Probability Density Estimation

Overview

- ❑ Parametric Methods
  - ✓ Maximum Likelihood Estimation Method
  - ✓ Bayesian Inference Method
- ❑ Non-parametric Methods
  - ✓ Histogram
  - ✓ Kernel-based Approaches
  - ✓ K-Nearest Neighbor (KNN) Approach
- ❑ Kullback-Leibler (KL) Distance
- ❑ Semi-Parametric Methods : Mixture Models (Gaussian Mixture Models (GMMs))
  - ✓ Maximum Likelihood (Nonlinear Optimization)
  - ✓ The EM Algorithm

Overview (Cont..)

- Parametric Methods
  - ✓ Specific functional form is chosen
  - ✓ Evaluation with new data is faster
- Non-parametric Methods
  - ✓ No specific functional form
  - ✓ Number of parameters grows with number of data points
  - ✓ Very slow for evaluation of new data
- Semi-Parametric Methods
  - ✓ No specific functional form
  - ✓ Number of parameters grows with complexity
  - ✓ Computationally intensive

Parametric Methods

Ex: Normal or Gaussian distribution

Estimator of parameters of a Model

- Maximum likelyhood estimation Method
- Bayesian Inference Method

Normal Distribution (Single Variable)

$$p(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \; ; \; \int_{-\infty}^{\infty} p(x) = 1$$

μ — Mean ; σ = Standard Deviation ; σ^2 = variance

$$\mu = E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx \Rightarrow \text{Expectation of } x$$

$$\sigma^2 = E\left[(x-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx$$

$E[\cdot]$ = Expectation

Parametric Methods (Cont..)

MULTIVARIATE NORMAL DISTRIBUTION

$$P(x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

$\Sigma \rightarrow$ co-variance Matrix of size $d \times d$

$\mu \rightarrow$ Mean vector of size d

$$\mu = E(x) \; ; \quad \Sigma = E\left[(x-\mu)(x-\mu)^T\right]$$

$\Sigma \rightarrow$ Symmetric Matrix

Independent parameters $= (\mu, \Sigma)$

$$= d + \frac{d(d+1)}{2} = \frac{d(d+3)}{2}$$

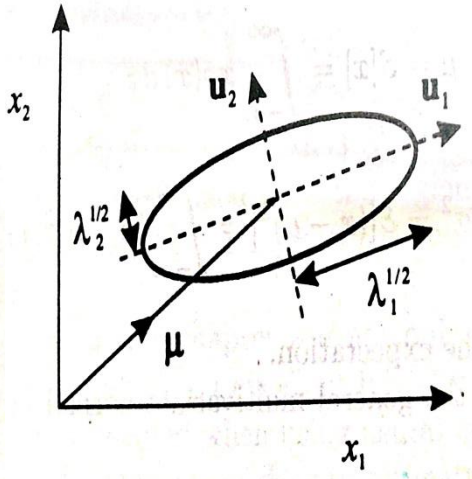
$\Delta^2 = (x-\mu)^T \Sigma^{-1} (x-\mu) \rightarrow$ Mahalanobis distance from x to μ

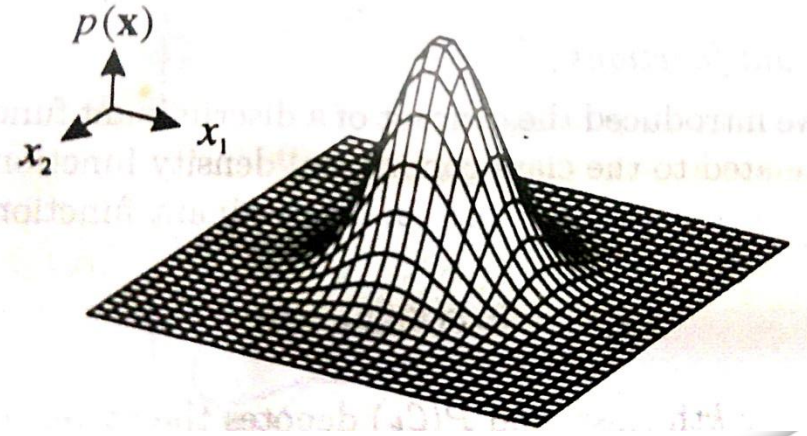
Hyper-ellipsoids (HE)

Principle axes of HE are Eigen vectors

$$\Sigma u_i = \lambda_i u_i$$

$\lambda_i \rightarrow$ Eigen values.





Simplification of parametric Models

Components of x are statistically independent

$\Rightarrow \Sigma \rightarrow$ diagonal Matrix

parameters $= (\mu, \Sigma) \Rightarrow 2d$

$$p(x) = \prod_{i=1}^{d} p(x_i)$$

$\sigma_j = \sigma$ for all j

parameters $= d+1$

$\Rightarrow$ Hyper spheres

Parametric Methods (Cont..)

MAXIMUM LIKELIHOOD ESTIMATION

optimum values of the parameters by maximizing the likelihood function derived from training data.

parameter set $\theta = [\theta_1, \theta_2, \dots \theta_M]$

Data set $X = [x^1, x^2, \dots x^N]$

pdf $x = p(x|\theta)$

$$pdf(X) = \prod_{n=1}^{N} p(x^n|\theta) = L(\theta)$$

$L(\theta)$ = likelihood of θ given X

ML function: $p(X|\theta) = \prod_{n=1}^{N} p(x^n|\theta) = L(\theta)$

Error function $E = -\ln L(\theta) = \sum_{n=1}^{N} \ln p(x^n|\theta)$

$\frac{\partial E}{\partial \theta} = 0 \Rightarrow$ optimal parameters

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x^n \; ; \; \hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x^n - \hat{\mu})^T (x^n - \hat{\mu})$$

Parametric Methods (Cont..)

BAYESIAN INFERENCE

Parameters are represented by pdfs

Prior pdfs are assumed

Posterior parameter pdfs are estimated by observing the dataset.

$$p(x|X) = \int p(x, \theta|X)\, d\theta$$

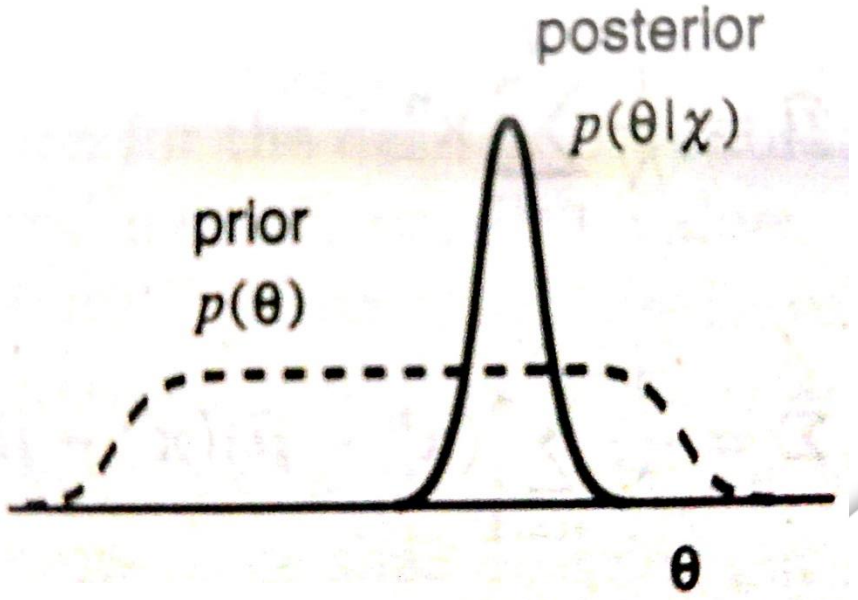
$$p(x, \theta|X) = p(x|\theta, X)\, p(\theta|X)$$

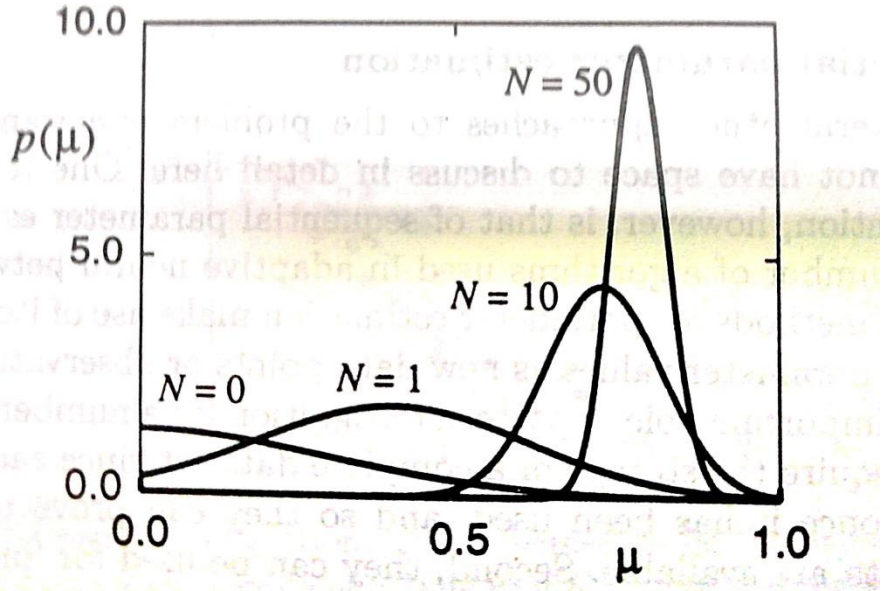
As $p(x|\theta, X)$ is independent of X

$$p(x, \theta|X) = p(x|\theta)\, p(\theta|X)$$

$$p(x|X) = \int p(x|\theta)\, p(\theta|X)\, d\theta$$

pdf of x is weighted average over all potential probabilities of θ.





Parametric Methods (Cont..)

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} = \frac{\prod_{n=1}^{N} p(x^n|\theta)\,p(\theta)}{p(x)} = \frac{p(\theta)}{p(x)} \prod_{n=1}^{N} p(x^n|\theta)$$

Conjugate Priors: $\{p(x|\theta), p(\theta)\} \Longleftrightarrow p(\theta|x)$

Initial pdf of μ = $p_0(\mu) = \frac{1}{(2\pi\sigma_0^2)^{\frac{1}{2}}} \exp\left[\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$

Posterior pdf of μ = $p_N(\mu|x) = \frac{p_0(\mu)}{p(x)} \prod_{n=1}^{N} p(x^n|\mu)$

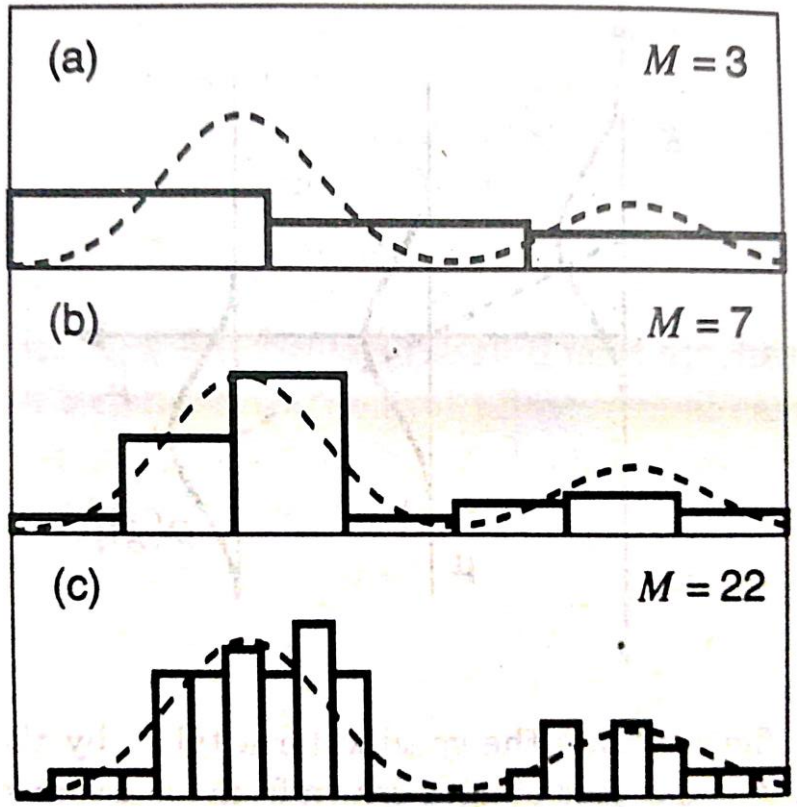
$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x^n$$

Non-Parametric Methods

- Histograms
- Kernel-based approaches
- K-Nearest Neighbor (KNN)
- Histograms (Simple Models)
  - Role of smoothing parameters
    - ✓ Smoothing parameter → Number of bins
    - ✓ Lower number of bins (smooth out crucial details)
    - ✓ Higher number of bins (spiky& noisy)
  - Advantages
    - ✓ Can be constructed sequentially, entire data is not required at one shot. No storage issues.
  - Disadvantages
    - ✓ Estimated density is not smooth
    - ✓ Curse of dimensionality with high-dimensionality data



Non-Parametric Methods : Density Estimation

Density Estimation in General

P – prob of vector x drawn from unknown density $p(x)$ will fall inside some region R of x-space is

$$P = \int_R p(x')\,dx'$$

Prob of k points fall within region R out of N points drawn

$$P_r(k) = \frac{N!}{k!\,(N-k)!}\,P^k\,(1-P)^{N-k} \Rightarrow \text{Binomial law}$$

Mean of k points fall in there region is given by

$$\text{Mean} = E\left[k/N\right] = P \simeq \frac{k}{N}$$

$$\text{Variance} = E\left[\left(\frac{k}{N} - P\right)^2\right] = \frac{P(1-P)}{N}$$

if $p(x)$ is continuous $\Rightarrow P = \int_R p(x')\,dx' \simeq p(x)\,V$

$$\Rightarrow \frac{k}{N} = p(x)\,V \Rightarrow \boxed{p(x) = \frac{k}{NV}}$$

Non-Parametric Methods : Density Estimation

Practical Density Estimation

$P \simeq \frac{k}{N}$ => Region "R" to be relatively Large => P will be large.

$P \simeq p(x)V$ => $p(x) = \frac{P}{V} = \frac{k}{NV}$ => $p(x)$ is almost constant when "R" is Relatively Small.

Choice of R for the best estimate of $p(x)$

k- Nearest Neighbor (kNN)

Fix the value of k => Det the corresponding volume "V" from data

Kernel-based density

Fix the volume V => Det "k" from data

For infinite "N" => Both kNN & Kernel based Methods will converge.

Non-Parametric Methods : Kernel-based Method

Kernel Based Methods

Region "R" to be hypercube with sides of length "h" centered on the point x in d-dimensional space

Volume of the hypercube $V = h^d$

Kernel function $H(u) = \begin{cases} 1 & |u_j| < \frac{1}{2} \quad j = 1, 2, \ldots d. \\ 0 & \text{otherwise} \end{cases}$

$H\left[\frac{x - x^n}{h}\right] = 1$; if x^n falls within hypercube

$= 0$ otherwise

$$K = \sum_{n=1}^{N} H\left[\frac{x - x^n}{h}\right]$$

$$\hat{p}(x) = \frac{K}{NV} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^d} H\left[\frac{x - x^n}{h}\right]$$

Estimated Density = Superposition of N cubes of side "h", with each cube centered on one of the data points

Cubes Vs bins in case of Histogram

Non-Parametric Methods : Kernel-based Method

Kernel - Based Methods (cont...)

Smoothing the density by choosing different forms of kernel function $H(u)$

Multivariate Normal (Gaussian) kernel

$$\tilde{p}(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{d/2}} \exp\left[-\frac{\|x - x^n\|^2}{2h^2}\right]$$

$$E\left[\tilde{p}(x)\right] = \frac{1}{N} \sum_{n=1}^{N} E\left[\frac{1}{h^d} H\left(\frac{x - x'}{h}\right)\right]$$

$$= E\left[\frac{1}{h^d} H\left(\frac{x - x'}{h}\right)\right]$$

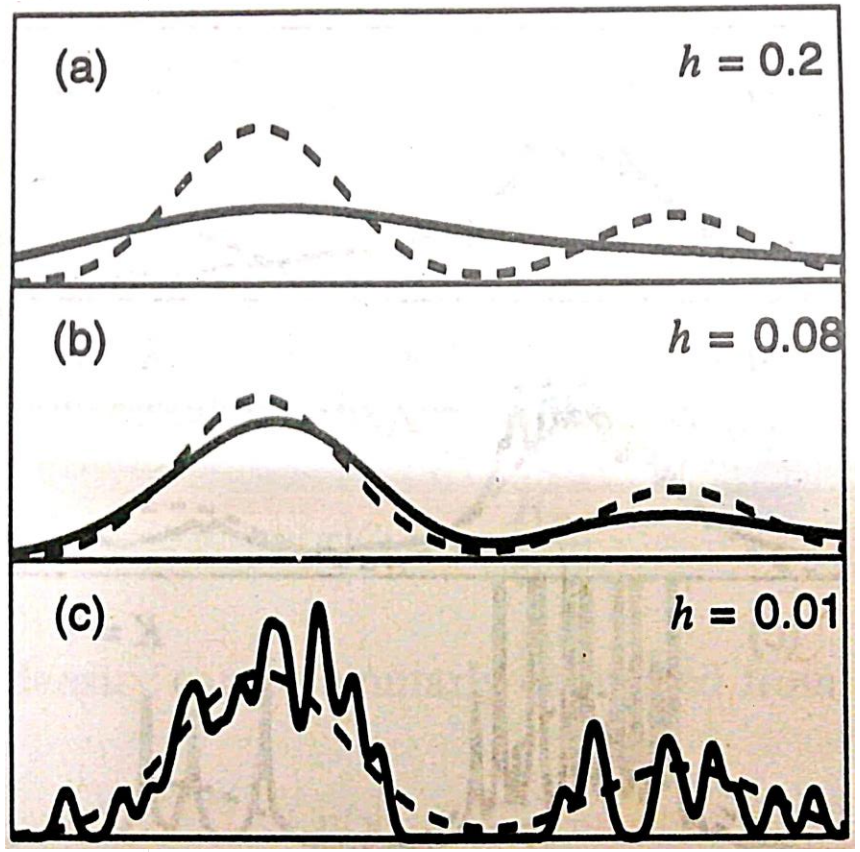
$$= \int \frac{1}{h^d} H\left(\frac{x - x'}{h}\right) p(x')\, dx'$$

Convolution of true density with kernel function

=> Smoothed version of true density.

$h \to 0$ => kernel fn. is delta fn. => $\tilde{p}(x) = p(x)$

For finite sample size, small values of h leads to noisy representation of $\tilde{p}(x)$



Non-Parametric Methods : Kernel-based Method

- Drawbacks
  - ✓ All data points to be stored
  - ✓ Evaluation of density is very slow if # data points is large
  - ✓ It gives biased estimate of density
- Solution to fast density estimation
  - ✓ Use of fewer kernel functions
  - ✓ Adapt their positions and width as per the data

Non-Parametric Methods : K-Nearest Neighbor

- Problem with Kernel-based approach → Fixed “h” for all data points
- Remedy → KNN : Fix “K” value & allow “V” to vary
- Disadvantages of K-Nearest Neighbor (KNN)
  - ✓ Estimated PD is not true density, because integral over all X-space diverges.
  - ✓ All the training data points must be retained.
  - ✓ Large computer storage is required to store the whole data
  - ✓ Require large amount of processing to evaluate the density at new data points.

Non-Parametric Methods : KNN-Classifier

KNN : classifier

N – Total number of data points
N_k – # points of class C_k

$$N = \sum_{k=1}^{K} N_k$$

For estimation of $p(x|C_k)$

generate a hyper-sphere around "x" which contains "K" points

Among "K" points, K_k corresponds to class C_k

$$p(x|C_k) = \frac{K_k}{N_k V} \quad \text{— class conditional density}$$

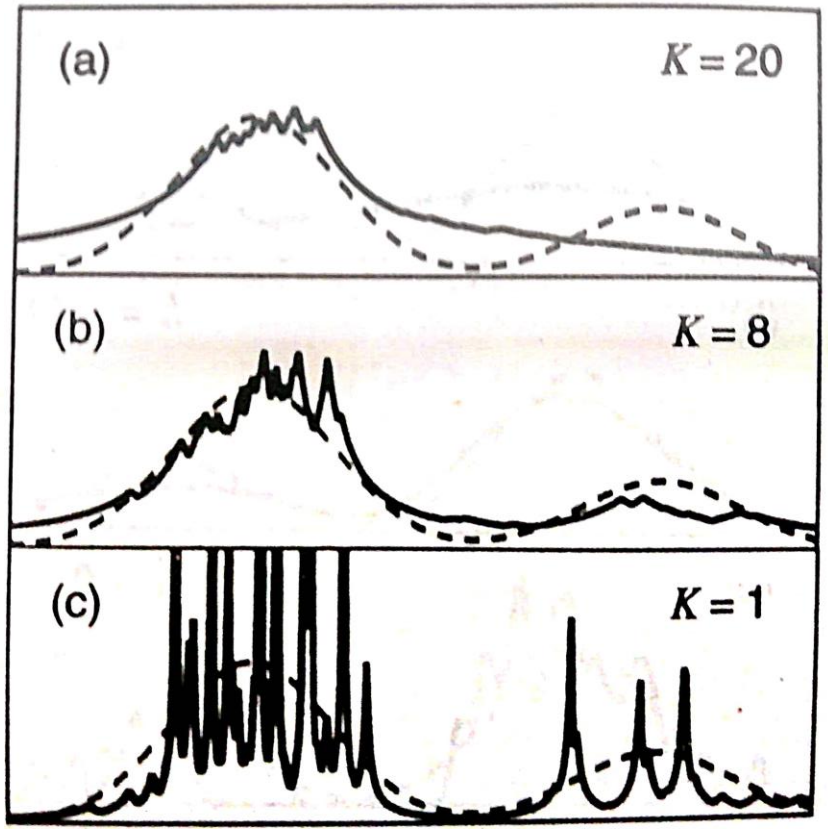
$$p(x) = \frac{K}{NV} \quad \text{— uncconditional density}$$

$$p(C_k) = \frac{N_k}{N}$$

$$p(C_k|x) = \frac{p(x|C_k)\,p(C_k)}{p(x)} = \frac{\frac{K_k}{N_k V} \cdot \frac{N_k}{N}}{\frac{K}{NV}} = \frac{K_k}{K}$$

To minimize the prob of misclassification, the vector X is assigned to class "C_k" for which the ratio of $\frac{K_k}{K}$ is largest

K=1 => Nearest Neighbour



Non-Parametric Methods : Smoothing Parameters

Smoothing Parameter

Histogram — width of the bins & # bins

Kernel Methods — width of the kernel "h"

KNN — Value of K

oversmoothed — high bias — poor estimator

Insufficient smoothing — high variance — Noisy density —

Terms in a polynomial & # hidden units in NN,

$$L(h) = \prod_{n=1}^{N} p(x^n \mid h ; x^1, x^2, \ldots x^N)$$

$L(h)$ is maximized as $h \to 0$ (impulse at each data point)

$$E\left[-\ln L(h)\right] = -\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \ln \tilde{p}(x^n) = -\int p(x) \ln \tilde{p}(x)\, dx$$

Distance bet true & estimated density

$$L = \int p(x) \ln p(x) - \int p(x) \ln \tilde{p}(x) = -\int p(x) \ln \frac{\tilde{p}(x)}{p(x)}\, dx$$

K-L distance : Kullback - Leibler distance

Distance is weighted by $p(x)$

Semi-Parametric Models : Mixture Models

- Parametric Methods
  - ✓ Specific functional form is chosen
  - ✓ Evaluation with new data is faster
- Non-parametric Methods
  - ✓ No specific functional form
  - ✓ Number of parameters grows with number of data points
  - ✓ Very slow for evaluation of new data
- Semi-Parametric Methods
  - ✓ No specific functional form
  - ✓ Number of parameters grows with complexity
  - ✓ Computationally intensive

Semi-Parametric Models : Mixture Models (Cont..)

Mixture Models (cont..)

$$p(x) = \sum_{j=1}^{M} p(x|j)\,p(j)$$ — linear combination of component densities

$$\sum_{j=1}^{M} p(j) = 1 \qquad 0 \le p(j) \le 1$$

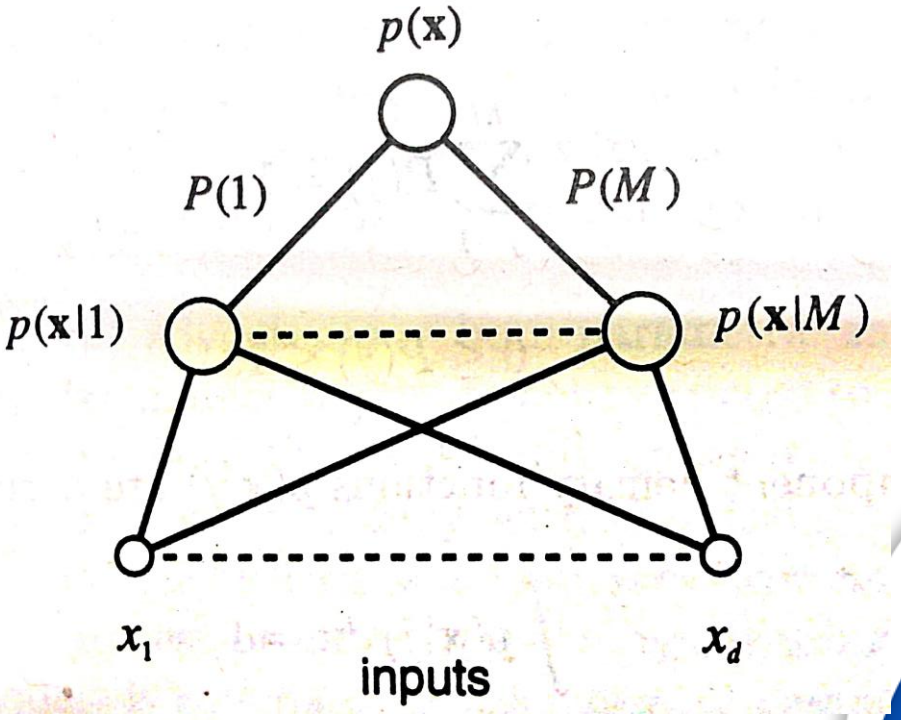
$$\int p(x|j) = 1$$

$p(x)$ — Mixture distribution

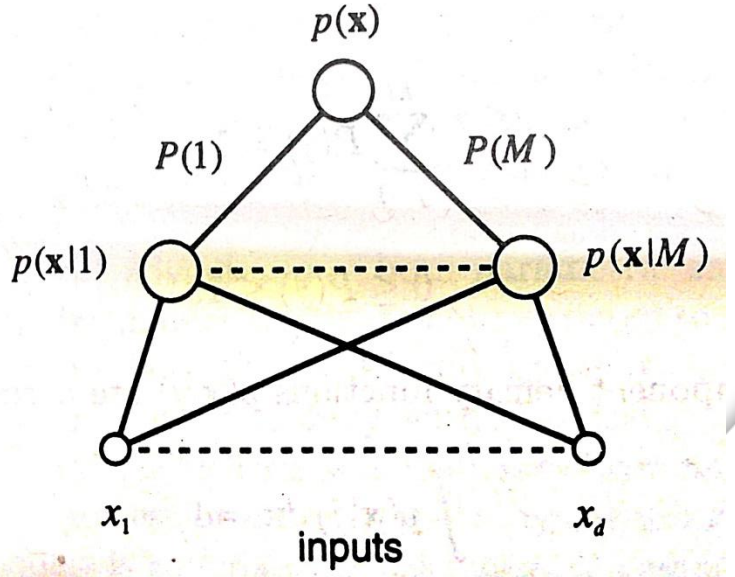
$p(j)$ — Mixing parameter (weight)

$p(x|j)$ — Component density

Property: For many choices of component density fns, they can approximate any continuous density to arbitrary accuracy provided models sufficiently large components & provided the parameters of the model are correct



Semi-Parametric Models : Mixture Models (Cont..)



Incomplete Data : Component labels are not known

Classification: $p(x|c_k)$ — class specific model

From Bayes Theorem: $P(j|x) = \frac{p(x|j)\,p(j)}{p(x)}$

$$\sum_{j=1}^{M} p(j|x) = 1$$

$$p(x|j) = \frac{1}{(2\pi\sigma_j^2)^{\frac{d}{2}}} \exp\left\{-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right\}$$

Mixture Models : Maximum Likelihood Technique

Maximum Likelihood Technique

$$E = -\ln(L) = -\sum_{n=1}^{N} \ln p(x^n) = -\sum_{n=1}^{N} \ln \left[\sum_{j=1}^{M} p(x^n | j)\, p(j) \right]$$

Maximizing Likelihood => Minimizing the Error fn.

$\exists$ parameters => Likelihood (L) $\to \infty$

$x = \mu_j$; $\sigma_j = 0$ => $L \to \infty$

Singular Solutions

Local Minima : Small group of pts close together
poor rep of true distribution

Singular Solns Can be avoided by
- Constrain the Components have equal co-variance
- Shrinking gaussian replaced by Large ones during iteration

Standard Non-linear optimization Method for computing the minima — gradient info.

Mixture Models : ML Technique (Cont..)

ML Technique (cont..)

$$\frac{\partial E}{\partial \mu_j} = \sum_{n=1}^{N} P(j|x^n)\frac{(\mu_j - x^n)}{\sigma_j^2}$$

$$\frac{\partial E}{\partial \sigma_j} = \sum_{n=1}^{N} P(j|x^n)\left\{\frac{d}{\sigma_j} - \frac{\|x^n - \mu_j\|^2}{\sigma_j^3}\right\}$$

$$P(j) = \frac{\exp(r_j)}{\sum_{k=1}^{M}\exp(r_k)} \qquad -\infty \le r_j \le \infty$$

$$\frac{\partial E}{\partial r_j} = -\sum_{n=1}^{N}\{P(j|x^n) - P(j)\}$$

$$\frac{\partial E}{\partial \mu_j} = 0 \Rightarrow \hat{\mu}_j = \sum_n P(j|x^n)x^n \Big/ \sum_n P(j|x^n)$$

$$\frac{\partial E}{\partial \sigma_j} = 0 \Rightarrow \sigma_j^2 = \frac{1}{d}\left[\frac{\sum_n P(j|x^n)\|x^n - \hat{\mu}_j\|^2}{\sum_n P(j|x^n)}\right]$$

$$\frac{\partial E}{\partial r_j} = 0 \Rightarrow \hat{P}_j = \frac{1}{N}\sum_n P(j|x^n)$$

Mixture Models : The EM Algorithm

The EM Algorithm

No direct soln to compute parameters $\hat{\mu}_J$, $\hat{\sigma}_J^2$ & $\hat{P}(J)$

Highly non-linear coupled eqns => Difficult to solve.

Iterative Scheme to find minimum of E

- Initialize with (guess) some parameters known as "old"
- Evaluate RHS of $\hat{\mu}_J$, $\hat{\sigma}_J^2$ & $\hat{P}(J)$ => New parameters
- New parameters => Error fn ↓
- New parameters ⟶ old parameters

Repeat above steps until local min is reached.

$$E^{new} - E^{old} = -\sum_n \ln\left[\frac{p^{new}(x^n)}{p^{old}(x^n)}\right]$$

$p^{new}(x)$ & $p^{old}(x)$ => Densities evaluated using new & old parameters respectively.

$$E^{new} - E^{old} = -\sum_n \ln\left[\frac{\sum_J p^{new}(J)\, p^{new}(x^n|J)}{p^{old}(x^n)} \cdot \frac{p^{old}(J|x^n)}{p^{old}(J|x^n)}\right]$$

Mixture Models : The EM Algorithm (Cont..)

The EM Algorithm (cont..)

using Jensen's inequality $\lambda_J \geq 0$ & $\sum_J \lambda_J = 1$

$$\ln\left(\sum_J \lambda_J x_J\right) \geq \sum_J \lambda_J \ln(x_J)$$

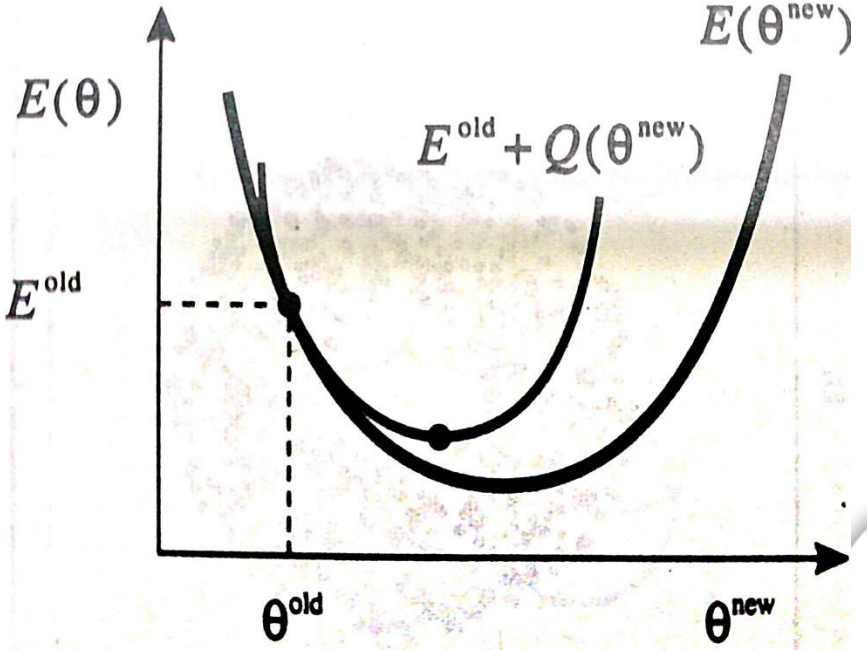
Applying JE on $E^{new} - E^{old}$

$$E^{new} - E^{old} \leq -\sum_n \sum_J p^{old}(J/x^n) \ln\left[\frac{p^{new}(J)\, p^{new}(x^n/J)}{p^{old}(x^n)\, p^{old}(J/x^n)}\right]$$

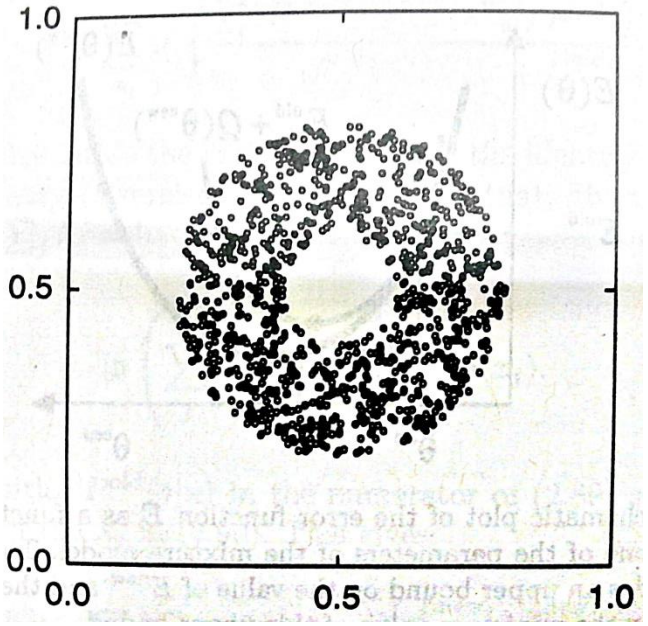
$$\mu_J^{new} = \frac{\sum_n p^{old}(J/x^n)\, x^n}{\sum_n p^{old}(J/x^n)}$$

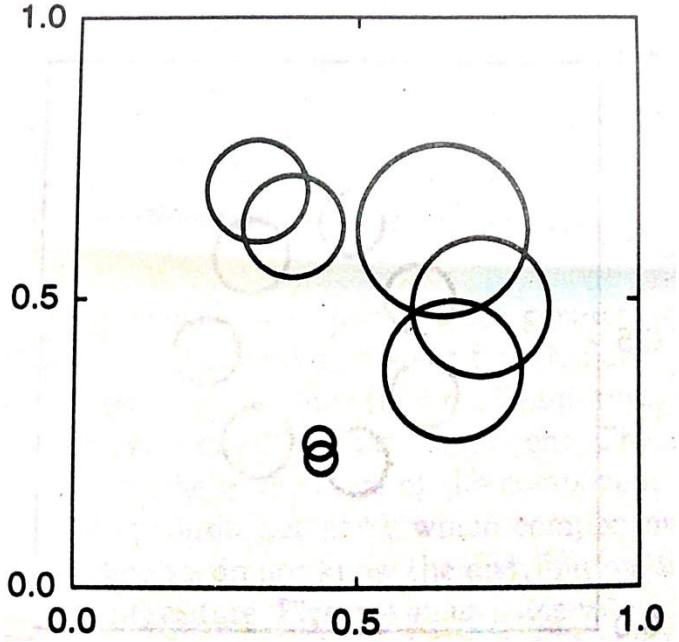
$$(\sigma_J^{new})^2 = \frac{1}{d}\left[\frac{\sum_n p^{old}(J/x^n)\, \|x^n - \mu_J^{new}\|^2}{\sum_n p^{old}(J/x^n)}\right]$$

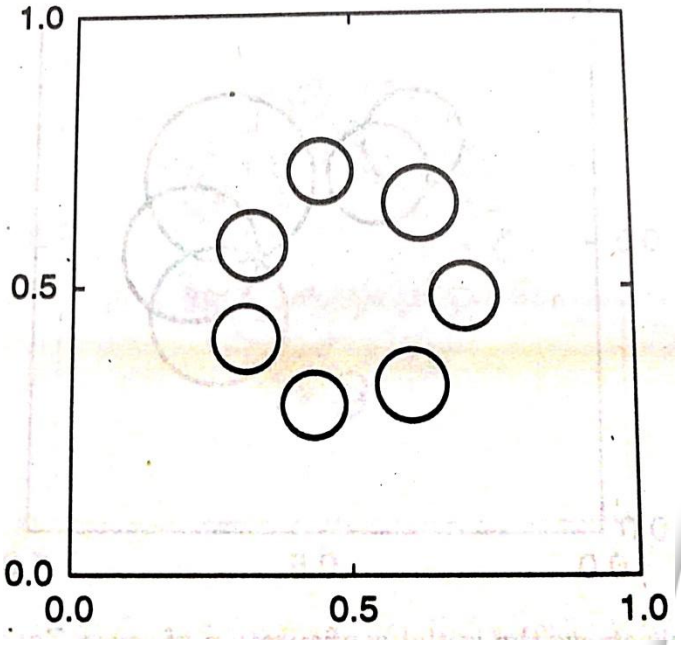
$$P^{new}(J) = \frac{1}{N}\sum_n p^{old}(J/x^n)$$

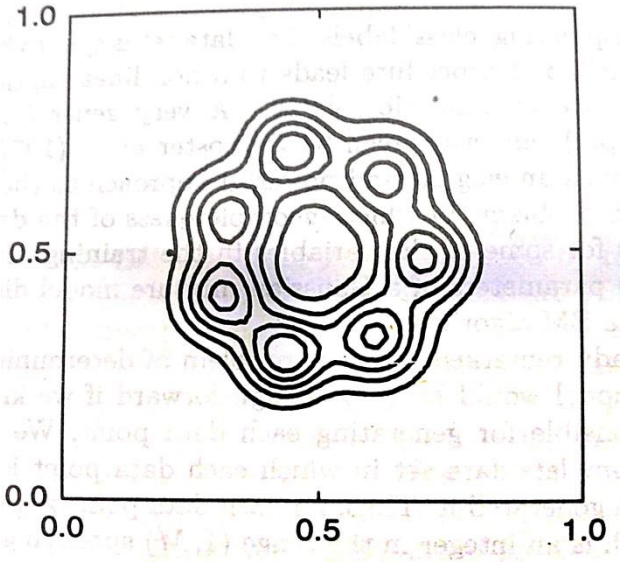


Mixture Models : The EM Algorithm (Example)









Mixture Models : The EM Algorithm (Incomplete Data)

EM Algorithm for Incomplete Data

Complete
Hypothetical Dataset $\Rightarrow [x^n : z^i]$; $n = 1, 2, \cdots N$; $i = 1, 2, \cdots M$

$z^n \rightarrow$ Label of the Mixture component $(1, M)$

$$E^{Comp} = -\ln L^{Comp} = \sum_{n=1}^{N} -\ln p^{new}(x^n, z^n)$$

$$= -\sum_{n=1}^{N} \ln \left[p^{new}(z^n)\, p^{new}(x^n / z^n) \right]$$

$\therefore$ we don't have the PD of z^n, we estimate by EM algo

- Guess some parameters of Mixture Model (old values)
- use the above parameters & Baye's Theorem find PD of z^n
- Then Complete Expectation of E^{Comp} $E[E^{Comp}]$

E-Step $\Rightarrow$ $E[E^{Comp}]$

- New values of parameters are determined by minimizing $E[E^{Comp}]$ w.r.t parameters

Mixture Models : The EM Algorithm (Incomplete Data)

EM Algorithm – Incomplete Data (cont.)

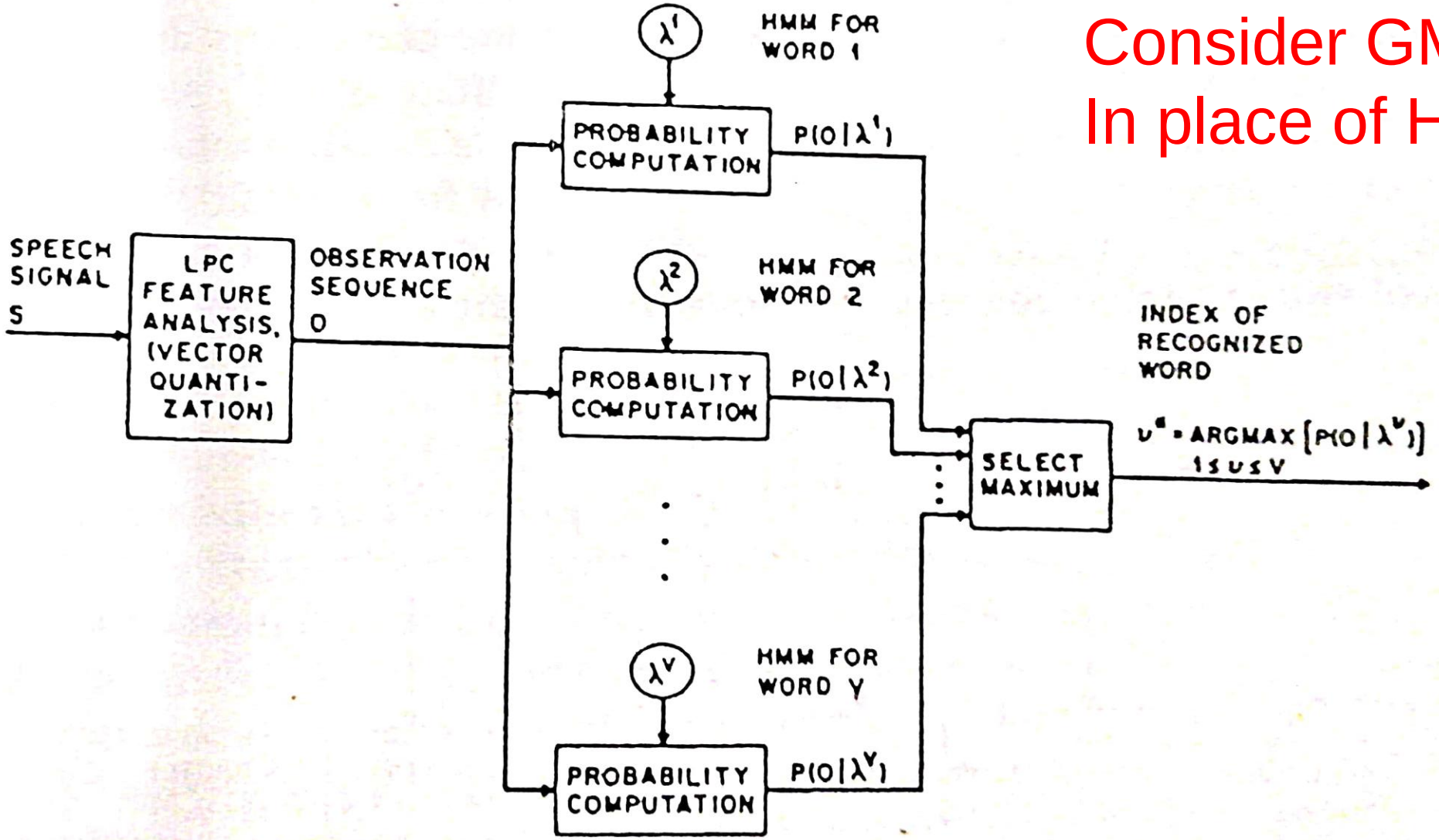
Minimization of $E\left[E^{comp}\right]$ => Maximization of $E\left[L^{comp}\right]$

M-Step => Maximization of $E\left[L^{comp}\right]$

$$E\left[E^{comp}\right] = \sum_{z^1=1}^{M} \sum_{z^2=1}^{M} \cdots \sum_{z^n=1}^{M} E^{comp} \prod_{n=1}^{N} p^{old}\left(z^n \mid x^n\right)$$

$$= -\sum_{n=1}^{N} \sum_{j=1}^{M} p^{old}\left(j \mid x^n\right) \ln\left[p^{new}(j)\, p^{new}\left(x^n \mid j\right)\right]$$

Illustration of Use of GMMs : Isolated Word Rec



Consider GMMs
In place of HMMs