Lecture Notes on Factor Analysis and I-Vectors

Man-Wai MAK

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
enmwmak@polyu.edu.hk

Abstract

This document provides the detailed formulations of factor analysis (FA) models in which the observed vectors are assumed to follow a mixture of Gaussians and prior of the latent factors follows a Gaussian distribution. The application of FA to i-vector extraction will be discussed.

Please cite this document as: M.W. Mak, "Lecture Notes on Factor Analysis and I-vectors", *Technical Report and Lecture Note Series, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University*, Feb 2016.

Keywords: Factor analysis, I-vectors, General linear model

1. Factor Analysis

1.1. Generative Model

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 assume to follow a linear model:

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

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 an 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 $\boldsymbol{\Sigma}$.

Given the factor analysis model in Eq. 1, it can be shown that the

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

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \\ &= \int \mathcal{N}(\mathbf{x}|\mathbf{m}+\mathbf{V}\mathbf{z},\mathbf{\Sigma})\mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I})d\mathbf{z} \\ &= \mathcal{N}(\mathbf{x}|\mathbf{m},\mathbf{V}\mathbf{V}^{\mathsf{T}}+\mathbf{\Sigma}). \end{aligned} \tag{2}$$

Eq. 2 can be obtained by convolution of Gaussian or by noting that $p(\mathbf{x})$ is a Gaussian. For the latter, we take the expectation of $\mathbf{x}$ in Eq. 1 to obtain:

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

1.2. EM Formulation

Although each iteration of EM should be started with the E-step followed by the M-step, notationally, it is more conveniently to describe the M-step first, assuming that the posterior expectations have already been computed in the E-step.

1.2.1. M-Step: Maximizing the Expectation of Complete Likelihood

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

$$\begin{aligned} Q(\boldsymbol{\omega}'|\boldsymbol{\omega}) &= \mathbb{E}_{\mathcal{Z}}\{\ln p(\mathcal{X},\mathcal{Z}|\boldsymbol{\omega}')|\mathcal{X},\boldsymbol{\omega}\} \\ &= \mathbb{E}_{\mathcal{Z}}\left\{\sum\nolimits_i \ln\left[p\left(\mathbf{x}_i|\mathbf{z}_i,\boldsymbol{\omega}'\right)p(\mathbf{z}_i)\right]\middle|\mathcal{X},\boldsymbol{\omega}\right\} \\ &= \mathbb{E}_{\mathcal{Z}}\left\{\sum\nolimits_i \ln\left[\mathcal{N}\left(\mathbf{x}_i|\mathbf{m}'+\mathbf{V}'\mathbf{z}_i,\mathbf{\Sigma}'\right)\mathcal{N}(\mathbf{z}_i|\mathbf{0},\mathbf{I})\right]\middle|\mathcal{X},\boldsymbol{\omega}\right\}, \end{aligned} \tag{3}$$

To simplify notations, we drop the symbol ($'$) in Eq. 33 and ignore the constant terms independent on the model parameters, which results in

$$\begin{aligned} Q(\boldsymbol{\omega}) &= -\sum_i \mathbb{E}_{\mathcal{Z}} \left\{ \frac{1}{2} \log |\boldsymbol{\Sigma}| + \frac{1}{2} (\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i) \right\} \\ &= \sum_i \left[-\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}_i - \mathbf{m})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m}) \right] \\ &\quad + \sum_i (\mathbf{x}_i - \mathbf{m})^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbf{V} \left\langle \mathbf{z}_i | \mathbf{x}_i \right\rangle - \frac{1}{2} \left[\sum_i \left\langle \mathbf{z}_i^\mathsf{T} \mathbf{V}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbf{V} \mathbf{z}_i | \mathbf{x}_i \right\rangle \right]. \end{aligned} \tag{4}$$

Using the following properties of matrix derivatives

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

we have

$$\frac{\partial Q}{\partial \mathbf{V}} = \sum_i \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m}) \langle \mathbf{z}_i^\mathsf{T} | \mathbf{x}_i \rangle - \sum_i \boldsymbol{\Sigma}^{-1} \mathbf{V} \left\langle \mathbf{z}_i \mathbf{z}_i^\mathsf{T} | \mathbf{x}_i \right\rangle. \tag{5}$$

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

$$\sum_i \mathbf{V} \left\langle \mathbf{z}_i \mathbf{z}_i^\mathsf{T} | \mathbf{x}_i \right\rangle = \sum_i (\mathbf{x}_i - \mathbf{m}) \langle \mathbf{z}_i^\mathsf{T} | \mathbf{x}_i \rangle \tag{6}$$

$$\mathbf{V} = \left[\sum_i (\mathbf{x}_i - \mathbf{m}) \langle \mathbf{z}_i | \mathbf{x}_i \rangle^\mathsf{T} \right] \left[\sum_i \left\langle \mathbf{z}_i \mathbf{z}_i^\mathsf{T} | \mathbf{x}_i \right\rangle \right]^{-1}. \tag{7}$$

To find $\boldsymbol{\Sigma}$, we evaluate

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

Note that according to Eq. 6, we have

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

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

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

Rearranging, we have

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

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 $\sum_i \langle \mathbf{z}_i|\mathbf{x}_i\rangle \approx \mathbf{0}$ 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., $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$.

1.2.2. E-Step: Computing the Posterior Expectations

In the E-step, we compute the posterior means $\langle \mathbf{z}_i|\mathbf{x}_i\rangle$ and posterior moments $\langle \mathbf{z}_i\mathbf{z}_i^\mathsf{T}|\mathbf{x}_i\rangle$. Let's express the following posterior density in terms of

its likelihood and prior [1]:

$$
\begin{aligned}
& p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\omega}) \\
& \propto p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\omega}) p(\mathbf{z}_i) \\
& \propto \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i)^\mathsf{T} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{m} - \mathbf{V}\mathbf{z}_i) - \frac{1}{2}\mathbf{z}_i^\mathsf{T}\mathbf{z}_i\right\} \\
& = \exp\left\{\mathbf{z}_i^\mathsf{T}\mathbf{V}^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{m}) - \frac{1}{2}\mathbf{z}_i^\mathsf{T}\left(\mathbf{I} + \mathbf{V}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{V}\right)\mathbf{z}_i\right\}.
\end{aligned} \tag{8}
$$

Consider the following property of Gaussian distribution with mean $\boldsymbol{\mu}_z$ and covariance $\mathbf{C}_z$

$$
\begin{aligned}
\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \mathbf{C}_z) &\propto \exp\left\{-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_z)^\mathsf{T}\mathbf{C}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z)\right\} \\
&\propto \exp\left\{\mathbf{z}^\mathsf{T}\mathbf{C}_z^{-1}\boldsymbol{\mu}_z - \frac{1}{2}\mathbf{z}^\mathsf{T}\mathbf{C}_z^{-1}\mathbf{z}\right\}.
\end{aligned} \tag{9}
$$

Comparing Eq. 8 and Eq. 9, we obtain the posterior mean and moment as follows:

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

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

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

$$
\boxed{
\begin{aligned}
&\textbf{E-step: } \langle \mathbf{z}_i|\mathbf{x}_i\rangle = \mathbf{L}^{-1}\mathbf{V}^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{m}); \quad \langle \mathbf{z}_i\mathbf{z}_i^\mathsf{T}|\mathbf{x}_i\rangle = \mathbf{L}^{-1} + \langle \mathbf{z}_i|\mathbf{x}_i\rangle\langle \mathbf{z}_i|\mathbf{x}_i\rangle^\mathsf{T} \\
&\qquad \mathbf{L} = \mathbf{I} + \mathbf{V}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{V} \\
&\textbf{M-step: } \mathbf{V}' = \left[\sum\nolimits_i (\mathbf{x}_i - \mathbf{m}')\langle \mathbf{z}_i|\mathbf{x}_i\rangle^\mathsf{T}\right]\left[\sum\nolimits_i \langle \mathbf{z}_i\mathbf{z}_i^\mathsf{T}|\mathbf{x}_i\rangle\right]^{-1}; \quad \mathbf{m}' = \frac{1}{N}\sum\nolimits_i \mathbf{x}_i \\
&\qquad \boldsymbol{\Sigma}' = \frac{1}{N}\left\{\sum\nolimits_{i=1}^N \left[(\mathbf{x}_i - \mathbf{m}')(\mathbf{x}_i - \mathbf{m}')^\mathsf{T} - \mathbf{V}'\langle \mathbf{z}_i|\mathbf{x}_i\rangle(\mathbf{x}_i - \mathbf{m}')^\mathsf{T}\right]\right\}
\end{aligned}
} \tag{10}
$$

1.3. 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, then $\mathbf{L} \to \frac{1}{\sigma^2}\mathbf{V}^\mathsf{T}\mathbf{V}$ and

$$\begin{aligned}\langle \mathbf{z}_i|\mathbf{x}_i\rangle &\to \frac{1}{\sigma^2}\sigma^2(\mathbf{V}^\mathsf{T}\mathbf{V})^{-1}\mathbf{V}^\mathsf{T}(\mathbf{x}_i - \mathbf{m}) \\ &= \mathbf{V}^{-1}(\mathbf{V}^\mathsf{T})^{-1}\mathbf{V}^\mathsf{T}(\mathbf{x}_i - \mathbf{m}) \\ &= \mathbf{V}^\mathsf{T}(\mathbf{x}_i - \mathbf{m}).\end{aligned} \tag{11}$$

Note that Eq. 11 is equivalent to PCA projection and that the posterior covariance ($\mathbf{L}^{-1}$) of $\mathbf{z}$ becomes 0.

The analysis above suggests that PCA is a specific case of factor analysis in which the covariances of residue noise $\boldsymbol{\epsilon}$ collapse to zero.[1] This means that in PCA, for a given $\mathbf{x}_i$ and $\mathbf{V}$ in Eq. 1, there is a *deterministic* $\mathbf{z}_i$ and a unique residue $\boldsymbol{\epsilon}_i$. On the other hand, in factor analysis, an $\mathbf{x}_i$ can be generated by infinite combinations of $\mathbf{z}_i$ and $\boldsymbol{\epsilon}_i$, as evident in Eq. 2.

The covariance matrix ($\mathbf{V}\mathbf{V}^\mathsf{T} + \mathbf{\Sigma}$) of observed data in FA suggests that FA treats covariances and variances of observed data separately. More specifically, the D column vectors (factor loadings) of $\mathbf{V}$ in FA capture most of the covariances while the variances of the individual components of $\mathbf{x}$'s are captured by the diagonal elements of $\mathbf{\Sigma}$. Note that the diagonal elements in $\mathbf{\Sigma}$ can be different, providing extra flexibility in modelling variances. On the other hand, because $\mathbf{\Sigma} \to \mathbf{0}$ in PCA, the D principal components attempt to capture most of the variabilities, including covariances and variances. As a result, FA is more flexible in terms of data modeling.

2. I-vectors

2.1. Gaussian Mixture Model

I-vectors are based on factor analysis in which the acoustic features (typically MFCC and log-energy plus their 1st and 2nd derivatives) are generated by a Gaussian mixture model (GMM). Given the i-th utterance, we denote

[1] Note that it does not mean that $\boldsymbol{\epsilon}_i = \mathbf{0}$, $\forall i$.

$\mathcal{O}_i = \{\mathbf{o}_{i1}, \ldots, \mathbf{o}_{iT_i}\}$ as a set of F-dimensional observed vectors, which are assumed to be generated by a GMM, i.e.,

$$p(\mathbf{o}_{it}) = \sum_{c=1}^{C} \lambda_c \mathcal{N}(\mathbf{o}_{it}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \quad t = 1, \ldots, T_i, \tag{12}$$

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.

2.2. Factor Analysis Model

In the i-vector framework [3], the GMM-supervector representing the i-th utterance is assumed to be generated by the following factor analysis model [2]:[2]

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}^{(b)} + \mathbf{T}\mathbf{w}_i + \boldsymbol{\epsilon}_i \tag{13}$$

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, $\mathbf{w}_i$ is the latent factor of dimension D, and $\boldsymbol{\epsilon}_i$ is the residual noise following a zero-mean Gaussian distribution. In practice, the covariance matrix of $\boldsymbol{\epsilon}_i$ is approximated by the UBM's covariance matrix $\boldsymbol{\Sigma}^{(b)}$. Note that Eq. 13 can also be written in a component-wise form:

$$\boldsymbol{\mu}_{ic} = \boldsymbol{\mu}_c^{(b)} + \mathbf{T}_c\mathbf{w}_i + \boldsymbol{\epsilon}_{ic}, \quad c = 1, \ldots, C \tag{14}$$

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 $\mathbf{T}$.

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 may express the joint posterior density of $\mathbf{w}_i$ and indicator variables $y_{i,t,c}$'s, where $t = 1, \ldots, T_i$ and $c = 1, \ldots, C$. $y_{i,t,c}$ specifies which of the C Gaussians generates $\mathbf{o}_{it}$. More specifically, $y_{i,t,c} = 1$ if the c-th mixture generates $\mathbf{o}_{it}$; otherwise, $y_{i,t,c} = 0$. The joint posterior density can be expressed

[2]Some authors, e.g. [3], omit the residue $\boldsymbol{\epsilon}_i$ in Eq. 13. In that case, $\boldsymbol{\mu}_i$ is considered as a vector produced by adding an offset $\mathbf{T}\mathbf{w}_i$ to the UBM's means. To be in line with the notation in Eq. 1, we considered $\boldsymbol{\mu}_i$ as an observed vector so that $\boldsymbol{\epsilon}_i$ represents the error of the linear model in producing $\boldsymbol{\mu}_i$.

in terms of $\mathbf{w}_i$ using the Bayes rule:

$$\begin{aligned} p(\mathbf{w}_i, y_{i,\cdot,\cdot}|\mathcal{O}_i) &\propto p(\mathcal{O}_i|\mathbf{w}_i, y_{i,\cdot,\cdot} = 1)p(y_{i,\cdot,\cdot})p(\mathbf{w}_i) \\ &= \prod_{t=1}^{T}\prod_{c=1}^{C}\left[\lambda_c p(\mathbf{o}_{it}|y_{i,t,c}=1,\mathbf{w}_i)\right]^{y_{i,t,c}} p(\mathbf{w}_i) \quad \text{[2, Eq. 9.38]} \\ &= \underbrace{p(\mathbf{w}_i)\prod_{t=1}^{T}\prod_{c=1}^{C}\left[\mathcal{N}(\mathbf{o}_{it}|\boldsymbol{\mu}_c^{(b)} + \mathbf{T}_c\mathbf{w}_i, \boldsymbol{\Sigma}_c^{(b)})\right]^{y_{i,t,c}}}_{\propto p(\mathbf{w}_i|\mathcal{O}_i)} \lambda_c^{y_{i,t,c}}, \end{aligned} \tag{15}$$

where we have used the fact that for each (i,t)-pair, only one of the $y_{i,t,c}$ for $c = 1,\ldots,C$ equals to 1, the rest are zeros. Extracting terms depending on $\mathbf{w}_i$ from Eq. 15, we obtain

$$\begin{aligned} p(\mathbf{w}_i|\mathcal{O}_i) &\propto \exp\left\{-\frac{1}{2}\sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)} - \mathbf{T}_c\mathbf{w}_i)^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1}(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)} - \mathbf{T}_c\mathbf{w}_i) - \frac{1}{2}\mathbf{w}_i^{\mathsf{T}}\mathbf{w}_i\right\} \\ &= \exp\left\{\mathbf{w}_i^{\mathsf{T}}\sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}\mathbf{T}_c^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1}(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}) - \frac{1}{2}\mathbf{w}_i^{\mathsf{T}}\left(\mathbf{I} + \sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}\mathbf{T}_c^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1}\mathbf{T}_c\right)\mathbf{w}_i\right\}, \end{aligned} \tag{16}$$

where $\mathcal{H}_{ic}$ comprises the frame indexes for which $\mathbf{o}_{it}$ aligned to mixture c. Comparing Eq. 16 with Eq. 9, we obtain the following posterior expectations:

$$\begin{aligned} \langle\mathbf{w}_i|\mathcal{O}_i\rangle &= \mathbf{L}_i^{-1}\sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}\mathbf{T}_c^{\mathsf{T}}\left(\boldsymbol{\Sigma}_c^{(b)}\right)^{-1}(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}) \\ &= \mathbf{L}_i^{-1}\sum_{c=1}^{C}\mathbf{T}_c^{\mathsf{T}}\left(\boldsymbol{\Sigma}_c^{(b)}\right)^{-1}\sum_{t\in\mathcal{H}_{ic}}(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}) \end{aligned} \tag{17}$$

$$\langle\mathbf{w}_i\mathbf{w}_i^{\mathsf{T}}|\mathcal{O}_i\rangle = \mathbf{L}_i^{-1} + \langle\mathbf{w}_i|\mathcal{O}_i\rangle\langle\mathbf{w}_i^{\mathsf{T}}|\mathcal{O}_i\rangle \tag{18}$$

where

$$\mathbf{L}_i = \mathbf{I} + \sum_{c=1}^{C}\sum_{t\in\mathcal{H}_{ic}}\mathbf{T}_c^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1}\mathbf{T}_c. \tag{19}$$

In Eq. 17 and Eq. 19, the sum over $\mathcal{H}_{ic}$ can be evaluated in two ways:

1. *Hard Decisions.* 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

$$\begin{aligned}\gamma_c(\mathbf{o}_{it}) &\equiv \Pr(\text{Mixture} = c|\mathbf{o}_{it}) \\ &= \frac{\lambda_c^{(b)}\mathcal{N}(\mathbf{o}_{it}|\boldsymbol{\mu}_c^{(b)}, \boldsymbol{\Sigma}_c^{(b)})}{\sum_{j=1}^{C}\lambda_j^{(b)}\mathcal{N}(\mathbf{o}_{it}; \boldsymbol{\mu}_j^{(b)}, \boldsymbol{\Sigma}_j^{(b)})}, \quad c = 1, \ldots, C\end{aligned} \tag{20}$$

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

2. *Soft Decisions.* Each frame is aligned to all of the mixtures with degree of alignment according to the posterior probabilities $\gamma_c(\mathbf{o}_{it})$. Then, we have

$$\begin{aligned}\sum_{t\in\mathcal{H}_{ic}} 1 &= \sum_{t=1}^{T_i}\gamma_c(\mathbf{o}_{it}) \\ \sum_{t\in\mathcal{H}_{ic}}(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}) &= \sum_{t=1}^{T_i}\gamma_c(\mathbf{o}_{it})(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}).\end{aligned} \tag{21}$$

Eq. 21 results in the following Baum-Welch statistics:

$$N_{ic} \equiv \sum_{t=1}^{T_i}\gamma_c(\mathbf{o}_{it}) \quad \text{and} \quad \tilde{\mathbf{f}}_{ic} \equiv \sum_{t=1}^{T_i}\gamma_c(\mathbf{o}_{it})(\mathbf{o}_{it} - \boldsymbol{\mu}_c^{(b)}). \tag{22}$$

Substituting Eq. 22 into Eq. 17 and Eq. 19, we have the following expression for i-vectors:

$$\begin{aligned}\mathbf{x}_i \equiv \langle \mathbf{w}_i|\mathcal{O}_i\rangle &= \mathbf{L}_i^{-1}\sum_{c=1}^{C}\mathbf{T}_c^{\mathsf{T}}\left(\boldsymbol{\Sigma}_c^{(b)}\right)^{-1}\tilde{\mathbf{f}}_{ic} \\ &= \mathbf{L}_i^{-1}\mathbf{T}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\tilde{\mathbf{f}}_i\end{aligned} \tag{23}$$

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

$$\mathbf{L}_i = \mathbf{I} + \sum_{c=1}^{C} N_{ic}\mathbf{T}_c^{\mathsf{T}}(\boldsymbol{\Sigma}_c^{(b)})^{-1}\mathbf{T}_c = \mathbf{I} + \mathbf{T}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\mathbf{N}_i\mathbf{T} \tag{24}$$

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.

2.3. Model Training

Model training involves estimating the total variability matrix $\mathbf{T}$ using the EM algorithm. The formulation can be derived based on the EM steps in Section 1. Specifically, using Eqs. 10, 17 and 18, the M-step for estimating $\mathbf{T}$ is

$$\mathbf{T}_c = \left[\sum\nolimits_i \tilde{\mathbf{f}}_{ic}\langle \mathbf{w}_i|\mathcal{O}_i\rangle^{\mathsf{T}}\right]\left[\sum\nolimits_i N_{ic}\langle \mathbf{w}_i\mathbf{w}_i^{\mathsf{T}}|\mathcal{O}_i\rangle\right]^{-1}, \quad c = 1, \ldots, C. \tag{25}$$

2.4. Relationship with MAP Adaptation in GMM-UBM

It can be shown that MAP adaptation in GMM-UBM [5] is a special case of the factor analysis model in Eq. 13. Specifically, when $\mathbf{T}$ is a $CF \times CF$ diagonal matrix denoted as $\mathbf{D}$, the posterior mean of latent factor $\mathbf{z}_i$ is

$$\langle \mathbf{z}_i|\mathcal{O}_i\rangle = \mathbf{L}_i^{-1}\mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\tilde{\mathbf{f}}_i,$$

where $\mathcal{O}_i$ comprises the MFCC vectors of utterance i. Define τ as the relevance factor in MAP adaptation such that $\tau\mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\mathbf{D} = \mathbf{I}$. Then, Eq. 24 becomes

$$\begin{aligned}\mathbf{L}_i &= \tau\mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\mathbf{D} + \mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\mathbf{N}_i\mathbf{D} \\ &= \mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}(\tau\mathbf{I} + \mathbf{N}_i)\mathbf{D}.\end{aligned}$$

As a result, the offset to the UBM's supervector $\boldsymbol{\mu}^{(b)}$ for utterance i is

$$\begin{aligned}\mathbf{D}\langle \mathbf{z}_i|\mathcal{O}_i\rangle &= \mathbf{D}\left[\mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}(\tau\mathbf{I} + \mathbf{N}_i)\mathbf{D}\right]^{-1}\mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\tilde{\mathbf{f}}_i \\ &= \mathbf{D}\left[\mathbf{D}^{-1}(\tau\mathbf{I} + \mathbf{N}_i)^{-1}(\boldsymbol{\Sigma}^{(b)})\mathbf{D}^{-\mathsf{T}}\right]\mathbf{D}^{\mathsf{T}}(\boldsymbol{\Sigma}^{(b)})^{-1}\tilde{\mathbf{f}}_i \\ &= (\tau\mathbf{I} + \mathbf{N}_i)^{-1}\tilde{\mathbf{f}}_i.\end{aligned} \tag{26}$$

Recall that given utterance i, the MAP adaptation of mixture c in GMM-

UBM is [5]:

$$\begin{aligned}\boldsymbol{\mu}_{i,c} &= \alpha_{i,c}E_c(\mathcal{X}_i) + (1-\alpha_{i,c})\boldsymbol{\mu}_c^{(b)} \\ &= \frac{N_{i,c}}{N_{i,c}+\tau}\frac{\mathbf{f}_{i,c}}{N_{i,c}} + \boldsymbol{\mu}_c^{(b)} - \frac{N_{i,c}}{N_{i,c}+\tau}\boldsymbol{\mu}_c^{(b)} \\ &= \boldsymbol{\mu}_c^{(b)} + \frac{1}{N_{i,c}+\tau}(\mathbf{f}_{i,c} - N_{i,c}\boldsymbol{\mu}_c^{(b)}) \\ &= \boldsymbol{\mu}_c^{(b)} + (\tau\mathbf{I} + \mathbf{N}_{i,c})^{-1}\tilde{\mathbf{f}}_{i,c}\end{aligned} \tag{27}$$

where τ is the relevance factor, $N_{i,c}$ is the zeroth order sufficient statistics of mixture c, and $\mathbf{N}_{i,c} = \text{diag}\{N_{i,c}, \ldots, N_{i,c}\}$ is the c-th block of matrix $\mathbf{N}_i$. Note that the offset to UBM's means in Eq. 27 is equivalent to Eq. 26.

2.5. DNN I-Vectors

Recently, Ferrer et al. [6] proposed to replace the UBM in the i-vector extractor by a deep neural network (DNN). The idea is to replace the zeroth order sufficient statistics (Eq. 20) by the outputs of a DNN trained for speech recognition. An advantage of this method is that it enables each test frame to be compared with the training frames for the same phonetic content.

One important characteristic of this method is that the acoustic features for speech recognition in the DNN are not necessary the same as the features for the i-vector extractor. Specifically, denote the acoustic features of utterance i for the DNN and i-vector extractor as $\mathbf{a}_{it}$ and $\mathbf{o}_{it}$, respectively, where $t = 1, \ldots, T_i$. Then, the Baum-Welch statistics in Eq. 22 can be rewritten as:

$$N_{ic} \equiv \sum_{t=1}^{T_i} \gamma_c^{\text{DNN}}(\mathbf{a}_{it}) \quad \text{and} \quad \tilde{\mathbf{f}}_{ic} \equiv \sum_{t=1}^{T_i} \gamma_c^{\text{DNN}}(\mathbf{a}_{it})(\mathbf{o}_{it} - \boldsymbol{\mu}_c), \tag{28}$$

where $\gamma_c^{\text{DNN}}(\mathbf{a}_{it})$'s are the senone posterior probabilities obtained from the DNN's outputs and $\boldsymbol{\mu}_c$'s are the probabilistic weighted sum of speaker recognition features:

$$\boldsymbol{\mu}_c = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i}\gamma_c^{\text{DNN}}(\mathbf{a}_{it})\mathbf{o}_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T_i}\gamma_c^{\text{DNN}}(\mathbf{a}_{it})}.$$

Similar strategy can also be applied to Eq. 24:

$$\mathbf{L}_i = \mathbf{I} + \sum_{c=1}^{C} N_{ic}\mathbf{T}_c^{\mathsf{T}}(\boldsymbol{\Sigma}_c)^{-1}\mathbf{T}_c = \mathbf{I} + \mathbf{T}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_i\mathbf{T}, \tag{29}$$

where N_{ic} is obtained in Eq. 28 and

$$\boldsymbol{\Sigma}_c = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i}\gamma_c^{\text{DNN}}(\mathbf{a}_{it})\mathbf{o}_{it}\mathbf{o}_{it}^{\mathsf{T}}}{\sum_{i=1}^{N}\sum_{t=1}^{T_i}\gamma_c^{\text{DNN}}(\mathbf{a}_{it})} - \boldsymbol{\mu}_c\boldsymbol{\mu}_c^{\mathsf{T}}. \tag{30}$$

Note that $\mathbf{T}_c$ in Eq. 29 should be estimated by replacing $\tilde{\mathbf{f}}_{ic}$ in Eq. 25 with the first-order statistics in Eq. 28. Similarly, the posterior expectation in Eq. 17 should also be computed as follows:

$$\langle \mathbf{w}_i|\mathcal{O}_i\rangle = \mathbf{L}_i^{-1}\sum_{c=1}^{C}\mathbf{T}_c^{\mathsf{T}}\boldsymbol{\Sigma}_c^{-1}\tilde{\mathbf{f}}_{ic}, \tag{31}$$

where $\mathbf{L}_i$ and $\boldsymbol{\Sigma}_c$ are obtained from Eqs. 29 and 30, respectively.

3. Probabilistic LDA

The Gaussian Probabilistic LDA (PLDA) is the supervised version of FA. Consider a data set comprising D-dim vectors $\mathcal{X} = \{\mathbf{x}_{ij}; i = 1,\ldots,N; j = 1,\ldots,H_i\}$ obtained from N classes such that the i-th class contains H_i vectors (e.g., i-vectors). The class-specific vectors should share the same latent factor:

$$\tilde{\mathbf{x}}_i = \tilde{\mathbf{m}} + \tilde{\mathbf{V}}\mathbf{z}_i + \tilde{\boldsymbol{\epsilon}}_i, \qquad \tilde{\mathbf{x}}_i, \tilde{\mathbf{m}} \in \Re^{DH_i},\ \tilde{\mathbf{V}} \in \Re^{DH_i\times M},\ \ \tilde{\boldsymbol{\epsilon}}_i \in \Re^{DH_i}, \tag{32}$$

where $\tilde{\mathbf{x}}_i = [\mathbf{x}_{i1}^{\mathsf{T}}\ \cdots\ \mathbf{x}_{iH_i}^{\mathsf{T}}]^{\mathsf{T}}$, $\tilde{\mathbf{m}} = [\mathbf{m}^{\mathsf{T}}\ \cdots\ \mathbf{m}^{\mathsf{T}}]^{\mathsf{T}}$, $\tilde{\mathbf{V}} = [\mathbf{V}^{\mathsf{T}}\ \cdots\ \mathbf{V}^{\mathsf{T}}]^{\mathsf{T}}$, and $\tilde{\boldsymbol{\epsilon}}_i = [\boldsymbol{\epsilon}_{i1}^{\mathsf{T}}\ \cdots\ \boldsymbol{\epsilon}_{iH_i}^{\mathsf{T}}]^{\mathsf{T}}$.

The E- and M-steps iteratively evaluate and maximize the expectation of the complete likelihood:

$$\begin{aligned} Q(\boldsymbol{\omega}'|\boldsymbol{\omega}) &= \mathbb{E}_{\mathcal{Z}}\{\ln p(\mathcal{X},\mathcal{Z}|\boldsymbol{\omega}')|\mathcal{X},\boldsymbol{\omega}\} \\ &= \mathbb{E}_{\mathcal{Z}}\left\{\sum\nolimits_{ij}\ln\left[p\left(\mathbf{x}_{ij}|\mathbf{z}_i,\boldsymbol{\omega}'\right)p(\mathbf{z}_i)\right]\middle|\mathcal{X},\boldsymbol{\omega}\right\} \\ &= \mathbb{E}_{\mathcal{Z}}\left\{\sum\nolimits_{ij}\ln\left[\mathcal{N}\left(\mathbf{x}_{ij}|\mathbf{m}'+\mathbf{V}'\mathbf{z}_i,\boldsymbol{\Sigma}'\right)\mathcal{N}(\mathbf{z}_i|\mathbf{0},\mathbf{I})\right]\middle|\mathcal{X},\boldsymbol{\omega}\right\}, \end{aligned} \tag{33}$$

where the notation $\sum_{ij}$ is a short-hand form of $\sum_{i=1}^{N}\sum_{j=1}^{H_i}$. It can be shown

that the EM algorithm for PLDA is as follows [4]:

E-step:

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

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

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

M-step:

$$\mathbf{V}' = \left[\sum\nolimits_{ij}(\mathbf{x}_{ij}-\mathbf{m}')\langle \mathbf{z}_i|\mathcal{X}\rangle^{\mathsf{T}}\right]\left[\sum\nolimits_{ij}\langle \mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathcal{X}\rangle\right]^{-1};$$

$$\mathbf{m}' = \frac{\sum_{ij}\mathbf{x}_{ij}}{\sum_i H_i}$$

$$\boldsymbol{\Sigma}' = \frac{1}{\sum_{i=1}^{N}H_i}\left\{\sum_{i=1}^{N}\sum_{j=1}^{H_i}\left[(\mathbf{x}_{ij}-\mathbf{m}')(\mathbf{x}_{ij}-\mathbf{m}')^{\mathsf{T}} - \mathbf{V}'\langle \mathbf{z}_i|\mathcal{X}\rangle(\mathbf{x}_{ij}-\mathbf{m}')^{\mathsf{T}}\right]\right\} \tag{34}$$

References

[1] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling with Sparse Training Data" *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005.

[2] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[4] M.W. Mak and J.T. Chien, "PLDA and Mixture of PLDA Formulations", Supplementary Materials for "Mixture of PLDA for Noise Robust I-Vector Speaker Verification", IEEE/ACM Trans. on Audio Speech and Language Processing, vol. 24, No. 1, pp. 130-142, Jan. 2016.", http://bioinfo.eie.polyu.edu.hk/mPLDA/SuppMaterials.pdf.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[6] L. Ferrer, Yun Lei, M. McLaren and N. Scheffer, "Study of Senone-Based Deep Neural Network Approaches for Spoken Language Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 105-116, Jan. 2016.