VARIATIONAL AUTO ENCODER: A DERIVATION OF GAUSSIAN AND INVERSE-WISHART EVIDENCE LOWER BOUND'S (ELBO'S)

Kenneth Ezukwoke*
École des Mines de Saint-Etienne
Univ. Clermont Auvergne, CNRS UMR 6158 LIMOS
Mathematics and Industrial Engineering
Henri FAYOL institute, F-42023, Saint-Etienne, France
{ifeanyi.ezukwoke}@emse.fr

ABSTRACT

Bayesian Inference has recently gained traction as n efficient method for text and image generation. Much of the success of this model is attributed to its rather unique loss function used for the generative process. In this paper we present the mathematical details behind its inference method and derive the Evidence lower bounds (ELBO) both for Gaussian distributions and Inverse-Wishart distributions.

***K*eywords** Variational AutoEncoder; Evidence Lower Bound; Gaussian distribution; Inverse-Wishart distribution.

1 Introduction

2 Variational AutoEncoder (VAE)

VAE [1] is a generative model and a derivative of Bayesian inference. Unlike autoencoder, VAE are more suited for textual data and have been used for unsupervised abstractive sentence summarization [2] and long and coherent Text generation [3] amongst others. VAE specifies a prior distribution $p(z)$ to which an encoder with posterior probability $q_\phi(z|x)$ should map all x by maximizing a variational lower bound on a decoder posterior probability $p_\theta(x|z)$. ϕ and θ are the weights of the encoder and decoder respectively. Given an observation x_i, the marginal distribution over an unknown distribution is given as,

$$p_\theta(x_i) = \int_z p_\theta(x_i|z)p_\theta(z)dz, \tag{1}$$

The integral over z is computationally intractable for large number of z variables. We proceed by introducing the posterior distribution of the encoder in order to solve that challenge.

$$p_\theta(x_i) = \int_z p_\theta(x_i|z)p_\theta(z)\frac{q_\phi(z|x_i)}{q_\phi(z|x_i)}dz, \tag{2}$$

$$p_\theta(x_i) = \int_z q_\phi(z|x_i)\frac{p_\theta(x_i|z)p_\theta(z)}{q_\phi(z|x_i)}dz, \tag{3}$$

This is can be transformed into an expectation form as,

$$p_\theta(x_i) = \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\frac{p_\theta(x_i|z)p_\theta(z)}{q_\phi(z|x_i)}\right], \tag{4}$$

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

We take the natural log of both sides as follows,

$$\ln p_\theta(x_i) = \ln \left(\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)} \left[\frac{p_\theta(x_i|z)p_\theta(z)}{q_\phi(z|x_i)} \right] \right), \tag{5}$$

From Jensen inequality, we know that $\mathbb{E}[\ln X] \leq \ln[\mathbb{E}(X)]$, hence,

$$\ln p_\theta(x_i) \geq \mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)} \left[\ln \frac{p_\theta(x_i|z)p_\theta(z)}{q_\phi(z|x_i)} \right], \tag{6}$$

$$\ln p_\theta(x_i) \geq - \mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)} \left[\ln \frac{q_\phi(z|x_i)}{p_\theta(z)} \right] + \mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)} [\ln p_\theta(x_i|z)], \tag{7}$$

$$\ln p_\theta(x_i) \geq \mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)} [\ln p_\theta(x_i|z)] - D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)), \tag{8}$$

The R.H.S of equation 8 is referred to as the **Evidence Lower Bound** (ELBO) since it is upper-bounded by the logarithm of the evidence $p_\theta(x_i)$. $\mathbb{E}[\ln p_\theta(x_i|z)]$ is the reconstruction loss while $D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z))$ is a tractable regulation term known as Kullback-Leibler Divergence which ensures that the approximate posterior of the encoder $q_\phi(z|x_i)$ does not deviate far from its prior $p(z)$.

the loss or objective function to be maximized is given by,

$$\mathcal{L}_{VAE}(\theta, \phi) = \mathop{E}_{z\sim q_\phi(z|x_i)} [\log p_\theta(x_i|z)] - D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) \tag{9}$$

For two multivariate normal distribution the closed-form for the objective loss function of VAE is given by,

$$\mathcal{L}_{VAE}(\theta, \phi) = \mathop{E}_{z\sim q_\phi(z|x_i)} [\log p_\theta(x_i|z)] + \frac{1}{2} \left[\ln \left| \frac{\Sigma_1}{\Sigma_0} \right| + m - tr(\Sigma_1 \Sigma_0^{-1}) - (\mu_1 - \mu_0)^T \Sigma_0^{-1} (\mu_1 - \mu_0) \right] \tag{10}$$

We train the VAE to find the optimal parameters

$$(\theta^*, \phi^*) = argmax_{(\theta,\phi)} \mathcal{L}_{VAE}(\theta, \phi), \tag{11}$$

Due to the lack of statistical independence of the observation in the latent space generated by VAE, we will adopt the Inverse-Wishart(IW) prior for latent space disentanglement, on the covariance matrix [4]. The inverse-Wishart distribution $\mathcal{W}^{-1}(\Psi, \nu)$ is parameterized by a positive-definite scale matrix $\phi \in \mathbb{R}^{p\times p}$ and a degrees of freedom (DoF) $\nu \geq p - 1$. An inverse-Wishart distributed random matrix $X \in \mathbb{R}^{n\times m}$ has probability density function given by,

$$p(X) = \frac{\Psi^{\frac{v}{2}}}{2^{\frac{vp}{2}} \Gamma_p(\frac{v}{2})} X^{-\frac{(v+p+1)}{2}} \exp^{-\frac{1}{2}} tr(\Psi X^{-1}), \tag{12}$$

where $\Gamma_p(.)$ is the multivariate gamma function. For the inference update,

$$p(\sigma) \sim \mathcal{W}_p^{-1}(\Sigma|\Psi, v), \tag{13}$$

$$p(z|\sigma) \sim \mathcal{N}(z|0, \Sigma), \tag{14}$$

Where Σ is the covariance matrix of the matrix X. The

2.1 Transformer-Variational Autoencoder (Trans-VAE)

The objective of using the transformer-based variational autoencoder is to reduce the dimension of the original space (noise or non-keyword removal) of the data whilst preserving context.

We elaborate on the derivations of the VAE loss functions using Guassian and Wishart prior and posteriors.

2.2 Loss function: Gaussian distribution

We recall the objective function for the VAE as,

$$\mathcal{L}_{VAE}(\theta,\phi) = -\underbrace{D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z))}_{①regularizer} + \underbrace{\underset{z\sim q_\phi(z|x_i)}{E}[\log p_\theta(x_i|z)]}_{②reconstruction} \tag{15}$$

We assume that the latent prior and the approximate posterior both follow a Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$ respectively. We can compute the analytic closed form solutions for the two terms in the loss.

The Kullback-Liebler divergence $D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z))$ is given as,

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln \frac{q_\phi(z|x_i)}{p_\theta(z)}\right] \tag{16}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}[\ln q_\phi(z|x_i) - \ln p_\theta(z)] \tag{17}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}[\ln \mathcal{N}(\mu_1, \Sigma_1) - \ln \mathcal{N}(\mu_0, \Sigma_0)] \tag{18}$$

$$\begin{aligned} D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = & \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_1|^{\frac{1}{2}}} e^{-\frac{1}{2}\|x-\mu_1\|^2_{\Sigma_1^{-1}}}\right] \\ D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) & - \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_0|^{\frac{1}{2}}} e^{-\frac{1}{2}\|x-\mu_0\|^2_{\Sigma_0^{-1}}}\right] \end{aligned} \tag{19}$$

$$\begin{aligned} D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = & \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\|x-\mu_1\|^2_{\Sigma_1^{-1}}\right] \\ D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) & - \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_0| - \frac{1}{2}\|x-\mu_0\|^2_{\Sigma_0^{-1}}\right] \end{aligned} \tag{20}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\frac{1}{2}\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - \frac{1}{2}\|x-\mu_1\|^2_{\Sigma_1^{-1}} + \frac{1}{2}\|x-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{21}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \frac{1}{2}\underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \|x-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{22}$$

Introducing the trace operator, we have that,

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \frac{1}{2}\underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - tr[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)] + \|x-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{23}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \frac{1}{2}\underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - tr[(x-\mu_1)(x-\mu_1)^T\Sigma_1^{-1}] + \|x-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{24}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \frac{1}{2}\underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - tr[\Sigma_1\Sigma_1^{-1}] + \|x-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{25}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \frac{1}{2}\underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - tr[I] + \|x-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{26}$$

$$D_{KL}(q_\phi(z|x_i) \parallel p_\theta(z)) = \frac{1}{2}\left[\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - m\right] + \frac{1}{2}\underset{z\sim q_\phi(z|x_i)}{\mathbb{E}}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] \tag{27}$$

We expand the second term on the right hand side (R.H.S) as follows,

$$\frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) = \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr\left[(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)\right]\right] \tag{28}$$

$$\frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) = \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr\left[(x_i-\mu_1+\mu_1-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_1+\mu_1-\mu_0)\right]\right] \tag{29}$$

$$\begin{aligned}\frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) &= \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr\left[(x_i-\mu_1)^T\Sigma_0^{-1}(x_i-\mu_1)\right]\right]\\ \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) &\quad + \mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr\left[(x_i-\mu_1)^T\Sigma_0^{-1}(\mu_0-\mu_1)\right]\right]\\ \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) &\quad + \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr\left[(\mu_1-\mu_0)^T\Sigma_0^{-1}(\mu_1-\mu_0)\right]\right]\end{aligned} \tag{30}$$

Since $\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}[(x_i-\mu_1)]$ is zero, we have that

$$\begin{aligned}\frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right]\frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] &= \frac{1}{2}tr\left[\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[(x_i-\mu_1)(x_i-\mu_1)^T\Sigma_0^{-1}\right]\right]\\ &\quad + \frac{1}{2}tr\left[\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[\underbrace{(\mu_1-\mu_0)^T\Sigma_0^{-1}(\mu_1-\mu_0)}_{constant}\right]\right]\end{aligned} \tag{31}$$

$$\frac{1}{2}\mathop{\mathbb{E}}_{z\sim q_\phi(z|x_i)}\left[tr[\|x-\mu_0\|^2_{\Sigma_0^{-1}}]\right] = \frac{1}{2}[tr(\Sigma_1\Sigma_0^{-1}) + \|\mu_1-\mu_0\|^2_{\Sigma_0^{-1}}] \tag{32}$$

By substituting equation 32 into 27, the KL divergence can be rewritten as,

$$D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) = \frac{1}{2}\left[\ln\left|\frac{\Sigma_0}{\Sigma_1}\right| - m + tr(\Sigma_1\Sigma_0^{-1}) + \|\mu_1-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{33}$$

The negative KL is given as

$$-D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) = \frac{1}{2}\left[\ln\left|\frac{\Sigma_1}{\Sigma_0}\right| + m - tr(\Sigma_1\Sigma_0^{-1}) - \|\mu_1-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{34}$$

However, if $p_\theta(z)\sim\mathcal{N}(z;0,I)$, then the negative KL reduces to,

$$-D_{KL}(q_\phi(z|x_i)\parallel p_\theta(z)) = \frac{1}{2}\left[\ln|\Sigma_1| + m - tr(\Sigma_1) - (\mu_1)^T(\mu_1)\right] \tag{35}$$

Finally, the objective functions can be writen as,

$$\mathcal{L}^{\mathcal{G}}_{VAE}(\theta,\phi) = \mathop{E}_{z\sim q_\phi(z|x_i)}[\log p_\theta(x_i|z)] + \frac{1}{2}\left[\ln\left|\frac{\Sigma_1}{\Sigma_0}\right| + m - tr(\Sigma_1\Sigma_0^{-1}) - \|\mu_1-\mu_0\|^2_{\Sigma_0^{-1}}\right] \tag{36}$$

2.3 Loss function: Inverse-Wishart distribution

As previously shown for Gaussian distribution, we derive similarly the KL-divergence $D_{KL}(\mathcal{W}^{-1}(\nu_0,\Phi_0)\parallel\mathcal{W}^{-1}(\nu_1,\Phi_1))$ for two Inverse-Wishart distributions. We note that for a random variable X which follows an Inverse-Wishart distribution, $\mathbb{E}[\ln X] = \ln|\Phi| - m\ln 2 - \phi_m(\frac{\nu}{2})$ and $\mathbb{E}[X] = \nu\Phi^{-1}$. The KL is given as,

$$D_{KL}(\mathcal{W}_1^{-1}(\nu_1,\Phi_1)\parallel\mathcal{W}_0^{-1}(\nu_0,\Phi_0)) = \mathop{\mathbb{E}}_{z\sim\mathcal{W}^{-1}}\left[\ln\frac{\mathcal{W}^{-1}(\nu_1,\Phi_1)}{\mathcal{W}^{-1}(\nu_0,\Phi_0)}\right] \tag{37}$$

We denote $D_{KL}(\mathcal{W}_1^{-1}(\nu_1,\Phi_1)\parallel\mathcal{W}_0^{-1}(\nu_0,\Phi_0))$ using the compact form $D_{KL}(\mathcal{W}_1^{-1}\parallel\mathcal{W}_0^{-1})$, so that,

$$D_{KL}(\mathcal{W}_1^{-1}\parallel\mathcal{W}_0^{-1}) = \mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\underbrace{\left[\ln\mathcal{W}^{-1}(\nu_1,\Phi_1)\right]}_{①term} - \underbrace{\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}[\mathcal{W}^{-1}(\nu_0,\Phi_0)]}_{②term} \tag{38}$$

We expand the first term as follow,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\ln\mathcal{W}^{-1}(\nu_1,\Phi_1)\right] = \mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\frac{\nu_1}{2}\ln|\Phi_1|\right] - \mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\frac{\nu_1+m+1}{2}\ln|X|\right] - \mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\frac{1}{2}tr(\Phi_1X^{-1})\right] - \mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\frac{\nu_1 m}{2}\ln 2 + \ln\Gamma_m(\frac{\nu_1}{2})\right] \tag{39}$$

Since the $\mathbb{E}[k]=k$, where k is a constant and $\mathbb{E}(tr[M])=tr(\mathbb{E}[M])=\mu_m$ where μ_m is the average of the entries of matrix M. Hence,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\ln\mathcal{W}^{-1}(\nu_1,\Phi_1)\right] = \frac{\nu_1}{2}\ln|\Phi_1| - \ln\Gamma_m(\frac{\nu_1}{2}) - \left(\frac{\nu_1+m+1}{2}\right)\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}[\ln|X|] - tr\left[\frac{1}{2}\left(\Phi_1\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}[X^{-1}]\right)\right] \tag{40}$$

The expectation of the log of the determinant of a matrix is expanded as follow,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\ln\mathcal{W}^{-1}(\nu_1,\Phi_1)\right] = \frac{\nu_1}{2}\ln|\Phi_1| - \ln\Gamma_m(\frac{\nu_1}{2}) - \frac{\nu_1+m+1}{2}\left[\ln\frac{|\Phi_1|}{2^m} - \psi_m(\frac{\nu_1}{2})\right] \tag{41}$$

$$- tr\left[\frac{1}{2}\left(\Phi_1\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}[X^{-1}]\right)\right] \tag{42}$$

For a positive definite matrix X, its Inverse-Wishart expectation is given by $\frac{\Phi}{\nu-m-1}$ for $\nu > m+1$, hence,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\ln\mathcal{W}^{-1}(\nu_1,\Phi_1)\right] = \frac{\nu_1}{2}\ln|\Phi_1| - \ln\Gamma_m(\frac{\nu_1}{2}) - \frac{\nu_1+m+1}{2}\left[\ln|\Phi_1| - \psi_m(\frac{\nu_1}{2})\right] - tr\left[\frac{1}{2}\left(\Phi_1\frac{\Phi_1^{-1}}{(\nu_1-m-1)^{-1}}\right)\right] \tag{43}$$

$$\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\ln\mathcal{W}^{-1}(\nu_1,\Phi_1)\right] = \frac{\nu_1}{2}\ln|\Phi_1| - \ln\Gamma_m(\frac{\nu_1}{2}) - \frac{\nu_1+m+1}{2}\left[\ln|\Phi_1| - \psi_m(\frac{\nu_1}{2})\right] - tr\left[\frac{1}{2}\left(\frac{I_m}{(\nu_1-m-1)^{-1}}\right)\right] \tag{44}$$

$$\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\ln\mathcal{W}^{-1}(\nu_1,\Phi_1)\right] = \frac{\nu_1}{2}\ln|\Phi_1| - \frac{\nu_1 p}{2} - \ln\Gamma_m(\frac{\nu_1}{2}) - \frac{\nu_1+m+1}{2}\left[\ln|\Phi_1| - \psi_m(\frac{\nu_1}{2})\right] - \frac{(\nu_1-m-1)}{2}m \tag{45}$$

Similarly, we expand the second term to obtain,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{W}_1^{-1}}\left[\ln\mathcal{W}^{-1}(\nu_0,\Phi_0)\right] = \frac{\nu_0}{2}\ln|\Phi_0| - \ln\Gamma_m(\frac{\nu_0}{2}) - \frac{\nu_1+m+1}{2}\left[\ln|\Phi_1| - \psi_m(\frac{\nu_1}{2})\right] - \frac{1}{2}tr\left(\frac{\Phi_0\Phi_1^{-1}}{(\nu_0-m-1)^{-1}}\right) \tag{46}$$

If $\nu_1=\nu_0$ and the degree of freedom is very large indicating an informative prior i.e $\nu \gg m$, then we can combine equations (45) and (46) to obtain the following,

$$D_{KL}(\mathcal{W}_1^{-1}\parallel\mathcal{W}_0^{-1}) = \frac{\nu_0}{2}\ln|\Phi_1\Phi_0^{-1}| + \frac{(\nu_0-m-1)}{2}tr(\Phi_0\Phi_1^{-1}) - \frac{(\nu_1-m-1)}{2}m + \ln\frac{\Gamma_m(\frac{\nu_0}{2})}{\Gamma_m(\frac{\nu_1}{2})} - \frac{\nu_1+m+1}{2}\left[\ln|\Phi_1| - \psi_m(\frac{\nu_1}{2})\right] \tag{47}$$

$$D_{KL}(\mathcal{W}_1^{-1}\parallel\mathcal{W}_0^{-1}) = \frac{\nu_0}{2}\ln|\Phi_1\Phi_0^{-1}| + \frac{\nu_1}{2}(tr(\Phi_0\Phi_1^{-1}) - m) + \underbrace{\frac{\Gamma_m(\frac{\nu_0}{2})}{\Gamma_m(\frac{\nu_1}{2})} - \frac{\nu_1}{2}\left[\ln|\Phi_1| - \psi_m(\frac{\nu_1}{2})\right]}_{constant} \tag{48}$$

and since the last two terms are constant, we can easily reduce the KL to

$$D_{KL}(\mathcal{W}_1^{-1} \parallel \mathcal{W}_0^{-1}) = \frac{\nu_0}{2} \ln |\Phi_1 \Phi_0^{-1}| + \frac{\nu_1}{2}(tr(\Phi_0 \Phi_1^{-1}) - m) \tag{49}$$

We can further expand the R.H.S of equation (49) as follow,

$$\begin{aligned} D_{KL}(\mathcal{W}_1^{-1} \parallel \mathcal{W}_0^{-1}) = \frac{\nu_0}{2} \ln |\Phi_1| + \frac{\nu_1}{2} tr(\Phi_0 \Phi_1^{-1}) \\ \underbrace{-\frac{\nu_1}{2} \ln |\Phi_0^{-1}| - \frac{\nu_1}{2} m}_{constant} \end{aligned} \tag{50}$$

$$D_{KL}(\mathcal{W}_1^{-1} \parallel \mathcal{W}_0^{-1}) = \frac{\nu_0}{2} \ln |\Phi_1| + \frac{\nu_1}{2} tr(\Phi_0 \Phi_1^{-1}) \tag{51}$$

Finally, the loss function for the Inverse-Wishart VAE can be given as,

$$\mathcal{L}_{VAE}^{\mathcal{W}^{-1}} = \underset{z \sim q_\phi(z|x_i)}{E} [\log p_\theta(x_i|z)] - \frac{\nu_0}{2} \ln |\Phi_1| - \frac{\nu_1}{2} tr(\Phi_0 \Phi_1^{-1}) \tag{52}$$

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[2] Raphael Schumann. Unsupervised abstractive sentence summarization using length controlled variational autoencoder, 2018.

[3] Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. Towards generating long and coherent text with multi-level latent variable models, 2019.

[4] Abdul Fatir Ansari and Harold Soh. Hyperprior induced unsupervised disentanglement of latent representations, 2019.