Using EM To Estimate A Probability Density With A Mixture Of Gaussians

Aaron A. D'Souza
adsouza@usc.edu

1 Introduction

The problem we are trying to address in this note is simple. Given a set of data points $\mathbf{X} = \{\mathbf{x}_i\}_1^N$, we wish to determine the underlying probability distribution $p(\mathbf{x})$, that generates this data. In general, the distribution $p(\mathbf{x})$ could be any real-valued, scalar function with the following constraints:

$$p(\mathbf{x}) \geq 0, \forall \mathbf{x} \text{ and } \int_{-\infty}^{\infty} p(\mathbf{x}) d\mathbf{x} = 1$$

The framework of mixture modeling using Gaussians makes the following assumptions:

- The data was generated using a *set* of K probability distributions.
- Each of the individual probability distributions is a Gaussian:

  $$\mathbf{x}; \boldsymbol{\theta}_k \sim \text{Normal}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$ [*]

  i.e., the probability of generating a data point $\mathbf{x}$ under the m^{th} model is given according to a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$.
- Each data point is generated according to the following algorithm:

  for $i = 1$ to N **do**
  $m \leftarrow$ index of one of the K models randomly selected
  according to the prior probability vector $\boldsymbol{\pi}$
  Randomly generate $\mathbf{x}_i$ according to the distribution $\text{Normal}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$
  end for

Representing probabilistic systems as graphical models is rapidly becoming a useful tool in Bayesian analysis. The graphical model corresponding to our formulation of the data generation process is shown in fig. 1. We adopt the convention that circular nodes correspond to random variables in the model, while rectangular nodes correspond to variables that parameterize the distributions of these random variables. Circular nodes with double borders indicate observed variables in our model (the data).

[*] We use the notation $\text{Normal}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ to denote the multivariate Gaussian distribution which is mathematically defined as:

$$\text{Normal}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = (2\pi)^{-d/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where d is the dimensionality of $\mathbf{x}$.

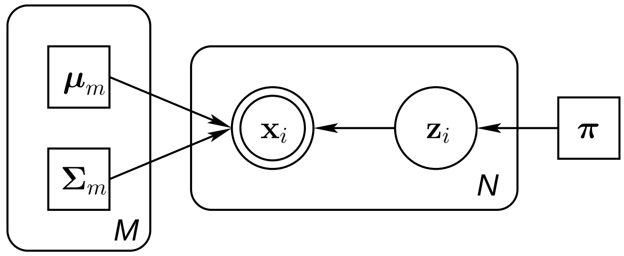


Figure 1: Graphical model for maximum likelihood density estimation using a mixture of Gaussians

In this model we have introduced an additional variable $\mathbf{z}_i$ associated with each $\mathbf{x}_i$. The $\mathbf{z}_i$ variables are *indicator* variables that are multinomially distributed according to the parameter vector $\boldsymbol{\pi}$, and indicate which component generates the corresponding $\mathbf{x}_i$. It is easiest to think of each $\mathbf{z}_i$ as an K dimensional vector with a 1 in the element corresponding to the selected mixture component, and 0's in all other elements. The probability of the m^{th} element being 1, is w_k.

$$\mathbf{z}_i = \underbrace{\begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 \end{bmatrix}^T}_{K \text{ elements}}$$

Since we do not know the corresponding $\mathbf{z}_i$ for each $\mathbf{x}_i$ (if we did, then we simply group the $\mathbf{x}_i$ according to their $\mathbf{z}_i$, and fit a single Gaussian to each group), these variables are called *hidden* variables.

Our problem is now reduced to finding the values of the model parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ for each of the K models, as well as the prior probability vector $\boldsymbol{\pi}$, which when plugged into the generative model, is most likely to generate the observed data distribution. In other words we are interested in maximizing the *likelihood* $\mathcal{L}(\boldsymbol{\theta}) = p(\mathbf{X}; \boldsymbol{\theta})$ of generating the observed data given the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k\}_1^K$. This approach is called the Maximum Likelihood (ML) framework since it finds the parameter settings that maximize the likelihood of observing the data.

Although the ML approach is an intuitively appealing solution, we often find that maximizing the expressions for likelihood w.r.t. the parameters $\boldsymbol{\theta}$ are often analytically intractable. The Expectation Maximization (EM) algorithm can be used to simplify the math considerably.

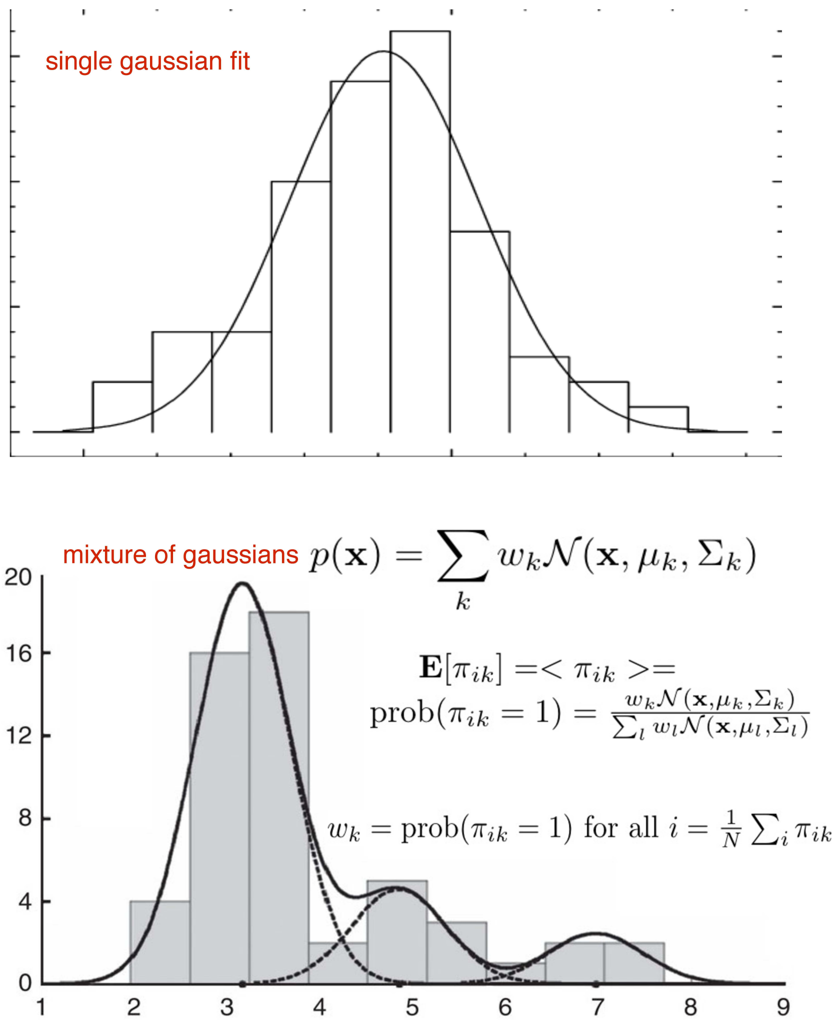


Figure 2: Mixtures equations

2 Estimating the Model Parameters using EM

Instead of attempting to maximize the likelihood of the observed data $p(\mathbf{X}; \boldsymbol{\theta})$, we attempt instead to maximize the likelihood of the joint distribution of $\mathbf{X}$ and $\mathbf{Z} = \{\mathbf{z}_i\}_1^N$, $p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$. For the purposes of maximization we can also work with the logarithm of this quantity, $\log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$. This quantity is also known as the *complete* log-likelihood. Since we cannot observe the values of the random variables $\mathbf{z}_i$ we must work with the expectation of this quantity w.r.t. some distribution $Q(\mathbf{Z})$.

The log of the complete data likelihood can be written as follows:

$$\begin{aligned} l_c(\boldsymbol{\theta}) &= \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) \\ &= \log \prod_i^N p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}) \end{aligned}$$

$$= \log \prod_{i}^{N} \prod_{k}^{K} \left[p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) p(\pi_{ik} = 1) \right]^{\pi_{ik}}$$

$$= \sum_{i=1}^{N} \sum_{k}^{K} \pi_{ik} \log p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) + \pi_{ik} \log w_k$$

Since we have assumed that each of the individual models is a Gaussian, the quantity $p(\mathbf{x}_i|m, \boldsymbol{\theta})$ is simply the conditional probability of generating $\mathbf{x}_i$ given that the m^{th} model is chosen:

$$\log p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \tag{1}$$

Taking expectations w.r.t. $Q(\mathbf{Z})$ we get:

$$\langle l_c(\boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})} = \sum_{i=1}^{N} \sum_{k}^{K} \langle \pi_{ik} \rangle \log p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) + \langle \pi_{ik} \rangle \log w_k \tag{2}$$

2.1 The M step

The "M" step in EM takes the expected complete log-likelihood as defined in eq. (2) and maximizes it w.r.t. the parameters that are to be estimated; in this case w_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$.

Differentiating eq. (2) w.r.t. $\boldsymbol{\mu}_k$ we get:

$$\frac{\partial \langle l_c(\boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})}}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} \langle \pi_{ik} \rangle \frac{\partial}{\partial \boldsymbol{\mu}_k} \log p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) = \mathbf{0} \tag{3}$$

We can compute $\frac{\partial}{\partial \boldsymbol{\mu}_k} \log p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta})$ using eq. (1) as follows:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k} \log p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\mu}_k} \log \left\{ \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \right\} \\
&= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \\
&= (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1 \dagger}
\end{aligned}$$

Substituting this result into eq. (3), we get:

$$\sum_{i=1}^{N} \langle \pi_{ik} \rangle (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} = \mathbf{0}$$

giving us the update equation:

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle \mathbf{x}_i}{\sum_{i=1}^{N} \langle \pi_{ik} \rangle} \tag{4}$$

Differentiating eq. (2) w.r.t. $\boldsymbol{\Sigma}_k^{-1}$ we get:

$$\frac{\partial \langle l_c(\boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})}}{\partial \boldsymbol{\Sigma}_k^{-1}} = \sum_{i=1}^{N} \langle \pi_{ik} \rangle \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \log p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) = \mathbf{0} \tag{5}$$

†Where we have used the relation $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{x}^T (\mathbf{A} + \mathbf{A}^T)$

We can compute $\frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \log p(\mathbf{x}_i|\pi_{ik}=1;\boldsymbol{\theta})$ using eq. (1) as follows:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \log p(\mathbf{x}_i|\pi_{ik}=1;\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \log \left\{ \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_k\right|^{1/2}} \exp\left\{ -\frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)\right\}\right\} \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \left\{ \frac{1}{2}\log\left|\boldsymbol{\Sigma}_k^{-1}\right| - \frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)\right\} \\
&= \frac{1}{2}\boldsymbol{\Sigma}_k - \frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)^T \,\ddagger
\end{aligned}$$

Substituting this result into eq. (5), we get:

$$\sum_{i=1}^{N} \langle \pi_{ik} \rangle \left(\frac{1}{2}\boldsymbol{\Sigma}_k - \frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)^T \right) = \mathbf{0}$$

giving us the update equation:

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle \left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)^T}{\sum_{i=1}^{N} \langle \pi_{ik} \rangle} \tag{6}$$

In order to maximize the expected log-likelihood in eq. (2) w.r.t. w_k, we have to keep in mind that the maximization has the constraint that $\sum_k^K w_k = 1$. In order to enforce this constraint we use the Lagrange multiplier λ, and augment eq. (2) as follows:

$$L'(\boldsymbol{\theta}) = \langle l_c(\boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})} - \lambda \left(\sum_k^K w_k - 1 \right) \tag{7}$$

We now differentiate this new expression w.r.t. each w_k giving us:

$$\frac{\partial}{\partial w_k} \langle l_c(\boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})} - \lambda = 0 \qquad \text{for } 1 \leq m \leq K$$

Using eq. (2) we get:

$$\left. \begin{aligned} \frac{1}{w_k}\sum_{i=1}^{N} \langle \pi_{ik} \rangle - \lambda = 0 \\ \text{or equivalently } \sum_{i=1}^{N} \langle \pi_{ik} \rangle - \lambda w_k = 0 \end{aligned} \right\} \qquad \text{for } 1 \leq m \leq K \tag{8}$$

Summing eq. (8) over all K models we get:

$$\sum_k^K \sum_{i=1}^{N} \langle \pi_{ik} \rangle - \lambda \sum_k^K w_k = 0$$

But since $\sum_k^K w_k = 1$ we get:

$$\lambda = \sum_k^K \sum_{i=1}^{N} \langle \pi_{ik} \rangle = N \tag{9}$$

‡Where we have used the relation $\frac{\partial}{\partial \mathbf{X}} \log|\mathbf{X}| = \left(\mathbf{X}^{-1}\right)^T$ and $\frac{\partial}{\partial \mathbf{A}} \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}\mathbf{x}^T$

Substituting this result back into eq. (8) we get the following update equation:

$$w_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle}{N} \tag{10}$$

which preserves the constraint that $\sum_k^K w_k = 1$.

2.2 The E step

Now that we have derived the update equations that maximize the expected *complete* log-likelihood $\langle \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) \rangle$, we wish to ensure that we are indeed also maximizing the *incomplete* log-likelihood $p(\mathbf{X}; \boldsymbol{\theta})$ (which is the quantity that we are truly interested in maximizing).

As we mentioned earlier in section 2, we are guaranteed to maximize the incomplete log-likelihood only when the expectation is taken w.r.t. the posterior distribution of $\mathbf{Z}$, namely $p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta})$. Hence each of the expectations $\langle \pi_{ik} \rangle$ that appear in the update equations derived in the previous section (section 2.1), should be computed as follows:

$$\begin{aligned}
\langle \pi_{ik} \rangle_{p(\mathbf{Z}|\mathbf{X};\boldsymbol{\theta})} &= 1 \cdot p(\pi_{ik} = 1 | \mathbf{x}_i; \boldsymbol{\theta}) + 0 \cdot p(\pi_{ik} = 0 | \mathbf{x}_i; \boldsymbol{\theta}) \\
&= p(\pi_{ik} = 1 | \mathbf{x}_i; \boldsymbol{\theta}) \\
&= \frac{p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) p(\pi_{ik} = 1)}{\sum_{k=1}^{K} p(\mathbf{x}_i | z_{ik} = 1; \boldsymbol{\theta}) p(z_{ik} = 1)} \\
&= \frac{p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) w_k}{\sum_{k=1}^{K} p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) w_k} \\
&= \frac{p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) w_k}{notmalized_over_k}
\end{aligned}$$

3 Summary

To summarize, given a set of data points $\mathbf{X}$, if we wish to estimate the underlying probability distribution using EM to fit K Gaussians, we apply Algorithm 1, iterating until convergence of the model parameters.

4 Why does this work? (A brief review of EM theory)

Knowing that $\log p(\mathbf{X}; \boldsymbol{\theta})$ is difficult to maximize analytically, we (seemingly arbitrarily) chose to maximize the expected *complete* log-likelihood $\langle \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})}$ in the hope that this also increases the *incomplete* log-likelihood $\log p(\mathbf{X}; \boldsymbol{\theta})$ (the quantity we are *really* interested in). This section will justify this choice and prove that we are indeed maximizing $\log p(\mathbf{X}; \boldsymbol{\theta})$

Let us rewrite $\log p(\mathbf{X}|\boldsymbol{\theta})$ as follows:

Initialize: all $\langle \pi_{ik} \rangle$, w_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$

repeat

  for $i = 1$ to N **do** *//The E step*

    for $m = 1$ to K **do**

$$p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) = (2\pi)^{-d/2} \left|\boldsymbol{\Sigma}_k\right|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}$$

$$\langle \pi_{ik} \rangle = \frac{p(\mathbf{x}_i | \pi_{ik} = 1; \boldsymbol{\theta}) w_k}{normalized\_over\_k}$$

    end for

  end for

  for $m = 1$ to K **do** *//The M step*

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \langle \pi_{ik} \rangle}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle \mathbf{x}_i}{\sum_{i=1}^{N} \langle \pi_{ik} \rangle}$$

$$w_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle}{N}$$

  end for

until model parameters converge

Algorithm 1: Estimate $w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ for $1 \leq m \leq K$

$$\begin{aligned}
\log p(\mathbf{X}; \boldsymbol{\theta}) &= \log \int p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\mathbf{Z} \\
&= \log \int Q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})}{Q(\mathbf{Z})} d\mathbf{Z} && (11) \\
&\geq \int Q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})}{Q(\mathbf{Z})} d\mathbf{Z} \qquad \text{(Jensen's inequality)} && (12) \\
&= \int Q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\mathbf{Z} - \int Q(\mathbf{Z}) \log Q(\mathbf{Z}) d\mathbf{Z} && (13)
\end{aligned}$$

Hence we have the following lower-bound to $\log p(\mathbf{X}; \boldsymbol{\theta})$:

$$\log p(\mathbf{X}; \boldsymbol{\theta}) \geq \underbrace{\langle \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})}}_{\text{exp. comp. log-lik.}} + \underbrace{\mathcal{H}\left[Q(\mathbf{Z})\right]}_{\text{entropy of } Q(\mathbf{Z})} = \mathcal{F}(Q, \boldsymbol{\theta}) \tag{14}$$

Since $Q(\mathbf{Z})$ is an arbitrary distribution, it is independent of $\boldsymbol{\theta}$. Hence in order to maximize the functional $\mathcal{F}(Q, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$, it suffices to simply maximize $\langle \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})}$. (Hence the M-step).

Does this maximization achieve our aim? Eq. (14) shows that the functional $\mathcal{F}(Q, \boldsymbol{\theta})$ is a *lower bound* to the quantity we are interested in. In which case, maximizing $\mathcal{F}(Q, \boldsymbol{\theta})$ does not guarantee that we are improving $\log p(\mathbf{X}; \boldsymbol{\theta})$ at all! If however, we set $Q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta})$ in Eq. (12), then we see that *the lower bound in fact becomes an equality.*

$$
\begin{aligned}
\int Q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})}{Q(\mathbf{Z})} d\mathbf{Z} &= \int p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}) \log \frac{p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta})} d\mathbf{Z} \\
&= \int p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}) \log \frac{p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}) p(\mathbf{X}; \boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta})} d\mathbf{Z} \\
&= \int p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}) \log p(\mathbf{X}; \boldsymbol{\theta}) d\mathbf{Z} \\
&= \log p(\mathbf{X}; \boldsymbol{\theta}) \int p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}) d\mathbf{Z} \\
&= \log p(\mathbf{X}; \boldsymbol{\theta})
\end{aligned}
$$

This means that when computing the expected complete log-likelihood $\langle \log p(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) \rangle_{Q(\mathbf{Z})}$, the expectation should be taken w.r.t. the true posterior $p(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta})$ of the hidden variables (the E step).

5 Examples

Figures 3, 4 and 5 show the result of fitting 2, 3, and 4 Gaussians respectively to a set of data points. In each figure the first plot shows the positions of the means and the relative covariances of each Gaussian, while the second shows the resulting estimated distribution obtained by marginalizing over the models as follows:

$$
p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k}^{K} p(\mathbf{x}|z_k = 1; \boldsymbol{\theta}) p(z_k = 1)
$$

Figure 6 demonstrates that one can model rather arbitrary non-gaussian distributions provided we have a sufficient number of mixture components. The tradeoff here is that too few components will fail to model the structure of the data, while too many will "overfit" the data. The *model selection* problem in this context is determining an appropriate compromise.

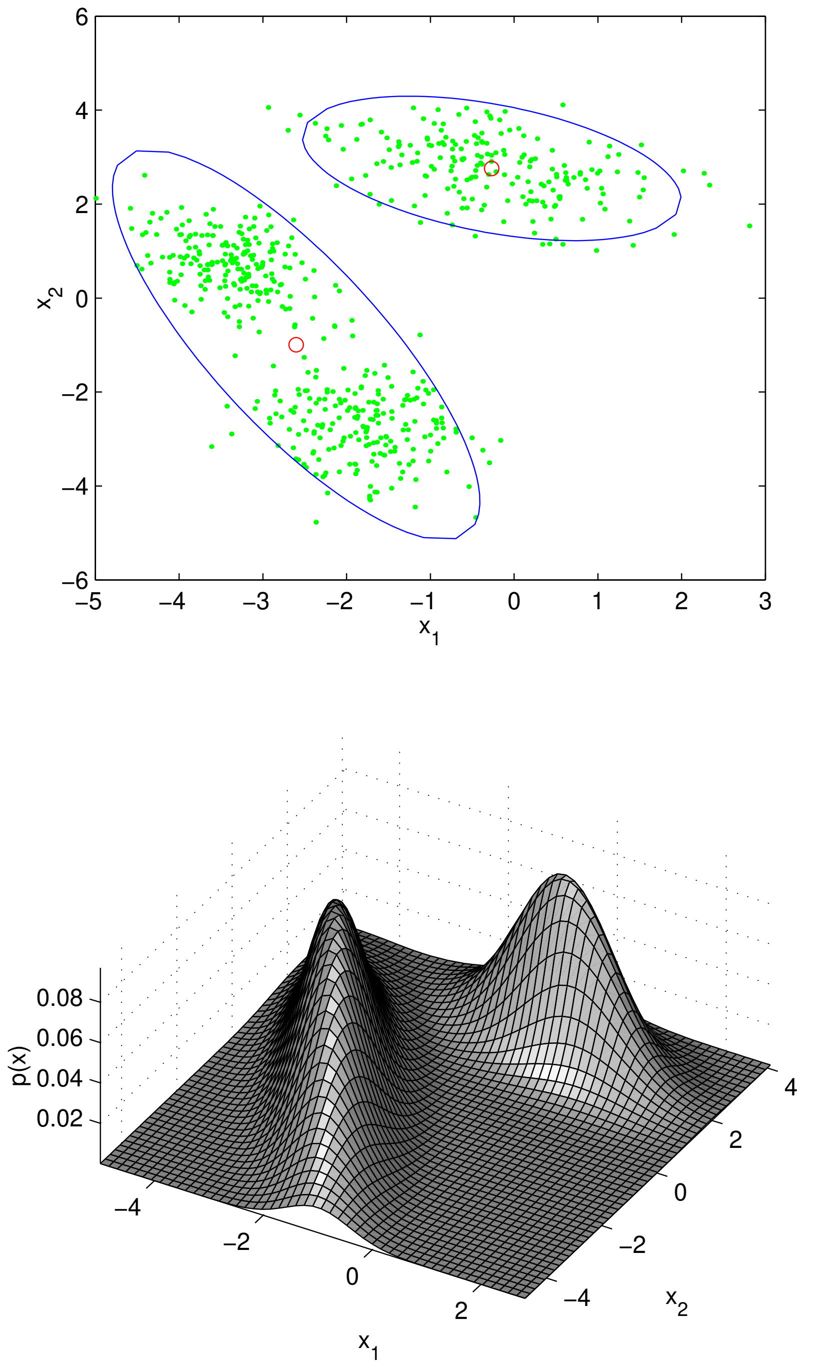


Figure 3: Fitting with 2 Gaussians

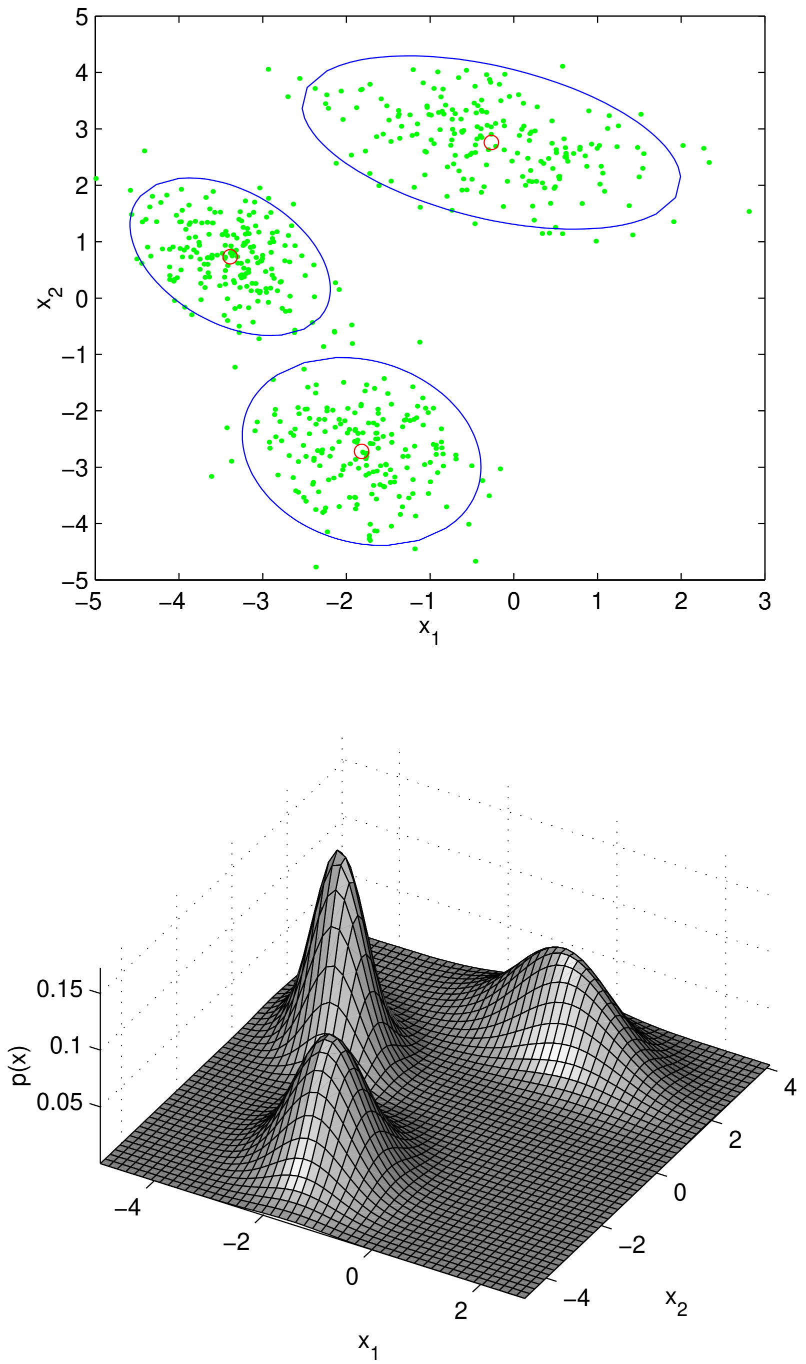


Figure 4: Fitting with 3 Gaussians

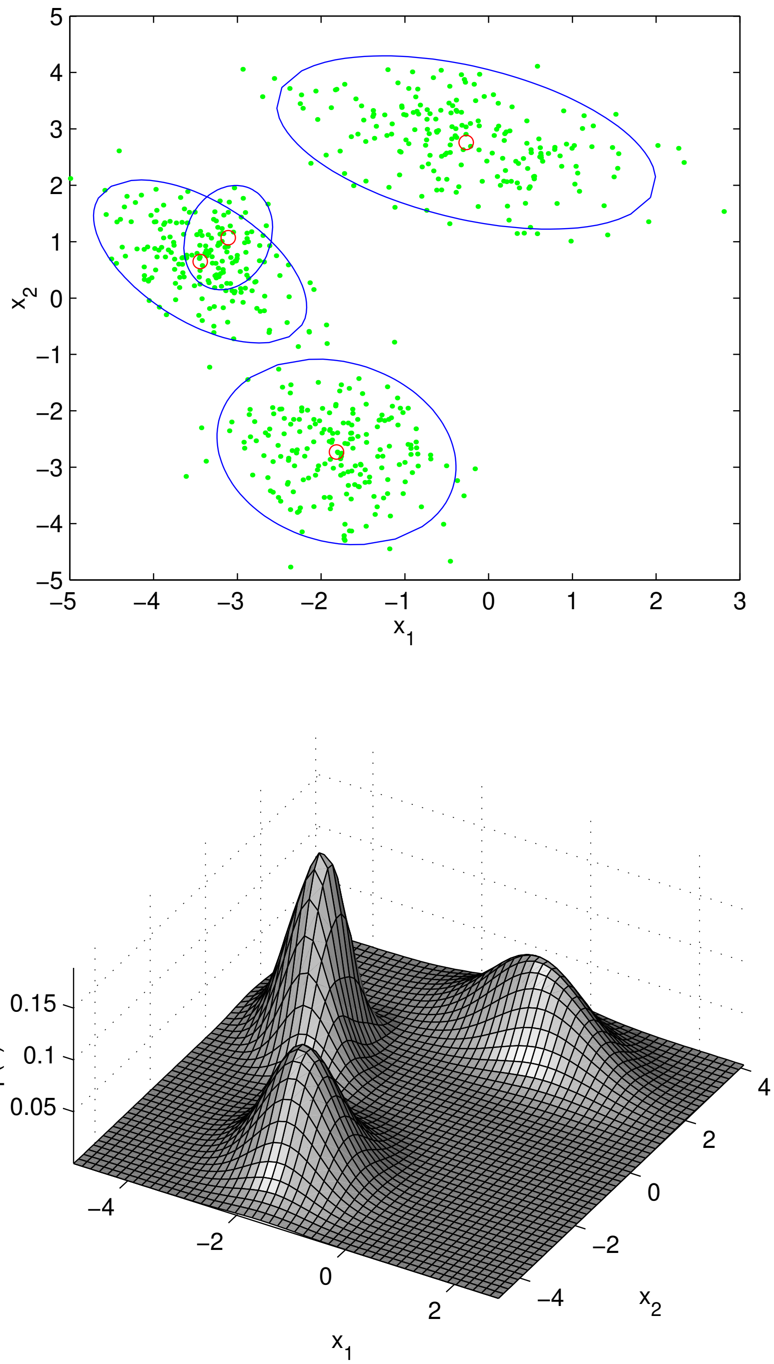
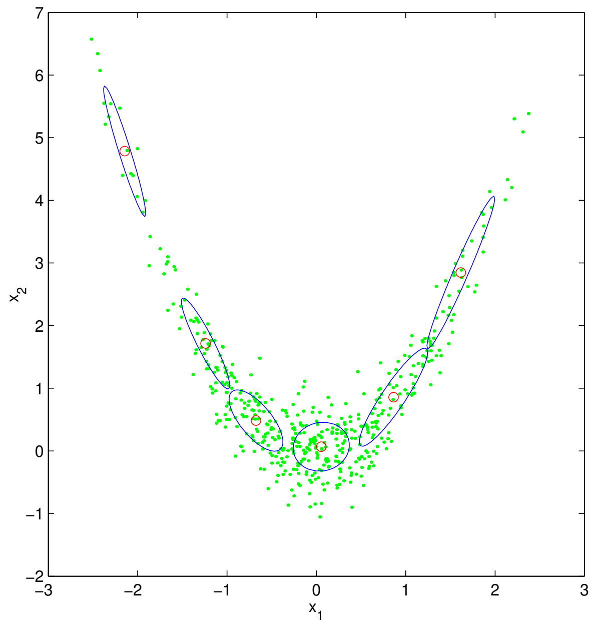


Figure 5: Fitting with 4 Gaussians



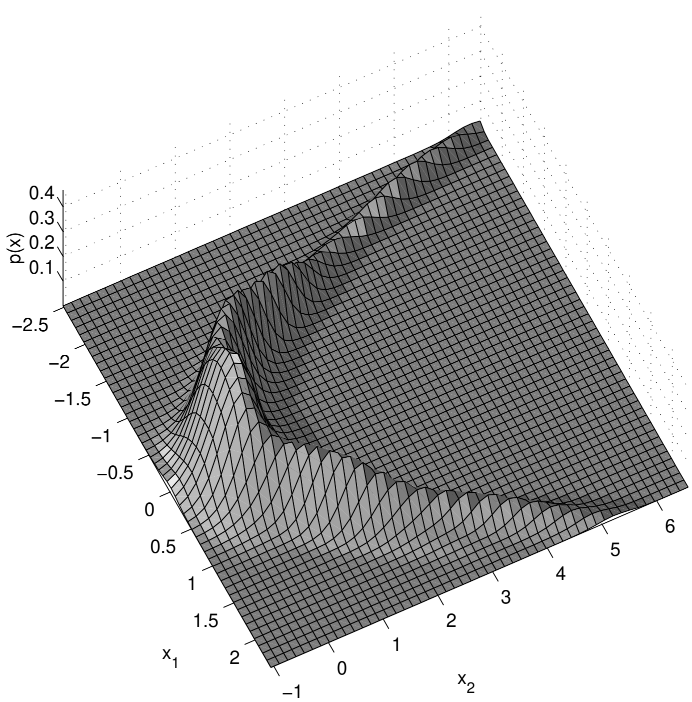


Figure 6: A sufficient number of mixture components can model arbitrary distributions