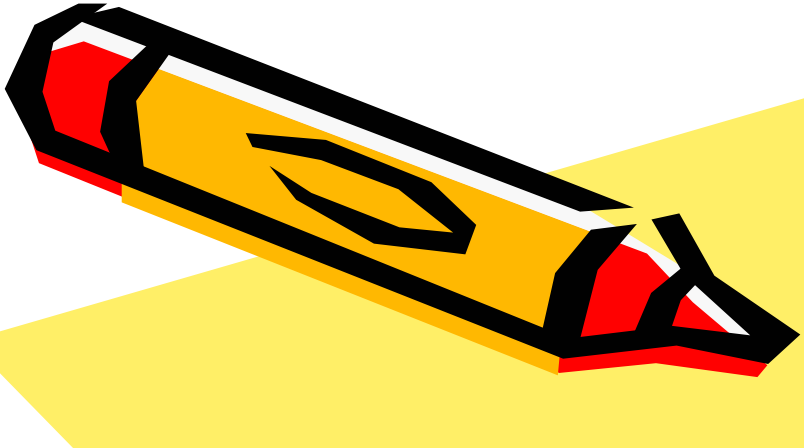
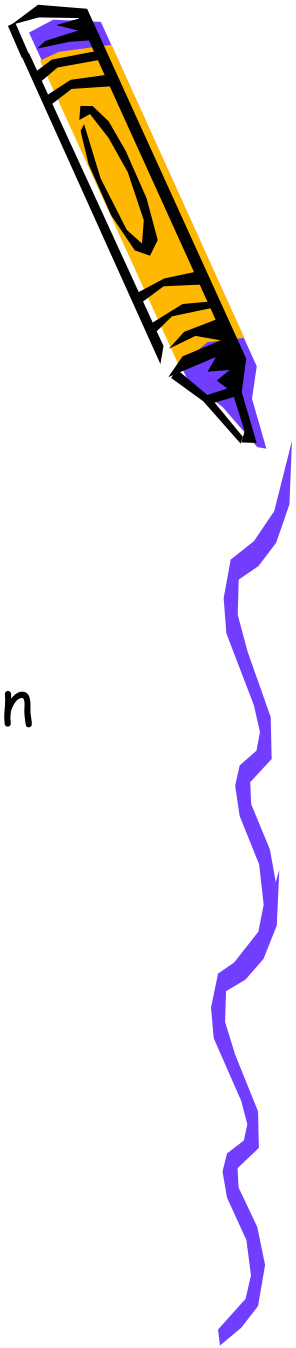


Adaptation Techniques for Acoustic Models

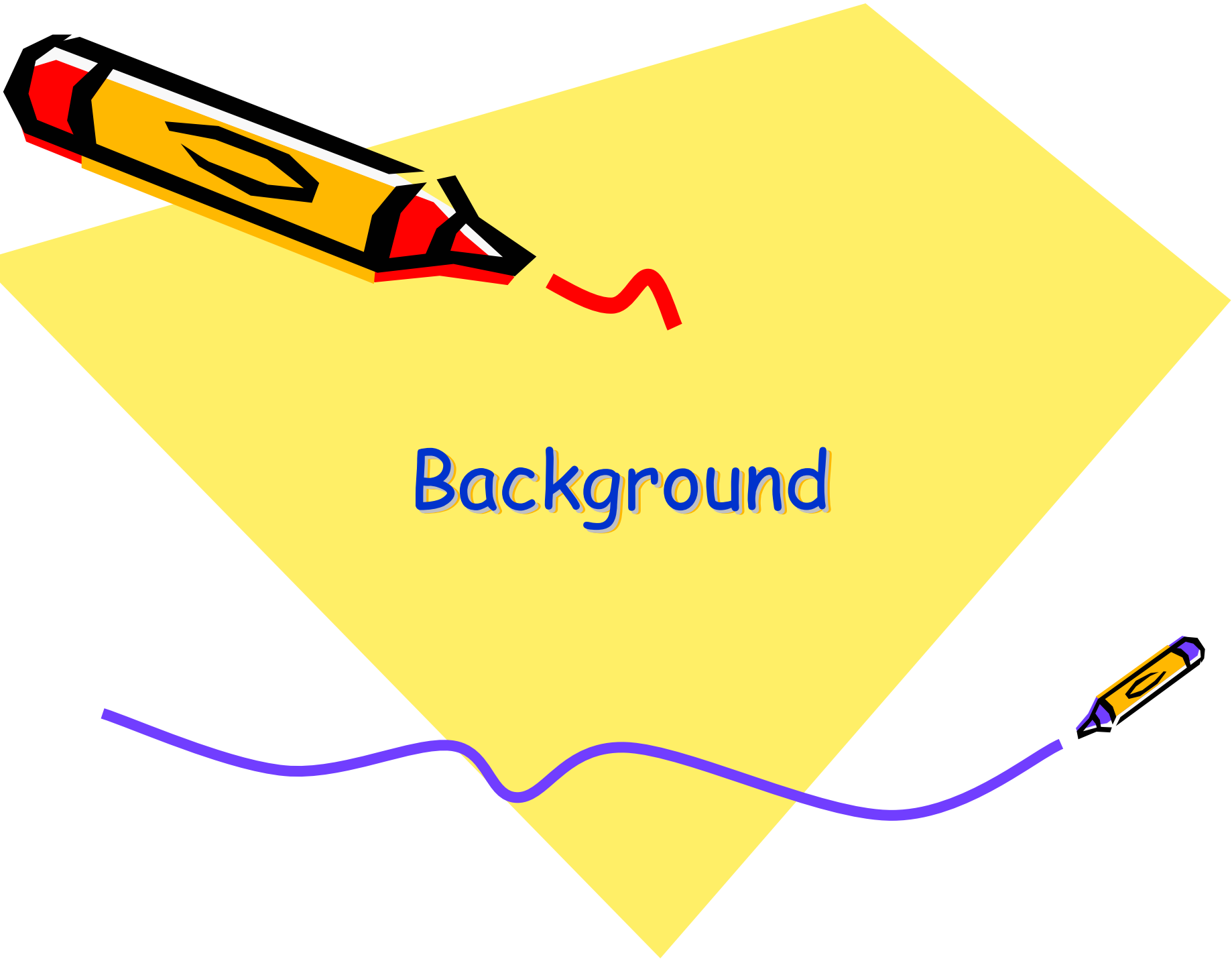
Jen-Wei Roger Kuo

Speech Lab, CSIE, NTNU
rogerkuo@csie.ntnu.edu.tw

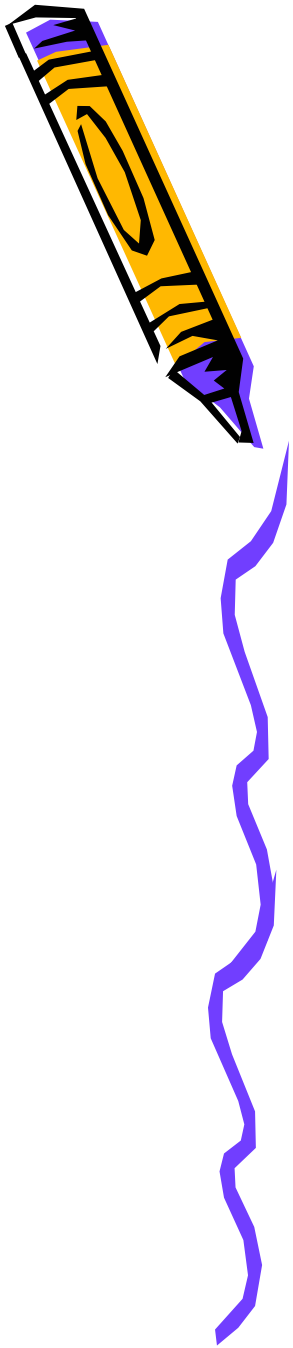
Presentation Outline



- Background
- Maximum a Posterior
- Maximum Likelihood Linear Regression
- MLLR another approach
- Constrained Maximum Likelihood Linear Regression
- Maximum a Posterior Linear Regression
- Structural Maximum a Posterior
- Joint MAP and MLLR
- Appendix — Matrix Calculus


Background

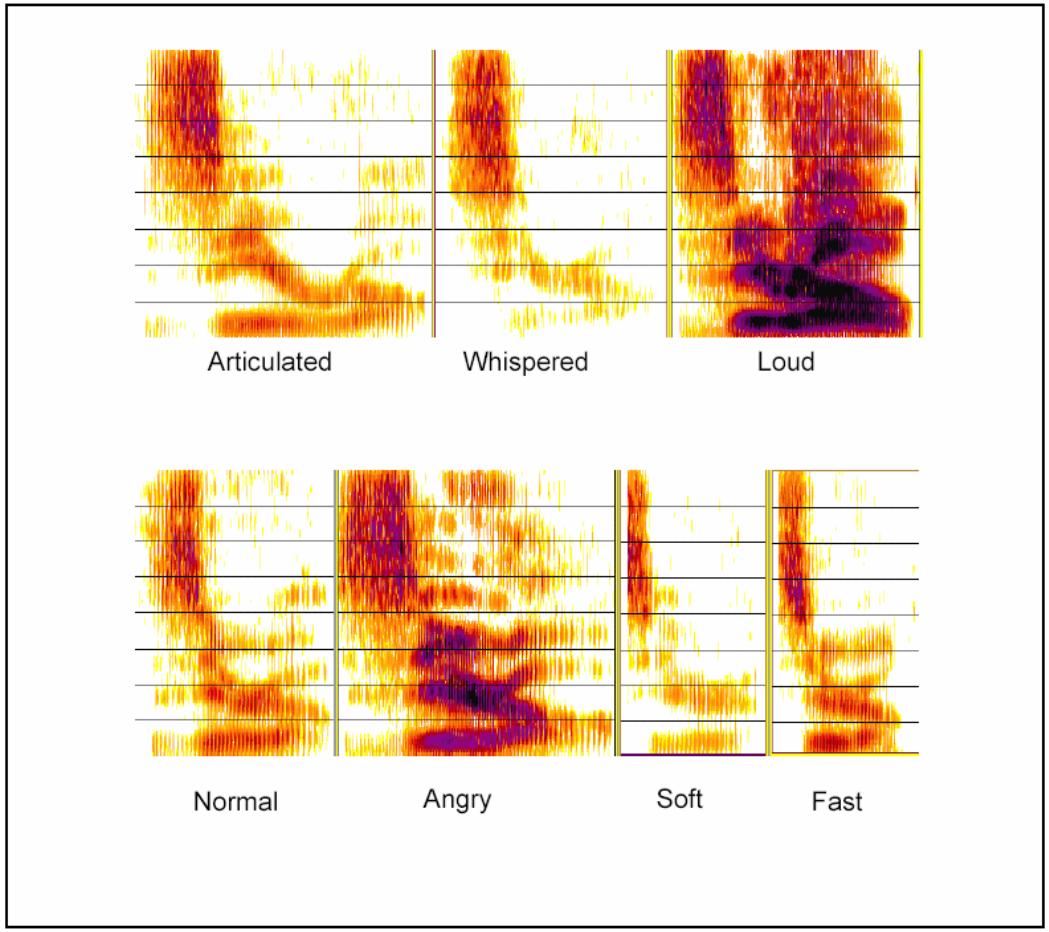
Reference



- Adaptive Methods for Speech and Speaker Recognition – PSTL Jean-Claude Junqua and Roland Kuhn

Background

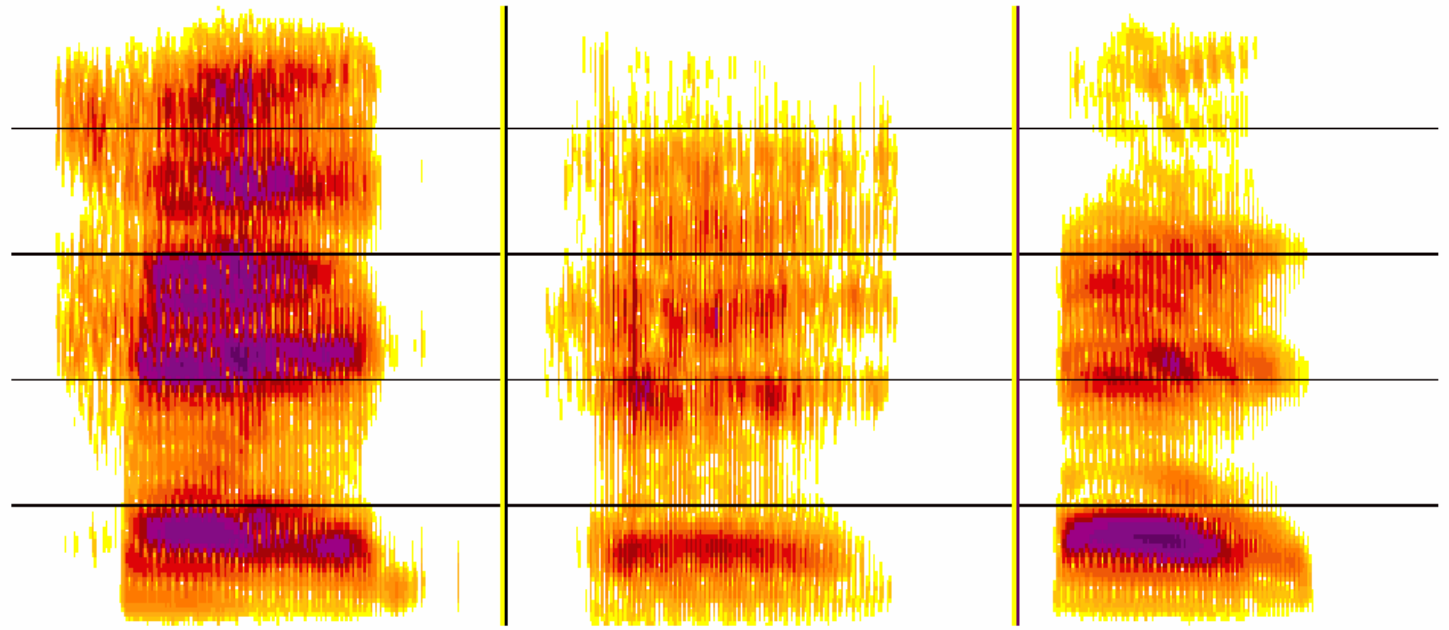
- Intra-Speaker Variability



Word "Zero" produced in different speaking styles

Background

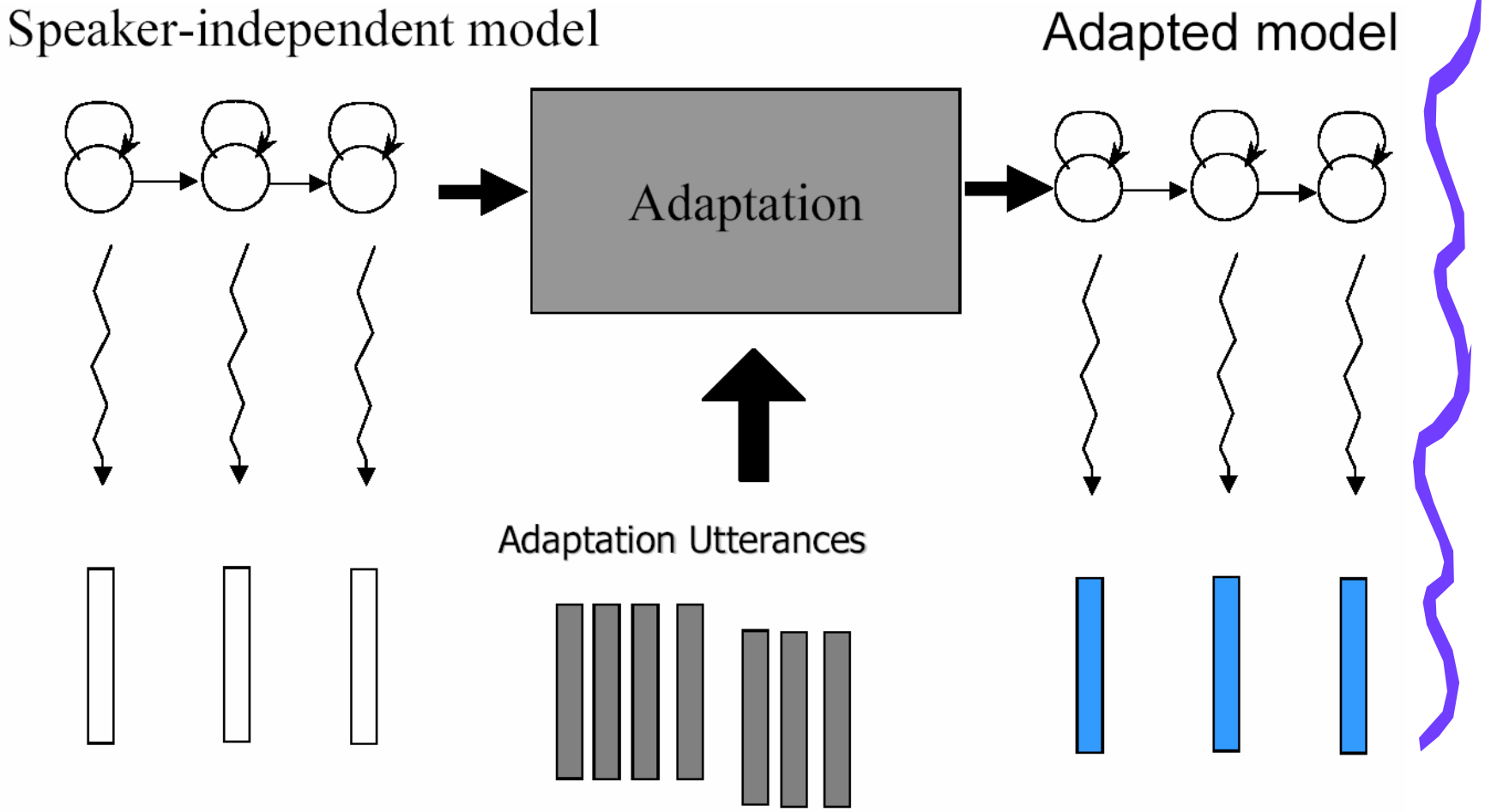
- Inter-Speaker Variability



Word "Head"

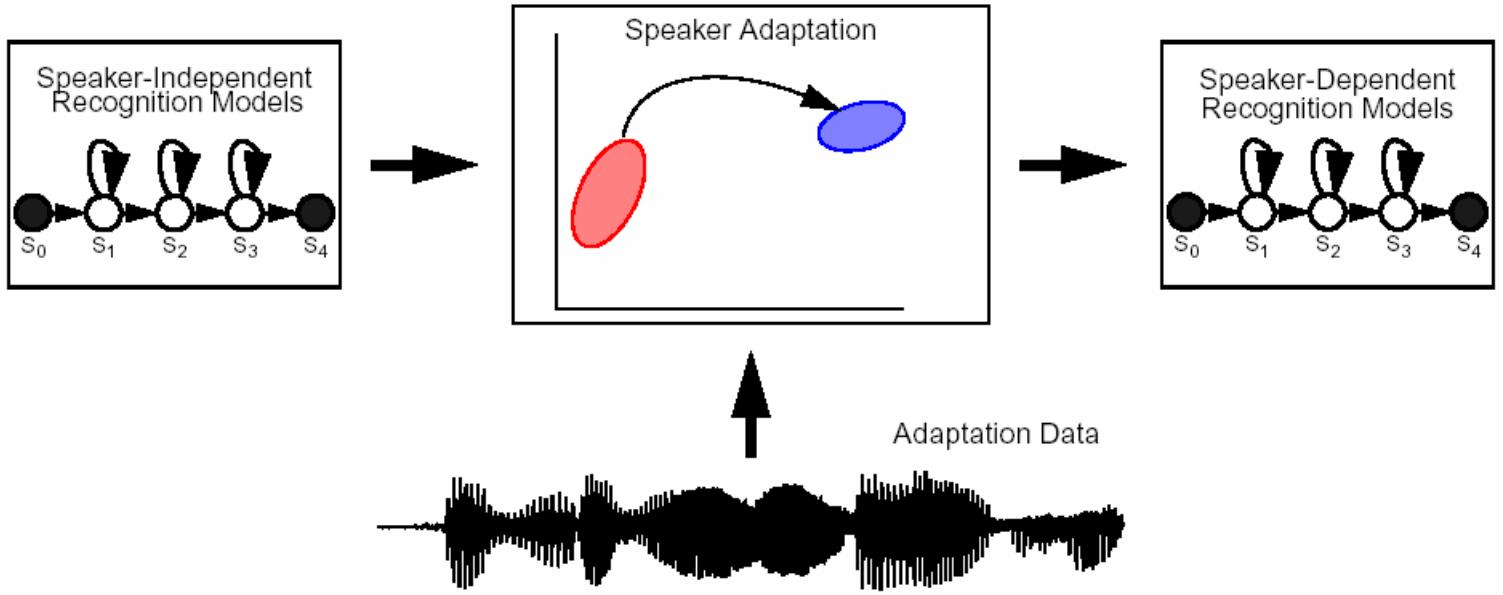
Background

- Adaptive System

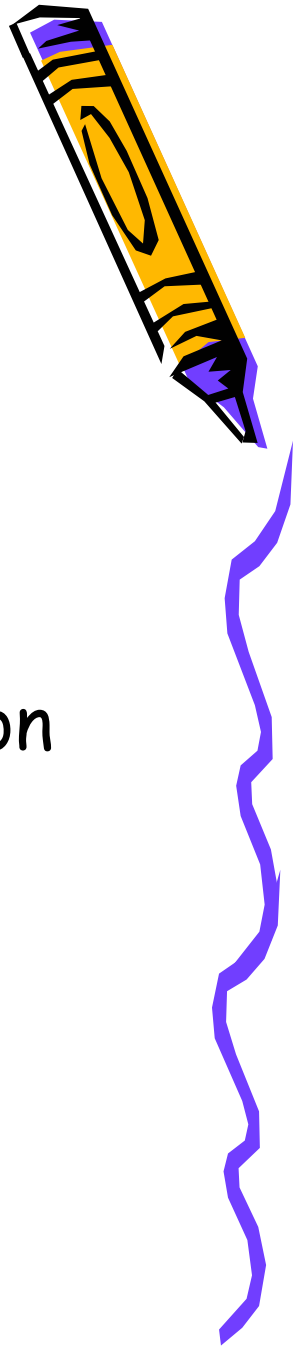


Background

- Goal
  - Produce a system with Speaker-Dependent (SD) performance given small amount of data from new speaker.



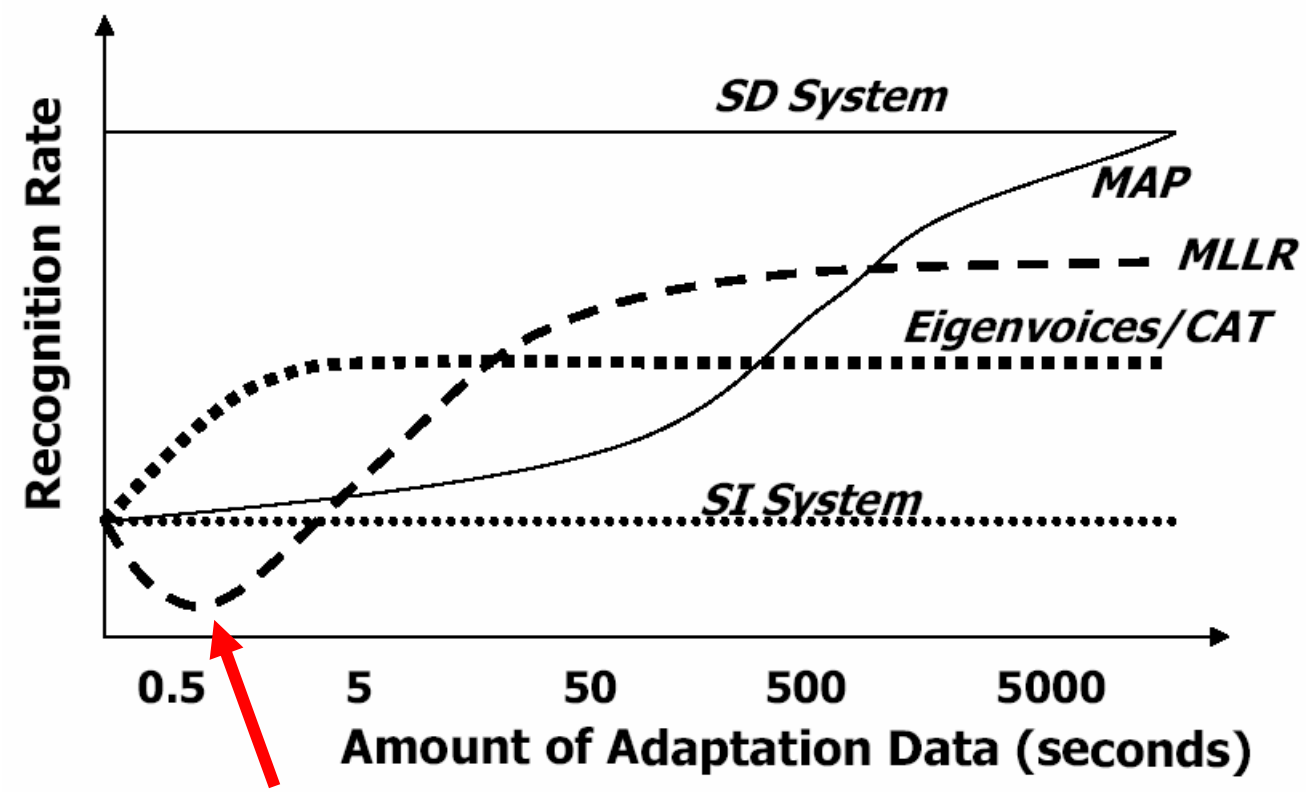
Background



- Main Speaker Adaptation Methods
  - Bayesian
  - Transformation-based
  - Based on Clustering and Model selection

Background

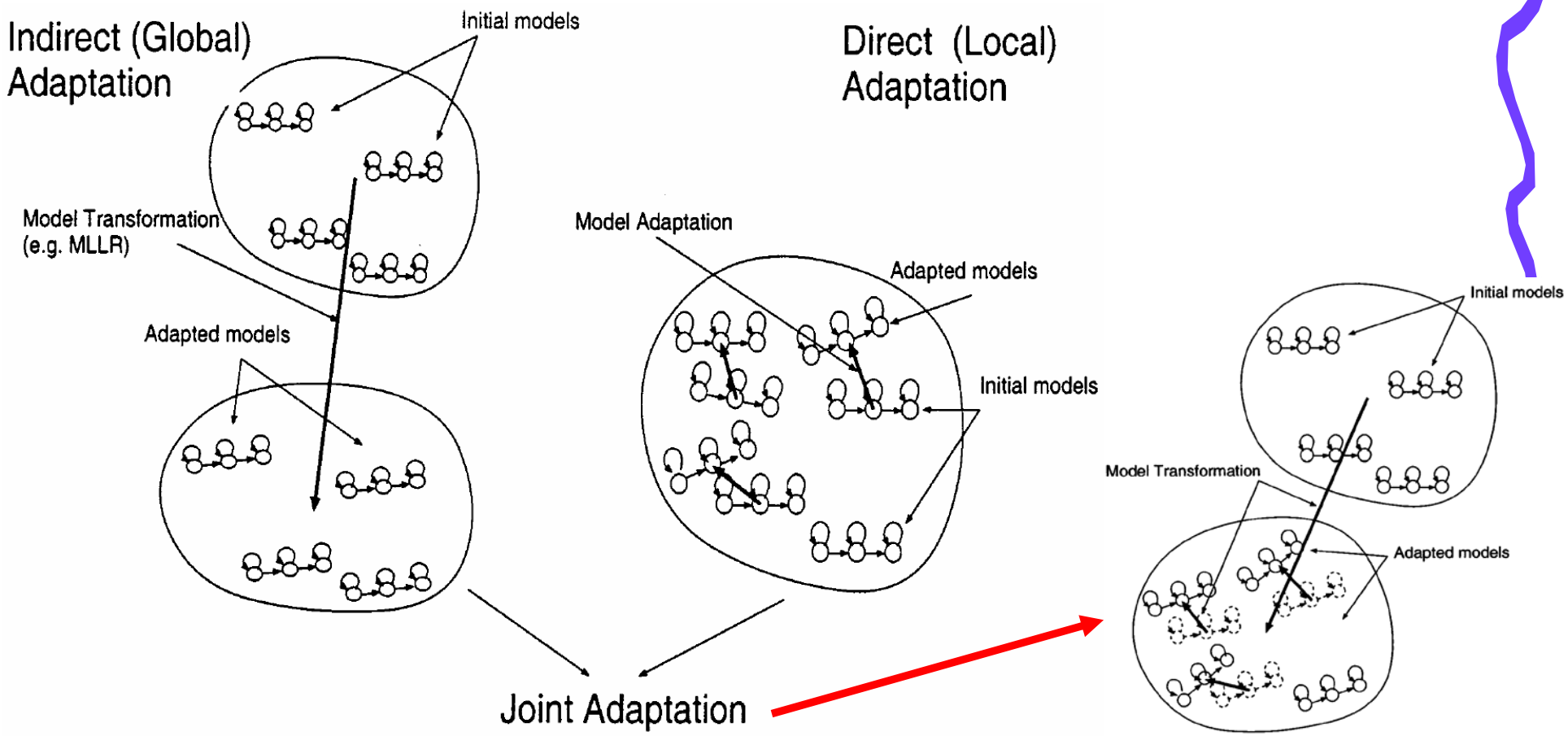
- Comparison of Methods



crazy model

Background

- Direct and indirect adaptation



Approaches

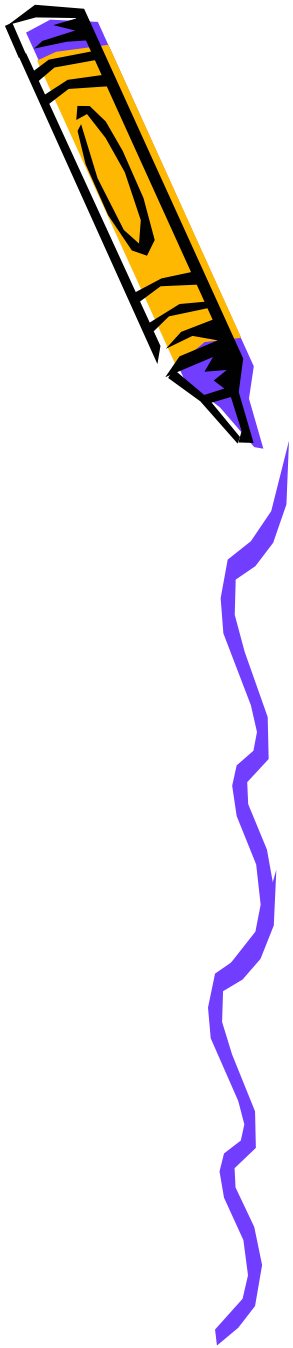
	Direct Adaptation	Indirect Adaptation (Transformation)	Eigen Approach	
			Direct	Indirect
Maximum Likelihood (ML)	Baum-Welch Training	MLLR CMLLR	EigenVoice	Eigen-MLLR
Maximum a Posterior (MAP)	MAP	MAPLR-Mean MAPLR-Covariance		Eigen-MAPLR
On-Line(QB)	QB	QBLR		
Structural Hierarchy	SMAP	SMAPLR EMAPLR		
Maximum Classification Error	MCE	MCELR		


Maximum a Posterior

Maximum a Posterior

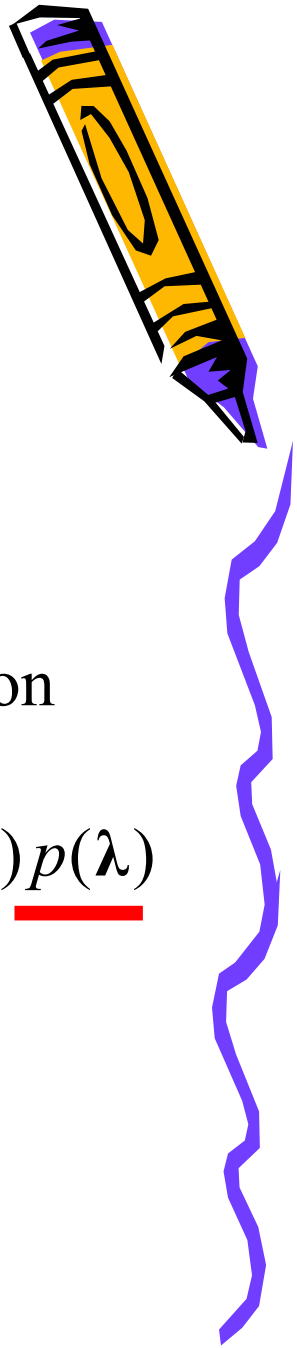
- Reference:
  - A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models – SAP'91 C.-H. Lee
  - Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition – TR'92 Q. Huo
  - Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains – SAP'94 J.-L. Gauvain
  - Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition – SAP'95 Q. Huo
  - On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition – Proceedings of IEEE'00 C.-H. Lee

Maximum a Posterior



- Introduction
- MAP estimate for Multinomial
  - Prior →Dirichlet
- MAP estimate for DHMM
  - Prior →Dirichlet
- MAP estimate for SCHMM
  - Prior →Dirichlet + normal-Wishart
- MAP estimate for CDHMM

Introduction



$\mathbf{X}$: observation data for adaptation

$\boldsymbol{\lambda}$: current acoustic model

$\overline{\boldsymbol{\lambda}}$: reestimated acoustic model

$$\overline{\boldsymbol{\lambda}}_{ML} = \max_{\boldsymbol{\lambda}} p(\mathbf{X} \mid \boldsymbol{\lambda}) \quad \rightarrow \text{Maximum Likelihood Estimation}$$

$$\overline{\boldsymbol{\lambda}}_{MAP} = \max_{\boldsymbol{\lambda}} p(\boldsymbol{\lambda} \mid \mathbf{X}) = \max_{\boldsymbol{\lambda}} \frac{p(\mathbf{X} \mid \boldsymbol{\lambda}) p(\boldsymbol{\lambda})}{p(\mathbf{X})} = \max_{\boldsymbol{\lambda}} p(\mathbf{X} \mid \boldsymbol{\lambda}) p(\boldsymbol{\lambda})$$

$$\downarrow$$

Maximum a Posterior Estimation

Introduction

To estimate the mean of a particular Gaussian :

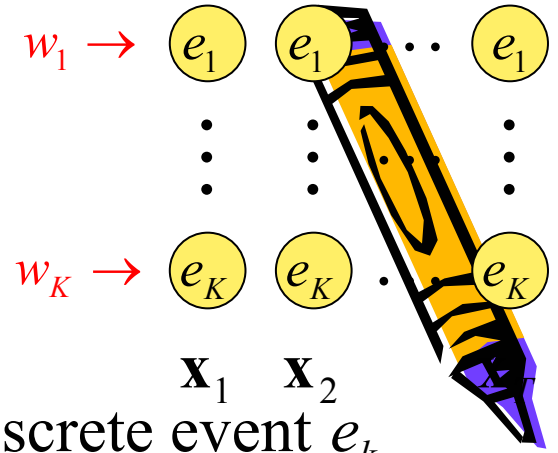
$$\text{Use MLE} \rightarrow \bar{\boldsymbol{\mu}}_{ML} = \frac{\sum_{t=1}^{T} \gamma(t)\mathbf{x}_t}{\sum_{t=1}^{T} \gamma(t)}$$

$$\text{Use MAP} \rightarrow \bar{\boldsymbol{\mu}}_{MAP} = \frac{\tau\boldsymbol{\mu}_{nwi} + \sum_{t=1}^{T} \gamma(t)\mathbf{x}_t}{\tau + \sum_{t=1}^{T} \gamma(t)}$$

$\boldsymbol{\mu}_{nwi}$ is the mean of prior Gaussian.

τ is strength of faith in prior, if it's big, movement away from prior will be slow.

Multinomial



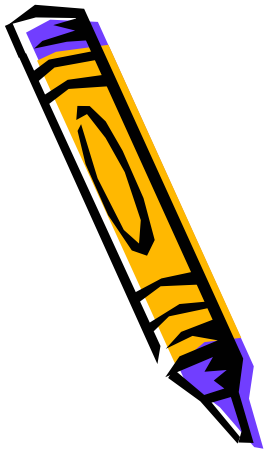
Let w_k be the probability of observing the kth discrete event e_k among a set of K possible outcomes $\{e_k \mid k = 1, \cdots, K\}$ and $\sum_{k=1}^{K} w_k = 1$.

Then, the probability of observing a sequence of i.i.d discrete observations $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ follows a multinomial distribution

$$p(\mathbf{x}_1, \cdots, \mathbf{x}_T \mid w_1, \cdots, w_K) = \frac{(n_1 + \cdots + n_k)!}{n_1! n_2! \cdots n_k!} \prod_{k=1}^{K} w_k^{n_k} = \frac{T!}{n_1! n_2! \cdots n_k!} \prod_{k=1}^{K} w_k^{n_k}$$

where $n_k = \sum_{t=1}^{T} 1(x_t = e_k)$ is the number of occurrence observing the kth event in the sequence with $1(\cdot)$ is the indicator function.

Multinomial



- Many useful random variables used in speech recognition and language processing, including N-grams, histograms, mixture gains, and discrete HMM probabilities, can be modeled this way.

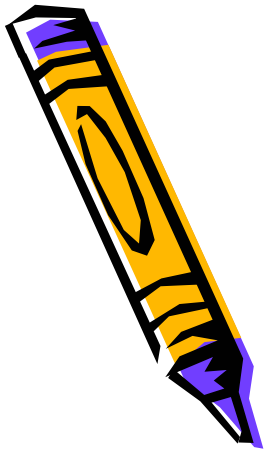
The prior density of $(w_1, \cdots, w_K)$ can be assumed as a Dirichlet density which is a conjugate prior for the parameters of a multinomial density.

$$p(w_1, ..., w_K \mid \nu_1, \cdots \nu_K) = \frac{\Gamma(\nu_1 + ... + \nu_K)}{\Gamma(\nu_1)...\Gamma(\nu_K)} w_1^{\nu_1 - 1} ... w_K^{\nu_K - 1}$$

where $\{\nu_k > 0 \mid k = 1, \cdots, K\}$ is the set of hyperparameters.

So, $p(w_1, ..., w_K \mid \nu_1, \cdots \nu_K) \propto \prod_{k=1}^{K} w_k^{\nu_k - 1}$

Multinomial



So, $p(w_1,...,w_K \mid \mathbf{x}_1,...,\mathbf{x}_T) \propto \prod_{k=1}^{K} w_k^{n_k+\nu_k-1}$

$$\Rightarrow \log p(w_1,...,w_K \mid \mathbf{x}_1,...,\mathbf{x}_T) = \Psi + \left[\sum_{k=1}^{K}(n_k+\nu_k-1)\log w_k\right] + l\left(\sum_{j=1}^{K} w_j\right)$$

Differentiate w.r.t w_k then

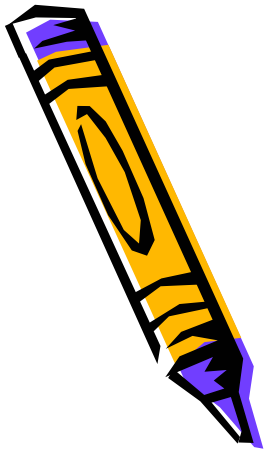
$$\frac{1}{w_k}\times(n_k+\nu_k-1)+l=0 \Rightarrow w_k = -\frac{n_k+\nu_k-1}{l}$$

$$\because \sum_{k=1}^{K} w_k = \sum_{k=1}^{K} -\frac{n_k+\nu_k-1}{l} = 1 \Rightarrow l = -\sum_{k=1}^{K}(n_k+\nu_k-1)$$

$$\therefore w_k = \frac{n_k+\nu_k-1}{\sum_{k=1}^{K}(n_k+\nu_k-1)}$$

Apply Lagrange Multiplier

Discrete HMM



Definition :

$$p(\mathbf{X} \mid \boldsymbol{\lambda}) = \sum_{\mathbf{S}} p(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\lambda}) = \sum_{\mathbf{S}} \left[\pi_{s_1} b_{s_1}(\mathbf{x}_1) \prod_{t=2}^{T} a_{s_{t-1} s_t} b_{s_t}(\mathbf{x}_t) \right]$$

$\mathbf{S}$ is a state sequence , where $\mathbf{S} = \{s_1, \cdots, s_T\}$ and $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$

Consider N - state DHMM with parameter vector $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$

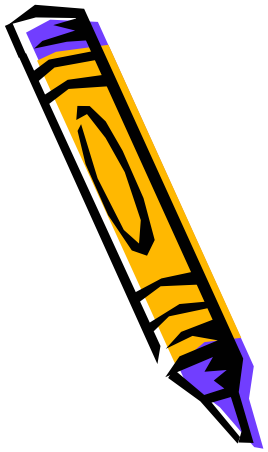
where $\boldsymbol{\pi} = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_N \end{bmatrix}$ is the initial state probability vector.

$\mathbf{A} = [a_{ij}]$ where $i, j = 1, \cdots, N$ is the transition probability matrix.

$\mathbf{B} = [b_{jk}]$ where $j = 1, \cdots, N$, $k = 1, \cdots, K$ with b_{jk} being the probability of observing symbol v_k in state j

The observation symbol set is denoted as $V = \{v_1, \cdots, v_K\}$

Discrete HMM



Q-function :

In order to maximize $\log p(\mathbf{X} \mid \bar{\lambda})$, we only need to maximize

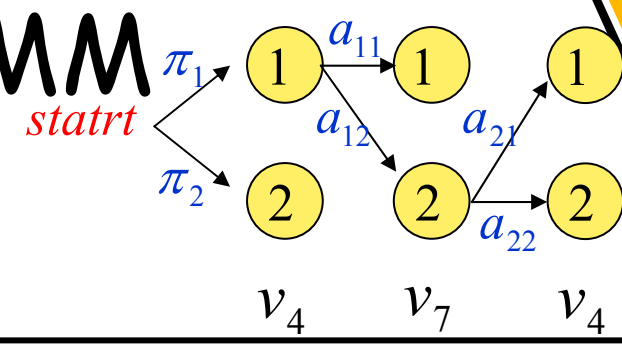
$$Q(\bar{\lambda} \mid \lambda) = E\left[\log P(\mathbf{X}, \mathbf{S} \mid \bar{\lambda}) \mid \mathbf{X}, \lambda\right] = \sum_{\mathbf{S}} \left[P(\mathbf{S} \mid \mathbf{X}, \lambda) \log P(\mathbf{X}, \mathbf{S} \mid \bar{\lambda})\right]$$

$$= \sum_{\mathbf{S}} \left[\frac{P(\mathbf{X}, \mathbf{S} \mid \lambda)}{P(\mathbf{X} \mid \lambda)} \log P(\mathbf{X}, \mathbf{S} \mid \bar{\lambda}) \right]$$

$$= \sum_{\mathbf{S}} \left[\frac{P(\mathbf{X}, \mathbf{S} \mid \lambda)}{\sum_{\mathbf{S}'} P(\mathbf{X}, \mathbf{S}' \mid \lambda)} \log \left[\bar{\pi}_{s_1} \bar{b}_{s_1}(\mathbf{x}_1) \prod_{t=2}^{T} \bar{a}_{s_{t-1} s_t} \bar{b}_{s_t}(\mathbf{x}_t) \right] \right]$$

$$= \sum_{\mathbf{S}} \left[\frac{\pi_{s_1} b_{s_1}(\mathbf{x}_1) \prod_{t=2}^{T} a_{s_{t-1} s_t} b_{s_t}(\mathbf{x}_t)}{\sum_{\mathbf{S}'} P(\mathbf{X}, \mathbf{S}' \mid \lambda)} \left[\log \bar{\pi}_{s_1} + \sum_{t=2}^{T} \log \bar{a}_{s_{t-1} s_t} + \sum_{t=2}^{T} \bar{b}_{s_t}(\mathbf{x}_t) \right] \right]$$

Discrete HMM



A simple example :

	$p(\mathbf{X}, \mathbf{s} \mid \boldsymbol{\lambda})$	$\log p(\mathbf{X}, \mathbf{s} \mid \bar{\boldsymbol{\lambda}})$
1	$\pi_1 \cdot b_{1,4} \cdot a_{11} \cdot b_{1,7} \cdot a_{11} \cdot b_{1,4}$	$\log \bar{\pi}_1 + \log \bar{b}_{1,4} + \log \bar{a}_{11} + \log \bar{b}_{1,7} + \log \bar{a}_{11} + \log \bar{b}_{1,4}$
2	$\pi_1 \cdot b_{1,4} \cdot a_{11} \cdot b_{1,7} \cdot a_{12} \cdot b_{2,4}$	$\log \bar{\pi}_1 + \log \bar{b}_{1,4} + \log \bar{a}_{11} + \log \bar{b}_{1,7} + \log \bar{a}_{12} + \log \bar{b}_{2,4}$
3	$\pi_1 \cdot b_{1,4} \cdot a_{12} \cdot b_{2,7} \cdot a_{21} \cdot b_{1,4}$	$\log \bar{\pi}_1 + \log \bar{b}_{1,4} + \log \bar{a}_{12} + \log \bar{b}_{2,7} + \log \bar{a}_{21} + \log \bar{b}_{1,4}$
4	$\pi_1 \cdot b_{1,4} \cdot a_{12} \cdot b_{2,7} \cdot a_{22} \cdot b_{2,4}$	$\log \bar{\pi}_1 + \log \bar{b}_{1,4} + \log \bar{a}_{12} + \log \bar{b}_{2,7} + \log \bar{a}_{22} + \log \bar{b}_{2,4}$
5	$\pi_2 \cdot b_{2,4} \cdot a_{21} \cdot b_{1,7} \cdot a_{11} \cdot b_{1,4}$	$\log \bar{\pi}_2 + \log \bar{b}_{2,4} + \log \bar{a}_{21} + \log \bar{b}_{1,7} + \log \bar{a}_{11} + \log \bar{b}_{1,4}$
6	$\pi_2 \cdot b_{2,4} \cdot a_{21} \cdot b_{1,7} \cdot a_{12} \cdot b_{2,4}$	$\log \bar{\pi}_2 + \log \bar{b}_{2,4} + \log \bar{a}_{21} + \log \bar{b}_{1,7} + \log \bar{a}_{12} + \log \bar{b}_{2,4}$
7	$\pi_2 \cdot b_{2,4} \cdot a_{22} \cdot b_{2,7} \cdot a_{21} \cdot b_{1,4}$	$\log \bar{\pi}_2 + \log \bar{b}_{2,4} + \log \bar{a}_{22} + \log \bar{b}_{2,7} + \log \bar{a}_{21} + \log \bar{b}_{1,4}$
8	$\pi_2 \cdot b_{2,4} \cdot a_{22} \cdot b_{2,7} \cdot a_{22} \cdot b_{2,4}$	$\log \bar{\pi}_2 + \log \bar{b}_{2,4} + \log \bar{a}_{22} + \log \bar{b}_{2,7} + \log \bar{a}_{22} + \log \bar{b}_{2,4}$

Total 8 paths

Discrete HMM

A simple example :

$$all = 1+2+3+4+5+6+7+8 (\text{all paths})$$

$$\left[\frac{1+2+3+4}{all}\right]\log\bar{\pi}_1 + \left[\frac{5+6+7+8}{all}\right]\log\bar{\pi}_2 = \gamma_1(1)\cdot\log\bar{\pi}_1 + \gamma_2(1)\cdot\log\bar{\pi}_2$$

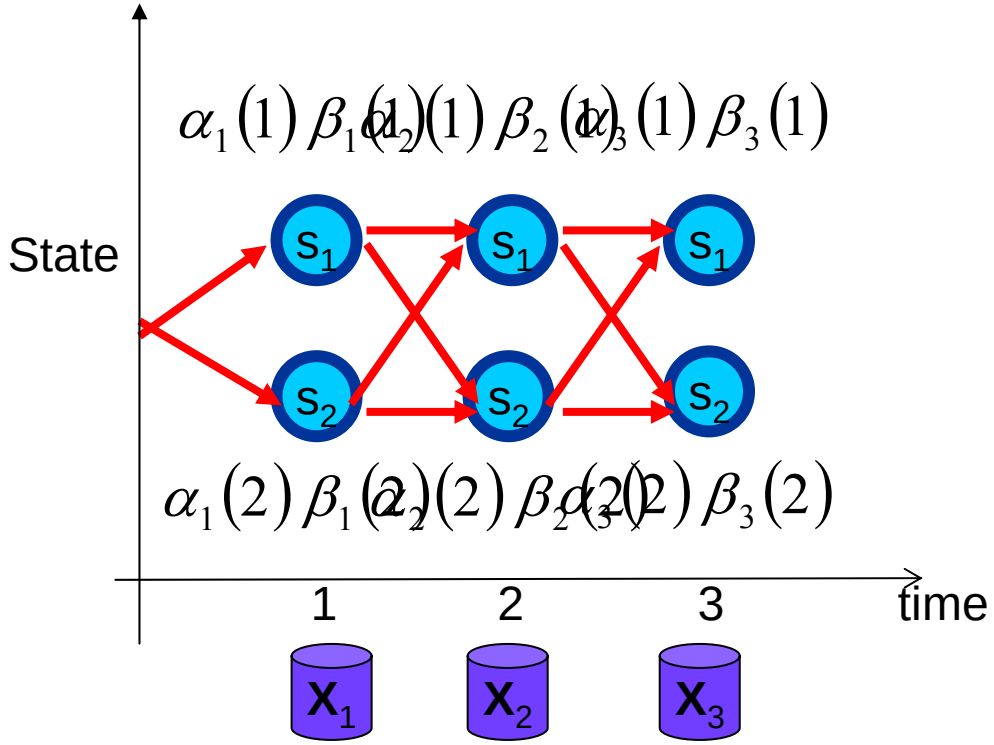
$t=1$ $t=2$ $j=1$ $j=2$ $i=1$ $i=2$

$$\left[\left[\frac{1+2}{all}+\frac{1+5}{all}\right]\log\bar{a}_{11} + \left[\frac{3+4}{all}+\frac{2+6}{all}\right]\log\bar{a}_{12}\right] +$$

$$\left[\left[\frac{5+6}{all}+\frac{3+7}{all}\right]\log\bar{a}_{21} + \left[\frac{7+8}{all}+\frac{4+8}{all}\right]\log\bar{a}_{22}\right]$$

Discrete HMM

The Forward/Backward Procedure



$$\gamma_t(i) = \frac{P(s_t = i, \boldsymbol{X}|\boldsymbol{\lambda})}{P(\boldsymbol{X}|\boldsymbol{\lambda})}$$

$$= \frac{P(s_t = i, \boldsymbol{X}|\boldsymbol{\lambda})}{\sum_{j=1}^{N} P(s_t = j, \boldsymbol{X}|\boldsymbol{\lambda})}$$

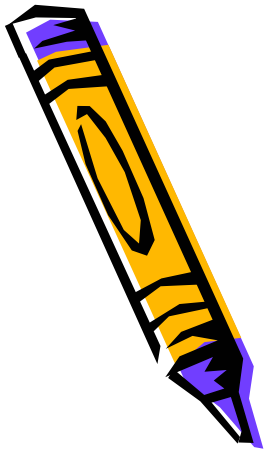
$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$$

$$\xi_t(i,j) = \frac{P(s_t = i, s_{t+1} = j, \boldsymbol{X}|\boldsymbol{\lambda})}{P(\boldsymbol{X}|\boldsymbol{\lambda})}$$

$$= \frac{P(s_t = i, s_{t+1} = j, \boldsymbol{X}|\boldsymbol{\lambda})}{\sum_{i=1}^{N}\sum_{j=1}^{N} P(s_t = i, s_{t+1} = j, \boldsymbol{X}|\boldsymbol{\lambda})}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)a_{ij}b_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j)}$$

Discrete HMM



Q-function :

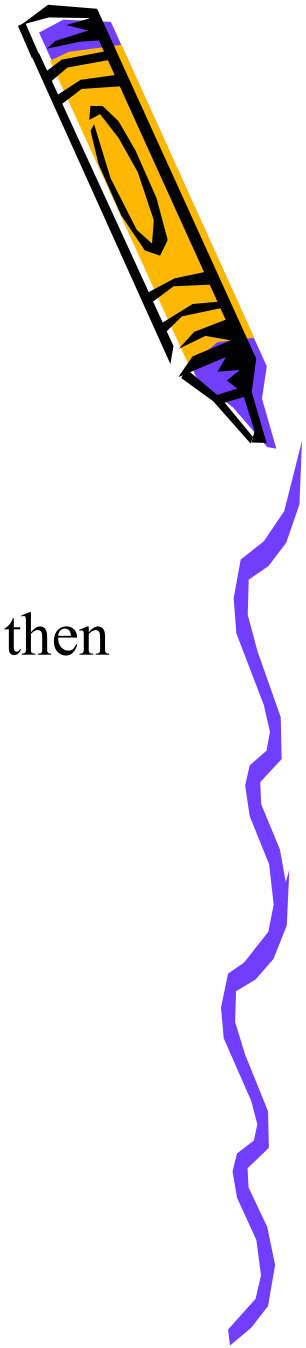
$$Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = \sum_{\mathbf{S}} \left[\frac{p(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\lambda})}{\sum_{\mathbf{S}} p(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\lambda})} \cdot \log p(\mathbf{X}, \mathbf{S} \mid \bar{\boldsymbol{\lambda}}) \right]$$

$$= \left[\sum_{i=1}^{N} \gamma_i(1) \log \bar{\pi}_i \right] + \left[\sum_{i=1}^{N} \sum_{j=1}^{N} \left(\sum_{t=1}^{T-1} \xi_{ij}(t) \right) \log \bar{a}_{ij} \right]$$

$$+ \left[\sum_{j=1}^{N} \sum_{k=1}^{K} \left(\sum_{t:x_t \sim v_k} \Pr(s_t = j, \mathbf{x}_t \sim v_k \mid \mathbf{X}, \boldsymbol{\lambda}) \right) \log \bar{b}_{jk} \right]$$

$$\frac{\sum_{t=1}^{T} \gamma_t(j) \cdot 1(\mathbf{x}_t = v_k)}{\sum_{t=1}^{T} \gamma_t(j)}$$

Discrete HMM



R-function :

For simplicity, prior independence of π, $\mathbf{A}$ and $\mathbf{B}$ is assumed.

The prior density for $\boldsymbol{\lambda}$ is then

$$p(\boldsymbol{\lambda}) = p(\boldsymbol{\pi}) \cdot p(\mathbf{A}) \cdot p(\mathbf{B})$$

and their densities assume the form of Dirichlet distributions then

$$p(\boldsymbol{\lambda}) = K_c \left[\prod_{i=1}^{N} \pi_i^{\eta_i - 1} \right] \left[\prod_{i=1}^{N} \prod_{j=1}^{N} a_{ij}^{\eta_{ij} - 1} \right] \left[\prod_{i=1}^{N} \prod_{k=1}^{K} b_{ik}^{\nu_{ik} - 1} \right]$$

where η_i, η_{ij} and $\nu_{ik} > 1$

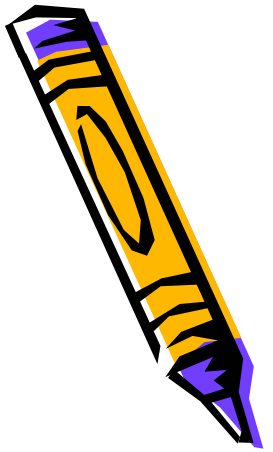
$$\bar{\boldsymbol{\lambda}}_{MAP} = \max_{\boldsymbol{\lambda}} \log p(\boldsymbol{\lambda} \mid \mathbf{X}) = \max_{\boldsymbol{\lambda}} \log \left[p(\mathbf{X} \mid \boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \right]$$

$$= \max_{\boldsymbol{\lambda}} \left[\log p(\mathbf{X} \mid \boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}) \right]$$

$$= \max_{\bar{\boldsymbol{\lambda}}} \left[Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) + \log p(\bar{\boldsymbol{\lambda}}) \right]$$

We define the auxiliary function $R(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) + \log p(\bar{\boldsymbol{\lambda}})$

Discrete HMM



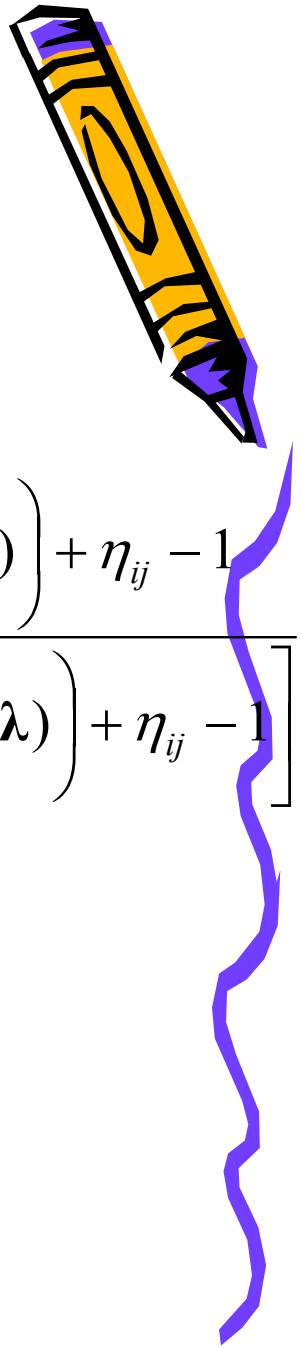
$$\therefore R(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = \Psi(constant) +$$

$$\left[\sum_{i=1}^{N} \left(\Pr(s_1 = i \mid \mathbf{X}, \boldsymbol{\lambda}) + \eta_i - 1 \right) \log \bar{\pi}_i \right] +$$

$$\left[\sum_{i=1}^{N} \sum_{j=1}^{N} \left(\left(\sum_{t=1}^{T-1} \Pr(s_t = i, s_{t+1} = j \mid \mathbf{X}, \boldsymbol{\lambda}) \right) + \eta_{ij} - 1 \right) \log \bar{a}_{ij} \right] +$$

$$\left[\sum_{j=1}^{N} \sum_{k=1}^{K} \left(\left(\sum_{t: x_t \sim v_k} \Pr(s_t = i, \mathbf{x}_t = v_k \mid \mathbf{X}, \boldsymbol{\lambda}) \right) + \nu_{ik} - 1 \right) \log \bar{b}_{jk} \right]$$

Discrete HMM

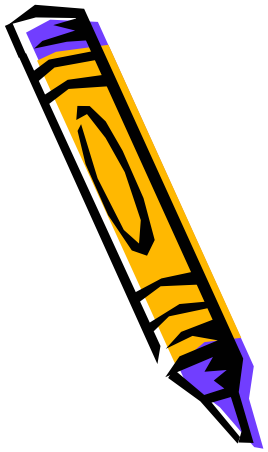


So, we can obtain

$$\overline{\pi}_i = \frac{\Pr(s_1 = i \mid \mathbf{X}, \boldsymbol{\lambda}) + \eta_i - 1}{\sum_{i=1}^{N} \left[\Pr(s_1 = i \mid \mathbf{X}, \boldsymbol{\lambda}) + \eta_i - 1\right]} \quad \overline{a}_{ij} = \frac{\left(\sum_{t=1}^{T-1} \Pr(s_t = i, s_{t+1} = j \mid \mathbf{X}, \boldsymbol{\lambda})\right) + \eta_{ij} - 1}{\sum_{j=1}^{N} \left[\left(\sum_{t=1}^{T-1} \Pr(s_t = i, s_{t+1} = j \mid \mathbf{X}, \boldsymbol{\lambda})\right) + \eta_{ij} - 1\right]}$$

$$\overline{b}_{jk} = \frac{\left(\sum_{t:x_t \sim v_k} \Pr(s_t = i, \mathbf{x}_t = v_k \mid \mathbf{X}, \boldsymbol{\lambda})\right) + \nu_{ik} - 1}{\sum_{k=1}^{K} \left[\left(\sum_{t:x_t \sim v_k} \Pr(s_t = i, x_t = v_k \mid \mathbf{X}, \boldsymbol{\lambda})\right) + \nu_{ik} - 1\right]}$$

Discrete HMM



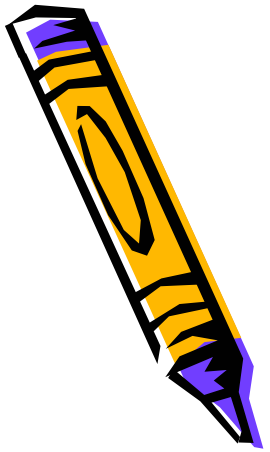
- How to choose the initial estimate for π_i , a_{ij} *and* b_{jk} ?
- One reasonable choice of the initial estimate is the mode of the prior density.

$$\pi_i^{(0)} = \frac{\eta_i - 1}{\sum_{p=1}^{N} (\eta_p - 1)} \quad i = 1, \cdots, N$$

$$a_{ij}^{(0)} = \frac{\eta_{ij} - 1}{\sum_{p=1}^{N} (\eta_{ip} - 1)} \quad i, j = 1, \cdots, N$$

$$b_{jk}^{(0)} = \frac{\nu_{jk} - 1}{\sum_{p=1}^{K} (\nu_{jp} - 1)} \quad j = 1, \cdots, N \text{ and } k = 1, \cdots, K$$

Discrete HMM



- What's the mode ?

  If $\boldsymbol{\lambda}_{mode}$ is the mode of the prior density

  $$\Rightarrow \boldsymbol{\lambda}_{mode} = \max_{\boldsymbol{\lambda}} p(\boldsymbol{\lambda})$$

  – So applying Lagrange Multiplier we can easily derive above modes.

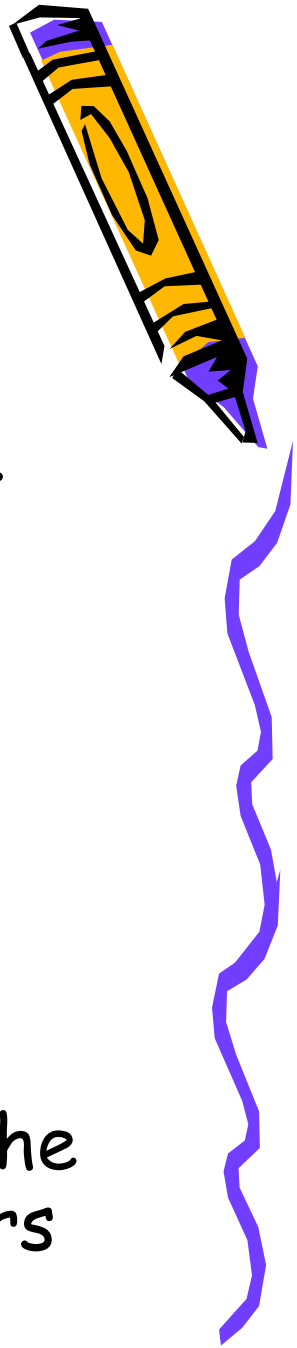
  – Example : $p(\pi_1,\cdots,\pi_N) \propto \prod_{i=1}^{N} \pi_i^{\eta_i - 1} \Rightarrow \log p(\pi_1,\cdots,\pi_N) = \Psi + \sum_{i=1}^{N} (\eta_i - 1)\log \pi_i$

  $$\frac{\partial \log p(\pi_1,\cdots,\pi_N)}{\partial \pi_i} = (\eta_i - 1)\times\frac{1}{\pi_i} + l(\sum_{p=1}^{N} \pi_p - 1) = 0$$

  $$\Rightarrow \pi_i = \frac{\eta_i - 1}{-l} \text{ but } \sum_{p=1}^{N} \pi_p = 1 \therefore \sum_{p=1}^{N} \frac{\eta_p - 1}{-l} = 1$$

  $$\therefore -l = \sum_{p=1}^{N} (\eta_p - 1) \Rightarrow \pi_i = \frac{\eta_i - 1}{\sum_{p=1}^{N} (\eta_p - 1)}$$

Discrete HMM



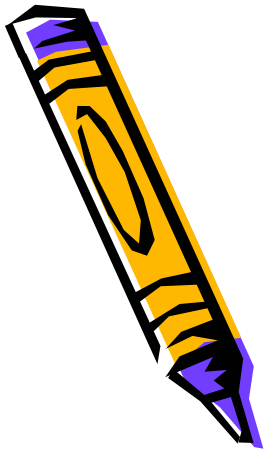
- Another reasonable choice of the initial estimate is the mean of the prior density.

$$\pi_i^{(0)} = \frac{\eta_i}{\sum_{p=1}^{N} \eta_p} \quad i = 1, \cdots, N \qquad a_{ij}^{(0)} = \frac{\eta_{ij}}{\sum_{p=1}^{N} \eta_{ip}} \quad i, j = 1, \cdots, N$$

$$b_{jk}^{(0)} = \frac{\nu_{jk}}{\sum_{p=1}^{K} \nu_{jp}} \quad j = 1, \cdots, N \text{ and } k = 1, \cdots, K$$

- Both are some kind of summarization of the available information about the parameters before any data are observed.

SCHMM



Likelihood $\Rightarrow$ Semi – Continuous HMM

Prior $\Rightarrow$ Dirichlet + normal – Wishart

Let $p(\mathbf{X} \mid \Lambda) = \sum_{\mathbf{S}} \left[\pi_{s_1} b_{s_1}(\mathbf{x}_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} b_{s_t}(\mathbf{x}_t) \right]$ be the likelihood

where $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ and $b_i(\mathbf{x}_t) = \sum_{k=1}^{K} w_{ik} N(\mathbf{x}_t \mid \mathbf{m}_k, \mathbf{r}_k)$

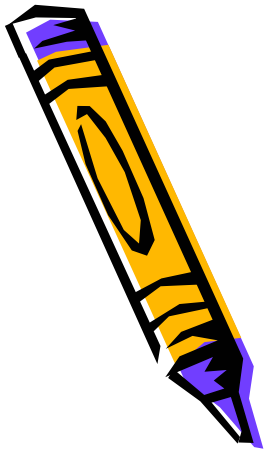
$\Lambda = \{\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_M, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$ where M is the total HMMs number

$\boldsymbol{\lambda}_m = \{\pi_i^{(m)}, a_{ij}^{(m)}, w_{ik}^{(m)} \mid i, j = 1, ..., N(\text{state number}), k = 1, \cdots, K\}$

and $\boldsymbol{\theta}_k = \{\mathbf{m}_k, \mathbf{r}_k\}$ $k = 1, .., K(\text{mixture number})$

where $N(\mathbf{x} \mid \mathbf{m}_k, \mathbf{r}_k) = (2\pi)^{-D/2} \mid \mathbf{r}_k \mid^{1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{r}_k (\mathbf{x}-\mathbf{m}_k)}$

SCHMM



The prior density for Λ is assumed to be :

$$g(\mathbf{\Lambda}) = \left[\prod_{m=1}^{M} g(\boldsymbol{\lambda}_m)\right]\left[\prod_{k=1}^{K} g(\mathbf{m}_k, \mathbf{r}_k)\right]$$

independent

where $g(\boldsymbol{\lambda}_m) \propto K_c\left[\prod_{i=1}^{N} \pi_i^{\eta_i - 1}\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{N} a_{ij}^{\eta_{ij}-1}\right]\left[\prod_{i=1}^{N}\prod_{k=1}^{K} w_{ik}^{\nu_{ik}-1}\right]$

If $\mathbf{r}_k$ is a full precision matrix then $g(\mathbf{m}_k, \mathbf{r}_k)$ is assumed as a

normal – Wishart $\Rightarrow g(\mathbf{m}_k, \mathbf{r}_k) \propto |\mathbf{r}_k|^{\frac{\alpha_k - D}{2}} e^{-\frac{\tau_k}{2}(\mathbf{m}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\gamma}_k (\mathbf{m}_k - \boldsymbol{\mu}_k)} e^{-\frac{1}{2} tr(\mathbf{u}_k \mathbf{r}_k)}$

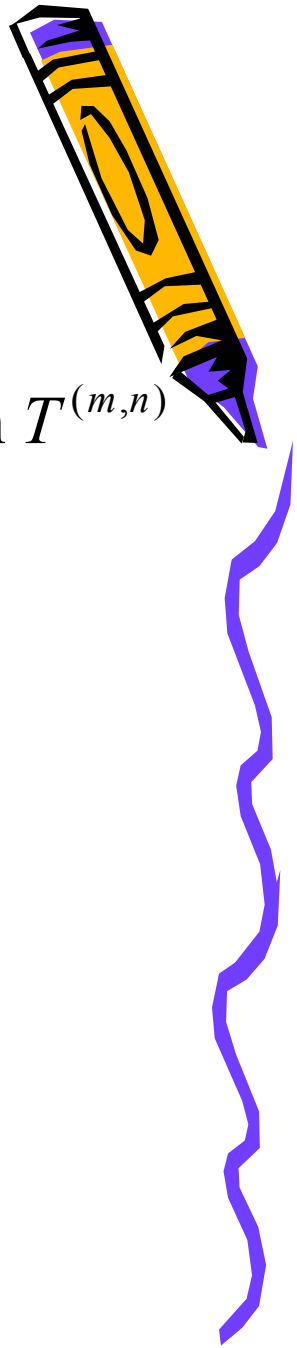
$\alpha_k > D - 1$, $\tau_k > 0$, $\boldsymbol{\mu}_k$ is a vector of dimension D

and $\mathbf{u}_k$ is a $D \times D$ positive defintite matrix

If $\mathbf{r}_k$ is a diagonal precision matrix then $g(\mathbf{m}_k, \mathbf{r}_k)$ is assumed as a

product of normal – gamma $\Rightarrow g(\mathbf{m}_k, \mathbf{r}_k) \propto \prod_{d=1}^{D} r_{kd}^{\frac{\alpha_{kd} - 1/2}{2}} e^{-\frac{\tau_{kd}}{2} r_{kd} (m_{kd} - \mu_{kd})^2} e^{-\beta_{kd} r_{kd}}$

SCHMM



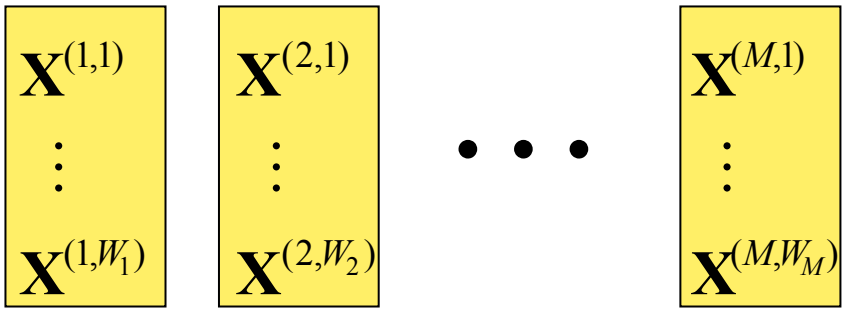
Let $\mathrm{X}^{(m,n)}$ denote the nth $o$bservation sequence of length $T^{(m,n)}$ associated with model m

and each model m has W_m observation sequences.

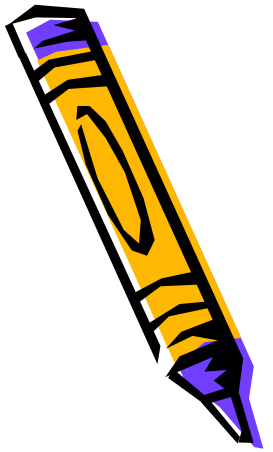
the MAP estimates of $\mathbf{\Lambda}$ can be obtained by

$$\mathbf{\Lambda}_{MAP} = \arg\max_{\mathbf{\Lambda}} \left[\left(\prod_{m=1}^{M} \prod_{n=1}^{Wm} f(\mathbf{X}^{(m,n)} \mid \boldsymbol{\lambda}_m, \mathbf{\Theta}) \right) g(\mathbf{\Lambda}) \right]$$

Model 1 Model 2 Model M



SCHMM



Q-function :

Define a Q – function as

$$Q(\overline{\Lambda} \mid \Lambda) = \sum_{m=1}^{M} \sum_{n=1}^{W_m} E\left[\log f(\mathbf{X}^{(m,n)}, \mathbf{S}, \mathbf{L} \mid \overline{\Lambda}) \mid \mathbf{X}^{(m,n)}, \Lambda\right]$$

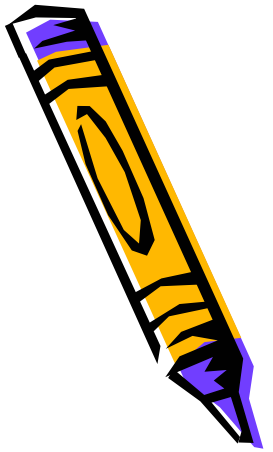
$$= \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{\mathbf{S}^{(m,n)}} \sum_{\mathbf{L}^{(m,n)}} f(\mathbf{S}^{(m,n)}, \mathbf{L}^{(m,n)} \mid \mathbf{X}^{(m,n)}, \Lambda) \log f(\mathbf{X}^{(m,n)}, \mathbf{S}^{(m,n)}, \mathbf{L}^{(m,n)} \mid \overline{\Lambda})$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{\mathbf{S}^{(m,n)}} \sum_{\mathbf{L}^{(m,n)}} \frac{f(\mathbf{S}^{(m,n)}, \mathbf{L}^{(m,n)}, \mathbf{X}^{(m,n)} \mid \Lambda)}{f(\mathbf{X}^{(m,n)} \mid \Lambda)} \log f(\mathbf{X}^{(m,n)}, \mathbf{S}^{(m,n)}, \mathbf{L}^{(m,n)} \mid \overline{\Lambda})$$

Where $f(\mathbf{X}^{(m,n)}, \mathbf{S}^{(m,n)}, \mathbf{L}^{(m,n)} \mid \overline{\Lambda})$

$$= \pi_{s_1} w_{s_1,l_1} N(\mathbf{x}_1 \mid \mathbf{m}_{l_1}, \mathbf{r}_{l_1}) \prod_{t=2}^{T^{(m,n)}} \left[a_{s_{t-1}s_t} w_{s_t l_t} N(\mathbf{x}_t^{(m,n)} \mid \mathbf{m}_{l_t}, \mathbf{r}_{l_t})\right]$$

SCHMM



Q-function :

$\therefore$ Q – function can be decomposed in

$$Q(\overline{\boldsymbol{\Lambda}} \mid \boldsymbol{\Lambda}) = \sum_{m=1}^{M}\sum_{i=1}^{N}\left(\sum_{n=1}^{W_m}\gamma_1^{(m,n)}(i)\right)\log\overline{\pi}_i^{(m)} + \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\gamma_t^{(m,n)}(i,j)\right)\log\overline{a}_{ij}^{(m)} +$$

$$\sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\xi_t^{(m,n)}(i,k)\right)\log\overline{w}_{ik}^{(m)} + \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\left(\xi_t^{(m,n)}(k)\log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)\right)$$

where $\gamma_t^{(m,n)}(i,j) = \Pr(s_t^{(m,n)} = i, s_{t+1}^{(m,n)} = j \mid \mathbf{X}^{(m,n)}, \boldsymbol{\lambda}_m)$

$$\gamma_t^{(m,n)}(i) = \Pr(s_t^{(m,n)} = i \mid \mathbf{X}^{(m,n)}, \boldsymbol{\lambda}_m)$$

$$\xi_t^{(m,n)}(i,k) = \Pr(s_t^{(m,n)} = i, l_t^{(m,n)} = k \mid \mathbf{X}^{(m,n)}, \boldsymbol{\lambda}_m)$$

$$\xi_t^{(m,n)}(k) = \Pr(l_t^{(m,n)} = k \mid \mathbf{X}^{(m,n)}, \boldsymbol{\lambda}_m)$$

and $\xi_t^{(m,n)}(i,k) = \gamma_t^{(m,n)}(i) \cdot \dfrac{w_{ik}^{(m)} N(\mathbf{x}_t^{(m,n)} \mid \mathbf{m}_k, \mathbf{r}_k)}{\sum_{k=1}^{K} w_{ik}^{(m)} N(\mathbf{x}_t^{(m,n)} \mid \mathbf{m}_k, \mathbf{r}_k)}$

SCHMM

$$\log g(\overline{\Lambda}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \left(\eta_i^{(m)} - 1\right) \log \overline{\pi}_i^{(m)} + \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\eta_{ij}^{(m)} - 1\right) \log \overline{a}_{ij}^{(m)}$$

$$+ \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\nu_{jk}^{(m)} - 1\right) \log \overline{w}_{jk}^{(m)} + \sum_{k=1}^{K} \log g(\overline{m}_k, \overline{r}_k) + Constant$$

$$R(\overline{\Lambda} \mid \Lambda) = Q(\overline{\Lambda} \mid \Lambda) + \log g(\Lambda) = \sum_{m=1}^{M} \sum_{i=1}^{N} \left\{ \left[\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i) \right) + \eta_i^{(m)} - 1 \right] \log \overline{\pi}_i^{(m)} \right\}$$

$$+ \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\{ \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1 \right] \log \overline{a}_{ij}^{(m)} \right\}$$

$$+ \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} \left\{ \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,k) \right) + \nu_{jk}^{(m)} - 1 \right] \log \overline{w}_{jk}^{(m)} \right\}$$

$$+ \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \left(\xi_t^{(m,n)}(k) \log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k) \right) + \sum_{k=1}^{K} \log g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k) + Constant$$

SCHMM Initial probability

- Differentiating $R(\overline{\Lambda}\,|\,\Lambda)$ w.r.t $\overline{\pi}_i^{(m)}$ and equating it to zero.

$$\frac{\partial R(\overline{\Lambda}\,|\,\Lambda)}{\partial \overline{\pi}_i^{(m)}} = 0 \Rightarrow \frac{\partial}{\partial \overline{\pi}_i^{(m)}} \sum_{m=1}^{M} \sum_{i=1}^{N} \left\{ \left[\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i) \right) + \eta_i^{(m)} - 1 \right] \log \overline{\pi}_i^{(m)} \right\} = 0$$

$$\Rightarrow \frac{\partial}{\partial \overline{\pi}_i^{(m)}} \left[\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i) \right) + \eta_i^{(m)} - 1 \right] \log \overline{\pi}_i^{(m)} + l \left(\sum_{j=1}^{N} \overline{\pi}_j^{(m)} \right) = 0$$

$$\frac{1}{\overline{\pi}_i^{(m)}} \left[\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i) \right) + \eta_i^{(m)} - 1 \right] + l = 0 \Rightarrow \overline{\pi}_i^{(m)} = \frac{\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i) \right) + \eta_i^{(m)} - 1}{-l}$$

$$\sum_{j=1}^{N} \overline{\pi}_j^{(m)} = 1 \Rightarrow \sum_{j=1}^{N} \frac{\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(j) \right) + \eta_j^{(m)} - 1}{-l} = 1 \Rightarrow -l = \sum_{j=1}^{N} \left[\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(j) \right) + \eta_j^{(m)} - 1 \right]$$

$$\therefore \overline{\pi}_i^{(m)} = \frac{\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i) \right) + \eta_i^{(m)} - 1}{\sum_{j=1}^{N} \left[\left(\sum_{n=1}^{W_m} \gamma_1^{(m,n)}(j) \right) + \eta_j^{(m)} - 1 \right]} = \frac{\eta_i^{(m)} - 1 + \sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i)}{\sum_{j=1}^{N} \eta_j^{(m)} - N + \sum_{j=1}^{N} \sum_{n=1}^{W_m} \gamma_1^{(m,n)}(j)}$$

SCHMM

Transition probability

- Differentiating $R(\overline{\Lambda}\,|\,\Lambda)$ w.r.t $\overline{a}_{ij}^{(m)}$ and equating it to zero.

$$\frac{\partial R(\overline{\Lambda}\,|\,\Lambda)}{\partial \overline{a}_{ij}^{(m)}} = 0 \Rightarrow \frac{\partial}{\partial \overline{a}_{ij}^{(m)}} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\{ \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1 \right] \log \overline{a}_{ij}^{(m)} \right\} = 0$$

$$\Rightarrow \frac{\partial}{\partial \overline{a}_{ij}^{(m)}} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1 \right] \log \overline{a}_{ij}^{(m)} + l \left(\sum_{j=1}^{N} \overline{a}_{ij}^{(m)} \right) = 0$$

$$\frac{1}{\overline{a}_{ij}^{(m)}} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1 \right] + l = 0 \Rightarrow \overline{a}_{ij}^{(m)} = \frac{\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1}{-l}$$

$$\sum_{j=1}^{N} \overline{a}_{ij}^{(m)} = 1 \Rightarrow \sum_{j=1}^{N} \frac{\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1}{-l} = 1 \Rightarrow -l = \sum_{j=1}^{N} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1 \right]$$

$$\therefore \overline{a}_{ij}^{(m)} = \frac{\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1}{\sum_{j=1}^{N} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j) \right) + \eta_{ij}^{(m)} - 1 \right]} = \frac{\eta_{ij}^{(m)} - 1 + \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j)}{\sum_{j=1}^{N} \eta_{ij}^{(m)} - N + \sum_{j=1}^{N} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j)}$$

SCHMM

Mixture weight

- Differentiating $R(\overline{\Lambda} \mid \Lambda)$ w.r.t $\overline{w}_{ik}^{(m)}$ and equate it to zero.

$$\frac{\partial R(\overline{\Lambda} \mid \Lambda)}{\partial \overline{w}_{ik}^{(m)}} = 0 \Rightarrow \frac{\partial}{\partial \overline{w}_{ik}^{(m)}} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} \left\{ \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,k) \right) + v_{ik}^{(m)} - 1 \right] \log \overline{w}_{ik}^{(m)} \right\} = 0$$

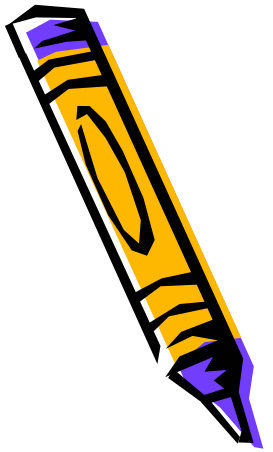
$$\Rightarrow \frac{\partial}{\partial \overline{w}_{ik}^{(m)}} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,k) \right) + v_{ik}^{(m)} - 1 \right] \log \overline{w}_{ik}^{(m)} + l \left(\sum_{j=1}^{K} \overline{w}_{ij}^{(m)} \right) = 0$$

$$\frac{1}{\overline{w}_{ik}^{(m)}} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,k) \right) + v_{ik}^{(m)} - 1 \right] + l = 0 \Rightarrow \overline{w}_{ik}^{(m)} = \frac{\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,k) \right) + v_{ik}^{(m)} - 1}{-l}$$

$$\sum_{j=1}^{K} \overline{w}_{ij}^{(m)} = 1 \Rightarrow \sum_{j=1}^{N} \frac{\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,j) \right) + v_{ij}^{(m)} - 1}{-l} = 1 \Rightarrow -l = \sum_{j=1}^{K} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,j) \right) + v_{ij}^{(m)} - 1 \right]$$

$$\therefore \overline{w}_{ik}^{(m)} = \frac{\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,k) \right) + v_{ik}^{(m)} - 1}{\sum_{j=1}^{K} \left[\left(\sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,j) \right) + v_{ij}^{(m)} - 1 \right]} = \frac{v_{ik}^{(m)} - 1 + \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,k)}{\sum_{j=1}^{K} v_{ij}^{(m)} - K + \sum_{j=1}^{K} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(i,j)}$$

SCHMM



- Differentiating $R(\overline{\Lambda} \mid \Lambda)$ w.r.t $\overline{\mathbf{m}}_k^{(m)}$ and equating it to zero.

$$\left[\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\left(\xi_t^{(m,n)}(k)\frac{\partial \log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{\mathbf{m}}_k}\right)\right] + \frac{\partial \log g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{\mathbf{m}}_k} = 0 \quad (55)$$

- Differentiating $R(\overline{\Lambda} \mid \Lambda)$ w.r.t $\overline{\mathbf{r}}_k^{(m)}$ and equating it to zero.

$$\left[\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\left(\xi_t^{(m,n)}(k)\frac{\partial \log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{\mathbf{r}}_k}\right)\right] + \frac{\partial \log g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{\mathbf{r}}_k} = 0 \quad (56)$$

SCHMM

Full Covariance

- Full Covariance matrix case :

$$\frac{\partial \log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{\mathbf{m}}_k} = \left[-\frac{1}{2} (\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)^T \overline{\mathbf{r}}_k (\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k) \right]'$$

$$= (-\frac{1}{2}) \times (\overline{\mathbf{r}}_k + \overline{\mathbf{r}}_k^T)(\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k) \times (-1)$$

$$= \overline{\mathbf{r}}_k (\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)$$

$$\frac{\partial \log g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{\mathbf{m}}_k} = \frac{1}{g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)} \times \left[|\overline{\mathbf{r}}_k|^{\frac{\alpha_k - D}{2}} e^{-\frac{\tau_k}{2}(\overline{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T \overline{\mathbf{r}}_k (\overline{\mathbf{m}}_k - \boldsymbol{\mu}_k)} e^{-\frac{1}{2} tr(\mathbf{u}_k \overline{\mathbf{r}}_k)} \right]'$$

$$= \frac{1}{g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)} \times |\overline{\mathbf{r}}_k|^{\frac{\alpha_k - D}{2}} e^{-\frac{1}{2} tr(\mathbf{u}_k \overline{\mathbf{r}}_k)} e^{-\frac{\tau_k}{2}(\overline{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T \overline{\mathbf{r}}_k (\overline{\mathbf{m}}_k - \boldsymbol{\mu}_k)} \times -\frac{\tau_k}{2} (\overline{\mathbf{m}}_k - \boldsymbol{\mu}_k)(\overline{\mathbf{r}}_k + \overline{\mathbf{r}}_k)$$

$$= -\tau_k \overline{\mathbf{r}}_k (\overline{\mathbf{m}}_k - \boldsymbol{\mu}_k)$$

SCHMM

Full Covariance

- Full Covariance matrix case :

$$\left[\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\left(\xi_t^{(m,n)}(k)\bar{\mathbf{r}}_k(\mathbf{x}_t^{(m,n)}-\bar{\mathbf{m}}_k)\right)\right]-\tau_k\bar{\mathbf{r}}_k(\bar{\mathbf{m}}_k-\boldsymbol{\mu}_k)=0$$

$$\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\xi_t^{(m,n)}(k)\bar{\mathbf{r}}_k\mathbf{x}_t^{(m,n)}-\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\xi_t^{(m,n)}(k)\bar{\mathbf{r}}_k\bar{\mathbf{m}}_k-\tau_k\bar{\mathbf{r}}_k\bar{\mathbf{m}}_k+\tau_k\bar{\mathbf{r}}_k\boldsymbol{\mu}_k=0$$

$$\left[\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\xi_t^{(m,n)}(k)+\tau_k\right]\bar{\mathbf{m}}_k=\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\xi_t^{(m,n)}(k)\mathbf{x}_t^{(m,n)}+\tau_k\boldsymbol{\mu}_k$$

$$\therefore\ \bar{\mathbf{m}}_k=\frac{\tau_k\boldsymbol{\mu}_k+\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\xi_t^{(m,n)}(k)\mathbf{x}_t^{(m,n)}}{\tau_k+\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\xi_t^{(m,n)}(k)}$$

SCHMM

Full Covariance

- Full Covariance matrix case :

$$\frac{\partial \log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{\mathbf{r}}_k} = \left[\log \mid \mathbf{r}_k \mid^{1/2}\right]' + \left[-\frac{1}{2}(\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)^T \overline{\mathbf{r}}_k (\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)\right]'$$

$$= \mid \mathbf{r}_k \mid^{-1/2} \times \frac{1}{2} \times \mid \mathbf{r}_k \mid^{-1/2} \left[\mid \mathbf{r}_k \mid\right]' - \frac{1}{2}\left[(\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)^T \overline{\mathbf{r}}_k (\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)\right]'$$

$$= \frac{1}{2}\left[\overline{\mathbf{r}}_k^{-1} - (\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)(\mathbf{x}_t^{(m,n)} - \overline{\mathbf{m}}_k)^T\right]$$

SCHMM

Full Covariance

- Full Covariance matrix case :

$$\frac{\partial \log g(\bar{\mathbf{m}}_k, \bar{\mathbf{r}}_k)}{\partial \bar{\mathbf{r}}_k} = \frac{1}{g(\bar{\mathbf{m}}_k, \bar{\mathbf{r}}_k)} \times \Big[\underbrace{|\bar{\mathbf{r}}_k|^{\frac{\alpha_k - D}{2}}}_{(1)} \times \underbrace{e^{-\frac{\tau_k}{2}(\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T \bar{\mathbf{r}}_k (\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)}}_{(2)} \underbrace{e^{-\frac{1}{2} tr(\mathbf{u}_k \bar{\mathbf{r}}_k)}}_{(3)} \Big]$$

$$= \begin{bmatrix} \frac{\alpha_k - D}{2} |\bar{\mathbf{r}}_k|^{\frac{\alpha_k - D - 1}{2}} |\bar{\mathbf{r}}_k| \times \bar{\mathbf{r}}_k^{-1} \times (2) \times (3) \\ + (1) \times (3) \times (2) \times -\frac{\tau_k}{2} (\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)(\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T \\ + (1) \times (2) \times (3) \times -\frac{1}{2} \mathbf{u}_k \end{bmatrix}$$

$$= \frac{\alpha_k - D}{2} \bar{\mathbf{r}}_k^{-1} - \frac{\tau_k}{2} (\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)(\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T - \frac{1}{2} \mathbf{u}_k$$

SCHMM

Full Covariance

- Full Covariance matrix case :

$$\therefore \left[\sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \left(\xi_t^{(m,n)}(k) \frac{1}{2} \left[\bar{\mathbf{r}}_k^{-1} - (\mathbf{x}_t^{(m,n)} - \bar{\mathbf{m}}_k)(\mathbf{x}_t^{(m,n)} - \bar{\mathbf{m}}_k)^T \right] \right) \right]$$

$$+ \left[\frac{\alpha_k - D}{2} \bar{\mathbf{r}}_k^{-1} - \frac{\tau_k}{2} (\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)(\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T - \frac{1}{2} \mathbf{u}_k \right] = 0$$

$$\Rightarrow \bar{\mathbf{r}}_k^{-1} \left\{ \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) + \alpha_k - D \right\}$$

$$= \mathbf{u}_k + \tau_k (\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)(\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k)(\mathbf{x}_t^{(m,n)} - \bar{\mathbf{m}}_k)(\mathbf{x}_t^{(m,n)} - \bar{\mathbf{m}}_k)^T$$

$$\Rightarrow \bar{\mathbf{r}}_k^{-1} = \frac{\mathbf{u}_k + \tau_k (\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)(\bar{\mathbf{m}}_k - \boldsymbol{\mu}_k)^T + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k)(\mathbf{x}_t^{(m,n)} - \bar{\mathbf{m}}_k)(\mathbf{x}_t^{(m,n)} - \bar{\mathbf{m}}_k)^T}{\sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) + \alpha_k - D}$$

SCHMM

Full Covariance

- The initial estimate can be chosen as the mode of the prior PDF

$$\pi_i^{(m)}, a_{ij}^{(m)}, w_{ik}^{(m)} \quad \text{same to DHMM}$$

and

$$\mathbf{m}_k = \boldsymbol{\mu}_k$$

$$\mathbf{r}_k = (\alpha_k - D)\mathbf{u}_k^{-1}$$

- And also can be chosen as the mean of the prior PDF

$$\pi_i^{(m)}, a_{ij}^{(m)}, w_{ik}^{(m)} \quad \text{same to DHMM}$$

and

$$\mathbf{m}_k = \boldsymbol{\mu}_k$$

$$\mathbf{r}_k = \alpha_k \mathbf{u}_k^{-1}$$

SCHMM

Diagonal Covariance

- Diagonal Covariance matrix case :
- Then

$$N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k) \propto \left(\prod_{d=1}^{D} \overline{r}_{kd} \right)^{1/2} e^{-\frac{1}{2}\sum_{d=1}^{D} (x_{td}^{(m,n)} - \overline{m}_{kd})^2 r_{kd}}$$

and

$$g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k) \propto \prod_{d=1}^{D} \left(\overline{r}_{kd}^{\alpha_{kd}-1/2} e^{-\frac{1}{2}\tau_{kd}\overline{r}_{kd}(\overline{m}_{kd}-\mu_{kd})^2} e^{-\beta_{kd}\overline{r}_{kd}} \right)$$

$$\log g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k) = \sum_{d=1}^{D} \log\left(\overline{r}_{kd}^{\alpha_{kd}-1/2} e^{-\frac{1}{2}\tau_{kd}\overline{r}_{kd}(\overline{m}_{kd}-\mu_{kd})^2} e^{-\beta_{kd}\overline{r}_{kd}} \right) + C$$

SCHMM

Diagonal Covariance

- Diagonal Covariance matrix case :

$$\frac{\partial \log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{m}_{kd}} = \frac{1}{N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}$$

$$\times \left(\prod_{d=1}^{D} \bar{r}_{kd} \right)^{1/2} \times e^{-\frac{1}{2}\sum_{d=1}^{D} \bar{r}_{kd}(x_{td}^{(m,n)} - \overline{m}_{kd})^2} \times (-\frac{1}{2}) \times (\bar{r}_{kd} \times 2)(x_{td}^{(m,n)} - \overline{m}_{kd}) \times (-1)$$

$$= \bar{r}_{kd}(x_{td}^{(m,n)} - \overline{m}_{kd})$$

$$\frac{\partial \log g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)}{\partial \overline{m}_{kd}} = \frac{1}{g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)} \times \left[\bar{r}_{kd}^{\alpha_{kd}-1/2} e^{-\frac{1}{2}\tau_{kd}\bar{r}_{kd}(\overline{m}_{kd}-\mu_{kd})^2} e^{-\beta_{kd}\bar{r}_{kd}} \right]'$$

$$= \frac{1}{g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)} \times \bar{r}_{kd}^{\alpha_{kd}-1/2} e^{-\frac{1}{2}\tau_{kd}\bar{r}_k(\overline{m}_{kd}-\mu_{kd})^2} e^{-\beta_{kd}\bar{r}_{kd}} \times -\frac{\tau_{kd}}{2}(\overline{m}_{kd} - \mu_{kd})(\bar{r}_{kd} \times 2)$$

$$= -\tau_{kd}\bar{r}_{kd}(\overline{m}_{kd} - \mu_{kd})$$

SCHMM

Diagonal Covariance

- Diagonal Covariance matrix case :

$$\therefore \left[\sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \left(\xi_t^{(m,n)}(k) \bar{r}_{kd} (x_{td}^{(m,n)} - \bar{m}_{kd}) \right) \right] - \tau_{kd} \bar{r}_{kd} (\bar{m}_{kd} - \mu_{kd}) = 0$$

$$\Rightarrow \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) \bar{m}_{kd} + \tau_{kd} \bar{m}_{kd} = \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) x_{td}^{(m,n)} + \tau_{kd} \mu_{kd}$$

$$\Rightarrow \bar{m}_{kd} = \frac{\tau_{kd} \mu_{kd} + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) x_{td}^{(m,n)}}{\tau_{kd} + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k)}$$

SCHMM

Diagonal Covariance

- Diagonal Covariance matrix case :

$$\frac{\partial \log N(\mathbf{x}_t^{(m,n)} \mid \bar{\mathbf{m}}_k, \bar{\mathbf{r}}_k)}{\partial \bar{r}_{kd}} = \frac{1}{N(\mathbf{x}_t^{(m,n)} \mid \bar{\mathbf{m}}_k, \bar{\mathbf{r}}_k)}$$

$$\times \left[\begin{array}{l} \frac{1}{2}\left(\prod_{d=1}^{D} r_{kd}\right)^{-1/2}\left(\prod_{i\neq d} r_{ki}\right) \times e^{-\frac{1}{2}\sum_{d=1}^{D}(x_{td}^{(m,n)}-\bar{m}_{kd})^2 \bar{r}_{kd}} \\ + e^{-\frac{1}{2}\sum_{d=1}^{D}(x_{td}^{(m,n)}-\bar{m}_{kd})^2 \bar{r}_{kd}} \times(-\frac{1}{2}) \times (x_{td}^{(m,n)} - \bar{m}_{kd}) \end{array} \right]$$

$$= \frac{1}{2}\left[\bar{r}_k^{-1} - (x_{td}^{(m,n)} - \bar{m}_{kd})^2\right]$$

SCHMM

Diagonal Covariance

- Diagonal Covariance matrix case :

$$\frac{\partial \log g(\bar{\mathbf{m}}_k, \bar{\mathbf{r}}_k)}{\partial \bar{r}_{kd}} = \frac{1}{g(\bar{\mathbf{m}}_k, \bar{\mathbf{r}}_k)} \times \left[|\bar{r}_{kd}|^{\alpha_{kd}-1/2} \times e^{-\frac{1}{2}\tau_{kd}\bar{r}_{kd}(\bar{m}_{kd}-\mu_{kd})^2} e^{-\beta_{kd}\bar{r}_{kd}} \right]$$

$$= \frac{1}{g(\bar{\mathbf{m}}_k, \bar{\mathbf{r}}_k)} \times \left[\begin{array}{l} (\alpha_{kd} - 1/2) \times \bar{r}_{kd}^{\alpha_{kd}-3/2} \times (2) \times (3) \\ + (1) \times (3) \times (2) \times -\frac{\tau_{kd}}{2} (\bar{m}_{kd} - \mu_{kd})^2 \\ + (1) \times (2) \times (3) \times (-\frac{1}{2} \beta_{kd}) \end{array} \right]$$

$$= (\alpha_{kd} - 1/2) \times \bar{r}_{kd}^{-1} - \frac{\tau_{kd}}{2} (\bar{m}_{kd} - \mu_{kd})^2 - \beta_{kd}$$

SCHMM

Diagonal Covariance

- Diagonal Covariance matrix case :

$$\therefore \left[\sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \left(\xi_t^{(m,n)}(k) \times \frac{1}{2} \left[\bar{r}_{kd}^{-1} - (x_{td}^{(m,n)} - \bar{m}_{kd})^2 \right] \right) \right]$$

$$+ (\alpha_{kd} - 1/2) \times \bar{r}_{kd}^{-1} - \frac{\tau_{kd}}{2} (\bar{m}_{kd} - \mu_{kd})^2 - \beta_{kd} = 0$$

$$\sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) \bar{r}_{kd}^{-1} + (2\alpha_{kd} - 1) \times \bar{r}_{kd}^{-1}$$

$$= 2\beta_{kd} + \tau_{kd} (\bar{m}_{kd} - \mu_{kd})^2 + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) (x_{td}^{(m,n)} - \bar{m}_{kd})^2$$

$$\bar{r}_{kd}^{-1} = \frac{2\beta_{kd} + \tau_{kd} (\bar{m}_{kd} - \mu_{kd})^2 + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k) (x_{td}^{(m,n)} - \bar{m}_{kd})^2}{(2\alpha_{kd} - 1) + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \xi_t^{(m,n)}(k)}$$

SCHMM

Diagonal Covariance

- The initial estimate can be chosen as the mode of the prior PDF

$$\pi_i^{(m)}, a_{ij}^{(m)}, w_{ik}^{(m)} \quad \text{same to DHMM}$$

and

$$m_{kd} = \mu_{kd}$$

$$r_{kd} = \frac{(\alpha_{kd} - 1/2)}{\beta_{kd}}$$

- And also can be chosen as the mean of the prior PDF $\pi_i^{(m)}, a_{ij}^{(m)}, w_{ik}^{(m)}$ same to DHMM

and

$$m_{kd} = \mu_{kd}$$

$$r_{kd} = \frac{\alpha_{kd}}{\beta_{kd}}$$

CDHMM

- Continuous Density HMM case:

Then

$$b_i(\mathbf{x}_t) = \sum_{k=1}^{K} w_{ik} N(\mathbf{x}_t \mid \mathbf{m}_k, \mathbf{r}_k)$$

$$\Downarrow$$

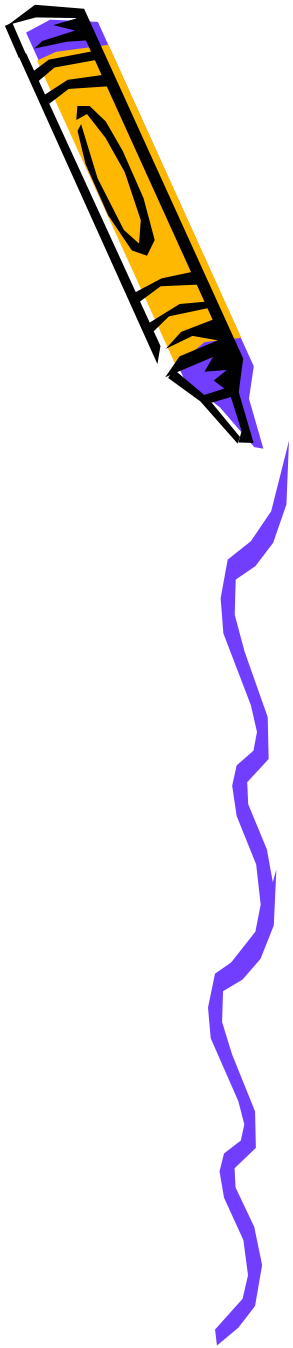
$$b_i(\mathbf{x}_t) = \sum_{k=1}^{K} w_{ik} N(\mathbf{x}_t \mid \mathbf{m}_{ik}, \mathbf{r}_{ik})$$

and

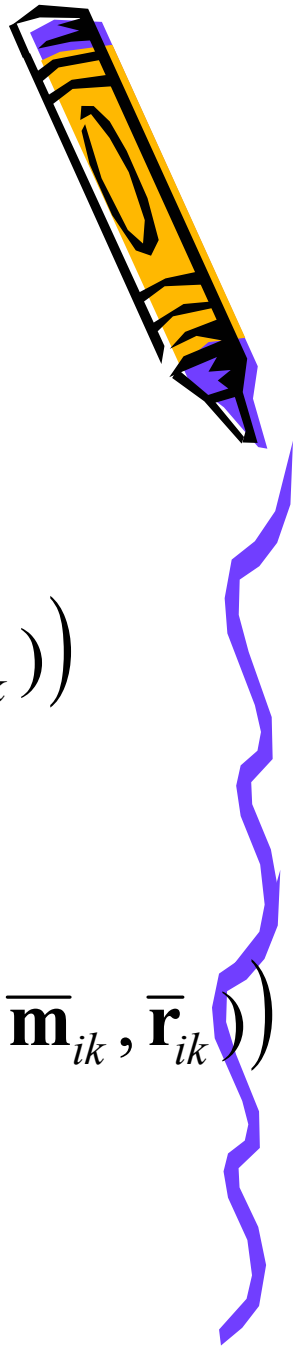
$$\text{where } N(\mathbf{x}_t \mid \mathbf{m}_k, \mathbf{r}_k) = (2\pi)^{-D/2} \mid \mathbf{r}_k \mid^{1/2} e^{-\frac{1}{2}(\mathbf{x}_t - \mathbf{m}_k)^T \mathbf{r}_k (\mathbf{x}_t - \mathbf{m}_k)}$$

$$\Downarrow$$

$$\text{where } N(\mathbf{x}_t \mid \mathbf{m}_{ik}, \mathbf{r}_{ik}) = (2\pi)^{-D/2} \mid \mathbf{r}_{ik} \mid^{1/2} e^{-\frac{1}{2}(\mathbf{x}_t - \mathbf{m}_{ik})^T \mathbf{r}_{ik} (\mathbf{x}_t - \mathbf{m}_{ik})}$$



CDHMM



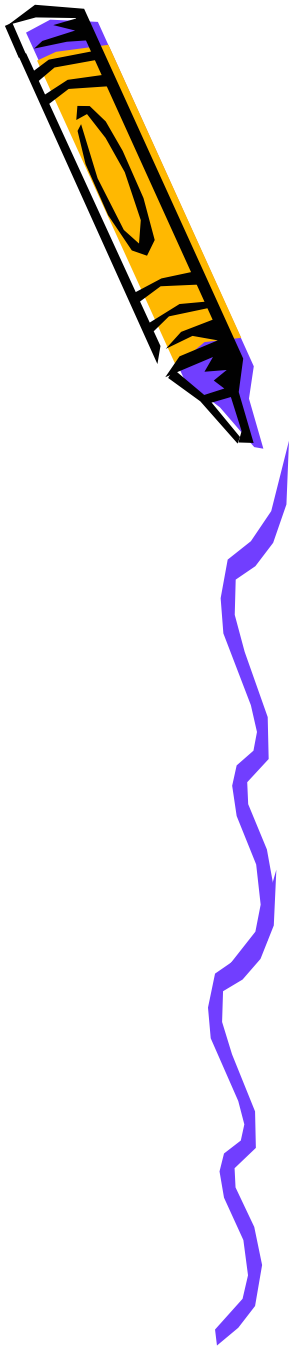
In Q – function

$$\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\left(\xi_t^{(m,n)}(k)\log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)\right)$$

$$\Downarrow$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\left(\xi_t^{(m,n)}(i,k)\log N(\mathbf{x}_t^{(m,n)} \mid \overline{\mathbf{m}}_{ik}, \overline{\mathbf{r}}_{ik})\right)$$

CDHMM



$$\ln \log g(\overline{\mathbf{\Lambda}}) \qquad \sum_{k=1}^{K} \log g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k)$$

$$\Downarrow$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \log g(\overline{\mathbf{m}}_{ik}, \overline{\mathbf{r}}_{ik})$$

$$\text{and} \quad g(\overline{\mathbf{m}}_k, \overline{\mathbf{r}}_k) \propto |\mathbf{r}_k|^{\frac{\alpha_k - D}{2}} e^{-\frac{\tau_k}{2}(\mathbf{m}_k - \boldsymbol{\mu}_k)^T \gamma_k (\mathbf{m}_k - \boldsymbol{\mu}_k)} e^{-\frac{1}{2} tr(\mathbf{u}_k \mathbf{r}_k)}$$

$$\Downarrow \text{(Full covariance case)}$$

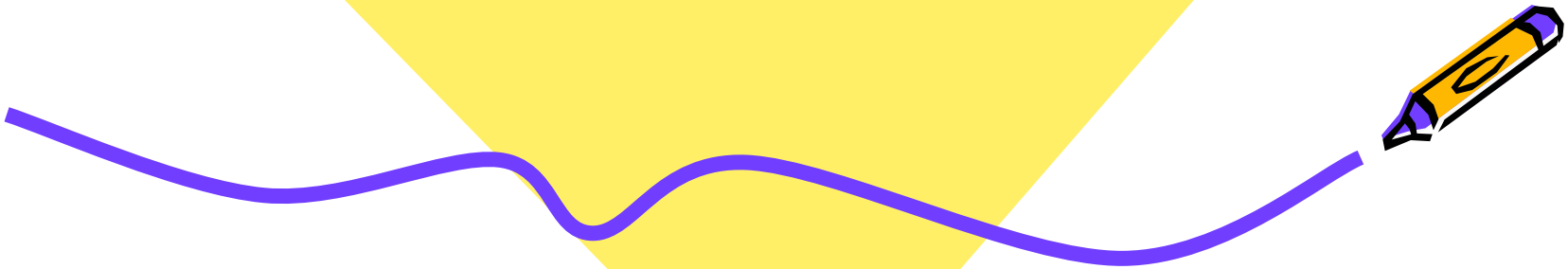
$$g(\overline{\mathbf{m}}_{ik}, \overline{\mathbf{r}}_{ik}) \propto |\mathbf{r}_{ik}|^{\frac{\alpha_{ik} - D}{2}} e^{-\frac{\tau_{ik}}{2}(\mathbf{m}_{ik} - \boldsymbol{\mu}_{ik})^T \boldsymbol{\gamma}_{ik} (\mathbf{m}_{ik} - \boldsymbol{\mu}_{ik})} e^{-\frac{1}{2} tr(\mathbf{u}_{ik} \mathbf{r}_{ik})}$$

$$\text{and} \quad g(\mathbf{m}_k, \mathbf{r}_k) \propto \prod_{d=1}^{D} r_{kd}^{\frac{\alpha_{kd} - 1/2}{2}} e^{-\frac{\tau_{kd}}{2} r_{kd} (m_{kd} - \mu_{kd})^2} e^{-\beta_{kd} r_{kd}}$$

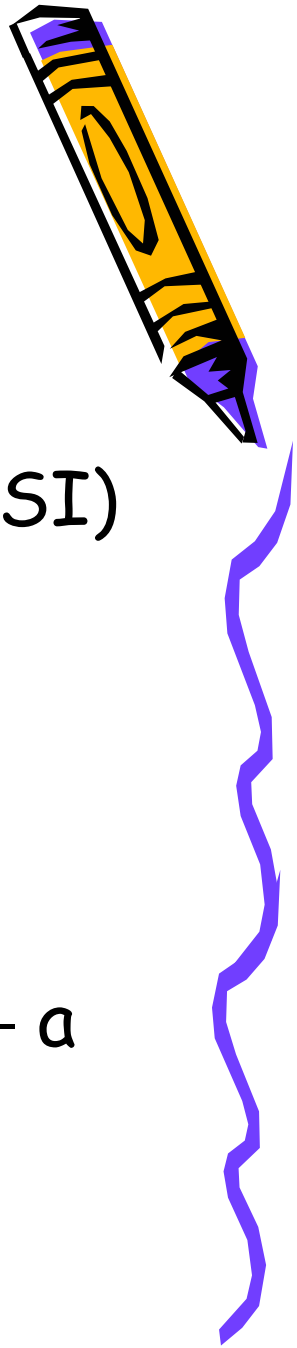
$$\Downarrow \text{(Diagonal covariance case)}$$

$$g(\mathbf{m}_{ik}, \mathbf{r}_{ik}) \propto \prod_{d=1}^{D} r_{ikd}^{\frac{\alpha_{ikd} - 1/2}{2}} e^{-\frac{\tau_{ikd}}{2} r_{ikd} (m_{ikd} - \mu_{ikd})^2} e^{-\beta_{ikd} r_{ikd}}$$

Maximum Likelihood Linear Regression



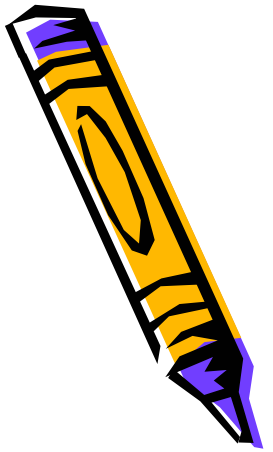
MLLR



Background

- Linear transformation of original model (SI) to maximize likelihood of adaptation
- MLLR is multiplicative; MAP is additive
- MLLR much less conservative than MAP – a few sec. of data may change model dramatically.

MLLR

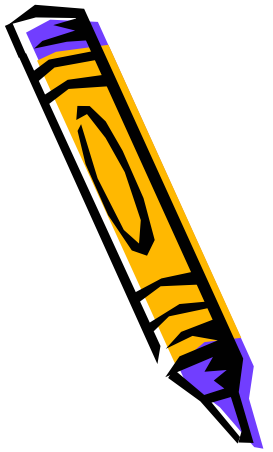


Reference :

- Speaker Adaptation of HMMs Using Linear Regression – TR'94 Leggetter and Woodland
- Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models – CSL'95 Leggetter and Woodland
- MLLR:A Speaker Adaptation Technique for LVCSR – Hamaker

MLLR

Single Gaussian Case



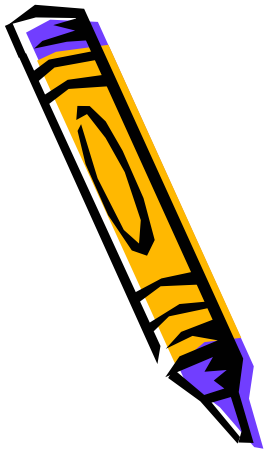
- The regression transform is first derived for the single Gaussian distribution pre state, and later extended to the general case of Gaussian mixtures.
- So, the p.d.f for the state s is

$$b_j(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} \mid \mathbf{C}_j \mid^{1/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu}_j)^T \mathbf{C}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)}$$

$\boldsymbol{\mu}_j$ is the mean and $\mathbf{C}_j$ is the covariance matrix

MLLR

Single Gaussian Case



If $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}$ is the mean, then we define $\boldsymbol{\xi} = \begin{bmatrix} \omega \\ \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}$

where ω is the offset term for the regression

The estimate of the adapted mean is $\bar{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} = \mathbf{W}\boldsymbol{\xi}$

where $\mathbf{W} = (\mathbf{A}, \mathbf{b})$ is the linear transform (an $D \times (D+1)$ matrix)

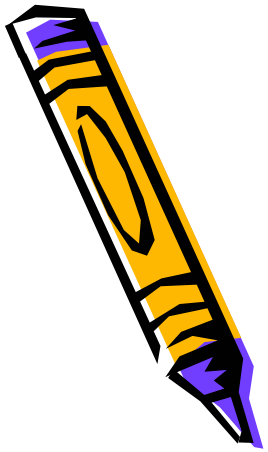
If $\omega = 1 \Rightarrow$ include an offset in the regression

If $\omega = 0 \Rightarrow$ ignore the offsets

So $b_j(\mathbf{x}) = \dfrac{1}{(2\pi)^{D/2} \, |\mathbf{C}_j|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{W}_j\boldsymbol{\xi}_j)^T \mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{W}_j\boldsymbol{\xi}_j)}$

MLLR

Single Gaussian Case



- A more general approach is adopted in which the same transformations matrix is used for several distributions. ←Regression Class

- If some of the distributions are not observed in the adaptation data, a transformation may still be applied.← Models would update whether correspond adaptation data observed or not.

MLLR

Single Gaussian Case

- MLLR estimates the regression matrices to maximize the likelihood of the adapted models generation the adaptation data.← Maximize the likelihood to obtain the regression matrices.

- Full and Diagonal covariance cases will be discussed.

MLLR

Single Gaussian Case

Assume the adaptation data, X, is a series of T observations.

$$\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$$

Denote the current set of model parameters by λ and a re-estimated set of model parameters as $\bar{\lambda}$

Current extended mean → ξ

Re-estimated mean → $\bar{\mu}$

MLLR

Single Gaussian Case

The total likelihood is

$$f(\mathbf{X} \mid \boldsymbol{\lambda}) = \sum_{\mathbf{S}} f(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\lambda})$$

$f(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\lambda})$ is the likelihood of generating $\mathbf{X}$ using the state sequence $\mathbf{S}$ given model $\boldsymbol{\lambda}$

The quantity $f(\mathbf{X} \mid \boldsymbol{\lambda})$ is the objective function to be maximized during adaptation.

MLLR

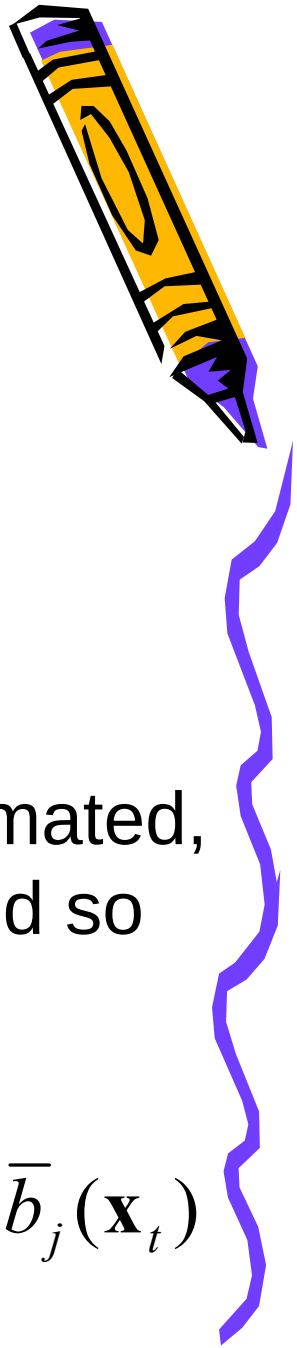
Single Gaussian Case

We define the auxiliary function

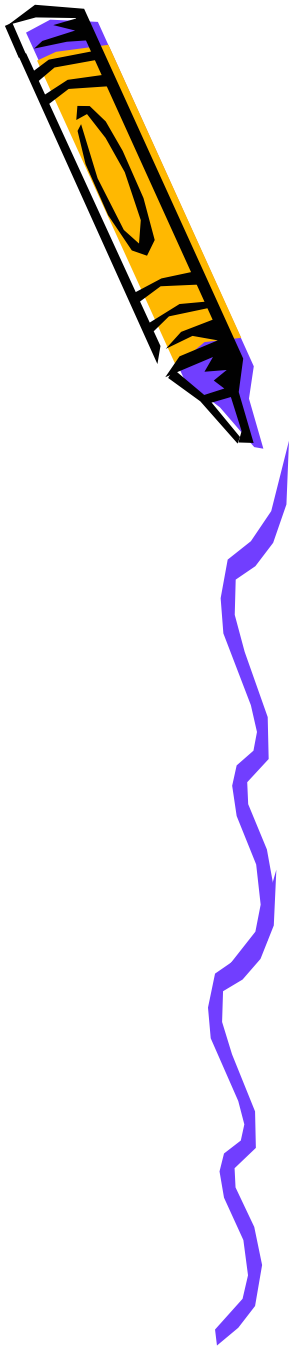
$$Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = \sum_{\mathbf{S}} f(\mathbf{S} \mid \mathbf{X}, \boldsymbol{\lambda}) \log\left[f(\mathbf{X}, \mathbf{S} \mid \bar{\boldsymbol{\lambda}})\right]$$

Since only the transformations $\mathbf{W}_j$ are re-estimated, only the output distributions $b_j(\mathbf{x}_t)$ are affected so the auxiliary function can be written as

$$Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = constant + \sum_{\mathbf{S}} \sum_{t=1}^{T} f(s_t = j \mid \mathbf{X}, \boldsymbol{\lambda}) \log \bar{b}_j(\mathbf{x}_t)$$



MLLR



Single Gaussian Case

$$\text{We define} \quad \gamma_j(t) = \sum_{\mathbf{s}} f(s_t = j \mid \mathbf{X}, \boldsymbol{\lambda})$$

So…The Q-function can be rewritten as

$$Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = constant + \sum_{t=1}^{T} \gamma_j(t) \log b_j(\mathbf{x}_t)$$

MLLR

Single Gaussian Case

Expanding $\log b_j(\mathbf{x}_t)$ then the auxiliary function is

$$Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = constant - \frac{1}{2} \times \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \left[D \log(2\pi) + \log |\mathbf{C}_j| + h(\mathbf{x}_t, j) \right]$$

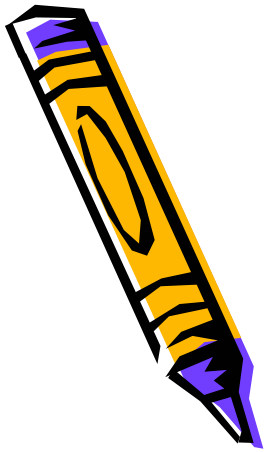
where $h(\mathbf{x}_t, j) = (\mathbf{x}_t - \overline{\mathbf{W}}_j \boldsymbol{\xi}_j)^T \mathbf{C}_j^{-1} (\mathbf{x}_t - \overline{\mathbf{W}}_j \boldsymbol{\xi}_j)$

The differential of $Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda})$ w.r.t $\overline{\mathbf{W}}_s$ is

$$\frac{\partial Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda})}{\partial \overline{\mathbf{W}}_s} = -\frac{1}{2} \frac{\partial}{\partial \overline{\mathbf{W}}_s} \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \left[D \log(2\pi) + \log |\mathbf{C}_j| + h(\mathbf{x}_t, j) \right]$$

MLLR

Single Gaussian Case



The differential of $h(\mathbf{x}_t, j)$ w.r.t $\overline{\mathbf{W}}_j$ is

$$\frac{\partial h(\mathbf{x}_t, j)}{\partial \overline{\mathbf{W}}_j} = \frac{\partial}{\partial \overline{\mathbf{W}}_j}(\mathbf{x}_t - \overline{\mathbf{W}}_j \boldsymbol{\xi}_j)^T \mathbf{C}_j^{-1}(\mathbf{x}_t - \overline{\mathbf{W}}_j \boldsymbol{\xi}_j)$$

$$= \frac{\partial}{\partial \overline{\mathbf{W}}_j}(\mathbf{x}_t^T - \boldsymbol{\xi}_j^T \overline{\mathbf{W}}_j^T)\mathbf{C}_j^{-1}(\mathbf{x}_t - \overline{\mathbf{W}}_j \boldsymbol{\xi}_j)$$

$$= \frac{\partial}{\partial \overline{\mathbf{W}}_j}\left[\mathbf{x}_t^T \mathbf{C}_j^{-1}\mathbf{x}_t - \boldsymbol{\xi}_j^T \overline{\mathbf{W}}_j^T \mathbf{C}_j^{-1}\mathbf{x}_t - \mathbf{x}_t^T \mathbf{C}_j^{-1}\overline{\mathbf{W}}_j \boldsymbol{\xi}_j + \boldsymbol{\xi}_j^T \overline{\mathbf{W}}_j^T \mathbf{C}_j^{-1}\overline{\mathbf{W}}_j \boldsymbol{\xi}_j\right]$$

$$= \frac{\partial}{\partial \overline{\mathbf{W}}_j}\left[-\boldsymbol{\xi}_j^T \overline{\mathbf{W}}_j^T \mathbf{C}_j^{-1}\mathbf{x}_t - \left(\mathbf{C}_j^{-T}\mathbf{x}_t\right)^T \overline{\mathbf{W}}_j \boldsymbol{\xi}_j + \boldsymbol{\xi}_j^T \overline{\mathbf{W}}_j^T \mathbf{C}_j^{-1}\overline{\mathbf{W}}_j \boldsymbol{\xi}_j\right]$$

$$= -\mathbf{C}_j^{-1}\mathbf{x}_t \boldsymbol{\xi}_j^T - \mathbf{C}_j^{-T}\mathbf{x}_t \boldsymbol{\xi}_j^T + \mathbf{C}_j^{-T}\overline{\mathbf{W}}_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T + \mathbf{C}_j^{-1}\overline{\mathbf{W}}_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T \left[\because \mathbf{C}_j^{-1} = \mathbf{C}_j^{-T}\right]$$

$$= -2\mathbf{C}_j^{-1}\left[\mathbf{x}_t - \overline{\mathbf{W}}_j \boldsymbol{\xi}_j\right]\boldsymbol{\xi}_j^T$$

MLLR

Single Gaussian Case

Then complete the differentiation, and equating to zero.

$$\frac{\partial}{\partial \overline{\mathbf{W}}_j} Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = \sum_{t=1}^{T} \gamma_j(t) \mathbf{C}_j^{-1} \left[\mathbf{x}_t - \overline{\mathbf{W}}_j \boldsymbol{\xi}_j \right] \boldsymbol{\xi}_j^T = 0$$

$$\therefore \sum_{t=1}^{T} \gamma_j(t) \mathbf{C}_j^{-1} \mathbf{x}_t \boldsymbol{\xi}_j^T = \sum_{t=1}^{T} \gamma_j(t) \mathbf{C}_j^{-1} \overline{\mathbf{W}}_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T$$

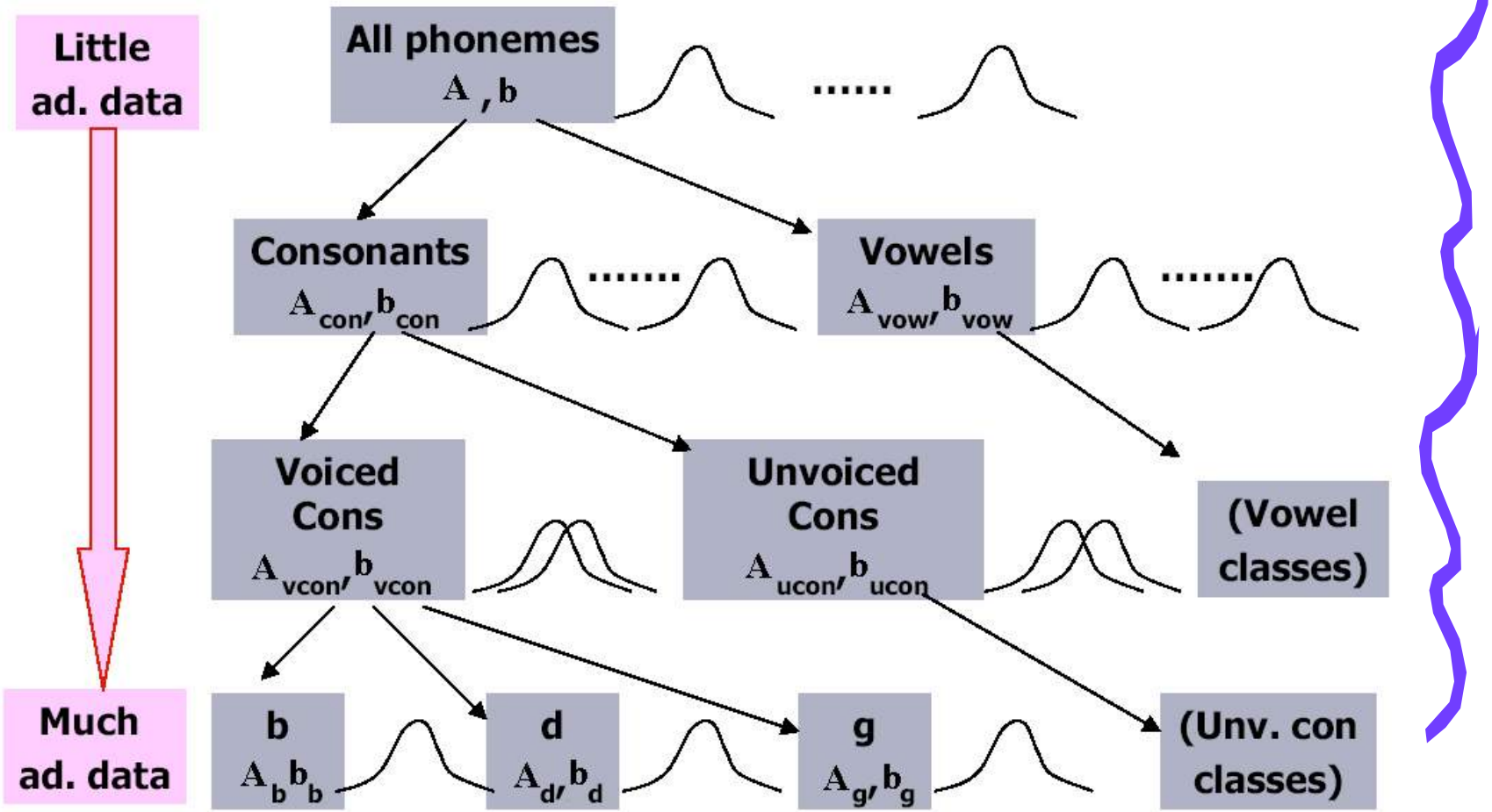
$$\Rightarrow \mathbf{C}_j^{-1} \left(\sum_{t=1}^{T} \gamma_j(t) \mathbf{x}_t \right) \boldsymbol{\xi}_j^T = \mathbf{C}_j^{-1} \overline{\mathbf{W}}_j \left(\sum_{t=1}^{T} \gamma_j(t) \right) \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T$$

$$\Rightarrow \sum_{t=1}^{T} \gamma_j(t) \mathbf{x}_t = \overline{\mathbf{W}}_j \left(\sum_{t=1}^{T} \gamma_j(t) \right) \boldsymbol{\xi}_j \qquad \therefore \bar{\boldsymbol{\mu}}_j = \overline{\mathbf{W}}_j \boldsymbol{\xi}_j = \frac{\sum_{t=1}^{T} \gamma_j(t) \mathbf{x}_t}{\sum_{t=1}^{T} \gamma_j(t)}$$

MLLR

Tied Regression Matrices

- Regression Class Tree for MLLR



MLLR

Tied Regression Matrices

Consider the sth regression class $RC^{(s)} = \{s_1, \ldots, s_R\}$

If W_s is shared by the states in the regression class $RC^{(s)}$, then

$$\sum_{t=1}^{T} \sum_{r=1}^{R} \gamma_{s_r}(t) \mathbf{C}_{s_r}^{-1} \mathbf{x}_t \boldsymbol{\xi}_{s_r}^T = \sum_{t=1}^{T} \sum_{r=1}^{R} \gamma_{s_r}(t) \mathbf{C}_{s_r}^{-1} \overline{\mathbf{W}}_s \boldsymbol{\xi}_{s_r} \boldsymbol{\xi}_{s_r}^T$$

$$\Rightarrow \sum_{r=1}^{R} \left\{ \mathbf{C}_{s_r}^{-1} \left(\sum_{t=1}^{T} \gamma_{s_r}(t) \mathbf{x}_t \right) \boldsymbol{\xi}_{s_r}^T \right\}_{D \times (D+1)}$$

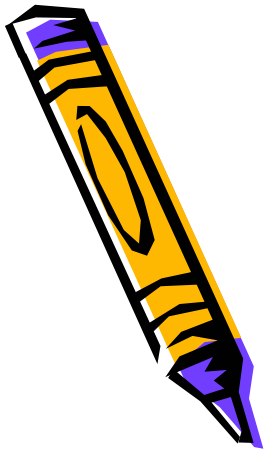
$$= \sum_{r=1}^{R} \left\{ \left[\left(\sum_{t=1}^{T} \gamma_{s_r}(t) \right) \mathbf{C}_{s_r}^{-1} \right]_{D \times D} \left[\overline{\mathbf{W}}_s \right]_{D \times (D+1)} \left[\boldsymbol{\xi}_{s_r} \boldsymbol{\xi}_{s_r}^T \right]_{(D+1) \times (D+1)} \right\}$$

$$[\mathbf{Z}]_{D \times (D+1)} = \left[\sum_{r=1}^{R} \mathbf{V}^{(r)} \overline{\mathbf{W}}_s \mathbf{D}^{(r)} \right]_{D \times (D+1)} \quad \text{where } \mathbf{Z} = \sum_{r=1}^{R} \left\{ \mathbf{C}_{s_r}^{-1} \left(\sum_{t=1}^{T} \gamma_{s_r}(t) \mathbf{x}_t \right) \boldsymbol{\xi}_{s_r}^T \right\}_{D \times (D+1)}$$

$$\mathbf{V}^{(r)} = \left[\left(\sum_{t=1}^{T} \gamma_{s_r}(t) \right) \mathbf{C}_{s_r}^{-1} \right]_{D \times D} \quad \mathbf{D}^{(r)} = \left[\boldsymbol{\xi}_{s_r} \boldsymbol{\xi}_{s_r}^T \right]_{(D+1) \times (D+1)}$$

MLLR

Tied Regression Matrices



If right hand side is denoted by the $D \times (D+1)$ matrix $\mathbf{Y}$

$$[\mathbf{Z}]_{D\times(D+1)} = [\mathbf{Y}]_{D\times(D+1)} \Rightarrow [\mathbf{Z}]_{ij} = [\mathbf{Y}]_{ij}$$

$$\text{where } \mathbf{Y} = \left[\sum_{r=1}^{R} [\mathbf{V}^{(r)}]_{D\times D} [\overline{\mathbf{W}}_s]_{D\times(D+1)} [\mathbf{D}^{(r)}]_{(D+1)\times(D+1)} \right]_{D\times(D+1)}$$

$$[\mathbf{V}^{(r)}\overline{\mathbf{W}}_s]_{ij} = \sum_{k=1}^{D} [\mathbf{V}^{(r)}]_{ik} [\overline{\mathbf{W}}_s]_{kj}$$

$$[\mathbf{Y}]_{ij} = \sum_{r=1}^{R} \sum_{q=1}^{D+1} \left\{ [\mathbf{V}^{(r)}\overline{\mathbf{W}}_s]_{iq} [\mathbf{D}^{(r)}]_{qj} \right\} = \sum_{r=1}^{R} \sum_{q=1}^{D+1} \left\{ \left[\sum_{p=1}^{D} [\mathbf{V}^{(r)}]_{ip} [\overline{\mathbf{W}}_s]_{pq} \right] [\mathbf{D}^{(r)}]_{qj} \right\}$$

$$= \sum_{p=1}^{D} \sum_{q=1}^{D+1} [\overline{\mathbf{W}}_s]_{pq} \left[\sum_{r=1}^{R} [\mathbf{V}^{(r)}]_{ip} [\mathbf{D}^{(r)}]_{qj} \right]$$

MLLR

Tied Regression Matrices

If the covariance matrix is diagonal $\Rightarrow \mathbf{V}^{(r)}$ is diagonal and $\mathbf{D}^{(r)}$ is symmetric

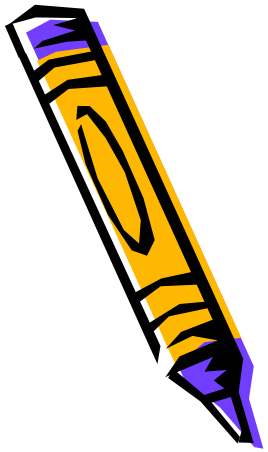
$$\sum_{r=1}^{R} \left[\mathbf{V}^{(r)}\right]_{ip}\left[\mathbf{D}^{(r)}\right]_{qj} = \begin{cases} \sum_{r=1}^{R} \left[\mathbf{V}^{(r)}\right]_{ip}\left[\mathbf{D}^{(r)}\right]_{jq} & i = p \\ 0 & i \neq p \end{cases}$$

$$\therefore [\mathbf{Z}]_{ij} = [\mathbf{Y}]_{ij} = \sum_{p=1}^{D}\sum_{q=1}^{D+1} \left[\overline{\mathbf{W}}_s\right]_{pq} \left[\sum_{r=1}^{R} \left[\mathbf{V}^{(r)}\right]_{ip}\left[\mathbf{D}^{(r)}\right]_{qj}\right] \rightarrow [\mathbf{G}]_{pq}^{(i,j)}$$

$$= \sum_{q=1}^{D+1} \left[\overline{\mathbf{W}}_s\right]_{iq} \left[\sum_{r=1}^{R} \left[\mathbf{V}^{(r)}\right]_{ii}\left[\mathbf{D}^{(r)}\right]_{jq}\right] = \sum_{q=1}^{D+1} \left[\overline{\mathbf{W}}_s\right]_{iq} [\mathbf{G}]_{jq}^{(i)}$$

MLLR

Tied Regression Matrices

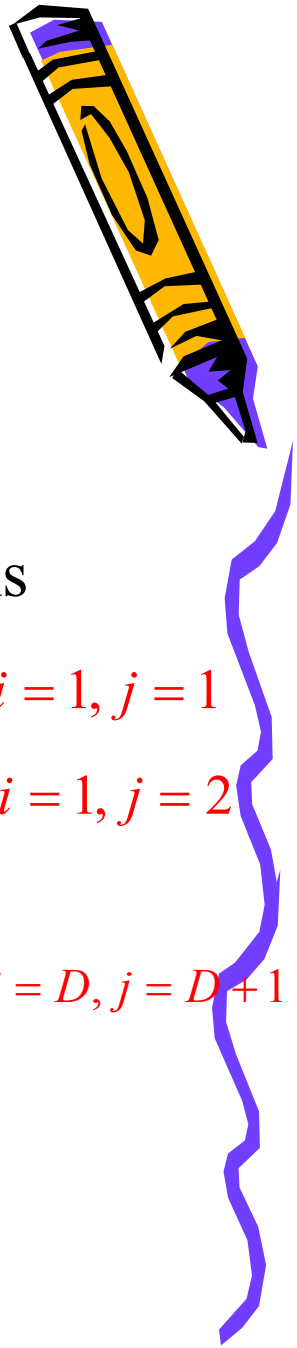


Then we can obtain a row i of $\left[\overline{\mathbf{W}}_s\right]$ by solving below linear equations

$$
\begin{cases}
\left[\mathbf{G}\right]^{(i)}_{1,1}\left[\overline{\mathbf{W}}_s\right]_{i,1} + \left[\mathbf{G}\right]^{(i)}_{1,2}\left[\overline{\mathbf{W}}_s\right]_{i,2} + \cdots + \left[\mathbf{G}\right]^{(i)}_{1,D+1}\left[\overline{\mathbf{W}}_s\right]_{i,D+1} = \left[\mathbf{Z}\right]_{i,1} & \leftarrow j=1 \\
\left[\mathbf{G}\right]^{(i)}_{2,1}\left[\overline{\mathbf{W}}_s\right]_{i,1} + \left[\mathbf{G}\right]^{(i)}_{2,2}\left[\overline{\mathbf{W}}_s\right]_{i,2} + \cdots + \left[\mathbf{G}\right]^{(i)}_{2,D+1}\left[\overline{\mathbf{W}}_s\right]_{i,D+1} = \left[\mathbf{Z}\right]_{i,2} & \leftarrow j=2 \\
\vdots & \\
\left[\mathbf{G}\right]^{(i)}_{D+1,1}\left[\overline{\mathbf{W}}_s\right]_{i,1} + \left[\mathbf{G}\right]^{(i)}_{D+1,2}\left[\overline{\mathbf{W}}_s\right]_{i,2} + \cdots + \left[\mathbf{G}\right]^{(i)}_{D+1,D+1}\left[\overline{\mathbf{W}}_s\right]_{i,D+1} = \left[\mathbf{Z}\right]_{i,D+1} & \leftarrow j=D+1
\end{cases}
$$

MLLR

Tied Regression Matrices



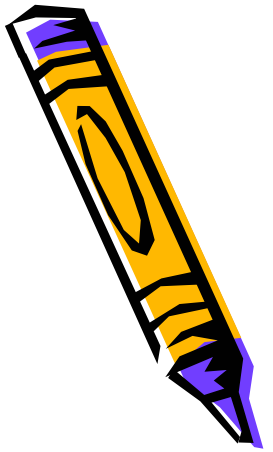
If the covariance matrix is still full,

we could obtain $\left[\overline{\mathbf{W}}_s\right]$ by solving below linear equations

$$\begin{cases} \left[\mathbf{G}\right]_{1,1}^{(1,1)}\left[\overline{\mathbf{W}}_s\right]_{1,1} + \cdots + \left[\mathbf{G}\right]_{D,D+1}^{(1,1)}\left[\overline{\mathbf{W}}_s\right]_{D,D+1} = \left[\mathbf{Z}\right]_{1,1} & \leftarrow i=1, j=1 \\ \left[\mathbf{G}\right]_{1,1}^{(1,2)}\left[\overline{\mathbf{W}}_s\right]_{1,1} + \cdots + \left[\mathbf{G}\right]_{D,D+1}^{(1,2)}\left[\overline{\mathbf{W}}_s\right]_{D,D+1} = \left[\mathbf{Z}\right]_{1,2} & \leftarrow i=1, j=2 \\ \vdots & \\ \left[\mathbf{G}\right]_{1,1}^{(D,D+1)}\left[\overline{\mathbf{W}}_s\right]_{1,1} + \cdots + \left[\mathbf{G}\right]_{D,D+1}^{(D,D+1)}\left[\overline{\mathbf{W}}_s\right]_{D,D+1} = \left[\mathbf{Z}\right]_{D,D+1} & \leftarrow i=D, j=D+1 \end{cases}$$

MLLR

Mixture Gaussian Case



- Then the p.d.f for the state *j* would be

$$b_j(\mathbf{x}) = \sum_{k=1}^{K} w_{jk} \frac{1}{(2\pi)^{D/2} |\mathbf{C}_{jk}|^{1/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu}_{jk})^T \mathbf{C}_{jk}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{jk})}$$

$\boldsymbol{\mu}_{jk}$ is the mean and $\mathbf{C}_{jk}$ is the covariance matrix and w_{jk} is the mixture weight

- Then likelihood

$$f(\mathbf{X}|\boldsymbol{\lambda}) = \sum_{\mathbf{S}} \sum_{\mathbf{L}} f(\mathbf{X}, \mathbf{S}, \mathbf{L}|\boldsymbol{\lambda})$$

where **S** is one possible state sequence

and **L** is one possible mixture sequence

MLLR

Mixture Gaussian Case

- Then Q-function will be

$$Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = \sum_{\mathbf{S}} \sum_{\mathbf{L}} f(\mathbf{S}, \mathbf{L} \mid \mathbf{X}, \boldsymbol{\lambda}) \log f(\mathbf{X}, \mathbf{S}, \mathbf{L} \mid \bar{\boldsymbol{\lambda}})$$

Only consider the term which dependent on the regression transform.

$$\therefore Q_b(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = \sum_{\mathbf{S}} \sum_{\mathbf{L}} \sum_{t=1}^{T} f(s_t = j, l_t = k \mid \mathbf{X}, \boldsymbol{\lambda}) \log \bar{b}_{jk}(\mathbf{x}_t)$$

$$= \sum_{t=1}^{T} \gamma_{jk}(t) \log \bar{b}_{jk}(\mathbf{x}_t)$$

where $\gamma_{jk}(t) = \sum_{\mathbf{S}} \sum_{\mathbf{L}} f(s_t = j, l_t = k \mid \mathbf{X}, \boldsymbol{\lambda})$

- The derivation is the same as single Gaussian case, just substitute for $\gamma_j(t)$ $\gamma_{jk}(t)$

MLLR

Least Squares Regression

- If all the covariance of the distributions tied to the same transformation are the same← a special case of MLLR

- Then

$$\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)\mathbf{C}_{s_r}^{-1}\mathbf{x}_t\xi_{s_r}^{T} = \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)\mathbf{C}_{s_r}^{-1}\overline{\mathbf{W}}_s\xi_{s_r}\xi_{s_r}^{T}$$

can be rewritten as

$$\Rightarrow \quad \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)\mathbf{x}_t\xi_{s_r}^{T} = \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)\overline{\mathbf{W}}_s\xi_{s_r}\xi_{s_r}^{T}$$

MLLR

Least Squares Regression

- If each frame is assigned to exactly one distribution (Viterbi alignment)

$$\gamma_{s_r}(t) = \begin{cases} 1 & \text{if } \mathbf{x}_t \text{ is assigned to state distribution } s_r \\ 0 & \text{otherwise} \end{cases}$$

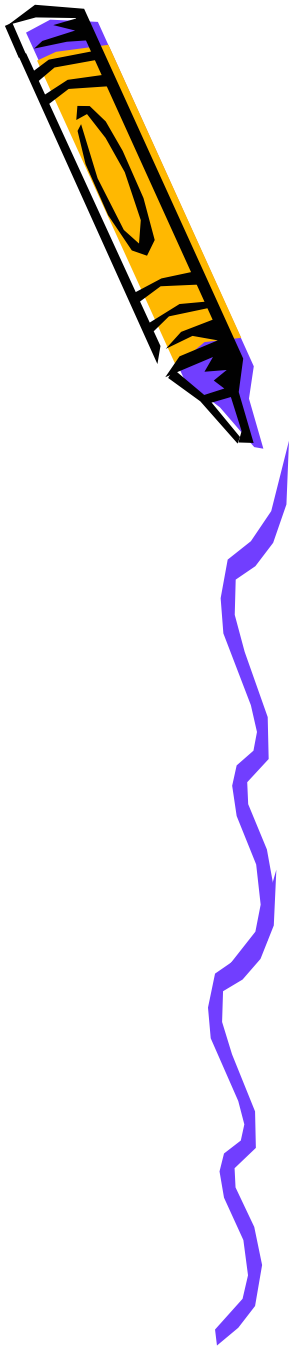
- Then $\sum_{t=1}^{T} \delta_{RC^{(n)},s_t} \mathbf{x}_t \xi_{s_t}^T = \sum_{t=1}^{T} \delta_{RC^{(n)},s_t} \overline{\mathbf{W}}_s \xi_{s_t} \xi_{s_t}^T$

$$\text{where} \quad \delta_{RC^{(n)},s_t} = \begin{cases} 1 & s_t \in RC^{(n)} \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \sum_{t=1}^{T} \mathbf{x}_t \xi_{s_t}^T \delta_{RC^{(n)},s_t} = \overline{\mathbf{W}}_s \sum_{t=1}^{T} \xi_{s_t} \xi_{s_t}^T \delta_{RC^{(n)},s_t}$$

MLLR

Least Squares Regression



Define matrices $\mathbf{X}, \mathbf{Y}$ as

$$\mathbf{X} = \left[\xi_{s_1}, \xi_{s_2}, \cdots, \xi_{s_T}\right]$$

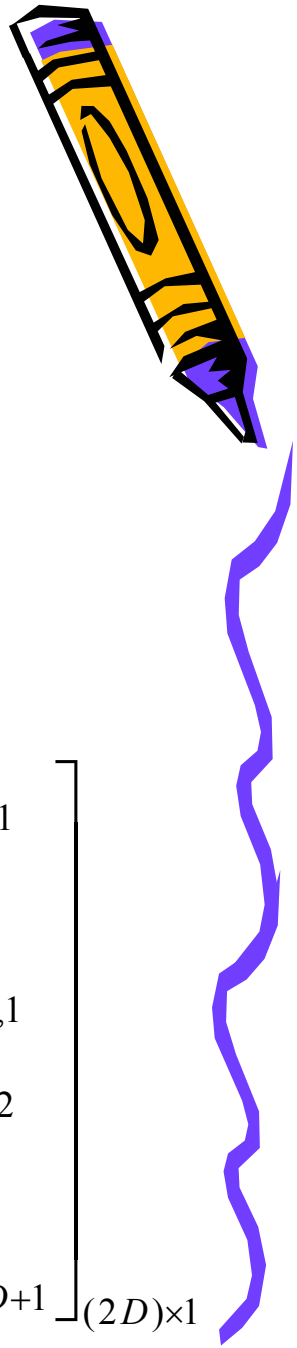
$$\mathbf{Y} = \left[\mathbf{x}_1 \delta_{RC^{(n)},s_1}, \mathbf{x}_2 \delta_{RC^{(n)},s_2}, \cdots, \mathbf{x}_T \delta_{RC^{(n)},s_T}\right]$$

then $\overline{\mathbf{W}}_s \mathbf{X}\mathbf{X}^T = \mathbf{Y}\mathbf{X}^T$

$$\overline{\mathbf{W}}_s = \mathbf{Y}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}$$

MLLR

Single Variable Linear Regression



- If the scaling portion of the regression matrix is assumed to be diagonal, the computation can be vastly reduced.

It means that $\overline{\mu}_i = x + y\mu_i$

$$\therefore \overline{\mathbf{W}}_s = \begin{bmatrix} w_{1,1} & w_{1,2} & 0 & \cdots & 0 \\ w_{2,1} & 0 & w_{2,3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{D,1} & 0 & 0 & \cdots & w_{D,D+1} \end{bmatrix}_{D\times(D+1)} \Rightarrow \overline{\mathbf{w}}_s = \begin{bmatrix} w_{1,1} \\ \vdots \\ w_{D,1} \\ w_{1,2} \\ \vdots \\ w_{D,D+1} \end{bmatrix}_{(2D)\times 1}$$

MLLR

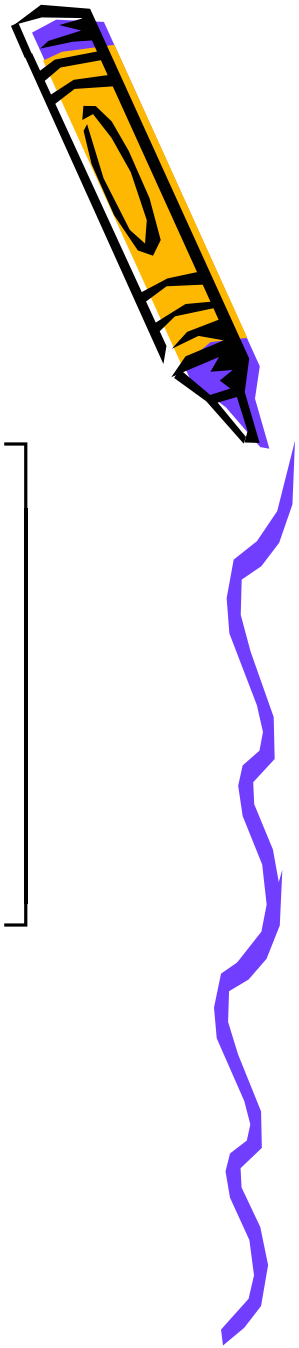
Single Variable Linear Regression

And define an $D \times 2D$ *matrix* $\mathbf{D}_s$

$$\mathbf{D}_s = \begin{bmatrix} \omega & 0 & \cdots & 0 & 0 & \mu_1 & 0 & \cdots & 0 & 0 \\ 0 & \omega & \cdots & 0 & 0 & 0 & \mu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \omega & 0 & 0 & 0 & \cdots & \mu_{D-1} & 0 \\ 0 & 0 & \cdots & 0 & \omega & 0 & 0 & \cdots & 0 & \mu_D \end{bmatrix}$$

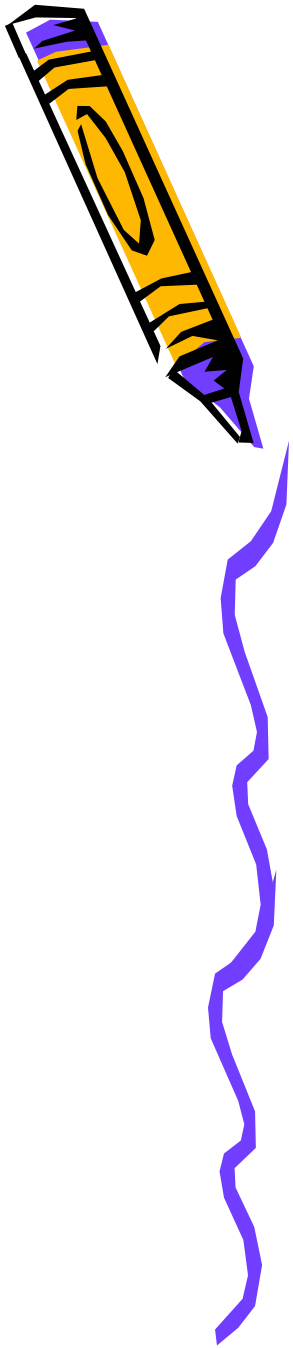
$$h(\mathbf{x}_t, s) = (\mathbf{x}_t - \overline{\mathbf{W}}_s \boldsymbol{\xi}_s)^T \mathbf{C}_s^{-1} (\mathbf{x}_t - \overline{\mathbf{W}}_s \boldsymbol{\xi}_s)$$

$$= (\mathbf{x}_t - \mathbf{D}_s \overline{\mathbf{w}}_s)^T \mathbf{C}_s^{-1} (\mathbf{x}_t - \mathbf{D}_s \overline{\mathbf{w}}_s)$$



MLLR

Single Variable Linear Regression



$$\frac{\partial h(\mathbf{x}_t, s)}{\partial \overline{\mathbf{w}}_s} = \frac{\partial}{\partial \overline{\mathbf{w}}_s}(\mathbf{x}_t - \mathbf{D}_s\overline{\mathbf{w}}_s)^T\mathbf{C}_s^{-1}(\mathbf{x}_t - \mathbf{D}_s\overline{\mathbf{w}}_s)$$

$$= \frac{\partial}{\partial \overline{\mathbf{w}}_s}\left[\mathbf{x}_t^T\mathbf{C}_s^{-1}\mathbf{x}_t - \mathbf{x}_t^T\mathbf{C}_s^{-1}\mathbf{D}_s\overline{\mathbf{w}}_s - \overline{\mathbf{w}}_s^T\mathbf{D}_s^T\mathbf{C}_s^{-1}\mathbf{x}_t - \overline{\mathbf{w}}_s^T\mathbf{D}_s^T\mathbf{C}_s^{-1}\mathbf{D}_s\overline{\mathbf{w}}_s\right]$$

$$= 0 - \left[\mathbf{x}_t^T\mathbf{C}_s^{-1}\mathbf{D}_s\right]^T - \mathbf{D}_s^T\mathbf{C}_s^{-1}\mathbf{x}_t - \left(\mathbf{D}_s^T\mathbf{C}_s^{-1}\mathbf{D}_s + \left[\mathbf{D}_s^T\mathbf{C}_s^{-1}\mathbf{D}_s\right]^T\right)\overline{\mathbf{w}}_s$$

$$= -2\mathbf{D}_s^T\mathbf{C}_s^{-1}\left(\mathbf{x}_t - \mathbf{D}_s\overline{\mathbf{w}}_s\right)$$

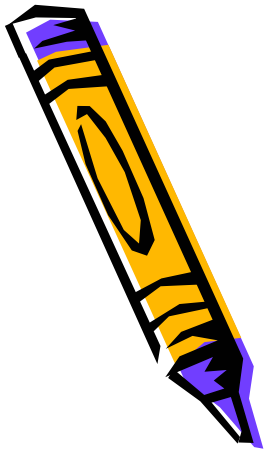
$$\therefore \frac{\partial}{\partial \overline{\mathbf{w}}_s}Q(\overline{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda}) = \sum_{t=1}^{T}\gamma_s(t)\mathbf{D}_s^T\mathbf{C}_s^{-1}\left(\mathbf{x}_t - \mathbf{D}_s\overline{\mathbf{w}}_s\right) = 0$$

$$\Rightarrow \mathbf{D}_s^T\mathbf{C}_s^{-1}\left[\sum_{t=1}^{T}\gamma_s(t)\mathbf{x}_t\right] = \left[\sum_{t=1}^{T}\gamma_s(t)\right]\mathbf{D}_s^T\mathbf{C}_s^{-1}\mathbf{D}_s\overline{\mathbf{w}}_s$$

$$\overline{\mathbf{w}}_s = \left[\left[\sum_{t=1}^{T}\gamma_s(t)\right]\mathbf{D}_s^T\mathbf{C}_s^{-1}\mathbf{D}_s\right]^{-1}\left[\mathbf{D}_s^T\mathbf{C}_s^{-1}\left[\sum_{t=1}^{T}\gamma_s(t)\mathbf{x}_t\right]\right]$$

MLLR

Single Variable Linear Regression



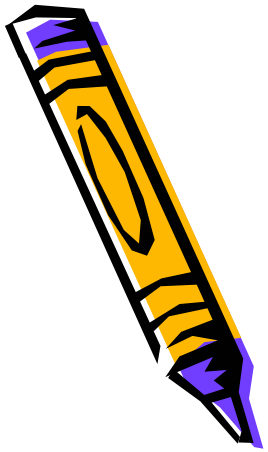
The extension to the tied regression matrix case :

$$\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)\mathbf{D}_{s_r}^{T}\mathbf{C}_{s_r}^{-1}\mathbf{x}_t=\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)\mathbf{D}_{s_r}^{T}\mathbf{C}_{s_r}^{-1}\mathbf{D}_{s_r}\overline{\mathbf{w}}_s$$

$$\Rightarrow \overline{\mathbf{w}}_s=\left[\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)\mathbf{D}_{s_r}^{T}\mathbf{C}_{s_r}^{-1}\mathbf{D}_{s_r}\right]^{-1}\left[\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)\mathbf{D}_{s_r}^{T}\mathbf{C}_{s_r}^{-1}\mathbf{x}_t\right]$$

MLLR

Defining Regression Classes



- Two approaches were considered:
  - 1. based on broad phonetic classes.
    - Models which represent the same broad phonetic class were placed in the same regression class.
  - 2. based on clustering of mixture components.
    - The mixture components were compared using a likelihood measure and similar components placed in the same regression class.
    - The data driven approach was found to be more appropriate for defining large numbers of classes.

MLLR

Variance Adapted

Reference :

- Variance Compensation Within the MLLR Framework for Robust Speech Recognition and Speaker Adaptation –ICSLP'96 Gales
- Mean and variance adaptation within the MLLR framework – CSL'96 Gales and Woodland
- MLLR:A Speaker Adaptation Technique for LVCSR – Hamaker

MLLR

Variance Adapted

Single Gaussian Case

- We apply Cholesky Decomposition to the inverse of covariance matrix:

$$\mathbf{C}_s^{-1} = \mathbf{L}_s \mathbf{L}_s^T \quad \text{where } \mathbf{L}_s \text{ is a lower triangular matrix}$$

$$\therefore \mathbf{C}_s = \mathbf{L}_s^{-T} \mathbf{L}_s^{-1}$$

- We can observe that $\left[\mathbf{C}_s^{-1}\right]_{ij} = \sum_{d=1}^{D} \left[\mathbf{L}_s\right]_{id} \left[\mathbf{L}_s\right]_{jd}$
- Now the inverse of covariance matrix is updated by

$$\overline{\mathbf{C}}_s^{-1} = \mathbf{L}_s \mathbf{H}_s^{-1} \mathbf{L}_s^T \quad \text{where } \mathbf{H}_s \text{ is the linear transformation}$$

- So

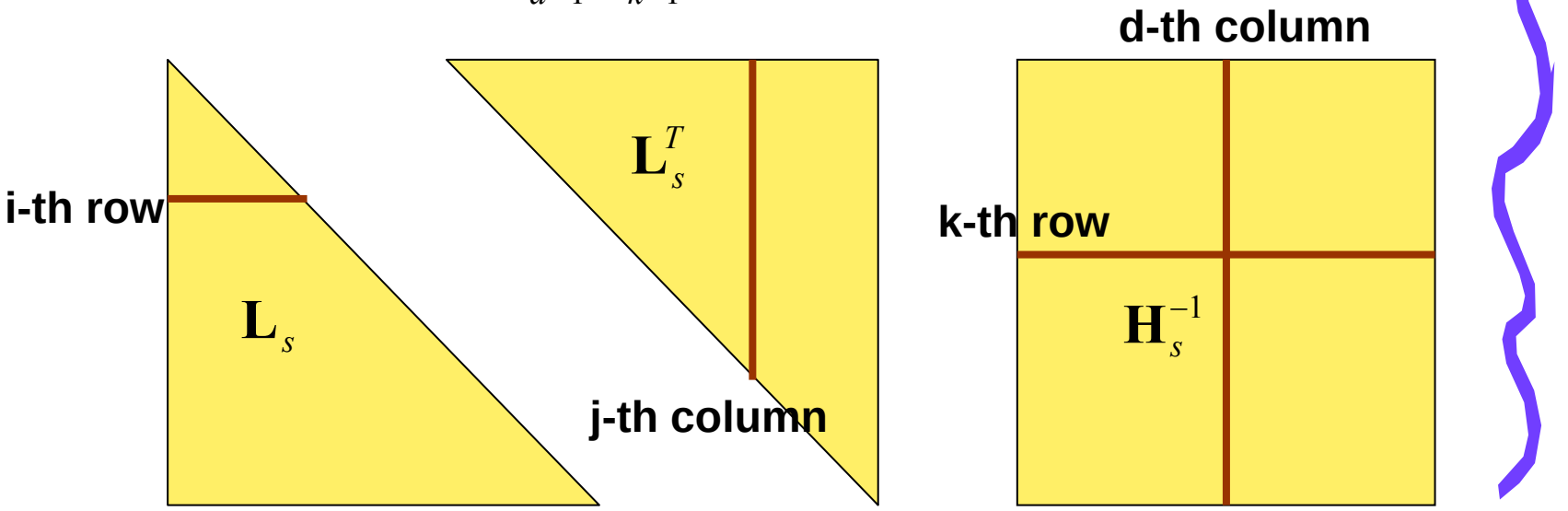
$$\overline{\mathbf{C}}_s = \mathbf{L}_s^{-T} \mathbf{H}_s \mathbf{L}_s^{-1}$$

MLLR

Variance Adapted

Single Gaussian Case

- What does the transformation mean ?
  - Origin : $\left[\mathbf{C}_s^{-1}\right]_{ij} = \sum_{d=1}^{D} \left[\mathbf{L}_s\right]_{id} \left[\mathbf{L}_s\right]_{jd}$
  - New : $\left[\overline{\mathbf{C}}_s^{-1}\right]_{ij} = \sum_{d=1}^{D} \sum_{k=1}^{D} \left[\mathbf{L}_s\right]_{ik} \left[\mathbf{L}_s\right]_{jd} \left[\mathbf{H}_s^{-1}\right]_{kd}$



MLLR

Variance Adapted

Single Gaussian Case

- The auxiliary can be obtained

transition probability

$$Q(\bar{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda})$$

$$= \boxed{constant} - \frac{1}{2} \times \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \left[D \log(2\pi) + \log \mid \overline{\mathbf{C}}_j \mid + (\mathbf{x}_t - \boldsymbol{\mu}_j)^T \overline{\mathbf{C}}_j^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_j) \right]$$

$$= -\frac{1}{2} \times \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \left[D \log(2\pi) + \log \mid \mathbf{L}_j^{-T} \overline{\mathbf{H}}_j \mathbf{L}_j^{-1} \mid + (\mathbf{x}_t - \boldsymbol{\mu}_j)^T \mathbf{L}_j \overline{\mathbf{H}}_j^{-1} \mathbf{L}_j^T (\mathbf{x}_t - \boldsymbol{\mu}_j) \right]$$

$$= -\frac{1}{2} \times \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \left[D \log(2\pi) + \log \left[\mid \mathbf{L}_j^{-T} \mid \cdot \mid \overline{\mathbf{H}}_j \mid \cdot \mid \mathbf{L}_j^{-1} \mid \right] + (\mathbf{L}_j^T \mathbf{x}_t - \mathbf{L}_j^T \boldsymbol{\mu}_j)^T \overline{\mathbf{H}}_j^{-1} (\mathbf{L}_j^T \mathbf{x}_t - \mathbf{L}_j^T \boldsymbol{\mu}_j) \right]$$

$$\because \mid \mathbf{L}_j^{-T} \mid \cdot \mid \overline{\mathbf{H}}_j \mid \cdot \mid \mathbf{L}_j^{-1} \mid = \mid \mathbf{L}_j^{-T} \mid \cdot \mid \mathbf{L}_j^{-1} \mid \cdot \mid \overline{\mathbf{H}}_j \mid = \mid \mathbf{L}_j^{-T} \mathbf{L}_j^{-1} \mid \cdot \mid \overline{\mathbf{H}}_j \mid = \mid \mathbf{C}_j \mid \cdot \mid \overline{\mathbf{H}}_j \mid$$

$$= -\frac{1}{2} \times \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \left[D \log(2\pi) + \log \mid \mathbf{C}_j \mid + \log \mid \overline{\mathbf{H}}_j \mid + (\mathbf{L}_j^T \mathbf{x}_t - \mathbf{L}_j^T \boldsymbol{\mu}_j)^T \overline{\mathbf{H}}_j^{-1} (\mathbf{L}_j^T \mathbf{x}_t - \mathbf{L}_j^T \boldsymbol{\mu}_j) \right]$$

MLLR Variance Adapted

Single Gaussian Case

- Differentiate Q-function w.r.t $\mathbf{H}_j$ and set it to zero then...

$$\frac{\partial}{\partial \mathbf{H}_j}\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_j(t)\left[\log|\overline{\mathbf{H}}_j|+(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)^T\overline{\mathbf{H}}_j^{-1}(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)\right]=0$$

$$\sum_{t=1}^{T}\gamma_j(t)\left[\frac{1}{|\overline{\mathbf{H}}_j|}\times|\overline{\mathbf{H}}_j|\times\overline{\mathbf{H}}_j^{-T}-\overline{\mathbf{H}}_j^{T}(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)^T\mathbf{H}_j^T\right]=0$$

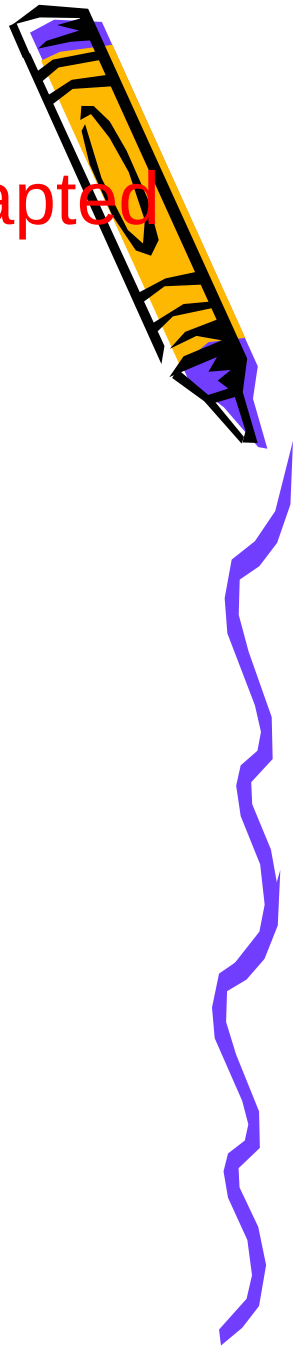
$$\sum_{t=1}^{T}\gamma_j(t)\times\overline{\mathbf{H}}_j^{-T}=\sum_{t=1}^{T}\gamma_j(t)\overline{\mathbf{H}}_j^{-T}(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)^T\mathbf{H}_j^{-T}$$

$$\sum_{t=1}^{T}\gamma_j(t)\times\overline{\mathbf{H}}_j^{T}=\sum_{t=1}^{T}\gamma_j(t)(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)^T$$

$$\overline{\mathbf{H}}_j^{T}=\frac{\sum_{t=1}^{T}\gamma_j(t)(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)(\mathbf{L}_j^T\mathbf{x}_t-\mathbf{L}_j^T\boldsymbol{\mu}_j)^T}{\sum_{t=1}^{T}\gamma_j(t)}$$

MLLR

Variance Adapted



Single Gaussian Case

$$\therefore \overline{\mathbf{H}}_j^T = \frac{\sum_{t=1}^{T} \gamma_j(t)(\mathbf{L}_j^T\mathbf{x}_t - \mathbf{L}_j^T\boldsymbol{\mu}_j)(\mathbf{L}_j^T\mathbf{x}_t - \mathbf{L}_j^T\boldsymbol{\mu}_j)^T}{\sum_{t=1}^{T} \gamma_j(t)}$$

$$= \frac{\mathbf{L}_j^T\left[\sum_{t=1}^{T} \gamma_j(t)(\mathbf{x}_t - \boldsymbol{\mu}_j)(\mathbf{x}_t - \boldsymbol{\mu}_j)^T\right]\mathbf{L}_j}{\sum_{t=1}^{T} \gamma_j(t)}$$

We can observe that $\overline{\mathbf{H}}_j^T$ is symmetric.

$$\therefore \overline{\mathbf{H}}_j = \overline{\mathbf{H}}_j^T$$

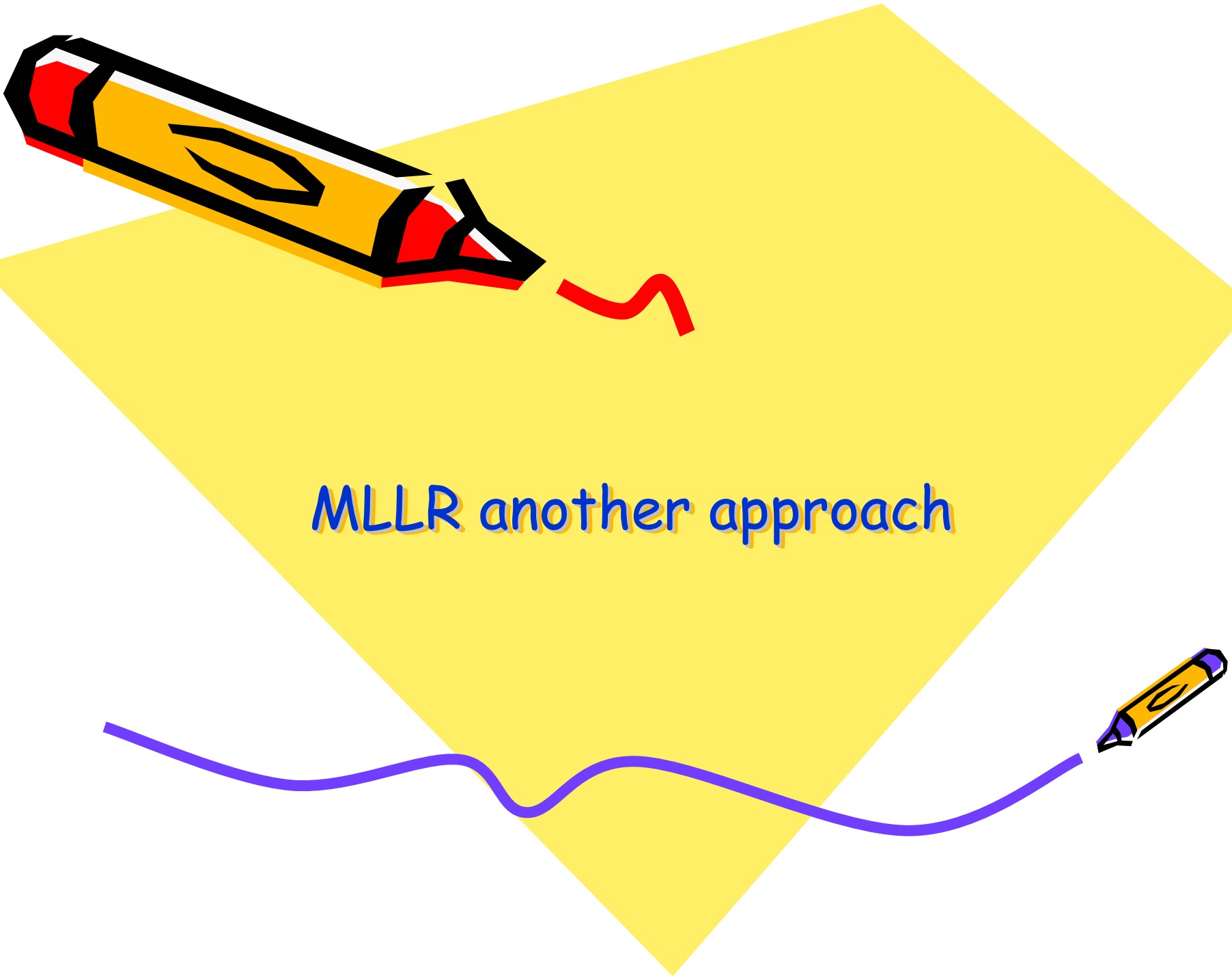
MLLR

Variance Adapted

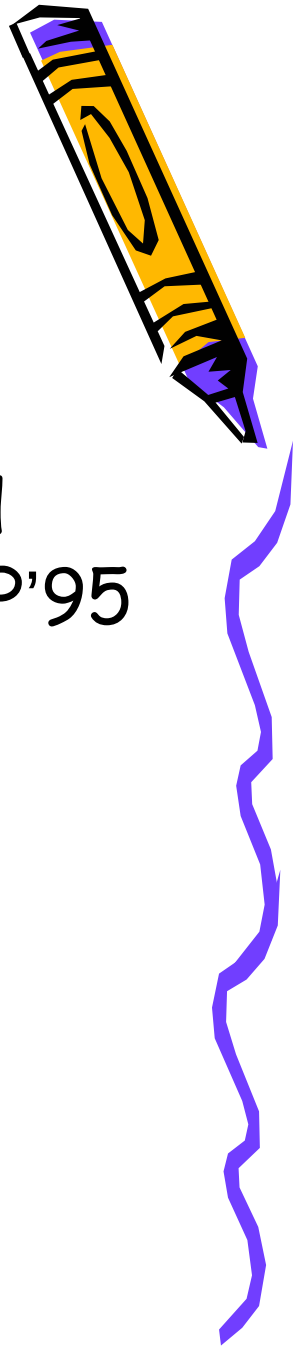
Tied Regression Matrices Case

If $\mathbf{H}_s$ is shared by R states $\{s_1, \cdots, s_R\}$ then

$$\overline{\mathbf{H}}_s = \frac{\sum_{r=1}^{R} \left\{ \mathbf{L}_{s_r}^T \left[\sum_{t=1}^{T} \gamma_{s_r}(t)(\mathbf{x}_t - \boldsymbol{\mu}_{s_r})(\mathbf{x}_t - \boldsymbol{\mu}_{s_r})^T \right] \mathbf{L}_{s_r} \right\}}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{s_r}(t)}$$

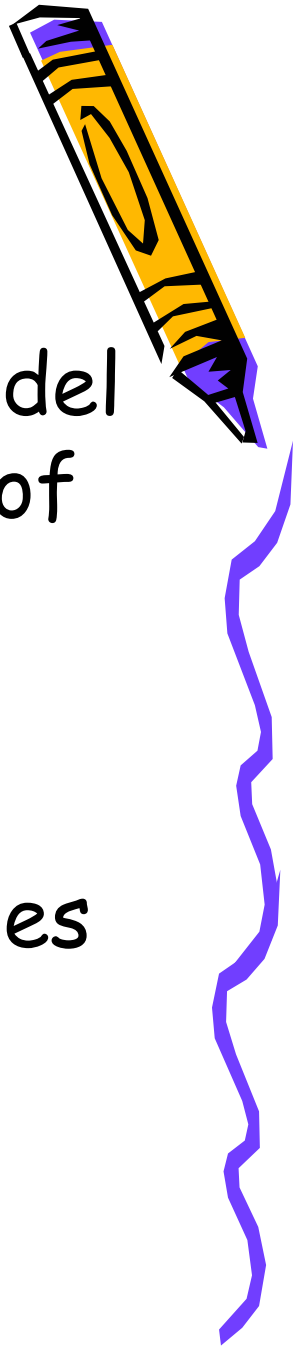

MLLR another approach

MLLR another approach



- Reference:
  - Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures – SAP'95 Vassilios V. Digalakis

Introduction



- This approach is an extension of model space MLLR where the covariances of the Gaussian components are constrained to share the same transforms as the means.
- The transformed means and variances and are given as a function of the transform parameters:

$$\overline{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\overline{\boldsymbol{\Sigma}} = \mathbf{A}\boldsymbol{\Sigma}A^T$$

Single Gaussian Case

- Assume the adaptation data, X, is a series of T observations.

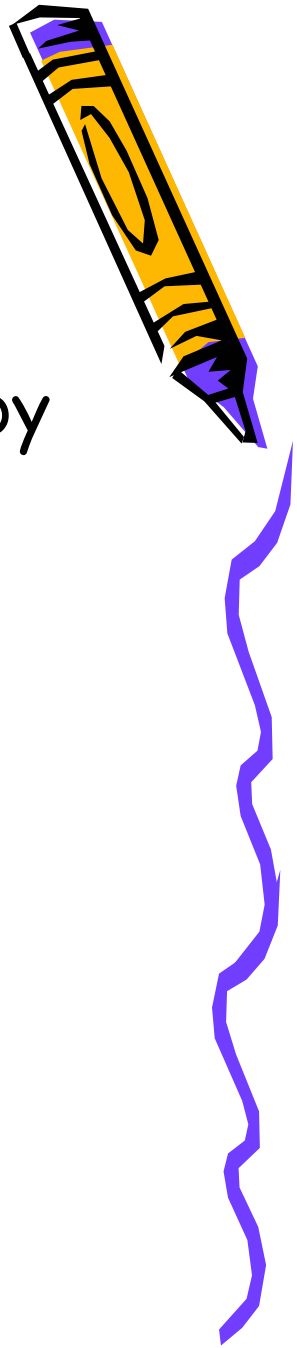
$$\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$$

- For each state *s*
- Denote the initial model by $\boldsymbol{\lambda}_s^{(0)} = (\boldsymbol{\mu}_s^{(0)}, \boldsymbol{\Sigma}_s^{(0)})$

- Current set of model parameters by applying the transformation $\mathbf{A}_s$

$$\boldsymbol{\lambda}_s = (\mathbf{A}_s \boldsymbol{\mu}_s^{(0)} + \mathbf{b}_s, \mathbf{A}_s \boldsymbol{\Sigma}_s^{(0)} \mathbf{A}_s^T)$$

Single Gaussian Case



- Re-estimated set of model parameters by applying the transformation $\overline{\mathbf{A}}_s$

$$\overline{\boldsymbol{\lambda}}_s = (\overline{\mathbf{A}}_s \boldsymbol{\mu}_s^{(0)} + \overline{\mathbf{b}}_s, \overline{\mathbf{A}}_s \boldsymbol{\Sigma}_s^{(0)} \overline{\mathbf{A}}_s^T)$$

- We denote the parameter set

$$\boldsymbol{\Lambda} = \{\mu_1^{(0)}, \mu_2^{(0)}, \cdots, \mu_{N_s}^{(0)}, \Sigma_1^{(0)}, \Sigma_2^{(0)}, \cdots, \Sigma_{N_s}^{(0)}\}$$

$$\boldsymbol{\eta} = \{\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_{N_s}, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{N_s}\}$$

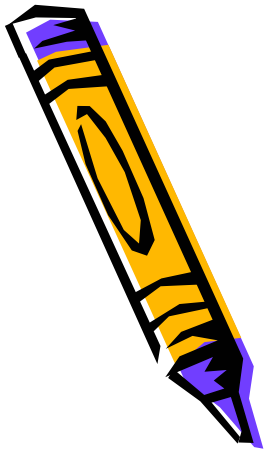
N_s is the total state number

Single Gaussian Case

$$\overline{\boldsymbol{\eta}}_{MAP} = \arg\max_{\boldsymbol{\eta}} p(\boldsymbol{\eta} \mid \boldsymbol{\Lambda}, \mathbf{X}) = \arg\max_{\boldsymbol{\eta}} p(\mathbf{X} \mid \boldsymbol{\eta}, \boldsymbol{\Lambda}) g(\boldsymbol{\eta})$$

where $p(\mathbf{X} \mid \boldsymbol{\eta}, \boldsymbol{\Lambda}) = \sum_{\mathbf{S}} p(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\eta}, \boldsymbol{\Lambda})$

and $p(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\eta}, \boldsymbol{\Lambda})$ is the likelihood of generating $\mathbf{X}$ using the state sequence $\mathbf{S}$ given model $\boldsymbol{\Lambda}$ and transformation $\boldsymbol{\eta}$



Single Gaussian Case

- We define the auxiliary function

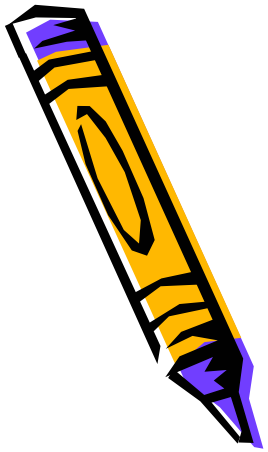
$$Q(\bar{\boldsymbol{\eta}} \mid \boldsymbol{\eta}) = \sum_{\mathbf{S}} p(\mathbf{S} \mid \mathbf{X}, \boldsymbol{\eta}, \boldsymbol{\Lambda}) \log\left[p(\mathbf{X}, \mathbf{S} \mid \bar{\boldsymbol{\eta}}, \boldsymbol{\Lambda})\right]$$

- Since only the transformations $\mathbf{A}_s, \mathbf{b}_s$ are re-estimated, only the output distributions $b_s(\mathbf{x}_t)$ are affected so the auxiliary function can be written as

$$Q(\bar{\boldsymbol{\Lambda}} \mid \boldsymbol{\Lambda}) = constant + \left[\sum_{\mathbf{S}} \sum_{t=1}^{T} \gamma_j(t) \log \bar{b}_j(\mathbf{x}_t)\right]$$

where $\gamma_j(t) = p(s_t = j \mid \mathbf{X}, \boldsymbol{\Lambda})$

Single Gaussian Case



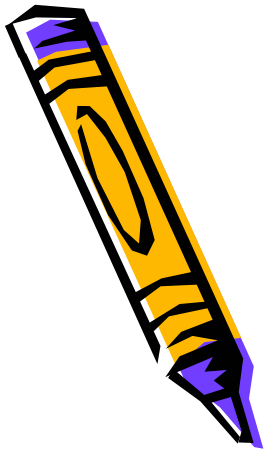
- Where

$$\bar{b}_j(\mathbf{x}_t) = (2\pi)^{-\frac{D}{2}} \left|\bar{\boldsymbol{\Sigma}}_j\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_t-\bar{\boldsymbol{\mu}}_j)^T \bar{\boldsymbol{\Sigma}}_j (\mathbf{x}_t-\bar{\boldsymbol{\mu}}_j)}$$

$$\log \bar{b}_j(\mathbf{x}_t) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left|\bar{\boldsymbol{\Sigma}}_j\right| - \frac{1}{2}(\mathbf{x}_t - \bar{\boldsymbol{\mu}}_j)^T \bar{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{x}_t - \bar{\boldsymbol{\mu}}_j)$$

$$= -\frac{1}{2}\left[D\log(2\pi) + \log\left|\bar{\boldsymbol{\Sigma}}_j\right| + (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_j)^T \bar{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{x}_t - \bar{\boldsymbol{\mu}}_j)\right]$$

Single Gaussian Case



- Expanding $\log b_j(\mathbf{x}_t)$ then the auxiliary function is

$$Q(\overline{\boldsymbol{\eta}} \mid \boldsymbol{\eta}) = constant - \frac{1}{2} \times \sum_{j=1}^{N_s} \sum_{t=1}^{T} \gamma_j(t) \left[D \log(2\pi) + \log \left| \overline{\mathbf{A}}_j \boldsymbol{\Sigma}_j^{(0)} \overline{\mathbf{A}}_j^T \right| + h(\mathbf{x}_t, j) \right]$$

$$\text{where } h(\mathbf{x}_t, j) = (\mathbf{x}_t - \overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)} - \overline{\mathbf{b}}_j)^T \left(\mathbf{A}_j \boldsymbol{\Sigma}_j^{(0)} \mathbf{A}_j^T \right)^{-1} (\mathbf{x}_t - \overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)} - \overline{\mathbf{b}}_j)$$

Assume that $\left[\boldsymbol{\Sigma}_j^{(0)}\right]_{pq}$ is the pth row and qth column element of matrix $\boldsymbol{\Sigma}_j^{(0)}$ and $\left[\mathbf{A}_j^{(0)}\right]_{pq}$ is the pth row and qth column element of matrix $\mathbf{A}_j^{(0)}$

Single Gaussian Case

$$\nabla_{\overline{\mathbf{A}}_j} Q(\overline{\boldsymbol{\eta}} \mid \boldsymbol{\eta}) = 0$$

$$\Rightarrow \nabla_{\overline{\mathbf{A}}_j} \sum_{t=1}^{T} \gamma_j(t) \left[\log \left| \overline{\mathbf{A}}_j \boldsymbol{\Sigma}_j^{(0)} \overline{\mathbf{A}}_j^T \right| + (\mathbf{x}_t - \overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)} - \mathbf{b}_j)^T \left(\overline{\mathbf{A}}_j \boldsymbol{\Sigma}_j^{(0)} \overline{\mathbf{A}}_j^T \right)^{-1} (\mathbf{x}_t - \overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)} - \mathbf{b}_j) \right] = 0$$

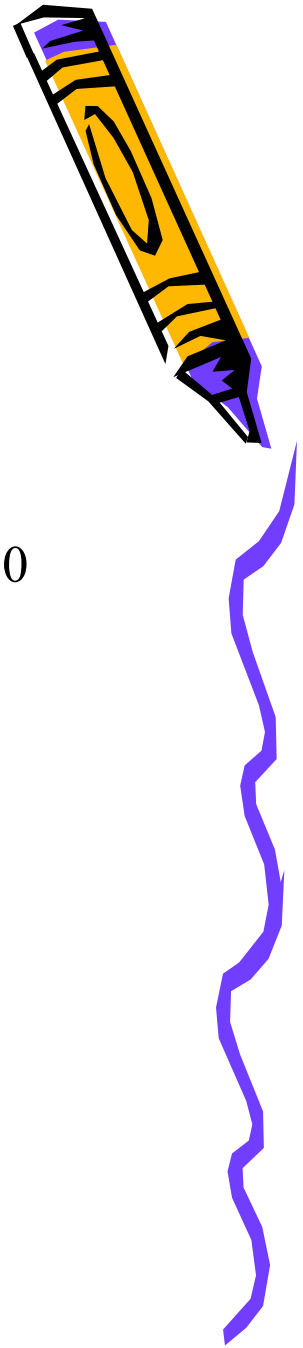
$$\Rightarrow \nabla_{\overline{\mathbf{A}}_j} \sum_{t=1}^{T} \gamma_j(t) \begin{bmatrix} \log \left| \overline{\mathbf{A}}_j \boldsymbol{\Sigma}_j^{(0)} \overline{\mathbf{A}}_j^T \right| \\ + (\mathbf{x}_t - \mathbf{b}_j)^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)} \right]^{-1} \overline{\mathbf{A}}_j^{-1} (\mathbf{x}_t - \mathbf{b}_j) \\ - (\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)})^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)} \right]^{-1} \overline{\mathbf{A}}_j^{-1} (\mathbf{x}_t - \mathbf{b}_j) \\ - (\mathbf{x}_t - \mathbf{b}_j)^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)} \right]^{-1} \overline{\mathbf{A}}_j^{-1} (\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)}) \\ + (\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)})^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)} \right]^{-1} \overline{\mathbf{A}}_j^{-1} (\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)}) \end{bmatrix} = 0$$

Single Gaussian Case

$$\Rightarrow \nabla_{\overline{\mathbf{A}}_j} \sum_{t=1}^{T} \gamma_j(t) \begin{bmatrix} \log|\overline{\mathbf{A}}_j \boldsymbol{\Sigma}_j^{(0)} \overline{\mathbf{A}}_j^T| \\ +(\mathbf{x}_t - \mathbf{b}_j)^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \overline{\mathbf{A}}_j^{-1} (\mathbf{x}_t - \mathbf{b}_j) \\ -(\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)})^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \overline{\mathbf{A}}_j^{-1} (\mathbf{x}_t - \mathbf{b}_j) \\ -(\mathbf{x}_t - \mathbf{b}_j)^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \overline{\mathbf{A}}_j^{-1} (\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)}) \\ +(\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)})^T \overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \overline{\mathbf{A}}_j^{-1} (\overline{\mathbf{A}}_j \boldsymbol{\mu}_j^{(0)}) \end{bmatrix} = 0$$

$$\Rightarrow \sum_{t=1}^{T} \gamma_j(t) \begin{bmatrix} 2\overline{\mathbf{A}}_j^{-T} \\ -\left[\overline{\mathbf{A}}_j^{-T} \left(\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} + \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-T}\right) \overline{\mathbf{A}}_j^{-1} (\mathbf{x}_t - \mathbf{b}_j)(\mathbf{x}_t - \mathbf{b}_j)^T \overline{\mathbf{A}}_j^{-T}\right] \\ +\left[\overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-T} \boldsymbol{\mu}_j^{(0)} (\mathbf{x}_t - \mathbf{b}_j)^T \overline{\mathbf{A}}_j^{-T}\right] \\ +\left[\overline{\mathbf{A}}_j^{-T} \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \boldsymbol{\mu}_j^{(0)} (\mathbf{x}_t - \mathbf{b}_j)^T \overline{\mathbf{A}}_j^{-T}\right] \end{bmatrix} = 0$$

Single Gaussian Case

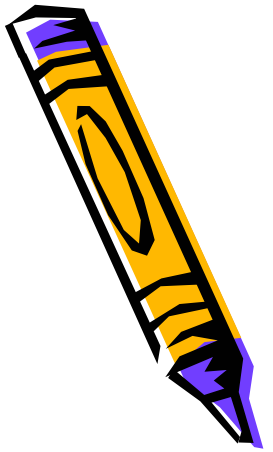


$$\sum_{t=1}^{T}\gamma_j(t)\begin{bmatrix} 2\overline{\mathbf{A}}_j^{-T} \\ -\left[\overline{\mathbf{A}}_j^{-T}\left(\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}+\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-T}\right)\overline{\mathbf{A}}_j^{-1}(\mathbf{x}_t-\mathbf{b}_j)(\mathbf{x}_t-\mathbf{b}_j)^T\overline{\mathbf{A}}_j^{-T}\right] \\ +\left[\overline{\mathbf{A}}_j^{-T}\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-T}\boldsymbol{\mu}_j^{(0)}(\mathbf{x}_t-\mathbf{b}_j)^T\overline{\mathbf{A}}_j^{-T}\right] \\ +\left[\overline{\mathbf{A}}_j^{-T}\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}\boldsymbol{\mu}_j^{(0)}(\mathbf{x}_t-\mathbf{b}_j)^T\overline{\mathbf{A}}_j^{-T}\right] \end{bmatrix}=0$$

multiply $\overline{\mathbf{A}}_j^{T}$ in the right and in the left side

$$\sum_{t=1}^{T}\gamma_j(t)\begin{bmatrix} 2\overline{\mathbf{A}}_j^{T} \\ -\left(\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}+\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-T}\right)\overline{\mathbf{A}}_j^{-1}(\mathbf{x}_t-\mathbf{b}_j)(\mathbf{x}_t-\mathbf{b}_j)^T \\ +\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-T}\boldsymbol{\mu}_j^{(0)}(\mathbf{x}_t-\mathbf{b}_j)^T \\ +\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}\boldsymbol{\mu}_j^{(0)}(\mathbf{x}_t-\mathbf{b}_j)^T \end{bmatrix}=0$$

Single Gaussian Case



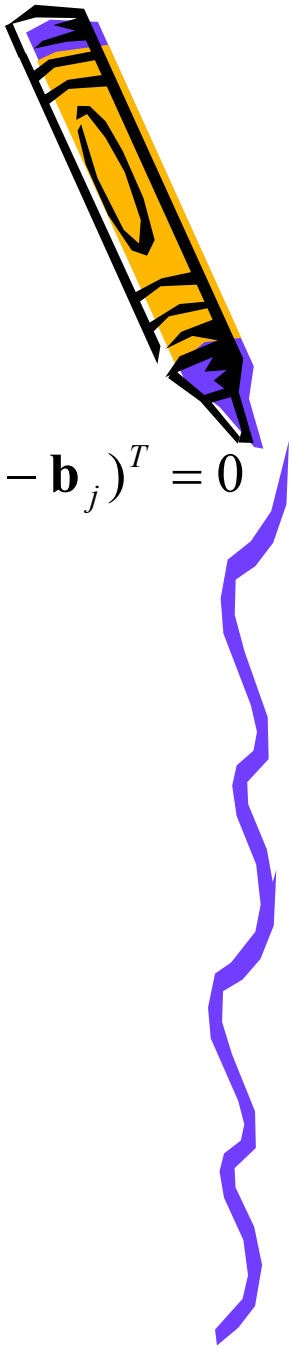
$$\because \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \text{ is symmetric}$$

$$\sum_{t=1}^{T} \gamma_j(t)\left[2\overline{\mathbf{A}}_j^T - 2\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}\overline{\mathbf{A}}_j^{-1}(\mathbf{x}_t - \mathbf{b}_j)(\mathbf{x}_t - \mathbf{b}_j)^T + 2\left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}\boldsymbol{\mu}_j^{(0)}(\mathbf{x}_t - \mathbf{b}_j)^T\right] = 0$$

$$\sum_{t=1}^{T} \gamma_j(t)\left[\overline{\mathbf{A}}_j^T - \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}\overline{\mathbf{A}}_j^{-1}(\mathbf{x}_t - \mathbf{b}_j)(\mathbf{x}_t - \mathbf{b}_j)^T + \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1}\boldsymbol{\mu}_j^{(0)}(\mathbf{x}_t - \mathbf{b}_j)^T\right] = 0$$

- Since covariance is diagonal, means that there is no correlation between dimensions. Hence, the transformation is assumed to be diagonal.

Single Gaussian Case (ML)



$$\overline{\mathbf{A}}_j^T \overline{\mathbf{A}}_j^T - \sum_{t=1}^{T} \gamma_j(t) \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} (\mathbf{x}_t - \mathbf{b}_j)(\mathbf{x}_t - \mathbf{b}_j)^T + \overline{\mathbf{A}}_j^T \sum_{t=1}^{T} \gamma_j(t) \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \boldsymbol{\mu}_j^{(0)} (\mathbf{x}_t - \mathbf{b}_j)^T = 0$$

Let $\sum_{t=1}^{T} \gamma_j(t) \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} (\mathbf{x}_t - \mathbf{b}_j)(\mathbf{x}_t - \mathbf{b}_j)^T = \mathbf{D}$

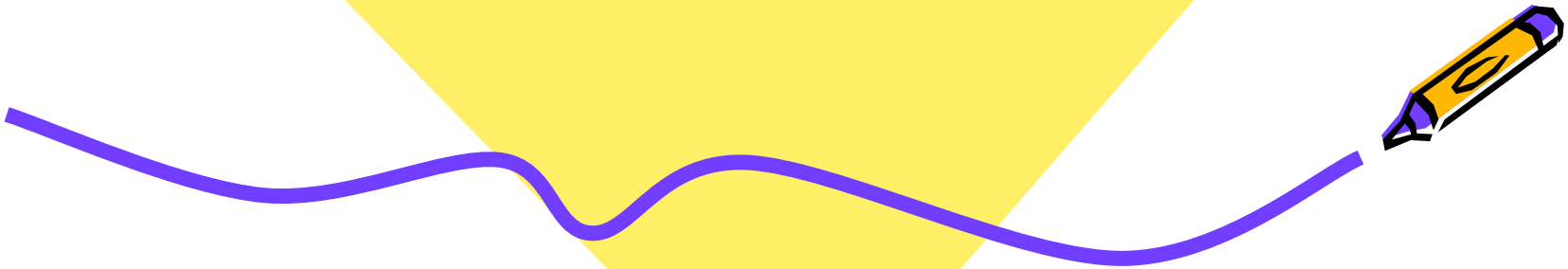
$$\sum_{t=1}^{T} \gamma_j(t) \left[\boldsymbol{\Sigma}_j^{(0)}\right]^{-1} \boldsymbol{\mu}_j^{(0)} (\mathbf{x}_t - \mathbf{b}_j)^T = \mathbf{Z}$$

$$\left[\overline{\mathbf{A}}_j^T\right]^2 + \overline{\mathbf{A}}_j^T \mathbf{Z} - \mathbf{D} = 0$$

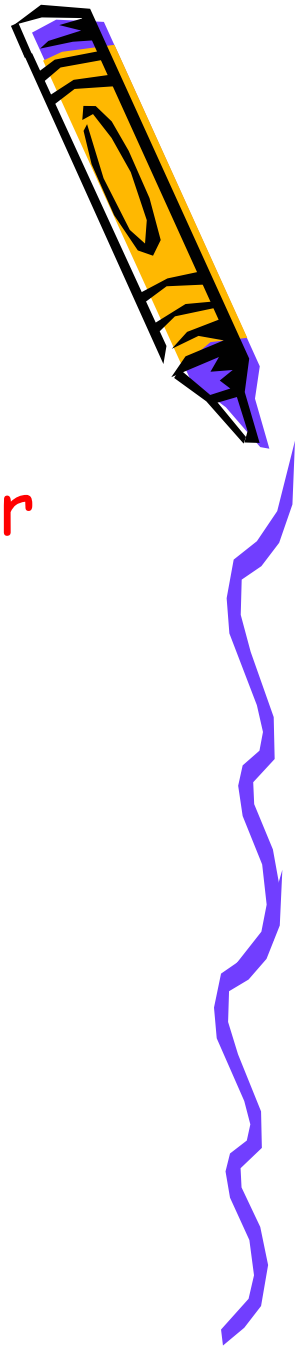
$$\overline{\mathbf{A}}_j = diag\left[a_1^{(j)}, \cdots, a_D^{(j)}\right]$$

$$\therefore \left(a_p^{(j)}\right)^2 + z_p^{(j)} a_p^{(j)} - d_{pp}^{(j)} = 0$$

Constrained Maximum Likelihood Linear Regression

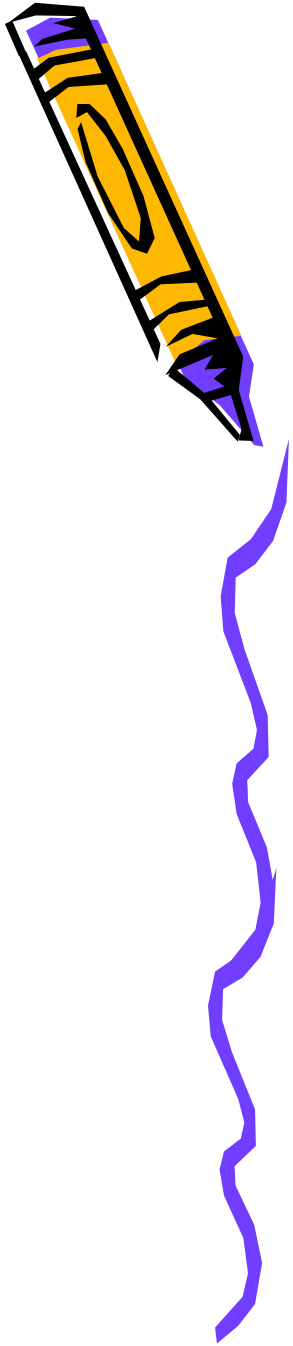


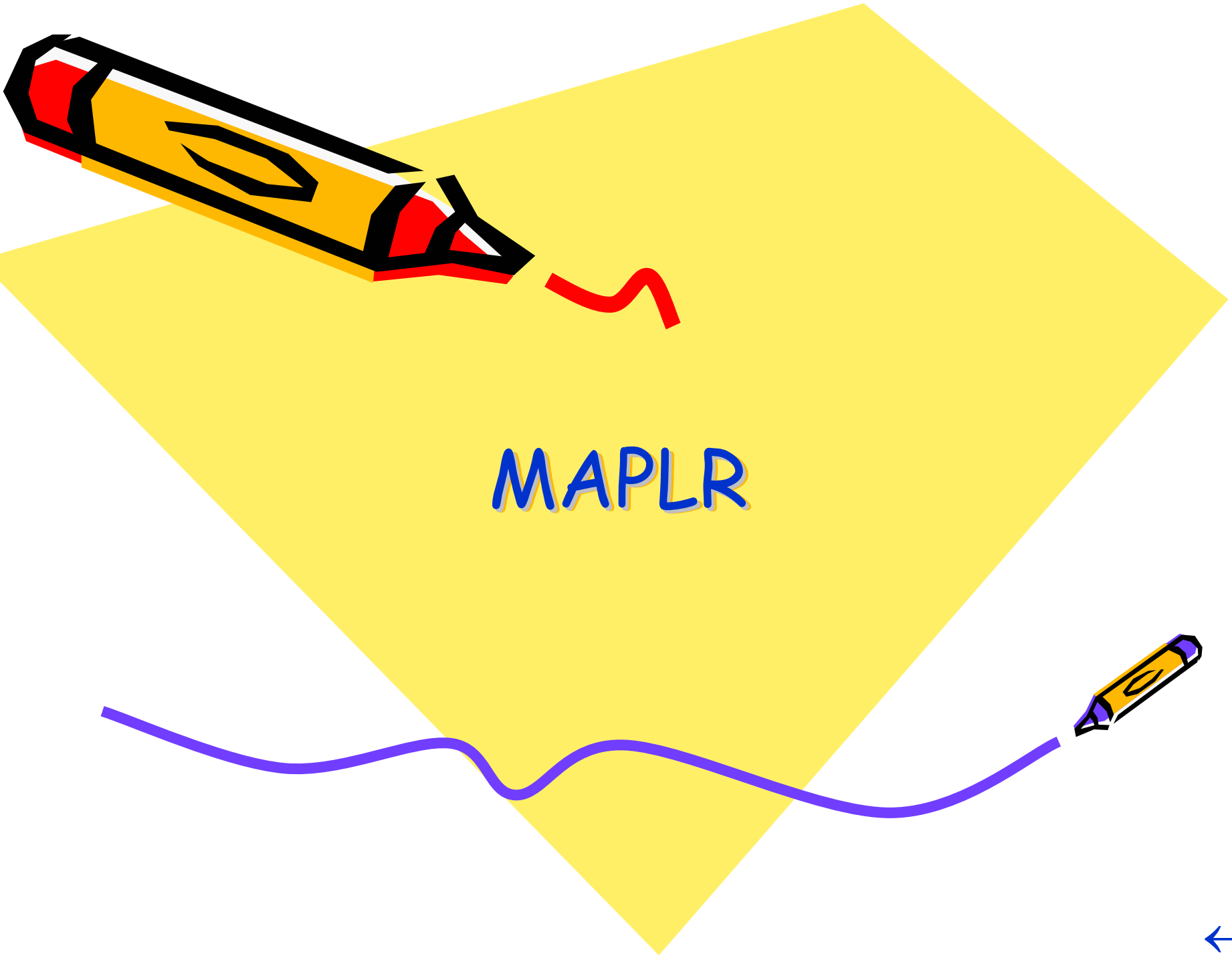
CMLLR



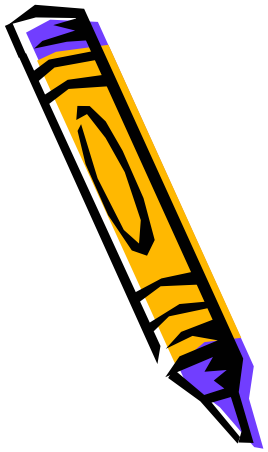
- Reference:
  - Constrained Maximum Likelihood Linear Regression for Speaker Adaptation – ICSLP'00 Mohamed Afify and Olivier Siohan

CMLLR



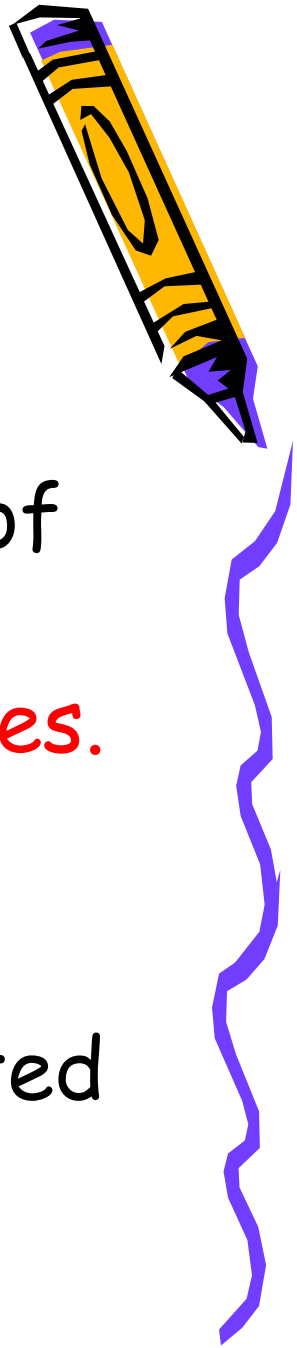

MAPLR

MAPLR



- Reference:
  - Hidden Markov Model Adaptation Using Maximum a Posteriori Linear Regression – In Workshop on Robust Methods for Speech Recognition in Adverse Conditions '99
  - Maximum a Posteriori Linear Regression for Hidden Markov Model Adaptation – EuroSpeech'99 C. Chesta
  - Maximum a Posterior Linear Regression with Elliptically Symmetric Matrix Variate Priors – EuroSpeech'99 W. Chou

MAPLR



- It is necessary to introduce some constraints on the possible values of the transformation parameters to avoid getting unreasonable estimates.

- A bayesian counterpart of the well known MLLR adaptation is formulated based on MAP estimation.

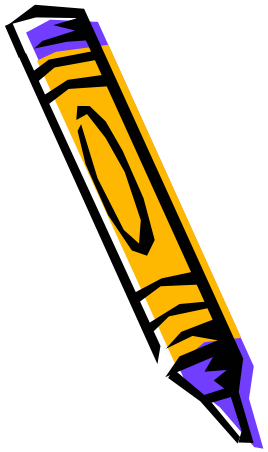
MAPLR

- Let Λ be a set of SI hidden Markov models.
- Some transformations $F_{\boldsymbol{\eta}}(\cdot)$ applied to various clusters of HMM parameters.
- Denote $\mathbf{X}$ be some adaptation data.
- Then

$$\hat{\boldsymbol{\eta}}_{MAP} = \arg\max_{\boldsymbol{\eta}} p(\boldsymbol{\eta} \mid \mathbf{X}, \boldsymbol{\Lambda}) \propto \arg\max_{\boldsymbol{\eta}} p(\mathbf{X} \mid \boldsymbol{\eta}, \boldsymbol{\Lambda}) p(\boldsymbol{\eta})$$

- Rather than carrying out the estimation using ML, we derive an estimate of $\boldsymbol{\eta} = (\mathbf{A}, \mathbf{b})$ using MAP.

MAPLR



- In a given state j , the pdf of an observation vector $\mathbf{x}$ is modeled by a mixture of K Gaussian distributions :

$$p(\mathbf{x} \mid s = j) = \sum_{k=1}^{K} w_{j,k} N\left(\mathbf{x} \mid \boldsymbol{\mu}_{j,k}, \mathbf{R}_{j,k}\right)$$

where $N\left(\mathbf{x} \mid \boldsymbol{\mu}_{j,k}, \mathbf{R}_{j,k}\right)$ is a Normal distribution of mean $\boldsymbol{\mu}_{j,k}$ and precision matrix $\mathbf{R}_{j,k}$

$$N\left(\mathbf{x} \mid \boldsymbol{\mu}_{j,k}, \mathbf{R}_{j,k}\right) \propto |\mathbf{R}_{j,k}|^{\frac{1}{2}} \exp\left[-\frac{1}{2} tr\left((\mathbf{x}-\boldsymbol{\mu}_{j,k})(\mathbf{x}-\boldsymbol{\mu}_{j,k})^{T} \mathbf{R}_{j,k}\right)\right]$$

MAPLR

- A mean vector $\boldsymbol{\mu}_{j,k} \in \mathbf{R}^D$ is adapted using an affine transformation $\boldsymbol{\eta} = \{\mathbf{A}, \mathbf{b}\}$

$$\overline{\boldsymbol{\mu}}_{j,k} = \mathbf{W}\boldsymbol{\xi}_{j,k}$$

- Where $\mathbf{W} = (\mathbf{A}, \mathbf{b})$ and $\boldsymbol{\xi}_{j,k} = (1, \boldsymbol{\mu}_{j,k})$
- So the Gaussian distribution ..

$$N\left(\mathbf{x}_t \mid \mathbf{A}_c, \mathbf{b}_c, \boldsymbol{\mu}_{s_t l_t}, \mathbf{R}_{s_t, l_t}\right) \propto$$

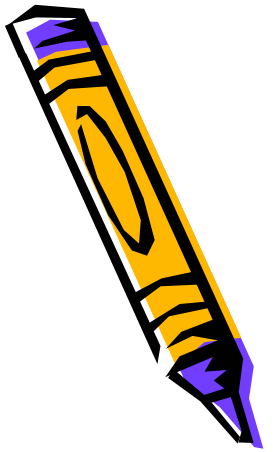
$$|\mathbf{R}_{s_t l_t}|^{\frac{1}{2}} \exp\left[-\frac{1}{2} tr\left((\mathbf{x}_t - \mathbf{A}_c\boldsymbol{\mu}_{s_t l_t} - \mathbf{b}_c)(\mathbf{x}_t - \mathbf{A}_c\boldsymbol{\mu}_{s_t l_t} - \mathbf{b}_c)^T \mathbf{R}_{s_t l_t}\right)\right]$$

- Clusters of mean vectors are also defined so that all mean vectors from the same cluster c share the same transformation $\mathbf{W}_c$

MAPLR

- How to select the prior density $p(\mathbf{W})$?
  - Unlike MAP estimation of HMM parameters, no obvious conjugate prior densities would be found in our case.
  - If the prior of matrix is chosen as the product of a Normal-Wishart density, no closed-form solution could be obtained for the square transformation matrix.

MAPLR



- Let $\mathbf{X} = \{\mathbf{x}_t\}$ be an adaptation utterance used to derive $\mathbf{W}$
- Define an auxiliary function

$$Q(\bar{\boldsymbol{\eta}}_c \mid \boldsymbol{\eta}_c) = E\left[\log p(\mathbf{X}, \mathbf{S}, \mathbf{L} \mid \boldsymbol{\Lambda}, \bar{\boldsymbol{\eta}}_c) + \log p(\bar{\boldsymbol{\eta}}_c) \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}_c\right]$$

$$= \left[\sum_S \sum_L p(\mathbf{S}, \mathbf{L} \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}_c) \log p(\mathbf{X}, \mathbf{S}, \mathbf{L} \mid \boldsymbol{\Lambda}, \bar{\boldsymbol{\eta}}_c)\right] + \log p(\bar{\boldsymbol{\eta}}_c)$$

Hidden data

Complete data

where $\mathbf{S} = \{s_t\}$ *state sequence*

$\mathbf{L} = \{l_t\}$ *mixture sequence*

MAPLR

$$Q(\overline{\boldsymbol{\eta}}_c \mid \boldsymbol{\eta}_c)$$

$$= \sum_{\mathbf{S}} \sum_{\mathbf{L}} p(\mathbf{S}, \mathbf{L} \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}_c) \sum_{t=1}^{T} \left[\log a_{s_{t-1}, s_t} + \log w_{s_t, l_t} + \log p(\mathbf{x}_t \mid \overline{\boldsymbol{\eta}}_c, \boldsymbol{\mu}_{s_t, l_t}, \mathbf{R}_{s_t, l_t}) \right]$$

$$+ \log p(\overline{\boldsymbol{\eta}}_c)$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \log p(\mathbf{x}_t \mid \overline{\boldsymbol{\eta}}_c, \boldsymbol{\mu}_{s_t, l_t}, \mathbf{R}_{s_t, l_t}) \right] + \log p(\overline{\boldsymbol{\eta}}_c) + \Psi$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left[-\frac{1}{2} tr\left((\mathbf{x}_t - \overline{\mathbf{A}}_c \boldsymbol{\mu}_{s_t, l_t} - \overline{\mathbf{b}}_c)(\mathbf{x}_t - \overline{\mathbf{A}}_c \boldsymbol{\mu}_{s_t, l_t} - \overline{\mathbf{b}}_c)^T \mathbf{R}_{s_t, l_t} \right) \right] \right]$$

$$+ \log p(\overline{\boldsymbol{\eta}}_c) + \Psi'$$

$$where\ \gamma_t(j,k) = P(s_t = j, l_t = k \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}_c)$$

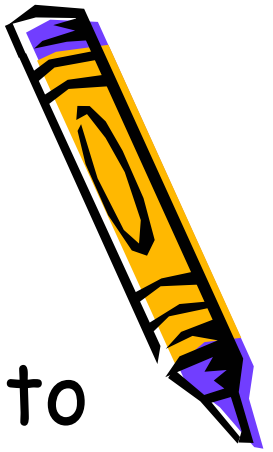
Ψ : all terms independent of $\overline{\boldsymbol{\eta}}_c$

MAPLR

- We assume that $p(\overline{\mathbf{W}}_c) = p(\overline{\mathbf{A}}_c, \overline{\mathbf{b}}_c) = p(\overline{\mathbf{A}}_c)p(\overline{\mathbf{b}}_c)$
- So we can maximize the transformation and bias individually.
- we will differentiate the auxiliary function w.r.t the bias firstly.
- Assuming the distribution of the bias is modeled by a Gaussian distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Xi}$

$$p(\overline{\mathbf{b}}_c) \propto |\boldsymbol{\Xi}_c|^{1/2} \exp\left[-\frac{1}{2} tr\left((\overline{\mathbf{b}}_c - \boldsymbol{\beta}_c)(\overline{\mathbf{b}}_c - \boldsymbol{\beta})^T \boldsymbol{\Xi}_c\right)\right]$$

MAPLR



- Differentiating w.r.t $\overline{\mathbf{b}}_c$ and equate to zero

$$\frac{\partial}{\partial \overline{\mathbf{b}}_c} Q(\overline{\mathbf{A}}_c, \overline{\mathbf{b}}_c \mid \mathbf{A}_c, \mathbf{b}_c) = 0$$

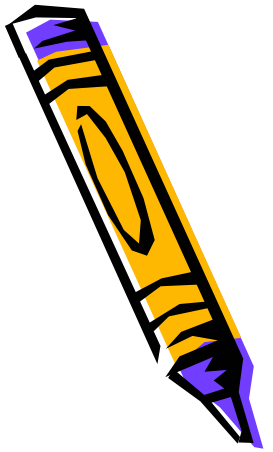
$$\therefore \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \times$$

$$\left[-\frac{1}{2} \times \left(-2 \times \mathbf{R}_{j,k} \times (\mathbf{x}_t - \overline{\mathbf{A}}_c \boldsymbol{\mu}_{j,k}) + 2 \times \mathbf{R}_{j,k} \times \overline{\mathbf{b}}_c \right) + \left(-\frac{1}{2} \right) \left(-2 \times \boldsymbol{\Xi}_c \times \overline{\mathbf{b}}_c + 2 \times \boldsymbol{\Xi}_c \times \boldsymbol{\beta}_c \right) \right] = 0$$

$$\Rightarrow \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) R_{j,k} (\mathbf{x}_t - \overline{\mathbf{A}}_c \boldsymbol{\mu}_{n,m} - \overline{\mathbf{b}}_c) \right] - \boldsymbol{\Xi}_c (\overline{\mathbf{b}}_c - \boldsymbol{\beta}_c) = 0$$

where $\gamma_t(j,k) = \Pr(s_t = j, l_t = k \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}_c)$

MAPLR



- A closed form solution can be easily obtained.

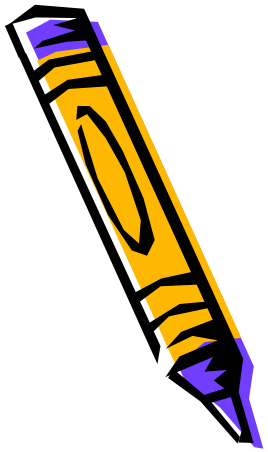
$$\left[\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(j,k)\mathbf{R}_{j,k}(\mathbf{x}_t-\overline{\mathbf{A}}_c\boldsymbol{\mu}_{j,k}-\overline{\mathbf{b}}_c)\right]-\boldsymbol{\Xi}_c(\overline{\mathbf{b}}_c-\boldsymbol{\beta}_c)=0$$

$$\left[\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{k=1}^{K}\left(\gamma_t(j,k)\mathbf{R}_{j,k}(\mathbf{x}_t-\overline{\mathbf{A}}_c\boldsymbol{\mu}_{j,k})-\gamma_t(j,k)\mathbf{R}_{j,k}\overline{\mathbf{b}}_c\right)\right]+\boldsymbol{\Xi}_c\boldsymbol{\beta}_c=\boldsymbol{\Xi}_c\overline{\mathbf{b}}_c$$

$$\left[\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(j,k)\mathbf{R}_{j,k}(\mathbf{x}_t-\overline{\mathbf{A}}_c\boldsymbol{\mu}_{j,k})\right]+\boldsymbol{\Xi}_c\boldsymbol{\beta}_c=\left[\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(j,k)\mathbf{R}_{j,k}\overline{\mathbf{b}}_c\right]+\boldsymbol{\Xi}_c\overline{\mathbf{b}}_c$$

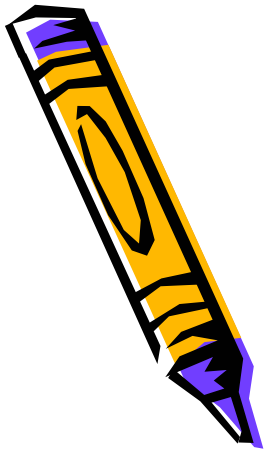
$$\therefore\overline{\mathbf{b}}_c=\left[\left(\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(j,k)\mathbf{R}_{j,k}\right)+\boldsymbol{\Xi}_c\right]^{-1}\left[\left(\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(j,k)\mathbf{R}_{j,k}(\mathbf{x}_t-\overline{\mathbf{A}}_c\boldsymbol{\mu}_{j,k})\right)+\boldsymbol{\Xi}_c\boldsymbol{\beta}_c\right]$$

MAPLR



- Second, we'll differentiate the auxiliary function w.r.t the transformation
- But how to choose the prior distribution for the transformation $\overline{\mathbf{A}}_c$?
- Unlike MAP estimation of HMM parameters, no obvious conjugate prior densities could be found.
- If we assume that the transformation is modeled by Normal-Wishart density, we would get a equation with no closed-form solution.
- The derivation will be shown in the following :

MAPLR



- Prior : $p(\overline{\mathbf{A}}_c) \propto |\overline{\mathbf{A}}_c|^{\frac{\alpha-p-1}{2}} e^{-\frac{1}{2}tr(\boldsymbol{\tau}_c \overline{\mathbf{A}}_c)}$
- Differentiating w.r.t $\overline{\mathbf{A}}_c$ and equate to zero

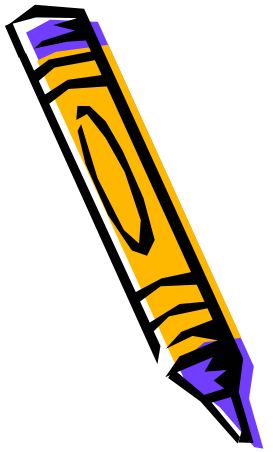
$$\frac{\partial}{\partial \overline{\mathbf{A}}_c} Q(\overline{\mathbf{A}}_c, \overline{\mathbf{b}}_c \mid \mathbf{A}_c, \mathbf{b}_c) =$$

$$\left\{ \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \times \left[-\frac{1}{2} \times \left(2 \times \mathbf{R}_{j,k} (\mathbf{x}_t - \overline{\mathbf{A}}_c \boldsymbol{\mu}_{j,k} - \overline{\mathbf{b}}_c) \times (-1) \times \boldsymbol{\mu}_{j,k}^T \right) \right] \right\}$$

$$+ \frac{1}{|\overline{\mathbf{A}}_c|^{\frac{\alpha-p-1}{2}} e^{-\frac{1}{2}tr(\boldsymbol{\tau}_c \overline{\mathbf{A}}_c)}} \left[\begin{array}{l} \frac{\alpha-p-1}{2} |\overline{\mathbf{A}}_c|^{\frac{\alpha-p-2}{2}} \times |\overline{\mathbf{A}}_c| \times \overline{\mathbf{A}}_c^{-T} \times e^{-\frac{1}{2}tr(\boldsymbol{\tau}_c \overline{\mathbf{A}}_c)} \\ + e^{-\frac{1}{2}tr(\boldsymbol{\tau}_c \overline{\mathbf{A}}_c)} \times \left(-\frac{1}{2} \times \boldsymbol{\tau}_c^T \right) \times |\overline{\mathbf{A}}_c|^{\frac{\alpha-p-1}{2}} \end{array} \right] = 0$$

$$where\ \gamma_t(j,k) = \Pr(s_t = j, l_t = k \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}_c)$$

MAPLR



$$\therefore \left\{ \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \times \left[\mathbf{R}_{j,k} (\mathbf{x}_t - \overline{\mathbf{A}}_c \boldsymbol{\mu}_{j,k} - \overline{\mathbf{b}}_c) \boldsymbol{\mu}_{j,k}^T \right] \right\}$$

$$+ \frac{1}{|\overline{\mathbf{A}}_c|^{\frac{\alpha-p-1}{2}} e^{-\frac{1}{2}tr(\boldsymbol{\tau}_c \overline{\mathbf{A}}_c)}} |\overline{\mathbf{A}}_c|^{\frac{\alpha-p-1}{2}} e^{-\frac{1}{2}tr(\boldsymbol{\tau}_c \overline{\mathbf{A}}_c)} \left[\frac{\alpha-p-1}{2} \times \overline{\mathbf{A}}_c^{-T} - \frac{1}{2} \boldsymbol{\tau}_c^T \right] = 0$$

$$\Rightarrow \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_t(j,k) \times \left[\mathbf{R}_{j,k} (\mathbf{x}_t - \overline{\mathbf{A}}_c \boldsymbol{\mu}_{j,k} - \overline{\mathbf{b}}_c) \boldsymbol{\mu}_{j,k}^T \right] + \frac{\alpha-p-1}{2} \times \overline{\mathbf{A}}_c^{-T} - \frac{1}{2} \boldsymbol{\tau}_c^T = 0$$

$$\Rightarrow \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \times \left[\mathbf{R}_{j,k} (\mathbf{x}_t - \overline{\mathbf{b}}_c) \boldsymbol{\mu}_{j,k}^T - \mathbf{R}_{j,k} \overline{\mathbf{A}}_c \boldsymbol{\mu}_{j,k} \boldsymbol{\mu}_{j,k}^T \right] + \frac{\alpha-p-1}{2} \times \overline{\mathbf{A}}_c^{-T} - \frac{1}{2} \boldsymbol{\tau}_c^T = 0$$

$$\Rightarrow \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \mathbf{R}_{j,k} \overline{\mathbf{A}}_c \boldsymbol{\mu}_{j,k} \boldsymbol{\mu}_{j,k}^T - \frac{\alpha-p-1}{2} \times \overline{\mathbf{A}}_c^{-T}$$

$$= \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \mathbf{R}_{j,k} (\mathbf{x}_t - \overline{\mathbf{b}}_c) \boldsymbol{\mu}_{j,k}^T - \frac{1}{2} \boldsymbol{\tau}_c^T$$

where $\gamma_t(j,k) = \Pr(s_t = j, l_t = k \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}_c)$

....No Closed - form solution can be obtained

MAPLR

- Chou suggests to select $p(\overline{\mathbf{W}})$ from a family of elliptically symmetric distributions.
- Here, a special case of elliptical distribution can be seen as a matrix version of a multivariate normal distribution.

$$p(\overline{\mathbf{W}}) \propto |\boldsymbol{\Sigma}|^{\frac{-(p+1)}{2}} |\boldsymbol{\Phi}|^{\frac{-p}{2}} \exp\left[-\frac{1}{2} tr\left((\overline{\mathbf{W}}-\mathbf{M})^T \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{W}}-\mathbf{M}) \boldsymbol{\Phi}^{-1}\right)\right]$$

$$\text{where } \overline{\mathbf{W}}, \mathbf{M} \in \mathbf{R}^{p\times(p+1)}$$

$$\boldsymbol{\Sigma} \in \mathbf{R}^{p\times p}, \boldsymbol{\Sigma} \geq 0 \quad \boldsymbol{\Phi} \in \mathbf{R}^{(p+1)\times(p+1)}, \boldsymbol{\Phi} \geq 0$$

MAPLR

- So, we can maximize the Q-function w.r.t. transformation $\overline{\mathbf{A}}$ and bias $\overline{\mathbf{b}}$ simultaneously. (i.e. with respect to $\overline{\mathbf{W}}$)
- The Q-function can be rewrite as

$$Q(\overline{\boldsymbol{\eta}}_c \mid \boldsymbol{\eta}_c)$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left[-\frac{1}{2} tr\left((\mathbf{x}_t - \overline{\mathbf{W}}_c \boldsymbol{\xi}_{s_t,l_t})(\mathbf{x}_t - \overline{\mathbf{W}}_c \boldsymbol{\xi}_{s_t,l_t})^T \mathbf{R}_{s_t,l_t} \right) \right] \right]$$

$$+ \log p(\overline{\boldsymbol{\eta}}_c)$$

MAPLR

- Differentiating Q-function w.r.t $\overline{\mathbf{W}}_c$ and set it to zero.

$$\frac{\partial}{\partial \overline{\mathbf{W}}_c} Q(\overline{\boldsymbol{\eta}}_c \mid \boldsymbol{\eta}_c) = 0$$

$$\therefore \frac{\partial}{\partial \overline{\mathbf{W}}_c} \left\{ \begin{array}{l} \left[\left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left[-\frac{1}{2} tr\left((\mathbf{x}_t - \overline{\mathbf{W}}_c \boldsymbol{\xi}_{j,k})(\mathbf{x}_t - \overline{\mathbf{W}}_c \boldsymbol{\xi}_{j,k})^T \mathbf{R}_{j,k} \right) \right] \right] \\ -\frac{1}{2} tr\left((\overline{\mathbf{W}}_c - \mathbf{M}_c)^T \boldsymbol{\Sigma}_c^{-1} (\overline{\mathbf{W}}_c - \mathbf{M}_c) \boldsymbol{\Phi}_c^{-1} \right) \end{array} \right\} = 0$$

- In the following we drop subscript *c*. It should be clear that the summation should be performed only on acoustic units belonging to the cluster *c*.

MAPLR

$$\frac{\partial}{\partial \overline{\mathbf{W}}_c} Q(\overline{\boldsymbol{\eta}}_c \mid \boldsymbol{\eta}_c)$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left[-\frac{1}{2} \times 2 \times \mathbf{R}_{j,k} \times (\mathbf{x}_t - \overline{\mathbf{W}} \boldsymbol{\xi}_{j,k}) \times (-1) \times \boldsymbol{\xi}_{j,k}^T \right] \right]$$

$$-\frac{1}{2} \times \left[\boldsymbol{\Sigma}^{-T} (\overline{\mathbf{W}} - \mathbf{M}) \boldsymbol{\Phi}^{-T} + \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{W}} - \mathbf{M}) \boldsymbol{\Phi}^{-1} \right] = 0$$

$$\Rightarrow \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left[\mathbf{R}_{j,k} (\mathbf{x}_t - \overline{\mathbf{W}} \boldsymbol{\xi}_{j,k}) \boldsymbol{\xi}_{j,k}^T \right]$$

$$= \frac{1}{2} \left[\boldsymbol{\Sigma}^{-T} (\overline{\mathbf{W}} - \mathbf{M}) \boldsymbol{\Phi}^{-T} + \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{W}} - \mathbf{M}) \boldsymbol{\Phi}^{-1} \right]$$

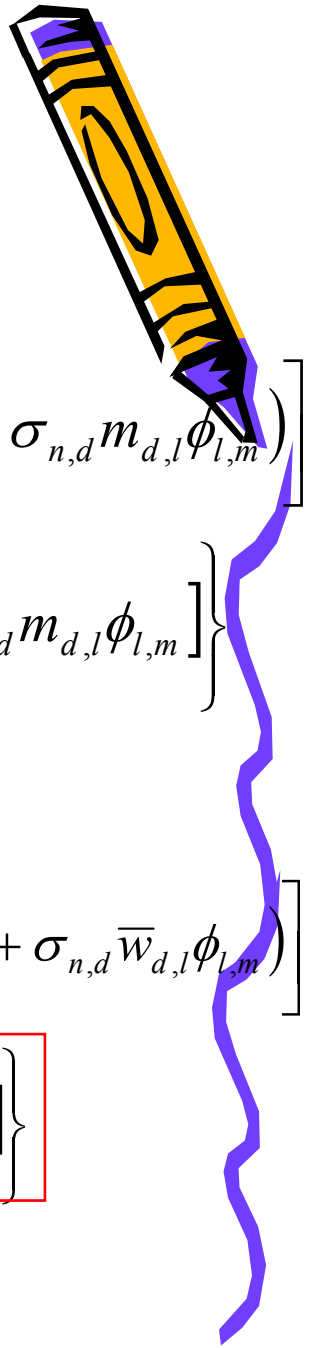
$$\Rightarrow \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \mathbf{R}_{j,k} \mathbf{x}_t \boldsymbol{\xi}_{j,k}^T + \frac{1}{2} \boldsymbol{\Sigma}^{-T} \mathbf{M} \boldsymbol{\Phi}^{-T} + \frac{1}{2} \boldsymbol{\Sigma}^{-1} \mathbf{M} \boldsymbol{\Phi}^{-1}$$

$$= \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \mathbf{R}_{j,k} \overline{\mathbf{W}} \boldsymbol{\xi}_{j,k} \boldsymbol{\xi}_{j,k}^T + \frac{1}{2} \boldsymbol{\Sigma}^{-T} \overline{\mathbf{W}} \boldsymbol{\Phi}^{-T} + \frac{1}{2} \boldsymbol{\Sigma}^{-1} \overline{\mathbf{W}} \boldsymbol{\Phi}^{-1}$$

$[Z]_{D\times(D+1)}$

$[H]_{D\times(D+1)}$

MAPLR



$$[\mathbf{Z}]_{n,m}$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left(\sum_{d=1}^{D} r_{n,d}^{(j,k)} x_d^{(t)} \xi_m^{(j,k)} \right) \right] + \frac{1}{2} \left[\sum_{d=1}^{D} \sum_{l=1}^{D+1} \left(\sigma_{d,n} m_{d,l} \phi_{m,l} + \sigma_{n,d} m_{d,l} \phi_{l,m} \right) \right]$$

$$= \sum_{d=1}^{D} \sum_{l=1}^{D+1} \left\{ \left[\sum_{j=1}^{N} \sum_{k=1}^{K} \left(\sum_{t=1}^{T} \gamma_t(j,k) x_d^{(t)} \right) r_{n,d}^{(j,k)} \xi_m^{(j,k)} \right] + \frac{1}{2} \left[\sigma_{d,n} m_{d,l} \phi_{m,l} + \sigma_{n,d} m_{d,l} \phi_{l,m} \right] \right\}$$

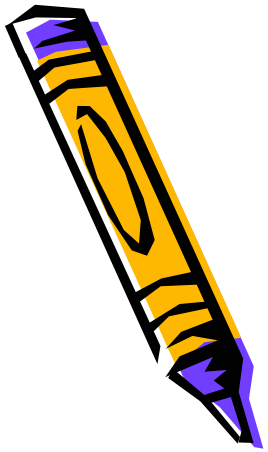
$$[H]_{n,m} =$$

$$\left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left(\sum_{d=1}^{D} \sum_{l=1}^{D+1} r_{n,d}^{(j,k)} \overline{w}_{d,l} \xi_l^{(j,k)} \xi_m^{(j,k)} \right) \right] + \frac{1}{2} \left[\sum_{d=1}^{D} \sum_{l=1}^{D+1} \left(\sigma_{d,n} \overline{w}_{d,l} \phi_{m,l} + \sigma_{n,d} \overline{w}_{d,l} \phi_{l,m} \right) \right]$$

$$= \sum_{d=1}^{D} \sum_{l=1}^{D+1} \overline{w}_{d,l} \left\{ \left[\sum_{j=1}^{N} \sum_{k=1}^{K} \left(\sum_{t=1}^{T} \gamma_t(j,k) \right) r_{n,d}^{(j,k)} \xi_l^{(j,k)} \xi_m^{(j,k)} \right] + \frac{1}{2} \left[\sigma_{d,n} \phi_{m,l} + \sigma_{n,d} \phi_{l,m} \right] \right\}$$

$$g_{d,l}^{(n,m)}$$

MAPLR

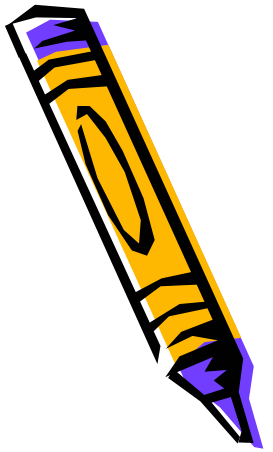


- The matrix $\overline{\mathbf{W}}$ can be obtained by solving the system of $D \times (D+1)$ linear equations:

$$\sum_{d,l} g_{d,l}^{(n,m)} \overline{w}_{d,l} = [\mathbf{Z}]_{n,m} \quad n = 1, \cdots, D \quad m = 1, \cdots, D+1$$

$$\begin{cases} \sum_{d,l} g_{d,l}^{(1,1)} \overline{w}_{d,l} = [\mathbf{Z}]_{1,1} \\ \vdots \\ \sum_{d,l} g_{d,l}^{(1,D+1)} \overline{w}_{d,l} = [\mathbf{Z}]_{1,D+1} \\ \vdots \\ \sum_{d,l} g_{d,l}^{(D,1)} \overline{w}_{d,l} = [\mathbf{Z}]_{D,1} \\ \vdots \\ \sum_{d,l} g_{d,l}^{(D,D+1)} \overline{w}_{d,l} = [\mathbf{Z}]_{D,D+1} \end{cases}$$

MAPLR



- If we assume that $\mathbf{R}_{j,k}$ and $\boldsymbol{\Sigma}$ are diagonal then

$$[\mathbf{Z}]_{n,m} = \sum_{l=1}^{D+1} \left\{ \left[\sum_{j=1}^{N} \sum_{k=1}^{K} \left(\sum_{t=1}^{T} \gamma_t(j,k) x_n^{(t)} \right) r_{n,n}^{(j,k)} \xi_m^{(j,k)} \right] + \frac{1}{2} \left[\sigma_{n,n} m_{n,l} \phi_{m,l} + \sigma_{n,n} m_{n,l} \phi_{l,m} \right] \right\}$$

$$[\mathbf{H}]_{n,m} = \sum_{l=1}^{D+1} \overline{w}_{n,l} \left\{ \left[\sum_{j=1}^{N} \sum_{k=1}^{K} \left(\sum_{t=1}^{T} \gamma_t(j,k) \right) r_{n,n}^{(j,k)} \xi_l^{(j,k)} \xi_m^{(j,k)} \right] + \frac{1}{2} \left[\sigma_{n,n} \phi_{m,l} + \sigma_{n,n} \phi_{l,m} \right] \right\}$$

- So the matrix $\overline{\mathbf{W}}$ can be obtained by $g_l^{(n,m)}$ solving D system $D+1$ linear equations.

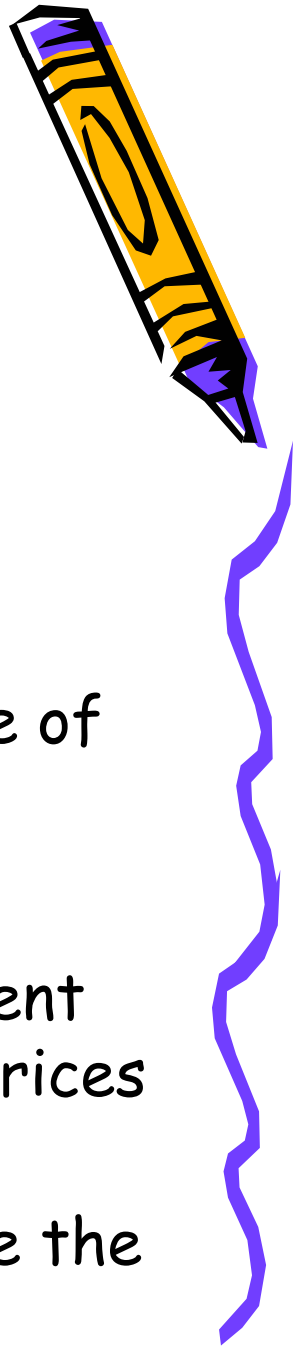
MAPLR

$$\sum_{l=1}^{D+1} \overline{w}_{n,l} g_l^{(n,m)} = [\mathbf{Z}]_{n,m} \quad m = 1, \cdots, D+1 \quad \text{for each } n\text{th system}$$

- When $n = 3$

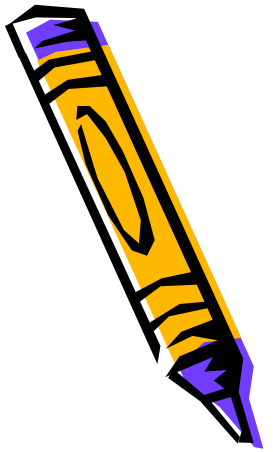
$$\begin{cases} \overline{w}_{3,1} g_1^{(3,1)} + \cdots + \overline{w}_{3,D+1} g_{D+1}^{(3,1)} = [\mathbf{Z}]_{3,1} \\ \vdots \\ \overline{w}_{3,D+1} g_{D+1}^{(3,D+1)} + \cdots + \overline{w}_{3,D+1} g_{D+1}^{(3,D+1)} = [\mathbf{Z}]_{3,D+1} \end{cases}$$

MAPLR



- Hyper-parameter Estimation :
  - Basic principle :
    - First generate a set of N transformation matrices.
    - Then use N matrices to derive an estimate of the hyper-parameters.
  - Another way:
    - Select a set of training data for N different speakers and estimate the regression matrices for each speaker using a MLLR approach.
    - Drawback: Some data is required to derive the hyper-parameter.

MAPLR



- Hyper-parameter Estimation :
  - Another attractive technique:
    - Collect the mean vectors from the SI models associated to a given regression class.
    - Assign the mean vectors into N subsets $\mathbf{S}_i \; i = 1, \cdots, N$
    - For each subset $\mathbf{S}_i$, the subset can be further partitioned into 2 subsets $\mathbf{L}_i$ and $\mathbf{R}_i$
    - Each mean vectors from $\mathbf{L}_i$ can be projected onto each mean vectors from $\mathbf{R}_i$ by applying the transformation.
    - $\mathbf{L}_i$ can be easily estimated using MLLR by assuming that $\mathbf{R}_i$ is a set of data vectors and corresponds to the models.

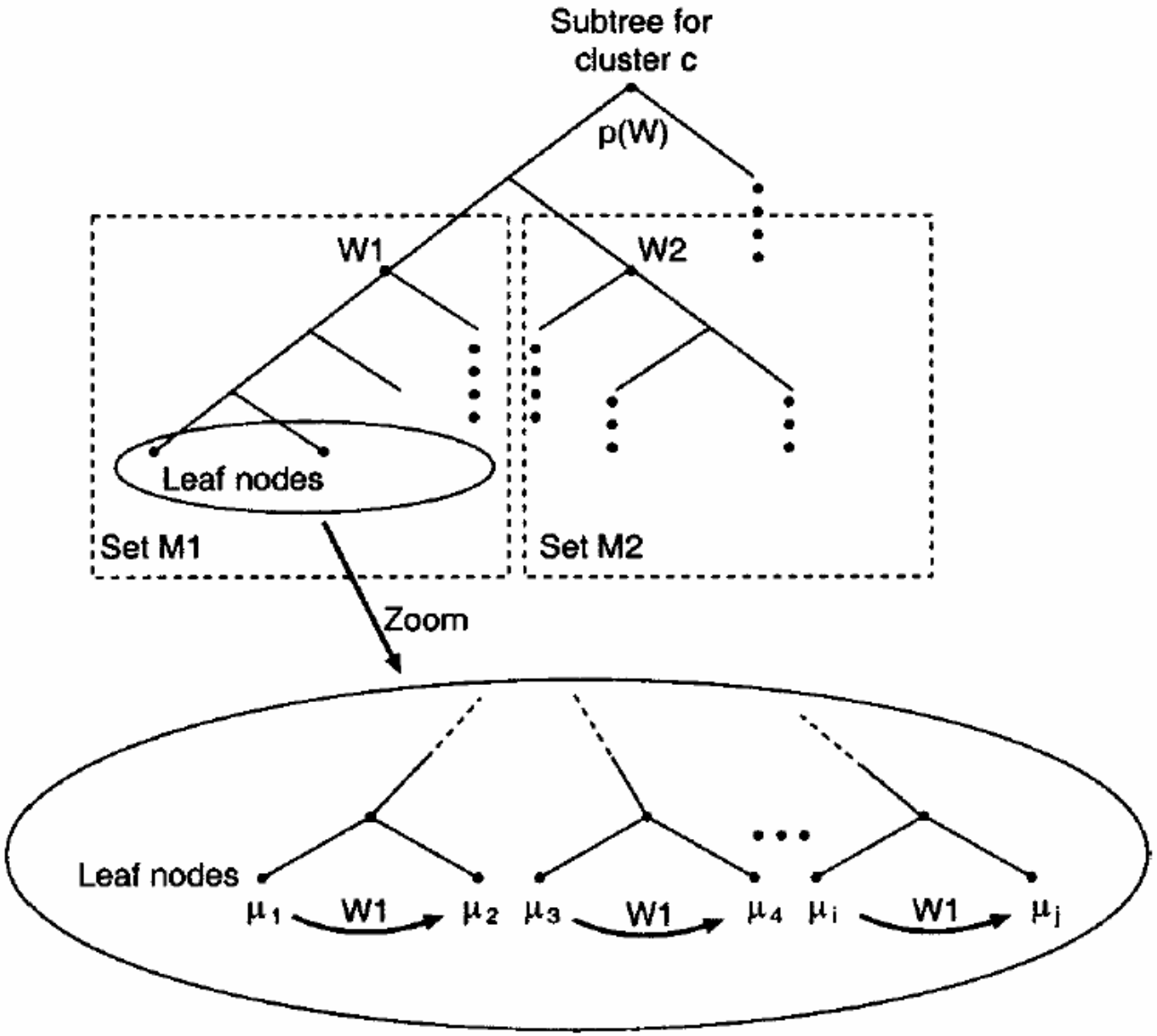


Fig. 5. Estimation of the prior densities from the tree of SI Gaussian distributions. K sets $\mathcal{M}_1, \cdots \mathcal{M}_K$ are defined for each tree associated to a cluster c. The mean vectors $\boldsymbol{\mu}_j$ associated to each set $\mathcal{M}_i$ are clustered into two subsets $\mathcal{L}_i$ and $\mathcal{R}_i$. A projection matrix $\mathbf{W}_i$ is derive using $\mathcal{L}_i$ as data and $\mathcal{R}_i$ as model.

MAPLR

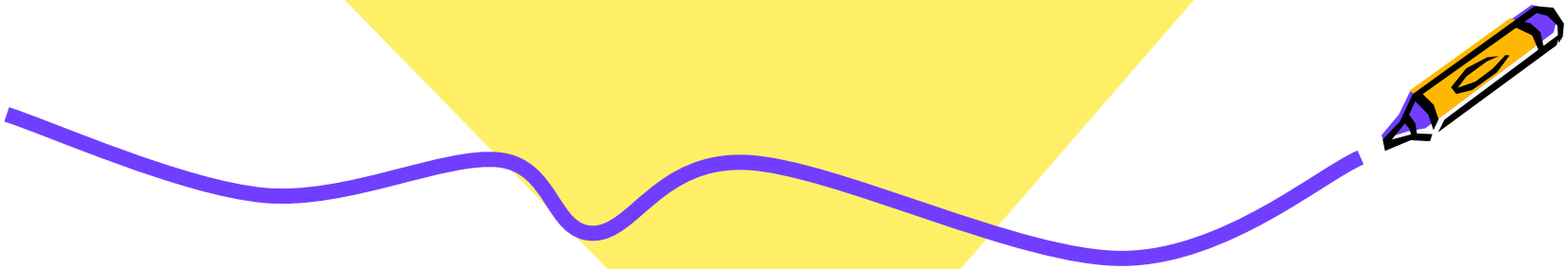
- Hyper-parameter Estimation :
  - Once these N transformation matrices $\mathbf{W}_i$ have been obtained, we can use a maximum likelihood approach to derive the $\mathbf{M}, \boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$

$$\mathbf{M} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{W}_i$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{W}_i - \mathbf{M})\boldsymbol{\Phi}^{-1}(\mathbf{W}_i - \mathbf{M})^T$$

$\boldsymbol{\Phi}$ can be chosen as the identity matrix

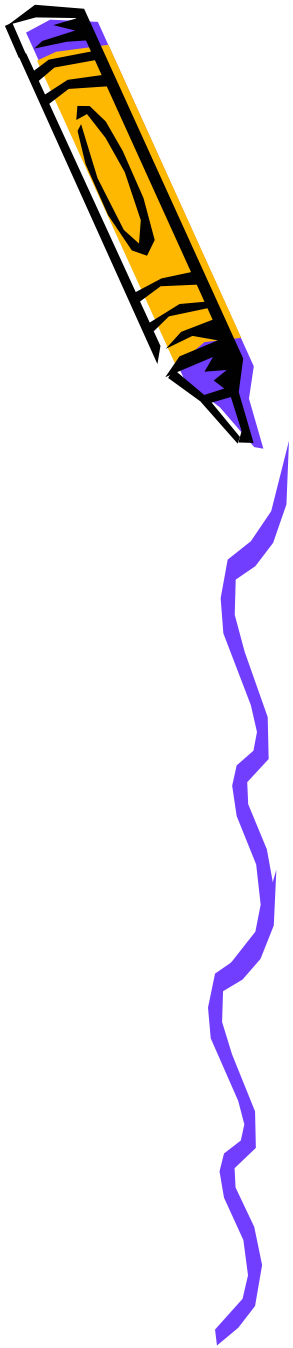
MAPLR Variance Adaptation



Reference

- Maximum a Posterior Linear Regression (MAPLR) Variance Adaptation for Continuous Density HMMs – EuroSpeech'03 Wu Chou and Xiaodong He

MAPLR Variance Adaptation



$$\boldsymbol{\Sigma}_n = \mathbf{B}_n^T \mathbf{B}_n = \boldsymbol{\Sigma}_n^{\frac{1}{2}} \boldsymbol{\Sigma}_n^{\frac{1}{2}} \quad \text{(Choleski factorization)}$$

$$\overline{\boldsymbol{\Sigma}}_n = \mathbf{B}_n^T \mathbf{H}_n \mathbf{B}_n$$

$$\text{ML criterion} \rightarrow \mathbf{H}_{\text{n}} = \arg\max_{\mathbf{H}} \left[\prod_k p(\mathbf{O}_m^k \mid \lambda_m) \right]$$

$$\text{MAP criterion} \rightarrow \mathbf{H}_{\text{n}} = \arg\max_{\mathbf{H}} \left[\prod_k p(\mathbf{O}_m^k \mid \lambda_m) \right] p(\mathbf{H})$$

Here, $\mathbf{H}_{\text{n}}$ is assumed to be diagonal

$$\mathbf{H}_{\text{n}} = diag\left[h_1^2, \cdots, h_D^2\right] \quad \text{let } \mathbf{h} = \left[h_1^2, \cdots, h_D^2\right]$$

$$p(\mathbf{h}) = (2\pi)^{-\frac{D}{2}} \left|\boldsymbol{\Sigma}_{\mathbf{h}}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{h}-\boldsymbol{\mu}_{\mathbf{h}})^{\mathbf{T}} \boldsymbol{\Sigma}_{\mathbf{h}}^{-1} (\mathbf{h}-\boldsymbol{\mu}_{\mathbf{h}})}$$

$$\boldsymbol{\mu}_{\mathbf{h}} = \left[\boldsymbol{\mu}_{\mathbf{h}}(1), \cdots, \boldsymbol{\mu}_{\mathbf{h}}(D)\right], \boldsymbol{\Sigma}_{\mathbf{h}} = \left[\boldsymbol{\sigma}_{\mathbf{h}}(1), \cdots, \boldsymbol{\sigma}_{\mathbf{h}}(D)\right]$$

MAPLR Variance Adaptation

$$Q(\overline{\mathbf{M}} \mid \mathbf{M})$$

$$= -\frac{1}{2}\sum_{t,n,m}\gamma_t(n,m)\left[\log\left|\overline{\boldsymbol{\Sigma}}_m\right| + (\mathbf{o}_t - \boldsymbol{\mu}_m)^T \overline{\boldsymbol{\Sigma}}_m^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_m)\right]$$

$$+ \log p(\mathbf{h})$$

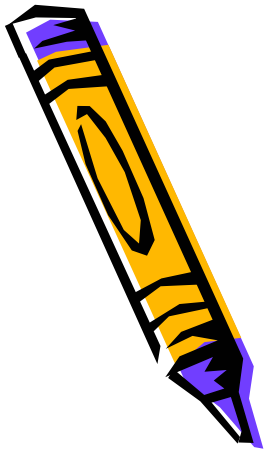
$$= -\frac{1}{2}\sum_{t,n,m}\sum_{d=1}^{D}\gamma_t(n,m)\left[\log\left(\sigma_m^{(d)}\right)^2 + \log\left(h^{(d)}\right)^2 + \frac{\left(h^{(d)}\right)^{-2}}{\left(\sigma_m^{(d)}\right)^2}(o_t^{(d)} - \mu_m^{(d)})^2\right]$$

$$-\frac{1}{2}\log 2\pi\sigma_{\mathbf{h}}^2(d) - \frac{(h^{(d)} - \mu_h(d))^2}{2\sigma_{\mathbf{h}}^2(d)}$$

Set $\dfrac{\partial Q(\overline{\mathbf{M}} \mid \mathbf{M})}{\partial h^{(d)}} = 0$ for each d

$$-\sum_{t,n,m}\gamma_t(n,m)\left[2 \times \frac{1}{\left(h^{(d)}\right)} + (-2) \times \frac{(o_t^{(d)} - \mu_m^{(d)})^2}{\left(\sigma_m^{(d)}\right)^2}\left(h^{(d)}\right)^{-3}\right] - \frac{1}{\sigma_{\mathbf{h}}^2(d)} \times 2 \times (h^{(d)} - \mu_h(d)) = 0$$

MAPLR Variance Adaptation

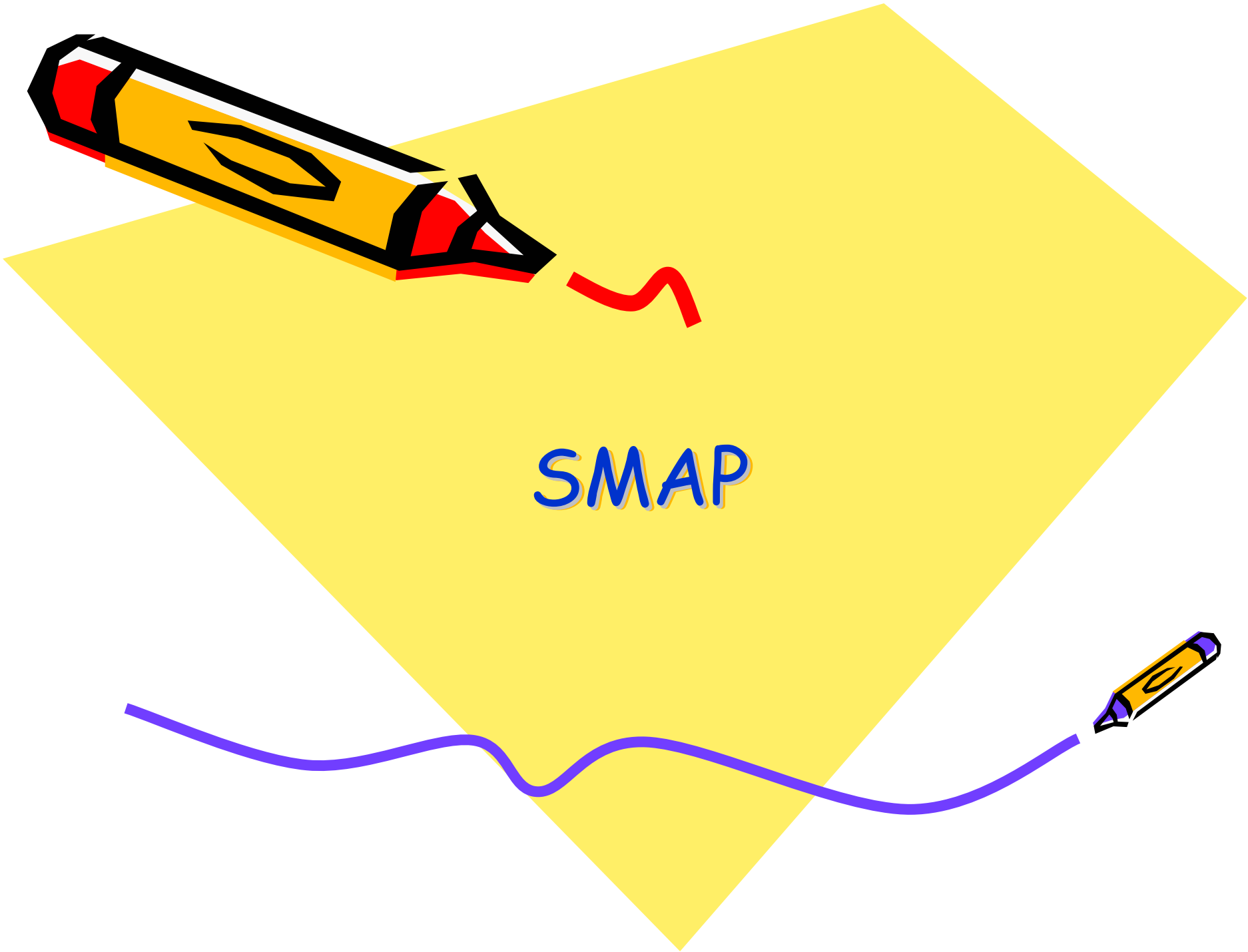


$$-\sum_{t,n,m}\gamma_t(n,m)\left[2\times\frac{1}{\left(h^{(d)}\right)}+(-2)\times\frac{(o_t^{(d)}-\mu_m^{(d)})^2}{\left(\sigma_m^{(d)}\right)^2}\left(h^{(d)}\right)^{-3}\right]-\frac{1}{\sigma_{\mathbf{h}}^2(d)}\times 2\times(h^{(d)}-\mu_h(d))=0$$

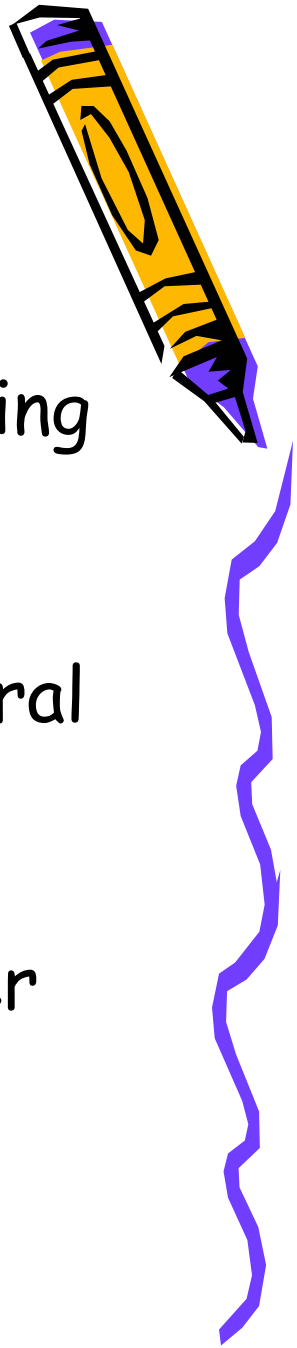
$$-\sum_{t,n,m}\gamma_t(n,m)\left[\left(h^{(d)}\right)^{-1}-\left(h^{(d)}\right)^{-3}\frac{(o_t^{(d)}-\mu_m^{(d)})^2}{\left(\sigma_m^{(d)}\right)^2}\right]-\frac{h^{(d)}}{\sigma_{\mathbf{h}}^2(d)}+\frac{\mu_h(d)}{\sigma_{\mathbf{h}}^2(d)}=0$$

$$\sum_{t,n,m}\gamma_t(n,m)\left[-\left(h^{(d)}\right)^2+\frac{(o_t^{(d)}-\mu_m^{(d)})^2}{\left(\sigma_m^{(d)}\right)^2}\right]-\frac{\left(h^{(d)}\right)^4}{\sigma_{\mathbf{h}}^2(d)}+\frac{\mu_h(d)}{\sigma_{\mathbf{h}}^2(d)}\left(h^{(d)}\right)^3=0$$

$$-\frac{1}{\sigma_{\mathbf{h}}^2(d)}\left(h^{(d)}\right)^4+\frac{\mu_h(d)}{\sigma_{\mathbf{h}}^2(d)}\left(h^{(d)}\right)^3-\sum_{t,n,m}\gamma_t(n,m)\left(h^{(d)}\right)^2+\sum_{t,n,m}\gamma_t(n,m)\frac{(o_t^{(d)}-\mu_m^{(d)})^2}{\left(\sigma_m^{(d)}\right)^2}=0$$

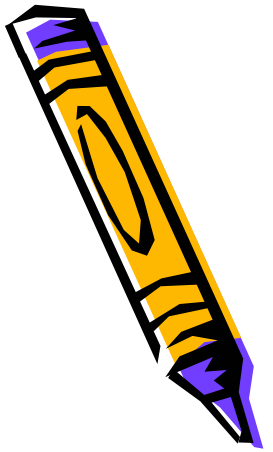

SMAP

Reference



- Structural MAP Speaker Adaptation Using Hierarchical Priors – ASRU'97 Koichi Shinoda and Chin-Hui Lee
- Unsupervised Adaptation Using Structural Bayes Approach – ICASSP'98 Koichi Shinoda and Chin-Hui Lee
- A Structural Bayes Approach to Speaker Adaptation – SAP'01 Koichi Shinoda and Chin-Hui Lee

Tree Structure



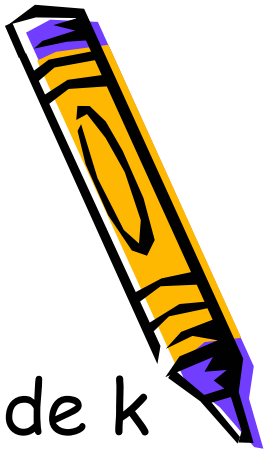
- Distance measure between two gaussian components:

$$d(m,n) = \int g_m(x) \log \frac{g_m(x)}{g_n(x)} dx + \int g_n(x) \log \frac{g_n(x)}{g_m(x)} dx$$

$$= \sum_i \left[\frac{\sigma_m^2(i) - \sigma_n^2(i) + \left(\mu_n(i) - \mu_m(i)\right)^2}{\sigma_n^2(i)} + \frac{\sigma_n^2(i) - \sigma_m^2(i) + \left(\mu_n(i) - \mu_m(i)\right)^2}{\sigma_m^2(i)} \right]$$

  - The sum of the Kullback-Leibler divergence

Tree Structure



- Approximate single gaussian for each node k

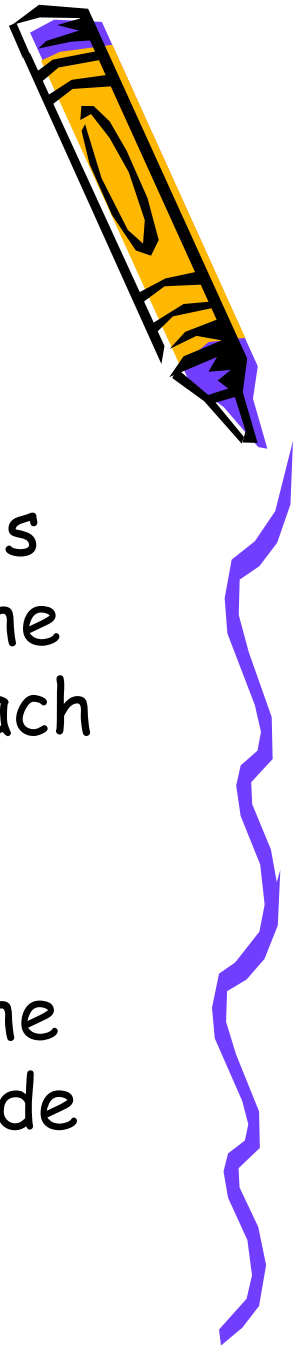
$$\mu_k(i) = \frac{1}{M_k}\sum_{m=1}^{M_k} E\left[x_m^{(k)}(i)\right] = \frac{1}{M_k}\sum_{m=1}^{M_k} \mu_m^{(k)}(i)$$

$$\sigma_k^2(i) = \frac{1}{M_k}\sum_{m=1}^{M_k} E\left[\left(x_m^{(k)}(i) - \mu_k(i)\right)^2\right]$$

$$= \frac{1}{M_k}\left[\sum_{k=1}^{M_k}\left(\sigma_m^{(k)}(i)\right)^2 + \sum_{m=1}^{M_k}\left(\mu_m^{(k)}(i)\right)^2 - M_k\mu_k^2(i)\right]$$

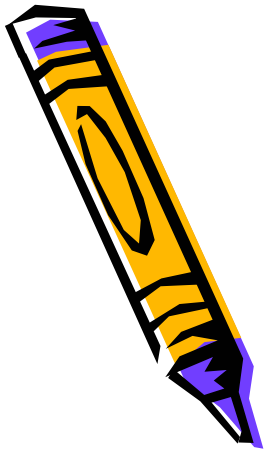
where M_k is the mixture number in node k

Tree Structure



- How to build a tree structure
  - The structure of the tree structure is designed; the number of layers and the number of branches from a node in each layer are determined.
  - Set the root node to be node k and the set G to be set G_{now}. Calculate the node pdf for the root node.

Tree Structure



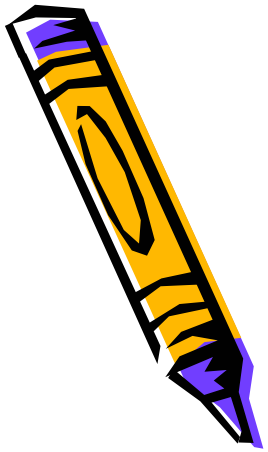
- How to build a tree structure
  - If node k has no child nodes, stop clustering. Otherwise, give the initial pdf for each child node using the minimax method that is described as follows.

$g^{(k)}(\cdot)$ is the node pdf for node k

P_k is the number of child nodes of node k

$g^{(c_p)}$ is the node pdf for child node c_p , $p = 1 \ldots P_k$

Tree Structure



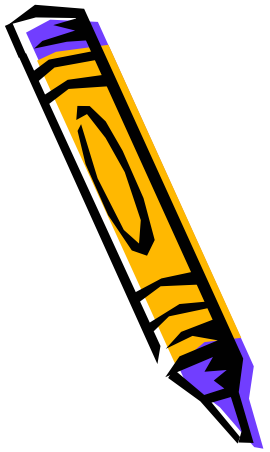
- Minimax method :
  - First :

    Choose among the set G_{now} the mixture component $\hat{m}$ that has the largest distance to $g^{(k)}(\cdot)$ and set it as node pdf for child c_1, i.e., $g^{(c_1)}(\cdot) = g_{\hat{m}}(\cdot)$
  - Second:

    Choose mixture components for c_p successively from $p = 2$ to $p = P_k$ and set those to the node pdfs for child nodes as follows :

$$\hat{m} = \arg\max_{m} \min_{1 \le i \le p-1} d(m, c_i) \qquad g^{(c_P)}(\cdot) = g_{\hat{m}}(\cdot)$$

Tree Structure



- Minimax method :
  - Third:

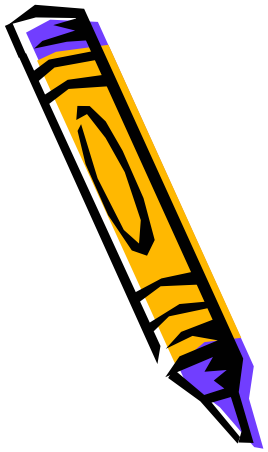
The node pdf for each child node c_p and the node pdf for k is interpolated and resulting pdf are set to be the node pdf for c_p as follows :

$$\mu'_{c_p}(i) = (1-\alpha)\mu_k(i) + \alpha\mu_{c_p}$$

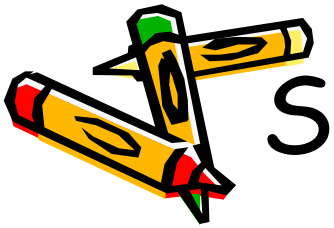
$$\sigma'^2_{c_p}(i) = (1-\alpha)\left(\sigma_k^2(i) + \mu_k^2(i)\right) + \alpha\left(\sigma_{c_p}^2(i) + \mu_{c_p}^2(i)\right) - \mu'_{c_p}$$

where $0 \le \alpha \le 1$

Tree Structure



- How to build a tree structure
  - Repeat the following k-means procedure until the grand sum of distances converges.
    - A) For each mixture in G_{now}, calculate the distance from it to each child node pdf, and assign each mixture to the nearest child node.
    - B) Recalculate the child node pdf
    - C) Calculate the sum of distances from each child node to each of its mixture and then obtain the grand sum.
  - Set each child node to be node k.



Summarization of Gaussian Distributions

Let $X = \{x_1, ..., x_T\}$ denote a given set of adaptation data.

each sample vector x_t is transformed into a vector y_{mt}

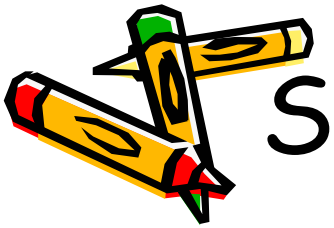
$$y_{mt} = \Sigma_m^{-1/2}(x_t - \mu_m) \quad t = 1, ..., T \quad m = 1, ..., M$$

When there is no mismatch between the training and adaptation data, the pdf for $\mathbf{Y}_m = \{y_{m1}, ..., y_{mT}\}$ is the standard normal distribution $N(\mathbf{Y} \mid \vec{0}, \mathbf{I})$

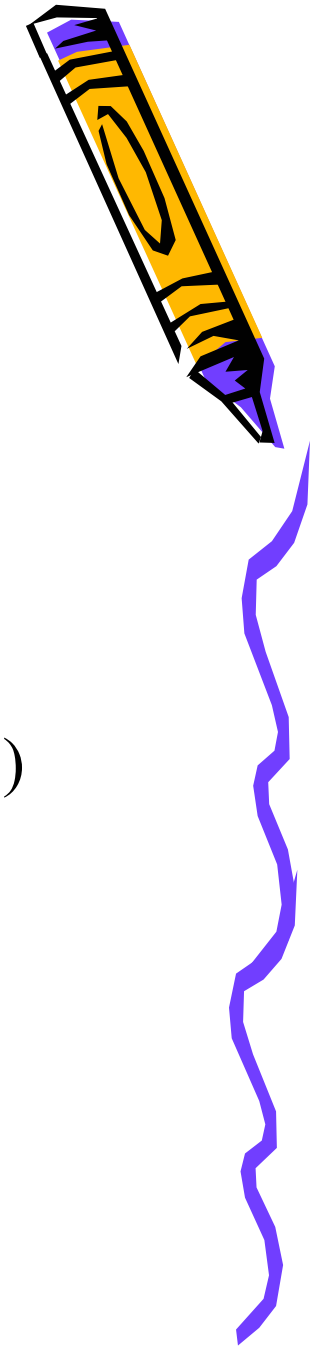
When there is a mismatch between them, the pdf for $\mathbf{Y}$ is different from $N(\mathbf{Y} \mid \vec{0}, \mathbf{I})$

Here the pdf for Y is assumed to be $N(\mathbf{Y} \mid \nu, \eta)$

where $\nu \neq \vec{0}$ and $\eta \neq \mathbf{I}$



Summarization of Gaussian Distributions



G which is the whole set of mixtures

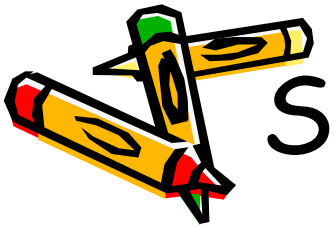
is divided into subsets $\{G_1,...,G_P\}$

where P is the total number of subsets

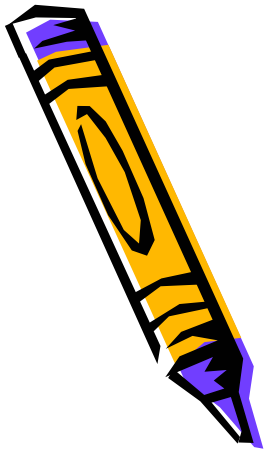
One common normalized pdf, $h^{(p)}(\cdot) = N(\mathbf{Y} \mid \nu^{(p)}, \eta^{(p)})$

is shared by all the mixture in each subset

$G_p = \{g_1^{(p)}(\cdot),..., g_m^{(p)}(\cdot),..., g_{M^{(p)}}^{(p)}(\cdot)\}$ where $M^{(p)}$ is the number of the mixture in subset G_p



Summarization of Gaussian Distributions



Then the ML estimates of the parameters $(\nu^{(p)}, \eta^{(p)})$

$$\tilde{\nu}^{(p)} = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M^{(p)}} \gamma_{mt}^{(p)} y_{mt}^{(p)}}{\sum_{t=1}^{T} \sum_{m=1}^{M^{(p)}} \gamma_{mt}^{(p)}}$$

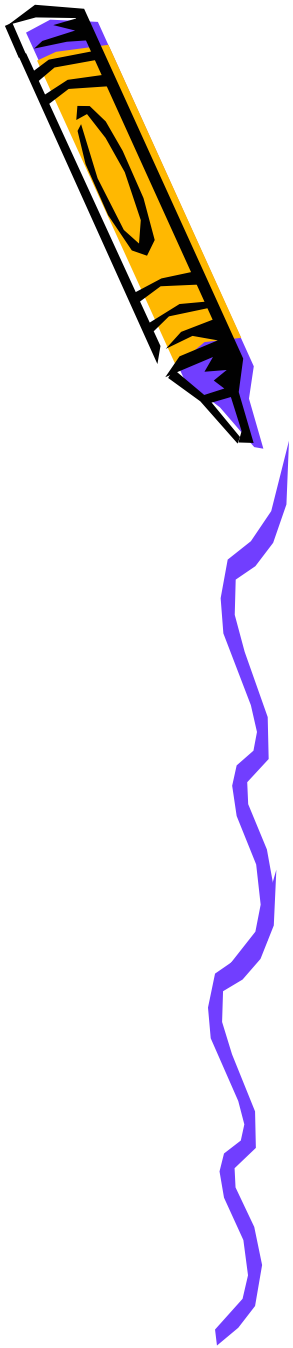
$$\tilde{\eta}^{(p)} = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M^{(p)}} \gamma_{mt}^{(p)} \left(y_{mt}^{(p)} - \tilde{\nu}^{(p)}\right)\left(y_{mt}^{(p)} - \tilde{\nu}^{(p)}\right)^{T}}{\sum_{t=1}^{T} \sum_{m=1}^{M^{(p)}} \gamma_{mt}^{(p)}}$$

Then HMM parameters are updated

$$\tilde{\mu}_m^{(p)} = \bar{\mu}_m^{(p)} + \left(\bar{\Sigma}_m^{(p)}\right)^{1/2} \tilde{\nu}^{(p)}$$

$$\tilde{\Sigma}_m^{(p)} = \left(\bar{\Sigma}_m^{(p)}\right)^{1/2} \tilde{\eta}^{(p)} \left[\left(\bar{\Sigma}_m^{(p)}\right)^{1/2}\right]^{T}$$

Hierarchical Prior



We have developed the tree structure representation

$$\hat{\lambda}_{k-1}^{(p)} = \left(\tilde{\nu}_{k-1}^{(p)}, \tilde{\eta}_{k-1}^{(p)}\right)$$

Now for each $k = 1,\ldots,K$, with given λ_{k-1}

$$\hat{\lambda}_k = \arg\max_{\lambda_k} p\left(\lambda_k \mid \hat{\lambda}_{k-1}, \mathbf{Y}\right)$$

$$= \arg\max_{\lambda_k} \frac{p\left(\mathbf{Y} \mid \lambda_k, \hat{\lambda}_{k-1}\right) p\left(\lambda_k \mid \hat{\lambda}_{k-1}\right)}{p(\mathbf{Y})}$$

$$\hat{\lambda}_0 = \lambda_0 = \left(\vec{0}, \mathbf{I}\right)$$

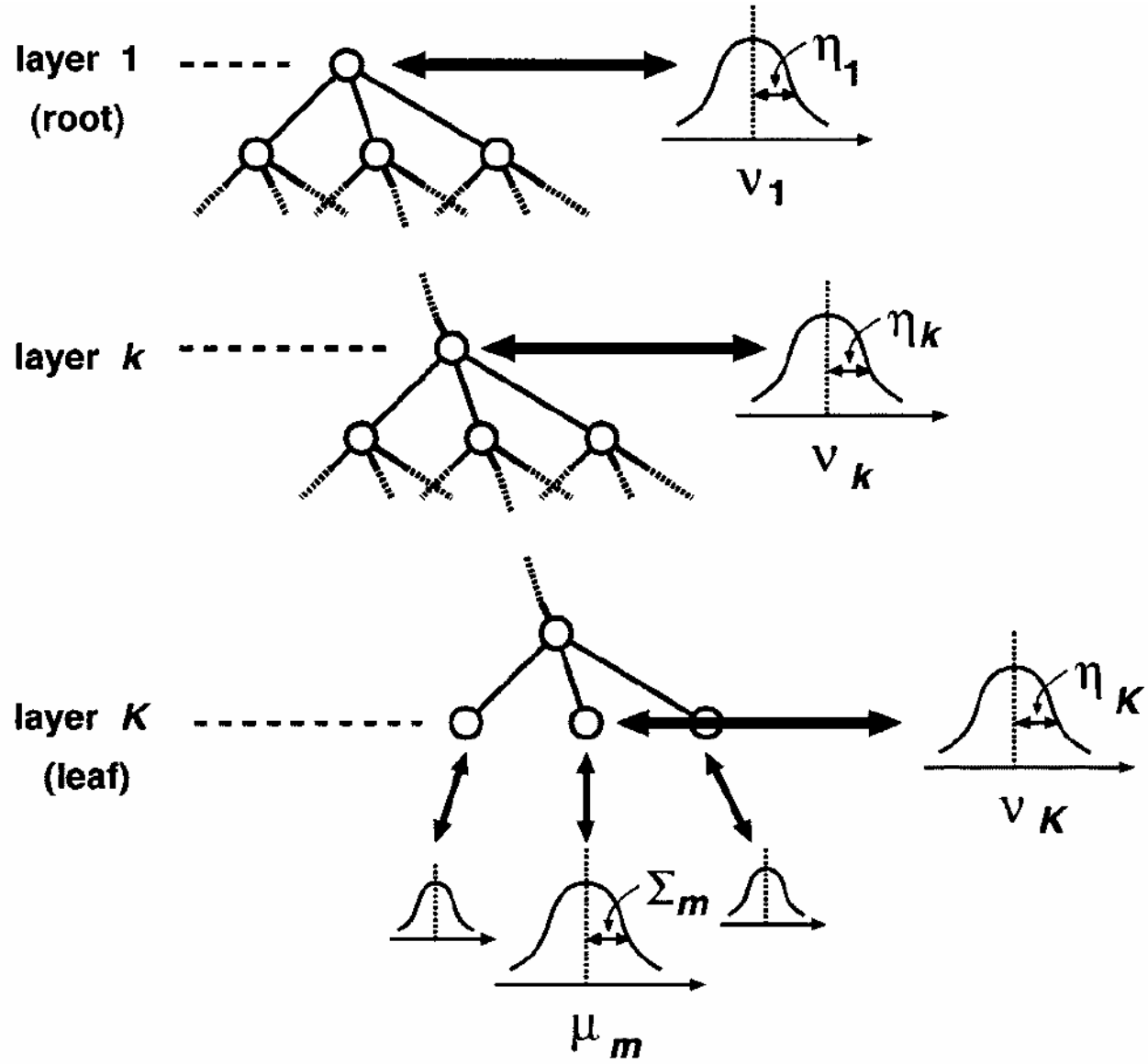
$$\Rightarrow \hat{\lambda}_k = \arg\max_{\lambda_k} p\left(\mathbf{Y} \mid \lambda_k\right) p\left(\lambda_k \mid \hat{\lambda}_{k-1}\right)$$

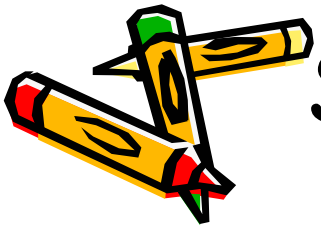
$p\left(\lambda_k \mid \hat{\lambda}_{k-1}\right)$ is assumed to be a normal - Wishart

$$p\left(\lambda_k \mid \hat{\lambda}_{k-1}\right) = g\left(\nu_k, \eta_k \mid \hat{\nu}_{k-1}, \hat{\eta}_{k-1}, \xi_k, \tau_k\right)$$

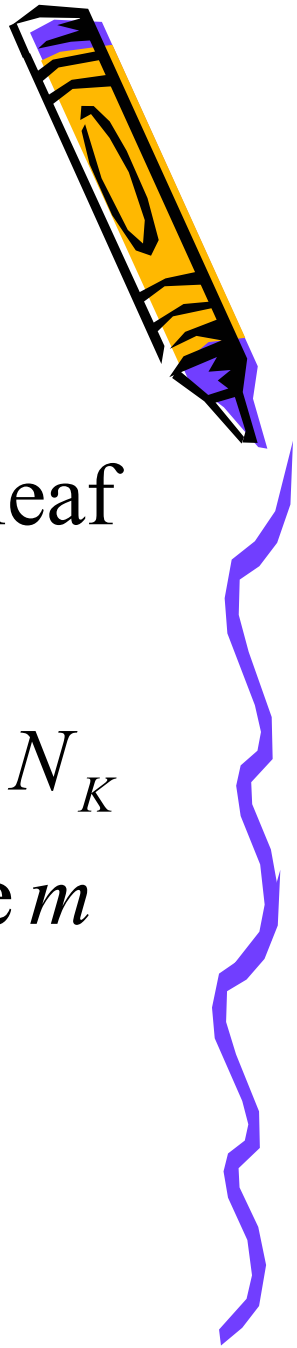
$$\propto \left|\eta_k\right|^{-\xi_k/2} \exp\left[-\frac{\tau_k}{2}\left(\nu_k - \hat{\nu}_{k-1}\right)^T \eta_k^{-1}\left(\nu_k - \hat{\nu}_{k-1}\right)\right] \exp\left[-\frac{1}{2} tr\left(\hat{\eta}_{k-1}\eta_k^{-1}\right)\right]$$

SMAP Adaptation Using Hierarchical Priors

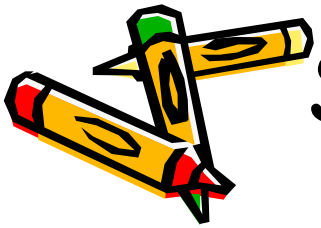




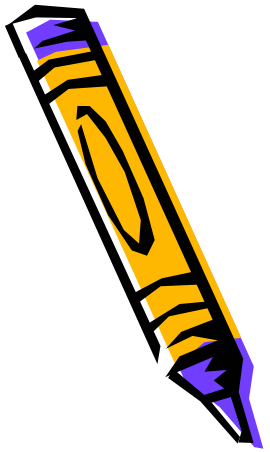
SMAP Adaptation Using Hierarchical Priors



Let the node sequence from the root to the leaf corresponding to the mth mixture be $\{N_1,...,N_k,...,N_K\}$ where N_1 is the root and N_K is the leaf node directly attached to mixture m



SMAP Adaptation Using Hierarchical Priors



Since $p(\lambda_k \mid \mathbf{Y}) = \int p(\lambda_k \mid \lambda_{k-1}, \mathbf{Y}) p(\lambda_{k-1} \mid \mathbf{Y}) d\lambda_{k-1} \quad \forall k = 1,..,K$

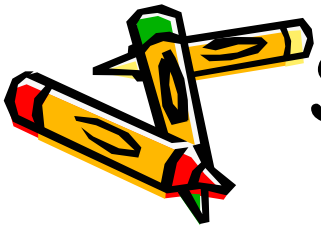
$$p(\lambda_K \mid \mathbf{Y}) = \int \cdots \int p(\lambda_K \mid \lambda_{K-1}, \mathbf{Y}) \cdots p(\lambda_k \mid \lambda_{k-1}, \mathbf{Y}) \cdots$$

$$(\lambda_1 \mid \lambda_0, \mathbf{Y}) p(\lambda_0 \mid \mathbf{Y}) d\lambda_0 d\lambda_1 \cdots d\lambda_{K-1} \quad \text{with} \int \mathrm{p}(\lambda_0 \mid \mathbf{Y}) d\lambda_0 = 1$$

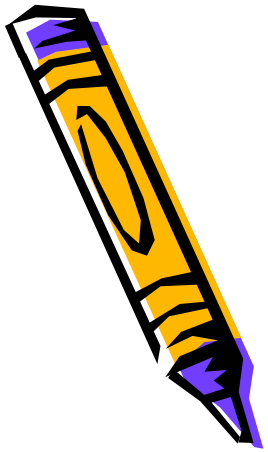
Because it is difficult to maximize directly, a key step is to be assume

$$\int p(\lambda_1 \mid \lambda_0 \mid \mathbf{Y}) p(\lambda_0 \mid \mathbf{Y}) d\lambda_0 \approx p(\lambda_1 \mid \hat{\lambda}_0, \mathbf{Y})$$

$$\int p(\lambda_{k+1} \mid \lambda_k \mid \mathbf{Y}) p(\lambda_k \mid \hat{\lambda}_{k-1}, \mathbf{Y}) d\lambda_k \approx p(\lambda_{k+1} \mid \hat{\lambda}_k, \mathbf{Y}) \quad k = 1,..., K-1$$



SMAP Adaptation Using Hierarchical Priors



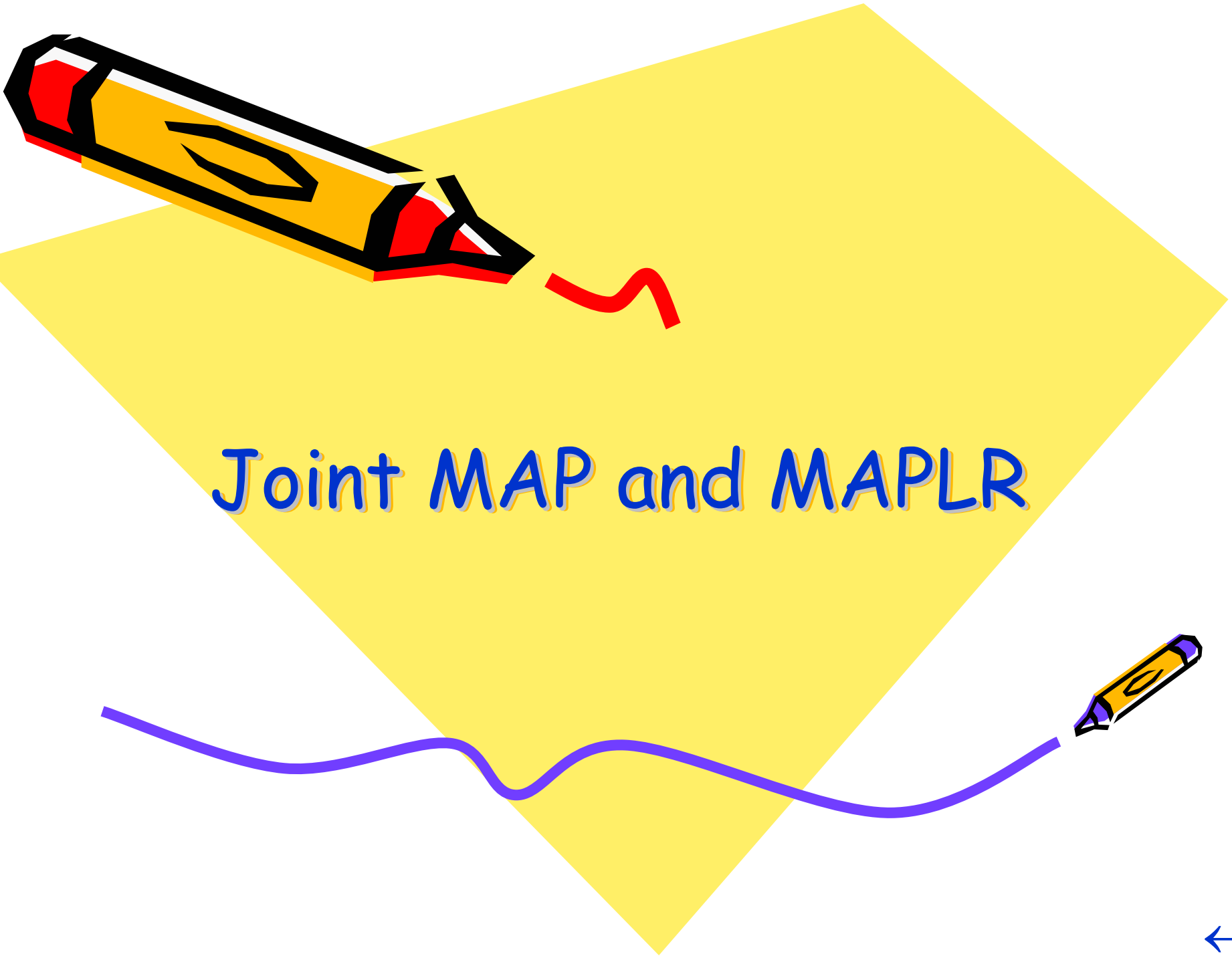
$$\therefore p(\lambda_K \mid \mathbf{Y}) \approx \prod_{k=1}^{K-1} p\left(\lambda_{k+1} \mid \hat{\lambda}_k, \mathbf{Y}\right)$$

Then

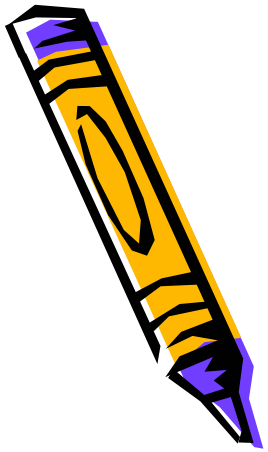
$$\hat{\nu}_k = \frac{\Gamma_k \tilde{\nu}_k + \tau_k \hat{\nu}_{k-1}}{\Gamma_k + \tau_k}$$

$$\hat{\eta}_k = \frac{\hat{\eta}_{k-1} + \Gamma_k \tilde{\eta}_k + \dfrac{\tau_k \Gamma_k}{\tau_k + \Gamma_k} (\tilde{\nu}_k - \hat{\nu}_{k-1})(\tilde{\nu}_k - \hat{\nu}_{k-1})^T}{\xi_k + \Gamma_k}$$

$$\therefore \hat{\nu}_K = \sum_{k=1}^{K} \omega_k \tilde{\nu}_k \quad \text{where} \quad \omega_k = \frac{\Gamma_k}{\Gamma_k + \tau_k} \prod_{i=k+1}^{K} \frac{\tau_i}{\Gamma_i + \tau_i}$$

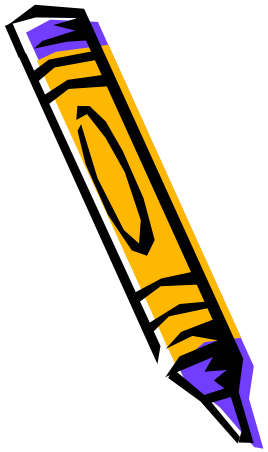

Joint MAP and MAPLR

Joint MAP and MAPLR



- Reference:
  - Speaker Adaptation Using combined Transformation and Bayesian Methods – SAP'96 Vassilios V. Digalakis
  - Joint Maximum a Posteriori Adaptation of Transformation and HMM Parameters – ICASSP'00 and SAP'01 O. Siohan

Joint MAP and MAPLR

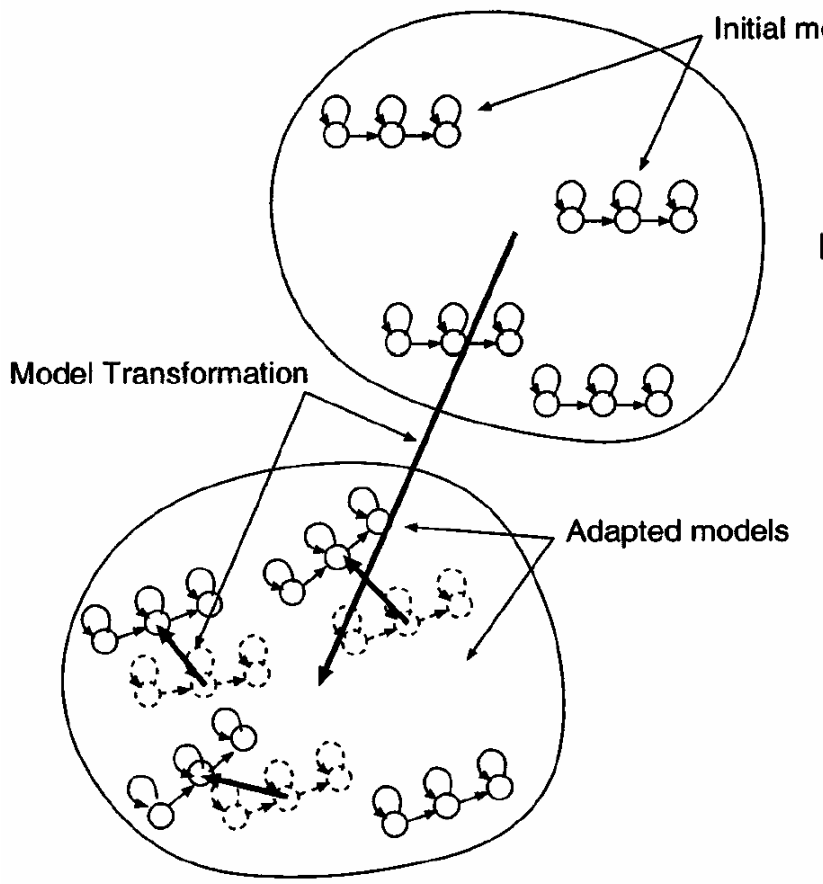


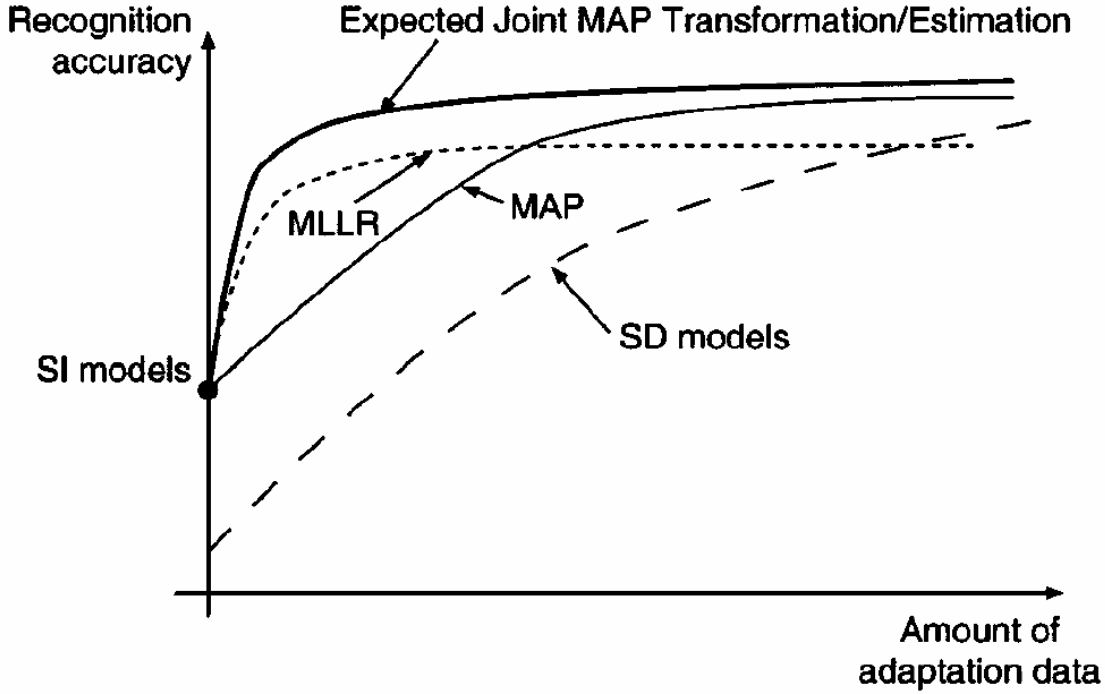
- MLLR is quite efficient when the amount of adaptation data is limited.
- MAP has nice asymptotic properties.
- We jointly optimize a direct and indirect adaptation to take advantage of both approaches.
- We use MAP criterion

$$\left(\overline{\boldsymbol{\Lambda}}_{MAP}, \overline{\boldsymbol{\eta}}_{MAP}\right) = \arg\max_{\boldsymbol{\Lambda},\boldsymbol{\eta}} p(\mathbf{X} \mid \boldsymbol{\Lambda}, \boldsymbol{\eta}) p(\boldsymbol{\Lambda}, \boldsymbol{\eta})$$

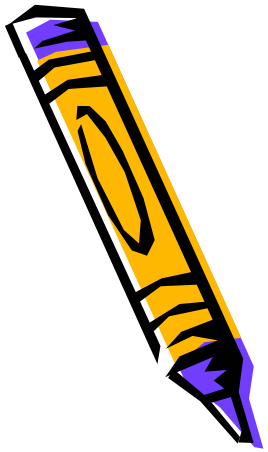
where $p(\mathbf{X} \mid \boldsymbol{\Lambda}, \boldsymbol{\eta})$ denotes the likelihood funtion

Joint MAP and MAPLR





Joint MAP and MAPLR



- Denote $\mathbf{S} = \{s_t\}$: state sequence

  $\mathbf{L} = \{l_t\}$: mixture sequence

  $\mathbf{X} = \{\mathbf{x}_t\}$: observation sequence

- So the criterion can be rewritten as

$$\left(\overline{\boldsymbol{\Lambda}}_{MAP}, \overline{\boldsymbol{\eta}}_{MAP}\right) = \arg\max_{\boldsymbol{\Lambda},\boldsymbol{\eta}} \sum_{\mathbf{S}} \sum_{\mathbf{L}} p(\mathbf{X}, \mathbf{S}, \mathbf{L} \mid \boldsymbol{\Lambda}, \boldsymbol{\eta}) p(\boldsymbol{\Lambda}, \boldsymbol{\eta})$$

- Define the auxiliary function

$$Q\left(\overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\eta}} \mid \boldsymbol{\Lambda}, \boldsymbol{\eta}\right) = E\left[\log p(\mathbf{X}, \mathbf{S}, \mathbf{L} \mid \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\eta}}) + \log p(\overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\eta}}) \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}\right]$$

- Assume that Λ and η are independent, then

$$p(\overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\eta}}) = p(\overline{\boldsymbol{\Lambda}}) p(\overline{\boldsymbol{\eta}})$$

$$\therefore Q\left(\overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\eta}} \mid \boldsymbol{\Lambda}, \boldsymbol{\eta}\right) = E\left[\log p(\mathbf{X}, \mathbf{S}, \mathbf{L} \mid \overline{\boldsymbol{\Lambda}}, \overline{\boldsymbol{\eta}}) + \log p(\overline{\boldsymbol{\Lambda}}) + \log p(\overline{\boldsymbol{\eta}}) \mid \mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\eta}\right]$$

Joint MAP and MAPLR

- The M-step can then be sub-divided to jointly optimize Λ and η via an iterative approach leading to the following M-step:
  - 1.Initialization: $\Lambda^{(0)} = \Lambda \;, i = 1$
  - 2.Setp i:
    - Maximize Q function w.r.t $\overline{\eta}^{(i)}$ given $\Lambda^{(i-1)}$

$$\overline{\eta}^{(i)} = \arg\max_{\eta} E\left[\log p(\mathbf{X},\mathbf{S},\mathbf{L} \mid \Lambda^{(i-1)}, \overline{\eta}^{(i)}) + \log p(\overline{\eta}^{(i)}) \mid \mathbf{X}, \Lambda^{(i-1)}\right]$$

    - Transform the model using $\eta^{(i)}$: $\widetilde{\Lambda}^{(i)} = F_{\overline{\eta}^{(i)}}(\Lambda^{(i-1)})$
    - Maximize Q-function w.r.t $\overline{\Lambda}^{(i)}$:

$$\overline{\Lambda}^{(i)} = \arg\max_{\Lambda} E\left[\log p(\mathbf{X},\mathbf{S},\mathbf{L} \mid \overline{\Lambda}^{(i)}) + \log p(\overline{\Lambda}^{(i)}) \mid \mathbf{X}, \widetilde{\Lambda}^{(i)}\right]$$

  - 3.If a fixed point is not reached, $i++$ goto step 2
  - 4.Termination: $\Lambda = \Lambda^{(i)}, \eta = \eta^{(i)}$

Joint MAP and MAPLR

$\Lambda^{(0)} = \Lambda$

(initial model)

MAPLR find $\boldsymbol{\eta}^{(1)}$

$\tilde{\Lambda}^{(1)} = F_{\boldsymbol{\eta}^{(1)}}(\Lambda^{(0)})$

(intermediate model)

MAP adapt $\tilde{\Lambda}^{(1)}$

$\Lambda^{(1)}$

(final model in step 1)

Step 1

$\Lambda^{(1)}$

MAPLR find $\boldsymbol{\eta}^{(2)}$

$\tilde{\Lambda}^{(2)} = F_{\boldsymbol{\eta}^{(2)}}(\Lambda^{(1)})$

(intermediate model)

MAP adapt $\tilde{\Lambda}^{(2)}$

$\Lambda^{(2)}$

(final model in step 2)

Step 2

$\Lambda^{(2)}$

MAPLR find $\boldsymbol{\eta}^{(3)}$

$\tilde{\Lambda}^{(3)} = F_{\boldsymbol{\eta}^{(3)}}(\Lambda^{(2)})$

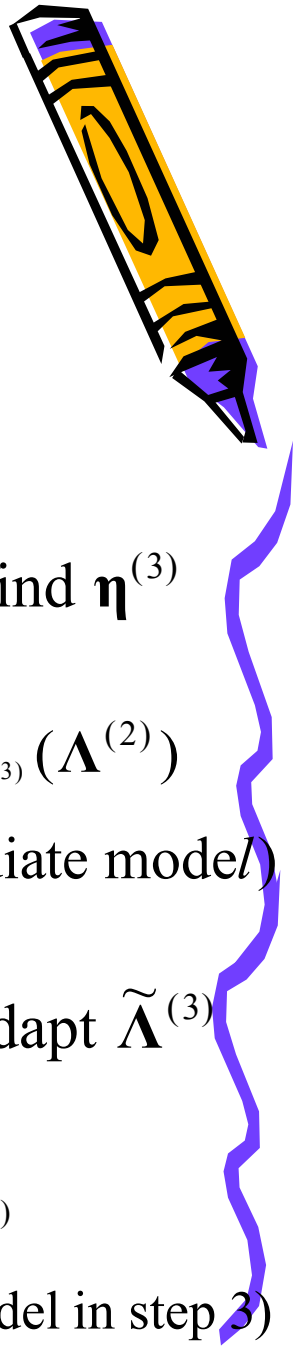
(intermediate model)

MAP adapt $\tilde{\Lambda}^{(3)}$

$\Lambda^{(3)}$

(final model in step 3)

Step 3



Joint MAP and MAPLR

- MAP estimation of the transformation matrix

  →MAPLR estimation

$$\overline{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{X} \mid \mathbf{W}, \boldsymbol{\Lambda}) p(\mathbf{W})$$

- MAP estimation of the HMM parameters

  →MAP estimation

$$\overline{\boldsymbol{\Lambda}} = \arg\max_{\tilde{\boldsymbol{\Lambda}}} p(\mathbf{X} \mid \mathbf{W}, \tilde{\boldsymbol{\Lambda}}) p(\tilde{\boldsymbol{\Lambda}})$$

Joint MAP and MAPLR

MAPLR

$$Q(\tilde{\mathbf{\Lambda}}_c \mid \mathbf{\Lambda}_c)$$

$$= \sum_{\mathbf{S}} \sum_{\mathbf{L}} p(\mathbf{S}, \mathbf{L} \mid \mathbf{X}, \mathbf{\Lambda}_c) \sum_{t=1}^{T} \left[\log a_{s_{t-1}, s_t} + \log w_{s_t, l_t} + \log p(\mathbf{x}_t \mid \overline{\boldsymbol{\eta}}_c, \boldsymbol{\mu}_{s_t, l_t}, \mathbf{R}_{s_t, l_t}) \right]$$

$$+ \log p(\overline{\mathbf{\Lambda}}_c) + \Psi$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \log p(\mathbf{x}_t \mid \overline{\boldsymbol{\eta}}_c, \boldsymbol{\mu}_{j,k}, \mathbf{R}_{j,k}) \right] + \log p(\overline{\mathbf{\Lambda}}_c) + \Psi'$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \left[-\frac{1}{2} tr\left((\mathbf{x}_t - \overline{\mathbf{W}}_c \boldsymbol{\xi}_{j,k})(\mathbf{x}_t - \overline{\mathbf{W}}_c \boldsymbol{\xi}_{j,k})^T \mathbf{R}_{j,k} \right) \right] \right]$$

$$+ \log p(\overline{\mathbf{W}}_c) + \Psi''$$

- Differentiating w.r.t $\overline{\mathbf{W}}_c$ and set it to zero

  then we can obtain the estimate

Joint MAP and MAPLR

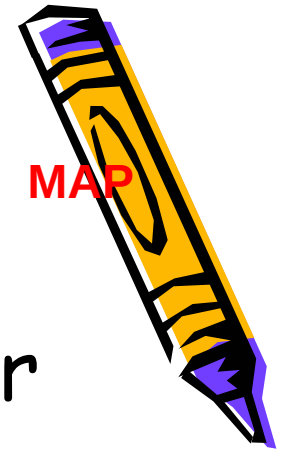
- How to choose $p(\tilde{\Lambda})$
  - It might happen that $p(\tilde{\Lambda})$ and $p(\Lambda)$ no longer belongs to the same family of distributions.
  - We only consider the reestimation of the mean vector $\bar{\boldsymbol{\mu}}$ and assume that the conjugate prior density of $\boldsymbol{\mu}$ is a Normal distribution

$$p(\boldsymbol{\mu}_{j,k}) = p(\boldsymbol{\mu}_{j,k} \mid \mathbf{m}_{j,k}, \boldsymbol{\tau}_{j,k}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\mu}_{j,k} - \mathbf{m}_{j,k})^T \boldsymbol{\tau}_{j,k} (\boldsymbol{\mu}_{j,k} - \mathbf{m}_{j,k})\right]$$

where $\mathbf{m}_{j,k}$ and $\boldsymbol{\tau}_{j,k}$ are all hyperparameters

$\mathbf{m}_{j,k}$ is the mean and $\boldsymbol{\tau}_{j,k}$ is the precision matrix

Joint MAP and MAPLR



- Then it is easy to show that the prior distribution of $\tilde{\boldsymbol{\mu}}_{j,k}$ is also a Normal distribution

$$p(\tilde{\boldsymbol{\mu}}_{j,k} \mid \tilde{\mathbf{m}}_{j,k}, \tilde{\boldsymbol{\tau}}_{j,k}) \propto \exp\left\{-\frac{1}{2}\left(\tilde{\boldsymbol{\mu}}_{j,k} - \tilde{\mathbf{m}}_{j,k}\right)^T \tilde{\boldsymbol{\tau}}_{j,k}\left(\tilde{\boldsymbol{\mu}}_{j,k} - \tilde{\mathbf{m}}_{j,k}\right)\right\}$$

$$\text{where} \quad \tilde{\mathbf{m}}_{j,k} = \mathbf{A}\mathbf{m}_{j,k} + \mathbf{b} \quad \text{and} \quad \tilde{\boldsymbol{\tau}}_{j,k} = (\mathbf{A}^{-1})^T \boldsymbol{\tau}_{j,k} \mathbf{A}^{-1}$$

Joint MAP and MAPLR

$$Q(\overline{\mathbf{\Lambda}}_c \mid \widetilde{\mathbf{\Lambda}}_c)$$

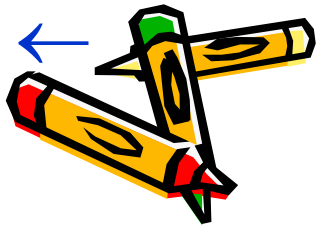
$$= \sum_{\mathbf{S}} \sum_{\mathbf{L}} p(\mathbf{S}, \mathbf{L} \mid \mathbf{X}, \widetilde{\mathbf{\Lambda}}) \sum_{t=1}^{T} \left[\log a_{s_{t-1}, s_t} + \log w_{s_t, l_t} + \log p(\mathbf{x}_t \mid \overline{\boldsymbol{\mu}}_{s_t, l_t}, \mathbf{R}_{s_t, l_t}) \right]$$

$$+ \log p(\overline{\mathbf{\Lambda}}_c)$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \log p(\mathbf{x}_t \mid \overline{\boldsymbol{\mu}}_{j,k}, \mathbf{R}_{j,k}) \right] + \log p(\overline{\boldsymbol{\mu}}_{j,k}) + \Psi$$

$$= \left[\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_t(j,k) \log p(\mathbf{x}_t \mid \overline{\boldsymbol{\mu}}_{j,k}, \mathbf{R}_{j,k}) \right]$$

$$- \frac{1}{2} \left(\overline{\boldsymbol{\mu}}_{j,k} - \widetilde{\mathbf{m}}_{j,k} \right)^T \widetilde{\boldsymbol{\tau}}_{j,k} \left(\overline{\boldsymbol{\mu}}_{j,k} - \widetilde{\mathbf{m}}_{j,k} \right) + \Psi'$$



Joint MAP and MAPLR

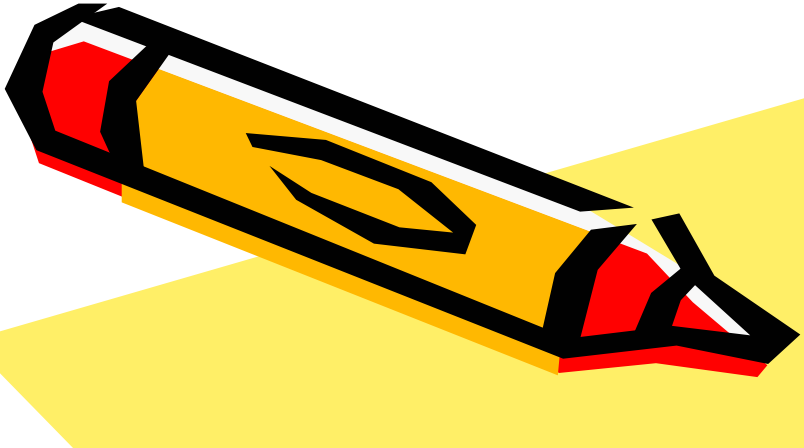
- Differentiate w.r.t $\bar{\boldsymbol{\mu}}_{j,k}$ and set it to zero, we can obtain the estimate [MAP]

$$\bar{\boldsymbol{\mu}}_{j,k} = \left[\tilde{\boldsymbol{\tau}}_{j,k} + \sum_{t=1}^{T} \gamma_t(j,k) \mathbf{R}_{j,k} \right]^{-1} \left(\tilde{\boldsymbol{\tau}}_{j,k} \tilde{\mathbf{m}}_{j,k} + \sum_{t=1}^{T} \gamma_t(j,k) \mathbf{R}_{j,k} \mathbf{x}_t \right)$$

$$\because \begin{pmatrix} \tilde{\mathbf{m}}_{j,k} = \mathbf{A}\mathbf{m}_{j,k} + \mathbf{b} \\ \tilde{\boldsymbol{\tau}}_{j,k} = \mathbf{A}^{-T} \boldsymbol{\tau}_{j,k} \mathbf{A}^{-1} \end{pmatrix}$$

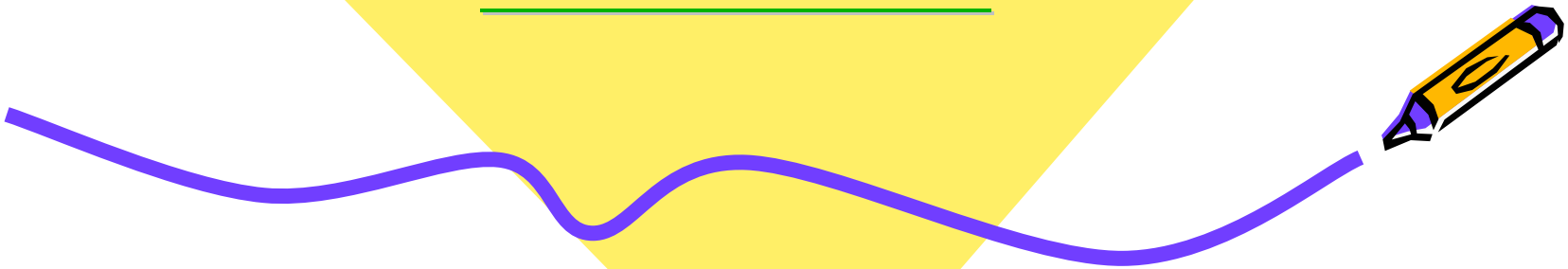
$$\tilde{\boldsymbol{\mu}}_{j,k} = \left[\left(\mathbf{A}^{-1} \right)^T \boldsymbol{\tau}_{j,k} \mathbf{A}^{-1} + \sum_{t=1}^{T} \gamma_t(j,k) \mathbf{R}_{j,k} \right]^{-1}$$

$$\left((\mathbf{A}^{-1})^T \boldsymbol{\tau}_{j,k} \mathbf{A}^{-1} (\mathbf{A}\mathbf{m}_{j,k} + \mathbf{b}) + \sum_{t=1}^{T} \gamma_t(j,k) \mathbf{R}_{j,k} \mathbf{x}_t \right)$$

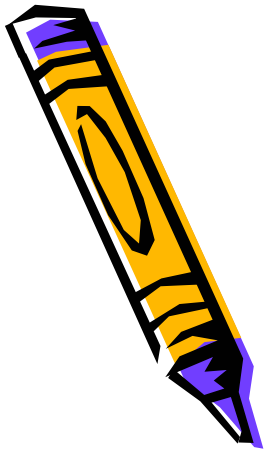


Appendix

Matrix Calculus



Appendix-Matrix Calculus(1)



- Notation: $\mathbf{a}$, $\mathbf{x}$, $\mathbf{b}$ are vectors of dimension n

  $\mathbf{a}_i$, $\mathbf{x}_i$, $\mathbf{b}_i$ be the ith element in vector $\mathbf{a}$, $\mathbf{x}$ and $\mathbf{b}$

  $\mathbf{B}$, $\mathbf{X}$ is a matrix of dimension $n \times n$

  $\mathbf{B}_{ij}$, $\mathbf{X}_{ij}$ be the ith row , jth column element in matrix $\mathbf{B}$ and $\mathbf{X}$

  $\frac{d\mathbf{y}}{dx}$ is a vector whose ith element is $\frac{d\mathbf{y}_i}{dx}$

  $\frac{dy}{d\mathbf{x}}$ is a vector whose ith element is $\frac{dy}{d\mathbf{x}_i}$

  $\frac{d\mathbf{y}^{\mathbf{T}}}{d\mathbf{x}}$ is a matrix whose (i, j) element is $\frac{d\mathbf{y}_j}{d\mathbf{x}_i}$

  $\frac{d\mathbf{X}}{dx}$ is a matrix whose (i, j) element is $\frac{d\mathbf{X}_{ij}}{dx}$

  $\frac{dy}{d\mathbf{X}}$ is a matrix whose (i, j) element is $\frac{dy}{d\mathbf{X}_{ij}}$

Appendix-Matrix Calculus(2)

- Properties 1: $\frac{d\,(\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{b})}{d\mathbf{X}} = \mathbf{a}\mathbf{b}^{\mathrm{T}}$
  - proof

$$\because \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \sum_{i=1}^{n}\left[\mathbf{b}_i \sum_{j=1}^{n}\left(\mathbf{a}_j \mathbf{X}_{ji}\right)\right] \quad \therefore \frac{\partial(\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{b})}{\partial \mathbf{X}_{kt}} = \frac{\partial\left\{\sum_{i=1}^{n}\left[\mathbf{b}_i \sum_{j=1}^{n}\left(\mathbf{a}_j \mathbf{X}_{ji}\right)\right]\right\}}{\partial \mathbf{X}_{kt}} = \mathbf{a}_k \mathbf{b}_t$$

$$\therefore \frac{d(\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{b})}{d\mathbf{X}} = \mathbf{a}\mathbf{b}^{\mathrm{T}}$$

- Properties 1— Extension: $\frac{d\,(\mathbf{a}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{b})}{d\mathbf{X}} = \mathbf{b}\mathbf{a}^{\mathrm{T}}$
  - proof

$$\because \mathbf{a}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{b} = \sum_{i=1}^{n}\left[\mathbf{b}_i \sum_{j=1}^{n}\left(\mathbf{a}_j \mathbf{X}_{ij}\right)\right] \quad \therefore \frac{\partial(\mathbf{a}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{b})}{\partial \mathbf{X}_{kt}} = \frac{\partial\left\{\sum_{i=1}^{n}\left[\mathbf{b}_i \sum_{j=1}^{n}\left(\mathbf{a}_j \mathbf{X}_{ij}\right)\right]\right\}}{\partial \mathbf{X}_{kt}} = \mathbf{b}_k \mathbf{a}_t$$

$$\therefore \frac{d(\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{b})}{d\mathbf{X}} = \mathbf{b}\mathbf{a}^{\mathrm{T}}$$

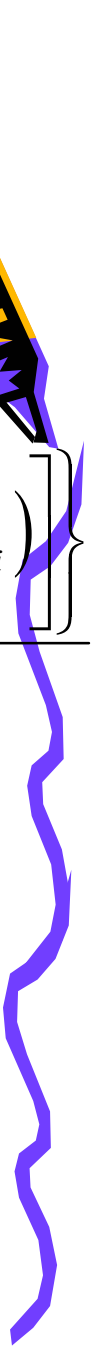
Appendix-Matrix Calculus(3)

- Properties 2: $\frac{d\,(\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}{d\mathbf{x}} = \left(\mathbf{C}+\mathbf{C}^{\mathrm{T}}\right)\mathbf{x}$
  - proof

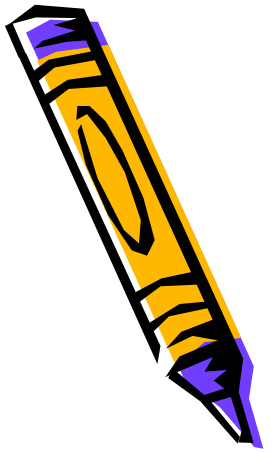
$$\because \mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} = \sum_{i=1}^{n}\left[\mathbf{x}_i \sum_{j=1}^{n}\left(\mathbf{x}_j \mathbf{C}_{ji}\right)\right] \quad \therefore \frac{\partial(\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}{\partial \mathbf{X}_k} = \frac{\partial\left\{\sum_{i=1}^{n}\left[\mathbf{x}_i \sum_{j=1}^{n}\left(\mathbf{x}_j \mathbf{C}_{ji}\right)\right]\right\}}{\partial \mathbf{X}_k}$$

$$= \begin{cases} 2\mathbf{x}_k \mathbf{C}_{kk} & i=j=k \\ \sum_{t\neq k} \mathbf{C}_{tk}\mathbf{x}_t & i=k\neq j \\ \sum_{t\neq k} \mathbf{C}_{kt}\mathbf{x}_t & j=k\neq i \end{cases} = \sum_{t=1}^{n} \mathbf{C}_{kt}\mathbf{x}_t + \sum_{t=1}^{n} \mathbf{C}_{tk}\mathbf{x}_t$$

$$\Rightarrow \frac{d(\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}{d\mathbf{X}} = \left(\mathbf{C}+\mathbf{C}^{\mathrm{T}}\right)\mathbf{x}$$



Appendix-Matrix Calculus(4)



- Properties 3: $\frac{d\left[tr\left(\mathbf{B}\mathbf{X}^{\mathbf{T}}\right)\right]}{d\mathbf{X}} = \mathbf{B}$
  - proof

$$\because \left(\mathbf{B}\mathbf{X}^{\mathbf{T}}\right)_{ij} = \sum_{k=1}^{n} \mathbf{B}_{ik}\mathbf{X}_{jk}$$

$$\therefore tr\left(\mathbf{B}\mathbf{X}^{\mathbf{T}}\right) = \sum_{k=1}^{n}\sum_{t=1}^{n} \mathbf{B}_{kt}\mathbf{X}_{kt}$$

$$\Rightarrow \frac{\partial\left[tr\left(\mathbf{B}\mathbf{X}^{\mathbf{T}}\right)\right]}{\partial \mathbf{X}_{ij}} = \sum_{k=1}^{n}\sum_{t=1}^{n} \mathbf{B}_{kt} \cdot 1(k = i\ ,\ t = j) = \mathbf{B}_{ij}$$

$$\therefore \frac{d\left[tr\left(\mathbf{B}\mathbf{X}^{\mathbf{T}}\right)\right]}{d\mathbf{X}} = \mathbf{B}$$

Appendix-Matrix Calculus(5)

- Properties 4: $\frac{d[\det(\mathbf{X})]}{d\mathbf{X}} = \det(\mathbf{X}) \cdot \mathbf{X}^{-\mathbf{T}}$
  - proof

$$\because \frac{d[\det(\mathbf{X})]}{d\mathbf{X}} = [adj(\mathbf{X})]^T$$

$$and \ \mathbf{X} \cdot adj(\mathbf{X}) = \det(\mathbf{X}) \cdot \mathbf{I}$$

$$\therefore \frac{d[\det(\mathbf{X})]}{d\mathbf{X}} = [adj(\mathbf{X})]^T = [\mathbf{X}^{-1} \cdot \det(\mathbf{X}) \cdot \mathbf{I}]^T = \det(\mathbf{X}) \cdot \mathbf{X}^{-T}$$

Appendix-Matrix Calculus(6)

- Properties 5: $\frac{d\left[\mathbf{a}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}\mathbf{b}\right]}{d\mathbf{X}}=\mathbf{C}^{\mathrm{T}}\mathbf{X}\mathbf{a}\mathbf{b}^{\mathrm{T}}+\mathbf{C}\mathbf{X}\mathbf{b}\mathbf{a}^{\mathrm{T}}$
  - proof

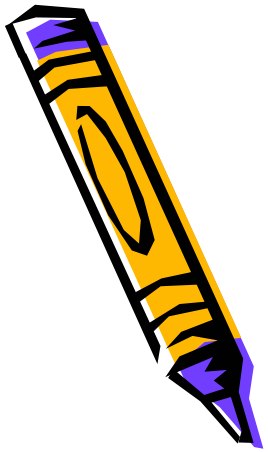
$$Let\ \mathbf{u}=\mathbf{X}\mathbf{a}\ \ and\ \ \mathbf{v}=\mathbf{X}\mathbf{b}\ \ \therefore \mathbf{u}_i=\sum_{k=1}^{n}\mathbf{X}_{ik}\mathbf{a}_k\quad and\quad \mathbf{v}_i=\sum_{k=1}^{n}\mathbf{X}_{ik}\mathbf{b}_k$$

$$\mathbf{a}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}\mathbf{b}=\mathbf{u}^{\mathrm{T}}\mathbf{C}\mathbf{v}=\sum_{i=1}^{n}\left(\mathbf{v}_i\sum_{j=1}^{n}\mathbf{u}_j\mathbf{C}_{ji}\right)=\sum_{i=1}^{n}\left(\left(\sum_{k=1}^{n}\mathbf{X}_{ik}\mathbf{b}_k\right)\sum_{j=1}^{n}\left(\sum_{k=1}^{n}\mathbf{X}_{jk}\mathbf{a}_k\right)\mathbf{C}_{ji}\right)$$

$$\therefore \frac{\partial\left[\mathbf{a}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}\mathbf{b}\right]}{\partial\mathbf{X}_{uv}}\overset{k=v}{=}\begin{cases}\dfrac{d\left(\mathbf{X}_{uv}\mathbf{b}_v\left(\mathbf{X}_{uv}\mathbf{a}_v\right)\mathbf{C}_{ju}\right)}{d\mathbf{X}_{uv}} & i=u,\ j=u\\ \dfrac{\partial\left(\mathbf{X}_{uv}\mathbf{b}_v\left(\sum_{j\neq u}^{n}\left(\mathbf{X}_{jv}\mathbf{a}_v\right)\mathbf{C}_{ju}\right)\right)}{\partial X_{uv}} & i=u,\ j\neq u\\ \dfrac{\partial\left(\sum_{i\neq u}^{n}\left(\mathbf{X}_{iv}\mathbf{b}_v\mathbf{X}_{uv}\mathbf{a}_v\mathbf{C}_{ui}\right)\right)}{\partial\mathbf{X}_{uv}} & i\neq u,\ j=u\end{cases}$$

$$=\begin{cases}2\mathbf{a}_v\mathbf{b}_v\mathbf{C}_{ju}\mathbf{X}_{uv} & i=u,\ j=u\\ b_v\sum_{j\neq u}^{n}\left(\mathbf{X}_{jv}\mathbf{a}_v\right)\mathbf{C}_{ju} & i=u,\ j\neq u\\ \sum_{i\neq u}^{n}\left(\mathbf{X}_{iv}\mathbf{b}_v\mathbf{a}_v\mathbf{C}_{ui}\right) & i\neq u,\ j=u\end{cases}=\mathbf{a}_v\mathbf{b}_v\sum_{j=1}^{n}\left(\mathbf{C}_{ju}\mathbf{X}_{jv}\right)+\mathbf{a}_v\mathbf{b}_v\sum_{i=1}^{n}\left(\mathbf{C}_{ui}\mathbf{X}_{iv}\right)=\mathbf{C}^{\mathrm{T}}\mathbf{X}\mathbf{a}\mathbf{b}^{\mathrm{T}}+\mathbf{C}\mathbf{X}\mathbf{b}\mathbf{a}^{\mathrm{T}}$$

Appendix-Matrix Calculus(7)



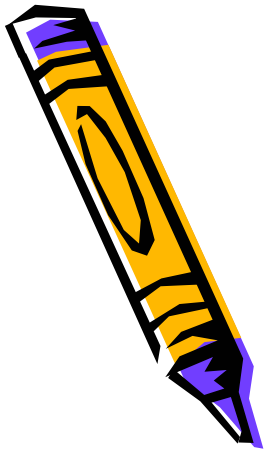
- Properties 6: $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = tr\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{A}\right)$
  - proof

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \left[\sum_{j=1}^{n}\left(\sum_{k=1}^{n}\mathbf{x}_k\mathbf{A}_{kj}\right)\mathbf{x}_j\right] = \sum_{j=1}^{n}\sum_{k=1}^{n}\mathbf{x}_j\mathbf{x}_k\mathbf{A}_{kj}$$

$$tr(\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{A}) = \sum_{i=1}^{n}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{A}\right)_{ii} = \sum_{i=1}^{n}\left(\sum_{k=1}^{n}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)_{ik}\mathbf{A}_{ki}\right)_{ii} = \sum_{i=1}^{n}\sum_{k=1}^{n}\mathbf{x}_i\mathbf{x}_k\mathbf{A}_{ki}$$

$$\therefore \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = tr(\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{A})$$

Appendix-Matrix Calculus(8)



- Some other properties :

  - $\frac{d(\mathbf{x}^{\mathrm{T}}\mathbf{a})}{d\mathbf{x}} = \frac{d(\mathbf{a}^{\mathrm{T}}\mathbf{x})}{d\mathbf{x}} = \mathbf{a}$
  - $\frac{d\left(tr(\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}\mathbf{B})\right)}{d\mathbf{X}} = \mathbf{A}^{T}\mathbf{X}\mathbf{B}^{T} + \mathbf{A}\mathbf{X}\mathbf{B}$
  - $\frac{d\left(\mathbf{a}^{T}\mathbf{X}^{-1}\mathbf{b}\right)}{d\mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^{T}\mathbf{X}^{-T}$
  - ?? $\frac{d\left[tr\left(\mathbf{X}^{-1}\mathbf{C}^{-1}\mathbf{X}^{-T}\mathbf{A}\right)\right]}{d\mathbf{X}} = -\mathbf{X}^{-T}(\mathbf{A}+\mathbf{A}^{T})(\mathbf{X}^{-1}\mathbf{C}^{-1}\mathbf{X}^{-T})$

    where $\mathbf{C}$ is symmetric