Suyash P. Awate:-

CS215: part-2:-

- Topics for the sem:

1) Transformation of random variables
   - univariate
   - multivariate
2) Multivariate statistics
   - Multivariate Gaussian
     - functional form, properties
   - Relationship with Linear algebra
     - singular value decomposition (SVD), of matrices.
     - Eigen decomposition of matrices.
3) Principle Component Analysis (PCA)
   - Relies on covariance of multiple random variables.
     - need not be multivariate gaussian.
   - Special properties when data is multivariate Gaussian.
4) Bayesian Statistics
   - improving over ML estimation.
   - prior models
   - Some concepts from information theory
5) Measuring quality of estimators
   efficency.
   - Fischer information
   - Cramer-Rao lower bound
   - Bayesian Cramer-Rao lower bound.

2 Assignments
1 Quiz
1 endsem.

1) Transformation of random variables:-

$X \rightarrow$ random variable. ; $p(x)$ is the PDF of x.

$g(\cdot)$ is the transformation function!

$$Y = g(X)$$

we need $q(y)$; the PDF of y.

— take $g(x)$ to strictly increasing; for now.

(we will generalize later).

Principle of probability mass conservation:-

$$P(a < X < b) = P(g(a) < Y < g(b))$$

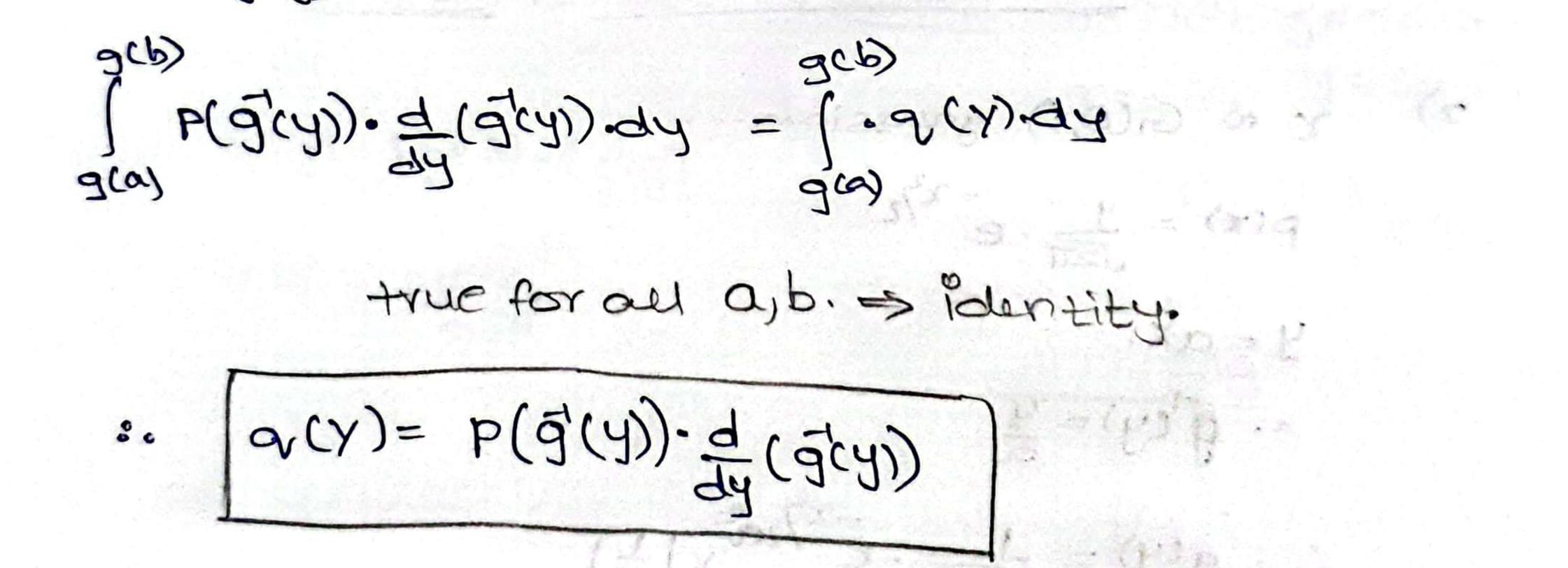
$$\int_a^b p(x)\,dx = \int_{g(a)}^{g(b)} q(y)\,dy$$

$x = g^{-1}(y)$

$$\int_{g(a)}^{g(b)} P(g^{-1}(y)) \cdot \frac{d}{dy}(g^{-1}(y))\,dy = \int_{g(a)}^{g(b)} \cdot\, q(y)\,dy$$

true for all a, b. $\Rightarrow$ identity.

$\therefore$ $$\boxed{q(y) = P(g^{-1}(y)) \cdot \frac{d}{dy}(g^{-1}(y))}$$



if $g(x)$ is strictly decreasing?...

$$q(y) = p(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

→ works for increasing $g(y)$ too!

Examples :-

1) $x \sim U(0,1)$

$y = -\frac{1}{\lambda} \cdot \log(x)$

$q(y) = ?$

$g^{-1}(y) = e^{-\lambda y}$

$\therefore q(y) = p(e^{-\lambda y}) \cdot |-\lambda \cdot e^{-\lambda y}|$

$= \frac{1}{1} \cdot \lambda \cdot e^{-\lambda y}$

$\therefore$ $y \rightarrow$ exponential pdf, $\mu_y = 1/\lambda$

2) $x \sim G(0,1)$ gaussian.....

$p(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}$

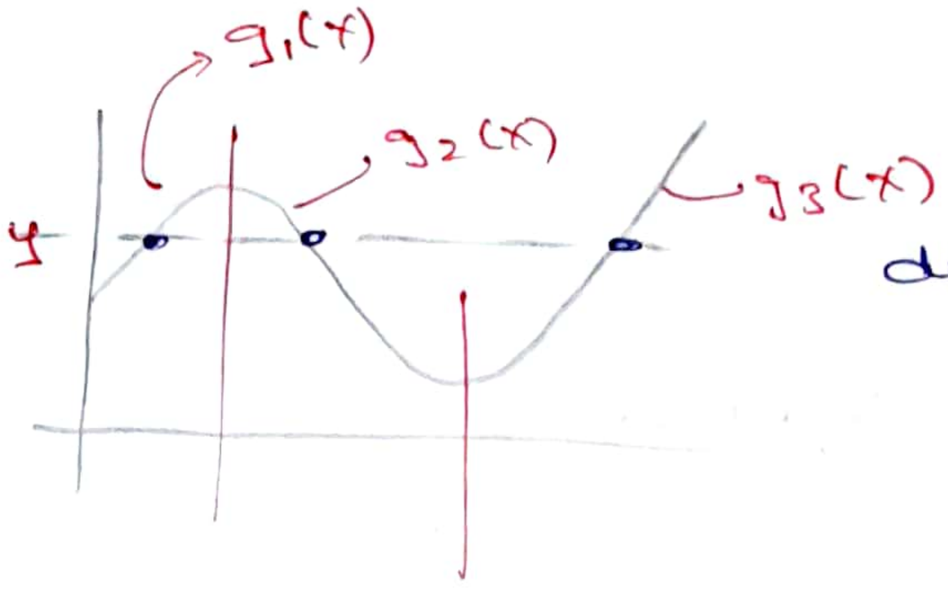
$y = ax.$

$\therefore g^{-1}(y) = \frac{y}{a}.$

$\therefore q(y) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2a^2} \times \left|\frac{1}{a}\right|$

$q(y) = G(0, a^2)$

* General monotonous functions:-



divide g(x) into monotonous functions.

& add their respective $q_i(y)$.

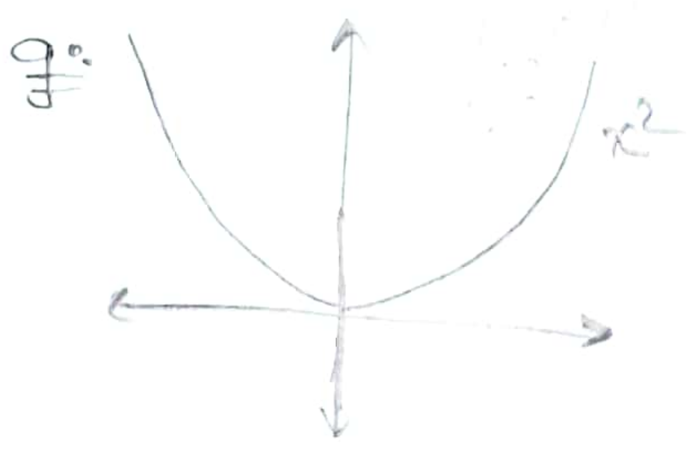
Eg: P(x) is: 0 for $x \le -1$

0.5 for $x \in [-1, 0]$

$0.5e^x$ for $x > 0$.

& $\underline{Y = X^2}$

$q(y) = ?$



case-1: $x > 0$; $y \in (0, \infty)$

$x = \sqrt{y}$

$\therefore q(y) = p(\sqrt{y}) \cdot \left|\frac{1}{2\sqrt{y}}\right|$

$q(y) = \frac{0.5 \cdot e^{-\sqrt{x}}}{2\sqrt{x}}$

case-2: $x \in (-1, 0)$ ~~$y \in \sqrt{-x}$~~ $y \in (0, 1)$

$x = -\sqrt{y}$

$\therefore q(y) = p(-\sqrt{y}) \left|\frac{-1}{2\sqrt{y}}\right|$

$g(x) = \frac{0.5}{2\sqrt{y}}$

$\therefore$ i) for $y \in (0, 1)$

$q(y) = \frac{1}{4\sqrt{y}}(1 + e^{-\sqrt{y}})$

(jump discontinuity).

ii) for $y \ge 1$

$q(y) = \frac{1}{4\sqrt{y}} \cdot e^{-\sqrt{y}}$

Eg 2:

$X \sim G(0,1)$

$$p(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

$\underline{y = x^2}$

what is p(y) (chi-square.)

1) $x > 0;\ y > 0$

$x = \sqrt{y}$.

$$\therefore q(y) = p(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}}$$

$$= \frac{1}{\sqrt{2\pi}} \cdot e^{-y/2} \cdot \frac{1}{2\sqrt{y}}$$

2) $x < 0;\ y > 0$

$x = -\sqrt{y}$

$$q(y) = p(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}}$$

$$= \frac{1}{\sqrt{2\pi}} \cdot e^{-y/2} \cdot \frac{1}{2\sqrt{y}}$$

$\therefore$ for $y > 0$; $\underline{q_{net}(y) = \frac{1}{\sqrt{y}\sqrt{2\pi}} \cdot e^{-y/2}}$

→ Gamma pdf :- (not poission pdf)
discrete pdf.

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \cdot e^{-\beta x} \equiv \text{Gamma}(x \mid \alpha, \beta)$$

(whole $\int_0^\infty$ is 1.)

- α (shape parameter) > 0, $x > 0$
- $\Gamma(\alpha) = \int_0^\infty e^{-x} \cdot x^{\alpha-1}\, dx$; α is any complex number.
- gamma(n) $= (n-1)!$ for int n.

say $Y = 1/X$.

$Y \longrightarrow$ inverse gamma pdf.

then $pdf(Y) = \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot y^{-\alpha-1} \cdot e^{-\beta/y}$

---

you got to do

$$\frac{d}{dy} g^{-1}(y) = \frac{-1}{y^2}$$

not

$$\frac{d}{dy} p(x). = (\text{big})$$

take care;
be present!

classmate
Date
Page

1. Transformation of variables:-

$Y = f(x)$, $P_Y(f(y)) = P_X(\bar{f}^{-1}(y)) \cdot \frac{d}{dy}\bar{f}(y)$

terrific!

conservation of probability mass.

2. Multivariate Statistics:-

→ Multivariate Gaussian:-

* vector random variable X:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_D \end{bmatrix}$$

joint pdf:

- the rand. var. X has a multivariate gaussian PDF if ∃ a finite set of iid univariate standard-normal RVS $w_1, w_2, \dots w_n$ such that $\mathcal{N}(0,1)$ (with $n \geq D$)

each x_d can be expressed as

$$x_d = \mu_d + \sum_{1}^{n} A_{dk} \cdot w_k$$

$$\Rightarrow \boxed{X = \mu + A \cdot W}$$ where W is $\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$

A: non-singular!

rank = D! if $A \neq N$.

only for m.v. gaussian; not any m.v.r.v.

if $n > D$; then we can always reduce to D.

If $n < D$, easy to fill null rows in A to increase n to $n \geq d$.

but $|A| = 0$ then!

* if $\mu \equiv 0$ & $A \equiv \mathbb{I}_{d \times d}$; then $x_d = w_d$

0 mean, isometric/spherical gaussian.

$$\therefore p_x(X = w) = p_W(W = w) = \prod_d p_{w_d}(w_d = w_d)$$ (Joint probability)

$$= \frac{1}{(2\pi)^{D/2}} \cdot e^{(-0.5 \cdot w^T \cdot w)}$$

a column vec.

→ **level sets:-**

locus of points in D-dimension; which have the same $p_x(x)$

↳ x is a datum.

So; a column vector! $x = (x_1, x_2, \dots x_d)$

$L_c(f) = \{(x_1, x_2, \dots x_n) \mid f(x_1, x_2, \dots x_n) = c\}$

* Hence, $\mu = 0$; $A = I_{D \times D}$;

then

level sets are hyper spheres.

∴ isometric/spherical gaussian.

---

* what is pdf; for $\mu \neq 0$ & A = non-singular, ~~diagonal~~ matrix.

Say $X = A \cdot W + \mu$

then we need pdf $(X = x)$.

level sets: Hyper ellipsoid

Step 1) if $X = x$ & $X = AW + \mu$. then $W = ?$

$W = A^{-1}(x - \mu)$

Step 2) Apply law of conservation of probability mass.

* but what will be 'volume' factor? scaling. : jacobian! of A^{-1}!

volume factor scaling $= |A^{-1}|$

$\therefore \text{pdf}_X(x) = \text{pdf}_W(A^{-1}(x-\mu)) \cdot |A^{-1}|$

$\because \begin{bmatrix} x_{new} \\ y_{new} \end{bmatrix} = A^{-1} \cdot \begin{bmatrix} x_{old} \\ y_{old} \end{bmatrix}$

for proof + why volume scaling $= |A^{-1}|$

$y_1 = x_1 a_{11} + x_2 a_{21} + \dots$

$d y_1 = a_{11} dx_1 + a_{12} dx_2 + \dots$

$\Rightarrow$

if $\mu = 0$:-

$\therefore q(x) = p(A^{-1}x) \cdot \frac{1}{|A|}$ don't forget.

$= \frac{1}{(2\pi)^{D/2} \cdot |A|} \cdot \exp\hat{}\left(-0.5 (A^{-1}x)^T \cdot (A^{-1}x)\right)$

$= \frac{1}{(2\pi)^{D/2} \cdot |A|} \cdot \exp\hat{}\left(-0.5\, x^T \cdot (A^{-1})^T (A^{-1}) \cdot x\right)$

$= \frac{1}{(2\pi)^{D/2} \cdot |A|} \cdot \exp\hat{}\left(-0.5\, x^T (A \cdot A^T)^{-1} \cdot x\right)$

let $(C = A \cdot A^T)$ We will see, C is covariance matrix of X.

$\therefore q(x) = \frac{1}{(2\pi)^{D/2} \cdot |C|^{1/2}} \cdot \exp\hat{}\left(-0.5\, x^T C^{-1} x\right)$

---

if $\mu \neq 0$:-

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

we are in mean centered world...

---

* Facts:- A is a square matrix.

- is A transformation! yea...

also A is non-singular ⇒ Gram-Schmidt orthogonalisation is possible.

∴ make $A^{-1} \longrightarrow A^{-1}_{\text{orthogonal columned matrix.}}$

now; rotate $A^{-1}_{\text{orthogonal}}$; to have its vectors along cardinal axis.

$A^{-1} \xrightarrow{A^{-1}_{\text{ortho}}}$ no $\rightarrow$ diagonal matrix. we know factor is $|A|$ now itself.

* if Y is multivariate gaussian;

$Z = B \cdot Y + C$ is also multivariate gaussian.

(coz; $Y = AW + \mu$, $Z = (B \cdot A) W + (C + B\mu)$)

fits into definition of multivariate gaussian

* $X = AW + \mu$.

then $\text{mean}(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_D) \end{bmatrix} = \mu$

• for general multivariate r.v X;

$$E_{P_x}(X) = \begin{bmatrix} E_{P_x}(X_1) \\ E_{P_x}(X_2) \\ \vdots \end{bmatrix} \equiv \begin{bmatrix} E_{P_{X_1}}(X_1) \\ E_{P_{X_2}}(X_2) \\ \vdots \\ E_{P_{X_n}}(X_n) \end{bmatrix}$$

important to note, which pdf, we are in.

→ covariance matrix, for a general vetor R.V. Y:-

$$C = E_{P(Y)}\left[\underset{D\times 1}{(Y - E(Y))^*} \underset{1\times D}{(Y - E(Y))^T} \right]$$

Square matrix $D \times D$.

$$\therefore C_{ij} = E_{P(Y)}\left[(Y_i - E(Y_i))(Y_j - E(Y_j)) \right]$$

$$\equiv E_{P(Y_i, Y_j)}\left[(Y_i - E(Y_i))(Y_j - E(Y_j)) \right]$$

$$\equiv \text{cov}(Y_i, Y_j)$$

* for $X = AW + \mu$; $C(X) = A \cdot A^T$

- For a general vector RV = X; need not be m.v.gaussian.
  let C be covariance matrix of X

1) $C = \underset{\text{sq.mtrx}}{E(X \cdot X^T)} - \underset{D\times 1}{E[X]} \cdot \underset{1\times D}{(E[X])^T} \rightsquigarrow C_{ij} = E(X_i \cdot X_j) - E(X_i) \cdot E(X_j) = \text{cov}(X_i, X_j)$ ✓

   $D \times D$ ✓

2) C is symmetric. (as $C_{ij} = \text{Var}(X_i, X_j) = \text{Var}(X_j, X_i) = C_{ji}$)

3) C is positive semi definite.

⇒ for any non zero $a \in \mathbb{R}^{D\times 1}$; $a^T \cdot C \cdot a \geq 0$

since; $a^T \cdot C a = a^T \cdot E\left((X - E(X)) \cdot (X - E(X))^T \right) a$

$= E\left(a^T (X - E(X)) \cdot (X - E(X))^T a \right) = E\left(([\quad](X))^2 \right) \geq 0$

→ terms:-

1) diagonal matrix.

2) orthogonal matrix.

if $Q \cdot Q^T = Q^T Q = I$.

then Q → orthogonal

$|Q| = \pm 1$.

3) Rotation matrix...

Q is orthogonal; and

$|Q| = 1$ then rotational matrix.

Improperly, can also take

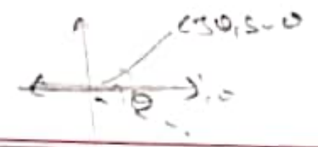
$|Q| = -1$ to be rotational;

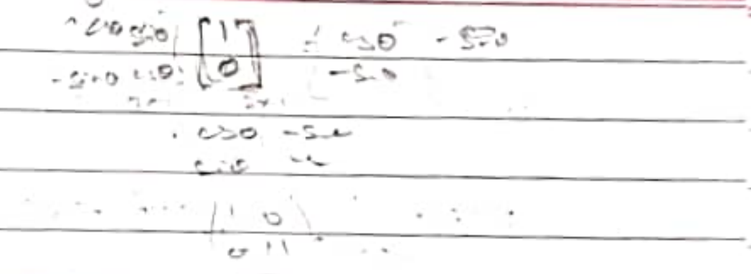
But its rotation + reflection...

4) reflection:-

A symmetric orthogonal matrix is reflection matrix.

$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$





lets discuss level sets :-

- Spectral theorem says;

a real symmetric matrix (a self-adj matrix).

has a orthonormal basis of eigen vectors.

⇒ there exist orthogonal matrix Q;

such that

$C \cdot Q = Q \cdot D$

D → diag. matrix of eigen values.

$\therefore C = QDQ^{-1} = QDQ^T$

* covariance matrix C

is Symmetric & positive definite.

⇒ all its eigen values are real & positive.

there exist orthogonal Q;

such that

$Q^{-1} \cdot C \cdot Q = D$

$\therefore C = Q \cdot D \cdot Q^{-1}$

D is all positive.

rotation matrix — Scaling matrix — for level sets.

Now;

$$P(x) = \frac{1}{(2\pi)^{D/2}|C|^{1/2}} \exp\left[-0.5\,(x-\mu)^T \cdot C^{-1} \cdot (x-\mu)\right]$$

$$= \frac{1}{(2\pi)^{D/2}|C|^{1/2}} \cdot \exp\left(-0.5\,\underbrace{(x-\mu)^T \cdot Q \cdot D^{-1} \cdot Q^T}_{!}(x-\mu)\right)$$

if eigenvalue is more; variance is max; ellipsoid axis's length more.

let $Q^T \cdot (x-\mu) = y$

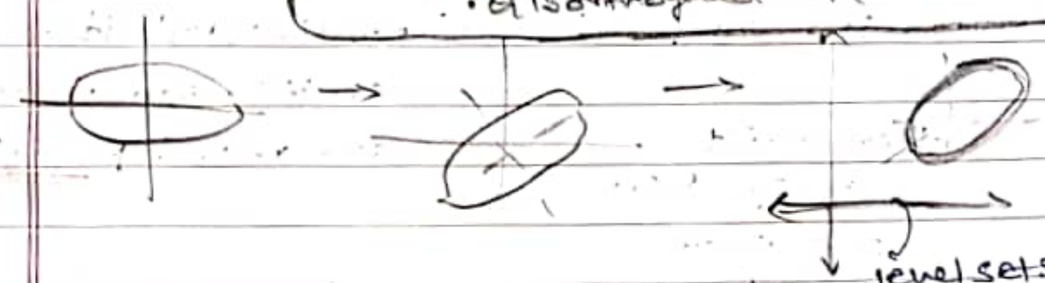
∴ in y's frame of axis;

level sets are hyperellipsoid at origin; with coaligned axis as axis.

now; $x = Q \cdot y + \mu$

↳ translation of hyperellipsoid.

↳ rotation of hyperellipsoid. ∵ Q is orthogonal matrix



* say $n > d$. !

$X = A \cdot W + \mu$.

$D \times 1$ $D \times n$ $n \times 1$ $D \times 1$.

now also; $C = A \cdot A^T$.

also $C = Q \cdot D \cdot Q^T$

$= Q \cdot (\sqrt{D})(\sqrt{D})^T \cdot Q^T$

∴ possibly;

$Q \cdot \sqrt{D} = A$.

no maths, only feel...

A is invertible ⇒ gram-schmidt on the rows

$A(e_{11}, e_{ij}, \ldots)$ ⇒ eigen values

$A \cdot (\,) \cdot Q^T$ = diagonal...

like

$$\text{newA} = \begin{bmatrix} 4 & 0 & 0 & 4 & 0 & 0 \\ 0 & 2 & 0 & 0 & 5 & 6 \\ 0 & 0 & 3 & 0 & 0 & 0 \end{bmatrix}$$

NOW;

$(W_1 + W_4)$ is a new W_{n1}

$(W_2 + W_3 + W_6)$ is new W_{n2}

↓

we can reduce n to D.

!

add^n of $a_1 \cdot N(0,1) + b_1 \cdot N(0,1)$

independent

$\simeq A \cdot N(0,1)$ again gaussian!

so $n \to d$ ✓

* when $n = d$;

if A is not diagonal;

since A is non-singular;

gram-schmit orthogonalisation possible.

& then rotate A ortho, to get diagonal matrix D.

then

$A_{n \times n} \cdot (B)_{n \times n} = D_{n \times n}$ $\quad |B| \neq 0$

& $A = D \cdot B^{-1}$ $\quad |B^{-1}| \neq 0$

∴ $X = D \cdot B^{-1}(W) + \mu$ $\quad B^{-1} \to$ rank $= D$ ✓

Hence, anyways; we can again have single! instead of linear comb^n.

$x_1 = \mu_1 + A_1 W_1$; $x_2 = \mu_2 + A_2 W_2$ ....

→ **Marginal PDFs:-**

meaning; Pdf_{x_1} or .. Pdf_{x_2,x_1} ....;

* the 1D marginal PDF is univariate gaussian.
for any multivariate gaussian X.

* marginal pdf of multivargaussian Y in N dimensions, (joint).
over any chosen subset of variables of size M
is multivariate gaussian pdf.

By definition;

$N \geq D$; not $N = D$.

if $N > D$; then we can "reduce" N to D.

Proof:-

let $X = \underset{N \times 1}{A}{}_{N \times N} W + \mu$ ← invertible.

B = projection matrix. such that
B is M×N
& B·X is the subset.

rank(B) = M
& A is invertible.
⇒ rank(AB) = M

$BX = BAW + B\mu$.

$\underset{M \times 1}{Y} = (BA)W + (B\mu)_{M \times 1}$

rank = M. ✓ hence multivar gaussian!.

∴ B will have each row with only one non-zero entry 1

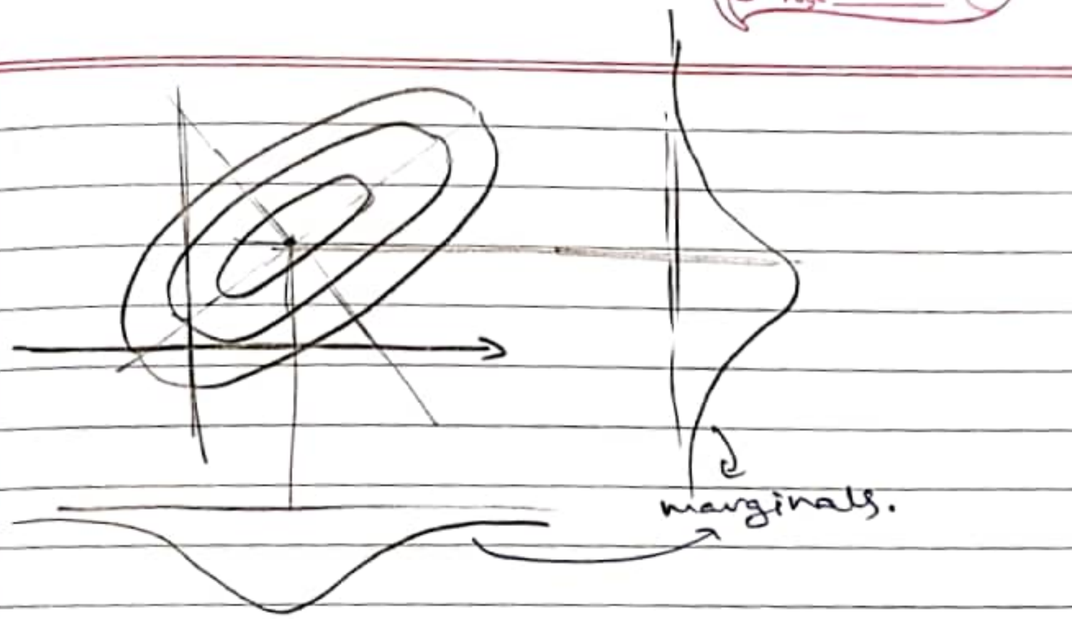
$B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}_{2 \times 4}$ or $\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}_{2 \times 4}$.

* But marginals being gaussian, won't imply joint is gaussian.

if $X_1 = w_1$
$X_2 = w_1$
then; $X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

rank of 2

Hence; X is not gaussian.

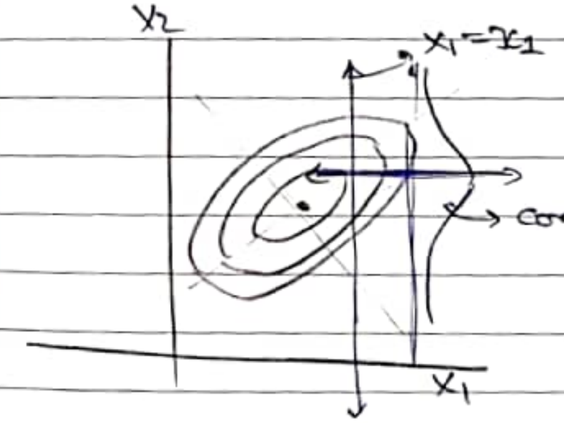


→ conditional pdfs:

let $X = [[x_1] \; [x_2]]$
gaussian multivariate

then $P(X_1 = x_i \mid X_2 = x_2)$ is also a m.v. gaussian.
conditional

$$P(X_1 \mid X_2 = x_2) = \frac{P(X_1, X_2 = x_2)}{P(X_2 = x_2)}$$ a multivariate gaussian
normalisation factor hide.



conditional pdf $(X_2 \mid X_1 = x_1)$
$= \frac{\text{pdf}(X_2, X_1 = x_1)}{P(X_1 = x_1)}$

$\vec{C}$; bcoz we do $P_W(\vec{g}(y))$ · <scaling factor>.

→ Maximum Likelihood estimation:- for general! (NOT for gaussian)

$$p(y) = \frac{1}{(2\pi)^{N/2}\,|C|^{1/2}} \cdot \exp\{-0.5\,(y-\mu)^T \cdot \vec{C}^{-1}(y-\mu)\}$$

$\vec{C}^{-1} = Q \cdot D^{-1} Q^T$

data is $\{y_1, y_2, \dots y_N\}$

these datums are vectors.

1) * MLE for mean vector.... (matrix calculus)

$$LL = \sum (y_i-\mu)^T \cdot \vec{C}^{-1}(y_i-\mu) + \{\text{const wrt } \mu\}$$

$$\frac{\partial LL}{\partial \mu} = 0. \qquad \frac{\partial}{\partial \mu}(x-\mu)^T \cdot \vec{C}^{-1}(x-\mu) = 2\vec{C}^{-1}(x-\mu)$$

(μ: $N\times 1$; $(x-\mu)^T$: $1\times N$; $\vec{C}^{-1}$: $N\times N$; $(x-\mu)$: $N\times 1$; $\vec{C}^{-1}(x-\mu)$: $N\times 1$)

→ $\sum 2\vec{C}^{-1}(y_i-\mu) = 0$

→ $\mu = \frac{\text{Sample}}{\text{mean of } y_i}$

∴ $\hat{\mu}$ = Sample mean

a vector random variable.

Note:-

scalar fcn. $f(a)$ of multiple variables in column vec(a)

Jacobian; $\frac{df}{da}$ will be a row vector of same length as a.

∵ $\underset{1\times 1}{df} = \underset{1\times N}{\frac{df}{da}} * \underset{N\times 1}{da}$ ✓✓

symmetric positive definite.

2) MLE for $\vec{C}$:- = Sample covariance.

biased one! ✓✓/n (not /n-1) ✗

$$LL = \sum_i \frac{(y_i-\mu_i)^T \cdot \vec{C}^{-1} \cdot (y_i-\mu_i)}{2} + \frac{\log|C|}{2} + \text{const} \dots$$

$$\frac{\partial LL}{\partial C} = 0$$

→ partial diff wrt C:- for a scalar fun^n, $f(C)$

1) $C_{N\times N}$ to column vector $a_{N^2\times 1}$

2) find $\frac{\partial f}{\partial a}$ · a row vector $1\times N^2$

3) reshape into $N\times N$.

∴ $df = \frac{\partial f}{\partial C} .^* dC$ (element-wise product.)

$$\frac{d}{dC}(x-\mu)^T \cdot \vec{C}^{-1} \cdot (x-\mu) = -\vec{C}^{-T} \cdot (x-\mu)(x-\mu)^T \cdot \vec{C}^{-T}$$

$$\frac{d}{dC}\log|C| = \frac{1}{|C|} \cdot |C| \cdot \vec{C}^{-T} = \vec{C}^{-T}.$$

∴ we get

$$-\vec{C}^{-T}\left(\sum (x_i-\mu)(x_i-\mu)^T\right) \cdot \vec{C}^{-T} + n \cdot \vec{C}^{-T} = 0$$

taking, μ is known.

sample covariance ∴

$$\underset{N\times N}{\vec{C}} = \frac{\sum \underset{N\times 1}{(x_i-\mu)} \underset{1\times N}{(x_i-\mu)^T}}{N}$$

x_i is a vector with x_{i1} x_{i2} !

→ Mahalanobis distance:- for multivariate gaussian.

the term $(y-\mu)^T \cdot C^{-1} \cdot (y-\mu)$ is mahalanobis distance of y from μ.
(squared).

$$d(y,\mu;C)^2 = (y-\mu)^T \cdot C^{-1} (y-\mu)$$

$C \longrightarrow C^{-1}$

* when $C = I$;

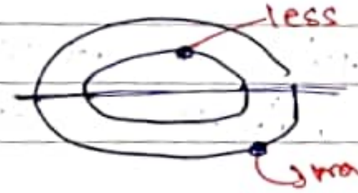
mahalanobis distance == euclidean distance.

or else;

$d(y,\mu;C)$ scales units along each dimension based on standard deviation along that marginal.
not variance.

eigen values = variance. (not σ).

locus is hyper ellipsoid
= level set for distribution.



* Mahalanobius distance; is a true distance metric.

distance metric funct^n $d(\cdot,\cdot) \to R$ satisfies:

(i) identity of indiscriminables:

$$d(x,y) = 0 \iff x = y.$$

(2) symmetric.

$$d(x,y) = d(y,x)$$

(3) triangle inequality.

$$d(x,y) <= d(x,z) + d(z,y)$$

(these three imply non-negativity!)

for $d(x,y;c)$ — (i) ✓ identity of indiscriminables

— (ii) ✓ symmetric.

(iii) let $C = D$. (no cog component).

& see $d(x,y) \leq d(x,z) + d(y,z)$

$$LHS = \sqrt{(u+v)^T \cdot C^{-1} (u+v)} \quad \text{let } x - z = u$$

$$RHS = \sqrt{u^T \cdot C^{-1} \cdot u} + \sqrt{v^T \cdot C^{-1} \cdot v} \quad z - y = v$$

$\sqrt{}$ is needed!!

$$LHS^2 = u^T C^{-1} u + v^T C^{-1} v + u^T C^{-1} v + v^T C^{-1} u.$$

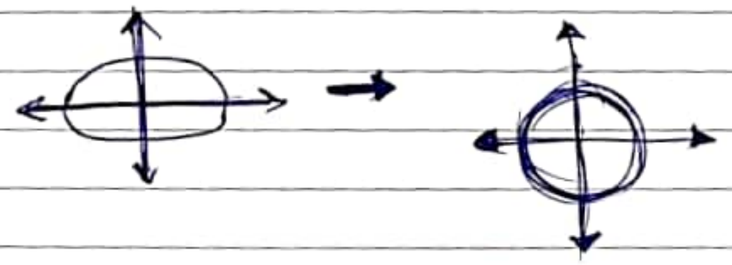
$$RHS^2 = u^T C^{-1} u + v^T C^{-1} v + 2 \cdot \sqrt{u^T C^{-1} u} \cdot \sqrt{v^T C^{-1} v}$$

Now, $C^{-1} = D^{-1}$

$$RHS^2 - LHS^2 = 2 \cdot \sqrt{\sum_d \frac{u_d^2}{\sigma_d^2}} \cdot \sqrt{\sum_d \frac{v_d^2}{\sigma_d^2}} - 2 \cdot \sum_d \frac{u_d v_d}{\sigma_d^2}$$

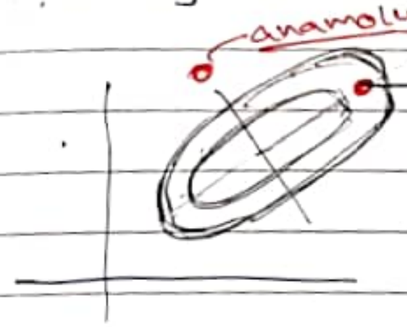
cauchy schwartz inequality.

* Mahalobis scales the coordinate plane



Multivariate Gaussian Applications:—

1) Anamoly detection: whether new datum fits in (or) no.



if mahanalobis distance low → fits in,
otherwise anamoly

2) Maximum Likelihood classification.

A datum given. belongs to which group?
compare the likelihood of it,
for each group individually
& choose max.

- makeobis dist. is not enough.
also see $\frac{1}{|C|^{1/2}}$ term.

decision boundary: $(x-m_1)^T C_1^{-1}(x-m_1) = (x-m_2)^T C_2^{-1}(x-m_2) + \log|C_1| + \log|C_2|$
decision surface.

hyper quadric

if $|C_1| = |C_2|$ & $C_1 = C_2 = C$; only mean shift is diff. then hyper plane decision surface

if $C_1 = C_2 = C$ & $\mu_1 \neq \mu_2$

decision surface is hyper plane.

Principle Component Analysis:-

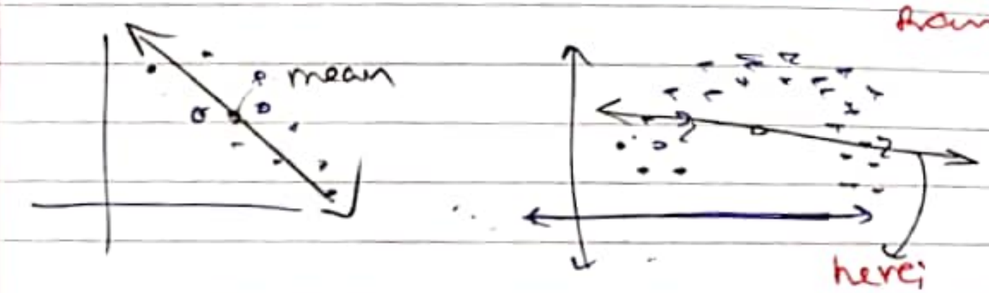
for general vector R.V.
not only gaussian.

- modes of variation:-

  Set of vectors (SO, both mag. & direct^n) that are used to depict the variation in population around the mean.

  ∴ these vectors are drawn from mean.

  "in mean-centered frame".



* let C = diagonal.

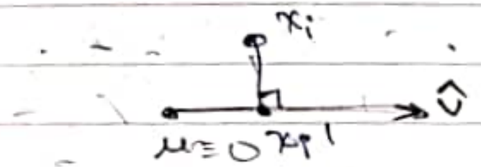
let data $\{x_i\}_{i=1,n}$.

& mean = μ.

let vector = V.

we want the 'direction' which maximizes variance of points incident on it; wrt mean.

"it captures max. dispersion of dataset!!



in N-dimensions.

now; $x_i' = \langle x_i, v \rangle v$; v is unit vector.

∴ var. of all projections

$$= \sum_{i=1}^{N} \frac{\langle x_i, v \rangle^2}{N}$$

$$= \frac{\langle x_i, v \rangle \langle x_i, v \rangle^T}{N}$$

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

$$= v^T \cdot C \cdot v \quad \text{(since C is diagonal)}$$

now; let $V = (V_1\ V_2 \cdots V_n)$ (no need of gaussian)

$\sum V_i^2 = 1$ (unit vector)

$\sum V_i^2 \cdot C_{ii} \rightarrow$ max.

∴ chose $V_i = 1$ for argmax C_{ii}

& $V_i = 0$ for others.

⇒ we choose the cardinal axis with max. eigen value.

(for diagonal C;
coordinate axes are eigen vec;
I is eigen matrix),
(the orthogonal Q),

∴ 1 pca direct^n = eigen vector of max eig.value.

* 2nd pca definition:-
direction, ⊥^{ar} to 1st pca;
& capturing max. dispersion.

then; .....
as predicted;
= eigen vec; with 2nd max eig.val.

* even if $C = Q \cdot D \cdot Q^T$;
$v^T \cdot C^{-1} \cdot v$
$v^T \cdot Q \cdot D^{-1} \cdot Q^T \cdot v$
$(Q^T v)^T \cdot D^{-1} \cdot (Q^T v)$ is max.

*

NOW;
for $C_{n \times n}$;
we have 'n' pcas.
but; if we have
15 observation
& 200 dimension;
as $C_{n \times n}$ is symm;
Only (14) pca can represent whole datum set.
n eigen vectors.
variance of projected points = eigen value (=).

* now; I want the D-dimensioned space; which captures max. dispersion.

≡ linear span of first D-pcas.

& variance is ..... sum of eigen values.
lets do!
$v_1, v_2 \dots v_d$.
$x_1 = \langle x_1, v_1 \rangle v_1 + \langle x_2, v_2 \rangle v_2 + \dots$
x_2
x_3
x_n
dist. from mean
$= \sum \langle x_1, v_1 \rangle^2 + \sum \langle x_2, v_2 \rangle^2 + \dots$
$\sigma_1 + \sigma_2 + \dots$

For gaussian multivariate:-

$X = AW + b$.

then $C = A \cdot A^{-1}$

Eg

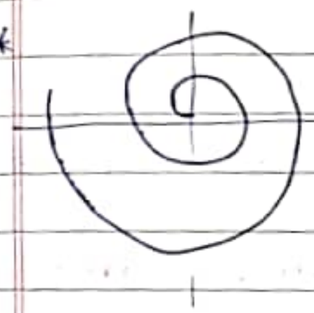
principle modes; are axis of hyperellipsoid. (level set).

4) Dimensionality reduction:-

intrinsic dimension:-

minimum no. of variables (degrees of freedom) required to represent signal.

*



even though every point is $f(x,y)$; it can be written as some function. as some $= (x,y)$

$f(0) = (x(0), y(0))$

∴ degree of freedom = 1.

Intrinsic dim. = 1.

representation dimension = 2.

* let $X = (x_1, x_2, \cdots x_N)$

& consider for $g(\cdot)$ & $M < N$ variables $a_1 a_2 \cdots a_M$.

such that

each $x \in X$ is writable as $g(a_1, a_2, \cdots a_M)$.

then; instead of 19200 variables; a fruit just needed 4 variables.

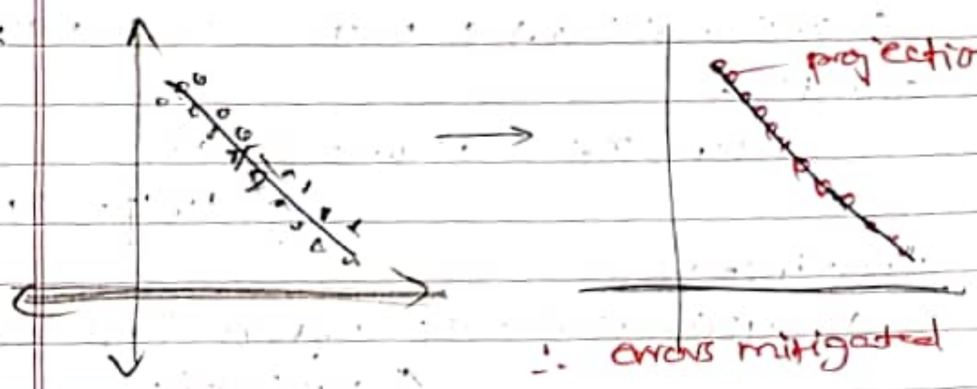
<Heart attacks>.

- Hence; by PCA;
  we find first 4 pca vectors
  &
  project the whole data onto this hyperplane.
  ∴ dimension reduced to!
  19200 → 4

* 1) Acquired data has errors.
  - this can make data seem of higher dims!

- pca performs linear dim. reduction.
  (fits hyperplanes only.)
  - So; if hyperquadric needs 5 dims hyperplane makes it 4.

*

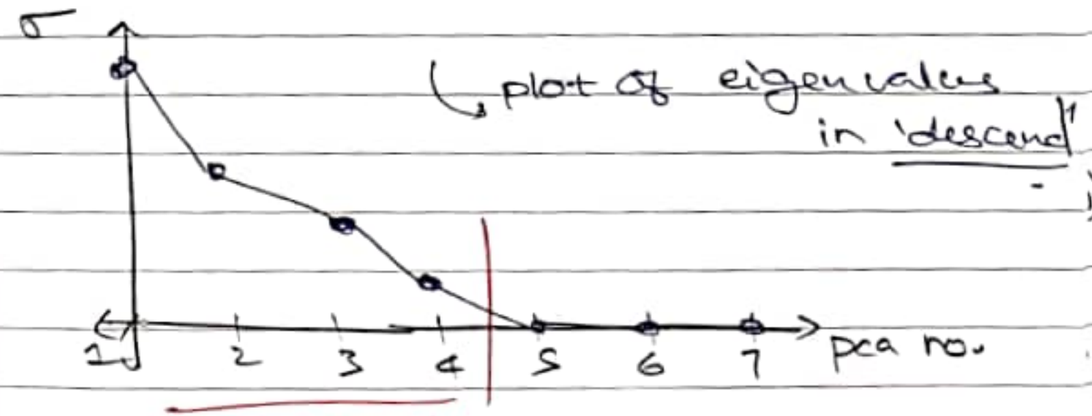


∴ errors mitigated

* How many dimensions should I reduce to?

Scree plot!

1) variance along a mode = eigen values.
2) we need to capture maximum variance of datas.
3) Hence; chose the dimensions; with have considerable eigen values.



∴ 4 pca; can capture most of datas variance.

A → linear operator.

U,S,VT → steps of the operator.

Singular Value Decomposition:-

(SVD)

* matrix factorisation.

- let A be M×N. our interest is M≤N.

But discussed generally.

When A is real valued, SVD of $A = \underset{M\times N}{U}\,\underset{M\times M}{S}\,\underset{M\times N}{V^T}$

V → orthogonal N×N

U → orthogonal M×M

S → rectangular diag matrix. M×N

↳ values of diag = Singular values.

* sing. values are non-negative, real.

Column vectors of U,V → singular vectors.

of U → left singular vectors

of V → right singular vectors.

When M≤N; then

$$A = \sum_{m=1}^{M} S_m \cdot U_m \cdot V_m^T$$

(S_m: scalar; U_m: M×1; V_m^T: 1×N)

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \end{bmatrix}_{3\times 5}$$

reals diagonal

* $A = USV^T$ · then $A \cdot z = U \cdot S \cdot (V^T \cdot z)$

rotation

scaling

rotation.

→ Matrix Norm:-

for $x \in \mathbb{R}^n$; 2-norm is $\|x\|_2$.

for matrix A; of M×N

definition:-

$$\|A\|_2 = \max_{x \neq 0} \left\{ \frac{\|Ax\|_2}{\|x\|_2} \right\} \geq 0$$

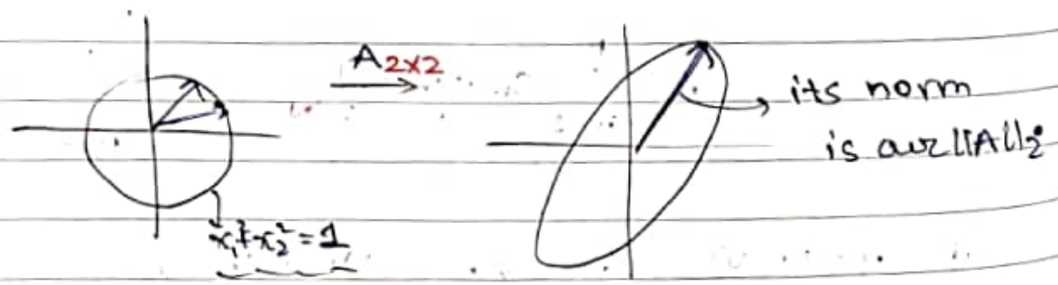
how to do manually?

x is N×1
∴ Ax is M×1

* Geometric interpretation is;
apply linear operator A to all unit vectors 'x' (N×1)

let $y = Ax$; for all such x.

then, pick the longest 'y', & report its 2-norm.



→ Existance of Singular value decomp:-

exists, for any real matrix $A_{M \times N}$.

- let $\sigma_1 = \|A\|_2$: $A_{M \times N}$

  i.e. $\exists V_1 \in \mathbb{R}^n$ such that $\|V_1\| = 1$

  $\& \sigma_1 u_1 = A \cdot V_1$

  $\|u_1\| = \sigma_1 / \sigma_1$

- consider orthogonal matrix U, a basis of $\mathbb{R}^M$ with first column $\frac{u_1}{\|u_1\|}$. not simply u_1.

also orthogonal matrix V; basis for $\mathbb{R}^n$ with first column V_1.

$\|V_1\| = 1$. ✓

then, $U^T \cdot A \cdot V = \begin{bmatrix} \sigma_{1\times 1} & [W^T]_{1\times N-1} \\ [0]_{M-1\times 1} & [B]_{M-1 \times N-1} \end{bmatrix} = S$

- we will show that $W = [0]_{N-1 \times 1}$.

Note: $V_{N\times 1}$ is a column vector; & $A_{N\times N}$ is orthogonal

$$\|V\|_2 = \|A \cdot V\|_2$$

for any V. ∵ orthogonal matrix = rotation.

* $S = U^T A V$

since; $\max_x \|Ax\| = \sigma_1 \; ; \; |x|=1$

$\Rightarrow \max \|A \cdot V \cdot x\| = \sigma_1$

$\|Vx\| = |x| = 1$

$\Rightarrow \max \|U^T A V x\| \equiv \|AVx\| = \sigma_1$

$\Rightarrow \max \|Sx\| = \sigma_1$

$\therefore \|S\|_2 = \|A\|_2 = \sigma_1$

(if say $\|S\|_2 = 2\cdot\sigma_1$

$\Rightarrow \exists x$ s.t. $\|x\|=1$ & $\|S\cdot x\| = 2\sigma_1$

$\Rightarrow \|U^T A V \cdot x\| = 2\sigma_1$

$\Rightarrow \|A \cdot (Vx)\| = 2\sigma_1$

but $\max \|Ay\| = \sigma_1$

contradict!!

$\therefore |S|_2 = |A|_2$)

Now; let $x = \dfrac{\begin{bmatrix}\sigma_1 \\ w\end{bmatrix}_{N\times 1}}{\sqrt{\sigma_1^2 + w^T w}}$

$\& \; \|Sx\| = \dfrac{1}{\sqrt{\sigma_1^2+w^Tw}} \cdot \left\| \begin{bmatrix}\sigma_1 & w^T \\ 0 & B\end{bmatrix} \cdot \begin{bmatrix}\sigma_1 \\ w\end{bmatrix} \right\|$

$= \dfrac{1}{\sqrt{\sigma_1^2+w^Tw}} \cdot \left\| \begin{bmatrix}\sigma_1^2 + w^Tw \\ B\cdot w\end{bmatrix}_{M\times 1} \right\|_{M\times 1}$

$\geq \dfrac{1}{\sqrt{\sigma_1^2+w^Tw}} \cdot (\sigma_1^2 + w^Tw)$

$\therefore \|Sx\|_{min} = \sqrt{\sigma_1^2 + w^Tw}$

max $= \sigma_1$

$\Rightarrow w^Tw = 0$

$\therefore w = [0]_{N-1\times 1}$

$\therefore$ now; $U^TAV = \underbrace{\begin{bmatrix}\sigma_1 & [0 \longrightarrow] \\ [0] & [\;\; B \;\;]\end{bmatrix}_{M\times N}}_{S}$.

claim: I can make S as diagonal. by doing same kind of decompions.

proof: via induction.

$M\times N \rightarrow M-1\times N-1 \rightarrow \cdots\cdots$

$1\times K$ or $K\times 1$

let $B = [a_1 \; a_2 \; \cdots \; a_K]$;

& $\|B\| = \sigma_1$

then $U^T \cdot B \cdot V = [\sigma_1 \; 0 \; 0 \; 0 \; 0 \; 0]$.

$\therefore B = U \cdot [\sigma_1 \; [0 \longrightarrow]] \cdot V^T$

$1\times 1 \quad 1\times K \quad K\times K$

diagonal ✓

Same for $K\times 1$.

Base case proved

Hence

$U_1 = A \cdot V_1$ ~~$\sigma_1 U_1 = A V_1$~~

U_i is unit norm.

let it hold for $M-1 \times N-1$

$\Rightarrow B = \underset{M-1 \times N-1}{U'} \cdot \underset{M-1 \times N-1}{D} \cdot \underset{N-1 \times N-1}{V'^T}$ (orthogonal) $D \rightarrow$ diagonal

$$\therefore U^T \cdot A \cdot V = \begin{bmatrix} \sigma_1 & [0 \longrightarrow] \\ \begin{bmatrix} 0 \\ \vdots \end{bmatrix} & \left[U' \cdot D \cdot V'^T\right] \end{bmatrix}_{M \times N}$$

$$= \underset{M \times M}{\begin{bmatrix} 1 & [0 \longrightarrow] \\ \begin{bmatrix} 0 \\ \vdots \end{bmatrix} & \left[U_1'\right] \end{bmatrix}} \cdot \underset{M \times N}{\begin{bmatrix} \sigma_1 & [0 \longrightarrow] \\ 0 & \\ \vdots & \left[D\right] \end{bmatrix}} \cdot \underset{N \times N}{\begin{bmatrix} 1 & 0 \\ 0 & V_1'^T \end{bmatrix}}$$

diagonal!

$\therefore U^T A V = U_1 \cdot D \cdot V_1^T$

$A = \underbrace{U \cdot U_1}_{\text{product of orthogonal matrices is orthogonal matrix.}} \cdot D \cdot V_1^T \cdot V^T$

$\therefore \boxed{A = U' \cdot D \cdot (V')^T}$ the singular value decomp.

* Now; $A = U \cdot S \cdot V^T$ exists.

What is AV_i? $= S_{ii} \cdot U_i$

this is a column vector.

$\therefore AV_i$ is along U_i, & orthogonal to all other column vectors of U.

$\Rightarrow$ orthogonal basis in $\mathbb{R}^n$ $\xrightarrow[\text{linear operator.}]{A_{m \times n}}$ orthogonal basis in $\mathbb{R}^m$.

* properties:-

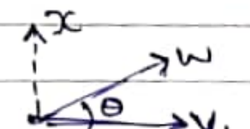
$\|A\|_2 = \sigma_1$ is unique.

but; the right singular vector V_1
s.t. $\|AV_1\| = \sigma_1$

is it unique? (upto sign). $\because |V_1|_2 = 1$

let there be w; such that $\|Aw\| = \sigma_1$
$|w| = 1$
linearly independent with V_1.

$x \rightarrow$ unit vector in direction of $w - \langle w, v_1 \rangle v_1$;



$\therefore \|Ax\| \leq \sigma_1$ (by defn of σ_1)

but lemme show ...

$\|Ax\| = \sigma_1$

$w = \cos\theta V_1 + \sin\theta (x)$
$= cV_1 + Sx$

$\sigma_1^2 = \|Aw\|^2 = (Aw)^T (Aw)$
$= (cAV_1 + SAx)^T (cAV_1 + SAx)$
$= c^2 \cdot \sigma_1^2 + S^2 \cdot |Ax|^2 + 2CS \cdot (Ax)^T (Av_1)$

if $x \perp V_1$. does $Ax \perp Av_1$?

$A = USV^T$. $(Ax)^T (Av_1)$
$= (USV^T x)^T (USV^T V_1)$
$= (x^T V S^T S V^T V)$
$= \left(\begin{bmatrix} 0 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T\right)_{1\times N} \cdot S^T S \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{N\times 1} = [0\ z_1\ z_2 \cdots] \cdot \begin{bmatrix} \sigma_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$

$\therefore |Ax| = \sigma_1$

$\perp V_1$

* Hence; $x \perp V_1$ & $|AV_1| = \sigma_1$
$|Ax| = \sigma_1$

∴ x must be 2nd right Singular vector of A.
with σ_1 as sing. value.

Seen as:-

$\because V^T x$ is right sing. vector for B.

$$\sigma_1 = \|Ax\| = |USV^T x|$$
$$= |S \cdot V^T x|$$
$$= \left| S \begin{bmatrix} 0 \\ y \end{bmatrix} \right|$$
$$= \left| \begin{bmatrix} \sigma_1 & 0 \\ 0 & B \end{bmatrix} \cdot \begin{bmatrix} 0 \\ y \end{bmatrix} \right|$$
$$= \left| \begin{bmatrix} 0 \\ By \end{bmatrix} \right|$$
$$= \underline{|B \cdot y|}$$

& ∴ $\|B\|_2 = \sigma_1$ ($\because \|B\|_2 \le \|A\|_2$) & now we have =).
first sing. value.

∴ if y is the right sing. vector for B; then this would transform to $V\begin{pmatrix} 0 \\ y \end{pmatrix}$ for A.

if $|B| = \beta$ then $\exists y$ s.t. $|B \cdot y| = \beta$
construct $y' = \begin{bmatrix} 0 \\ y \end{bmatrix}$
then $|S \cdot y'| = |By|$
then $|S \cdot V^T (Vy')| = \beta$
$\Rightarrow |A \cdot (Vy')| = \beta$
∴ $\|A\| \ge \|B\|$

$B = U_B \begin{bmatrix} \sigma_1 & \cdot \\ 0 & B_{n-1} \end{bmatrix} V_1^T$

* hence;
if $\sigma_1 = \|A\|_2$;

(i) if V_1 isn't unique (upto $\pm$); then the singular value σ_1 is repeated in diagonal-S notation. not 'simple!

(ii) if σ_1 is simple (no multiplicity) then V_1 is unique (upto sign).

* once σ_1, u_1, v_1 are determined; the remaining SVD is determined by linear operation A on space orthogonal to V_1.
(again see $|Ax|_{max}$!)

∴ we can build $U_1 V_1$.
∵ if $x \perp V_1$
then $Ax \perp V_1$
$\Rightarrow U_x \perp U_1$
∴ orthogonal matrices...

* SVD in multigaussians:-

$X_{M\times M} = A_{M\times N} W_{N\times N}$ $M \le N$ by definition.

$A = U \cdot S \cdot V^T$] possible for any real A.

$\therefore X = U \cdot S \cdot V^T \cdot W$ where comp of W are indepen
$N\times N \quad N\times 1$

$= U \cdot S \cdot W'$ ⟶ where comp. of W'
$M\times N \quad N\times 1$ are also independent!

$X \simeq (U \cdot S')\, W''$
$M\times M \quad M\times 1$

first M rows of S — first M elements of W'

$\boxed{X = A' \cdot W''}$

N>M to N=M

or; $X = W$; $A = I$

covar(x)
$= I \cdot I$
⇒ independent

$X = V^T \cdot W$; $A = V^T$

covar(x) $= A \cdot A^T$
$= V^T \cdot V = I$
⇒ independent
congrats!

& Standard normal (high!)

∴ its sum of Gaussians;

& i) standard normal;;

$\mu_{net} = \sum \mu_i p_i$
$= \sum 0 \cdot p_i$
$= 0$

$\sigma_{net} = \sqrt{\sum_{i=1}^{n} \sigma_i^2 \cdot p_i^2}$
$= \sqrt{\sum 1 \cdot p_i^2}$
$= \sqrt{1}$ ∵ orthogonal matrix.

(ii) independent;

Structure of an orthogonal matrix.

$\langle u_i, u_j \rangle = 0 \; \forall \; i \neq j$

$C_{old} = \underset{M\times N}{A} \cdot \underset{N\times M}{A^T} = M\times M = U S V^T \cdot V S^T U^T$

$= U S S^T U^T$

$C_{new} = \underset{M\times M}{U} \cdot \underset{M\times N}{S'} \cdot \underset{N\times M}{S'^T} \cdot \underset{M\times M}{U^T}$

$= U \cdot S \cdot S'^T \cdot U^T$

$= C_{old}$

Here $S \cdot S^T$ is a diagonal matrix.

$\therefore C = U \cdot (SS^T) \cdot U^T$

$= Q \cdot D \cdot Q^T$ form.

SS^T is the eigen vector values of C.

= square of sing. values of A

univariate → multivariate Bayesian → univariate.

Baysian Statistics:- for univariate cases. phew

Bayes theorem: $P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$

↳ very imp to statistics.

Now;

$$P(\mu=\mu_0 / data) = \frac{P(data/\mu=\mu_0) \cdot P(\mu=\mu_0)}{P(data)}$$

posterior — likelihood — prior — evidence

$$= \int_{\mu=-\infty}^{\infty} P(data/\mu) \, f(\mu)$$

X = random var. (discrete)

∴ modelling some unknown parameter.

Y = cont./disc r.v. modelling observed data

1) Likelihood: $P(Y=y \mid X=x)$
   - conditional

2) Evidence: $P(Y) = \sum_x P(Y=y, X=x)$
   - marginal pdf. (In the ways, the data could hav been generated) doesn't depend on parameter X.

3) Prior ∵ $P(X)$
   - Before data is observed; our beliefs about X.
   didn't have this before. was uniform.

4) Posterior: $P(X=x \mid Y=y)$
   - After data is observed, our beliefs are updated.

Improvement over MLE.

* ML estimation:-

-this assumes that prior info is not available
hence; uniform from $-\infty$ to ∞:
(or). over the domain of x.

$P(x=x_0|data) \propto \underbrace{P(data|x=x_0)}_{\text{Likelihood function.}}$

& we take the mode (argmax(LL)) of posterior function as our estimate of x.

* if we have prior belief (& thus a pdf) about x's value; then
the probability of true value would be even amplified; benifiting us.

∴ "improvement over MLE".

"only when we take a good prior fn!
not a non-sense one."

Eg: data $\{x_i\}$ drawn from Gaussian known var. = σ^2
unknown mean = μ

prior belief: μ is drawn from $N(\mu_0, \sigma_0^2)$

∴ posterior

$$P(\mu|data) = \frac{P(data|\mu)\cdot P(\mu)}{\text{(const) normalizing.}}$$

$$P(\mu) = c_0 \cdot e^{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0^2}\right)}$$

$$P(data|\mu) = c_0 \cdot e^{-0.5\left(\sum \frac{x_i-\mu}{\sigma^2}\right)}$$

joint pdf.

LL function.

$$= c_0 \cdot e^{-0.5\left(\frac{\mu^2 N - \mu 2(\sum x_i)}{\sigma^2}\right)}$$

$$= e^{-\frac{0.5}{\sigma^2/N}\left(\mu^2 - 2\mu(\bar{x})\right)} + \text{const}$$

$$\therefore P(data|\mu) \propto e^{\left(\frac{-1}{2}\cdot\frac{(\mu-\bar{x})^2}{\sigma^2/N}\right)}$$

$$\therefore P(\mu|data) = C \times e^{\left(\frac{-1}{2}\cdot\frac{(\mu-\bar{x})^2}{\sigma_N^2}\right)} \cdot e^{-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)}$$

$$= e^{\left(\frac{-1}{2}\cdot\frac{(\mu-\bar{\mu})^2}{c^2}\right)}$$

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

→ the MAP maximum a posterior estimate

$\hat{\mu}$: mean | mode for gaussian.

Now

$N \to \infty$: law of large nos. ignore prior.

$\sigma_0 \to \infty$: weak prior. see data

$\sigma_0 \to 0$: strong prior. see prior

* if asked MAP; then posterior middle.

what if asked something else...

given $\{x_i\}_{i=1 \text{ to } n}$ from $P(x|\theta^*)$

posterior pdf

$$P(\theta|\{x_i\}) = \frac{P(\{x_i\}|\theta) \cdot P(\theta)}{P(\{x_i\})}$$

Now, find $\hat{\theta}^*$ to minimize the MSD

$$E_{P(\theta|\{x_i\})} (\hat{\theta}^* - \theta)^2$$

would be $\hat{\theta}^* = E_{P(\theta|\{x_i\})} \theta$

could have been any 'risk function'

$\Rightarrow$ mean of Posterior PDF

= mode; if Gaussian.

→ Loss & risk function:-

if true value was θ; our estimator produces $\hat{\theta}$,

we incur a 'loss' based on

loss function $L(\hat{\theta}|\theta)$

But! θ is distributed as per posterior pdf

$P(\theta|\{x_i\})$

∴ Risk function

$R(\hat{\theta})$ = expectation of loss

$$= E_{P(\theta|\{x_i\})} L(\hat{\theta}, \theta)$$

Goal: choose optimal $\hat{\theta}$ $(= \hat{\theta}^*)$ to minimize risk function. $R(\hat{\theta})$

again point estimate.

• we can choose $L(\hat{\theta}|\theta)$ appropiately;

Such that

$\hat{\theta}$ = mode of $P(\theta|\{x_i\})$ → $L(\hat{\theta}|\theta) = I(\hat{\theta} \neq \theta)$ (indicator fn.)

$\hat{\theta}$ = mean of $P(\theta|\{x_i\})$ → $L(\hat{\theta}|\theta) = (\hat{\theta} - \theta)^2$ MSD

$\hat{\theta}$ = median of $P(\theta|\{x_i\})$ → $L(\hat{\theta}|\theta) = |\hat{\theta} - \theta|$ absolute dev.

order:-

1) take θ_{true}
2) take dataset X
3) do slope of LL at $\theta = \theta_{true}$.
4) ~~expectation over all X~~

→ Fisher Information:-

1) If likelihood function, $P(data|\theta)$ changes sharply wrt change in θ, around $\theta = \theta_{true}$; then easier to estimate θ_{true}.

2) If $P(data|\theta)$ has large spread wrt $\Delta\theta$, around θ_{true}; it will take many N-sized observations to get a close estimate to θ_{true}

1) $E_{P(data|\theta)_{true}}\left[\frac{\partial}{\partial\theta}\log(P(data|\theta))\Big|_{\theta_{true}}\right]$ (log likelihood!)

$= 0$ (expectation, over all possible observations)

2) $E_{P(data|\theta_{true})}\left[\left(\frac{\partial}{\partial\theta}\log P(data|\theta)\right)^2\Big|_{\theta_{true}}\right]$

$\neq 0$

$= I(\theta_{true})$

Fischer information for θ_{true}. (can be any no. so, just θ.)

$$I(\theta) = E_{P(x|\theta)}\left(\frac{\partial}{\partial\theta}\log(P(x|\theta))\right)^2 = -E_{P(x|\theta)}\frac{\partial^2}{\partial\theta^2}\log(P(x|\theta))$$

for fisher info, of a parameter expectation over all possible observations.

∴ more fisher info ⇒ larger negatives of $\frac{\partial^2}{\partial\theta^2}$ —

⇒ the likelihood is more concave

means; more certain about θ_{true}..

* Gaussian; σ^2 known

μ → determine!

then $P(x|\theta) = K \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$\therefore \log P(x|\theta) = -\frac{(x-\mu)^2}{2\sigma^2} + c.$

$\frac{\partial}{\partial\theta}(\ ) = \frac{(x-\mu)}{\sigma^2}$

$\frac{\partial^2}{\partial\theta^2}(\ ) = \frac{-1}{\sigma^2}$

$\therefore \boxed{I(\mu) = \frac{1}{\sigma^2}}$ → independent of μ!

(μ position parameter)

if N observations!

$\boxed{I(\mu) = \frac{N}{\sigma^2}}$

ANY unbiased estimator for θ.

Cramer-Rao Lower bound:-

How good can an estimator for θ get?

- good → see MSE for estimator.
- estimator → lets see unbiased estimators for now.

* let $\hat{\theta}(x)$ be an unbiased estimator for θ.

$$\Rightarrow E_{P(x|\theta)}\left(\hat{\theta}(x) - \underset{\text{true}}{\theta}\right) = 0.$$

we show that;

$$\boxed{\text{Var}(\hat{\theta}(x)) \geq (I(\theta))^{-1}} \; !$$

Same as MSE; since unbiased.

* for gaussian mean; MLE is unbiased.

$$\text{Var} = \left(\frac{1}{\sigma^2}\right)^{-1} \; ; \; I(\theta) = \frac{1}{\sigma^2}$$

$\therefore$ MLE mean is unbiased, minimum variance estimator.

Bayesian Cramer-Rao lower bound:-

for biased, unbiased estimators.

$$E_{Q(\theta|\alpha)}\left[\underbrace{E_{p(x|\theta)}(\hat{\theta}(x)-\theta)^2}_{\text{MSE.}}\right] \geq \left(E_{\theta(\theta|\alpha)}(I_p[\theta]) + \underset{Q}{J}(\theta)\right)^{-1}$$

where $Q(\theta|\alpha)$ is a prior; $\alpha \rightarrow$ hyper parameter.

$$J_Q(\theta) = \left(E_{Q(\theta|\alpha)}\left(\frac{\partial}{\partial\theta}\log Q(\theta|\alpha)\right)\right)$$

not α.

→ Jeffry's prior :-

take $Q(\theta) \propto \sqrt{I(\theta)}$.

then, if your friend takes $(\beta = g(\theta))$ ← transformation

$Q(\beta) \propto \sqrt{I(\beta)}$

Both will be in sync!

∴ making the prior "non-informative".

$$Q(\beta) = Q_\theta(g^{-1}(\beta)) \cdot \frac{\partial}{\partial \beta} g^{-1}(\beta)$$

if $Q_\theta(\theta) \propto \sqrt{I(\theta)}$,

then, we will get

$Q(\beta) \propto \sqrt{I(\beta)}$. ← same 'functional form'

Harmony everywhere.