Slides modified from:

PATTERN RECOGNITION AND MACHINE LEARNING

CHRISTOPHER M. BISHOP

and:

Computer vision: models, learning and inference.

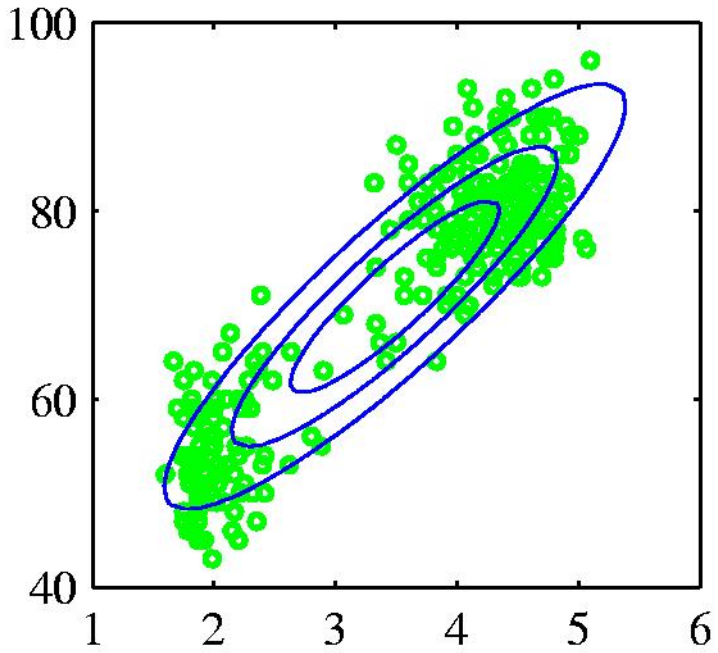
Outline for this week and the following

Mixture models and expectation-maximization (EM-) algorithm

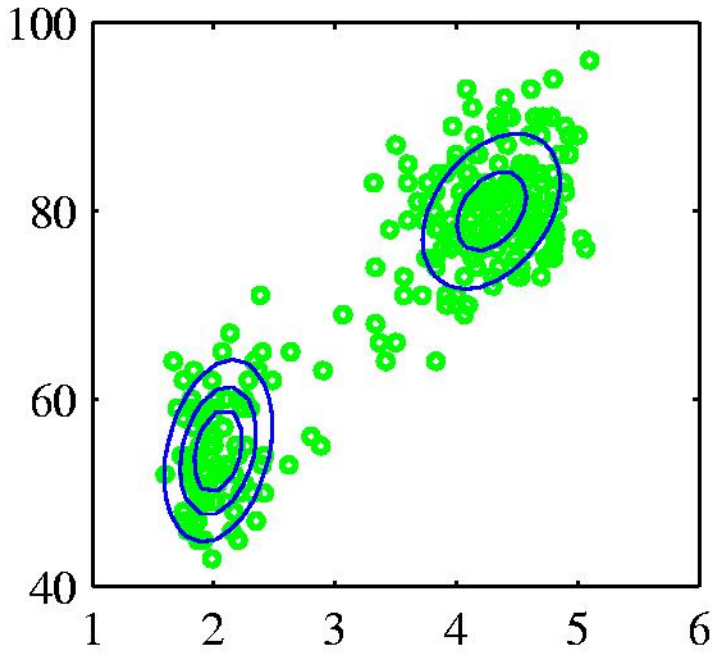
- K-Means
- Mixture of Gaussians
- Formal and general treatment of EM

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian



Mixture of two Gaussians

Mixtures of Gaussians (2)

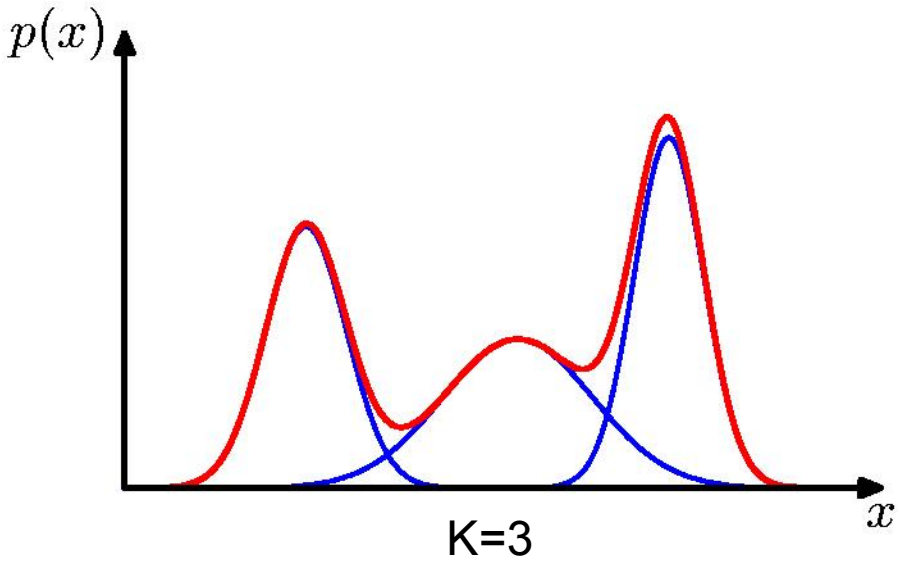
Combine simple models into a complex model:

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

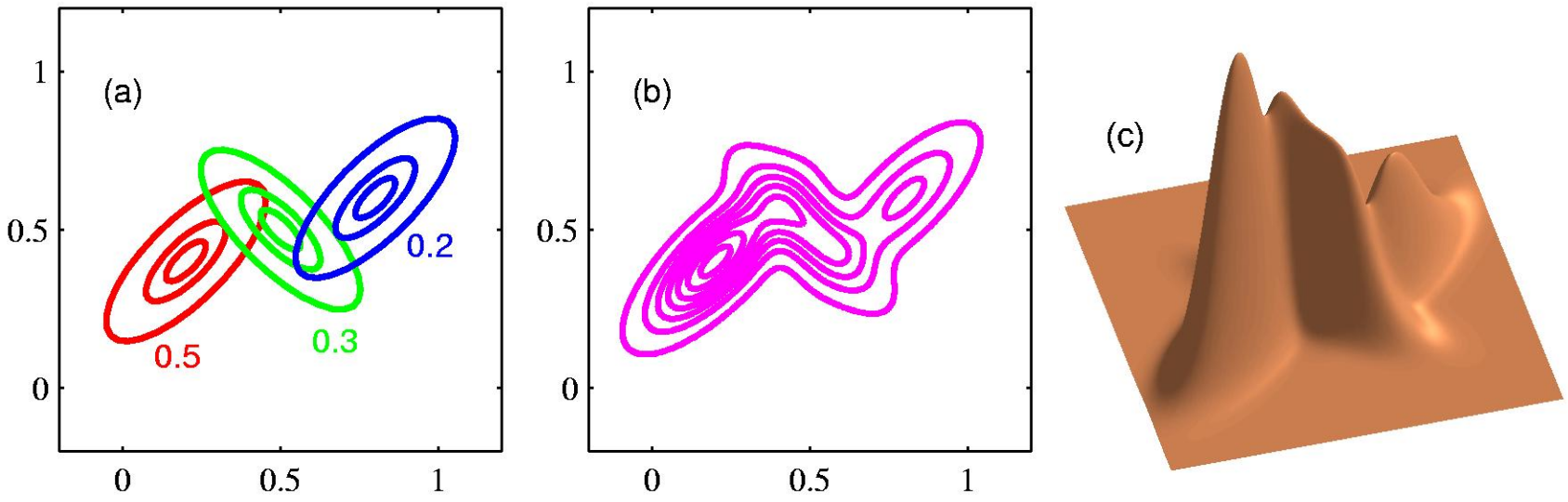
Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



Mixtures of Gaussians (3)



K-means Clustering

Suppose we have N observations of a random D-dim Euclidean variable x: $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Goal: partition data into K (fixed) clusters

Intuitively, cluster: inter-point distances smaller to points outside of the cluster.

K-means Clustering

Formalize by introducing D-dim vectors $\boldsymbol{\mu}_k$ with k = 1, . . . , K

$\boldsymbol{\mu}_k$ is a 'prototype' associated with k^{th} cluster (will turn out to be centers of the cluster)

Goals:

- find assignment of data points to clusters
- as well as a set of vectors $\{\boldsymbol{\mu}_k\}$

such that sum of the squares of the distances of each data point to its closest vector $\boldsymbol{\mu}_k$, is minimal.

K-means Clustering

Objective function (*distortion measure*)

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

with binary indicator variable $r_{nk} = 1$, if data point x_n is assigned to cluster k, and $r_{nj} = 0$ for $j \neq k$

Goal: to find values for the $\{r_{nk}\}$ and the $\{\mu_k\}$ so as to minimize J.

K-means Clustering

Solve by iterative procedure, two successive steps:

- Choose some initial values for the $\boldsymbol{\mu}_k$
- Minimize J with respect to r_{nk}, keeping $\boldsymbol{\mu}_k$ fixed (E)
- Minimize J with respect to $\boldsymbol{\mu}_k$, keeping r_{nk} fixed (M)
- Repeat until convergence.

E (expectation) and M (maximization) steps of EM algorithm

K-means Clustering

Expectation step:

- *J* is linear function of r_{nk}
- Terms involving different *n* are independent, optimize for each *n* separately
- Choose r_{nk} to be 1 for whichever value of *k* gives the minimum of $\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

- I.e. simply assign the n^{th} data point to the closest cluster centre.

K-means Clustering

Maximization step:

Derivative of J with respect to $\boldsymbol{\mu}_k$:

$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

Which gives:

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

Denominator is number of points assigned to cluster k

Thus, $\boldsymbol{\mu}_k$ is mean of data points x_n assigned to cluster k

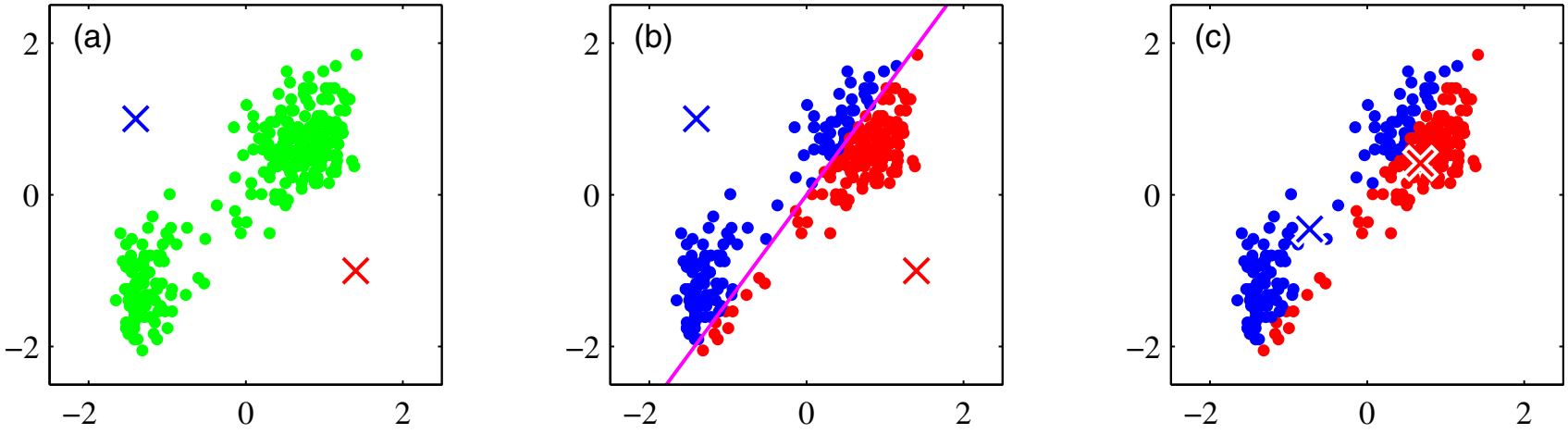
K-means Clustering

- Both phases, E and M reduce value of objective function J
- Thus, convergence of the algorithm is assured
- However, it may converge to a local rather than global minimum of J
- Note: ‘hard’ assignment of data points, uniquely to one cluster
- Later: ‘soft’ assignment by probabilistic approach

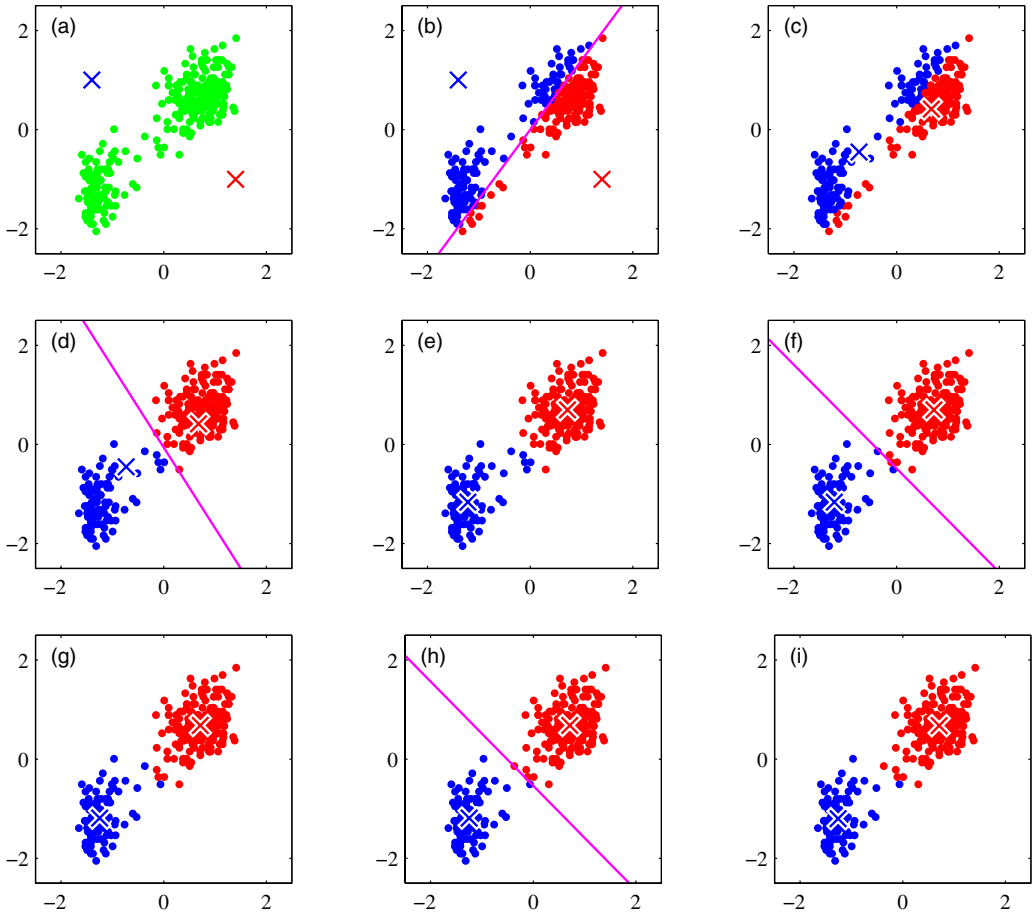
K-means Clustering

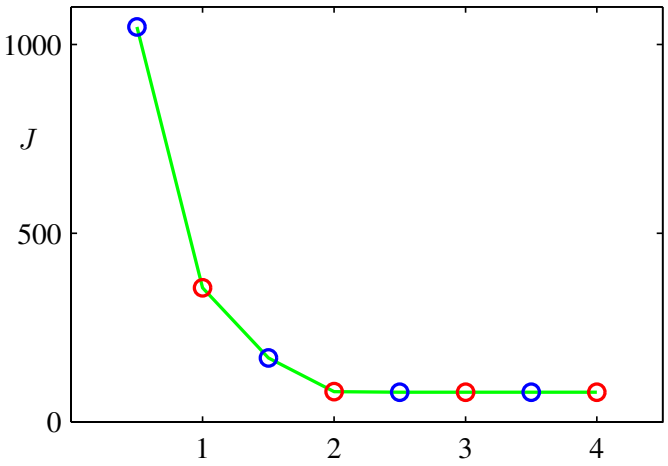
E step

M step



K-means Clustering





E step (blue points)
M step (red points)

K-means Clustering

- Note: generalize the K-means algorithm by introducing general dissimilarity measure

$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$

- M step is potentially more complex
- Restrict each cluster prototype to be equal to one of the data vectors assigned to that cluster

Application: image segmentation

- Goal of segmentation: partition image into objects or parts of objects
- Each pixel is point in 3-dimensional space, intensities of red, blue, green channels
- After k-means: Replace each pixel vector with {R, G, B} intensity triplet assigned $\boldsymbol{\mu}_k$

Application: image segmentation

Mixtures of Gaussians

Gaussian mixture distribution: a rich and important class of density models

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

Formulation of Gaussian mixtures in terms of discrete *latent* variables **z.**

Mixtures of Gaussians

- **z**: K-dimensional binary random variable
- One element z_k is equal to 1 and all other elements equal to 0.
- Thus, **z** has K possible states
- Joint distribution p(**x**, **z**)
- Marginal distribution p(**z**)
- Conditional distribution p(**x**|**z**)

Mixtures of Gaussians

- Marginal distribution over **z** is specified by mixing coefficients

$$p(z_k = 1) = \pi_k$$

with $0 \leqslant \pi_k \leqslant 1$ and $\sum_{k=1}^{K} \pi_k = 1$

- Thus, we can write $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$

Mixtures of Gaussians

Conditional distribution of **x** given a particular value for **z**:

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Or

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Marginal distribution of **x**: sum joint distribution over **z**:

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

Thus, marginal distribution is Gaussian mixture distrib.

Mixtures of Gaussians

Comments:

- If we have several observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, for every $\mathbf{x}_i$ there is a corresponding latent variable $\mathbf{z}_i$
- Equivalent formulation of the Gaussian mixture with explicit latent variable
- Joint distribution $p(\mathbf{x}, \mathbf{z})$ easier to handle!

Mixtures of Gaussians

Conditional probability of **z** given **x** (Bayes' theorem):

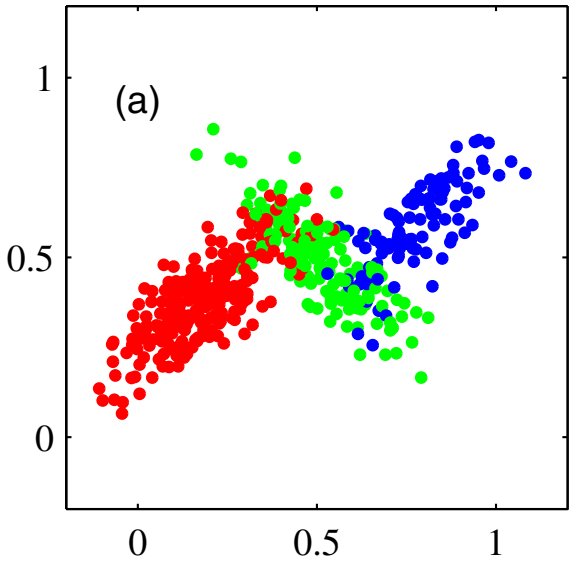
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x})$$

$$= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

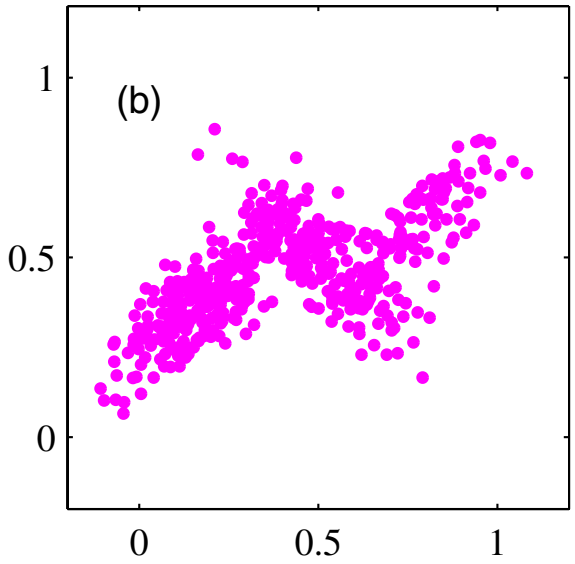
- π_k is prior probability of $z_k = 1$
- $\gamma(z_k)$ corresponding posterior probability once **x** is observed
- $\gamma(z_k)$ is *responsibility* that component k takes for 'explaining' the observation **x**

Mixtures of Gaussians

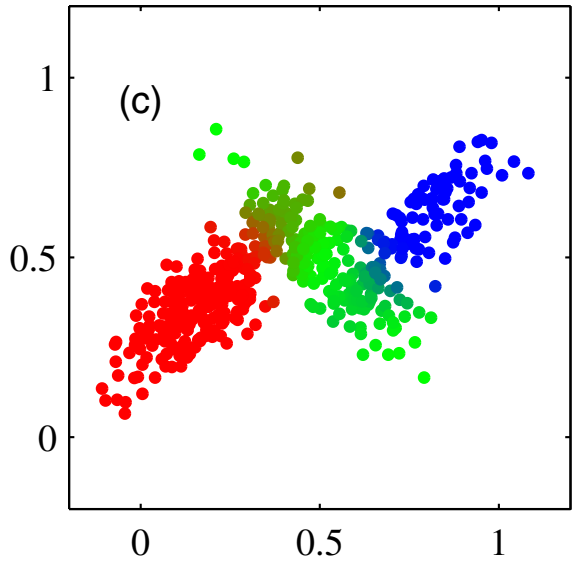
Joint distribution $p(\mathbf{x}, \mathbf{z})$



Marginal distribution $p(\mathbf{x})$



Responsibilities $\gamma(z_k)$



Mixtures of Gaussians

Learn parameters of mixture of Gaussians model?

Maximum likelihood!

Likelihood

Consider probability distribution depending on parameter θ
Likelihood:

$$L(\theta|x) = P(x|\theta)$$

The likelihood of parameter value θ given an observed (fixed) outcome x is equal to the probability of x given the parameter value θ

Example

- “Given that I have flipped a coin 100 times and it is a fair coin, what is the *probability* of it landing heads-up every time?”
- "Given that I have flipped a coin 100 times and it has landed heads-up 100 times, what is the *likelihood* that the coin is fair?"

Maximum Likelihood (ML)

Consider probability distribution depending on parameter θ
Likelihood:

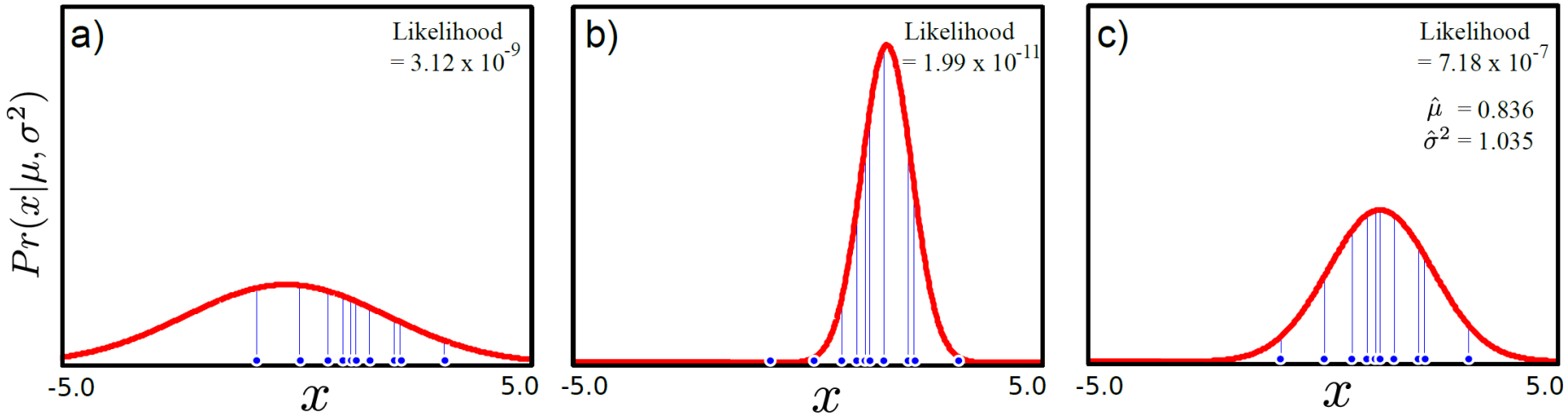
$$L(\theta|x) = P(x|\theta)$$

The likelihood of parameter value θ given an observed (fixed) outcome x is equal to the probability of x given the parameter value θ

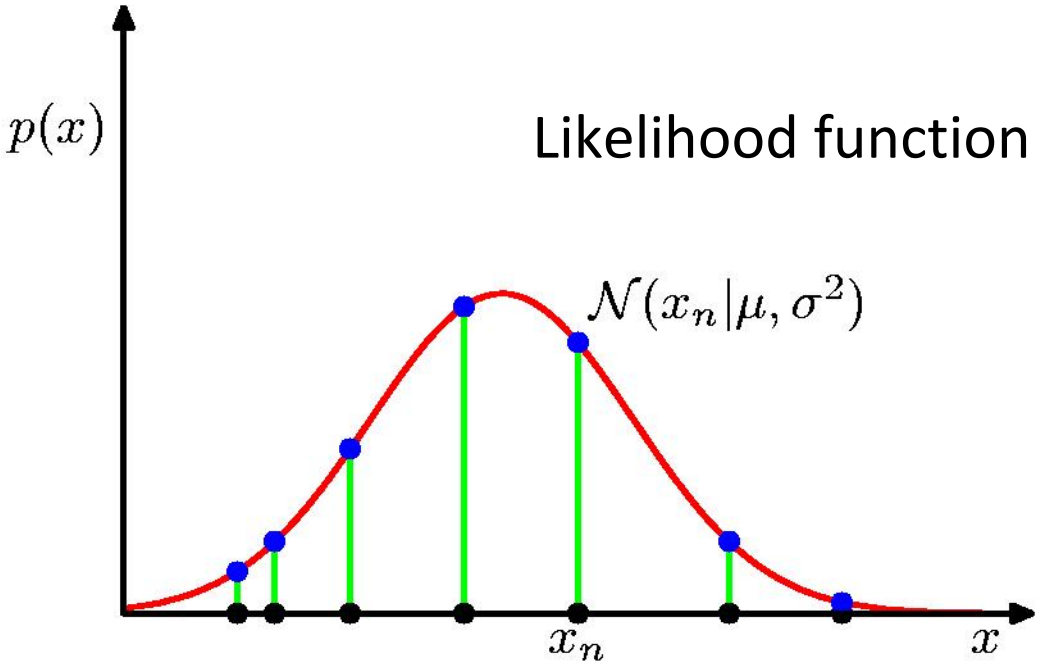
What is the most likely value of the parameter θ, given the outcome x?

Fitting normal distribution: ML

$$
\begin{aligned}
Pr(x_{1\ldots I}|\mu,\sigma^2) &= \prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2) \\
&= \prod_{i=1}^{I} \mathrm{Norm}_{x_i}[\mu,\sigma^2] \\
&= \frac{1}{(2\pi\sigma^2)^{I/2}} \exp\left[-0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right]
\end{aligned}
$$



Gaussian Parameter Estimation



$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

Likelihood for the Gaussian

Assume σ is known. Given i.i.d. data $\mathbf{x} = \{x_1, \ldots, x_N\}$, the likelihood function for μ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

This has a Gaussian shape as a function of μ (but it is *not* a distribution over μ).

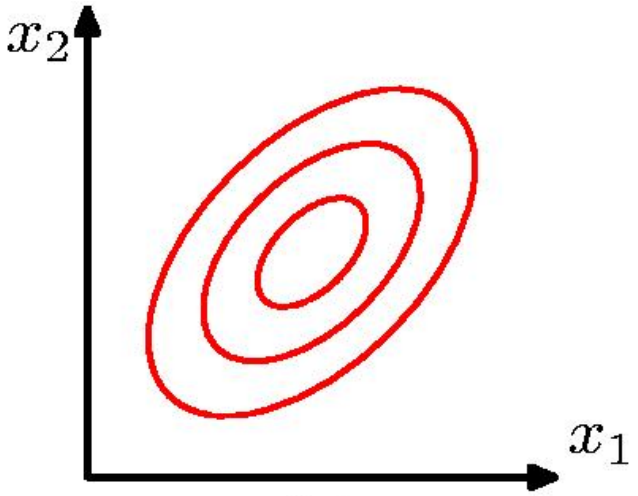
Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

The Multivariate Gaussian

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



Maximum Likelihood for the Multivariate Gaussian

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

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n.$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

Mixtures of Gaussians

Maximum likelihood:

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

With matrix **X** in which the n^{th} row is $\mathbf{x}^T_n$

- No closed-form solution

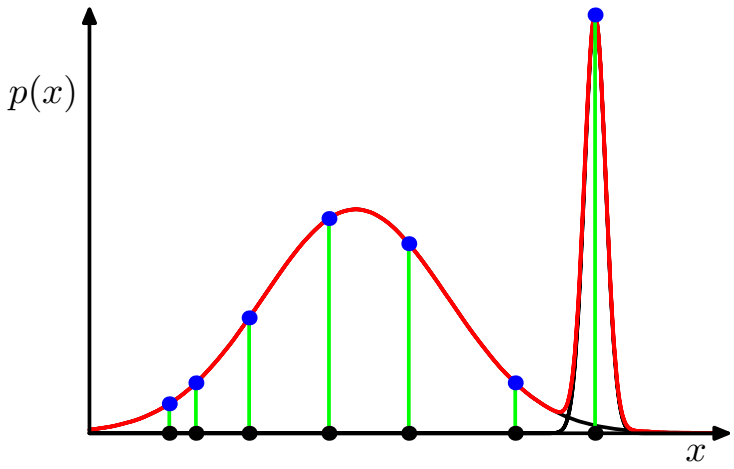
Mixtures of Gaussians

Maximum likelihood problem:

If $\boldsymbol{\mu}_j = \mathbf{x}_n$ for some value of n, this point contributes

$$\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2\mathbf{I}) = \frac{1}{(2\pi)^{1/2}}\frac{1}{\sigma_j}$$

Goes to infinity for σ_j-> 0



EM for Mixtures of Gaussians

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

with respect to the means $\boldsymbol{\mu}_k$

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

Multiplying by by Σ_n^{-1} yields $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$ with effective number assigned to cluster k $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

Interpretation: Average weighted by responsibility

EM for Mixtures of Gaussians

Set derivative with respect to $\boldsymbol{\Sigma}_{\text{k to zero}}$ yields

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

Same form as the corresponding result for a single Gaussian but again weighted by responsibilities and effective number of points in component

EM for Mixtures of Gaussians

Maximize with respect to mixing coefficients π_k

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

giving $$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

Multiply both sides by π_k and sum over k yields $\lambda = -N$

Eliminate λ and rearrange: $$\pi_k = \frac{N_k}{N}$$

Mixing coefficient is given by the average responsibility

EM for Mixtures of Gaussians

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.

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

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

EM for Mixtures of Gaussians

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

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\
\pi_k^{\text{new}} &= \frac{N_k}{N}
\end{aligned}$$

where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}).$$

4. Evaluate the log likelihood

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

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

EM for Mixtures of Gaussians