Lecture Notes on Image Classification

Hervé Delingette

October 2022

1 Evidence lower bound

We consider the generic case of a set X of N observed random variables X_n, $n \in [1, \ldots N]$. It is assumed that those observation can be explained by generic probabilistic model $p(X|Z)$ where Z is a set of P latent (unobserved) random variables. Note that in its generic form, the size of the latent variable P may differ from the dimension N of the observed variable X The associated inverse problem is described by graphical model of Fig.1.

The latent prior $p(Z)$, the likelihood $p(X|Z)$, the posterior $p(Z|X)$, and the marginal likelihood or evidence $p(X)$ are related through the Bayes law and law of total probability :

$$\begin{aligned}
p(X|Z) &= \frac{p(X,Z)}{p(Z)} \\
p(Z|X) &= \frac{p(X|Z)p(Z)}{p(X)} \\
p(X) &= \sum_Z p(X|Z)p(Z) = \sum_Z p(X,Z) \quad \text{if Z discrete} \\
p(X) &= \int p(X|Z)p(Z)\ dZ = \int p(X,Z)\ dZ \quad \text{if Z continuous}
\end{aligned}$$

The estimation of the posterior probability $p(Z|X)$, requires the computation of the evidence $p(X)$, the probability of observed variables, which often cannot be computed in closed form.

Then, an approximation of the posterior distribution is introduced as variable $U(Z) \approx p(Z|X)$. The evidence lower bound (ELBO) is linking the evidence $p(X)$ with its lower bound defined as function of the surrogate variable U. We derive the ELBO for various configurations of the latent space Z, depending on its discrete or continuous nature.

1.1 Categorical latent variables

We assume that Z is a set of P categorical variables that can take K values. We use one-hot-encoding for the latent variable Z_p which means that Z_p is a vector of K binary values $Z_{pk} \in \{0,1\}$ such that only one of such values is 1, i.e. $\sum_{k=1}^{K} Z_{pk} = 1$. We write e_i^K the canonical one-hot vector of dimension K such that $e_i^K[i] = 1$ and $e_i^K[j \neq i] = 0$. The probability that variable Z_p is in class k writes as follows : $p(Z_p = e_k^K) = p(Z_{pk} = 1)$.

The surrogate posterior variable U follows the same structure as Z, i.e. it is a collection of P categorical variables U_p that can take K values. Furthermore, we write $p(U_{pk} = 1) = u_{pk}$ such that u_p is a vector of dimension K $u_p = (u_{p1}, \dots u_{pK})^T \in [0,1]^K$, such that : $\sum_{k=1}^{K} u_{pk} = 1$. Then we have :

$$\begin{aligned}\log p(X) &= (\sum_{k=1}^{K} u_{pk}) \log p(X) \\ &= \frac{1}{P} (\sum_{p=1}^{P} \sum_{k=1}^{K} u_{pk}) \log p(X)\end{aligned}$$

Based on the definition of the conditional probability, we have $p(X, Z_{pk}) = p(Z_{pk}|X)p(X)$. Therefore we can write $p(X) = \frac{U_{pk}}{p(Z_{pk}|X)} \frac{p(X,Z_{pk})}{1} \frac{1}{U_{pk}}$ leading to :

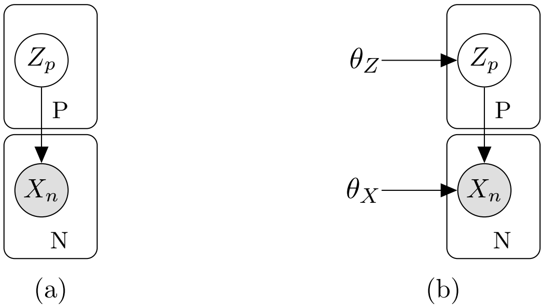
$$\begin{aligned}\log p(X) &= \frac{1}{P} \sum_{p=1}^{P} \sum_{k=1}^{K} U_{pk} \log \left(\frac{U_{pk}}{p(z_{pk}|X)} \right) + \sum_{p=1}^{N} \sum_{k=1}^{K} U_{pk} \log p(X, z_{pk}) - \sum_{p=1}^{N} \sum_{k=1}^{K} U_{pk} \log U_{pk} \\ &= D_{KL}(U||p(Z|X)) + \mathbb{E}_U(\log p(X,Z)) + \mathbb{H}(U)\end{aligned}$$

where

- $D_{KL}(U||p(Z|X))$ is the Kullback-Leibler divergence between the surrogate probability distribution U and the posterior $p(Z|X)$. This divergence is positive or null and is null only if $U = p(Z|X)$.
- $\mathbb{E}_U(\log p(X,Z)) = \sum_{n=1}^{N} \sum_{k=1}^{K} U_{nk} \log p(x_n, z_n)$ is the expectation of the complete log likelihood $p(X,Z)$ with respect to variable U.
- $\mathbb{H}(U) = -\sum_{n=1}^{N} \sum_{k=1}^{K} U_{nk} \log U_{nk}$ is Shannon entropy of variable U

Since the Kullback-Leibler divergence is positive, the quantity $\mathcal{L}_{VI} = \mathbb{E}_U(\log p(X,Z)) + \mathbb{H}(U)$ is the **evidence lower bound** :

$$\log p(X) \geq \mathcal{L}_{VI} = \mathbb{E}_U(\log p(X,Z)) + \mathbb{H}(U) = \log p(X) - D_{KL}(U||p(Z|X))$$



(a) (b)

Figure 1: (a) Graphical model of a generic inverse problem with observed values x_n explained by latent variable Z_n;(b) Graphical model of a generic inverse problem for the Expectation-Maximization where the parameters θ_X, θ_Z are estimated jointly with the latent variable Z.

The opposite of the lower bound is the **variational free energy** : $\mathcal{L}_{VFE} = -\mathcal{L}_{VI}$.

The lower bound can take several forms for instance including the likelihood $p(X|Z)$ and prior probabilities $p(Z)$ since $p(X, Z) = p(X|Z)p(Z)$:

$$\begin{aligned} \mathcal{L}_{VI} &= -D_{KL}(U||p(X, Z)) && (1) \\ &= \mathbb{E}_U(\log p(X|Z)) - D_{KL}(U||p(Z)) && (2) \end{aligned}$$

This last expression used in the variational autoencoder literature can be interpreted in the following way. The first term $\mathbb{E}_U(\log p(X|Z))$ is such that the log likelihood is maximized (equivalent to the goodness of fit) while the second term $-D_{KL}(U||p(Z))$ penalizes the complexity of surrogate function U with respect to the prior $p(Z)$.

Continuous case . If X and Z are continuous variables, we write $\beta \in \mathbb{R}^P$ the integration variables for probability density function $Z(\beta)$. Then we can show similarly to the categorical case that the same relations holds for the lower bound. In this case the surrogate function $U(\beta) \in [0, 1]$ approximating the posterior sums to unity ($\int_{\mathbb{R}^P} U(\beta)\ d\beta = 1$) and we have the following relations:

$$
\begin{aligned}
D_{KL}(U||p(Z|X)) &= \int_{\mathbb{R}^l} U(\beta)\ \log\left(\frac{U(\beta)}{p(Z=\beta|X)}\right)\ d\beta \\
\mathbb{E}_U(\log p(X,Z)) &= \sum_{n=1}^{N}\int_{\mathbb{R}^l} U(\beta)\ \log p(x_n, Z=\beta)\ d\beta \\
\mathbb{H}(U) &= -\int_{\mathbb{R}^l} U(\beta)\log U(\beta)\ d\beta
\end{aligned}
$$

Note the lower bound expression is very generic as the latent variable Z may in fact include different random variables some of them commonly considered as parameters and other as hidden variables. The joint probability $p(X,Z)$ is usually expanded to reveal to true relationships between variables.

2 Expectation-Maximization algorithm

2.1 Generic Algorithm

The EM algorithm applies when one wants to estimate the parameters $\theta = \{\theta_X, \theta_Z\}$ associated with the likelihood function $p(X|Z,\theta_X)$ or the prior $p(Z|\theta_Z)$.

The second condition to apply the EM algorithm is that the marginal likelihood (*aka* the evidence) $p(X|\theta)$, and the posterior probability $p(Z|X,\theta)$ can be computed in closed form.

In this case, the estimation of the parameters θ is normally done by the maximization of the marginal log likelihood $\log p(X|\theta)$. Yet, this maximization is often difficult to perform in closed form.

The EM algorithm instead aims at replacing the maximization of the log marginal likelihood $\theta^{\text{opt}} = \arg\min_\theta \log p(X|\theta)$ by the maximization of its lower bound or equivalently by the minimization of the variational free energy :

$$(\theta^{\text{opt}}, U^{\text{opt}}) = \arg\min_{\theta,U} \mathcal{L}_{VI}(U,\theta) = \log p(X|\theta) - D_{KL}(U||p(Z|X,\theta))$$

By choosing $U = p(Z|X,\theta)$, the Kullback Leibler divergence becomes null and the lower bound is equal to the marginal log likelihood. By introducing the additional variable $U(Z)$, the maximization is relaxed and proceeds by iterating between these two steps :

- **Expectation-Step**. The lower bound $\mathcal{L}_{VI}(U,\theta) = \log p(X|\theta) - D_{KL}(U||p(Z|X,\theta))$ is optimized with respect to U by choosing $U = p(Z|X,\theta)$.
- **Maximization-Step**. The lower bound $\mathcal{L}_{VI}(U,\theta) = \mathbb{E}_U(\log p(X,Z|\theta)) + \mathbb{H}(U)$ is maximized with respect to θ. More precisely, only the conditional expectation depends on θ and therefore $\theta^{t+1} = \arg\max_\theta \mathbb{E}_U(\log p(X,Z|\theta))$. This optimization with respect to $\mathbb{E}_U(\log p(X,Z|\theta))$ is provably easier to perform than with respect to the log marginal likelihood.

Therefore the EM-algorithm can be seen as the relaxation of a maximization problem by introducing an extra variable which is solved by an alternate optimization. The iterative maximization of the lower bound is displayed in Fig. 2.

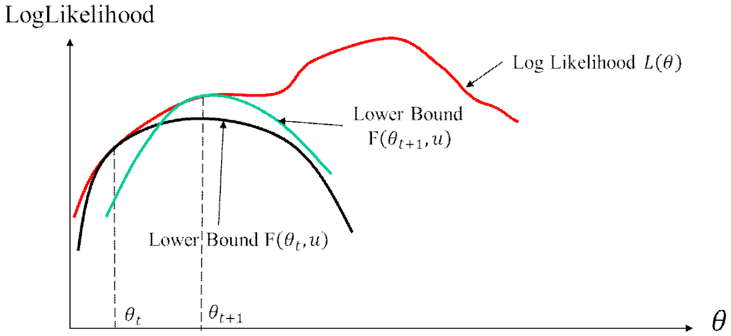


Figure 2: Illustration of the EM algorithm as the maximization of a lower bound; at each iteration the E-step consists in creating the lower bound which is touching the log likelihood at θ^t; the M-step consists in optimizing the lower bound at θ^{t+1}

The EM algorithm iterative increase the log-likelihood but is not guaranteed to converge towards a global maximum.

2.2 Application 1 : Gaussian Mixture Model

We show how the EM algorithm applies in the case of a Gaussian Mixture model. In such case, the observed random variable $X_n \in \mathbb{R}$ is continuous and the latent variable Z_n is a categorical variable belonging to K class. Furthermore, there are as many latent variables as observed ones ($P = N$), and the observations X_N are conditionally independent knowing the latent variable Z_n. Then for each class k, the X_n value is assumed to follow a

Gaussian distribution characterized by its mean value μ_k and variance σ_k^2 such that $\theta_X = \{\mu_k, \sigma_k^2\}$:

$$p(X_n|Z_{nk} = 1) = \mathcal{N}(x_n; \mu_k, \sigma_k^2)$$

The prior on the label $p(Z)$ is constant for all samples and parameterized by a multivariate Bernoulli having parameters $\theta_Z = \pi_k$ such that $\sum_{k=1}^{K} \pi_k = 1$. Thus we have :

$$p(Z_{nk} = 1) = \pi_k$$

The associated graphical model is displayed in Fig. 3(a).

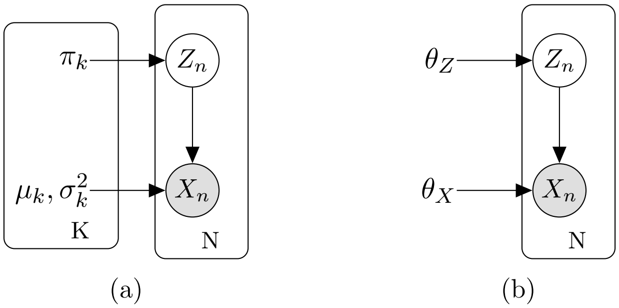


Figure 3: (a) Graphical model of a Gaussian mixture model with observed values X_n explained by latent variable Z_n;(b)

If this case, the marginal likelihood can be written in closed form :

$$p(X) = \sum_{k=1}^{K} p(X_n|Z_{nk})p(Z_{nk}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X_n; \mu_k, \sigma_k^2)$$

and the posterior probability :

$$p(Z_{nk} = 1|X_n) = \frac{\pi_k \mathcal{N}(X_n; \mu_k, \sigma_k^2)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(X_n; \mu_l, \sigma_l^2)} \tag{3}$$

The EM algorithm aims to maximize the log marginal likelihood $\log p(X|\theta)$ by introducing a surrogate probability function U_{nk} and by maximizing a lower bound $\mathcal{L}_{VI}(U, \theta)$ of the log marginal likelihood. In this case, we have $p(X_n, Z_n) = p(X_n|Z_n)p(Z_n)$ which implies that :

$$\begin{aligned}\mathcal{L}_{VI}(U, \theta) &= \log p(X|\theta) - D_{KL}(U||p(Z|X, \theta)) \\ &= \mathbb{E}_U(\log p(X, Z|\theta)) + \mathbb{H}(U) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} U_{nk} \left(\log(\pi_k \mathcal{N}(X_n; \mu_k, \sigma_k^2)) - \log U_{nk}\right)\end{aligned}$$

The 2 steps of the EM-algorithm then becomes :

- **E-step**. Determine $U_{nk} = p(Z_{nk} = 1|X_n)$ as the posterior probability given by equation 3.
- **M-step**. Maximize $\mathcal{L}_{VI}(U, \theta)$ with respect to $\theta_Z = \{\pi_k\}$ and $\theta_X = \{\mu_k, \sigma_k^2\}$ giving :

$$\begin{aligned}
\pi_k &= \frac{\sum_{n=1}^{N} U_{nk}}{N} \\
\mu_k &= \frac{\sum_{n=1}^{N} U_{nk} X_n}{\sum_{n=1}^{N} U_{nk}} \\
\sigma_k^2 &= \frac{\sum_{n=1}^{N} U_{nk} (X_n - \mu_k)^2}{\sum_{n=1}^{N} U_{nk}}
\end{aligned}$$

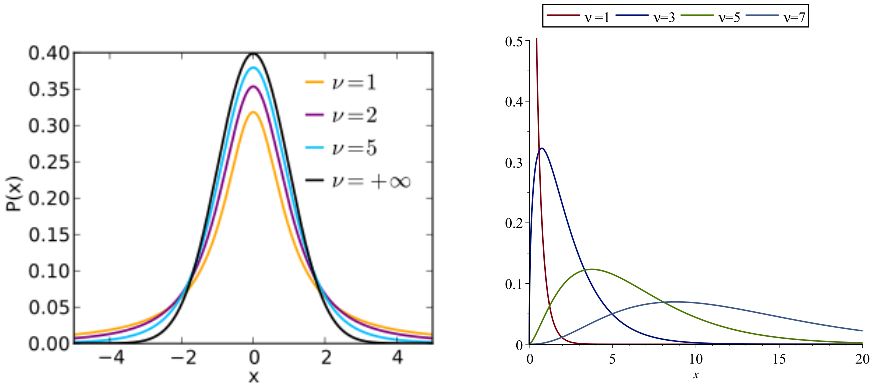
2.3 Application 2 : Student-t distribution

In this case, we consider fitting a Student-t distribution on a set of observations $\{X_n\}$. A Student-t is a generalization of a Gaussian distribution where an additional parameter ν, the degrees of freedom, is introduced. When $\nu \to +\infty$, then the Student-t $S(x; \mu, \sigma^2, \nu) \to \mathcal{N}(x; \mu, \sigma^2)$ converges towards a Gaussian distribution. For finite values of ν the Student-t distribution has an heavy tail, meaning that it makes values away from the mean more probable than for a normal distribution. The probability density function of a student is :

$$S(x; \mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{2\pi\sigma^2}} \left(1 + \frac{(x-\mu)^2}{\sigma^2 \nu}\right)^{-\frac{\nu+1}{2}} \tag{4}$$

Given N observations $X_n \in \mathbb{R}$ the problem is to estimate the 3 parameters θ of the Student-t distribution i.e. , the mean μ, variance σ^2 and degrees of freedom ν. The maximization of the log likelihood $\log p(X|\theta) = \sum_{n=1}^{N} p(X_n|\theta)$ does not lead to any closed form expression unlike the Gaussian case. Instead of resorting to the joint non linear optimization $\theta^{\text{opt}} = \arg\max_\theta \log p(X|\theta)$, one can use an EM algorithm where μ and σ^2 (but not ν) can be estimated in closed form.

This EM fitting approach is possible because a Student-t distribution may be seen as a Gaussian scale mixture, i.e. as an infinite mixture of Gaussian distributions having the same mean but varying variance. More precisely, the precision (inverse of the variance) is following a Gamma law



(a) (b)

Figure 4: (a) Student-t distribution of zero mean and unit standard deviation and various values of the number of degrees of freedom ν. For $\nu = \infty$, the distribution becomes a Gaussian distribution.(b) Gamma distribution $\mathrm{G}a(x; \frac{\nu}{2}, \frac{\nu}{2})$ for various value of ν. Each distribution has mean 1 and variance $\frac{2}{\nu}$.

parameterized by $\frac{\nu}{2}$:

$$S(x; \mu, \sigma^2, \nu) = \int_0^\infty \mathcal{N}(x; \mu, \sigma^2/\tau)\ \mathrm{G}a(\tau; \frac{\nu}{2}, \frac{\nu}{2})\ d\tau \tag{5}$$

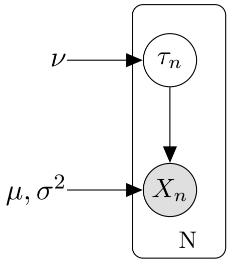
The Gamma distribution $\mathrm{G}a(\tau; \alpha, \beta) = \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)}$ (see Fig 4(b)) applies on positive scalars and is classically defined by the shape α and rate β parameters (the scale parameters $\theta = 1/\beta$ is also often used).

We can then formulate the problem of fitting the parameters $\theta = \{\mu, \sigma^2, \nu\}$ of a Student-t distribution from data, as solving with the EM algorithm an inverse problem involving the hidden variables τ_n and the parameters θ.

More precisely, writing $\mathrm{T} = \{\tau_n\}$ the hidden variables, and $\theta_X = \{\mu, \sigma^2\}$ the Gaussian parameters, we have :

$$p(X_n|\theta_X, \tau_n) = \mathcal{N}(X_n; \mu, \sigma^2/\tau_n)$$
$$p(\tau_n|\nu) = \mathrm{G}a(\tau_n; \frac{\nu}{2}, \frac{\nu}{2})$$

The marginal log likelihood is same as the likelihood associated with the Student-t $p(X_n|\theta) = \int_0^\infty p(X_n|\tau_n)\ p(\tau_n|\nu)\ d\tau_n = S(X_n; \nu, \sigma^2, \nu)$.



(a)

Figure 5: (a) Graphical model for fitting a Student-t with observed values X_n following a Gaussian distribution and precision variables τ_n following a Gamma distribution;(b)

The posterior distribution of the hidden variable τ used in the E-step can be written as a Gamma distribution :

$$\begin{aligned} p(\tau_n|X_n) &\propto p(X_n|\tau_n)\ p(\tau_n) \\ &= \mathrm{G}a(\tau_n; a_n, b_n) \end{aligned}$$

with $a_n = \frac{\nu+1}{2}$ and $b_n = \frac{\nu}{2} + \frac{(X_n-\mu)^2}{2\sigma^2}$.

The complete log-likelihood involved in the lower bound is written as :

$$\begin{aligned} \log p(X_n, \tau_n|\theta) &= \log p(X_n|\tau_n, \theta_X) + \log p(\tau_n|\nu) \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 + \frac{1}{2}\log\tau_n - \frac{\tau_n}{2\sigma^2}(X_n-\nu)^2 \\ &\quad - \log\Gamma(\frac{\nu}{2}) + \frac{\nu}{2}\log\frac{\nu}{2} + (\frac{\nu}{2}-1)\log\tau_n - \frac{\nu}{2}\tau_n \end{aligned}$$

The EM-algorithm proceeds by introducing N continuous variables $U_n(\tau)$. In the E-step, those variables are set to $U_n(\tau) = p(\tau_n|X_n,\theta) = \mathrm{G}a(\tau_n; a_n, b_n)$.

The M-step relies on the evidence lower bound $\mathcal{L}_{VI}(U,\theta) = \mathbb{E}_U(\log p(X,Z|\theta)) + \mathbb{H}(U)$. Since, the entropy term $\mathbb{H}(U)$ is independent of θ, we concentrate on the expectation term :

$$\begin{aligned} \mathbb{E}_U(\log p(X,Z|\theta)) &= \sum_{n=1}^{N}\int_0^\infty U_n(\tau_n)\log p(X_n,\tau_n|\theta)\ d\tau_n \\ &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 + \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[\log\tau_n] - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(X_n-\mu)^2\mathbb{E}[\tau_n] \\ &\quad - N\log\Gamma(\frac{\nu}{2}) + N\frac{\nu}{2}\log\frac{\nu}{2} + (\frac{\nu}{2}-1)\sum_{n=1}^{N}\mathbb{E}[\log\tau_n] - \frac{\nu}{2}\sum_{n=1}^{N}\mathbb{E}[\tau_n] \end{aligned}$$

where $\mathbb{E}[\tau_n] = \int_0^\infty \tau_n \text{G}a(\tau_n; a_n, b_n)$ and $\mathbb{E}[\log \tau_n] = \int_0^\infty \log \tau_n \text{G}a(\tau_n; a_n, b_n)$. Since the expectation of a Gamma function $\text{G}a(\tau_n; \alpha, \beta)$ is $\frac{\alpha}{\beta}$, we have $\mathbb{E}[\tau_n] = \frac{a_n}{b_n}$. The second term can be also computed in closed form and gives : $\mathbb{E}[\log \tau_n] = \psi(a_n) - \log b_n$ where $\psi(x)$ is the digamma function. We introduce the quantity $\hat{\tau_n} = \mathbb{E}[\tau_n] = \frac{a_n}{b_n} = \frac{\nu+1}{\nu+\frac{(X_n-\mu)^2}{\sigma^2}}$ as the expectation of the normalized precision.

The M-step consists in finding the mean μ, the variance σ^2 and the degrees of freedom which maximizes $\mathbb{E}_U(\log p(X, Z|\theta))$. The maximization gives with respect to μ and σ^2 gives two closed form relations:

$$\frac{\partial \mathbb{E}_U(\log p(X, Z|\theta))}{\partial \mu} = 0 \Rightarrow \mu = \frac{\sum_{n=1}^N \hat{\tau_n} X_n}{\sum_{n=1}^N \hat{\tau_n}}$$

$$\frac{\partial \mathbb{E}_U(\log p(X, Z|\theta))}{\partial \sigma^2} = 0 \Rightarrow \sigma^2 = \frac{1}{N}(\sum_{n=1}^N (X_n - \mu)^2 \hat{\tau_n})$$

We see that $\hat{\tau_n}$ acts as a weight associated with each data X_n. When the Mahalanobis distance $\frac{(X_n-\mu)^2}{\sigma^2}$ is larger than 1, then the expected normalized precision $\hat{\tau_n} = \frac{\nu+1}{\nu+\frac{(X_n-\mu)^2}{\sigma^2}}$ becomes less than 1 and therefore are less taken into account than those data points that are closer to the mean value.

For the optimization of the degrees of freedom ν, no closed-form solution can be obtained and numerical optimization must be performed :

$$\frac{\partial \mathbb{E}_U(\log p(X, Z|\theta))}{\partial \nu} = 0 \Rightarrow \psi(\frac{\nu}{2}) - \log \frac{\nu}{2} = 1 + \frac{1}{N} \sum_{n=1}^N (\psi(a_n) - \log b_n - \hat{\tau_n})$$

$$\psi(\frac{\nu+1}{2}) - \psi(\frac{\nu}{2}) + \log \frac{\nu}{2} - \log \frac{\nu+1}{2} = \frac{1}{N} \sum_{n=1}^N (\hat{\tau_n} - \log \hat{\tau_n} - 1)$$

3 Mean Field Approximation

3.1 Generic Result

We consider the generic case of section 1 of observed random variables X of dimension N and latent variables Z of dimension P. The objective of the mean field approximation is to approximate the posterior probability $p(Z|X)$ as a function $q(Z)$ which factorizes over its components. Assuming that $Z \in \mathbb{R}^P$ is a continuous variable and that $\beta \in \mathbb{R}^P$ is the integration variable such that $\int_{\mathbb{R}^P} p(Z=\beta)\ d\beta = 1$, then we assume that

$$q(\beta) = \prod_{p=1}^{P} q_p(\beta_p)$$

where $\beta_p \in \mathbb{R}$ is the p component of vector β. In other words, we assume that $q(\beta)$ is a product of P univariate functions. This hypothesis drastically simplifies the estimation of the posterior probability but at the same time is a very crude approximation of the true posterior.

The evidence lower bound presented in section 1, can be written as follows :

$$D_{KL}(U||p(Z|U)) = \log p(X) - \mathbb{E}_U(\log p(X,Z)) - \mathbb{H}(U)$$

We now replace $U(\beta)$ with $q(\beta) = \prod_{p=1}^{P} q_p(\beta_p)$ in this expression. We get :

$$D_{KL}(U||p(Z|U)) = \log p(X) + \int_{\mathbb{R}^P} \prod_{p=1}^{P} q_p(\beta_p) \left(-\log p(X,\beta) + \sum_{p=1}^{P} \log q_p(\beta_p) \right)\ d\beta$$

Now we consider that the $q_p(\beta_p)$, $p \neq j$ are known for a specific index j. We write the condition on $q_j(\beta_j)$ to minimize the Kullback-Leibler divergence.

$$\begin{aligned} D_{KL}(q||p(Z|q)) = \text{cst} + &\int_{\mathbb{R}} q_j(\beta_j) \log q_j(\beta_j)\ d\beta_j - \\ &\int_{\mathbb{R}} q_j(\beta_j) \left(\int_{\mathbb{R}^{P-1}} \prod_{p\neq j} q_p(\beta_p) \log p(X,\beta)\ d\beta_{\neq j} \right) d\beta_j \\ &= \text{cst} + D_{KL}(q_j||\tilde{q}_j) \end{aligned}$$

where by construction we have :

$$\log \tilde{q}_j = \int_{\mathbb{R}^{P-1}} \prod_{p\neq j} q_p(\beta_p) \log p(X,\beta)\ d\beta_{\neq j} + \text{cst} \tag{6}$$

Thus for $q_j(\beta_j)$ to minimize the discrepancy between $p(Z|X)$ and $q(Z)$, it is necessary that $q_j(\beta_j) = \tilde{q}_j(\beta_j)$ where:

$$\tilde{q}_j(\beta_j) = \frac{\exp\left(\int_{\mathbb{R}^{P-1}} \prod_{p\neq j} q_p(\beta_p) \log p(X,\beta)\ d\beta_{\neq j}\right)}{\int_{\mathbb{R}} \exp\left(\int_{\mathbb{R}^{P-1}} \prod_{p\neq j} q_p(\beta_p) \log p(X,\beta)\ d\beta_{\neq j}\right)\ d\beta_j} \tag{7}$$

The mean-field algorithm thus proceeds by optimizing each approximate marginal distributions $q_j(\beta_j)$ separately and by iterating other all marginals. Algorithm 1 provides a sketch of the mean field algorithm.

Algorithm 1: Mean Field approximation of posterior $p(Z|X)$

```
input  : Joint probability function $p(X, \beta) = p(X|Z = \beta)p(Z = \beta)$
output: Approximate marginal distribution
        $q(\beta) = \prod_{p=1}^{P} q_p(\beta_p) \approx p(Z|X)$
/* Initialize the approximate marginals to $q_j^0(\beta_j)$          */
for $p \leftarrow 1$ to $P$ do
    $q_p(\beta_p) \leftarrow q^0(\beta_p)$
/* Loop until the change in the distribution $q(\beta_j)$ is
   smaller than a threshold                                  */
do
    $q^{\text{old}} \leftarrow \prod_{p=1}^{P} q_p(\beta_p)$
    /* Update the marginals one after the other              */
    for $p \leftarrow 1$ to $P$ do
        Update $q_p(\beta_p)$ according to Eq.9
while $\|q^{\text{old}} - \prod_{p=1}^{P} q_p(\beta_p)\| < \epsilon$
```

3.2 EM algorithm

3.3 Hidden Potts Model

We extend the previous work in Gaussian Mixture Model by modifying the hypothesis about the prior on the label. In section2.2, the prior on the labels $p(Z_{nk} = 1)$ was supposed to be constant for all samples, i.e. $p(Z_{nk} = 1) = \pi_k$. In the Hidden Potts Model, we make a less stringent hypothesis by assuming that the prior of a label in a graph (and more precisely in an image), depends on its neighbors. Let $\mathcal{O}(n)$ be set of all voxel neighbors to voxel n. Then the Potts model assumes that the probability of a label depends on the label of its neighbors as follows:

$$p(Z_{nk}|Z_{\mathcal{O}(n)}) \propto \pi_k \exp\left(-\alpha \sum_{i \in \mathcal{O}(n)} (2Z_{ik}Z_{nk} - 1)\right)$$

where $Z_{\mathcal{O}(n)}$ is the set of random label variables on the neighbors of site n, and α is a positive scalar (often written as the inverse of a temperature) and π_k is a prior label probability. Since Z_n is a vector of binary variables, the product $2Z_{ik}Z_{nk} - 1$ is equal to 1 if Z_n and Z_i belong to the class k and is equal to -1 otherwise. Another way to write the prior probability is by writing the log prior label probability:

$$\begin{aligned}\log p(Z|\theta_Z) &= \sum_{n=1}^{N}\sum_{k=1}^{K}\log p(Z_{nk})\\ &= \left(\sum_{\mathcal{E}(n,m)}\sum_{k=1}^{K} -\alpha(2Z_{nk}Z_{mk}-1) + N\sum_{k=1}^{K}\log\pi_k\right) - \log D_Z\end{aligned}$$

where D_Z is a normalizing constant, $\mathcal{E}(n,m)$ is the set of edges connecting two neighboring nodes and $\theta_Z = \{\alpha\}\cup\{\pi_k\}$. The joint probability then writes as :

$$\begin{aligned}\log p(X,Z) &= \log p(X|Z) + \log p(Z)\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{nk}\log\mathcal{N}(x_n;\mu_k,\sigma_k^2) + \log(Z)\\ &= \sum_{\mathcal{E}(n,m)}\sum_{k=1}^{K} -\alpha(2Z_{nk}Z_{mk}-1) + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{nk}\left(\log\pi_k\mathcal{N}(x_n;\mu_k,\sigma_k^2)\right) - \log D_Z\end{aligned}$$

We approximate the posterior $p(Z|X)$ as the factorized function $q(Z) = \prod_{n=1}^{N} q_n(Z_n)$ where $q_n(Z_n)$ follows a multivariate Bernoulli distribution (*aka multinoulli distribution,* aka categorical distribution) parameterized by the vector $\hat{q}_n = [\hat{q}_{nk}] \in \mathbb{R}^K$ such that $q_n(e_k^K) = \hat{q}_{nk}$ and $\sum_{k=1}^{K}\hat{q}_{nk} = 1$. Therefore the approximation is fully determined by the matrix of size $N\times K$ of $\hat{q}_{nk}$. In this case, the posterior approximation of a variable Z_n is considered to be independent from the other variables Z_m which is a strong hypothesis.

To get a good approximation $q(Z)$, it is necessary to minimize the Kullback-Leibler divergence $D_{KL}(q||p(Z|X))$ which leads to the mean-field update of equation 8. Writing this update on discrete latent variables gives:

$$\begin{aligned}\log q_j(Z_j) &= \sum_{Z_1=e_1}^{e_K}\dots\sum_{Z_{j-1}=e_1}^{e_K}\sum_{Z_{j+1}=e_1}^{e_K}\dots\sum_{Z_N=e_1}^{e_K}\prod_{p\neq j} q_p(Z_p)\log p(X,Z_1,\dots,Z_N) + \text{cst}\\ &= \sum_{\tilde{Z}\in Z_{-j}}\prod_{p\neq j} q_p(\tilde{Z}_p)\log p(X,\tilde{Z}\cup Z_j) + \text{cst}\end{aligned}$$

where e_i (dropping the subscript K) is a one-hot encoded vector of size K where $e_i[i] = 1$ and $e_i[j\neq i] = 0$. Z_{-j} is the set of all possible latent variables of dimension $N-1$ which does not include Z_j. The cardinality of

Z_{-j} is therefore K^{N-1}. Thus $\tilde{Z}$ is a vector of random variables Z_p of size $N-1$ and $\tilde{Z} \cup Z_j$ is a latent vector of size N which is built by inserting Z_j into $\tilde{Z}$. In this equation, we are only interested in the functions of Z_j, the rest being store in a const which will be eliminated by the normalization process. Therefore, it is important to isolate in $\log p(X, \tilde{Z} \cup Z_j)$ the terms that depend on Z_j. We get :

$$\log p(X, \tilde{Z} \cup Z_j) = \sum_{n \in \mathcal{O}(j)} \sum_{k=1}^{K} -\alpha(2Z_{nk}Z_{jk} - 1) + \sum_{k=1}^{K} Z_{jk} \left(\log \pi_k \mathcal{N}(x_j; \mu_k, \sigma_k^2)\right) + \text{cst}$$

Furthermore we note that $\sum_{Z_p=e_1}^{e_K} q_p(Z_p) = 1$ because $\sum_{k=1}^{K} \hat{q}_{nk} = 1$. Therefore the sum over Z_{-j} can be restricted to the sum over node j and its neighbors in $\mathcal{O}(j)$. Furthermore, we can discard all sums over latent variables that are not involved in a term of $\log p(X, \tilde{Z} \cup Z_j)$. Finally, we get :

$$\log q_j(Z_{jl} = 1) = \hat{q}_{jl} = \sum_{n \in \mathcal{O}(j)} -\alpha(2\hat{q}_{nl} - 1) + \left(\log \pi_l \mathcal{N}(x_j; \mu_l, \sigma_l^2)\right) + \text{cst}$$

This leads to the following relationship after normalization:

$$\hat{q}_{jl} = \frac{\exp(\sum_{n \in \mathcal{O}(j)} -\alpha(2\hat{q}_{nl} - 1))\pi_l \mathcal{N}(x_j; \mu_l, \sigma_l^2)}{\sum_{k=1}^{K} \exp(\sum_{n \in \mathcal{O}(j)} -\alpha(2\hat{q}_{nk} - 1))\pi_k \mathcal{N}(x_j; \mu_k, \sigma_k^2)}$$