Lecture 20

Monday, September 7, 2020 4:07 PM

* **Gaussian Distribution :-** $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- Standard Normal Distribution $\Rightarrow \boxed{\sigma = 1}$

  Verify by $\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)^2 = \int_{-\infty}^{\infty}\int e^{-(x^2+y^2)} dx\,dy$ and convert to polar.

- The CDF for $\mu=0$, $\sigma=1$ is:- $F_x(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$

  Error Function is defined as $\frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

  $\rightarrow F_x(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)$

  Why? idk.

  (What is Error function?)

- **MGF :-** $\phi_x(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$ —— (**)

  Proof:- Solve for $N(0,1)$ to get $e^{t^2/2}$ as MGF

  But $N(0,1)$ is just replacing X with $Y = \frac{X-\mu}{\sigma} \Rightarrow X = \mu + \sigma Y$

  $$\phi_{\sigma Y+\mu}(t) = E\left[e^{(\sigma Y+\mu)t}\right] = E\left[e^{\mu t} \cdot e^{Y(\sigma t)}\right] = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

$\hookrightarrow$ Notice that calculating mean and Variance from MGF becomes Easy!

Mean $= E[\phi'_x(0)] = E[(e^0)(\mu+0)] = \mu$!

Similarly for Variance!

Lecture 21

Monday, September 7, 2020 10:36 PM

* **Central Limit Theorm:-** MATLAB Example

$X_1, X_2, \ldots X_n$ be n independent and identically distributed variables; all with μ, σ. Define new Y s.t

$$\boxed{Y_n = \sqrt{n}\left[\frac{\sum X_i}{n} - \mu\right]}$$

Lindberg's Central Limit theorm (not in Syllabus!)

As $n \to \infty$; Distribution of $Y_n \to N(0, \sigma^2)$

"Y_n converges in Distribution to $N(0,\sigma^2)$"

But we can see that it still is Gaussian.

– We can usually model Errors as Gaussian as they're independent!

Emperically; Lindberg's holds when σ of some X_i is not HUGE compared to rest.

* **LAW OF LARGE NUMBERS & CLT:-**

– We can tell that they're related by the fact :- for $Y = X\sqrt{n} \Rightarrow X = N(0, \frac{\sigma^2}{n})$

$Var \to 0$ as $n \to \infty$

Proof of CLT

Rephrase as:- $\lim_{n\to\infty} P\left(\underbrace{\frac{\sum X_i - n\mu}{\sqrt{n}\sigma}}_{Z_n} \le x\right) \to \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt$

$\to N(0,1)$

As all X_i are i.i.d; let $X' = X - \mu$

$$\Rightarrow \phi_{\sum X_i - n\mu} = \left[\phi_{X'}(t)\right]^n \text{ ;- As } \phi_{X+Y}(t) = \phi_X(t)\cdot\phi_Y(t)$$

Now; $Z_n = \left(X'/\sigma\sqrt{n}\right) \Rightarrow \phi_{Z_n}(t) = \left[\phi_{X'}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$ from prop.

Rephrasing further; RTP:- $\lim_{n\to\infty} n\cdot \log\left[\phi_{X'}\left(\frac{t}{\sigma\sqrt{n}}\right)\right] = t^2/2$

$$n \to 1/x^2 \Rightarrow \lim_{x\to 0} \frac{\log\left[\phi_{X'}\left(\frac{tx}{\sigma}\right)\right]}{x^2} \Rightarrow \frac{\frac{d}{dx}\phi_{X'}\left(\frac{tx}{\sigma}\right)}{\phi_{X'}\left(\frac{tx}{\sigma}\right)\cdot 2x}$$

$$\Rightarrow E\left[\frac{tX}{\sigma} e^{\frac{txX'}{\sigma}}\right] \Big/ \phi_{X'}\left(\frac{tx}{\sigma}\right)\cdot 2x$$

Still (0/0)

In brief; convert to X' and apply L'hospital to prove MGFs of Both sides are the same.

$$\Rightarrow E\left[\frac{tX}{\sigma} e^{\frac{\text{[illegible]}}{\sigma}}\right] \Big/ \phi_{X'}\left(\frac{tx}{\sigma}\right)\cdot 2x \qquad \text{Still } (0/0)$$

$$\Rightarrow \frac{E\left[\left(\frac{tX'}{\sigma}\right)^2 e^{\frac{t\,x\,X'}{\sigma}}\right]}{2\phi_{X'}\left[\frac{tx}{\sigma}\right] + 2x\cdot E\left[\left(\frac{tX}{\sigma}\right. e^{\frac{t\,x\,X'}{\sigma}}\right]} \Rightarrow \frac{t^2}{2\phi_{X'}(0)} = \frac{t^2}{2} \qquad (\text{RHS!})$$

Lecture 22

Thursday, September 10, 2020 4:53 PM

- Application of CRT :- 5200 heads in 10000 tosses

CDF of Gauss.

Apply CRT and get μ, σ; Find prob for $\Phi(\mu - n\sigma)$ to $\Phi(\mu + n\sigma)$

* Tail Bound for Gaussian :-

Defined as $P(X>x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \leq \int_x^\infty \left(\frac{1}{x\sqrt{2\pi}}\right) t e^{-t^2/2} dt$ $\rightarrow \boxed{x>0}$

$$\Rightarrow \boxed{P(X>x) \leq \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}} \quad (**)$$

* Distribution of Sample Mean :-

$$Y = \frac{\sum X_i}{n} \Rightarrow \left.\begin{aligned} E(Y) &= \mu \\ Var(Y) &= \frac{\sigma^2}{n} \end{aligned}\right] \rightarrow$$

All X_i independent

X_i are i.i.d with σ, μ

* X_i - Random Normal Variable ; Prove for $\bar{X}$

* Distribution of Sample Variance :-

$$Y = S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1} = \frac{\sum (X_i^2) - n\bar{X}^2}{n-1}$$

RV as well!

- $E(S^2) = \frac{1}{n-1} \cdot \left[nE(X^2) - nE(\bar{X}^2) \right] = \frac{n}{n-1}\left[\sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)\right]$

$$\Rightarrow E(S^2) = \sigma^2$$ — Reason for $(n-1)$ in defn of Var.

- We now learn χ^2-Dist to understand this better.

* χ^2-Distribution :-

- $X = Z_1^2 + Z_2^2 + \dots + Z_n^2$ is χ^2-dist. with n-degrees of freedom

  $\Downarrow$

  $X \sim \chi_n^2$ $\quad\hookrightarrow Z_i = $ independant $N(0,1)$ (Standard Normal)

- $f_X(x) = \dfrac{x^{\frac{n}{2}-1} e^{-x/2}}{2^{n/2}\,\Gamma(n/2)}$

$\Gamma(x) = (x-1)!$ if $x = \mathbb{Z}$

$= \int_0^\infty t^{x-1} e^{-t} dt$ for $x = \mathbb{R}$

$\downarrow$

Derive for χ_1^2 ; pretty brain dead ...

Lecture 23

Thursday, September 10, 2020 5:36 PM

* MGF of χ^2_n – $\phi_x(t) = (1-2t)^{-n/2}$ defined only for $t < 1/2$!

Derivation:- $\phi_x(t) = (e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f_x(x)\,dx$ where $f_x(x) = c \cdot x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$

$\phi_x(t) = c\int_0^{\infty} x^{\frac{n}{2}-1} e^{-x/2} e^{tx}\,dx$ → χ^2_n always +ve by defn.

– proceed; Sub $\int$ with $\Gamma(n/2)$

– Simplify.

* Properties:-

1) $\chi^2_t = \chi^2_n + \chi^2_m$ and $t = m+n$ – (*) Additive prop.

↳ only if X_n & X_m are independant!

– Going back to distribution of S^2;-

$$(n-1)S^2 = \sum (x_i - \bar{x})^2 = \sum (x_i-\mu)^2 - n(\mu-\bar{x})^2$$

↓

Rewrite as,

$$\sum\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{\sum(x_i-\bar{x})^2}{\sigma^2} + \left[\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\right]^2$$

1 2 3

1:- Sum of 'n' $N(0,1)$ ↓ χ^2_n ; 3:- is a $N(0,1)$ itself by CLT. ↓ χ^2_1

– Turns out 1 and 3 are independent. 2 → χ^2_{n-1}

• Notice that 2 is $\frac{n-1}{\sigma^2} S^2$!

* Uniform Distribution:- $f_x(x) = 1/b-a$ if $x \in (a,b)$

0 otherwise.

- $E(x) = a+b/2$
- $Var(x) = (b-a)^2/12$
- MGF :- $\phi(x) = \begin{cases} \dfrac{e^{tb}-e^{ta}}{(b-a)t} & t \neq 0 \\ 1 & t = 0 \end{cases}$

Simple Calculations

- Application :- Random permutation of Set :-

↳ All possible sets have same prob.

$A \equiv$ Original Set $\qquad B_K =$ Subset of length k

$A_i \equiv i^{th}$ Element $\qquad I_K = \begin{cases} 1 & \text{if } A_i \in B_K \\ 0 & \text{otherwise} \end{cases}$

↳ For any B_K; $P(I_1 = 1) = K/n \implies \left. \begin{aligned} P(I_2 = 1 | I_1 = 1) &= K-1/n-1 \\ P(I_2 = 1 | I_1 = 0) &= K/n-1 \end{aligned} \right] - P(I_2 = 1/I_1) = \dfrac{K - I_1}{n-1}$

↳ From this, we can prove $\boxed{P(I_j | I_1, \dots I_{j-1}) = \dfrac{k - \sum_{i=1}^{j-1} I_i}{n-j+1}}$ ← (Try to prove yourself!)

Lecture 24

Sunday, September 13, 2020 12:27 PM

* **Poisson Distribution:-** Sequence of independant Bernoulli trails

- Simply put, Binomial for $n \to \infty$. $\boxed{P(x=i) = \frac{\lambda^i}{i!} e^{-\lambda}}$; $\boxed{\lambda = np}$ $\Longrightarrow$ As we $\uparrow n$; p drops!

  Poisson Limit theorm, $\boxed{i \geq 0}$

- λ = number of Expected outcomes (Constant)

- $E(x) = \lambda$; $MGF = e^{\lambda(e^t - 1)}$ $\Rightarrow$ Variance $= \sigma^2 = \lambda$

$$MGF \equiv E(e^{tX}) = \sum_{i=0}^{\infty} e^{ti} \cdot \frac{\lambda^i}{i!} e^{-\lambda} = e^{-\lambda} \sum_{i=0}^{\infty} e^{ti} \cdot \frac{\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!} = e^{-\lambda} e^{\lambda e^t}$$

discrete dist..

– Mode derivation –

- For large Values of λ; $\frac{X - \lambda}{\sqrt{\lambda}} \sim N(0,1)$ $\Rightarrow$ Shown Easily by MGF

- X, Y are variables with λ_1, λ_2; $Z = X + Y$ is also poisson with $\lambda_1 + \lambda_2$.

- Used to model rare occurances, "Law of Small numbers"

* Thinning of a Poisson Random Variable:-

- If $X \sim Poisson(\lambda)$ and $P(Y|x=i) = Binomial(i, p)$; then $Y = Poisson(\lambda p)$

Lecture 25

Sunday, September 13, 2020 1:20 PM

- Let λ = Avg. Successes per unit time, for poisson.

  u = time until first success.

- Work out Distribution of $P(u)$:-

  $P(u) = Poiss(\lambda u > 0) = 1 - Poiss(\lambda u = 0)$

  $\Rightarrow P(u) = 1 - e^{-\lambda u} \rightarrow$ CDF

  $\Rightarrow$ PDF $\equiv \lambda e^{\lambda u} = f_x(u)$

* Exponential Distribution :- Continuous Distr, non-negative Values only!

  - $f_x(t) = \lambda e^{-\lambda t} \Rightarrow$ only one parameter

  - MGF $\Rightarrow \phi(t) = \lambda/(\lambda - t)$; $\mu = 1/\lambda$ and Variance $= 1/\lambda^2$.

  $E(e^{tx}) = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t}$

  - Mode $\Rightarrow x = 0$; Median $\Rightarrow x = \frac{\ln 2}{\lambda}$;

  - "Memoryless" in Nature :- $P(x > u+s \mid x > u) = P(x > s)$

    ** Not ind.

  - $x_1, x_2, \ldots, x_n$ be Exponential; - $\min(x_1, \ldots, x_n)$ is also Exp. distribution.

Lecture-26

Friday, September 18, 2020 4:35 AM

- We know that X is of some family, now we're trying to get its parameters.

- Let X be a R.V with pdf $f_X(x;\theta)$ where θ is parameter(s)

  * Likelihood :- If X takes x_1; we say $f_X(x_1;\theta)$ is Likelihood

  * Joint Likelihood :- Repeat exp 'n' times $\Rightarrow X_1, \ldots, X_n$ are i.i.d. → independant is compulsory!
  i.d nah...

  Let X_i take $x_i \Rightarrow$ Joint pdf :- $f(x_1, x_2, \ldots, x_n; \theta) \Rightarrow$ Joint Likelihood.

- This is a function in θ; find $\hat{\theta}$ for which is max. ⇒ Maximum Likelihood Estimate

  "often easier to calculate for $\log(f)$ than f."

* ML for Bernoulli :-

$X_i = 1$ with p; otherwise 0.

$$\Rightarrow \boxed{f(x_1, x_2, \ldots, x_n; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}}$$

$$\log f = \sum_{i=1}^{n} x_i \log p + (1-x_i)\log(1-p) \Rightarrow \sum_{i=1}^{n} \frac{x_i}{p} + \left(\frac{1-x_i}{p-1}\right) = \frac{\cancel{p x_i} - x_i + p - \cancel{p x_i}}{p(p-1)}$$

$$\Rightarrow \sum_{i=1}^{n} \frac{p-x_i}{p(p-1)} = \frac{np - n\bar{x}}{p(p-1)} = 0 \Rightarrow \boxed{\hat{p} = \bar{x}}$$

② ML for Poisson :-

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \Rightarrow f(x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

$$\Rightarrow \log f = \sum_{i=1}^{n} x_i \log\lambda - \lambda - \log(x_i!) \Rightarrow \sum \frac{x_i}{\lambda} - 1 = 0$$

$$\boxed{\hat{\lambda} = \frac{\sum x_i}{n}}$$

③ ML for gaussian :-

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow \log(f) = \sum_{i=1}^{n} -\frac{(x-\mu)^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) \Rightarrow -n\log(\sigma\sqrt{2\pi}) - \sum \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \log f}{\partial \mu} \Rightarrow \sum \frac{\cancel{2}(x_i-\mu)}{\cancel{2}\sigma^{\cancel{2}}} = 0 \Rightarrow \sum x_i = n\mu \Rightarrow \boxed{\hat{\mu} = \sum x_i / n}$$

$$\frac{\partial \log f}{\partial \sigma} \Rightarrow -\frac{n}{\sigma} + \sum \frac{(x_i-\mu)^2}{\cancel{2}} \cdot \frac{\cancel{2}}{\sigma^3} = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum (x_i-\mu)^2}{n} \Rightarrow \boxed{\hat{\sigma}^2 = \frac{\sum (x_i-\hat{\mu})^2}{n}}$$

this $\mu \neq \sum x_i/n$!

but we replace with $\hat{\mu}$ if we don't know μ to get $\hat{\sigma}$

Lecture-27

Friday, September 18, 2020 5:15 AM

④ ML for Uniform :-

***

$$f_X(x) = \begin{cases} \frac{1}{k} & \text{when } x \in [0,k] \\ 0 & \text{otherwise} \end{cases} \Rightarrow f(x_1, \dots, x_n) = \begin{cases} \frac{1}{k^n} & \text{if } \forall i,\ x_i \in [0,k] \\ 0 & \text{otherwise.} \end{cases}$$

To maximize f; we need smallest 'k' s.t holds.

$$\Rightarrow \boxed{\hat{k} = \max\{x_1, \dots, x_n\}}$$

⑤ Linear Regression formula :-

Let $\boxed{y_i = mx_i + c + \epsilon_i}$ where x_i – accurately known

y_i – noisy with ϵ_i

$\hookrightarrow \epsilon = N(0, \sigma^2)$

$\Rightarrow y_i - (mx_i + c) \sim N(0, \sigma^2) \Rightarrow y_i = N(mx_i + c, \sigma^2)$ → All y_i are indep. but not indentically distr.!

$$\boxed{\hat{m} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \text{ and } \hat{c} = \sum_{i=1}^{n} \frac{y_i}{n} - \hat{m} \sum_{i=1}^{n} \frac{x_i}{n} \Rightarrow \boxed{\hat{c} = \bar{y} - m\bar{x}}$$

$$\Rightarrow Y_i = N(mx_i + c, \sigma^2)$$

$$f_{Y_i}(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i - mx_i - c)^2}{2\sigma^2}} \Rightarrow f(y_1, \dots, y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i - mx_i - c)^2}{2\sigma^2}}$$

$$\rightarrow \log f = \sum_{i=1}^{n} - \frac{(y_i - mx_i - c)^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) = 0$$

$$\Rightarrow \log f = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{2\sigma^2} \Rightarrow \frac{\partial f}{\partial m} = 0 + \sum_{i=1}^{n} \frac{\cancel{-2}(y_i - mx_i - c)(x_i)}{\cancel{2\sigma^2}}$$

$$\rightarrow c\sum x_i + m\sum x_i^2 - \sum y_i x_i \text{ ——— } (1)$$

$$\Rightarrow c\sum x_i + m\sum x_i^2 = \sum y_i x_i \quad \text{———} \quad (1)$$

$\frac{\partial l}{\partial c}$ ⇒ Get 2^{nd} Eqn ⇒ Solve both!

Lecture-28

Friday, September 18, 2020 11:22 AM

- Let $X_1, \ldots, X_n$ be n i.i.d variables, with parameter θ.
  - ↳ Let θ' be a Estimate of θ. How to know how good θ' is?

* <u>Mean Squared Error</u> :- $\boxed{E[(\theta-\theta')^2] \text{ should be less.}}$ "MSE"

- Firstly, notice that θ' is a R.V because it depends on X_i

- <u>Bias of Estimator</u> :- "b"

  θ' is biased if $E(\theta') \neq \theta \Rightarrow \boxed{\text{Bias} = E(\theta') - \theta}$

* <u>Variance</u> :- $\sigma^2 = \text{Var}(\theta') = E\left[(\theta' - E(\theta'))^2\right] = E(\theta'^2) - E(\theta')^2$

- If θ' varies wildly with $X_i \Rightarrow \sigma^2 \uparrow$

$$MSE = \sigma^2 + b^2 \Rightarrow \underset{MSE}{E[(\theta-\theta')^2]} = \underset{\sigma^2}{E[\theta'^2 - E(\theta')^2]} + \underset{b^2}{(E(\theta') - E(\theta))^2}$$

— Calculating Bias & Variance —

(1) <u>ML for μ, σ^2 of Gaussian</u>

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

$b(\mu) = E(\hat{\mu}) - E(\mu) = 0$ — <u>Unbiased</u>!

Silly mistake --; don't do $\text{Var}(ai) = a\,\text{Var}(i)$; its $\boxed{a^2}\,\text{Var}(i)$

$$\text{Var}(\hat{\mu}) = \text{Var}\left[\sum\left(\frac{x_i}{n}\right)\right] = \sum \text{Var}\left(\frac{x_i}{n}\right) = \sum\left(\frac{\sigma}{n}\right)^2 = \sigma^2/n$$

→ Imp! Possible only because i.i.d.

$E(\hat{\sigma}^2) \Rightarrow E\left(\frac{1}{n}\sum(x_i - \hat{\mu})^2\right)$ ⟶ if we replace $\hat{\mu}$ with μ; $E(\hat{\sigma}^2) = E(\sigma^2)$ → Unbiased!

$\Rightarrow$ Not replacing :- $\frac{1}{n}\sum E\left((x_i - \hat{\mu})^2\right) = \frac{1}{n}\sum E\left(x_i^2 + \hat{\mu}^2 - 2x_i\hat{\mu}\right)$

$$= \frac{1}{n}\sum E(x_i^2) + \frac{1}{n}\sum E(\hat{\mu}^2) - \frac{2}{n} E\left(\hat{\mu}\sum x_i\right)$$

$$= \frac{1}{n}\sum\left[\sigma^2+\mu^2\right] + E(\hat{\mu}^2) - 2E(\hat{\mu}^2)$$

$$= (\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right) = \frac{n-1}{n}\sigma^2$$

indep.

$$Var(\hat{\sigma}^2) \Rightarrow Var\left[\frac{1}{n}\sum(x_i-\hat{\mu})^2\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var\left[(x_i-\hat{\mu})^2\right] \longrightarrow \text{if } \hat{\mu}=\mu;\ \text{then } Var(\hat{\sigma}^2) = \sigma^2/n^2$$

(2) Uniform :-

Let $d_1(\theta) = \frac{2}{n}\sum X_i$ (not ML!)

$$E(d_1(\theta)) = E\left[\frac{2}{n}\sum X_i\right] = \frac{2}{n}E\left[\sum X_i\right] = \frac{2}{n}\cdot\frac{\theta n}{2} = \theta \longrightarrow d_1(\theta) \text{ is unbiased!}$$

$$Var(d_1(\theta)) = Var\left[\frac{2}{n}\sum x_i\right] = \frac{4}{n^2}Var\left[\sum x_i\right] = \frac{4}{n^2}\sum Var(x_i)$$

$$= \frac{4}{n^2}\sum\left(E(x_i^2) - \frac{\theta^2}{4}\right)$$

$$= \frac{4}{n^2}\sum\left[\frac{\theta^2}{3} - \frac{\theta^2}{4}\right] = \frac{4}{n}\cdot\frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

$MSE = b^2 + \sigma^2 = \theta^2/3n \longrightarrow$ for d_1; unbiased.

* For d_2 :-

$X_1, \dots, X_n \Rightarrow$ We know $\theta_2 = \max\{x_i\}$ $\Rightarrow$ Lets look at dist. of θ_2

$P(\theta_2 \le x) \Rightarrow P(\max\{x_i\} < x) = (x/\theta)^n$ when $x \le \theta$. because x_i - indep.

$$\Rightarrow f(x;\theta) = \frac{nx^{n-1}}{\theta^n} \Rightarrow E(\theta_2) = \int_0^\theta \theta_2\cdot\frac{n\theta_2^{n-1}}{\theta^n}\,d\theta_2 = \frac{n}{n+1}\theta \Rightarrow \text{Biased!}$$

$Var(x)$ where x follows θ_2 :- $Var(x) = E(x^2) - E^2(x)$

$$E(x^2) = \int_0^\theta x^2\cdot\frac{nx^{n-1}}{\theta^n}\,dx = \int_0^\theta \frac{nx^{n+1}}{\theta^n}\,dx = \frac{n}{n+2}\theta^2 \Rightarrow Var(x) = \left[\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right]\theta^2$$

* Estimator Consistency :-

– for any $\epsilon > 0$;- $\lim_{n\to\infty} P\left[|\hat{\theta}-\theta| > \epsilon\right] = 0$

where θ = parameter and $\hat{\theta}$ is estimator.

* "Consistency" and "unbiased" are two terms with different meanings.

** MLE is consistent as long as parameter doesn't depend on 'n'. **

** No consistent estimator has MSE lower than MLE **

Lecture-28

Tuesday, September 22, 2020 3:52 PM

* Confidence Intervals :-

- When we find values of ML Estimate, we'd like it to be near the actual value of the parameter.
- We construct an interval around estimate ($\hat{\mu}$) and show that μ lies in this interval with high probability.

∘ For Gaussian :- By CLT $\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \to N(0,1)$ where X_i are indep.

σ ← known!

$$\Rightarrow \sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \in [-2.5, 2.5] \longrightarrow P = 99\%$$

Solve to get

Two Side 99% Confidence Interval $\mu \in \left[\bar{X} - \frac{2.5\sigma}{\sqrt{n}}, \bar{X} + \frac{2.5\sigma}{\sqrt{n}}\right] \to P = 99\%$

* If X, Y are independent $\Rightarrow X+Y=Z$ has pdf :-

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx$$

→ if X_i gaussian, above holds perfectly

otherwise, it holds approx

- Now Lets Look at intervals for S^2 :-

We already know $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

define some k :- such that $P\left(\frac{n-1}{\sigma^2} S^2 \geq k\right) = \frac{\alpha}{2}$

Lets rep. k as $\chi^2_{\frac{\alpha}{2}, n-1}$

$$\rightarrow \underset{A}{P\left(\frac{n-1}{\sigma^2} S^2 \geq \chi^2_{\frac{\alpha}{2}, n-1}\right) = \frac{\alpha}{2}} \quad \text{and} \quad \underset{B}{P\left(\frac{n-1}{\sigma^2} S^2 \geq \chi^2_{1-\frac{\alpha}{2}, n-1}\right) = 1-\frac{\alpha}{2}}$$

$$P\left(\sigma^2 \le \frac{n-1}{\chi^2_{\frac{\alpha}{2}}} S^2\right) = \frac{\alpha}{2} \quad \text{and} \quad P\left(\sigma^2 \le \frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}} S^2\right) = 1-\frac{\alpha}{2}$$

Now, let $\chi^2_{\frac{\alpha}{2}} > \chi^2_{1-\frac{\alpha}{2}} \Rightarrow \frac{n-1}{\chi^2_{\frac{\alpha}{2}}} S^2 < \frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}} S^2$

$\Rightarrow$ Subtract to get :- $P\left(\frac{n-1}{\chi^2_{\frac{\alpha}{2}}} S^2 \le \sigma^2 \le \frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}} S^2\right) = 1-\alpha$

Lecture-29

Thursday, September 24, 2020 10:29 AM

Non-parametric density estimation

- Take $\{X_i\}_{i=1\to n} \sim p(x)$ are i.i.d

Assump For simplicity, assume $X_i \in [0,1]$ and $|p'(x)| \le K$ Smooth bounded

- Consider a Histogram with M bins;
  - Any $x \in B_L$; the density estimate is $\hat{P}_n(x) = \dfrac{\# \text{ of obsv in } B_L}{n \cdot \text{Bin width}}$ for getting $\int_0^1 \hat{P}(x)dx = 1$

$$\Rightarrow \hat{P}(x) = \frac{\sum_{i=1}^{n} I(x_i \in B_L)}{n} \cdot M \qquad I = 1 \text{ or } 0$$

Estimate

* Now, lets find b, σ^2, MSE of this estimate :-

(1) $E[\hat{P}(x)] = \frac{M}{n}\sum_{i=1}^{n} E[I(x_i \in B_L)] = \frac{M}{n}\sum_{i=1}^{n}[1 \cdot P(x_i \in B_L) + 0 \cdot P(x_i \notin B_L)]$

$E[\hat{P}(x)] = M \cdot P(x \in B_L)$ where P is obtained using true pdf.

$= M \cdot [F(l/M) - F(\frac{l-1}{M})]$ where F is true cdf.

But we don't know how this compares with true pdf.

Rewrite RHS as:- $\dfrac{F(l/M) - F(\frac{l-1}{M})}{\frac{l}{M} - \frac{l-1}{M}}$ $\Rightarrow$ by LMVT this equals $p(x^*)$

where p = true pdf and $x^* \in B_L$

$\therefore$ Proved that $E[\hat{P}(x)] = p(x^*) \Rightarrow \text{bias} = p(x^*) - p(x)$

Again, by LMVT:- $p(x^*) - p(x) = p'(x_1)(x^* - x)$

Again, by LMVT:- $p(x^*) - p(x) = p'(x_1)(x^* - x)$

$$\leq \underline{\underline{K/M}} \Rightarrow \text{Bias is bounded!}$$

(2) $\mathrm{Var}[\hat{p}_n(x)] = \frac{M^2}{N^2} \mathrm{Var}\left[\sum I(x_i \in B_L)\right]$

$$= \frac{M^2}{N^2} \sum \mathrm{Var}[I(x_i \in B_L)] = \frac{M^2}{N} P(x_i \in B_L) \cdot [1 - P(x_i \in B_L)]$$

I is a bernoulli RV

From above:- $\mathrm{Var} = \frac{1}{N} p(x^*)(M - p(x^*)) = \frac{M}{N} P(x^*) - \frac{P^2(x^*)}{N}$

$$\leq \frac{M}{N} P(x^*) + \frac{P^2(x^*)}{N}$$

Doing this to gett upper bound

* Mathematical Statistics:-

- Application of mathematics to statistics for data analysis and interpretation.

- We would be dealing mostly with continuous RV.

* Transformation of Random Variables -

Let X be an RV with pdf $p(x)$. Let $g(x)$ be a strictly ↑ function

Now define $Y = g(X)$. we wish to find this pdf.

- Principle of Probability mass conservation

In simple terms; $\boxed{P(a \le X \le b) = P(g(a) \le Y \le g(b))}$

$$\Rightarrow \int_a^b p(x)dx = \int_{g(a)}^{g(b)} q(y)dy$$

$$\Rightarrow \int_{g(a)}^{g(b)} p(g^{-1}(y))\left[\frac{d}{dy} g^{-1}(y)\right]dy = \int_{g(a)}^{g(b)} q(y)dy \qquad (\text{put } x = g^{-1}(y))$$

This holds for all intervals ⇒ $\boxed{q(y) = p(g^{-1}(y)) \cdot \left|\frac{d}{dy} g^{-1}(y)\right|}$ → modulus present to take care of ↓ g

- If g wasn't strictly monotonic;- split it into piecewise monotonic and apply conservation of probability mass.

• If our pdfs were multivariate and the transformation function 'g' was multidimensional, derivative is replaced with determinant of the Jacobian matrix.

Multivariate Gaussian :-

Consider a vector RV $X = [x_1, .., x_D]$ of length 'D'.

Definition X has a multivariate joint gaussian pdf if $\exists$ finite set of i.i.d univariate standard normal RVs $W_1, ..., W_n$ $(n \geq D)$ such that each x_d can be represented as

$$x_d = \mu_d + \sum_n A_{nd} W_n \, .$$

Example :- Zero mean + Isotropic/Spherical gaussian

Defined as $\mu = 0$, $A = I_{D \times D}$ $\Rightarrow X = W$ (all are independant!)

$(D=N)$

$$\Rightarrow p(X) = p(W) = \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2}(\sum \omega_i^2)} = \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2}(\omega \times \omega^T)}$$

Definition * Level set — Essentially a contour; $L_c(f) = \{(x_1, .., x_n) \mid f(x_1, .., x_n) = c\}$

- Therefore, the level set of above gaussian is when $\sum \omega_i^2 = k \Rightarrow$ in 3D we get a sphere — hence the name!

- We shall get to the most general case by making our analysis more "wide".

Generalization 1 - 'A' is Non Singular and Diagonal

- A is not singular $\Rightarrow$ no diagonal element is zero.

$X = \mu + AW \Rightarrow x_i = \mu_i + A_{ii}\omega_i \Rightarrow$ As all ω_i are independent with $\mu = 0, \sigma = 1$

x_i are independant with $\mu = \mu_i$, $\sigma^2 = A_{ii}^2$

$$\Rightarrow P(x) = \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{(\pi A_{ii})} \cdot \exp\left[-0.5 \sum \left(\frac{x_i - \mu_i}{A_{ii}}\right)^2\right]$$

This P(x) is a Hyper-Ellipsoid with mean at μ; Axes aligned with cardinal axes.

- In two dimensions, we get an ellipse centered at (μ_1, μ_2); with major/minor axes' lengths being $2A_{11}/2A_{22}$

Generalization 2 - 'A' is non-Singular, $\mu = 0$

$X = AW \Rightarrow X = g(W) \Rightarrow$ transformation of variables!

$g^{-1}(W) = A^{-1}W \Rightarrow$ in 2D, this depends on the magnitude of derivative of g^{-1}.

We measured how g^{-1} scaled the values.

In general, this would depend on the determinant of the Jacobian of g^{-1}.

and in 3D, g^{-1} refers to how volumes are scaled between the "axes",

Reminder Jacobian $\Rightarrow$ $\nabla f = \dfrac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$

In 2D :- $dx \rightarrow dy \Rightarrow$ Linear to Linear

3D :- $dx \cdot dy \rightarrow dx' \cdot dy' \Rightarrow$ Cube to parallelopiped

nD :- $dx_{\dots} \rightarrow dx'_{\dots} \Rightarrow$ Hyper Cube to hyper parallelopiped.

- To calculate the value of $p(X)$, we would need to find $|\nabla A^{-1}|$. However, understand that this is just a scaling factor, between an infinite-simal Hyper-Parallelopiped and an infinitesimal Hypercube.

- Without proof, we state that the volume of a Hyper-parallelopiped is determinant of of the sides of the hyper-parallelopiped.

(prove by gram-Schmidt and rotate to form Hyper-Cube)

- Therefore; $p(X) = p(A^{-1}W) \cdot \text{Scaling} = p(A^{-1}W) \cdot \dfrac{1}{\det A} = \dfrac{1}{(2\pi)^{D/2} |\det A|} \cdot \exp\left(-0.5\, X^T A^{-1^T} A^{-1} X\right)$

- For simplicity, take $C = AA^T \Rightarrow p(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |C|^{1/2}} \exp\left[-0.5\, x^T C^{-1} x\right]$

$|C| = |A|^2$

* We can easily extend this to Singular A, non-zero μ

Let $Y = X + \mu \Rightarrow$ $\boxed{p(Y) = \frac{1}{(2\pi)^{D/2} |\det C|^{1/2}} \exp\left[-0.5\,(Y-\mu)^T C^{-1} (Y-\mu)\right]}$

Lemma: If Y is a multi-variate gaussian; $Z = AY + c$ is also a multi-variate gaussian.

Definition :- Mean of a multi-variate gaussian $X = AW + \mu \Rightarrow \mu$,

Covariance Matrix of $X = AW + \mu$ is given by $C = AA^T$

Properties :- $C = E\left[xx^T\right] - E[x]\, E[x]^T$

C is symmetric (obviously!)

C is positive - Semi Definite matrix.

- C is said to be positive Semidefinite iff $\forall$ column vectors V,

  $V^T C V \geq 0$.

- If $V^T C V > 0 \Rightarrow C$ is positive definite.

- Now that we know the joint pdf of X ; we are interested in its Level Sets.

We first define a few terms.

Definition Orthogonal Matrix

- 'A' when $AA^T = A^TA$ = Identity matrix

- If $|A| = +1$; it is called as a Rotation matrix, models rotation

  $|A| = -1$; called as Reflection matrix, models reflection + rotation. They are also Symmetric.

- Lets find the Level sets for the multivariate gaussian. We start from special cases and build upto general cases.

Case 1 – $\mu = 0$; A is orthogonal

$$\Rightarrow X = AW \Rightarrow p(x) = \frac{1}{(2\pi)^{D/2}} \exp\left(-0.5 x^T x\right) \Rightarrow \text{same as } W!$$

- This is also a "zero mean isotropic multivariate gaussian". The pdf is unchanged, because A can either rotate/reflect p(W). But because p(W) is spherical, it remains unchanged.

Case 2 – $\mu = 0$, A is Square diagonal with +ve entries

$$X = AW \Rightarrow p(x) = \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{|\det A|} \exp\left(-0.5 x^T A^{-2} x\right) \quad \text{from the formula}$$

- Graphically, the value of $x_i = A_{ii} W_i$; meaning each dimension is amplified by a factor of A_{ii}. Therefore, the pdf is Zero mean anisotropic in nature.

- We can extend this case further. Suppose $A = RS$ where S is diagonal and R is orthogonal in nature.

- Without solving, we can see that $p(X)$ is just rotating the $p(x')$ where $X' = SW$ by R !

  Hence, it is still Zero mean Anisotropic in nature.

- However, if $S = kI \rightarrow p(x')$ is circular $\Rightarrow p(x) = p(x')$!

Case 3 :- General case

- We've already stated that $C = AA^T$ is symmetric, and positive semi-definite in nature.

  However, we shall look at the cases where C is positive definite.

  (when C is semi-def; C is not invertible, causing problems)

Recall :- $Av = \lambda v \Rightarrow$ for a column vector v; λ is called the corresponding eigen value.

- This is possible iff A is diagonizable $\Rightarrow$ it is similar to a diagonal matrix

  $\Rightarrow \exists P$ which is invertible and diagonal D

  s.t $P^{-1}AP = D$

- If A is diagonizable, but is invertible, it is then called as a "Defective Matrix".

Theorem :- Every real symmetric matrix is diagonizable by an orthogonal matrix. It has N real Eigen values with N-linearly independant Eigen vectors — Spectral Theorem

– Applies for C; as it is real & symmetric.

– In mathematical terms;-

If C is real Symmetric $\rightarrow \exists V$, $V^TV = VV^T = I$ and $V^TCV =$ Diagonal matrix

also; N-Eigen values, N-LI-Eigen Vectors.

Extending Spectral $\Rightarrow$ If C is a positive definite matrix, all the Eigen values are positive.

- Returning to the original question of finding Level sets;-

$$p(x) = \frac{1}{(2\pi)^{D/2} \; |C|^{1/2}} \exp\left(-0.5 (x-\mu)^T C^{-1} (x-\mu)\right)$$

- From the spectral theorem, $C = V^T D V \Rightarrow C^{-1} = V^T D^{-1} V \rightarrow C^{-1}$ is PD

- Every Level set has $(x-\mu)^T C^{-1} (x-\mu) = \text{Constant} \geq 0$ as C^{-1} is also PD

$$\Rightarrow (x-\mu)^T V^T D^{-1} V (x-\mu) \Rightarrow \left[V(x-\mu)\right]^T D^{-1} \left[V(x-\mu)\right] = a \geq 0$$

V is orthogonal $\Rightarrow V(x-\mu) = x' - \mu'$ by changing the axes

$$\Rightarrow (x'-\mu')^T D^{-1} (x'-\mu') = a$$

- The center is at μ' in the new rotated system.

In the new system, the half-lengths are root of diagonal Elements of D^{-1}

Define A as diagonal square with $A_{ii} = (D_{ii})^{1/2}$ and write pdf to get this pdf again!

Marginal PDF for Multivariate gaussian

- The marginal pdf along any dimension shall be univariate gaussian in nature. This can be seen from the definition of $X = AW + \mu$.
- More generally, we can also say that the multivariate pdf of a subset of random variables of the gaussian, would also be gaussian!

– However, marginal pdfs having gaussian distribution dont imply joint pdf is gaussian!

* Conditional pdf for multivariate gaussian

- Defined similarly as before; $P(x_1|x_2=x) = \dfrac{P(x_1, x_2=x)}{P(x_2=x)}$. The condition could even be $P(x_1, x_2;\ Ax_1 + Bx_2 = C)$
- The conditional probability is gaussian as well!

* ML Estimation for multivariate gaussian

- The method for calculating $\hat{\mu}$ and $\hat{C}$ remains the same.
- Upon calulating, the value of $\hat{\mu}$ comes out to be the sample mean.

You'll have to use the formula – $\frac{\partial}{\partial \mu} (x-\mu)^T C^{-1} (x-\mu) = 2C^{-1}(x-\mu)$

- Co-variance matrix can also be found using:- $\frac{\partial}{\partial C} (x-\mu)^T C^{-1} (x-\mu) = -C^{-T}(x-\mu)(x-\mu)^T C^{-T}$

$$\frac{\partial}{\partial C} \log|C| = C^{-T}$$

* Mahalanobis Distance :-

Definition $d(y,\mu;C)^2 = (y-\mu)^T C^{-1}(y-\mu) \Rightarrow$ M.d of y from μ. (Exponent part of pdf!)

- It defines Euclidean distance in a multidimensional space. When C is identity, 'd' reduces to Euclidean distance.

Property :- Mahalanobis distance is a $\boxed{\text{true distance metric.}}$

Implies 1) Identity of indiscernibles $\Rightarrow d(x,y)=0 \rightarrow x=y$

2) Symmetry $\Rightarrow d(x,y) = d(y,x)$

3) Triangle inequality $\Rightarrow d(x,y) + d(y,z) \geq d(x,z)$

* Application - Decision boundaries

- We know that all points with the same Mahalanobis distance correspond to a level set.
- Given two pdfs - $P_1(x;\mu_1,C_1)$ and $P_2(x;\mu_2,C_2)$, we wish to find the nature of the curve $P_1(x;\mu_1,C_1) = P_2(x;\mu_2,C_2) = k$

$$\Rightarrow \log\left(\frac{P_1}{P_2}\right) = 0$$

Substituting the value of P_1, P_2 :- $(x-\mu_1)^T C_1^{-1}(x-\mu_1) - \log|C_1| = (x-\mu_2)^T C_2^{-1}(x-\mu_2) - \log|C_2|$

$$\Rightarrow \underbrace{(x-\mu_1)^T C_1^{-1}(x-\mu_1) - (x-\mu_2)^T C_2^{-1}(x-\mu_2)}_{\text{Quadratic}} = \underbrace{\log\frac{|C_1|}{|C_2|}}_{\text{Constant}}$$

- In 2D, this is similar to the $ax^2+2hxy+by^2+\ldots$ of conic Section.
  - This corresponding equation is reffered as "HyperQuadratic Equation".
- When $C_1 = C_2$, the constant terms cancel out, yielding a "Hyper Plane".

(plane, but in multi dim.)

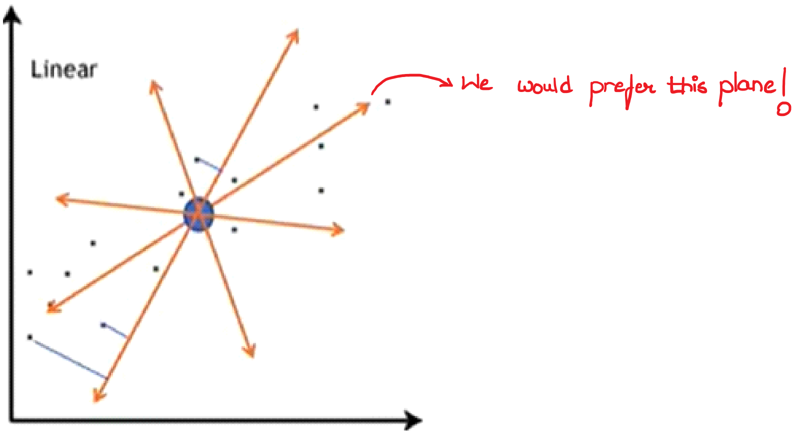
* Principal Component Analysis :-

- Set of vectors are used to depict variation of data, around the mean.

Method :-

- Consider a multivariate variable X, with pdf P(X); mean μ, co-variance matrix C.

– Find a vector v, which along with μ defines a 1D line. The vector should be such that the variance of the projected data set is the maximum.



Mathematical Analysis

Number of data sets = N, Co-variance matrix = C

Mean = μ $\Rightarrow$ WLOG put $\mu = 0$ by origin shifting

Reqd. to maximize — $\sum \frac{\langle x_i, v\rangle^2}{N}$, $\|v\| = 1$ (v is unit vector)

$$\Rightarrow \sum \frac{(x_i^T v)^2}{N}, \quad \|v\| = 1 \qquad \langle a, b\rangle = a\cdot b = a^T b \text{ — Linear Algebra}$$

$$\Rightarrow \sum \frac{(x_i^T v)^T (x_i^T v)}{N} = \sum \frac{v^T x_i x_i^T v}{N}$$

$$= \boxed{v^T \sum \frac{x_i x_i^T}{N} v = v^T C v}$$

C = Covariance Matrix

– Therefore, we need to maximize $v^T C v$. Like we've done so many times before, we look at special cases then generalize.

* C is a diagonal matrix :-

- The problem reduces to maximizing $val = \sum_d C_{dd}(v_d)^2$, $\sum v_d^2 = 1$

Let C_{ii} be the greatest element.

$\Rightarrow$ this is maximized when $v_i = 1$
$v_j = 0 \quad j \neq i$

$\Rightarrow$ In this case the vector is in the dimension with largest diagonal Element/ Eigen Value.

- If we wanted another vector u, orthogonal to v, and maximizing variance?

 · Simply, we can see that $val = \frac{\sum \langle x_i, u \rangle}{N} = u^T C u$, $u \perp v$ is to be max

(or Eigen value)

With the same argument; $u_i = 1$ for the second largest diagonal Element!

This is called as the "Second mode of variation".

* Generalizing

 · Let C be a psd matrix. We have already shown that $C = Q \lambda Q^T$ where λ = diagonal

Q = Adjoint

Given vector v, reqd to maximize $v^T C v = v^T Q \lambda Q^T v$

$= (Q^T v)^T \lambda (Q^T v) = u^T \lambda u$ done already!

Therefore, the max-mode direction is given by max diagonal element, in Q-Space!

* What about projecting onto a plane?

A plane is defined by u, v where $\|u\| = \|v\| = 1$ and $\langle u \cdot v \rangle = 0$

Reqd to minimize $u^T C u + v^T C v = \sum C_{ii} (v_i^2 + u_i^2)$ is to be maximized.

(taking C to be diag.)

(Finish this proof Later)

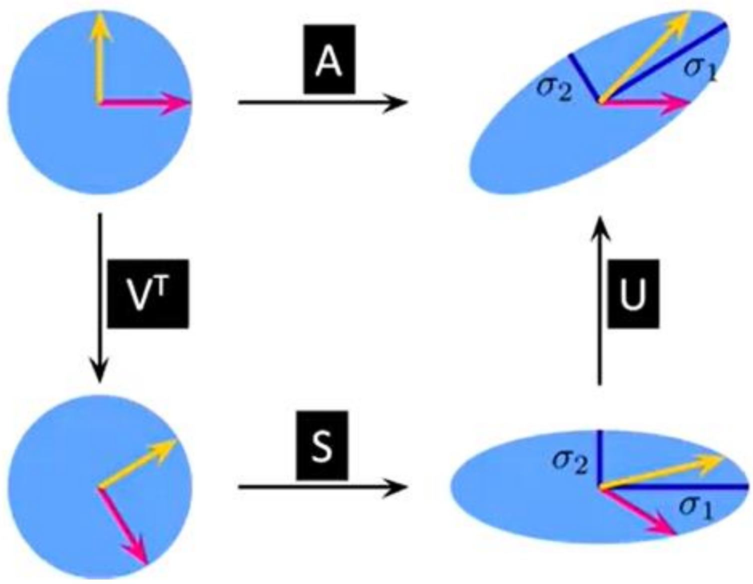
- In summary, if C is the co-variance matrix of a multivariate gaussian, and if we wish to represent it via 'N'-dimensional space;
  - The space is given by the eigen vectors of N-largest Eigen values.
  - The variance in this space is sum of these eigen values.

Singular Value Decomposition

- Let A be an $M \times N$ matrix, if A is real valued, then $A = USV^T$ ← SVD of A
  - V is orthogonal of size $N \times N$ ($V = I$ if A is complex)
  - U is orthogonal of size $M \times M$ ($U = I$ if A is complex)
  - S is a diagonal of $M \times N$:- diagonal values → Singular values - always non-negative Real
  - If $M \le N$; we can write $A = \sum_{m=1}^{M} s_m u_m v_m^T$

    s_m - diagonal element ; u_m, v_m - m^{th} column in U, V

- For $X = AW$, with SVD of A being USV^T; the effect of A on W can be understood as shown below.



* Matrix Norm :-

- Let 2-norm of a vector x be :- $\|x\|_2$

  For a matrix of size $M \times N$, the norm is given by $\|A\|_2 = \max_{x \neq 0} \left(\frac{\|Ax\|_2}{\|x\|_2} \right) \ge 0$

- Geometrically, we can take $\|x\|_2 = 1$ wlog. By SVD, we can think that A would convert the "circle" $\|x\|_2 = 1$ to an ellipse. The max distance of a point on this ellipse from orign is the value of $\|A\|_2$.

- $A = USV^T$, Let i^{th} column be given by u_i, v_i respectively.

  It can be seen clearly that for every i; Av_i is in the direction of u_i, scaled by S_{ii}

  $\Rightarrow$ The right singular vectors can be converted into left singular vectors.

$$Av_i = S_{ii} \cdot u_i$$

Bayesian Statistics

Definition — Bayes theorem (discrete)

- X – discrete RV, Y – discrete/cont. RV (modeling observed data)
  - Likelihood $\Rightarrow P(Y=y \mid X=x)$
  - Evidence $\Rightarrow P(Y=y) = \sum_x P(X=x, Y=y)$ the data from obsv
  - Prior $\Rightarrow P(X=x)$ bfr obsv
  - Posterior $\Rightarrow P(X=x \mid Y=y)$ after observation.
- Here, Y is known from experiments. We try to model X from Y.

- Notice that Posterior = (Likelihood) · (Prior) / $P(Y=y)$. — normalising factor

– In case of a continuous X;
  - Likelihood $\Rightarrow P(Y=y \mid X=x)$
  - Evidence $\Rightarrow P(Y=y) = \int_x P(X=x, Y=y)$ the data from obsv
  - Prior $\Rightarrow P(X=x)\,dx$
  - Posterior $\Rightarrow P(X=x \mid Y=y)\,dx$

- Bayesian Analysis uses prior as "previous knowledge", and is used when the data set is small and finite. Having a good prior knowledge is paramount.

Example — Lets say we have $\{x_i\}_1^N$ drawn from Gaussian with known variance and unknown mean.

Bayesian strategy :- mean M is drawn from a gaussian with μ_0, σ_0^2

$\therefore$ The model is:- draw μ from prior $P(M)$, then data from $P(X \mid M=\mu)$

Maximum A Posteriori Estimate :-

- Prior – $P(M=\mu)$, Likelihood = $P(\text{data} \mid M=\mu) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

upon calculation,

$$\boxed{\hat{\mu} = \frac{\bar{x}\,\sigma_0^2 + \mu_0\,\sigma^2/N}{\sigma_0^2 + \sigma^2/N}}$$

$\rightarrow$ Weighted mean of ML & max. priori est.

i.e. we find the value of μ which maximizes the Posterior prob. distribution

- Because we're maximizing Posterior Likelihood, it is also known as Posterior mode.

- Instead, we can find the mean as follows :-

* "Posterior mean" to minimize expected square error.

i.e., if $\{x_i\}_{i=1}^{N}$, and we have a prior θ;

$$\text{posterior} = \frac{P(x/\theta)\,P(\theta)}{\int_\theta P(x,\theta)\,d\theta} \Rightarrow \text{we wish to minimize } E\left[(\hat{\theta}-\theta)^2\right]$$

- $E\left[(\hat{\theta}-\theta)^2\right]$ is a function of $\hat{\theta}$, minimize to find $\underline{\hat{\theta}^*}$!

$\Rightarrow$ Baye's posterior mean $= E[\theta]$ where $P(\theta) =$ posterior distribution.

Tool for easy calc!

Product of gaussians $\equiv G_1(\mu_1, \sigma_1^2) \cdot G_2(\mu_2, \sigma_2^2) \propto G_3(\mu_3, \sigma_3^2)$

$$\mu_3 = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_2^2 + \sigma_1^2}, \quad \sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

* Loss function

- If $\hat{\theta}$ was estimate of θ, we say that we incur a "Loss" based on Loss function $L(\hat{\theta}/\theta)$

- Note that θ is an RV of the posterior.

- $E\left[L(\hat{\theta}/\theta)\right]$ is called the Risk function, and this is what we wish to minimize.

- Notice that if L was a squared Loss function $\Rightarrow L(\hat{\theta}/\theta) = (\hat{\theta}-\theta)^2$ — Estimate which minimizes risk function would be $\hat{\theta} =$ posterior mean.

(a) Zero One Loss function

$$L(\hat{\theta}/\theta) = I(\hat{\theta} \neq \theta) \Rightarrow \text{Risk} = E\left[I(\theta \neq \hat{\theta})\right] = 1 - P\left(\hat{\theta} = \theta/\text{data}\right)$$

$\rightarrow$ posterior probability given data

- Similarly, for the continuous case, let Loss be 1 if $\hat{\theta} \in \left[\theta - \frac{\epsilon}{2}, \theta + \frac{\epsilon}{2}\right]$

$$\therefore \text{Risk} = 1 - \int_{\hat{\theta}-\epsilon/2}^{\hat{\theta}+\epsilon/2} P(\theta)\,d\theta \Rightarrow \text{we wish to find } \hat{\theta} \text{ for this to be min.}$$

posterior

Put $\epsilon \to 0 \Rightarrow$ If $\hat{\theta}$ is the mode, then Risk is minimized as area under the curve is largest.

$P(\Theta)\big|_{\Theta=\hat{\Theta}}$ is max value.

(3) Absolute Error Loss – $L(\hat{\theta}/\theta) = |\hat{\theta}-\theta|$

$P(\theta)$ = posterior

$$\text{Risk} = E\left[|\hat{\theta}-\theta|\right] = \int_{-\infty}^{\hat{\theta}} (\hat{\theta}-\theta)\, P(\theta)\, d\theta + \int_{\hat{\theta}}^{\infty} (\theta-\hat{\theta})\, P(\theta)\, d\theta$$

Leibnitz :- $\frac{\partial}{\partial a} \int_{L(a)}^{u(a)} f(x,a)\, dx = \int_{L(a)}^{u(a)} \frac{\partial f}{\partial a} dx + f(u(a),a)\frac{\partial u}{\partial a} - f(L(a),a)\frac{\partial L}{\partial a}$ $\Rightarrow$ use this to diff. and find $\hat{\theta}^*$

differentiating, $\int_{-\infty}^{\hat{\theta}} P(\theta)\, d\theta - \int_{\hat{\theta}}^{\infty} P(\theta)\, d\theta = 0 \Rightarrow \int_{-\infty}^{\hat{\theta}} P(\theta)\, d\theta = \int_{\hat{\theta}}^{\infty} P(\theta)\, d\theta$ $\Rightarrow \hat{\theta}^*$ is median of $P(\theta)$

Fisher Information

- Informs about the amount of information is conveyed by given data about an unknown parameter, quantitatively.

Observations

1. It is easier to estimate a parameter θ from given data if the graph of Likelihood $P(\text{data}/\theta)$ versus θ peaks sharply for small change in θ.

   $\Rightarrow |dL/d\theta|$ should be large.

   – Notice that if the prior has a large variance, the reliability of our estimate is reduced. Try to draw 5 points from two gaussians with different variances. If σ is high, data will be all over, making estimates inaccurate.

2. If the likelihood doesn't "peak" properly wrt changes in θ, more data samples are to be drawn.

   (follows from 1)

- Assume that θ_{true} is known. As stated earlier, we might have to repeat the expt few times to get a good estimate of θ_{true}. Let x_i be the value of Likelihood at $\theta = \theta_{true}$ for the i^{th} experiment.

- The expected value of the slope of the log-Likelihood function at $\theta = \theta_{true}$ over all the experiments is 0.

$$\Rightarrow \quad E\left[\frac{d}{d\theta}\left(\log P(\text{data}/\theta_t)\right)\right] = 0$$

(Calculating is easy enough...)

Definition

- Because the expected value is zero, the variance of slope is given as:-

$$\sigma^2 = E\left[\left(\frac{d}{d\theta}\log P(\text{data}/\theta_t)\right)^2\right] = I(\theta_{true}) \rightarrow \text{Fisher information} \geq 0$$

i.e, if the graphs were sharp, then variance would be high as well.

Alternate Defn.

Instead of variance, we could look at second derivative of the log-Likelihood function, as it tells us about the "peak-ness" of the graph.

Turns out, $E\left[\frac{\partial^2}{\partial\theta^2}\log\left(P(\text{data}/\theta_t)\right)\right] = -I(\theta_t)$ -ve sign because the graphs need to be concave - $\frown$

* Cramer Rao Lower bound – applicable for unbiased estimators only.

- Tells us how good a class of estimators can ever be.

  Let $\hat{\theta}$ be an unbiased estimator of θ.

  $$\Rightarrow Var(\hat{\theta}(x)) \geq I(\theta)^{-1}$$

- An unbiased estimator whose variance equals $I(\theta)^{-1}$ is called as Minimum variance unbiased estimator. MVUE

* Bayesian Cramer-Rao Lower bound :-

  Let X be the model of a dataset

  consider $P(x/\theta)$ be likelihood with parameter θ

  Let prior$(\theta) = q(\theta/\alpha)$ where α is a known hyperparameter.

$$E_{q(\theta/\alpha)}\left[E_{P(x/\theta)}\left[(\hat{\theta}-\theta)^2\right]\right] \geq \left(E_{q(\theta/\alpha)}\left[I_P(\theta)\right] + J_q(\theta)\right)^{-1}$$

$$J_q(\theta) = \int_a^b q(\theta/\alpha)\left[\frac{\partial}{\partial\theta}\log q(\theta/\alpha)\right]^2 d\theta \rightarrow \text{Prior information}$$

Expected value of the square of slope for $\log q(\theta/\alpha)$ vs θ graph.

- For this to be valid, a few assumptions are needed :-
  - $q(\theta/\alpha)$ has to be defined in a finite interval (a,b) ; with $q(\theta/\alpha) \rightarrow 0$ as $\theta \rightarrow a$ or $\theta \rightarrow b$.

* Jefferey's Prior

- Analyzes how the prior changes when re-parametrization is done.
- A prior is said to be Jeffery's prior if it is invariant wrt reparametrization.

  that is, the prior should be the same for a new $\beta = f(\theta)$ where f is monotonic.

  $P(\theta) \propto \sqrt{I(\theta)}$ is a Jefferey's prior.

* Conjugate Prior :-

- A prior is said to be conjugate if the posterior and prior both belong to the same family. The prior and posterior are called conjugate pdfs.
- Having a conjugate prior ensures that the denominator is integrable and that it has a closed form expression.

Example :- See from slides tmrw.