MA-324

Statistical Inference & Multivariate Analysis

Moodle Pwd: ma324202

We have ♡data♡

We want to analyze, find distribution, etc,-

Graphical Methods
- Histogram
  - Symmetric
  - Peakness
  - Tail thickness
- Q-Q plot

Normal → not good for modelling non-negative values

{Weibull, Gamma, Exp}

- Empirical Sample estimator (To find empirical CDF from data)

i> Order sample $X_{(1)} \le X_{(2)} \cdots \le X_{(n)}$

ii> $P[X_i = X_{(i)}] = 1/n$

iii> $F(x) = P(X \le x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \le x)$

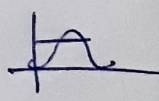
$= \begin{cases} 0 & ; x < X_{(1)} \\ \vdots & \\ 1 & ; x > X_{(n)} \end{cases}$

Another method,

$\hat{f}(x) \approx \frac{F(x+h) - F(x-h)}{2h}$

$\hat{f}(x)$ → freq density (kinda like PDF) at x

$= \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}(x-h \le X_i \le x+h)}{2h}$



$\{X_1, \ldots X_n\}$ (each element is 1D)

Assumptions — univariate, independent

Aim: What is best fitted distribution?

- **Kernels** → some function $K(\cdot)$ following 1) $K() \ge 0$ 2) $\int_{-\infty}^{\infty} K(x)dx = 1$

↳ used for approximating in an interval

Ex- $\mathbb{I}$, $K\left(\frac{X_i - x}{h}\right)$ ← gaussian kernel

Kernel density estimation → $\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\cdot)$

- **Q-Q plot**

  Quantile Quantile plot

Characteristics of distributions:
- Skewness / Symmetric ✱
- Peakness of dist. / Kurtosis ✱
- Existence of outliers / Measure of heavy tail

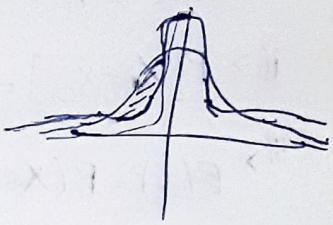
Dist
- Normal
- t-dist
  - Double exponential ↓

$$\frac{1}{2b} e^{-\left|\frac{x-\mu}{b}\right|}$$

$$V(X) = E(X-\mu)^2$$

$$S(x) = E\left[\frac{X-\mu_x}{\sigma_x}\right]^3$$

$$K(x) = E(X-\mu)^4$$, If $K(x) > 3$, high kurtosis



Taking standard dist^n's,

Normal $\rightarrow f(x) \propto e^{-x^2/2}$

DoubleExp $\rightarrow f(x) \propto e^{-|x|}$ $\longrightarrow$ Heavier tail than normal

t-dist $\rightarrow f(x) \propto \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}} \simeq \frac{1}{x^{\nu}}$ $\rightarrow$ Heavier tail than normal

Pareto $\rightarrow f(x) \propto \frac{1}{(1+x)^{\alpha}}$ $\rightarrow$ Heavier tail than double exp

Heavy tail measurement must be with reference.
And,
Both mean & variance of both must be same

- **Pareto**

- **Slowly Varying Function** $\rightarrow$ if $\frac{f(tx)}{f(x)} \to 1$ as $x \to \infty$

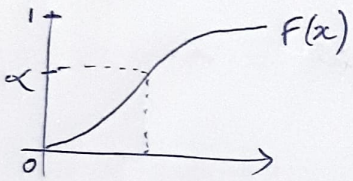
- **Cauchy distribution**

- **Mixture Models**

projects?
- Quantile regression
- VaR

week 2

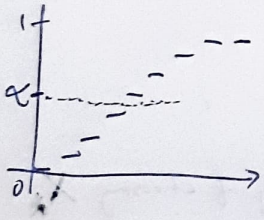
Quantile of a distribution



α^{th} quantile point is $F^{-1}(\alpha) = \{x : F(x) = \alpha\}$

but only holds in case of cts functions

for discrete,



Then, α^{th} quantile pt. is $F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}$

for example: $\hat{F}(x) = \begin{cases} 0 & , x < x_{(1)} \\ 1/n & , x_{(1)} \leq x < x_{(2)} \\ 2/n & , \\ \vdots & \vdots \\ 1 & , x \geq x_{(n)} \end{cases}$

Then, for $0 \leq \alpha < 1/n$, α^{th} quantile is $\inf\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\} = x_{(1)}$

for any general α,

$$F^{-1}(\alpha) = \begin{cases} x_{(\lfloor n\alpha \rfloor)} & , n\alpha \text{ is an integer} \\ x_{(\lfloor n\alpha \rfloor + 1)} & , n\alpha \text{ is not an integer} \end{cases}$$

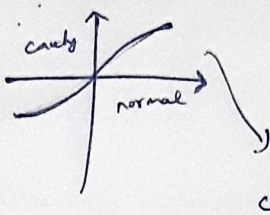
$\rightarrow$ This is true when our 'origin' is $x_{(1)}$

if 'origin' is diff then,

$F^{-1}(\alpha) = \begin{cases} x_{\lfloor n\alpha \rfloor + \text{origin} - 1} & , n\alpha \text{ int} \\ x_{\lfloor n\alpha \rfloor + \text{origin}} & , n\alpha \text{ not int} \end{cases}$

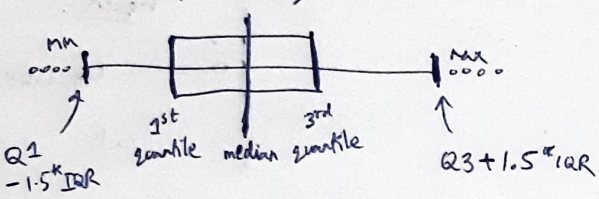
- **QQ plot** → Quantiles of 2 diff distributions along diff axes

	Normal (0,1)	Cauchy (0,1)
$\alpha = 0.999$		
$\alpha = 0.98$		
$\alpha = 0.90$		
$\alpha = 0.85$		
⋮		



Based on this, can compare kurtosis / heavy tail ness, etc of 2 diff distributions

- **Box Plot**



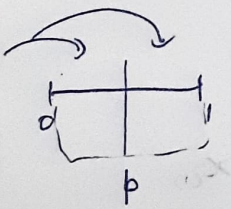
Going from univariate → multivariate

- Extension of Binomial distribution in higher dimension → Multinomial

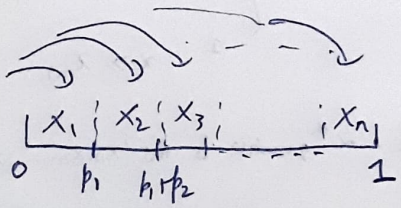
$$(X_1, X_2, \dots X_5) \sim \text{Multinomial}\ (n, p_1, p_2, \dots p_5)$$

$$\sum_{i=1}^{n} X_i = n \quad ; \quad \sum_{i=1}^{n} p_i = 1$$

Generating elements from these dist^{n}s



Bihomial



Multinomial

- $P[X_1 = x_1, \dots X_k = x_k]$

$$= \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

- $P[X_k = x_k] \sim \text{Bin}(n, p_k)$

(proof using total prob thm)

Probability of choosing X_k

$$P\left[\sum_{i=1}^{k-1} p_i < U_i \le \sum_{i=1}^{k} p_i\right]$$

$$= P\left[U_i \le \sum_{i=1}^{k} p_i\right] - P\left[U_i \le \sum_{i=1}^{k-1} p_i\right]$$

$$= \sum_{i=1}^{k} p_i - \sum_{i=1}^{k} p_i$$

$$= p_k \quad //$$

gaze
gays
guaze
guess
garge
fis

• Multinormal

$\{z_1, \dots z_n\} \overset{iid}{\sim} N(0,1)$ ← n independent normals

$$f(\underline{z}) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n} z_i^2}, \quad -\infty \le z_i \le \infty$$

($f(\underline{z})$: multinormal)

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\underline{z}'\underline{z}}$$

note:- for normal dist^n, $cov = 0 \Leftrightarrow$ independance

But not in general

let

$\underline{z} \sim N(0, I)$

$\underline{x} \sim N(\underline{\mu}, \Sigma)$ Then, $E(\underline{x}) = \mu$, $V(\underline{x}) = \Sigma$

Since Σ is symmetric,

we can write $\Sigma = \sum_{i=1}$

⋮

Then, pdf of multivariate normal is, $\{z_1, z_2, \dots z_d\} \overset{iid}{\sim} N(0,1)$

$$f(\underline{x}, \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})}$$

$\underline{x}, \underline{\mu} \to d \times 1$

$\Sigma \to d \times d$

$\underline{z}^{d\times 1} \sim N(0, I)$

$\underline{X} = \underline{\mu} + \Sigma^{1/2}\underline{z}$

$\underline{z} \to$ p.d matrix

We see properties of multivariate normal dist^n now,

i) Moment generating func

$$M_X(t) = E_X\left(e^{t'X}\right)$$

$$= \quad \Big| \quad X = \mu + \Sigma^{1/2} Z \quad , \quad t^{*'} = t'\Sigma^{1/2}$$

If $X \sim N(\mu, \sigma^2)$

$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

Let $= \mu_{t+1}$

If $X \sim N(\mu, \Sigma)$

MGF of X is $E(e^{t'X}) = e^{t'\mu + \frac{1}{2}t'\Sigma t}$

- $Y = AX \sim N(A\mu, A\Sigma A')$

Then, $E[t'Y] =$

Ex: If $X^{d\times 1} = \begin{pmatrix} X_1^{d_1\times 1} \\ X_2^{(d-d_1)\times 1} \end{pmatrix}$, then find dist^n of X_1.

Ans: Let $X_1 = \underbrace{(I^{d_1\times d_1}; O)}_{\text{Take as } A} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$

Take $\mu = \begin{pmatrix} \mu_1^{d_1\times 1} \\ \mu_2^{(d-d_1)\times 1} \end{pmatrix}$.

$$\Sigma = \begin{pmatrix} \Sigma_{11}^{d_1\times d_1} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then, $X_1^{d\times 1} \sim N(\mu_1^{d\times 1}, \Sigma_{11})$

ii) $\underline{X}_1$ and $\underline{X}_2$ are ind $\iff \Sigma_{12} = \Sigma_{21} = 0$

$$M_X(\underline{t}) = E(e^{\underline{t}'X}) = E\left(e^{\underline{t}_1'X_1 + \underline{t}_2'X_2}\right)$$

$$= e^{(\underline{t}_1', \underline{t}_2')\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} + (\underline{t}_1', \underline{t}_2')\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\begin{pmatrix}t_1\\ t_2\end{pmatrix}}$$

$$\vdots$$

$$= M_{X_1}(t_1) \cdot M_{X_2}(t_2) \quad //$$

- $(X_1, X_2) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$

Then, $X_1 | X_2 = x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_2}{\sigma_1}(x_2 - \mu_2), \sigma_2^2(1-\rho^2)\right)$

Given $(X_1, X_2) \sim N(\mu, \Sigma)$

$$\Rightarrow (X_1 | X_2 = x_2) \sim N\left(\mu_{1|2} + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

Proof:

$$W = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$$

$$Y = X_2$$

Then, $\begin{pmatrix}W\\ Y\end{pmatrix} = AX = \begin{pmatrix}I^{d_1 \times d_1} & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I^{\overline{d-d_1}\times\overline{d-d_1}}\end{pmatrix}\begin{pmatrix}X_1^{d_1\times 1}\\ X_2^{\overline{d-d_1}\times 1}\end{pmatrix}$

$$V\begin{pmatrix}W\\ Y\end{pmatrix} = A\Sigma A^T = \begin{pmatrix}I & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I\end{pmatrix}\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}I & 0\\ -\Sigma_{12}\Sigma_{22}^{-1} & I\end{pmatrix} = \begin{pmatrix}\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0\\ 0 & \Sigma_{22}\end{pmatrix}$$

Var??

$$E\begin{pmatrix}W\\ Y\end{pmatrix} = \begin{pmatrix}\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2\\ \mu_2\end{pmatrix}$$

Since, W and Y are ind, dist^n of $W|Y$ is same as W.

So, $(W|Y) \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

Putting $X_2 = x_2$

$$(X_1 \mid X_2 = x_2) \sim N\left(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12} \cdots, \cdots\right)$$

$$\sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

→ Statistical Inference
- Parameter-based → distributional assumption
- Non-parametric → no assuming dist^n

Parameter-based approaches

- Frequentist based (θ is fixed)
  - Maximum likelihood estimate
  - method of moments
  - Minimax algo
  - Uniformly minimum variance unbiased estimate
- Bayesian approach (θ is random)

Hogg & Craig

Lee notes

- Loss/Risk for estimation

$L(\hat{\theta}, \theta)$ ←

loss

for example:

$$L(\hat{\theta}, \theta) = \left(\hat{\theta}(x_1, \cdots, x_n) - \theta\right)^2$$

$$|\hat{\theta} - \theta| = \begin{cases} 1, & |\hat{\theta} - \theta| > \varepsilon \\ 0, & |\hat{\theta} - \theta| \leq \varepsilon \end{cases}$$

$$\text{Risk} = E\left(L(\hat{\theta}, \theta)\right)$$

$\{X_1, \ldots, X_n\} \sim F(x, \theta)$ — cdf

$\theta \to$ fixed (Frequentist setup)

we need to guess θ.

$$\text{Risk} = \mathbb{E}\left(L(\hat{\theta}(X), \theta)\right)$$

$$= \int_{\underline{x}} L(\hat{\theta}(\underline{x}), \theta)\, f(\underline{x} \mid \theta)\, d\underline{x}$$

f: func of θ

But when $\theta \to$ random, how to take average of loss (∵ we have 2 rv's, X and θ)

↓

- **Bayesian Approach**

$$P(B_i \mid A) \propto P(A \mid B_i)\, P(B_i)$$

posterior — likelihood function — prior

[chances of observing the available info]

relate this with $\pi f(x_i \mid \theta)$

Bayes Thm

$$P(B_i \mid A) = \frac{P(A \mid B_i) P(B_i)}{\sum P(A \mid B_i) P(B_i)}$$

- **Maximum Likelihood Estimate (MLE)**

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\ln L(\theta)\right]$$

$$p(\theta \mid x) \propto p(x \mid \theta)$$

Let $\pi(\theta)$: prior dist^n

$$p(\theta \mid \underline{x}) = \frac{f(x \mid \theta)\, \pi(\theta)}{\int f(\underline{x} \mid \theta)\, \pi\theta\, d\theta}$$

($\underline{x}$: observed data)

$$\propto f(x \mid \theta)\, \pi(\theta)$$

(f: likelihood, π: prior)

$$\text{Risk} = \underset{(\theta \mid \underline{x})}{\mathbb{E}}\left(L(\hat{\theta}, \theta)\right) \quad (\because \theta \text{ is random})$$

Now, we can minimise risk and find a θ.

If loss is MSE, (and under frequentist setup)

$$E(L(\hat{\theta},\theta)) = E_x\left(\hat{\theta}-\theta\right)^2$$

$$= E_x\left(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta\right)^2$$

$$= E\left(\hat{\theta}-E(\hat{\theta})\right)^2 + \left(E(\hat{\theta})-\theta\right)^2 = \text{Variance} + \text{Bias}^2$$

$\downarrow$

Bias $= 0 \Rightarrow$ set of unbiased estimator

$\Downarrow$

Choose $\hat{\theta}$ $\rightarrow$ variance becomes minimum

$\Downarrow$

Uniformly minimum variance unbiased estimator (UMVUE)

- Maximum likelihood

If $\{X_1,\dots,X_n\} \overset{iid}{\sim} N(\mu,1)$

Then $L(\theta) = f(x_1,\dots,x_n|\mu)$

$= \prod f(x_i|\mu) = \frac{1}{(2\pi)^{n/2}} e^{\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$

Ex:

$\{X_1,\dots X_n\} \sim f(x;\theta)$

relapse time $\rightarrow$ $Y_i = \begin{cases} c, & X_i > c \\ X_i, & X_i < c \end{cases}$ Find maximum likelihood of Y.

$\downarrow$

$$L(\theta) = \prod_{i=1}^{n} f^{\delta_i}(x_i,\theta)\, P[X_i>c]^{(1-\delta_i)}$$

$$\delta_i = \begin{cases} 0, & X_i > c \\ 1, & X_i < c \end{cases}$$

$$\frac{\partial \ln L(\theta)}{\partial \mu} = \frac{\partial}{\partial \mu}\left[\frac{-1}{2}\sum (x_i-\mu)^2\right] = 0$$

$\Rightarrow \hat{\mu} = \bar{x}$

!! this is truncation on time, consider ex. of truncation on # of obs.

Then, $Y_i = \begin{cases} X_i, & i<r \\ X_n, & i \geq r \end{cases}$

Ex: Weibull — β, θ

$$(\hat{\beta}, \hat{\theta}) = \underset{\beta,\theta}{\operatorname{argmax}} \ln L(\beta, \theta)$$

$$\ln L(f(x_i, \beta), \beta) = \text{profile log likelihood for } \beta$$

EX:

$$f(x;\theta) = \begin{cases} 1/\theta, & 0 \le x_i \le \theta \\ 0, & \text{o.w.} \end{cases}$$

order as $x_{(1)} < \cdots < x_{(n)}$

Then,

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) = \begin{cases} 1/\theta^n, & 0 \le x_{(1)} \le \cdots < x_{(n)} \le \theta \\ 0, & \text{o.w.} \end{cases}$$

Now, $L(\theta)$ is ↓sing func.

So, max likelihood is at min θ.

To see clearly, let $I(a,b) = \begin{cases} 1, & a \le b \\ 0, & \text{o.w.} \end{cases}$ bcz $0 \le x_{(1)} \Leftrightarrow x_i \ge 0 \ \forall i$

$x_{(n)} \le \theta \Leftrightarrow x_i < c \ \forall i$

Then, $L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) = \frac{1}{\theta^n} I(0, x_{(1)}) \, I(x_{(n)}, \theta)$

EX: $f(x, \theta) = \begin{cases} 1, & \theta \le x \le \theta + 1 \\ 0, & \text{o.w.} \end{cases}$

$$\prod_{i=1}^{n} f(x_i, \theta) = \begin{cases} 1, & \theta \le x_{(1)} < \cdots < x_{(n)} \le \theta + 1 \\ 0, & \text{o.w.} \end{cases}$$

$$= I(\theta; x_{(1)}) \, I(x_{(n)}; \theta + 1)$$

So, MLE is $\in$, $x_{(n)} - 1 \le \theta \le x_{(1)}$ → so MLE may not be unique.

Ex: $X \sim bin(n, p)$

Tha, $L($

Designing perspective

it down
it down
it down

political aspect

surveillance system

Arabin is always correct
- Anant

depending on evolution

- EM algorithm

① Introduce 'missing observations' within the system

② $\underset{z_i | x, \theta^t}{E} f_\theta(x_i, z_i) =$ ~~complete~~ psuedo likelihood func $= f(x, \theta, \theta^t)$

z_i: unknown / hidden

- $p(z_i | x, \theta^t) \equiv$ posterior of z given $x, \theta^t \rightarrow$ known

Expectation step

$$Q(\theta, \theta^t) = \underset{z|x,\theta^t}{E}(f(x, z|\theta))$$

Maximization step

$$\hat{\theta}_{EM} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^t)$$

$\hat{\theta} \rightarrow \theta_{true}$

$q(x, \theta)$

So, $L(\theta) = \ln f(x, \theta) =$ likelihood based on incomp. data

$$\ln L(\theta) - \ln L(\theta^t) = \ln \frac{f(x|\theta)}{f(x|\theta^t)}$$

↑ we show this ≥ 0 at $\theta = \hat{\theta}$

$$= \ln \frac{\int f(x, z|\theta)\, dz}{f(x; \theta^t)}$$

$$= \ln \int \frac{f(x, z|\theta)}{f(x|\theta^t)}\, dz$$

$z \rightarrow$ unknown RV

$q \rightarrow$ dist^n of z given x

$$= \ln \int \left[\frac{f(x, z|\theta)\, q(z|x, \theta^t)}{f(x, \theta^t)\, q(x, z|\theta^t)}\right] dz$$

$$\geq \int \ln \left[\frac{f(x, z|\theta)\, q(z|x, \theta^t)}{f(x, z|\theta^t)}\right] dz = \Delta L(\theta, \theta^t)$$

(By jensen's ineq)

Since log is concave func

$\Rightarrow \ln L(\theta) - \ln L(\theta^t) \geq \Delta L(\theta, \theta^t)$

Say $q \propto f(x, z | \theta^t)$

everything is concave - Me

Then, $\int \left[\ln \frac{f(x,z|\theta)}{f(x,z|\theta^t)} \right] f(z|x,\theta^t)\, dz$

$= \int \ln f(x,z|\theta)\, q(z|x,\theta^t)\, dz - \int \ln f(x,z|\theta^t)\, q(z|x,\theta^t)\, dz$

↑ This term θ is not random, so when taking derivative, it goes to 0.

$\Rightarrow$

- Gaussian Mixture model as clustering framework
- K-Means
- Spectral based cluster

will provide assignment
can explore beyond also

} for project

→ EM algo in "estimating the parameter of two mixture of normal distribution
to figure out $p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$.

$$f(x; p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p f_1(x; \mu_1, \sigma_1^2) + (1-p) f_2(x; \mu_2, \sigma_2^2)$$

$X \rightarrow$ may be from f_1 or f_2

let $Z_i = \begin{cases} 1, & X_i \sim f_1 \\ 0, & X_i \sim f_2 \end{cases}$ Then, (X_i, Z_i), $i = 1, 2, \ldots, n$

↑ complete observations

Then,

$$f(x_i, z_i) = f_i^{z_i}(x_i)\, f_2(x_i)^{(1-z_i)}$$

↖ but this is not the likelihood

↓

the chance of observing x_i

So,

$$P[x_i \le X_i \le x_i + \Delta x_i,\ Z = z_i] = P[x_i \le X \le x_i + \Delta x_i \mid Z_i = z] \cdot P[Z_i = z_i]$$

$$= p_i f(x_i)$$

So, actual likelihood is,

$$f(x_i, z_i) = \left(p f_1(x_i)\right)^{z_i} \left((1-p) f_2(x_i)\right)^{(1-z_i)}$$

Complete likelihood is (for 1 to n) $= \prod_{i=1}^{n} f(x_i, z_i) = \prod_{i=1}^{n} \left[\left(p f_1(x_i)\right)^{z_i} \left((1-p) f_2(x_i)\right)^{1-z_i}\right]$

Then taking log,

$$\ln L = \sum_{i=1}^{n} z_i \left[\ln p + \ln f_1(x_i, \mu_1, \sigma_1^2)\right] + (1 - z_i)\left[\ln(1-p) + \ln f_2(x_i, \mu_2, \sigma_2^2)\right]$$

$$\text{Pseudolikelihood} = E_{(z \mid x, \theta^t)}\left(\ln L(\theta)\right)$$

$$= \sum_{i=1}^{n} (\ln p) \underbrace{E_{z \mid x, \theta^t}(z_i)}_{\text{I}} + \underbrace{\sum_{i=1}^{n} E_{z \mid x, \theta^t}\left(z_i \left(\ln f_1(x_i; \mu_1, \sigma_1^2)\right)\right)}_{\text{II}} + \sum_{i=1}^{n} \ln(1-p) \underbrace{E_{z \mid x, \theta^t}(1 - z_i)}_{\text{I}} + \underbrace{\sum E\left((1-z_i) \ln f_2(x_i; \mu_2, \sigma_2^2)\right)}_{\text{III}}$$

Then,

$$\underbrace{E(z_i)}_{I} = E\left[\sum_{i=1}^{n} z_i\, P[z_i \mid X, \theta^t]\right] \leftarrow \text{(from total probability theorem)}$$

$$= E\left[\sum_i z_i\, P[z_i \mid x_i, \theta^t]\right] \quad \text{(depends on only } x_i \text{, independent of others)}$$

$$= \sum_i P[z_i = 1 \mid x_i = x, \theta^t]$$

$$= \sum_i \frac{p f_1(x; \mu_1, \sigma_1^2)}{p f_1(x; \mu_1, \sigma_1^2) + (1-p) f_2(x; \mu_2, \sigma_2^2)} \rightarrow m_i^t$$

(Here $\theta^t = (p^t, \mu_1^t, \sigma_1^{2t}, \mu_2^t, \sigma_2^{2t})$ ↓ all are known. So, this expectation (x) is known)

So, pseudolikelihood $= \sum_{i=1}^{n} m_i^t \ln p + \sum_{i=1}^{n} m_i^t [\ln f_1(\ \ )]$

$+ \sum_{i=1}^{n} (1-m_i^t) \ln(1-p) + \sum_{i=1}^{n} (1-m_i^t) \ln f_2(\ \ )$

Now, take derivative wrt $p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$, and maximise to get,

$$\hat{p}^{t+1} = \frac{1}{n} \sum_{i=1}^{n} m_i^t$$

$$\hat{\mu}_1^{t+1} = \frac{1}{\sum m_i^t} \sum_{i=1}^{n} m_i^t x_i$$

$$\hat{\sigma}_1^{2\,t+1} = \frac{1}{\sum m_i^t} \sum_{i=1}^{n} m_i^t (x_i - \hat{\mu}_1^{t+1})^2$$

$$\hat{\mu}_2^{t+1} = \frac{1}{\sum (1-m_i^t)} \sum_{i=1}^{n} (1-m_i^t) x_i$$

$$\hat{\sigma}_2^{2\,t+1} = \frac{1}{\sum (1-m_i^t)} \sum_{i=1}^{n} (1-m_i^t)(x_i - \hat{\mu}_2^{t+1})^2$$

→ This can be used for estimating & clustering

One iteration formula has been converted to single dimensional problem

This EM algo will converge rapidly even for far-away starting points

---

- Calculation of MLE in case of Multivariate Normal dist^n

$\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$, $\mathcal{D} = \{ \underset{\sim}{x}_1^{d\times 1}, \ldots, \underset{\sim}{x}_n^{d\times 1} \}$ are iid

d^2+d unknown parameters

(atleast these many samples needed)

Arabic

catch sleeper nurture scratch

→ Vector calculus tangent for 30 mins

$f: \mathbb{R}^n \to \mathbb{R}$, $f(\underset{\sim}{x}) = x^T A x$

$$\frac{df(x)}{d\underset{\sim}{x}^{n\times 1}} = \left[\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_n}\right]$$

$$\frac{\partial f}{\partial x_i} = (x^T A + x^T A^T)_{x_i} = x^T A + x^T A^T$$

$$= 2x^T A \quad (A \text{ is sym})$$

Matrix calculus

$f : \mathbb{R}^{m \times n} \to \mathbb{R}$

Result 1 $\quad \frac{d\, tr(AB)}{dA} = B$

Derivative will be,

Result 2

If A is symmetric, $\frac{d \ln|A|}{dA} = A^{-1}$

- MLE for multivariate normal dist^n

→ pdf is $f(x_i) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{N/2}} e^{\frac{-1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$

$$L(\underline{\mu}, \Sigma; x) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{N/2}} e^{-1/2 \sum_{i=1}^{N} (x_i-\underline{\mu})^T|\Sigma|^{-1}(\underline{x_i}-\underline{\mu})}$$

$$\ln L(\underline{\mu}, \Sigma) = \ln\left[\frac{1}{(2\pi)^{N/2}|\Sigma|^{N/2}} e^{-\frac{1}{2}\sum_{i=1}^{N}(\underline{x_i}-\underline{\mu})^T\Sigma^{-1}(\underline{x_i}-\underline{\mu})}\right]$$

wrt μ,

$$\Rightarrow \frac{\partial \ln L(\underline{\mu}, \Sigma)}{d\mu} = 0$$

$$\Rightarrow \frac{-1}{2} \times 2 \sum_{i=1}^{N} (x_i-\mu)^T\Sigma^{-1} = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \underline{x_i}$$ //

Wrt Σ, take derivative wrt Σ^{-1} instead

$$\Rightarrow \frac{\partial \ln L(\mu, \Sigma)}{\partial \Sigma^{-1}} = 0$$

$$\Rightarrow \frac{N}{2}\Sigma - \frac{d}{d\Sigma^{-1}} \frac{1}{2}\sum_{i=1}^{N} tr\left(\Sigma^{-1}(x-\mu)(x-\mu)^T\right) = 0$$

$$\Rightarrow \frac{N}{2}\Sigma - \frac{1}{2} \times \sum_{i=1}^{N} (x_i-\hat{\mu})(x_i-\hat{\mu})^T = 0$$

$$\Rightarrow \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\underline{x_i}-\underline{\hat{\mu}})(\underline{x_i}-\underline{\hat{\mu}})^T$$ //

→ Gaussian mixture Model (GMM)

$$\underline{X} \sim \sum_{k=1}^{K} \pi_k f(\underline{x}, \mu_k, \Sigma_k)$$

- For univariate GMM,

$$f(x; \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \pi f_1(x; \mu_1, \sigma_1^2) + (1-\pi) f_2(x; \mu_2, \sigma_2^2)$$

for $\underline{x} = \{x_1, \ldots, x_n\}$

$\underline{z} = \{z_1, \ldots, z_n\}$ then $\hat{\pi}, \hat{\mu_i}, \hat{\sigma_i^2}, \ldots$

For higher dimensional GMM's,

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} m_{ik}^t \quad , \quad z_{ik} = \begin{cases} 1, & x_i \sim f_k \\ 0, & \text{o.w.} \end{cases}$$

$$\underset{\sim}{\mu}_k = \frac{1}{\sum_{i=1}^{N} m_{ik}^t} \sum_{i=1}^{N} m_{ik}^t \underset{\sim}{x}_i$$

$$\hat{\Sigma}_k = \frac{1}{\sum_{i=1}^{N} m_{ik}} \sum_{i=1}^{N} m_{ik}^t (\underset{\sim}{x}_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

- GMM can be used for clustering also.

→ K-Means Algorithm

↳ assumption: # of clusters = K

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & K \end{array} \rightarrow \text{clusters} \\ \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 \; 1 \end{bmatrix} \end{array}$$

↑ obs

Let $\gamma_{ik} = \begin{cases} 1, & \text{if } x_i \text{ is in } k^{th} \text{ cluster} \\ 0, & \text{o.w.} \end{cases}$

Then objective func is,

$$J = \frac{1}{N} \sum_i \sum_k \gamma_{ik} \| \underset{\sim}{x}_i - \underset{\sim}{\mu}_k \|^2$$

Algo

2. Fix γ, get best μ to minimise J
1. Fix μ, compute γ (by minimizing J)

(To get μ, (1)

$$\frac{\partial J}{\partial \underset{\sim}{\mu}_k} = 2 \sum_{i=1}^{\hat{}} \gamma_{ik} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_k)^T = 0 \Rightarrow \hat{\mu}_k = \frac{1}{\sum_{i=1}^{\hat{}} \gamma_{ik}} \sum_{i=1}^{\hat{}} \gamma_{ik} \underset{\sim}{x}_i$$

Exam

How is K-means connected to EM algo / some MLE ?

↕

(expression is similar to GMM, where $m_{ij} \equiv \gamma_{ij}$)

GMM → soft thresholding
K-means → 0/1

So, in GMM, can take cluster as maximum m_{ij} value

book - Bishop

K-Means	GMM
① Minimizing MSE	minimizing log likelihood
② Hard thresholding	Soft thresholding
③ Good for spherical clusters	Good for non-spherical clusters ← important point for project (Extend more on this) ↳ elliptic shapes, spectral cluster, etc. ☆ (will do later)

← for project also

- MLE

Thm
- $\hat{\theta}$ is an MLE of θ
- $g(\cdot)$ is an onto function
- Then, $g(\hat{\theta})$ is an MLE for $g(\theta)$

~~Statistical Inference~~ (for bernoulli) (see later)

Proof: (Imp)

Since $g(\cdot)$ is onto,

$\theta \in H \rightarrow$ domain

$$\Lambda_y = \{\theta : g(\theta) = y\}, \quad H = \bigcup_y \Lambda_y$$

Now,

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left[\max_{\theta \in \Lambda_y} L(\theta) \right] = \underset{\theta \in H}{\operatorname{argmax}} L(\theta)$$

If $\hat{\theta}$ is an MLE of $\theta \Rightarrow \hat{\theta} = \underset{\theta \in H}{\operatorname{argmax}} L(\theta)$

Now,

$$\hat{y} = \hat{g}(\theta) \iff g(\hat{\theta}) = y^* \Rightarrow \theta \in \Lambda_{y^*}$$

(‖ same as)

where $\Lambda_{y^*} = \{\theta : g(\theta) = y^*\}$

claim: ① $L(\hat{\theta}) = \max_{\theta \in \Lambda_{y^*}} L(\theta)$, then $\Rightarrow$ $\hat{\theta}$ which is arbitrary point in Λ_{y^*} gives the max of L

↳ this shows $y^* = \hat{y}$, so then $g(\hat{\theta}) = y^*$

Proof of claim ①,

$$L(\hat{\theta}) \geq \max_{\theta \in \Lambda_{y^*}} L(\theta)$$ ← from definition of $L(\theta)$ and $\hat{\theta}$

$$L(\hat{\theta}) \leq \max_{\theta \in \Lambda_{y^*}} L(\theta)$$ ← since $\hat{\theta}$ is an arbitrary pt in Λ_{y^*}

- Information Function

$$I(\theta) = E_x\left[-\frac{\partial^2}{\partial \theta^2} \ln f(x;\theta)\right] \overset{?}{=} E\left(\frac{\partial}{\partial \theta} \ln f(x,\theta)\right)^2 = V\left(\frac{\partial}{\partial \theta} \ln f(x;\theta)\right)$$

= expected value of hessian of $\ln f(x;\theta)$

$\frac{\partial}{\partial \theta} \ln f(x,\theta)$ called 'score' function of f

Finding,

$$E\left(\frac{\partial}{\partial \theta} \ln f(x;\theta)\right)$$

$$= \int_x \left[\frac{\partial}{\partial \theta} \ln f(x;\theta)\right] f(x;\theta)\, dx$$

$$= \int_x \frac{1}{f(x;\theta)} \left(\frac{\partial}{\partial \theta} f(x;\theta)\right) f(x;\theta)\, dx$$

$$= \frac{\partial}{\partial \theta}\left[\int_x f(x;\theta)\, dx\right]$$ ↳ always 1

$$= 0$$

(under some regularity(?) conditions, we can interchange $\int$ and $\frac{\partial}{\partial x}$)

↓

① $f(x;\theta)$ is bounded func in x and domain of x doesn't depend on θ

② $f(x;\theta)$ has ∞ support, cts diff$^\Delta$ and integral uniformly converges for all θ

Then, $\int_x \left[\frac{\partial^2}{\partial \theta^2} \ln f(x;\theta)\right] f(x;\theta)\, dx =$

- If $X_1, \cdots, X_n \overset{iid}{\sim} f(x;\theta)$

Then, $I_n(\theta) = n I(\theta)$

Proof:

$$I_n(\theta) = E\left[\frac{-\partial^2}{\partial\theta^2} \sum_{i=1}^{n} \ln f(x;\theta)\right]$$

$$= \sum_{i=1}^{n} E\left(\frac{-\partial^2}{\partial\theta^2} \ln f(x;\theta)\right)$$

$$= \sum_{i=1}^{n} I(\theta) = n\,I(\theta) \quad //$$

$$V(\hat{\theta})_{MLE} \approx \frac{1}{n\,I(\theta)} \quad \text{for large } n$$

$\nexists\ T(X) = \hat{\theta}_{MLE}$

$$E(T(X)) = g(\theta) \quad \forall\theta$$

$T(X)$ is an unbiased estimator

Then, $V(T(X)) \geq \dfrac{g'(\theta)^2}{n\,I(\theta)}$ (inequality true only in case of MLE)