9. Expectation-Maximization Algorithm

- background on ML
- two-component mixture model
- EM algorithm in general
- applications

Maximum likelihood estimates

suppose $y^{(1)}, y^{(2)}, \ldots, y^{(N)}$ be available samples from a distribution $f(y;\theta)$

the maximum likelihood estimate is obtained by

$$\hat{\theta}_{\mathrm{ml}} = \mathrm{argmax}_{\theta} \log f(y^{(1)}, y^{(2)}, \ldots, y^{(N)}; \theta)$$

- $\hat{\theta}_{\mathrm{ml}}$ gives the distribution that most agrees with the data
- many distributions have a closed-form expression of ML estimate
- most of ML estimates are very intuitive and natural, *e.g.*,
  - Gaussian: $\hat{\mu}$ is the sample mean and $\hat{\sigma}^2$ is the sample variance
  - binomial: $X \in \{0, 1\}$ where parameter is $p = P(X = 1)$

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \quad \text{(portion of samples that are equal to } 1\text{)}$$

ML estimation of Gaussian distribution

let $Y \sim \mathcal{N}(\mu, \Sigma)$ and we have samples $\{y^{(i)}\}_{i=1}^N$

log-likelihood function of one Gaussian sample is

$$\log f(y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)^T$$

for i.i.d. samples, the log-likelihood function of $\{y^{(i)}\}_{i=1}^N$ is the sum of individuals:

$$\mathcal{L}(\Theta) = -\frac{nN}{2}\log(2\pi) + \frac{N}{2}\log\det\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^N (y^{(i)}-\mu)^T\Sigma^{-1}(y^{(i)}-\mu)^T$$

if we define $C = \frac{1}{N}\sum_{i=1}^N (y^{(i)}-\mu)(y^{(i)}-\mu)^T$

the log-likelihood function (up to constant) and to be maximized is

$$\mathcal{L}(\Theta) = \frac{N}{2} \log\det \Sigma^{-1} - \frac{N}{2} \,\mathbf{tr}(C\Sigma^{-1})$$

the zero gradient conditions are

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{i=1}^{N} \Sigma^{-1}(y^{(i)} - \mu) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \Sigma^{-1}} = \Sigma - C = 0 \quad \text{(use } \frac{\partial \log\det X}{\partial X} = X^{-1} \text{ and } \frac{\partial\, \mathbf{tr}(A^T X)}{\partial X} = A \text{)}$$

we can solve for the ML estimates as

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}, \quad \hat{\Sigma} = C = \frac{1}{N}\sum_{i=1}^{N} (y^{(i)} - \hat{\mu})(y^{(i)} - \hat{\mu})^T$$

(ML estimates are sample mean and (a biased) sample covariance)

ML estimation of multinomial distribution

two possible ways of explaining $X \sim \text{Multinomial}(\phi)$

- $X = (X_1, X_2, \ldots, X_m)$ where a sample of X is

$$X = (0, \ldots, 0, \underbrace{1}_{k^{\text{th}}}, 0, \ldots, 0) \quad \text{with probability} \quad \phi_k, \quad k = 1, 2, \ldots, m$$

$$p(x) = \phi_1^{x_1} \phi_2^{x_2} \cdots \phi_m^{x_m}$$

- $X \in \{1, 2, \ldots, m\}$ where $P(X = k) = \phi_k$ for $k = 1, 2, \ldots, m$

$$p(x) = \phi_1^{I\{x=1\}} \phi_2^{I\{x=2\}} \cdots \phi_m^{I\{x=m\}}$$

where $I\{x \in C\}$ is an indicator function that returns 1 if $x \in C$ and 0 otherwise to obtain ML estimate of ϕ, we maximize the cost function:

$$g(\phi) = \log p(x; \phi) - \lambda(\phi_1 + \phi_2 + \cdots \phi_m - 1)$$

(constrained optimization due to the constraint: $\sum_i \phi_i = 1$)

- suppose we have data $\{x^{(i)}\}_{i=1}^N$ available
- first form of X:

$$\log p(x^{(1)}, \ldots, x^{(N)}; \phi) = \sum_{i=1}^N x_1^{(i)} \log \phi_1 + x_2^{(i)} \log \phi_2 + \cdots + x_m^{(i)} \log \phi_m$$

$$\frac{\partial g}{\partial \phi_j} = \sum_{i=1}^N \frac{x_j^{(i)}}{\phi_j} - \lambda = 0 \quad \Rightarrow \quad \phi_j = \frac{1}{\lambda} \sum_{i=1}^N x_j^{(i)}$$

then we can the summation over j

$$1 = \sum_{j=1}^{m} \phi_j = \frac{1}{\lambda} \sum_{i=1}^{N} \sum_{j=1}^{m} x_j^{(i)} = \frac{N}{\lambda} \quad \Rightarrow \quad \lambda = N$$

the ML estimate of ϕ_j is then the portion of $x_j^{(i)} = 1$ out of N samples

$$\phi_j = \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)}, \quad j = 1, 2, \ldots, m$$

- second form of X:

$$\log p(x^{(1)}, \ldots, x^{(N)}; \phi) = \sum_{i=1}^{N} I\{x^{(i)} = 1\} \log \phi_1 + \cdots + I\{x^{(i)} = m\} \log \phi_m$$

$$\log p(x^{(1)}, \ldots, x^{(N)}; \phi) = \sum_{i=1}^{N} I\{x^{(i)} = 1\} \log \phi_1 + \cdots + I\{x^{(i)} = m\} \log \phi_m$$

$$\frac{\partial g}{\partial \phi_j} = \sum_{i=1}^{N} \frac{I\{x^{(i)} = j\}}{\phi_j} - \lambda = 0 \quad \Rightarrow \quad \phi_j = \frac{1}{\lambda} \sum_{i=1}^{N} I\{x^{(i)} = j\}$$

then we can the summation over j in the same way and obtain $\lambda = N$

$$\phi_j = \frac{1}{N} \sum_{i=1}^{N} I\{x^{(i)} = j\}, \quad j = 1, 2, \ldots, m$$

the result is the same: ϕ_j is the portion of $x^{(i)} = j$ from N samples

Bayes rule

from Bayes rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

let Z be latent variable, Y be data measurement, and θ be model parameter

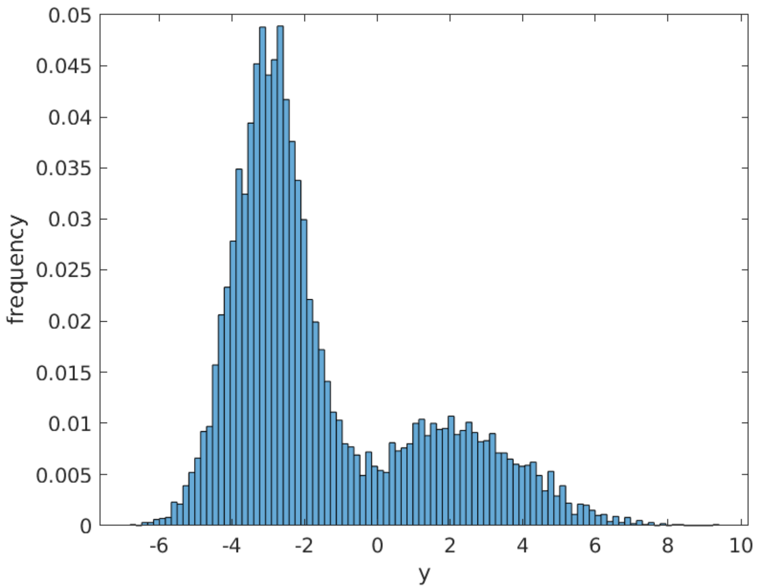
one important identity used in EM algorithm is

$$\begin{aligned} P(Z|Y;\theta) &= \frac{P(Y|Z;\theta)P(Z;\theta)}{P(Y;\theta)} \\ &= \frac{P(Y|Z;\theta)P(Z;\theta)}{\sum_z P(Y|Z;\theta)P(Z;\theta)} \end{aligned}$$

the latter is obtained from the total probabilities

Two-component mixture model

we explain a density estimation of mixture model as an example of EM



mixture model

$$
\begin{aligned}
Y_1 &\sim \mathcal{N}(\mu_1, \Sigma_1) \\
Y_2 &\sim \mathcal{N}(\mu_1, \Sigma_1) \\
Y &= (1 - Z)Y_1 + ZY_2
\end{aligned}
$$

$Z \in \{0, 1\}$ with $P(Z = 1) = \pi$

- bi-modal shape in histogram suggests us to use a mixture model instead of a Gaussian
- the problem is to estimate $\Theta = (\mu_1, \Sigma_1, \mu_2, \Sigma_2, \pi)$ where Z is unobservable

suppose data $Y = \{y^{(i)}\}_{i=1}^N$ are available

- density function of Y is followed from

$$F_Y(y) = \underbrace{(1-\pi)P(Y_1 \leq y)}_{Z=0} + \underbrace{\pi P(Y_2 \leq y)}_{Z=1}, \quad f_Y(y) = (1-\pi)f_1(y) + \pi f_2(y)$$

- loglikelihood function of Θ:

$$\mathcal{L}(Y;\Theta) = \sum_{i=1}^N \log\left[(1-\pi)f_1(y^{(i)}) + \pi f_2(y^{(i)})\right]$$

difficult to solve ML even numerically due to the sum of the term inside $\log(\cdot)$

assumption: if Z is **known**

- density function of (Y, Z) is

$$f(Y, Z; \Theta) = f(Y|Z; \Theta) f(Z; \Theta), \quad f_{Y|Z} \text{ is normal and } f(Z; \Theta) = \pi^z (1-\pi)^{1-z}$$

- loglikelihood function is

$$\mathcal{L}(Y, Z; \Theta) = \sum_{i=1}^{N} \left[(1 - z^{(i)}) \log f_1(y^{(i)}) + z^{(i)} \log f_2(y^{(i)}) \right] + \sum_{i=1}^{N} \left[(1 - z^{(i)}) \log(1 - \pi) + z^{(i)} \log \pi \right]$$

- ML estimate of (μ_i, Σ_i): sample mean and covariance
- ML estimate of π: the portion of $z^{(i)} = 1$

ML estimate of Θ when Z is assumed to be measurable

$$\hat{\pi} = \frac{1}{N}\sum_{i=1}^{N} I\{z^{(i)} = 1\}$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N} I\{z^{(i)} = 0\} y^{(i)}}{\sum_{i=1}^{N} I\{z^{(i)} = 0\}}, \quad \hat{\Sigma}_1 = \frac{\sum_{i=1}^{N} I\{z^{(i)} = 0\}(y^{(i)} - \hat{\mu}_1)(y^{(i)} - \hat{\mu}_1)^T}{\sum_{i=1}^{N} I\{z^{(i)} = 0\}}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{N} I\{z^{(i)} = 1\} y^{(i)}}{\sum_{i=1}^{N} I\{z^{(i)} = 1\}}, \quad \hat{\Sigma}_2 = \frac{\sum_{i=1}^{N} I\{z^{(i)} = 1\}(y^{(i)} - \hat{\mu}_2)(y^{(i)} - \hat{\mu}_2)^T}{\sum_{i=1}^{N} I\{z^{(i)} = 1\}}$$

note that $I\{X\}$ is the indicator function that returns 1 if the event X holds and returns 0 otherwise

conclusion: ML estimate is very natural and easy to obtain when Z is known

EM algorithm of two-mixture model

since Z is actually **unknown**, we propose an iterative EM algorithm

1. E-step: guess the values of $Z^{(i)}$ by its expected value

$$\gamma_i(\Theta) = \mathbf{E}[Z^{(i)} \mid \Theta, Y] = P(Z^{(i)} = 1 \mid \Theta, Y), \quad i = 1, 2, \ldots, N$$

   γ_i is called **responsibility** of model 2 for observation i

2. M-step: update the estimates using weight from responsibilities

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y^{(i)}}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \quad \hat{\Sigma}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y^{(i)} - \hat{\mu}_1)(y^{(i)} - \hat{\mu}_1)^T}{\sum_{i=1}^N (1-\hat{\gamma}_i)}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y^{(i)}}{\sum_{i=1}^N \hat{\gamma}_i}, \quad \hat{\Sigma}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y^{(i)} - \hat{\mu}_2)(y^{(i)} - \hat{\mu}_2)^T}{\sum_{i=1}^N \hat{\gamma}_i}$$

   $\hat{\pi} = \frac{1}{N} \sum_{i=1}^N \hat{\gamma}_i$

- we iterate E and M-steps until convergence
- the responsibilities can be computed by Bayes rule on page 9-9:

$$
\begin{aligned}
P(Z=1 \mid \Theta, Y) &= \frac{f(Y|Z=1;\Theta)P(Z=1;\Theta)}{f(Y|Z=1;\Theta)P(Z=1;\Theta) + f(Y|Z=0;\Theta)P(Z=0;\Theta)} \\
&= \frac{\pi f_2(y)}{\pi f_2(y) + (1-\pi) f_1(y)} \\
\gamma_i &= \frac{\hat{\pi} f_2(y^{(i)})}{\hat{\pi} f_2(y^{(i)}) + (1-\hat{\pi}) f_1(y^{(i)})}
\end{aligned}
$$

(soft guess for the values of $z^{(i)}$ instead of using the indicator function)

- initial guess of Θ is needed for the first iteration
- we try several initial guesses to find many local maxima solutions

Mixtures of M Gaussians

we can extend to mixture of M Gaussians with the setting:

- Z is a random variable with sample space of $\{1, 2, \ldots, M\}$
- $Z \sim$ multinomial(ϕ) with $P(Z = j) = \phi_j, \quad j = 1, 2, \ldots, M$

$$\phi \succeq 0, \quad \mathbf{1}^T \phi = 1$$

- when $Z = j$, Y is drawn from $\mathcal{N}(\mu_j, \Sigma_j)$
- samples $\{y^{(i)}\}_{i=1}^N$ are generated by random hidden variables $z^{(i)}$

problem:

- only Y are observed but Z is latent (hidden) variable
- we aim to estimate $\Theta = (\mu_1, \Sigma_1, \ldots, \mu_M, \Sigma_M, \phi)$

- loglikelihood function

$$\mathcal{L}(Y;\Theta) = \sum_{i=1}^{N} \log f(y^{(i)};\Theta) = \sum_{i=1}^{N} \log \sum_{z^{(i)}=1}^{M} f(y^{(i)}|z^{(i)};\mu,\Sigma) f(z^{(i)};\phi)$$

(difficult to find ML estimate in closed-form)

- if Z was known, the log-likelihood function and ML estimate would be

$$\mathcal{L}(Y,Z;\Theta) = \sum_{i=1}^{N} \log f(y^{(i)}|z^{(i)};\mu,\Sigma) + \log f(z^{(i)};\phi)$$

$$\hat{\phi}_j = (1/N)\sum_{i=1}^{N} I\{z^{(i)}=j\}, \quad \hat{\mu}_j = \frac{\sum_{i=1}^{N} I\{z^{(i)}=j\} y^{(i)}}{\sum_{i=1}^{N} I\{z^{(i)}=j\}}$$

$$\hat{\Sigma}_j = \frac{\sum_{i=1}^{N} I\{z^{(i)}=j\}(y^{(i)}-\hat{\mu}_j)(y^{(i)}-\hat{\mu}_j)^T}{\sum_{i=1}^{N} I\{z^{(i)}=j\}}$$

EM algorithm for M-mixture model

1. E-step: for each i, j guess the values of $Z^{(i)}$ by its expected value

$$\gamma_j^{(i)} = P(Z^{(i)} = j \mid \Theta, y^{(i)}), \quad j = 1, 2, \ldots, M$$

   (posterior probability of $Z^{(i)}$ given $y^{(i)}$ using the current estimate of Θ)

2. M-step: update the estimates using **soft guess** of $Z^{(i)}$

$$\hat{\phi}_j = \frac{1}{N}\sum_{i=1}^{N} \gamma_j^{(i)}, \quad \hat{\mu}_j = \frac{\sum_{i=1}^{N} \gamma_j^{(i)} y^{(i)}}{\sum_{i=1}^{N} \gamma_j^{(i)}}, \hat{\Sigma}_j = \frac{\sum_{i=1}^{N} \gamma_j^{(i)} (y^{(i)} - \hat{\mu}_j)(y^{(i)} - \hat{\mu}_j)^T}{\sum_{i=1}^{N} \gamma_j^{(i)}}$$

   for $j = 1, 2, \ldots, M$

3. repeat 1) and 2) until the convergence

notes on EM algorithm for mixture models

- the difference in the M-step
  - $Z^{(i)}$ were known: use hard guess as the indicator function
  - $Z^{(i)}$ is not known: use soft guess as the posterior probability
- in the E-step, we calculate γ_j using Bayes rule on page 9-9

$$
\begin{aligned}
P(Z=j|y;\Theta) &= \frac{f(y|Z=j;\mu,\Sigma)P(Z=j;\phi)}{\sum_{j=1}^{M} \underbrace{f(y|Z=j;\mu,\Sigma)}_{\text{Gaussian density}} \underbrace{P(Z=j;\phi)}_{\phi_j}} \\
\gamma_j^{(i)} &= \frac{\phi_j f_j(y^{(i)})}{\sum_{j=1}^{M} \phi_j f_j(y^{(i)})}, \quad i=1,2,\ldots,N
\end{aligned}
$$

where f_j is the Gaussian density function of the jth model governed by the current estimate of μ_j, Σ_j

EM algorithm in general

applied to maximum likelihood estimation problems with **latent** variables

problem assumptions:

- (Y, Z) are random variables; only Y is observed but Z is a latent
- log-likelihood function is

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \log f(y^{(i)}; \Theta) = \sum_{i=1}^{N} \log \sum_{z} f(y^{(i)}, z^{(i)}; \Theta)$$

explicit ML estimate is hard to obtained; but easy when $z^{(i)}$ were **observed**

Ingredients in EM

Jensen's inequality: if X is and RV and $\phi(\cdot)$ is a convex function then

$$\mathbf{E}[\phi(X)] \geq \phi(\mathbf{E}[X])$$

let $\phi(x) = \log(x)$ (concave) and let $f(y, z; \theta)$ and $q(z)$ be *any* density functions

$$\log\left(\sum_z q(z)\frac{f(y,z;\theta)}{q(z)}\right) \geq \sum_z q(z) \log \frac{f(y,z;\theta)}{q(z)}$$

(here it is the expectation of f/q and is w.r.t. to distribution q)

- if $f(y, z; \theta)/q(z)$ does not depend on z (constant) then ineq. becomes **tight**
- this is achieved when $q(z) = f(z \mid y; \theta)$ (sufficient choice)

$$\text{since we can choose } \; q(z) = \frac{f(y,z;\theta)}{\sum_z f(y,z;\theta)} = \frac{f(y,z;\theta)}{f(y;\theta)} = f(z \mid y; \theta)$$

Expectation and Maximization steps

we start with the exact loglikelihood function (to be maximized) on page 9-20

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \log \sum_{z} f(y^{(i)}, z^{(i)}; \Theta)$$

and its lower bound using Jensen's inequality

$$\mathcal{L}(\Theta) \geq \sum_{i=1}^{N} \sum_{z^{(i)}} q_i(z^{(i)}) \log \frac{f(y^{(i)}, z^{(i)}; \Theta)}{q_i(z^{(i)})}$$

- **E-step**: for each i, set q_i's to be the posterior of $z^{(i)}$ given $y^{(i)}$ and current Θ

$$q_i(z^{(i)}) = f(z^{(i)} \mid y^{(i)}; \Theta)$$

  and the inequality becomes equality (LB is the expectation w.r.t. q_i)

- **M-step**: maximize the lower bound w.r.t. Θ

$$\Theta^{+} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{z^{(i)}} q_i(z^{(i)}) \log \frac{f(y^{(i)}, z^{(i)}; \Theta)}{q_i(z^{(i)})}$$

monotonic property: let Θ and Θ^{+} be updates from successive iterations

we can show that EM always **monotonically** improve the log-likelihood

$$\mathcal{L}(\Theta^{+}) \geq \mathcal{L}(\Theta)$$

convergence test is to check if small improvement in $\mathcal{L}(\Theta)$ (set by a threshold)

Proof of monotonic property

- when we start with Θ in E-step, we choose $q_i(z^{(i)}) = f(z^{(i)}|y^{(i)}; \Theta)$

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \sum_{z^{(i)}} q_i(z^{(i)}) \log \frac{f(y^{(i)}, z^{(i)}; \Theta)}{q_i(z^{(i)})} \quad \text{(Jensen's ineq holds with eq.)}$$

- recall Jensen's inequality also holds with

$$\begin{aligned} \mathcal{L}(\Theta^+) &\geq \sum_{i=1}^{N} \sum_{z^{(i)}} q_i(z^{(i)}) \log \frac{f(y^{(i)}, z^{(i)}; \Theta^+)}{q_i(z^{(i)})} \\ &\geq \sum_{i=1}^{N} \sum_{z^{(i)}} q_i(z^{(i)}) \log \frac{f(y^{(i)}, z^{(i)}; \Theta)}{q_i(z^{(i)})} = \mathcal{L}(\Theta) \end{aligned}$$

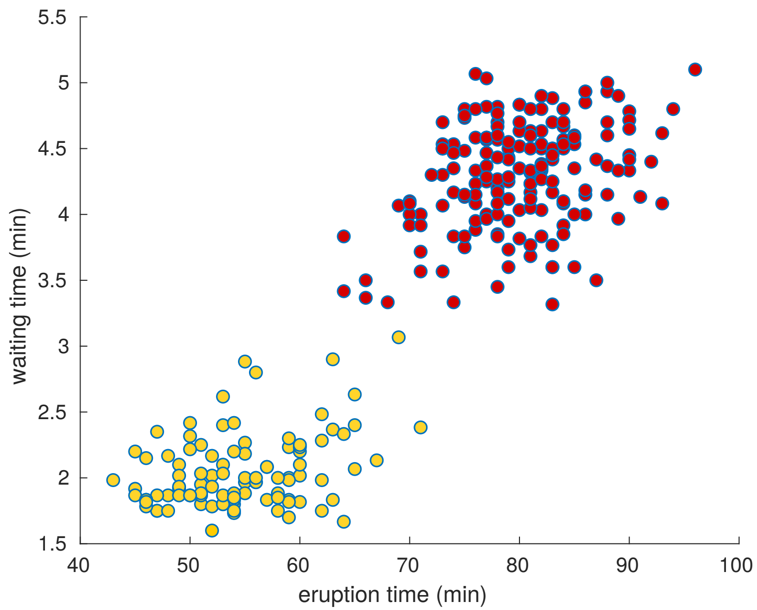
(since Θ^+ maximizes the RHS of ineq when Θ is treated as a dummy variable)

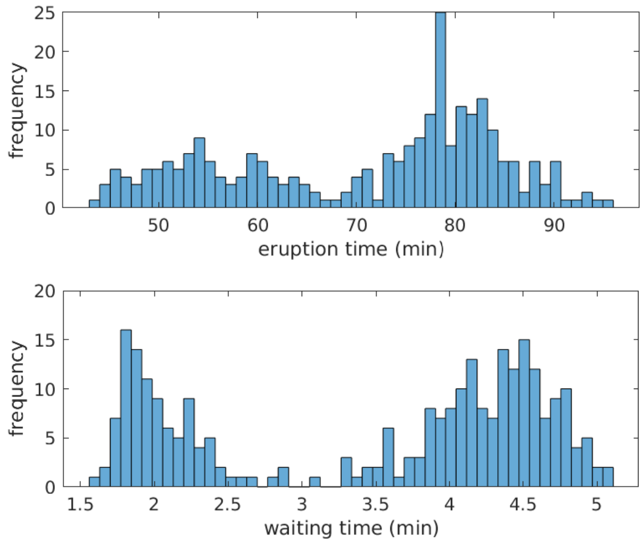
Application on fitting mixture model

geyser at Yellowstone national park, U.S

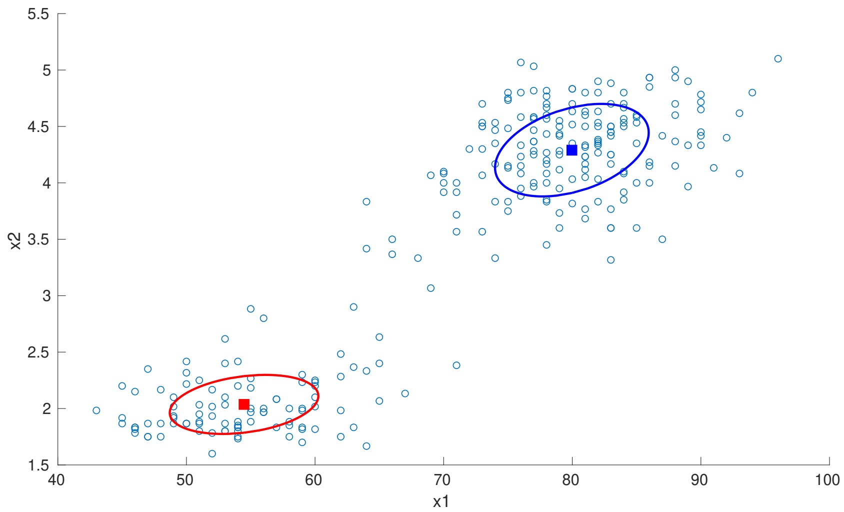
data are eruption time and waiting time

bimodel shapes are apparent





results of fitting two Gaussian mixture model using EM



- Gaussian parameters are

$$\mu_1 = \begin{bmatrix} 79.97 \\ 4.29 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 36.04 & 0.94 \\ 0.04 & 0.17 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 54.48 \\ 2.04 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 33.7 & 0.44 \\ 0.44 & 0.07 \end{bmatrix}$$

- MATLAB (file exchange) codes by Mo Chen
- the result can be compared with k-mean clustering

References

Chapter 7 in

T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition, Springer, 2009

A. Ng, *CS229 Lecture notes on the EM algorithm*

Chapter 9 in

C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006