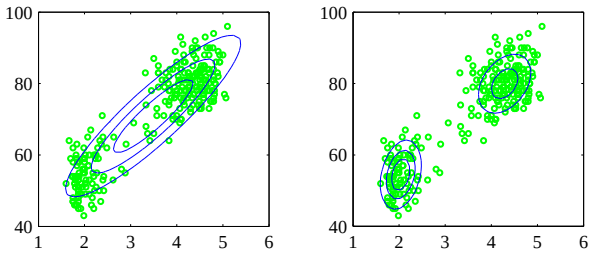
Lecture Outline

Reading: Section 2.3.9, Section 9.2

- Gaussian Mixture Model (GMM)
- ML estimation of GMM parameters using Expectation Maximization (EM) algorithm

2.3.9 Mixtures of Gaussians

While the Gaussian distribution has some important analytical properties, it suffers from significant limitations when it comes to modeling real data sets. Consider the example shown in Figure 21. This is known as the 'Old Faithful' data set (see Appendix A). Each 2D measurement comprises the duration of the eruption in minutes (x_1) and the time in minutes to the next eruption (x_2). We see that the data set forms two dominant clumps, and that a simple Gaussian distribution is unable to capture this structure, whereas a linear superposition of two Gaussians gives a better characterization of the data set.



(a) (b)

Figure 21: Independent variable (horizontal axis) duration of eruption and dependent variable (vertical axis) time in minutes to next eruption.

Such superpositions, formed by taking linear combinations of more basic distributions such as Gaussians, can be formulated as probabilistic models known as *mixture densities*. In Figure 22 we see that a linear combination of Gaussians can give rise to very complex densities. By using a sufficient number of gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy.

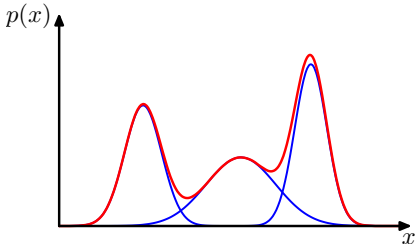


Figure 22:

We therefore consider a superposition of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

which is called a Gaussian Mixture Model (GMM). Each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is called a *component* of the mixture and has its own mean and covariance. The parameters $0 \leq \pi_k \leq 1$ are called *mixing coefficients* or *weights* and

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2}$$

since both the GMM and component densities are normalized.

The marginal density is given by

$$\begin{aligned} p(\mathbf{x}) &= \sum_{k=1}^{K} p(\mathbf{x}, k) \quad ;\text{sum rule} \\ &= \sum_{k} p(k)p(\mathbf{x}|k) \quad ;\text{product rule} \\ &= \sum_{k} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned} \tag{3}$$

which is equivalent to (1) in which we can view $\pi_k = p(k)$ as the *prior* probability of picking the kth component, and the component density, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x}|k)$ as the probability of $\mathbf{x}$ conditioned on k.

Contour and surface plots for a GMM having 3 components are shown in Figure 23.

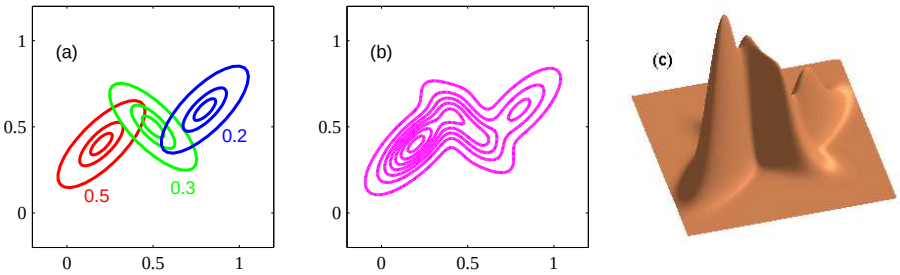


Figure 23:

An important role is played by the *posterior* probabilities, $p(k|\mathbf{x})$. From Bayes' theorem,

$$\begin{aligned} p(k|\mathbf{x}) &= \frac{p(\mathbf{x}|k)p(k)}{p(\mathbf{x})} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \end{aligned} \tag{4}$$

The *posterior* probabilities, $p(k|\mathbf{x})$ can be viewed as the *responsibility* that component k takes for 'explaining' the observation $\mathbf{x}$.

The distributional parameters of the GMM are $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ where $\boldsymbol{\pi} \equiv \{\pi_1, \ldots, \pi_K\}$, $\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, and $\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$. Of obvious interest will be estimation of $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ for a given set of training data.

Example: Suppose we have a GMM with $D = 20$ dimensional vectors and $K = 100$ components. The GMM has 42,100 distributional parameters: K priors π_k, $D \times K$ mean vector elements, and $D^2 \times K$ covariance matrix elements.

Estimation of GMM Parameters via Maximum Likelihood

Suppose we have a set of observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and we wish to model this data using a mixture of Gaussians. Thus from the data and a given K we must find the distributional parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ of our model.

The likelihood function is given by

$$p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5}$$

and the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{6}$$

Maximizing the log likelihood function for a GMM turns out to be a more complex problem than for the case of a single Gaussian. The difficulty arises from the presence of the summation over k that appears inside the log, so that the log function no longer acts directly on the Gaussian. There is currently no known closed-form solution for $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ via ML.

9.2.2 Expectation Maximization for Gaussian Mixtures

An elegant and powerful method for finding maximum likelihood solutions for mixture models is called the *expectation-maximization* algorithm or the EM algorithm. Let us begin by writing down the conditions that must be satisfied at a maximum of the likelihood function.

Condition 1: Setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ in (6) with respect to each of the mean vectors, $\boldsymbol{\mu}_k$ of the Gaussian components to zero we obtain

$$0 = \sum_{n=1}^{N} p(k|\mathbf{x}_n) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k), \quad 1 \leq k \leq K. \tag{7}$$

The optimal $\{\boldsymbol{\mu}_k\}$ must satisfy this.

Multiplying by $\boldsymbol{\Sigma}_k$, we obtain

$$\sum_{n=1}^{N} p(k|\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{8}$$

which gives

$$\sum_{n=1}^{N} p(k|\mathbf{x}_n)\mathbf{x}_n = \boldsymbol{\mu}_k \sum_{n=1}^{N} p(k|\mathbf{x}_n) \tag{9}$$

or

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} p(k|\mathbf{x}_n)\mathbf{x}_n \tag{10}$$

where

$$N_k = \sum_{n=1}^{N} p(k|\mathbf{x}_n). \tag{11}$$

We can interpret N_k as the effective number of data points assigned to component k. We see that the mean $\boldsymbol{\mu}_k$ for the kth Gaussian component is obtained by taking a weighted mean of all the points in the data set, in which the weighting factor for data point $\mathbf{x}_n$ is given by the posterior probability that component k was *responsible* for generating $\mathbf{x}_n$. We call $p(k|\mathbf{x}_n)$ the "kth responsibility for datapoint $\mathbf{x}_n$" and later in the text will be denoted as $\gamma(z_{nk})$

Condition 2: If we set the derivative of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}_k$ to zero we obtain

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} p(k|\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T. \tag{12}$$

Condition 3: Finally, if we set the derivative of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to π_k to zero we obtain

$$\pi_k = \frac{N_k}{N} = \frac{\sum_{n=1}^{N} p(k|\mathbf{x}_n)}{N} \tag{13}$$

so the mixing coefficient for the kth component density is given by the average responsibility which that component takes for explaining the data points.

It is worth emphasizing that the results in (10), (12), and (13) do not constitute a closed-form solution for the parameters of the GMM because the responsibilities $p(k|\mathbf{x}_n)$ depend on those parameters, $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, in a complex way through (4). However, these results suggest a simple iterative scheme for finding a solution to the maximum likelihood problem, i.e. EM algorithm for GMM.

EM Algorithm for Gaussian Mixtures

1. Initialize[1] the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$, and mixing coefficients π_k, and evaluate the initial value of the log likelihood for the given data set of observations, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

2. **E step**. Evaluate the responsibilities using the current parameter values

$$p(k|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{14}$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} p(k|\mathbf{x}_n)\mathbf{x}_n \tag{15}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} p(k|\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T. \tag{16}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{17}$$

[1] A popular initialization is $\pi_k = 1/K$, $\boldsymbol{\Sigma}_k = \mathbf{I}$, and $\boldsymbol{\mu}_k$ set to randomly chosen data points.

where

$$N_k = \sum_{n=1}^{N} p(k|\mathbf{x}_n). \tag{18}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{19}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

Notes:

1. We can also terminate the EM algorithm after a fixed number of iterations.
2. To track the convergence of the log likelihood (and hence the parameters) we first compute the improvement in log likelihood from iteration $^{(m)}$ to iteration $^{(m+1)}$ with

$$\text{improvement} = \frac{\ln p(\mathbf{X}|\boldsymbol{\pi}^{(m+1)}, \boldsymbol{\mu}^{(m+1)}, \boldsymbol{\Sigma}^{(m+1)}) - \ln p(\mathbf{X}|\boldsymbol{\pi}^{(m)}, \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})}{\ln p(\mathbf{X}|\boldsymbol{\pi}^{(m+1)}, \boldsymbol{\mu}^{(m+1)}, \boldsymbol{\Sigma}^{(m+1)})} \tag{20}$$

   and check to see if this improvement is below a threshold.

Figure 24: Text Figure 9.8