The Expectation-Maximization Algorithm

Mihaela van der Schaar

Department of Engineering Science
University of Oxford

Latent Variables and Marginal Likelihoods

- Many probabilistic models have hidden variables that are not observable in the dataset $\mathcal{D}$: these models are known as *latent variable models*.
- **Examples:** Hidden Markov Models & Mixture Models.
- **How would MLE be carried out for such models?**
  - Each data point is drawn from a **joint** distribution $\mathbb{P}_\theta(X, Z)$.
  - For a realization $((X_1, Z_1), \ldots, (X_n, Z_n))$, we only observe the variables in the dataset $\mathcal{D} = (X_1, \ldots, X_n)$.
  - **Complete-data likelihood:**

$$\mathbb{P}_\theta((X_1, Z_1), \ldots, (X_n, Z_n)) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i, Z_i).$$

  - **Marginal likelihood:**

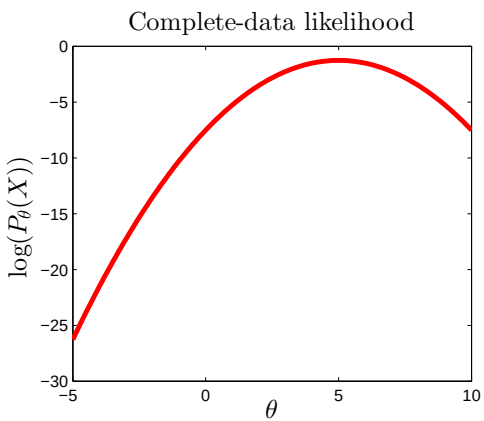
$$\mathbb{P}_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} \sum_{z} \mathbb{P}_\theta(X_i,\ Z_i = z).$$

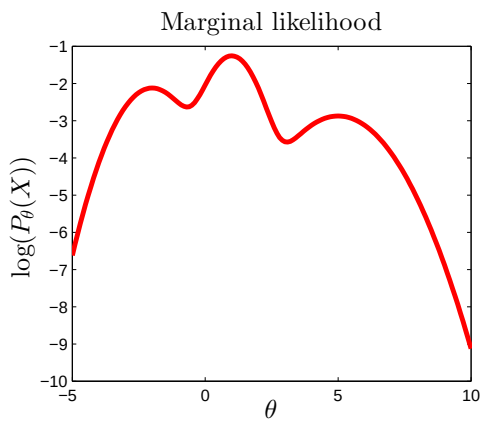
The Hardness of Maximizing Marginal Likelihoods (I)

- The MLE is obtained by maximizing the marginal likelihood:

$$\hat{\theta}_n^* = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log \left(\sum_z \mathbb{P}_\theta(X_i,\, Z_i = z) \right).$$

- **Solving this optimization problem is often a hard task!**
  - Non-convex.
  - Many local maxima.
  - No analytic solution.





The Hardness of Maximizing Marginal Likelihoods (II)

- The MLE for θ is obtained by maximizing the marginal log likelihood function:

$$\hat{\theta}_n^* = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log \left(\sum_z \mathbb{P}_\theta(X_i, Z_i = z) \right).$$

- **Solving this optimization problem is often a hard task!**
- The methods used in the previous lecture would not work.
- Need a simpler **approximate** procedure!
- The Expectation-Maximization is an *iterative algorithm* that computes an approximate solution for the MLE optimization problem.

Exponential Families (I)

- The EM algorithm is well-suited for exponential family distributions.

Exponential Family

A single-parameter exponential family is a set of probability distributions that can be expressed in the form

$$\mathbb{P}_\theta(X) = h(X) \cdot \exp\left(\eta(\theta) \cdot T(X) - A(\theta)\right),$$

where $h(X), A(\theta)$ and $T(X)$ are known functions. An alternative, equivalent form often given as

$$\mathbb{P}_\theta(X) = h(X) \cdot g(\theta) \cdot \exp\left(\eta(\theta) \cdot T(X)\right).$$

The variable θ is called the parameter of the family.

Exponential Families (II)

- Exponential family distributions:

$$\mathbb{P}_\theta(X) = h(X) \cdot \exp\left(\eta(\theta) \cdot T(X) - A(\theta)\right).$$

- $T(X)$ **is a sufficient statistic of the distribution**
  - The sufficient statistic is a function of the data that fully summarizes the data X within the density function $\mathbb{P}_\theta(X)$.
  - This means that for any data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, the density function is the same if $T(\mathcal{D}_1) = T(\mathcal{D}_2)$. This is true even if $\mathcal{D}_1$ and $\mathcal{D}_2$ are quite different.
  - The sufficient statistic of a set of independent identically distributed data observations is simply the sum of individual sufficient statistics, i.e. $T(\mathcal{D}) = \sum_{i=1}^{n} T(X_i)$.

Exponential Families (III)

- Exponential family distributions:

$$\mathbb{P}_\theta(X) = h(X) \cdot \exp\left(\eta(\theta) \cdot T(X) - A(\theta)\right).$$

- $\eta(\theta)$ **is called the natural parameter**
  - The set of values of $\eta(\theta)$ for which the function $\mathbb{P}_\theta(X)$ is finite is called the natural parameter space.
- $A(\theta)$ **is called the log-partition function**
  - The mean, variance and other moments of the sufficient statistic $T(X)$ can be derived by differentiating $A(\theta)$.

Exponential Families (IV)

- **Exponential Family Example: Normal Distribution**

$$\begin{aligned}\mathbb{P}_\theta(X) &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{-(X-\mu)^2}{2\,\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{X^2 - 2\,X\,\mu + \mu^2}{2\,\sigma^2} - \log(\sigma)\right) \\ &= \frac{\exp\left(\left\langle \left[\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right]^T, \left[X, X^2\right]^T \right\rangle - \left(\frac{\mu^2}{2\sigma^2} + \log(\sigma)\right)\right)}{\sqrt{2\pi}}.\end{aligned}$$

$$\eta(\theta) = \left[\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right]^T, \quad h(X) = (2\pi)^{-\frac{1}{2}}$$

$$T(X) = \left[X, X^2\right]^T, \quad A(\theta) = \left(\frac{\mu^2}{2\sigma^2} + \log(\sigma)\right).$$

Exponential Families (V)

Properties of Exponential Families

- Exponential families have sufficient statistics that can summarize arbitrary amounts of independent identically distributed data using a fixed number of values.
- Exponential families have conjugate priors (an important property in Bayesian statistics).
- The posterior predictive distribution of an exponential-family random variable with a conjugate prior can always be written in closed form.

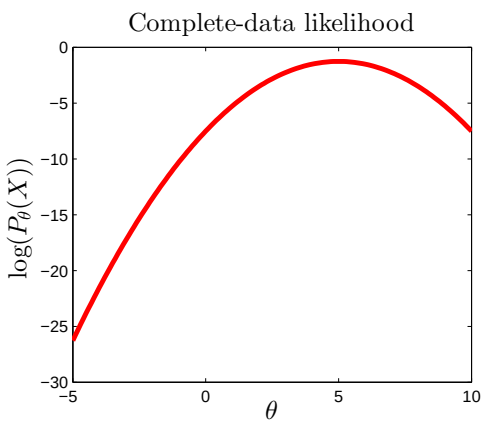
Exponential Families (VI)

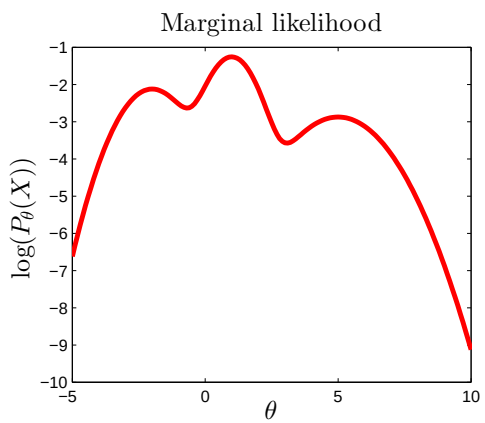
The Canonical Form of Exponential Families

- If $\eta(\theta) = \theta$, then the exponential family is said to be in canonical form.
- The canonical form is non-unique, since $\eta(\theta)$ can be multiplied by any nonzero constant, provided that $T(X)$ is multiplied by that constant's reciprocal, or a constant c can be added to $\eta(\theta)$ and $h(X)$ multiplied by $\exp(-c \cdot T(x))$ to offset it.

Expectation-Maximization (I)

- **Two unknowns:**
  - The latent variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$.
  - The parameter θ.
- Complications arise because we don't know the latent variables $(Z_1, \ldots, Z_n) \rightarrow$ **maximizing $\mathbb{P}_\theta((X_1, Z_1), \ldots, (X_n, Z_n))$ is often a simpler task!**
- Recall that maximizing the complete-data likelihood is often simpler than maximizing the marginalized likelihood!





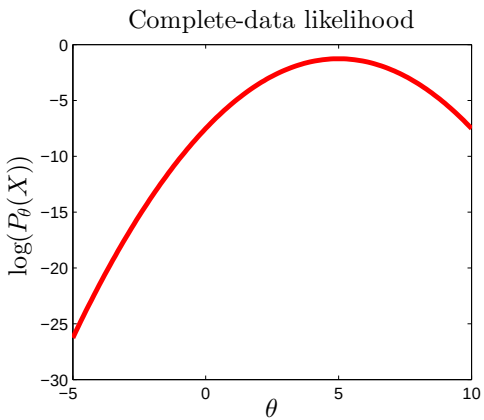
Expectation-Maximization (II)

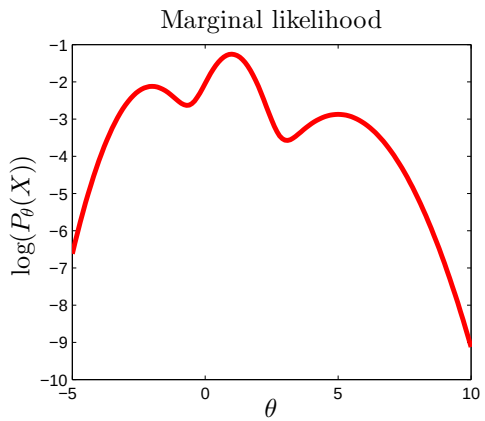
The EM Algorithm

1. Start with an initial guess $\hat{\theta}^{(o)}$ for θ.
   For every iteration t, do the following:
2. **E-Step:** $Q(\theta, \hat{\theta}^{(t)}) = \sum_{\mathbf{z}} \log\left(\mathbb{P}_\theta(\mathbf{Z} = \mathbf{z}, \mathcal{D})\right) \cdot \mathcal{P}(\mathbf{Z} | \mathcal{D}, \hat{\theta}^{(t)})$
3. **M-Step:** $\hat{\theta}^{(t+1)} = \arg\max_{\theta \in \Theta} Q(\theta, \hat{\theta}^{(t)})$
4. **Go to step 2 if stopping criterion is not met.**

Expectation-Maximization (III)

- **Two unknowns:**
  - The latent variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$.
  - The parameter θ.
- **"Expected" Likelihood:** $\sum_{\mathbf{z}} \log\left(\mathbb{P}_\theta(\mathbf{Z} = \mathbf{z}, \mathcal{D})\right) \cdot \mathbb{P}(\mathbf{Z} | \mathcal{D}, \theta)$.
  - Here the logarithm acts directly on the complete-data likelihood, so the corresponding M-step will be tractable.





Expectation-Maximization (III)

- **Two unknowns:**
  - The latent variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$.
  - The parameter θ.
- **"Expected" Likelihood:** $\sum_{\mathbf{z}} \log\left(\mathbb{P}_\theta(\mathbf{Z} = \mathbf{z}, \mathcal{D})\right) \cdot \mathbb{P}(\mathbf{Z} | \mathcal{D}, \theta)$.
  - Here the logarithm acts directly on the complete-data likelihood, so the corresponding M-step will be tractable.
  - But we still have two terms ($\log\left(\mathbb{P}_\theta(\mathbf{Z} = \mathbf{z}, \mathcal{D})\right)$ & $\mathbb{P}(\mathbf{Z} | \mathcal{D}, \theta)$) that depend on the two unknowns $\mathbf{Z}$ and θ.
- **The EM algorithm:**
  - **E-step:** Fix the posterior $\mathbf{Z} | \mathcal{D}, \theta$ by conditioning on the current guess for θ, i.e. $\mathbf{Z} | \mathcal{D}, \theta^{(t)}$.
  - **M-step:** Update the guess for θ by solving a tractable optimization problem.
- The EM algorithm breaks down the intractable MLE optimization problem into simpler, tractable iterative steps.

EM for Exponential Family (I)

- The critical points of the marginal likelihood function: $\frac{\partial \log(\mathbb{P}_\theta(\mathcal{D}))}{\partial \theta} = \frac{1}{\mathbb{P}_\theta(\mathcal{D})} \frac{\sum_z \partial \mathbb{P}_\theta(\mathcal{D}, \mathbf{Z}=z)}{\partial \theta} = 0.$

$$\frac{\partial \log\left(\mathbb{P}_\theta(\mathcal{D}, \mathbf{Z})\right)}{\partial \theta} = \frac{\partial}{\partial \theta} \log \left(\underbrace{\exp\left(\langle \eta(\theta), T(\mathcal{D}, \mathbf{Z}) \rangle - A(\theta)\right) \cdot h(\mathcal{D}, \mathbf{Z})}_{\text{Canonical form of exponential family}} \right).$$

For $\eta(\theta) = \theta$, we have that

$$\frac{\partial \mathbb{P}_\theta(\mathcal{D}, \mathbf{Z})}{\partial \theta} = \left(T(\mathcal{D}, \mathbf{Z}) - \frac{\partial}{\partial \theta} A(\theta) \right) \mathbb{P}_\theta(\mathcal{D}, \mathbf{Z}).$$

EM for Exponential Family (II)

- For exponential families: $\mathbb{E}_\theta[T(\mathcal{D}, \mathbf{Z})] = \frac{\partial}{\partial\theta} A(\theta)$.

$$\frac{\partial \mathbb{P}_\theta(\mathcal{D}, \mathbf{Z})}{\partial \theta} = (T(\mathcal{D}, \mathbf{Z}) - \mathbb{E}_\theta[T(\mathcal{D}, \mathbf{Z})])\, \mathbb{P}_\theta(\mathcal{D}, \mathbf{Z}).$$

- Since $\frac{1}{\mathbb{P}_\theta(\mathcal{D})} \sum_z \frac{\partial \mathbb{P}_\theta(\mathcal{D}, \mathbf{Z}=z)}{\partial\theta} = 0$, we have that

$$\frac{1}{\mathbb{P}_\theta(\mathcal{D})} \sum_z (T(\mathcal{D}, \mathbf{Z}=z) - \mathbb{E}_\theta[T(\mathcal{D}, \mathbf{Z})])\, \mathbb{P}_\theta(\mathcal{D}, \mathbf{Z}=z) = 0$$

$$\sum_z T(\mathcal{D}, \mathbf{Z}=z) \frac{\mathbb{P}_\theta(\mathcal{D}, \mathbf{Z}=z)}{\mathbb{P}_\theta(\mathcal{D})} - \mathbb{E}_\theta[T(\mathcal{D}, \mathbf{Z})] = 0$$

$$\mathbb{E}_\theta[T(\mathcal{D}, \mathbf{Z}) | \mathcal{D}] - \mathbb{E}_\theta[T(\mathcal{D}, \mathbf{Z})] = 0$$

EM for Exponential Family (III)

- For the critical values of θ, the following condition is satisfied:

$$\mathbb{E}_\theta \left[T(\mathcal{D}, \mathbf{Z}) | \mathcal{D} \right] = \mathbb{E}_\theta \left[T(\mathcal{D}, \mathbf{Z}) \right].$$

- **How is this related to the EM objective $Q(\theta, \hat{\theta}^{(t)})$?**

$$\begin{aligned} Q(\theta, \hat{\theta}^{(t)}) &= \sum_{\mathbf{z}} \log \left(\mathbb{P}_\theta(\mathbf{Z} = \mathbf{z}, \mathcal{D}) \right) \cdot \mathcal{P}_{\hat{\theta}^{(t)}}(\mathbf{Z} | \mathcal{D}) \\ &= \theta \, \mathbb{E}_{\hat{\theta}^{(t)}} \left[T(\mathcal{D}, \mathbf{Z}) | \mathcal{D} \right] - A(\theta) + \text{Constant} \\ &= \theta \, \mathbb{E}_{\hat{\theta}^{(t)}} \left[T(\mathcal{D}, \mathbf{Z}) | \mathcal{D} \right] - \mathbb{E}_\theta \left[T(\mathcal{D}, \mathbf{Z}) \right] + \text{Constant}. \end{aligned}$$

- $\frac{\partial Q(\theta, \hat{\theta}^{(t)})}{\partial \theta} = 0 \rightarrow \mathbb{E}_{\hat{\theta}^{(t)}} \left[T(\mathcal{D}, \mathbf{Z}) | \mathcal{D} \right] = \mathbb{E}_\theta \left[T(\mathcal{D}, \mathbf{Z}) \right].$
- Since it is difficult to solve the above equation analytically, the EM algorithm solves for θ via *successive approximations*, i.e. solve the following for $\hat{\theta}^{(t+1)}$:

$$\mathbb{E}_{\hat{\theta}^{(t)}} \left[T(\mathcal{D}, \mathbf{Z}) | \mathcal{D} \right] = \mathbb{E}_{\hat{\theta}^{(t+1)}} \left[T(\mathcal{D}, \mathbf{Z}) \right].$$

Example: Multivariate Gaussian Mixtures

- **Parameters for a mixture of K Gaussians:** mixture proportions $\{\pi_k\}_{k=1}^K$, mean vectors and covariance matrices $\{(\mu_k, \Sigma_k)\}_{k=1}^K$.

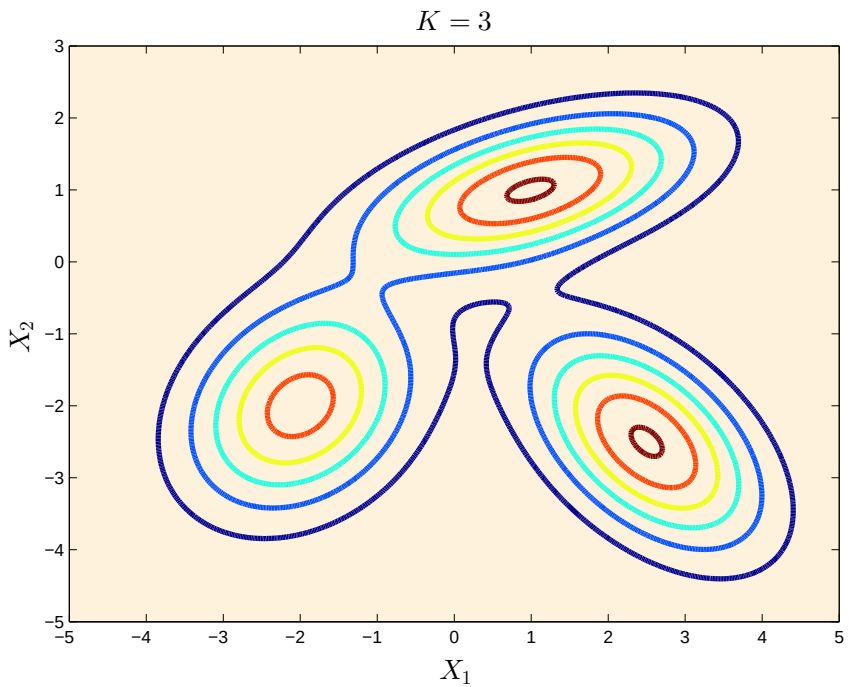


Figure: Contour plot for the density of a mixture of 3 bivariate Gaussian distributions.

The Generative Process

- $Z_i = z \sim \text{Categorical}(\pi_1, \ldots, \pi_K)$, and $X_i \sim \mathcal{N}(\mu_z, \Sigma_z)$.

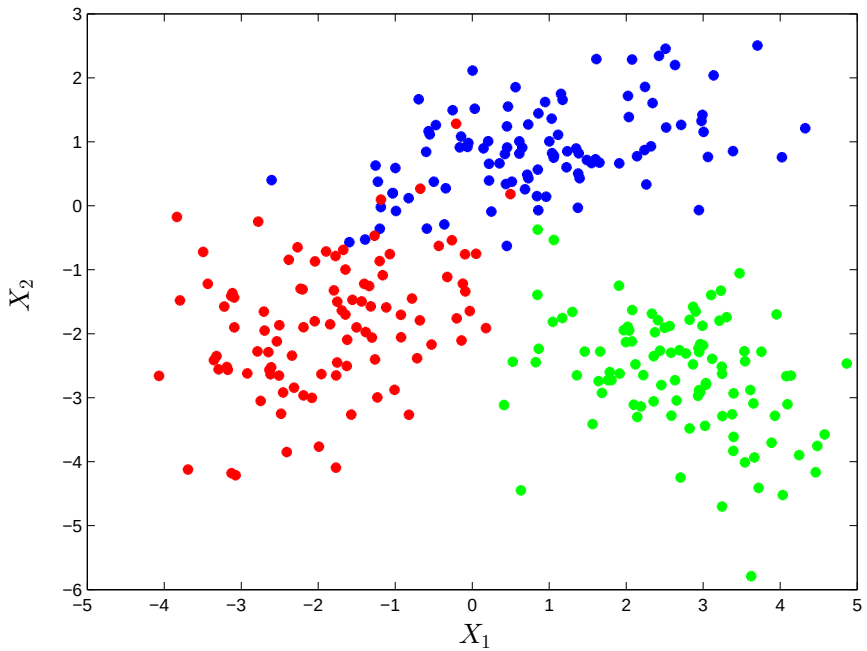
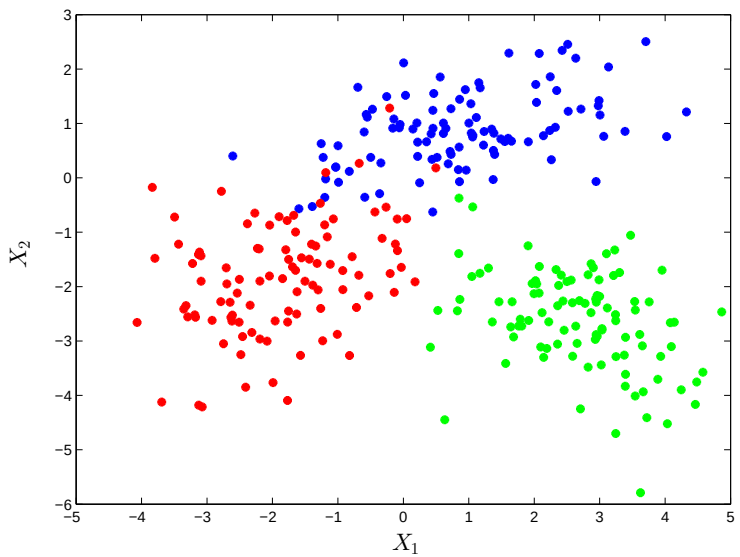


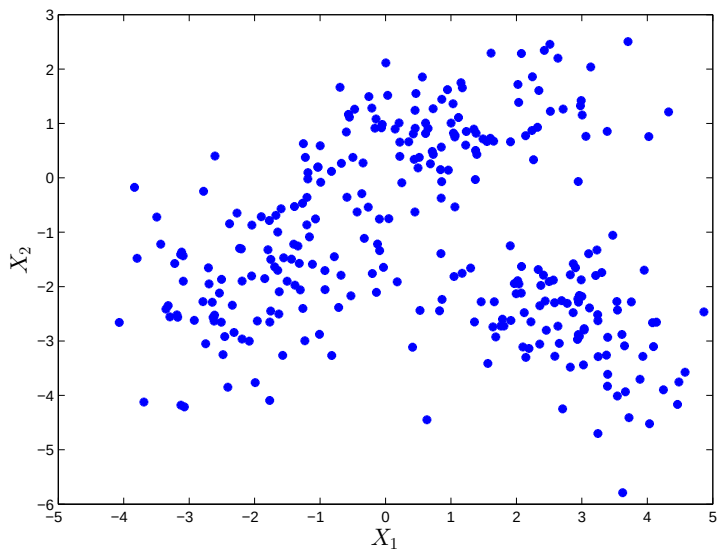
Figure: A sample from a mixture model: every data point is colored according to its component membership.

The Dataset

- Need to learn the parameters $(\pi_k, \mu_k, \Sigma_k)_{k=1}^K$ from the data points $\mathcal{D} = (X_1, \ldots, X_n)$ that are not "colored by the component memberships", i.e. we do not observe the latent variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$.



(a) $(\mathcal{D}, \mathbf{Z})$: the data points and their component memberships.



(b) $\mathcal{D}$: the dataset with the observed data points (component memberships are latent).

MLE for the Gaussian Mixture Models

- The complete-data likelihood function is given by

$$\mathbb{P}_\theta(\mathcal{D}, \mathbf{Z}) = \prod_{i=1}^{n} \pi_{z_i} \cdot \mathcal{N}(X_i|\mu_{z_i}, \Sigma_{z_i}).$$

- The marginal likelihood function is

$$\mathbb{P}_\theta(\mathcal{D}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(X_i|\mu_k, \Sigma_k).$$

- The MLE can be obtained by maximizing the marginal log likelihood function:

$$\hat{\theta}_n^* = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(X_i|\mu_k, \Sigma_k) \right).$$

- **Exercise:** Is the objective function above concave?

Implementing EM for the Gaussian Mixture Model (I)

- The *expected* complete-data log likelihood function is

$$\mathbb{E}_{\mathbf{Z}}\left[\mathbb{P}_\theta(\mathcal{D}, \mathbf{Z})\right] = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma(k, X_i|\theta)\left(\log(\pi_k) + \log\left(\mathcal{N}(X_i|\mu_k, \Sigma_k)\right)\right)$$

$$\gamma(k, X_i|\theta) = \mathbb{P}_\theta(Z_i = k|X_i).$$

- $\gamma(k, X_i|\theta)$ is called the **responsibility** of component k towards data point X_i

$$\gamma(k, X_i|\theta) = \frac{\pi_k \cdot \mathcal{N}(X_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(X_i|\mu_j, \Sigma_j)}$$

- **Try to work out the derivation above yourself!**

Implementing EM for the Gaussian Mixture Model (II)

- **(E-step)** Approximate expected complete-data likelihood by fixing the responsibilities $\gamma(k, X_i|\theta)$ using the parameter estimates obtained from the previous iteration.

$$Q(\theta, \hat{\theta}^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma(k, X_i|\hat{\theta}^{(t)}) \left(\log(\pi_k) + \log\left(\mathcal{N}(X_i|\mu_k, \Sigma_k)\right)\right)$$

$$\gamma(k, X_i|\hat{\theta}^{(t)}) = \frac{\hat{\pi}_k^{(t)} \cdot \mathcal{N}(X_i|\hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)})}{\sum_{j=1}^{K} \hat{\pi}_j^{(t)} \cdot \mathcal{N}(X_i|\hat{\mu}_j^{(t)}, \hat{\Sigma}_j^{(t)})}.$$

- **(M-step)** Solve a tractable optimization problem

$$(\hat{\pi}^{(t+1)}, \hat{\mu}^{(t+1)}, \hat{\Sigma}^{(t+1)}) =$$

$$\arg\max_{(\pi,\mu,\Sigma)} \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma(k, X_i|\hat{\theta}^{(t)}) \left(\log(\pi_k) + \log\left(\mathcal{N}(X_i|\mu_k, \Sigma_k)\right)\right)$$

Implementing EM for the Gaussian Mixture Model (III)

- The **(M-step)** yields the following parameter updating equations

$$\hat{\pi}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma(k, X_i | \hat{\theta}^{(t)})$$

$$\hat{\mu}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} X_i \cdot \gamma(k, X_i | \hat{\theta}^{(t)})$$

$$\hat{\Sigma}_k^{(t+1)} = \sum_{i=1}^{n} \frac{\gamma(k, X_i | \hat{\theta}^{(t)})}{\sum_{j=1}^{n} \gamma(k, X_j | \hat{\theta}^{(t)})} (X_i - \hat{\mu}_k^{(t+1)})(X_i - \hat{\mu}_k^{(t+1)})^T$$

- **Try to work out the updating equations by yourself!**

EM in Practice

Consider a Gaussian mixture model with $K = 3$, and the following parameters:

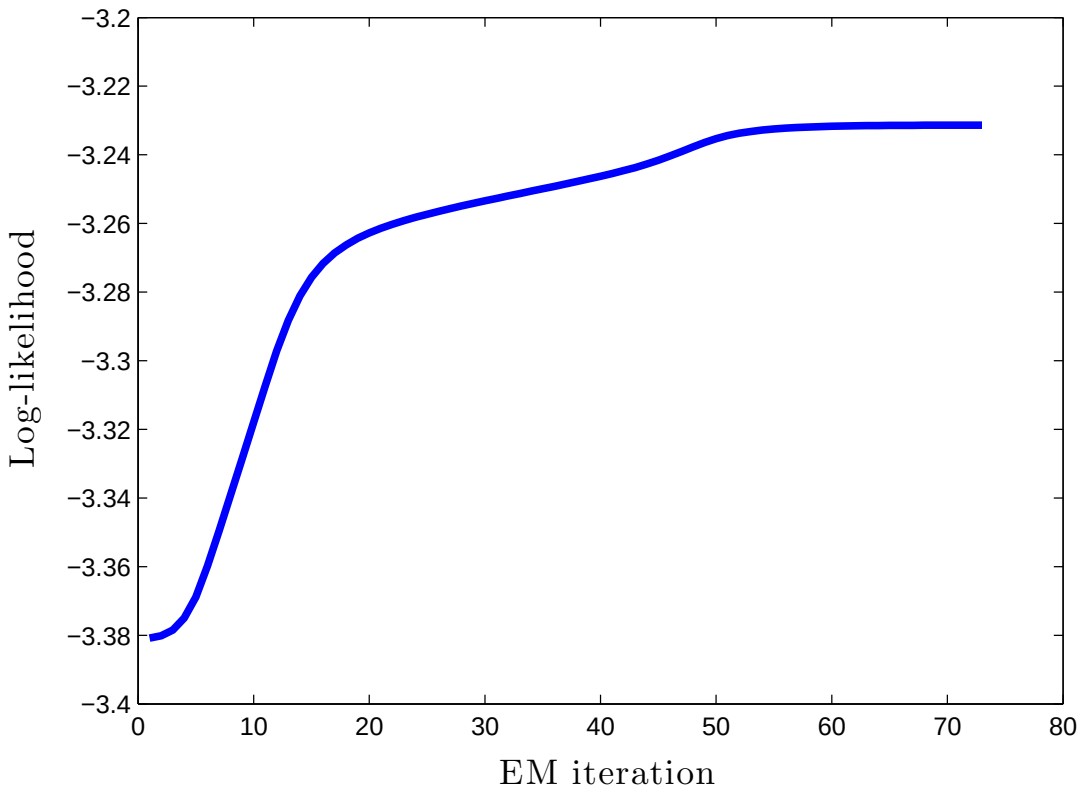
- $\pi_1 = 0.6, \pi_2 = 0.05$, and $\pi_3 = 0.35$.
- $\mu_1 = [-1.4, 1.8]^T, \mu_2 = [-1.4, -2.8]^T, \mu_3 = [-1.9, 0.55]^T$.

$$\Sigma_1 = \begin{bmatrix} 0.8 & -0.8 \\ -0.8 & 4 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1.2 & 2.3 \\ 2.3 & 5.2 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.4 & -0.01 \\ -0.01 & 0.35 \end{bmatrix}$$

- **Try writing a MATLAB code that generates a random dataset of 5000 data points drawn from the model specified above, and implement the EM algorithm to learn the model parameters from this dataset.**

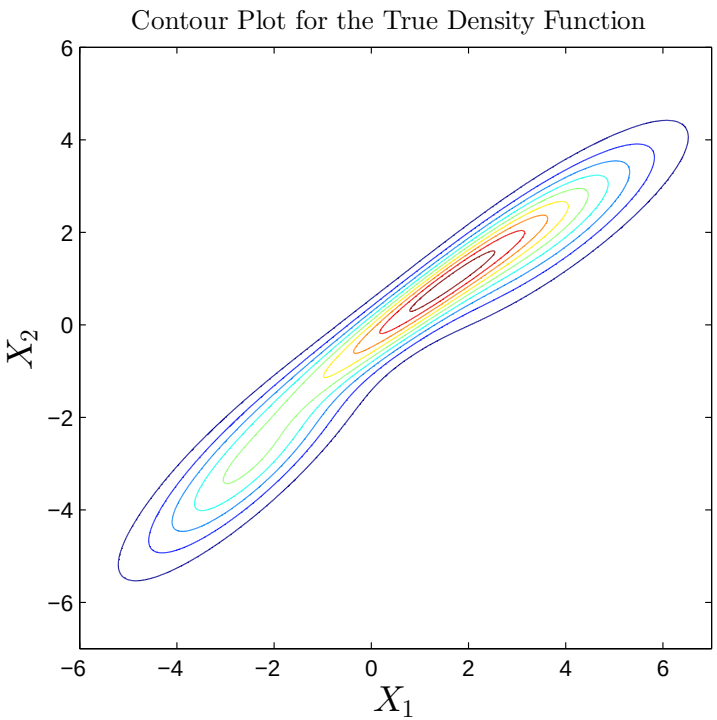
EM in Practice

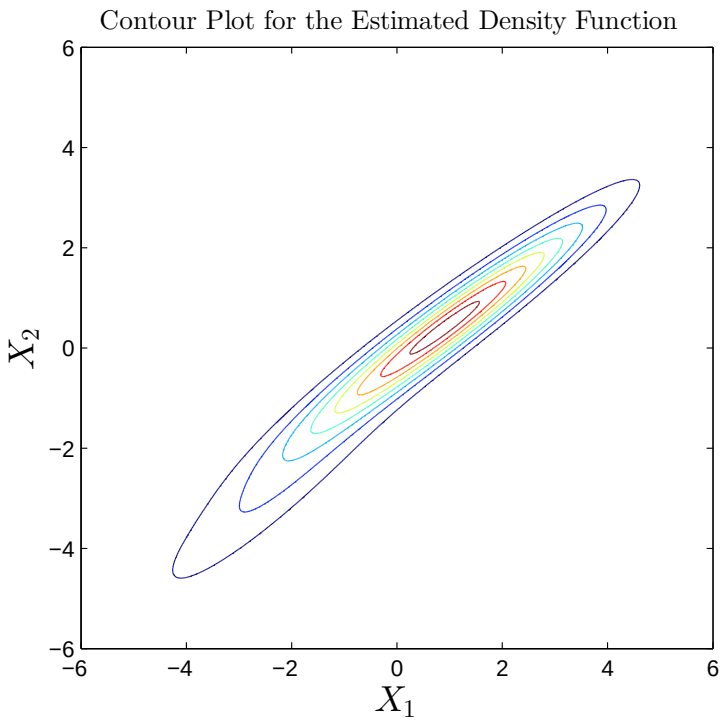
- The complete-data log likelihood increases after every EM iteration!
- This means that every new iteration finds a "better" estimate!



EM in Practice

- Compare the true density function with the estimated one.





What Does EM Guarantee?

- The EM algorithm does not guarantee that $\hat{\theta}^{(t)}$ will converge to $\hat{\theta}_n^*$.
- **EM guarantees the following:**
  - $\hat{\theta}^{(t)}$ **always converges (to a local optimum).**
  - **Every iteration improves the marginal likelihood** $\mathbb{P}_{\hat{\theta}^{(t)}}(\mathcal{D})$.

Does the Initial Value Matter?

1. The initial value $\theta^{(o)}$ affects the speed of convergence and the value of $\theta^{(\infty)}$! Smart initialization methods are often needed.
2. The K-means algorithm is often used to initialize the parameters in a Gaussian mixture model before applying the EM algorithm.

References

1. Robert W Keener, "Statistical theory: notes for a course in theoretical statistics," 2006.
2. Robert W Keener, "Theoretical Statistics: Topics for a Core Course," 2010.
3. Christopher Bishop, "Pattern Recognition and Machine Learning," 2007.