Essential Estimation Techniques in Data Science

Dr. Arun Kumar

Indian Institute of Technology Ropar

August 22, 2025

TABLE OF CONTENTS

Maximum Likelihood Estimation

Point Estimation

Point Estimation

In **point estimation**, we estimate an unknown parameter of the population using a single number that is calculated from the sample data which is a subset of the population.

Example

Based on sample results, we estimate that p, the proportion of all Indian Voters who are in favor of a particular party, is 0.6.

Maximum Likelihood Estimation (MLE)

In the maximum likelihood estimation (MLE) the parameters of a probability distribution are estimated by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

Mathematically, let $X_1, X_2, \cdots, X_n$ be a random sample from a population with density function f and unknown parameter θ. The MLE of θ is such that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(\theta | x_1, x_2, \ldots, x_n).$$

MLE for Bernoulli distribution

- Data = $\{H, T, T, T, H, \ldots\}$.
- $\mathbb{P}(\textit{Heads}) = p$ and $\mathbb{P}(\textit{Tails}) = 1 - p$, where p is unknown parameter.
- MLE: Choose p that maximizes the probability of observed data, that is

$$\begin{aligned}
\hat{p} &= \operatorname*{argmax}_p \mathbb{P}(\textit{Data}|p) \\
&= \operatorname*{argmax}_p \mathbb{P}(x_1, x_2, \ldots, x_n|p) \\
&= \operatorname*{argmax}_p p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\
&= \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.
\end{aligned}$$

Thus, the MLE for p is the ratio of number of heads with the total number of trials.

Simple Linear Regression parameters estimation using MLE

Our data are (x_i, y_i) having n observations with one explanatory variable and one response variable. The model is

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim \mathit{Normal}(0, \sigma^2).$$

The likelihood function is

$$\begin{aligned} L(\alpha, \beta, \sigma^2 | x, y) &= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n \epsilon_i^2} \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}. \end{aligned}$$

The log-likelihood function is

$$\begin{aligned} l(\alpha, \beta, \sigma^2 | x, y) &= \log L(\alpha, \beta, \sigma^2 | x, y) \\ &= -n \log(\sigma) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2. \end{aligned}$$

Regression parameters estimation using MLE contd...

Thus, maximizing the likelihood function for the parameters α and β is equivalent to minimizing

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

The parameters can be obtained by differentiating S with respect to α and β and putting these equations equal to 0.

AR(1) parameters estimation using MLE

- A stationary Gaussian process has the form

$$X_t = c + \rho X_{t-1} + \epsilon_t, \ t = 1, 2, \cdots, \text{ where } |\rho| < 1.$$

- $\mathbb{E}X_1 = \frac{1}{1-\rho}$ and $\text{Var}(X_1) = \frac{\sigma^2}{1-\rho^2}$.
- The errors are Gaussian and hence $X_1 \sim N\left(\frac{c}{1-\rho}, \frac{\sigma^2}{1-\rho^2}\right)$.
- $X_2 = c + \rho X_1 + \epsilon_2$, and hence $(X_2|X_1 = x_1) \sim N\left(c + \rho x_1, \sigma^2\right)$
- Similarly, $X_3 = c + \rho X_2 + \epsilon_3$, and hence $(X_3|X_2 = x_2, X_1 = x_1) = (X_3|X_2 = x_2) \sim N\left(c + \rho x_2, \sigma^2\right)$.
- The likelihood of the complete sample can thus be calculated as

$$f_{X_n, X_{n-1}, X_{n-2}, \cdots, X_1}(x_n, x_{n-1}, x_{n-2}, \cdots, x_1) = f_{X_1}(x_1) \prod_{j=2}^{n} f_{X_j|X_{j-1}}(x_j|x_{j-1}$$

Least Square Estimation

Least Square Estimation

Suppose, we have a data set $(x_i, y_i), i = 1, 2, \cdots, n$ and we are interested to fit a best function f such that

$$y_i = f(x_i) + \epsilon_i,$$

which describes the relationship between y_i and x_i. We can use the following steps:

- Define the model depending on your data. For example $y_i = f(x_i) + \epsilon_i$.
- Formulate the least square problem by minimizing the residual square i.e.

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

Least square estimation contd...

Many times we use parametric model for example $y_i = \alpha + \beta X_i + \epsilon_i$. Then, the least square estimation become to estimate α and β which minimizes

$$S = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

We also consider the multiple linear regression in the matrix form

$$Y = AX + \epsilon.$$

Weighted Least Square

WLS

- The linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

  where the random errors are iid $N(0, \sigma^2)$.

- What happens if the ϵ_i's are independent but with unequal variance i.e. $\epsilon_i \sim N(0, \sigma_i^2)$? Also called heteroscedasticity.

- The ordinary least squares (OLS) estimates for β_j are unbiased, but no longer have the minimum variance.

- The weighted Least Squares (WLS) fixes the problem of heteroscedasticity

WLS

- For the model

$$y_i = \beta_0 + \beta_1 1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma_i^2)$?.

- The weighted Least Squares (WLS) is finding the estimates of β_j such that

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2}{\sigma_i^2}.$$

- In WLS, we focus more on minimizing errors of observation with smaller variances and focus less on minimizing errors of observations with larger variances.

- In general WLS is finding the estimates of β_j such that

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} w_i (y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2,$$

where the weights $w_1, w_2, \cdots, w_n$ are known and $w_i > 0$ for

WLS estimates

The WLS estimates of β_j which minimize

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} w_i (y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2,$$

is

$$\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W Y,$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, W = \begin{pmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_n \end{pmatrix}$$

Maximum A Posteriori Estimation

Bayesian Statistics

being, relating to, or involving statistical methods that assign probabilities or distributions to events (such as rain tomorrow) or parameters (such as a population mean) based on experience or best guesses before experimentation and data collection and that apply Bayes' theorem to revise the probabilities and distributions after obtaining experimental data

- Webster's Revised Unabridged Dictionary

Prior/Posterior

Example

Suppose you have 4 coins. Three of them are fair coins but the fourth is a weighted coin which has 80% chance of landing heads up. You take a coin randomly out of your pocket and toss it. Suppose it lands heads up. What is the probability that the coin is the weighted coin ?

Prior: π(fair) = 0.75, π(weighted) = 0.25.
Likelihood: $\mathbb{P}$(head|fair) = 0.5 and $\mathbb{P}$(head|weighted) = 0.8.
Posterior:

$$\mathbb{P}(\textit{weighted}|\textit{head}) = \frac{\mathbb{P}(\textit{head}|\textit{weighted})\mathbb{P}(\textit{weighted})}{\mathbb{P}(\textit{head}|\textit{fair})\mathbb{P}(\textit{fair}) + \mathbb{P}(\textit{head}|\textit{weighted})\mathbb{P}(\textit{weigh}}$$
$$= 0.348.$$

Bayesian Inference

Prior

The prior is the probability distribution that represents your uncertainty over θ **before** you have sampled any data.

Posterior

The posterior is the probability distribution that represents your uncertainty over θ **after** you have sampled data.

We have

$$\pi(\Theta|X) = \frac{f_{X|\Theta}(x|\theta)\pi_\Theta(\theta)}{f_X(x)}.$$

Classical vs Bayesian Statistics

Classical Statistics

In classical statistics, the unknown parameter is a fixed constant, just we don't know it.

Bayesian Statistics

In Bayesian statistics, the unknown parameter is assumed to be a random variable.

Different representations

Discrete parameter discrete data

$$p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)}$$

Discrete parameter continuous data

$$p_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{f_X(x)}$$

Continuous parameter discrete data

$$f_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{p_X(x)}$$

Example

Example

Find the posterior distribution of probability of heads denoted by θ on a coin. Suppose x heads are observed in n tosses.
Solution: Suppose we assume a uniform prior for Θ i.e. $f_\Theta(\theta) = 1,\ 0 < \theta < 1$. Also

$$p_{X|\Theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

We then have

$$\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta} \\ &= \frac{(n+1)!\theta^x(1-\theta)^{n-x}}{x!(n-x)!}. \end{aligned}$$

Maximum a posteriori probability (MAP) Estimator

We choose the estimator $\hat{\Theta}$ such that posterior probability $p_{\Theta|X}(\theta|x)$ or $f_{\Theta|X}(\theta|x)$ is maximize. In other words the MAP estimate of θ is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f_{\Theta|X}(\theta|x) \text{ or } \underset{\theta}{\operatorname{argmax}}\, p_{\Theta|X}(\theta|x).$$

Example

In the coin toss example, the MAP estimate will be

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \frac{(n+1)!\theta^x(1-\theta)^{n-x}}{x!(n-x)!} = \frac{x}{n},$$

which is same as the maximum likelihood estimate (MLE).

Example

Let X be a continuous random variable with the following pdf

$$f_X(x) = \begin{cases} 2x \text{ if } 0 < x < 1 \\ 0 \quad \text{otherwise} \end{cases}$$

Also, suppose that

$$Y|X = x \sim \text{Geometric}(x).$$

Find the MAP estimate of X given $Y = 3$.
Hint: $p_{Y|X}(y|x) = x(1-x)^{y-1},\ y = 1, 2, \cdots$

MAP estimation: Gaussian distribution

Example

Suppose $f_\Theta(\theta) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta - x_0}{\sigma_0}\right)^2}$. Further, $X_1, X_2, \ldots, X_n$ which conditioned on the value $\Theta = \theta$, are independent Gaussian with mean θ and variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$. Find the MAP estimate of Θ.

Solution: We have

$$f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \ldots x_n|\theta) = ce^{-\frac{1}{2}\left(\frac{x_1 - \theta}{\sigma_1}\right)^2} e^{-\frac{1}{2}\left(\frac{x_2 - \theta}{\sigma_2}\right)^2} \cdots e^{-\frac{1}{2}\left(\frac{x_n - \theta}{\sigma_n}\right)^2}.$$

Using, Bayes' theorem, we have

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{f_X(x)},$$

or

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) f_\Theta(\theta),$$

since $f_X(x)$ is independent of θ, it doesn't play any role in

MAP estimation: Gaussian distribution contd...

$$f_{X|\Theta}(x|\theta)f_{\Theta}(\theta) = c_1 e^{-\frac{1}{2}\left(\frac{x_0-\theta}{\sigma_0}\right)^2} e^{-\frac{1}{2}\left(\frac{x_1-\theta}{\sigma_1}\right)^2} e^{-\frac{1}{2}\left(\frac{x_2-\theta}{\sigma_2}\right)^2} \cdots e^{-\frac{1}{2}\left(\frac{x_n-\theta}{\sigma_n}\right)^2}.$$

Observe that

$$e^{-\frac{1}{2}\left(\frac{x_0-\theta}{\sigma_0}\right)^2} e^{-\frac{1}{2}\left(\frac{x_1-\theta}{\sigma_1}\right)^2} = e^{-\frac{1}{2}\left(\frac{x_0^2}{\sigma_0^2}+\frac{x_1^2}{\sigma_1^2}\right)} e^{-\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2}+\frac{\theta_1^2}{\sigma_1^2}-2\theta\left(\frac{x_0}{\sigma_0^2}+\frac{x_1}{\sigma_1^2}\right)\right)}$$

$$= Be^{-\frac{1}{2}\frac{\left(\theta-\frac{x_0/\sigma_0^2+x_1/\sigma_1^2}{1/\sigma_0^2+1/\sigma_1^2}\right)^2}{\frac{1}{1/\sigma_0^2+1/\sigma_1^2}}},$$

which is minimize at

$$\theta = \frac{x_0/\sigma_0^2 + x_1/\sigma_1^2}{1/\sigma_0^2 + 1/\sigma_1^2}.$$

Thus MAP estimate is

$$\hat{\theta} = \frac{x_0/\sigma_0^2 + x_1/\sigma_1^2 + \cdots + x_n/\sigma_n^2}{1/\sigma_0^2 + 1/\sigma_1^2 + \cdots + 1/\sigma_n^2}.$$

EM Algorithm

Expectation Maximization Algorithm

- **Expectation-Maximization (EM)** algorithm is a general iterative algorithm for model parameter estimation by maximizing the likelihood in presence of missing data.
- An alternative to numerical optimization of the likelihood function, the EM algorithm was introduced in 1977 by Dempster, Laird and Rubin.
- Applications: Estimating Gaussian mixture models (GMMs), estimating Hidden Markov models (HMMs) and model based data clustering.

1

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM Algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.

Example: Coin Flipping

- Two coins A and B, chosen at random and flipped 10 times.
- $\mathbb{P}(\text{H from coin A}) = \theta_A$, $\mathbb{P}(\text{H from coin B}) = \theta_B$.
- Repeat the process 5 times.
- Aim: To estimate the unknown parameters θ_A and θ_B.

Example Contd...

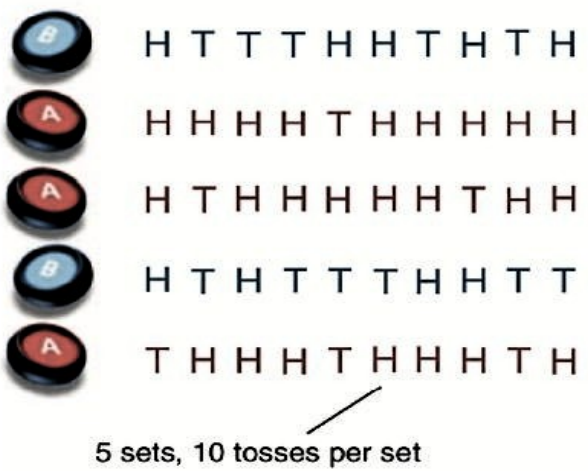
EM algorithm

$$\hat{\theta}_A = \frac{\text{\# of heads using coin A}}{\text{total \# of flips using coin A}}$$

$$\hat{\theta}_B = \frac{\text{\# of heads using coin B}}{\text{total \# of flips using coin B}}$$

Example: the 2 coin problem.
Scenario 1: no missing value:

a Maximum likelihood



Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

2

[2] Do, B.C. and Batzoglou, S. (2008). What is the expectation maximization algorithm? Nature Biotechnology, 26, 897–899.

Example Contd...

Mathematically, this can be formulated as follows:

$$X_{ij} = \begin{cases} 1 & \text{if } \textit{jth} \text{ flip in the } \textit{ith} \text{ set is H} \\ 0 & \text{if } \textit{jth} \text{ flip in the } \textit{ith} \text{ set is T,} \end{cases}$$

where $1 \leq i \leq 5$ and $1 \leq j \leq 10$.

- Y_i takes value A or B based on the coin used for *ith* set.
- Let $X_i = \sum_{j=1}^{10} X_{ij}$ be the number of heads in the *ith* set.
- Also, $X_i | Y_i = k \sim \textit{Binomial}(10, \theta_k)$, where $k = \{\text{A}, \text{B}\}$.
- Based on the information, the MLEs for parameters $\hat{\theta}_A = 0.8$ and $\hat{\theta}_B = 0.45$.

MLE in presence of hidden data

- Suppose for the same dataset the coin identities are not given i.e. $Y_i's$ are hidden.
- We have “incomplete data" consisting of X values only.
- In this case, the unknown parameters can not be estimated by maximizing the likelihood or log-likelihood function.
- However, the EM algorithm first incorporates the hidden information and then maximize the expectation of the complete data log-likelihood.

Coin Flipping Example: EM algorithm

- The hidden variable $Y_i \in \{A, B\}$
- Take $\pi_A = \mathbb{P}(Y_i = A)$ and $\pi_B = \mathbb{P}(Y_i = B)$.
- Initial guess for $\pi_A^{(0)} = 0.5$, $\pi_B^{(0)} = 0.5$, $\theta_A^{(0)} = 0.6$ and $\theta_B^{(0)} = 0.5$.

- E-step: For each data X_i, compute the probabilities that it is from coin A and coin B and denote it by w_{iA} and w_{iB} respectively. That is, we guess the value of $Y_i's$ using posterior probability distribution $\mathbb{P}(Y_i = c|X_i)$, where $c \in \{A, B\}$.
- Set $w_{ic} = \mathbb{P}(Y_i = c|X_i, \theta_A^{(0)}, \theta_B^{(0)}, \pi_A^{(0)}, \pi_B^{(0)})$,

$$
\begin{aligned}
w_{ic} &= \frac{\mathbb{P}(X_i|Y_i = c)\mathbb{P}(Y_i = c)}{\mathbb{P}(X_i|Y_i = A)\mathbb{P}(Y_i = A) + \mathbb{P}(X_i|Y_i = B)\mathbb{P}(Y_i = B)} \\
&= \frac{\mathbb{P}(X_i|Y_i = c)\pi_A^{(0)}}{\mathbb{P}(X_i|Y_i = A)\pi_A^{(0)} + \mathbb{P}(X_i|Y_i = B)\pi_B^{(0)}} \qquad (1)
\end{aligned}
$$

Continued...

For set 1 i.e. $i = 1$ and coin A,

$$w_{1A} = \frac{\theta_A^{(0)5}(1-\theta_A^{(0)})^5\pi_A^{(0)}}{\theta_A^{(0)5}(1-\theta_A^{(0)})^5\pi_A^{(0)} + \theta_B^{(0)5}(1-\theta_B^{(0)})^5\pi_B^{(0)}}$$
$$= \frac{0.5 \times (0.6)^5 \times (0.4)^5}{0.5 \times (0.6)^5 \times (0.4)^5 + 0.5 \times (0.5)^5 \times (0.5)^5} \approx 0.45$$

For set 1 i.e. $i = 1$ and coin B,

$$w_{1B} = \frac{\theta_B^{(0)5}(1-\theta_B^{(0)})^5\pi_B^{(0)}}{\theta_A^{(0)5}(1-\theta_A^{(0)})^5\pi_A^{(0)} + \theta_B^{(0)5}(1-\theta_B^{(0)})^5\pi_B^{(0)}}$$
$$= \frac{0.5 \times (0.5)^5 \times (0.5)^5}{0.5 \times (0.6)^5 \times (0.4)^5 + 0.5 \times (0.5)^5 \times (0.5)^5} \approx 0.55$$

Continued...

For set 2 i.e. $i = 2$ and coin A,

$$w_{2A} = \frac{\theta_A^{(0)9}(1-\theta_A^{(0)})\pi_A^{(0)}}{\theta_A^{(0)9}(1-\theta_A^{(0)})\pi_A^{(0)} + \theta_B^{(0)9}(1-\theta_B^{(0)})\pi_B^{(0)}}$$
$$= \frac{0.5 \times (0.6)^9 \times (0.4)}{0.5 \times (0.6)^9 \times (0.4) + 0.5 \times (0.5)^9 \times (0.5)} \approx 0.80.$$

For set 2 i.e. $i = 2$ and coin B,

$$w_{2B} = \frac{\theta_B^{(0)9}(1-\theta_B^{(0)})\pi_B^{(0)}}{\theta_A^{(0)9}(1-\theta_A^{(0)})\pi_A^{(0)} + \theta_B^{(0)9}(1-\theta_B^{(0)})\pi_B^{(0)}}$$
$$= \frac{0.5 \times (0.5)^9 \times (0.5)}{0.5 \times (0.6)^9 \times (0.4) + 0.5 \times (0.5)^9 \times (0.5)}$$
$$\approx 0.2$$

Continued...

- M-step: Now, we use the membership weights from E-step to update new parameter values.
- Let $n_A = w_{1A} + w_{2A}$ and $n_B = w_{1B} + w_{2B}$.
- Update the parameters as follows:

$$\pi_A^{(1)} = \frac{n_A}{n} = \frac{1.25}{2}$$
$$\pi_B^{(1)} = \frac{n_B}{n} = \frac{0.75}{2}.$$

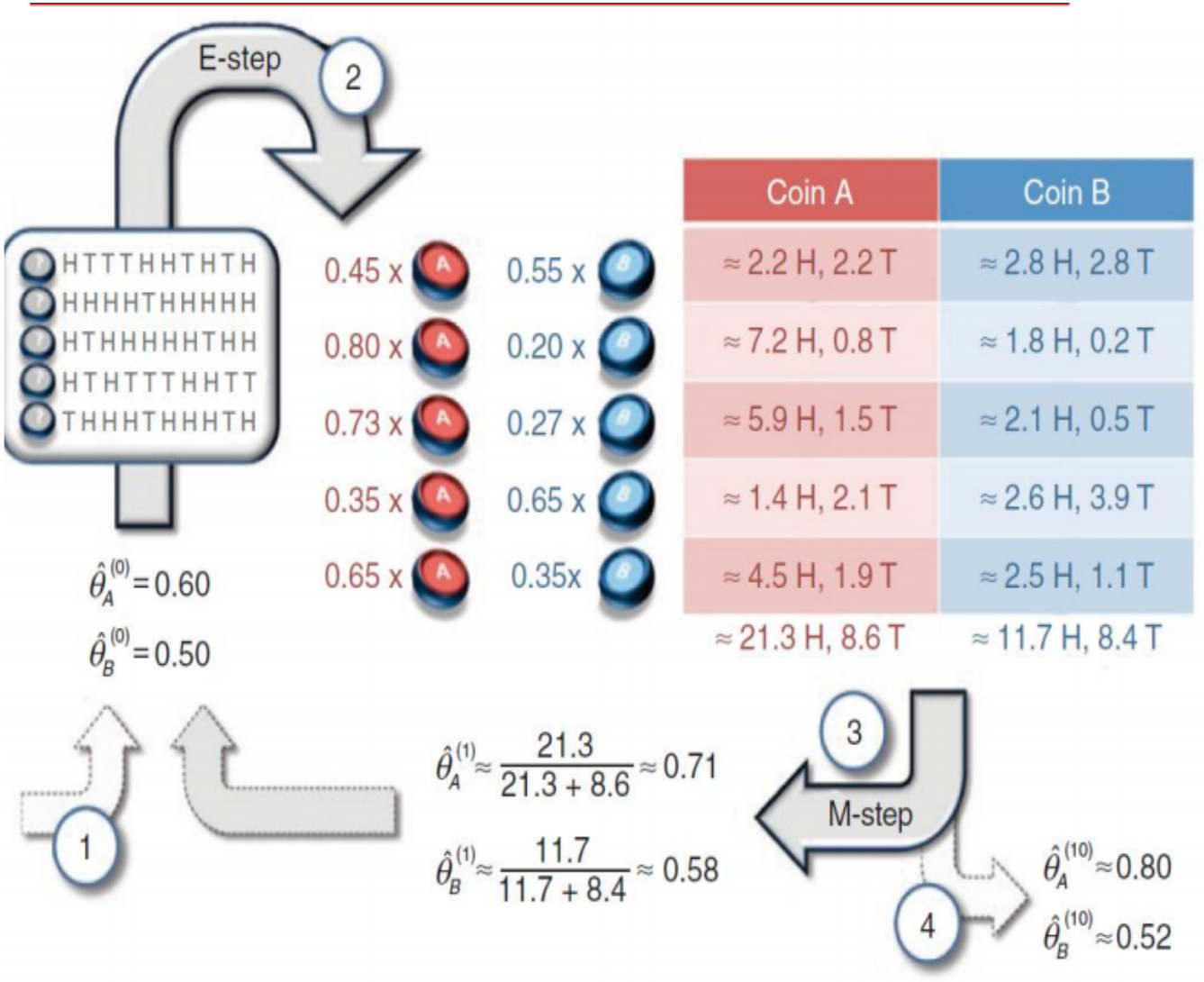
Continued...

Update

$$\theta_A^{(1)} = \frac{w_{1A}X_1 + w_{2A}X_2}{w_{1A}X_1 + w_{2A}X_2 + (10 - X_1)w_{1A} + (10 - X_2)w_{2A}}$$
$$= \frac{0.45 \times 5 + 0.80 \times 9}{0.45 \times 5 + 0.80 \times 9 + 5 \times 0.45 + 1 \times 0.80}$$
$$\approx 0.76.$$

$$\theta_B^{(1)} = \frac{w_{1B}X_1 + w_{2B}X_2}{w_{1B}X_1 + w_{2B}X_2 + (10 - X_1)w_{1B} + (10 - X_2)w_{2B}}$$
$$= \frac{0.55 \times 5 + 0.2 \times 9}{0.55 \times 5 + 0.2 \times 9 + 0.55 \times 5 + 0.2 \times 1}$$
$$\approx 0.61.$$

Continued...



[3] Do, B.C. and Batzoglou, S. (2008). What is the expectation maximization algorithm? Nature Biotechnology, 26, 897–899.

Notations

- $\Theta =$ set of parameters.
- $\theta =$ unknown parameter.
- $X =$ observed data.
- $Z =$ unknown\missing data.
- $f(X, Z|\Theta) =$ complete data density.
- $f(Z|X, \Theta) =$ conditional unobserved data density.
- $f(X|\Theta) =$ marginal observed data density.
- $f(Z|\Theta) =$ marginal unobserved data density.
- $E_{Z|X,\Theta}(g(Z)|x, \Theta) = \int_{\mathcal{Z}} g(z)f(z|x, \Theta)dz$, where $\mathcal{Z}$ denote the domain for the conditional density function $f(Z|X, \Theta)$.

EM Algorithm

- We can define a likelihood function such that

$$L(\Theta|X, Z) = f(X, Z|\Theta).$$

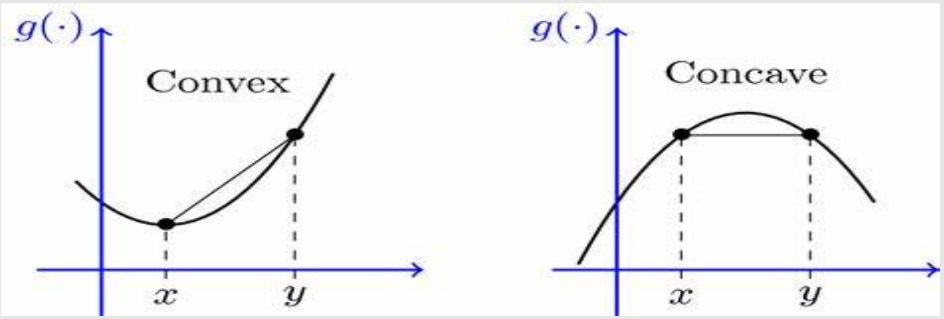
- This function is a random variable since the missing information Z is unknown and random governed by an underlying distribution.
- The original likelihood $L(\Theta|X)$ is referred to as the incomplete-data likelihood function.
- The EM algorithm first finds the expected value of the complete-data log-likelihood $\log f(X, Z|\Theta)$ with respect to the unknown data Z given observed data X and current parameter estimates.

Jensen's Inequality

Definition (Concave Function)

A function g is said to be concave on $\mathbb{R}$ if

$$g(\alpha x + (1-\alpha)y) \geq \alpha g(x) + (1-\alpha)g(y) \quad \text{for all } \alpha \in [0,1].$$



Jensen's Inequality: Let f be a concave function on $\mathbb{R}$ and suppose $\mathbb{E}[X] < \infty$ and $\mathbb{E}[f(X)] < \infty$. Then $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$.

Example

Note that $\log$ is a concave function and hence for a random variable X and function g, we have

EM Algorithm Contd...

For the given data vector X, our goal is to find the ML estimate such that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta|X) = \underset{\theta}{\operatorname{argmax}}\, f(X|\theta) = \underset{\theta}{\operatorname{argmax}} \log f(X|\theta) = l(\theta|X)$$

The key idea of EM algorithm is such as follows:

$$\begin{aligned} l(\theta|X) = \log f(X|\theta) &= \log \int_z f(x,z|\theta)dz = \log \int_z p(z|x,\theta^{(k)}) \frac{f(x,z|\theta)}{p(z|x,\theta^{(k)})} \\ &= \log \mathbb{E}_{Z|X,\theta^{(k)}} \left[\frac{f(x,Z|\theta)}{p(Z|x,\theta^{(k)})} \right] \geq \mathbb{E}_{Z|X,\theta^{(k)}} \left[\log \frac{f(x,Z|\theta)}{p(Z|x,\theta^{(k)})} \right] \\ &= \int_z p(z|x,\theta^{(k)}) \left[\log \frac{f(x,z|\theta)}{p(z|x,\theta^{(k)})} \right] dz \\ &= \int_z p(z|x,\theta^{(k)}) \log f(x,z|\theta)dz - \int_z p(z|x,\theta^{(k)}) \log p(z|x,\theta^{(k)})d \end{aligned}$$

EM Algorithm Contd...

- Thus we have

$$
\begin{aligned}
l(\theta|X) &= \log f(X|\theta) \\
&\geq \int_z p(z|x,\theta^{(k)}) \log f(x,z|\theta) dz - \int_z p(z|x,\theta^{(k)}) \log p(z|x,\theta^{(} \\
&= Q(\theta|\theta^{(k)}) - \int_z p(z|x,\theta^{(k)}) \log p(z|x,\theta^{(k)}) dz.
\end{aligned}
$$

- Rather than maximizing $l(\theta|X)$ the EM algorithm will maximize the lower bound given by

$$
Q(\theta|\theta^{(k)}) - \int_z p(z|x,\theta^{(k)}) \log p(z|x,\theta^{(k)}) dz.
$$

- The term $\int_z p(z|x,\theta^{(k)}) \log p(z|x,\theta^{(k)}) dz$ is independent of θ and hence it will maximize

$$
Q(\theta|\theta^{(k)}) = \mathbb{E}_{Z|X}[\log f(x,Z|\theta)].
$$

EM Algorithm Contd...

- Let $k = 0$. Start with initial estimate for θ as $\theta^{(k)}$.
- Find the conditional density $f(Z|X, \theta^{(k)})$ for the completion variables.
- Calculate the conditional expected log likelihood or "Q"-function:

$$Q(\theta|\theta^{(k)}) = \mathbb{E}[\log f(X, Z|\theta)|X, \theta^{(k)}].$$

  Here, expectation is with respect to the conditional distribution of Z given X and $\theta^{(k)}$.

- Find the θ that maximizes the function $Q(\theta|\theta^{(k)})$ and this new estimate will be $\theta^{(k+1)}$ i.e.

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(k)}).$$

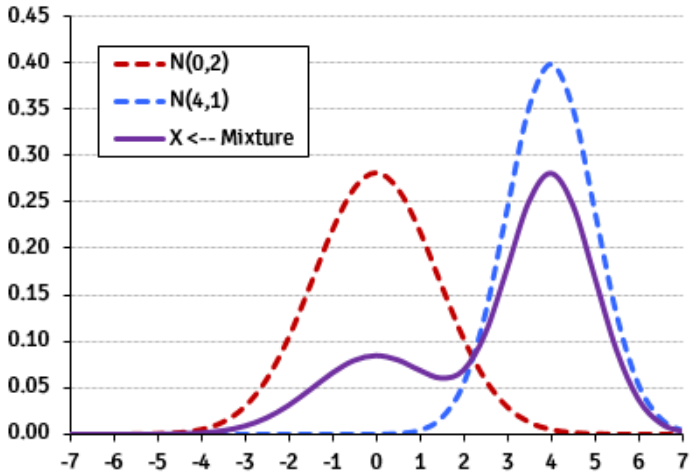
- Let $k := k + 1$ and repeat until convergence; standard criteria for convergence is to iterate until the estimate stops changing i.e. $\|\theta^{(k+1)} - \theta^{(k)}\| < \epsilon$, for the chosen tolerance ϵ.

Mixture distributions

A random vector X is said to arise from a parametric finite mixture distribution if, for all $x \in \text{Supp}(X)$, we can write its density as:

$$f_X(x|\nu) = \sum_{g=1}^{G} \pi_g p_g(x|\theta_g),$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are called mixing proportions, $p_g(x|\theta_g)$ are component densities, and $\nu = (\pi_1, \pi_2, ..., \pi_G, \theta_1, \theta_2, ..., \theta_G)$.



4

[4] Image Source: Internet

EM for Gaussian Mixture Model

- Consider the data $X = \{X_1, X_2, ..., X_n\}$ from Gaussian mixture model with $C = 2$ components. The latent variable for model is $Z_i \in \{1, 2\}$.
- The marginal probability distribution of X_i is of the form $P(X_i = x) = \sum_{c=1}^{2} \pi_c P(X_i = x | Z_i = c)$, where π_c is the mixture proportion representing the probability that X_i belongs to the c-th mixture component.
- The conditional distribution $X_i | Z_i = c \sim N(\mu_c, \sigma_c^2)$.

- E-step: For each data X_i, compute the probability that it belongs to cluster c and denote it by w_{ic}. That is, we guess the values of $Z_i's$ using posterior probability distribution $\mathbb{P}(Z_i = c|X_i)$.

- Set $w_{ic} = \mathbb{P}(Z_i = c|X_i, \mu^{(k)}, (\sigma^2)^{(k)})$,

$$w_{ic} = \frac{\mathbb{P}(X_i|Z_i = c)\mathbb{P}(Z_i = c)}{\sum_{c=1}^{2} \mathbb{P}(X_i|Z_i = c)\mathbb{P}(Z_i = c)}$$
$$= \frac{N(\mu_c^{(k)}, (\sigma_c^2)^{(k)})\pi_c^{(k)}}{\sum_{c=1}^{2} N(\mu_c^{(k)}, (\sigma_c^2))^{(k)}\pi_c^{(k)}}$$

Continued...

- M-step: Now, we use the membership weights w_{ic} and data to update the new parameter values.
- Let $n_c = \sum_{i=1}^{n} w_{ic}$ for each component c.
- Update the parameters as follows:

$$\pi_c^{(k+1)} = \frac{n_c}{n}$$

$$\mu_c^{(k+1)} = \frac{\sum_{i=1}^{n} w_{ic} X_i}{n_c}$$

$$(\sigma_c^2)^{(k+1)} = \frac{\sum_{i=1}^{n} w_{ic} (X_i - \mu_c^{(k+1)})^2}{n_c}.$$

Advantages

- The main advantage of the EM algorithm is that it is easy to implement.
- Its memory requirements tend to be modest.
- It is always guaranteed that likelihood will increase with each iteration and the algorithm will reach a local optima.
- The EM simultaneously optimize a large number of parameters.

Disadvantages

- The main disadvantage of the EM algorithm is its very slow linear convergence in some cases.
- Not guaranteed to get globally optimal estimate, generally converge to local optima.
- In practice multiple local maxima of the likelihood function are frequent and the algorithm converges to a local maxima, the initial estimates can greatly affect the final results.
- The slow linear convergence can be improved by some acceleration sachems but that will increase the complexity of the algorithm.

Example 2: Gaussian Mixture Model (Numeric)

- Data: $x = \{1.0, 1.2, 0.8, 3.0, 3.2, 2.8\}$.
- Model: two Gaussians with equal variance $\sigma^2 = 0.1$, unknown means μ_1, μ_2 and mixing weights π_1, π_2.
- Initialize: $\mu_1^{(0)} = 1$, $\mu_2^{(0)} = 3$, $\pi_1 = \pi_2 = 0.5$.

Gaussian Mixture: E-step (Numeric)

Compute responsibilities for component 1:

$$w_{i1} = \frac{\pi_1 \phi(x_i; \mu_1, \sigma)}{\pi_1 \phi(x_i; \mu_1, \sigma) + \pi_2 \phi(x_i; \mu_2, \sigma)}$$

where $\phi(x; \mu, \sigma)$ is normal pdf.

x_i	w_{i1}	$w_{i2} = 1 - w_{i1}$
1.0	0.99	0.01
1.2	0.99	0.01
0.8	0.99	0.01
3.0	0.01	0.99
3.2	0.01	0.99
2.8	0.01	0.99

Gaussian Mixture: M-step (Numeric)

Update:

$$\pi_1^{new} = \frac{\sum w_{i1}}{6} \approx 0.5, \quad \pi_2^{new} = 1 - \pi_1^{new}.$$

$$\mu_1^{new} = \frac{\sum w_{i1} x_i}{\sum w_{i1}} \approx 1.0, \quad \mu_2^{new} \approx 3.0.$$

- As expected, clusters separate near means 1 and 3.
- Continue iterating until convergence if starting parameters differ more.

Example 3: Poisson Mixture (Count Data)

- Data: counts $x = (2, 3, 4, 0, 1, 5)$.
- Model: two Poisson components with rates λ_1, λ_2, weights π_1, π_2.
- Initialize: $\lambda_1^{(0)} = 1, \lambda_2^{(0)} = 4, \pi_1 = 0.5$.

Poisson Mixture: E-step (Numeric)

$$w_{i1} = \frac{\pi_1 e^{-\lambda_1} \lambda_1^{x_i} / x_i!}{\pi_1 e^{-\lambda_1} \lambda_1^{x_i} / x_i! + \pi_2 e^{-\lambda_2} \lambda_2^{x_i} / x_i!}.$$

x_i	w_{i1}	w_{i2}
2	0.80	0.20
3	0.35	0.65
4	0.10	0.90
0	0.99	0.01
1	0.95	0.05
5	0.05	0.95

Poisson Mixture: M-step (Numeric)

Updated parameters:

$$\pi_1 = \frac{\sum w_{i1}}{6} \approx 0.54, \quad \pi_2 = 0.46.$$

$$\lambda_1 = \frac{\sum w_{i1} x_i}{\sum w_{i1}} \approx 1.2, \quad \lambda_2 \approx 4.2.$$

Iterate until convergence.

References

1. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM Algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.

2. Do, B.C. and Batzoglou, S. (2008). What is the expectation maximization algorithm? Nature Biotechnology, 26, 897–899.

3. McLachlan, G. J. and Krishnan, T. (2008). The EM Algorithm and Extensions. 2nd edition. Wiley.

Thank You