Gaussian Mixture Models

Andrea Casalino

December 24, 2019

1 Gaussian mixture distribution

A Gaussian Mixture Model (GMM) is a probability distribution function, often adopted to approximate an unknown distribution. GMM is basically a mixture (Section 1.2)of Gaussians (Section 1.1). In order to explain all the GMM properties, the expectation $\mathbb{E}$ operator must be introduced. The expectation $\mathbb{E}$ of a distribution w.r.t a function $g(X)$ defined over the same domain of f is equal to:

$$\mathbb{E}_f[g(X)] = \int_{\mathcal{X}} f(X)g(X)dx_{1,\cdots,n} \tag{1}$$

1.1 Gaussian distributions

The generic probability density function (aka pdf) f of a continuous random variable X, having $\mathcal{X} \in \mathbb{R}^n$ as domain, is a function defined as follows:

$$f : \mathcal{X} \to [0,1] \tag{2}$$

$$\int_{\mathcal{X}} f(X)dx_1 \cdot \cdots \cdot dx_n = 1 \tag{3}$$

Among all the possible distributions, the normal distributions are one of the most important. The mono-variate Normal distribution is a pdf defined as follows:

$$\phi_{(\mu,\Sigma)} \quad : \quad \mathcal{X} \to [0,1] \tag{4}$$

$$\phi_{(\mu,\Sigma)}(x) \quad = \quad \frac{1}{\sqrt{2\pi\Sigma}} exp\bigg(-\frac{1}{2}(x-\mu)\cdot\frac{1}{\Sigma}\cdot(x-\mu)\bigg) \tag{5}$$

where $X \subset \mathbb{R}$.
Σ and μ are the covariance and mean of the distribution and can be obtained as the following expectations:

$$\mu \quad = \quad \mathbb{E}_f[x] = \int_{\mathcal{X}} f(x)xdx \tag{6}$$

$$\Sigma \quad = \quad \mathbb{E}_f[(x-\mu)^2] = \int_{\mathcal{X}} f(x)(x-\mu)^2dx \tag{7}$$

The multivariate version is built generalizing the uni-variate one. Indeed, every multivariate Gaussian distribution can be seen, in a proper space, as a simple union of n independent Gaussian distributions. Consider n independent Gaussians having a null means and the variances equal to $\Sigma_{1,\cdots,n}$. They form a multivariate Gaussian distribution $\Phi_{(0,\Sigma)}$, with a covariance matrix Σ equal to the following diagonal matrix:

$$\Sigma = \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_n \end{bmatrix} \tag{8}$$

Since all the Gaussians in Φ are independent, the expression of the density function is obtained as the following product:

$$\Phi_{(0,\Sigma)}(Y) \quad = \quad \phi_{(0,\Sigma_1)}(y_1)\cdot\cdots\cdot\phi_{(0,\Sigma_n)}(y_n) \tag{9}$$

$$\Phi_{(0,\Sigma)}(Y) \quad = \quad \frac{1}{\sqrt{2\pi\Sigma_1}} exp\bigg(-\frac{1}{2}\frac{y_1^2}{\Sigma_1}\bigg) \cdot\cdots\cdot \frac{1}{\sqrt{2\pi\Sigma_n}} exp\bigg(-\frac{1}{2}\frac{y_n^2}{\Sigma_n}\bigg) \tag{10}$$

$$\Phi_{(0,\Sigma)}(Y) \quad = \quad \frac{1}{\sqrt{(2\pi)^n\cdot\Sigma_1\cdot\cdots\cdot\Sigma_n}} exp\bigg(-\frac{1}{2}\sum_{i=1}^{n}\frac{y_i^2}{\Sigma_i}\bigg) \tag{11}$$

$$\Phi_{(0,\Sigma)}(Y) \quad = \quad \frac{1}{\sqrt{(2\pi)^n\cdot|\Sigma|}} exp\bigg(-\frac{1}{2}Y^T \begin{bmatrix} \frac{1}{\Sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\Sigma_n} \end{bmatrix} Y\bigg) \tag{12}$$

$$\Phi_{(0,\Sigma)}(Y) \quad = \quad \frac{1}{\sqrt{(2\pi)^n\cdot|\Sigma|}} exp\bigg(-\frac{1}{2}Y^T\Sigma^{-1}Y\bigg) \tag{13}$$

Notice that:

$$\int_{\mathcal{Y}} \Phi_{(0,\Sigma)}(Y) dy_{1,\cdots,n} = 1 \tag{14}$$

is ensured due to the fact that the following single integrals are all equal to 1 (equation (5)):

$$\Phi_{(0,\Sigma)}(Y) = \left(\int_{\mathcal{Y}} \frac{1}{\sqrt{2\pi\Sigma_1}} exp\left(-\frac{1}{2}\frac{y_1^2}{\Sigma_1}\right) dy_1 \right) \cdot \cdots \cdot \left(\int_{\mathcal{Y}} \frac{1}{\sqrt{2\pi\Sigma_n}} exp\left(-\frac{1}{2}\frac{y_n^2}{\Sigma_n}\right) dy_1 \right) \tag{15}$$

In the domain of $\mathcal{Y}$, the multivariate Gaussian $\Phi(Y)$ is made of many independent Gaussians. Anyway, when performing a change of variables, the same distribution become a multivariate correlated Gaussian. Consider to set $Y = R \cdot X^{'}$, with R that is a rotation matrix, i.e. $R^{-1} = R^T$. X is in turn described by a multivariate Gaussian and his pdf is defined as follows:

$$\Phi_{(0,\Sigma)}(Y) = \frac{1}{\sqrt{(2\pi)^n \cdot |\Sigma|}} exp\left(-\frac{1}{2} X^{'T} R^T \Sigma^{-1} R X^{'} \right) \tag{16}$$

$$\Phi_{(0,\Sigma^{'})}(X) = \frac{1}{\sqrt{(2\pi)^n \cdot |\Sigma^{'}|}} exp\left(-\frac{1}{2} X^{'T} (\Sigma^{'})^{-1} X^{'} \right) \tag{17}$$

The above simplifications are due to the fact that:

$$\Sigma^{'} = R^T \Sigma R \tag{18}$$

$$(\Sigma^{'})^{-1} = R^{-1}\Sigma^{-1}R = R^T\Sigma^{-1}R \tag{19}$$

Moreover, since R is a rotation matrix, $\Sigma^{'}$ is symmetric positive definite (since $\Sigma_{1,\cdots,n} \geq 0$). Additionally $|\Sigma| = \left|\Sigma^{'}\right|$. A distribution with a non zero mean can be obtained by applying a simple transformation of the form $X = X^{'} + \mu$. X is the generic multivariate Gaussian, with a pdf equal to:

$$\Phi_{(\mu,\Sigma^{'})}(X) = \frac{1}{\sqrt{(2\pi)^n \cdot |\Sigma^{'}|}} exp\left(-\frac{1}{2} (X-\mu)^T (\Sigma^{'})^{-1} (X-\mu) \right) \tag{20}$$

1.2 Mixture models

Mixture models are in general a way to define a probability density function as the combination of a certain number of simpler ones.
It is possible to define a generic mixture model, by combining N probability densities $f_{1,\cdots,N}$ satisfying equation (3) and sharing the same domain $\mathcal{X}$. Indeed, considering N weights $\lambda_{1,\cdots,N}$, the density of the mixture f_{mix} is defined as follows:

$$f_{mix}(x) = \sum_{i=1}^{N} \lambda_i f_i(x) \tag{21}$$

To ensure that f_{mix} is in turn a valid probability density function satisfying equation 3, it is necessary to impose that the combination expressed in equation (21) should be convex, meaning that:

$$\sum_{i=1}^{N} \lambda_i = 1 \tag{22}$$

In fact, when the above specification holds, it is true that:

$$\begin{aligned} \int_X f_{mix}(x)dx &= \int_X [\lambda_1 f_1(x) + \cdots + \lambda_N f_N(x)]dx \\ &= \int_X \lambda_1 f_1(x)dx + \cdots + \int_X \lambda_N f_N(x)dx \\ &= \lambda_1 \int_X f_1(x)dx + \cdots + \lambda_N \int_X f_N(x)dx \\ \int_X f_{mix}(x)dx &= \lambda_1 + \cdots + \lambda_N = \sum_{i=1}^{N} \lambda_i = 1 \end{aligned} \tag{23}$$

The simplifications made in the above equation are justified by the fact that every function f_i is a probability distribution function satisfying equation (3):

$$\begin{cases} \int_X f_1(x)dx = 1 \\ \vdots \\ \int_X f_N(x)dx = 1 \end{cases} \tag{24}$$

It is worth noticing that no particular hypothesis were posed about the combining distributions $f_{1,\cdots,N}$. We need only to require they are valid probability density functions defined over the same domain. The above considerations are true also when considering multivariate distributions.

1.3 Gaussian Mixture models

Gaussian mixture models are particular mixture models combining N multivariate Gaussian distributions. The distribution of a GMM, f_{GMM}, is defined in this way:

$$f_{GMM}(x_{1,\cdots,n}) = \sum_{i=1}^{N} \lambda_i \Phi_{(\mu_i,\Sigma_i)}(x_{1,\cdots,n}) \tag{25}$$

where in the above equation Σ_i and μ_i are the variance and mean of the i^{th} Gaussian distribution in the mixture. GMM can be also adopted for approximating complex unknown distributions. Indeed, considering the proper number of clusters, any kind of probability distribution can be approximated by a GMM.
As an example, consider to approximate the uniform distribution $U(0,1)$ with a GMM, f_{approx}, made of n_{mix} components defined as follows:

$$\begin{aligned} f_{approx}(x|n_{mix}) &= \sum_{i=1}^{n_{mix}} \Phi_{(\mu_i,\Sigma_i)} \\ \Sigma_i &= \left(\frac{1}{n_{mix}}\right)^2 \quad \mu_i = \frac{1}{n_{mix}}\left(i - 0.5\right) \end{aligned} \tag{26}$$

In Figure 1, the density of $U(0,1)$ is compared with the one of f_{approx}, varying the number of approximating clusters. As can be visually appreciated, the approximation error decreases when increasing the number of clusters in the model. This phenomenon is not always verified when considering other kind of distributions.

1.4 Learning of GMM using Expectation maximization

The values of the weights $\lambda_{1,\cdots,N}$, as well the parameters characterizing every single Φ_i in the mixture, i.e. the covariances and the means, are tuned through a learning process which considers as training set $\langle X^1, \cdots, X^M \rangle$, *i.e.* M realizations of f_{mix}. Such learning process is performed using the expectation maximization algorithm. EM, is essentially an iterative algorithm that starts from an initial guess for the parameters of the clusters $\theta = \{\cdots \lambda_i, \Sigma_i, \mu_i, \cdots\}$, and then adjusts the model values until the convergence to a maximum for the likelihood $L(\theta|X)$.
EM is considered as an unsupervised algorithm, since only the number of clusters must be specified when performing learning, i.e.how many clusters consider for the mixture, omitting the labels [1] of the elements in the training set.
The expectation-maximization algorithm is in general used to learn the optimal parameter θ of a model having some observed variables X and also some latent ones Z. Z are variables whose values are hidden, but are linked in a probabilistic way to those observed, *i.e.* X. Learning is done according to a training set $\langle X^1, \cdots, X^M \rangle$, made of realizations of X: the corresponding values for $Z^{1,\cdots,M}$ are not known. EM is an iterative algorithm, which starts from an initial guess θ_0 and iteratively improves it. At every iteration, an Expectation and a Maximization are preformed, explaining the name of the algorithm. The Expectation step is performed for taking into account the expectation of the likelihood w.r.t. Z, in order to maximise the likelihood of the training set, no matter the values for Z, which are, in a certain sense, eliminated.
As usually done, learning aims at maximizing a likelihood function involving the training set. In this case, we would like to find those θ maximising $L(X|\theta)$ [2]. $L(X|\theta)$ can be computed considering how the joint conditioned distribution of X, Z is factorizable:

$$\begin{aligned} \mathbb{P}(X, Z|\theta) &= \mathbb{P}(Z|X,\theta)\mathbb{P}(X|\theta) \\ \mathbb{P}(X|\theta) &= \frac{\mathbb{P}(X, Z|\theta)}{\mathbb{P}(Z|X,\theta)} \end{aligned} \tag{27}$$

[1] The Gaussian in the mixture that produced that sample
[2] For the moment assume to have a single sample X in the training set

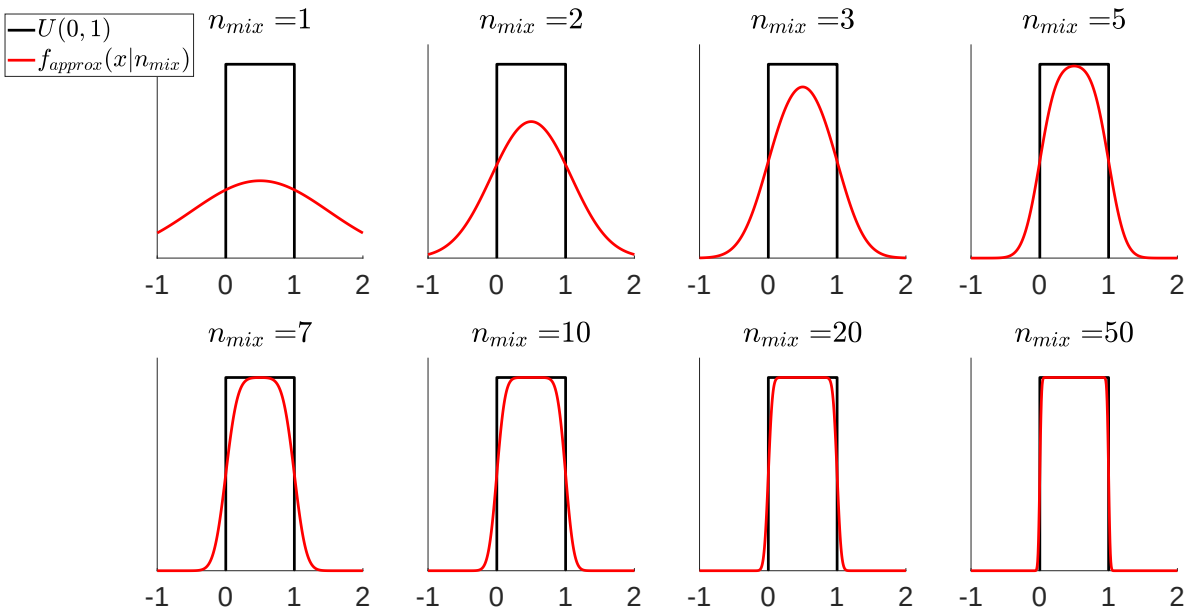

Figure 1: The uniform density between 0 and 1 is compared with an approximating mixture, varying the number of components.

Passing to the logarithms we obtain:

$$log(\mathbb{P}(X|\theta)) = log(\mathbb{P}(X,Z|\theta)) - log(\mathbb{P}(Z|X,\theta)) \tag{28}$$

We are now in position to describe the Expectation step of EM algorithm. Right hand side of equation (28) is a function of Z, which is unfortunately unknown. For this reason, we want to marginalize Z, by passing to the expectations w.r.t to density $\mathbb{P}(Z|X,\theta_k)$, where θ_k are the values of the parameter at step k:

$$\begin{aligned}
\sum_Z \mathbb{P}(Z|X,\theta_k) log(\mathbb{P}(X|\theta)) &= \sum_Z \mathbb{P}(Z|X,\theta_k) log(\mathbb{P}(X,Z|\theta)) + \cdots \\
&- \sum_Z \mathbb{P}(Z|X,\theta_k) log(\mathbb{P}(Z|X,\theta)) \\
log(\mathbb{P}(X|\theta)) \sum_Z \mathbb{P}(Z|X,\theta_k) &= \sum_Z \mathbb{P}(Z|X,\theta_k) log(\mathbb{P}(X,Z|\theta)) + \cdots \\
&- \sum_Z \mathbb{P}(Z|X,\theta_k) log(\mathbb{P}(Z|X,\theta))
\end{aligned} \tag{29}$$

Setting:

$$\begin{aligned}
Q(\theta|\theta_k) &= \sum_Z \mathbb{P}(Z|X,\theta_k) log(\mathbb{P}(X,Z|\theta)) \\
H(\theta_k|\theta_k) &= -\sum_Z \mathbb{P}(Z|X,\theta_k) log(\mathbb{P}(Z|X,\theta))
\end{aligned} \tag{30}$$

leads to [3]

$$log(\mathbb{P}(X|\theta)) = Q(\theta|\theta_k) + H(\theta|\theta_k) \tag{31}$$

Considering the difference $log(\mathbb{P}(X|\theta)) - log(\mathbb{P}(X|\theta_k))$ and equation (31) leads to:

$$log(\mathbb{P}(X|\theta) - log(\mathbb{P}(X|\theta_k) = Q(\theta|\theta_k) - Q(\theta_k|\theta_k) + H(\theta|\theta_k) - H(\theta_k|\theta_k) \tag{32}$$

[3] Considering that $\sum_Z \mathbb{P}(Z|X,\theta_k) = 1$.

At this point we can apply the Gibbs inequality, prescribing that in case of two distributions $f_{1,2}$ defined over the same domain applies what follows:

$$-\sum_x f_1(x)log(f_1(x)) \leq -\sum_x f_1(x)log(f_2(x)) \tag{33}$$

Setting $f_1 = \mathbb{P}(Z|X,\theta_k)$ and $f_2 = \mathbb{P}(Z|X,\theta)$ the inequalities in equation (33) allows us to state that:

$$H(\theta|\theta_k) - H(\theta_k|\theta_k) \geq 0 \tag{34}$$

and consequently that:

$$log(\mathbb{P}(X|\theta)) - log(\mathbb{P}(X|\theta_k)) \geq Q(\theta|\theta_k) - Q(\theta_k|\theta_k) \tag{35}$$

For this reason, the Maximization step of the algorithm computes θ_{k+1} in order to increase Q, which leads indirectly (equation (35)) to an increase of the quantity of interest, *i.e.* $log(\mathbb{P}(X|\theta))$. To be more precise, θ_{k+1} is computed as follows:

$$\theta_{k+1} = argmax_\theta Q(\theta|\theta_k) \tag{36}$$

It is not difficult to prove that function Q, when considering a training set made of a certain number of independent samples, is a summation of terms:

$$Q(\theta|\theta_k) = \sum_i Q_i(\theta|\theta_k) = \sum_i \sum_Z \mathbb{P}(Z|X^i,\theta_k) log(\mathbb{P}(X^i,Z|\theta)) \tag{37}$$

1.4.1 Learning of Gaussian Mixture Models

In case of GMM, the latent variables Z are the labels specifying which cluster produced every sample X^i. Let be $\gamma_j^i = \mathbb{P}(X^i \in Cluster_j)$ (see Section **??**). γ_j^i is a function of the model parameters and therefore varies along the iterations of the EM algorithm, *i.e.* γ_{jk}^i. Let n_{jk} be the sum of γ over samples in the training set, *i.e.* $n_{jk} = \sum_i \gamma_{jk}^i$. When considering GMM, it is true what follows [4]:

$$\mathbb{P}(Z = j|X^i,\theta_k) = \gamma_{jk}^i \tag{38}$$

$$\begin{aligned} \mathbb{P}(X^i, Z = j|\theta) &= \mathbb{P}(X^i, |Z = j,\theta)\mathbb{P}(Z = j|\theta) \\ &= \lambda_j \left(\sqrt{2\pi\,|\Sigma_j|^n}\right)^{-1} exp\Big(-0.5(X^i-\mu_j)^T\Sigma_j^{-1}(X^i-\mu_j)\Big) \end{aligned} \tag{39}$$

In the second equation we exploited the fact that in a mixture model, weights λ expresses the a priori probability of a sample being generated by the corresponding cluster. Considering equations (38) (39), function Q in case of GMM is computable as follows:

$$\begin{aligned} Q(\theta|\theta_k) &= \sum_i \sum_j^N \gamma_{jk}^i \Big(log(\lambda_j) - 0.5log(|\Sigma_j|) - 0.5(X^i-\mu_j)^T\Sigma_j^{-1}(X^i-\mu_j)\Big) \\ &= \sum_j^N n_{jk}\Big(log(\lambda_j) - 0.5log(|\Sigma_j|)\Big) + \cdots \\ \cdots &+ \sum_i \sum_j^N -0.5(X^i-\mu_j)^T\Sigma_j^{-1}(X^i-\mu_j) \end{aligned} \tag{40}$$

The maximization step described before, has to solve a constrained maximization problem, considering Q as objective function and $\sum_j \lambda_j = 1$ as a constraint, since GMM are mixture models (equation (**??**)). Since we deal with an equality constraints, we consider the Lagrangian function $Q^{'}$:

$$Q^{'}(\theta) = Q(\theta|\theta_k) + \xi(\sum_j \lambda_j - 1) \tag{41}$$

[4] The same notation of Section **??** was assumed

where ξ is the lagrangian multiplier. The maximum of $Q^{'}$ is obtained by finding those combinations of values for which the gradient is null.

Imposing the gradient of $Q^{'}$ w.r.t. the generic weight λ_j equal to 0 leads to:

$$\begin{aligned} \frac{\partial}{\partial \lambda_j} &= \frac{n_{jk}}{\lambda_j} + \xi = 0 \\ \lambda_j &= -\frac{n_{jk}}{\xi} \end{aligned} \tag{42}$$

In order to let the constraint $\sum_j \lambda_j = 1$ be satisfied, we have to prescribe that:

$$\sum_j \lambda_j = \sum_j \frac{n_{jk}}{\xi} \Rightarrow \xi = -\sum_j n_{jk} \tag{43}$$

Therefore, substituting into equation (42) leads to:

$$\lambda_{j\ k+1} = \frac{n_{jk}}{\sum_{j=1}^{N} n_{jk}} \tag{44}$$

The gradient of $Q^{'}$ w.r.t. the generic weight μ_j is equal to:

$$\begin{aligned} \frac{\partial}{\partial \mu_j} &= \sum_i \gamma_{jk}^i \frac{\partial}{\partial \mu_j} \Big(-0.5 (X^i - \mu_j)^T \Sigma_j^{-1} (X^i - \mu_j) \Big) \\ &= \sum_i \gamma_{jk}^i \Sigma_j^{-1} (\mu_j - X^i) \\ &= \Sigma_j^{-1} \Big(n_{jk} \mu_j - \sum_i \gamma_{jk}^i X^i \Big) \end{aligned} \tag{45}$$

Imposing $n_{jk}\mu_j - \sum_i \gamma_{jk}^i X^i = 0$, ensures the entire gradient is null. Therefore, it applies what follows:

$$\mu_{j\ k+1} = \frac{\sum_i \gamma_{jk}^i X^i}{n_{jk}} \tag{46}$$

Finally, the gradient w.r.t. to Σ_k is equal to (here the properties reported in [**?**] are exploited):

$$\begin{aligned} \frac{\partial}{\partial \Sigma_j} &= -0.5 n_{jk} \Sigma_j^{-1} + 0.5 \sum_i \gamma_{jk}^i \Sigma_j^{-1} (X^i - \mu_j)^T (X^i - \mu_j) \Sigma_j^{-1} \\ &= 0.5 \Sigma_j^{-1} \Big(-n_{jk} I + \big(\sum_i \gamma_{jk}^i (X^i - \mu_j)^T (X^i - \mu_j) \big) \Sigma_j^{-1} \Big) \end{aligned} \tag{47}$$

Imposing $-n_{jk} I + \big(\sum_i \gamma_{jk}^i (X^i - \mu_j)^T (X^i - \mu_j) \big) \Sigma_j^{-1} = 0$ leads to:

$$\Sigma_{j\ k+1} = \frac{\sum_i \gamma_{jk}^i (X^i - \mu_j)(X^i - \mu_j)^T}{n_{jk}} \tag{48}$$

At every step k of EM, every γ_{jk}^i is recomputed and equations (44), (46) and (48) are applied for updating the parameters of the mixture. After all the simplifications it turns out that the updating equations of EM, in case of training a GMM, have an heuristic interpretation. Indeed, equation (44), the new value of the weight of a cluster simply consider the importance if that cluster w.r.t. to all the others, *i.e.* the summation of the probabilities that samples in the training set belongs to that cluster.
The new means of the clusters are computed as a weighted mean, equation (46), which gives more importance to those samples having an high probability to belongs to the cluster for which the mean is re evaluated. A similar consideration holds for the covariance of clusters, equation (48).

1.5 Classification

The functions characterizing the mixture, equation (26), can be interpreted as clusters. In such cases, the values of weights $\lambda_{1,\cdots,N}$ are priors for the probability that a certain value $\overline{X}$ is part of a certain cluster. The classification of a value $\overline{X}$, according to the Bayes formula, can be done as follows:

$$\begin{aligned}\mathbb{P}(\overline{x} \in Cluster_i) = \mathbb{P}(Cluster_i|\overline{X}) &\propto L(\overline{X}|Cluster_i)\mathbb{P}_{prior}(Cluster_i) \\ &\propto f_i(\overline{X})\lambda_i \\ &= \frac{f_i(\overline{X})\lambda_i}{\sum_{j=1}^{N} f_j(\overline{X})\lambda_j} = \frac{\Phi_{(\mu_i,\Sigma_i)}(\overline{X})\lambda_i}{\sum_{j=1}^{N} \Phi_{(\mu_j,\Sigma_j)}(\overline{X})\lambda_j}\end{aligned} \tag{49}$$

1.6 Drawing samples from a GMM

From what discussed in Section 1.1 (which leads to obtain equation (5)), it is always possible to find a linear transformation to obtain a multivariate Gaussian with a desired mean μ and covariance Ψ, starting from an isotropic Gaussian $\Phi_{(0,I)}(Y)$, with I that is the identity matrix. Consider the change of variable $X = A \cdot Y + T$. If X is multivariate Gaussian having a zero mean and a covariance equal to the identity matrix, then Y is a multivariate Gaussian with a covariance $\Sigma = A \cdot A$ and a mean $\mu = T$. The point here is how to compute A in order to obtain that desired Σ. Since Σ must be a symmetric positive definite matrix, it can be factorized as follows (due to the spectral theorem):

$$\Sigma = U \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix} U^T \tag{50}$$

$$= \left(U \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}\right) \cdot \left(\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} U^T\right) \tag{51}$$

$$= A \cdot A^T \tag{52}$$

where U are the eigenvectors of Σ and $\sigma^2_{1,\cdots,n}$ their eigenvalues. From the analysis equation (52), it is evident that:

$$A = U \cdot \left(\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}\right. \tag{53}$$

In order to sample from a Gaussian distribution having Σ and μ as covariance and mean, it is sufficient to draw samples from a standard distribution having I and 0 as covariance and mean, also called isotropic Gaussian, i.e. take samples from n independent uni-variate Gaussians having a zero mean and a unitary variance and then transform such samples. Suppose to have sampled M values $\langle Y^1, \cdots, Y^M \rangle$ from the isotropic Gaussian. Applying the following distribution to each sample:

$$X^1 = U \cdot \left(\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}\right. Y^1 + \mu \tag{54}$$

M samples distributed as the desired Gaussian multivariate are obtained.

The above procedure is able to produce samples from a single Gaussian. When dealing with a GMM, a similar procedure can be followed. For each sample to draw, the cluster to consider for producing that sample must be firstly determined. The i^{th} cluster is selected with a probability equal to the corresponding weight λ_i [5] and then, a sample is draw from the i^{th} Gaussian by following the strategy described above.

1.7 Divergence of Kullback Leibler of GMMs

The divergence of two general pdfs f_1 and f_2 is a measure of how much that two distributions differ from each other. Indeed, in case that two distributions are defined over the same domain $\mathcal{X}$, the divergence of f_2 w.r.t f_1 is equal to:

$$\mathbb{E}_{f_1}\left[log\left(\frac{f_1(X)}{f_2(X)}\right)\right] = \int_{\mathcal{X}} f_1(X) log\left(\frac{f_1(X)}{f_2(X)}\right) dX_{1,\cdots,n} \tag{55}$$

[5] The index of the cluster to consider is sampled from a discrete distribution having $\lambda_{1,\cdots,N}$

It is important to remark that the divergence is not symmetric, i.e. $D_{KL}(f_1||f_2) \neq D_{KL}(f_2||f_1)$. Such measure can be used to compare tow high dimensional multivariate distributions.
In case of GMM, the Kullback-Leibler divergence is not computable in a closed form. On the opposite, a Monte Carlo approach can be followed. Indeed, M samples $\langle X^1, \cdots, X^M \rangle$ from f_1 can be drawn (as described in Section 1.6) and the following mean is assumed as an approximation of the expectation in equation (55):

$$D_{KL}(f_1||f_2) \cong \frac{1}{M} \sum_{i=1}^{M} log\left(\frac{f_1(X^i)}{f_2(X^i)}\right) \tag{56}$$

Other approaches exist, providing an upper and lower bound for the real value of the divergence.