Variational Bayes Factor Analysis for i-Vector Extraction

Jesús Villalba

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

May 11, 2012

1 Introduction

In this document we are going to derive the equations needed to implement a Variational Bayes i-vector extractor. This can be used to extract longer i-vectors reducing the risk of overfitting or to adapt an i-vector extractor from a database to another with scarce development data. This work is based on [1] and [2].

2 The Model

2.1 JFA

Joint Factor Analysis for i-vector extraction is a linear generative model represented in Figure 1.

![Figure 1: BN for i-vector extractor.](image)

This model assumes that speech frames are generated by a special type of mixture of factor analysers. An speech frame \(x_{it}\) of a session \(i\) and generated by the component \(k\) of the mixture model can be written as:

\[
x_{it} = m_k + W_k y_i + \epsilon_{itk}
\]  

where \(m_k\) is a session independent term, \(W_k\) is a low-rank factor loading matrix, \(y_i\) is the factor vector, and \(\epsilon_{itk}\) is a residual term. The prior distribution for the variables:

\[
y_i \sim \mathcal{N}(y_i | 0, I)
\]

\[
\epsilon_{itk} \sim \mathcal{N}(\epsilon_{itk} | 0, \Lambda_k^{-1})
\]

where \(\mathcal{N}\) denotes a Gaussian distribution.
This model differs from a standard mixture of FA in the way in which the factors are tied. In traditional FA, we have a different value of \( y \) for each frame and each component of the mixture of the session. On the contrary, in this model we share the same value of \( y \) for all the frames and mixture components of the same session.

We can define the session mean vector for component \( k \) as

\[
M_{ik} = m_k + W_k y_i .
\]  

(4)

In this manner, each frame is a session mean plus the residual term:

\[
x_{it} = M_{ik} + \epsilon_{itk} .
\]  

(5)

We find convening stacking the means and factor loading matrices of all components to form a mean supervector:

\[
M_i = m + WY_i
\]  

(6)

For this work, we are going to assume that \( m \) and \( \Lambda \) are given. We estimate them by EM-iterations of simple GMM. Besides, we assume that \( P(z_{it}) \) are known and fixed. In practice, we compute them using the GMM.

### 2.2 Notation

We define:

- Let \( X_i \) be the frames of session \( i \).
- Let \( X \) be the frames of all sessions.
- Let \( Y \) be the factors of all sessions.
- Let \( d \) be the features dimension.
- Let \( n_y \) be the factor dimension.
- Let \( K \) be the number of components of the mixture of FA.
- Let \( \Sigma_k = \Lambda_k^{-1} \).

### 3 Sufficient statistics

We define the statistics for segment \( i \) and component \( k \) as:

\[
N_{ik} = \sum_{t} P(z_{itk} = 1) F_{ik} = \sum_{t} P(z_{itk} = 1) x_{it}
\]  

(7)

We define the normalized sufficient statistics for component \( k \) as:

\[
\overline{F}_{ik} = \sum_{t} P(z_{itk} = 1) \Lambda_k^{1/2} (x_{it} - m_k) = \Lambda_k^{1/2} (F_{ik} - N_{ik} m_k)
\]  

(8)

If we normalize the sufficient statistics in mean and variance it is the same as having a FA model with \( m = 0 \) and \( \Sigma = I \). As we assume that \( m \) and \( \Sigma \) are fixed, doing that we can simplify the equations.

We define, too:

\[
N_i = \begin{bmatrix}
N_{i1}I_d & 0 & \cdots & 0 \\
0 & N_{i2}I_d & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & N_{iK}I_d
\end{bmatrix}, \quad \overline{F}_i = \begin{bmatrix}
\overline{F}_{i1} \\
\vdots \\
\overline{F}_{iK}
\end{bmatrix}
\]  

(9)
where $I_d$ is the identity matrix of dimension $d$.

We define the global normalized statistic:

$$
\mathbf{S}_k = \frac{1}{N_i} \sum_{t=1}^{T_i} P(z_{itk} = 1) (x_{it} - m_k) \Lambda_k (x_{it} - m_k)
$$

(10)

4 Conditional likelihood

The likelihood of the data of session $i$ given the latent variables is

$$
\ln P(X_i | Y_i, W, m, \Lambda) = -\frac{K}{2} \sum_{k=1}^{K} \left( \sum_{i=1}^{N_i} \mathbf{S}_{ik} \right) + \frac{1}{2} \sum_{i=1}^{N_i} y_i^T W y_i - \frac{1}{2} y_i^T W N_i W y_i
$$

(11)

5 Variational inference with Gaussian-Gamma priors

5.1 Model priors

We introduce a hierarchical prior $P(W | \alpha)$ over the matrix $W$ 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 $W$ through a conditional Gaussian distribution of the form:

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

(12)

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

5.2 Variational distributions

We write the joint distribution of the latent variables:

$$
P(X, Y, W, \alpha | m, \Lambda, a, b) = P(X | Y, W, m, \Lambda) P(Y) P(W | \alpha) P(\alpha | a, b)
$$

(14)

Following, the conditioning on $(m, \Lambda, a, b)$ will be dropped for convenience.

Now, we consider the partition of the posterior:

$$
P(Y, W, \alpha | X) \approx q(Y, W, \alpha) = q(Y) q(W) q(\alpha)
$$

(15)

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

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

$$
= \mathbb{E}_W [\ln P(X | Y, W)] + \ln P(Y) + \text{const}
$$

$$
= \sum_{i=1}^{H} y_i^T E[W]^T F_i - \frac{1}{2} y_i^T \left( I + \sum_{k=1}^{K} N_i E[W_k]^T W_k \right) y_i + \text{const}
$$

(16)
Therefore \( q^* (Y) \) is a product of Gaussian distributions.

\[
q^* (Y) = \prod_{i=1}^{H} \mathcal{N} (y_i | \mathbf{y}_i, \mathbf{L}_y^{-1}) \tag{19}
\]

\[
\mathbf{L}_{y_i} = \mathbf{I} + \sum_{k=1}^{K} N_{ik} \mathbb{E} \left[ \mathbf{W}_k^T \mathbf{W}_k \right] \tag{20}
\]

\[
\mathbf{y}_i = \mathbf{L}^{-1}_{y_i} \mathbb{E} \left[ \mathbf{W} \right]^T \mathbf{F}_i \tag{21}
\]

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

\[
\ln q^* (\mathbf{W}) = \mathbb{E}_{Y, \alpha} \left[ \ln P (\mathbf{X}, \mathbf{Y}, \mathbf{W}, \alpha) \right] + \text{const} \tag{22}
\]

\[
= \mathbb{E}_{Y} \left[ \ln P (\mathbf{X}| \mathbf{Y}, \mathbf{W}) \right] + \mathbb{E}_{\alpha} \left[ \ln P (\mathbf{W}| \alpha) \right] + \text{const} \tag{23}
\]

\[
= \sum_{i=1}^{H} \left( \mathbb{E} [y_i]^T \mathbf{W}^T \mathbf{F}_i - \frac{1}{2} \mathbb{E} \left[ y_i^T \mathbf{W}^T \mathbf{N}_i \mathbf{W} y_i \right] \right) - \frac{1}{2} \sum_{q=1}^{n_y} \mathbb{E} \left[ \alpha_q \right] \mathbf{w}_q^T \mathbf{w}_q + \text{const} \tag{24}
\]

\[
= \text{tr} \left( \mathbf{W}^T \mathbf{C} - \frac{1}{2} \sum_{k=1}^{K} \mathbf{W}_k^T \mathbf{W}_k \mathbf{R}_k \right) - \frac{1}{2} \sum_{q=1}^{n_y} \mathbb{E} \left[ \alpha_q \right] \mathbf{w}_q^T \mathbf{w}_q + \text{const} \tag{25}
\]

\[
= \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( \mathbf{w}_{kr}^T \mathbf{C}_k - \frac{1}{2} \mathbf{W}_k^T \mathbf{W}_k \mathbf{R}_k \right) - \frac{1}{2} \sum_{r=1}^{d} \mathbf{w}_{kr}^T \text{diag} (\mathbb{E} [\alpha]) \mathbf{w}_{kr} + \text{const} \tag{26}
\]

\[
= \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( \mathbf{w}_{kr}^T \mathbf{C}_k - \frac{1}{2} \mathbf{w}_{kr}^T \mathbf{w}_{kr} (\mathbb{E} [\alpha] + \mathbf{R}_k) \right) + \text{const} \tag{27}
\]

where \( \mathbf{w}_{kr} \) is a column vector containing the \( r^{th} \) row of \( \mathbf{W}_k \),

\[
\mathbf{w}_{kr}^T = \mathbf{W}_k^T \tag{28}
\]

\[
\mathbf{C} = \sum_{i=1}^{H} \mathbf{F}_i \mathbb{E} [y_i]^T \tag{29}
\]

\[
\mathbf{R}_k = \sum_{i=1}^{H} N_{ik} \mathbb{E} [y_i y_i^T] \tag{30}
\]

and \( \mathbf{C}_{kr} \) is the \( r^{th} \) block of \( \mathbf{C} \) corresponding to component \( k \) (row \((k-1)*d + r\)).

Then \( q^* (\mathbf{W}) \) is a product of Gaussian distributions

\[
q^* (\mathbf{W}) = \prod_{k=1}^{K} \prod_{r=1}^{d} \mathcal{N} (\mathbf{w}_{kr} | \mathbf{w}_{kr}^T, \mathbf{L}_{\mathbf{w}_k}^{-1}) \tag{31}
\]

\[
\mathbf{L}_{\mathbf{w}_k} = \mathbb{E} [\alpha] + \mathbf{R}_k \tag{32}
\]

\[
\mathbf{w}_{kr}^T = \mathbf{L}_{\mathbf{w}_k}^{-1} \mathbf{C}_{kr}^T \tag{33}
\]

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

\[
\ln q^* (\alpha) = \mathbb{E}_{\mathbf{Y}, \mathbf{W}} \left[ \ln P (\mathbf{X}, \mathbf{Y}, \mathbf{W}, \alpha) \right] + \text{const} \tag{34}
\]

\[
= \mathbb{E}_{\mathbf{W}} \left[ \ln P (\mathbf{W}| \alpha) \right] + \ln P (\alpha|a, b) + \text{const} \tag{35}
\]

\[
= \sum_{q=1}^{n_y} \frac{K d}{2} \ln \alpha_q - \frac{1}{2} \alpha_q \mathbb{E} \left[ \mathbf{w}_q^T \mathbf{w}_q \right] + (a - 1) \ln \alpha_q - b \alpha_q + \text{const} \tag{36}
\]

\[
= \sum_{q=1}^{n_y} \left( \frac{K d}{2} + a - 1 \right) \ln \alpha_q - \alpha_q \left( b + \frac{1}{2} \mathbb{E} \left[ \mathbf{w}_q^T \mathbf{w}_q \right] \right) + \text{const} \tag{37}
\]
Then

\[ q^* (\alpha) = \prod_{q=1}^{n_y} G (\alpha_q | a', b_q') \]  (38)

\[ a' = a + \frac{Kd}{2} \]  (39)

\[ b_q' = b + \frac{1}{2} \mathbb{E} [w_q^T w_q] \]  (40)

We evaluate the expectations:

\[ \mathbb{E} \left[ \alpha \right] = a' \]  (41)

\[ \mathbb{E} \left[ W \right] = \begin{bmatrix} \tilde{w}_1^T \\ \tilde{w}_2^T \\ \vdots \\ \tilde{w}_{Kd}^T \end{bmatrix} \]  (42)

\[ \mathbb{E} \left[ w_q^T w_q \right] = \sum_{k=1}^{K} \sum_{r=1}^{d} \mathbb{E} \left[ w_{krq}^T w_{krq} \right] = \sum_{k=1}^{K} \sum_{r=1}^{d} \tilde{w}_{krq}^2 + \sum_{r=1}^{d} \tilde{w}_{krq}^2 \]  (43)

\[ \mathbb{E} \left[ w_k^T w_k \right] = \mathbb{E} \left[ \tilde{w}_k^T \tilde{w}_k^T \right] = d \mathbb{L}_{\tilde{w}_k} - \mathbb{E} \left[ \tilde{w}_k^T \right] \]  (46)

\[ = d \mathbb{L}_{\tilde{w}_k} + \mathbb{E} \left[ \tilde{w}_k \right] \]  (47)

5.3 Variational lower bound

The lower bound is given by

\[ \mathcal{L} = \mathbb{E}_{Y,W} \left[ \ln P (X | Y, W) \right] + \mathbb{E}_{Y} \left[ \ln P (Y) \right] + \mathbb{E}_{W, \alpha} \left[ \ln P (W | \alpha) \right] + \mathbb{E}_{\alpha} \left[ \ln P (\alpha) \right] \]

\[ - \mathbb{E}_{Y} \left[ \ln q (Y) \right] - \mathbb{E}_{W} \left[ \ln q (W) \right] - \mathbb{E}_{\alpha} \left[ \ln q (\alpha) \right] \]  (48)

The term \( \mathbb{E}_{Y,W} \left[ \ln P (X | Y, W) \right] \):

\[ \mathbb{E}_{Y,W} \left[ \ln P (X | Y, W) \right] = - \sum_{k=1}^{K} N_k d \frac{1}{2} \log(2\pi) - \frac{1}{2} \text{tr} \left( \sum_{k=1}^{K} \mathbb{S}_k \right) \]

\[ + \sum_{i=1}^{H} \mathbb{E} \left[ y_i^T \right] \mathbb{E} \left[ W \right] \mathbb{F}_i - \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{H} \text{tr} \left( N_{ik} \mathbb{E} \left[ W_k^T W_k \right] \mathbb{E} \left[ y_i y_i^T \right] \right) \]  (49)

\[ = - \frac{1}{2} \text{tr} \left( \sum_{k=1}^{K} \mathbb{S}_k \right) \]

\[ - \frac{1}{2} \text{tr} \left( -2 \mathbb{E} \left[ W \right] \mathbb{C} + \sum_{k=1}^{K} \mathbb{E} \left[ W_k^2 W_k \right] \mathbb{R}_k \right) \]  (50)

The term \( \mathbb{E}_{Y} \left[ \ln P (Y) \right] \):

\[ \mathbb{E}_{Y} \left[ \ln P (Y) \right] = - \frac{H n_y}{2} \ln(2\pi) - \frac{1}{2} \text{tr} \left( \sum_{i=1}^{H} \mathbb{E} \left[ y_i y_i^T \right] \right) \]  (51)

\[ = - \frac{H n_y}{2} \ln(2\pi) - \frac{1}{2} \text{tr} (\mathbb{P}) \]  (52)
where

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

(53)

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

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

(54)

where

\[ E[\ln \alpha_q] = \psi(a') - \ln b' \]  

(55)

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

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

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

(56)

\[ = n_y (a \ln b - \ln \Gamma(a)) + (a - 1) \sum_{q=1}^{n_y} E[\ln \alpha_q] - b \sum_{q=1}^{n_y} E[\alpha_q] \]  

(57)

The term \( E_{\mathbf{Y}} \ln q(\mathbf{Y}) \):

\[ E_{\mathbf{Y}} \ln q(\mathbf{Y}) = - \frac{H n_y}{2} \ln(2\pi) + \frac{1}{2} \sum_{i=1}^{H} \ln |L_{y_i}| - \frac{1}{2} \text{tr} \left( L_{y_i} E \left[ (y_i - y_i') (y_i - y_i')^T \right] \right) \]  

(58)

\[ = - \frac{H n_y}{2} \ln(2\pi) + \frac{1}{2} \sum_{i=1}^{H} \ln |L_{y_i}| - \frac{1}{2} \sum_{i=1}^{H} \text{tr} \left( L_{y_i} E \left[ (y_i - y_i') (y_i - y_i')^T \right] \right) \]  

(59)

\[ = - \frac{H n_y}{2} \ln(2\pi) + \frac{1}{2} \sum_{i=1}^{H} \ln |L_{y_i}| - \frac{1}{2} \sum_{i=1}^{H} \text{tr}(I) \]  

(60)

\[ = - \frac{H n_y}{2} (\ln(2\pi) + 1) + \frac{1}{2} \sum_{i=1}^{H} \ln |L_{y_i}| \]  

(61)

The term \( E_{\mathbf{W}} \ln q(\mathbf{W}) \):

\[ E_{\mathbf{W}} \ln q(\mathbf{W}) = - \frac{K d n_y}{2} \ln(2\pi) + \frac{d}{2} \sum_{k=1}^{K} \ln |L_{w_k}| \]  

\[ - \frac{1}{2} \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( L_{w_k} E \left[ (w_k' - \mathbf{w}_k') (w_k' - \mathbf{w}_k')^T \right] \right) \]  

(62)

\[ = - \frac{K d n_y}{2} (\ln(2\pi) + 1) + \frac{d}{2} \sum_{k=1}^{K} \ln |L_{w_k}| \]  

(63)
The term $E_\alpha [\ln q(\alpha)]$:

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

(64)

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

(65)

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

(66)

5.4 Hyperparameter optimization

We can set the Hyperparameters manually or estimate them from the development data maximizing the lower bound.

We derive for $a$:

$$
\frac{\partial L}{\partial a} = n_y (\ln b - \psi(a)) + \sum_{q=1}^{n_y} E[\ln \alpha_q] = 0 \quad \Rightarrow
$$

(67)

$$
\psi(a) = \ln b + \frac{1}{n_y} \sum_{q=1}^{n_y} E[\ln \alpha_q]
$$

(68)

We derive for $b$:

$$
\frac{\partial L}{\partial b} = \frac{n_y a}{b} \sum_{q=1}^{n_y} E[\alpha_q] = 0 \quad \Rightarrow
$$

(69)

$$
b = \left( \frac{1}{n_y a} \sum_{q=1}^{n_y} E[\alpha_q] \right)^{-1}
$$

(70)

We solve these equations with the procedure described in [3]. We write

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

(71)

$$
b = \frac{a}{d}
$$

(72)

where

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

(73)

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

(74)

Then

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

(75)

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

$$
a_{new} = a - \frac{f(a)}{f'(a)} =
$$

(76)

$$
= a \left( 1 - \frac{\psi(a) - \ln a + \ln d - c}{a \psi'(a) - 1} \right)
$$

(77)
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 $a = \exp(\tilde{a})$.

$$\tilde{a}_{\text{new}} = \tilde{a} - \frac{f(\tilde{a})}{f'(\tilde{a})} = \tilde{a} - \frac{\psi(a) - \ln a + \ln d - c}{\psi'(a)a - 1}$$ (78)

Taking exponential in both sides:

$$a_{\text{new}} = a \exp\left(\frac{-\psi(a) - \ln a + \ln d - c}{\psi'(a)a - 1}\right)$$ (80)

### 5.5 Minimum divergence

We assume a more general prior for the hidden variables:

$$P(y) = \mathcal{N}(y|\mu_y, \Lambda_y^{-1})$$ (81)

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

$$L(\mu_y, \Lambda_y) = \sum_{i=1}^{H} E_Y [\ln \mathcal{N}(y|\mu_y, \Lambda_y^{-1})]$$ (82)

The, we get

$$\mu_y = \frac{1}{H} \sum_{i=1}^{M} E_Y [y_i]$$ (83)

$$\Sigma_y = \Lambda_y^{-1} = \frac{1}{H} \sum_{i=1}^{H} E_Y [(y_i - \mu_y)(y_i - \mu_y)^T]$$ (84)

$$= \frac{1}{H} \sum_{i=1}^{H} E_Y [y_i y_i^T] - \mu_y \mu_y^T$$ (85)

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

$$y = \mu_y + (\Sigma_y^{1/2})^T y'$$ (86)

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

$$\bar{w}_{kr} \leftarrow \Sigma_y^{1/2} \bar{w}_{kr}$$ (87)

$$L_{W_k}^{-1} \leftarrow \Sigma_y^{1/2} L_{W_k}^{-1} (\Sigma_y^{1/2})^T$$ (88)

$$L_{W_k} \leftarrow (\Sigma_y^{1/2})^{-1} L_{W_k} (\Sigma_y^{1/2})^{-1}$$ (89)

### 6 Variational inference with full covariance priors

#### 6.1 Model priors

Let's assume that we compute the posterior of $W$ given a development database with a large amount of data. If we want to compute the posterior $W$ for a small database we could use the posterior given the large database as prior. Thus, we take a prior distribution for $W$

$$P(W) = \prod_{k=1}^{K} \prod_{r=1}^{d} \mathcal{N}(\bar{w}_{kr}^'|\bar{w}_{0kr}', L_{W_k}^{-1})$$ (90)

where $\bar{w}_{0kr}', L_{W_k}^{-1}$ are parameters computed with the large dataset.
6.2 Variational distributions

The joint distribution of the latent variables:

\[ P(X, Y, W) = P(X|Y, W, m, \Lambda) P(Y) P(W) \]  

We approximate the posterior by:

\[ P(Y, W|X) \approx q(Y, W) = q(Y) q(W) \]  

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

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

\[
\ln q^*(W) = E_Y \left[ \ln P(X|Y, W) \right] + \text{const}
\]

\[
= \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( w'_{kr} C_{kr} - \frac{1}{2} w'_{kr} w'_{kr}^T R_k \right) - \frac{1}{2} (w'_{kr} - w'_{0kr})^T L_{W_{0kr}} (w'_{kr} - w'_{0kr}) + \text{const}
\]

\[
= K \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( w'_{kr} \left( L_{W_{0kr}} L_{W_{0kr}} + C_{kr} \right) - \frac{1}{2} w'_{kr} w'_{kr}^T \left( L_{W_{0kr}} + R_k \right) \right) + \text{const}
\]

Therefore, the \( q^*(W) \) is, again, a product of Gaussian distributions:

\[ q^*(W) = \prod_{k=1}^{K} \prod_{r=1}^{d} \mathcal{N}(w'_{kr}|w'_{0kr}, L_{W_{kr}}^{-1}) \]

\[ L_{W_k} = L_{W_{0k}} + R_k \]

\[ w'_{kr} = L_{W_{0kr}}^{-1} \left( L_{W_{0kr}} w'_{0kr} + C_{kr} \right) \]

6.3 Variational lower bound

The lower bound is given by

\[
\mathcal{L} = E_{X,Y,W} \left[ \ln P(X|Y, W) \right] + E_Y \left[ \ln P(Y) \right] + E_W \left[ \ln P(W) \right] - E_Y \left[ \ln q(Y) \right] - E_W \left[ \ln q(W) \right]
\]
The term $E_w \ln P(W)$:

$$E_w \ln P(W) = -\frac{n_y K d}{2} \ln(2\pi) + \frac{d}{2} \sum_{k=1}^{K} \ln |L_{W_{ok}}|$$

$$- \frac{1}{2} \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( L_{W_{ok}} E \left[ (w'_{kr} - w'_{0kr}) (w'_{kr} - w'_{0kr})^T \right] \right)$$

$$= -\frac{n_y K d}{2} \ln(2\pi) + \frac{d}{2} \sum_{k=1}^{K} \ln |L_{W_{ok}}|$$

$$- \frac{1}{2} \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( L_{W_{ok}} (L_{W_{ok}}^{-1} w'_{kr} w'_{kr} - w'_{0kr} w'_{0kr} - w'_{kr} w'_{0kr} + w'_{0kr} w'_{0kr}) \right)$$  \hspace{1cm} (102)

$$= -\frac{n_y K d}{2} \ln(2\pi) + \frac{d}{2} \sum_{k=1}^{K} \ln |L_{W_{ok}}|$$

$$- \frac{d}{2} \sum_{k=1}^{K} \text{tr} \left( L_{W_{ok}} L_{W_{ok}}^{-1} \right) - \frac{1}{2} \sum_{k=1}^{K} \sum_{r=1}^{d} (w'_{kr} - w'_{0kr})^T L_{W_{ok}} (w'_{kr} - w'_{0kr})$$  \hspace{1cm} (103)

$$= -\frac{n_y K d}{2} \ln(2\pi) + \frac{d}{2} \sum_{k=1}^{K} \ln |L_{W_{ok}}|$$

$$- \frac{d}{2} \sum_{k=1}^{K} \text{tr} \left( L_{W_{ok}} L_{W_{ok}}^{-1} \right) - \frac{1}{2} \sum_{k=1}^{K} \sum_{r=1}^{d} (w'_{kr} - w'_{0kr})^T L_{W_{ok}} (w'_{kr} - w'_{0kr})$$  \hspace{1cm} (104)

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

### 7 Variational inference with Gaussian-Gamma priors for high rank W

The amount of memory needed for the factor analyser grows quadratically with the dimension of the factor vector $n_y$. Due to that, we are limited to use small i-vectors ($n_y < 1000$). We are going to modify the variational partition function so that the memory grows linearly with the number of factors. We derive the equations for the case of Gaussian-Gamma prior for $W$.

#### 7.1 Variational distributions

We choose the partition function:

$$P(Y, W, \alpha | X) \approx q(Y, W, \alpha) = \prod_{p=1}^{P} q(Y^{(p)}) q(W^{(p)}) q(\alpha)$$  \hspace{1cm} (106)

where

$$Y = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(P)} \end{bmatrix}$$  \hspace{1cm} (107)

We define the blocks $W_{k}, W^{(p)}$ and $W_{k}^{(p)}$ of $W$ such as

$$W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_K \end{bmatrix} = \begin{bmatrix} W^{(1)} \\ W^{(2)} \\ \vdots \\ W^{(P)} \end{bmatrix} = \begin{bmatrix} W_{1}^{(1)} & W_{1}^{(2)} & \ldots & W_{1}^{(P)} \\ W_{2}^{(1)} & W_{2}^{(2)} & \ldots & W_{2}^{(P)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{K}^{(1)} & W_{K}^{(2)} & \ldots & W_{K}^{(P)} \end{bmatrix}$$  \hspace{1cm} (108)
This partition function assumes that there are groups of components of the i-vectors that are independent between them in the posterior. For example, the components in \( Y_1 \) would be independent from the components in \( Y_2 \) but the components inside \( Y_i \) would be dependent between them. We are going to assume that every group has the same number of components \( n_y = n_y / P \).

The optimum for \( q^\ast(Y^{(p)}) \):

\[
\ln q^\ast(Y^{(p)}) = E_{W,Y^{(p)}} \left[ \ln P(X,Y,W) \right] + \text{const} \tag{109}
\]

\[
= E_{W,Y^{(p)}} \left[ \ln P(X|Y,W) \right] + \ln P(Y^{(p)}) + \text{const} \tag{110}
\]

\[
= \sum_{i=1}^{H} E_{Y^{(p)}} [y_i]^T E[W]^T f_i
\]

\[
- \frac{1}{2} \sum_{k=1}^{K} \frac{N_{ik}}{N} E_{W,Y^{(p)}} \left[ y_i^{(n)} y_i^{(m)} \right] - \frac{1}{2} \gamma(y^{(p)} y^{(p)}) + \text{const} \tag{111}
\]

\[
= \sum_{i=1}^{H} y_i^{(p)} E[W_i] f_i - \frac{1}{2} y_i^{(p)} \left( \mathbf{I} + \sum_{k=1}^{K} N_{ik} E \left[ W_k^{(p)} W_k^{(p)} \right] \right) y_i^{(p)} + \text{const} \tag{112}
\]

\[
= \sum_{i=1}^{H} y_i^{(p)} F_i - \frac{1}{2} y_i^{(p)} \left( \mathbf{I} + \sum_{k=1}^{K} N_{ik} E \left[ W_k^{(p)} W_k^{(p)} \right] \right) y_i^{(p)} + \text{const} \tag{113}
\]

\[
= \sum_{i=1}^{H} y_i^{(p)} \sum_{k=1}^{K} \left( E \left[ W_k^{(p)} \right] f_{ik} - N_{ik} \sum_{n \neq p} E \left[ W_k^{(p)} W_n^{(n)} \right] y_i^{(n)} \right)
\]

\[
- \frac{1}{2} y_i^{(p)} \left( \mathbf{I} + \sum_{k=1}^{K} N_{ik} E \left[ W_k^{(p)} W_k^{(p)} \right] \right) y_i^{(p)} + \text{const} \tag{114}
\]

\[
= \sum_{i=1}^{H} y_i^{(p)} \sum_{k=1}^{K} \left( E \left[ W_k^{(p)} \right] f_{ik} - N_{ik} \sum_{n \neq p} E \left[ W_k^{(p)} W_n^{(n)} \right] y_i^{(n)} \right)
\]

\[
- \frac{1}{2} y_i^{(p)} \left( \mathbf{I} + \sum_{k=1}^{K} N_{ik} E \left[ W_k^{(p)} W_k^{(p)} \right] \right) y_i^{(p)} + \text{const} \tag{115}
\]

\[
= \sum_{i=1}^{H} y_i^{(p)} \left( F_i - N_{i} \sum_{n \neq p} E \left[ W_n^{(n)} \right] y_i^{(n)} \right)
\]

\[
- \frac{1}{2} y_i^{(p)} \left( \mathbf{I} + \sum_{k=1}^{K} N_{ik} E \left[ W_k^{(p)} W_k^{(p)} \right] \right) y_i^{(p)} + \text{const} \tag{116}
\]

Therefore \( q^\ast(Y^{(p)}) \) is a product of Gaussian distributions.

\[
q^\ast(Y^{(p)}) = \prod_{i=1}^{H} \mathcal{N} \left( y_i^{(p)} | \mathbf{y}_i^{(p)}, \mathbf{L}_i^{(p)} \right) \tag{117}
\]

\[
\mathbf{L}_i^{(p)} = \mathbf{I} + \sum_{k=1}^{K} N_{ik} E \left[ W_k^{(p)} W_k^{(p)} \right] \tag{118}
\]

\[
\mathbf{y}_i^{(p)} = \mathbf{y}_i^{(p)} - \mathbf{F}_i - N_{i} \sum_{n \neq p} E \left[ W_n^{(n)} \right] y_i^{(n)} \tag{119}
\]
The optimum for $q^* (\mathbf{W}^{(p)})$:

$$
\ln q^* (\mathbf{W}^{(p)}) = E_{X,Y} [\ln P (X,Y,W,\alpha)] + \text{const}
$$

$$
= E_{X,Y} [\ln P (X|Y,W)] + E_{W^{(p)},\alpha} [\ln P (W|\alpha)] + \text{const}
$$

$$
= \sum_{i=1}^{H} \left( E \left[ y_i^{(p)} \right]^{T} \mathbf{W}^{(p)} \mathbf{F}_i \right) - \frac{1}{2} E_{X,Y} [E_{W^{(p)}} \left[ y_i^{T} W^T N_i W y_i \right]]
$$

$$
- \frac{1}{2} \sum_{q=1}^{n_p} E \left[ \alpha_{q}^{(p)} \right] w_q^{(p)} w_q^{(p)} + \text{const}
$$

$$
= \sum_{i=1}^{H} \left( E \left[ y_i^{(p)} \right]^{T} \mathbf{W}^{(p)} \mathbf{F}_i \right) - \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{P} E_{W^{(p)},\alpha} \left[ y_i^{(n)}^{T} W_k^{(n)} W_k^{(n)} y_i^{(n)} \right]
$$

$$
- \frac{1}{2} \sum_{q=1}^{n_p} E \left[ \alpha_{q}^{(p)} \right] w_q^{(p)} w_q^{(p)} + \text{const}
$$

$$
= \text{tr} \left( \mathbf{W}^{(p)} \sum_{i=1}^{H} \left( \mathbf{F}_i - N_i \sum_{n \neq p} E \left[ W^{(n)} \right] E \left[ y_i^{(n)} \right] \right) E \left[ y_i^{(p)} \right]^{T} \right)
$$

$$
- \frac{1}{2} \sum_{k=1}^{K} W_k^{(p)} W_k^{(p)} \sum_{i=1}^{H} N_i E \left[ y_i^{(p)} y_i^{(p)T} \right]
$$

$$
- \frac{1}{2} \sum_{q=1}^{n_p} E \left[ \alpha_{q}^{(p)} \right] w_q^{(p)} w_q^{(p)} + \text{const}
$$

$$
= \text{tr} \left( \mathbf{W}^{(p)} \mathbf{C}^{(p)} - \frac{1}{2} \sum_{k=1}^{K} W_k^{(p)} W_k^{(p)} R_k^{(p)} \right)
$$

$$
- \frac{1}{2} \sum_{q=1}^{n_p} E \left[ \alpha_{q}^{(p)} \right] w_q^{(p)} w_q^{(p)} + \text{const}
$$

$$
= \sum_{k=1}^{K} \text{tr} \left( \mathbf{W}_k^{(p)} \mathbf{C}_k^{(p)} - \frac{1}{2} \mathbf{W}_k^{(p)} \mathbf{W}_k^{(p)} R_k^{(p)} \right)
$$

$$
- \frac{1}{2} \sum_{k=1}^{d} \sum_{r=1}^{d} \mathbf{w}_{kr}^{(p)T} \text{diag} \left( E \left[ \alpha_{q}^{(p)} \right] \right) \mathbf{w}_{kr}^{(p)} + \text{const}
$$

$$
= \sum_{k=1}^{K} \sum_{r=1}^{d} \text{tr} \left( \mathbf{w}_{kr}^{(p)} \mathbf{C}_{kr}^{(p)} - \frac{1}{2} \mathbf{w}_{kr}^{(p)} \mathbf{w}_{kr}^{(p)T} \left( E \left[ \alpha_{q}^{(p)} \right] + R_k^{(p)} \right) \right) + \text{const}
$$

where $\mathbf{w}_{kr}^{(p)}$ is a column vector containing the $r^{th}$ row of $\mathbf{W}_k^{(p)}$,

$$
\mathbf{C}^{(p)} = \sum_{i=1}^{H} \left( \mathbf{F}_i - N_i \sum_{n \neq p} E \left[ W^{(n)} \right] E \left[ y_i^{(n)} \right] \right) E \left[ y_i^{(p)} \right]^{T}
$$

$$
\mathbf{R}_k^{(p)} = \sum_{i=1}^{H} N_i E \left[ y_i^{(p)} y_i^{(p)T} \right]
$$

and $\mathbf{C}_{kr}^{(p)}$ is the $r^{th}$ of the block of $\mathbf{C}^{(p)}$ corresponding to component $k$ (row $(k-1) \times d + r$).
Then \( q^* (W^{(p)}) \) is a product of Gaussian distributions:

\[
q^* (W^{(p)}) = \prod_{k=1}^{K} \prod_{r=1}^{d} \mathcal{N} (w_{kr}^{(p)}, \mathbb{W}_{kr}^{(p)}, L_{W_{kr}}^{(p)-1})
\]  

\[
L_{W_{kr}}^{(p)} = \mathbb{E} [\alpha_k^{(p)}] + R_{k}^{(p)}
\]

\[
\mathbb{W}_{kr}^{(p)} = \mathbb{L}_{W_{kr}}^{(p)-1} C_{kr}^{(p)T}
\]

The optimum for \( q^* (\alpha) \) is the same as in equation (35).

We need to evaluate the expectations:

\[
\mathbb{E} [w_q^{(p)T} w_q^{(p)}] = \sum_{k=1}^{K} dL_{W_{kq}}^{(p)-1} + \sum_{r=1}^{d} \mathbb{W}_{rkq}^{(p)^2}
\]

\[
\mathbb{E} [W_k^{(p)T} W_k^{(p)T}] = \mathbb{L}_{W_{k}}^{(p)-1} + \mathbb{E} [W_k^{(p)}]^{T} \mathbb{E} [W_k^{(p)}]
\]

### 7.2 Variational lower bound

The lower bound is given by

\[
\mathcal{L} = \mathbb{E}_{Y,W} [\ln P (X|Y, W)] + \mathbb{E}_{Y} [\ln P (Y)] + \mathbb{E}_{W, \alpha} [\ln P (W|\alpha)] + \mathbb{E}_{\alpha} [\ln P (\alpha)]
\]

\[
- \sum_{p=1}^{P} \mathbb{E}_{Y^{(p)}} [\ln q (Y^{(p)})] - \sum_{p=1}^{P} \mathbb{E}_{W^{(p)}} [\ln q (W^{(p)})] - \mathbb{E}_{\alpha} [\ln q (\alpha)]
\]

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

\[
\mathbb{E}_{Y,W} [\ln P (X|Y, W)] = - \sum_{k=1}^{K} \frac{N_k d}{2} \log (2\pi) - \frac{1}{2} \text{tr} \left( \sum_{k=1}^{K} \mathbb{S}_k \right) + \sum_{i=1}^{H} \mathbb{E} [y_i]^{T} \mathbb{E} [W]^{T} \mathbb{F}_i
\]

\[
- \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{H} \sum_{n=1}^{P} \sum_{m=1}^{P} \text{tr} \left( N_{ik} \mathbb{E} [W_{k}^{(n)T} W_{k}^{(m)}] \mathbb{E} [y_i^{(m)} y_i^{(n)T}] \right)
\]

\[
= - \sum_{k=1}^{K} \frac{N_k d}{2} \log (2\pi) - \frac{1}{2} \text{tr} \left( \sum_{k=1}^{K} \mathbb{S}_k \right) + \text{tr} (\mathbb{E} [W]^{T} \mathbb{C})
\]

\[
- \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{P} \text{tr} \left( \mathbb{E} [W_{k}^{(n)T} W_{k}^{(n)}] \right) R_{k}^{(n)}
\]

\[
+ 2 \sum_{m=n+1}^{P} \mathbb{E} [W_{k}^{(n)}]^{T} \mathbb{E} [W_{k}^{(m)}] R_{k}^{(m,n)}
\]

where

\[
R_{k}^{(m,n)} = \sum_{i=1}^{H} N_{ik} \mathbb{E} [y_i^{(m)}] \mathbb{E} [y_i^{(n)T}]
\]

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

\[
\mathbb{E}_{Y} [\ln P (Y)] = - \frac{H n_2}{2} \ln (2\pi) - \frac{1}{2} \text{tr} \left( \sum_{i=1}^{H} \mathbb{E} [y_i y_i^{T}] \right)
\]

\[
= - \frac{H n_2}{2} \ln (2\pi) - \frac{1}{2} \sum_{p=1}^{P} \text{tr} \left( \sum_{i=1}^{H} \mathbb{E} [y_i^{(p)} y_i^{(p)T}] \right)
\]

\[
= - \frac{H n_2}{2} \ln (2\pi) - \frac{1}{2} \sum_{p=1}^{P} \text{tr} (P^{(p)})
\]
where

\[ P^{(p)} = \sum_{i=1}^{H} E \left[ y_i^{(p)} y_i^{(p)T} \right] \] (143)

The term \( E_{Y^{(p)}} [\ln q (Y^{(p)})] \):

\[ E_{Y^{(p)}} [\ln q (Y^{(p)})] = - \frac{H \tilde{n}_y}{2} (\ln(2\pi) + 1) + \frac{1}{2} \sum_{i=1}^{H} \ln |L_{y_i}^{(p)}| \] (144)

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

\[ E_{W^{(p)}} [\ln q (W^{(p)})] = - \frac{K \tilde{n}_w}{2} (\ln(2\pi) + 1) + \frac{d}{2} \sum_{k=1}^{K} \ln |L_{W_k}^{(p)}| \] (145)

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

References

[1] Patrick Kenny, “Joint factor analysis of speaker and session variability : Theory and algorithms - Technical report CRIM-06/08-13,” Tech. Rep., CRIM, Montreal, 2005.

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

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