Joint Factor Analysis for Speaker Verification

Mengke HU

ASPITRG Group, ECE Department
Drexel University

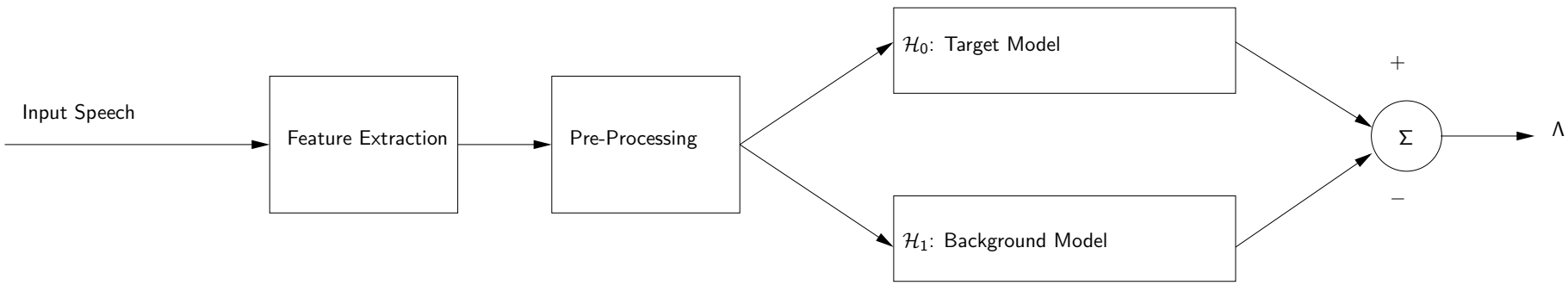
mengke.hu@gmail.com

October 12, 2012

Outline

1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

Baseline System



Baseline System

1. Given a speech segment X, we test 2 hypotheses:

   $\mathcal{H}_0 : X$ is from claimed target speaker S (GMM)

   $\mathcal{H}_1 : X$ is not from speaker S , it is from the background (UBM).

2. Decision Rule

$$Score = \log \frac{p(\mathbf{X}|TargetModel)}{p(\mathbf{X}|UBM)} \underset{H_1}{\overset{H_0}{\gtrless}} Threshold$$

Note: $Score = \log p(\mathbf{X}|TargetModel) - \log p(\mathbf{X}|UBM)$

Baseline Experiment

1. Feature Extraction (MFCC)
2. Train the UBM model
3. Obtain Adapted GMM model for target speaker model
4. Test trials against 2 hypotheses
5. Scoring
6. DET(Detection Error Tradeoff) curve false accept VS. false reject

Problem

How to cancel the channel effect?

Outline

1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

Session Variation

- Inter-Speaker Variation: Two utterances are from different speaker

Session Variation

- Inter-Speaker Variation: Two utterances are from different speaker
- Inter-Session Variation: Two utterances are from the same speaker
  - Channel effects: Utterances are recorded from different channels
  - Intra-Speaker Variation: Utterances varies with speaker's health or emotional state etc. .

Outline

1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

Gaussian Mixture Model Review

- Recall GMM:

$$p(\mathbf{s}|\mathbf{z}_s) = \prod_{i=1}^{M_s} \mathcal{N}(\mathbf{s}; \mu_i^s, \Sigma_i^s)^{z_{s,i}}$$

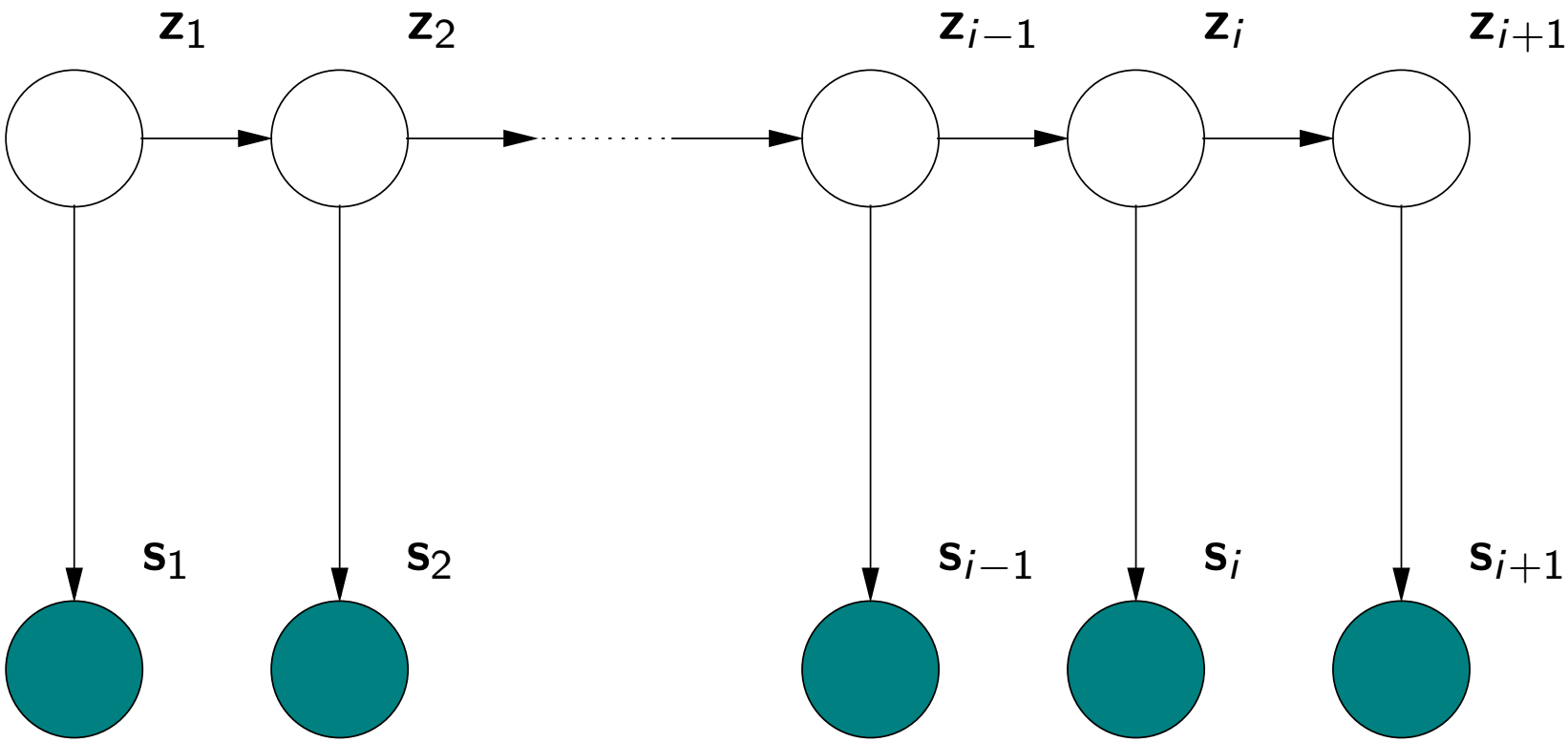
$$p(\mathbf{s}) = \sum_{\mathbf{z}_s} p(\mathbf{z}_s) p(\mathbf{s}|\mathbf{z}_s) = \sum_{i=1}^{M_s} \pi_i^{\mathbf{s}} \mathcal{N}(\mathbf{s}; \mu_i^s, \Sigma_i^s)^{z_{s,i}}$$

$$p(\mathbf{z}_s) = \prod_{i=1}^{M_s} \pi_i^{z_{s,i}}$$

- $\mathbf{z}_s$ is a hidden variable indicate which Gaussian mixture component is active.
- Remark: $\{z_{s,i}\}_{i \in 1 \ldots M_s}$ are independent

Hidden Markov Model

Graphic Model



$$P(\mathbf{z}_n|\mathbf{z}_{n-1},\ldots,\mathbf{z}_1) = P(\mathbf{z}_n|\mathbf{z}_{n-1})$$

HMM is often used in speaker recognition.

Hidden Markov Model

We have the following joint probability:

$$p(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{z}_1|\pi)\left(\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A})\right)\prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m, \phi)$$

where $\mathbf{A}$ is transition probability matrix and

$$p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{n,k}}$$

$$p(\mathbf{z}_1|\pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}}, \quad \sum_k \pi_k = 1$$

$$p(\mathbf{x}_n|\mathbf{z}_n, \phi) = \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}$$

Outline

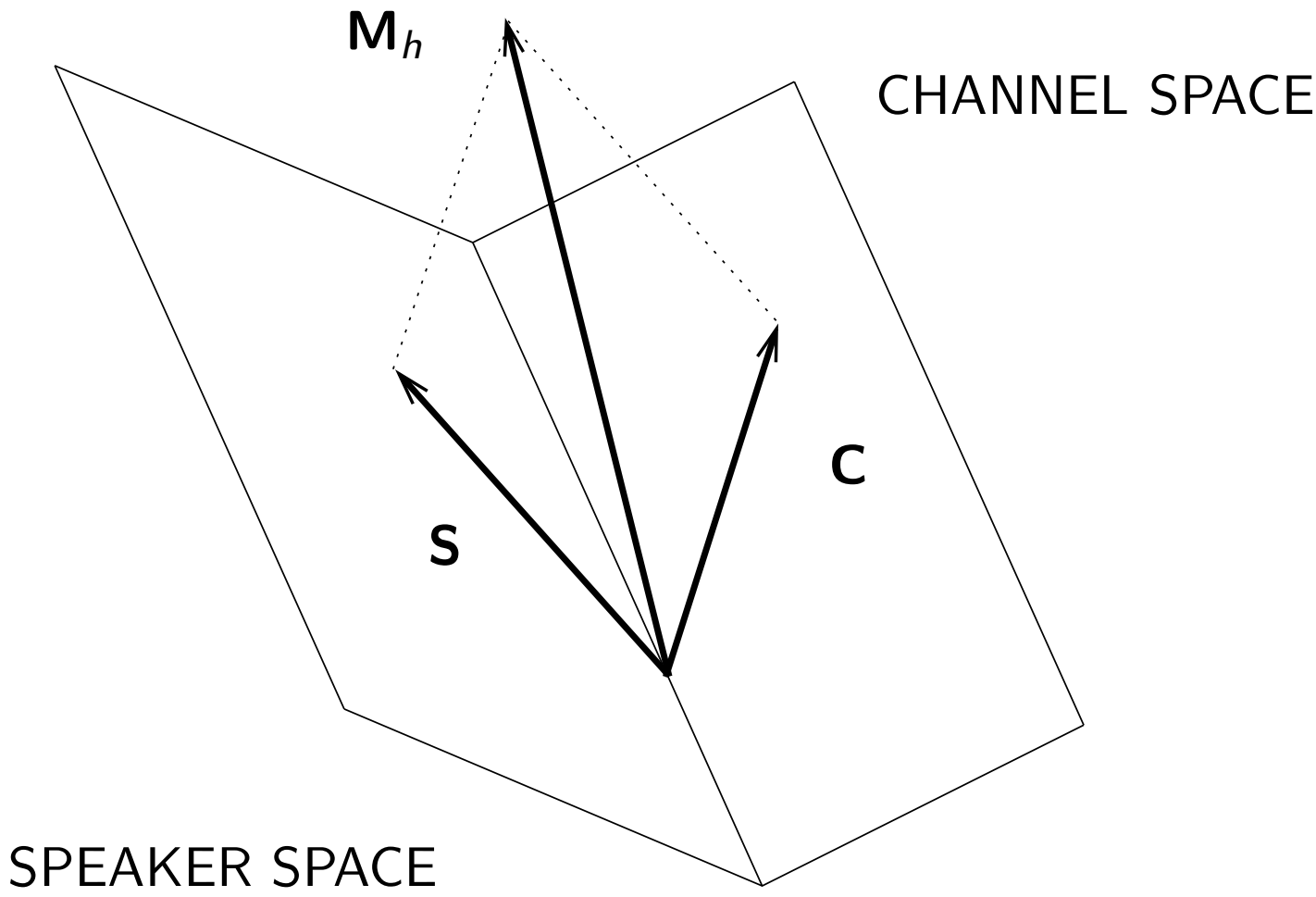
1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

Supervector definition

Given the GMM mean vector $(\mathbf{m}_c)_{F\times 1}$, $c \in \{1, \ldots, C\}$, C is the total number of mixture components, F is the dimension of feature vector
Supervector is:

$$\mathbf{m}_{CF\times 1} = (\mathbf{m}_1^T, \ldots, \mathbf{m}_c^T)$$

Speaker and Channel Dependent Supervector $\mathbf{M}_h$



$\mathbf{M}_h$ is the speaker-and channel-dependent supervector

Notations

- S: speaker ID
- **Speaker factors**: components of $\mathbf{y}(s)$
- **Channel factors**: components of $\mathbf{x}_h(s)$
- **Speaker space**: affine translating the range of $\mathbf{vv}^*$ by $\mathbf{m}$
- **Channel space**: the range of $\mathbf{uu}^*$
- **Loading matrix for speaker factors and channel factors**: $\mathbf{v}$ and $\mathbf{u}$
- $h = 1, \cdots, H(s)$: one index from set of recordings for a speaker s
- C: total number of mixture components for a fixed GMM structure
- F: dimension of the acoustic feature vectors
- R_C: channel rank
- R_S: speaker rank
- $\Sigma(s)$: given speaker s and recording h, the covariance of the observation from GMM
- $\mathbf{d}$: given speaker s, the covariance of the observation from GMM

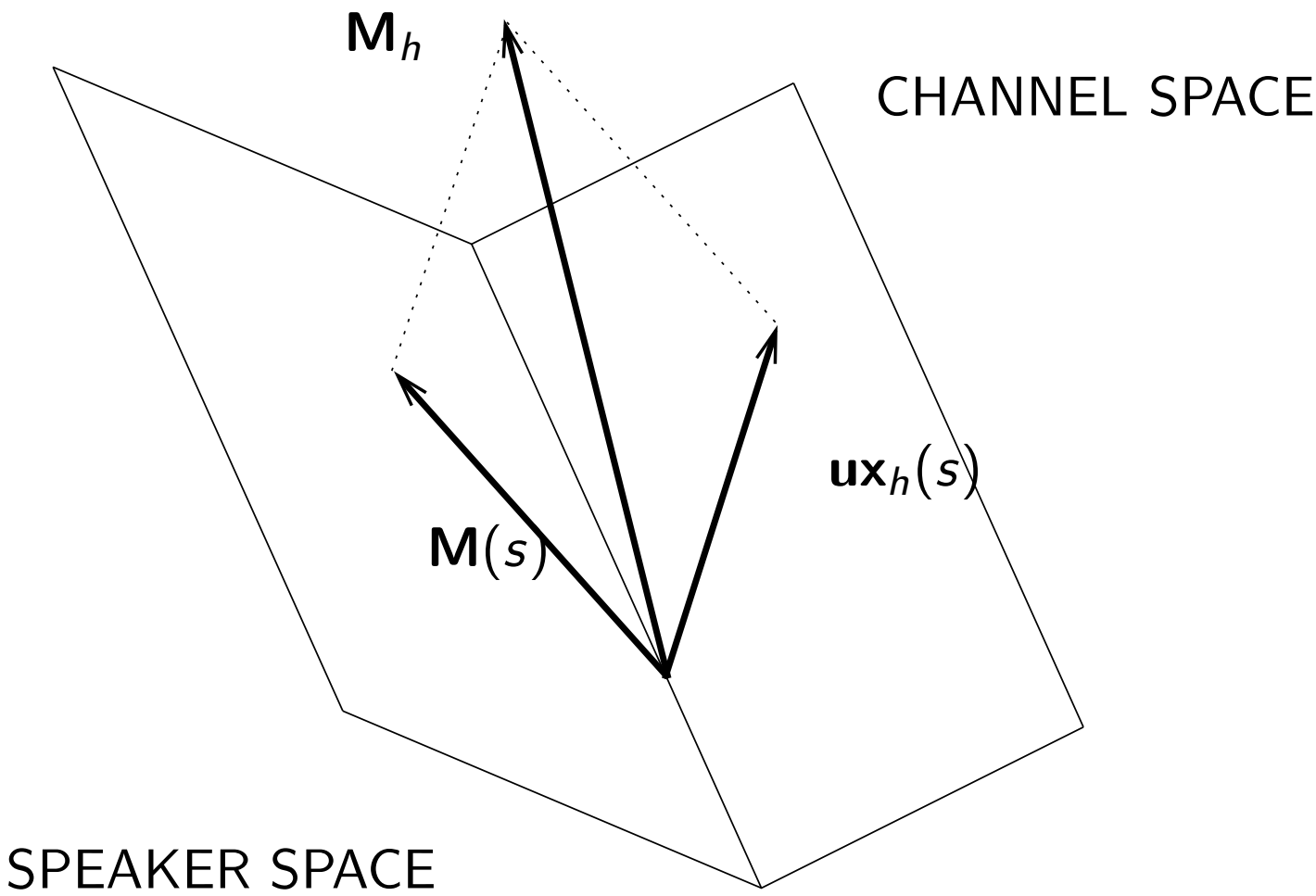
Joint Factor Analysis Model

JFA model

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{vy}(s) + \mathbf{dz}(s)$$
$$\mathbf{M}_h(s) = \mathbf{M}(s) + \mathbf{ux}_h(s)$$

- $\mathbf{m}_{C\times F}$: Given a HMM/GMM structure with C mixture components, we concatenate the mean vectors $m_1, \ldots, m_C$ together then obtain $\mathbf{m}$
- $\mathbf{M}(s)$: single speaker-dependent supervector
- $\mathbf{M}_h(s)$: speaker-and-channel dependent
- $\mathbf{u}$ and $\mathbf{v}$ are speaker independent
- $\mathbf{d}$ is a block diagonal matrix
- $\mathbf{z}$ is normal

JFA model


$\mathbf{M}_h$
CHANNEL SPACE
$\mathbf{ux}_h(s)$
$\mathbf{M}(s)$
SPEAKER SPACE

Problem

- Purpose: estimate the hyperparameters $\mathbf{\Lambda} = (\mathbf{m}, \mathbf{u}, \mathbf{v}, \mathbf{d}, \mathbf{\Sigma})$.
- The number of GMM component is large. $C = 2048$
- The dimension of the feature vector is $F = 39$
- $C \times F = 79872 \Longrightarrow \mathbf{m}_{79872}$ and $\Sigma_{79872 \times 79872}$.

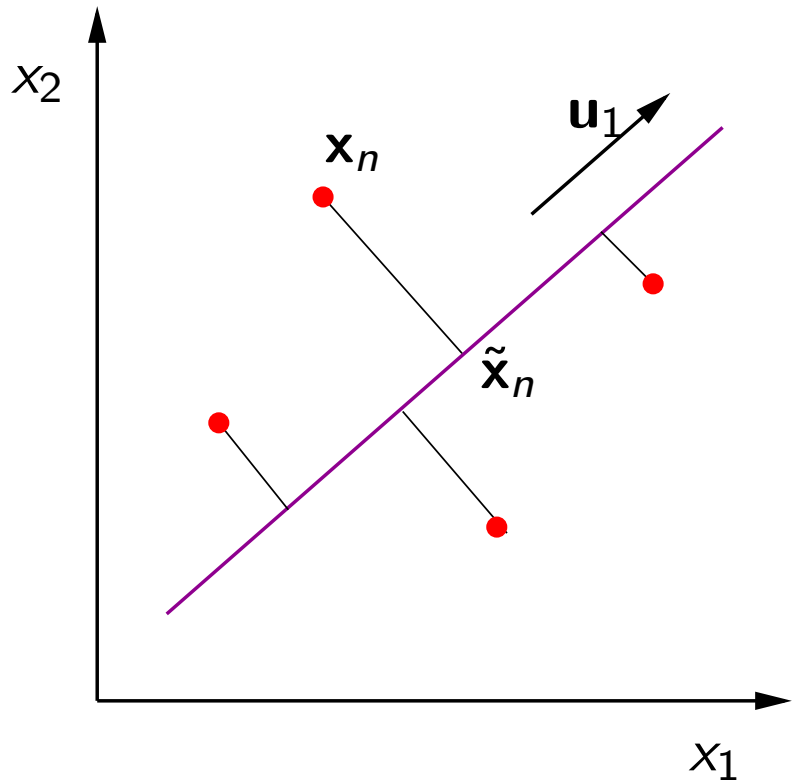
Problem

Σ is very large and it is not full rank, how to estimate?

Outline

1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

Principle Components Analysis Technique



PCA technique is to find a principal subspace (magenta line), s.t. the variance of the projected points ($\tilde{\mathbf{x}}_n$) are maximized.

Maximum Variance Formulation

Find the principle components for principle subspace

- Given feature vectors as observations $\{(\mathbf{x}_n)_{N\times 1}\}$, $n = 1, \ldots, N$, we want to find the principle subspace with M basis, $M < N$

Maximum Variance Formulation

Find the principle components for principle subspace

- Given feature vectors as observations $\{(\mathbf{x}_n)_{N\times 1}\}$, $n = 1, \ldots, N$, we want to find the principle subspace with M basis, $M < N$
- Sample mean $\bar{\mathbf{x}}$ and sample covariance $\mathbf{S}$:

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

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

- Let the M basis for the principle subspace be $\mathbf{u}_1, \ldots, \mathbf{u}_M$ and $\mathbf{u}_i^T\mathbf{u} = 1,\ i \in [M]$

$$\mathbf{P}_{N\times M} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]_{N\times M}$$

Maximum Variance Formulation

- Optimization problem find the 1^{st} principle component

$$\max \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$$
$$\mathbf{u}_1^T \mathbf{u}_1 = 1$$

- By Lagrange methode: maximize:

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)$$

take derivative

$$\frac{\partial \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)}{\partial \mathbf{u}_1} = (\mathbf{S} + \mathbf{S}^T)\mathbf{u}_1 + \lambda_1(-2\mathbf{u}_1)$$
$$= 2\mathbf{S}\mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0 \Longrightarrow \mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

- Solution: λ_1 is the largest eigenvalue of $\mathbf{S}$, the correspond $\mathbf{u}_1$ is the first principle component

Maximum Variance Formulation

- Find M principle components: find M largest eigenvalues, and their correspond eigenvectors $\mathbf{u}_i, \quad i \in [M]$, such that:

$$\mathbf{u}_i^T \mathbf{u}_i = 1, \quad i \in [M]$$
$$\mathbf{u}_i \perp \mathbf{u}_j, \quad i \neq j$$

- Eigen-decomposition $\mathbf{S}$, find the M largest eigenvalues, decreasing sorted. Then, find the $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M$.
- Remark:

$$[\mathbf{u}_1^T, \mathbf{u}_2^T, \ldots, \mathbf{u}_M^T]^T \mathbf{S} [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M] = \mathbf{P}^T \mathbf{S} \mathbf{P}$$

Outline

1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

Probabilistic Model

PCA model

$D >> M$

$$\mathbf{x}_{D\times 1} = \mathbf{W}_{D\times M}\mathbf{z}_{M\times 1} + \mu + \epsilon$$

- $\mathbf{x}$ is D-dimension observation vector
- $\mathbf{z}$ is M-dimension hidden variable

We are given the following probability distributions:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$
$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I})$$

Probabilistic PCA

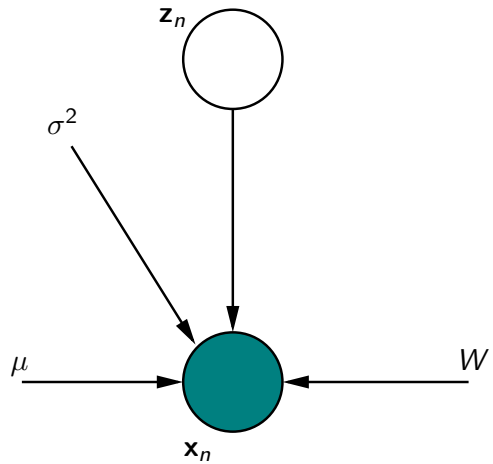
- $p(\mathbf{x})$ is Gaussian

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \mathcal{N}(\mathbf{x}|\mu, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

- mean and variance of $p(\mathbf{x})$

$$\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \mathbb{E}[\mathbf{W}\mathbf{z} + \mu + \epsilon] = \mu \\
cov[\mathbf{x}] &= \mathbb{E}[(\mathbf{W}\mathbf{z} + \epsilon)(\mathbf{W}\mathbf{z} + \epsilon)^T] \\
&= \mathbb{E}[\mathbf{W}\mathbf{z}\mathbf{z}_T\mathbf{W}^T] + \mathbb{E}[\epsilon\epsilon^T] \\
&= \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}
\end{aligned}$$

Probabilistic PCA



- The graph shows for each observation $\mathbf{x}_n$ is associate with a value of latent variable $\mathbf{z}_n$
- $\mathbf{x}_n$ can be obtained by marginalization over $\mathbf{z}_n$.
- Using EM algorithm to estimate the parameters in PCA model (Train PCA model)

Outline

1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

5 steps for JFA Speaker Verification System

1. Train the UBM model
2. Train JFA/PCA model: estimate speaker independent hyperparameters $\Lambda = (\mathbf{m}, \mathbf{u}, \mathbf{v}, \mathbf{d}, \Sigma)$ from a large database in which each speaker is recorded in multiple sessions
3. Adapt Λ from one speaker population to another
4. Enrolling a speaker: estimate the speaker-independent hyperparameters $\Lambda(s) = (\mathbf{m}(s), \mathbf{u}(s), \mathbf{v}(s), \mathbf{d}(s), \Sigma(s))$
5. Test: Given test utterance χ and hypothesized speaker

$$\log \frac{P_{\Lambda(s)}(\mathcal{X})}{P_{\Lambda}(\mathcal{X})}$$

, where $\mathcal{X}$ are observations.

Outline

1. Speaker Verification
   - Baseline System
   - Session Variation
2. Joint Factor Analysis
   - Hidden Markov Model
   - Factor Analysis Model
   - Principal Components Analysis (PCA)
   - Probabilistic PCA
3. JFA for Speaker Verification
   - General Steps
   - Hyperparameter estimation

Train the JFA/PCA model

Estimate Λ

- Training set: several speakers with multiple recordings for each speaker
- Use EM algorithms to estimate Λ
  - Maximum Likelihood Approach (slow)
  - Divergence minimization approach (faster, well initialized)
  - Both algorithm are to fit entire collection of speakers in the training data
- Total likelihood $\prod_s P_\Lambda(\mathcal{X}(s))$, s ranges over the speakers in the training set. It increases from 1 iteration to the next.

Adapt from one speaker population to another

- Adaptation is necessary since data set is limit. For a given speaker, there are at most 2 recordings.
- Keep channel space related hyperparameters fixed ($\mathbf{u}$ and Σ_h), re-estimate only the speaker space hyperparameters ($\mathbf{m}$,$\mathbf{v}$,$\mathbf{d}$).
- Remark: Assume channel space related hyperparameters are speaker independent

Enroll a target speaker

Estimate $\Lambda(s)$

Recall JFA model:

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{vy}(s) + \mathbf{dz}(s)$$
$$\mathbf{M}_h(s) = \mathbf{M}(s) + \mathbf{ux}_h(s)$$

- Calculate the posterior distribution $\mathbf{M}(s)$
- Adjusting the $\Lambda(s)$ to fit this posterior
- Adopt minimum divergence approach

Likelihood Function

Hyperparameters $\mathbf{\Lambda} = (\mathbf{m}, \mathbf{u}, \mathbf{v}, \mathbf{d}, \mathbf{\Sigma})$.

$$P_{\mathbf{\Lambda}}(\underline{\mathcal{X}}(s)) = \int P_{\mathbf{\Lambda}}(\underline{\mathcal{X}}(s)|\underline{\mathbf{X}})\mathcal{N}(\underline{\mathbf{X}}|\underline{\mathbf{0}}, \underline{\mathbf{I}})d\underline{\mathbf{X}}$$

where:

- $\underline{\mathcal{X}}(s)$ (**observable**) is the collections of labeled frames for recording h

$$\underline{\mathcal{X}}(s) = \left(\mathcal{X}_1(s), \ldots, \mathcal{X}_{H(s)}(s)\right)^T$$

- $\underline{\mathbf{X}}(s)$ (**unobservable**) is the vector of hidden variables

$$\underline{\mathbf{X}}(s) = \left(\mathbf{x}_1(s), \ldots, \mathbf{x}_{H(s)}, \mathbf{y}(s), \mathbf{z}(s)\right)^T$$

- $\mathcal{N}(\underline{\mathbf{X}}|\underline{\mathbf{0}}, \underline{\mathbf{I}})$ is the standard Gaussian kernel

$$\mathcal{N}(\underline{\mathbf{X}}|\underline{\mathbf{0}}, \underline{\mathbf{I}}) = \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \mathbf{I}) \ldots \mathcal{N}(\mathbf{x}_{H(s)}|\mathbf{0}, \mathbf{I})\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{I})\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

Likelihood ratio

- Given speech data $\mathcal{X}$ uttered by speaker t
- Test $\mathcal{H}_0 = \{t = s\}$ against $\mathcal{H}_1 = \{t \neq s\}$
-

$$\frac{1}{T} \log \frac{P_{\mathbf{\Lambda}_s}(\mathcal{X})}{P_{\mathbf{\Lambda}}(\mathcal{X})}$$