Unsupervised Adaptation of SPLDA

Jesús Villalba

Communications Technology Group (GTC),
Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
villalba@unizar.es
June 19, 2013

1 Introduction

In this document we present a Variational Bayes solution to adapt a SPLDA [1] model to a new domain by using unlabelled data. We assume that we count with a labelled dataset (for example Switchboard) to initialise the model.

2 The Model

2.1 SPLDA

SPLDA is a linear generative model where an i-vector $\phi_j$ of speaker $i$ can be written as:

$$\phi_j = \mu + Vy_i + \epsilon_j \quad (1)$$

where $\mu$ is a speaker independent mean, $V$ is the eigen-voices matrix, $y_i$ is the speaker factor vector, and $\epsilon$ is a channel offset.

We assume the following priors for $y$ and $\epsilon$:

$$y_i \sim N(y_i|0, I) \quad (2)$$

$$\epsilon_j \sim N(\epsilon_j|0, W^{-1}) \quad (3)$$

where $N$ denotes a Gaussian distribution; $W$ is the within class precision matrix.

Figure 1 shows the case where the development dataset is split into two parts: one part where the speaker labels are known (supervised) and another with unknown labels (unsupervised).

We introduce the variables involved:

- Let $\Phi_d$ be the i-vectors of the supervised dataset.
- Let $\Phi$ be the i-vectors of the unsupervised dataset.
- Let $\Phi_i$ be the i-vectors belonging to the speaker $i$.
- Let $Y_d$ be the speaker identity variables of the supervised dataset.
- Let $Y$ be the speaker identity variables of the unsupervised dataset.
- Let $\theta_d$ be the labelling of the supervised dataset. It partitions the $N_d$ i-vectors into $M_d$ speakers.
- Let $\theta$ be the labelling of the unsupervised dataset. It partitions the $N$ i-vectors into $M$ speakers.

$\theta_j$ is a latent variable comprising a 1-of-$M$ binary vector with elements $\theta_{ji}$ with $i = 1, \ldots, M$. This
variable is equivalent to the cluster occupations of a GMM. The conditional distribution of \( \theta \) given the weights of the mixture is:

\[
P(\theta|\pi) = \prod_{j=1}^{N} \prod_{i=1}^{M} \pi_{\theta_{ji}}.
\] (4)

- Let \( \pi_{\theta} \) be the weights of the mixture. We choose a Dirichlet prior for the weights:

\[
P(\pi_{\theta}|\tau_0) = \text{Dir}(\pi_{\theta}|\tau_0) = C(\tau_0) \prod_{i=1}^{M} \pi_{\theta_{i}}^{\tau_0 - 1}
\] (5)

where by symmetry we have chosen the same parameter \( \tau_0 \) for each of the components, and \( C(\tau_0) \) is the normalisation constant for the Dirichlet distribution defined as

\[
C(\tau_0) = \frac{\Gamma(M\tau_0)}{\Gamma(\tau_0)^{M}}
\] (6)

and \( \Gamma \) is the Gamma function.

- Let \( d \) be the i-vector dimension.

- Let \( n_y \) be the speaker factor dimension.

- Let \( \mathcal{M} = (\mu, V, W) \) be the set of all the SPLDA parameters. In the most general case, we can assume that the parameters of the model are also hidden variables with prior and posterior distributions.

### 2.2 Sufficient Statistics

We define some statistics for speaker \( i \) in the unsupervised dataset:

\[
N_i = \sum_{j=1}^{N} \theta_{ji}
\] (7)

\[
F_i = \sum_{j=1}^{N} \theta_{ji} \phi_j
\] (8)

\[
S_i = \sum_{j=1}^{N} \theta_{ji} \phi_j^T \phi_j.
\] (9)
We define the centered statistics as

\[ F_i = F_i - N_i \mu \]  \hspace{1cm} (10)

\[ S_i = \sum_{j=1}^N \theta_{ji} (\phi_j - \mu) (\phi_j - \mu)^T = S_i - \mu F_i^T - F_i \mu^T + N_i \mu \mu^T. \]  \hspace{1cm} (11)

We define the global statistics

\[ N = \sum_{i=1}^M N_i \]  \hspace{1cm} (12)

\[ F = \sum_{i=1}^M F_i \]  \hspace{1cm} (13)

\[ F = \sum_{i=1}^M F_i \]  \hspace{1cm} (14)

\[ S = \sum_{i=1}^M S_i \]  \hspace{1cm} (15)

\[ S = \sum_{i=1}^M S_i. \]  \hspace{1cm} (16)

Equally, we can define statistics for the supervised dataset: \( N_d, F_d, S_d \), etc.

### 2.3 Data conditional likelihood

The likelihood of the data given the hidden variables for speaker \( i \) is

\[
\ln P(\Phi_i|y_i, \theta, \mu, V, W) = \sum_{j=1}^N \theta_{ji} \ln N(\phi_j|\mu + Vy_i, W^{-1})
\]

\[
= \frac{N_i}{2} \ln \frac{W}{2\pi} - \frac{1}{2} \sum_{j=1}^N \theta_{ji}(\phi_j - \mu - Vy_i)^T W (\phi_j - \mu - Vy_i)
\]

\[
= \frac{N_i}{2} \ln \frac{W}{2\pi} - \frac{1}{2} \text{tr} (WS_i) + y_i^T V^T WF_i - \frac{N}{2} y_i^T V^T W V y_i. \]  \hspace{1cm} (18)

We can also write this likelihood as:

\[
\ln P(\Phi_i|y_i, \theta, \mu, V, W) = \frac{N_i}{2} \ln \frac{W}{2\pi} - \frac{1}{2} \text{tr} (WS_i) + y_i^T V^T WF_i - \frac{N}{2} y_i^T V^T W V y_i. \]  \hspace{1cm} (19)

If we define:

\[
\tilde{y}_i = \begin{bmatrix} y_i \\ 1 \end{bmatrix}, \quad \tilde{V} = \begin{bmatrix} V & \mu \end{bmatrix}
\]

we can write it as

\[
\ln P(\Phi_i|y_i, \theta, \mu, V, W) = \sum_{j=1}^N \theta_{ji} \ln N(\phi_j|\tilde{\Phi}_i, W^{-1})
\]

\[
= \frac{N_i}{2} \ln \frac{W}{2\pi} - \frac{1}{2} \text{tr} (WS_i - 2F_i \mu^T + N_i \mu \mu^T)
\]

\[-2(F_i - N_i \mu) y_i^T V^T + N_i V y_i y_i^T V^T. \]  \hspace{1cm} (21)

If we define:

\[
\tilde{y}_i = \begin{bmatrix} y_i \\ 1 \end{bmatrix}, \quad \tilde{V} = \begin{bmatrix} V & \mu \end{bmatrix}
\]

we can write it as

\[
\ln P(\Phi_i|y_i, \theta, \mu, V, W) = \sum_{j=1}^N \theta_{ji} \ln N(\phi_j|\tilde{\Phi}_i, W^{-1})
\]

\[
= \frac{N_i}{2} \ln \frac{W}{2\pi} - \frac{1}{2} \text{tr} (WS_i - 2F_i \mu^T + N_i \mu \mu^T)
\]

\[-2(F_i - N_i \mu) y_i^T V^T + N_i V y_i y_i^T V^T. \]  \hspace{1cm} (24)
3 Variational Inference with Point Estimates of $\mu$, $V$ and $W$

As first approximation, we assume a simplified model where we take point estimates of the parameters $\mu$, $V$ and $W$. In this case, the graphical model simplifies to the one in Figure 2.

In this model, $y_i$, $y_{di}$ and $\theta_{ij}$ are the only hidden variables. $V$, $\mu$ and $W$ are hyperparameters that can be obtained by maximising the VB lower bound.

3.1 Variational Distributions

We write the joint distribution of the observed and latent factors:

$$P(\Phi, Y, Y_d, \theta, \pi) = P(\Phi | Y, \theta, \mu, V, W) P(Y) P(\theta | \pi) P(\pi | \tau_0).$$

Following, the conditioning on $(\theta_d, \tau_0, \mu, V, W)$ will be dropped for convenience.

Now, we consider the partition of the posterior:

$$P(Y, Y_d, \theta, \pi|\Phi, \Phi_d) \approx q(Y, Y_d, \theta) = q(Y, Y_d) q(\theta) q(\pi).$$

The optimum for $q^*(Y, Y_d)$:

$$\ln q^*(Y, Y_d) = E_{\Phi, \pi_d} [\ln P(\Phi, \Phi_d, Y, Y_d, \theta, \pi_d)] + \text{const}$$

$$= E_{\theta} [\ln P(\Phi | Y, \theta)] + \ln P(Y) + \ln P(\Phi_d | Y_d) + \ln P(Y_d) + \text{const}$$

$$= \sum_{i=1}^{M} y_i^T V^T W E [F_i] - \frac{1}{2} y_i^T \left( I + E[N_i] V^T W V \right) y_i$$

$$- \sum_{i=1}^{M_d} y_{di}^T V^T W E_{di} - \frac{1}{2} y_{di}^T \left( I + N_{di} V^T W V \right) y_{di} + \text{const}$$

Figure 2: BN for SPLDA with point estimates of the model parameters.
Therefore $q^* (Y, Y_d)$ is a product of Gaussian distributions.

$$q^* (Y, Y_d) = \prod_{i=1}^{M} N (y_i | \bar{y}_i, L_{y_i}^{-1}) \prod_{i=1}^{M} N (y_d_i | \bar{y}_{d_i}, L_{y_d_i}^{-1})$$  \hspace{1cm} (30)$$

$$L_{y_i} = I + E [N] V^T W V$$  \hspace{1cm} (31)$$

$$\bar{y}_i = L_{y_i}^{-1} V^T W E [F_i]$$  \hspace{1cm} (32)$$

$$E [N_i] = \sum_{j=1}^{N} E [\theta_{ji}]$$  \hspace{1cm} (33)$$

$$E [F_i] = \sum_{j=1}^{N} E [\theta_{ji}] (\phi_{j} - \mu)$$  \hspace{1cm} (34)$$

$$L_{y_d_i} = I + N_d_i V^T W V$$  \hspace{1cm} (35)$$

$$\bar{y}_{d_i} = L_{y_d_i}^{-1} V^T W F_{d_i}$$  \hspace{1cm} (36)$$

The optimum for $q^* (\theta)$:

$$\ln q^* (\theta) = E_Y Y_d, \pi_\theta \left[ \ln P (\Phi, \Phi_d, Y, Y_d, \theta, \pi_\theta) \right]$$  \hspace{1cm} (37)$$

$$= E_Y \ln P (\Phi | Y, \theta) + E_{\pi_\theta} \ln P (\theta | \pi_\theta) + \text{const}$$  \hspace{1cm} (38)$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{M} \theta_{ji} \left[ \frac{1}{2} \ln \frac{W}{2\pi} - \frac{1}{2} E [(\phi_{j} - \mu - V y_i)^T W (\phi_{j} - \mu - V y_i)] + E [\ln \pi_\theta] \right] + \text{const}$$  \hspace{1cm} (39)$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{M} \theta_{ji} \left[ \frac{1}{2} \ln \frac{W}{2\pi} - \frac{1}{2} (\phi_{j} - \mu)^T W (\phi_{j} - \mu) + E [y_i]^T V^T W (\phi_{j} - \mu) \right.$$  \hspace{1cm} (40)$$

$$- \frac{1}{2} \text{tr} (V^T W V E [y_i, y_i^T]) + E [\ln \pi_\theta] \right] + \text{const}.$$

Taking exponentials in both sides:

$$q^* (\theta) = \prod_{j=1}^{N} \prod_{i=1}^{M} r_{ji}^{\theta_{ji}}$$  \hspace{1cm} (41)$$

where

$$r_{ji} = \frac{\theta_{ji}}{\sum_{j=1}^{N} \theta_{ji}}$$  \hspace{1cm} (42)$$

$$\ln \theta_{ji} = \frac{1}{2} \ln \frac{W}{2\pi} - \frac{1}{2} (\phi_{j} - \mu)^T W (\phi_{j} - \mu) + E [y_i]^T V^T W (\phi_{j} - \mu)$$  \hspace{1cm} (43)$$

$$- \frac{1}{2} \text{tr} (V^T W V E [y_i, y_i^T]) + E [\ln \pi_\theta]$$  \hspace{1cm} (44)$$

The optimum for $q^* (\pi_\theta)$:

$$\ln q^* (\pi_\theta) = E_Y Y_d, \theta \left[ \ln P (\Phi, \Phi_d, Y, Y_d, \theta, \pi_\theta) \right]$$  \hspace{1cm} (45)$$

$$= E_{\theta} \ln P (\theta | \pi_\theta) + \ln P (\pi_\theta | \tau_0) + \text{const}$$  \hspace{1cm} (46)$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{M} E [\theta_{ji}] \ln \pi_{\theta_i} + (\tau_0 - 1) \sum_{i=1}^{M} \ln \pi_{\theta_i} + \text{const}$$  \hspace{1cm} (47)$$

$$= \sum_{i=1}^{M} (E [N_i] + \tau_0 - 1) \ln \pi_{\theta_i}.$$  \hspace{1cm} (48)$$

Thus:

$$q^* (\pi_\theta) = \text{Dir}(\pi_\theta | \tau) = C(\tau) \prod_{i=1}^{M} \pi_{\theta_i}^{\tau - 1}$$  \hspace{1cm} (49)$$
\[ \tau_i = \mathbb{E}[N_i] + \tau_0 \]  
\[ C(\tau) = \frac{\Gamma\left(\sum_{i=1}^{M} \tau_i\right)}{\Pi_{i=1}^{M} \Gamma(\tau_i)} \]  

Finally, we evaluate the expectations:

\[ \mathbb{E}[y_i] = \varphi_i \]  
\[ \mathbb{E}[y_i y_i^T] = \mathbf{L}^{-1} + \varphi_i \varphi_i^T \]  
\[ \mathbb{E}[\tilde{y}_i \tilde{y}_i^T] = \begin{bmatrix} \mathbb{E}[y_i y_i^T] & \mathbb{E}[y_i] \\ \mathbb{E}[y_i^T] & 1 \end{bmatrix} \]  
\[ \mathbb{E}[\theta_{ji}] = r_{ji} \]  
\[ \mathbb{E}[\pi_{\theta_i}] = \frac{\tau_i}{\sum_{i=1}^{M} \tau_i} \]  
\[ \mathbb{E}[\ln \pi_{\theta_i}] = \psi(\tau_i) - \psi\left(\sum_{i=1}^{M} \tau_i\right) \]  

3.1.1 Distributions with deterministic annealing

If we use annealing, for a parameter \( \kappa \), we have

\[ q^*(Y, Y_d) = \prod_{i=1}^{M} \mathcal{N}(y_i|\varphi_i, 1/\kappa \mathbf{L}_{\varphi_i}^{-1}) \prod_{i=1}^{M} \mathcal{N}(y_{d,i}|\varphi_{d,i}, 1/\kappa \mathbf{L}_{\varphi_{d,i}}^{-1}) \]  

\[ q^*(\theta) = \prod_{j=1}^{N} \prod_{i=1}^{M} \theta_{ji}^{r_{ji}} \]  

where

\[ r_{ji} = \frac{g_{ji}^{\theta}}{\sum_{i=1}^{M} \tau_i} \]  
\[ q^*(\pi_{\theta}) = \text{Dir}(\pi_{\theta}|\tau) = C(\tau) \prod_{i=1}^{M} \pi_{\theta_i}^{\tau_i-1} \]  

where

\[ \tau_i = \kappa(\mathbb{E}[N_i] + \tau_0 - 1) + 1 \]

3.2 Variational lower bound

The lower bound is given by:

\[ \mathcal{L} = \mathbb{E}_{\Phi, \theta}[\ln P(\Phi, \theta)] + \mathbb{E}_Y[\ln P(Y)] + \mathbb{E}_{\theta, \pi_{\theta}}[\ln P(\theta|\pi_{\theta})] + \mathbb{E}_{\sigma_{\pi}}[\ln P(\pi_{\theta})] \]
\[ + \mathbb{E}_{Y_d}[\ln P(\Phi_d|Y_d)] + \mathbb{E}_{Y_d}[\ln P(Y_d)] \]
\[ - \mathbb{E}_Y[\ln q(Y)] - \mathbb{E}_{\theta}[\ln q(\theta)] - \mathbb{E}_{\sigma_{\pi}}[\ln q(\pi_{\theta})] - \mathbb{E}_{Y_d}[\ln q(Y_d)] \]  

The term \( \mathbb{E}_{\Phi, \theta}[\ln P(\Phi|Y, \theta)] \):

\[ \mathbb{E}_{\Phi, \theta}[\ln P(\Phi|Y, \theta)] = \frac{\mathbb{E}[N]}{2} \ln \left| \frac{W}{2\pi} \right| \]
\[ - \frac{1}{2} \text{tr} \left( W \left( E[S] - \sum_{i=1}^{M} E[F_i] E[y_i]^{T} \hat{V}^{T} + \hat{V} \sum_{i=1}^{M} E[N_i] E[y_i^{i}] \hat{V}^{T} \right) \right) \]
We define

\[ C_{\bar{y}} = \sum_{i=1}^{M} E[F_i] E[\bar{y}_i]^T \]  

(65)

\[ R_{\bar{y}} = \sum_{i=1}^{M} E[N_i] E[\bar{y}_i \bar{y}_i^T] \]  

(66)

Then

\[ E_{Y,\theta}[\ln P(\Phi|Y, \theta)] = \frac{E[N]}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \text{tr} \left( W \left( E[S] - 2C_{\bar{y}} \bar{V}^T + \bar{V} R_{\bar{y}} \bar{V}^T \right) \right). \]  

(67)

The term \( E_{Y,d}[\ln P(\Phi_d|Y_d)] \):

\[ E_{Y,d}[\ln P(\Phi_d|Y_d)] = \frac{N_d}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \text{tr} \left( W \left( S_d - 2C_{\bar{y}_d} \bar{V}_d^T + \bar{V}_d R_{\bar{y}_d} \bar{V}_d^T \right) \right). \]  

(68)

where

\[ C_{\bar{y}_d} = \sum_{i=1}^{M_d} F_d E[\bar{y}_d_i]^T \]  

(69)

\[ R_{\bar{y}_d} = \sum_{i=1}^{M_d} N_d E[\bar{y}_d_i \bar{y}_d_i^T] \]  

(70)

The term \( E_Y[\ln P(Y)] \):

\[ E_Y[\ln P(Y)] = - \frac{Mn_y}{2} \ln(2\pi) - \frac{1}{2} \text{tr} (P_Y) \]  

(71)

where

\[ P = \sum_{i=1}^{M} E[y_i y_i^T] \]  

(72)

The term \( E_{Y,d}[\ln P(Y_d)] \):

\[ E_{Y,d}[\ln P(Y_d)] = - \frac{M_d n_d}{2} \ln(2\pi) - \frac{1}{2} \text{tr} (P_{Y,d}) \]  

(73)

where

\[ P_{Y,d} = \sum_{i=1}^{M_d} E[y_d_i y_d_i^T] \]  

(74)

The term \( E_{\theta,\pi_o}[\ln P(\theta|\pi_o)] \):

\[ E_{\theta,\pi_o}[\ln P(\theta|\pi_o)] = \sum_{j=1}^{N} \sum_{i=1}^{M} r_{ji} E[\ln \pi_o_i] \]  

(75)

The term \( E_{\pi_o}[\ln P(\pi_o)] \):

\[ E_{\pi_o}[\ln P(\pi_o)] = \ln C(\tau_0) + (\tau_0 - 1) \sum_{i=1}^{M} E[\ln \pi_o_i] \]  

(76)

The term \( E_Y[\ln q(Y)] \):

\[ E_Y[\ln q(Y)] = - \frac{Mn_y}{2} \ln(2\pi + 1) + \frac{1}{2} \sum_{i=1}^{M} E[\ln L_{y_i}] \]  

(77)
The term $E_{Y_d} [\ln q (Y_d)]$:

$$E_{Y_d} [\ln q (Y_d)] = - \frac{M_d n_d}{2} \left( \ln(2\pi) + 1 \right) + \frac{1}{2} \sum_{i=1}^{M_d} \ln |L_{y_i}|$$

(78)

The term $E_\theta [\ln q (\theta)]$:

$$E_\theta [\ln q (\theta)] = \sum_{j=1}^N \sum_{i=1}^M r_{ji} \ln r_{ji}$$

(79)

The term $E_{\pi_\theta} [\ln q (\pi_\theta)]$:

$$E_{\pi_\theta} [\ln q (\pi_\theta)] = \ln C(\tau) + \sum_{i=1}^M (\tau_i - 1) E [\ln \pi_{\theta_i}]$$

(80)

### 3.3 Hyperparameter optimisation

We can obtain the hyperparameters ($\tau_0, \mu, V, W$) by maximising the lower bound. We control the weight of each of the databases on the estimation by introducing the parameter $\eta \leq 1$ into the lower bound expression:

$$\mathcal{L}(\mu, V, W, \tau_0) = E_{Y, \theta} [\ln P (\Phi | Y, \theta)] + E_{\pi_\theta} [\ln P (\pi_\theta)] + \eta E_{Y_d} [\ln P (\Phi_d | Y_d)] + \text{const}$$

(81)

We derive for $\tilde{V}$:

$$\frac{\partial \mathcal{L}}{\partial \tilde{V}} = C_{\tilde{\Psi}} + \eta C_{\tilde{\Psi}d} - \tilde{V} (R_{\tilde{\Psi}} + \eta R_{\tilde{\Psi}d}) = 0 \implies \tilde{V} = C_{\tilde{\Psi}} R_{\tilde{\Psi}}^{-1}$$

(82)

(83)

where

$$C_{\tilde{\Psi}} = C_{\Psi} + \eta C_{\Psi_d}$$

(84)

$$R_{\tilde{\Psi}} = R_{\Psi} + \eta R_{\Psi_d}$$

(85)

We derive for $W$:

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{E [N] + \eta N_d}{2} \left( 2 W^{-1} - \text{diag} (W^{-1}) \right) - \frac{1}{2} (K + K^T - \text{diag} (K))$$

(86)

where

$$K = E [S] + \eta S_d - 2 C_{\tilde{\Psi}} \tilde{V} + \tilde{V} R_{\tilde{\Psi}} \tilde{V}^T$$

(87)

Then

$$W^{-1} = \frac{1}{E [N] + \eta N_d} \frac{K + K^T}{2}$$

(88)

We derive for $\tau_0$:

$$\frac{\partial \mathcal{L}}{\partial \tau_0} = M \left( \psi (M \tau_0) - \psi (\tau_0) \right) + \sum_{i=1}^M E [\ln \pi_{\theta_i}] = 0$$

(89)

We define $\tau_0 = \exp(\tilde{\tau}_0)$ and

$$f(\tau_0) = \psi (M \tau_0) - \psi (\tau_0) + g = 0$$

(90)

$$g = \frac{1}{M} \sum_{i=1}^M E [\ln \pi_{\theta_i}] = 0 .$$

(91)
We can solve for $\tilde{\tau}_0$ by Newton-Rhapson iterations:

$$\tilde{\tau}_{0\text{new}} = \tilde{\tau}_0 - \frac{f(\tilde{\tau}_0)}{f'(\tilde{\tau}_0)}$$

$$= \tilde{\tau}_0 - \frac{\psi(M\tilde{\tau}_0) - \psi(\tau_0) + g}{\tau_0 (\psi'(M\tilde{\tau}_0) - \psi'(\tau_0))}$$

(92)

Taking exponentials in both sides:

$$\tau_{0\text{new}} = \tau_0 \exp \left( - \frac{\psi(M\tilde{\tau}_0) - \psi(\tau_0) + g}{\tau_0 (\psi'(M\tilde{\tau}_0) - \psi'(\tau_0))} \right)$$

(93)

### 3.4 Minimum divergence

We assume a more general prior for the hidden variables:

$$P(y) = \mathcal{N}(y|\mu_y, \Lambda_y^{-1})$$

(95)

Then we maximise

$$\mathcal{L}(\mu_y, \Lambda_y) = \sum_{i=1}^{M} \mathbb{E}_Y \left[ \ln \mathcal{N}(y|\mu_y, \Lambda_y^{-1}) \right] + \eta \sum_{i=1}^{M_d} \mathbb{E}_Y \left[ \ln \mathcal{N}(y_d|\mu_y, \Lambda_y^{-1}) \right]$$

$$= \frac{M + \eta M_d}{2} \ln |\Lambda_y|$$

$$- \frac{1}{2} \text{tr} \left( \Lambda_y \left( \sum_{i=1}^{M} \mathbb{E} \left[ (y_i - \mu_y)(y_i - \mu_y)^T \right] + \eta \sum_{i=1}^{M_d} \mathbb{E} \left[ (y_{d_i} - \mu_y)(y_{d_i} - \mu_y)^T \right] \right) \right) + \text{const}$$

(96)

We derive for $\mu_y$:

$$\frac{\partial \mathcal{L}(\mu_y, \Lambda_y)}{\partial \mu_y} = \frac{1}{2} \sum_{i=1}^{M} \Lambda_y \mathbb{E} [y_i - \mu_y] + \frac{\eta}{2} \sum_{i=1}^{M_d} \Lambda_y \mathbb{E} [y_{d_i} - \mu_y] = 0 \quad \implies \quad \mu_y = \frac{1}{M + \eta M_d} \left( \sum_{i=1}^{M} \mathbb{E} [y_i] + \eta \sum_{i=1}^{M_d} \mathbb{E} [y_{d_i}] \right)$$

(97)

(98)

We derive for $\Lambda_y$:

$$\frac{\partial \mathcal{L}(\mu_y, \Lambda_y)}{\partial \Lambda_y} = \frac{M + \eta M_d}{2} (2\Lambda_y^{-1} - \text{diag}(\Lambda_y^{-1})) - \frac{1}{2} (2S - \text{diag}(S)) = 0$$

(99)

(100)

where

$$S = \sum_{i=1}^{M} \mathbb{E} \left[ (y_i - \mu_y)(y_i - \mu_y)^T \right] + \eta \sum_{i=1}^{M_d} \mathbb{E} \left[ (y_{d_i} - \mu_y)(y_{d_i} - \mu_y)^T \right]$$

(101)

Then

$$\Sigma_y = \Lambda_y^{-1} = \frac{1}{M + \eta M_d} (P_y + \eta P_{y_d}) - \mu_y \mu_y^T$$

(102)

To obtain a standard prior for $y$ we transform $\mu$ and $V$ by using

$$\mu' = \mu + V \mu_y$$

$$V' = V (\Sigma_y^{-1}/2)^T$$

(103)

(104)
3.5 Determining the number of speakers

To determine the number of speakers we initialise the algorithm assuming that there is a large number of speakers and after some iterations we eliminate speakers based on heuristics:

- Each i-vector belongs only to one speaker.
- Each speaker has an integer number of i-vectors.
- If several i-vectors have similar $E[\theta]$ for several speakers we can merge the speakers.
- Compare the lower bound for different values of $M$ to determine the best number of speakers.

3.6 Initialise the VB

- The values of $\mu$, $V$ and $W$ can be initialised using the supervised dataset.
- $q(\pi_{\theta})$ can be initialised assuming that all the speakers have the same number of i-vectors.
- $q(\theta)$ can be initialised using AHC or some simple algorithm based on the pairwise scores computed evaluating the initial PLDA model. We should also initialise $q(\theta)$ with the oracle labels and check that the partition does not degrade itself as the algorithm iterates. This will provide an upper bound for the performance of the algorithm.
- Instead of initialising $q(\theta)$ we can initialise $q(Y)$ sampling random speakers from the standard distribution and afterwards, compute $q(\theta)$ given $q(Y)$.

3.7 Combining VB and sampling methods

I am interested in Dan’s idea of combining VB and sampling methods. Instead of computing the i-vector statistics as shown in Equations (33) and (34), we can draw samples $\hat{\theta}_{jk}$, $k = 1, \ldots, K$ from $q(\theta)$. Then, compute $K$ i-vector statistics for speaker $i$ as:

$$N_{ik} = \sum_{j=1}^{N} \hat{\theta}_{jki}$$

$$F_{ik} = \sum_{j=1}^{N} \hat{\theta}_{jki} \phi_j$$

(105)

Thus, the statistics are computed in a way that each i-vector only belongs to one speaker while in the standard VB formulation i-vectors are shared between several clusters. Then, we can follow several strategies:

- Select the sample $k^*$ that maximises the lower bound.
- For sample $k$, obtain the accumulators needed to compute $\mu$, $V$ and $W$ ($R_{\phi}$, $C_{\phi}$, etc), average the accumulators of all the samples and compute the model.
- For each sample $k$, compute a model and average the models. However, I think that averaging the accumulators is more correct.

The drawback of this method is that the computational cost grows linearly with $K$, and we may need a large $K$ to make it work.

4 Variational inference with Gaussian-Gamma priors for $V$, Gaussian for $\mu$ and non-informative prior for $W$

4.1 Model priors

We chose the model priors based on the Bishop’s paper about VB PPCA [2]. We introduce a hierarchical prior $P(V|\alpha)$ over the matrix $V$ governed by a $n_y$ dimensional vector of hyperparameters where $n_y$ is
the dimension of the factors. Each hyperparameter controls one of the columns of the matrix \(V\) through a conditional Gaussian distribution of the form:

\[
P(V|\alpha) = \prod_{q=1}^{n_y} \left(\frac{\alpha_q}{2\pi}\right)^{d/2} \exp\left(-\frac{1}{2} \alpha_q v_q^T v_q\right)
\]

(106)

where \(v_q\) are the columns of \(V\). Each \(\alpha_q\) controls the inverse variance of the corresponding \(v_q\). If a particular \(\alpha_q\) has a posterior distribution concentrated at large values, the corresponding \(v_q\) will tend to be small, and that direction of the latent space will be effectively ‘switched off’.

We define a prior for \(\alpha\):

\[
P(\alpha) = \prod_{q=1}^{n_y} \mathcal{G}(\alpha_q|a, b)
\]

(107)

where \(\mathcal{G}\) denotes the Gamma distribution. Bishop defines broad priors setting \(a = b = 10^{-3}\).

We place a Gaussian prior for the mean \(\mu\):

\[
P(\mu) = \mathcal{N}(\mu|\mu_0, \text{diag}(\beta)^{-1})
\]

(108)

We will consider the case where each dimension has different precision and the case with isotropic precision (\(\text{diag}(\beta) = \beta I\)).

Finally, we use a non-informative prior for \(W\) like in [3].

\[
P(W) = \lim_{k \to 0} W(WW_0/k, k)
\]

(109)

\[
= \alpha |W|^{-(d+1)/2}
\]

(110)

### 4.2 Variational distributions

We write the joint distribution of the observed and latent variables:

\[
P(\Phi, \Phi_d, \mathbf{y}, \mathbf{y}_d, \theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha|\theta_d, \tau_0, \mu_0, \beta, a, b) = P(\Phi|\mathbf{y}, \theta, \mu, \mathbf{V}, \mathbf{W}) P(\mathbf{Y}) P(\theta|\pi_\theta) P(\pi_\theta|\tau_0)
\]

\[
P(V|\alpha) P(\alpha|a, b) P(\mu|\mu_0, \beta) P(W)
\]

(111)

Following, the conditioning on \((\theta_d, \tau_0, \mu_0, \beta, a, b)\) will be dropped for convenience.

Now, we consider the partition of the posterior:

\[
P(\mathbf{y}, \mathbf{y}_d, \theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha|\Phi, \Phi_d) \approx q(\mathbf{y}, \mathbf{y}_d, \theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha)
\]

\[
= q(\mathbf{y}, \mathbf{y}_d) q(\theta) q(\pi_\theta) \prod_{r=1}^{d} q(v_r) q(W) q(\alpha)
\]

(112)

where \(v_r\) is a column vector containing the \(r\)th row of \(\mathbf{V}\). If \(W\) were a diagonal matrix the factorisation \(\prod_{r=1}^{d} q(v_r)\) is not necessary because it arises naturally when solving the posterior. However, for full covariance \(W\), the posterior of \(\text{vec}(\mathbf{V})\) is a Gaussian with a huge full covariance matrix. We force the factorisation to make the problem tractable.

The optimum for \(q^*(\mathbf{y}, \mathbf{y}_d)\):

\[
\ln q^*(\mathbf{y}, \mathbf{y}_d) = \mathbb{E}_{\theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha} [\ln P(\Phi, \Phi_d, \mathbf{y}, \mathbf{y}_d, \theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha)] + \text{const}
\]

(113)

\[
= \mathbb{E}_{\theta, \mu, \mathbf{V}, \mathbf{W}} [\ln P(\Phi|\mathbf{y}, \theta, \mu, \mathbf{V}, \mathbf{W})] + \ln P(\mathbf{Y})
\]

\[
+ \mathbb{E}_{\mu, \mathbf{V}, \mathbf{W}} [\ln P(\Phi_d|\mathbf{y}_d, \mu, \mathbf{V}, \mathbf{W})] + \ln P(\mathbf{Y}_d) + \text{const}
\]

(114)

\[
= \sum_{i=1}^{M} y_i^T \mathbb{E} [\nabla^T W (F_i - N_i, \mu)] - \frac{1}{2} y_i^T (\mathbf{I} + \mathbb{E} [N_i] \mathbb{E} [\nabla^T W W]) y_i
\]

\[
+ \sum_{i=1}^{M_d} y_{di}^T \mathbb{E} [\nabla^T W (F_{di} - N_{di}, \mu)] - \frac{1}{2} y_{di}^T (\mathbf{I} + \mathbb{E} [N_{di}] \mathbb{E} [\nabla^T W W]) y_{di} + \text{const}
\]

(115)
Therefore $q^* (Y, Y_d)$ is a product of Gaussian distributions.

\[
q^* (Y, Y_d) = \prod_{i=1}^{M} \mathcal{N}(y_i | \bar{y}_i, L_{y_i}^{-1}) \prod_{i=1}^{M_d} \mathcal{N}(y_d_i | \bar{y}_{d_i}, L_{y_{d_i}}^{-1})
\]  

(116)

\[
L_{y_i} = I + E[N_i] E[V^TWV]
\]  

(117)

\[
\bar{y}_i = L_{y_i}^{-1} \left( E[V]^T E[W] E[F_i] - E[N_i] E[V^T W \mu] \right)
\]  

(118)

\[
E[N_i] = \sum_{j=1}^{N} E[\theta_{ji}]
\]  

(119)

\[
E[F_i] = \sum_{j=1}^{N} E[\theta_{ji}] \phi_j
\]  

(120)

\[
L_{y_{d_i}} = I + N_d_i E[V^TWV]
\]  

(121)

\[
\bar{y}_{d_i} = L_{y_{d_i}}^{-1} \left( E[V]^T E[W] F_{d_i} - N_d_i E[V^T W \mu] \right)
\]  

(122)

The optimum for $q^* (\theta)$:

\[
\ln q^* (\theta) = E_{Y, Y_d, \pi, \mu, V, W, \alpha} \left[ \ln P (\Phi, \Phi_d, Y, Y_d, \theta, \pi_\theta, \mu, V, W, \alpha) \right] + \text{const}
\]  

(123)

\[
= E_{Y, \mu, V, W} \left[ \ln P (\Phi|Y, \theta, \mu, V, W) \right] + E_{\pi_\theta} \left[ \ln P (\theta|\pi_\theta) \right] + \text{const}
\]  

(124)

\[
= \sum_{j=1}^{N} \sum_{i=1}^{M} \theta_{ji} \left[ \frac{1}{2} E[\ln |W|] - \frac{d}{2} \ln(2\pi) - \frac{1}{2} E \left( (\phi_j - \bar{V}y_i)^T W (\phi_j - \bar{V}y_i) \right) + E[\ln \pi_\theta] \right] + \text{const}
\]  

(125)

Taking exponentials in both sides:

\[
q^* (\theta) = \prod_{j=1}^{N} \prod_{i=1}^{M} r_{ji}^{\theta_{ji}}
\]  

(126)

where

\[
r_{ji} = \frac{\theta_{ji}}{\sum_{i=1}^{M} \theta_{ji}}
\]  

(127)

\[
\ln \theta_{ji} = \frac{1}{2} E[\ln |W|] - \frac{d}{2} \ln(2\pi) - \frac{1}{2} E \left( (\phi_j - \bar{V}y_i)^T W (\phi_j - \bar{V}y_i) \right) + E[\ln \pi_\theta]
\]  

(128)

The optimum for $q^* (\pi_\theta)$:

\[
q^* (\pi_\theta) = \text{Dir}(\pi_\theta | \tau) = C(\tau) \prod_{i=1}^{M} \pi_{\theta_i}^{\tau_i-1}
\]  

(129)

where

\[
\tau_i = E[N_i] + \tau_0
\]  

(130)

\[
C(\tau) = \frac{\Gamma \left( \sum_{i=1}^{M} \tau_i \right)}{\prod_{i=1}^{M} \Gamma (\tau_i)}
\]  

(131)

To compute the optimum for $q^* (\bar{V}_i)$, we, again, introduce the parameter $\eta$ to control the weight of
the supervised dataset.

\[
\ln q^* (\tilde{\mathbf{v}}_r') = E_{\mathbf{Y}, \mathbf{Y}_d, \theta, \pi_e, \mathbf{W}, \alpha, \tilde{\mathbf{v}}_{s, r}} [\ln P (\Phi, \Phi_d, \mathbf{Y}, \mathbf{Y}_d, \theta, \pi_e, \mu, \mathbf{V}, \mathbf{W}, \alpha)] + \text{const} \tag{132}
\]

\[
= E_{\mathbf{Y}, \theta, \mathbf{W}, \tilde{\mathbf{v}}_{s, r}} [\ln P (\Phi | \mathbf{Y}, \theta, \mu, \mathbf{V}, \mathbf{W})]
+ \eta E_{\mathbf{Y}_d, \mathbf{W}, \tilde{\mathbf{v}}_{s, r}} [\ln P (\Phi_d | \mathbf{Y}_d, \mu, \mathbf{V}, \mathbf{W})]
+ E_{\alpha, \tilde{\mathbf{v}}_{s, r}} [\ln P (\mathbf{V} | \alpha)] + E_{\alpha, \tilde{\mathbf{v}}_{s, r}} [\ln P (\mu)] + \text{const}
\tag{133}
\]

\[
= -\frac{1}{2} \text{tr} \left( -2\tilde{\mathbf{v}}'_r \mathbf{C}_r + \sum_{s \neq r} \overline{\mathbf{r}}_{rs} \left( \mathbf{C}_s - E [\tilde{\mathbf{v}}'_s]^T \mathbf{R}'_s \right) + \beta_r \bar{\mu}_0^T \right)
+ \tilde{\mathbf{v}}'_r \tilde{\mathbf{v}}'^T_r \left( \text{diag} (\overline{\alpha}_r) + \overline{\mathbf{r}}_{rs} \mathbf{R}'_s \right)
\tag{134}
\]

where \(\overline{\mathbf{r}}_{rs}\) is the element \(r, s\) of \(E [\mathbf{W}]\):

\[
\mathbf{C}_r = \sum_{i=1}^{M} E [\mathbf{F}_i] E [\tilde{\mathbf{y}}_i]^T
\tag{135}
\]

\[
\mathbf{R}_r = \sum_{i=1}^{M} E [\mathbf{N}_i] E [\tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T]
\tag{136}
\]

\[
\mathbf{C}_{\tilde{y}_d} = \sum_{i=1}^{M_d} \mathbf{F}_d, E [\tilde{\mathbf{y}}_d_i]^T
\tag{137}
\]

\[
\mathbf{R}_{\tilde{y}_d} = \sum_{i=1}^{M_d} \mathbf{N}_d, E [\tilde{\mathbf{y}}_d_i \tilde{\mathbf{y}}_d_i^T]
\tag{138}
\]

\[
\overline{\mathbf{r}}_{rs} = \frac{E [\alpha]}{\beta_r}, \quad \bar{\mu}_0 = \frac{0_{n_s \times 1}}{\mu_0_r}
\tag{141}
\]

and \(\mathbf{C}_r\) is the \(r^{th}\) row of \(\mathbf{C}_r\).

Then \(q^* (\tilde{\mathbf{v}}_r')\) is a Gaussian distribution:

\[
q^* (\tilde{\mathbf{v}}_r') = \mathcal{N} \left( \tilde{\mathbf{v}}'_r | \overline{\mathbf{v}}'_r, \mathbf{L}_{\tilde{\mathbf{v}}_r}^{-1} \right)
\tag{142}
\]

\[
\mathbf{v}'_r = \mathbf{L}_{\tilde{\mathbf{v}}_r} \left( \overline{\mathbf{r}}_{rs} \mathbf{C}_r^T + \sum_{s \neq r} \overline{\mathbf{r}}_{rs} \left( \mathbf{C}_s^T - \mathbf{R}_s \tilde{\mathbf{v}}'_{s, r} \right) + \beta_r \bar{\mu}_0 \right)
\tag{144}
\]

The optimum for \(q^* (\alpha)\):

\[
\ln q^* (\alpha) = E_{\mathbf{Y}, \mathbf{Y}_d, \theta, \pi_e, \mu, \mathbf{V}, \mathbf{W}} [\ln P (\Phi, \Phi_d, \mathbf{Y}, \mathbf{Y}_d, \theta, \pi_e, \mu, \mathbf{V}, \mathbf{W}, \alpha)] + \text{const}
\tag{145}
\]

\[
= E_{\tilde{\mathbf{v}}_r} [\ln P (\mathbf{V} | \alpha)] + \ln P (\alpha | a_\alpha, b_\alpha) + \text{const}
\tag{146}
\]

\[
= \sum_{q=1}^{n_s} \left( \frac{d}{2} + a_\alpha - 1 \right) \ln \alpha_q - \alpha_q \left( b_\alpha + \frac{1}{2} E [\tilde{\mathbf{v}}'_q v_q] \right) + \text{const}
\tag{147}
\]
Then \( q^* (\alpha) \) is a product of Gammas:

\[
q^* (\alpha) = \prod_{q=1}^{n_y} \mathcal{G} \left( \alpha_q | \alpha'_q, b'_\alpha_q \right) \tag{149}
\]

\[
a'_q = a_q + \frac{d}{2} \tag{150}
\]

\[
b'_\alpha = b_\alpha + \frac{1}{2} \mathbb{E} \left[ v_q^T v_q \right] \tag{151}
\]

The optimum for \( q^* (W) \):

\[
\ln q^* (W) = \mathbb{E}_{Y, Y_d, \theta, \pi_\theta, \mu, V, W} \left[ \ln P(\Phi, \Phi_d, Y, Y_d, \theta, \pi_\theta, \mu, V, W, \alpha) \right] + \text{const} \tag{152}
\]

\[
= \mathbb{E}_{Y, \theta, \mu, V} \left[ \ln P(\Phi | Y, \theta, \mu, V, W) \right] + \eta \mathbb{E}_{Y_d, \mu, V} \left[ \ln P(\Phi_d | Y_d, \mu, V, W) \right] + \ln P(W) + \text{const} \tag{153}
\]

\[
= \frac{N'}{2} \ln |W| - \frac{d + 1}{2} \ln |W| - \frac{1}{2} \text{tr} (WK) + \text{const} \tag{154}
\]

where

\[
N' = \mathbb{E}[N] + \eta N_d \tag{155}
\]

\[
K = \mathbb{E}[S] + \eta S_d - C^T \mathbb{E} \left[ V V^T \right] - \mathbb{E} \left[ V \right] C^T + \mathbb{E} \left[ V R_s V^T \right] \tag{156}
\]

Then \( q^* (W) \) is Wishart distributed:

\[
q^* (W) = W \left( W | K^{-1}, N' \right) \quad \text{if } N' > d . \tag{157}
\]

Finally, we evaluate the expectations:

\[
\mathbb{E} [y_i] = \bar{y}_i \tag{158}
\]

\[
\mathbb{E} [y_i y_i^T] = L^{-1}_y + \bar{y} \bar{y}^T \tag{159}
\]

\[
\mathbb{E} [y_i y_i^T] = \begin{bmatrix} \mathbb{E} [y_i y_i^T] & \mathbb{E} [y_i] \\ \mathbb{E} [y_i] & 1 \end{bmatrix} \tag{160}
\]

\[
\mathbb{E} [\theta_{ji}] = r_{ji} \tag{161}
\]

\[
\mathbb{E} [\pi_{\theta}] = \frac{\tau_i}{\sum_{i=1}^{M} \tau_i} \tag{162}
\]

\[
\mathbb{E} [\ln \pi_{\theta}] = \psi (\tau_i) - \psi \left( \sum_{i=1}^{M} \tau_i \right) \tag{163}
\]

\[
\mathbb{E} [\alpha_q] = \frac{\alpha'_q}{b'_{\alpha_q}} \tag{164}
\]

\[
\bar{V} = \mathbb{E} \left[ V \right] = \begin{bmatrix} \bar{v}_1^T \\ \bar{v}_2^T \\ \vdots \\ \bar{v}_d^T \end{bmatrix} \tag{165}
\]

\[
\bar{W} = \mathbb{E} [W] = N' \bar{K}^{-1} \tag{166}
\]
\[ E \left[ v_q^T v_q \right] = \sum_{r=1}^{d} E \left[ v_{rq}^T v_{rq} \right] \]
\[ = \sum_{r=1}^{d} \mathbf{L}_{rr}^{-1} + \mathbf{V}_{rq}^2 \]  
(167)

\[ E \left[ \mathbf{V}^T \mathbf{V} \right] = E \left[ \mathbf{V} \mathbf{W} \mathbf{V}^T \right] \]
\[ = \sum_{r=1}^{d} \sum_{s=1}^{d} \mathbf{w}_{rs} E \left[ v_r v_s^T \right] \]
\[ = \sum_{r=1}^{d} \mathbf{w}_{rr} \mathbf{V}_{r} + \mathbf{V}^T \mathbf{W} \mathbf{V} \]  
(169)

\[ E \left[ \mathbf{V}^T \mathbf{W} \right] = \sum_{r=1}^{d} \mathbf{w}_{rr} \mathbf{V}_{r} + \mathbf{V}^T \mathbf{W} \mathbf{V} \]  
(170)

\[ E \left[ \mathbf{V}^T \mathbf{W} \mu \right] = \sum_{r=1}^{d} \mathbf{w}_{rr} \mathbf{V}_{r} + \mathbf{V}^T \mathbf{W} \mathbf{V} \]  
(171)

\[ E \left[ \mathbf{V}^T \mathbf{W} \right] = \sum_{r=1}^{d} \mathbf{w}_{rr} \mathbf{V}_{r} + \mathbf{V}^T \mathbf{W} \mathbf{V} \]  
(172)

\[ E \left[ \mathbf{V}^T \mathbf{W} \right] = \sum_{r=1}^{d} \mathbf{w}_{rr} \mathbf{V}_{r} + \mathbf{V}^T \mathbf{W} \mathbf{V} \]  
(173)

\[ E \left[ (\phi_j - \mathbf{V} \tilde{y}_i)^T \mathbf{W} (\phi_j - \mathbf{V} \tilde{y}_i) \right] = \phi_j^T \mathbf{W} \phi_j - 2 \phi_j^T \mathbf{W} \mathbf{V} \tilde{y}_i + \text{tr} \left( E \left[ \mathbf{V}^T \mathbf{W} \mathbf{V} \right] E \left[ \tilde{y}_i \tilde{y}_i^T \right] \right) \]
\[ E \left[ \mathbf{V}^T \tilde{y}_i \tilde{y}_i^T \right] = \sum_{r=1}^{d} \sum_{s=1}^{d} \mathbf{R}_{rs} E \left[ \mathbf{v}_r v_s^T \right] \]
\[ = \mathbf{V}^T \mathbf{V} + \text{diag} (\rho) \]  
(174)

\[ \mathbf{V}_{r}^2 = \left[ \begin{array}{c} \mathbf{V}_{r} \mathbf{V}_{r}^T \\ \mathbf{V}_{r} \mathbf{V}_{r}^T \\ \vdots \\ \mathbf{V}_{r} \mathbf{V}_{r}^T \end{array} \right] \]
\[ = \mathbf{L}_{rr}^{-1} \]  
(175)

\[ \mathbf{V}_{r}^2 = \left[ \begin{array}{c} \mathbf{V}_{r} \mathbf{V}_{r}^T \\ \mathbf{V}_{r} \mathbf{V}_{r}^T \\ \vdots \\ \mathbf{V}_{r} \mathbf{V}_{r}^T \end{array} \right] \]
\[ = \mathbf{L}_{rr}^{-1} \]  
(176)

where
\[ \mathbf{V}_r = \left[ \begin{array}{ccc} \mathbf{V}_{r}^2 & \mathbf{V}_{r}^T \mathbf{V}_{r} & \vdots \\ \mathbf{V}_{r}^T \mathbf{V}_{r} & \mathbf{V}_{r}^2 & \vdots \\ \vdots & \vdots & \ddots \end{array} \right] \]
\[ = \mathbf{L}_{rr}^{-1} \mathbf{V}_{r} \]  
(177)

\[ \rho = \left[ \begin{array}{c} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_d \end{array} \right]^T \]  
(178)

\[ \rho_i = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} \left( \mathbf{R}_{rs} \circ \mathbf{L}_{r}^{-1} \right) \]  
(179)

and \( \circ \) is the Hadamard product.

### 4.2.1 Distributions with deterministic annealing

If we use annealing, for a parameter \( \kappa \), we have
\[ q^* (\mathbf{Y}, \mathbf{Y}_d) = \prod_{i=1}^{M} \mathcal{N} \left( \mathbf{y}_i | \mathbf{v}_i, 1/\kappa \mathbf{L}^{-1}_{\mathbf{y}_i} \right) \prod_{i=1}^{M} \mathcal{N} \left( \mathbf{y}_d, \mathbf{v}_d, 1/\kappa \mathbf{L}^{-1}_{\mathbf{y}_d} \right) \]
\[ q^* (\theta) = \prod_{j=1}^{N} \prod_{i=1}^{M} \mathcal{N} \left( \mathbf{v}_i | \mathbf{v}_i, 1/\kappa \mathbf{L}^{-1}_{\mathbf{v}_i} \right) \]
\[ q^* (\pi) = \text{Dir}(\pi | \tau) = C(\tau) \prod_{i=1}^{M} \pi_{\theta_i}^{-1} \]  
(180)

\[ q^* (\pi) = \text{Dir}(\pi | \tau) = C(\tau) \prod_{i=1}^{M} \pi_{\theta_i}^{-1} \]  
(181)

\[ q^* (\pi) = \text{Dir}(\pi | \tau) = C(\tau) \prod_{i=1}^{M} \pi_{\theta_i}^{-1} \]  
(182)

\[ q^* (\pi) = \text{Dir}(\pi | \tau) = C(\tau) \prod_{i=1}^{M} \pi_{\theta_i}^{-1} \]  
(183)
where
\[ \tau_i = \kappa(N_i + \tau_0 - 1) + 1 \quad (184) \]
\[ q^*(\nu') = N(\nu', \nu, 1/\kappa \nu^{-1}) \quad (185) \]
\[ q^*(W) = W[W|1/\kappa K^{-1}, \kappa(N' - d - 1) + d + 1] \quad \text{if } \kappa(N' - d - 1) + 1 > 0 \quad (186) \]
\[ q^*(\alpha_0) = \prod_{q=1}^{n_u} \mathcal{G}(\alpha_0 | a'_0, b'_{\alpha_0}) \quad (187) \]
\[ a'_0 = \kappa \left( a_0 + \frac{d}{2} - 1 \right) + 1 \quad (188) \]
\[ b'_{\alpha_0} = \kappa \left( b_0 + \frac{1}{2} E[V_q^T V_q] \right) \quad (189) \]

4.3 Variational lower bound

The lower bound is given by:
\[ \mathcal{L} = E_{Y, \theta, \mu, V, W} [\ln P(\Phi|Y, \theta, \mu, V, W)] + E_{Y} [\ln P(Y)] + E_{\theta, \pi_0} [\ln P(\theta|\pi_0)] + E_{\pi_0} [\ln P(\pi_0)] + E_{V, \alpha} [\ln P(V|\alpha)] + E_{\alpha} [\ln P(\alpha)] + E_{\mu} [\ln P(\mu)] + E_{W} [\ln P(W)] + \eta E_{Y_d, \mu, V, W} [\ln P(\Phi_d|Y_d, \mu, V, W)] + \eta E_{Y_d} [\ln P(Y_d)] - E_{\psi} [\ln q(\nu) - E_{\alpha} [\ln q(\alpha)] - E_{W} [\ln q(W)] - \eta E_{Y_d} [\ln q(Y_d)]. \quad (190) \]

The term \( E_{Y, \theta, \mu, V, W} [\ln P(\Phi|Y, \theta, \mu, V, W)] \):
\[ E_{Y, \theta, \mu, V, W} [\ln P(\Phi|Y, \theta, \mu, V, W)] = \frac{E[N]}{2} E[\ln |W|] - \frac{E[N] d}{2} \ln(2\pi) \]
\[ - \frac{1}{2} \text{tr} \left( W \left( E[S] - 2C_0 \bar{V}_0^T + E \left[ \bar{V} R_{\bar{Y}_0} \bar{V}^T \right] \right) \right) \quad (191) \]
\[ = \frac{E[N]}{2} \ln W - \frac{E[N] d}{2} \ln(2\pi) - \frac{1}{2} \text{tr} (W E[S]) \]
\[ - \frac{1}{2} \text{tr} \left( -2 \bar{V}^T WC_0 \bar{Y}_0 + E \left[ \bar{V}^T W\bar{V} \right] R_{\bar{Y}_0} \right) \quad (192) \]

where
\[ \ln W = E[\ln |W|] \quad (193) \]
\[ = \sum_{i=1}^{d} \psi \left( \frac{N' + 1 - i}{2} \right) + d \ln 2 + \ln |K^{-1}| \quad (194) \]
and \( \psi \) is the digamma function.

The term \( E_{Y_d, \theta, \mu, V, W} [\ln P(\Phi_d|Y_d, \mu, V, W)] \):
\[ E_{Y_d, \theta, \mu, V, W} [\ln P(\Phi_d|Y_d, \theta, \mu, V, W)] = \frac{N_d}{2} E[\ln |W|] - \frac{N_d d}{2} \ln(2\pi) \]
\[ - \frac{1}{2} \text{tr} \left( W \left( S_d - 2C_{\bar{Y}_d, 0} \bar{V}_d^T + E \left[ \bar{V} R_{\bar{Y}_d} \bar{V}^T \right] \right) \right) \quad (195) \]
\[ = \frac{N_d}{2} \ln W - \frac{N_d d}{2} \ln(2\pi) - \frac{1}{2} \text{tr} (WS_d) \]
\[ - \frac{1}{2} \text{tr} \left( -2 \bar{V}_d^T WC_{\bar{Y}_d} \bar{Y}_d + E \left[ \bar{V}_d^T W\bar{V}_d \right] R_{\bar{Y}_d} \right) \quad (196) \]
The term $E_{V, \alpha} [\ln P (V | \alpha)]$:

$$
E_{V, \alpha} [\ln P (V | \alpha)] = \frac{-n_y d}{2} \ln (2\pi) + \frac{d}{2} n_y \sum_{q=1}^{n_y} E [\ln \alpha_q] - \frac{1}{2} \sum_{q=1}^{n_y} E [\alpha_q] E [v_q^T v_q]
$$

(197)

where

$$
E [\ln \alpha_q] = \psi (a'_\alpha) - \ln b'_\alpha.
$$

(198)

The term $E_{\alpha} [\ln P (\alpha)]$:

$$
E_{\alpha} [\ln P (\alpha)] = n_y (a_\alpha \ln b_\alpha - \ln \Gamma (a_\alpha)) + (a_\alpha - 1) \sum_{q=1}^{n_y} E [\ln \alpha_q] - b_\alpha \sum_{q=1}^{n_y} E [\alpha_q]
$$

(199)

The term $E_{\mu} [\ln P (\mu)]$:

$$
E_{\mu} [\ln P (\mu)] = -\frac{d}{2} \ln (2\pi) + \frac{1}{2} \sum_{r=1}^{d} \ln \beta_r - \frac{1}{2} \sum_{r=1}^{d} \beta_r \left( \Sigma_{\mu_r} + E [\mu_r]^2 - 2\mu_0 E [\mu_r] + \mu_0^2 \right)
$$

(200)

The term $E_{W} [\ln P (W)]$:

$$
E_{W} [\ln P (W)] = -\frac{d+1}{2} \ln W
$$

(201)

The term $E_{\tilde{V}} \left[ \ln q \left( \tilde{V} \right) \right]$:

$$
E_{\tilde{V}} \left[ \ln q \left( \tilde{V} \right) \right] = -\frac{d(n_y + 1)}{2} \ln (2\pi + 1) + \frac{1}{2} \sum_{r=1}^{d} \ln |L_{\tilde{V}_r}|
$$

(202)

The term $E_{\alpha} [\ln q (\alpha)]$:

$$
E_{\alpha} [\ln q (\alpha)] = -\sum_{q=1}^{n_y} H [q (\alpha_q)]
$$

(203)

$$
= n_y ((a'_\alpha - 1) \psi (a'_\alpha) - a'_\alpha - \ln \Gamma (a'_\alpha)) + \sum_{q=1}^{n_y} \ln b'_\alpha
$$

(204)

The term $E_{W} [\ln q (W)]$:

$$
E_{W} [\ln q (W)] = -H [q (W)]
$$

(205)

$$
= \ln B (K^{-1}, N) + \frac{N - d - 1}{2} \ln W - \frac{Nd}{2}
$$

(206)

where

$$
B(A, N) = \frac{1}{2^{N_d/2} Z_{Nd}} |A|^{-N/2}
$$

(207)

$$
Z_{Nd} = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma ((N + 1 - i)/2)
$$

(208)

The expressions for the terms $E_{Y} [\ln P (Y)]$, $E_{Y|d} [\ln P (Y|d)]$, $E_{\theta, \pi} [\ln P (\theta | \pi_\theta)]$, $E_{\pi_\theta} [\ln P (\pi_\theta)]$, $E_{Y} [\ln q (Y)]$, $E_{Y|d} [\ln q (Y|d)]$, $E_{\theta} [\ln q (\theta)]$ and $E_{\pi_\theta} [\ln q (\pi_\theta)]$ are the same as the ones in Section 3.2.

17
4.4 Hyperparameter optimisation

We can set the Hyperparameters \((\tau_0, \mu_0, \beta, a, b)\) manually or estimate them from the development data maximising the lower bound.

\(\tau_0\) can be derived as shown in Section 3.3.

we derive for \(a\):

\[
\frac{\partial L}{\partial a} = n_y \left( \ln b_\alpha + \frac{1}{n_y} \sum_{q=1}^{n_y} \mathbb{E}[\ln \alpha_q] \right) = 0 \implies (209)
\]

\[
\psi(a) = \ln b_\alpha + \frac{1}{n_y} \sum_{q=1}^{n_y} \mathbb{E}[\ln \alpha_q] \tag{210}
\]

We derive for \(b\):

\[
\frac{\partial L}{\partial b_\alpha} = \frac{n_y a_\alpha}{b} - \sum_{q=1}^{n_y} \mathbb{E}[\alpha_q] = 0 \implies (211)
\]

\[
b_\alpha = \left( \frac{1}{n_y a_\alpha} \sum_{q=1}^{n_y} \mathbb{E}[\alpha_q] \right)^{-1} \tag{212}
\]

We solve these equations with the procedure described in [4]. We write

\[
\psi(a) = \ln b + c \tag{213}
\]

\[
b = \frac{a}{d} \tag{214}
\]

where

\[
c = \frac{1}{n_y} \sum_{q=1}^{n_y} \mathbb{E}[\ln \alpha_q] \tag{215}
\]

\[
d = \frac{1}{n_y} \sum_{q=1}^{n_y} \mathbb{E}[\alpha_q] \tag{216}
\]

Then

\[
f(a) = \psi(a) - \ln a + \ln d - c = 0 \tag{217}
\]

We can solve for \(a\) using Newton-Raphson iterations:

\[
a_{new} = a - \frac{f(a)}{f'(a)} \tag{218}
\]

\[
= a \left( 1 - \frac{\psi'(a) - \ln a + \ln d - c}{a\psi''(a) - 1} \right) \tag{219}
\]

This algorithm does not assure that \(a\) remains positive. We can put a minimum value for \(a\). Alternatively we can solve the equation for \(\hat{a}\) such as \(a = c \exp(\hat{a})\).

\[
\hat{a}_{new} = \hat{a} - \frac{f(\hat{a})}{f'(\hat{a})} = \tag{220}
\]

\[
= \hat{a} - \frac{\psi(\hat{a}) - \ln a + \ln d - c}{\psi'(\hat{a})a - 1} \tag{221}
\]

Taking exponential in both sides:

\[
a_{new} = a \exp \left( -\frac{\psi(a) - \ln a + \ln d - c}{\psi'(a)a - 1} \right) \tag{222}
\]
We derive for $\mu_0$:

$$\frac{\partial L}{\partial \mu_0} = 0 \implies \mu_0 = \mathbb{E}[\mu] \tag{223}$$

We derive for $\beta$:

$$\frac{\partial L}{\partial \beta} = 0 \implies \beta^{-1} = \Sigma_{\mu_r} + \mathbb{E}[\mu_r]^2 - 2\mu_0, \mathbb{E}[\mu_r] + \mu_0^2 \tag{225}$$

If we take an isotropic prior for $\mu$:

$$\beta^{-1} = \frac{1}{d} \sum_{r=1}^{d} \Sigma_{\mu_r} + \mathbb{E}[\mu_r]^2 - 2\mu_0, \mathbb{E}[\mu_r] + \mu_0^2 \tag{226}$$

4.5 Some ideas

What we expect from this model is:

- We expect that taking into account the full posterior of the parameters of the SPLDA, we will obtain a better estimation of the labels and the number of speakers.
- The variances of $V$ and $W$ decrease as the number of speakers and segments, respectively, grow. Thus, we expect a larger improvement in cases where we have scarce adaptation data.
- We can analyse, how the labels affect the posteriors of the parameters. I have the intuition that if the labels are wrong the variance of $V$ should be larger than if the labels are right.
- From $q(\alpha)$, we can infer the best value for $n_y$. If the $\mathbb{E}[\alpha_q]$ (prior precision of $v_q$) is large, $v_q$ will tend to be small as can be seen in Equation (106).

References

[1] Jesús Villalba, “SPLDA,” Tech. Rep., University of Zaragoza, Zaragoza, July 2011.

[2] Christopher Bishop, “Variational principal components,” in Proceedings of the 9th International Conference on Artificial Neural Networks, ICANN 99, Edinburgh, Scotland, Sept. 1999, IET, vol. 1, pp. 509–514.

[3] Jesús Villalba, “Fully Bayesian Two-Covariance Model,” Tech. Rep., University of Zaragoza, Zaragoza, Spain, 2010.

[4] Matthew J. Beal, Variational algorithms for approximate Bayesian inference, Ph.D. thesis, University College London, 2003.