Manifold Learning for Signal and Visual Processing
Lecture 8: Gaussian Mixtures, the EM Algorithm, Model-Based Clustering

Radu Horaud
INRIA Grenoble Rhone-Alpes, France
Radu.Horaud@inrialpes.fr
`http://perception.inrialpes.fr/`

Outline of Lecture 8

- Probabilities, densities;
- Expectations, covariance, and correlation;
- The Gaussian distribution – univariate case;
- What is the curse of dimensionality?
- The multivariate Gaussian distribution;
- The Gaussian mixture model (GMM);
- The Expectation-Maximization algorithm for Gaussian mixtures.

Material for This Lecture

- C. Fraley and A. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. J. of the Am. Stat. Ass. June 2002
- C. M. Bishop. Pattern Recognition and Machine Learning. 2006. (Chapters 1, 2, & 9)
- R. Horaud, F. Forbes, M. Yguel, G. Dewaele, and J. Zhang. Rigid and Articulated Point Registration with Expectation Conditional Maximization. IEEE Trans. on Patt. An. and Mach. Intell. March 2011. (the derivation of EM is based on this paper).
- Software: MCLUST package (in R) `http://www.stat.washington.edu/mclust/`

Probability Theory (Discrete Random Variables)

- Sum rule:

$$P(X) = \sum_i P(X, Y_i)$$

- Product rule:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

- Bayes:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_i P(X|Y_i)P(Y_i)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalization}}$$

Probability Densities (Continuous Random Variables)

- Probability that x lies in an interval:

$$p(x \in (a,b)) = \int_a^b p(x)dx$$

- $p(x)$ is called the *probability density* over x.
- $p(x) \geq 1$, $p(x \in (-\infty, \infty)) = 1$
- nonlinear change of variable $x = g(y)$:

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right|$$

- *cumulative distribution function*: $P(z) = p(x \in (-\infty, z))$
- sum and product rules extend to probability densities.

Expectation of a Scalar Function

- Expectation: the average value of some function $f(x)$ under a probability distribution $p(x)$;
- The continuous case: $E[f] = \int p(x) f(x) dx$
- The discrete case: $E[f] = \sum_i p(x_i) f(x_i)$
- Empirical expectation:

$$\overline{f} = E[f] \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

- Functions of several variables: $E_x[f] = \sum_i p(x_i) f(x_i, y)$
- Expectation over two variables: $E_{x,y}[f(x,y)] = \sum_i \sum_j p(x_i, y_j) f(x_i, y_j)$
- Conditional expectation: $E_x[f|y] = \sum_i p(x_i|y) f(x_i)$

Variance and Covariance

- *Variance* of $f(x)$: a measure of the variations of $f(x)$ around $E[f]$.
- Definition: $var[f] = E[(f(x) - E[f(x)])^2] = E[f^2] - E[f]^2$
- The variance of a scalar random variable x: $var[x] = E[x^2] - E[x]^2$
- The standard deviation: $\sigma_x = \sqrt{var[x]}$
- *Covariance* for two random variables: $cov[x, y] = E[(x - E[x])(y - E[y])] = E[xy] - E[x]E[y]$

Covariance Matrices

Consider two vectors of random variables, $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^{D'}$. The corresponding covariance matrices are:

- a *non symmetric* $D \times D'$ (rectangular) matrix:

$$\begin{aligned}\boldsymbol{\Sigma}_{xy} = cov[\boldsymbol{x},\boldsymbol{y}] &= E_{x,y}[(\boldsymbol{x} - E[\boldsymbol{x}])(\boldsymbol{y}^\top - E[\boldsymbol{y}^\top])] \\ &= E_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - E[\boldsymbol{x}]E[\boldsymbol{y}^\top]\end{aligned}$$

- notice that:

$$\boldsymbol{\Sigma}_{yx} = \boldsymbol{\Sigma}_{xy}^\top$$

- and a semi-definite positive symmetric $D \times D$ matrix:

$$\begin{aligned}\boldsymbol{\Sigma}_{xx} = cov[\boldsymbol{x},\boldsymbol{x}] &= E_{x,x}[(\boldsymbol{x} - E[\boldsymbol{x}])(\boldsymbol{x}^\top - E[\boldsymbol{x}^\top])] \\ &= E_{x,x}[\boldsymbol{x}\boldsymbol{x}^\top] - E[\boldsymbol{x}]E[\boldsymbol{x}^\top]\end{aligned}$$

Covariance Matrices

With the notations: $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_i \dots \boldsymbol{x}_n]$, $\mathbf{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_j \dots \boldsymbol{y}_n]$ and $\overline{\boldsymbol{x}} = E[\boldsymbol{x}], \overline{\boldsymbol{y}} = E[\boldsymbol{y}]$, we have:

$$\boldsymbol{\Sigma}_{xy} = \frac{1}{n}\mathbf{X}\mathbf{Y}^\top - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{y}}^\top$$

$$\boldsymbol{\Sigma}_{yx} = \frac{1}{n}\mathbf{Y}\mathbf{X}^\top - \overline{\boldsymbol{y}}\,\overline{\boldsymbol{x}}^\top$$

$$\boldsymbol{\Sigma}_{xx} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^\top$$

Joint Covariance Matrix

The covariance of the *joint* random variable $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$ of dimension $D + D'$:

$$\boldsymbol{\Sigma}_{xyxy} = E\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\begin{pmatrix} \boldsymbol{x}^\top & \boldsymbol{y}^\top \end{pmatrix}\right] - E\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] E\left[\begin{pmatrix} \boldsymbol{x}^\top & \boldsymbol{y}^\top \end{pmatrix}\right]$$

It is the symmetric $(D + D') \times (D + D')$ matrix:

$$\boldsymbol{\Sigma}_{xyxy} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}$$

Correlation

- Standardization of a scalar random variable x:

$$\hat{x} = \frac{x - E[x]}{\sigma_x}$$

- The covariance $cov[\hat{x}\hat{y}] = E[\hat{x}\hat{y}]$ is a measure of the correlation between the two variables:

$$\rho_{xy} = corr[x,y] = E_{x,y}\left[\frac{x - E[x]}{\sigma_x}\frac{y - E[y]}{\sigma_y}\right]$$

- Let $x_1 \dots x_i \dots x_n$ and $y_1 \dots y_i \dots y_n$ be n realizations of the random variables, then:

$$\rho_{xy} = \sum_i \hat{x}_i \hat{y}_i$$

Properties of Correlation

- Standardized data: $E[\hat{x}] = 0$ and $var[\hat{x}] = 1$
- The value ρ_{xy} is also known as the Pearson correlation coefficient.
- The following three conditions are equivalent (Taylor & Cristianini 2004):

  $\rho_{xy} = 1$; $\hat{x} = \hat{y}$; $y = ax + b$ for some $a > 0$ and b.
- $\rho_{xy} = -1$ if and only if $\hat{x} = -\hat{y}$;
- $\rho_{xy} = 0$ if the two variables are linearly uncorrelated.

The Correlation Matrix

- For two random vector variables $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^{D'}$ we obtain a $D \times D'$ correlation matrix with entries:

$$\rho_{ij} = \frac{1}{n} \sum_{k=1}^{n} \frac{x_{ki} - E[x_i]}{\sigma_{x_i}} \frac{y_{kj} - E[y_j]}{\sigma_{y_j}}$$

- with:

$$\rho_{ij} = \begin{cases} 1 & if \quad i = j \\ \in [-1; +1] & if \quad i \neq j \end{cases}$$

The Gaussian Distribution

- The Gaussian distribution of a single real-valued variable x:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- The mean: $E[x] = \mu$
- The variance: $\text{var}[x] = \sigma^2$
- in D dimensions: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) : \mathbb{R}^D \rightarrow \mathbb{R}$

The Maximum Likelihood Estimator

- $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a set of n observations of the SAME scalar random variable x
- Assume that this data set is *independent and identically distributed* (iid):

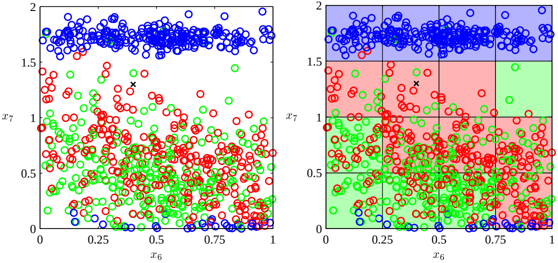
$$p(x_1, \ldots, x_n | \mu, \sigma^2) = \prod_{i=1}^{n} p(x_i | \mu, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(x_i | \mu, \sigma^2)$$

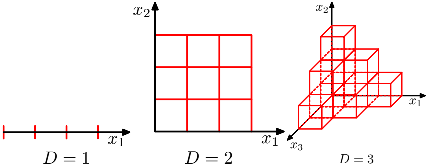
- $\max p$ is equivalent to $\max \ln(p)$ or $\min(-\ln(p))$
- $-\ln p(x_1, \ldots, x_N | \mu, \sigma^2) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \ldots$
- maximum likelihood solution: μ_{ML} and σ^2_{ML}
- MLE underestimates the variance: bias

The Curse of Dimensionality (from Bishop'06)

- curse: malédiction, fléau ...
- *Not all the intuitions developed in spaces of low dimensionality will generalize to spaces of many dimensions*
- Example 1: Training a classifier in high dimensions
- Example 2: How empty is a hypersphere?

Classification

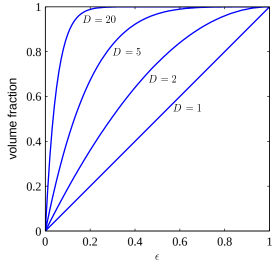




The “Empty” Hypersphere

- The volume of a sphere with radius r: $V_D(r) = K_D r^D$
- The fraction of the volume lying in between r and $r - \epsilon$:

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$



The Multivariate Gaussian Distribution

- The Gaussian distribution for a D-dimensional vector $\boldsymbol{x}$:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{D/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- The D-dimensional mean: $\boldsymbol{\mu}$
- The $D \times D$ covariance matrix: $\boldsymbol{\Sigma}$
- $|\boldsymbol{\Sigma}|$ denotes the determinant
- $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) : \mathbb{R}^D \rightarrow \mathbb{R}$

Analytical Properties (I)

- The quadratic form $\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is called the *Mahalanobis distance*.
- We already provided a geometric interpretation in Lecture #2. Assuming that $\boldsymbol{\Sigma}$ is non singular, we have the following spectral decomposition of its inverse:

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^\top \text{ and: } \Delta^2 = \sum_{i=1}^{D} \left(\frac{\boldsymbol{u}_i^\top (\boldsymbol{x} - \boldsymbol{\mu})}{\lambda^{1/2}} \right)^2$$

The quadratic form and hence the Gaussian density will be constant on hyper-ellipsoids with half eccentricities $\lambda_1^{1/2}, \ldots, \lambda_D^{1/2}$

Analytical Properties (II)

- With the change of variable $y_i = \boldsymbol{u}_i^\top (\boldsymbol{x} - \boldsymbol{\mu})$
- The quadratic form becomes: $\Delta^2 = \sum_{i=1}^D y_i^2 / \lambda_i$
- The covariance and its determinant: $\mathbf{\Sigma}_Y = \mathbf{\Lambda}, |\mathbf{\Sigma}_Y| = \prod_{i=1}^D \lambda_i$
- The multivariate Gaussian distribution writes:

$$\mathcal{N}(\boldsymbol{z}|\mathbf{\Lambda}) = \prod_{i=1}^{D} \frac{1}{(2\pi\lambda_i)^{1/2}} \exp\left(-\frac{1}{2}\frac{y_i^2}{\lambda_i}\right)$$

- which is the product of D *independent* univariate centred Gaussian distributions with variances λ_i.
- The number of free parameters of a Gaussian distribution: $D + D(D+1)/2 = D(D+3)/2$ hence it grows quadratically with D.
- Spherical or isotropic covariance: $\mathbf{\Lambda} = \lambda \mathbf{I} \leftrightarrow \mathbf{\Sigma} = \lambda \mathbf{I}$;

Maximum Likelihood (Multivariate Case)

- Consider a dataset $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_n]^\top$ in which the observations $\{\boldsymbol{x}_j\}$ are drawn independently from a multivariate Gaussian. The *negative log-likelihood function* writes:

$$-\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{nD}{2}\ln(2\pi) + \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})$$

- By taking the derivatives with respect to mean and covariance and setting these derivatives to zero, we obtain the ML mean:

$$\boldsymbol{\mu}_{ML} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$$

- as well as the *unbiased* covariance:

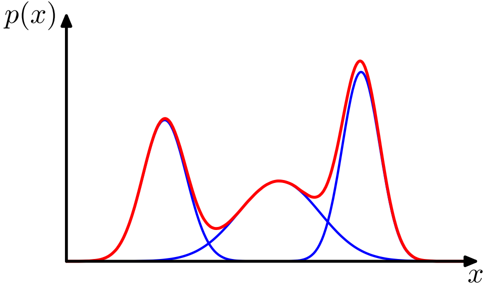
$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_j - \boldsymbol{\mu}_{ML})^\top$$

Gaussian Mixtures

- A mixture distribution: linear combinations of basic distributions, such as Gaussians.
- Consider a superposition of m Gaussian densities:

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

- Each Gaussian density k is a *component* of the mixture with its own mean and covariance.



The Mixing Coefficients

- If we integrate $p(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ and note that both $p(\boldsymbol{x})$ and the individual Gaussian components are normalized, we obtain:
$$\sum_{k=1}^{m} \pi_k = 1$$
- From $p(\boldsymbol{x}) \geq 0$ and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \geq 0$ we obtain that
$$0 \leq \pi_k \leq 1$$
- Therefore, the mixing coefficients are probabilities, i.e., π_k is the prior probability of the k-th component.

Probabilistic Interpretation

- Using the *sum* and *product* rules:

$$p(\boldsymbol{x}) = \sum_{k=1}^{m} p(\boldsymbol{x}, k) = \sum_{k=1}^{m} p(\boldsymbol{x}|k)p(k)$$

- By identification with the Gaussian mixture:

$$\begin{aligned} p(k) &= \pi_k \\ p(\boldsymbol{x}|k) &= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

- The posterior probabilities or *responsabilities*:

$$p(k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|k)p(k)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|k)p(k)}{\sum_{l=1}^{m} p(\boldsymbol{x}|l)p(l)}$$

The EM algorithm

- Introduced by Dempster, Laird, and Rubin in 1977 – **33,500 citations (22,800 in 2011)**
- EM is a maximum-likelihood estimator. In the case of Gaussian mixtures it estimates: (i) the mean and covariance of each component and (ii) the assignment of an observation $\boldsymbol{x}_j$ to a component (or a cluster) k.
- EM alternates between two steps:
  1. *E-step:* the conditional expectation of the complete-data log-likelihood given the the observed data and the current parameter estimates is computed;
  2. *M-step:* parameters that maximize the expected log-likelihood from the E-step are determined
- Under fairly mild regularity conditions, EM converges to a local maximum of the observed-data likelihood. These conditions do not always hold in practice.

Complete-Data: Observed Data + Missing Data

- The observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_n$;
- The missing (or unobserved) data: $\boldsymbol{Z} = (z_1, \ldots, z_j, \ldots, z_n)$ with:

$$z_j = \begin{cases} k & \text{if } \boldsymbol{x}_j \text{ belongs to group k} \\ 0 & \text{otherwise.} \end{cases}$$

- New notations:

$$\begin{aligned} P(z_j = k) &= \pi_k \\ P(\boldsymbol{x}_j | z_j = k) &= \mathcal{N}(\boldsymbol{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ P(z_j = k | \boldsymbol{x}_j) &= \frac{P(\boldsymbol{x}_j | z_j = k) P(z_j = k)}{\sum_{l=1}^{m} P(\boldsymbol{x}_j | z_j = l) P(z_j = l)} = \gamma_{jk} \end{aligned}$$

The Observed-Data Log-Likelihood

- The observed data are independent and identically distributed: $P(\mathbf{X}) = P(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = P(\boldsymbol{x}_1) \dots P(\boldsymbol{x}_n)$
- The probability of one observation:

$$\begin{array}{rl} P(\boldsymbol{x}_j) = & \sum_{k=1}^{m} P(\boldsymbol{x}_j | z_j = k) P(z_j = k) \\ = & \sum_{k=1}^{m} \pi_k \mathcal{N}(\boldsymbol{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{array}$$

- The Observed-data log-likelihood:

$$\ln P(\mathbf{X}) = \sum_{j=1}^{n} \ln \sum_{k=1}^{m} \pi_k \mathcal{N}(\boldsymbol{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- The direct maximization of this function is not a well posed problem because of the presence of singularities (See Bishop'06, page 432–435).

The Complete-Data Log Likelihood

- Compute $\ln P(\mathbf{X}, \boldsymbol{Z})$ instead of $\ln P(\mathbf{X})$:

$$\ln P(\mathbf{X}, \boldsymbol{Z}) = \ln \prod_{j=1}^{n} P(\boldsymbol{x}_j, z_j) = \ln \prod_{j=1}^{n} P(\boldsymbol{x}_j | z_j) P(z_j)$$

- with ($\delta_k(z_j) = 1$ if $z_j = k$ and 0 otherwise):

$$P(\boldsymbol{x}_j | z_j) P(z_j) = \prod_{k=1}^{m} \left(P(\boldsymbol{x}_j | z_j = k) \pi_k \right)^{\delta_k(z_j)}$$

- Finally:

$$\ln P(\mathbf{X}, \boldsymbol{Z}) = \sum_{j=1}^{n} \sum_{k=1}^{m} \delta_k(z_j) \underbrace{\left(\ln \pi_k + \ln P(\boldsymbol{x}_j | z_j = k) \right)}_{\alpha}$$

The Conditional Expectation

- Let's compute the *conditional expectation of the complete-data log-likelihood given the observed data*:

$$E[\ln P(\mathbf{X}, \boldsymbol{Z})|\mathbf{X}] = \sum_{j=1}^{n} \sum_{k=1}^{m} \alpha E[\delta_k(z_j)|\mathbf{X}]$$

- From the formula for the *conditional expectation* we obtain:

$$E[\delta_k(z_j)|\mathbf{X}] = \sum_{l=1}^{m} \delta_j(l) P(z_j = l|\boldsymbol{x}_j) = \gamma_{jk}$$

- Finally, $E[\ln P(\mathbf{X}, \boldsymbol{Z})|\mathbf{X}]$ becomes:

$$-\frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{m} \gamma_{jk} \left((\boldsymbol{x}_j - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_k) + \ln \pi_k + \ln |\boldsymbol{\Sigma}_k| \right)$$

The EM Algorithm for Gaussian Mixtures

- It **maximizes** the conditional **expectation** of the complete-data log-likelihood, in short: **expectation maximization**;
- It converges to a local maximum of the observed-data log-likelihood;
- In practice we minimize the negative expectation.

An EM iteration

1. Initialize the means $\boldsymbol{\mu}_k$, the covariance matrices $\boldsymbol{\Sigma}_k$ and the mixing coefficients π_k.
2. *E-step:* Evaluate the responsibilities using the current parameter values:

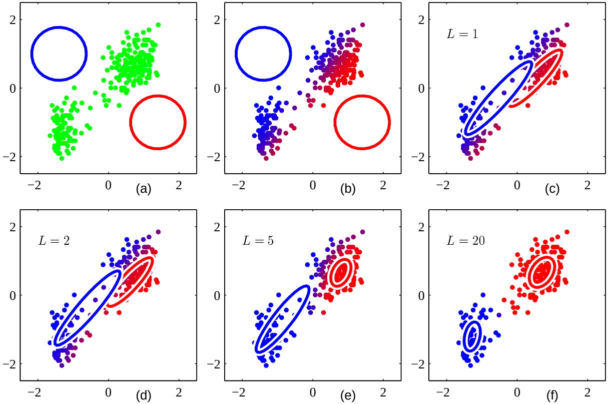
$$\gamma_{jk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{m} \pi_l \mathcal{N}(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

3. *M-step:* Re-estimate the parameters using the current responsibilities: $\boldsymbol{\mu}_k^{\text{new}}$, $\boldsymbol{\Sigma}_k^{\text{new}}$ and π_k^{new}.
4. Evaluate the log-likelihood and check for convergence of either the parameters or the log-likelihood. If not, return to step 2.

The parameters

$$
\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{n_k}\sum_{j=1}^{n}\gamma_{jk}\boldsymbol{x}_j \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{n_k}\sum_{j=1}^{n}\gamma_{jk}(\boldsymbol{x}_j-\boldsymbol{\mu}_k^{\text{new}})(\boldsymbol{x}_j-\boldsymbol{\mu}_k^{\text{new}})^\top \\
\pi_k^{\text{new}} &= \frac{n_k}{n} \\
n_k &= \sum_{j=1}^{n}\gamma_{jk}
\end{aligned}
$$

An example



EM in Practice

- Covariance models
- How to initialize the algorithm?
    - Hierarchical clustering
    - K-means
- How to choose the number of clusters (model selection)?
- How to deal with non-Gaussian data (outliers)?

Covariance Models

- Spherical covariance, same or different size for each cluster: $\mathbf{\Sigma}_k = \sigma \mathbf{I}, \mathbf{\Sigma}_k = \sigma_k \mathbf{I}$
- Constant covariance accross the clusters: $\mathbf{\Sigma}_k = \mathbf{\Sigma}$
- The *equal shape model*: $\mathbf{\Sigma}_k = \mathbf{U}_k \mathbf{\Lambda} \mathbf{U}_k^\top$
- The *equal orientation model*: $\mathbf{\Sigma}_k = \mathbf{U} \mathbf{\Lambda}_k \mathbf{U}^\top$
- etc.

Hierarchical clustering

- The classification-likelihood:

$$P_{CL}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m; z_1, \ldots, z_n | \mathbf{X}) = \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{x}_j | \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j})$$

- The presence of *class labels* introduces a combinatorial aspect making exact maximization impractical.
- Model-based agglomerative hierarchical clustering (C. Fraley. "Algorithms for model-based Gaussian hierarchical clustering". SIAM J. Sci. Comput. 1998):
  - Successively merging pairs of clusters corresponding to the greatest increase in the classification-likelihood,
  - Starts with considering each observation as a *singleton* cluster with spherical covariance

K-means Clustering

See Bishop'2006 (pages 424–428) for more details.

- What is a cluster: a group of points whose inter-point distance are small compared to distances to points outside the cluster.
- Cluster centers: $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m$.
- Goal: find an assignment of points to clusters as well as a set of mean-vectors $\boldsymbol{\mu}_k$.
- Notations: For each point $\boldsymbol{x}_j$ there is a *binary indicator variable* $r_{jk} \in \{0, 1\}$.
- Objective: minimize the following *distorsion measure*:

$$J = \sum_{j=1}^{n} \sum_{k=1}^{m} r_{jk} \|\boldsymbol{x}_j - \boldsymbol{\mu}_k\|^2$$

The K-means Algorithm

1. Initialization: Choose m and initial values for $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_m$.
2. First step: Assign the j-th point to the closest cluster center:

$$r_{jk} = \begin{cases} 1 & \text{if } k = \arg\min_l \|\boldsymbol{x}_j - \mu_l\|^2 \\ 0 & \text{otherwise} \end{cases}$$

3. Second Step: Minimize J to estimate the cluster centers:

$$\boldsymbol{\mu}_k = \frac{\sum_{j=1}^n r_{jk}\boldsymbol{x}_j}{\sum_{j=1}^n r_{jk}}$$

4. Convergence: Repeat until no more change in the assignments.

How Many Clusters?

- Let M_m denote the "model" associated with m clusters, this also corresponds to a parameter set:

$$\mathbf{\Theta}_m = (\boldsymbol{\mu}_1, \mathbf{\Sigma}_1, \ldots, \boldsymbol{\mu}_m, \mathbf{\Sigma}_m)$$

- The posterior probability of a model given the data:

$$P(M_m|\mathbf{X}) \propto P(\mathbf{X}|M_m)P(M_m)$$

- The *integrated-likelihood* of a model M_m:

$$P(\mathbf{X}|M_m) = \int_{\Theta_m} P(\mathbf{X}|\mathbf{\Theta}_m, M_m)P(\mathbf{\Theta}_m|M_m)d\mathbf{\Theta}_m$$

Bayes Factor

- Choose the model that is the most likely a posteriori. If $P(M_1) = \ldots = P(M_m)$, this amounts to choosing the model with the highest integrated-likelihood.
- Bayes factor:

$$B_{12} = \frac{P(\mathbf{X}|M_1)}{P(\mathbf{X}|M_2)}$$

There is strong evidence for M_1 if $B_{12} > 100$.

The Bayesian Information Criterion (BIC)

- The main difficulty in using Bayes factors is the evaluation of the integrated-likelihood.
- BIC approximation:

$$\begin{aligned} \mathrm{BIC}_m &= 2 \ln P(\mathbf{X}|\widehat{\mathbf{\Theta}}_m, M_m) - \nu \ln(n) \\ &\approx 2 \ln P(\mathbf{X}|M_m) \end{aligned}$$

- Good performance in the case of Gaussian mixtures, but **Do not expect a miracle!**

Dealing with Outliers

- Add a *uniform component* to the mixture likelihood:

$$p(\boldsymbol{x}) = \sum_{k=1}^{m} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \pi_{m+1}\mathcal{U}(\boldsymbol{x}|a, b)$$

- This introduces an additional prior, modifies the posterior probabilities **and nothing else**:

$$\pi_{m+1} = 1 - \sum_{l=1}^{m} \pi_l$$

$$\gamma_{jk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{m} \pi_l \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \pi_{m+1}\emptyset}$$