Bayesian Two-Covariance Model Integrating Out the Speaker Space Distribution

Jesús Villalba, Niko Brümmer

October 20, 2010

1 Introduction

The two-covariance model presented in [1] for speaker verification is a two steps process:

- Given a development database, make a ML point estimate of the 2-covariance model by EM iterations.
- Given a train and test segment and the estimated model, calculate the target vs non-target posterior probability.

In this work, we intend to do a Bayesian treatment of the two-covariance model as described on [2]. Using the posterior of the model given the data, we are going to integrate out the parameters describing the speakers distribution.

2 The Model

2.1 Notation

We are going to introduce some notation:

- Let $\mathbf{\Phi}_\mathrm{d}$ be the development i-vectors dataset.
- Let $\mathbf{\Phi}_\mathrm{t} = \{l, r\}$ be the test i-vectors.
- Let $\overline{\mathbf{\Phi}} = \mathbf{\Phi}_\mathrm{d} \cup \mathbf{\Phi}_\mathrm{t}$
- Let $\mathbf{\Phi}$ be any of the previous datasets.
- Let θ_d be the labelling of the development dataset. It partitions the N_d i-vectors into M_d speakers.
- Let θ_t be the labelling of the test set, so that $\theta_\mathrm{t} \in \{\mathcal{T}, \mathcal{N}\}$, where $\mathcal{T}$ is the hypothesis that l and r belong to the same speaker and $\mathcal{N}$ is the hypothesis that they belong to different speakers.
- Let $\overline{\theta} = \theta_\mathrm{d} \wedge \theta_\mathrm{t}$ the labelling of $\overline{\mathbf{\Phi}}$
- Let θ be any of the previous labellings.
- Let $\mathcal{S}_i$ be the i-vectors belonging to the speaker i.
- Let $\mathbf{Y}_\mathrm{d}$ be the speaker identity variables of the development set. We will have as many identity variables as speakers.
- Let $\mathbf{Y}_\mathrm{t}$ be the speaker identity variables of the test set.
- Let $\overline{\mathbf{Y}} = \mathbf{Y}_\mathrm{d} \cup \mathbf{Y}_\mathrm{t}$
- Let $\mathbf{Y}$ be any of the previous speaker identity variables sets.
- Let $\pi = (P_\mathcal{T}, P_\mathcal{N})$ be the hypothesis prior where $P_\mathcal{T}$ is the prior probability of a target trial and $P_\mathcal{N} = 1 - P_\mathcal{T}$ be the prior probability of a non target trial.
- Let $\mathcal{M} = (\mu, \mathbf{B}, \mathbf{W})$ be the set of all the parameters of the model and $\mathcal{M}_y = (\mu, \mathbf{B})$

2.2 The two-covariance model

We take a linear-Gaussian generative model $\mathcal{M}$. We suppose that an i-vector ϕ of speaker s can be written as:

$$\phi = \mathbf{y}_s + \mathbf{z} \tag{1}$$

where $\mathbf{y}_s$ is the speaker identity variable and $\mathbf{z}$ is a channel offset.

We assume the following probability distributions:

$$P(\mathbf{y}|\mathcal{M}) = \mathcal{N}\left(\mathbf{y}|\mu, \mathbf{B}^{-1}\right) \tag{2}$$

$$P(\phi|\mathbf{y}, \mathcal{M}) = \mathcal{N}\left(\phi|\mathbf{y}, \mathbf{W}^{-1}\right) \tag{3}$$

where $\mathcal{N}$ denotes a Gaussian distribution; μ is the speakers mean; $\mathbf{B}^{-1}$ is the between class covariance matrix, $\mathbf{W}^{-1}$ is the within class covariance matrix; and $\mathbf{B}$ and $\mathbf{W}$ are the precision matrices.

3 Bayesian two-covariance model

Following a Bayesian treatment, instead of assuming fixed values for μ and $\mathbf{B}$ we are going to work with a probability distribution for model parameters. For this, we need a prior for the model $P(\mathcal{M}_y|\Pi)$ and calculate the posterior distribution of the model given the data, the labelling, $\mathbf{W}$ and the prior:

$$P(\mathcal{M}_y|\mathbf{\Phi}, \theta, \mathbf{W}, \Pi) \tag{4}$$

In Figure 1 we show the graphical representation of this model. $\mathbf{\Phi}_{\mathrm{d}}$, $\mathbf{\Phi}_{\mathrm{t}}$ and θ_{d} are observed variables; μ, $\mathbf{B}$ and θ_{t} are hidden variables; $\mathbf{W}$ is assumed to be known; and π and Π are the priors over the hidden variables. We can use the rules described in [3] on this graphical model to determine the dependencies between variables.

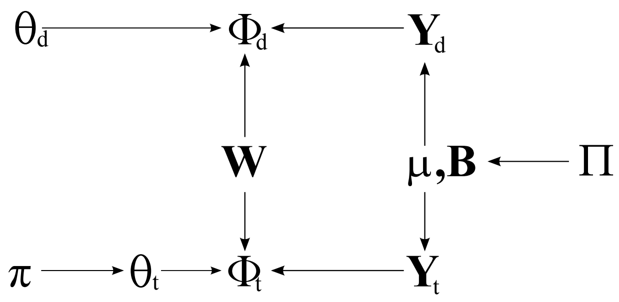

Figure 1: Graphical model of the Bayesian two-covariance model.

We are going to use a non informative prior (Jeffreys prior) for the parameters μ and $\mathbf{B}$ of the speaker Gaussian distribution as in [4]. We are assuming that $\mathbf{W}$ is known for simplicity, we will use the ML point estimate on the experiments.

4 Likelihood ratio given a known model

Given a model $\mathcal{M}$ we can calculate the ratio of the posterior probabilities of target and non target as shown in [1]:

$$\frac{P(\mathcal{T}|\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M}, \pi)}{P(\mathcal{N}|\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M}, \pi)} = \frac{P_{\mathcal{T}}}{P_{\mathcal{N}}} \frac{P(\mathbf{\Phi}_{\mathrm{t}}|\mathcal{T}, \mathcal{M})}{P(\mathbf{\Phi}_{\mathrm{t}}|\mathcal{N}, \mathcal{M})} = \frac{P_{\mathcal{T}}}{P_{\mathcal{N}}} R(\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M}) \tag{5}$$

where we have defined the plug-in likelihood ratio $R(\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M})$. To get this ratio we need to calculate $P(\mathbf{\Phi}|\theta, \mathcal{M})$. Given a model $\mathcal{M}$, the $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M \in \mathbf{Y}$ are sampled independently from $P(\mathbf{y}|\mathcal{M})$.

Besides, given the $\mathcal{M}$ and a speaker i the set $\mathcal{S}_i$ of i-vectors produced by that speaker are drawn independently from $P(\mathbf{\Phi}|\mathbf{y}_i, \mathcal{M})$. Using these independence assumptions we can write:

$$P(\mathbf{\Phi}|\theta, \mathcal{M}) = \prod_{i=1}^{M} P(\mathcal{S}_i|\mathcal{M}) \tag{6}$$

$$P(\mathcal{S}_i|\mathbf{y}, \mathcal{M}) = \prod_{\phi\in\mathcal{S}_i} P(\phi|\mathbf{y}, \mathcal{M}) \tag{7}$$

Using Bayes rule we can write:

$$P(\mathcal{S}_i, \mathbf{y}_0|\mathcal{M}) = P(\mathcal{S}_i|\mathbf{y}_0, \mathcal{M}) P(\mathbf{y}_0|\mathcal{M}) = P(\mathbf{y}_0|\mathcal{S}_i, \mathcal{M}) P(\mathcal{S}_i|\mathcal{M}) \tag{8}$$

and then

$$P(\mathcal{S}_i|\mathcal{M}) = \frac{P(\mathcal{S}_i|\mathbf{y}_0, \mathcal{M}) P(\mathbf{y}_0|\mathcal{M})}{P(\mathbf{y}_0|\mathcal{S}_i, \mathcal{M})} \tag{9}$$

where $\mathbf{y}_0$ is whatever value we want as long as the denominator is not zero. Note that $\mathbf{y}_0$ is not in the LHS of the equation so the result does not depend on it.

Then, for the complete dataset:

$$P(\mathbf{\Phi}|\theta, \mathcal{M}) = \prod_{i=1}^{M} \frac{P(\mathcal{S}_i|\mathbf{y}_0, \mathcal{M}) P(\mathbf{y}_0|\mathcal{M})}{P(\mathbf{y}_0|\mathcal{S}_i, \mathcal{M})} = K(\mathbf{\Phi}) L(\theta|\mathbf{\Phi}) \tag{10}$$

where $K(\mathbf{\Phi}) = \prod_{i=1}^{N} P(\phi_j|\mathbf{y}_0, \mathcal{M})$ is a term that only dependent on the dataset, not θ, so it vanishes when doing the ratio and we do not need to calculate it. What we need to calculate is:

$$L(\theta|\mathbf{\Phi}) = \prod_{i=1}^{M} Q(\mathcal{S}_i) \tag{11}$$

$$Q(\mathcal{S}_i) = \frac{P(\mathbf{y}_0|\mathcal{M})}{P(\mathbf{y}_0|\mathcal{S}_i, \mathcal{M})} \tag{12}$$

and the likelihood ratio is:

$$R(\mathbf{\Phi}_\mathrm{t}, \mathcal{M}) = \frac{Q(\{l, r\})}{Q(\{l\}) Q(\{r\})} \tag{13}$$

Distribution 2 is a conjugate prior for distribution 3 so the posterior is again Gaussian distributed:

$$P(\mathbf{y}|\mathcal{S}, \mathcal{M}) = \mathcal{N}\left(\mathbf{y}|\mathbf{L}^{-1}\gamma, \mathbf{L}^{-1}\right) \tag{14}$$

$$\mathbf{L} = \mathbf{B} + n\mathbf{W} \tag{15}$$

$$\gamma = \mathbf{B}\mu + \mathbf{W}\sum_{\phi\in\mathcal{S}} \phi \tag{16}$$

where n is the number of i-vectors of $\mathcal{S}$.

Making $\mathbf{y}_0 = 0$ we can use this posterior to calculate $Q(\mathcal{S})$

$$\ln Q(\mathcal{S}) = \frac{1}{2}\left(\ln|\mathbf{B}| - \mu^T\mathbf{B}\mu - \ln|\mathbf{L}| + \gamma^T\mathbf{L}^{-1}\gamma\right) \tag{17}$$

5 Bayesian likelihood ratio

In the Bayesian case, given the development and test data and $\mathbf{W}$ we can calculate the ratio of the posterior probabilities of target and non target as shown in [1]:

$$\frac{P(\mathcal{T}|\mathbf{\Phi}_\mathrm{t}, \mathbf{\Phi}_\mathrm{d}, \theta_\mathrm{d}, \mathbf{W}, \Pi, \pi)}{P(\mathcal{N}|\mathbf{\Phi}_\mathrm{t}, \mathbf{\Phi}_\mathrm{d}, \theta_\mathrm{d}, \mathbf{W}, \Pi, \pi)} = \frac{P_\mathcal{T}}{P_\mathcal{N}} \frac{P(\mathbf{\Phi}_\mathrm{t}|\mathcal{T}, \mathbf{\Phi}_\mathrm{d}, \theta_\mathrm{d}, \mathbf{W}, \Pi)}{P(\mathbf{\Phi}_\mathrm{t}|\mathcal{N}, \mathbf{\Phi}_\mathrm{d}, \theta_\mathrm{d}, \mathbf{W}, \Pi)} = \frac{P_\mathcal{T}}{P_\mathcal{N}} R\left(\overline{\mathbf{\Phi}}, \theta_\mathrm{d}, \mathbf{W}, \Pi\right) \tag{18}$$

where $R\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right)$ is the Bayesian likelihood ratio.

Using Bayes rule we can write

$$\begin{aligned} P\left(\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M}_y | \theta_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = & \\ & = P\left(\mathbf{\Phi}_{\mathrm{t}} | \mathcal{M}_y, \theta_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) P\left(\mathcal{M}_y | \theta_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) \\ & = P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \overline{\theta}, \mathbf{W}, \Pi\right) P\left(\mathbf{\Phi}_{\mathrm{t}} | \theta_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) \end{aligned} \tag{19}$$

Using the conditional independence assumptions given by the graphic model we have:

$$P\left(\mathbf{\Phi}_{\mathrm{t}} | \mathcal{M}_y, \theta_{\mathrm{t}}, \mathbf{W}\right) P\left(\mathcal{M}_y | \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \overline{\theta}, \mathbf{W}, \Pi\right) P\left(\mathbf{\Phi}_{\mathrm{t}} | \theta_{\mathrm{t}}, \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) \tag{20}$$

Then, we can write the ratio in 18 like:

$$R\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = \frac{P\left(\mathbf{\Phi}_{\mathrm{t}} | \mathcal{M}_y, \mathcal{T}, \mathbf{W}\right)}{P\left(\mathbf{\Phi}_{\mathrm{t}} | \mathcal{M}_y, \mathcal{N}, \mathbf{W}\right)} \frac{P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{N}, \mathbf{W}, \Pi\right)}{P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{T}, \mathbf{W}, \Pi\right)} \tag{21}$$

$$= R\left(\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M}_y, \mathbf{W}\right) \frac{P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{N}, \mathbf{W}, \Pi\right)}{P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{T}, \mathbf{W}, \Pi\right)} \tag{22}$$

So the Bayesian likelihood ratio is equal to the plug-in likelihood ratio multiplied by a correction factor. Note that the LHS of the equation does not dependent on the model $\mathcal{M}_y$ used in the computation so we can use whatever $\mathcal{M}_y$ that we find convenient.

Therefore, now we need to calculate $P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \overline{\theta}, \mathbf{W}, \Pi\right)$ that is

$$P\left(\mathcal{M}_y | \overline{\mathbf{\Phi}}, \overline{\theta}, \mathbf{W}, \Pi\right) = \int_{\mathcal{Y}} P\left(\mathcal{M}_y | \overline{\mathbf{Y}}, \Pi\right) P\left(\overline{\mathbf{Y}} | \overline{\mathbf{\Phi}}, \overline{\theta}, \mathbf{W}, \Pi\right) \, \mathrm{d}\overline{\mathbf{Y}} \tag{23}$$

Solving this integral is complicated so we are going to use the Variational Bayes method to estimate the posterior distribution of $\mathcal{M}_y$.

6 VB likelihood ratio

6.1 Likelihood ratio approximation

The Variational Bayes approach [3] is based on the approximation:

$$P\left(\mathcal{M}_y, \mathbf{Y} | \mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) \approx q\left(\mathcal{M}_y, \mathbf{Y}\right) = q\left(\mathcal{M}_y\right) q\left(\mathbf{Y}\right) \tag{24}$$

We need to write $R\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right)$ as a function of $P\left(\mathcal{M}_y, \mathbf{Y} | \mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right)$. Again we use Bayes rule:

$$P\left(\mathcal{M}_y, \mathbf{Y} | \mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) = P\left(\mathbf{Y} | \mathcal{M}_y, \mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) P\left(\mathcal{M}_y | \mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) \tag{25}$$

Simplifying with the conditional independence rules:

$$P\left(\mathcal{M}_y, \mathbf{Y} | \mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) = P\left(\mathbf{Y} | \mathcal{M}_y, \mathbf{\Phi}, \theta, \mathbf{W}\right) P\left(\mathcal{M}_y | \mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) \tag{26}$$

If we use 26 in 21:

$$R\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = R\left(\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M}_y, \mathbf{W}\right) \frac{P\left(\mathcal{M}_y, \overline{\mathbf{Y}} | \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{N}, \mathbf{W}, \Pi\right)}{P\left(\mathcal{M}_y, \overline{\mathbf{Y}} | \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{T}, \mathbf{W}, \Pi\right)} \frac{P\left(\overline{\mathbf{Y}} | \mathcal{M}_y, \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{T}, \mathbf{W}\right)}{P\left(\overline{\mathbf{Y}} | \mathcal{M}_y, \overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{N}, \mathbf{W}\right)} \tag{27}$$

Now we can use the fact that, given the model $\mathcal{M}_y$, the $\mathbf{y}$ of each speaker are drown independently among then:

$$P\left(\mathbf{Y} | \mathcal{M}_y, \mathbf{\Phi}, \theta, \mathbf{W}\right) = \prod_{i=1}^{N} P\left(\mathbf{y}_i | \mathcal{M}_y, \mathcal{S}_i, \mathbf{W}\right) \tag{28}$$

and

$$P\left(\overline{\mathbf{Y}} | \mathcal{M}_y, \overline{\mathbf{\Phi}}, \overline{\theta}, \mathbf{W}\right) = P\left(\mathbf{Y}_{\mathrm{t}} | \mathcal{M}_y, \mathbf{\Phi}_{\mathrm{t}}, \theta_{\mathrm{t}}, \mathbf{W}\right) P\left(\mathbf{Y}_{\mathrm{d}} | \mathcal{M}_y, \mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}\right) \tag{29}$$

If we plug 29 into 27 we have

$$R\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = R\left(\mathbf{\Phi}_{\mathrm{t}}, \mathcal{M}_y, \mathbf{W}\right) \frac{P\left(\mathcal{M}_y, \overline{\mathbf{Y}}|\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{N}, \mathbf{W}, \Pi\right)}{P\left(\mathcal{M}_y, \overline{\mathbf{Y}}|\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{T}, \mathbf{W}, \Pi\right)} \frac{P\left(\mathbf{Y}_{\mathrm{t}}|\mathcal{M}_y, \mathbf{\Phi}_{\mathrm{t}}, \mathcal{T}, \mathbf{W}\right)}{P\left(\mathbf{Y}_{\mathrm{t}}|\mathcal{M}_y, \mathbf{\Phi}_{\mathrm{t}}, \mathcal{N}, \mathbf{W}\right)} \tag{30}$$

Note again that the LHS does not depend on $\mathcal{M}_y$ or $\overline{\mathbf{Y}}$ so we can use whatever we want. Now we make $\mathbf{y} = \mathbf{y}_0$ for all speakers in $\overline{\mathbf{\Phi}}$, what we denote as $\overline{\mathbf{Y}} = \overline{\mathbf{Y}}_0$. We take equation 28 and the plug-in likelihood in equation 11 and substitute them in 30. Finally simplifying we get:

$$R\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = \frac{1}{P\left(\mathbf{y}_0|\mathcal{M}_y\right)} \frac{P\left(\mathcal{M}_y, \overline{\mathbf{Y}}_0|\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{N}, \mathbf{W}, \Pi\right)}{P\left(\mathcal{M}_y, \overline{\mathbf{Y}}_0|\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{T}, \mathbf{W}, \Pi\right)} \tag{31}$$

and the Variational Bayes version:

$$R\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) \approx R_{VB}\left(\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = \frac{1}{P\left(\mathbf{y}_0|\mathcal{M}_y\right)} \frac{q_{\mathcal{N}}\left(\mathcal{M}_y, \overline{\mathbf{Y}}_0\right)}{q_{\mathcal{T}}\left(\mathcal{M}_y, \overline{\mathbf{Y}}_0\right)} \tag{32}$$

where $q_{\mathcal{T}}\left(\mathcal{M}_y, \overline{\mathbf{Y}}_0\right)$ and $q_{\mathcal{N}}\left(\mathcal{M}_y, \overline{\mathbf{Y}}_0\right)$ are the variational posteriors assuming that $\theta_{\mathrm{t}} = \{\mathcal{T}\}$ and $\theta_{\mathrm{t}} = \{\mathcal{N}\}$ respectively.

6.2 Variational distributions

In order to formulate the variational version of this model we need to write down the joint distribution of the random variables:

$$P\left(\mathbf{\Phi}, \mathcal{M}_y, \mathbf{Y}|\theta, \mathbf{W}, \Pi\right) = P\left(\mathbf{\Phi}|\mathcal{M}_y, \mathbf{Y}, \theta, \mathbf{W}, \Pi\right) P\left(\mathbf{Y}|\mathcal{M}_y, \theta, \mathbf{W}, \Pi\right) P\left(\mathcal{M}_y|\theta, \mathbf{W}, \Pi\right) \tag{33}$$

and simplifying:

$$P\left(\mathbf{\Phi}, \mathcal{M}_y, \mathbf{Y}|\theta, \mathbf{W}, \Pi\right) = P\left(\mathbf{\Phi}|\mathbf{Y}, \theta, \mathbf{W}\right) P\left(\mathbf{Y}|\mathcal{M}_y\right) P\left(\mathcal{M}_y|\Pi\right) \tag{34}$$

Now we consider the variational distribution:

$$q\left(\mathcal{M}_y, \mathbf{Y}\right) = q\left(\mathcal{M}_y\right) q\left(\mathbf{Y}\right) \tag{35}$$

The optimum for the factor $q\left(\mathbf{Y}\right)$ is given by

$$\ln q^*\left(\mathbf{Y}\right) = \mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathbf{\Phi}, \mathcal{M}_y, \mathbf{Y}|\theta, \mathbf{W}, \Pi\right)\right] + \mathrm{const} \tag{36}$$

Using decomposition 34 we have:

$$\ln q^*\left(\mathbf{Y}\right) = \mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathbf{\Phi}|\mathbf{Y}, \theta, \mathbf{W}\right)\right] + \mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathbf{Y}|\mathcal{M}_y\right)\right] + \mathrm{const} \tag{37}$$

Substituting equations 2, 3 and 7 into 37 and absorbing any term that does not depend on $\mathbf{Y}$ into the additive constant we have

$$\ln q^*(\mathbf{Y}) =$$

$$= \sum_{i=1}^{M} \sum_{\phi \in \mathcal{S}_i} \mathrm{E}_{\mathcal{M}_y} \left[\ln P(\phi | \mathbf{y}_i, \mathbf{W}) \right] + \sum_{i=1}^{M} \mathrm{E}_{\mathcal{M}_y} \left[\ln P(\mathbf{y}_i | \mathcal{M}_y) \right] + \mathrm{const} \tag{38}$$

$$= -\frac{1}{2} \sum_{i=1}^{M} \sum_{\phi \in \mathcal{S}_i} \mathrm{E}_{\mathcal{M}_y} \left[(\phi - \mathbf{y}_i)^T \mathbf{W} (\phi - \mathbf{y}_i) \right] - \frac{1}{2} \sum_{i=1}^{M} \mathrm{E}_{\mathcal{M}_y} \left[(\mathbf{y}_i - \mu)^T \mathbf{B} (\mathbf{y}_i - \mu) \right] + \mathrm{const} \tag{39}$$

$$= \sum_{i=1}^{M} \mathbf{y}_i^T \mathbf{W} \sum_{\phi \in \mathcal{S}_i} \phi - \frac{1}{2} \sum_{i=1}^{M} n_i \mathbf{y}_i^T \mathbf{W} \mathbf{y}_i \tag{40}$$

$$- \frac{1}{2} \sum_{i=1}^{M} \mathbf{y}_i^T \mathrm{E}_{\mathcal{M}_y} [\mathbf{B}] \, \mathbf{y}_i + \sum_{i=1}^{M} \mathbf{y}_i^T \mathrm{E}_{\mathcal{M}_y} [\mathbf{B} \mu] + \mathrm{const}$$

$$= \sum_{i=1}^{M} -\frac{1}{2} \mathbf{y}_i^T \left(\mathrm{E}_{\mathcal{M}_y} [\mathbf{B}] + n_i \mathbf{W} \right) \mathbf{y}_i + \mathbf{y}_i^T \left(\mathrm{E}_{\mathcal{M}_y} [\mathbf{B} \mu] + \mathbf{W} \sum_{\phi \in \mathcal{S}_i} \phi \right) + \mathrm{const} \tag{41}$$

Equation 41 has the form of the sum of logarithms of Gaussian distributions so finally:

$$q^*(\mathbf{Y}) = \prod_{i=1}^{M} q^*(\mathbf{y}_i) \tag{42}$$

$$q^*(\mathbf{y}_i) = \mathcal{N} \left(\mathbf{y}_i | \mathbf{L}_i^{-1} \gamma_i, \mathbf{L}_i^{-1} \right) \tag{43}$$

$$\mathbf{L}_i = \mathrm{E}_{\mathcal{M}_y} [\mathbf{B}] + n_i \mathbf{W} \tag{44}$$

$$\gamma_i = \mathrm{E}_{\mathcal{M}_y} [\mathbf{B} \mu] + \mathbf{W} \sum_{\phi \in \mathcal{S}_i} \phi \tag{45}$$

The optimum for the factor $q(\mathcal{M}_y)$ is given by

$$\ln q^*(\mathcal{M}_y) = \mathrm{E}_{\mathbf{Y}} \left[\ln P(\mathbf{\Phi}, \mathcal{M}_y, \mathbf{Y} | \theta, \mathbf{W}, \Pi) \right] + \mathrm{const} \tag{46}$$

Using decomposition 34 we have:

$$\ln q^*(\mathcal{M}_y) = \mathrm{E}_{\mathbf{Y}} \left[\ln P(\mathbf{Y} | \mathcal{M}_y) \right] + \mathrm{E}_{\mathbf{Y}} \left[\ln P(\mathcal{M}_y | \Pi) \right] + \mathrm{const} \tag{47}$$

Substituting equations 2, and 187 into 47 and absorbing any term that does not depend on μ or $\mathbf{B}$ into the additive constant we have

$$\ln q^*(\mathcal{M}_y) =$$

$$= \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} \left[\ln P(\mathbf{y}_i | \mathcal{M}_y) \right] + \mathrm{E}_{\mathbf{Y}} \left[\ln P(\mu, \mathbf{B} | \Pi) \right] + \mathrm{const} \tag{48}$$

$$= \frac{M}{2} \ln |\mathbf{B}| - \frac{1}{2} \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} \left[(\mathbf{y}_i - \mu)^T \mathbf{B} (\mathbf{y}_i - \mu) \right] + \frac{1}{2} \ln |\mathbf{B}| - \frac{d+1}{2} \ln |\mathbf{B}| + \mathrm{const} \tag{49}$$

Now we define:

$$\overline{\mathbf{y}} = \frac{1}{M} \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} [\mathbf{y}_i] \tag{50}$$

$$\mathbf{S}_y = \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} \left[(\mathbf{y}_i - \overline{\mathbf{y}}) (\mathbf{y}_i - \overline{\mathbf{y}})^T \right] = \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} \left[\mathbf{y}_i \mathbf{y}_i^T \right] - M \overline{\mathbf{y}} \overline{\mathbf{y}}^T \tag{51}$$

Now we can write 49 as

$$
\begin{aligned}
&\ln q^* (\mathcal{M}_y) = \\
&= \frac{M}{2} \ln |\mathbf{B}| - \frac{1}{2}\mathrm{tr} \left(\mathbf{B} \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} \left[(\mathbf{y}_i - \mu) (\mathbf{y}_i - \mu)^T \right] \right) + \frac{1}{2} \ln |\mathbf{B}| - \frac{d+1}{2} \ln |\mathbf{B}| + \text{const} && (52) \\
&= \frac{M}{2} \ln |\mathbf{B}| - \frac{1}{2}\mathrm{tr} \left(\mathbf{B} \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} \left[((\mathbf{y}_i - \overline{\mathbf{y}}) - (\mu - \overline{\mathbf{y}})) ((\mathbf{y}_i - \overline{\mathbf{y}}) - (\mu - \overline{\mathbf{y}}))^T \right] \right) && (53) \\
&\quad + \frac{1}{2} \ln |\mathbf{B}| - \frac{d+1}{2} \ln |\mathbf{B}| + \text{const} && (54) \\
&= \frac{M}{2} \ln |\mathbf{B}| - \frac{1}{2}\mathrm{tr} \left(\mathbf{B} \left(\sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}} \left[(\mathbf{y}_i - \overline{\mathbf{y}}) (\mathbf{y}_i - \overline{\mathbf{y}})^T \right] + M (\mu - \overline{\mathbf{y}}) (\mu - \overline{\mathbf{y}})^T \right) \right) && (55) \\
&\quad + \frac{1}{2} \ln |\mathbf{B}| - \frac{d+1}{2} \ln |\mathbf{B}| + \text{const} \\
&= \left[\frac{1}{2} \ln |\mathbf{B}| - \frac{M}{2} (\mu - \overline{\mathbf{y}})^T \mathbf{B} (\mu - \overline{\mathbf{y}}) \right] + \left[\frac{M}{2} \ln |\mathbf{B}| - \frac{d+1}{2} \ln |\mathbf{B}| - \frac{1}{2}\mathrm{tr} (\mathbf{B}\mathbf{S}_y) \right] + \text{const} && (56)
\end{aligned}
$$

If we compare equation 56 with equation 204 we can see that $q^* (\mathcal{M}_y)$ Gaussian-Wishart distributed:

$$
q^* (\mathcal{M}_y) = \mathcal{N} \left(\mu | \overline{\mathbf{y}}, (M\mathbf{B})^{-1} \right) \mathcal{W} \left(\mathbf{B} | \mathbf{S}_y^{-1}, M \right) \quad \text{if } M > d \tag{57}
$$

In order to estimate the parameters of the factoring distributions, we need to evaluate the expectations $\mathrm{E}_{\mathcal{M}_y} [\mathbf{B}]$, $\mathrm{E}_{\mathcal{M}_y} [\mathbf{B}\mu]$, $\mathrm{E}_{\mathbf{Y}} [y_i]$ and $\mathrm{E}_{\mathbf{Y}} \left[y_i y_i^T \right]$. Using the properties of the Gaussian and Wishart distributions [3] we have

$$
\mathrm{E}_{\mathcal{M}_y} [\mathbf{B}] = M \mathbf{S}_y^{-1} \tag{58}
$$

$$
\begin{aligned}
\mathrm{E}_{\mathcal{M}_y} [\mathbf{B}\mu] &= \int_{\mathbf{B}} \int_{\mu} \mathbf{B} \mu \mathcal{N} \left(\mu | \overline{\mathbf{y}}, (M\mathbf{B})^{-1} \right) \mathcal{W} \left(\mathbf{B} | \mathbf{S}_y^{-1}, M \right) \, \mathrm{d}\mu \, \mathrm{d}\mathbf{B} && (59) \\
&= \int_{\mathbf{B}} \mathbf{B} \int_{\mu} \mu \mathcal{N} \left(\mu | \overline{\mathbf{y}}, (M\mathbf{B})^{-1} \right) \, \mathrm{d}\mu \, \mathcal{W} \left(\mathbf{B} | \mathbf{S}_y^{-1}, M \right) \, \mathrm{d}\mathbf{B} && (60) \\
&= \int_{\mathbf{B}} \mathbf{B} \mathcal{W} \left(\mathbf{B} | \mathbf{S}_y^{-1}, M \right) \, \mathrm{d}\mathbf{B} \, \overline{\mathbf{y}} && (61) \\
&= M \mathbf{S}_y^{-1} \overline{\mathbf{y}} && (62)
\end{aligned}
$$

$$
\mathrm{E}_{\mathbf{Y}} [\mathbf{y}_i] = \mathbf{L}_i^{-1} \gamma_i \tag{63}
$$

$$
\begin{aligned}
\mathrm{E}_{\mathbf{Y}} \left[\mathbf{y}_i \mathbf{y}_i^T \right] &= \mathbf{L}_i^{-1} + \mathrm{E}_{\mathbf{Y}} [\mathbf{y}_i] \, \mathrm{E}_{\mathbf{Y}} [\mathbf{y}_i]^T && (64) \\
&= \mathbf{L}_i^{-1} + \mathbf{L}_i^{-1} \gamma_i \gamma_i^T \mathbf{L}_i^{-1} && (65)
\end{aligned}
$$

6.3 Variational lower bound

In this section, we are going to evaluate the variational lower bound to use to check the convergence of our algorithm.

$$
\begin{aligned}
\mathcal{L} &= \int_{\mathbf{Y}} \int_{\mathbf{B}} \int_{\mu} q (\mu, \mathbf{B}, \mathbf{Y}) \ln \left(\frac{P (\mathbf{\Phi}, \mu, \mathbf{B}, \mathbf{Y} | \theta, \mathbf{W}, \Pi)}{q (\mu, \mathbf{B}, \mathbf{Y})} \right) \mathrm{d}\mu \, \mathrm{d}\mathbf{B} \, \mathrm{d}\mathbf{Y} && (66) \\
&= \mathrm{E}_{\mathcal{M}_y, \mathbf{Y}} [\ln P (\mathbf{\Phi}, \mu, \mathbf{B}, \mathbf{Y} | \theta, \mathbf{W}, \Pi)] - \mathrm{E}_{\mathcal{M}_y, \mathbf{Y}} [\ln q (\mu, \mathbf{B}, \mathbf{Y})] && (67) \\
&= \mathrm{E}_{\mathbf{Y}} [\ln P (\mathbf{\Phi} | \mathbf{Y}, \theta, \mathbf{W})] + \mathrm{E}_{\mathcal{M}_y, \mathbf{Y}} [\ln P (\mathbf{Y} | \mathcal{M}_y)] + \mathrm{E}_{\mathcal{M}_y} [\ln P (\mathcal{M}_y | \Pi)] && (68) \\
&\quad - \mathrm{E}_{\mathcal{M}_y} [\ln q (\mu, \mathbf{B})] - \mathrm{E}_{\mathbf{Y}} [\ln q (\mathbf{Y})]
\end{aligned}
$$

The simplifications that are done in the following equations assume that the lower bound is calculated after the $q(\mathcal{M}_y)$ re-estimation. We are going to define:

$$\overline{\phi}_i = \frac{1}{n_i}\sum_{\phi\in\mathcal{S}_i}\phi \tag{69}$$

$$\mathbf{R}_\phi = \sum_{j=1}^{N}\phi_j\phi_j^T \tag{70}$$

$$\mathbf{S}_\phi = \sum_{i=1}^{M}\sum_{\phi\in\mathcal{S}_i}\mathrm{E}_{\mathbf{Y}}\left[(\phi-\mathbf{y}_i)(\phi-\mathbf{y}_i)^T\right] \tag{71}$$

$$= \mathbf{R}_\phi + \sum_{i=1}^{M} n_i\left(\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] - \mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i]\,\overline{\phi}_i^T - \overline{\phi}_i\mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i]^T\right) \tag{72}$$

We evaluate $\mathrm{E}_{\mathbf{Y}}[\ln P(\mathbf{\Phi}|\mathbf{Y},\theta,\mathbf{W})]$:

$$\mathrm{E}_{\mathbf{Y}}[\ln P(\mathbf{\Phi}|\mathbf{Y},\theta,\mathbf{W})] = \sum_{i=1}^{M}\sum_{\phi\in\mathcal{S}_i}\mathrm{E}_{\mathbf{Y}}[\ln P(\phi|\mathbf{y}_i,W)] \tag{73}$$

$$= \frac{N}{2}\ln|\mathbf{W}| - \frac{Nd}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{M}\sum_{\phi\in\mathcal{S}_i}\mathrm{E}_{\mathbf{Y}}\left[(\phi-\mathbf{y}_i)^T\mathbf{W}(\phi-\mathbf{y}_i)\right] \tag{74}$$

$$= \frac{N}{2}\ln|\mathbf{W}| - \frac{Nd}{2}\ln(2\pi) - \frac{1}{2}\mathrm{tr}\left(\mathbf{W}\sum_{i=1}^{M}\sum_{\phi\in\mathcal{S}_i}\mathrm{E}_{\mathbf{Y}}\left[(\phi-\mathbf{y}_i)(\phi-\mathbf{y}_i)^T\right]\right) \tag{75}$$

$$= \frac{N}{2}\ln|\mathbf{W}| - \frac{Nd}{2}\ln(2\pi) - \frac{1}{2}\mathrm{tr}(\mathbf{W}\mathbf{S}_\phi) \tag{76}$$

Now, we evaluate $\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln P(\mathbf{Y}|\mathcal{M}_y)]$:

$$\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln P(\mathbf{Y}|\mathcal{M}_y)] = \sum_{i=1}^{M}\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln P(\mathbf{y}_i|\mu,\mathbf{B})] \tag{77}$$

$$= \frac{M}{2}\mathrm{E}_{\mathbf{B}}[\ln|\mathbf{B}|] - \frac{Md}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{M}\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}\left[(\mathbf{y}_i-\mu)^T\mathbf{B}(\mathbf{y}_i-\mu)\right] \tag{78}$$

$$= \frac{M}{2}\mathrm{E}_{\mathbf{B}}[\ln|\mathbf{B}|] - \frac{Md}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{M}\big(\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}\left[\mathbf{y}_i^T\mathbf{B}\mathbf{y}_i\right] \tag{79}$$
$$-2\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}\left[\mathbf{y}_i^T\mathbf{B}\mu\right] + \mathrm{E}_{\mathcal{M}_y}\left[\mu^T\mathbf{B}\mu\right]\big)$$

$$= \frac{M}{2}\mathrm{E}_{\mathbf{B}}[\ln|\mathbf{B}|] - \frac{Md}{2}\ln(2\pi) + M\overline{\mathbf{y}}^T\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}\mu] \tag{80}$$
$$-\frac{1}{2}\mathrm{tr}\left(\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}]\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] + M\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\mu^T\right]\right)$$

To evaluate 80 we need to calculate some new expectations. Using the properties of the Gaussian and Wishart distributions [3] we have

$$\ln\tilde{\mathbf{B}} \equiv \mathrm{E}_{\mathbf{B}}[\ln|\mathbf{B}|] = \sum_{i=1}^{d}\psi\left(\frac{M+1-i}{2}\right) + d\ln 2 + \ln\left|\mathbf{S}_y^{-1}\right| \tag{81}$$

$$\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\mu^T\right] = \int_{\mathbf{B}}\int_{\mu} \mathbf{B}\mu\mu^T \mathcal{N}\left(\mu|\overline{\mathbf{y}}, (M\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right)\, \mathrm{d}\mu\, \mathrm{d}\mathbf{B} \tag{82}$$

$$= \int_{\mathbf{B}} \mathbf{B} \int_{\mu} \mu\mu^T \mathcal{N}\left(\mu|\overline{\mathbf{y}}, (M\mathbf{B})^{-1}\right) \mathrm{d}\mu\, \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right)\, \mathrm{d}\mathbf{B} \tag{83}$$

$$= \int_{\mathbf{B}} \mathbf{B} \left((M\mathbf{B})^{-1} + \overline{\mathbf{y}}\overline{\mathbf{y}}^T\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right)\, \mathrm{d}\mathbf{B} \tag{84}$$

$$= M^{-1}\mathbf{I} + \int_{\mathbf{B}} \mathbf{B}\mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right)\, \mathrm{d}\mathbf{B}\, \overline{\mathbf{y}}\overline{\mathbf{y}}^T \tag{85}$$

$$= M^{-1}\mathbf{I} + M\mathbf{S}_y^{-1}\overline{\mathbf{y}}\overline{\mathbf{y}}^T \tag{86}$$

Now we can plug equations 51, 58, 62, 81 and 86 into 80

$$\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}\left[\ln P\left(\mathbf{Y}|\mathcal{M}_y\right)\right] = \frac{M}{2}\ln\tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) + M\overline{\mathbf{y}}^T M\mathbf{S}_y^{-1}\overline{\mathbf{y}} \tag{87}$$

$$- \frac{1}{2}\mathrm{tr}\left(M\mathbf{S}_y^{-1}\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] + M\left(M^{-1}\mathbf{I} + M\mathbf{S}_y^{-1}\overline{\mathbf{y}}\overline{\mathbf{y}}^T\right)\right)$$

$$= \frac{M}{2}\ln\tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) \tag{88}$$

$$- \frac{1}{2}\mathrm{tr}\left(\mathbf{I} + M\mathbf{S}_y^{-1}\left(\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] - M\overline{\mathbf{y}}\overline{\mathbf{y}}^T\right)\right)$$

$$= \frac{M}{2}\ln\tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) - \frac{1}{2}\mathrm{tr}\left(\mathbf{I} + M\mathbf{S}_y^{-1}\mathbf{S}_y\right) \tag{89}$$

$$= \frac{M}{2}\ln\tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) - \frac{1}{2}\mathrm{tr}\left((M+1)\mathbf{I}\right) \tag{90}$$

$$= \frac{M}{2}\ln\tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) - \frac{(M+1)d}{2} \tag{91}$$

Now, we evaluate $\mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathcal{M}_y|\Pi\right)\right]$

$$\mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathcal{M}_y|\Pi\right)\right] = \ln\alpha - \frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln\tilde{\mathbf{B}} \tag{92}$$

Now, we evaluate $\mathrm{E}_{\mathcal{M}_y}\left[\ln q\left(\mu, \mathbf{B}\right)\right]$

$$\mathrm{E}_{\mathcal{M}_y}\left[\ln q\left(\mu, \mathbf{B}\right)\right] = \mathrm{E}_{\mathcal{M}_y}\left[\ln\mathcal{N}\left(\mu|\overline{\mathbf{y}}, (M\mathbf{B})^{-1}\right)\right] + \mathrm{E}_{\mathbf{B}}\left[\ln\mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right)\right] \tag{93}$$

$$= \frac{d}{2}\ln\left(\frac{M}{2\pi}\right) + \frac{1}{2}\ln\tilde{\mathbf{B}} - \frac{M}{2}\mathrm{E}_{\mathcal{M}_y}\left[(\mu - \overline{\mathbf{y}})^T\mathbf{B}(\mu - \overline{\mathbf{y}})\right] - \mathrm{H}\left[q\left(\mathbf{B}\right)\right] \tag{94}$$

where $\mathrm{H}\left[q\left(\mathbf{B}\right)\right]$ is the Entropy of the Wishart distribution [3]

$$\mathrm{H}\left[q\left(\mathbf{B}\right)\right] = \mathrm{H}\left[\mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right)\right] \tag{95}$$

$$= -\ln B\left(\mathbf{S}_y^{-1}, M\right) - \frac{M-d-1}{2}\ln\tilde{\mathbf{B}} + \frac{Md}{2} \tag{96}$$

$$B(\mathbf{W}, N) = \frac{1}{2^{Nd/2}Z_{Nd}}\left|\mathbf{W}\right|^{-N/2} \tag{97}$$

and

$$\mathrm{E}_{\mathcal{M}_y}\left[(\mu - \overline{\mathbf{y}})^T\mathbf{B}(\mu - \overline{\mathbf{y}})\right] = \mathrm{E}_{\mathcal{M}_y}\left[\mu^T\mathbf{B}\mu\right] - 2\overline{\mathbf{y}}^T\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\right] + \overline{\mathbf{y}}^T\mathrm{E}_{\mathbf{B}}\left[\mathbf{B}\right]\overline{\mathbf{y}} \tag{98}$$

$$= \mathrm{tr}\left(\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\mu^T\right]\right) - \overline{\mathbf{y}}^T\mathrm{E}_{\mathbf{B}}\left[\mathbf{B}\right]\overline{\mathbf{y}} \tag{99}$$

$$= \mathrm{tr}\left(M^{-1}\mathbf{I} + M\mathbf{S}_y^{-1}\overline{\mathbf{y}}\overline{\mathbf{y}}^T\right) - \overline{\mathbf{y}}^T M\mathbf{S}_y^{-1}\overline{\mathbf{y}} \tag{100}$$

$$= dM^{-1} \tag{101}$$

If we plug 101 in 94

$$\mathrm{E}_{\mathcal{M}_y}\left[\ln q\left(\mu,\mathbf{B}\right)\right]=\frac{d}{2}\ln\left(\frac{M}{2\pi}\right)+\frac{1}{2}\ln\tilde{\mathbf{B}}-\frac{d}{2}-\mathrm{H}\left[q\left(\mathbf{B}\right)\right] \tag{102}$$

Now, we evaluate $\mathrm{E}_{\mathbf{Y}}\left[\ln q\left(\mathbf{Y}\right)\right]$

$$\mathrm{E}_{\mathbf{Y}}\left[\ln q\left(\mathbf{Y}\right)\right]=\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\ln\mathcal{N}\left(\mathbf{y}_i|\mathbf{L}_i^{-1}\gamma_i,\mathbf{L}_i^{-1}\right)\right] \tag{103}$$

$$=-\frac{Md}{2}\ln(2\pi)+\frac{1}{2}\sum_{i=1}^{M}\ln\left|\mathbf{L}_i\right|-\frac{1}{2}\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[(\mathbf{y}_i-\mathbf{L}_i^{-1}\gamma_i)^T\mathbf{L}_i(\mathbf{y}_i-\mathbf{L}_i^{-1}\gamma_i)\right] \tag{104}$$

$$=-\frac{Md}{2}\ln(2\pi)+\frac{1}{2}\sum_{i=1}^{M}\ln\left|\mathbf{L}_i\right|-\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left(\mathbf{L}\mathrm{E}_{\mathbf{Y}}\left[\left(y_i-\mathbf{L}_i^{-1}\gamma_i\right)\left(y_i-\mathbf{L}_i^{-1}\gamma_i\right)^T\right]\right) \tag{105}$$

$$=-\frac{Md}{2}\ln(2\pi)+\frac{1}{2}\sum_{i=1}^{M}\ln\left|\mathbf{L}_i\right|-\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left(\mathbf{L}\mathbf{L}^{-1}\right) \tag{106}$$

$$=-\frac{Md}{2}\ln(2\pi)-\frac{Md}{2}+\frac{1}{2}\sum_{i=1}^{M}\ln\left|\mathbf{L}_i\right| \tag{107}$$

Finally, simplifying

$$\mathcal{L}=\frac{N}{2}\ln\left|\mathbf{W}\right|-\frac{1}{2}\mathrm{tr}\left(\mathbf{W}\mathbf{S}_\phi\right)-\frac{1}{2}\sum_{i=1}^{M}\ln\left|\mathbf{L}_i\right|+\frac{M}{2}\ln\left|\mathbf{S}_y^{-1}\right| \tag{108}$$
$$-\frac{Nd}{2}\ln(2\pi)-\frac{d}{2}\ln M+\ln Z_{Md}+\frac{Md}{2}(1+\ln 2)$$

7 VB likelihood ratio with conjugate priors

7.1 Likelihood ratio approximation

The approach taken in section 6 has a great computational cost due to the fact that in each iteration we need to re-estimate the $q\left(\mathbf{y}_i\right)$ distributions for all the speakers of the development database. In this section, we are going to follow a different approximation for not having to use the development i-vectors on the calculus of the likelihood ratio. We intend to estimate the posterior distribution of the model $\mathcal{M}_y$ given $\mathbf{\Phi}_\mathrm{d}$, θ_d, $\mathbf{W}$ and the non-informative prior Π; and then, using this posterior as a prior Π_d for the calculus of the likelihood ratio. To estimate this posterior we are going to use the following approximation:

$$P\left(\mathcal{M}_y|\Pi_\mathrm{d}\right)=P\left(\mathcal{M}_y|\mathbf{\Phi}_\mathrm{d},\theta_\mathrm{d},\mathbf{W},\Pi\right)\approx q_\mathrm{d}\left(\mathcal{M}_y\right) \tag{109}$$

where $q_\mathrm{d}\left(\mathcal{M}_y\right)$ is the variational factor of the $\mathcal{M}_y$ given only the development data. This factor is Gaussian-Wishart distributed as we got in equation 57.

$$q_\mathrm{d}\left(\mathcal{M}_y\right)=\mathcal{N}\left(\mu|\overline{\mathbf{y}}_\mathrm{d},\left(\beta_\mathrm{d}\mathbf{B}\right)^{-1}\right)\mathcal{W}\left(\mathbf{B}|\mathbf{S}_{\mathrm{d}y}^{-1},\nu_\mathrm{d}\right)\quad\text{if }\nu_\mathrm{d}>d \tag{110}$$

where $\beta_\mathrm{d}=\nu_\mathrm{d}=M_\mathrm{d}$.

So with this approximation the likelihood ratios is

$$R\left(\mathbf{\Phi}_\mathrm{t},\mathbf{W},\Pi_\mathrm{d}\right)=\frac{P\left(\mathbf{\Phi}_\mathrm{t}|\mathcal{T},\mathbf{W},\Pi_\mathrm{d}\right)}{P\left(\mathbf{\Phi}_\mathrm{t}|\mathcal{N},\mathbf{W},\Pi_\mathrm{d}\right)} \tag{111}$$

$$=\frac{P\left(\mathbf{\Phi}_\mathrm{t}|\mathcal{M}_y,\mathcal{T},\mathbf{W}\right)}{P\left(\mathbf{\Phi}_\mathrm{t}|\mathcal{M}_y,\mathcal{N},\mathbf{W}\right)}\frac{P\left(\mathcal{M}_y|\mathbf{\Phi}_\mathrm{t},\mathcal{N},\mathbf{W},\Pi_\mathrm{d}\right)}{P\left(\mathcal{M}_y|\mathbf{\Phi}_\mathrm{t},\mathcal{T},\mathbf{W},\Pi_\mathrm{d}\right)} \tag{112}$$

$$=R\left(\mathbf{\Phi}_\mathrm{t},\mathcal{M}_y,\mathbf{W}\right)\frac{P\left(\mathcal{M}_y|\mathbf{\Phi}_\mathrm{t},\mathcal{N},\mathbf{W},\Pi_\mathrm{d}\right)}{P\left(\mathcal{M}_y|\mathbf{\Phi}_\mathrm{t},\mathcal{T},\mathbf{W},\Pi_\mathrm{d}\right)} \tag{113}$$

$$=R\left(\mathbf{\Phi}_\mathrm{t},\mathcal{M}_y,\mathbf{W}\right)\frac{P\left(\mathcal{M}_y,\mathbf{Y}_\mathrm{t}|\mathbf{\Phi}_\mathrm{t},\mathcal{N},\mathbf{W},\Pi_\mathrm{d}\right)}{P\left(\mathcal{M}_y,\mathbf{Y}_\mathrm{t}|\mathbf{\Phi}_\mathrm{t},\mathcal{T},\mathbf{W},\Pi_\mathrm{d}\right)}\frac{P\left(\mathbf{Y}_\mathrm{t}|\mathcal{M}_y,\mathbf{\Phi}_\mathrm{t},\mathcal{T},\mathbf{W}\right)}{P\left(\mathbf{Y}_\mathrm{t}|\mathcal{M}_y,\mathbf{\Phi}_\mathrm{t},\mathcal{N},\mathbf{W}\right)} \tag{114}$$

Now we make $\mathbf{y} = \mathbf{y}_0$ for all speakers in $\mathbf{\Phi}_{\mathrm{t}}$, what we denote as $\mathbf{Y}_{\mathrm{t}} = \mathbf{Y}_{\mathrm{t}0}$. We take equation 28 and the plug-in likelihood in equation 11 and substitute them in 114. Finally simplifying we get:

$$R\left(\mathbf{\Phi}_{\mathrm{t}}, \mathbf{W}, \Pi_{\mathrm{d}}\right) = \frac{1}{P\left(\mathbf{y}_0 | \mathcal{M}_y\right)} \frac{P\left(\mathcal{M}_y, \mathbf{Y}_{\mathrm{t}0} | \mathbf{\Phi}_{\mathrm{t}}, \mathcal{N}, \mathbf{W}, \Pi_{\mathrm{d}}\right)}{P\left(\mathcal{M}_y, \mathbf{Y}_{\mathrm{t}0} | \mathbf{\Phi}_{\mathrm{t}}, \mathcal{T}, \mathbf{W}, \Pi_{\mathrm{d}}\right)} \tag{115}$$

and the Variational Bayes version:

$$R\left(\mathbf{\Phi}_{\mathrm{t}}, \mathbf{W}, \Pi_{\mathrm{d}}\right) \approx R_{VB}\left(\mathbf{\Phi}_{\mathrm{t}}, \mathbf{W}, \Pi_{\mathrm{d}}\right) = \frac{1}{P\left(\mathbf{y}_0 | \mathcal{M}_y\right)} \frac{q_{\mathcal{N}}\left(\mathcal{M}_y, \mathbf{Y}_{\mathrm{t}0}\right)}{q_{\mathcal{T}}\left(\mathcal{M}_y, \mathbf{Y}_{\mathrm{t}0}\right)} \tag{116}$$

where $q_{\mathcal{T}}\left(\mathcal{M}_y, \mathbf{Y}_{\mathrm{t}0}\right)$ and $q_{\mathcal{N}}\left(\mathcal{M}_y, \mathbf{Y}_{\mathrm{t}0}\right)$ are the variational posteriors assuming that $\theta_{\mathrm{t}} = \{\mathcal{T}\}$ and $\theta_{\mathrm{t}} = \{\mathcal{N}\}$ respectively.

7.2 Variational distributions

We are going to get the variational factors using the conjugate prior. We write again the joint distribution of the random variables:

$$P\left(\mathbf{\Phi}, \mathcal{M}_y, \mathbf{Y} | \theta, \mathbf{W}, \Pi_{\mathrm{d}}\right) = P\left(\mathbf{\Phi} | \mathbf{Y}, \theta, \mathbf{W}\right) P\left(\mathbf{Y} | \mathcal{M}_y\right) P\left(\mathcal{M}_y | \Pi_{\mathrm{d}}\right) \tag{117}$$

Now, we consider the variational distribution:

$$q\left(\mathcal{M}_y, \mathbf{Y}\right) = q\left(\mathcal{M}_y\right) q\left(\mathbf{Y}\right) \tag{118}$$

The optimum for the factor $q\left(\mathbf{Y}\right)$ is the same as in the previous section

$$q^*\left(\mathbf{Y}\right) = \prod_{i=1}^{M} q^*\left(\mathbf{y}_i\right) \tag{119}$$

$$q^*\left(\mathbf{y}_i\right) = \mathcal{N}\left(\mathbf{y}_i | \mathbf{L}_i^{-1}\gamma_i, \mathbf{L}_i^{-1}\right) \tag{120}$$

$$\mathbf{L}_i = \mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\right] + n_i \mathbf{W} \tag{121}$$

$$\gamma_i = \mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\right] + \mathbf{W} \sum_{\phi \in \mathcal{S}_i} \phi \tag{122}$$

The optimum for the factor $q\left(\mathcal{M}_y\right)$ is given by

$$\ln q^*\left(\mathcal{M}_y\right) = \mathrm{E}_{\mathbf{Y}}\left[\ln P\left(\mathbf{\Phi}, \mathcal{M}_y, \mathbf{Y} | \theta, \mathbf{W}, \Pi_{\mathrm{d}}\right)\right] + \mathrm{const} \tag{123}$$

Using decomposition 117 we have:

$$\ln q^*\left(\mathcal{M}_y\right) = \mathrm{E}_{\mathbf{Y}}\left[\ln P\left(\mathbf{Y} | \mathcal{M}_y\right)\right] + \mathrm{E}_{\mathbf{Y}}\left[\ln P\left(\mathcal{M}_y | \Pi_{\mathrm{d}}\right)\right] + \mathrm{const} \tag{124}$$

Substituting equations 2, and 110 into 124 and absorbing any term that does not depend on μ or $\mathbf{B}$ into the additive constant we have

$$\ln q^*\left(\mathcal{M}_y\right) =$$

$$= \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}}\left[\ln P\left(\mathbf{y}_i | \mathcal{M}_y\right)\right] + \mathrm{E}_{\mathbf{Y}}\left[\ln P\left(\mu, \mathbf{B} | \Pi_{\mathrm{d}}\right)\right] + \mathrm{const} \tag{125}$$

$$= \frac{M}{2} \ln |\mathbf{B}| - \frac{1}{2} \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}}\left[(\mathbf{y}_i - \mu)^T \mathbf{B} (\mathbf{y}_i - \mu)\right] \tag{126}$$

$$+ \frac{1}{2} \ln |\mathbf{B}| - \frac{\beta_{\mathrm{d}}}{2} (\mu - \overline{\mathbf{y}}_{\mathrm{d}})^T \mathbf{B} (\mu - \overline{\mathbf{y}}_{\mathrm{d}}) + \frac{\nu_{\mathrm{d}} - d - 1}{2} \ln |\mathbf{B}| - \frac{1}{2} \mathrm{tr}\left(\mathbf{B}\mathbf{S}_{\mathrm{d}y}\right) + \mathrm{const} \tag{127}$$

Now, we define

Now we define:

$$\overline{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\right] \tag{128}$$

$$\mathbf{S}_y = \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}}\left[\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)^T\right] = \sum_{i=1}^{M} \mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i \mathbf{y}_i^T\right] - M\overline{\mathbf{y}}\overline{\mathbf{y}}^T \tag{129}$$

$$\beta' = \beta_{\mathrm{d}} + M \tag{130}$$

$$\nu' = \nu_{\mathrm{d}} + M \tag{131}$$

$$\overline{\mathbf{y}}' = \frac{1}{\beta'}\left(\beta_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}} + M\overline{\mathbf{y}}\right) \tag{132}$$

$$\mathbf{S}'_y = \mathbf{S}_{\mathrm{d}y} + \mathbf{S}_y + \frac{\beta_{\mathrm{d}} M}{\beta'}\left(\overline{\mathbf{y}} - \overline{\mathbf{y}}_{\mathrm{d}}\right)\left(\overline{\mathbf{y}} - \overline{\mathbf{y}}_{\mathrm{d}}\right)^T \tag{133}$$

Now we can write 49 as

$$\begin{aligned}
&\ln q^*\left(\mathcal{M}_y\right) = \\
&= \frac{M}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)^T\right] + M\left(\mu - \overline{\mathbf{y}}\right)\left(\mu - \overline{\mathbf{y}}\right)^T\right)\right) && (134)\\
&\quad + \frac{1}{2}\ln|\mathbf{B}| - \frac{\beta_{\mathrm{d}}}{2}(\mu - \overline{\mathbf{y}}_{\mathrm{d}})^T\mathbf{B}(\mu - \overline{\mathbf{y}}_{\mathrm{d}}) + \frac{\nu_{\mathrm{d}} - d - 1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\mathbf{S}_{\mathrm{d}y}\right) + \mathrm{const}\\
&= -\frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(M\left(\mu - \overline{\mathbf{y}}\right)\left(\mu - \overline{\mathbf{y}}\right)^T + \beta_{\mathrm{d}}\left(\mu - \overline{\mathbf{y}}_{\mathrm{d}}\right)\left(\mu - \overline{\mathbf{y}}_{\mathrm{d}}\right)^T\right)\right) && (135)\\
&\quad + \frac{1}{2}\ln|\mathbf{B}| + \frac{\nu_{\mathrm{d}} + M - d - 1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(\mathbf{S}_{\mathrm{d}y} + \mathbf{S}_y\right)\right) + \mathrm{const}\\
&= -\frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(M\left(\mu\mu^T - \mu\overline{\mathbf{y}}^T - \overline{\mathbf{y}}\mu^T + \overline{\mathbf{y}}\overline{\mathbf{y}}^T\right) + \beta_{\mathrm{d}}\left(\mu\mu^T - \mu\overline{\mathbf{y}}_{\mathrm{d}}^T - \overline{\mathbf{y}}_{\mathrm{d}}\mu^T + \overline{\mathbf{y}}_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}^T\right)\right)\right) && (136)\\
&\quad + \frac{1}{2}\ln|\mathbf{B}| + \frac{\nu' - d - 1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(\mathbf{S}_{\mathrm{d}y} + \mathbf{S}_y\right)\right) + \mathrm{const}\\
&= -\frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(\left(M + \beta_{\mathrm{d}}\right)\left(\mu\mu^T - \frac{1}{M + \beta_{\mathrm{d}}}\mu\left(M\overline{\mathbf{y}} + \beta_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}\right)^T - \frac{1}{M + \beta_{\mathrm{d}}}\left(M\overline{\mathbf{y}} + \beta_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}\right)\mu^T\right.\right.\right.\\
&\quad \left.\left.\left. + \frac{1}{M + \beta_{\mathrm{d}}}\left(M\overline{\mathbf{y}}\overline{\mathbf{y}}^T + \beta_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}^T\right)\right)\right)\right) && (137)\\
&\quad + \frac{1}{2}\ln|\mathbf{B}| + \frac{\nu' - d - 1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(\mathbf{S}_{\mathrm{d}y} + \mathbf{S}_y\right)\right) + \mathrm{const}\\
&= \frac{1}{2}\ln|\mathbf{B}| - \frac{\beta'}{2}(\mu - \overline{\mathbf{y}}')^T\mathbf{B}(\mu - \overline{\mathbf{y}}') && (138)\\
&\quad - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(M\overline{\mathbf{y}}\overline{\mathbf{y}}^T + \beta_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}^T - \frac{1}{M + \beta_{\mathrm{d}}}\left(M\overline{\mathbf{y}} + \beta_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}\right)\left(M\overline{\mathbf{y}} + \beta_{\mathrm{d}}\overline{\mathbf{y}}_{\mathrm{d}}\right)^T\right)\right)\\
&\quad + \frac{\nu' - d - 1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(\mathbf{S}_{\mathrm{d}y} + \mathbf{S}_y\right)\right) + \mathrm{const}\\
&= \frac{1}{2}\ln|\mathbf{B}| - \frac{\beta'}{2}(\mu - \overline{\mathbf{y}}')^T\mathbf{B}(\mu - \overline{\mathbf{y}}') && (139)\\
&\quad + \frac{\nu' - d - 1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\left(\mathbf{S}_{\mathrm{d}y} + \mathbf{S}_y + \frac{\beta_{\mathrm{d}} M}{M + \beta_{\mathrm{d}}}\left(\overline{\mathbf{y}} - \overline{\mathbf{y}}_{\mathrm{d}}\right)\left(\overline{\mathbf{y}} - \overline{\mathbf{y}}_{\mathrm{d}}\right)^T\right)\right) + \mathrm{const}\\
&= \left[\frac{1}{2}\ln|\mathbf{B}| - \frac{\beta'}{2}(\mu - \overline{\mathbf{y}}')^T\mathbf{B}(\mu - \overline{\mathbf{y}}')\right] + \left[\frac{\nu' - d - 1}{2}\ln|\mathbf{B}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{B}\mathbf{S}'_y\right)\right] + \mathrm{const} && (140)
\end{aligned}$$

see that $q^*\left(\mathcal{M}_y\right)$ Gaussian-Wishart distributed:

$$q^*\left(\mathcal{M}_y\right) = \mathcal{N}\left(\mu|\overline{\mathbf{y}}', \left(\beta'\mathbf{B}\right)^{-1}\right)\mathcal{W}\left(\mathbf{B}|\mathbf{S}'^{-1}_y, \nu'\right) \quad \text{if } \nu' > d \tag{141}$$

In order to estimate the parameters of the factoring distributions, we need to evaluate the expectations $\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}]$, $\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}\mu]$, $\mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i]$ and $\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right]$. Using the properties of the Gaussian and Wishart distributions [3] we have

$$\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}] = \nu'\mathbf{S}_y'^{-1} \tag{142}$$

$$\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}\mu] = \nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}' \tag{143}$$

$$\mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i] = \mathbf{L}_i^{-1}\gamma_i \tag{144}$$

$$\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] = \mathbf{L}_i^{-1} + \mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i]\,\mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i]^T \tag{145}$$

7.3 Variational lower bound

In this section, we are going to evaluate Ade variational lower bound to use to check the convergence of our algorithm.

$$\begin{aligned}
\mathcal{L} &= \int_{\mathbf{Y}}\int_{\mathbf{B}}\int_{\mu} q(\mu,\mathbf{B},\mathbf{Y}) \ln\left(\frac{P(\mathbf{\Phi},\mu,\mathbf{B},\mathbf{Y}|\theta,\mathbf{W},\Pi)}{q(\mu,\mathbf{B},\mathbf{Y})}\right) \mathrm{d}\mu\,\mathrm{d}\mathbf{B}\,\mathrm{d}\mathbf{Y} && (146)\\
&= \mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln P(\mathbf{\Phi},\mu,\mathbf{B},\mathbf{Y}|\theta,\mathbf{W},\Pi)] - \mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln q(\mu,\mathbf{B},\mathbf{Y})] && (147)\\
&= \mathrm{E}_{\mathbf{Y}}[\ln P(\mathbf{\Phi}|\mathbf{Y},\theta,\mathbf{W})] + \mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln P(\mathbf{Y}|\mathcal{M}_y)] + \mathrm{E}_{\mathcal{M}_y}[\ln P(\mathcal{M}_y|\Pi)] && (148)\\
&\quad - \mathrm{E}_{\mathcal{M}_y}[\ln q(\mu,\mathbf{B})] - \mathrm{E}_{\mathbf{Y}}[\ln q(\mathbf{Y})]
\end{aligned}$$

We are going to define:

$$\begin{aligned}
\overline{\phi}_i &= \frac{1}{n_i}\sum_{\phi\in\mathcal{S}_i}\phi && (149)\\
\mathbf{R}_\phi &= \sum_{j=1}^{N}\phi_j\phi_j^T && (150)\\
\mathbf{S}_\phi &= \sum_{i=1}^{M}\sum_{\phi\in\mathcal{S}_i}\mathrm{E}_{\mathbf{Y}}\left[(\phi-\mathbf{y}_i)(\phi-\mathbf{y}_i)^T\right] && (151)\\
&= \mathbf{R}_\phi + \sum_{i=1}^{M} n_i\left(\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] - \mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i]\,\overline{\phi}_i^T - \overline{\phi}_i\mathrm{E}_{\mathbf{Y}}[\mathbf{y}_i]^T\right) && (152)
\end{aligned}$$

We evaluate $\mathrm{E}_{\mathbf{Y}}[\ln P(\mathbf{\Phi}|\mathbf{Y},\theta,\mathbf{W})]$:

$$\mathrm{E}_{\mathbf{Y}}[\ln P(\mathbf{\Phi}|\mathbf{Y},\theta,\mathbf{W})] = \frac{N}{2}\ln|\mathbf{W}| - \frac{Nd}{2}\ln(2\pi) - \frac{1}{2}\mathrm{tr}(\mathbf{W}\mathbf{S}_\phi) \tag{153}$$

Now, we evaluate $\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln P(\mathbf{Y}|\mathcal{M}_y)]$:

$$\begin{aligned}
\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}[\ln P(\mathbf{Y}|\mathcal{M}_y)] &= \frac{M}{2}\mathrm{E}_{\mathbf{B}}[\ln|\mathbf{B}|] - \frac{Md}{2}\ln(2\pi) + M\overline{\mathbf{y}}^T\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}\mu] && (154)\\
&\quad - \frac{1}{2}\mathrm{tr}\left(\mathrm{E}_{\mathcal{M}_y}[\mathbf{B}]\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] + M\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\mu^T\right]\right)
\end{aligned}$$

To evaluate 154 we need to calculate some new expectations. Using the properties of the Gaussian and Wishart distributions [3] we have

$$\ln \tilde{\mathbf{B}} \equiv \mathrm{E}_{\mathbf{B}}\left[\ln |\mathbf{B}|\right] = \sum_{i=1}^{d} \psi\left(\frac{\nu' + 1 - i}{2}\right) + d \ln 2 + \ln \left|\mathbf{S}_y'^{-1}\right| \tag{155}$$

$$\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\mu^T\right] = \int_{\mathbf{B}} \int_{\mu} \mathbf{B}\mu\mu^T \mathcal{N}\left(\mu|\overline{\mathbf{y}}', (\beta'\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu'\right) \mathrm{d}\mu\, \mathrm{d}\mathbf{B} \tag{156}$$

$$= \int_{\mathbf{B}} \mathbf{B} \int_{\mu} \mu\mu^T \mathcal{N}\left(\mu|\overline{\mathbf{y}}', (\beta'\mathbf{B})^{-1}\right) \mathrm{d}\mu\, \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu'\right) \mathrm{d}\mathbf{B} \tag{157}$$

$$= \int_{\mathbf{B}} \mathbf{B}\left((\beta'\mathbf{B})^{-1} + \overline{\mathbf{y}}'\overline{\mathbf{y}}'^T\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu'\right) \mathrm{d}\mathbf{B} \tag{158}$$

$$= \beta'^{-1}\mathbf{I} + \int_{\mathbf{B}} \mathbf{B}\mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu'\right) \mathrm{d}\mathbf{B}\, \overline{\mathbf{y}}'\overline{\mathbf{y}}'^T \tag{159}$$

$$= \beta'^{-1}\mathbf{I} + \nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}'\overline{\mathbf{y}}'^T \tag{160}$$

Now we can plug equations 129, 142, 143, 155 and 160 into 154

$$\mathrm{E}_{\mathcal{M}_y,\mathbf{Y}}\left[\ln P\left(\mathbf{Y}|\mathcal{M}_y\right)\right] = \frac{M}{2}\ln \tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) + M\overline{\mathbf{y}}^T\nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}' - \frac{1}{2}\mathrm{tr}\left(\nu'\mathbf{S}_y'^{-1}\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] + M\left(\beta'^{-1}\mathbf{I} + \nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}'\overline{\mathbf{y}}'^T\right)\right) \tag{161}$$

$$= \frac{M}{2}\ln \tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) + M\overline{\mathbf{y}}^T\nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}' - \frac{1}{2}\mathrm{tr}\left(\frac{M}{\beta'}\mathbf{I} + \nu'\mathbf{S}_y'^{-1}\left(\sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\mathbf{y}_i\mathbf{y}_i^T\right] + M\overline{\mathbf{y}}'\overline{\mathbf{y}}'^T\right)\right) \tag{162}$$

$$= \frac{M}{2}\ln \tilde{\mathbf{B}} - \frac{Md}{2}\ln(2\pi) - \frac{Md}{2\beta'} - \frac{M\nu'}{2}(\overline{\mathbf{y}}' - \overline{\mathbf{y}})^T\mathbf{S}_y'^{-1}(\overline{\mathbf{y}}' - \overline{\mathbf{y}}) - \frac{\nu'}{2}\mathrm{tr}\left(\mathbf{S}_y'^{-1}\mathbf{S}_y\right) \tag{163}$$

Now, we evaluate $\mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathcal{M}_y|\Pi_{\mathrm{d}}\right)\right]$

$$\mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathcal{M}_y|\Pi_{\mathrm{d}}\right)\right] = \mathrm{E}_{\mathcal{M}_y}\left[\ln \mathcal{N}\left(\mu|\overline{\mathbf{y}}_{\mathrm{d}}, (\beta_{\mathrm{d}}\mathbf{B})^{-1}\right)\right] + \mathrm{E}_{\mathbf{B}}\left[\ln \mathcal{W}\left(\mathbf{B}|\mathbf{S}_{\mathrm{d}y}^{-1}, \nu_{\mathrm{d}}\right)\right] \tag{164}$$

$$= \frac{d}{2}\ln\left(\frac{\beta_{\mathrm{d}}}{2\pi}\right) + \frac{1}{2}\ln \tilde{\mathbf{B}} - \frac{\beta_{\mathrm{d}}}{2}\mathrm{E}_{\mathcal{M}_y}\left[(\mu - \overline{\mathbf{y}}_{\mathrm{d}})^T\mathbf{B}(\mu - \overline{\mathbf{y}}_{\mathrm{d}})\right] + \ln B\left(\mathbf{S}_{\mathrm{d}y}^{-1}, \nu_{\mathrm{d}}\right) + \frac{\nu_{\mathrm{d}} - d - 1}{2}\ln \tilde{\mathbf{B}} - \frac{1}{2}\mathrm{tr}\left(\mathbf{S}_{\mathrm{d}y}\mathrm{E}_{\mathcal{M}_y}\left[B\right]\right) \tag{165}$$

where

$$\mathrm{E}_{\mathcal{M}_y}\left[(\mu - \overline{\mathbf{y}}_{\mathrm{d}})^T\mathbf{B}(\mu - \overline{\mathbf{y}}_{\mathrm{d}})\right] = \mathrm{E}_{\mathcal{M}_y}\left[\mu^T\mathbf{B}\mu\right] - 2\overline{\mathbf{y}}_{\mathrm{d}}^T\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\right] + \overline{\mathbf{y}}_{\mathrm{d}}^T\mathrm{E}_{\mathbf{B}}\left[\mathbf{B}\right]\overline{\mathbf{y}}_{\mathrm{d}} \tag{166}$$

$$= \mathrm{tr}\left(\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\mu^T\right]\right) - 2\overline{\mathbf{y}}_{\mathrm{d}}^T\nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}' + \overline{\mathbf{y}}_{\mathrm{d}}^T\nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}_{\mathrm{d}} \tag{167}$$

$$= \mathrm{tr}\left(\beta'^{-1}\mathbf{I} + \nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}'\overline{\mathbf{y}}'^T\right) - 2\overline{\mathbf{y}}_{\mathrm{d}}^T\nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}' + \overline{\mathbf{y}}_{\mathrm{d}}^T\nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}_{\mathrm{d}} \tag{168}$$

$$= \frac{d}{\beta'} + \nu'(\overline{\mathbf{y}}' - \overline{\mathbf{y}}_{\mathrm{d}})^T\mathbf{S}_y'^{-1}(\overline{\mathbf{y}}' - \overline{\mathbf{y}}_{\mathrm{d}}) \tag{169}$$

If we plug 169 in 165

$$\mathrm{E}_{\mathcal{M}_y}\left[\ln P\left(\mathcal{M}_y|\Pi_{\mathrm{d}}\right)\right] = \frac{d}{2}\ln\left(\frac{\beta_{\mathrm{d}}}{2\pi}\right) + \frac{1}{2}\ln \tilde{\mathbf{B}} - \frac{d\beta_{\mathrm{d}}}{2\beta'} - \frac{\beta_{\mathrm{d}}\nu'}{2}(\overline{\mathbf{y}}' - \overline{\mathbf{y}}_{\mathrm{d}})^T\mathbf{S}_y'^{-1}(\overline{\mathbf{y}}' - \overline{\mathbf{y}}_{\mathrm{d}}) + \ln B\left(\mathbf{S}_{\mathrm{d}y}^{-1}, \nu_{\mathrm{d}}\right) + \frac{\nu_{\mathrm{d}} - d - 1}{2}\ln \tilde{\mathbf{B}} - \frac{\nu'}{2}\mathrm{tr}\left(\mathbf{S}_{\mathrm{d}y}\mathbf{S}_y'^{-1}\right) \tag{170}$$

Now, we evaluate $\mathrm{E}_{\mathcal{M}_y}\left[\ln q\left(\mu, \mathbf{B}\right)\right]$

$$\begin{aligned}
\mathrm{E}_{\mathcal{M}_y}\left[\ln q\left(\mu, \mathbf{B}\right)\right] &= \mathrm{E}_{\mathcal{M}_y}\left[\ln \mathcal{N}\left(\mu|\overline{\mathbf{y}}', \left(\beta'\mathbf{B}\right)^{-1}\right)\right] + \mathrm{E}_{\mathbf{B}}\left[\ln \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu'\right)\right] && (171)\\
&= \frac{d}{2}\ln\left(\frac{\beta'}{2\pi}\right) + \frac{1}{2}\ln\tilde{\mathbf{B}} - \frac{\beta'}{2}\mathrm{E}_{\mathcal{M}_y}\left[(\mu-\overline{\mathbf{y}}')^T\mathbf{B}(\mu-\overline{\mathbf{y}}')\right] - \mathrm{H}\left[q\left(\mathbf{B}\right)\right] && (172)
\end{aligned}$$

where $\mathrm{H}\left[q\left(\mathbf{B}\right)\right]$ is the Entropy of the Wishart distribution [3]

$$\begin{aligned}
\mathrm{H}\left[q\left(\mathbf{B}\right)\right] &= \mathrm{H}\left[\mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu'\right)\right] && (173)\\
&= -\ln B\left(\mathbf{S}_y'^{-1}, \nu'\right) - \frac{\nu'-d-1}{2}\ln\tilde{\mathbf{B}} + \frac{\nu' d}{2} && (174)\\
B(\mathbf{W}, N) &= \frac{1}{2^{Nd/2} Z_{Nd}}\left|\mathbf{W}\right|^{-N/2} && (175)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}_{\mathcal{M}_y}\left[(\mu-\overline{\mathbf{y}}')^T\mathbf{B}(\mu-\overline{\mathbf{y}}')\right] &= \mathrm{E}_{\mathcal{M}_y}\left[\mu^T\mathbf{B}\mu\right] - 2\overline{\mathbf{y}}'^T\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\right] + \overline{\mathbf{y}}'^T\mathrm{E}_{\mathbf{B}}\left[\mathbf{B}\right]\overline{\mathbf{y}}' && (176)\\
&= \mathrm{tr}\left(\mathrm{E}_{\mathcal{M}_y}\left[\mathbf{B}\mu\mu^T\right]\right) - \overline{\mathbf{y}}'^T\mathrm{E}_{\mathbf{B}}\left[\mathbf{B}\right]\overline{\mathbf{y}}' && (177)\\
&= \mathrm{tr}\left(\beta'^{-1}\mathbf{I} + \nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}'\overline{\mathbf{y}}'^T\right) - \overline{\mathbf{y}}'^T\nu'\mathbf{S}_y'^{-1}\overline{\mathbf{y}}' && (178)\\
&= d\beta'^{-1} && (179)
\end{aligned}$$

If we plug 179 in 172

$$\mathrm{E}_{\mathcal{M}_y}\left[\ln q\left(\mu, \mathbf{B}\right)\right] = \frac{d}{2}\ln\left(\frac{\beta'}{2\pi}\right) + \frac{1}{2}\ln\tilde{\mathbf{B}} - \frac{d}{2} - \mathrm{H}\left[q\left(\mathbf{B}\right)\right] \quad (180)$$

Now, we evaluate $\mathrm{E}_{\mathbf{Y}}\left[\ln q\left(\mathbf{Y}\right)\right]$

$$\begin{aligned}
\mathrm{E}_{\mathbf{Y}}\left[\ln q\left(\mathbf{Y}\right)\right] &= \sum_{i=1}^{M}\mathrm{E}_{\mathbf{Y}}\left[\ln \mathcal{N}\left(\mathbf{y}_i|\mathbf{L}_i^{-1}\gamma_i, \mathbf{L}_i^{-1}\right)\right] && (181)\\
&= -\frac{Md}{2}\ln(2\pi) - \frac{Md}{2} + \frac{1}{2}\sum_{i=1}^{M}\ln\left|\mathbf{L}_i\right| && (182)
\end{aligned}$$

Finally, simplifying

$$\begin{aligned}
\mathcal{L} = &\frac{N}{2}\ln\left|\mathbf{W}\right| - \frac{1}{2}\mathrm{tr}\left(\mathbf{W}\mathbf{S}_\phi\right) - \frac{1}{2}\sum_{i=1}^{M}\ln\left|\mathbf{L}_i\right| - \frac{\nu'}{2}\mathrm{tr}\left(\mathbf{S}_y'^{-1}(\mathbf{S}_y + \mathbf{S}_{\mathrm{d}y})\right) && (183)\\
&- \frac{M\nu'}{2}(\overline{\mathbf{y}}' - \overline{\mathbf{y}})^T\mathbf{S}_y'^{-1}(\overline{\mathbf{y}}' - \overline{\mathbf{y}}) - \frac{\beta_{\mathrm{d}}\nu'}{2}(\overline{\mathbf{y}}' - \overline{\mathbf{y}}_{\mathrm{d}})^T\mathbf{S}_y'^{-1}(\overline{\mathbf{y}}' - \overline{\mathbf{y}}_{\mathrm{d}})\\
&- \frac{\nu_{\mathrm{d}}}{2}\ln\left|\mathbf{S}_{\mathrm{d}y}^{-1}\right| + \frac{\nu'}{2}\ln\left|\mathbf{S}_y'^{-1}\right|\\
&+ \frac{d}{2}\ln\left(\frac{\beta_{\mathrm{d}}}{\beta'}\right) - \frac{Nd}{2}\ln(2\pi) - \ln Z_{\nu_{\mathrm{d}}d} + \ln Z_{\nu' d} + \frac{Md}{2}(1+\ln 2) + \frac{\nu' d}{2}
\end{aligned}$$

A Inferring a Gaussian distribution

In [4], we can find a detailed explanation of how to estimate a Gaussian distribution using non informative priors. Here, we are going to write some those equations using Precision matrices instead of Covariance matrices. A Gaussian distribution is defined by:

$$P\left(\mathbf{x}|\mathbf{m}, \mathbf{\Lambda}\right) = \mathcal{N}\left(\mathbf{x}|\mathbf{m}, \mathbf{\Lambda}^{-1}\right) = \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{\Lambda}(\mathbf{x}-\mathbf{m})\right) \quad (184)$$

We assume a non informative prior Π for $\mathbf{m}$ and $\mathbf{\Lambda}$ (Jeffrey's Prior):

$$P(\mathbf{m}, \boldsymbol{\Lambda}|\Pi) = P(\mathbf{m}|\boldsymbol{\Lambda}, \Pi)\, P(\boldsymbol{\Lambda}|\Pi) \tag{185}$$

$$= \lim_{k\to 0} \mathcal{N}\left(\mathbf{m}|\mathbf{m}_0, (k\boldsymbol{\Lambda})^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_0/k, k) \tag{186}$$

$$= \alpha \left|\frac{\boldsymbol{\Lambda}}{2\pi}\right|^{1/2} |\boldsymbol{\Lambda}|^{-(d+1)/2} \tag{187}$$

Now, we get the posteriors for $\mathbf{m}$ and $\boldsymbol{\Lambda}$, given a set observations $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$

$$P(\mathbf{m}, \boldsymbol{\Lambda}|\mathbf{X}, \Pi) = \frac{P(\mathbf{m}, \boldsymbol{\Lambda}|\Pi)}{P(\mathbf{X}|\Pi)} \left|\frac{\boldsymbol{\Lambda}}{2\pi}\right|^{N/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})^T\boldsymbol{\Lambda}(\mathbf{x}_i - \mathbf{m})\right) \tag{188}$$

$$= \frac{P(\mathbf{m}, \boldsymbol{\Lambda}|\Pi)}{P(\mathbf{X}|\Pi)} \left|\frac{\boldsymbol{\Lambda}}{2\pi}\right|^{N/2} \exp\left(-\frac{N}{2}(\mathbf{m} - \overline{\mathbf{x}})^T\boldsymbol{\Lambda}(\mathbf{m} - \overline{\mathbf{x}})\right) \exp\left(-\frac{1}{2}\mathrm{tr}\,(\mathbf{S}\boldsymbol{\Lambda})\right) \tag{189}$$

where

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{190}$$

$$\mathbf{S} = \sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \tag{191}$$

The marginal posterior for the precision $\boldsymbol{\Lambda}$:

$$P(\boldsymbol{\Lambda}|\mathbf{X}, \Pi) = \int_{\mathbf{m}} P(\mathbf{m}, \boldsymbol{\Lambda}|\mathbf{X}, \Pi)\ \mathrm{d}\mathbf{m} \tag{192}$$

$$= \frac{1}{P(\mathbf{X}|\Pi)} \frac{\alpha}{|\boldsymbol{\Lambda}|^{(d+1)/2}} \frac{1}{N^{d/2}} \left|\frac{\boldsymbol{\Lambda}}{2\pi}\right|^{N/2} \exp\left(-\frac{1}{2}\mathrm{tr}\,(\mathbf{S}\boldsymbol{\Lambda})\right) \int_{\mathbf{m}} \left|\frac{N\boldsymbol{\Lambda}}{2\pi}\right|^{1/2} \exp\left(-\frac{N}{2}(\mathbf{m} - \overline{\mathbf{x}})^T\boldsymbol{\Lambda}(\mathbf{m} - \overline{\mathbf{x}})\right)\ \mathrm{d}\mathbf{m} \tag{193}$$

$$= \frac{1}{P(\mathbf{X}|\Pi)} \frac{\alpha}{|\boldsymbol{\Lambda}|^{(d+1)/2}} \frac{1}{N^{d/2}} \left|\frac{\boldsymbol{\Lambda}}{2\pi}\right|^{N/2} \exp\left(-\frac{1}{2}\mathrm{tr}\,(\mathbf{S}\boldsymbol{\Lambda})\right) \tag{194}$$

$$= \frac{1}{P(\mathbf{X}|\Pi)} P(\mathbf{X}|\boldsymbol{\Lambda})\, P(\boldsymbol{\Lambda}|\Pi) \tag{195}$$

Now, we use

$$\int_{\mathbf{V}\geq 0} |\mathbf{V}|^{-k-(d+1)/2} \exp\left(-\mathrm{tr}\left(\mathbf{A}\mathbf{V}^{-1}\right)\right)\ \mathrm{d}\mathbf{V} = \int_{\mathbf{W}\geq 0} |\mathbf{W}|^{k-(d+1)/2} \exp\left(-\mathrm{tr}\,(\mathbf{A}\mathbf{W})\right)\ \mathrm{d}\mathbf{W} \tag{196}$$

$$= |\mathbf{A}|^{-k}\, \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left(k - (i-1)/2\right) \tag{197}$$

$$\text{if } k > (d-1)/2 \textit{ and } |\mathbf{A}| > 0$$

to get

$$P(\mathbf{X}|\Pi) = \int_{\boldsymbol{\Lambda}} \frac{\alpha}{|\boldsymbol{\Lambda}|^{(d+1)/2}} \frac{1}{N^{d/2}} \left|\frac{\boldsymbol{\Lambda}}{2\pi}\right|^{N/2} \exp\left(-\frac{1}{2}\mathrm{tr}\,(\mathbf{S}\boldsymbol{\Lambda})\right)\ \mathrm{d}\boldsymbol{\Lambda} \tag{198}$$

$$= \begin{cases} \alpha\frac{\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((N+1-i)/2)}{N^{d/2}|\pi\mathbf{S}|^{N/2}} & N > d \\ \alpha\left(\pi\left|\mathbf{S}_0^{-1}\mathbf{S}\right|_+ |\mathbf{S}_0|\right)^{-N/2} \quad \text{(see [4])} & N \leq d \end{cases} \tag{199}$$

where $\mathbf{S}_0 = (\overline{\mathbf{x}} - \mathbf{m}_0)(\overline{\mathbf{x}} - \mathbf{m}_0)^T + \mathbf{V}_0$

We can plug 199 into 194 to get the marginal posterior of $\mathbf{\Lambda}$:

$$P(\mathbf{\Lambda}|\mathbf{X},\Pi) = \frac{1}{Z_{Nd}\left|\mathbf{\Lambda}\right|^{(d+1)/2}}\left|\frac{\mathbf{S\Lambda}}{2}\right|^{N/2}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{S\Lambda}\right)\right) \quad \text{if } N > d \tag{200}$$

$$= \mathcal{W}\left(\mathbf{\Lambda}|\mathbf{S}^{-1}, N\right) \tag{201}$$

where $Z_{Nd} = \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\left((N+1-i)/2\right)$ and plug-in 199 into 189 we have the join posterior for $\mathbf{m}$ and $\mathbf{\Lambda}$:

$$P(\mathbf{m},\mathbf{\Lambda}|\mathbf{X},\Pi) = \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{1/2}\left|\mathbf{\Lambda}\right|^{-(d+1)/2}\frac{N^{d/2}\left|\pi\mathbf{S}\right|^{N/2}}{Z_{Nd}}\left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{N/2}\exp\left(-\frac{N}{2}(\mathbf{m}-\overline{\mathbf{x}})^{T}\mathbf{\Lambda}(\mathbf{m}-\overline{\mathbf{x}})\right)\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{S\Lambda}\right)\right) \tag{202}$$

$$= \frac{1}{Z_{Nd}\left|\mathbf{\Lambda}\right|^{(d+1)/2}}\left|\frac{N}{\pi\mathbf{S}}\right|^{1/2}\left|\frac{\mathbf{S\Lambda}}{2}\right|^{(N+1)/2}\exp\left(-\frac{N}{2}(\mathbf{m}-\overline{\mathbf{x}})^{T}\mathbf{\Lambda}(\mathbf{m}-\overline{\mathbf{x}})\right)\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{S\Lambda}\right)\right) \tag{203}$$

$$= \left[\left|\frac{N\mathbf{\Lambda}}{2\pi}\right|^{1/2}\exp\left(-\frac{N}{2}(\mathbf{m}-\overline{\mathbf{x}})^{T}\mathbf{\Lambda}(\mathbf{m}-\overline{\mathbf{x}})\right)\right]\left[\frac{1}{Z_{Nd}\left|\mathbf{\Lambda}\right|^{(d+1)/2}}\left|\frac{\mathbf{S\Lambda}}{2}\right|^{N/2}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{S\Lambda}\right)\right)\right] \tag{204}$$

$$= \mathcal{N}\left(\mathbf{m}|\overline{\mathbf{x}},(N\mathbf{\Lambda})^{-1}\right)\mathcal{W}\left(\mathbf{\Lambda}|\mathbf{S}^{-1},N\right) \quad \text{if } N > d \tag{205}$$

$$\tag{206}$$

So finally, summing up:

$$P(\mathbf{m},\mathbf{\Lambda}|\mathbf{X},\Pi) = P(\mathbf{m}|\mathbf{\Lambda},\mathbf{X},\Pi)\,P(\mathbf{\Lambda}|\mathbf{X},\Pi) \tag{207}$$

$$P(\mathbf{m}|\mathbf{\Lambda},\mathbf{X},\Pi) = \mathcal{N}\left(\mathbf{m}|\overline{\mathbf{x}},(N\mathbf{\Lambda})^{-1}\right) \tag{208}$$

$$P(\mathbf{\Lambda}|\mathbf{X},\Pi) = \mathcal{W}\left(\mathbf{\Lambda}|\mathbf{S}^{-1},N\right) \tag{209}$$

$$\text{if } N > d$$

Now, we can calculate the marginal posterior for the mean $\mathbf{m}$

$$
\begin{aligned}
P(\mathbf{m}|\mathbf{X},\Pi) &= \int_{\mathbf{\Lambda}} P(\mathbf{m},\mathbf{\Lambda}|\mathbf{X},\Pi)\,\mathrm{d}\mathbf{\Lambda} && (210)\\
&= \frac{1}{P(\mathbf{X}|\Pi)} \int_{\mathbf{\Lambda}} \frac{\alpha}{|\mathbf{\Lambda}|^{(d+1)/2}} \frac{1}{N^{d/2}} \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{N/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{\Lambda})\right) && (211)\\
&\qquad \left|\frac{N\mathbf{\Lambda}}{2\pi}\right|^{1/2} \exp\left(-\frac{N}{2}(\mathbf{m}-\overline{\mathbf{x}})^T\mathbf{\Lambda}(\mathbf{m}-\overline{\mathbf{x}})\right)\mathrm{d}\mathbf{\Lambda}\\
&= \frac{1}{P(\mathbf{X}|\Pi)} \int_{\mathbf{\Lambda}} \frac{\alpha}{|\mathbf{\Lambda}|^{(d+1)/2}} \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{(N+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\left(\mathbf{S}+N(\mathbf{m}-\overline{\mathbf{x}})(\mathbf{m}-\overline{\mathbf{x}})^T\right)\mathbf{\Lambda}\right)\right)\mathrm{d}\mathbf{\Lambda}\\
& && (212)\\
&= \frac{1}{P(\mathbf{X}|\Pi)} \frac{\alpha Z_{(N+1)d}}{\pi^{(N+1)d/2}} \left|\mathbf{S}+N(\mathbf{m}-\overline{\mathbf{x}})(\mathbf{m}-\overline{\mathbf{x}})^T\right|^{-(N+1)/2} && (213)\\
&= \frac{Z_{(N+1)d}\,N^{d/2}}{Z_{Nd}\,\pi^{(N+1)d/2}} |\pi\mathbf{S}|^{N/2} \left|\mathbf{S}+N(\mathbf{m}-\overline{\mathbf{x}})(\mathbf{m}-\overline{\mathbf{x}})^T\right|^{-(N+1)/2} && (214)\\
&= \frac{\Gamma((N+1)/2)\;N^{d/2}}{\Gamma((N+1-d)/2)\,,\pi^{d/2}} |\mathbf{S}|^{-1/2} \left|\mathbf{I}+N\mathbf{S}^{-1}(\mathbf{m}-\overline{\mathbf{x}})(\mathbf{m}-\overline{\mathbf{x}})^T\right|^{-(N+1)/2} && (215)\\
&= \frac{\Gamma((N+1)/2)\;N^{d/2}}{\Gamma((N+1-d)/2)\,,\pi^{d/2}} |\mathbf{S}|^{-1/2} \left|\mathbf{I}+N\mathbf{S}^{-1}(\mathbf{m}-\overline{\mathbf{x}})(\mathbf{m}-\overline{\mathbf{x}})^T\right|^{-(N+1)/2} && (216)\\
&= \frac{\Gamma((N+1)/2)}{\Gamma((N+1-d)/2)} \left|\frac{N\mathbf{S}^{-1}}{\pi}\right|^{1/2} \left(1+N(\mathbf{m}-\overline{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{m}-\overline{\mathbf{x}})\right)^{-(N+1)/2} && (217)\\
&= \mathcal{T}_{\mathrm{M}}(\mathbf{m}|\overline{\mathbf{x}},\mathbf{S}/N,N+1) && (218)\\
&= \mathcal{T}\left(\mathbf{m}|\overline{\mathbf{x}},\mathbf{S}/N^2,N+1-d\right) && (219)\\
&\quad \text{if } N>d
\end{aligned}
$$

where we have used the matrix relation in [8]

$$|\mathbf{I}+\mathbf{BC}| = |\mathbf{I}+\mathbf{CB}| \tag{220}$$

So the mean marginal posterior is Student's T distributed. We have found several definitions of Student's T distribution in the literature.

This is the definition in [4]

$$\mathcal{T}_{\mathrm{M}}\left(\mathbf{x}|\mathbf{m},\mathbf{\Lambda}^{-1},N\right) = \frac{\Gamma(N/2)}{\Gamma((N-d)/2)} \left|\frac{\mathbf{\Lambda}}{\pi}\right|^{1/2} \left(1+(\mathbf{x}-\mathbf{m})^T\mathbf{\Lambda}(\mathbf{x}-\mathbf{m})\right)^{-N/2} \tag{221}$$

This is the definition in [3] that seems more widely spread

$$
\begin{aligned}
\mathcal{T}\left(\mathbf{x}|\mathbf{m},\mathbf{\Lambda}^{-1},N\right) &= \frac{\Gamma((N+d)/2)}{\Gamma(N/2)} \left|\frac{\mathbf{\Lambda}}{N\pi}\right|^{1/2} \left(1+\frac{1}{N}(\mathbf{x}-\mathbf{m})^T\mathbf{\Lambda}(\mathbf{x}-\mathbf{m})\right)^{-(N+d)/2} && (222)\\
&= \mathcal{T}_{\mathrm{M}}\left(\mathbf{x}|\mathbf{m},N\mathbf{\Lambda}^{-1},N+d\right) && (223)
\end{aligned}
$$

References

[1] Niko; Brummer and Edward De Villiers, "The Speaker Partitioning Problem," in *Oddyssey Speaker and Language Recognition Workshop*, 2010.

[2] Niko Brummer, "Bayesian PLDA," 2010.

[3] Christopher Bishop, *Pattern Recognition and Machine Learning*, 2006.

[4] Thomas Minka, "Inferring a Gaussian distribution," 1998.

[5] Kaare Petersen and Michael Pedersen, *The Matrix Cookbook*, Technical University of Denmark, 2008.

[6] H.V. Henderson and S.R. Searle, "On Deriving the Inverse of a Sum of Matrices," *Society for Industrial and Applied Mathematics*, vol. 23, no. 1, pp. 53–60, 1981.

[7] M. Brookes, "The Matrix Reference Manual," 2005.

[8] Thomas Minka, "Old and New Matrix Algebra Useful for Statistics," 2000.