BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session 15

(Maximum Likelihood Estimation & Gaussian Mixture Model)

16th and 17th Sep 2023

IMP Note to Self

Maximum Likelihood Estimation (MLE)

Estimation is the process of estimating unknown true values of population parameters using their corresponding best sample statistics (good estimators) in an optimum manner.

An estimator is said to be a good if it is

- unbiased,
- consistent,
- efficient and
- sufficient while estimating its parameter.

Maximum Likelihood Estimation (MLE)

- Method of Maximum Likelihood Estimation is the best and most popular one among all methods to obtain an almost good or best estimator for a population parameter.
- It is a method of obtaining an estimator which most (maximum) likely estimates the true value of the parameter i.e., finding an estimator that can give most likely nearer value for the unknown true value of parameter.
- The corresponding estimator is called maximum likelihood estimator (MLE).

Maximum Likelihood Estimation (MLE)

Suppose we have a random sample $x_1, x_2, .., x_n$ whose assumed probability distribution depends on some unknown parameter θ.

Ex:

1) For Binomial unknown parameters are n, p.

2) For Poisson unknown parameter is λ.

3) For Normal unknown parameters are $\mu \; and \; \sigma^2$.

Our goal is to find good estimate of θ(population parameter) using sample and which can be done with the help of MLE.

Maximum Likelihood

- It is observed that a good estimate of unknown parameter θ would be the value of θ that maximizes the probability
- i.e. the likelihood of getting the data we observed (this is reason, why we called as likelihood function)

Maximum Likelihood function

- Let $x_1, x_2, .., x_n$ be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ.
- The goal of MLE to maximize likelihood function

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$$
$$= f(x_1|\theta) * f(x_2|\theta) \ldots f(x_n|\theta)$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

Maximum Likelihood Estimation (MLE)

- The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes likelihood(θ).

It is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i / \theta)$$

For maximization, we have

$$\frac{dL}{d\theta} = 0 \; ; \quad \frac{d^2L}{d\theta^2} < 0$$

Maximum Likelihood Estimation (MLE)

If L and $\log_e L$ are not differentiable or integrable or principle of maxima-minima fails then in such case direct method of finding the estimator of the parameter which maximizes L or $\log_e L$ is applied using order statistic principle empirically.

Maximum Likelihood Estimation (MLE)

MLEs are:

- Consistent
- Efficient
- Sufficient
- MLEs May (or may not) be unbiased
- MLEs are Asymptotically normally distributed
- Asymptotically tend to have least variance.

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P\left(H = 61 \mid p = \frac{1}{3}\right) = \binom{100}{61}\left(\frac{1}{3}\right)^{61}\left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P\left(H = 61 \mid p = \frac{1}{2}\right) = \binom{100}{61}\left(\frac{1}{2}\right)^{61}\left(\frac{1}{2}\right)^{39} = 0.007$$

$$P\left(H = 61 \mid p = \frac{2}{3}\right) = \binom{100}{61}\left(\frac{2}{3}\right)^{61}\left(\frac{1}{3}\right)^{39} = 0.040$$

p.m.f.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$0 \le p \le 1$$
$$x = 0, 1, 2, \ldots, n;$$

Since $P\left(H = 61 \mid p = \frac{2}{3}\right)$ is maximum and hence MLE is $p = \frac{2}{3}$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Solution: Since the distribution follow is Binomial distribution, with parameter p. Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61}(1-p)^{39}$$

For maximization

$$\frac{d}{dp} P(H = 61|p) = 0$$

$$\Rightarrow \binom{100}{61} [61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0$$

$$P\left(H = 61 \middle| p = \frac{61}{100}\right) = \binom{100}{61}\left(\frac{61}{100}\right)^{61}\left(\frac{39}{100}\right)$$

$$P(H = 61|p = 1) = 0$$

Since $P\left(H = 61 \middle| p = \frac{61}{100}\right)$ is maximum and hence $p = \frac{61}{100}$ is the MLE.

Maximum Likelihood for a Binomial distribution

- Suppose we wish to find the maximum likelihood estimate (MLE) of θ for a Binomial distribution,

$$p_k(k,\theta) = nC_k\theta^k(1-\theta)^{n-k}$$

$$\log p_k(k,\theta) = \log(nC_k) + k\log(\theta) + (n-k)\log((1-\theta)$$

$$\frac{\partial \log p_k(k,\theta)}{\partial\theta} = 0 \Rightarrow 0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$k - k\theta = n\theta - k\theta \Rightarrow \theta = \frac{k}{n}$$

Maximum Likelihood Example 3:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.

Soln :

Maximum Likelihood Example 3:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.

Soln :

L(p)=P[X=0] P[X=1] P[X=0] P[X=0] P[X=1] P[X=0]

=(1-p) p (1-p) (1-p) p (1-p)

$=(1-p)^3p^2$.

Log L(p)=log[$(1-p)^3p^2$.]=log[$(1-p)^3$]+log[$(p^{2)}$.]=3log(1-p)+2logp

$$\frac{\partial LogL(p)}{\partial p} = 0 \text{ means,} \qquad \frac{-3}{1-p} + \frac{2}{p} = 0 \Rightarrow \frac{-3p+2-2p}{p(1-p)} = 0 \Rightarrow p = 2/5$$

That is , there is 1/3 chance to observe this sample if we believe the population to be Binomial distributed .

MLE for Binomial distribution parameter P

Let $X_1, X_2, \ldots, X_N \in R$ be samples obtained from a Binomially Distribution.

Binomial Distribution is used to model 'x' successes in 'n' Bernoulli trials. Its p.d.f. is given by:

$$f(x) = \frac{n!}{x!(n-x)!} p^{x} (1-p)^{n-x}$$

The likelihood function L(p) is given by:

$$L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{N} \frac{n!}{x_i!(n-xi)!} p^{x_i} (1-p)^{n-x_i}$$

The log-likelihood is:

$$lnL(p) = \sum_{i=1}^{N} ln(n!) - \sum_{i=1}^{N} ln(x_i!) - \sum_{i=1}^{N} ln(n-x_i!) + \sum_{i=1}^{N} xi.\, ln(p) + (n - \sum_{i=1}^{N} xi).\, ln(1-p)$$

Setting its derivative with respect to p to zero,

$$\frac{d}{dp} lnL(p) = \frac{1}{p} . \sum_{i=1}^{N} xi - \frac{1}{1-p} \sum_{i=1}^{N} (n - xi) = 0$$

which implies,

$$\frac{1}{p} . \sum_{i=1}^{N} xi = (\frac{1}{1-p})(N.n - \sum_{i=1}^{N} xi)$$

giving,

$$\hat{p} = \frac{1}{N}\left(\frac{\sum_{i=1}^{N} x_i}{n}\right) = \frac{1}{N}\left(\frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_N}{n}\right) = \frac{\overline{X}}{n}$$

which is the maximum likelihood estimate.

MLE for Poisson Distribution Parameter

Let $X_1, X_2, \ldots, X_n \in R$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$; where x =0,1,2,...

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i}$$

The log-likelihood is:

$$lnL(\lambda) = -\lambda n + \sum_{i=1}^{n} xi.\, ln(\lambda) - ln(\prod_{i=1}^{n} xi)$$

Setting its derivative with respect to λ to zero, we have:

$$\frac{d}{d\lambda} lnL(\lambda) = -n + \sum_{i=1}^{n} xi. \frac{1}{\lambda} = 0$$

giving,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$$

which is the maximum likelihood estimate

MLEs for Normal Distribution Parameters

Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal distribution with unknown mean μ and variance σ^2. Find maximum likelihood estimators of mean μ and variance σ^2.

In finding the estimators, the first thing we'll do is write the probability density function as a function of $\theta_1 = \mu$ and $\theta_2 = \sigma^2$:

$$f(x_i;\theta_1,\theta_2) = \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}}\exp\left[-\frac{(x_i-\theta_1)^2}{2\theta_2}\right]$$

for $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. We do this so as not to cause confusion when taking the derivative of the likelihood with respect to σ^2. Now, that makes the likelihood function:

$$L(\theta_1,\theta_2) = \prod_{i=1}^{n} f(x_i;\theta_1,\theta_2) = \theta_2^{-n/2}(2\pi)^{-n/2}\exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right]$$

and therefore the log of the likelihood function:

$$\log L(\theta_1,\theta_2) = -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i-\theta_1)^2}{2\theta_2}$$

Now, upon taking the partial derivative of the log likelihood with respect to θ_1, and setting to 0, we see that a few things cancel each other out, leaving us with:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \frac{-\cancel{2}\sum(x_i - \theta_1)\cancel{(-1)}}{\cancel{2}\theta_2} \overset{\text{SET}}{\equiv} 0$$

Now, multiplying through by θ_2, and distributing the summation, we get:

$$\sum x_i - n\theta_1 = 0$$

Now, solving for θ_1, and putting on its hat, we have shown that the maximum likelihood estimate of θ_1 is:

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Now for θ_2. Taking the partial derivative of the log likelihood with respect to θ_2, and setting to 0, we get:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{\sum(x_i - \theta_1)^2}{2\theta_2^2} \overset{\text{SET}}{\equiv} 0$$

Multiplying through by $2\theta_2^2$:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \left[-\frac{n}{2\theta_2} + \frac{\sum(x_i - \theta_1)^2}{2\theta_2^2} \overset{\text{SET}}{\equiv} 0\right] \times 2\theta_2^2$$

Hence the MLEs for the parameters mean and variance of the normal model are respectively:

we get:

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

And, solving for θ_2, and putting on its hat, we have shown that the maximum likelihood estimate of θ_2 is:

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

MLEs for the parameters of Uniform distribution

Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from a Uniform population with probability function

$f(X, a, b) = 1/(b-a)$ for $a <= X <= b$.

The likelihood function is expressed as follows:

$L = f(x_1, a, b)\, f(x_2, a, b) \ldots f(x_n, a, b)$

$= 1/(b-a) \times 1/(b-a) \times \ldots \times 1/(b-a)$

$= 1/(b-a)^n = (b-a)^{-n}$

Log L = -n Log (b-a)

d Log L/da = 0

n/(b-a) = 0

n = 0 (Contradiction)

Also, d Log L/db = 0

-n/(b-a) = 0

-n = 0 (Contradiction)

Therefore, principle of maxima fails to give MLEs.

Now, we use order statistic principle. By ordering the sample observations, we obtain

$a <= x_{(1)} < x_{(2)} < \ldots < x_{(n)} <= b$ since $a <= X <= b$

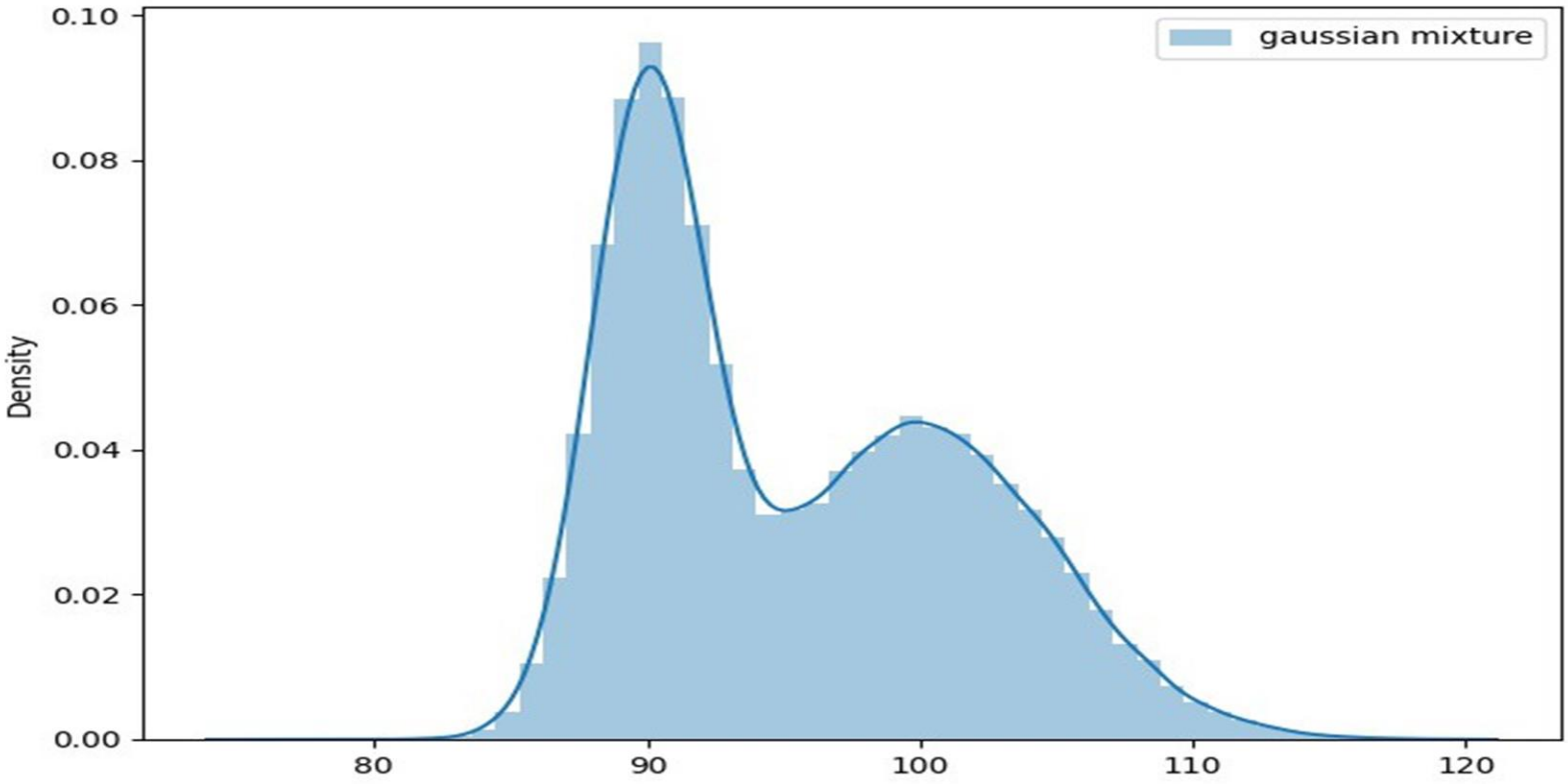
Where $x_{(1)} = \text{Min}\{x_1, x_2, \ldots, x_n\}$ and $x_{(n)} = \text{Max}\{x_1, x_2, \ldots, x_n\}$

As $x_{(1)}$ and $x_{(n)}$ fall nearest to a and b respectively in the inside interval [a, b] then $x_{(1)}$ and $x_{(n)}$ are considered as MLEs for the respective parameters a and b using order statistic principle.

EM alternates between performing an expectation E-step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization M-step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E-step. The parameters found on the M-step are then used to begin another E-step, and the process is repeated until convergence.

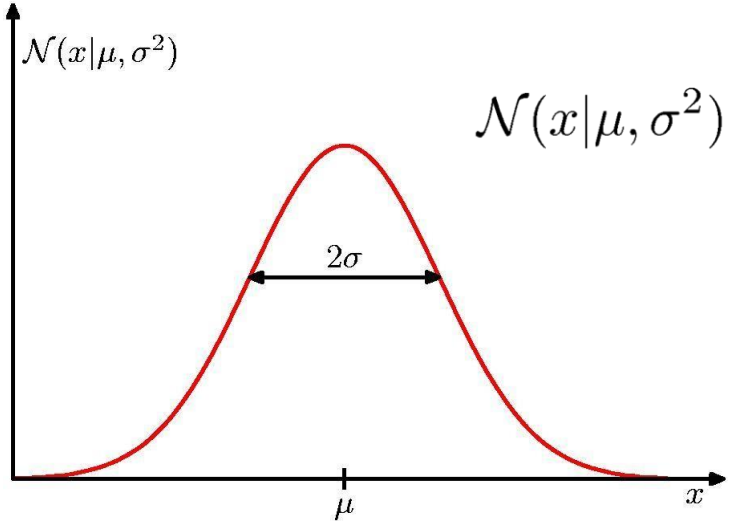
Gaussian Mixture Model

- Suppose Company A share price is normally distributed with mean price Rs. 100 and standard deviation of price Rs.2 with 1000 sample points
- Company B share price is normally distributed with mean price Rs. 100 and standard deviation of price Rs.2 with 800 sample points
- Now, both samples are mixed, then we obtain Normal (Gaussian) mixture model.

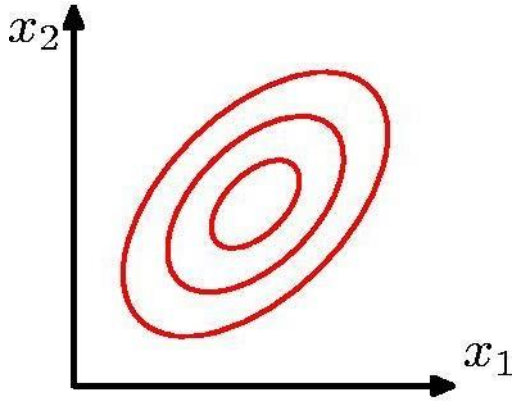


So after mixing the processes together, we have the dataset that we see on the plot. We can notice 2 peaks: around 90 and 100, but for many of the points in the middle of the peaks it is ambiguous to which distribution they were drawn from. So how should we approach this problem?

Gaussian Distribution



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



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

Mixtures of Gaussians

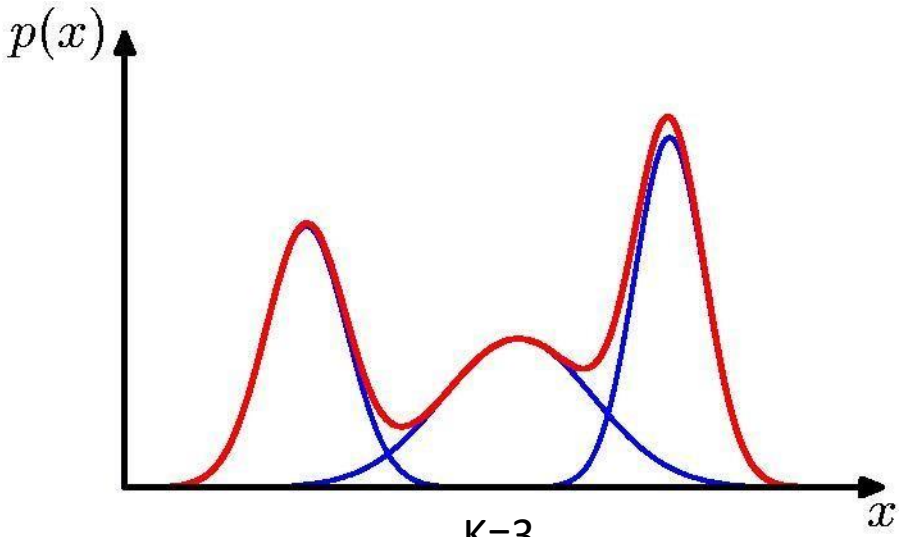
- Combine simple models into a complex model:

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

Component

Mixing coefficient

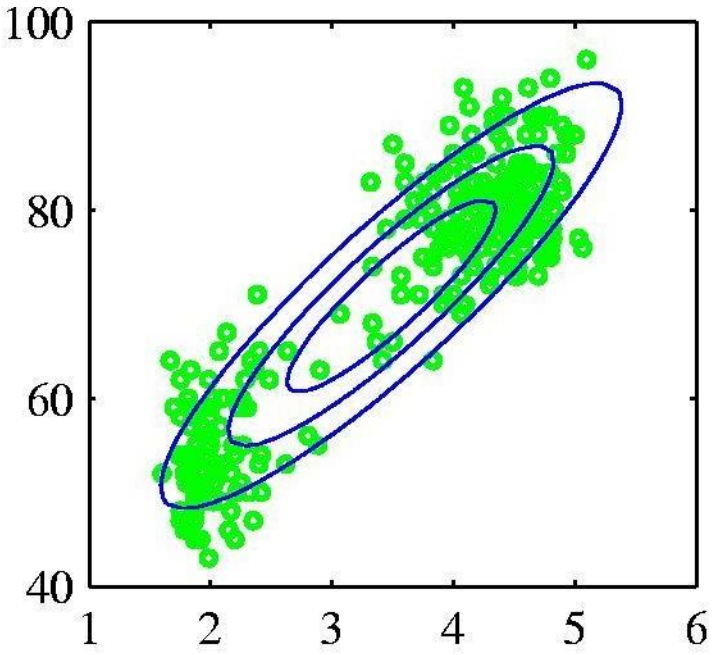
$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



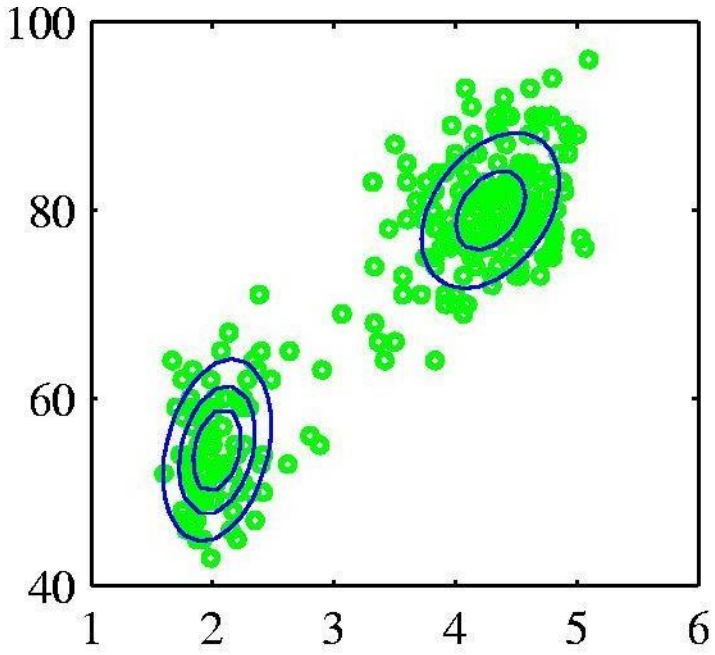
K=3

➢ Find parameters through EM (Expectation Maximization) algorithm

Probabilistic version: Mixtures of Gaussians



Single Gaussian

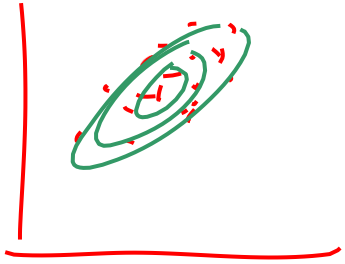


Mixture of two Gaussians

Gaussian Mixture Model

- Data set

$$D = \{\mathbf{x}_n\} \quad n = 1, \ldots, N$$



- Consider first a single Gaussian
- Assume observed data points generated independently

$$p(D|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

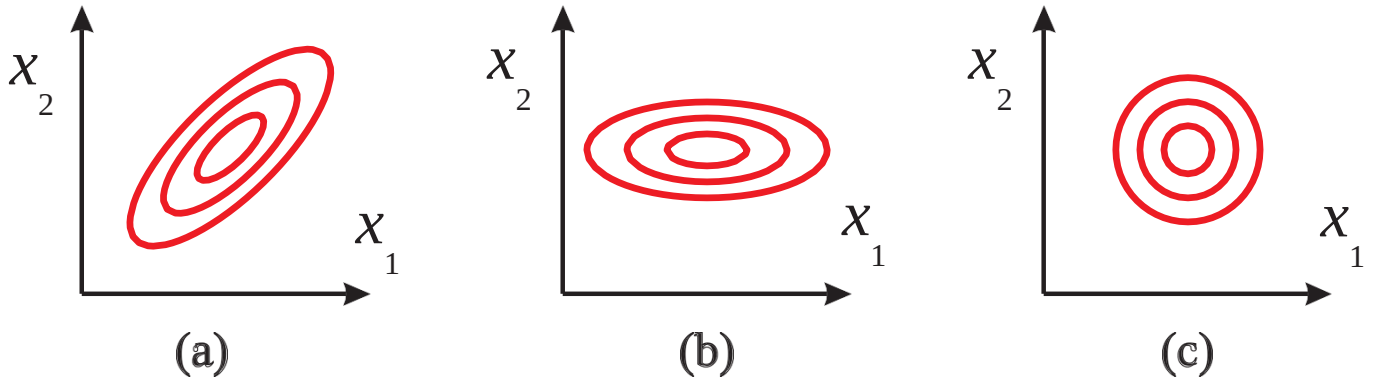
- Viewed as a function of the parameters, this is known as the *likelihood function*

The Gaussian Distribution

- Multivariate Gaussian

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

mean covariance



Gaussian Mixture Model

- K-dimensional binary random variable z having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0.
- The values of z_k therefore satisfy $z_k \in \{0,1\}$
- K possible states for the vector z according to which element is nonzero.
- Joint distribution p(x,z) in terms of a marginal distribution p(z) and a conditional distribution p(x|z),
- Marginal distribution over z is specified in terms of the mixing coefficients π_k, such that $p(z_k = 1) = \pi_k$

Fitting the Gaussian Mixture

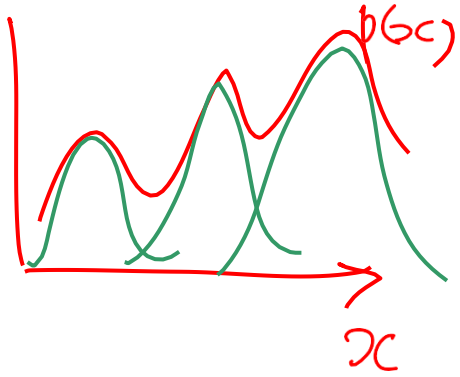
- We wish to invert this process – given the data set, find the corresponding parameters:
  - mixing coefficients
  - means
  - covariances
- If we knew which component generated each data point, the maximum likelihood solution would involve fitting each component to the corresponding cluster

- Problem: the data set is unlabelled
- We shall refer to the labels as *latent* (= hidden) variables

Gaussian Mixture Model

- Linear super-position of Gaussians

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



- Normalization and positivity require

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

- Can interpret the mixing coefficients as prior probabilities

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

Gaussian Mixture Model

- z uses a 1-of-K representation, we can also write this distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- Conditional distribution of x given a particular value for z is a Gaussian

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

- Joint distribution is given by p(z)p(x|z), and the marginal distribution of x is then obtained by summing the joint distribution over all possible states of z to given

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

Gaussian Mixture Model

- Conditional probability of z given x
- use $\gamma(z_k)$ to denote $p(z_k = 1 \mid x)$, whose value can be found using 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 as the prior probability of $z_k = 1$, and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed x.

Maximum Likelihood

Log of likelihood function:

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

- Maximizing the log likelihood function for a Gaussian mixture model turns out to be a more complex problem than for the case of a single Gaussian.

- The difficulty arises from the presence of the summation over k that appears inside the logarithm, so that the logarithm function no longer acts directly on the Gaussian.

GMM Problems and Solutions

- How to maximize the log likelihood
  - solved by expectation-maximization (EM) algorithm
- How to avoid singularities in the likelihood function
  - solved by a Bayesian treatment
- How to choose number *K* of components
  - also solved by a Bayesian treatment

Expectation Maximization (EM) Algorithm

- Conditions for MLE: Setting the derivatives of ln p(X|π,μ,Σ)

in with respect to the means μk of the Gaussian components to zero, we obtain

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

rearranging we obtain:

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

where we have defined $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

Expectation Maximization (EM) Algorithm

- μ_k for the kth Gaussian component is obtained by taking a weighted mean of all of the points in the data set
- Weighting factor for data point x_n is given by the posterior probability $\gamma(z_{nk})$ that component k was responsible for generating xn
- If we set the derivative of ln p(X|π,μ,Σ) with respect to Σk to 0, and follow a similar line of reasoning, making use of the result for the maximum likelihood solution for the covariance matrix of a single Gaussian

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

Expectation Maximization (EM) Algorithm

- Maximize ln p(X|π,μ,Σ) with respect to the mixing coefficients π_k with constraint

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

- Using a Lagrange multiplier and maximizing the following quantity

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

which gives

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

- If we now multiply both sides by πk and sum over k, we find λ = −N.
- Rearranging we obtain

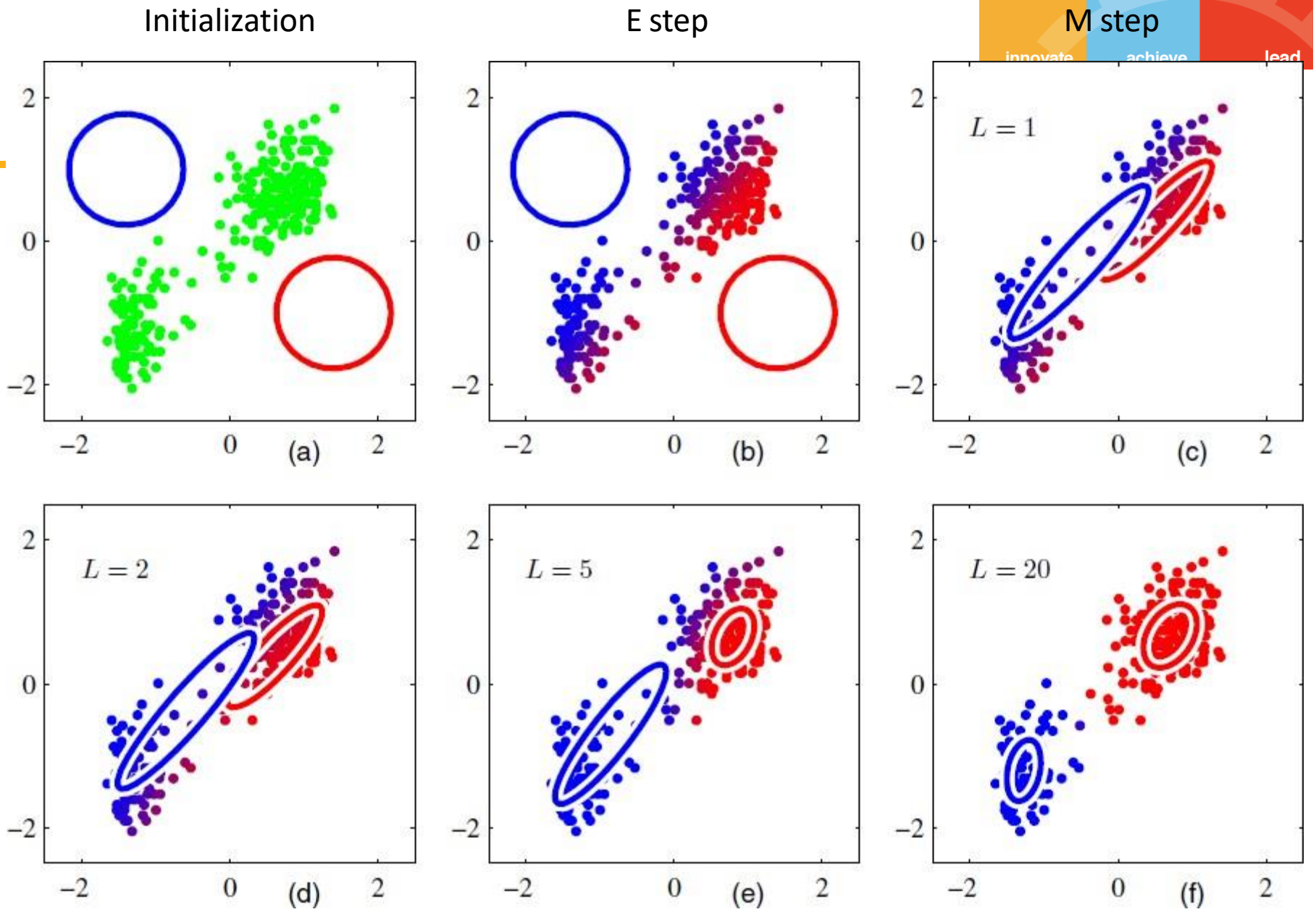
$$\pi_k = \frac{N_k}{N}$$

Expectation Maximization (EM) Algorithm

- We first choose some initial values for the means, covariances, and mixing coefficients.
- Then we alternate between the following two updates that we shall call the E step and the M step
- In the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities,
- We then use these probabilities in the maximization step, or M step, to re-estimate the means, covariances, and mixing
- In practice, the algorithm is deemed to have converged when the change in the log likelihood function, or alternatively in the parameters, falls below some threshold

EM Algorithm – Informal Derivation

- The solutions are not closed form since they are coupled
- Suggests an iterative scheme for solving them:
  - make initial guesses for the parameters
  - alternate between the following two stages:
    - E-step: evaluate responsibilities
    - M-step: update parameters using ML results
- Each EM cycle guaranteed not to decrease the likelihood



Figures from Chris Bishop

EM algorithm for GMM

1. Initialize the means μk, covariances Σk and mixing coefficients πk, 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 algorithm for GMM

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

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

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

$$\pi_k^{\text{new}} = \frac{N_k}{N} \qquad \text{where} \quad N_k=\sum_{n=1}^{N}\gamma(z_{nk})$$

4. Evaluate the log likelihood

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

IMP Note to Self

Thanks