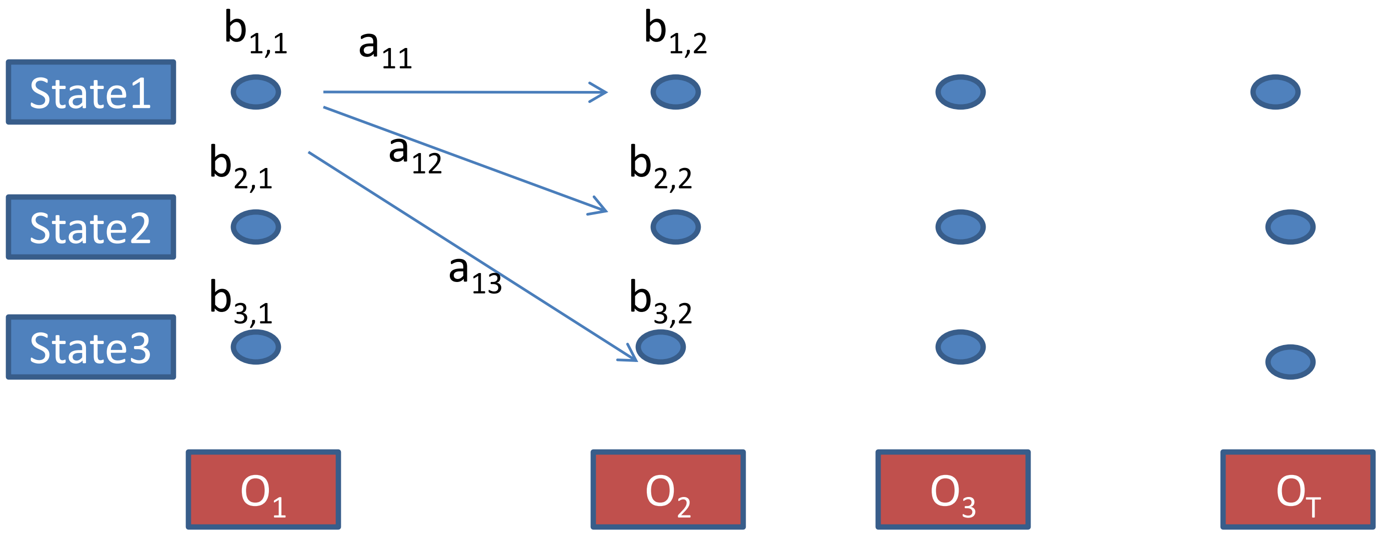
- 1 Phoneme = 1 HMM
- 3 state Left-Right HMM
- Each state Density ~ GMM



$$\Pi = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

Learning Model Parameters: Baum Welch



- Traditional EM Approach: Baum Welch Algorithm
  - E-Step: Evaluate Posterior Probability ~ $P(q_t = s_i \mid O_1, O_2, ..., O_T)$
  - M-Step: Update Parameters using Posterior ~ (π,A,B)

Learning Approach: Gibbs Sampling

- Purpose:
  - To generate samples from a joint distribution
- Idea:
  - generate a sequence of samples from conditional distributions

  ⟹ Stationary Markov Chain to approximate joint distribution

- Distribution to sample: $P(a,b,c)$

  Algorithm:

$$a_i \sim P(a \mid b_{i-1}, c_{i-1})$$
$$b_i \sim P(b \mid a_i, c_{i-1})$$
$$c_i \sim P(c \mid a_i, b_i)$$

  With Initializations: $b = b_0, c = c_0$

Sample from Bivariate Gaussian

- Task: To generate samples from a 2-D Gaussian
- Given: (μ,***σ***)

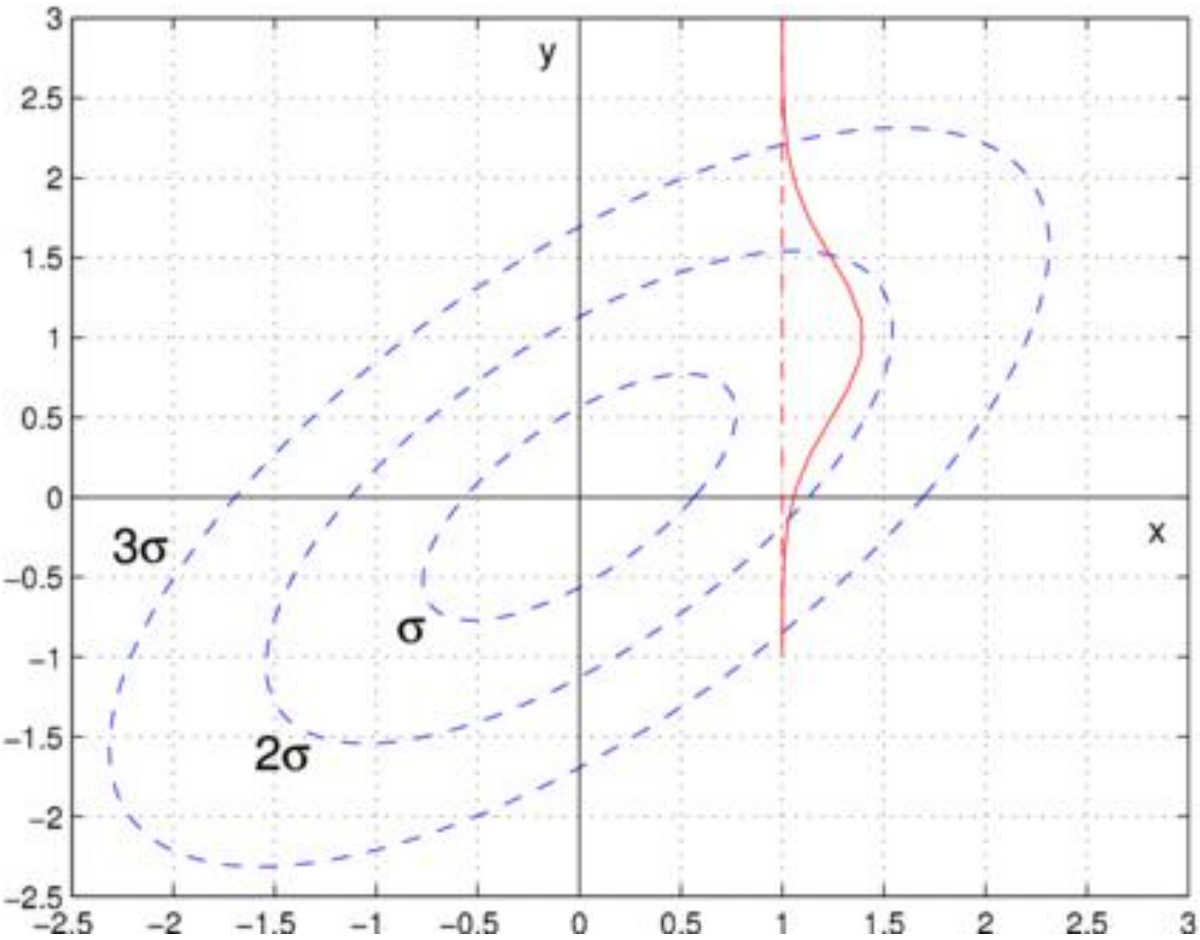
$$P(x_1, x_2) = \mathrm{N}(X; \mu, \Sigma)$$

$$\mu = [\mu_1, \mu_2]$$

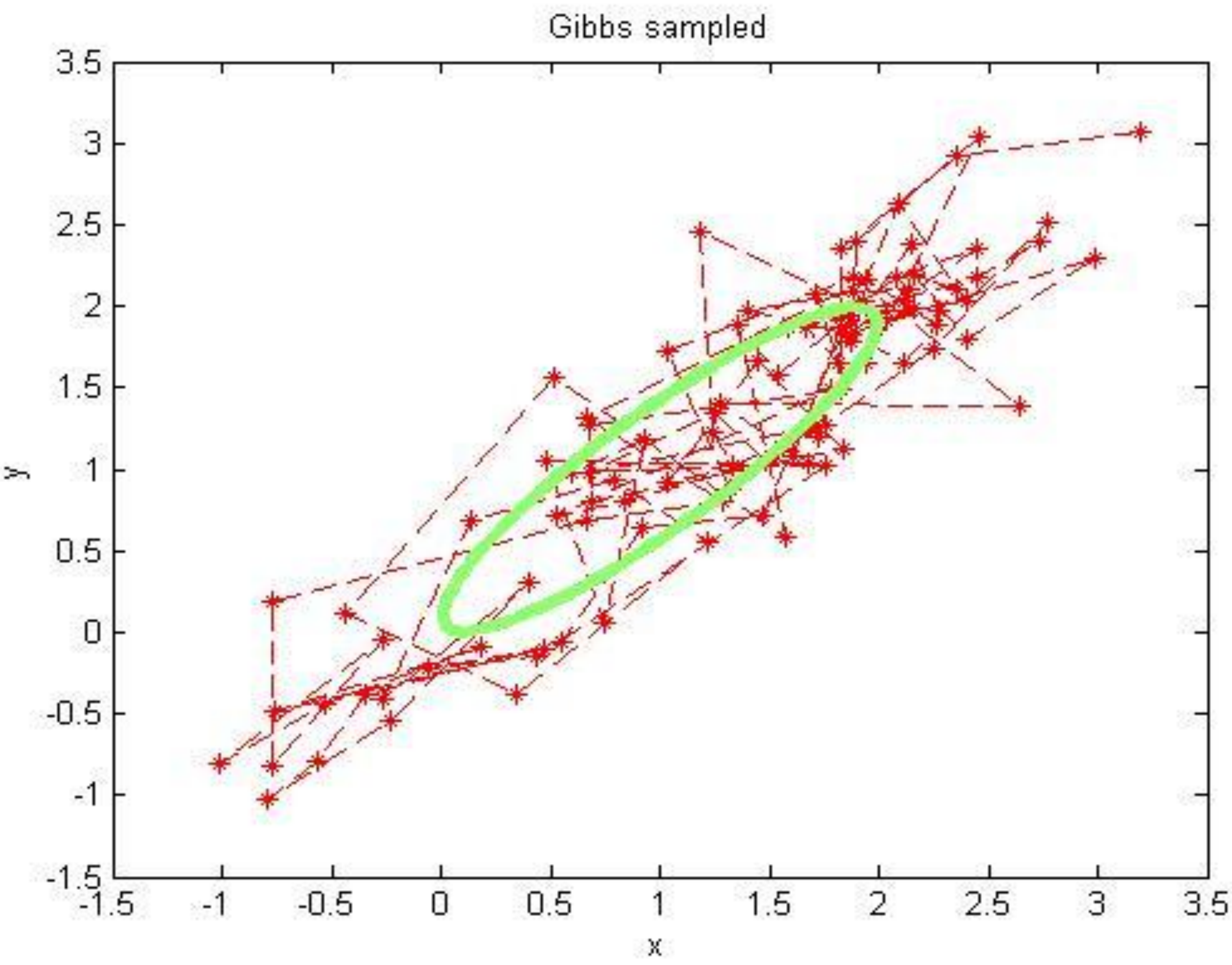
$$\Sigma = \begin{pmatrix} \delta_{1,1}^2 & \delta_{1,2}^2 \\ \delta_{2,1}^2 & \delta_{2,2}^2 \end{pmatrix}$$

$$P(x_i \mid x_j) = \mathrm{N}(x_i; \tilde{\mu}_i, \tilde{\delta}_i)$$

Sample from Bivariate Gaussian: Process



Sample from Bivariate Gaussian: Actual Samples



Sampling Parameters of Gaussian

- Given: Data (X)
- To determine: (μ,λ) for X ~ $\mathcal{N}$(X; μ, λ)
- Assume Normal-Gamma Prior on P(μ, λ) ,i.e.,

$$P(\mu,\lambda) \sim NG(\mu,\lambda \mid \mu_0,\kappa_0,\alpha_0,\beta_0)$$

$$NG(\mu,\lambda \mid \mu_0,\kappa_0,\alpha_0,\beta_0) = \mathrm{N}(\mu \mid \mu_0,(\kappa_0\lambda)^{-1})Ga(\lambda \mid \alpha_0,\beta_0)$$

$$P(\mu \mid \lambda, X) = \mathrm{N}(\mu; \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n}, ((\kappa_0 + n)\lambda^{-1})$$

$$P(\lambda \mid X) = Ga(\lambda; \alpha_0 + n/2, \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\kappa_0 n(\bar{x} - \mu_0)^2}{2(\kappa_0 + n)})$$

HMMs

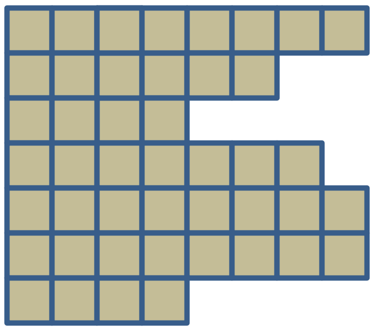
HMM1

HMM2

HMM3

HMM P

Observation Sequences

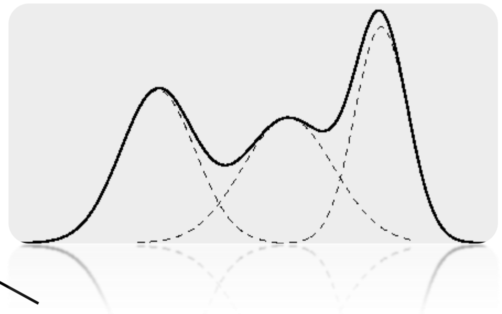


Observation State

1	s1
1	s2
2	s3
2	s4
3	s5
3	s6
3	s7
3	s8

GMM

c1	c2	c3	c4



Parameters:

L_k, Υ, N_p

$s_i,\ a_{ij}, \eta_{ij}$

$\mu_m, \lambda_m, \pi_m, c_n, \beta_m$

Sample HMM Parameters

- Getting HMM Labels for each sequence

SEQ 1 →	HMM 3
SEQ 2 →	HMM 1
SEQ 34 →	HMM 5
SEQ 56 →	HMM P
SEQ 209 →	HMM 7

$\{L_1, L_2, \ldots, L_{34}, \ldots, L_{56}, \ldots, L_{209} \ldots\} \in [1,2,3,\ldots\ldots,P]$

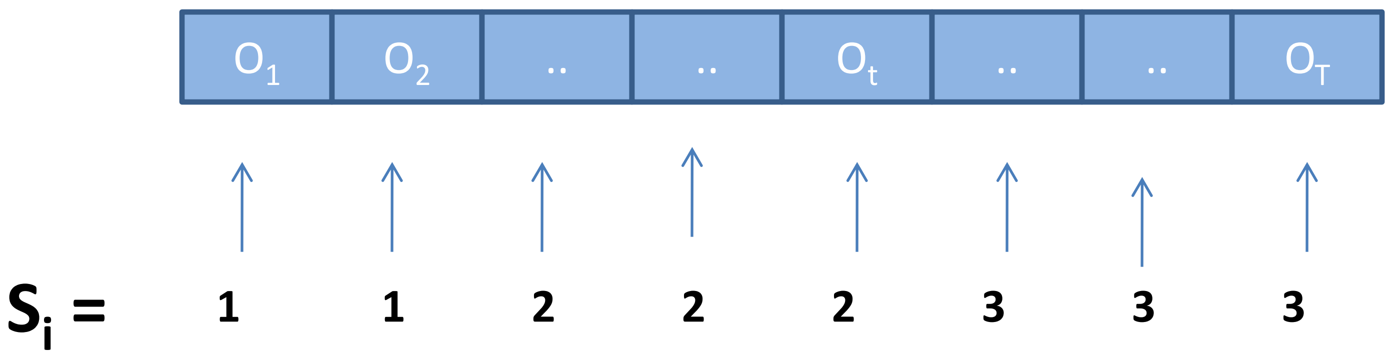
Sample HMM Parameters

- Sampling Equation:

$$P(L_k = p \mid \ldots) = \frac{(N_p + \gamma / P)}{K - 1 + \gamma} P(O_k \mid \theta_p)$$

Multiple States: Sample State Labels

- Getting State Labels for each sample within a Sequence
- Updating transition probabilities



Multiple States: Sample State Labels

- 3 states

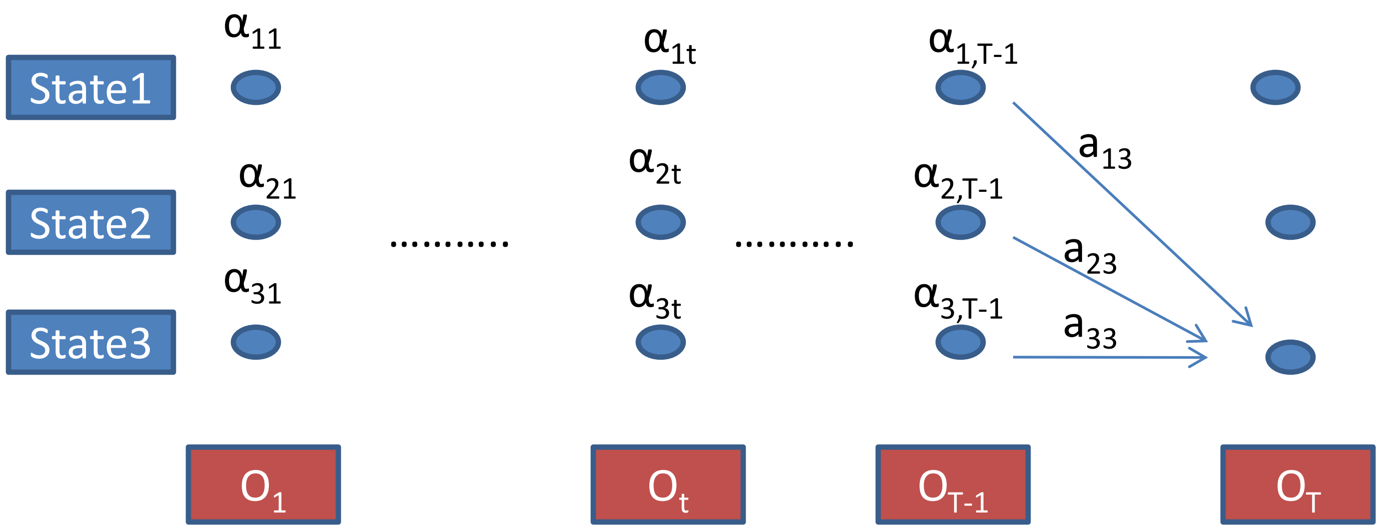
- Block sampling state labels within sequence: s_i
  - Calculate $\alpha_{ti} = P(O_1, O_2, \ldots\ldots, O_t, q_t = s_i \mid \theta_p)$
  - Back Track using transition probabilities to get states

- Transition probabilities $\{a_{ij}\}$

Prior $\longrightarrow$ $P(a_{ij}) \sim Dir(a_{ij}; \eta_0)$

Posterior $\longrightarrow$ $P(a_{ij} \mid \ldots) \sim Dir(a_{ij}; \eta_0 + \eta_{ij})$

Multiple States: Sample State Labels



- Sample S_i from Probabilities:

$$\mathbf{S_i} \sim \begin{bmatrix} \alpha_{1,T-1} * a_{13} \\ \alpha_{2,T-1} * a_{23} \\ \alpha_{3,T-1} * a_{33} \end{bmatrix}$$

Sample GMM Parameters

- Each state represents one GMM
- Sample mixture Labels c_n
- Update $\boldsymbol{\mu}_m, \boldsymbol{\lambda}_m, \pi_m$

	O_1	O_2	..	..	O_T
$c_n =$	m-1	2	m	2	1

Sample GMM Parameters

- Parameters to be sampled: ($\boldsymbol{\mu}$,$\boldsymbol{\lambda}$,c_1,c_2,…….,c_T)
- Priors:

$$P(\mu_m, \lambda_m) \sim NG(\mu_m, \lambda_m \mid \mu_0, \kappa_0, \alpha_0, \beta_0)$$

$$P(\pi) \sim Dir(\pi; \beta)$$

$$P(c_n \mid \pi) \sim Cat(c_n \mid \pi)$$

- Posteriors:

$$P(c_n = m \mid \pi_1, \pi_2, ..., \pi_m, \mu_1, \mu_2, ..., \mu_m, \lambda_1, \lambda_2, ..., \lambda_m, X) =$$

$$\frac{P(X \mid c_n = m) P(c_n = m)}{P(X)}$$

$$P(\pi \mid c_1, c_2, ..., c_n, \mu_1, \mu_2, ..., \mu_m, \lambda_1, \lambda_2, ..., \lambda_m, X) = Dir(\pi, \beta + N_m)$$

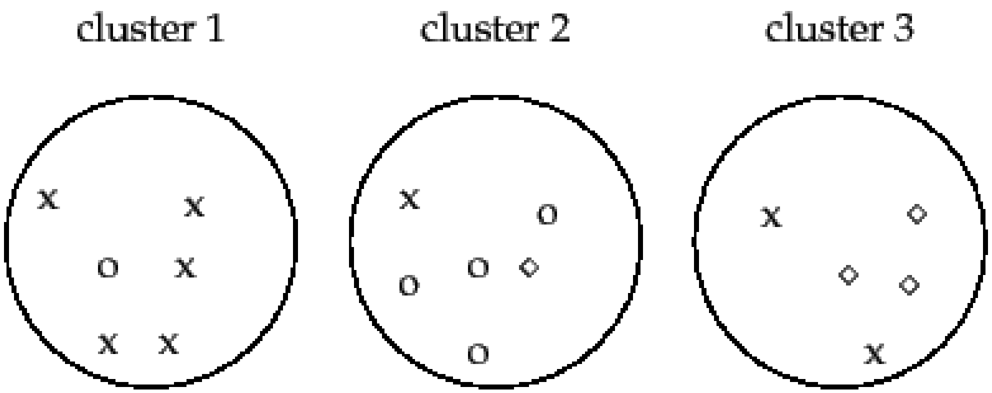
Outlook: Non Parametric

- Number of HMMs unknown

- Each iteration:
  1. Calculate Posterior for existing HMMs
  2. Sample new HMM parameter set $\boldsymbol{\theta}_{new}$
  3. Calculate Posterior for this new HMM
  4. Sample HMM label (L_i) for each sequence
     - ➢ If assigned Label ~ Newly sampled HMM, keep HMM parameters
     - ➢ Else continue with old HMM count

Results

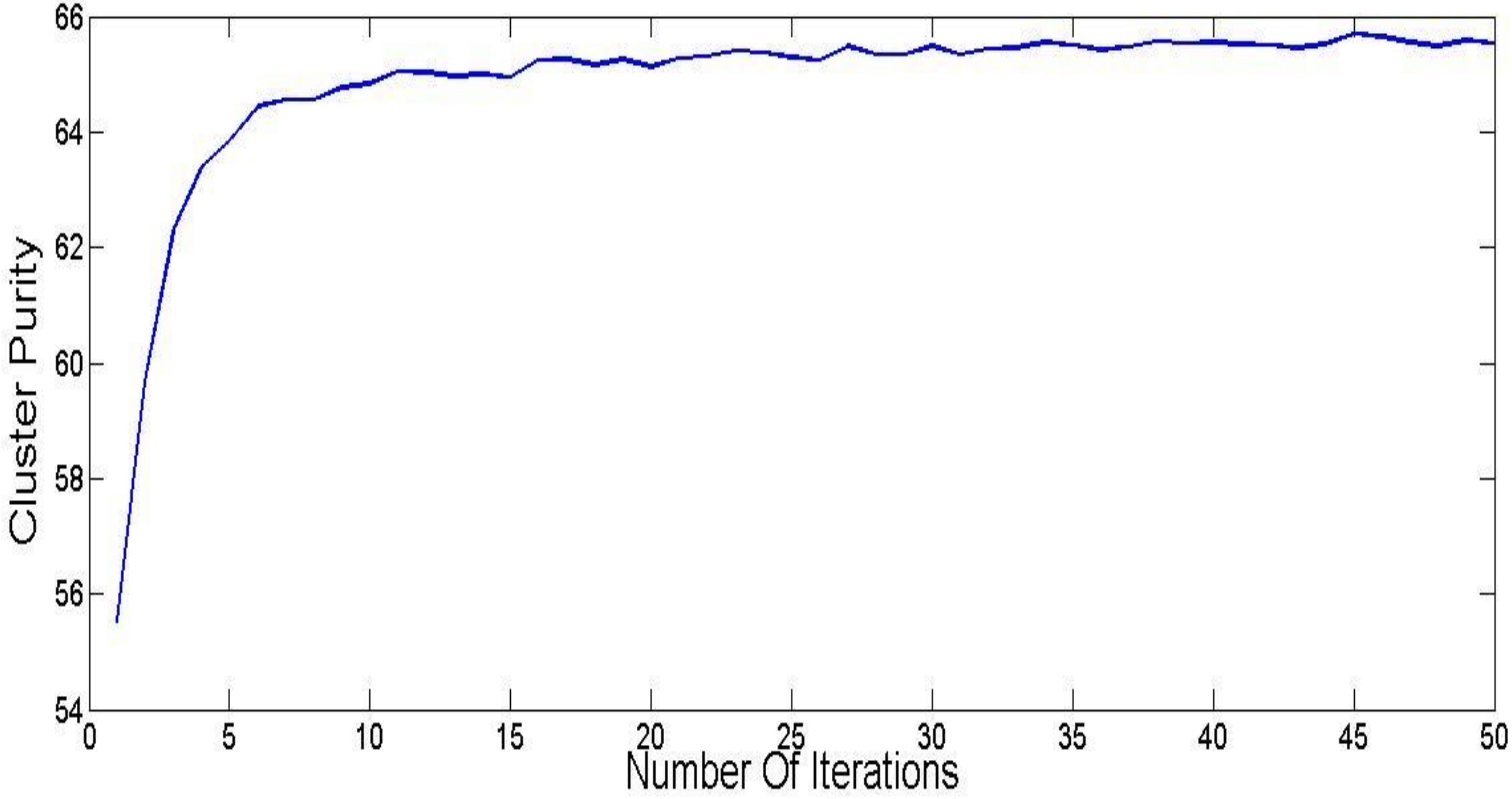
- Evaluation Measure: Cluster Purity



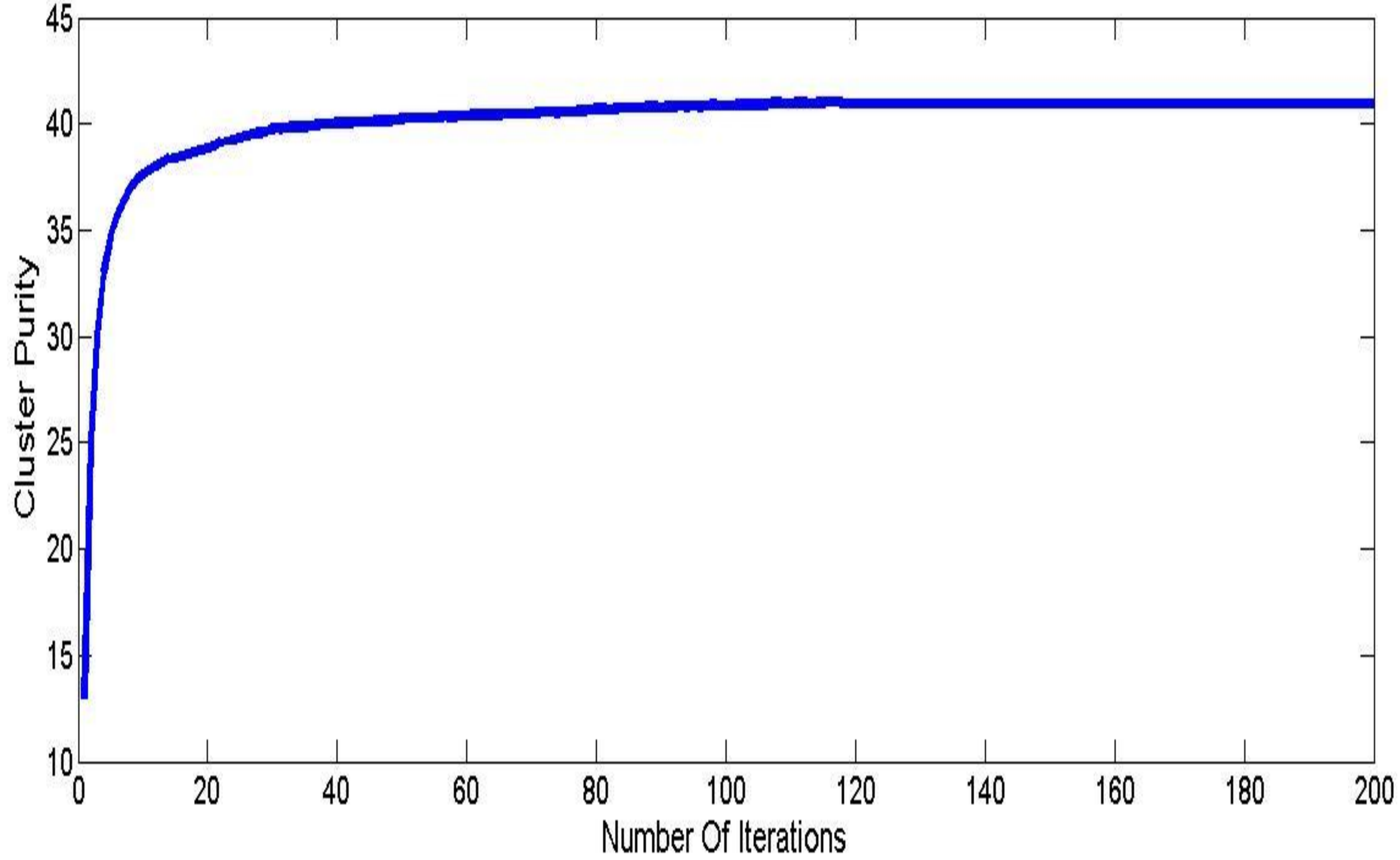
$$Purity = \frac{1}{N}\sum_{k} \max |\omega_k \cap c_j|$$

$$Purity = \frac{1}{17} \times (5+4+3) = 0.71$$

Supervised: Known Number of Classes



Unsupervised: Known Number of Classes



Future Work

- Extension to Non Parametric Case .i.e., unknown number of HMMs.
- Use of Un-segmented Data.

Thank You !

Questions ?