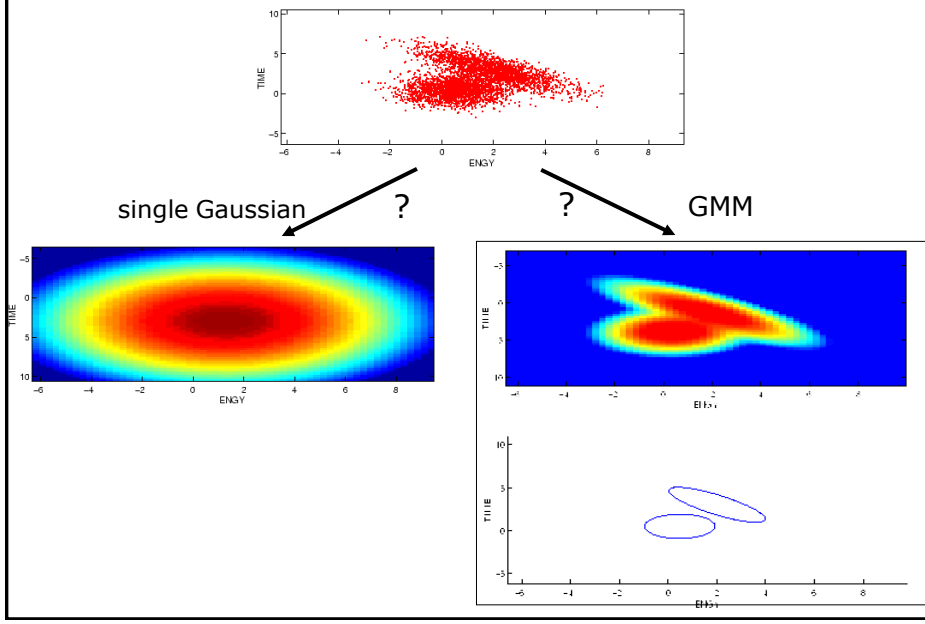
Density Estimation

- Parametric techniques
  - Maximum Likelihood
  - Maximum A Posteriori
  - Bayesian Inference
  - **G**aussian **M**ixture **M**odels (GMM)
    - EM-Algorithm
- Non-parametric techniques
  - Histogram
  - Parzen Windows
  - k-nearest-neighbor rule

GMM Applications



GMM Applications

Density estimation

Observed data from a complex but unknown probability distribution.

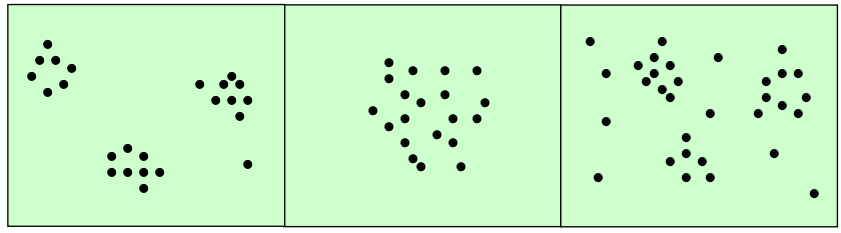
Can we describe this data with a few parameters ?

Which (new) samples are unlikely to come from this unknown distribution (Outlier detection)?

GMM Applications

Clustering

Observations from K classes. Each class produces samples from a multivariate normal distribution. Which observations belong to which class ?



Sometimes easy

Sometimes impossible

Often possible but not clear-cut

GMM: Definition

- Mixture models are linear combinations of densities:

$$p(x \mid \Theta) = \sum_{i=1}^{K} c_i \, p(x \mid \Theta_i)$$

$$\text{with} \quad \sum_{i=1}^{K} c_i = 1 \quad , \quad \int_x p(x \mid \Theta_i) dx = 1$$

  - Capable of approximating almost any complex and irregularly shaped distributions (*K might get big*)!

- For Gaussian mixtures:

$$\Theta_i = \{\mu_i, \Sigma_i\}, \quad \Rightarrow \quad p(x \mid \Theta_i) \equiv N(\mu_i, \Sigma_i)$$

Sampling a GMM

- How to generate a random variable according to a known GMM $p(x) = \sum_i^K c_i \, N(\mu_i, \Sigma_i)$

Assume that each data point is generated according to the following recipe:

1. Pick a component ($i \in [\, 1 \,..\, K \,]$) at random. Choose component i with probability c_i .
2. Sample data point ~ $N(\mu_i, \Sigma_i)$.

In the end, we might not know which data points came from which component (unless someone kept track during the sampling process)!

Learning a GMM

Recall ML-estimation

We have:

A density function $p(\cdot\,; \Theta)$ governed by a set of unknown parameters Θ.
A data set of size N drawn from this distribution
$X = \{x_1, \ldots, x_N\}$

We wish:

to obtain the parameters best explaining data X
by maximizing the log-likelihood function:

$$L(\Theta) = \ln p(X;\Theta)$$
$$\Theta^* = \arg\max_{\Theta} L(\Theta)$$

Learning a GMM

- For a single Gaussian distribution this is simple to solve. We have an analytical solution.
- Unfortunately for many problems (including GMM) it is not possible to find analytical expressions.
  - Resort to classical optimization techniques ?
  - Possible but there is a better way:

    EM – Algorithm (Expectation-Maximization)

Expectation Maximization (**EM**)

- General method for finding ML-estimates in the case of incomplete or missing data (GMM's are one application).
- Usually used when:
  - the observation is actually incomplete; some values are missing from the data set.
  - the likelihood function is analytically intractable but can be simplified by assuming the existence of additional but missing (so-called hidden/latent) parameters.

The latter technique is used for GMMs. Think of each data point as having a hidden label specifying the component it belongs to. These component labels are the latent parameters.

General EM procedure

The EM setting

Observed data set (*incomplete*): X

Assume a complete data set exists: $\mathcal{Z} = (\mathrm{X}, \mathrm{Y})$

$\mathcal{Z}$ has a joint density function:

$$p(\mathbf{z}\,|\,\Theta) \equiv p(\mathbf{x},\mathbf{y}\,|\,\Theta) \;= p(\mathbf{y}\,|\,\mathbf{x},\Theta)\cdot p(\mathbf{x}\,|\,\Theta)$$

Define the *complete-data* log-likelihood function:

$$\mathrm{L}(\Theta\,|\,\mathcal{Z}) = \mathrm{L}(\Theta\,|\,\mathrm{X},\mathrm{Y}) = \ln p(\mathrm{X},\mathrm{Y}\,|\,\Theta)$$

Our aim is to find a Θ that maximizes this function.

General EM procedure

- But: We cannot simply maximize $\mathrm{L}(\Theta \mid \mathrm{X},\mathrm{Y}) = ln\,\mathrm{p}(\mathrm{X},\mathrm{Y} \mid \Theta)$ because Y is not known.
  - $L\,(\Theta|\mathrm{X},\mathrm{Y})$ is in fact a random variable:
    - Y can be assumed to come from some distribution $f(\mathbf{y}|\,\mathrm{X},\Theta)$
    - That is, $L(\Theta|\mathrm{X},\mathrm{Y})$ can be interpreted as a function where X and Θ are constant and Y is a random variable.
- The EM will compute a new, auxiliary function, based on L, that can be maximized instead.
- Let's assume we already have a reasonable estimate for the parameters: $\Theta^{(i-1)}$.

General EM procedure

- EM uses an auxiliary function:

$$Q(\Theta,\Theta^{(i-1)}) = E\left[\ln p(\mathrm{X},\mathrm{Y} \mid \Theta) \,\middle|\, \mathrm{X},\Theta^{(i-1)}\right]$$

How to read this:

- X and $\Theta^{(i-1)}$ are constants,
- Θ is a simple variable (the function argument),
- Y is a random variable governed by distribution *f* .

- The task is to rewrite Q and perform some calculations to make it a **fully determined** function.
- Q is the **expected value** of the complete-data log-likelihood w.r.t. to missing data (Y), observed data (X) and current parameter estimates ($\Theta^{(i-1)}$).

This is called the **E-step** (*expectation-step*)

General EM procedure

- Q can be rewritten by means of the *marginal* distribution *f*:

If y is a continuous random variable:

$$Q(\Theta,\Theta^{(i-1)}) = E\left[\ln p(\mathrm{X},\mathrm{Y}\mid\Theta)\,\middle|\,\mathrm{X},\Theta^{(i-1)}\right] = \int_{\mathbf{y}} \ln p(\mathrm{X},\mathbf{y}\mid\Theta)\cdot f(\mathbf{y}\mid\mathrm{X},\Theta^{(i-1)})\,dy$$

If y is a discrete random variable:

$$Q(\Theta,\Theta^{(i-1)}) = E\left[\ln p(\mathrm{X},\mathrm{Y}\mid\Theta)\,\middle|\,\mathrm{X},\Theta^{(i-1)}\right] = \sum_{y} \ln p(\mathrm{X},\mathbf{y}\mid\Theta)\, f(\mathbf{y}\mid\mathrm{X},\Theta^{(i-1)})$$

Think of this as the expected value of a function of Y

E[g(Y)]

Evaluate $f(\boldsymbol{y} \mid \mathrm{X}, \Theta^{(i-1)})$, using the current estimate $\Theta^{(i-1)}$.

Now Q is fully determined and we can use it!

General EM procedure

- In a **second** step Q is used to obtain a better set of parameters Θ:

$$\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta,\Theta^{(i-1)})$$

This is called the **M-step** (*maximization-step*)

- Both E- and M-steps are repeated until convergence.
  - In each E-Step, we find a new auxiliary function Q
  - In each M-Step, we find a new parameter set Θ

General EM algorithm

Summary of the general EM algorithm (see also Bishop, p.440)

1. Choose an initial setting for the parameters $\Theta^{(i-1)}$.

2. E-step: evaluate $f(y \mid X, \Theta^{(i-1)})$,
   plug it into $Q(\Theta, \Theta^{(i-1)}) = \int_y f(y \mid X, \Theta^{(i-1)}) \ln p(X, Y \mid \Theta) dy$
   to obtain a fully determined auxiliary function

3. M-step: evaluate $\Theta^{(i)}$ given by $\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)})$

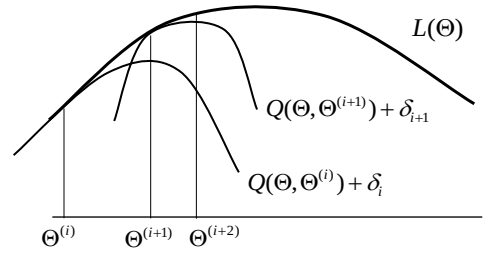
4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let $\Theta^{(i-1)} \leftarrow \Theta^{(i)}$ and return to step 2.

General EM Illustration

<u>Iterative majorisation</u>

Aim of EM: Find local maximum of function $L(\Theta)$ by using auxiliary function $Q(\Theta, \Theta^{(i)})$.

How does this work?



- Q touches L at point $[\Theta^{(i)}, L(\Theta^{(i)})]$ and lies everywhere below L .
- Maximize auxiliary function.
- The position of the maximum $\Theta^{(i+1)}$ gives a value of L which is greater than in the previous iteration.
- Repeat this scheme with new auxiliary function until convergence.

General EM Summary

- Iterative algorithm for ML-estimation of systems with hidden/missing values.
- Calculates expectance for hidden values based on observed data and joint distribution.
- Slow but guaranteed convergence.
- May get „stuck" in local maximum.
- There is no general EM implementation. The details of both steps depend very much on the particular application.

Application: EM for Mixture Models

- Our probabilistic model is now:

$$p(x \mid \Theta) = \sum_{i=1}^{M} c_i \, p_i(x \mid \theta_i)$$

with parameters: $\Theta = (c_1, \ldots, c_M, \theta_1, \ldots, \theta_M)$

such that: $\sum_{i=1}^{M} c_i = 1$

- That is, we have M component densities p_i (of the same family) combined through M mixing coefficients c_i.

EM for Mixture Models

- The incomplete-data log-likelihood becomes (remember we assume X is iid):

$$L(\Theta \mid X) = \ln \prod_{i=1}^{N} p(x_i \mid \Theta) = \sum_{i=1}^{N} \ln \left(\sum_{j=1}^{M} c_j \, p_j(x_i \mid \theta_j) \right)$$

- Difficult to optimize with log of sum

- Now let's try the EM-trick:
  - Consider X as incomplete.
  - Introduce unobserved data $Y = \{y_i\}_{i=1}^{N}$ whose values indicate which component of the MM generated each data item.
  - That is, $y_i \in 1,\ldots,M$ and $y_i = k$ if the i-th sample stems from the k-th component.

EM for Mixture Models

- If we knew the values of Y, the log likelihood would simplify to:

$$L(\Theta \mid X,Y) = \ln p(X,Y \mid \Theta)$$
$$= \sum_{i=1}^{N} \ln \left(p(x_i \mid y_i, \Theta) \, p(y_i \mid \Theta) \right) = \sum_{i=1}^{N} \ln \left(c_{y_i} p_{y_i}(x_i \mid \theta_{y_i}) \right)$$

Could apply standard optimization techniques

- But we don't know Y, so we follow the EM-procedure:
  1. Start with an initial guess of the mixture parameters:
     $\Theta^g = (c_1^g, \ldots, c_M^g, \theta_1^g, \ldots, \theta_M^g)$
  2. Find an expression for the marginal density function of the unobserved data $p(\boldsymbol{y} \mid X, \Theta)$:

EM for Mixture Models

$$p(\mathbf{y} \mid X, \Theta^g) = \prod_{i=1}^{N} p(y_i \mid x_i, \Theta^g)$$

y_i is the (unknown) component label of data point x_i.

Using Bayes's rule, we get:

$$p(y_i \mid x_i, \Theta^g) = \frac{p(x_i \mid y_i, \Theta^g)\, p(y_i)}{p(x_i \mid \Theta^g)}$$

$$= \frac{c_{y_i} p_{y_i}(x_i \mid \theta^g_{y_i})}{p(x_i \mid \Theta^g)} = \frac{c_{y_i} p_{y_i}(x_i \mid \theta^g_{y_i})}{\sum_{k=1}^{M} c_k p_k(x_i \mid \theta^g_k)}$$

- Using guessed parameters, we obtained the desired marginal density function.
- This can now be substituted in Q (i.e. in the E-step).

EM for **Gaussian** Mixtures

- For our mixture model, the **E-step** is:

$$Q(\Theta, \Theta^g) = \sum_{\mathbf{y}} \ln(L(\Theta \mid X, \mathbf{y}))\, p(\mathbf{y} \mid X, \Theta^g)$$

Substituted marginal hidden data density

- The **M-step** is to find a parameter set Θ^{new} that maximizes Q.
- But for Gaussian mixtures, there is no need to deal with Q in the above form!

$$\Theta^{new} = \underset{\Theta}{\operatorname{argmax}} \sum_{y} L(\Theta \mid X, y)\, p(y \mid X, \Theta^g)$$

Here it is not necessary to deal with this directly

- Instead, a set of simple formulas for updating Θ can be used.

EM for **Gaussian** Mixtures

3. Compute parameters Θ^{new} , using update formulas (perform E- and M-step simultaneously):

Update formulas

$$c_k^{new} = \frac{1}{N}\sum_{i=1}^{N} p(k \mid x_i, \Theta^g) \qquad \mu_k^{new} = \frac{\sum_{i=1}^{N} x_i\, p(k \mid x_i, \Theta^g)}{\sum_{i=1}^{N} p(k \mid x_i, \Theta^g)}$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^{N} p(k \mid x_i, \Theta^g)(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T}{\sum_{i=1}^{N} p(k \mid x_i, \Theta^g)}$$

Plug in the expression found in previous step
(k = label of the k-th component)

These formulas are derived from $Q(\Theta, \Theta^g)$.

EM for Mixture Models

Derivation of the update formulas

Q in its initial form:

$$Q(\Theta, \Theta^g) = \sum_{\mathbf{y}} \ln(\mathrm{L}(\Theta \mid \mathcal{X}, \mathbf{y}))\, p(\mathbf{y} \mid \mathcal{X}, \Theta^g)$$

Substituted marginal hidden data density

- After a lot of simplification we arrive at an equation where the c_k and θ_k are expressed independently:

$$Q(\Theta, \Theta^g) = \sum_{k=1}^{M}\sum_{i=1}^{N} \ln(c_k)\, p(k \mid x_i, \Theta^g) + \sum_{k=1}^{M}\sum_{i=1}^{N} \ln(p_k(x_i \mid \theta_k))\, p(k \mid x_i, \Theta^g)$$

Get formula for c_k from this part

Get formulas for θ_k from this part

Formula for c_k, after further simplification

with $$c_k = \frac{1}{N}\sum_{i=1}^{N} p(k \mid x_i, \Theta^g)$$

EM for Gaussian Mixtures

- Formula for c_k (previous slide) is valid for any mixture model, not just Gaussian.
- Formulas for θ_k will be specific to the Gaussian mixture.
- For a d-dimensional Gaussian component, use

$$\theta = (\mu, \Sigma) \qquad p_k(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

Plug this into the expression on the previous slide

- Take the derivatives of the resulting expression with respect to μ_k and Σ_k *(very technical).*
- Set the derivatives to zero, then solve for μ_k and Σ_k.

  → The results are the update formulas for μ_k^{new} and Σ_k^{new}.

EM for Gaussian Mixture Models

Summary of the algorithm for GMM (see Bishop, p.438):

1. Initialize the parameters $\Theta^{old} = (c_1 \ldots c_M, \mu_1 \ldots \mu_M, \Sigma_1 \ldots \Sigma_M)$

2. E-step: evaluate the responsibilities of each component for all data points:

$$p(k \mid x_i, \Theta^{old}) = \frac{c_k p_k(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{M} c_j p_j(x_i \mid \mu_j, \Sigma_j)},$$

Responsibility of the k-th component for the i-th data point

No need to compute $Q(\Theta, \Theta^{(i-1)})$ explicitly!

EM for Gaussian Mixture Models

3. E-step/M-step: Update the parameters

$$c_k^{new} = \frac{1}{N}\sum_{i=1}^{N} p(k \mid x_i, \Theta^{old}) \qquad \mu_k^{new} = \frac{\sum_{i=1}^{N} p(k \mid x_i, \Theta^{old}) x_i}{\sum_{i=1}^{N} p(k \mid x_i, \Theta^{old})}$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^{N} p(k \mid x_i, \Theta^{old})(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T}{\sum_{i=1}^{N} p(k \mid x_i, \Theta^{old})}$$

4. Evaluate the log likelihood

$$\ln p(X \mid \Theta) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{M} c_k p_k(x_i \mid \mu_k, \Sigma_k)\right)$$

and check it for convergence. If the convergence criterion is not satisfied, return to step 2.

Relation to k-means

- Let $c_k = 1/M$ and $\Sigma_k = \sigma^2 I$
- k-means procedure:
  1. Random initialize M cluster centers.
  2. Assign each data point to a cluster according to the minimum distance criterion:

$$p(k \mid x_i) = \begin{cases} 1 & \text{if } \forall j \| x_i - \mu_k \| \le \| x_i - \mu_j \| \\ 0 & \text{otherwise} \end{cases}$$

  3. Re-calculate cluster centers:

$$\mu_k^{new} = \frac{\sum_{i=1}^{N} p(k \mid x_i) x_i}{\sum_{i=1}^{N} p(k \mid x_i)}$$

  4. Go to step 2 until no change in cluster centers.

Relation to k-means

- GMM is referred to as soft clustering
  - Probability $p(k \mid x_i, \Theta)$ indicates the responsibility of the k-th component for the i-th observation (i.e. the posterior prob. that the i-th observation comes from the k-th component).
  - For each point x_i, GMM produces smooth posterior. From these, one can find cluster label

    for each x_i: $C(x_i) = \underset{k}{\operatorname{argmax}}\, p(k \mid x_i, \Theta)$

- k-means is a hard clustering method
  - The responsibility of can only be 1 or 0 .

GMM: Open questions

- How many components are required ?
  - Answer is highly problem dependent.
  - One possibility: Try different numbers, then choose model (number) which gives best performance on a validation data set .
- Which initial parameters to use ?
  - Same here: in general we don't know where to look for global maximum.
  - Obvious approaches:
    1. Perform k-means to obtain initial μ's.
    2. Try different random values and choose the ones which lead to maximal likelihood.

GMM/EM Resources

- **J. A. Bilmes et al : A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models (1998)**
- **GMMBAYES - Gaussian Mixture Model Methods Matlab-Toolbox**
  http://www.it.lut.fi/project/gmmbayes/
- **Gaussian Mixtures Demo Applet**
  http://lcn.epfl.ch/tutorial/english/gaussian/html/index.html