Bayesian SPLDA

Jesús Villalba

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

May 15, 2012

1 Introduction

In this document we are going to derive the equations needed to implement a Variational Bayes estimation of the parameters of the SPLDA model [1]. This can be used to adapt the SPLDA from one database to another with few development data or to implement the fully Bayesian recipe [2]. Our approach is similar to Bishop’s VB PPCA in [3].

2 The Model

2.1 SPLDA

SPLDA is a linear generative model represented in Figure 1.

An i-vector $\phi$ of speaker $i$ can be written as:

$$\phi_{ij} = \mu + Vy_i + \epsilon_{ij} \tag{1}$$

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

We assume the following priors for the variables:

$$y_i \sim N(y_i|0, I) \tag{2}$$

$$\epsilon_{ij} \sim N(\epsilon_{ij}|0, W^{-1}) \tag{3}$$

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

We are going to introduce some notation:

- Let $\Phi_d$ be the development i-vectors dataset.
- Let $\Phi_t = \{l, r\}$ be the test i-vectors.
- Let $\Phi$ be any of the previous datasets.
- Let $\theta_d$ be the labelling of the development dataset. It partitions the $N_d$ i-vectors into $M_d$ speakers.
- Let $\theta_t$ be the labelling of the test set, so that $\theta_t \in \{T, N\}$, where $T$ is the hypothesis that $l$ and $r$ belong to the same speaker and $N$ is the hypothesis that they belong to different speakers.
- Let $\theta$ be any of the previous labellings.
- Let $\phi_i$ be the i-vectors belonging to the speaker $i$.
- Let $Y_d$ be the speaker identity variables of the development set. We will have as many identity variables as speakers.
- Let $Y_t$ be the speaker identity variables of the test set.
- Let $Y$ be any of the previous speaker identity variables sets.
- 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 parameters.

3 Sufficient statistics

We define the sufficient statistics for speaker $i$. The zero-order statistic is the number of observations of speaker $i$ $N_i$. The first-order and second-order statistics are

$$
F_i = \sum_{j=1}^{N_i} \phi_{ij} \tag{4}
$$

$$
S_i = \sum_{j=1}^{N_i} \phi_{ij} \phi_{ij}^T \tag{5}
$$

We define the centered statistics as

$$
\bar{F}_i = F_i - N_i \mu \tag{6}
$$

$$
\bar{S}_i = \sum_{j=1}^{N_i} (\phi_{ij} - \mu)(\phi_{ij} - \mu)^T = S_i - \mu F_i^T - F_i \mu^T + N_i \mu \mu^T \tag{7}
$$
We define the global statistics

\[ N = \sum_{i=1}^{M} N_i \] (8)

\[ F = \sum_{i=1}^{M} F_i \] (9)

\[ F = \sum_{i=1}^{M} F_i \] (10)

\[ S = \sum_{i=1}^{M} S_i \] (11)

\[ S = \sum_{i=1}^{M} S_i \] (12)

4 Data conditional likelihood

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

\[
\ln P(\Phi_i | y_i, \mu, V, W) = \sum_{j=1}^{N_i} \ln N \left( \phi_{ij} | \mu + V y_i, W^{-1} \right) 
\] (13)

\[
= \frac{N_i}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \sum_{j=1}^{N_i} (\phi_{ij} - \mu - V y_i)^T W (\phi_{ij} - \mu - V y_i) 
\] (14)

\[
= \frac{N_i}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \text{tr} \left( W S_i \right) + y_i^T V^T W F_i - \frac{N_i}{2} y_i^T V^T W V y_i 
\] (15)

We can write this likelihood in another form:

\[
\ln P(\Phi_i | y_i, \mu, V, W) = \frac{N_i}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \text{tr} \left( W (S_i - 2F_i \mu^T + N_i \mu \mu^T - 2(F_i - N_i \mu) y_i^T V^T + N_i \bar{y}_i y_i^T V^T) \right) 
\] (16)

We can write this likelihood in another form if we define:

\[
\tilde{y}_i = \begin{bmatrix} y_i \\ 1 \end{bmatrix}, \quad \tilde{V} = \begin{bmatrix} V & \mu \end{bmatrix} 
\] (18)

Then

\[
\ln P(\Phi_i | y_i, \mu, V, W) = \sum_{j=1}^{N_i} \ln N \left( \phi_{ij} | \tilde{V} \tilde{y}_i, W^{-1} \right) 
\] (19)

\[
= \frac{N_i}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \sum_{j=1}^{N_i} (\phi_{ij} - \tilde{V} \tilde{y}_i)^T W (\phi_{ij} - \tilde{V} \tilde{y}_i) 
\] (20)

\[
= \frac{N_i}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \text{tr} \left( W S_i \right) + \tilde{y}_i^T \tilde{V}^T W F_i - \frac{N_i}{2} \tilde{y}_i^T \tilde{V}^T W \tilde{y}_i 
\] (21)

\[
= \frac{N_i}{2} \ln \left| \frac{W}{2\pi} \right| - \frac{1}{2} \text{tr} \left( W \left( S_i - 2F_i \tilde{y}_i^T \tilde{V}^T + N_i \tilde{V} \tilde{y}_i \tilde{y}_i^T \tilde{V}^T \right) \right) 
\] (22)

3
5 Variational inference with Gaussian-Gamma priors for V, Gaussian for µ and Wishart for W (informative and non-informative)

5.1 Model priors

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:

$$P(V|\alpha) = \prod_{q=1}^{n_y} \mathcal{G} \left( \frac{\alpha_q}{2\pi} \right)^{d/2} \exp \left( -\frac{1}{2\alpha_q \mathbf{v}_q^T \mathbf{v}_q} \right)$$

(23)

where $\mathbf{v}_q$ are the columns of $V$. Each $\alpha_q$ controls the inverse variance of the corresponding $\mathbf{v}_q$. If a particular $\alpha_q$ has a posterior distribution concentrated at large values, the corresponding $\mathbf{v}_q$ will tend to be small, and that direction of the latent space will be effectively 'switched off'.

We define a prior for $\alpha$:

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

(24)

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

We place a Gaussian prior for the mean $\mu$:

$$P(\mu) = \mathcal{N}(\mu|0, \text{diag}(\beta)^{-1})$$

(25)

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

Finally, we put a Wishart prior on $W$,

$$P(W) = \mathcal{W}(W|\Psi_0, \nu_0)$$

(26)

We can make the Wishart prior non-informative like in [10].

$$P(W) = \lim_{k \to 0} \mathcal{W}(W|W_0/k, k)$$

(27)

$$= a |W|^{-(d+1)/2}$$

(28)

5.2 Variational distributions

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

$$P(\Phi, \Phi', \mathbf{Y}, \mu, V, W, \alpha|\mu_0, \beta, a_0, b_0) = P(\Phi|\Phi', \mu, V, W) P(\Phi) P(\Phi|\alpha) P(\alpha|a, b) P(\mu|\mu_0, \beta) P(W)$$

(29)

Following, the conditioning on $(\mu_0, \beta, a_0, b_0)$ will be dropped for convenience.

Now, we consider the partition of the posterior:

$$P(\mathbf{Y}, \mu, V, W, \alpha|\Phi) \approx q(\mathbf{Y}, \mu, V, W, \alpha) = q(\mathbf{Y}) \prod_{r=1}^{d} q(\mathbf{v}_r^\prime) q(W) q(\alpha)$$

(30)

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

The optimum for $q^*$ ($\mathbf{Y}$):

$$\ln q^*(\mathbf{Y}) = \mathbb{E}_{\mu, V, W, \alpha} [\ln P(\Phi, \Phi', \mu, V, W, \alpha)] + \text{const}$$

(31)

$$= \mathbb{E}_{\mu, V, W} [\ln P(\Phi|\Phi', \mu, V, W)] + \ln P(\mathbf{Y}) + \text{const}$$

(32)

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

(33)
Therefore \( q^* (Y) \) is a product of Gaussian distributions.

\[
q^* (Y) = \prod_{i=1}^{M} N (y_i | \tilde{y}_i, L_{\tilde{y}_i}^{-1})
\]  

(34)

\[
\begin{aligned}
L_{\tilde{y}_i} &= I + N_i E \left[ V^T W V \right] \\
\tilde{y}_i &= L_{\tilde{y}_i}^{-1} E \left[ V^T W (F_i - N_i \mu) \right] \\
&= L_{\tilde{y}_i}^{-1} \left( E [V^T E [W] F_i - N_i E [V^T W \mu] \right)
\end{aligned}
\]  

(35)

(36)

(37)

The optimum for \( q^* (\tilde{y}_i') \):

\[
\ln q^* (\tilde{y}_i') = E_{Y,W,\alpha,\tilde{y}_i' | x} \left[ \ln p (\Phi, Y, \mu, V, \omega, \alpha) \right] + \text{const}
\]

(38)

\[
= E_Y, W, \tilde{y}_i' [\ln P (\Phi | Y, \mu, V, W) + E_{\alpha, \tilde{y}_i' | x} [\ln P (\alpha)] + E_{\mu, \tilde{y}_i' | x} [\ln P (\mu)] + \text{const}
\]

(39)

\[
= -\frac{1}{2} \sum_{i=1}^{M} \text{tr} \left( E [W] \left( -2 F_i E [\tilde{y}_i] + E [\tilde{y}_i'] [\tilde{V}] + N_i E [\tilde{y}_i'] [\tilde{V}^T] \right) \right)
\]

(40)

\[
= -\frac{1}{2} \sum_{i=1}^{M} \text{tr} \left( E [W] \left( -2 CE [\tilde{y}_i'] [\tilde{V}] + E [\tilde{y}_i'] [\tilde{V}^T] \right) \right)
\]

(41)

\[
= -\frac{1}{2} \sum_{i=1}^{M} E [\tilde{y}_i'] [\tilde{V}]^T E [W] C + E [\tilde{y}_i'] [\tilde{V}^T] E [W] \tilde{V}
\]

(42)

\[
= -\frac{1}{2} \sum_{i=1}^{d} \tilde{y}_i E [\tilde{y}_i'] C + \sum_{s \neq r} \tilde{y}_s E [\tilde{y}_s'] R_{\tilde{y}} + \tilde{y}_r E [\tilde{y}_r'] R_{\tilde{y}}
\]

(43)

\[
= -\frac{1}{2} \sum_{i=1}^{d} \tilde{y}_i E [\tilde{y}_i'] C + \sum_{s \neq r} \tilde{y}_s E [\tilde{y}_s'] R_{\tilde{y}} + \tilde{y}_r E [\tilde{y}_r'] R_{\tilde{y}}
\]

(44)

\[
= -\frac{1}{2} \sum_{i=1}^{d} \tilde{y}_i E [\tilde{y}_i'] C + \sum_{s \neq r} \tilde{y}_s E [\tilde{y}_s'] R_{\tilde{y}} + \tilde{y}_r E [\tilde{y}_r'] R_{\tilde{y}}
\]

(45)

\[
= -\frac{1}{2} \sum_{i=1}^{d} \tilde{y}_i E [\tilde{y}_i'] C + \sum_{s \neq r} \tilde{y}_s E [\tilde{y}_s'] R_{\tilde{y}} + \tilde{y}_r E [\tilde{y}_r'] R_{\tilde{y}}
\]

(46)
where $\pi_{rs}$ is the element $r,s$ of $E[W],$
\[
C = \sum_{i=1}^{M} F_i E[y_i]^T
\]
(47)
\[
R_\tilde{y} = \sum_{i=1}^{M} N_i E[\tilde{y}_i \tilde{y}_i^T]
\]
(48)
\[
E[\tilde{y}_i \tilde{y}_i^T] = \begin{bmatrix}
E[y_i y_i^T] & E[y_i] \\
E[y_i] & 1
\end{bmatrix}
\]
(49)
\[
\bar{v}_{\alpha r} = \begin{bmatrix}
E[\alpha] \\
\beta_r
\end{bmatrix}
\]
\[\tilde{\mu}_0 = \begin{bmatrix}0_{n_y \times 1}\end{bmatrix}
\]
(50)
and $C_r$ is the $r^{th}$ row of $C$.

Then $q^*(\tilde{v}^r)$ is a Gaussian distribution:
\[
q^*(\tilde{v}^r) = N(\tilde{v}^r | \overline{v}_r, L^{-1}_r)
\]
(51)
\[
L_{\overline{v}_r} = \text{diag}(\overline{a}_r) + \overline{w}_{rr} R_{\tilde{y}}
\]
(52)
\[
\overline{v}_r = L^{-1}_r \left( \overline{w}_{rr} C_r^T + \sum_{s \neq r} \overline{w}_{rs} \left( C_r^T - R_{\tilde{y}} \overline{v}_s \right) + \beta_r \tilde{\mu}_0 \right)
\]
(53)

The optimum for $q^*(\alpha)$:
\[
\ln q^*(\alpha) = E_{\tilde{Y}, \mu, \nu, W}[\ln P(\Phi, Y, \mu, \nu, W, \alpha)] + \text{const}
\]
(54)
\[
= E_{\tilde{Y}}[\ln P(\nu | \alpha)] + \ln P(\alpha | a_\alpha, b_\alpha) + \text{const}
\]
(55)
\[
= \sum_{q=1}^{n_y} \frac{d}{2} \ln \alpha_q - \frac{1}{2} a_\alpha E[v_q^T v_q] + (a_\alpha - 1) \ln \alpha_q - b_\alpha \alpha_q + \text{const}
\]
(56)
\[
= \sum_{q=1}^{n_y} \left( \frac{d}{2} + a_\alpha - 1 \right) \ln \alpha_q - \alpha_q \left( b_\alpha + \frac{1}{2} E[v_q^T v_q] \right) + \text{const}
\]
(57)

Then $q^*(\alpha)$ is a product of Gammas:
\[
q^*(\alpha) = \prod_{q=1}^{n_y} \mathcal{G}(\alpha_q | a'_\alpha, b'_\alpha)
\]
(59)
\[
a'_\alpha = a_\alpha + \frac{d}{2}
\]
(60)
\[
b'_\alpha = b_\alpha + \frac{1}{2} E[v_q^T v_q]
\]
(61)

The optimum for $q^*(W)$ in the non-informative case:
\[
\ln q^*(W) = E_{\tilde{Y}, \mu, \nu, \alpha}[\ln P(\Phi, Y, \mu, \nu, W, \alpha)] + \text{const}
\]
(62)
\[
= E_{\tilde{Y}, \mu, \nu}[\ln P(\Phi, Y, \mu, \nu, W)] + \ln P(W) + \text{const}
\]
(63)
\[
= N \frac{d}{2} \ln |W| - \frac{d+1}{2} \ln |W| - \frac{1}{2} \text{tr}(WK) + \text{const}
\]
(64)
where
\[
K = \sum_{i=1}^{M} E \left[ S_i - F_i \tilde{y}_i^T \tilde{y}_i^T - \tilde{y}_i \tilde{y}_i^T + N_i \tilde{y}_i \tilde{y}_i^T \tilde{y}_i^T \right]
\]
(65)
\[
= S - CE \tilde{V}^T - E \tilde{V} C^T + E_{\tilde{V}} [\tilde{V} R_{\tilde{y}} \tilde{V}^T]
\]
(66)
Then \( q^* (W) \) is Wishart distributed:

\[
P (W) = \mathcal{W} (W | \Psi, \nu) \quad \text{if } \nu > d
\]

(67)

\[
\Psi^{-1} = K
\]

(68)

\[
\nu = N
\]

(69)

The optimum for \( q^* (W) \) in the informative case:

\[
\ln q^* (W) = E_{Y, \mu, V, \alpha} [\ln P (\Phi, Y, \mu, V, W, \alpha)] + \text{const}
\]

(70)

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

(71)

\[
= \frac{N}{2} \ln |W| + \frac{\nu_0 - d - 1}{2} \ln |W| - \frac{1}{2} \text{tr} (W (\Psi^{-1}_0 + K)) + \text{const}
\]

(72)

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

\[
P (W) = \mathcal{W} (W | \Psi, \nu)
\]

(73)

\[
\Psi^{-1} = \Psi^{-1}_0 + K
\]

(74)

\[
\nu = \nu_0 + N
\]

(75)
Finally, we evaluate the expectations:

\[ E[\alpha_q] = \frac{a_{\alpha_q}'}{b_{\alpha_q}} \]  

(76)

\[ \tilde{V} = E[\tilde{V}] = \begin{bmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \\ \vdots \\ \tilde{V}_d^T \end{bmatrix} \]  

(77)

\[ \tilde{W} = E[\tilde{W}] = \nu \Psi \]  

(78)

\[ E[v_{rq}^Tv_{rq}] = \sum_{r=1}^d E[v_{rq}'v_{rq}'] \]  

(79)

\[ = \sum_{r=1}^d \tilde{V}_{rq}^1 + \tilde{V}_{rq}^2 \]  

(80)

\[ E[V^TWV] = E[V\tilde{W}V^T] \]  

(81)

\[ = \sum_{r=1}^d \sum_{s=1}^d \tilde{W}_{rs} E[v_{r}'v_{s}'] \]  

(82)

\[ = \sum_{r=1}^d \tilde{W}_{rr} \Sigma_v + \sum_{r=1}^d \sum_{s=1}^d \tilde{W}_{rs} v_{r}'v_{s}' \]  

(83)

\[ E[V^T\mu] = \sum_{r=1}^d \tilde{W}_{rr} \Sigma_v + \tilde{W}^T \tilde{W} \]  

(86)

\[ E[\tilde{V}R_\tilde{y}\tilde{V}^T] = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} \tilde{y}_r \tilde{y}_s \tilde{E}_r \tilde{E}_s^T \]  

(87)

\[ = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} r_{\tilde{y}r\tilde{y}s} E[\tilde{v}_r\tilde{v}_s']_d \times d \]  

(88)

\[ = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} r_{\tilde{y}r\tilde{y}s} [E[\tilde{v}_r\tilde{v}_s']_d \times d] \]  

(89)

\[ = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} r_{\tilde{y}r\tilde{y}s} [E[\tilde{v}_r'\tilde{v}_s']_d \times d] \]  

(90)

\[ = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} r_{\tilde{y}r\tilde{y}s} [E[\tilde{v}_r\tilde{v}_s']_d \times d] + \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} r_{\tilde{y}r\tilde{y}s} \text{diag}(\sigma_{\tilde{v}_{rs},d}) \]  

(91)

\[ = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} r_{\tilde{y}r\tilde{y}s} E[\tilde{v}_r] E[\tilde{v}_s']_d + \text{diag}(\rho) \]  

(92)

\[ = \tilde{V} \tilde{W} = \tilde{V}^T \tilde{W} + \text{diag}(\rho) \]  

(93)
where \( \hat{v}_ri \) is the \( i \)th element of \( \hat{v}_r \), \( \hat{v}_r' \) is the \( r \)th element of \( \hat{v}_i' \),

\[
\Sigma \hat{v}_r = \begin{bmatrix} \Sigma \hat{v}_r & \Sigma \hat{v}_r \\ \Sigma \hat{v}_r & \Sigma \hat{v}_r \end{bmatrix} = L^{-1} \hat{v}_r
\]

\[
\rho = \begin{bmatrix} \rho_1 & \rho_2 & \ldots & \rho_d \end{bmatrix}^T
\]

\[
\rho_i = \sum_{r=1}^{n_y} \sum_{s=1}^{n_y} \left( R_{\hat{y}} \circ L^{-1}_{\hat{v}_r} \right)_{rs}
\]

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

### 5.2.1 Distributions with deterministic annealing

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

\[
q^*(Y) = \prod_{i=1}^{M} N(y_i|\overline{y}_i, 1/\kappa L^{-1}_{\overline{y}_i})
\]

\[
q^*(\hat{v}_r') = N(\hat{v}_r'|\overline{v}_r', 1/\kappa L^{-1}_{\overline{v}_r'})
\]

\[
q^*(W) = W(W|1/\kappa \Psi, \kappa(\nu - d - 1) + d + 1) \quad \text{if} \quad \kappa(\nu - d - 1) + 1 > 0
\]

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

\[
a'_q = \kappa \left( a_q + \frac{d}{2} - 1 \right) + 1
\]

\[
b'_{aq} = \kappa \left( b_q + \frac{1}{2} E \left[ v_q^T v_q \right] \right)
\]

### 5.3 Variational lower bound

The lower bound is given by

\[
L = E_{\Phi, \mu, V, W} [\ln P(\Phi|Y, \mu, V, W)] + E_Y [\ln P(Y)] + E_{V, \alpha} [\ln P(V|\alpha)]
\]

\[
+ E_{\alpha} [\ln P(\alpha)] + E_\mu [\ln P(\mu)] + E_W [\ln P(W)]
\]

\[
- E_Y [\ln q(Y)] - E_V \left[ \ln q(\hat{V}) \right] - E_\alpha [\ln q(\alpha)] - E_W [\ln q(W)]
\]

The term \( E_{\Phi, \mu, V, W} [\ln P(\Phi|Y, \mu, V, W)] \):

\[
E_{\Phi, \mu, V, W} [\ln P(\Phi|Y, \mu, V, W)] = \frac{N}{2} E [\ln |W|] - \frac{Nd}{2} \ln(2\pi)
\]

\[
- \frac{1}{2} tr \left( \overline{W} \left( S - 2C^{-1} \overline{V} + E \left[ V^T \overline{V}^T \right] \right) \right)
\]

\[
= \frac{N}{2} \ln \overline{W} - \frac{Nd}{2} \ln(2\pi) - \frac{1}{2} tr \left( \overline{W} S \right)
\]

\[
- \frac{1}{2} tr \left( -2 \overline{V}^T \overline{W} C + E \left[ \overline{V}^T \overline{W} \overline{V} \right] R_{\hat{y}} \right)
\]

where

\[
\ln \overline{W} = E [\ln |W|]
\]

\[
= \sum_{i=1}^{d} \psi \left( \frac{\nu + 1 - i}{2} \right) + d \ln 2 + \ln |\Psi|
\]

\[9\]
and \( \psi \) is the digamma function.

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

\[
E_Y \ln P(Y) = -\frac{Mn_y}{2} \ln(2\pi) - \frac{1}{2} \text{tr} \left( \sum_{i=1}^{M} E [y_i y_i^T] \right)
\]

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

(108)

where

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

(110)

The term \( E_{\mathbf{V},\alpha} \ln P(\mathbf{V} | \alpha) \):

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

(111)

where

\[
E[\ln \alpha_q] = \psi(a'_q) - \ln b'_{\alpha_q}.
\]

(112)

The term \( E_\alpha \ln P(\alpha) \):

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

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

(113)

The term \( E_\mu \ln P(\mu) \):

\[
E_\mu \ln P(\mu) = -\frac{d}{2} \ln(2\pi) + \frac{1}{2} \sum_{r=1}^{d} \ln \beta_r - \frac{1}{2} \sum_{r=1}^{d} \beta_r \left( E[\mu_r^2] - 2\mu_{0_r} E[\mu_r] + \mu_{0_r}^2 \right)
\]

\[
= -\frac{d}{2} \ln(2\pi) + \frac{1}{2} \sum_{r=1}^{d} \ln \beta_r - \frac{1}{2} \sum_{r=1}^{d} \frac{\beta_r \left( \sum_{r=1}^{d} E[\mu_r^2] - 2\mu_{0_r} E[\mu_r] + \mu_{0_r}^2 \right)}{2}
\]

(115)

(116)

The term \( E_W \ln P(W) \) for the non-informative case:

\[
E_W \ln P(W) = -\frac{d + 1}{2} \ln \mathbf{W}.
\]

(117)

The term \( E_W \ln P(W) \) for the informative case:

\[
E_W \ln P(W) = \ln B(\Psi_0, \nu_0) + \frac{\nu_0 - d - 1}{2} \ln \mathbf{W} - \frac{\nu}{2} \text{tr} (\Psi_0^{-1} \Psi)
\]

(118)
The term $E_Y [\ln q(Y)]$:

$$E_Y [\ln q(Y)] = -\frac{Mn_y}{2} \ln(2\pi) + \frac{1}{2} M \sum_{i=1}^{M} \ln |L_y| - \frac{1}{2} \text{tr} \left( L_y E \left[ (y_i - \bar{y}_i)(y_i - \bar{y}_i)^T \right] \right)$$

$$= -\frac{Mn_y}{2} \ln(2\pi) + \frac{1}{2} M \sum_{i=1}^{M} \ln |L_y|$$

$$- \frac{1}{2} \sum_{i=1}^{M} \text{tr} (L_y \left( E[y_iy_i^T] - \bar{y}_i E[y_i] - E[y_i]^{T} + \bar{y}_i \bar{y}_i^{T} \right))$$

$$= -\frac{Mn_y}{2} \ln(2\pi) + \frac{1}{2} M \sum_{i=1}^{M} \ln |L_y| - \frac{1}{2} \sum_{i=1}^{M} \text{tr} (I)$$

$$= -\frac{Mn_y}{2} \ln(2\pi) + \frac{1}{2} \sum_{i=1}^{M} \ln |L_y|$$

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

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

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

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

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

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

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

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

$$= \ln B(\Psi, \nu) + \frac{\nu - d - 1}{2} \ln \mathcal{W} = \frac{\nu d}{2}$$

where

$$B(A, N) = \frac{1}{2^h d/2 Z_N d} |A|^{-N/2}$$

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

5.4 Hyperparameter optimization

We can set the Hyperparameters ($\mu_0, \beta, a_\alpha, b_\alpha$) manually or estimate them from the development data maximizing the lower bound.

we derive for $a_\alpha$

$$\frac{\partial L}{\partial a_\alpha} = n_y (\ln b_\alpha - \psi(a_\alpha)) + \sum_{q=1}^{n_y} E[\ln \alpha_q] = 0 \implies$$

$$\psi(a_\alpha) = \ln b_\alpha + \frac{1}{n_y} \sum_{q=1}^{n_y} E[\ln \alpha_q]$$
We derive for $b_\alpha$:

$$\frac{\partial L}{\partial b_\alpha} = \frac{n_y a_\alpha}{b_\alpha} - \sum_{q=1}^{n_y} E[\alpha_q] = 0 \implies (133)$$

$$b_\alpha = \left( \frac{1}{n_y a_\alpha} \sum_{q=1}^{n_y} E[\alpha_q] \right)^{-1} (134)$$

We solve these equations with the procedure described in [5]. We write

$$\psi(a) = \ln b + c \quad (135)$$

$$b = \frac{a}{d} \quad (136)$$

where

$$c = \frac{1}{n_y} \sum_{q=1}^{n_y} E[\ln \alpha_q] \quad (137)$$

$$d = \frac{1}{n_y} \sum_{q=1}^{n_y} E[\alpha_q] \quad (138)$$

Then

$$f(a) = \psi(a) - \ln a + \ln d - c = 0 \quad (139)$$

We can solve for $a$ using Newton-Raphson iterations:

$$a_{\text{new}} = a - \frac{f(a)}{f'(a)} = a \left( 1 - \frac{\psi(a) - \ln a + \ln d - c}{\psi'(a)a - 1} \right) \quad (140)$$

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

$$\tilde{a}_{\text{new}} = \tilde{a} - \frac{f(\tilde{a})}{f'(\tilde{a})} = \tilde{a} - \frac{\psi(a) - \ln a + \ln d - c}{\psi'(a)a - 1} (142)$$

Taking exponential in both sides:

$$a_{\text{new}} = a \exp \left( - \frac{\psi(a) - \ln a + \ln d - c}{\psi'(a)a - 1} \right) (144)$$

We derive for $\mu_0$:

$$\frac{\partial L}{\partial \mu_0} = 0 \implies \mu_0 = E[\mu] (145)$$

We derive for $\beta$:

$$\frac{\partial L}{\partial \beta} = 0 \implies \beta^{-1} = \sum_{r=1}^{d} \mu_r + E[\mu_r^2] - 2\mu_0 E[\mu_r] + \mu_0^2 (147)$$

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

$$\beta^{-1} = \frac{1}{d} \sum_{r=1}^{d} \mu_r + E[\mu_r^2] - 2\mu_0 E[\mu_r] + \mu_0^2 (149)$$
5.5 Minimum divergence

We assume a more general prior for the hidden variables:

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

To minimize the divergence we maximize the part of \( L \) that depends on \( \mu_y \):

\[ L(\mu_y, \Lambda_y) = \sum_{i=1}^{M} E_Y \left[ \ln \mathcal{N}(y_i | \mu_y, \Lambda_y^{-1}) \right] \]  

The, we get

\[ \mu_y = \frac{1}{M} \sum_{i=1}^{M} E_Y [y_i] \]  

\[ \Sigma_y = \Lambda_y^{-1} = \frac{1}{M} \sum_{i=1}^{M} E_Y \left[ (y_i - \mu_y)(y_i - \mu_y)^T \right] \]  

\[ = \frac{1}{M} \sum_{i=1}^{M} E_Y [y_i y_i^T] - \mu_y \mu_y^T \]  

We have a transform \( y = \phi(y') \) such as \( y' \) has a standard prior:

\[ y = \mu_y + (\Sigma_y^{1/2})^T y' \]  

we also can write that as

\[ \tilde{y} = J\tilde{y}' \]  

where

\[ J = \begin{bmatrix} (\Sigma_y^{1/2})^T & \mu_y \\ 0^T & 1 \end{bmatrix} \]  

Now, we get \( q(\tilde{V}_r') \) such as if we apply the transform \( y' = \phi^{-1}(y) \), the term \( E[\ln P(X|Y, W)] \) of \( L \) remains constant:

\[ \tilde{V}_r' \leftrightarrow J^T \tilde{V}_r' \]  

\[ L_{\tilde{V}_r'}^{-1} \leftrightarrow J^T L_{\tilde{V}_r'}^{-1} J \]  

\[ L_{\tilde{V}_r'} \leftrightarrow G^T L_{\tilde{V}_r} G \]  

where

\[ G = (J^T)^{-1} \]

\[ = \begin{bmatrix} (\Sigma_y^{1/2})^{-1} & 0 \\ -\mu_y^T (\Sigma_y^{1/2})^{-1} & 1 \end{bmatrix} \]

6 Variational inference with Gaussian-Gamma priors for \( V \), Gaussian for \( \mu \) and Gamma for \( W \)

6.1 Model priors

In section 5, we saw that if we use a full covariance \( W \) we had a full covariance posterior for \( \tilde{V} \). Then, to get a tractable solution, we forced independence between the the rows of \( \tilde{V} \) when choosing the variational partition function.
In this section, we are going to assume that we have applied a rotation to the data such as we can consider that $W$ is going to remain diagonal during the VB iteration.

Then we are going to place a broad Gamma prior over each element of the diagonal of $W$:

$$P(W) = \prod_{r=1}^{d} G(w_{rr}|a_w, b_w) \quad (163)$$

We also consider the case of an isotropic $W$ ($W = wI$). Then the prior is

$$P(w) = G(w|a_w, b_w) \quad (164)$$

### 6.2 Variational distributions

We write the joint distribution of the latent variables:

$$P(\Phi, Y, \mu, V, W, \alpha|\mu_0, \beta, a_0, b_0, a_w, b_w) = P(\Phi|Y, \mu, V, W) P(Y) P(V) P(\alpha)$$

$$P(\alpha|a, b) P(\mu|\mu_0, \beta) P(W|a_w, b_w) \quad (165)$$

Following, the conditioning on $\mu_0, \beta, a_0, b_0, a_w, b_w$ will be dropped for convenience.

Now, we consider the partition of the posterior:

$$P(Y, \mu, V, W, \alpha|\Phi) \approx q(Y, \mu, V, W, \alpha) = q(Y) q(\mu) q(V) q(W) q(\alpha) \quad (166)$$

The optimum for $q^*(Y)$ and $q^*(\alpha)$ are the same as in section 5.2.

The optimum for $q^*(\tilde{V})$:

$$\ln q^*(\tilde{V}) = E_{Y,\mu,\alpha}[\ln P(\Phi, \mu, V, W, \alpha)] + \text{const} \quad (167)$$

$$= E_{Y,\mu}[\ln P(\Phi|Y, \mu, V, W)] + E_{\alpha}[\ln P(V|\alpha)] + \ln P(\mu) + \text{const} \quad (168)$$

$$= -\frac{1}{2} \text{tr} \left(-2\tilde{V}^T E[W] C + \tilde{V}^T E[W] \tilde{V} R \tilde{\gamma} \right)$$

$$- \frac{1}{2} \sum_{r=1}^{d} \tilde{v}_{r}^{T} \text{diag}(E[\alpha]) \tilde{v}_{r} - \frac{1}{2} \sum_{r=1}^{d} \beta_r (\mu_r - \mu_{0r})^2 + \text{const} \quad (169)$$

$$= -\frac{1}{2} \sum_{r=1}^{d} \text{tr} \left(-2\tilde{v}_{r}^{T} (\overline{\alpha}_r C_r + \beta_r \tilde{\mu}_{0r}^T) + \tilde{v}_{r}^{T} \tilde{W}_r (\text{diag}(\overline{\alpha}_r) + \overline{\alpha}_r R \tilde{\gamma}) \right) \quad (170)$$

Then $q^*(\tilde{V})$ is a product of Gaussian distributions:

$$q^*(\tilde{V}) = \prod_{r=1}^{d} \mathcal{N}(\tilde{v}_{r}|\overline{\alpha}_r, L_{\tilde{V}_r}^{-1}) \quad (171)$$

$$L_{\tilde{V}_r} = \text{diag}(\overline{\alpha}_r) + \overline{\alpha}_r R \tilde{\gamma} \quad (172)$$

$$\overline{\alpha}_r = L_{\tilde{V}_r}^{-1} (\overline{\alpha}_r C_r + \beta_r \tilde{\mu}_{0r}) \quad (173)$$

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

$$\ln q^*(W) = E_{Y,\mu,\alpha}[\ln P(\Phi, Y, \mu, V, W, \alpha)] + \text{const} \quad (174)$$

$$= E_{Y,\mu}[\ln P(\Phi|Y, \mu, V, W)] + \ln P(W) + \text{const} \quad (175)$$

$$= \sum_{r=1}^{d} \frac{N}{2} \ln w_{rr} - \frac{1}{2} w_{rr} k_{rr} + (a_w - 1) \ln w_{rr} - b_w w_{rr} + \text{const} \quad (176)$$

$$= \sum_{r=1}^{d} \left( a_w + \frac{N}{2} - 1 \right) \ln w_{rr} - \left( b_w + \frac{1}{2} k_{rr} \right) w_{rr} + \text{const} \quad (177)$$
where
\[ K = \text{diag} \left( S - CE \left[ \tilde{V} \right]^T - E \left[ \tilde{V} \right] C^T + E_{\tilde{V}} \left[ \tilde{V} R_{\tilde{y}} \tilde{V}^T \right] \right) \] (178)

Then \( q^* (W) \) is a product of Gammas:
\[ q^* (W) = \prod_{r=1}^{d} G \left( w_{rr} | a'_{wr}, b'_{wr} \right) \] (179)
\[ a'_{wr} = a_w + \frac{Nd}{2} \] (180)
\[ b'_{wr} = b_w + \frac{1}{2}k_{rr} \] (181)

If we force an isotropic \( W \), the optimum \( q^* (W) \) is
\[ \ln q^* (W) = \frac{Nd}{2} \ln w - \frac{1}{2}w + (a_w - 1) \ln w - b_w w + \text{const} \] (182)
\[ = \left( a_w + \frac{Nd}{2} - 1 \right) \ln w - \left( b_w + \frac{1}{2}k \right) w + \text{const} \] (183)

where
\[ k = tr \left( S - CE \left[ \tilde{V} \right]^T - E \left[ \tilde{V} \right] C^T + E_{\tilde{V}} \left[ \tilde{V} R_{\tilde{y}} \tilde{V}^T \right] \right) \] (184)
\[ = \text{tr} \left( S - 2CE \left[ \tilde{V} \right]^T \right) + \text{tr} \left( E \left[ \tilde{V}^T \tilde{V} \right] R_{\tilde{y}} \right) \] (185)

Then \( q^* (W) \) is a Gamma distribution:
\[ q^* (W) = G \left( w | a'_w, b'_w \right) \] (186)
\[ a'_w = a_w + \frac{Nd}{2} \] (187)
\[ b'_w = b_w + \frac{1}{2}k \] (188)

Finally, we evaluate the expectations:
\[ E \left[ \tilde{V}^T \tilde{V} \right] = E \left[ \tilde{V}' \tilde{V}'^T \right] \] (189)
\[ = \sum_{r=1}^{d} L_{\tilde{V}r}^{-1} + E \left[ \tilde{V}' \right] E \left[ \tilde{V}' \right]^T \] (190)
\[ = \sum_{r=1}^{d} L_{\tilde{V}r}^{-1} + E \left[ \tilde{V}' \right] E \left[ \tilde{V}' \right] \] (191)

### 6.3 Variational lower bound

The lower bound is given by
\[ L = E_{Y, \mu, V, W} \left[ \ln P (\Phi | Y, \mu, V, W) \right] + E_Y \left[ \ln P (Y) \right] + E_{V, \alpha} \left[ \ln P (V | \alpha) \right] \]
\[ + E_{\alpha} \left[ \ln P (\alpha) \right] + E_{\mu} \left[ \ln P (\mu) \right] + E_W \left[ \ln P (W) \right] \]
\[ - E_Y \left[ \ln q (Y) \right] - E_{\tilde{V}} \left[ \ln q \left( \tilde{V} \right) \right] - E_{\alpha} \left[ \ln q \left( \alpha \right) \right] - E_W \left[ \ln q \left( W \right) \right] \] (192)

The term \( E_{Y, \mu, V, W} \left[ \ln P (\Phi | Y, \mu, V, W) \right] \):
\[ E_{Y, \mu, V, W} \left[ \ln P (\Phi | Y, \mu, V, W) \right] = \frac{N d}{2} \sum_{r=1}^{d} \ln w_{rr} - \frac{Nd}{2} \ln (2\pi) - \frac{1}{2} \text{tr} \left( \overline{\tilde{V}^T} \tilde{W} C + E \left[ \tilde{V}^T \tilde{W} \tilde{V} \right] R_{\tilde{y}} \right) \] (193)
where

\[
\ln w_{rr} = E[\ln |w_{rr}|] = \psi(a'_w) - \ln b'_w
\]

and \(\psi\) is the digamma function.

The term \(E_W[\ln P(W)]\) with non-isotropic \(W\):

\[
E_W[\ln P(W)] = \sum_{r=1}^{d} a_w \ln b_w + (a_w - 1) E[\ln w_{rr}] - b_w E[w_{rr}] - \ln \Gamma(a_w) = d(a_w \ln b_w - \ln \Gamma(a_w)) + (a_w - 1) \sum_{r=1}^{d} E[\ln w_{rr}] - b_w \sum_{r=1}^{d} E[w_{rr}]
\]

The term \(E_W[\ln P(W)]\) with isotropic \(W\):

\[
E_W[\ln P(W)] = a_w \ln b_w - \ln \Gamma(a_w) + (a_w - 1) E[\ln w] - b_w E[w] = \psi(a'_w) - \psi(a_w) - \ln b'_w - \ln \Gamma(a'_w) + \ln b_w
\]

The term \(E_W[\ln q(W)]\) with non-isotropic \(W\):

\[
E_W[\ln q(W)] = -d \sum_{r=1}^{d} H[q(w_{rr})] = -d(a'_w - 1) \psi(a'_w) - a'_w - \ln \Gamma(a'_w) + \ln b'_w
\]

The term \(E_W[\ln q(W)]\) with isotropic \(W\):

\[
E_W[\ln q(W)] = -H[q(w)] = (a'_w - 1) \psi(a'_w) - a'_w - \ln \Gamma(a'_w) + \ln b_w
\]

The rest of terms are the same as in section 6.3.

### 6.4 Hyperparameter optimization

We can estimate the parameters \((a_w, b_w)\) from the development data maximizing the lower bound.

For non-isotropic \(W\):

we derive for \(a_w\)

\[
\frac{\partial \mathcal{L}}{\partial a_w} = d \ln b_w - \psi(a_w) + \sum_{r=1}^{d} E[\ln w_{rr}] = 0 \implies \psi(a_w) = \ln b_w + \frac{1}{d} \sum_{r=1}^{d} E[\ln w_{rr}]
\]

We derive for \(b_w\):

\[
\frac{\partial \mathcal{L}}{\partial b_w} = \frac{d a_w}{b} - \frac{d}{b} \sum_{r=1}^{d} E[w_{rr}] = 0 \implies b_w = \left(\frac{1}{d a_w} \sum_{r=1}^{d} E[w_{rr}]\right)^{-1}
\]

For isotropic \(W\):

we derive for \(a_w\)

\[
\frac{\partial \mathcal{L}}{\partial a_w} = \ln b_w - \psi(a_w) + E[\ln w] = 0 \implies \psi(a_w) = \ln b_w + E[\ln w]
\]

16
We derive for $b_w$:

$$\frac{\partial L}{\partial b_w} = \frac{a_w}{b} - E[w] = 0 \implies b_w = \left( \frac{1}{a_w} E[w] \right)^{-1}$$

(209)

(210)

We can solve these equations by Newton-Rhapson iterations as described in section 5.4.

7 Variational inference with full covariance Gaussian prior for $V$ and $\mu$ and Wishart for $W$

7.1 Model priors

Let’s assume that we compute the posterior of model parameters given a development database with a large amount of data. If we want to compute the model posterior for a small database we could use the posterior given the large database as prior.

Thus, we take a prior distribution for $\tilde{V}$

$$P(\tilde{V}) = \prod_{r=1}^{d} N(\tilde{v}_r | \tilde{v}_{0r}, L_{\tilde{V}_{0r}})$$

(211)

The prior for $W$ is

$$P(W) = W(W | \Psi_0, \nu_0)$$

(212)

The parameters $\tilde{v}_{0r}, L_{\tilde{V}_{0r}}, \Psi_0, \nu_0$ are computed with the large dataset.

7.2 Variational distributions

The joint distribution of the latent variables:

$$P(\Phi, Y, \mu, V, W) = P(\Phi | Y, \mu, V, W) P(Y) P(\tilde{V}) P(W)$$

(213)

Now, we consider the partition of the posterior:

$$P(Y, \mu, V, W | \Phi) \approx q(Y, \tilde{V}, W) = \prod_{r=1}^{d} q(\tilde{v}_r) q(W)$$

(214)

The optimum for $q^\ast(Y)$ is the same as in section 5.2.

The optimum for $q^\ast(\tilde{v}_r)$:

$$\ln q^\ast(\tilde{v}_r) = \mathbb{E}_{Y, W, \tilde{V}, \tilde{v}_{s \neq r}} [\ln P(\Phi, Y, \mu, V, W)] + \text{const} \implies$$

$$= \mathbb{E}_{Y, W, \tilde{V}, \tilde{v}_{s \neq r}} [\ln P(\Phi | Y, \mu, V, W)] + \ln P(\tilde{V}) + \text{const}$$

$$= - \frac{1}{2} \text{tr} \left( -2\tilde{v}_r \left( \tilde{w}_{rr} C_r + \sum_{s \neq r} \tilde{w}_{rs} \left( C_s - E[\tilde{v}_s]^T R \tilde{y} \right) \right) + \tilde{v}_r \tilde{v}_r^T \tilde{w}_{rr} R \tilde{y} \right)$$

$$- \frac{1}{2} (\tilde{v}_r - \tilde{v}_{0r})^T L_{\tilde{V}_{0r}} (\tilde{v}_r - \tilde{v}_{0r}) + \text{const}$$

$$= - \frac{1}{2} \text{tr} \left( -2\tilde{v}_r \left( \tilde{w}_{0r} L_{\tilde{V}_{0r}} + \tilde{w}_{rr} C_r + \sum_{s \neq r} \tilde{w}_{rs} \left( C_s - E[\tilde{v}_s]^T R \tilde{y} \right) \right) \right.$$  

$$+ \tilde{v}_r \tilde{v}_r^T \left( L_{\tilde{V}_{0r}} + \tilde{w}_{rr} R \tilde{y} \right) \right)$$

(215)

(216)

(217)

(218)
Then $q^* (\tilde{\nu}'_r)$ is a Gaussian distribution:

$$q^* (\tilde{\nu}'_r) = \mathcal{N} \left( \tilde{\nu}'_r | \tilde{\nu}_r, L_{\tilde{\nu}_r}^{-1} \right)$$  \hspace{1cm} (219)

$$L_{\tilde{\nu}_r} = L_{\tilde{\nu}_{cr}} + \pi_{rr} R_y$$  \hspace{1cm} (220)

$$\tilde{\nu}'_r = L_{\tilde{\nu}_{cr}}^{-1} \left( L_{\tilde{\nu}_{cr}} \tilde{\nu}'_{cr} + \pi_{rr} C^T_s + \sum_{s \neq r} \pi_{rs} \left( C^T_s - R_y \tilde{\nu}'_s \right) \right)$$  \hspace{1cm} (221)

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

$$\ln q^* (W) = E_{Y, \mu, V, \alpha} \left[ \ln P (\Phi, Y, \mu, V, W, \alpha) \right] + \text{const}$$  \hspace{1cm} (222)

$$= E_{Y, \mu, V} \left[ \ln P (\Phi | Y, \mu, W) \right] + \ln P (W) + \text{const}$$  \hspace{1cm} (223)

$$= \frac{N}{2} \ln |W| + \frac{\nu_0 - d - 1}{2} \ln |W| - \frac{1}{2} \text{tr} \left( W \left( \Psi_0^{-1} + K \right) \right) + \text{const}$$  \hspace{1cm} (224)

where

$$K = S - CE \left[ \tilde{V} \right]^T - E \left[ \tilde{V} \right] C^T + E_{\tilde{V}} \left[ \tilde{V} R_y \tilde{V}^T \right]$$  \hspace{1cm} (225)

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

$$P (W) = \mathcal{W} (W | \Psi, \nu)$$  \hspace{1cm} (226)

$$\Psi^{-1} = \Psi_0^{-1} + K$$  \hspace{1cm} (227)

$$\nu = \nu_0 + N$$  \hspace{1cm} (228)

### 7.2.1 Distributions with deterministic annealing

If we use annealing, for a parameter $\kappa$, we have:

$$q^* (W) = \mathcal{W} (W | \kappa \Psi, \nu_0 + N - d - 1 + d + 1) \text{ if } \kappa (\nu_0 + N - d - 1 + d + 1) > 0$$  \hspace{1cm} (229)

### 7.3 Variational lower bound

The lower bound is given by

$$\mathcal{L} = E_{Y, \mu, V, W} \left[ \ln P (\Phi | Y, \mu, V, W) \right] + E_{Y} \left[ \ln P (Y) \right] + E_{\tilde{V}} \left[ \ln P (\tilde{V}) \right] + E_{W} \left[ \ln P (W) \right]$$

$$- E_{Y} \left[ \ln q (Y) \right] - E_{\tilde{V}} \left[ \ln q (\tilde{V}) \right] - E_{W} \left[ \ln q (W) \right]$$  \hspace{1cm} (230)
The term $E_{\mathcal{V}} \left[ \ln P \left( \mathcal{V} \right) \right]$:

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

$$- \frac{1}{2} \sum_{r=1}^{d} \text{tr} \left( L_{\mathcal{V}_r} \mathbb{E} \left[ (\mathbf{V}_r^\prime - \mathbf{V}_0^r) (\mathbf{V}_r^\prime - \mathbf{V}_0^r)^T \right] \right)$$

$$= -\frac{n_d}{2} \ln(2\pi) + \frac{1}{2} \sum_{r=1}^{d} \ln |L_{\mathcal{V}_r}|$$

$$- \frac{1}{2} \sum_{r=1}^{d} \text{tr} \left( L_{\mathcal{V}_r} \left( L_{\mathcal{V}_r}^{-1} + \mathbf{V}_r^\prime \mathbf{V}_r^\prime - \mathbf{V}_0^r \mathbf{V}_0^r - \mathbf{V}_r^\prime \mathbf{V}_0^r + \mathbf{V}_0^r \mathbf{V}_r^\prime \right) \right)$$

$$= -\frac{n_d}{2} \ln(2\pi) + \frac{1}{2} \sum_{r=1}^{d} \ln |L_{\mathcal{V}_r}|$$

$$- \frac{1}{2} \sum_{r=1}^{d} \text{tr} \left( L_{\mathcal{V}_r} \left( L_{\mathcal{V}_r}^{-1} - \frac{1}{2} \sum_{r=1}^{d} (\mathbf{V}_r - \mathbf{V}_0^r) (\mathbf{V}_r - \mathbf{V}_0^r)^T \right) \right)$$

$$= -\frac{n_d}{2} \ln(2\pi) + \frac{1}{2} \sum_{r=1}^{d} \ln |L_{\mathcal{V}_r}|$$

$$- \frac{1}{2} \sum_{r=1}^{d} \text{tr} \left( L_{\mathcal{V}_r} \left( L_{\mathcal{V}_r}^{-1} - \frac{1}{2} \sum_{r=1}^{d} (\mathbf{V}_r - \mathbf{V}_0^r) (\mathbf{V}_r - \mathbf{V}_0^r)^T \right) \right)$$

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

$$E_{\mathcal{W}} \left[ \ln P \left( \mathcal{W} \right) \right] = \ln B \left( \Psi_0, \nu_0 \right) + \frac{\nu_0 - d - 1}{2} \ln |\mathbf{W}| - \frac{\nu d}{2}$$

where

$$\ln |\mathbf{W}| = \mathbb{E} \left[ \ln |\mathbf{W}| \right]$$

$$= \sum_{i=1}^{d} \psi \left( \frac{\nu + 1 - i}{2} \right) + d \ln 2 + \text{ln} |\Psi|$$

and $\psi$ is the digamma function.

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

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

$$= \ln B \left( \Psi, \nu \right) + \frac{\nu - d - 1}{2} \ln |\mathbf{W}| - \frac{\nu d}{2}$$

The rest of terms are the same as the ones in section 5.3.

8 Variational inference with full covariance Gaussian prior for $\mathcal{V}$ and $\mu$ and Gamma for $\mathcal{W}$

8.1 Model priors

Thus, we take a prior distribution for $\tilde{\mathcal{V}}$

$$P \left( \tilde{\mathcal{V}} \right) = \prod_{r=1}^{d} \mathcal{N} \left( \mathbf{V}_r^\prime | \mathbf{V}_0^r, L_{\mathcal{V}_r}^{-1} \right)$$

(240)
The prior for non-isotropic $W$ is

$$P(W) = \prod_{r=1}^{d} \mathcal{G}(w_{rr}|a_{w}, b_{w}) \quad (241)$$

The prior for isotropic $W$ is

$$P(W) = \mathcal{G}(w|a_{w}, b_{w}) \quad (242)$$

The parameters $\nabla_{0r}, L_{V_{0r}}^{-1}, a_{w}$ and $b_{w}$ are computed with the large dataset.

### 8.2 Variational distributions

The joint distribution of the latent variables:

$$P(\Phi, Y, \mu, V, W) = P(\Phi|Y, \mu, V, W) P(Y) P(\tilde{V}) P(W) \quad (243)$$

Now, we consider the partition of the posterior:

$$P(Y, \mu, V, W|\Phi) \approx q(Y, \tilde{V}, W) = q(Y) q(\tilde{V}) q(W) \quad (244)$$

The optimum for $q^*(Y)$ is the same as in section 5.2.

Then optimum for $q^*(\tilde{V})$ is:

$$q^*(\tilde{V}) = \prod_{r=1}^{d} \mathcal{N}(\nabla_{\tilde{V}_{r}}, L_{\tilde{V}_{r}}^{-1}) \quad (245)$$

$$L_{\tilde{V}_{r}} = L_{\tilde{V}_{0r}} + \bar{w}_{rr} R_{\tilde{V}} \quad (246)$$

$$\nabla_{\tilde{V}_{r}} = L_{\tilde{V}_{0r}}^{-1}(L_{\tilde{V}_{0r}}, \nabla_{0r} + \bar{w}_{rr} C_{r}^{T}) \quad (247)$$

The optimum for $q^*(W)$ for non-isotropic $W$:

$$q^*(W) = \prod_{r=1}^{d} \mathcal{G}(w_{rr}|a'_{w}, b'_{w}) \quad (248)$$

$$a'_{w} = a_{w} + \frac{N}{2} \quad (249)$$

$$b'_{w} = b_{w} + \frac{1}{2} k_{rr} \quad (250)$$

where

$$K = \text{diag} \left( S - CE \begin{bmatrix} \tilde{V} \end{bmatrix}^{T} - E \begin{bmatrix} \tilde{V} \end{bmatrix} C^{T} + E \begin{bmatrix} \tilde{V} \end{bmatrix} R_{\tilde{V}} \tilde{V}^{T} \right) \quad (251)$$

The optimum for $q^*(W)$ for isotropic $W$:

$$q^*(W) = \mathcal{G}(w|a'_{w}, b'_{w}) \quad (252)$$

$$a'_{w} = a_{w} + \frac{Nd}{2} \quad (253)$$

$$b'_{w} = b_{w} + \frac{1}{2} k \quad (254)$$

where

$$k = \text{tr} \left( S - 2CE \begin{bmatrix} \tilde{V} \end{bmatrix}^{T} \right) + \text{tr} \left( E \begin{bmatrix} \tilde{V}^{T} \tilde{V} \end{bmatrix} R_{\tilde{V}} \right) \quad (255)$$
8.3 Variational lower bound

The lower bound is given by

\[
\mathcal{L} = E_{Y, \mu, V, W} \left[ \ln P(\Phi|Y, \mu, V, W) \right] + E_{Y} \left[ \ln P(Y) \right] + E_{V} \left[ \ln \left( \tilde{V} \right) \right] + E_{W} \left[ \ln P(W) \right] \\
- E_{Y} \left[ \ln q(Y) \right] - E_{V} \left[ \ln q(V) \right] - E_{W} \left[ \ln q(W) \right]
\]  

(256)

The term \( E_{W} \left[ \ln P(W) \right] \) with non-isotropic \( W \):

\[
E_{W} \left[ \ln P(W) \right] = \sum_{r=1}^{d} a_{w} \ln b_{w_{r}} + (a_{w} - 1) E \left[ \ln w_{rr} \right] - b_{w_{r}} E \left[ w_{rr} \right] - \ln \Gamma(a_{w})
\]

(257)

\[
= -d \ln \Gamma(a_{w}) + a_{w} \sum_{r=1}^{d} \ln b_{w_{r}} + (a_{w} - 1) \sum_{r=1}^{d} E \left[ \ln w_{rr} \right] - \sum_{r=1}^{d} b_{w_{r}} E \left[ w_{rr} \right]
\]

(258)

The terms \( E_{Y} \left[ \ln P(Y) \right], E_{Y} \left[ \ln q(Y) \right] \) and \( E_{V} \left[ \ln q(V) \right] \) are the same as in section 5.3.

The terms \( E_{Y, \mu, V, W} \left[ \ln P(\Phi|Y, \mu, V, W) \right], E_{W} \left[ \ln P(W) \right] \) with isotropic \( W \) and \( E_{W} \left[ \ln q(W) \right] \) are the same as in section 6.3.

The term \( E_{V} \left[ \ln \left( \tilde{V} \right) \right] \) is the same as in section 7.3.

References

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

[2] Jesús Villalba and Niko Brummer, “Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance,” in Interspeech 2011, Florence, 2011, pp. 28–31.

[3] C M Bishop, “Variational principal components,” 9th International Conference on Artificial Neural Networks ICANN 99, vol. 1, no. 470, pp. 509–514, 1999.

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

[5] Matthew J Beal, “Variational algorithms for approximate Bayesian inference,” Philosophy, vol. 38, no. May, pp. 1–281, 2003.