Factor Analysis and I-Vectors

Man-Wai MAK

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University

enmwmak@polyu.edu.hk
http://www.eie.polyu.edu.hk/∼mwmak

References:

- M.W. Mak and J.T. Chien, *Machine Learning for Speaker Recognition*, Cambridge University Press, 2020.
- S.J.D. Prince,*Computer Vision: Models Learning and Inference*, Cambridge University Press, 2012
- C. Bishop, "*Pattern Recognition and Machine Learning*", Appendix E, Springer, 2006.
- http://www.eie.polyu.edu.hk/∼mwmak/papers/FA-Ivector.pdf

May 31, 2019

Overview

1. Factor Analysis
   - What is Factor Analysis
   - Generative Model
   - EM Formulation

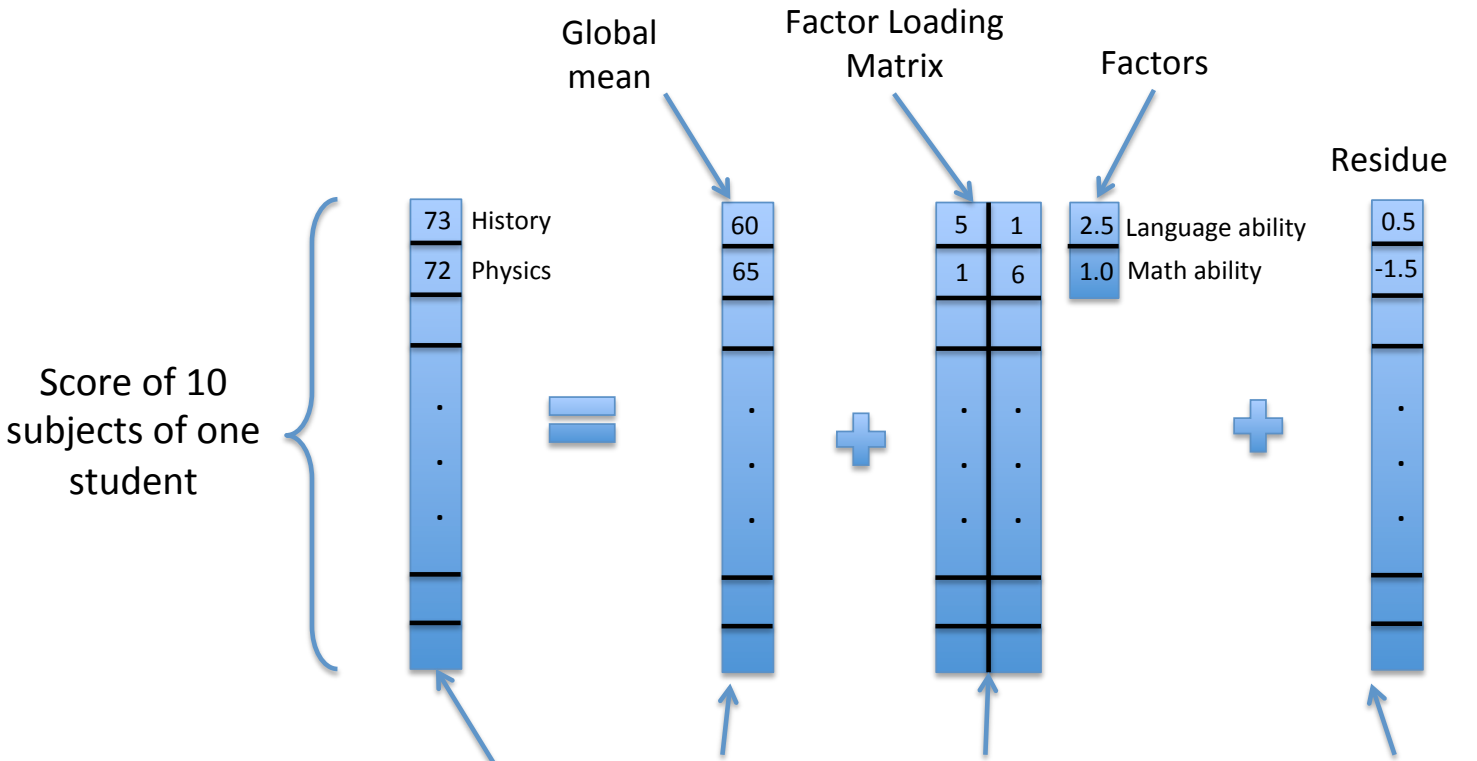
2. I-Vectors
   - Gaussian Mixture Models
   - Factor Analysis Model
   - Applications of I-Vectors

What is Factor Analysis?

- **Wiki**: "Factor analysis is a statistical method used to describe *variability* among observed, correlated variables in terms of a potentially lower number of unobserved variables called *factors*."
- It was originally introduced by psychologists to find the underlying *latent* factors that account for the correlations among a set of observations or measures.
- **Example:** The exam scores of 10 different subjects of 1,000 students may be explained by two latent factors (also called common factors):
  - Language ability
  - Math ability
- Each student has their own values for these two factors across all of the 10 subjects. His/her exam score for each subject is a linear weighted sum of these two factors plus the score mean and an error.

What is Factor Analysis?

- For each subject, all students share the same weights, which are referred to as the factor loadings, for this subject.



$$\mathbf{x}_i = \begin{bmatrix} m_1 \\ \vdots \\ m_{10} \end{bmatrix} + \begin{bmatrix} v_{1,1} & v_{1,2} \\ \vdots & \vdots \\ v_{10,1} & v_{10,2} \end{bmatrix} \begin{bmatrix} z_{i,1} \\ z_{i,2} \end{bmatrix} + \begin{bmatrix} \epsilon_{i,1} \\ \vdots \\ \epsilon_{i,10} \end{bmatrix}$$

$$= \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, 1000,$$

What is Factor Analysis?

- Denote $\mathbf{x}_i = [x_{i,1} \ \cdots \ x_{i,10}]^{\mathsf{T}}$ and $\mathbf{z}_i = [z_{i,1} \ \ z_{i,2}]^{\mathsf{T}}$ as the exam scores and common factors of the i-th student, respectively. Also, denote $\mathbf{m} = [m_1 \ \cdots \ m_{10}]^{\mathsf{T}}$ as the mean exam scores of these 10 subjects. Then, we have

$$x_{i,j} = m_j + v_{j,1} z_{i,1} + v_{j,2} z_{i,2} + \epsilon_{i,j}, \quad i = 1, \ldots, 1000 \text{ and } j = 1, \ldots, 10, \tag{1}$$

  where $v_{j,1}$ and $v_{j,2}$ are the factor loadings for the j-th subject and $\epsilon_{i,j}$ is an error term.

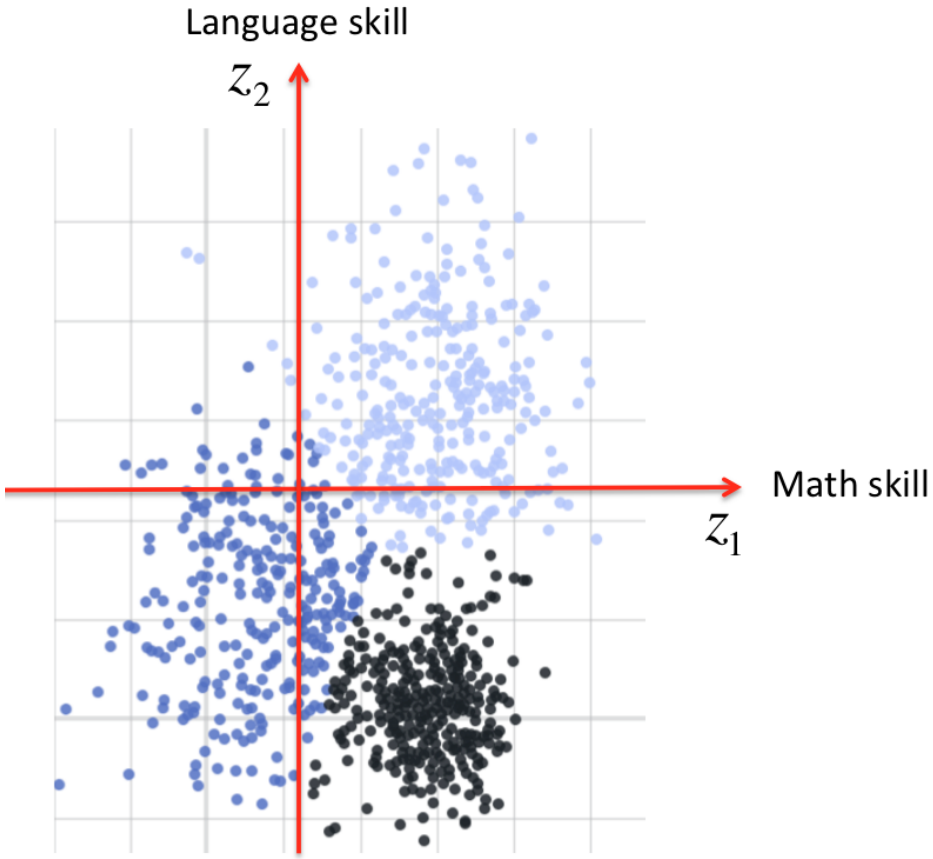
- Eq. 1 can also be written as:

$$\begin{aligned} \mathbf{x}_i &= \begin{bmatrix} m_1 \\ \vdots \\ m_{10} \end{bmatrix} + \begin{bmatrix} v_{1,1} & v_{1,2} \\ \vdots & \vdots \\ v_{10,1} & v_{10,2} \end{bmatrix} \begin{bmatrix} z_{i,1} \\ z_{i,2} \end{bmatrix} + \begin{bmatrix} \epsilon_{i,1} \\ \vdots \\ \epsilon_{i,10} \end{bmatrix} \\ &= \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, 1000, \end{aligned} \tag{2}$$

  where $\mathbf{V}$ is a 10×2 matrix comprising the factor loadings of the 10 subjects.

Example Usage of FA

- For the exam score example, we may apply clustering on the factors $\mathbf{z}_i$, for $i = 1, \ldots, 1000$.



- We may identify the students who are weak in both math and language.

FA versus PCA

- Both PCA and FA are dimension reduction techniques. But they are fundamentally different.
- Components in PCA are orthogonal to each other, whereas factor analysis does not require the columns of the loading matrix to be orthogonal, i.e., the correlation between the factors could be non-zero.
- PCA finds a linear combination of the observed variables to preserve the variance of the data, whereas FA predicts observed variables from theoretical latent factors.

$$\text{PCA}: \quad \mathbf{y}_i = \mathbf{W}^{\mathsf{T}}(\mathbf{x}_i - \mathbf{m}) \text{ and } \hat{\mathbf{x}}_i = \mathbf{m} + \mathbf{W}\mathbf{y}_i$$
$$\text{FA}: \quad \mathbf{x}_i = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_i$$

where $\mathbf{x}_i$'s are observed vectors, $\mathbf{y}_i$'s are PCA-projected vectors of lower dimension, and $\mathbf{z}_i$'s are factors.

Why is FA Important?

- **Finance**: Financial analysts use FA to find "what really drive performance of underling assets", allowing them to make better investment decisions.
- **Social Science**: FA reduces the number of variables to a smaller set of factors that facilitates our understanding of social problems.
- **Marketing**: How change in price (factor) affect the change in the sales (observations)
- **Engineering**: Many engineering applications need to determine the hidden factors that lead to the observations.

Generative Model of FA

- Denote $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ as a set of R-dimensional vectors. In factor analysis, $\mathbf{x}_i$'s are assumed to follow a linear model:

$$\mathbf{x}_i = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N \tag{3}$$

  where $\mathbf{m}$ is the global mean of vectors in $\mathcal{X}$, $\mathbf{V}$ is a low-rank $R \times D$ matrix, $\mathbf{z}_i$ is a D-dimensional latent factor with prior density $\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\epsilon}_i$ is the residual noise following a Gaussian density with zero mean and covariance matrix $\boldsymbol{\Sigma}$.

- The marginal distribution of $\mathbf{x}$ is given by

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \\ &= \int \mathcal{N}(\mathbf{x}|\mathbf{m} + \mathbf{V}\mathbf{z}, \boldsymbol{\Sigma}) \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) d\mathbf{z} \\ &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}) \end{aligned} \tag{4}$$

Generative Model of FA

- Eq. 4 can be obtained by noting that $p(\mathbf{x})$ is a Gaussian.
- We take the expectation of $\mathbf{x}$ in Eq. 3 to obtain:

$$\mathbb{E}\{\mathbf{x}\} = \mathbf{m}.$$

- We take the expectation of $(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathsf{T}}$ in Eq. 3 to obtain:

$$\begin{aligned}\mathbb{E}\{(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\} &= \mathbb{E}\{(\mathbf{V}\mathbf{z}+\boldsymbol{\epsilon})(\mathbf{V}\mathbf{z}+\boldsymbol{\epsilon})^{\mathsf{T}}\} \\ &= \mathbf{V}\mathbb{E}\{\mathbf{z}\mathbf{z}^{\mathsf{T}}\}\mathbf{V}^{\mathsf{T}} + \mathbb{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\ &= \mathbf{V}\mathbf{I}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma} \\ &= \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}.\end{aligned} \tag{5}$$

- Eq. 4 and Eq. 5 suggest that $\mathbf{x}$'s vary in a subspace of $\mathbb{R}^D$ with variability explained by the covariance matrix $\mathbf{V}\mathbf{V}^{\mathsf{T}}$. Any deviations away from this subspace are explained by $\boldsymbol{\Sigma}$.

Generative Model of FA

- Showing $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{m} + \mathbf{V}\mathbf{z}, \mathbf{\Sigma})$
  - The mean can be obtained by considering $\mathbf{z}$ deterministic (known):

$$\begin{aligned} \mathbb{E}\{\mathbf{x}|\mathbf{z}\} &= \mathbb{E}\{\mathbf{m} + \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon}|\mathbf{z}\} && (6)\\ &= \mathbf{m} + \mathbf{V}\mathbf{z} + \mathbb{E}\{\boldsymbol{\epsilon}\} && (7)\\ &= \mathbf{m} + \mathbf{V}\mathbf{z} && (8) \end{aligned}$$

  - Given that the mean is $\mathbf{m} + \mathbf{V}\mathbf{z}$ when we know $\mathbf{z}$, the covariance matrix of $p(\mathbf{x}|\mathbf{z})$ is

$$\begin{aligned} \mathbb{E}\{(\mathbf{x} - \mathbf{m} - \mathbf{V}\mathbf{z})(\mathbf{x} - \mathbf{m} - \mathbf{V}\mathbf{z})^{\mathsf{T}}\} &= \mathbb{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} && (9)\\ &= \mathbf{\Sigma} && (10) \end{aligned}$$

Generative Model of FA

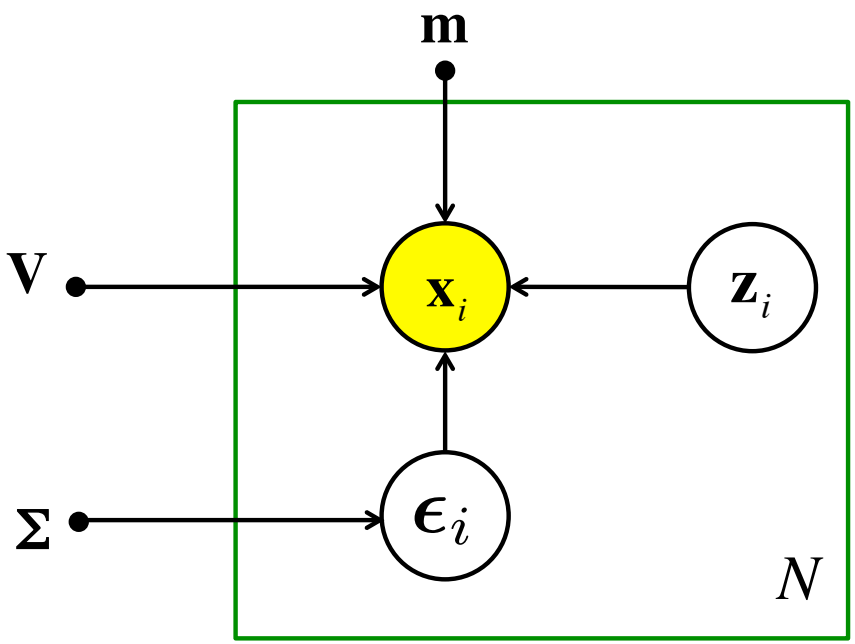
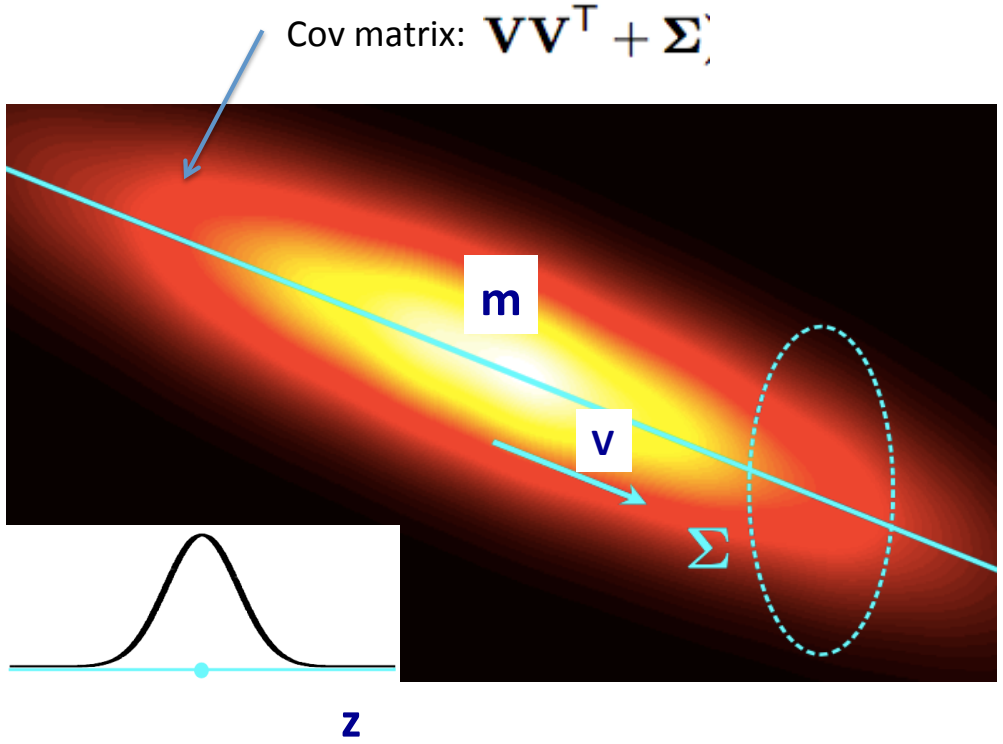


Figure: Graphical model of factor analysis.

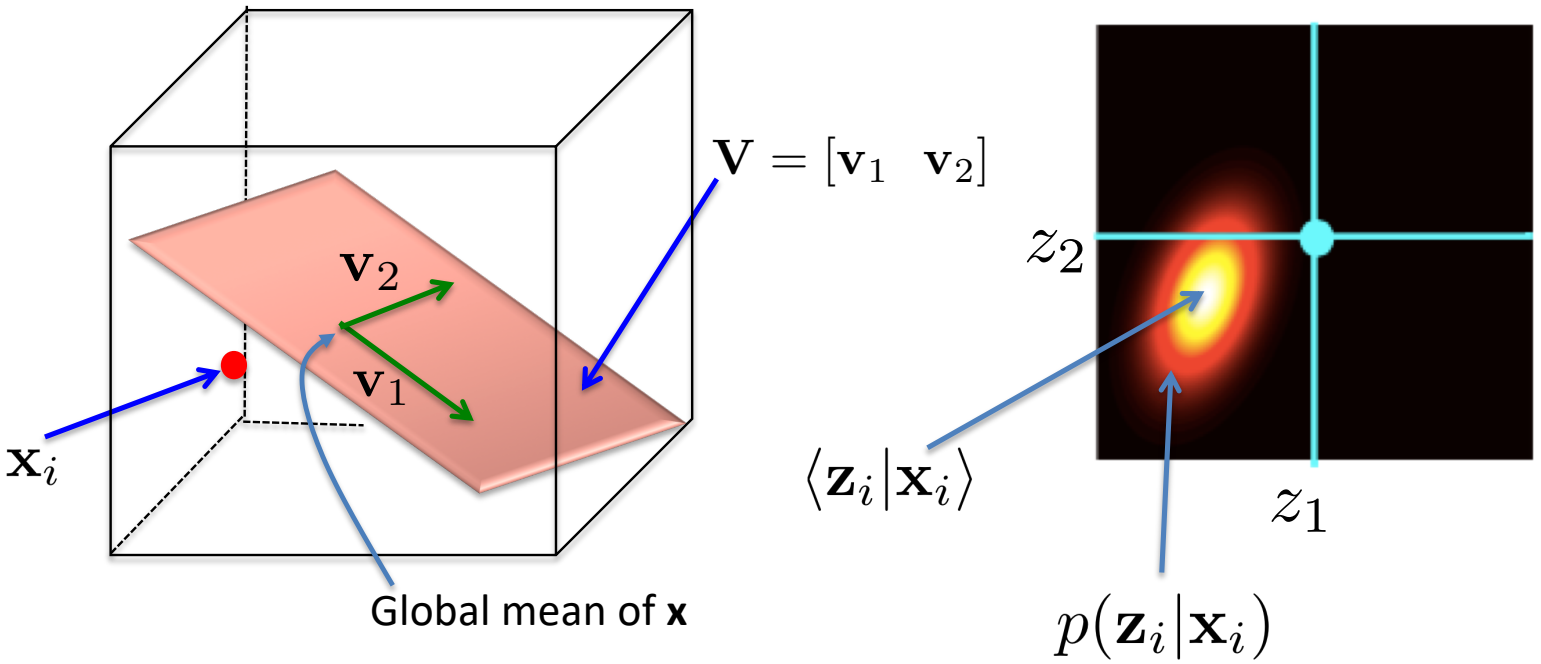
$$\mathbf{x}_i = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N$$

Generative Model of FA



- The column vectors in $\mathbf{V}$ define the directions (subspace) in which the data are most correlated.
- The covariance matrix $\mathbf{\Sigma}$ defines the remaining variation that cannot be captured by $\mathbf{V}\mathbf{V}^{\mathsf{T}}$.

Posterior Density of Latent Factor



- Recall that

$$\mathbf{x}_i = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_i.$$

Therefore, each vector $\mathbf{x}_i$ in the original space can be generated by an infinite number of latent vector $\mathbf{z}_i$ because $\boldsymbol{\epsilon}_i$ is a random vector.

EM Formulation: M-Step

- M-Step: Maximizing the Expectation of Complete Likelihood
- Denote $\boldsymbol{\omega}' = \{\mathbf{m}', \mathbf{V}', \boldsymbol{\Sigma}'\}$ as the new parameter sets. The E- and M-steps iteratively evaluate and maximize the expectation of the complete likelihood:

$$\begin{aligned} Q(\boldsymbol{\omega}'|\boldsymbol{\omega}) &= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\{\log p(\mathcal{X}, \mathcal{Z}|\boldsymbol{\omega}')|\mathcal{X}, \boldsymbol{\omega}\} \\ &= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\left\{\sum\nolimits_i \log\left[p\left(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\omega}'\right) p(\mathbf{z}_i)\right] \middle| \mathcal{X}, \boldsymbol{\omega}\right\} \\ &= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\left\{\sum\nolimits_i \log\left[\mathcal{N}\left(\mathbf{x}_i|\mathbf{m}' + \mathbf{V}'\mathbf{z}_i, \boldsymbol{\Sigma}'\right) \mathcal{N}(\mathbf{z}_i|\mathbf{0}, \mathbf{I})\right] \middle| \mathcal{X}, \boldsymbol{\omega}\right\}. \end{aligned} \tag{11}$$

- Students are suggested to compare this equation with Eq. 5 of the Clustering slides.

EM Formulation: M-Step

- Drop the symbol ($'$) in Eq. 11 and ignore the constant terms independent on the model parameters
- We obtain

$$
\begin{aligned}
Q(\boldsymbol{\omega}) &= -\sum_i \mathbb{E}_{\mathcal{Z}} \left\{ \frac{1}{2} \log |\boldsymbol{\Sigma}| + \frac{1}{2} (\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i) \right\} \\
&= \sum_i \left[-\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}_i - \mathbf{m})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m}) \right] \\
&\quad + \sum_i (\mathbf{x}_i - \mathbf{m})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V} \langle \mathbf{z}_i | \mathbf{x}_i \rangle - \frac{1}{2} \left[\sum_i \left\langle \mathbf{z}_i^{\mathsf{T}} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V} \mathbf{z}_i | \mathbf{x}_i \right\rangle \right].
\end{aligned}
\tag{12}
$$

EM Formulation: M-Step

- Using the properties of matrix derivatives:

$$\frac{\partial \mathbf{a}^{\mathsf{T}}\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^{\mathsf{T}}$$
$$\frac{\partial \mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{B}\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{B}^{\mathsf{T}}\mathbf{X}\mathbf{b}\mathbf{c}^{\mathsf{T}} + \mathbf{B}\mathbf{X}\mathbf{c}\mathbf{b}^{\mathsf{T}}$$
$$\frac{\partial}{\partial \mathbf{A}} \log |\mathbf{A}^{-1}| = -(\mathbf{A}^{-1})^{\mathsf{T}}$$

- We obtain

$$\frac{\partial Q}{\partial \mathbf{V}} = \sum_i \mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{m})\langle \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle - \sum_i \mathbf{\Sigma}^{-1}\mathbf{V} \left\langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \right\rangle . \qquad (13)$$

EM Formulation: M-Step

- Setting $\frac{\partial Q}{\partial \mathbf{V}} = 0$, we have

$$\sum_i \mathbf{V} \left\langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \right\rangle = \sum_i (\mathbf{x}_i - \mathbf{m}) \langle \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle \tag{14}$$

$$\mathbf{V} = \left[\sum_i (\mathbf{x}_i - \mathbf{m}) \langle \mathbf{z}_i | \mathbf{x}_i \rangle^{\mathsf{T}} \right] \left[\sum_i \left\langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \right\rangle \right]^{-1}. \tag{15}$$

- To find $\mathbf{\Sigma}$, we evaluate

$$\frac{\partial Q}{\partial \mathbf{\Sigma}^{-1}} = \frac{1}{2} \sum_i \left[\mathbf{\Sigma} - (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^{\mathsf{T}} \right] + \sum_i (\mathbf{x}_i - \mathbf{m}) \langle \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle \mathbf{V}^{\mathsf{T}} - \frac{1}{2} \sum_i \mathbf{V} \langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle \mathbf{V}^{\mathsf{T}}.$$

EM Formulation: M-Step

- Note that according to Eq. 14, we have

$$\frac{\partial Q}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2}\sum_i \left[\boldsymbol{\Sigma} - (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^{\mathsf{T}}\right] + \sum_i (\mathbf{x}_i - \mathbf{m})\langle \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle \mathbf{V}^{\mathsf{T}} - \frac{1}{2}\sum_i (\mathbf{x}_i - \mathbf{m})\langle \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle \mathbf{V}^{\mathsf{T}}.$$

- Therefore, setting $\frac{\partial Q}{\partial \boldsymbol{\Sigma}^{-1}} = 0$ we have

$$\sum_i \boldsymbol{\Sigma} = \sum_i \left[(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^{\mathsf{T}} - (\mathbf{x}_i - \mathbf{m})\langle \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle \mathbf{V}^{\mathsf{T}}\right].$$

- Rearranging, we have

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_i \left[(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^{\mathsf{T}} - \mathbf{V}\langle \mathbf{z}_i | \mathbf{x}_i \rangle (\mathbf{x}_i - \mathbf{m})^{\mathsf{T}}\right].$$

EM Formulation: M-Step

- To compute $\mathbf{m}$, we evaluate

$$\frac{\partial Q}{\partial \mathbf{m}} = -\sum_i \left(\mathbf{\Sigma}^{-1}\mathbf{m} - \mathbf{\Sigma}^{-1}\mathbf{x}_i\right) + \sum_i \mathbf{\Sigma}^{-1}\mathbf{V}\langle \mathbf{z}_i|\mathbf{x}_i\rangle.$$

- Setting $\frac{\partial Q}{\partial \mathbf{m}} = 0$, we have

$$\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

where we have used the property $\sum_{i=1}^{N}\langle \mathbf{z}_i|\mathbf{x}_i\rangle \approx \mathbf{0}$ when N is sufficiently large. We have this property because of the assumption that the prior of $\mathbf{z}$ follows a Gaussian distribution, i.e., $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$.

EM Formulation: E-Step

- In the E-step, we compute the posterior means $\langle \mathbf{z}_i | \mathbf{x}_i \rangle$ and posterior moments $\langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle$.
- Consider the posterior density:

$$\begin{aligned} & p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\omega}) \\ & \propto p(\mathbf{x}_i | \mathbf{z}_i, \boldsymbol{\omega}) p(\mathbf{z}_i) \\ & \propto \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i) - \frac{1}{2} \mathbf{z}_i^{\mathsf{T}} \mathbf{z}_i \right\} \\ & = \exp\left\{ \mathbf{z}_i^{\mathsf{T}} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m}) - \frac{1}{2} \mathbf{z}_i^{\mathsf{T}} \left(\mathbf{I} + \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V} \right) \mathbf{z}_i \right\}. \end{aligned} \tag{16}$$

- A Gaussian distribution can be expressed as

$$\begin{aligned} \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_z, \mathbf{C}_z) & \propto \exp\left\{ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu}_z)^{\mathsf{T}} \mathbf{C}_z^{-1} (\mathbf{z} - \boldsymbol{\mu}_z) \right\} \\ & \propto \exp\left\{ \mathbf{z}^{\mathsf{T}} \mathbf{C}_z^{-1} \boldsymbol{\mu}_z - \frac{1}{2} \mathbf{z}^{\mathsf{T}} \mathbf{C}_z^{-1} \mathbf{z} \right\}. \end{aligned} \tag{17}$$

EM Formulation: E-Step

- Comparing Eq. 16 and Eq. 17, we obtain the posterior mean and moment as follows:

$$\langle \mathbf{z}_i | \mathbf{x}_i \rangle = \mathbf{L}^{-1} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m})$$
$$\langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle = \mathbf{L}^{-1} + \langle \mathbf{z}_i | \mathcal{X} \rangle \langle \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle$$

  where $\mathbf{L}^{-1} = (\mathbf{I} + \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V})^{-1}$ is the posterior covariance matrix of $\mathbf{z}_i$, $\forall i$.

EM Formulation

In summary, we have the following EM algorithm for factor analysis:

E-step:

$$\langle \mathbf{z}_i | \mathbf{x}_i \rangle = \mathbf{L}^{-1} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m})$$

$$\langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle = \mathbf{L}^{-1} + \langle \mathbf{z}_i | \mathbf{x}_i \rangle \langle \mathbf{z}_i | \mathbf{x}_i \rangle^{\mathsf{T}}$$

$$\mathbf{L} = \mathbf{I} + \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V}$$

M-step:

$$\mathbf{V}' = \left[\sum\nolimits_i (\mathbf{x}_i - \mathbf{m}') \langle \mathbf{z}_i | \mathbf{x}_i \rangle^{\mathsf{T}} \right] \left[\sum\nolimits_i \langle \mathbf{z}_i \mathbf{z}_i^{\mathsf{T}} | \mathbf{x}_i \rangle \right]^{-1}$$

$$\mathbf{m}' = \frac{1}{N} \sum\nolimits_i \mathbf{x}_i$$

$$\boldsymbol{\Sigma}' = \frac{1}{N} \left\{ \sum\nolimits_{i=1}^{N} \left[(\mathbf{x}_i - \mathbf{m}')(\mathbf{x}_i - \mathbf{m}')^{\mathsf{T}} - \mathbf{V}' \langle \mathbf{z}_i | \mathbf{x}_i \rangle (\mathbf{x}_i - \mathbf{m}')^{\mathsf{T}} \right] \right\}$$

(18)

Relation to PCA

- Consider $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. Then, we have

$$\mathbf{L} = \mathbf{I} + \frac{1}{\sigma^2}\mathbf{V}^{\mathsf{T}}\mathbf{V}$$

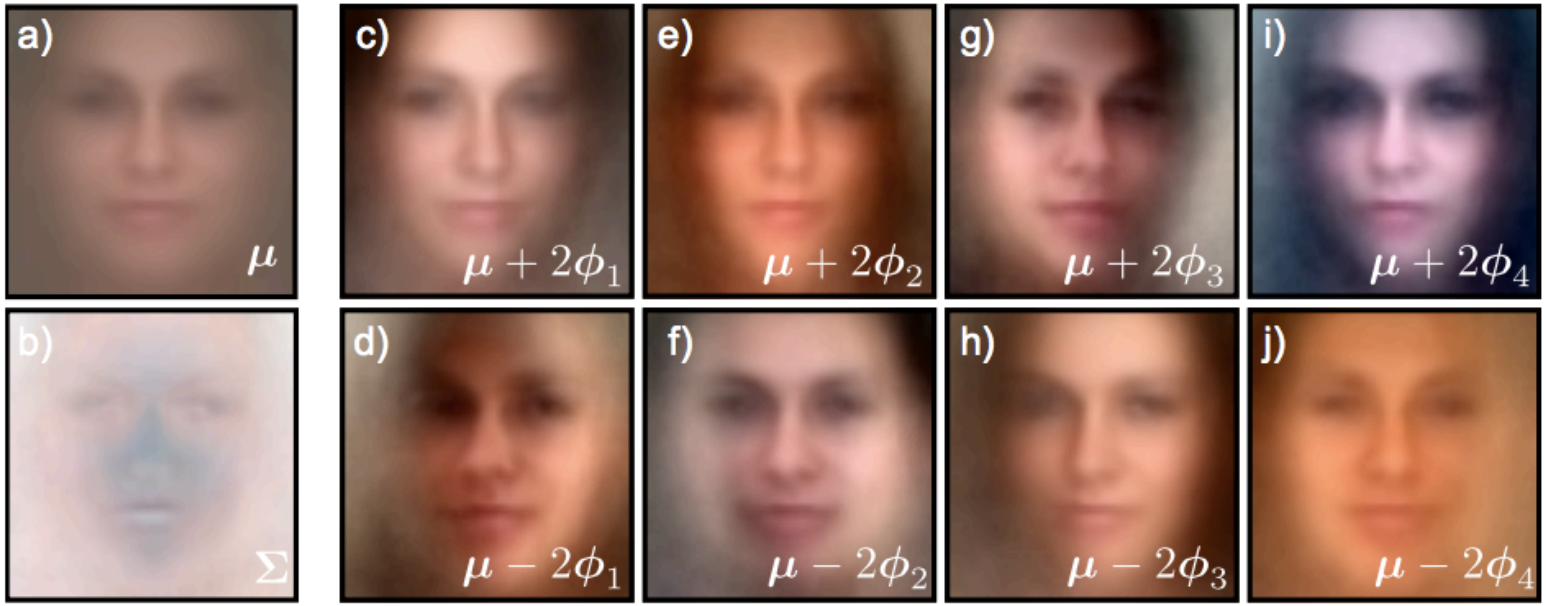
$$\langle \mathbf{z}_i | \mathbf{x}_i \rangle = \frac{1}{\sigma^2}\mathbf{L}^{-1}\mathbf{V}^{\mathsf{T}}(\mathbf{x}_i - \mathbf{m})$$

- When $\sigma^2 \to 0$ and $\mathbf{V}$ is an orthogonal matrix such that $\mathbf{V}^{-1} = \mathbf{V}^{\mathsf{T}}$, then $\mathbf{L} \to \frac{1}{\sigma^2}\mathbf{V}^{\mathsf{T}}\mathbf{V}$ and

$$\begin{aligned} \langle \mathbf{z}_i | \mathbf{x}_i \rangle &\to \frac{1}{\sigma^2}\sigma^2(\mathbf{V}^{\mathsf{T}}\mathbf{V})^{-1}\mathbf{V}^{\mathsf{T}}(\mathbf{x}_i - \mathbf{m}) \\ &= \mathbf{V}^{-1}(\mathbf{V}^{\mathsf{T}})^{-1}\mathbf{V}^{\mathsf{T}}(\mathbf{x}_i - \mathbf{m}) \\ &= \mathbf{V}^{\mathsf{T}}(\mathbf{x}_i - \mathbf{m}) \end{aligned} \tag{19}$$

- Note that Eq. 19 is equivalent to PCA projection and that the posterior covariance ($\mathbf{L}^{-1}$) of $\mathbf{z}$ becomes $\mathbf{0}$.

Applications of FA to Facial Images



Source: S.J.D. Prince, *Computer Vision: Models Learning and Inference*, Cambridge University Press, 2012

- Different column vectors in $\mathbf{V}$ encode different types of variability in the facial data, e.g., hue and pose.

I-Vectors

GMM I-Vectors

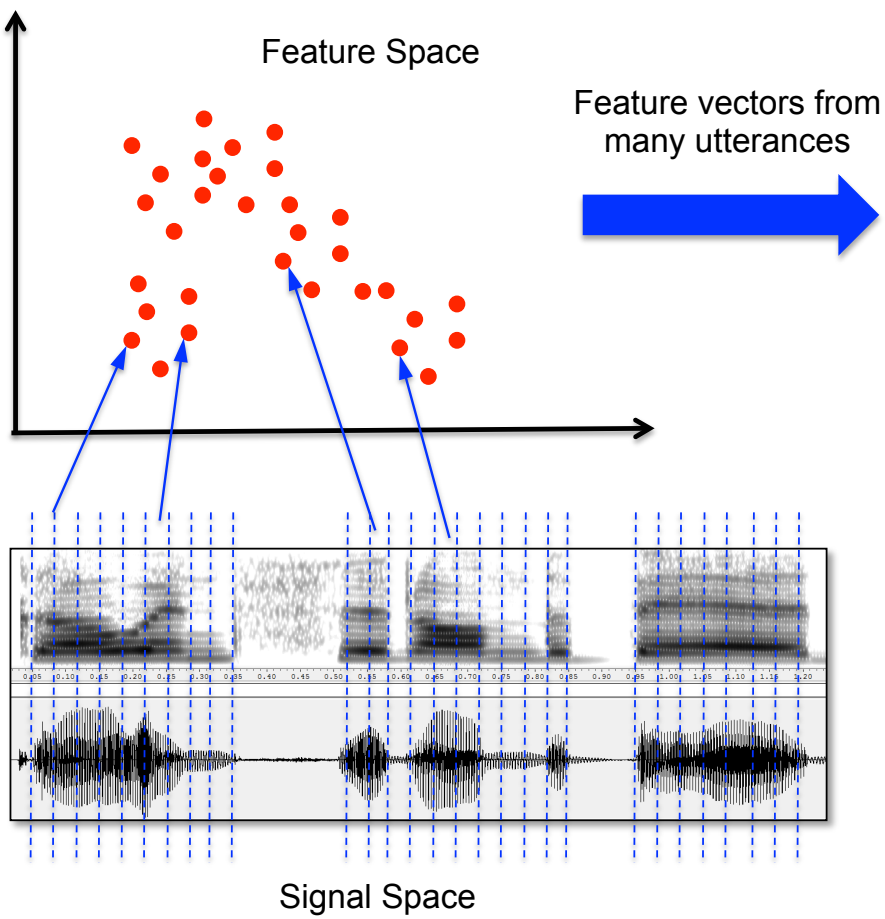
- I-vectors are based on factor analysis in which the acoustic features are generated by a Gaussian mixture model (GMM).
- Given the i-th utterance, we denote $\mathcal{O}_i = \{\mathbf{o}_{i1}, \ldots, \mathbf{o}_{iT_i}\}$ as a set of F-dimensional observed vectors, which are assumed to be generated by a GMM, i.e.,

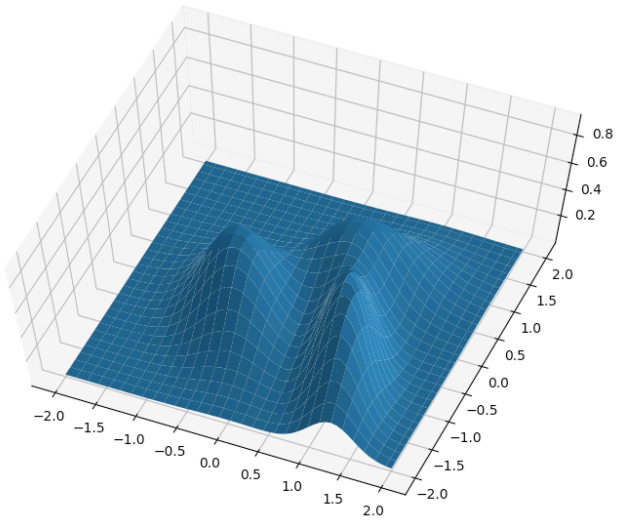
$$p(\mathbf{o}_{it}) = \sum_{c=1}^{C} \lambda_c \mathcal{N}(\mathbf{o}_{it}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \qquad t = 1, \ldots, T_i, \tag{20}$$

where $\{\lambda_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$ are the parameters of the GMM, C is the number of mixtures, and T_i is the number of frames in the utterance.

GMM I-Vectors

- Gaussian mixture model (GMM) for acoustic modeling
- Each acoustic frame in speech is represented by a low-dimensional acoustic vector



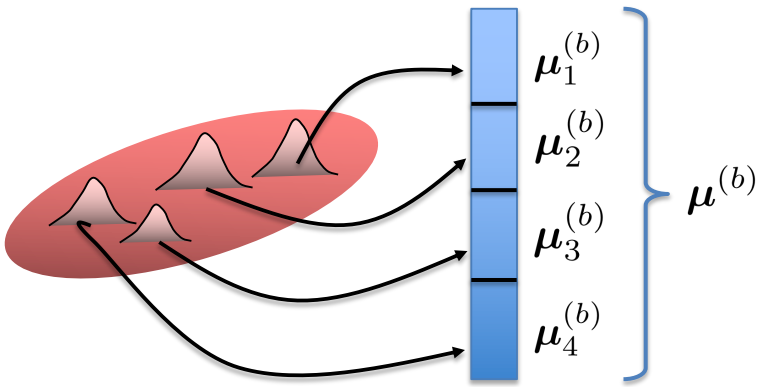


I-Vectors: FA Model

- The GMM-supervector representing the i-th utterance is assumed to be generated by the following factor analysis model:

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}^{(b)} + \mathbf{T}\mathbf{w}_i \tag{21}$$

where $\boldsymbol{\mu}^{(b)}$ is obtained by stacking the mean vectors of a universal background model (UBM), $\mathbf{T}$ is a $CF \times D$ low-rank total variability matrix modeling the speaker and channel variability, and $\mathbf{w}_i$ is the latent factor of dimension D.



I-Vectors: FA Model

- Eq. 21 can also be written in a component-wise form:

$$\boldsymbol{\mu}_{ic} = \boldsymbol{\mu}_c^{(b)} + \mathbf{T}_c \mathbf{w}_i, \qquad c = 1, \ldots, C \tag{22}$$

where $\boldsymbol{\mu}_{ic} \in \Re^F$ is the c-th sub-vector of $\boldsymbol{\mu}_i$ (similarly for $\boldsymbol{\mu}_c^{(b)}$) and $\mathbf{T}_c$ is an $F \times D$ sub-matrix of $\mathbf{T}$.

I-Vectors: FA Model

- Given an utterance with acoustic vectors $\mathcal{O}_i$, the i-vector $\mathbf{x}_i$ representing the utterance is the posterior mean of $\mathbf{w}_i$, i.e., $\mathbf{x}_i = \langle \mathbf{w}_i | \mathcal{O}_i \rangle$.
- To determine $\mathbf{x}_i$, we consider the joint posterior distribution:

$$
\begin{aligned}
p(\mathbf{w}_i, y_{i,\cdot,\cdot} | \mathcal{O}_i) &\propto p(\mathcal{O}_i | \mathbf{w}_i, y_{i,\cdot,\cdot} = 1) p(y_{i,\cdot,\cdot}) p(\mathbf{w}_i) \\
&= \prod_{t=1}^{T} \prod_{c=1}^{C} \left[\lambda_c p(\mathbf{o}_{it} | y_{i,t,c} = 1, \mathbf{w}_i) \right]^{y_{i,t,c}} p(\mathbf{w}_i) \\
&= \underbrace{p(\mathbf{w}_i) \prod_{t=1}^{T} \prod_{c=1}^{C} \left[\mathcal{N}(\mathbf{o}_{it} | \boldsymbol{\mu}_c^{(b)} + \mathbf{T}_c \mathbf{w}_i, \boldsymbol{\Sigma}_c^{(b)}) \right]^{y_{i,t,c}}}_{\propto p(\mathbf{w}_i | \mathcal{O}_i)} \lambda_c^{y_{i,t,c}},
\end{aligned}
\tag{23}
$$

I-Vectors: FA Model

- Extracting terms depending on $\mathbf{w}_i$ from Eq. 23, we obtain

$$\begin{aligned}\log p(\mathbf{w}_i|\mathcal{O}_i) \propto & -\frac{1}{2}\sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}(\mathbf{o}_{it}-\boldsymbol{\mu}_c^{(b)}-\mathbf{T}_c\mathbf{w}_i)^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1} \\ & (\mathbf{o}_{it}-\boldsymbol{\mu}_c^{(b)}-\mathbf{T}_c\mathbf{w}_i)-\frac{1}{2}\mathbf{w}_i^{\mathsf{T}}\mathbf{w}_i \\ = & \mathbf{w}_i^{\mathsf{T}}\sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}\mathbf{T}_c^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1}(\mathbf{o}_{it}-\boldsymbol{\mu}_c^{(b)}) \\ & -\frac{1}{2}\mathbf{w}_i^{\mathsf{T}}\left(\mathbf{I}+\sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}\mathbf{T}_c^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1}\mathbf{T}_c\right)\mathbf{w}_i,\end{aligned} \tag{24}$$

where $\mathcal{H}_{ic}$ comprises the frame indexes for which $\mathbf{o}_{it}$ aligned to mixture c.

I-Vectors: FA Model

- Comparing Eq. 24 with Eq. 17, we obtain the following posterior expectations:

$$\langle \mathbf{w}_i | \mathcal{O}_i \rangle = \mathbf{L}_i^{-1} \sum_{c=1}^{C} \sum_{t \in \mathcal{H}_{ic}} \mathbf{T}_c^{\mathsf{T}} \left(\mathbf{\Sigma}_c^{(b)} \right)^{-1} (\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)})$$

$$= \mathbf{L}_i^{-1} \sum_{c=1}^{C} \mathbf{T}_c^{\mathsf{T}} \left(\mathbf{\Sigma}_c^{(b)} \right)^{-1} \sum_{t \in \mathcal{H}_{ic}} (\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}) \quad (25)$$

$$\langle \mathbf{w}_i \mathbf{w}_i^{\mathsf{T}} | \mathcal{O}_i \rangle = \mathbf{L}_i^{-1} + \langle \mathbf{w}_i | \mathcal{O}_i \rangle \langle \mathbf{w}_i^{\mathsf{T}} | \mathcal{O}_i \rangle \quad (26)$$

where

$$\mathbf{L}_i = \mathbf{I} + \sum_{c=1}^{C} \sum_{t \in \mathcal{H}_{ic}} \mathbf{T}_c^{\mathsf{T}} (\mathbf{\Sigma}_c^{(b)})^{-1} \mathbf{T}_c. \quad (27)$$

Hard Decisions for Frame Alignment

- For each t, the posterior probabilities of $y_{i,t,c}$, for $c = 1, \ldots, C$, are computed. Then, $\mathbf{o}_{it}$ is aligned to mixture c^* when

$$c^* = \arg\max_c \gamma_c(\mathbf{o}_{it})$$

where

$$\begin{aligned} \gamma_c(\mathbf{o}_{it}) &\equiv \Pr(C_t = c|\mathbf{o}_{it}) \\ &= \frac{\lambda_c^{(b)} \mathcal{N}(\mathbf{o}_{it}|\boldsymbol{\mu}_c^{(b)}, \boldsymbol{\Sigma}_c^{(b)})}{\sum_{j=1}^{C} \lambda_j^{(b)} \mathcal{N}(\mathbf{o}_{it}; \boldsymbol{\mu}_j^{(b)}, \boldsymbol{\Sigma}_j^{(b)})}, \quad c = 1, \ldots, C \end{aligned} \tag{28}$$

are the posterior probabilities of $\mathbf{o}_{it}$.

Soft Decisions for Frame Alignment

- Each frame is aligned to all of the mixtures with degree of alignment according to the posterior probabilities $\gamma_c(\mathbf{o}_{it})$:

$$\begin{aligned} \sum_{t \in \mathcal{H}_{ic}} 1 &= \sum_{t=1}^{T_i} \gamma_c(\mathbf{o}_{it}) \\ \sum_{t \in \mathcal{H}_{ic}} (\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}) &= \sum_{t=1}^{T_i} \gamma_c(\mathbf{o}_{it})(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}). \end{aligned} \tag{29}$$

- Baum-Welch statistics:

$$N_{ic} \equiv \sum_{t=1}^{T_i} \gamma_c(\mathbf{o}_{it}) \quad \text{and} \quad \tilde{\mathbf{f}}_{ic} \equiv \sum_{t=1}^{T_i} \gamma_c(\mathbf{o}_{it})(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}). \tag{30}$$

I-Vector Extraction

Substituting Eq. 30 into Eq. 25 and Eq. 27, we have the following expression for i-vectors:

$$\begin{aligned}\mathbf{x}_i &= \langle \mathbf{w}_i | \mathcal{O}_i \rangle \\ &= \mathbf{L}_i^{-1} \sum_{c=1}^{C} \mathbf{T}_c^{\mathsf{T}} \left(\boldsymbol{\Sigma}_c^{(b)} \right)^{-1} \tilde{\mathbf{f}}_{ic} \\ &= \mathbf{L}_i^{-1} \mathbf{T}^{\mathsf{T}} (\boldsymbol{\Sigma}^{(b)})^{-1} \tilde{\mathbf{f}}_i\end{aligned} \tag{31}$$

where $\tilde{\mathbf{f}}_i = [\tilde{\mathbf{f}}_{i1}^{\mathsf{T}} \ \cdots \ \tilde{\mathbf{f}}_{iC}^{\mathsf{T}}]^{\mathsf{T}}$ and

$$\mathbf{L}_i = \mathbf{I} + \sum_{c=1}^{C} N_{ic} \mathbf{T}_c^{\mathsf{T}} (\boldsymbol{\Sigma}_c^{(b)})^{-1} \mathbf{T}_c = \mathbf{I} + \mathbf{T}^{\mathsf{T}} (\boldsymbol{\Sigma}^{(b)})^{-1} \mathbf{N}_i \mathbf{T}$$

where $\mathbf{N}_i$ is a $CF \times CF$ block diagonal matrix containing $N_{ic}\mathbf{I}$, $c = 1, \ldots, C$, as its block diagonal elements.

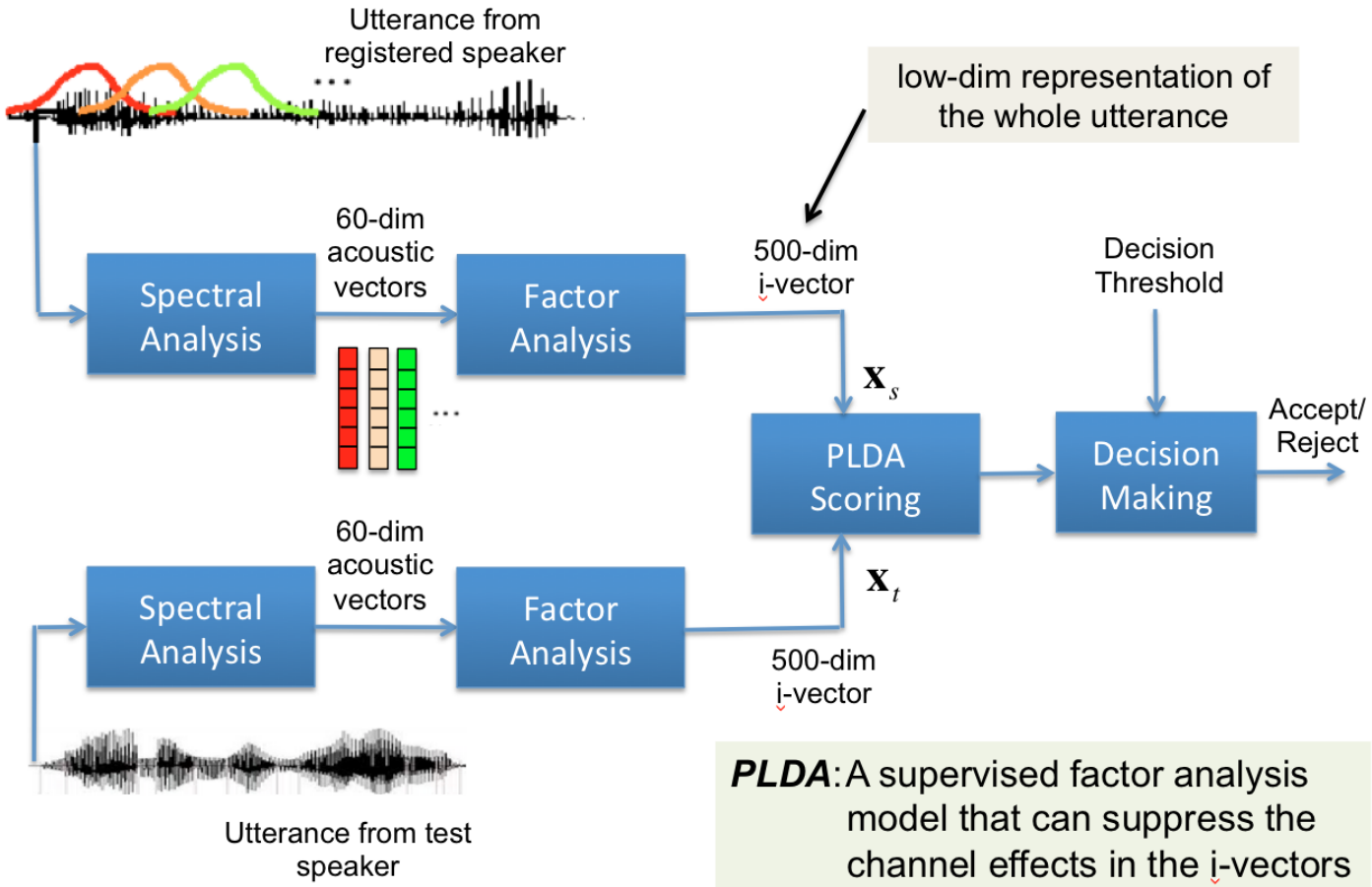
I-Vectors: Model Training

- Model training involves estimating the total variability matrix $\mathbf{T}$ using the EM algorithm.
- Using Eq. 18 and Eq. 24, the M-step for estimating $\mathbf{T}$ is

$$\mathbf{T}_c = \left[\sum\nolimits_i \tilde{\mathbf{f}}_{ic}\langle \mathbf{w}_i|\mathcal{O}_i\rangle^{\mathsf{T}}\right]\left[\sum\nolimits_i N_{ic}\langle \mathbf{w}_i\mathbf{w}_i^{\mathsf{T}}|\mathcal{O}_i\rangle\right]^{-1}, \quad c = 1,\ldots,C. \tag{32}$$

Applications of I-Vectors

- I-vectors have been applied to many domains, including speaker recognition, language recognition, speech recognition, and speaker diarization.
- The most popular application of i-vectors is speaker verification



Applications of I-Vectors

- Adapting DNNs for patient-dependent ECG classification

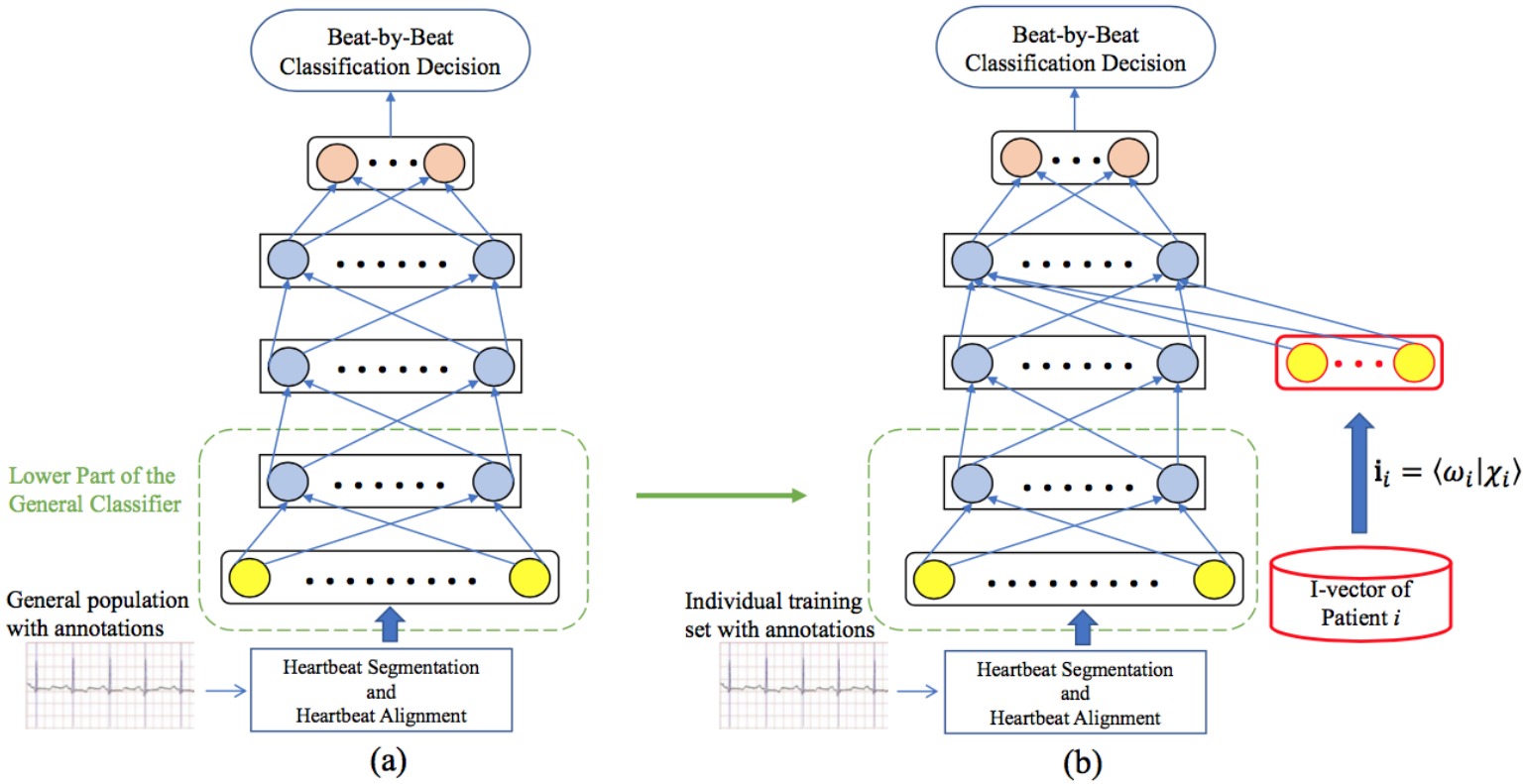


Fig. 2: I-vector adapted patient-specific DNN (iAP-DNN). (a) General classier. (b) Patient-specific classifier.

Source: S. Xu, M.W. Mak and C.C. Cheung, "I-Vector Based Patient Adaptation of Deep Neural Networks for Automatic Heartbeat Classification", IEEE Journal of Biomedical and Health Informatics, May 2019

Software Tools

- Factor Analysis
  - Python scikit-learn: sklearn.decomposition.FactorAnalysis
  - Matlab factoran
- I-Vectors
  - SIDEKIT: https://projets-lium.univ-lemans.fr/sidekit/overview/index.html
  - mPLDA: http://bioinfo.eie.polyu.edu.hk/mPLDA
  - Kaldi: https://kaldi-asr.org
  - MSR Identity Toolbox