Chapter 3

Mixture models

3.1 Mixture distributions and clustering

In Sections 1.3 and 2.4 we discussed Logistic Regression and Naive Bayes, two different approaches adopted in *classification*. Given a training set of observations $x_1, \dots, x_N$ we also observed $y_1, \dots, y_N$, labelling their membership to a class, and used them for the inference of the model parameters. This is why we call it *supervised* learning (or classification). In this chapter, we still assume that the data are grouped into K classes (or clusters), but the membership of the i-th observation to a class is now labelled by a *latent* random variable Z_i, not observed and that we aim (in some sense) to infer from the data. This approach to *unsupervised learning* is know as model based *clustering* and basically relies on mixture distributions. The features $\mathbf{x} := (x_1, \dots, x_N)$ are assumed to be realisations of i.i.d. random variables whose probability density function[1]

$$p(x_i|\Theta) = \sum_{k=1}^{K} \pi_k p_k(x_i|\theta_k), \tag{3.1}$$

where $\pi_1, \dots, \pi_k$ are the mixture proportions ($\pi_k \in (0,1)$ for all k and $\sum_{k=1}^{K} \pi_k = 1$), $p_k(\cdot|\theta_k)$ is the pdf of the k-th mixining component and θ_k is the corresponding parameter. Finally, we denote by Θ the set of all the model parameters: mixture proportions and mixing parameters. Apart from being

[1] In order to ease the exposition and without loss of generality, we assume here that x_i is a continuous random vector. However, everything we state remains true for discrete random vectors.

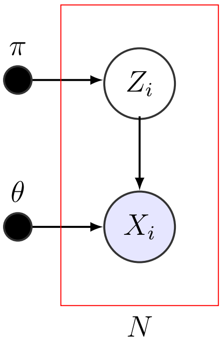

Figure 3.1: Graphical representation of a mixture model (i-th observation).

a well defined pdf (why?), $p(\cdot|\Theta)$ can be equally described by introducing a set of i.i.d. latent variables $\mathbf{Z} := (Z_1, \ldots, Z_N)$ such that $\mathbb{P}(Z_i = k) = \pi_k$ and

$$p(x_i|z_i = k, \theta) = p_k(x_i|\theta_k), \tag{3.2}$$

with $\theta := (\theta_1, \ldots, \theta_K)$ and $X_1, \ldots, X_N$ assumed to be *conditionally* independent given $Z_1, \ldots, Z_N$, although clearly no longer identically distributed. So, given the "prior" mass function $p(\mathbf{z}) = \prod_{i=1}^N p(z_i) := \prod_{i=1}^N \pi_{z_i}$ the joint distribution of $(\mathbf{X}, \mathbf{Z})$ is

$$p(\mathbf{x}, \mathbf{z}|\Theta) = p(\mathbf{x}|\mathbf{z}, \theta)p(\mathbf{z}|\pi) = \prod_{i=1}^N \prod_{k=1}^K (p_k(x_i|\theta_k)\pi_k)^{z_{ik}}, \tag{3.3}$$

where $\pi := (\pi_1, \ldots, \pi_K)$. The graphical model corresponding to the above joint density, also called *complete data* likelihood, is in Figure 3.1. Now, when marginalizing with respect to $\mathbf{Z}$, i.e. taking the sum on both sides of the above equation with respect to all possible values that $\mathbf{z}$ could take we obtain

$$p(\mathbf{x}|\Theta) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\Theta) = \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k p_k(x_i|\theta_k) \right), \tag{3.4}$$

where, in the parenthesis on the right hand side we recognize $p_k(x_i|\theta)$ in Eq. (3.1). Now, from a statistical perspective, two things are or particular interest for us: i) the ML estimate of Θ and ii) the posterior probability $p(z_i|x_i, \hat{\Theta}_{ML})$. The latter in particular allows one to answer the question: which is the most likely group for the i-th observation? I.e. to perform clustering. By the way, we need to computing $\hat{\Theta}_{ML}$. The main problem we have is that when computing the logarithm of the *observed data* log-likelihood, on the right hand side of Eq. (3.4), we come up with

$$l(\Theta) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k p_k(x_i|\theta_k) \right).$$

The presence of the sum inside the logarithm in the log-likelihood above makes that quantity not tractable analytically. We might proceed by looking for numerical solutions to the equation $\nabla_\Theta \left(l(\Theta) + \lambda(\sum_{k=1}^{K} \pi_k = 1) \right) = 0$ via gradient descent (i.e. ascent), as we did for the Logistic regression but there's a more efficient solution: the EM algorithm. In order illustrate how to estimate the model parameters as well as the most likely posterior cluster memberships via the EM algorithm, we consider henceforth and wlog the (maybe) most popular mixture model: the Gaussian mixture model.

Gaussian mixture model

We henceforth assume that x_i is a feature vector in $\mathbb{R}^D$, with $D \geq 1$. The following conditional density for x_i

$$p(x_i|z_i = k, \theta_k) := \phi(x_i; \mu_k, \Sigma_k), \tag{3.5}$$

where

$$\phi(x; \mu, \Sigma) := \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

denotes here the pdf of a multivariate Gaussian distributed random vector with mean $\mu \in \mathbb{R}^D$ and covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$. Thus, $\theta_k := (\mu_k, \Sigma_k)$, for all k in $\{1, \ldots, K\}$. In Figure 3.2 we see a simulated dataset of $N = 100$ observation in $\mathbb{R}^2$ spread into $K = 2$ clusters, each corresponding to a Gaussian, isotropic distribution. An estimated density of the mixture of two univariate Gaussian distributions can be seen in Figure 3.3.

Although the observed data log-likelihood is not tractable, in the light of the next section, it is useful to take a look at the log-likelihood of the complete data (logarithm of Eq. (3.3)) in this scenario:

$$l_c(\Theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} z_{ik} \left[-\frac{1}{2} \log(|\Sigma_k|) - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \pi_k + C \right]. \tag{3.6}$$

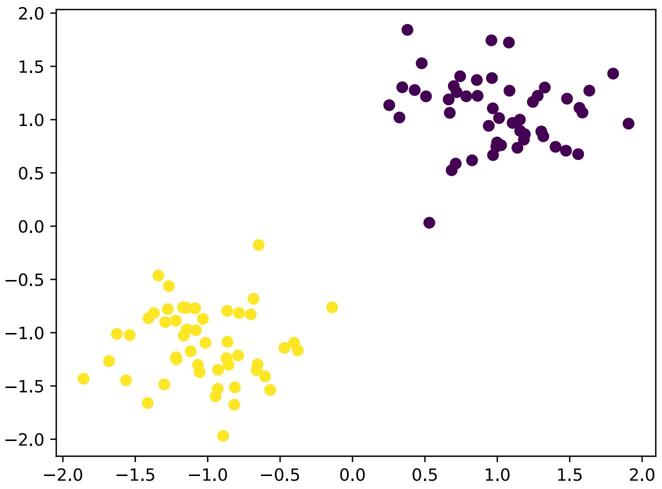

Figure 3.2: Two clusters corresponding to isotropic bivariate Gaussian distributions, sharing the same standard deviation of $\sigma_1 = \sigma_2 = 0.4$. The yellow cluster is centred in $\mu_1 = (-1, -1)$, the violet cluster in $\mu_2 = (1, 1)$. The mixing proportions are symmetric: $\pi = (\frac{1}{2}, \frac{1}{2})$.

Since the log-likelihood factorizes over k, *if we knew* $\mathbf{z}$, we could proceed as we did for the Naive Bayes classifier, i.e. class by class. It can be shown that the ML estimates in this case are:

$$\begin{aligned} \hat{\pi}_k &:= \frac{N_k}{N} \\ \hat{\mu}_k &:= \frac{1}{N_k} \sum_{i=1}^{N} z_{ik} x_i \\ \hat{\Sigma}_k &:= \frac{1}{N_k} \sum_{i=1}^{N} z_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T, \end{aligned} \tag{3.7}$$

where $N_k := \sum_{i=1}^{N} z_{ik}$ counts the number of observations in cluster k. The first two equations are straightforward to prove (**exercise**), the last one requires a bit more of infinitesimal matrix calculus. The interested reader can refer to (Bishop and Nasrabadi, 2006, Section 2.3.4 and Appendix C). However, it

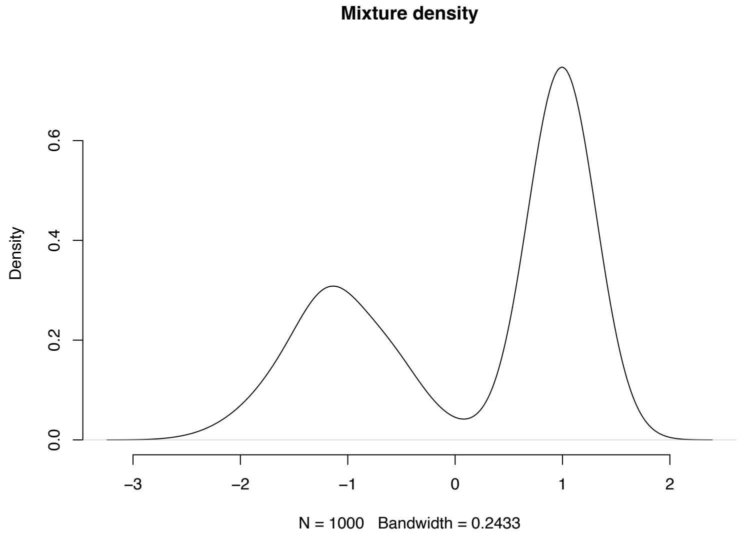

Figure 3.3: Estimated mixture density of two univariate Gaussian distributions $\mathcal{N}(-1, 0.71)$ and $\mathcal{N}(1, 0.41)$ with mixture proportions $(0.4, 0.6)$.

is important to remark how these estimators make perfectly sense! The ML estimate of π_k is given by the proportion of points in cluster k, and the ML estimates of μ_k and Σ_k are the empirical mean and covariance in the k-th cluster.

3.2 The EM algorithm

There is a main intuition allowing us to numerically compute the ML estimates of our GMM and massively used in variational inference (next chapter): although the log-likelihood of the observed data in not tractable we can replace it by a tractable *lower bound* and optimize it in place. In more details

$$
\begin{aligned}
l(\Theta) := \log p(\mathbf{x}|\Theta) &= \log \left(\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\Theta) \right) \\
&= \log \left(\sum_{\mathbf{z}} \frac{p(\mathbf{x}, \mathbf{z}|\Theta)}{q(\mathbf{z})} q(\mathbf{z}) \right) \\
&= \log \left(\mathbb{E}_{\mathbf{Z}\sim q} \left[\frac{p(\mathbf{x}, \mathbf{Z}|\Theta)}{q(\mathbf{Z}))} \right] \right) \\
&\geq \mathbb{E}_{\mathbf{Z}\sim q} \left[\log \left(\frac{p(\mathbf{x}, \mathbf{Z}|\Theta)}{q(\mathbf{Z}))} \right) \right] =: \mathcal{L}(q, \Theta).
\end{aligned}
\tag{3.8}
$$

In the sequence of equations above: i) "$\sum_{\mathbf{z}}$" denotes the sum taken over the support of $\mathbf{z}$ according to the prior distribution $p(\cdot)$ (here $\{1, \ldots, K\}^N$), ii) $q(\cdot)$ denotes *any* other distribution sharing the same support with $p(\cdot)$ and from the third line on the capital letter $\mathbf{Z}$ is used to emphasize that $\mathbf{z}$ is seen as a random vector and no longer as a realisation iii) the inequality comes from **Jensen**'s inequality[2], since the logarithm is a concave function and iv) $\mathcal{L}(q, \Theta)$ is introduced to denote the lower bound of the log-likelihood of the observed data. Henceforth, in order to keep the notation uncluttered, we use $\mathbb{E}_q := \mathbb{E}_{\mathbf{Z}\sim q}$. First, we observe that the lower bound involves an expectation w.r.t. $\mathbf{Z}$ of Eq. (3.6), which is linear in z_{ik} and thus analytically tractable. Second, the difference $l(\Theta) - \mathcal{L}(q, \Theta)$ can be quantified precisely thanks to

Proposition 2. *Given a fixed set of parameters Θ and a probability mass function $q(\cdot)$ with the same support of $p(\cdot)$, the following equality holds*

$$
l(\Theta) = \mathcal{L}(q, \Theta) + \mathbb{E}_q \left[\log \left(\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{x}, \Theta)} \right) \right], \tag{3.9}
$$

*where $p(\mathbf{z}|\mathbf{x}, \Theta)$ is the **posterior** distribution of $\mathbf{Z}$ given the data $\mathbf{x}$ and the model parameters Θ.*

Proof. Exercise. □

Henceforth, the abbreviated notation $p_{\mathbf{z}}(\cdot)$ will be used to denote the posterior distribution of $\mathbf{Z}$ given the data and the model parameters. The second term on the right hand side of Eq. (3.9) is the **Kullback-Leibler**

[2] https://en.wikipedia.org/wiki/Jensen%27s_inequality

divergence between $q(\cdot)$ and $p_{\mathbf{z}}(\cdot)$, denoted as $KL(q||p_{\mathbf{z}})$. Although it can be shown that the KL divergence is non negative and satisfies the triangular inequality, it is *not* symmetric and this is why it is not a distance. The main intrerst of Eq. (3.9) is that it makes it clear that whether $q = p_{\mathbf{z}}$, then $KL(q||p_{\mathbf{z}})$ vanishes and the lower bound equals the observed log-likelihood

$$\log p(\mathbf{x}|\Theta) = \mathbb{E}_{p_{\mathbf{z}}}\left[\log\left(\frac{p(\mathbf{x},\mathbf{Z}|\Theta)}{p_{\mathbf{z}}(\mathbf{Z}))}\right)\right].$$

Hence, *if* the posterior probability is tractable and the above expectation can be computed analytically the Expectation maximization (EM) algorithm consists of two steps

1. **Expectation.** For a fixed value of the model parameters, say Θ_c, one computes the lower bound

$$\mathcal{L}(p_{\mathbf{z}}, \Theta_c) = \mathbb{E}_{p_{\mathbf{z}}}\left[\log\left(\frac{p(\mathbf{x},\mathbf{Z}|\Theta_c)}{p_{\mathbf{z}}(\mathbf{Z})}\right)\right]$$

   which equals the observed log-likelihood at Θ_c. It is important to notice that the posterior probability used to compute the expectation is $p_{\mathbf{z}}(\cdot) = p(\cdot|\mathbf{x},\Theta_c)$, depending on the current value of the model parameters.

2. **Maximization.** The following (often tractable) maximization problem

$$\Theta_n := \arg\max_{\Theta} \mathcal{L}(p_{\mathbf{z}}, \Theta)$$

   is solved. Notice that here the posterior is considered as given since the expectation has already been computed in the E-step. In contrast, the lower bound is seen as a function of the models parameters Θ and Θ_c is updated to Θ_n.

Now, in force of Eq. 3.9, after the M-step the following inequalities hold

$$\begin{aligned}
l(\Theta_n) :&= \mathcal{L}(p_{\mathbf{z}}, \Theta_n) + \mathbb{E}_{p_{\mathbf{z}}}\left[\log\left(\frac{p_{\mathbf{z}}(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{x},\Theta_n)}\right)\right] \\
&\geq \mathcal{L}(p_{\mathbf{z}}, \Theta_c) + \mathbb{E}_{p_{\mathbf{z}}}\left[\log\left(\frac{p_{\mathbf{z}}(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{x},\Theta_c)}\right)\right] \\
&= \mathcal{L}(p_{\mathbf{z}}, \Theta_c) \\
&= l(\Theta_c),
\end{aligned}$$

where we stress once more that $p_{\mathbf{z}}(\cdot) = p(\cdot|\mathbf{x}, \Theta_c)$. Moreover the middle inequality turn into an equality iff Θ_c already is a stationary point of $\mathcal{L}(p_{\mathbf{z}}, \cdot)$ in which case $\Theta_n = \Theta_c$. In words: any step of the EM algorithm is guaranteed to increase the observed log-likelihood until a stationary point is reached. For a more detailed discussion about the convergence properties of the EM algorithm, the reader is referred to Wu (1983); Xu and Jordan (1996). Here we just to point out that, since the log-likelihood in mixture modes (Eq 3.1) is not concave, the stationary point reached via the EM algorithm might not be a global optimum (i.e. not the actual ML estimate).

Let us be back to the Gaussian mixture model (GMM) introduced in the previous section. What does the E and M steps correspond to in that case? First of all we notice that we are in a good shape since the posterior distribution of $\mathbf{z}$ given the data $\mathbf{x}$ and the model parameters is tractable:

$$\begin{aligned} p(\mathbf{z}|\mathbf{x}, \Theta) &= \frac{\prod_{i=1}^{N} \prod_{k=1}^{K} \left[\pi_k \phi(x_i; \mu_k, \Sigma_k)\right]^{z_{ik}}}{\prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Sigma_K)\right)} \\ &= \prod_{i=1}^{N} \left[\frac{\prod_{k=1}^{K} \left[\pi_k \phi(x_i; \mu_k, \Sigma_k)\right]^{z_{ik}}}{\sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Sigma_K)}\right] \\ &= \prod_{i=1}^{N} \frac{p(x_i, z_i|\Theta)}{p(x_i|\Theta)} \\ &=: \prod_{i=1}^{N} p(z_i|x_i, \Theta). \end{aligned}$$

So the posterior probability of $\mathbf{z}$ factorizes and

$$\tau_{ik} := \mathbb{P}\left(Z_{ik} = 1|x_i, \Theta\right) = \frac{\pi_k \phi\left(x_i; \mu_k, \Sigma_k\right)}{\sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Sigma_k)}. \tag{3.10}$$

In other terms Z_i's are independent categorical random vectors with posterior distribution of parameter $\tau_i := (\tau_{i1}, \dots, \tau_{iK})$, for all i. In particular the (posterior) marginal distribution of Z_{ik} is a Bernoulli of parameter τ_{ik}, for each k and $\mathbb{E}_{p_{\mathbf{z}}}[Z_{ik}] = \tau_{ik}$. From Eq. (3.6) and the definition of $\mathcal{L}$ (Eq. (3.8)) it follows that (**E-step**)

$$\mathcal{L}(p_{\mathbf{z}}, \Theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik}\left[-\frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k) + \log\pi_k + C\right] \\ - \sum_{i=1}^{N}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik} \tag{3.11}$$

where Θ denotes the current set of model parameters (the subscript c is no longer included to keep the notation uncluttered). Notice that this lower bound is essentially equivalent to the log-likelihood of the complete-data except for i) z_{ik} being replaced by τ_{ik} and ii) the last entropy term on the right hand side of the above equation. Next, we look at the above lower bound as a function of Θ and consider τ as being fixed. Clearly, the optimal updates of π_k, μ_k and Σ_k are (**M-step**):

$$\begin{aligned} \hat{\pi}_k &:= \frac{\tilde{N}_k}{N} \\ \hat{\mu}_k &:= \frac{1}{\tilde{N}_k}\sum_{i=1}^{N}\tau_{ik}x_i \\ \hat{\Sigma}_k &:= \frac{1}{\tilde{N}_k}\sum_{i=1}^{N}\tau_{ik}(x_i-\hat{u}_k)(x_i-\hat{\mu}_k)^T, \end{aligned} \tag{3.12}$$

where $\tilde{N}_k := \sum_{i=1}^{N}\tau_{ik}$. Take some time to compare these estimates with those in Eq. (3.7). The above two steps are iterated until the lower bound no longer increases (a stationary point is reached). The whole algorithm is summarized in Algorithm 2

3.3 Relation to K-means

In Gaussian mixture models, the final posterior probability mass function τ_i, estimated via the EM algorithm, is typically used to assign the i-th observation to a group (a.k.a. *clustering*). Another popular approach to data clustering is K-means, which minimizes the following objective function

$$J(r, \mu) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}||x_i - \mu_k||_2^2, \tag{3.13}$$

Algorithm 2 Pseudocode: EM - GMM

```
function Fit($\mathbf{x}$,$K$)
    $\tau \leftarrow$ Init($\mathbf{x}$, K, type)    ▷ type is “multiple random” or “k-means”
    First updates of $\Theta_c = \{\hat{\mu}_k, \hat{\pi}_k, \hat{\Sigma}_k\}_k$ via Eq (3.12)
    while $\mathcal{L}(p_{\mathbf{z}}, \Theta_c)$ increases do
        Update $\tau$ via Eq. (3.10)
        Compute the new $\mathcal{L}(\tau, \Theta_c)$    ▷ E-step
        Compute $\Theta_n$ via Eq. (3.12)    ▷ M-Step
        $\Theta_c = \Theta_n$
    end while
    return $(\tau, \Theta_c)$
end function
```

with $\mathbf{x} = (x_1, \dots, x_N)$ and $\mu = (\mu_1, \dots \mu_k)$ being the same as before, $r_i := (r_{i1}, \dots, r_{iK})$ being a binary vector whose k-th component is equal to one iff x_i is in cluster k, zero otherwise and $r = (r_1, \dots, r_N)$. First, notice that r here and $\mathbf{z}$ before have the same role: they label the membership of an observation to one and only one cluster (hard clustering). However, in GMM, $\mathbf{z}$ is seen as the outcome of a random vector $\mathbf{Z}$ whose posterior distribution we inspect, whereas in k-means r is a parameter we aim to estimate. That said, K-means clustering can be seen as a limit case of a Gaussian mixture model. Indeed, assume that Σ_k in Eq (3.5) is equal to ηI_D for all k with η being a small and positive hyper-parameter (i.e. fixed). The lower bound in Eq. (3.11) reduced to

$$\mathcal{L}(p_{\mathbf{z}}, \mu, \pi) = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \left[-\frac{D}{2} \log \eta - \frac{1}{2\eta} ||x_i - \mu_k||_2^2 + \log \pi_k + \log \tau_{ik} \right] + C,$$

where C includes the remaining constant terms. First of all we notice that maximizing this lower bound with respect to the model parameters (μ, π) is equivalent to maximize

$$\eta \mathcal{L}(p_{\mathbf{z},\mu,\pi}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \left[-\frac{D}{2} \eta \log \eta - \frac{1}{2} ||x_i - \mu_k||_2^2 + \eta \log(\pi_k \tau_{ik}) \right] + \eta C.$$

Then we observe that, since

$$\tau_{ik} = \frac{\pi_k \exp\left(-\frac{1}{2\eta}||x_i - \mu_k||_2^2\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\frac{1}{2\eta}||x_i - \mu_j||_2^2\right)},$$

as long as π_k is positive for all k (no empty cluster) we have

$$\lim_{\eta \to 0} \tau_{ik} = \begin{cases} 1 & \text{if} \quad k = \arg\min_{j \leq K} ||x_i - \mu_j||_2^2 \\ 0 & \text{otherwise} \end{cases}$$

and thus τ converges to r, for a given set of means μ. The last observation coupled with the fact that $\lim_{x \to 0}(x \log x) = 0$ allows us to conclude

$$\lim_{\eta \to 0} (\eta \mathcal{L}(p_z)\mu, \pi) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} ||x_i - \mu_k||_2^2$$

which is the negative cost of the k-means algorithm.

3.4 How many mixing components?

The unknown number of mixing components K was considered as given in our exposition. However, in real scenarios this assumption is by far too optimistic and we have to estimate K from the data. With respect to this point, model based approaches (such as GMM) have an advantage with respect to discriminative methods (such as K-means or spectral clustering): via the notion of *posterior* probability they allow one to estimate K. Let's see how. One target probability mass function, that one would like to maximize with respect to the pair $(\mathbf{z}, K)$ is

$$p(\mathbf{z}, K|\mathbf{x}) = \int p(\mathbf{z}, K|\mathbf{x}, \Theta)p(\Theta)d\Theta \tag{3.14}$$

where $\mathbf{z} := (z_1, \ldots, z_N)$, $\mathbf{x} := (x_1, \ldots, x_N)$ and the whole model parameters Θ are seen as random variables and integrated out[3]. Note that, from a full

[3] In passing we note the the term inside the integrand is proportional to the likelihood of the complete data and the reason why we need to integrate the model parameters out is that otherwise the higher K, the higher the dimension of Θ and so the higher is the likelihood.

Bayesian perspective, the number of clusters K is also viewed as a random variable in the above equation. Of course, this quantity in general *not* tractable. However, based on the Bayes rule

$$p(\mathbf{z}, K|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z}|K)p(K)}{p(\mathbf{x})}.$$

where $p(K)$ is a prior distribution over K and $p(\mathbf{x})$ is the marginal probability of $\mathbf{x}$ after integrating out everything! Since the denominator does not depend on $(\mathbf{z}, K)$, it holds that

$$\begin{aligned}\arg\max_{(\mathbf{z},K)} p(\mathbf{z}, K|\mathbf{x}) &= \arg\max_{(\mathbf{z},K)} \big(p(\mathbf{x}, \mathbf{z}|K)p(K)\ \big) \\ &= \arg\max_{K} \left(\arg\max_{\mathbf{z}|K} \big(p(\mathbf{x}, \mathbf{z}|K)\ \big)p(K)\right).\end{aligned}$$

In order to simplify the exposition, we can assume that K is uniformly distributed ($p(K) \propto 1$) and the above equation then reduces to

$$\arg\max_{(\mathbf{z},K)} p(\mathbf{z}, K|\mathbf{x}) = \arg\max_{(\mathbf{z},K)} p(\mathbf{x}, \mathbf{z}|K). \tag{3.15}$$

Thus, computing the maximum posterior (MAP) estimates of Z and K reduces to maximize the **integrated** log-likelihood of the complete data:

$$p(\mathbf{x}, \mathbf{z}|K) = \int p(\mathbf{x}, \mathbf{z}|\Theta, K)p(\Theta)d\Theta$$

where the first term on the left inside the integral is the same we computed in Eq. (3.6). Now, for a particular choice of $p(\Theta)$ (conjugated prior distribution) the above integral can be explicitly computed. However, this is not what is generally done. Instead, as asymptotic approximation of $p(\mathbf{x}, \mathbf{z}|K)$ is usually adopted and it is known as Integrated Classification Likelihood (**ICL**)

$$ICL_K := \max_{\Theta} \log p(\mathbf{x}, \mathbf{z}|\Theta, K) - \frac{\nu(K)}{2} \log N, \tag{3.16}$$

where $\nu(K)$ is the number of model parameters. This quantity was introduced in Biernacki et al. (2000) and the reader can refer to that paper to see how this asymptotic approximation is obtained. Concretely, for a given K, after running the EM algorithm we might compute a MAP estimate of Z as $\hat{\mathbf{z}} := (\hat{z}_1, \ldots, \hat{z}_N)$ where $\hat{z}_i = \arg\max_k \tau_{ik}$. Then, we replace $\mathbf{z}$ by $\hat{\mathbf{z}}$ in

Eq. (3.16) and solve the maximization problem by means of Eq. (3.7). In such a way, ICL_K can be computed for several values of K ranging from 1 to a certain $K_{\max}$ and the value maximizing Eq. (3.16) is finally retained.

A more widespread alternative to ICL is the Bayesian Information Criterion **BIC**

$$BIC_K := \max_{\Theta} \log p(\mathbf{x}|\Theta, K) - \frac{\nu(K)}{2} \log N. \tag{3.17}$$

As it can be seen (for GMMs) it only differs from ICL in replacing the log-likelihood of the complete data with the one of the observed data. Following the very same reasoning that we did for ICL, BIC can be seen as an approximation of $p(K|\mathbf{x})$ where both $\mathbf{z}$ and Θ are integrated out. The details of how such an approximation is obtained can be found in Lebarbier and Mary-Huard (2006). Notice that, since $\log p(\mathbf{x}|\Theta, K)$ is not tractable, in practices the first term on the right hand side of the equality is replaced by $\mathcal{L}(p_{\mathbf{z}}, \hat{\Theta})$ the final lower bound after the EM algorithm converged. ICL and BIC do not necessarily select the same number of components, in particular it can be shown that $ICL_K < BIC_K$ and that ICL is always more conservative than BIC (Baudry et al., 2010).