Basic concepes
Equally likely
$P(A) = |A| / |\Omega|$.
$|A|$: size
define probas measure

Two interpretations:
(1) subject belief of uncertainty
(2) long run frequency
If we repeat experiment a large number of times
$P(A)$: how often A happens

Random variables
Experiment:
Randomly sample a person from a population
X = height of this person
Event / statement: $X > 6$ ft

Sample space Ω = whole population
$= \{\omega_1, \dots, \omega_n\}$
$= \{\omega\}$.

For a random person $w \in \Omega$
height $X(w)$ $\Omega \to \mathbb{R}$
each person $\to$ height

$A: X > 6$
$A = \{ \cancel{w \in \Omega}\ w : X(w) > 6 \} \subset \Omega$
sub-population
$P(A) = P(X > 6) = P(\{w : X(w) > 6\})$
$= |A| / |\Omega|$.

Experiment $\to$ outcome $\to$ number (R.v.)
event $\uparrow$ $\uparrow$
statement $\to$ Equation / inequaliey about

special event
$X \le x$
$\downarrow$ $\downarrow$
Capital lower case: specific value

$F(x) = P(X \le x) = P(\{w : X(w) \le x\})$
cumulative distn function (cdf.)

2 types: discrete: gender; ...

continous: height, weight, age

Basic Events

discrete: $X = x$

$P(X = x) = p(x)$

prob. mass func.

$P(A) = \sum_{x \in A} p(x)$

$p(x)$: Population propotion of $X(\omega) = x$.

How often $X = x$.

continuous:

basic event: $X \in (x, x + \Delta x)$

$P(X \in (x, x + \Delta x))$:

population propotion of slice / ciration

How often $X \in (x, x + \Delta x)$.

$f(x) = \lim_{\Delta x \to 0} \frac{P(X \in (x, x+\Delta x))}{\Delta x}$ → pdf.

$\text{density} = \frac{\text{\# of people within neighborhood} / N}{\text{size of neigh.}}$

$= P(X(\omega) \in (x, x + \Delta x)) / \Delta x$.

Summary: $E(x) \begin{cases} \sum_x x\, p(x) \\ \int x f(x)\, dx. \end{cases}$

long run averange

population averange.

$P\left(\left| \frac{\Sigma X_i(w)}{n} - \frac{1}{2} \right| < \varepsilon \right) \xrightarrow[\infty]{n} 1$ (ε fixed)

— weak law of large #.

Strong law:

$P\left(\lim_{n\to\infty} \frac{\Sigma X_i(w)}{n} = \frac{1}{2} \right) = 1.$

Conditional prob.

$P(A|B) = P(A \cap B) / P(B).$

STATS 200

随机变量及其线性变换

Random variables

Experiment → outcome → numble

Event: statement about (outcome)

R.V.: $=, \le$ about (numble)

Random variable X

special event $X \le x$ $\{w : X(w) \le x\}$.

$F(x) = P(X \le x)$ cumulative distn. fun.

basic events:

discrete $X = x$ $P(x) = P(X = x)$ pmf.

continuous $X \in (x, x + \Delta x)$ pdf

Interpretation
- population distn.

$$F(x) = P(X \le x) = \begin{cases} \sum_{z \le x} P(z) \\ \int_{-\infty}^{x} f(x)\,dx. \end{cases}$$

Summary: $$E(X) = \begin{cases} \sum_x x\, p(x) \\ \int x f(x)\,dx. \end{cases}$$

- population average
- long run average

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow E(X)$

$(x_1, x_2, \dots, x_n) \overset{iid}{\sim} f(x)$

$$E(h(x)) = \begin{cases} \sum_x h(x)\,p(x) \\ \int h(x)\,f(x)\,dx. \end{cases}$$

h( ) : utiliey fun.

$E(h(x))$ vs. $E(h(Y))$.

$Y = h(X)$ $\quad X \sim f_X(x)$

$\quad Y \sim f_Y(y)$

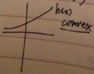
$E(Y) = \int y\, f_Y(y)\, dy = E(h(x)) = \int h(x) f_X(x)\, dx$

$Var(X) = E((X-\underset{EX}{\mu})^2). = EX^2 - (EX)^2.$

variation, fluation, volatiley, spread

linear transformation:

$E(aX+b) = aEX + b.$



$E(h(x)) \geq h(EX)$ convex

$Var(aX+b) = a^2 Var(X)$

Transformation

$X \sim f_X(x)$. $\quad Y = h(X)$. $\quad Y \sim f_Y(y)$.

$h(x)$: monouton increacing, continuing differentiable

$X = h^{-1}(Y) = g(Y)$.

$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx$

$F_Y(y) = P(Y \le y) = P(h(X) \le y)$

$= P(X \le g(y))$

$= \int_{-\infty}^{g(y)} f(x)\,dx = F_X(g(y))$.

$f_Y(y) = F_Y'(y) = \frac{d}{dy} F_X(g(y))$

$= F_X'(g(y))\, g'(y)$

$= f_X(g(y)) \times g'(y)$

$f_Y(y) = P(Y \in (y, y+\Delta y)) / \Delta y$

$= P(X \in (x, x+\Delta x)) / \Delta y$

$\cancel{= P(X \in (x, x+\Delta x)) / \Delta x}$

$= f_X(x) \cdot \Delta x / \Delta y$

$= f_X(x) \cdot g'(y)$

Symbollically,

$X \sim f_X(x)\, dx = f_X(g(y))\, d\,g(y)$.

$= f_X(g(y)) |g'(y)| dy = f_Y(y) dy \sim Y$.

NOTES 200A. OM OT OW OT OF OS OS DATE / /

S3

联合分布期望、方差…

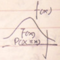
Special case of transformation

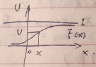
Inverse method

$f(x) \to F(x)$.

How to generate $X \sim f(x)$

- generate $U \sim Unif\ [0,1]$
- $X = F^{-1}(U)$.





$F'(x) = f(x)$ $\quad F(x) = U$.

$X = F^{-1}(U)$.

need to prove $P(X \le x) = F(x)$.

$\uparrow$

$F^{-1}(U)$.

$P(F^{-1}(U) \le x) = P(U \le F(x))$. $= F(x)$.

$U \sim Unif\ [0,1]$ $\quad P(U \le u) = u$.

pseudo-random #'s

linear congruential method

seed x_0

$x_{t+1} = a x_t + b \mod M$.

$(x_{t+1} = 7^5 x_t \mod 2^{31} - 1)$.

$U_t \sim Unif \{0, \frac{1}{M}, \frac{2}{M}, \ldots, \frac{M-1}{M}\}$

$\searrow$ M large

Unif $[0,1]$.

$X_t \sim \text{Unif}\ \{F^{-1}(0), F^{-1}(\frac{1}{M}), \ldots, F^{-1}(\frac{M-1}{M})\}$

p-value ~~to~~ Unif $[0, 1]$.

Two R. V's. X, Y.

basic statements $X = x$ & $Y = y$.

$X \in (x, x+\Delta x)$ & $Y \in (y, y+\Delta y)$

A B

$P(A \cap B)$

$$P(A \cap B) \begin{cases} p(x, y) & \text{discrete} \\ f(x, y)\Delta x \Delta y & \text{continous} \end{cases}$$

$$f(x, y) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} P(X \in (x, x+\Delta x)\ \&\ Y \in (y, y+\Delta y)) / \Delta x \Delta y$$

$$P(A) = \begin{cases} \sum_{(x,y) \in A} p(x, y) \\ \int_A f(x, y)\, dx\, dy. \end{cases}$$

$$E(h(X, Y)) = \frac{1}{M} \sum_{w=1}^{M} h(X(w), Y(w))$$

$$= \sum_{\text{cells } (x, x+\Delta x) \times (y, y+\Delta y)} h(x, y) f(x, y) \Delta x \Delta y$$

$$\xrightarrow[0]{\Delta x \Delta y} \int h(x, y) f(x, y)\, dx\, dy.$$

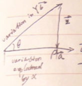
define $E(h(x,y)) = \int h(x,y) f(x,y)\,dx\,dy$.

$Var(h(x,y)) = E(h(x,y) - \mu_h)^2$

$Cov(X,Y) = E((X-\mu_X)(Y-\mu_Y))$.

$Corr(X,Y) = \dfrac{Cov(X,Y)}{\underbrace{\sqrt{Var X}}_{\sigma_X}\,\underbrace{\sqrt{Var Y}}_{\sigma_Y}}$

$= Cov\left(\dfrac{X-\mu_X}{\sigma_X}, \dfrac{Y-\mu_Y}{\sigma_Y}\right)$.



$\dfrac{\langle \vec{a}, \vec{b} \rangle}{|\vec{a}|\,|\vec{b}|} = \cos\theta \in [-1, 1]$.

$E(Y - \underbrace{\rho}_{\text{Corr.}} X)^2$ min.

ρ^2 = percentage of variation in Y explained by

= strength of linear relationship.

$Y = \rho X + Z$.

regression: go back $\rho < 1$.

§4 多元分布及线性变换 & 根分布

Multivariate distribution

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \underline{Var\ X = E((X-\mu)(X-\mu)^T)}$$

joint distribution = lim ...

Expectation:

generalize $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ & \cdots & & \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix}$

$$EX = \begin{pmatrix} Ex_{11} & Ex_{12} & \cdots & Ex_{1n} \\ Ex_{21} & Ex_{22} & \cdots & Ex_{2n} \\ & \cdots & & \end{pmatrix}$$

Back to general $X_{m\times n}$. 线性变换

linear operation $(A_{k\times m} X_{m\times n})_{k\times n}$.

$(X_{m\times n} B_{n\times 1})_{m\times 1}$.

Recall $\underline{E(AX) = AEX}$.

$$Var\ X = (E(x_i-\mu_i)(x_j-\mu_j))_{n\times n}$$
$$= \begin{pmatrix} Var\,x_1 & cov(x_1,x_2) & \cdots \\ & Var\,x_2 & \\ & & \ddots \\ & & & Var\,x_n \end{pmatrix}_{n\times n}.$$

Variance - Covariance Matrix.

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

$$\mathrm{cov}(X, Y) = E\left((X-\mu_X)(Y-\mu_Y)^T\right)_{n\times m}.$$

$$\mathrm{var}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mathrm{Var}\,X & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{Var}\,Y \end{pmatrix}.$$

Linear transformation

$$Y_{m\times 1} = A_{m\times n} X_{n\times 1}$$

$$EY = A\,EX.$$

$$\mathrm{Var}\,Y = A\,\mathrm{Var}(X)\,A^T.$$

Density $X \sim f_X(x)$. $Y \sim f_Y(y)$.

$$f_X(x) = \lim_{D_x \to x} \frac{P(X \in D_x)}{|D_x|}.$$

$$f_Y(y) = \lim_{D_y \to y} \frac{P(Y \in D_y)}{|D_y|} = \lim_{D_y \to y} \frac{P(X \in D_x)}{|D_y|}$$

$$x \in D_x \iff y \in D_y.$$

$$= \lim_{D_y \to y} \frac{f_X(x)\,|D_x|}{|D_y|} = f_X(x) / |\mathrm{Det}(A)|.$$

Multivariate Normal.

$$Z \sim N(0, I_n).$$

$$E(Z) = 0 \quad \mathrm{Var}(Z) = I_n.$$

$$Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \quad z_i \overset{iid}{\sim} N(0, 1).$$

$$f_Z(z) = \prod_{i=1}^n f(z_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n z_i^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2} z^T z}.$$

$X = AZ$.

$E(X) = 0$. $\quad Var\, X = AA^T = \Sigma$

$X \sim (0, \Sigma)$.

$$f_X(x) = f_Z(z) \,/\, |\det A|$$

$$= f_Z(A^{-1}x) \,/\, |\det A|$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(A^{-1}x)^T(A^{-1}x)} \Big/ |\det A|$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}x^T\Sigma^{-1}x} \Big/ |\det A|.$$

generalize $\quad X \sim N(\mu, \Sigma)$.

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \frac{1}{|\det A|}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

Eigen-decomposition.

$$\Sigma = Q \Lambda Q^T.$$

$Q = (\vec{q}_1, \dots \vec{q}_n)$. $\quad \langle q_i, q_j \rangle = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$

$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$

$X = y_1 \vec{q}_1 + \cdots + y_n \vec{q}_n = QY \qquad Y = (y_1 \cdots y_n)^T$

$y_i = \langle X, \vec{q}_i \rangle \qquad Y = (\langle X, \vec{q}_i \rangle)^T$

$= (q_i^T X)^T = Q^T X.$

$X = QY$, $\quad Y = Q^T X$, $\quad \Rightarrow \quad QQ^T = Q^TQ = I$

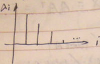
synthesis $\quad$ analysis

$X \sim N(0, \Sigma) \qquad Y = Q^T X$

$E(Y) = 0 \qquad Var\, Y = Q^T \, Var\, X \, Q = \Lambda$

$y_i \overset{indep}{\sim} N(0, \lambda_i)$.

principal component analysis



$\lambda_i \approx 0$ for $i > d$

$y_i \approx 0.$

dimension reduction $X = \sum_{i=1}^{d} y_i q_i$. $d \ll n$.

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2} x^T \Sigma^{-1} x}$$

$$x^T \Sigma^{-1} x = (QY)^T (Q \Lambda Q^T)^{-1} (QY)$$

$$= Y^T Q^T Q \Lambda^{-1} Q^T Q Y.$$

$$= Y^T \Lambda Y. = \sum_{i=1}^{n} y_i^2 / \lambda_i .$$

Contour plot

$$X^T \Sigma^{-1} X = \text{const}.$$

$$\Sigma\, y_i^2 / \lambda_i = \text{const}.$$

55 200A 10/16/13

多元正态分布 & 条件期望、方差

Multiverinte Gaussian

$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ scores of d subjects

- - - - d sports

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

Regression

$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, \Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix})$

前提： $Y = AX + \Sigma$ $\quad cov(X, \Sigma) = 0$

$cov(X, \Sigma) = cov(X, Y - AX)$

$= cov(X, Y) - cov(X, AX)$

$= \Sigma_{XY} - \Sigma_{XX} A^T$

$\Rightarrow \cancel{A = \Sigma_{XY} \Sigma_{XX}^{-1}}$. $A^T = \Sigma_{XX}^{-1} \Sigma_{XY}$

$Var(\Sigma) = Var(Y - AX) = Var(Y) - Var(AX)$.

$\cancel{Var(Y) =}$ $= \Sigma_{YY} - A \Sigma_{XX} A^T$.

$= \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$

Fix x, $X = x$

$Y = Ax + \Sigma$

$Y \sim N(Ax, \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY})$.

$[Y \mid X] \sim N(\Sigma_{YX} \Sigma_{XX}^{-1} x, \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY})$.

Conditioning

$(X, Y) \sim f(x, y)$

$X \downarrow$ age, $Y \downarrow$ height

$f(x, y)\Delta x \Delta y = P(X \in (x, x+\Delta x), Y \in (y, y+\Delta y))$.

basic event n.

$f_X(x)\Delta x = P(X \in (x, x+\Delta x))$

$= \sum_{\text{bins of } y} P(X \in (x, x+\Delta x), Y \in (y, y+\Delta y))$.

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

$f(y|x) = P(Y \in (y, y+\Delta y) \mid X \in (x, x+\Delta x)) / \Delta y$.

dist. of a slice of a population.

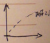
$E(Y|X=x) = \int y f(y|x)\, dy = h(x)$.

$Var(Y|X=x) = E((Y-h(x))^2 \mid X=x)$.

non-linear regression.

$Y = h(X) + \varepsilon$

within group variance



between group variance

$E(Y|X) = h(X)$

Adam : $\underline{E\,E(Y|X) = EY}$.

Eve : $\underline{Var\,Y = E(Var(Y|X)) + Var(E(Y|X))}$

$Y = h(X) + \varepsilon$, $\varepsilon = Y - h(X)$

$E(\varepsilon) = E(Y) - E(h(X)) = 0$.

$cov(\varepsilon, h(X)) = E(\varepsilon h(X)) = E\,E(\varepsilon h(X) \mid X) =$

$$
\begin{aligned}
Var(Y) &= Var(Z) + Var(h(X)) \\
&= E Z^2 + Var(E(Y|X)). \\
&= E(E(Z|X)) + \cdots \\
&= E(Var(Y|X)) + \cdots
\end{aligned}
$$

prediction of Y based on X

$$
r^2 = Var(E(Y|X)) / Var(Y).
$$

条件期望 / 方差.

S6 条件概率 & 贝叶斯公式 & 各种例子

Conditioning

$P(A|B) = P(A\cap B) / P(B)$.

- population ...

$P(A|B) \to$ as if sample from male pop.

condition = sub-population

~~$P(A\cap B)$~~

$P(A|B\cap C) = P(A\cap B|C) / P(B|C)$.

Chain rule: $P(A\cap B) = \underbrace{P(B)}_{\text{B happens}} \underbrace{P(A|B)}_{\text{A happens when B happens}}$

rule of total prob.

$P(A) = \sum_{i=1}^{n} P(A\cap B_i)$

$= \sum_{i=1}^{n} P(B_i) P(A|B_i)$.

Bayes rule

$P(B_i|A) = P(B_i\cap A) / P(A)$

$= P(B_i) P(A|B_i) / \sum P(A\cap B_i)$

$= P(B_i) P(A|B_i) / \sum P(B_i) P(A|B_i)$

B_i cause

$\downarrow$

A effect.

NOTES

Random Variables

- A, B, ...
- $X \le x$, $X \in (x, x+\Delta x)$ $\quad X = x$.

$P(X=x) = P(x)$ pmf.

$P(X \in (x, x+\Delta x)) = f(x)\,\Delta x$ pdf.

$P(X \le x) = F(x)$ cdf.

$P(Y=y \mid X=x) = P(Y=y \cap X=x) / P(X=x)$.

$P(y \mid x) = P(x,y) / P_x(x)$.

$P_{Y|X}$.

$P(Y \in (y, y+\Delta y) \mid X \in (x, x+\Delta x))$

$= P(X \in (x, x+\Delta x) \cap Y \in (y, y+\Delta y)) / P(X \in (x, x+\Delta x))$

$= f(x,y)\,\Delta x \Delta y / f_x(x)\,\Delta X$

$\Rightarrow f(y \mid x) = f(x,y) / f_x(x) \rightarrow \int f(x,y)\,dy$.

$f_{Y|X}$.

Given $f(x)$, $f(y \mid x)$

$f_Y(y) = \int f(x)\, f(y \mid x)\, dx$.

Bayes Rule: Given $P(x)$ $\quad P(y \mid x)$.

$P(x \mid y) = P(x,y) / P_Y(y)$

$= P(x)\,P(y \mid x) / \sum_x P(x)\,P(y \mid x)$

$f(x \mid y) = f(x,y) / f_Y(y) = f_X(x)\, f(y \mid x) / \int f_X(x)\, f(y \mid x)\, dx$

Example 1: X, Y are discrete

$x = \begin{cases} D & \text{disease} \\ N & \text{no disease} \end{cases}$

1% population has disease

$P(X = D) = 1\%$.

Medical test:

$Y = \begin{cases} + \\ - \end{cases}$ $\quad P(+|D) = 90\% \quad P(-|D) = 10\%$

$\quad P(+|N) = 10\% \quad P(-|N) = 90\%$.

$$P(D|+) = P(+|D) \cdot P(D) \Big/ P(+|D)P(D) + P(+|N)P(N)$$

$$= \frac{0.01 \times 0.9}{0.01 \times 0.9 + 0.99 \times 0.1} = \frac{1}{12}.$$

Example 2: X discrete, Y continuous

$x = \begin{cases} 0 & \text{female} \\ 1 & \text{male} \end{cases}$ $\quad P(X=1) = P_X(1) = \lambda$.

Y = height

$[Y | X = 1] \sim f_1(y) \quad [Y | X = 0] = f_0(y)$.

$$f(y) = P_X(1) f_1(y) + P_X(0) f_0(y)$$

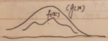
$$= (\overset{\lambda}{\cancel{1-\lambda}}) f_1(y) + (1-\lambda) f_0(y).$$ mixture model.

$$P(X=1|y) = P(1|y) = \frac{P_X(1) f(y|1)}{P_X(0) f(y|0) + P_X(1) f(y|1)}$$

$$= \lambda f_1(y) \big/ [\lambda f_1(y) + (1-\lambda) f_0(y)].$$

M people in population
male #: λM.
female #: $(1-\lambda)M$.
$P(Y \in (y, y+\Delta y)) = f(y)\Delta y$.

Example 3: X continuous, Y discrete
Rejection sampling
hard $X \sim f(x)$.
can generate $X \sim g(x)$ simple
$f(x) \le c\, g(x)$
envelop



Algorithm:
(1) generate $X \sim g(x)$
generate $U \sim U[0,1]$.
If $U \le \frac{f(x)}{c\,g(x)}$ then return X
otherwise go back to (1)

returned $\alpha \sim f(x)$

$$Y = \begin{cases} 1 & U \le f(x)/c\,g(x) \\ 0 & \text{otherwise} \end{cases}$$

$[X \mid Y=1] \sim f(x \mid 1) = \dfrac{f_X(x)\, P(1|x)}{\int f_X(x) P(1|x)\, dx}$

$$= g(x) \cdot \frac{f(x)}{c\,g(x)} \Big/ \int g(x) \frac{f(x)}{c\,g(x)}\,dx$$

$$= f(x) \Big/ \int f(x)\,dx = f(x).$$

x – horizontal $\quad U c\,g(x)$ – vertical

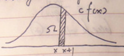
200 A ✓ DATE 23/10/13

57 拒绝模型 & 贝叶斯推理

Rejection Sampling

$X \sim f(x)$

thought experiment

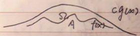


$P(\text{point falls into } A) = P(X \in (x, x+\Delta x))$

$= |A| / |\Omega| = \dfrac{c \cdot f(x)\,\Delta x}{c} = f(x) \cdot \Delta x.$

want $X \sim f(x)$

can $X \sim g(x)$ $\quad c\,g(x) \geq f(x)$



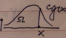
throw into Ω, only accept if point $\in A$.

random point (X, Y) under $c\,g(x)$

$X \sim g(x)$

$[Y \mid X = x] \sim \text{Unif}[0, c\,g(x)]$

$= c\,g(x) \cdot U \rightarrow \text{Unif}[0,1]$.



Algorithm: $X \sim g(x)$

$U \sim \text{Unif}[0,1]$

$Y = c\,g(x)\,U.$

into A ?

$Y \le f(x)$

Return X if $cg(x)\,U \le f(x)$

$U \le \frac{f(x)}{c\,g(x)}$.

otherwise go back to (*)

Acceptance : $P(\text{accept}) = \frac{|A|}{|\Omega|} = \frac{1}{c}$.

Missing data 例3

ask people's incomes

non-ignorable missing

ask 1000 people

☑ ☑ ☒ ... ☐

1 2 3 1000

$X \sim f(x)$ — true population density.

$Y = \begin{cases} 1 & \text{response (observed)} \\ 0 & \text{non-response (missing)}. \end{cases}$

$P(Y=1 \mid X=x) = p(x)$.

$[Y \mid X=x] \sim \text{Ber}(p(x))$.

$f(x \mid 1) = \frac{f(x,1)}{f(1)} = \frac{f(x)\,f(1,x)}{\int f(x) f(1,x)\,dx}$

$= \frac{f(x)\,p(x)}{\int f(x)\,p(x)\,dx}$.

density of observed x.

$w_i = 1 / \rho_i(x)$ $\quad \sum w_i x_i / \sum w_i$.

Algorithm:

$\left\{ \begin{array}{l} X \sim f(x) \\ \text{Accept } X \text{ with } \rho(x). \end{array} \right.$

Example: Bayecian inference 均连续

$\left\{ \begin{array}{l} X \sim N(\mu, \sigma^2). \\ Y \mid X = x \sim N(x, \tau^2) \end{array} \right.$

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

$$= \cdots$$

compromise between prior & data

$$[x \mid Y = y] = N\left(\frac{\mu/\sigma^2 + y/\tau^2}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\sigma^2 + 1/\tau^2} \right).$$

$\sigma^2 \to \infty$ $\qquad N(y, \tau^2)$

$\tau^2 \to \infty$ $\qquad N(\mu, \sigma^2)$

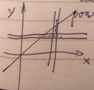
Bayesian Inference

$x_1, \ldots, x_n \sim N(\theta, \tau^2)$.

prior $\quad \theta \sim N(\mu, \sigma^2)$.

data $\quad \bar{x}|_\theta \sim N(\theta, \tau^2/n)$.

posterior $\quad [\theta \mid \bar{x}] \sim N\left(\frac{\frac{\mu}{\sigma^2} + \frac{n\bar{x}}{\tau^2}}{\frac{1}{\sigma^2} + \frac{n}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{n}{\tau^2}} \right)$



Another way:

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

$Y|_{X,\Sigma} \sim N(X, \tau^2)$

$Y = X + \Sigma$

Σ indep. of X

$\Sigma \sim N(0, \tau^2)$.

$$\begin{pmatrix} X \\ \Sigma \end{pmatrix} \xrightarrow{A=\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}} \begin{pmatrix} X \\ Y \end{pmatrix}$$

$$\downarrow$$

$$N\left[\begin{pmatrix} \mu \\ 0 \end{pmatrix} \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix}\right] \rightarrow N\left(A\begin{pmatrix} \mu \\ 0 \end{pmatrix}, A\begin{pmatrix} \sigma^2 & \\ & \tau^2 \end{pmatrix}\right.$$

$[\tilde{X} | \tilde{Y}] \sim N(\Sigma_{X\cancel{Y}} \Sigma_{YY}^{-1} \tilde{Y}, \ \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$

$\tilde{X} \downarrow X - \mu_X$ $\quad \tilde{Y} \downarrow Y - \mu_Y$.

Factor Analysis

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{pmatrix}$ ~~sc~~ scores on 10 sports decathon

$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ $x_1 \to$ strength, $x_2 \to$ speed, $x_3 \to$ endurance $\sim N(0, I)$.

$\underset{(10\times1)}{Y} = \underset{(10\times3)}{L}\underset{(3\times1)}{X} + \Sigma$

$\begin{pmatrix} X \\ Y \end{pmatrix} \rightarrow (X|Y)$.

Y \ X	discrete	continuous
discrete	rare disease	rejection missing data
continuous	mixture classification	normal factor analysis

S8 非线性变换

Non-linear transformation of multivariate R.V.

Recall linear transformation

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \sim f_X(x) = f_X(x_1, \dots, x_n)$$

$$Y = A_{n\times n} X_{n\times 1}$$

$$Y_{n\times 1} \sim f_Y(y) = f_Y(y_1, \dots y_n).$$

$$f_Y(y) = \frac{P(Y \in D(y))}{|D(y)|} = \frac{P(X \in D(x))}{|D(y)|} = \frac{f_X(x)\,|D(x)|}{|D(y)|}$$

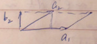
$$= f_X(A^{-1}y) \,/\, |\det(A)|$$

$$A = (\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n)$$

if $\vec{a}_i \perp \vec{a}_j$, $i \neq j$.

$$|\det A| = |\vec{a}_1|\,|\vec{a}_2| \cdots |\vec{a}_n|$$

In general orthogonaliation

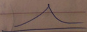


$$A \to B(b_1, b_2, \dots, b_n).$$

$$|\det(A)| = |b_1|\,|b_2| \cdots |b_n|.$$

Independent Component analysis

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad x_i \overset{indep}{\sim} p_i(x)$$



$$Y = AX \qquad W = A^{-1}$$

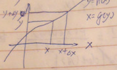
$f_Y(y) = f_X(A^{-1}y) \,/\, |\det A|$

$= f_X(Wy)\,|\det W|$

$\frac{\partial}{\partial W} \log|W| = W^{-1}$

(or $\log|\det(W)|$)

Recall non-linear transformation of one-dim



$$f_Y(y) = \frac{P(Y\in(y+\Delta y))}{\Delta y} = \frac{P(X\in(x+\Delta x))}{\Delta y} = \frac{f_X(x)\,\Delta x}{\Delta y} = f_X(g(y))\,g'(y)$$

$$x \sim f_X(x)\,dx \sim f_X(g(y))\,dg(y)$$

$$\sim \underbrace{f_X(g(y))\,|g'(y)|\,dy}_{f_Y(y)} \sim f_Y(y)\,dy$$

Example: $T \sim Exp(1)$ $\quad T \sim e^{-t}\,dt$

$R = \sqrt{2T}$ $\quad \sim e^{-\frac{r^2}{2}}\,d\frac{r^2}{2} \sim e^{-\frac{r^2}{2}}\cdot r\,dr$

Multivariate

$X_{n\times 1} \sim f_X(x)$

$Y_{n\times 1} \sim h(X) \Leftrightarrow g(Y) = X$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} h_1(x_1 \dots x_n) \\ h_2(x_1 \dots x_n) \\ \vdots \\ h_n(x_1 \dots x_n) \end{pmatrix}$$

$$f_Y(y) = f_X(g(y))\,|\det(g'(y))|$$

$$g'(y) = \begin{pmatrix} & j & \\ i & & \end{pmatrix}_{n\times n} \quad \frac{\partial h_i(x, \dots x_n)}{\partial x_j}$$

✓ Polar method.

$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, I_2)$.

$X, Y \overset{iid}{\sim} N(0,1)$.

$(X,Y) \sim f(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} dx\,dy$.

$\begin{cases} X = R\cos\theta \\ Y = R\sin\theta \end{cases}$

$$\begin{aligned}
(X,Y) &\sim \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} dx\,dy \\
&\sim \frac{1}{2\pi} e^{-\frac{r^2}{2}} r\,dr\,d\theta \\
&\sim \frac{1}{2\pi} e^{-t} dt\,d\theta \quad (t = \tfrac{r^2}{2}) \\
&\sim \frac{1}{2\pi}\, d e^{-t}\, d\theta \\
&\sim \frac{1}{2\pi}\, du\,d\theta \quad u \in [0,1] \\
&\sim du\,dv \quad u \sim \text{Unif}[0,1] \quad v \sim \text{Unif}[0,1]
\end{aligned}$$

$\theta = 2\pi V \quad t = -\log U \quad r = \sqrt{2t} = \sqrt{-2\log U}$.

$X = r\cos\theta \quad Y = r\sin\theta$.

$\begin{cases} X = \sqrt{-2\log U}\cos(2\pi V) \\ Y = \sqrt{-2\log U}\sin(2\pi V) \end{cases}$.

Multivariate things

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

$Y = AX \sim N(A\mu, A\Sigma A^T)$.

$$y = \vec{a}^T x = \langle x, \vec{a} \rangle \sim N(\vec{a}^T \mu, a^T \Sigma a).$$

caculus

$$f(x_1, x_2, \dots x_n) = f(x)$$

$$f'(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}. \qquad f''(x)_{n\times n} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 x_1} & & \\ & \frac{\partial^2 f}{\partial x_i \partial x_j} & \\ & & \ddots \end{pmatrix}_{n\times n}$$

$$f(x) = f(x_0) + f'(x)(x - x_0)^T + \frac{1}{2}(x - x_0)^T f''(x_0)(x - x_0) + \cdots$$

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \qquad f'(x) = \nabla f. \qquad f''(x) = \nabla^2 f = \nabla \nabla^T f.$$

$$g(t) = f(x_0 + t\vec{v}) \qquad |\vec{v}| = 1.$$

$$g(t) = g(0) + g'(0)t + \frac{1}{2} g''(0) t^2 + \cdots$$

$$f(x) = f(x_0) + \langle f'(x_0), \underbrace{\vec{v}t}_{x - x_0} \rangle + \cdots$$

Multivariate $f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix}$ $\qquad f(x) = \begin{pmatrix} f_1(x_1 \cdots x_n) \\ f_2(x_1 \cdots x_n) \\ \cdots \\ f_m(x_1 \cdots x_n) \end{pmatrix}$

$$f'(x)_{m\times n} = \left(\frac{\partial f_i}{\partial x_j}\right)_{m\times n} = f \cdot \nabla^T$$

S9

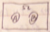
独立性 & 各种模型

Independence

$P(A\cap B) = P(A)P(B)$

$P(A|B) = P(A\cap B)/P(B) = P(A)$.

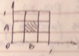
$P(B|A) = \cancel{P(A\cap B)}\ P(B)$.



disjoint

$P(A\cap B)=0$ not indep.

$P(A|B)=0$ $P(B|A)=0$.



$P(A\cap B) = |A||B| = P(A)P(B)$

$A \perp B$.

$(X, Y) \overset{indep}{\sim} Unif[0,1]$.

R.v.s

$X \perp Y$ $p(x,y) = p(x)p(y)$

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

$P(X\in A \;\&\; Y\in B) = P(X\in A)P(Y\in B)$.

Conditional Independence

$A \perp B \mid C$

$P(A\cap B|C) = P(A|C)P(B|C)$.

✓ $P(A|\underset{B\cap C}{B,C}) = P(A|C)$

$X\perp Y \mid Z$ $p(x,y|z) = p(x|z)p(y|z)$

$$p(x,y,z) = p(x,y\mid z)\,p(z) = p(z)\,p(y\mid z)\,p(x\mid z)$$

$$\begin{tikzcd} & z \arrow[dl] \arrow[dr] & \\ x & & y \end{tikzcd} \qquad x \to z \to y$$

Recall: Chain Rule: $p(x,y) = p(x)\,p(y\mid x)$.

Marginalize: $p(y) = \sum_x p(x,y)$.

Conditioning: $p(x\mid y) = \frac{p(x,y)}{p(y)} = \frac{p(x)\,p(y\mid x)}{\sum_x p(x)\,p(y\mid x)}$.

$p(x\mid y) = \underbrace{\frac{p(x,y)}{p(y)}}_{\text{normalizing constant}} \overset{\text{fix } y}{\propto} p(x,y)$ as function of x.

$\sum_x p(x\mid y) = 1$.

Some joint distributions

Multivariate Gaussian

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \sim N(0,\Sigma).$$

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}x^T\Sigma^{-1}x}$$

$$\propto e^{-\frac{1}{2}x^T A x} \qquad A = \Sigma^{-1} = (a_{ij})_{n\times n}.$$

$$= e^{-\frac{1}{2}\sum_{ij} a_{ij}x_i x_j}$$

a pair of $i \neq j$ $\quad a_{ij} = 0.$

$x_i \perp x_j \mid$ other variables

Gene regulation network

Ising model

one-dim $x_1, x_2, \dots, x_n$.

$x_i \in \{+1, -1\}$.

$$P(x) = \frac{1}{Z_{(a,b)}} e^{\sum_{i=1}^{n} a x_i + \sum_{i=1}^{n-1} b x_i x_{i+1}}$$

$$Z = \sum_{x} e^{\sum_{i=1}^{n} a x_i + \sum_{i=1}^{n-1} b x_i x_{i+1}}$$ partition function

2^n values

$$\sum_x p(x) = 1.$$

Gibbs distribution

Boltzmann

Exponential family model

~~Mrictor~~ Markov random field

undirected graphical model

Markov (conditional indep)

$n = \cancel{?}\ 5$

$$p(X) = \frac{1}{Z} e^{\sum_{i=1}^{5} a x_i + \sum_{i=1}^{4} b x_i x_{i+1}}$$

$$p(x_3 \mid x_1, x_2, x_4, x_5) \propto e^{a x_3 + b x_2 x_3 + b x_3 x_4}$$

$$e^{x_3 (a + b x_2 + b x_4)}.$$

$$= p(x_3 \mid x_2, x_4).$$

neighbors of site 3

$$p(x_3 = +1) \propto e^{a + b x_2 + b x_4} \Big/ e^{(\ )} + e^{-(\ )}$$

$$p(x_3 = -1) \propto e^{-a - b x_2 - b x_4} \Big/ e^{(\ )} + e^{-(\ )}.$$

$$p(x = +1) \propto \frac{1}{1 + e^{-2(\ )}}$$

auto logistic regression

General form

$$p(x) = \frac{1}{Z(\lambda)} e^{\lambda \phi(x)}.$$

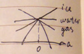
$$\frac{\partial}{\partial \lambda} \log Z(\lambda) = \frac{1}{Z(\lambda)} Z'(\lambda) = \frac{1}{Z(\lambda)} \frac{\partial}{\partial \lambda} \sum e^{\lambda \phi(x)}$$

$$= \frac{1}{Z(\lambda)} \sum e^{\lambda \phi(x)} \phi(x).$$

$$= \sum \phi(x) p(x) = E(\phi(x)).$$

phase transition

$$\frac{\partial}{\partial a} \log Z(a, \underset{\text{fixed}}{b}) = E(\textstyle\sum_i x_i).$$



Boltzmann machine

$x_i, y_i \in \{0, 1\}.$

$$p(x, y) = \frac{1}{Z} e^{\sum_i a_i x_i + \sum_j b_j y_j + \sum_{i,j} w_{ij} x_i y_j}$$

$$p(y|x) \propto e^{\sum_j b_j y_j + \sum_j y_j \sum_i w_{ij} x_i}$$

$$= e^{\sum_j y_j (\sum_i w_{ij} x_i + b_j)}$$

$$= \prod_{j=1}^{n} e^{y_j (\sum_i w_{ij} x_i + b_j)}$$

y_j are indep. given $x_1 \cdots x_n$

$$p(y_j = 1) \propto e^{\sum w_{ij} x_i + b_j}$$

$$p(y_j = 0) \propto 1.$$

$$p(y_j = 1) = \frac{e^{\cdot}}{1 + e^{\cdot}}.$$

logistic regression.

$$p(x) = \sum_{y \in 2^n} p(x, y) \propto \sum_y e^{\sum a_i x_i + \sum_j y_j (\sum_i w_{ij} x_i + b_j)}$$

$$= e^{\sum a_i x_i} \sum_{y_1 \cdots y_n} e^{\sum_j y_j (\quad)}$$

$$\cancel{p(y|x) = \frac{p(x,y)}{p(x)}}.$$

$$= e^{\sum a_i x_i} \prod_{j=1}^{n} \sum_{y_j} e^{y_j(\ )}$$

$$(1 + e^{\sum_j w_{ij} x_i + b_j}).$$

$$= e^{\sum a_i x_i} \prod_{j=1}^{n} (1 + e^{\sum_j w_{ij} x_i + b_j}).$$

S10 条件独立 & Causal effect & 亚洲模型

Conditional indep.

$$\begin{tikzcd} & Z \arrow[dl] \arrow[dr] & \\ X & & Y \end{tikzcd}$$

$$X \perp Y \mid Z$$

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z)$$

Example: health vs. smoking habit

$X = \begin{cases} 1 & \text{cigarett} \\ 0 & \text{pipe} \end{cases}$ $\quad Y = \begin{cases} 1 & \text{healthy} \\ 0 & \text{not.} \end{cases}$

$$P(Y=1 \mid X=1) = P(Y=1 \mid X=0).$$

$Z = \begin{cases} 1 & \text{young} \\ 0 & \text{old} \end{cases}$

It's possible $X \perp Y \mid Z$.

$$p(Z=0) = p(Z=1) = 1/2.$$

$p(X=1 \mid Z=1) = 0.8 \quad p(X=0 \mid Z=1) = 0.2$

$p(X=1 \mid Z=0) = 0.2 \quad p(X=0 \mid Z=0) = 0.8.$

$P(Y=1 \mid Z=1) = 0.9 \quad P(Y=0 \mid Z=\phi) = 0.1$

$P(Y=1 \mid Z=0) = 0.6 \quad P(Y=0 \mid Z=0) = 0.4.$

$$p(x, y) = \sum_z p(x, y, z) = \sum_z p(z)\, p(x, y \mid z)$$

$$= \sum_z p(z)\, p(x \mid z)\, p(y \mid z).$$

$P(1,1) = \frac{1}{2} \times 0.8 \times 0.9 + \frac{1}{2} \times 0.2 \times 0.6 = 0.42.$

$P(\phi, 0) = 0.08.$

$P(0,1) = 0.33$

$P(1,1) = 0.17.$

$P(Y=1 \mid X=1) = 0.84$ $\quad P(Y=1 \mid X=0) = 0.66$

How to define causal relationship ?

A population

Ω	cig	pipe	Causal Effect
1	✓ 80	? 60	20
2	? 70	✓ 60	10
- - -	- - -	- - -	
M	✓ 60	? 70	-10

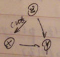
counter factual.

treatment assignment

(1) random assignment.

(2) depends on outcome.

observational study. May be affected by other factors



$P(Y=1 \mid X=1) - P(Y=1 \mid X=0)$. ← observed out.

$P(Y=1 \mid X \leftarrow 1) - P(Y=1 \mid X \leftarrow 0)$ ← assigned out.

$Z \leftarrow f_Z(\varepsilon_Z)$

$X \leftarrow f_X(Z, \varepsilon_X)$ $\quad \varepsilon_X, \varepsilon_Y, \varepsilon_Z$ indep.

$Y \leftarrow f_Y(X, Z, \varepsilon_Y)$

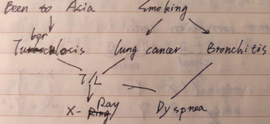
$\rightarrow$

$$Z \leftarrow f_Z(\varepsilon_Z)$$
$$X \leftarrow \tilde{f}_X(\tilde{\varepsilon}_X)$$
$$Y \leftarrow f_Y(X, Z, \varepsilon_Y).$$

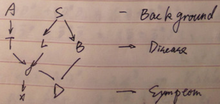
$\Rightarrow$

$$Z \leftarrow f_Z(\varepsilon_Z)$$
$$X \leftarrow 1.$$
$$Y \leftarrow f_Y(X, Z, \varepsilon_Y).$$

☆ Example : Asian Example



Directed Acylic Graph (DAG).



$s = 1 \quad x = 1, \quad D = 0.$

$P(T=1) = ?$

$P(T=1 \mid s=1,\ x=1,\ D=0).$

Joint prob.

$P(a, s, t, l, b, o, x, d)$

$= P(a)\,P(s)\,P(t \mid a)\,P(l \mid s)\,P(b \mid s)$

$\quad p(o \mid t, l)\,p(x \mid o)\,p(d \mid o, b).$

Belief Propogation algorithm.

Error correction code

$\underbrace{\square \cdots \square}_{\text{signal}} \quad \underbrace{\square \cdots \square}_{\text{extra}}$

original $\quad X \quad Y$

noisy $\quad \tilde{X} \quad \tilde{Y}.$

$P(X \mid \tilde{X}, \tilde{Y}).$

$$\begin{array}{ccc} X & - & Y \\ | & & | \\ \tilde{X} & & \tilde{Y}. \end{array}$$

§11 条件独立 11.6.13

Conditional Indep.

$$\begin{tikzcd} & Z \arrow[dl] \arrow[dr] & \\ X & & Y \end{tikzcd}$$

Z: age. X: smoking habit. Y: health.

$X \perp Y \mid Z$. (within each age group).

X and Y are dependent marginally (whole population)
sum aggregate.

Continuous

$$(X, Y \mid Z = z) \sim N\left(\begin{pmatrix} z \\ z \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right).$$

$X \perp Y \mid Z$.

$X \mid Z = z \sim N(z, \sigma^2)$.

$Y \mid Z = z \sim N(z, \sigma^2)$.

$$f(x, y \mid z) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-z)^2 + (y-z)^2}{2\sigma^2}}$$

$Z \sim N(\mu, \tau^2)$. $f(z)$.

$$f(x, y) = \int f(x, y \mid z) f(z)\, dz.$$

独立时

$$Cov(X, Y) = E(cov(X, Y \mid Z)) + Cov(E(X \mid Z), E(Y \mid Z))$$
$$= 0 + Cov(Z, Z) = \tau^2.$$

$$Var(X) = E(Var(X \mid Z)) + Var(E(X \mid Z))$$
$$= \sigma^2 + \tau^2$$

$$Corr(X, Y) = \frac{\tau^2}{\sigma^2 + \tau^2}.$$

$E(X) = E(Y) = 0$ otherwise $X \leftarrow X - \mu_X$

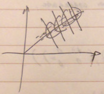
$Y \leftarrow Y - \mu_Y$

石例情况:

$Cov(X, Y) = E(XY) = E(E(XY|Z))$.

✓ $E(cov(X,Y|Z)) = E(E(X - E(X|Z))(Y - E(Y|Z))|Z)$
$= E(XY) - E[E(X|Z)E(Y|Z)]$

✓ $cov(E(X|Z), E(Y|Z)) = E(E(X|Z)E(Y|Z))$
$- E(E(X|Z))E(E(Y|Z))$.
$= E(E(X|Z)E(Y|Z)) - EX\,EY$.



Intervention

Counterfactual / potential outcomes / manipulation

Causality

Z: treatment   background

population

Ω	$Z=0$	$Z=1$	X
1	Y_{10}	Y_{11}	x_1
2	Y_{20}	Y_{21}	x_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
M	Y_{M0}	Y_{M0}	x_M

causal effect for w:
$Y_{w1} - Y_{w0}$.

causal effect for the whole population:
$\frac{1}{M}\sum_w Y_{w1} - \frac{1}{M}\sum_w Y_{w0}$.

treatment assignment

$Z_w \lessgtr {}^{1}_{0}$

Even if $Y_{w1} = Y_{w0}$ $\forall w$.

can find Z_w.

$$\frac{1}{\Sigma Z_w}\sum_w Y_{w1} Z_w - \frac{1}{\Sigma(1-Z_w)}\sum_w Y_{w0}(1-Z_w).$$

average of observed #'s.

observational study

Z_w depends on X_w. (which may affect Y_{w1}, Y_{w0}).

Experiment

Z_w: randomized.

indep. of any Z_w.

Causal inference estimate Δ based on observed Y_{w1}, Y_{w0}.

Graph / Structural equation.

$$\begin{tikzcd} & X \text{ age.} \arrow[dl] \arrow[d] \\ \text{smoking } Z \arrow[r] & Y \text{ health.} \end{tikzcd} \xrightarrow{\text{survey}} \begin{tikzcd} & X \arrow[d] \\ Z=z \arrow[r] & Y \end{tikzcd}$$

$P(Y \mid Z=z)$. $\downarrow$ actively assigned as

$P(Y \mid Z=z)$. $\downarrow$ passively observed as mechanism

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ Z \arrow[rr] & & Y \end{tikzcd}$$

$X \leftarrow f_X(\varepsilon_X)$

$Z \leftarrow f_Z(X, \varepsilon_Z)$

$Y \leftarrow f_Y(X, Z, \varepsilon_Y)$

ε's indep.

$P(Y \mid Z = z)$.

counterfactual

$$\begin{tikzcd} & X \arrow[dr] & \\ Z \leftarrow z \arrow[rr] & & Y \end{tikzcd}$$

$X \leftarrow f_X(\varepsilon_X)$

$Z \leftarrow z$. (can do it $\forall z$).

$Y \leftarrow f_Y(X, Z, \varepsilon_Y)$

$P(Y \mid Z \leftarrow z)$.

$$p(y \mid z) = \int p(y \mid z, x)\, p(x \mid z)\, dx.$$

$$p(y \mid z \leftarrow z) = \int p(y \mid z, x)\, p(x)\, dx.$$

$p(y \mid do(z)) = p(y \mid z)$ if $p(x \mid z) = p(x)$

$X \perp Z$.

(1) Z is indep of X.

random assignment experiment.

(2) X is indep. of Z.

subpopulation of $Z=0$ is the same as - - ~~ - - $Z=1$.

link between potential outcome & structure equation

$w = (\varepsilon_X, \varepsilon_Y, \varepsilon_Z)$

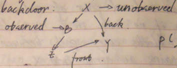
	$Z=0$	$Z=1$
w	Y_{w0}	Y_{w1}

$\Leftarrow$ $Z \leftarrow 1$

$Z \leftarrow 0$.

X may not be fully observed

backdoor:



$p(y \mid do(z))$ if $B = b$

$P(Y \mid \cancel{b}\, do(z), b) = P(y \mid z, b)$.

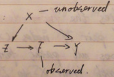
$p(y \mid do(z)) = \int p(y \mid do(z), b)\, \overset{p(b)}{p(b \mid do(z))}\, db$

$p(y) = \int p(y \mid b)\, p(b)\, db$.

✓ $P(Y \mid do(z)) = \int p(y \mid z, b)\, p(b)\, db$.

if $B = b$. adjust.

front:



$$p(y \mid do(z)) = \int p(y \mid do(z), f)\, p(f \mid do(z))\, df$$
$$= \int p(y \mid do(f))\, p(f \mid z)\, df.$$
$$= \int \left(\int p(y \mid f, \tilde{z})\, p(\tilde{z})\, d\tilde{z} \right) p(f \mid z)\, df.$$

S12   马[illegible]

Rubin's potential outcomes

Ω	$z=1$	$z=0$	X
1	v		
2	v		
$\vdots$			
M			

distribution of Y's in column $z=1$ $P(Y \mid z=1)$.

Pearl's diagram

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ Z \arrow[rr] & & Y \end{tikzcd}$$

(1) $X \leftarrow f_X(\varepsilon_X)$

(2) $Z \leftarrow f_Z(X, \varepsilon_Z)$.

(3) $Y \leftarrow f_Y(X, Z, \varepsilon_Y)$.

$$\begin{tikzcd} & X \arrow[dr] & \\ Z \arrow[rr] & & Y \end{tikzcd}$$

(1)

(2) $Z \leftarrow 1$ or $Z \leftarrow 0$

(3)

$P(Y \mid z \leftarrow 1)$ do (1)

distribution of "all" Y's in column $z=1$

define target / estimated.

Missing data mechanism.

Obs	X	Y	M
1			0
2			$\vdots$
$\vdots$			
n			

Missing completely at random (MCAR)

$M \perp (X, Y)$

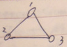
Missing at random

$M \perp Y \mid X$.

non-ignorable missing. M depends on Y, X

Markov Chains

random walk



At each step, with half prob stay, with prob $\frac{1}{4}$, go to one of the two neighbors

state space $\mathcal{X} = \{1, 2, 3\}$.

x_t = state at time t.

$x_0 = 1$ (or random state).

$P(x_{t+1} = y \mid x_t = x, x_{t-1}, \dots) = P(x_{t+1}=y \mid x_t)$

$= K(x, y)$

x \ y	1	2	3
1	1/2	1/4	1/4
2	1/4	1/2	1/4
3	1/4	1/4	1/2

graph

$x_0 \to x_1 \to \dots \to x_t \to x_{t+1}$

$x_{t+1} = f(x_t, \varepsilon_t)$. structural eqn, computer code

ε_t indep.

$p^{(t)}(x) = P(x_t = x)$

$p^{(t+1)}(y) = P(x_{t+1} = y) = \sum_x P(x_{t+1} = y \cap x_t = x)$

$= \sum_x P(x_t = x) P(x_{t+1} = y \mid x_t = x)$.

$= \sum_x p^{(t)}(x) K(x, y)$.

$p^{(t)} = (p^{(t)}(1), p^{(t)}(2), p^{(t)}(3))$

$p^{(t+1)} = \dots$

$K^{(2)}(x, y) = P(x_{t+2} = y \mid x_t = x)$.

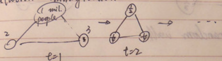
$= \sum_{z} p(x_{t+2}=y \cap x_{t+1}=z \mid x_t = x)$.

$= \sum_{z} p(x_{t+1}=z \mid x_t=x)\, p(x_{t+2}=y \mid x_{t+1}=z, x_t=x)$.

$= \sum_{z} K(x,z)\, K(z,y)$

$K^{(2)} = K^2$. $\qquad p^{(t)} = p^{(0)} K^t$.

Population immigration.



$p^{(t)} \xrightarrow[\infty]{t} \pi$ stationary distribution.

Markov Chain $(\mathcal{X}, K, \pi)$

~~Rere~~ Reversible chain (detailed balance)

$\forall x, y$

$\pi(x)\, K(x,y) = \pi(y)\, K(y,x)$ detailed balance

$\#$ of people $x \to y$ $\qquad$ $\#$ of people $y \to x$ $\qquad$ overall balance $\pi(y) = \sum_x \pi(x) K(x,y)$

Given π, what is K?

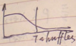
MCMC ← - Metropolis Alg. (Metropolis, Rosenbluth, R. Teller, T 1950)

Gibbs Sampler.

Markov Chain, Monte Carlo

Real life example of MCMC: card shuffling Unif (52! permutations)

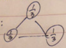
Diaconis $\|P^{(t)} - \pi\|$



Target distribution

base chain: random walk

π:



π:

K: random walk.

target : $\pi(x)$

base chain $B(x,y)$ $\qquad \pi(x)B(x,y) \quad \pi(y)B(y,x)$

$\pi(x)B(x,y) \geq \pi(y)B(y,x)$ $\dashrightarrow$ Propb. of accept $= \dfrac{\pi(y)B(y,x)}{\pi(x)B(x,y)}$

$$M(x,y) = \underbrace{B(x,y)}_{\text{purposal}} \min\Big(1, \underbrace{\frac{\pi(y)B(y,x)}{\pi(x)B(x,y)}}_{\text{acceptance.}}\Big)$$
$(y \neq x)$

$\pi(x)M(x,y) = \pi(y)M(y,x) = \min(\pi(x)B(x,y), \pi(y)B(y,x))$

Only need to know $\pi(x)$ in proportionality.

Original $B(x,y) = B(y,x)$

accept prob. $= \min\left(1, \frac{\pi(y)}{\pi(x)}\right)$

$\pi(x) = \frac{1}{Z} e^{-\varepsilon(x)/T}$.

high low

acceptance prob. $= \min\left(1, e^{\frac{\varepsilon(x)-\varepsilon(y)}{T}}\right)$.

NOTES 2010 A OM OT OW OT OF OS OS DATE 11/13/13

513 Gibbs Sampling & 连续时间

MCMC

Gibbs sampler

$(X, Y) \sim f(x, y)$

$[X|Y] \sim f(x|y)$

$[Y|X] \sim f(y|x)$.

Start from $(X^{(0)}, Y^{(0)})$

Iterate:

$X^{(t+1)} \sim f(x | y^{(t)})$

$Y^{(t+1)} \sim f(y | x^{(t+1)})$.

Example: $f(x, y) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$

$[X | Y = y] \sim N(\rho y, 1-\rho^2)$.

$[Y | X = x] \sim N(\rho x, 1-\rho^2)$.

Start $X^{(0)} = 0$, $Y^{(0)} = 0$.

Iterate: $X^{(t+1)} \sim N(\rho Y^{(t)}, 1-\rho^2)$.

$Y^{(t+1)} \sim N(\rho X^{(t+1)}, 1-\rho^2)$.

$\xrightarrow[\infty]{t} (X^{(t)}, Y^{(t)}) \sim p^{(t)}(x, y) \longrightarrow f(x, y)$

– stationary distr.

If $(X^{(t)}, Y^{(t)}) \sim f(x, y)$. 不动点原则.

$Y^{(t)} \sim f(y)$

$[X^{(t+1)} | Y^{(t)} = y, X^{(t)}] \sim f(x|y)$

$[X^{(t+1)}, Y^{(t)}] \sim f(y) f(x|y) = f(x, y)$

$X^{(t+1)} \sim f(x)$

$[Y^{(t+1)} \mid X^{(t+1)} = x, Y^{(t)}] \sim f(y \mid x)$

$[X^{(t+1)}, Y^{(t+1)}] \sim f(x) f(y \mid x) = f(x, y)$

Run Gibbs sampler for 10,000 iterations

$(X^{(0)}, Y^{(0)}), (X^{(1)}, Y^{(1)}) \dots (X^{(10000)}, Y^{(10000)})$

throw away first 5,000 iterations

$$I = \int h(x,y) f(x,y)\, dx\, dy$$

$$\approx \frac{1}{5000} \sum_{t=5001}^{10000} h(X^{(t)}, Y^{(t)}).$$

In general, $\vec{X} = (X_1, \dots X_n)^T \sim f(x_1, \dots x_n) = f(\vec{x})$

Iterate for $(i = 1$ to $n)$

sample x_i given X_{-i}

current values of all the other components except i

Gibbs distribution

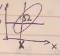
Ising model

$x_1 \quad x_2 \quad \cdots \quad x_n$

$$P(x) = \frac{1}{Z(\beta)} \exp\left(\sum_{i=1}^{n-1} \beta x_i x_{i+1}\right) \qquad x_i \in \{-1, +1\}$$

$$P(x_i \mid X_{-i}) = P(x_i \mid x_{i-1}, x_{i+1}). \qquad \text{Markov}$$

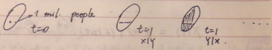
Population immigration

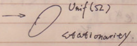


$f(x,y) = \text{Unif}(\Omega)$

$f(x|y) =$ Uniform on horizontal line segment

$f(y|x) = \ldots \ldots$ vertical $\ldots$





Why Monte Carlo ?

$I = \int h(x) f(x)\, dx.$

avoid curse of dimensionality

X is high-dim. deterministic methods break down.

$x_1, x_2, \ldots x_n \overset{iid}{\sim} \cancel{f}$

$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} h(x_i) \qquad x \sim f(x).$

$Var(\hat{I}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(h(x_i))$

$= \frac{1}{n^2} n \cdot Var(h(x)) = \frac{Var(h(x))}{n}.$

减小方差

Continuous time process

Interest rate. r

$0 \quad \delta t \quad 2\delta t \quad \cdots \quad t \quad \to t$

$X(0) = x. \qquad X(t) = ?$

NOTES

$X(t+\Delta t) = X(t)\,(1+r\Delta t)$

$\frac{dX(t)}{dt} = r X(t)$.

$X(\Delta t) = X(0) * (1+r\Delta t)$

$X(2\Delta t) = X(\Delta t)\,(1+r\Delta t) = X(0)(1+r\Delta t)^2$

$\cdots\cdots$

$X(t) = X(0)\,(1+r\Delta t)^{t/\Delta t} \xrightarrow{\Delta t \to 0} x e^{rt}$

Symbolically $1+\alpha \doteq e^{\alpha}$

$X(t) = X(0)\,(1+r\Delta t)^{t/\Delta t}$

$\to x e^{rt}$

Poisson process

$0 \quad \Delta t \quad 2\Delta t \quad \cdots \quad t$

Flip a coin within each period indep.

$P(\text{head in } (t, t+\Delta t)) = \lambda \Delta t$ frequency.

T = time until 1^{st} head

$P(T>t) = P(\text{all tails until } t)$

$= (1-\lambda\Delta t)^{\frac{t}{\Delta t}} \longrightarrow e^{-\lambda t}$

$P(T \in (t, t+\Delta t)) = (1-\lambda\Delta t)^{\frac{t}{\Delta t}}\,\lambda\Delta t$

$$f(t) = \frac{P(T \in (t, t+\Delta t))}{\Delta t}$$
$$= \lambda e^{-\lambda t}.$$

λ not constant. hazard rate.

$$P(\text{head in } (t, t+\Delta t)) = \lambda(t)\Delta t$$

survival analysis

$$P(T>t) = (1-\lambda(0)\Delta t)(1-\lambda(\Delta t)\Delta t)\cdots(1-\lambda(t-\Delta t)\Delta t)$$
$$= \prod_{s=0}^{t}(1-\lambda(s)\Delta t) \xrightarrow{\Delta t \to 0} \prod_{s=0}^{t} e^{-\lambda(s)\Delta t}$$
$$= e^{-\sum_0^t \lambda(s)\Delta t} \longrightarrow e^{-\int \lambda(t)\,dt}$$

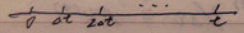
$$P(T>t) = e^{-\int \lambda(t)\,dt}.$$

Markov Chain.

$P(\text{stay}) = \frac{1}{2}$

$P(\text{move to each neighbor}) = \frac{1}{4}.$

Markov jump process



$$P(X_{t+\Delta t} = j \mid X_t = i) = a_{ij}\Delta t$$
$$j \neq i.$$

Among people in state i at time t.

# of people moving to j at one day (Δt).

$$K^{(\Delta t)}(i,i) = P(X_{t+\Delta t} = i \mid X_t = i) = 1 - \underbrace{\sum_{j\neq i} a_{ij}}_{a_{ii}}\Delta t.$$

NOTES

$$K^{(\Delta t)} = \begin{pmatrix} a_{11}\Delta t+1 & a_{12}\Delta t & a_{13}\Delta t \\ a_{21}\Delta t & a_{22}\Delta t+1 & a_{23}\Delta t \\ a_{31}\Delta t & a_{32}\Delta t & a_{33}\Delta t+1 \end{pmatrix}$$

$$= I + A\Delta t$$

$K^{(t)}(i,j) = P(X_t = j \mid X_0 = i)$

$K^{(t)} = (K^{(\Delta t)})^{t/\Delta t}$

(Recall discrete time)

$K^{(n)} = K^n$

$$K^{(t)} = (I + A\Delta t)^{t/\Delta t} \to e^{At}$$

$0 \quad t \quad t+\Delta t$

last step analysis

$K^{(t+\Delta t)} = K^{(t)} K^{(\Delta t)} = K^{(t)}(I + A\Delta t)$

$\frac{K^{(t+\Delta t)} - K^{(t)}}{\Delta t} = K^{(t)} A.$

$\xrightarrow{\Delta t \to 0} \frac{dK^{(t)}}{dt} = K^{(t)} A$ forward equ.

Kolmogorov equ.

$0 \quad \Delta t \quad t+\Delta t$

$K^{(t+\Delta t)} = K^{(\Delta t)} K^{(t)} = (I + A\Delta t) K^{(t)}$

$(K^{(t+\Delta t)} - K^{(t)})/\Delta t = A K^{(t)}$

$\xrightarrow{\Delta t \to 0} \frac{dK^{(t)}}{dt} = A K^{(t)}$ backward equ.

$K_{3\times 3}$ — noun: transition prob.
— verb: act on sth. operator

$g_{3\times 1} = K_{3\times 3}\, h_{3\times 1}$ — backward.

$f_{1\times 3} = P_{1\times 3}\, K_{3\times 3}$ — forward.

NOTES Doo A DATE / /

S14

Markov Transition

$K(x,y)$ K_{ij} $K_{3\times3}$

$P(x_{t+1}=y \mid x_t = x)$.

$\Downarrow$

$P(x_{t+2}=y \mid x_t = x) = K^{(2)}(x,y)$.

$= \sum_z P(x_{t+2}=y, x_{t+1}=z \mid x_t = x)$

$= \sum_z P(x_{t+1}=z \mid x_t = x) P(x_{t+2}=y \mid x_{t+1}=z)$

$= \sum K(x,z) K(z,y)$.

Continuous time

modeling one period. $0 \;\; \Delta t \;\; 2\Delta t \;\; \cdots \;\; t \rightarrow$

rate : $K^{(\Delta t)}(x,y) = P(x_{t+\Delta t}=y \mid x_t = x)$.

$= a(x,y)\Delta t$

$K_{ij}^{(\Delta t)} = P(x_{t+\Delta t}=j \mid x_t = i)$

$= a_{ij}\Delta t$ $i \neq j$.

$K^{(\Delta t)} = I + A\Delta t$

$A = (a_{xy})_{3\times3}$ = (jumping - into/out of - rate) x

$a_{xx} = -\sum_{y\neq x} a_{xy}$.

Cumulative consequence

$K^{(t)}(x,y) = P(x_t = y \mid x_0 = x)$.

$K^{(t)} = (K^{(\Delta t)})^{t/\Delta t} = (I + A\Delta t)^{t/\Delta t}$

$$\xrightarrow[0]{\Delta t} \quad e^{At} = \sum_{n=0}^{\infty} \frac{(At)^n}{n!}$$

$$K^{(t)} = e^{At}.$$

Semi-group: $K^{(t+s)} = K^{(t)} K^{(s)}$.

A: generator.

$$K^{(t+\Delta t)} = K^{(t)} K^{(\Delta t)} = K^{(\Delta t)} K^{(t)}.$$

$$= K^{(t)} (I + A\Delta t) = (I + A\Delta t) K^{(t)}.$$

$$dK^{(t)}/dt = \underset{\text{forward}}{K^{(t)} A} = \underset{\text{backward}}{A K^{(t)}}.$$

$$\frac{dK^{(t)}(x,y)}{dt} = \sum_z K^{(t)}(x,z)\, A(z,y).$$

$$= \sum_z A(x,z)\, K^{(t)}(z,y).$$

meaning of $K^{(s)} = P(X_{t+s} = y \mid X_t = x)$.

noun: $K^{(s)}$ = (transition probs).

verb: $h = (h(x)) = \begin{pmatrix} h(1) \\ h(2) \\ h(3) \end{pmatrix}_{3\times 1}$

$$K^{(s)}_{3\times 3} h_{3\times 1} = g_{3\times 1}$$

$$(K^{(s)} h)(x) = g(x)$$

$$g(x) = \sum_y K^{(s)}(x,y)\, h(y) = \sum_y P(X_s = y \mid X_0 = x)\, h(y)$$

$$= E(h(X_s) \mid X_0 = x)$$

$$p^{(0)}(x) = P(X_0 = x) \qquad p^{(t)}(x) = P(X_t = x)$$

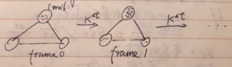
$p^{(t)} = (p^{(t)}(1),\ p^{(t)}(2),\ p^{(t)}(3))$.

$\cancel{p^{(t)} \times K^{(s)} = \Omega}$.

$p^{(t)} K^{(s)} = p^{(t+s)}$.

$$p^{(t+s)}(y) = P(X_{t+s}=y) = \sum_x P(X_{t+s}=y \,\&\, X_t = x)$$
$$= \sum_x P(X_t = x) P(X_{t+s}=y \mid X_t = x).$$
$$= \sum_x p^{(t)}(x) K^{(s)}(x,y).$$

Population immigration



$p^{(t)} K^{(s)} = p^{(t+s)}$ $\quad g_t = K^{(t)} h$.

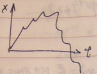
What if $p^{(t)} = \begin{pmatrix} p^{(t)}(1) \\ p^{(t)}(2) \\ p^{(t)}(3) \end{pmatrix}$

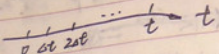
$K^{(s)T} p^{(t)} = p^{(t+s)}$.

Diffusion / Brownian motion.

Random walk on real line

X_t :





modeling: $X_{t+\Delta t} = X_t + \delta \varepsilon_t \sqrt{\Delta t}$

$\varepsilon_t \overset{iid}{\sim} E(\varepsilon_t) = 0$ Var

(x) { If $X_{t+\Delta t} = X_t + \delta \varepsilon_t \Delta t$

$\frac{X_{t+\Delta t} - X_t}{\Delta t} = \delta \varepsilon_t$ = velocity.

$\Delta t \downarrow \quad v \uparrow$

If $X_{t+\Delta t} = X_t + \delta \varepsilon_t \sqrt{\Delta t}$

$\frac{X_{t+\Delta t} - X_t}{\Delta t} = \delta \varepsilon_t / \sqrt{\Delta t} \xrightarrow[0]{\Delta t} \infty$

$X_t = \sum_t \delta \varepsilon_t \sqrt{\Delta t}$.

$E(X_t) = \sum_t \delta E(\varepsilon_t) \sqrt{\Delta t} = 0$.

$Var(X_t) = \sum_t \delta^2 \Delta t = \delta^2 t$.

δ ~ size of

dust movement: $\Delta t \sim 10^{-2}$ s.

math $\Delta t \to 0$ $X(t)$ continuous path

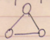
non-differentiable.

$\Omega = \{$all paths$\}$.

physics: $\Delta t = 10^{-12}$

Individual collisions

Jump process



$$K^{(\Delta t)}(x,y) = P(X_{t+\Delta t}=y \mid X_t = x) = a_{xy}\Delta t \quad (x \neq y)$$

$$K^{(\Delta t)}(x,x) = 1 - \sum_{y\neq x} a_{xy}\Delta t$$

$$K^{(\Delta t)} = I + A\Delta t \Rightarrow A = \frac{K^{(\Delta t)} - I}{\Delta t}$$

$$K^{(t)} = \left(K^{(\Delta t)}\right)^{\frac{t}{\Delta t}} = (I + A\Delta t)^{t/\Delta t} \to e^{tA}$$

First step analysis

$$K^{(t+\Delta t)} = K^{(\Delta t)} K^{(t)} = (I + A\Delta t) K^{(t)}$$

$$\frac{\partial K^{(t)}}{\partial t} = A K^{(t)} \quad \to \text{Backward}$$

Last step analysis

$$K^{(t+\Delta t)} = K^{(t)} K^{(\Delta t)} = K^{(t)}(I + A\Delta t)$$

$$\frac{\partial K^{(t)}}{\partial t} = K^{(t)} A \quad \to \text{Forward.}$$

noun: K transition prob.

A transition rate

verb: K smoothing, diffusing

A smoothing rate, diffusing rate

smoothing

$$K^{(t)} h = g$$

$$g(x) = \sum_y K^{(t)}(x,y)\, h(y) = \sum_y p(X_t = y \mid X_0 = x)\, h(y)$$

$$= E(h(X_t) \mid X_0 = x) = \text{local average around } x$$

$$Ah = \frac{(K^{\Delta t} - I)h}{\Delta t} = g$$

$$g(x) = \frac{E(h(X_{\Delta t}) \mid X_0 = x) - h(x)}{\Delta t}$$

$$\partial K^{(t)} h / \partial t = A K^{(t)} h. \Rightarrow \frac{\partial u(t,x)}{\partial t} = A\, u(t,x).$$

$$u(t,x) = (K^{(t)} h)(x). \qquad \frac{\partial u}{\partial t} = Au.$$

$$\Downarrow$$

$$u(t) = e^{tA} h.$$

Diffusion process

Brown motion.

$$X_{t+\Delta t} = X_t + \sigma \varepsilon_t \sqrt{\Delta t}. \qquad E(\varepsilon_t) = 0$$

$$\varepsilon_t \overset{iid}{\sim} f(\varepsilon) \qquad Var(\varepsilon_t) = 1.$$

Simplify: $\varepsilon_t = \begin{cases} +1 & prob = \frac{1}{2} \\ -1 & prob = \frac{1}{2} \end{cases}$

step size $\sigma \sqrt{\Delta t} = \Delta x$.

$-2\Delta x \quad -\Delta x \quad 0 \quad \Delta x \quad 2\Delta x \qquad x$

$$K^{(\Delta t)} = \begin{array}{c} \quad y \rightarrow \\ x \downarrow \begin{pmatrix} \ddots & & & \\ \frac{1}{2} & \_ & \frac{1}{2} & \\ & & & \ddots \end{pmatrix} \end{array} \qquad A = \frac{K^{(\Delta t)} - I}{\Delta t}$$

$$A = \frac{1}{\Delta t} \begin{pmatrix} \frac{1}{2} & -1 & \frac{1}{2} & & & \\ & \frac{1}{2} & -1 & \frac{1}{2} & & \\ & & & \ddots & & \\ & & & \frac{1}{2} & -1 & \frac{1}{2} \end{pmatrix}$$

$$g = Ah \qquad x\big\downarrow \begin{pmatrix} g(x) \end{pmatrix}_{g} = \frac{\delta^2}{2}\frac{1}{\Delta x^2}\begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & & \ddots & \end{pmatrix}\begin{pmatrix} \\ \end{pmatrix}_{h} \big\downarrow x$$

$$g(x) = \frac{\delta^2}{2}\frac{1}{\Delta x^2}\left[h(x-\Delta x) - 2h(x) + h(x+\Delta x)\right]$$

$$= \frac{\delta^2}{2}\frac{\frac{h(x+\Delta x)-h(x)}{\Delta x} - \frac{h(x)-h(x-\Delta x)}{\Delta x}}{\Delta x}$$

$$\xrightarrow[0]{\Delta x} \frac{\delta^2}{2}\frac{\partial^2}{\partial x^2}h(x).$$

$$A = \frac{\delta^2}{2}\frac{\partial^2}{\partial x^2}.$$

Backward : $\partial K^{(t)}(x,y) / \partial t = \frac{\delta^2}{2}\frac{\partial^2}{\partial x^2}K^{(t)}(x,y).$

Forward : $\partial K^{(t)}(x,y) / \partial t = \frac{\delta^2}{2}\frac{\partial^2}{\partial y^2}K^{(t)}(x,y).$

$$g = Ah \qquad g(x) = \frac{E(h(X_{t+\Delta t}) \mid X_t = x) - h(x)}{\Delta t}$$

$$\big\downarrow$$

$$\frac{\delta^2}{2}\frac{\partial^2}{\partial x^2}. \qquad = \frac{E(h(x+\delta\varepsilon_t\sqrt{\Delta t})) - h(x)}{\Delta t}$$

$$= \frac{E\left(h(x) + h'(x)\delta\varepsilon_t\sqrt{\Delta t} + \frac{1}{2}h''(x)\delta^2\varepsilon_t^2\Delta t\right) - h(x)}{\Delta t}$$

$$= \frac{1}{2}h''(x)\delta^2 \cancel{E(\varepsilon_t^2)}.$$

First step

$$K^{(t+\Delta t)}(x,y) = \int \overset{[y|\varepsilon]}{K^{(t)}(x+\delta\varepsilon_t\sqrt{\Delta t}, y)}\ \overset{[\varepsilon]}{f(\varepsilon)}\,d\varepsilon$$

$X_t = x$ $\quad X_{t+\Delta t} = y$

$0 \quad \Delta t \quad t+\Delta t \longrightarrow t$

$$\big\downarrow$$

$$X_{\Delta t} = x + \delta\varepsilon_t\sqrt{\Delta t}$$

$$= \int \Big(K^{(t)}(x,y) + \frac{\partial}{\partial x}K^{(t)}(x,y)\delta\varepsilon\sqrt{\Delta t}$$

$$+ \frac{1}{2}\frac{\partial}{\partial x^2}K^{(t)}(x,y)\delta^2\varepsilon^2\Delta t \Big) f(\varepsilon)\,d\varepsilon$$

$$= K^{(t)}(x,y) + \frac{1}{2}\frac{\partial^2}{\partial x^2} K^{(t)}(x,y)\, \delta^2 \Delta t.$$

$$\frac{\partial^{(t)}(x,y)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2} K^{(t)}(x,y).$$

Last step

$$K^{(t+\Delta t)}(x,y) = \int K^{(t)}(x, y - \delta\varepsilon_t\sqrt{\Delta t})\, f(\varepsilon)\, d\varepsilon.$$

$$= \int \Big[K^{(t)}(x,y) - \frac{\partial}{\partial y} K^{(t)}(x,y)(-\delta\varepsilon_t\sqrt{\Delta t}) + \frac{1}{2}\frac{\partial^2}{\partial y^2} K^{(t)}(x,y)\, \delta^2 \varepsilon_t^2 \Delta t \Big] f(\varepsilon)\, d\varepsilon$$

fixed

$$= K^{(t)}(x,y) + \frac{1}{2}\frac{\partial^2}{\partial y^2} K^{(t)}(x,y)\, \delta^2 \Delta t.$$

BE $\longrightarrow$ $$\frac{\partial K^{(t)}(x,y)}{\partial t} = \frac{\delta^2}{2}\frac{\partial^2}{\partial y^2} K^{(t)}(x,y).$$

$$K^{(t)}(x,y) = \frac{1}{\sqrt{2\pi\delta^2 t}}\, e^{-\frac{(y-x)^2}{2\delta^2 t}}$$

$$[x_t \mid x_0 = x] \sim N(x, \delta^2 t).$$

Limiting Th:

$x_1, \ldots, x_n \overset{iid}{\sim} f(x)$

$x \sim f(x)$ $\quad \mu = E(x) \quad \delta^2 = Var(x).$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad E(\bar{x}) = \mu \quad Var(\bar{x}) = \frac{\delta^2}{n}.$

$\bar{x} \to \mu. \qquad \xrightarrow[\infty]{n} 0.$

$Y = \sqrt{n}(\bar{x} - \mu) \quad E(Y) = 0 \quad Var(Y) = \delta^2$

$\xrightarrow{D} N(0, \delta^2).$

$$x_t = x + \sum^{t} \delta\sqrt{\Delta t}\, \varepsilon_t = x + \sum_{i=1}^{n} \delta \frac{\sqrt{t}}{\sqrt{n}} \varepsilon_i = x + \delta\frac{\sqrt{t}}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i$$

$$= x + \delta\sqrt{t}\sqrt{n}\,\bar{\varepsilon} \longrightarrow x + \delta\sqrt{t}\, N(0,1) = N(x, \delta^2 t)$$

Bond : $X_{t+\Delta t} = X_t + r X_t \Delta t$

volatility

stock : ~~bet~~ $X_{t+\Delta t} = X_t + a X_t \Delta t + \delta X_t \varepsilon_t \sqrt{\Delta t}$

derivative : bet on bet

strike price K

Black-shoes formula.

NOTES 200A DATE 11 22 13

516

Stochastic differential equation

difference equ.

$$\underset{0}{\overset{}{|}}\ \underset{\Delta t}{\overset{1}{|}}\ \underset{2\Delta t}{\overset{2}{|}}\ \overset{3}{|}\ \cdots\ \underset{t}{\overset{n}{|}}\ \underset{t+\Delta t}{|}$$

Interest rate

$$X_{t+\Delta t} = X_t + rX_t\Delta t = X_t(1+r\Delta t)$$

$$X_0 = x.$$

$$X_{\Delta t} = x(1+r\Delta t)$$

$$X_{2\Delta t} = X_{\Delta t}(1+r\Delta t)$$

$$\cdots$$

1

$$X_t = x(1+r\Delta t)^{t/\Delta t} \xrightarrow{\Delta t \to 0} xe^{rt}.$$

differential : $\frac{X_{t+\Delta t} - X_t}{\Delta t} = rX_t$

$\Delta t \to 0$ $\frac{dX_t}{dt} = rX_t.$ $X_t = xe^{rt}.$

$$\log X_{t+\Delta t} = \log X_t + \log(1+r\Delta t).$$

$$= \log X_t + r\Delta t$$

2

$$\log X_t = \log x_0 + \sum_t r\Delta t$$

$$= \log x_0 + rt.$$

$$X_t = x_0 e^{rt}.$$

Stock : $X_{t+\Delta t} = X_t + \mu X_t\Delta t + \sigma X_t\varepsilon_t\sqrt{\Delta t}.$

$$= X_t(1+\mu\Delta t + \sigma\varepsilon_t\sqrt{\Delta t}).$$

$$\log X_{t+\Delta t} = \log X_t + \log(1+\mu\Delta t + \sigma\varepsilon_t\sqrt{\Delta t}).$$

$$= \log X_t + \mu\Delta t + \sigma\varepsilon_t\sqrt{\Delta t} - \tfrac{1}{2}\sigma^2\varepsilon_t^2\Delta t$$

$$\Rightarrow \log X_t = \log x_0 + \mu t \cancel{- \frac{1}{\sqrt{2}} \sigma^2 \varepsilon_t^2 + \sigma^2 \varepsilon t} + \sigma\sqrt{t}\,\boxed{\left(\Sigma \varepsilon_i\right) / \sqrt{n}} - \frac{\sigma^2 t}{2} \boxed{\left(\Sigma \varepsilon_i^2\right) / n}$$

$N(0,1)$ 1 Ito calculus

$$= \log x + \mu t + N(0, \sigma^2 t) - \frac{\sigma^2 t}{2}.$$

$$\Rightarrow X_t = x\, e^{\mu t - \frac{\sigma^2 t}{2} + N(0, \sigma^2 t)}.$$

Generator A

$$K^{(\Delta t)} = I + A\Delta t$$

$$A = \frac{K^{(\Delta t)} - I}{\Delta t}$$

$$g = Ah \qquad g(x) = A h(x)$$

$$= \frac{K^{(\Delta t)}\{h\} - Ih}{\Delta t} \qquad = \frac{E(h(X_{t+\Delta t}) \mid X_t = x) - h(x)}{\Delta t}$$

rate of smoothing

$$\frac{E(h(x + \mu x \Delta t + \sigma x \varepsilon_t \sqrt{\Delta t}) - h(x)}{\Delta t}$$

$$= \frac{E(h(x) + h'(x)(\mu x \Delta t + \sigma x \varepsilon_t \sqrt{\Delta t}) + \frac{1}{2} h''(x) \sigma^2 x^2 \varepsilon_t^2 \Delta t) - h(x)}{\Delta t}$$

$$= h'(x)\mu x + \frac{1}{2} h''(x) \sigma^2 x^2$$

$$A = \mu x \frac{\partial}{\partial x} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2}{\partial x^2}.$$

$$\frac{\partial K^{(t)} h}{\partial t} = A K^{(t)} h.$$

$$u(t,x) = K^{(t)} h$$

$$\frac{\partial u}{\partial t} = A u \quad \nearrow \quad u = e^{tA} h$$

$$= K^{(t)} h$$

$$u(t,x) = E(h(X_t) \mid X_0 = x).$$

SDE PDE

a single person whole population

$X_{t+\Delta t} = X_t + \mu X_t \Delta t + \sigma X_t \varepsilon_t \sqrt{\Delta t}$

$X_t = x e^{(\mu - \frac{\sigma^2}{2})t + N(0, \sigma^2 t)}$

$\frac{\partial u(x,t)}{\partial t} = (\mu x \frac{\partial}{\partial x} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2}{\partial x^2}) u$

$u(t, x) = (K^{(t)} h)(x)$

$= E(h(X_t) \mid X_0 = x)$

Black - scholes

option at time T

right to buy at price K.

$(X_T - K)_+ = \max(X_T - K, 0) = C_T$.

price at time t $C(t, X_t)$ $C(T, X_T) = (X_T - K)_+$.

Investment S share of stock. bet on bet

sell 1 share option.

time t: $S X_t - C(t, X_t)$ investment

$t+\Delta t$: $S X_{t+\Delta t} - C(t+\Delta t, X_{t+\Delta t})$

Change: $S \Delta X_t - [C_t(t, X_t)\Delta t + C_x(t, X_t)\Delta X_t + \frac{1}{2} C_{xx}(t, X_t) \Delta X_t^2]$

$\Delta X_t^2 = \sigma^2 X_t^2 \varepsilon_t^2 \Delta t$, $\varepsilon_t^2 \approx E = 1$

hedge: $S = C_x(t, X_t)$

Change: $-C_t(t, X_t)\Delta t - \frac{1}{2} C_{xx}(t, X_t) \sigma^2 X_t^2 \Delta t$.

$= r\Delta t (S X_t - C(t, X_t))$.

$= r\Delta t (C_x(t, X_t) X_t - C(t, X_t))$ $\forall X_t$.

$-C_t - \frac{1}{2} C_{xx} \sigma^2 x^2 = r x C_x - rC$.

NOTES

$$\boxed{rxC_x + \frac{1}{2}\sigma^2 x^2 C_{xx} + C_t = rC}$$

Recall $A = \mu x \frac{\partial}{\partial x} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2}{\partial x^2}$

$A \cdot c + \frac{\partial}{\partial t} c = r \cdot c$

$\downarrow$

$(\mu = r)$ $\quad \frac{\partial}{\partial t} c = (r - A)c$

$\cancel{C_t} = C_T = e^{(r-A)(T-t)} C_t.$

$C_t = e^{\cancel{r-A}\,(A-r)(T-t)} C_T.$

$C_t = e^{-r(T-t)} e^{A(T-t)} C_T.$

time discount. $\quad K_{t \to T}$

$C(t, x) = e^{-r(T-t)} E\,(C_T(X_T) \mid X_t = x)$

$= e^{-r(T-t)} E_Q\,((X_T - K)_+ \mid \cancel{X_t} X_t = x).$

risk neutral

not μ

as if stock: $X_{t+\delta t} = X_t + rX_t\,\delta t$

$+ \sigma X_t \varepsilon_t \sqrt{\delta t}$

S17

Stochastic Differential Equ.

$$X_{t+\Delta t} = X_t + \mu X_t \Delta t + \sigma X_t \overbrace{\varepsilon_t \sqrt{\Delta t}}^{\Delta B_t}.$$

$$E(\varepsilon_t) = 0 \qquad Var(\varepsilon_t) = 1.$$

$$B_{t+\Delta t} = B_t + \underbrace{\varepsilon_t \sqrt{\Delta t}}_{\Delta B_t}.$$

$$\Delta t \to 0 \qquad dX_t = \mu X_t \, dt + \sigma X_t \, dB_t.$$

more generally. $dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dB_t$

measure over all continous paths X_t, $X(t)$

$$\underset{0}{} \quad \underset{\Delta t}{1} \quad \underset{2\Delta t}{2} \quad \underset{3\Delta t}{3} \quad \cdots \quad \underset{t}{} \quad \underset{t+\Delta t}{n} \longrightarrow t$$

$$\frac{1}{n}\sum \varepsilon_i^2 \xrightarrow{L.L.N} E(\varepsilon_i^2) = 1. \qquad \varepsilon_t^2 \perp 1.$$

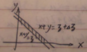
$$\frac{1}{\sqrt{n}}\sum \varepsilon_i \xrightarrow{C.L.T.} N(0,1). \qquad (dB_t)^2 \perp dt.$$

Sum of RV's

$$Z = X + Y.$$

$$(X, Y) \sim f(x,y) \overset{ind.}{=} f_X(x) f_Y(y).$$

$$\frac{P(Z \in (z, z+\Delta z))}{\Delta z} = \frac{P(X+Y \in (z, z+\Delta z))}{\Delta z}$$



$$= \frac{\sum_{cells} P(X+Y \in cell)}{\Delta z}$$

$$= \frac{\sum_x f(x, z-x)\,\Delta z\,\Delta x}{\Delta z} \longrightarrow \int f(x, z-x)\,dx$$

$$\overset{indep}{=} \int f_X(x) f_Y(z-x)\,dx. = f_Z(z).$$

$$= \int f_X(z-y) f_Y(y)\,dy.$$

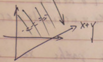
$$\underset{\mu_Z}{E(Z)} = \underset{\mu_X}{E(X)} + \underset{\mu_Y}{E(Y)}.$$

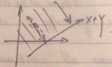
$$Var(Z) = Var(X) + Var(Y) + 2\,Cov(X, Y)$$

$$E((Z-\mu_Z)^2) = E\left(X+Y-(\mu_X+\mu_Y)\right)^2$$

$$= E\left((X-\mu_X)+(Y-\mu_Y)\right)^2$$

$$= Var\,X + Var\,Y + 2\,Cov(X, Y).$$



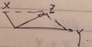


Ω	$\vec{X}$	$\vec{Y}$
1		
2		
$\vdots$	$X(\omega)$	$Y(\omega)$
M		

$\mu_X = \mu_Y = 0.$

$|\vec{X}|^2 = M\,Var(X)$

$|\vec{Y}|^2 = M\,Var(Y).$



iid $\quad X_1 \quad X_2 \quad \underline{iid} \quad f(x)$

$\bar{x} = \frac{X_1+X_2}{2} \qquad E(X_i) = \mu \qquad Var(X_i) = \sigma^2.$

$$E(\bar{x}) = E\left(\frac{X_1+X_2}{2}\right) = \frac{E(X_1)+E(X_2)}{2} = \mu$$

$$Var(\bar{x}) = Var\left(\frac{X_1+X_2}{2}\right) = \frac{Var(X_1)+Var(X_2)}{4} = \frac{\sigma^2}{2}.$$

averaging: (1) reducing variance $\rightarrow$ L.L.N.

(2) smoothing distr. $\rightarrow$ C.L.T.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n} \xrightarrow[\infty]{n} 0.$

$\bar{X} \to \mu.$

weak law of large #.

$P(|\bar{x} - \mu| > \underset{\substack{\downarrow \\ \text{fixed}}}{\varepsilon}) \longrightarrow 0.$

Markov Inequality.

$Z \geq 0. \quad t > 0.$

$P(Z > t) \leq \frac{E(Z)}{t}.$

Proof: $E(Z) = \int_0^\infty z f(z)\, dz.$

$> \int_t^\infty z f(z)\, dz \geq \int_t^\infty t f(z)\, dz.$

$= t\, P(Z > t).$

Chebychev inequ.

$Z = (X - \mu)^2.$

$P(|x - \mu| > t) = P((x-\mu)^2 > t^2).$

$\leq \frac{E((x-\mu)^2)}{t^2} = \frac{Var(X)}{t^2}.$

$P(|\bar{x} - \mu| > \varepsilon) \leq \frac{Var(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0.$

Central limit Th.

$E(X_i) = 0 \qquad Var(X_i) = 1.$

$E(\bar{x}) = 0 \qquad Var(\bar{x}) = \frac{1}{n}$.

microscope $\quad \sqrt{n}\,\bar{x} = Z \qquad E(Z) = 0 \quad Var(Z) = 1$.

$Z \xrightarrow{D} N(0, 1)$.

$P(Z \in (a, b)) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$.

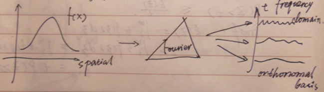
Characteristic function / Fourier transform

$X \sim f(x)$

$\varphi(t) = E(e^{itx})$

$= E(\cos tx + i \sin tx)$.

$= \int f(x) e^{itx} dx$.



$f(x) \propto \int e^{itx} \varphi(t) dt$.

$X \sim f_X(x) \qquad Y \sim f_Y(y)$

$\varphi_X(t) = E_X(e^{itx}) \qquad \varphi_Y(t) = E_Y(e^{itY})$.

$Z = X + Y \sim f_Z(z) = \int_{\infty} f_X(z-y) f_Y(y) dy = \cdots$

$\varphi_Z(t) = E(e^{itZ}) = E(e^{it(X+Y)})$

$= E(e^{itX} \cdot e^{itY})$.

$= E(e^{itX}) E(e^{itY}) = \varphi_X(t) \varphi_Y(t)$

$X \perp Y$.

$\Rightarrow \varphi_Z(t) = \varphi_X(t)\,\varphi_Y(t).$

$Z = \sum x_i / \sqrt{n}.$

$\varphi_Z(t) = E(e^{itZ}) = E(e^{it/\sqrt{n}\,\sum x_i}).$

$= E(\prod_{i=1}^{n} e^{i\frac{t}{\sqrt{n}}x_i})$

$\stackrel{\text{indep.}}{=} \prod_{i=1}^{n} E(e^{i\frac{t}{\sqrt{n}}x_i}) = \prod_{i=1}^{n} \varphi_X(\frac{t}{\sqrt{n}}) = \varphi_X^n(\frac{t}{\sqrt{n}})$

$= (\varphi_X(0) + \varphi_X'(0)\frac{t}{\sqrt{n}} + \frac{1}{2}\varphi_X''(0)\frac{t^2}{n} + \cdots)^n.$

$\varphi_X'(t) = E\, de^{itx}/dt) = E(ixe^{itx})$

$\varphi_X'(0) = E(ix) = 0.$

$\varphi_X''(t) = E(e^{itx}(ix)^2)$

$\varphi_X''(0) = -1. \qquad \varphi(0) = 0.$

$\varphi_Z(t) = (1 - \frac{1}{2}\frac{t^2}{n} + o(\frac{1}{n}))^n.$

$\longrightarrow e^{-\frac{t^2}{2}}.$

$$Z = \frac{\sum x_i}{\sqrt{n}} \xrightarrow{D} N(0,1) \quad \left\} \begin{array}{l}\text{uniqueness}\\ \text{continuity (Levy)}\end{array}\right.$$

$$\begin{array}{ccc} \downarrow & & \downarrow \\ \varphi_Z(t) & \longrightarrow & e^{-\frac{t^2}{2}} \end{array}$$

$Y \sim f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$

$\varphi(t) = E(e^{itY}) = \int e^{ity}\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\,dy.$

$= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y^2 - 2ity)}\,dy.$

$= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-it)^2} e^{-\frac{1}{2}t^2}\,dy.$

$= e^{-\frac{t^2}{2}}.$

Moment generating function

$X \sim f(x)$

$m(t) = E(e^{tX})$.

$m'(t) = E(e^{tX} X)$ $\quad m'(0) = E(X)$

$m''(0) = E(X^2) \quad \cdots \quad m^{(k)}(0) = E(X^k)$.

$P(Z > c) \le E(Z)/c$

Chernoff scheme.

$P(X > c) = P(e^{tX} > e^{tc})$.

$\le \min_t E(e^{tX})/e^{tc}$.

518

200 A 12/4/13

Recall Law of large #

$X_1, X_2, \dots, X_n$ iid $f(x)$

$E_f(X) = \mu$

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mu$

$P(|\bar{x} - \mu| > \varepsilon) \longrightarrow 1.$

$\frac{1}{n} \sum_{i=1}^{n} h(x_i) \longrightarrow E_f(h(x)) = \int h(x) f(x) dx$

$\frac{1}{n} \sum_{i=1}^{n} Y_i \longrightarrow E(Y_i)$

Shannon's equipaition properey in information theory

Example : Flip flipping

$X_1, X_2, \dots, X_n$ iid Bernoulli $(\frac{1}{2})$

	Tail	Head
X	0	1
P	$\frac{1}{2}$	$\frac{1}{2}$

$\underbrace{P(X_1, X_2, \dots, X_n)}_{\text{random sequence}} = 1/2^n.$

Example : non-uniform

X	A	B	C	D
Prob	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

$X_1, X_2, \dots, X_n$ iid $P(x)$

$P(X_1, X_2, \dots, X_n) = P(X_1) P(X_2) \dots P(X_n).$

$\Rightarrow \frac{1}{n} \log_2 P(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \log_2 P(X_i)$

$\longrightarrow E_p(\log_2 P(x))$

$= \sum_{A,B,C,D} P(x) \log_2 P(x) = -\frac{7}{4}.$

If n is large, $P(X_1, \dots, X_n) = 2^{-1.75n}.$

$X_1, X_2, \dots, X_n \approx 1.75\, n$ coin flippings

each $X_i \approx 1.75$ coin flippings

$P(X_1, X_2, \dots, X_n) \approx 2^{-n \overbrace{E_p(-\log_2 P(x))}^{H(p)}}$ $H(p)$ — entropy of p

$= 2^{-n H(p)}$ randomness of p

$|A| = 2^{nH(p)}$ ~ Uniform(A). $p(x) = \frac{1}{|A|} = 2^{-nH(p)}$.

$A_{n,\varepsilon} = \{ \underbrace{(x_1, x_2, \dots, x_n)}_{\text{lower case}} : 2^{-(nH(p)+\varepsilon)} \le p(x_1 \dots x_n) \le 2^{-n(H(p)-\varepsilon)} \}$

typical set

$\exists N, \forall n > N.\ P(A_{n,\varepsilon}) \ge 1-\delta$.

$P((x_1, x_2, \dots, x_n) \in A_{n,\varepsilon})$

$= P(|\frac{1}{n}\sum \log_2 p(x_i) - H(p)| < \varepsilon) \longrightarrow 1$.

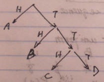
$|A_{n,\varepsilon}|$ $\quad 1-\delta < P(A_{n,\varepsilon}) < 1$

$1 > P(A_{n,\varepsilon}) > 2^{-n(H(p)+\varepsilon)} |A_{n,\varepsilon}|$

$1-\delta < P(A_{n,\varepsilon}) < 2^{-n(H(p)-\varepsilon)} |A_{n,\varepsilon}|$

$\Rightarrow (1-\delta)\, 2^{n(H(p)-\varepsilon)} < |A_{n,\varepsilon}| < 2^{n(H(p)+\varepsilon)}$

$P(x_1, x_2, \dots, x_n) \sim Unif(A)$ $\quad |A| = 2^{nH(p)}$.



E(# of coinflipping) = 1.75

general $E_p(-\log_2 p(x)) = H(p)$.

Code: A A B C D $\rightleftarrows$ 1101001000

Prefix code: the code of any letter is not the beginning part of code of any other letter

Reason: after generating a letter by coin flipping. STOP

$X_1 X_2 \dots X_n \overset{iid}{\sim} p(x)$

$\updownarrow$

0101100110 $\overset{iid}{\sim}$ Ber$(\frac{1}{2})$. $\nearrow$ shortest code, cannot be further compressed

$\searrow$ Unif(Ω) 2^N sequency

coin flipping experiment

Shorter code length to more frequent letters

Kolmogorov complexity

What's a "random" sequence?

We cannot find a shorter code to produce it.

No compressible.

If a sequence is not compressible, then it has all statistical properties of a seq. iid $Ber(\frac{1}{2})$.

Frequency of heads $= \frac{1}{2}$.

What if freq. of heads $= \frac{1}{3}$. it's compressible.

$(N = 3000$, 1000 heads, 2000 tails$)$.

$|\{$ all sequences with 1000 heads, 2000 tails $\}|$

$\sim 2^{3000}$

say 2^{1500} need 1500 bits.

Kullback-Leibler divergence

	X	A	B	C	D	
true	p	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\rightarrow$ coding length $= E_p(-\log_2 p(x))$
mistaken	q	$\frac{1}{8}$	$\frac{1}{8}$	ϕ	$\frac{1}{2}$	$\rightarrow$ coding length $= E_p(-\log_2 q(x))$

redundency: $E_p(-\log q(x)) - E_p(-\log p(x))$

$= E_p(\log \frac{p(x)}{q(x)}) = D(p\|q) \geq 0.$

Jensen Inequality

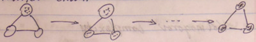
$E(\log(Z)) \leq \log E(Z).$

$Z = q(x)/p(x)$ $\quad E_p(\log \frac{q(x)}{p(x)}) \leq \log E_p(\frac{q(x)}{p(x)}) = 0.$

$E_p(\log \frac{p(x)}{q(x)}) \geq 0.$

Arrow of time

Markov Chain



a single person reversible

population not reversible.

$X_1, X_2, \dots X_t$ Markov Chain

$X_t \sim P^{(t)}$ $X_{t+1} \sim P^{(t+1)}$.

$\text{entropy}(P^{(t+1)}) \geq \text{entropy}(P^{(t)})$.

$\mathbb{P}(x,y): (X_t, X_{t+1}) \sim P^{(t)}(x)\, K(x,y) \sim P^{(t+1)} K_1(y,x)$

If $X_t \sim \pi$, $\pi(x,y): (X_t, X_{t+1}) \sim \pi(x)\, K(x,y)$.

$X_{t+1} \sim \pi$. $\sim \pi(y)\, K_0(y,x)$

$$D(P(x,y) \mid \pi(x,y)) = E_p\left(\log \frac{P(x,y)}{\pi(x,y)}\right) = E_p\left(\log \frac{P^{(t)}(x)\, K(x,y)}{\pi(x)\, K(x,y)}\right)$$

$$= D(P^{(t)} \mid \pi) = E_p\left(\log \frac{P^{(t+1)}(y)\, K_1(y,x)}{\pi(y)\, K_0(y,x)}\right).$$

$$= E_p(\log P^{(t+1)}(y)/\pi(y)) + E_p\left(\log \frac{K_1(y,x)}{K_0(y,x)}\right)$$ not reversible

$$= D(P^{(t+1)} \mid \pi) + D(K_1 \mid K_0).$$

$$D(P^{(t)} \mid \pi) \geq D(P^{(t+1)} \mid \pi).$$

π: uniform

entropy increases

519

Lec A

Law of large #

$x_1\ x_2\ \cdots\ x_n \overset{iid}{\sim} p(x)$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow \mu = E_p(x) = \int x\,p(x)\,dx$

Markov: $Z \geq 0 \quad P(Z > t) < E(Z)/t$

Proof: $E(Z) = \int_0^\infty z\,p(z)\,dz \geq \int_t^\infty z\,p(z)\,dz \geq t\int_t^\infty p(z)\,dz$
$= t\,P(Z > t)$

Chebchev:

$P(|x-\mu| > \varepsilon) = P((x-\mu)^2 > \varepsilon^2) \leq \frac{E((x-\mu)^2)}{\varepsilon^2} = \frac{Var(x)}{\varepsilon^2}$

$Var(\bar{x}) = \frac{S^2}{n} \qquad S^2 = Var(X)$

$P(|\bar{x}-\mu| > \varepsilon) \leq \frac{Var(\bar{x})}{\varepsilon^2} = S^2/n\varepsilon^2 \longrightarrow 0$

Chernoff method:

$P(X > t) = P(\underbrace{\boxed{e^{\lambda X}}}_{Z} > e^{\lambda t}) \leq \frac{E(e^{\lambda X})}{e^{\lambda t}}$ — $Z(\lambda)$: moment generating function

Recall characteristic fun $\varphi(t) = E(e^{itx})$. $\quad Z'(0) = E(X) \quad Z''(0) = Var(X)$

$P(\bar{x} > t) = P(\sum_{i=1}^{n} x_i > nt) = P(e^{\lambda\sum_{i=1}^{n} x_i} > e^{\lambda nt})$

$\leq E(e^{\lambda\sum x_i}) / e^{n\lambda t} = E(\prod_{i=1}^{n} e^{\lambda x_i}) / e^{n\lambda t}$

$= \prod_{i=1}^{n} E(e^{\lambda x_i}) / e^{n\lambda t} = Z(\lambda)^n / e^{n\lambda t} = (\frac{Z(\lambda)}{e^{\lambda t}})^n$

$= e^{-n\underbrace{[\lambda t - \log Z(\lambda)]}_{max}}$

$I(t) = \max_\lambda [\lambda t - \log Z(\lambda)] \geq 0$

$\lambda t - \log Z(\lambda)|_{\lambda=0} = 0$

$P(\bar{x} > t) \leq e^{-nI(t)}$

$P_\lambda(x) = \frac{1}{Z(\lambda)} e^{\lambda x} p(x)$ tilting $P = P_0$ to P_λ.

$Z(\lambda) = \int e^{\lambda x} p(x)\,dx = E_p(e^{\lambda x})$

(recall Ising model. exponential family)

$\frac{\partial}{\partial\lambda}\log Z(\lambda) = \frac{Z'(\lambda)}{Z(\lambda)} = \frac{1}{Z(\lambda)}\int x e^{\lambda x} p(x)\,dx = E_\lambda(x)$

$$\frac{\partial^2}{\partial \lambda^2} \log Z(\lambda) = Var_\lambda(x)$$

$$\frac{\partial^2}{\partial \lambda^2} \log Z(\lambda) = \int x \frac{1}{Z(\lambda)} e^{\lambda x} x\, p(x)\, dx - \int x \frac{1}{Z^2(\lambda)} Z'(\lambda) e^{\lambda x} p(x) dx$$

$$= \int x^2 p_\lambda(x)\, dx - \frac{Z'(\lambda)}{Z(\lambda)} \int x \frac{1}{Z(\lambda)} e^{\lambda x} p(x) dx$$

$$= E_\lambda(x^2) - E_\lambda^2(x) = Var_\lambda(x).$$

Hoeffding Inequility

$x \sim p(x)$

$E(x) = 0 \qquad x \in [a, b]$.

$$\log Z(\lambda) = \underbrace{\log Z(0)}_{0} + \underbrace{\frac{\partial}{\partial \lambda} \log Z(\lambda)\Big|_{\lambda=0}}_{0} \lambda + \frac{1}{2} \underbrace{\frac{\partial^2}{\partial \lambda^2} \log Z(\lambda)\Big|_{\lambda=\tilde{\lambda}}}_{Var_{\tilde{\lambda}}(x)} \lambda^2$$

$$Var_{\tilde{\lambda}}(x) = Var_{\tilde{\lambda}}\left(x - \frac{b+a}{2}\right) \leq E\left(\left|x - \frac{b+a}{2}\right|^2\right) \leq \left(\frac{b-a}{2}\right)^2$$

$$\Rightarrow \log Z(\lambda) \leq \frac{(b-a)^2}{8} \lambda^2$$

$$\lambda t - \log Z(\lambda) \geq \lambda t - \frac{(b-a)^2}{8} \lambda^2.$$

$$\lambda = \frac{4t}{(b-a)^2} \qquad \hookrightarrow \frac{2t^2}{(b-a)^2}$$

$$I(t) \geq \frac{2t^2}{(b-a)^2}.$$

$$P(\bar{x} > t) \leq \exp\left(-n \cdot \frac{2t^2}{(b-a)^2}\right).$$

If $E(X) = \mu$, $P(|\bar{x} - \mu| \geq \varepsilon) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$.

$\hookrightarrow$ concentration inequality

Chebchev: $P(|\bar{x} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$.

Random coinflipping

flip n times indeps.

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} Ber\left(\frac{1}{2}\right)$.

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i =$ frequency of heads

$P(|\bar{x} - \frac{1}{2}| > \varepsilon) \leq 2e^{-2n\varepsilon^2} \to 0$

Viewed in terms of $\bar{x}$

Viewed in $\Omega = \{2^n \text{ sequences}\}$

Count # of seq's

where frequency is $\in \left(\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right)$

$$\frac{\#\text{ of such seqs}}{2^n} > 1 - 2e^{-2n\varepsilon^2}.$$

A vast majority of seqs are typical

Unif (Ω) : random, concentrate on typical seq's
frequency of heads $\doteq \frac{1}{2}$ deterministic

Kolmogrov Complexiey
a non-compressible seq. must have typical property
random seq. freq(H) $\doteq \frac{1}{2}$.
If $|\bar{x} - \frac{1}{2}| > \varepsilon$, seq. $\in$ non-typical seq's
# of non-typical seq's $\leq 2^n \times (2e^{-2n\varepsilon^2})$
$\sim 2^{\rho n} \ (\rho < 1)$
minority
coded by ρn bits

Hoeffding: quardratic approxmating to $\log Z(\lambda)$.
concentration
Cramer: $P_\lambda(x) = \frac{1}{Z(\lambda)} e^{\lambda x} p(x)$. $\longrightarrow$ large deviation
$\max_\lambda$: $\lambda t - \log Z(\lambda)$.
$\frac{\partial}{\partial \lambda}$ $t - E_\lambda(x) = 0$
solution λ^* : $E_{\lambda^*}(x) = t$.

$I(t) = \lambda^* t - \log Z(\lambda^*) = D(P_{\lambda^*} \| P)$
$D(P_{\lambda^*} \| P) = E_{P_{\lambda^*}} \left(\log \frac{P_{\lambda^*}(x)}{P(x)} \right)$
$= E_{\lambda^*} (\lambda^* x - \log Z(\lambda^*))$
$= \lambda^* t - \log Z(\lambda^*)$.

$P(\bar{x} > t) \approx e^{-nD(P_{\lambda^*} \| P)}$
$\frac{1}{n} \log P(\bar{x} > t) \longrightarrow D(P_{\lambda^*} \| P)$.