Bayesian Factor Analysis

Carlos Sevilla-Salcedo*

February 2021

In this document I will extend and explain the bayesian formulation of two well known factor analysis models, Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA). The main objective of this document is to develop the equations and justify the complete distributions of the random variables of both these models.

One of the most used extensions of the probabilistic formulation is the bayesian approach. This does not only provide prior information that include previous information to the model, but also develops the posterior distribution of the observed data to sample from that distributions. This implies that we are not only estimating the value of each parameter but also the distribution that describes each one of them. The process to estimate the distribution of these variables is known as bayesian variational inference.

First of all, in order to work with bayesian variational inference we need to know the posterior distribution of the variables, $p(\Theta|\,\mathbf{X})$, where Θ is the group of all the variables. However, this posterior distribution is not tractable. Nevertheless, we can approximate it by a distribution $q(\Theta)$, being $q(\Theta)$ a group of treatable variables.

In order to adjust $q(\Theta)$, we need to minimise the following equation:

$$KL(q||p) \quad = \quad \int q(\Theta)\ln\left(\frac{q(\Theta)}{p(\Theta|\,\mathbf{X})}\right)d\Theta, \tag{1}$$

where KL symbolizes the Kullback-Leibler divergence, which measures the difference between two distributions. This implies that we minimise the distance between these two distributions. If we develop this equation, we have that

$$\begin{aligned} KL(q||p) &= \int q(\Theta)\ln\left(\frac{q(\Theta)}{\frac{p(\mathbf{X},\Theta)}{p(\mathbf{X})}}\right)d\Theta \\ &= \int q(\Theta)\left(\ln\left(\frac{q(\Theta)}{p(\mathbf{X},\Theta)}\right)+\ln(p(\mathbf{X}))\right)d\Theta \\ &= \int q(\Theta)\ln\left(\frac{q(\Theta)}{p(\mathbf{X},\Theta)}\right)d\Theta+\int q(\Theta)\ln(p(\mathbf{X}))d\Theta \\ &= \int q(\Theta)\ln\left(\frac{q(\Theta)}{p(\mathbf{X},\Theta)}\right)d\Theta+\ln(p(\mathbf{X})). \end{aligned} \tag{2}$$

This way, we can rearrange the previous equation to get that

$$L(q) \quad = \quad \ln(p(\mathbf{X}))-KL(q||p)\leq\ln(p(\mathbf{X})), \tag{3}$$

*Corresponding author. Email address: sevisal@tsc.uc3m.es

having that

$$L(q) = -\int q(\Theta) \ln\left(\frac{q(\Theta)}{p(\mathbf{X}, \Theta)}\right) d\Theta \tag{4}$$

is a lower bound of $\ln(p(\mathbf{X}))$. For this reason, maximising the bound implies minimising $KL(q||p)$, keeping in mind $L(q)$ will have a maximum value when $p = q$. This simple trick allows us to approximate the original variable distribution to another that can be calculated, e.g. using the mean field method.

The mean field method

The mean field method raises as an approximation to the posterior of the model variables previously explained. Specifically, it considers that one can approximate the posterior distribution of all the model variables $q(\Theta)$ by factorising over all the variables

$$q(\Theta) \;\; = \;\; \prod_i q(\Theta_i) = \prod_i q_i. \tag{5}$$

Hence, we can use the lower bound to determine the distribution $q(\Theta)$, so that it maximizes $L(q)$. To do so, let's substitute Equation (5) in (4)

$$\begin{aligned}
L(q_j) &= \int q(\Theta) \ln\left(\frac{p(\mathbf{X}, \Theta)}{q(\Theta)}\right) d\Theta = \int \prod_i q_i \left[\ln(p(\mathbf{X}, \Theta)) - \sum_i \ln(q_i)\right] d\Theta \\
&= \int \prod_i q_i \ln(p(\mathbf{X}, \Theta)) d\Theta - \int \prod_i q_i \sum_i \ln(q_i) d\Theta \\
&= \int q_j \prod_{i \neq j} q_i \ln(p(\mathbf{X}, \Theta)) d\Theta - \int q_j \prod_{i \neq j} q_i \left(\ln(q_j) + \sum_{i \neq j} \ln(q_i)\right) d\Theta \\
&= \int q_j \prod_{i \neq j} q_i \ln(p(\mathbf{X}, \Theta)) d\Theta - \int q_j \prod_{i \neq j} q_i \sum_{i \neq j} \ln(q_i) d\Theta - \int q_j \prod_{i \neq j} q_i \ln(q_j) d\Theta \\
&= \int q_j \left[\int \prod_{i \neq j} q_i \ln(p(\mathbf{X}, \Theta)) d\Theta_i\right] d\Theta_j - \int q_j \ln(q_j) d\Theta_j + \text{const} \\
&= \int q_j \ln(f_j) d\Theta_j - \int q_j \ln(q_j) d\Theta_j + \text{const}
\end{aligned} \tag{6}$$

where

$$\ln(f_j) \;\; = \;\; \mathbb{E}_{-q_j}[\ln(p(\mathbf{X}, \Theta))] + \text{const} \tag{7}$$

and $-q_j$ means that we calculate this on all the variables except the j-th variable. It can be now seen that Equation (6) is a negative KL between $q_j(\Theta_j)$ and f_j, thus to maximise $L(q_j)$ we need to minimise $KL(q_j||f_j)$. Therefore, we have that the optimum solution has the following expression:

$$\ln(q_j^*) \;\; = \;\; \mathbb{E}_{-q_j}[\ln(p(\mathbf{X}, \Theta))] + \text{const}. \tag{8}$$

This constitutes the basis of the variational inference and is the approximation we will use in the following approaches.

1 Bayesian PCA

Here we explain the bayesian extension of the PPCA (**??**) first presented in **?**. To do so, we adapt the probabilistic version by including a prior over the parameters of the model $(\mathbf{W}, \sigma^2)$. We will firstly define what is the generative model, including the variable distributions and the graphic model, to later present the variational inference result obtained.

Generative model

As we have seen before, PCA is a FA algorithm that combines the information of a set of observed data and determines a latent space. The particularity of this latent space is that its dimensions are orthogonal to one another. Keeping this in mind, we can start by defining the distribution over the observations marginalizing with bayes rules

$$p(\mathbf{X}) \sim \int p(\mathbf{X} \,|\, \mathbf{W}, \tau) p(\mathbf{W}, \tau |\, \mathbf{X}) d\mathbf{W} d\tau \tag{9}$$

where we have defined a new noise variable $\tau = \frac{1}{\sigma^2}$. This way, we can define our generative model presented in Figure 1, which includes the different variables of the model as well as the relation between them.

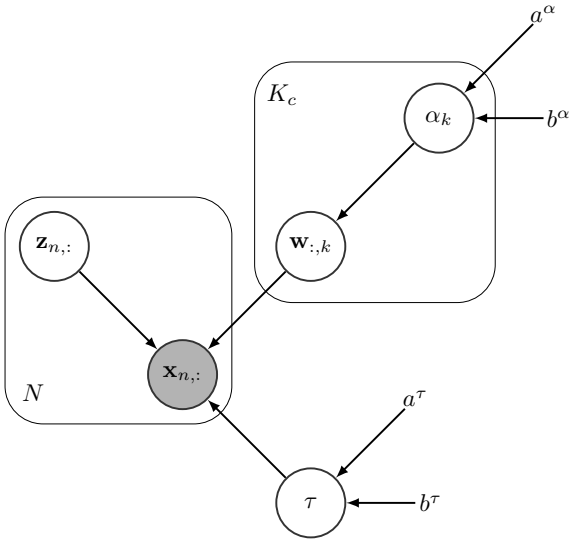

Figure 1: Plate diagram for the bayesian PCA graphical model. Gray circles denote observed variables, white circles unobserved random variables. Nodes without a circle correspond to the hyperparameters.

Furthermore, we can define the distribution of all the model variables included in the graphic model

$$\mathbf{z}_{\mathrm{n},:} \sim \mathcal{N}(0, \mathbf{I}_{\mathrm{K_c}}) \tag{10}$$

$$\mathbf{w}_{:,\mathrm{k}} \sim \mathcal{N}(0, \mathbf{I}_{\mathrm{K_c}}) \tag{11}$$

$$\mathbf{x}_{\mathrm{n},:} \,|\, \mathbf{z}_{\mathrm{n},:} \sim \mathcal{N}(\mathbf{z}_{\mathrm{n},:}\, \mathbf{W}^{\mathrm{T}}, \tau^{-1}\, \mathbf{I}_{\mathrm{D}}) \tag{12}$$

$$\tau \sim \Gamma(a^\tau, b^\tau) \tag{13}$$

where we are assuming the observed data to be independent, which implies that the random noise modeled is the same for every sample. Note that the distribution of the observed data, $\mathbf{x}_{\mathrm{n},:}$, connects the different random variables (rv).

¿Sería la distribución de $\mathbf{W}$ o de $\mathbf{w}_{\mathrm{d},:}$?

1.1 Variational inference

Making use of the variational inference definition previously stated, the problem can be defined to apply those techniques to design the model. First of all, considering Figure 1 we can extract the probability distribution of the model

$$p(\Theta|\,\mathbf{X}) \;\approx\; q(\mathbf{W})q(\tau)\prod_{n=1}^{N} q(\mathbf{z}_{\mathrm{n},:}) \tag{14}$$

This way, we can determine the optimum distribution of the different variables, applying Equation (8) and then calculate the lower bound defined in Equation (6).

In this section the different distributions specified in Equation (14) will be developed making use of the variational inference.

1.1.1 Calculation of W

$$\ln(q^*(\mathbf{W})) \;=\; \mathbb{E}_{\mathbf{Z},\tau}[\ln(p(\mathbf{X},\mathbf{W},\mathbf{Z},\tau))] = \mathbb{E}_{\mathbf{Z},\tau}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] + \mathbb{E}[\ln(p(\mathbf{W}))] + \text{const} \tag{15}$$

We can now evaluate both terms independently and then sum the results:

$$\begin{aligned}
\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau)) \;&\approx\; \sum_{\mathrm{n}=1}^{\mathrm{N}} \ln\Big(\mathcal{N}\Big(\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}},(\tau)^{-1}I\Big)\Big) + \text{const}\\
&=\; \sum_{\mathrm{n}=1}^{\mathrm{N}} \left(-\frac{1}{2}\ln\left|(\tau)^{-1}I\right| -\frac{1}{2}\big(\mathbf{x}_{\mathrm{n},:}-\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\big)\tau\big(\mathbf{x}_{\mathrm{n},:}-\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\big)^{T}\right) + \text{const}\\
&=\; -\frac{\tau}{2}\sum_{\mathrm{n}=1}^{\mathrm{N}}\big(\mathbf{x}_{\mathrm{n},:}\,\mathbf{x}_{\mathrm{n},:}^{\mathrm{T}} - 2\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} + \mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\big) + \text{const}\\
&=\; \sum_{\mathrm{n}=1}^{\mathrm{N}}\Big(\tau\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} - \frac{\tau}{2}\,\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\Big) + \text{const}\\
&=\; \sum_{\mathrm{n}=1}^{\mathrm{N}}\left(\tau\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} - \frac{\tau}{2}\sum_{\mathrm{d}=1}^{\mathrm{D}}\big(\mathbf{z}_{\mathrm{n},:}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\,\mathbf{w}_{\mathrm{d},:}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\big)\right) + \text{const}\\
&=\; \sum_{\mathrm{n}=1}^{\mathrm{N}}\left(\tau\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} - \frac{\tau}{2}\sum_{\mathrm{d}=1}^{\mathrm{D}}\big(\mathbf{w}_{\mathrm{d},:}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\,\mathbf{z}_{\mathrm{n},:}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\big)\right) + \text{const}
\end{aligned} \tag{16}$$

and applying the expectation we have that

$$\mathbb{E}_{\mathbf{Z},\tau}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] \;=\; \sum_{\mathrm{n}=1}^{\mathrm{N}}\left(\langle\tau\rangle\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\langle\mathbf{z}_{\mathrm{n},:}\rangle - \frac{\langle\tau\rangle}{2}\sum_{\mathrm{d}=1}^{\mathrm{D}}\big(\mathbf{w}_{\mathrm{d},:}\langle\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\,\mathbf{z}_{\mathrm{n},:}\rangle\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\big)\right) \tag{17}$$

Regarding the second, we have that

$$\mathbb{E}[\ln(p(\mathbf{w}_{\mathrm{d},:}))] \;=\; \ln(p(\mathbf{w}_{\mathrm{d},:})) = \sum_{\mathrm{d}=1}^{\mathrm{D}}\left(-\frac{1}{2}\,\mathbf{w}_{\mathrm{d},:}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\right) + \text{const} \tag{18}$$

Finally, joining Equations (17) and (18), we obtain the mean field approximation

$$\begin{aligned}
&\ln(q^*(\mathbf{W})) \\
&= \sum_{d=1}^{D}\left(-\frac{1}{2}\,\mathbf{w}_{d,:}\,\mathbf{w}_{d,:}^{T}\right) + \sum_{n=1}^{N}\left(\langle\tau\rangle\,\mathbf{x}_{n,:}\,\mathbf{W}\langle\mathbf{z}_{n,:}\rangle - \frac{\langle\tau\rangle}{2}\sum_{d=1}^{D}\left(\mathbf{w}_{d,:}\langle\mathbf{z}_{n,:}^{T}\,\mathbf{z}_{n,:}\rangle\,\mathbf{w}_{d,:}^{T}\right)\right) + \text{const} \\
&= \sum_{d=1}^{D}\left(-\frac{1}{2}\,\mathbf{w}_{d,:}\,\mathbf{w}_{d,:}^{T} + \sum_{n=1}^{N}\left(\langle\tau\rangle\,x_{n,d}\,\mathbf{w}_{d,:}\langle\mathbf{z}_{n,:}\rangle - \frac{\langle\tau\rangle}{2}\,\mathbf{w}_{d,:}\langle\mathbf{z}_{n,:}^{T}\,\mathbf{z}_{n,:}\rangle\,\mathbf{w}_{d,:}^{T}\right)\right) + \text{const} \\
&= \sum_{d=1}^{D}\left(-\frac{1}{2}\,\mathbf{w}_{d,:}\left(\mathbf{I} + \langle\tau\rangle\langle\mathbf{z}_{n,:}^{T}\,\mathbf{z}_{n,:}\rangle\right)\mathbf{w}_{d,:}^{T} + \langle\tau\rangle\,\mathbf{w}_{d,:}\langle\mathbf{Z}^{T}\rangle\,\mathbf{x}_{:,d}^{T}\right) + \text{const} \qquad (19)
\end{aligned}$$

This way, comparing the results with the normal distribution, we can identify terms, extracting the following conclusions:

$$q^*(\mathbf{W}) = \prod_{d=1}^{D}\left(\mathcal{N}\left(\mathbf{w}_{d,:}\,|\mu_{\mathbf{w}_{d,:}}, \Sigma_{\mathbf{W}}\right)\right) \qquad (20)$$

where the variance is

$$\Sigma_{\mathbf{W}}^{-1} = \mathbf{I} + \langle\tau\rangle\langle\mathbf{Z}^{T}\,\mathbf{Z}\rangle \qquad (21)$$

and the mean is

$$\mu_{\mathbf{w}_{d,:}} = \langle\tau\rangle\Sigma_{\mathbf{W}}\langle\mathbf{Z}\rangle^{T}\,\mathbf{x}_{:,d}^{T} \qquad (22)$$

or in a matricial way

$$\langle\mathbf{W}\rangle = \langle\tau\rangle\,\mathbf{X}^{T}\langle\mathbf{Z}\rangle\Sigma_{\mathbf{W}} \qquad (23)$$

1.1.2 Calculation of Z

$$\ln(q^*(\mathbf{Z})) = \mathbb{E}_{\mathbf{W},\tau}[\ln(p(\mathbf{X},\mathbf{W},\mathbf{Z},\tau))] = \mathbb{E}_{\mathbf{W},\tau}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] + \mathbb{E}[\ln(p(\mathbf{Z}))] \qquad (24)$$

We will again start by calculating the first term (which is similar to equation (17)

$$\mathbb{E}_{\mathbf{W},\tau}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] = \sum_{n=1}^{N}\left(\langle\tau\rangle\,\mathbf{x}_{n,:}\langle\mathbf{W}\rangle\,\mathbf{z}_{n,:}^{T} - \frac{\langle\tau\rangle}{2}\,\mathbf{z}_{n,:}\langle\mathbf{W}^{T}\,\mathbf{W}\rangle\,\mathbf{z}_{n,:}^{T}\right) + \text{const} \qquad (25)$$

As the associated prior of this variable is the same as the previous one, the result of the second term is equivalent

$$\mathbb{E}[\ln(p(\mathbf{Z}))] = \ln(p(\mathbf{Z})) = \sum_{n=1}^{N}\left(-\frac{1}{2}\,\mathbf{z}_{n,:}\,\mathbf{z}_{n,:}^{T}\right) + \text{const} \qquad (26)$$

Therefore, joining the different terms in Equations (25) and (26), the mean field approximation has the form

$$\ln(q^*(\mathbf{Z})) = \sum_{n=1}^{N}\left(-\frac{1}{2}\,\mathbf{z}_{n,:}\,\mathbf{z}_{n,:}^{T} + \langle\tau\rangle\,\mathbf{x}_{n,:}\langle\mathbf{W}\rangle\,\mathbf{z}_{n,:}^{T} - \frac{\langle\tau\rangle}{2}\,\mathbf{z}_{n,:}\langle\mathbf{W}^{T}\,\mathbf{W}\rangle\,\mathbf{z}_{n,:}^{T}\right) + \text{const} \qquad (27)$$

Again, if we identify terms from the previous equation we get that

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N} \left(\mathcal{N}\left(\mathbf{z}_{n,:} \,|\mu_{\mathbf{z}_{n,:}}, \Sigma_{\mathbf{Z}}\right)\right) \tag{28}$$

where the variance is

$$\Sigma_{\mathbf{Z}}^{-1} = I + \sum_{m=1}^{M} \left(\langle\tau\rangle \langle \mathbf{W}^{\mathrm{T}}\, \mathbf{W}\rangle\right) \tag{29}$$

and the mean is

$$\mu_{\mathbf{z}_{n,:}} = \sum_{m=1}^{M} (\langle\tau\rangle\, \mathbf{x}_{n,:} \langle \mathbf{W}\rangle \Sigma_{\mathbf{Z}}) \tag{30}$$

or in a matricial way

$$\langle \mathbf{Z}\rangle = \sum_{m=1}^{M} (\langle\tau\rangle\, \mathbf{X} \langle \mathbf{W}\rangle \Sigma_{\mathbf{Z}}) \tag{31}$$

1.1.3 Calculation of τ

Let's now calculate the approximate distribution of the noise variable, τ.

$$\ln(q^*(\tau)) = \mathbb{E}_{\mathbf{W},\mathbf{Z}}[\ln(p(\mathbf{X} \,|\, \mathbf{W}, \mathbf{Z}, \tau))] + \mathbb{E}[\ln(p(\tau))] \tag{32}$$

Similarly to what was done in equation (17) we can now calculate the first term, although the terms that will be constant for the expectation will vary

$$\begin{aligned}
\ln(p(\mathbf{X}|\, \mathbf{W}, \mathbf{Z}, \tau)) &= \sum_{n=1}^{N} \ln\left(\mathcal{N}\left(\mathbf{z}_{n,:}\, \mathbf{W}^{\mathrm{T}}, (\tau)^{-1} I\right)\right) + \text{const} \\
&= \sum_{n=1}^{N}\sum_{d=1}^{D} \left(\frac{1}{2}\ln|\tau| - \frac{\tau}{2}\left(\mathrm{x}_{n,d} - \mathbf{z}_{n,:}\, \mathbf{w}_{d,:}^{\mathrm{T}}\right)^2\right) + \text{const} \\
&= \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\sum_{n=1}^{N}\sum_{d=1}^{D}\left({\mathrm{x}_{n,d}}^2 - 2\,\mathbf{w}_{d,:}\, \mathbf{z}_{n,:}^{\mathrm{T}}\, \mathrm{x}_{n,d} + \left(\mathbf{z}_{n,:}\, \mathbf{w}_{d,:}^{\mathrm{T}}\right)^2\right) + \text{const} \\
&= \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{n=1}^{N}\sum_{d=1}^{D}{\mathrm{x}_{n,d}}^2 - 2\sum_{d=1}^{D_m}\mathbf{w}_{d,:}\, \mathbf{Z}^{\mathrm{T}}\, \mathbf{x}_{:,d} + \sum_{d=1}^{D}\mathbf{w}_{d,:}\, \mathbf{Z}^{\mathrm{T}}\, \mathbf{Z}\, \mathbf{w}_{d,:}^{\mathrm{T}}\right) + \text{const} \\
&= \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{n=1}^{N}\sum_{d=1}^{D}{\mathrm{x}_{n,d}}^2 - 2\,\mathrm{Tr}\{\mathbf{W}\, \mathbf{Z}^{\mathrm{T}}\, \mathbf{X}\} + \mathrm{Tr}\{\mathbf{W}\, \mathbf{Z}^{\mathrm{T}}\, \mathbf{Z}\, \mathbf{W}^{\mathrm{T}}\}\right) + \text{const}
\end{aligned}$$

We can now calculate the expectation with respect the rest of the random variables

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\ln(p(\mathbf{X} \,|\, \mathbf{W}, \mathbf{Z}, \tau))] \\
&= \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{n=1}^{N}\sum_{d=1}^{D}{\mathrm{x}_{n,d}}^2 - 2\,\mathrm{Tr}\{\langle\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\rangle\, \mathbf{X}\} + \mathrm{Tr}\{\langle\mathbf{W}^{\mathrm{T}}\, \mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\, \mathbf{Z}\rangle\}\right) + \text{const}
\end{aligned}$$

The second term can subsequently be determined as:

$$\mathbb{E}[\ln(p(\tau))] \;=\; \ln(p(\tau)) = -b_0^\tau \tau + (a_0^\tau - 1)\ln(\tau) + \text{const}$$

Joining them together, we have that:

$$\begin{aligned}\ln(q^*(\tau)) \;=\; & \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{d}=1}^{\mathrm{D}} {\mathrm{x}_{\mathrm{n,d}}}^2 - 2\,\mathrm{Tr}\big\{\langle\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\rangle\,\mathbf{X}\big\} + \mathrm{Tr}\big\{\langle\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}\rangle\big\}\right) \\ & - b_0^\tau\tau + (a_0^\tau - 1)\ln(\tau) + \text{const}\end{aligned}$$

So by identifying terms with the distribution, we would have that for this parameter:

$$q^*(\tau) \;=\; (Gamma(\tau|a^\tau, b^\tau)) \tag{33}$$

where the first parameter of the distribution is

$$a^\tau \;=\; \frac{DN}{2} + \alpha_0^\tau \tag{34}$$

and the second is

$$b^\tau \;=\; \beta_0^\tau + \frac{1}{2}\left(\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{d}=1}^{\mathrm{D}} {\mathrm{x}_{\mathrm{n,d}}}^2 - 2\,\mathrm{Tr}\big\{\langle\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\rangle\,\mathbf{X}\big\} + \mathrm{Tr}\big\{\langle\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}\rangle\big\}\right) \tag{35}$$

2 Bayesian PCA with ARD

2.1 Generative model

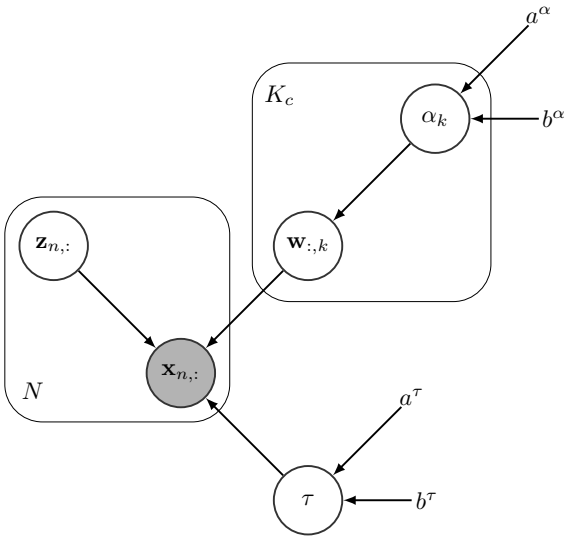

Figure 2: Plate diagram for the bayesian PCA with ARD graphical model. Gray circles denote observed variables, white circles unobserved random variables. Nodes without a circle correspond to the hyperparameters.

We can define the distribution over the observations by marginalizing using the bayes rule

$$p(\mathbf{X}) \sim \int p(\mathbf{X} \mid \mathbf{W}, \boldsymbol{\alpha}, \tau) p(\mathbf{W}, \boldsymbol{\alpha}, \tau \mid \mathbf{X}) d\,\mathbf{W}\, d\,\boldsymbol{\alpha}\, d\tau \tag{36}$$

Therefore, we can now establish the prior distributions over the different random variables in the problem

$$\begin{aligned}
\mathbf{z}_{\mathrm{n},:} &\sim \mathcal{N}(0, \mathbf{I}_{\mathrm{K_c}}) && (37)\\
\mathbf{w}_{:,\mathrm{k}} &\sim \mathcal{N}\left(0, {\alpha_k}^{-1}\, \mathbf{I}_{\mathrm{K_c}}\right) && (38)\\
\mathbf{x}_{\mathrm{n},:} \mid \mathbf{z}_{\mathrm{n},:} &\sim \mathcal{N}(\mathbf{z}_{\mathrm{n},:}\, \mathbf{W}^{\mathrm{T}}, \tau^{-1}\, \mathbf{I}_{\mathrm{D}}) && (39)\\
\alpha_k &\sim \Gamma(a^{\boldsymbol{\alpha}}, b^{\boldsymbol{\alpha}}) && (40)\\
\tau &\sim \Gamma(a^{\tau}, b^{\tau}) && (41)
\end{aligned}$$

2.2 Variational inference

Making use of the variational inference definition previously stated, the problem can be defined to apply those techniques to design the model. First of all, considering Figure **??** we can extract the probability distribution of the model

$$p(\Theta \mid \mathbf{X}) \approx \left(q(\mathbf{W}) q(\tau) \prod_{\mathrm{k}=1}^{\mathrm{K_c}} q(\alpha_k) \right) \prod_{\mathrm{n}=1}^{\mathrm{N}} q(\mathbf{z}_{\mathrm{n},:}) \tag{42}$$

This way, we can determine the optimum distribution of the different variables, applying Equation (8) and then calculate the lower bound defined in Equation (6).

In this section the different distributions specified in Equation (42) will be developed making use of the variational inference.

2.3 W

$$\begin{aligned}
\ln(q^*(\mathbf{W})) &= \mathbb{E}_{\mathbf{Z},\boldsymbol{\alpha},\tau}[\ln(p(\mathbf{X}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\alpha}, \tau))]\\
&= \mathbb{E}_{\mathbf{Z},\tau}[\ln(p(\mathbf{X} \mid \mathbf{W}, \mathbf{Z}, \tau))]\\
&\quad + \mathbb{E}_{\boldsymbol{\alpha}}[\ln(p(\mathbf{W} \mid \boldsymbol{\alpha}))] + \text{const}
\end{aligned} \tag{43}$$

We can now evaluate both terms independently and then sum the results:

$$\begin{aligned}
\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau)) &\approx \sum_{\mathrm{n}=1}^{\mathrm{N}} \ln\Big(\mathcal{N}\Big(\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}},(\tau)^{-1}I\Big)\Big) + \text{const} \\
&= \sum_{\mathrm{n}=1}^{\mathrm{N}} \left(-\frac{1}{2}\ln\left|\tau^{-1}I\right| - \frac{1}{2}\big(\mathbf{x}_{\mathrm{n},:}-\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\big)\tau\big(\mathbf{x}_{\mathrm{n},:}-\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\big)^{T}\right) + \text{const} \\
&= -\frac{\tau}{2}\sum_{\mathrm{n}=1}^{\mathrm{N}} \big(\mathbf{x}_{\mathrm{n},:}\,\mathbf{x}_{\mathrm{n},:}^{\mathrm{T}} - 2\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} + \mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\big) + \text{const} \\
&= \sum_{\mathrm{n}=1}^{\mathrm{N}} \Big(\tau\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} - \frac{\tau}{2}\,\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\Big) + \text{const} \\
&= \sum_{\mathrm{n}=1}^{\mathrm{N}} \left(\tau\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} - \frac{\tau}{2}\sum_{\mathrm{d}=1}^{\mathrm{D}}\big(\mathbf{z}_{\mathrm{n},:}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\,\mathbf{w}_{\mathrm{d},:}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\big)\right) + \text{const} \\
&= \sum_{\mathrm{n}=1}^{\mathrm{N}} \left(\tau\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} - \frac{\tau}{2}\sum_{\mathrm{d}=1}^{\mathrm{D}}\big(\mathbf{w}_{\mathrm{d},:}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\,\mathbf{z}_{\mathrm{n},:}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\big)\right) + \text{const}
\end{aligned} \tag{44}$$

This way the expectation can be calculated as:

$$\mathbb{E}_{\mathbf{Z},\tau}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] = \sum_{\mathrm{n}=1}^{\mathrm{N}} \left(\langle\tau\rangle\,\mathbf{x}_{\mathrm{n},:}\,\mathbf{W}\langle\mathbf{z}_{\mathrm{n},:}\rangle - \frac{\langle\tau\rangle}{2}\sum_{\mathrm{d}=1}^{\mathrm{D}}\big(\mathbf{w}_{\mathrm{d},:}\langle\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\,\mathbf{z}_{\mathrm{n},:}\rangle\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\big)\right) \tag{45}$$

Equivalently, the second term would be:

$$\begin{aligned}
\ln(p(\mathbf{W}\,|\,\boldsymbol{\alpha})) &= \sum_{\mathrm{d}=1}^{\mathrm{D}}\sum_{\mathrm{k}=1}^{\mathrm{K_c}} \ln(p(\mathrm{w}_{\mathrm{d,k}}\,|\,\alpha_k)) \\
&= \sum_{\mathrm{d}=1}^{\mathrm{D}}\sum_{\mathrm{k}=1}^{\mathrm{K_c}} \left(\frac{1}{2}\ln(\alpha_k) - \frac{\alpha_k}{2}{\mathrm{w}_{\mathrm{d,k}}}^2\right) + \text{const}
\end{aligned} \tag{46}$$

$$\mathbb{E}_{\boldsymbol{\alpha}}[\ln(p(\mathbf{W}\,|\,\boldsymbol{\alpha}))] = \sum_{\mathrm{d}=1}^{\mathrm{D}}\sum_{\mathrm{k}=1}^{\mathrm{K_c}} \left(-\frac{\langle\alpha_k\rangle}{2}{\mathrm{w}_{\mathrm{d,k}}}^2\right) + \text{const} \tag{47}$$

Finally, joining Equations (45) and (47), we would have that the optimum solution for

the variable $\mathbf{W}$ is:

$$\begin{aligned}
\ln(q^*(\mathbf{W})) &= \sum_{d=1}^{D}\sum_{k=1}^{K_c}\left(-\frac{\langle\alpha_k\rangle}{2}{w_{d,k}}^2\right) + \sum_{n=1}^{N}\Big(\langle\tau\rangle\, \mathbf{x}_{n,:}\, \mathbf{W}\langle\mathbf{z}_{n,:}\rangle \\
&\quad -\frac{\langle\tau\rangle}{2}\sum_{d=1}^{D}\big(\mathbf{w}_{d,:}\langle\mathbf{z}_{n,:}^T\, \mathbf{z}_{n,:}\rangle\, \mathbf{w}_{d,:}^T\big)\Big) + \text{const} \\
&= \sum_{d=1}^{D}\left(-\frac{1}{2}\,\mathbf{w}_{d,:}\, diag(\langle\alpha_k\rangle)\, \mathbf{w}_{d,:}^T + \sum_{n=1}^{N}\Big(\langle\tau\rangle\, x_{n,d}\, \mathbf{w}_{d,:}\langle\mathbf{z}_{n,:}\rangle\right. \\
&\quad \left.-\frac{\langle\tau\rangle}{2}\,\mathbf{w}_{d,:}\langle\mathbf{z}_{n,:}^T\, \mathbf{z}_{n,:}\rangle\, \mathbf{w}_{d,:}^T\Big)\right) + \text{const} \\
&= \sum_{d=1}^{D}\Big(-\frac{1}{2}\,\mathbf{w}_{d,:}\big(diag(\langle\alpha_k\rangle) + \langle\tau\rangle\langle\mathbf{z}_{n,:}^T\, \mathbf{z}_{n,:}\rangle\big)\, \mathbf{w}_{d,:}^T + \langle\tau\rangle\, \mathbf{w}_{d,:}\langle\mathbf{Z}^T\rangle\, \mathbf{x}_{:,d}^T\Big) + \text{const}
\end{aligned} \tag{48}$$

This way, comparing the results with the normal distribution, we can identify terms, extracting the following conclusions:

$$q^*(\mathbf{W}) = \prod_{d=1}^{D}\big(\mathcal{N}\big(\mathbf{w}_{d,:}\,|\mu_{\mathbf{w}_{d,:}}, \Sigma_{\mathbf{W}}\big)\big) \tag{49}$$

Where the variance would be:

$$\Sigma_{\mathbf{W}}^{-1} = diag(\langle\boldsymbol{\alpha}\rangle) + \langle\tau\rangle\langle\mathbf{Z}^T\, \mathbf{Z}\rangle \tag{50}$$

And the mean could be expressed as:

$$\begin{aligned}
\mu_{\mathbf{w}_{d,:}} &= \langle\tau\rangle\Sigma_{\mathbf{W}}\langle\mathbf{Z}\rangle^T\, \mathbf{x}_{:,d}^T \\
\langle\mathbf{W}\rangle &= \langle\tau\rangle\, \mathbf{X}^T\langle\mathbf{Z}\rangle\Sigma_{\mathbf{W}}
\end{aligned} \tag{51}$$

2.4 Calculation of Z

$$\ln(q^*(\mathbf{Z})) = \mathbb{E}_{\mathbf{W},\tau}[\ln(p(\mathbf{X},\mathbf{W},\mathbf{Z},\tau))] = \mathbb{E}_{\mathbf{W},\tau}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] + \mathbb{E}[\ln(p(\mathbf{Z}))] \tag{52}$$

The first term is similar to equation (17), changing the expectation dependency:

$$\mathbb{E}_{\mathbf{W},\tau}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] = \sum_{n=1}^{N}\left(\langle\tau\rangle\, \mathbf{x}_{n,:}\langle\mathbf{W}\rangle\, \mathbf{z}_{n,:}^T - \frac{\langle\tau\rangle}{2}\, \mathbf{z}_{n,:}\langle\mathbf{W}^T\, \mathbf{W}\rangle\, \mathbf{z}_{n,:}^T\right) + \text{const} \tag{53}$$

The second term could be calculated as:

$$\mathbb{E}[\ln(p(\mathbf{Z}))] = \ln(p(\mathbf{Z})) = \sum_{n=1}^{N}\left(-\frac{1}{2}\, \mathbf{z}_{n,:}\, \mathbf{z}_{n,:}^T\right) + \text{const} \tag{54}$$

Therefore, joining the different terms in Equations (53) and (54) , the optimum solution has the form:

$$\ln(q^*(\mathbf{Z})) \;=\; \sum_{\mathrm{n=1}}^{\mathrm{N}} \left(-\frac{1}{2}\, \mathbf{z}_{\mathrm{n},:}\, \mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} + \langle \tau \rangle\, \mathbf{x}_{\mathrm{n},:} \langle \mathbf{W} \rangle\, \mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} - \frac{\langle \tau \rangle}{2}\, \mathbf{z}_{\mathrm{n},:} \langle \mathbf{W}^{\mathrm{T}}\, \mathbf{W} \rangle\, \mathbf{z}_{\mathrm{n},:}^{\mathrm{T}} \right) + \mathrm{const} \tag{55}$$

This way, we can now identify terms, having that:

$$q^*(\mathbf{Z}) \;=\; \prod_{n=1}^{N} \left(\mathcal{N}\left(\mathbf{z}_{\mathrm{n},:} \,|\mu_{\mathbf{z}_{\mathrm{n},:}}, \Sigma_{\mathbf{Z}} \right) \right) \tag{56}$$

Where the variance would be:

$$\Sigma_{\mathbf{Z}}^{-1} \;=\; I + \langle \tau \rangle \langle \mathbf{W}^{\mathrm{T}}\, \mathbf{W} \rangle \tag{57}$$

And the mean could be expressed as:

$$\begin{aligned} \mu_{\mathbf{z}_{\mathrm{n},:}} \;&=\; \langle \tau \rangle\, \mathbf{x}_{\mathrm{n},:} \langle \mathbf{W} \rangle \Sigma_{\mathbf{Z}} \\ \langle \mathbf{Z} \rangle \;&=\; \langle \tau \rangle\, \mathbf{X} \langle \mathbf{W} \rangle \Sigma_{\mathbf{Z}} \end{aligned} \tag{58}$$

2.5 Calculation of $\boldsymbol{\alpha}$

$$\ln(q^*(\boldsymbol{\alpha})) \;=\; \mathbb{E}_{\mathbf{W}}[\ln(p(\mathbf{W} \,|\, \boldsymbol{\alpha}))] + \mathbb{E}[\ln(p(\boldsymbol{\alpha}))] \tag{59}$$

Similarly to what we have done before, we can develop both terms independently:

$$\begin{aligned} \ln(p(\mathbf{W} \,|\, \boldsymbol{\alpha})) \;&=\; \sum_{\mathrm{d=1}}^{\mathrm{D}} \sum_{\mathrm{k=1}}^{\mathrm{K_c}} \left(\frac{1}{2} \ln |\alpha_k| - \frac{1}{2}\, \alpha_k (\mathrm{w_{d,k}})^2 \right) + \mathrm{const} \\ &=\; \sum_{\mathrm{k=1}}^{\mathrm{K_c}} \left(\frac{D}{2} \ln(\alpha_k) - \frac{1}{2}\, \alpha_k\, \mathbf{w}_{:,\mathrm{k}}^{\mathrm{T}}\, \mathbf{w}_{:,\mathrm{k}} \right) + \mathrm{const} \end{aligned}$$

$$\mathbb{E}[\ln(p(\mathbf{W} \,|\, \boldsymbol{\alpha}))] \;=\; \sum_{\mathrm{k=1}}^{\mathrm{K_c}} \left(\frac{D}{2} \ln(\alpha_k) - \frac{1}{2}\, \alpha_k \langle \mathbf{w}_{:,\mathrm{k}}^{\mathrm{T}}\, \mathbf{w}_{:,\mathrm{k}} \rangle \right) + \mathrm{const} \tag{60}$$

The second term is deployed as follows:

$$\ln(p(\alpha_k)) \;=\; -\,\beta_0\, \alpha_k + (\alpha_0 - 1) \ln(\alpha_k) + \mathrm{const}$$

$$\mathbb{E}[\ln(p(\boldsymbol{\alpha}))] \;=\; \sum_{\mathrm{k=1}}^{\mathrm{K_c}} (\ln(p(\alpha_k))) = \sum_{\mathrm{k=1}}^{\mathrm{K_c}} (-\beta_0\, \alpha_k + (\alpha_0 - 1) \ln(\alpha_k)) + \mathrm{const} \tag{61}$$

Therefore, by joining both Equations (60) and (61) together, we can now determine the value f the expectation we were looking for

$$\ln(q^*(\boldsymbol{\alpha})) \;=\; \sum_{\mathrm{k=1}}^{\mathrm{K_c}} \left(\frac{D}{2} \ln(\alpha_k) - \frac{1}{2}\, \alpha_k \langle \mathbf{w}_{:,\mathrm{k}}^{\mathrm{T}}\, \mathbf{w}_{:,\mathrm{k}} \rangle - \beta_0\, \alpha_k + (\alpha_0 - 1) \ln(\alpha_k) \right) + \mathrm{const} \tag{62}$$

This way, we can now identify terms, having that:

$$q^*(\boldsymbol{\alpha}) \;=\; \prod_{\mathrm{k=1}}^{\mathrm{K_c}} (Gamma(\alpha_k \,|a_{\alpha_k}, b_{\alpha_k})) \tag{63}$$

Where the variance would be:

$$a_{\alpha_k} \;=\; \frac{D}{2} + \alpha_0 \tag{64}$$

And the mean could be expressed as:

$$b_{\alpha_k} \;=\; \beta_0 + \frac{1}{2}\langle \mathbf{w}_{:,\mathrm{k}}^{\mathrm{T}}\, \mathbf{w}_{:,\mathrm{k}} \rangle = \beta_0 + \frac{1}{2}\langle \mathbf{W}^{\mathrm{T}}\, \mathbf{W} \rangle_{k,k} \tag{65}$$

where

$$\langle \mathbf{w}_{:,\mathrm{k}}^{\mathrm{T}}\, \mathbf{w}_{:,\mathrm{k}} \rangle \;=\; \sum_{\mathrm{d=1}}^{\mathrm{D}} (\mathrm{w_{d,k}})^2 = \langle \mathbf{W}^{\mathrm{T}}\, \mathbf{W} \rangle_{k,k}$$

2.6 Calculation of τ

$$\ln(q^*(\tau)) \;=\; \mathbb{E}_{\mathbf{W},\mathbf{Z}}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] + \mathbb{E}[\ln(p(\tau))] \tag{66}$$

Similarly to what was done in equation (45):

$$\begin{aligned}
\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau)) \;&=\; \sum_{\mathrm{n=1}}^{\mathrm{N}} \ln\Big(\mathcal{N}\Big(\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}, (\tau)^{-1} I\Big)\Big) + \mathrm{const} \\
&=\; \sum_{\mathrm{n=1}}^{\mathrm{N}} \sum_{\mathrm{d=1}}^{\mathrm{D}} \left(\frac{1}{2}\ln|\tau| - \frac{\tau}{2}\big(\mathrm{x_{n,d}} - \mathbf{z}_{\mathrm{n},:}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\big)^2\right) + \mathrm{const} \\
&=\; \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{d=1}}^{\mathrm{D}} \left(\mathrm{x_{n,d}}^2 - 2\,\mathbf{w}_{\mathrm{d},:}\,\mathbf{z}_{\mathrm{n},:}^{\mathrm{T}}\,\mathrm{x_{n,d}} + \big(\mathbf{z}_{\mathrm{n},:}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\big)^2\right) + \mathrm{const} \\
&=\; \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{d=1}}^{\mathrm{D}} \mathrm{x_{n,d}}^2 - 2\sum_{\mathrm{d=1}}^{\mathrm{D}} \mathbf{w}_{\mathrm{d},:}\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{x}_{:,\mathrm{d}} + \sum_{\mathrm{d=1}}^{\mathrm{D}} \mathbf{w}_{\mathrm{d},:}\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}\,\mathbf{w}_{\mathrm{d},:}^{\mathrm{T}}\right) + \mathrm{const} \\
&=\; \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{d=1}}^{\mathrm{D}} \mathrm{x_{n,d}}^2 - 2\,\mathrm{Tr}\big\{\mathbf{W}\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{X}\big\} + \mathrm{Tr}\big\{\mathbf{W}\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}\,\mathbf{W}^{\mathrm{T}}\big\}\right) + \mathrm{const}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\ln(p(\mathbf{X}\,|\,\mathbf{W},\mathbf{Z},\tau))] \;=\; &\frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{d=1}}^{\mathrm{D}} \mathrm{x_{n,d}}^2 - 2\,\mathrm{Tr}\big\{\langle\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\rangle\,\mathbf{X}\big\}\right. \\
&\left. + \mathrm{Tr}\big\{\langle\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}\rangle\big\}\right) + \mathrm{const}
\end{aligned}$$

The second term could subsequently be determined as:

$$\mathbb{E}[\ln(p(\tau))] \;=\; \ln(p(\tau)) = -\beta_0^{\tau}\tau + (\alpha_0^{\tau} - 1)\ln(\tau) + \mathrm{const}$$

Joining them together, we have that:

$$\begin{aligned}\ln(q^*(\tau)) &= \frac{DN}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{d=1}}^{\mathrm{D}}{\mathrm{x_{n,d}}}^2\right. \\ &\quad \left. -2\,\mathrm{Tr}\big\{\langle\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\rangle\,\mathbf{X}\big\} + \mathrm{Tr}\big\{\langle\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}\rangle\big\}\right) \\ &\quad - \beta_0^{\tau}\tau + (\alpha_0^{\tau}-1)\ln(\tau) + \mathrm{const}\end{aligned}$$

So by identifying terms with the distribution, we would have that for this parameter:

$$q^*(\tau) = (Gamma(\tau|a_\tau, b_\tau)) \tag{67}$$

the variance would be:

$$a_\tau = \frac{DN}{2} + \alpha_0^{\tau} \tag{68}$$

and the mean could be expressed as:

$$b_\tau = \beta_0^{\tau} + \frac{1}{2}\left(\sum_{\mathrm{n=1}}^{\mathrm{N}}\sum_{\mathrm{d=1}}^{\mathrm{D}}{\mathrm{x_{n,d}}}^2 - 2\,\mathrm{Tr}\big\{\langle\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\rangle\,\mathbf{X}\big\} + \mathrm{Tr}\big\{\langle\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\rangle\langle\mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}\rangle\big\}\right) \tag{69}$$