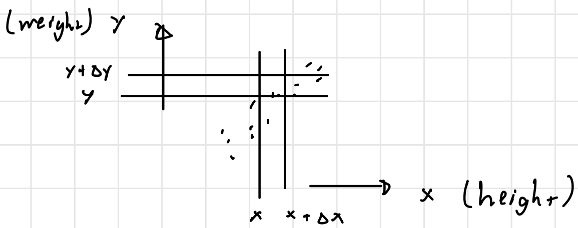
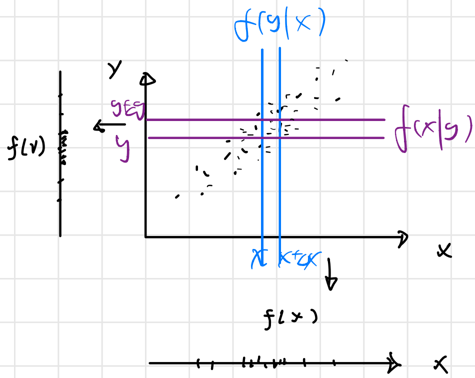
Lecture 11

- Probability:

  - $(x, y) \sim f(x, y)$
  - imagine a population of N ($\approx \infty$) equally likely possibilities
  - Population version scatterplot



- Counting:
  - joint: $f(x, y)$
  - marginal: $f(x)$, $f(y)$
  - conditional: $f(y|x)$, $f(x|y)$

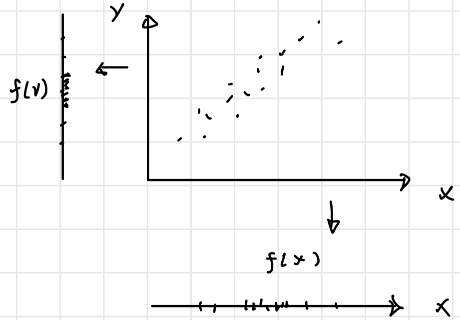


- 3 operations & 1 meta rule

(1) Marginalization :

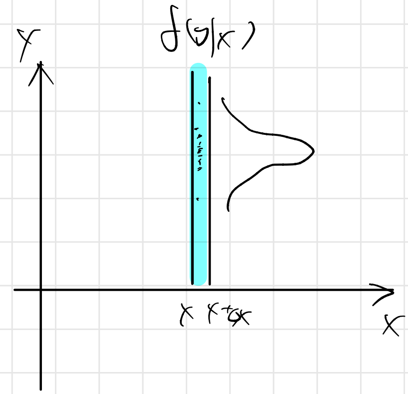
$$f(x) = \int f(x,y)\, dy$$

$$f(y) = \int f(x,y)\, dx$$

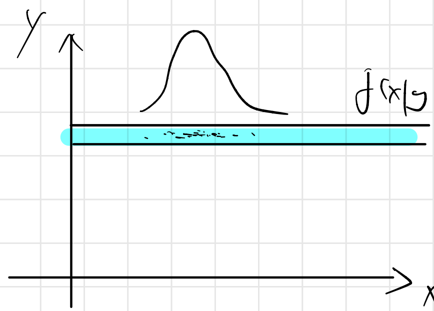


(2) Conditioning

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{f(x,y)}{\underbrace{\int f(x,y)\, dy}_{\text{normilization}}}$$



$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(x,y)}{\int f(x,y)\, dx}$$



(3) Factorization :

$$f(x,y) = f(x)\, f(y|x) = f(y)\, f(x|y)$$

- Meta Rule: $\underbrace{\text{Insert the same condition}}_{\text{count the same subpopulation}}$

- Example

(1) $$f(x \mid z) = \int f(x, y \mid z)\, dy$$ Marginalization

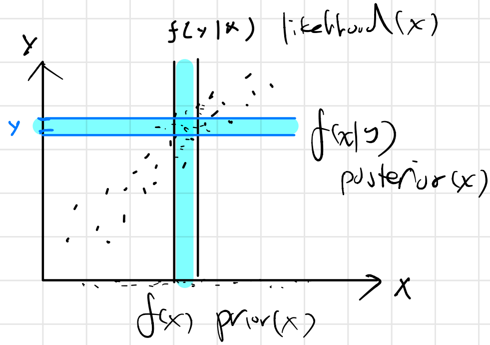
(2) $$f(y \mid x, z) = \frac{f(x, y \mid z)}{f(x \mid z)}$$ Conditioning

(3) $$f(x, y \mid z) = f(x \mid z)\, f(y \mid x, z)$$ factorization

- Discrete: $\int \longmapsto \Sigma$

- Baye's Rule :

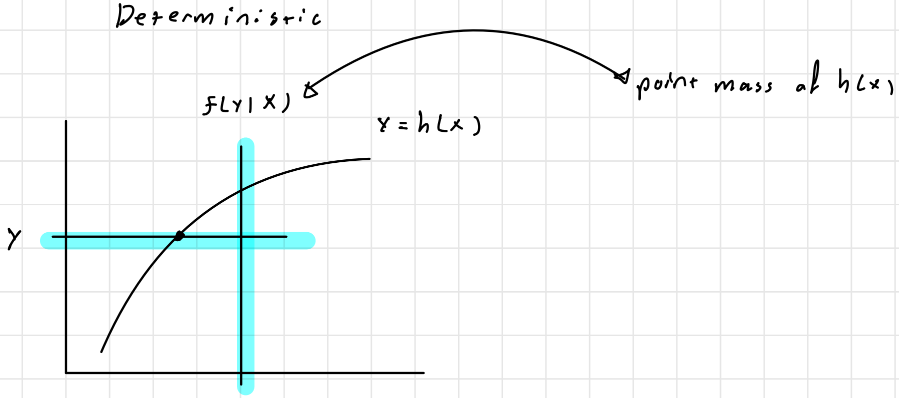
$$X \quad \text{cause (disease)}$$
$$f(x|y) \ \text{inverse/inference} \uparrow \quad \downarrow f(y|x)$$
$$Y \quad \text{effect (symptom)}$$



$$\begin{cases} X \sim f(x) & \text{prior}(x) \\ [Y \mid X] \sim f(y \mid x) & \text{likelihood}(x) \end{cases}$$

$$\to \quad [X \mid Y] \sim f(x \mid y) \quad \text{posterior}(x)$$

Deterministic



Solving : $y = h(x)$

Solution: $x = h^{-1}(y)$

- Baye's Rule:

$$f(x \mid y) = \frac{f(x,y)}{f(y)} \qquad \text{conditioning}$$

fixed

$$= \frac{f(x,y)}{\int f(x,y)\,dx} \qquad \text{marginalization}$$

$$= \frac{f(x)\,f(y \mid x)}{\int f(x)\,f(y \mid x)\,dx} \qquad \text{factorization}$$

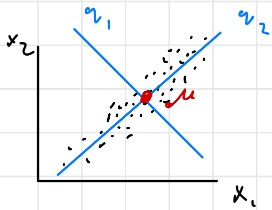
- Note: $\underbrace{f(x \mid y)}_{\text{fixed}} \propto \underbrace{f(x,y)}_{\text{fixed}} = \underbrace{f(x)}_{\text{prior}}\;\underbrace{f(y \mid x)}_{\text{likelihood}}$

as a function of x

- Multivariate Gaussian:

$$X \sim \mathcal{N}(\mu, \Sigma)$$

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



$N \approx \infty$

$\Sigma = Q \Lambda Q^T$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(0, \; \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

Special case when $\Sigma_{21} = \Sigma_{12} = 0$,
Then,

$$f(x_1, x_2) = f(x_1) f(x_2), \quad \text{i.e.} \quad x_1 \perp x_2$$

- In general diagonize

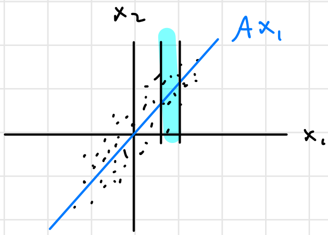
$$\begin{pmatrix} X_1 \\ \varepsilon \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 - AX_1 \end{pmatrix} = \begin{pmatrix} I & 0 \\ -A & I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix}\right)$$

$$A = \Sigma_{21}\Sigma_{11}^{-1}$$

$$X_2 = AX_1 + \varepsilon, \quad \varepsilon \perp X_1$$

$$[X_2 \mid \underset{\uparrow}{X_1}] \sim \mathcal{N}(AX_1, \ \Sigma_{22} - \dots)$$

fixed



- Change Notation:

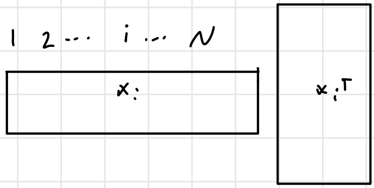
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right)$$

- popularion - version regression

1		
⋮		
i	x_i^T	y_i^T
⋮		
N	X	Y

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

$\frac{1}{N} X^T X =$



$$= \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \overset{1\,dim}{\equiv} \frac{1}{N} \sum x_i^2 = \frac{1}{N} |\vec{x}|^2$$

$$= \mathbb{E}\left((x-\mu)(x-\mu)^T \right) = Var(x) = \Sigma_{xx}$$

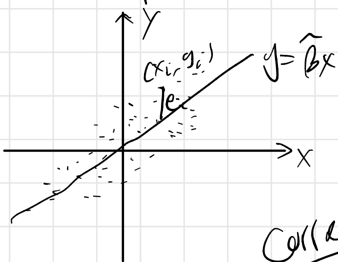
Similarly $\quad \frac{1}{N} X^T Y = \Sigma_{xy} \overset{1\,dim}{\equiv} \frac{1}{N} \langle \vec{x}, \vec{y} \rangle$

So $\quad \hat{\beta} = \Sigma_{xx}^{-1} \Sigma_{xy} = A^T$

1 dim

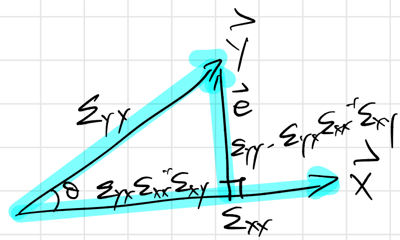
	$\vec{x}$	$\vec{y}$	$\vec{e}$
i	x_i	y_i	e_i
N	x_N	y_N	

Scatterplot



2 dim

Vector plot



Correlation

$$\rho = Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = Cov\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right)$$

$$= \frac{\frac{1}{N}\langle \vec{x}, \vec{y} \rangle}{\sqrt{\frac{1}{N}|\vec{x}|^2}\sqrt{\frac{1}{N}|\vec{y}|^2}} = \cos\theta$$

$$\frac{X-\mu_X}{\sigma_X} \rightarrow \quad \frac{Y-\mu_Y}{\sigma_Y} \rightarrow \quad \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Regression

$$[\underbrace{Y}_{\text{son height}} \mid \underbrace{X}_{\text{father height}}] \sim N(\rho X, \underbrace{1-\rho^2}_{<1})$$

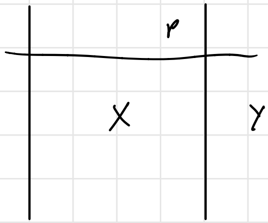
$$1-\rho^2 = \sin^2\theta = \frac{|\vec{e}|^2}{|\vec{Y}|^2} = \frac{Var(e)}{Var(y)}$$

Regression towards the mean

original meaning of "regression"

- Bayesian Regression:

Data



$$[Y \mid X, \beta] \sim N(X\beta, \sigma^2 I_n)$$

$$\overset{P(\beta)}{\text{prior}}: \quad \beta \sim N(0, \tau^2 I_p)$$

(as if β is sampled from population of N possibilities)

posterior:

$$P(\beta \mid X, Y) \propto \underbrace{P(\beta)}_{\text{prior}(\beta)} \underbrace{P(Y \mid X, \beta)}_{\text{likelihood}(\beta)}$$

Note: Gauss assumed $\tau^2 = \infty$, so we know nothing about β (non-informative prior).

- What about finite τ^2

$$P(\beta \mid X, Y) \propto \frac{1}{(2\pi)^{\frac{p}{2}} (\tau^2)^{p/2}} \exp\left(-\frac{1}{2\tau^2} \beta^T \beta\right)$$

$[Y \mid X, \beta]$ density

$$\rightarrow \frac{1}{(2\pi)^{n/2} (\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)\right)$$

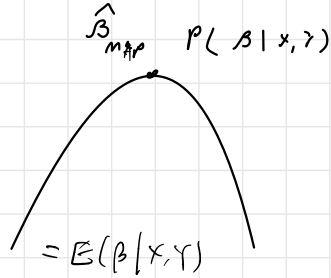
$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} |Y - X\beta|^2 + \frac{1}{\tau^2} |\beta|^2\right)\right)$$

Ridge-Regression.

- Maximum A Posterior: (MAP)

$$\min_\beta \frac{1}{\sigma^2} |Y - X\beta|^2 + \frac{1}{\tau^2} |\beta|^2$$

$$= \min_\beta |Y - X\beta|^2 + \underbrace{\frac{\sigma^2}{\tau^2}}_{\lambda > 0} |\beta|^2$$



- $[\beta \mid X, Y] \sim p(\beta \mid X, Y)$

$\downarrow$

basis for inference / prediction

$\downarrow$

Does not overfit

1, 2, ..., n	X	Y
0	x_0	$\hat{Y}_0 = x_0^T \beta$

account for uncertainty

$\rightarrow$ Sample $\beta^{(1)}, \dots, \beta^{(m)}, \dots \beta^{(M)} \sim p(\beta \mid X, Y)$

$\rightarrow$ $\hat{Y}_0^{(1)}, \dots, \hat{Y}_0^{(m)}, \dots, \hat{Y}_0^{(M)} \sim P(\hat{Y}_0 = x_0^T \beta \mid X, Y)$

$\rightarrow \begin{cases} E(\hat{Y}_0 \mid X, Y, x_0) \\ Var(\hat{Y}_0 \mid X, Y, x_0) \end{cases}$

can be derived directly

population of β $\rightarrow$ population of $\begin{pmatrix} Y \\ \hat{Y}_0 \end{pmatrix} \sim N\left(0, \quad \right)$

- Bayesian / Frequentist controversy

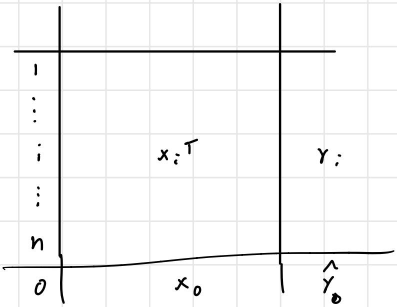
  - example: prior, speed of light $c \sim prior(c)$

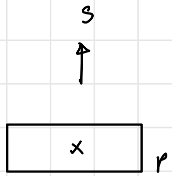
  - $c \sim p(c)$
  - feasible of repeated sampling ?

$c_1, c_2, \dots, c_m \overset{iid}{\sim} p(c)$

$\downarrow$ $\qquad \downarrow$

frequency $\rightarrow$ prob

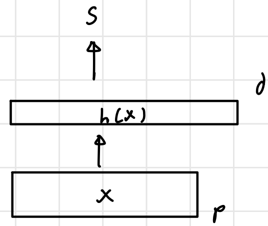




$s = f(x) = x^T \beta$

$\beta \sim N(0, \sigma^2 I_p)$

(Ridge)



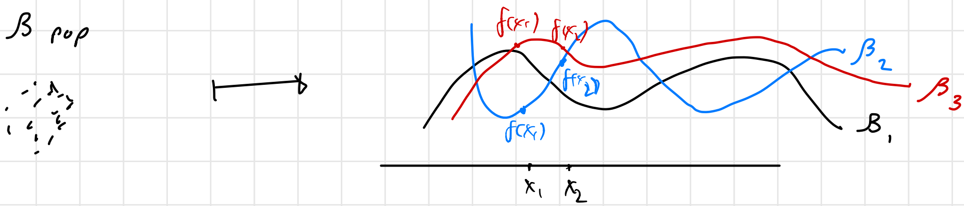
$s = f(x) = h(x)^T \beta$

$\beta \sim N(0, \sigma^2 I_d)$

(kernel)

- For simplicity assume X is 1d

population of β $\longrightarrow$ population of $f(x) = h(x)^T \beta$



So we say $f(x) = h(x)^T \beta \sim$ Gaussian Process

So $\mathrm{Cov}(f(x_1), f(x_2)) = \mathrm{cov}(h(x_1)^T \beta, h(x_2)^T \beta)$

Remember $\mu = 0$

$$= \mathbb{E}(h(x_1)^T \beta \beta^T h(x_2))$$, β is random, x_1 & x_2 is fixed

$$= h(x_1)^T \mathbb{E}(\beta\beta^T) h(x_2)$$

$$= h(x_1)^T \sigma^2 I_d \, h(x_2)$$

$$= \sigma^2 \langle h(x_1), h(x_2) \rangle$$

$$= \sigma^2 k(x_1, x_2)$$

So $f(x) \sim GP(0, \sigma^2 k)$

1		$f(x_1)$	y_1
$\vdots$			
i	x_i^T	$f(x_i)$	$y_i = f(x_i) + \varepsilon_i$
$\vdots$			
n		$f(x_n)$	y_n
0	x_0	$f(x_0)$	

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim N(0, \sigma^2 I)$$

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_i) \\ f(x_n) \\ \hline f(x_0) \end{pmatrix} \sim N\left(0, \begin{array}{c} \\ i \\ 0 \end{array} \begin{array}{c} \begin{array}{cc} j & 0 \end{array} \\ \left[\begin{array}{c|c} \sigma^2 k_{ij} & \sigma^2 k_{i0} \\ \hline \sigma^2 k_{0j} & \sigma^2 k_{00} \end{array}\right] \end{array}\right)$$

$$k_{ij} = k(x_i, x_j)$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \\ \hline \hat{y}_0 = f(x_0) \end{pmatrix} \sim N\left(0,\ \left[\begin{array}{c|c} \tau^2 K + \sigma^2 I_n & \tau^2 k_0^T \\ \hline \tau^2 k_0 & \tau^2 k_{00} \end{array}\right] \right)$$

$$\text{So } [\hat{y}_0 = f(x_0) \mid y, x] \sim N\big(\tau^2 k_0 (\tau^2 K + \sigma^2 I_n)^{-1} y,$$

$$\tau^2 k_{00} - \tau^2 k_0 (\tau^2 K + \sigma^2 I_n)^{-1} \tau^2 k_0^T \big)$$

Kernel Regression

$$k_0 \left(K + \frac{\sigma^2}{\tau^2} I_n\right)^{-1} y$$

$$= k_0 (K + \lambda I_n)^{-1} y$$

$$= k_0 c$$

$$= \begin{bmatrix} \quad k_0 \quad \end{bmatrix} \begin{bmatrix} \\ c \\ \\ \end{bmatrix}$$

$$f(x_0) = \sum_{i=1}^{n} c_i \, k(x_i, x_0)$$