9. Advanced Probability

Section 9.1 Moment and Correlation

Section 9.2 Probability Model (只教不考)

Section 9.3 Entropy

Section 9.4 Kullback-Leibler Divergence (KL Divergence)

Section 9.5 Basic Concepts of Random Process (只教不考)

Section 9.6 Independent Component Analysis (只教不考)

9.1 Moment and Correlation

9.1.1 Review for Probability

1. Discrete Case

(a) Probability (Probability Mass Function, PMF)

$$P_X(n) = P\{X = n\} = \frac{number\ of\ cases\ where\ X = n}{total\ number\ of\ cases}$$

Note: $\sum_n P_X(n) = 1$

(b) Joint Probability

$$P_{X,Y}(n,m) = P\{(X = n)\ and\ (Y = m)\}$$

$$= \frac{number\ of\ cases\ where\ X = n\ and\ Y = m}{total\ number\ of\ cases}$$

(c) Conditional Probability

$$P_{X|Y}(n\,|\,m) = P\{(X=n)\,|\,(Y=m)\}$$

$$= \frac{number\ of\ cases\ where\ X=n\ and\ Y=m}{number\ of\ cases\ where\ Y=m}$$

$$P_{X|Y}(n\,|\,m) = \frac{P_{X,Y}(n,m)}{P_Y(m)}$$

An Example of the Discrete Probability Distribution

Binomial Distribution

$$\binom{N}{n} = C^N_n = \frac{N!}{n!\,(N-n)!}$$

$$P_X(n) = \binom{N}{n} p^n (1-p)^{N-n} \quad \text{for } n = 0,\ 1,\ 2,\ \ldots,\ N$$

$$P_X(n) = 0$$

Other examples are shown in Section 9.2.1

2. Continuous Case

(a) Cumulative Distribution Function (CDF)

$$F_X(x) = Prob(X \le x)$$

(b) Probability Density Function (PDF)

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Note: $F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx$

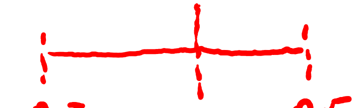
$$\lim_{x \to \infty} F_X(x) = \int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

$$f_X(x) = \lim_{\Delta \to 0} \frac{F_X(x+\Delta) - F_X(x-\Delta)}{2\Delta}$$

$$f_X(x) = \lim_{\Delta \to 0} \frac{Prob(x-\Delta < X \le x+\Delta)}{2\Delta}$$

for quantization error

$Prob(X \le x) = 0.5 + x$ if $-0.5 \le x \le 0.5$



$Prob(X \le x) = 0$ if $x < -0.5$

$Prob(X \le x) = 1$ if $x > 0.5$

CDF

$$f_X(x) \ne \text{Prob}\{X = x\}$$

$f_X(x) = 0$ if $x < -0.5$ or $x > 0.5$

$= 1$ if $-0.5 < x < 0.5$

(c) Joint Cumulative Distribution Function

$$F_{X,Y}(x,y) = Prob((X \le x)\ and\ (Y \le y))$$

(d) Joint Probability Density Function

$$f_{X,Y}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{X,Y}(x,y)$$

(e) Conditional Probability

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{X,Y}(x,y)}{\frac{d}{dy}F_Y(y)}$$

(f) Integral Probability

$$f_X(x) = \int f_{X,Y}(x,y)\,dy$$

$$f_Y(y) = \int f_{X,Y}(x,y)\,dx$$

Examples of the Probability Density Function in the Continuous Case

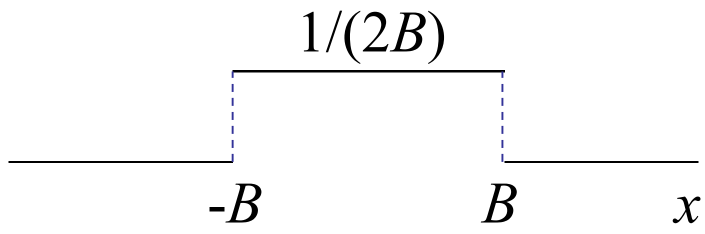
Uniform Distribution

(Quantization error is a special case of $B = 0.5$)

skewness = 0

$$f_X(x) = \frac{1}{2B} \quad \text{when } |x| < B$$

$$f_X(x) = 0 \quad \text{otherwise}$$



Normal Distribution

skewness = 0

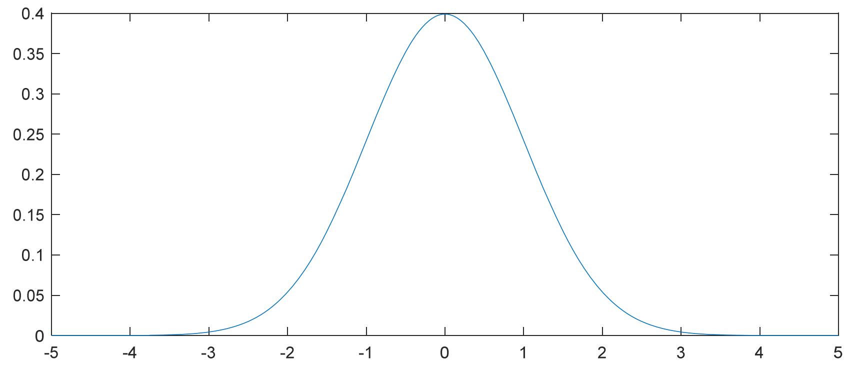
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$-\infty < x < \infty$

μ : mean

σ : standard deviation

when $\mu = 0$, $\sigma = 1$



Other examples are shown in Section 9.2.2

3. Expected Value

ex: P(X = 1) = 0.5 g[1] = 100
P(X = 0) = 0.5 g[0] = -50
E(g[X])
= 100 × 0.5 + (-50) × 0.5
= 25

(discrete case)

$$E(g[X]) = \sum_n g[n]P(X = n) = \sum_n g[n]P_X(n)$$

(continuous case)

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

It is usually written as

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)\, dF_X(x)$$

9.1.2 Moments

Mean

$$\mu_X = E(X) = \sum_n nP_X(n) \qquad \text{(discrete case)}$$

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \qquad \text{(continuous case)}$$

Variance (var)

$$var_X = E\left((X-\mu_X)^2\right) = \sum_n (n-\mu_X)^2 P_X(n) \qquad \text{(discrete case)}$$

$$var_X = E\left((X-\mu_X)^2\right) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\, dx \qquad \text{(continuous case)}$$

Standard Deviation (std)

$$\sigma_X = \sqrt{var_X} = \sqrt{E\left((X-\mu_X)^2\right)} \qquad \text{(discrete case)}$$

$$\sigma_X = \sqrt{var_X} = \sqrt{E\left((X-\mu_X)^2\right)} \qquad \text{(continuous case)}$$

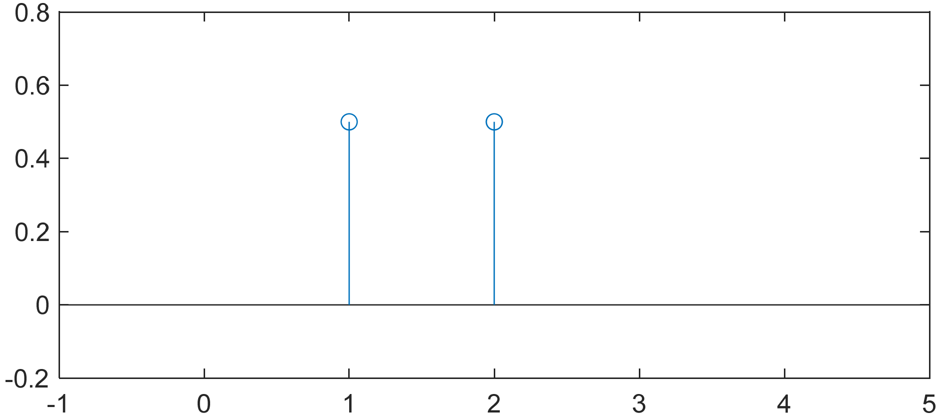
ex: for quantization

$\mu_x = \int_{-0.5}^{0.5} x \cdot 1\, dx = 0$

$var_x = \int_{-0.5}^{0.5} (x-0)^2 \cdot 1\, dx = \left.\frac{x^3}{3}\right|_{-0.5}^{0.5} = \frac{1}{12}$

[Example 1]

(1) $P_X(n)$

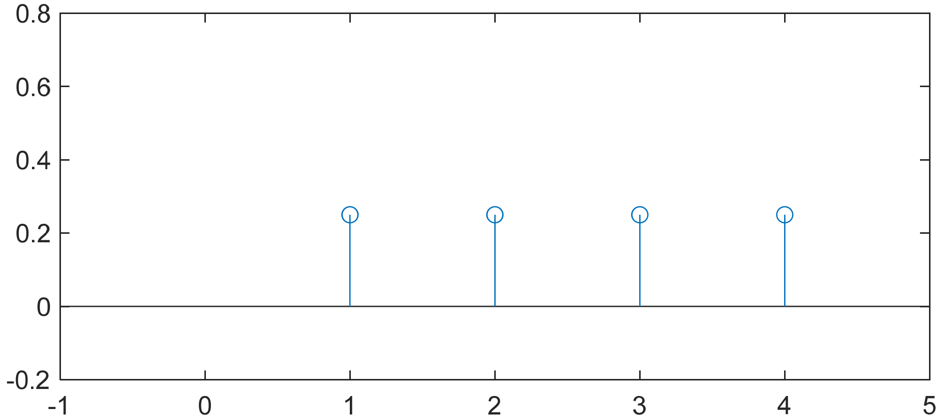


$$\mu_X = 1 \cdot 0.5 + 2 \cdot 0.5 = 1.5$$

$$var_X = 0.5(1-1.5)^2 + 0.5(2-1.5)^2 = 0.25$$

$$\sigma_X = \sqrt{var_X} = 0.5$$

(2) $P_X(n)$



$\mu_X = 2.5$ $\quad var_X = 1.25$ $\quad \sigma_X = \sqrt{5}/2$

(1) Moment (raw form)

$$\hat{m}_k = E\left(X^k\right) = \sum_n n^k P_X(n) \qquad \text{(discrete case)}$$

$$\hat{m}_k = E\left(X^k\right) = \int_{-\infty}^{\infty} x^k f_X(x)\,dx \qquad \text{(continuous case)}$$

(2) Moment (central form)

$$m_k = E\left((X-\mu_X)^k\right) = \sum_n (n-\mu_X)^k P_X(n) \qquad \text{(discrete case)}$$

$$m_k = E\left((X-\mu_X)^k\right) = \int_{-\infty}^{\infty} (x-\mu_X)^k f_X(x)\,dx \qquad \text{(continuous case)}$$

if $k=1$, $m_1 = E(X) - E(\mu_X)$

$= \mu_k - \mu_k = 0$

(3) Moment (standardized form)

$$v_k = \frac{m_k}{\sigma_X^k} = \frac{E\left((X-\mu_X)^k\right)}{\left[E\left((X-\mu_X)^2\right)\right]^{k/2}} = \frac{\sum_n (n-\mu_X)^k P_X(n)}{\left[\sum_n (n-\mu_X)^2 P_X(n)\right]^{k/2}}$$

σ_K: standard deviation

(discrete case)

$$v_k = \frac{m_k}{\sigma_X^k} = \frac{E\left((X-\mu_X)^k\right)}{\left[E\left((X-\mu_X)^2\right)\right]^{k/2}} = \frac{\int_{-\infty}^{\infty} (x-\mu_X)^k f_X(x)\,dx}{\left[\int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,dx\right]^{k/2}}$$

(continuous case)

Order	Moment (raw)	Moment (central)	Moment (standardized)
$k = 1$	Mean	0	0
$k = 2$		Variance	1
$k = 3$			Skewness 偏度
$k = 4$			Kurtosis 峰度
$k = 5$			Hyperskewness

Skewness (偏度)

It indicates the relative locations of the high-probability region and the long tail to the mean.

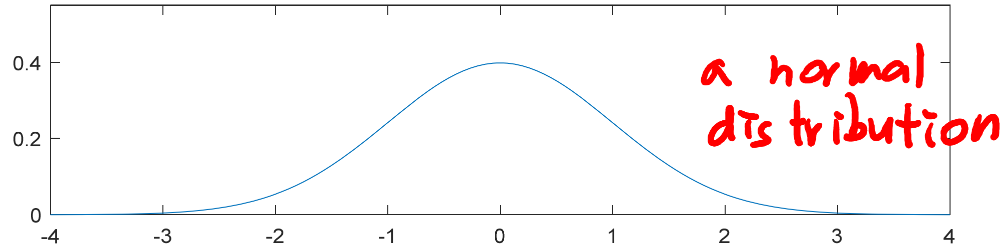
$$skewness = \frac{\sum_{n}(n-\mu_X)^3 P_X(n)}{\left[\sum_{n}(n-\mu_X)^2 P_X(n)\right]^{3/2}} \quad or \quad \frac{\int_{-\infty}^{\infty}(x-\mu_X)^3 f_X(x)\,dx}{\left[\int_{-\infty}^{\infty}(x-\mu_X)^2 f_X(x)\,dx\right]^{3/2}}$$

skewness > 0: The high-probability region is in the left of the mean.

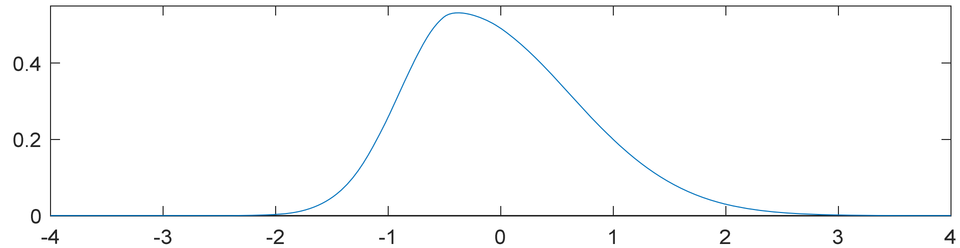
skewness < 0: The high-probability region is in the right of the mean.

skewness $= 0$: Symmetry

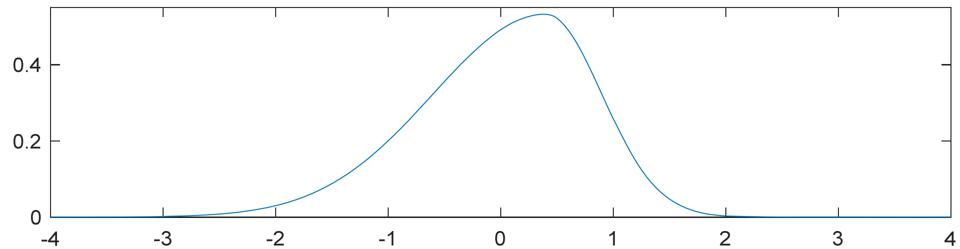
skewness = 0
mean = 0



skewness = 0.4992
mean = 0



skewness = -0.4992
mean = 0



Kurtosis (峰度)

It indicates how sharp the high-probability region is.

$$kurtosis = \frac{\sum_{n}(n-\mu_X)^4 P_X(n)}{\left[\sum_{x}(x-\mu_X)^2 P_X(x)\right]^2} \qquad or \quad \frac{\int_{-\infty}^{\infty}(x-\mu_X)^4 f_X(x)dx}{\left[\int_{-\infty}^{\infty}(x-\mu_X)^2 f_X(x)dx\right]^2}$$

kurtosis ≥ 0

large kurtosis: The high probability region is sharp.

$$f_X(x) = e^{-2|x|}$$

kurtosis = 6

std = 0.707

$$f_X(x) = \frac{1}{2} e^{-|x|}$$

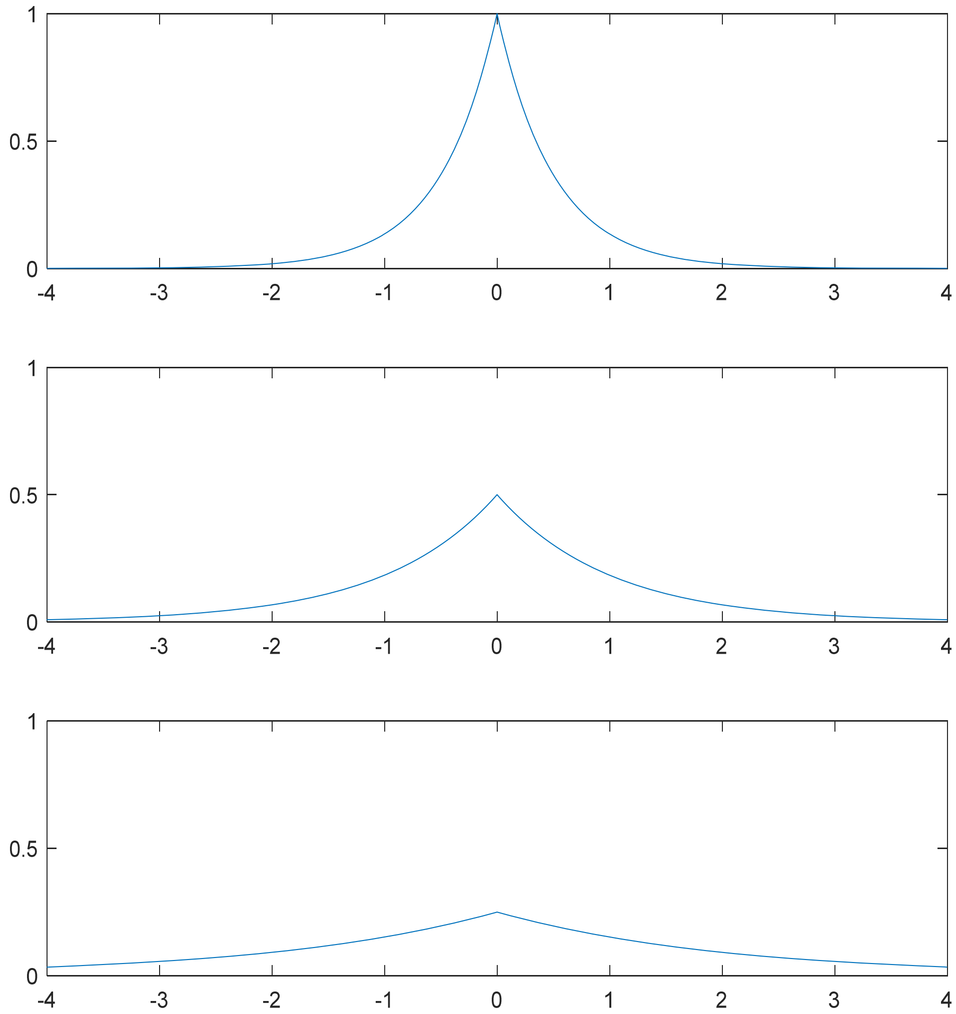
kurtosis = 5.86

std = 1.414

$$f_X(x) = \frac{1}{4} e^{-|x|/2}$$

kurtosis = 4.38

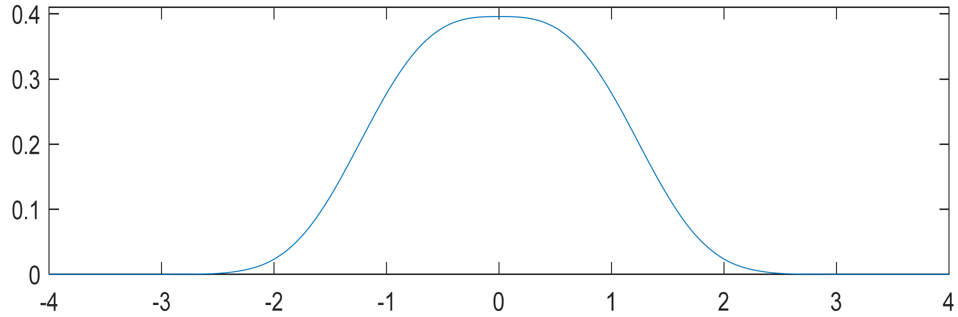
std = 2.647



$$f_X(x) = 0.396\, e^{-|x|^3/2^{3/2}}$$

kurtosis = 2.4184

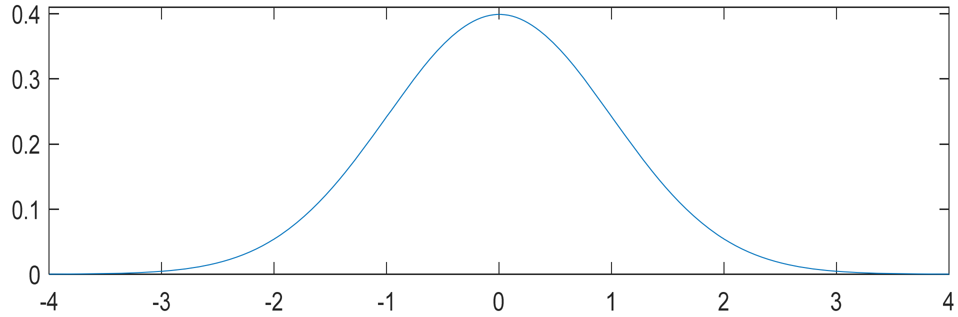
std = 0.864



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

kurtosis = 3

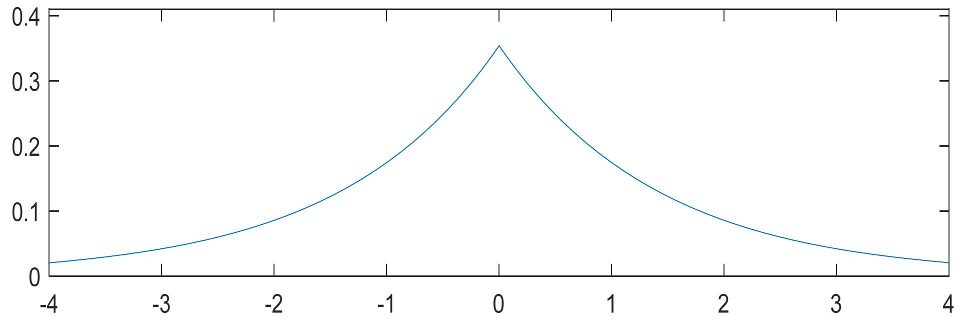
std = 1



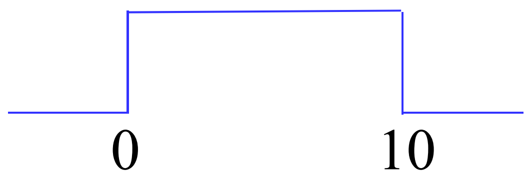
$$f_X(x) = \frac{1}{2\sqrt{2}} e^{-|x|/\sqrt{2}}$$

kurtosis = 5.2903

std = 1.9726



[Example 2] Suppose that X is uniformly distributed in $x \in [0, 10]$. Determine the central moments, the standardized moments, the variance, the skewness, and the kurtosis of X.



(Solution): $f_X(x) = \frac{1}{10} \quad for\ 0 < x < 10$

$$f_X(x) = 0 \quad otherwise$$

$$\mu_X = \int_0^{10} x \frac{1}{10} dx = 5$$

$$\sigma_X = \sqrt{\int_0^{10} (x-5)^2 \frac{1}{10} dx} = \sqrt{\frac{1}{10}\int_{-5}^{5} x^2\, dx} = \frac{5}{\sqrt{3}}$$

Central moment:

$$m_k = \int_0^{10} (x-5)^k \frac{1}{10} dx = \frac{1}{10}\int_{-5}^{5} x^k\, dx = \begin{cases} 0 & if\ k\ is\ odd \\ \dfrac{5^k}{k+1} & if\ k\ is\ even \end{cases}$$

Standardized moment:

$$v_k = \frac{m_k}{\sigma_X^k} = 3^{k/2}\frac{m_k}{5^k} = \begin{cases} 0 & \text{if } k \text{ is odd} \\ \frac{3^{k/2}}{k+1} & \text{if } k \text{ is even} \end{cases}$$

Variance: $var_X = m_2 = \frac{25}{3}$

Skewness = $v_3 = 0$

Kurtosis = $v_4 = \frac{9}{5}$

9.1.3 Correlation

Covariance

$$cov_{X,Y} = E\left((X-\mu_X)(Y-\mu_Y)\right)$$

$$cov_{X,Y} = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) P_{X,Y}(x,y)$$ (discrete case)

$$cov_{X,Y} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y)\,dxdy$$ (continuous case)

Note: (1) When $X = Y$,

$$cov_{X,X} = var_X$$

(2) $$cov_{X,Y} = E(XY) - \mu_X \mu_Y$$

Correlation

$$corr_{X,Y} = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X \sigma_Y} = \frac{cov_{X,Y}}{\sigma_X \sigma_Y}$$

$$corr_{X,Y} = \frac{\sum_x \sum_y (x-\mu_X)(y-\mu_Y)P_{X,Y}(x,y)}{\sqrt{\sum_x (x-\mu_X)^2 P_X(x) \sum_y (y-\mu_Y)^2 P_Y(y)}}$$

(discrete case)

$$corr_{X,Y} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y)\,dxdy}{\sqrt{\int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,dx \int_{-\infty}^{\infty} (y-\mu_Y)^2 f_Y(y)\,dy}}$$

(continuous case)

$$corr_{X,Y} = \frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X \sigma_Y}$$

Note: (1)

$$-1 \le corr_{X,Y} \le 1$$

(2) When $Y = cX + d$, c is a positive constant, d is a constant,

$$corr_{X,Y} = 1$$

(Proof): $Y = cX + d, \qquad \mu_Y = c\mu_X + d$

$$E\big((X-\mu_X)(Y-\mu_Y)\big) = E\big((X-\mu_X)(cX + d - c\mu_X - d)\big)$$

$$= E\left(c(X-\mu_X)^2\right) = c\,\mathrm{var}_X$$

$$\sigma_Y = \sqrt{E\left((cX + d - c\mu_X - d)^2\right)} = \sqrt{c^2 E\left((X-\mu_X)^2\right)} = |c|\,\sigma_X$$

$$corr_{X,Y} = \frac{c\,\mathrm{var}_X}{|c|\,\sigma_X \sigma_X} = \frac{c}{|c|}$$

(3) When $Y = cX + d$, c is a negative constant, d is a constant,

$$corr_{X,Y} = -1$$

(4) If Y is independent of X,

$$corr_{X,Y} = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X \sigma_Y} = \frac{E\left(X-\mu_X\right)E\left(Y-\mu_X\right)}{\sigma_X \sigma_Y} = 0$$

(i) Full Correlation:

$$\left|corr_{X,Y}\right| \geq 0.9$$

(ii) High Correlation:

$$0.6 \leq \left|corr_{X,Y}\right| < 0.9$$

(iii) Middle Correlation:

$$0.3 \leq \left|corr_{X,Y}\right| < 0.6$$

(iv) Low Correlation:

$$\left|corr_{X,Y}\right| < 0.3$$

(v) Positive Correlation:

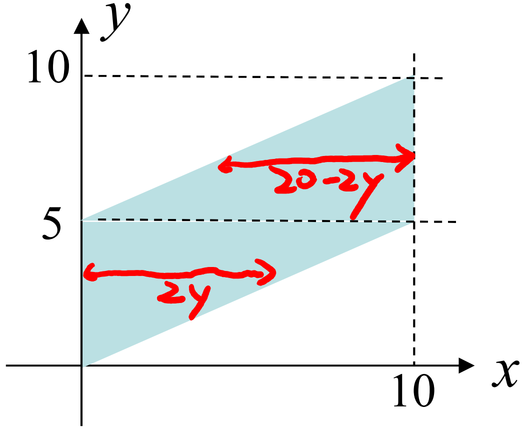
$$corr_{X,Y} > 0$$

(vi) Negative Correlation:

$$corr_{X,Y} < 0$$

[Example 3] Determine the Covariance and the Correlation of X and Y if

$$f_{X,Y}(x,y)=\begin{cases}\frac{1}{50} & 0<x<10, \quad \frac{x}{2}<y<\frac{x}{2}+5\\ 0 & otherwise\end{cases}$$



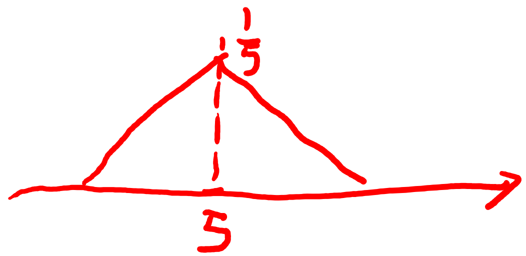
(Solution):

Note that

$$f_X(x)=\int f_{X,Y}(x,y)dy=\frac{1}{50}\int_{x/2}^{x/2+5}dy=\frac{1}{10}$$

$$f_Y(y)=\int f_{X,Y}(x,y)dx=\frac{1}{50}\int_{\max(0,2y-10)}^{\min(10,2y)}dx$$

$$f_Y(y)=\begin{cases}y/25 & for\ 0<y<5\\ (10-y)/25 & for\ 5<y<10\\ 0 & otherwise\end{cases}$$



$$\mu_X = \int x f_X(x)\,dx = \frac{1}{10}\int_0^{10} x\,dx = 5$$

$$\mu_Y = \int y f_Y(y)\,dy = \int_0^5 \frac{y^2}{25}\,dy + \int_5^{10} \frac{10y - y^2}{25}\,dy = 5$$

To determine the covariance,

$$\begin{aligned} cov_{X,Y} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y)\,dx\,dy \\ &= \frac{1}{50}\int_0^{10}\int_{x/2}^{x/2+5} (x-5)(y-5)\,dy\,dx \\ &= \frac{1}{50}\int_0^{10} (x-5)\left(\frac{5x-25}{2}\right)dx \\ &= \frac{25}{6} \end{aligned}$$

To determine the correlation, first, we determine the variance

$$\sigma_X^2 = \int (x-\mu_X)^2 f_X(x)\,dx = \frac{1}{10}\int_0^{10} (x-5)^2\,dx = \frac{25}{3}$$

$$\sigma_Y^2 = \int (y-\mu_Y)^2 f_Y(y)\,dy = \int_0^5 \frac{(y-5)^2 y}{25}\,dy + \int_5^{10} \frac{(y-5)^2(10-y)}{25}\,dy$$

$$= \int_0^5 \frac{(y-5)^2 y}{25}\,dy + \int_5^0 \frac{(y_1-5)^2 y_1}{25}(-dy_1) = 2\int_0^5 \frac{(y-5)^2 y}{25}\,dy$$

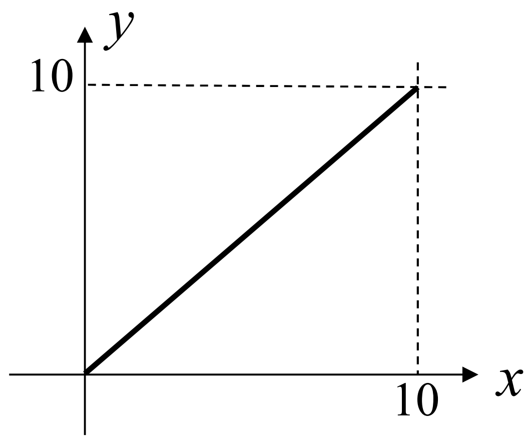
(set y_1 = 10-y)

$$= \frac{25}{6}$$

Therefore, $\sigma_X = \frac{5}{\sqrt{3}}$, $\sigma_Y = \frac{5}{\sqrt{6}}$

$$corr_{X,Y} = \frac{cov_{X,Y}}{\sigma_X \sigma_Y} = \frac{25}{6}\frac{\sqrt{18}}{25} = \frac{1}{\sqrt{2}} = 0.707$$

(X and Y are highly correlated)



[Example 4] Note that if

$$f_{X,Y}(x,y)=\begin{cases}\frac{1}{10}\delta(x-y) & 0<x<10\\ 0 & otherwise\end{cases}$$

then

$\delta(x-y)$ $\quad x-y\to\infty$ if $y=x$ $\quad x-y=0$ if $y\neq x$

$$f_X(x)=\frac{1}{10}\int_{0\ (-\infty)}^{10\ (\infty)}\delta(x-y)\,dy=\frac{1}{10}\quad 0<x<10 \qquad f_Y(y)=\frac{1}{10}\int_{0\ (-\infty)}^{10\ (\infty)}\delta(x-y)\,dx=\frac{1}{10}\quad 0<y<10$$

(page 347(1))

$$\mu_X=\frac{1}{10}\int_0^{10}x\,dx=5 \qquad \mu_Y=\frac{1}{10}\int_0^{10}y\,dy=5$$

$$cov_{X,Y}=\int_0^{10}\int_{-\infty}^{\infty}(x-5)(y-5)\frac{1}{10}\delta(x-y)\,dy\,dx$$

$\delta(y-x)$

(page 348(2))

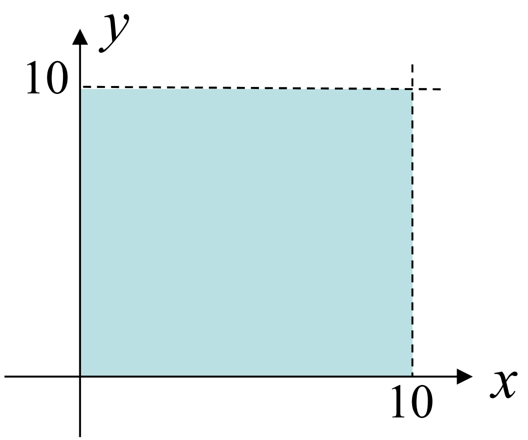
$$=\frac{1}{10}\int_0^{10}(x-5)^2\,dx=\frac{25}{3}$$

$$\sigma_X=\sqrt{\frac{1}{10}\int_0^{10}(x-5)^2\,dx}=\sqrt{\frac{25}{3}} \qquad \sigma_Y=\sqrt{\frac{1}{10}\int_0^{10}(y-5)^2\,dy}=\sqrt{\frac{25}{3}}$$

$$corr_{X,Y}=\frac{cov_{X,Y}}{\sigma_X\sigma_Y}=1$$

[Example 5] If

$$f_{X,Y}(x,y)=\begin{cases}\frac{1}{100} & 0<x<10 \text{ and } 0<y<10 \\ 0 & otherwise\end{cases}$$



then

$$f_X(x)=\int_0^{10}\frac{1}{100}dy=\frac{1}{10} \qquad f_Y(y)=\int_0^{10}\frac{1}{100}dx=\frac{1}{10}$$

$$\mu_X=\frac{1}{10}\int_0^{10}xdx=5 \qquad \mu_Y=\frac{1}{10}\int_0^{10}ydy=5$$

$$cov_{X,Y}=\int_0^{10}\int_0^{10}(x-5)(y-5)\frac{1}{100}dydx=0$$

(odd symmetry with respect to (5, 5))

$$corr_{X,Y}=\frac{cov_{X,Y}}{\sigma_X\sigma_Y}=0$$

9.2 Probability Model

9.2.1 Discrete Probability Model

(1) Uniform Distribution

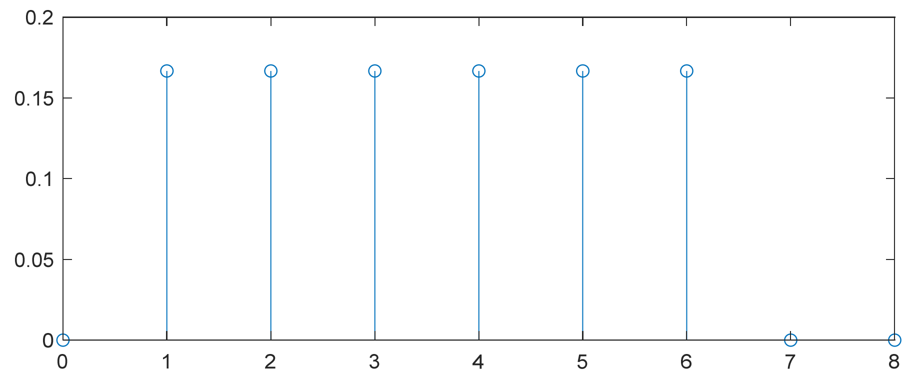
Probability Mass Function (PMF)

$$P_X(n) = \frac{1}{N} \qquad \text{for } n = a,\ a+1,\ \ldots,\ a+N-1$$

Mean: $\mu_X = a + \frac{N-1}{2}$

standard deviation: $\sigma_X = \sqrt{\frac{N^2-1}{12}}$

$$skewness = 0$$



$a = 1, N = 6$

(2) Binomial Distribution

Probability Mass Function (PMF)

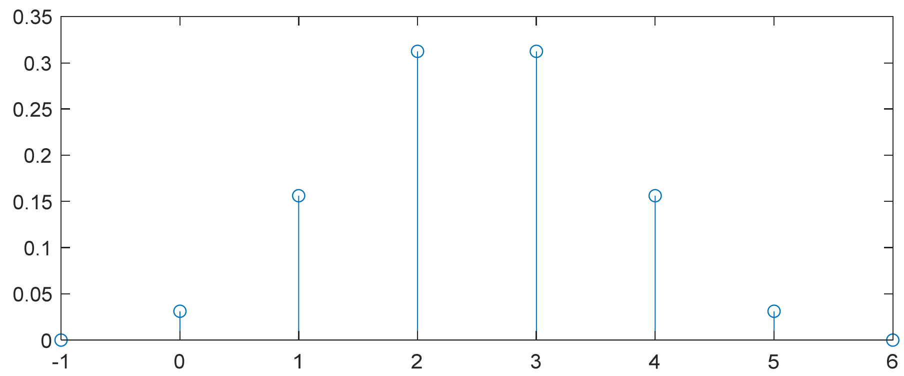
$$P_X(n) = \binom{N}{n} p^n (1-p)^{N-n} \quad \text{for } n = 0, 1, \ldots, N$$

[Physical Meaning]: If we perform a trial N times and for each time the successful rate is p, then $P_X(n)$ is the probability where the number of successful trials is n.

Mean: $\mu_X = Np$

standard deviation: $\sigma_X = \sqrt{Np(1-p)}$

$$skewness = \frac{1-2p}{\sqrt{Np(1-p)}}$$



$p = 1/2,\ N = 5$

When $N = 1$, the binomial distribution is called the Bernoulli distribution.

(3) Geometric Distribution

Probability Mass Function (PMF)

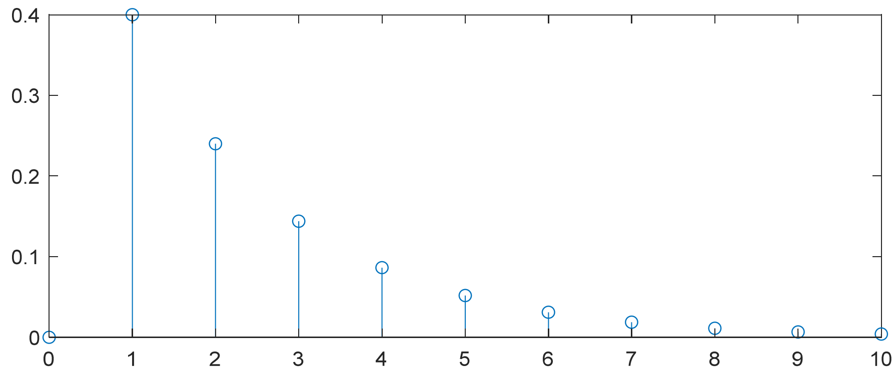
$$P_X(n) = p(1-p)^{n-1} \qquad \text{for } n = 1, 2, 3, \ldots\ldots$$

[Physical Meaning]: If each trial has the successful rate of p, then $P_X(n)$ is the probability where the first successful trial is the n^{th} trial.

Mean: $\mu_X = 1/p$

standard deviation: $\sigma_X = \sqrt{\dfrac{1-p}{p^2}}$

$$skewness = \frac{2-p}{\sqrt{1-p}}$$



$p = 0.4$

(4) Hypergeometric Distribution

Probability Mass Function (PMF)

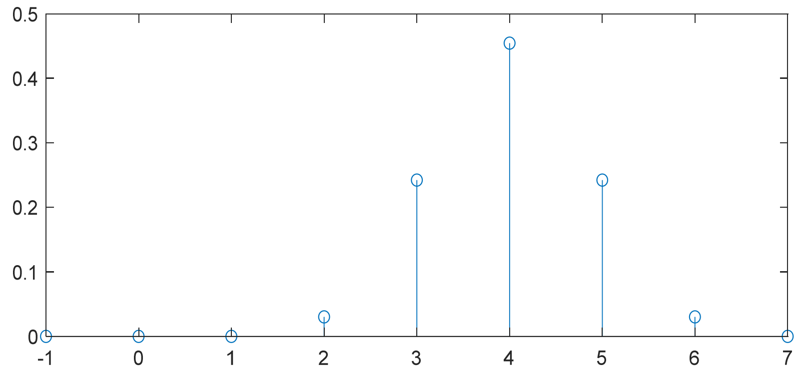
$$P_X(n) = \frac{\binom{K}{n}\binom{N-K}{m-n}}{\binom{N}{m}} \qquad \text{for } n = 0, 1, 2, \ldots\ldots, \min(m, K)$$

[Physical Meaning]: Suppose that there are N balls in a set. There is a subset which contains K balls. If we choose m balls from the set, then $P_X(n)$ means the probability that n of the balls are chosen from the subset.

Mean: $\mu_X = mK / N$

standard deviation: $\sigma_X = \sqrt{\dfrac{mK(N-K)(N-m)}{N^2(N-1)}}$

$$skewness = \frac{\sqrt{N-1}(N-2K)(N-2m)}{\sqrt{mK(N-K)(N-m)(N-2)}}$$



$N = 12, K = 8, m = 6$

(5) Poisson Distribution

Probability Mass Function (PMF)

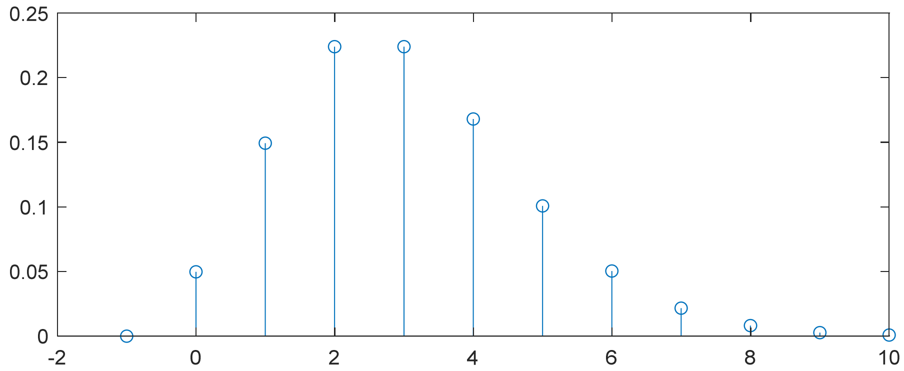
$$P_X(n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

[Physical Meaning]: Suppose that, within a certain time interval, an event will occur λ times in average. Then, $P_X(n)$ indicates the probability that the event occurs n times within the time interval.

mean: $\mu_X = \lambda$

standard deviation: $\sigma_X = \sqrt{\lambda}$

$skewness = 1/\sqrt{\lambda}$



$\lambda = 3$

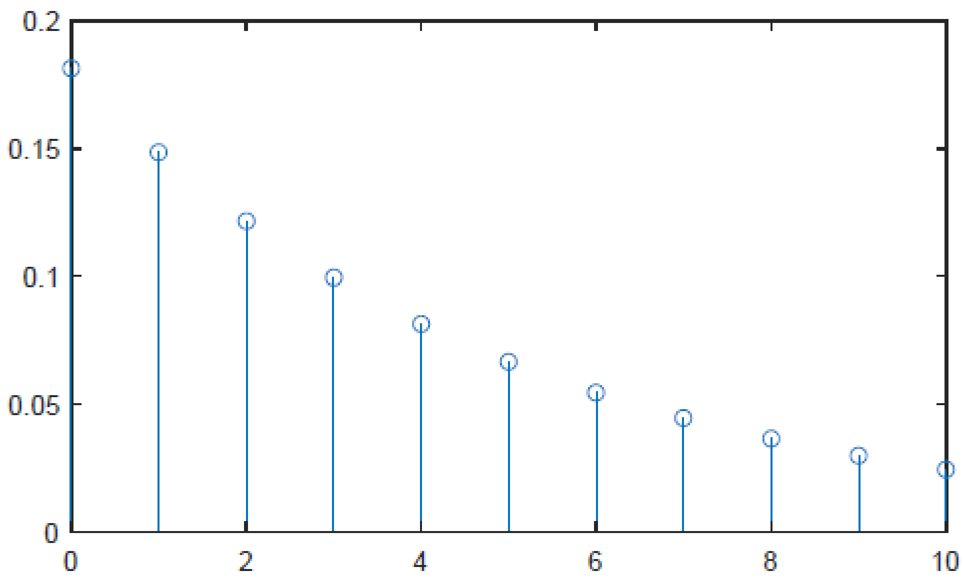
(6) Discrete Exponential Distribution

Probability Mass Function (PMF)

$$P_X(n) = \left(1 - e^{-\lambda}\right)\exp(-\lambda n) \qquad \text{for } n = 0,\ 1,\ 2,\ \ldots\ldots$$

Mean: $\mu_n = \dfrac{e^{-\lambda}}{1 - e^{-\lambda}}$

standard deviation: $\sigma_n = \dfrac{e^{-\lambda/2}}{1 - e^{-\lambda}}$



$\lambda = 2$

9.2.2 Continuous Probability Model

(1) Uniform Distribution

Probability Density Function (PDF)

$$f_X(x) = \frac{1}{b-a} \quad \text{for } a < x < b \qquad f_X(x) = 0 \quad \text{otherwise}$$

mean: $\mu_X = \frac{a+b}{2}$

standard deviation: $\sigma_X = \frac{b-a}{\sqrt{12}}$

$skewness = 0$

(2) Exponential Distribution

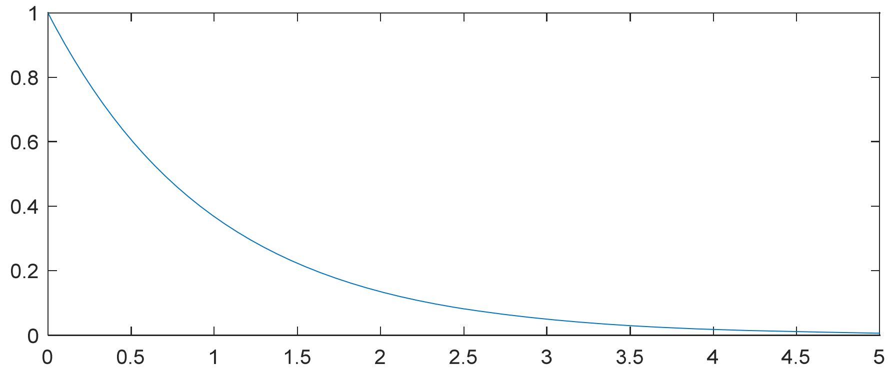
PDF:

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0 \qquad f_X(x) = 0 \quad \text{for } x < 0$$

mean: $\mu_X = \dfrac{1}{\lambda}$

standard deviation: $\sigma_X = \dfrac{1}{\lambda}$

$skewness = 2$



$\lambda = 1$

(3) Normal Distribution (Gaussian Distribution)

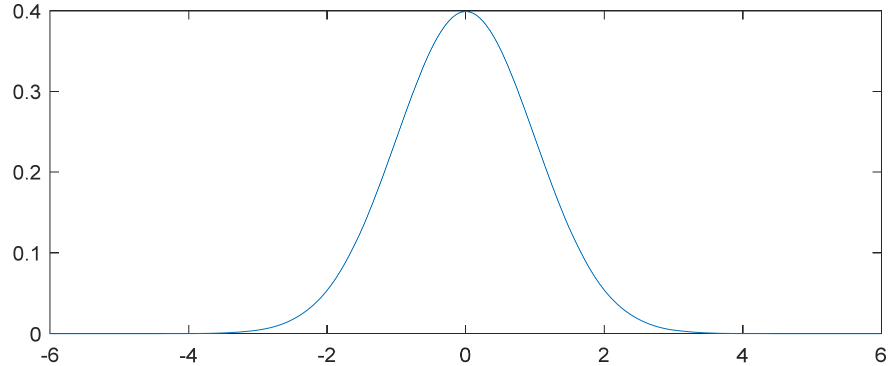
PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

mean: $\mu_X = \mu$

standard deviation: $\sigma_X = \sigma$

$skewness = 0$



$\mu = 0, \quad \sigma = 1$

The normal distribution is the most popular probability distribution.
However, is it reasonable?

Confidence Level (信心水準):

The confidence level is the probability where the data value is within some confidence interval (信賴區間)

$$confidence\ level = Prob\left(\underbrace{a \le X \le b}_{\text{confidence interval}}\right)$$

$$confidence\ level = F_X\left(b\right) - F_X\left(a\right)$$

Some confidence level for the normal distribution,

$$Prob\{|X-\mu|\le\sigma\}=68.2689\%$$

$$Prob\{|X-\mu|\le 2\sigma\}=95.4500\%$$

$$Prob\{|X-\mu|\le 3\sigma\}=99.7300\%$$

$$Prob\{|X-\mu|\le 4\sigma\}=99.9937\%$$

$$Prob\{|X-\mu|\le 5\sigma\}=99.99994\%$$

$$Prob\{|X-\mu|\le 6\sigma\}=99.9999998\%$$

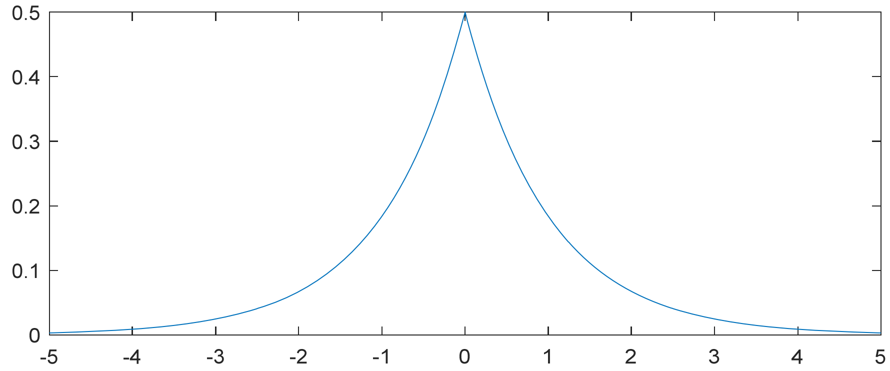
$$Prob\{|X-\mu|\le 7\sigma\}=99.9999999997\%$$

$$Prob\{|X-\mu|> 7\sigma\}=3\cdot 10^{-12}$$

(4) Laplace Distribution

PDF:

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x|}$$



$\lambda = 1$

mean: $\mu_X = 0$

standard deviation: $\sigma_X = \frac{\sqrt{2}}{\lambda}$

$skewness = 0$

(5) Hyper-Laplacian Distribution

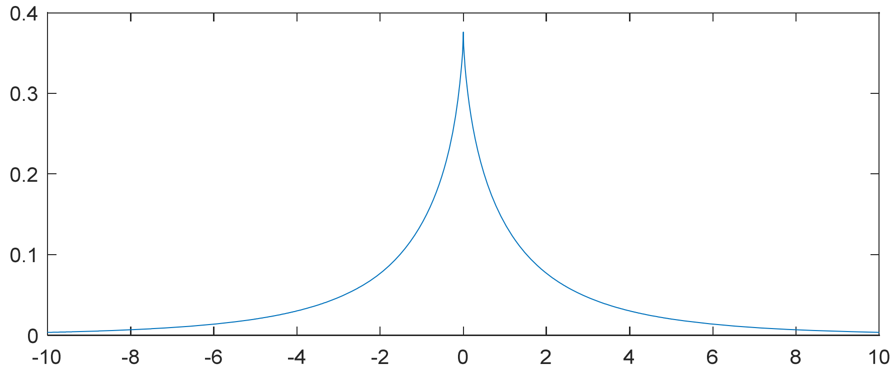
PDF:

$$f_X(x) = Ce^{-\lambda|x|^{\alpha}} \quad \text{where} \quad C = \frac{1}{2\int_0^{\infty} e^{-\lambda x^{\alpha}} dx}$$

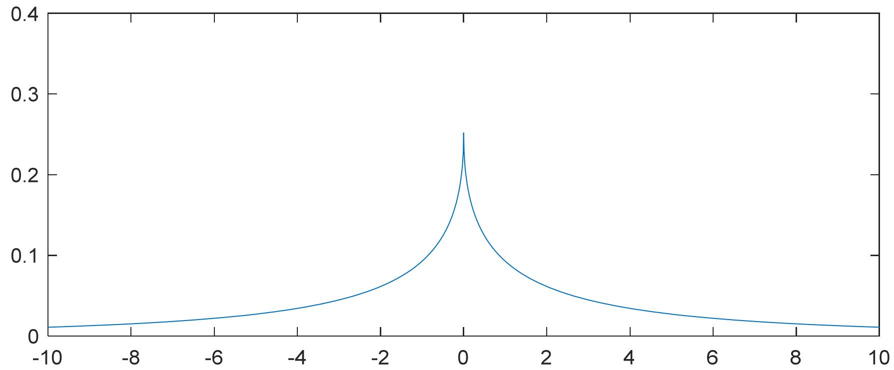
mean: $\mu_X = 0$ standard deviation: decreases with α

$skewness = 0$

$\lambda = 1, \quad \alpha = 2/3$



$\lambda = 1, \quad \alpha = 1/2$



(6) Log-Normal Distribution

PDF:

$$f_X(x) = \frac{1}{x\eta\sqrt{2\pi}} e^{-\frac{(\ln x - u)^2}{2\eta^2}}$$

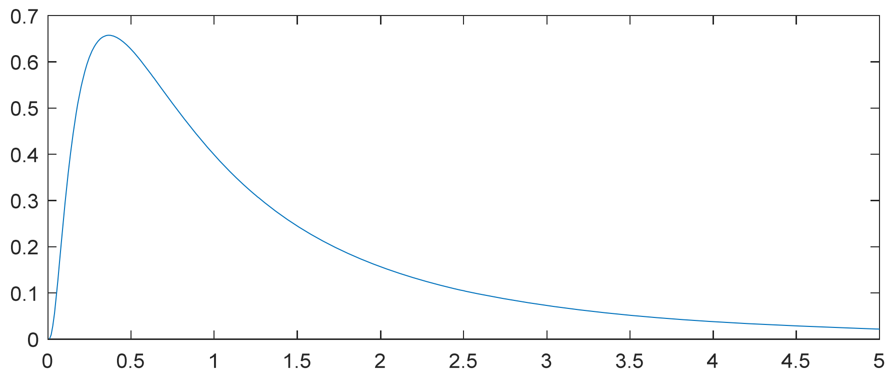
where $x > 0$

mean: $\mu_X = \exp\left(u + \eta^2 / 2\right)$

standard deviation:

$$\sigma_X = \sqrt{e^{\sigma^2} - 1} \exp\left(u + \frac{\eta^2}{2}\right)$$

$$skewness = \sqrt{e^{\sigma^2} - 1}\left(e^{\sigma^2} + 2\right)$$



$\eta = 1, \quad u = 0$

(7) Rayleigh Distribution

PDF:

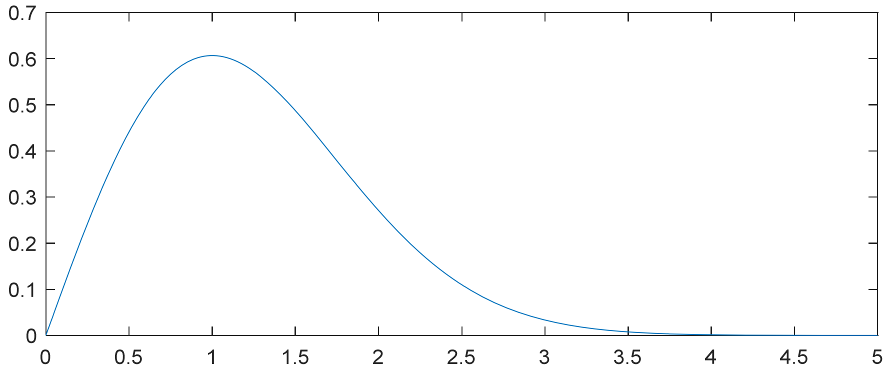
$$f_X(x) = \frac{x}{\eta^2} e^{-\frac{x^2}{2\eta^2}}$$

where $x > 0$

mean: $\mu_X = \eta\sqrt{\frac{\pi}{2}}$

standard deviation: $\sigma_X = \eta\sqrt{\frac{4-\pi}{2}}$

$$skewness = \frac{2\sqrt{\pi}(\pi - 3)}{(4-\pi)^{3/2}}$$



$\eta = 1$

(8) Pareto Distribution

PDF:

$$f_X(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}} \quad \text{when } x > x_0 \qquad f_X(x) = 0 \quad \text{otherwise}$$

where $x_0 > 0,\ \alpha > 0$

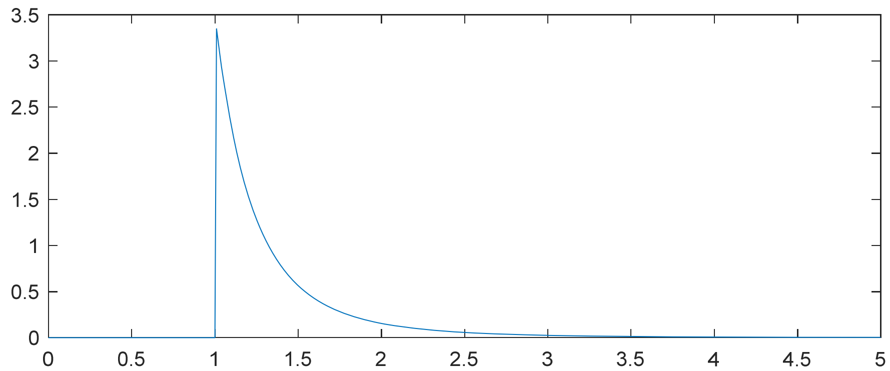
mean: $\mu_X = \dfrac{\alpha x_0}{\alpha - 1}$ when $\alpha > 1$

$\mu_X \to \infty$ when $\alpha \le 1$

standard deviation: $\sigma_X = \dfrac{x_0}{\alpha - 1}\sqrt{\dfrac{\alpha}{\alpha - 2}}$

when $\alpha > 2$

$\sigma_X \to \infty$ when $\alpha \le 2$



$x_0 = 1, \quad \alpha = 3.5$

$$skewness = \frac{2 + 2\alpha}{\alpha - 3}\sqrt{\frac{\alpha - 2}{\alpha}} \quad \text{when } \alpha > 3$$

$skewness \to \infty$ when $\alpha \le 3$

9.3 Entropy (熵[illegible]，亂度)

Discrete Case

Entropy of X can be denoted by $H(X)$

$$Entropy = -\sum_{n} P_X(n) \ln\left[P_X(n)\right]$$

In fact, $\quad Entropy = -E\left(\ln\left[P_X(n)\right]\right)$

Continuous Case

$$Entropy = -\int_{-\infty}^{\infty} f_X(x) \ln\left[f_X(x)\right] dx$$

In fact, $\quad Entropy = -E\left(\ln\left[f_X(x)\right]\right)$

Note:

(1) Since

$$-\ln\left[P_X(n)\right] \geq 0 \qquad -\ln\left[f_X(x)\right] \geq 0$$

we have

$$Entropy \geq 0$$

(2) In some literature, the entropy of X is denoted by

$$H(X)$$

(3) When $P_X(n) = 0$, we can set

$$P_X(n)\ln\left[P_X(n)\right] = 0$$

when calculating the entropy.

[Example 1] If

$$P_X(1) = 1, \quad P_X(n) = 0 \ \ otherwise$$

$-1 \cdot \ln(1) = 0$

entropy: $-\sum_n P_X(n) \ln(P_X(n))$

then

$$H(X) = -1 \cdot \ln(1) = 0$$

[Example 2] If

$$P_X(1) = 0.8, \quad P_X(2) = 0.2, \quad P_X(n) = 0 \ \ otherwise$$

then

$$H(X) = -0.8 \cdot \ln(0.8) - 0.2 \cdot \ln(0.2) = 0.5004$$

[Example 3] If

$$P_X(1) = 0.5, \quad P_X(2) = 0.5, \quad P_X(n) = 0 \ \ otherwise$$

then

$$H(X) = -0.5 \cdot \ln(0.5) - 0.5 \cdot \ln(0.5) = \ln(2) = 0.6931$$

[Example 4] If

$$P_X(1)=0.7,\quad P_X(2)=0.1,\quad P_X(3)=0.1,\quad P_X(4)=0.1,$$

$$P_X(n)=0\ \ otherwise$$

$$H(X)=-0.7\cdot\ln(0.7)-3\left(0.1\cdot\ln(0.1)\right)=0.9404$$

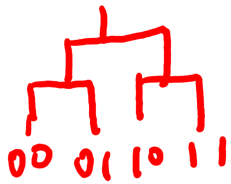
[Example 5] If

$$P_X(1)=P_X(2)=P_X(3)=P_X(4)=0.25,\quad P_X(n)=0\ \ otherwise$$

$$H(X)=-4\left(0.25\cdot\ln(0.25)\right)=1.3863$$

H(X): ln4

$\log_2 H(Y): \frac{\ln 4}{\ln 2} = 2$



Main Applications of Entropy

(a) Thermodynamics (熱力學)

(b) Information Theory

less entropy = more meaningful information

N possible cases
same probabilities
entropy = $N\left(-\frac{1}{N}\ln\left(\frac{1}{N}\right)\right)$
$= \ln N$

(c) Data Compression

$\log_2(entropy)$ = the number of bits for each input

(d) Optimization, Classification, Machine Learning

9.4 Kullback-Leibler Divergence

9.4.1 Definition

The Kullback-Leibler divergence (KL divergence, KL 散度，相對熵) is to determine the difference of two probability distributions.

In the discrete case, suppose that there are two probability distribution $P_X(n)$ and $P_Y(n)$. Then the KL divergence from $P_Y(n)$ to $P_X(n)$ is

$$D_{KL}\left(X \,\|\, Y\right) = \sum_{n} P_X\left(n\right) L_{X,Y}\left(n\right)$$

Approximated probability model ($P_Y(n)$)

Ture probability ($P_X(n)$)

where $L_{X,Y}\left(n\right) = \ln \frac{P_X\left(n\right)}{P_Y\left(n\right)}$ if $P_X(n) \neq 0$

$L_{X,Y}\left(n\right) = 0$ if $P_X(n) = 0$

$$D_{KL}(X \| Y) = \sum_n P_X(n) L_{X,Y}(n) \qquad L_{X,Y}(n) = \ln\frac{P_X(n)}{P_Y(n)} \quad \text{if } P_X(n) \neq 0$$

$$L_{X,Y}(n) = 0 \quad \text{if } P_X(n) = 0$$

Note:

(1) If $P_X(n) = P_Y(n)$ for all n, then

$$D_{KL}(X \| Y) = 0$$

(2) If it exists some n such that $P_Y(n) = 0$ but $P_X(n) \neq 0$, then

$$D_{KL}(X \| Y) \to \infty$$

(3) In fact,

$$D_{KL}(X \| Y) = -\sum_n P_X(n) \ln P_Y(n) - H(X)$$

(4) In usual, X is the true probability and Y is the probability model.

$$D_{KL}(X \| Y) = \sum_n P_X(n) L_{X,Y}(n) \qquad L_{X,Y}(n) = \ln \frac{P_X(n)}{P_Y(n)} \quad \text{if } P_X(n) \neq 0$$

$$L_{X,Y}(n) = 0 \quad \text{if } P_X(n) = 0$$

Note:

(5) In fact,

$$D_{KL}(X \| Y) \neq D_{KL}(Y \| X)$$

(6)

$$D_{KL}(X \| Y) \geq 0$$

In the continuous case, suppose that there are two probability distribution $f_X(x)$ and $f_Y(x)$. Then the KL divergence from $f_Y(x)$ to $f_X(x)$ is

$$D_{KL}(X \| Y) = \int f_X(x) L_{X,Y}(x)\,dx$$

where

$$L_{X,Y}(x) = \ln \frac{f_X(x)}{f_Y(x)} \qquad \text{if } f_X(x) \neq 0$$

$$L_{X,Y}(x) = 0 \qquad \text{if } f_X(x) = 0$$

Note: The properties of the KL divergence in the continuous case are the same those in the discrete case.

[Example 1] Suppose that

$$P_X(1)=0.15,\quad P_X(2)=0.15,\quad P_X(3)=0.35,\quad P_X(4)=0.35,$$

$$P_X(n)=0\quad otherwise$$

$$P_Y(n)=0.25\quad for\ n=1,2,3,4\qquad P_Y(n)=0\quad otherwise$$

$$P_Z(1)=0.1,\quad P_Z(2)=0.2,\quad P_Z(3)=0.3,\quad P_Z(4)=0.4,$$

$$P_Z(n)=0\quad otherwise$$

Determine the KL divergences from Y to X and from Z to X.

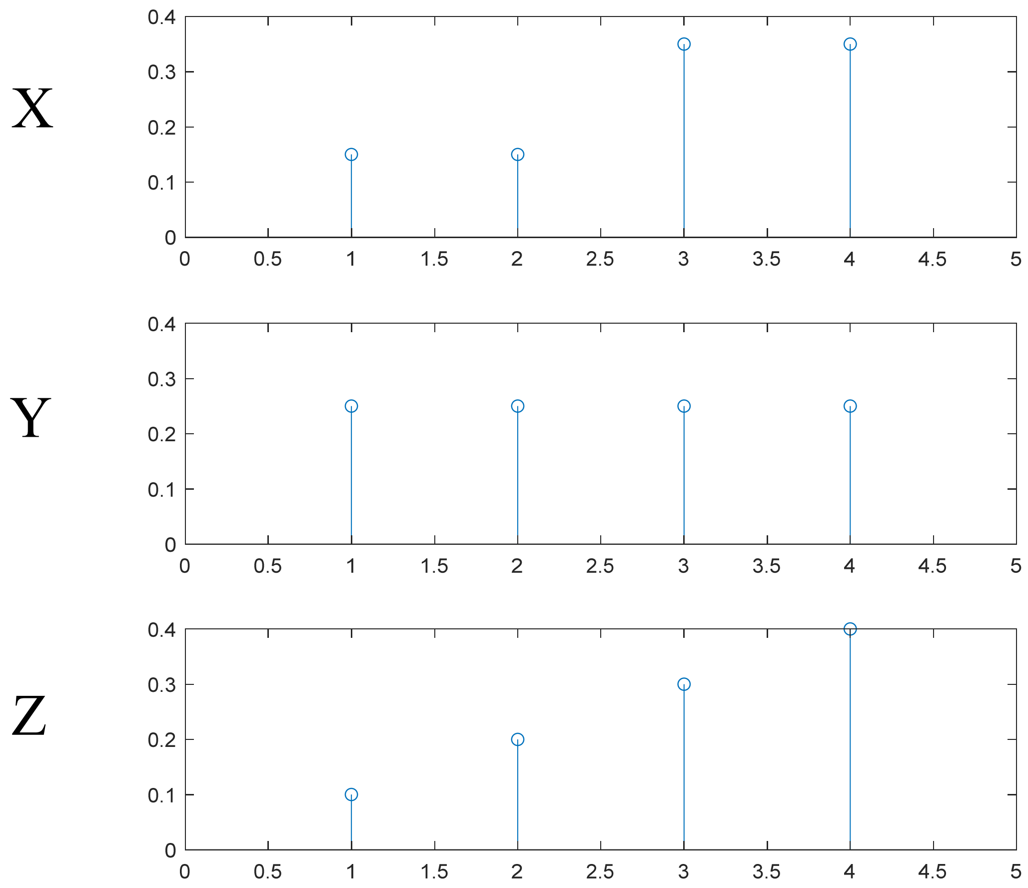
(Solution):

$$D_{KL}(X\,\|\,Y)=0.15\ln\frac{0.15}{0.25}+0.15\ln\frac{0.15}{0.25}+0.35\ln\frac{0.35}{0.25}+0.35\ln\frac{0.35}{0.25}$$

$$=0.0823$$

$$D_{KL}(X\,\|\,Z)=0.15\ln\frac{0.15}{0.1}+0.15\ln\frac{0.15}{0.2}+0.35\ln\frac{0.35}{0.3}+0.35\ln\frac{0.35}{0.4}$$

$$=0.0249$$



$P_X(n)$ is more similar to $P_Z(n)$ than $P_Y(n)$.

[Example 2] Suppose that X, Y, and Z distributes the same as those in Example 1.

Determine the KL divergences from X to Y and from X to Z.

(Solution):

$$D_{KL}(Y \| X) = 0.25 \ln \frac{0.25}{0.15} + 0.25 \ln \frac{0.25}{0.15} + 0.25 \ln \frac{0.25}{0.35} + 0.25 \ln \frac{0.25}{0.35}$$

$$= 0.0872$$

$$D_{KL}(Z \| X) = 0.1 \ln \frac{0.1}{0.15} + 0.2 \ln \frac{0.2}{0.15} + 0.3 \ln \frac{0.3}{0.35} + 0.4 \ln \frac{0.4}{0.35}$$

$$= 0.0242$$

Note that

$$D_{KL}(Y \| X) \neq D_{KL}(X \| Y)$$

$$D_{KL}(Z \| X) \neq D_{KL}(X \| Z)$$

[Example 3] Suppose that

$$f_X(x) = \frac{1}{10} \quad for\ |x| \le 5 \qquad \text{(uniform distribution)}$$

$$f_Y(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \qquad \text{(normal distribution with zero mean)}$$

Determine σ such that $f_Y(x)$ is most similar to $f_X(x)$

(Solution):

$D_{KL}(X \| Y) = 2.7830$ when $\sigma = 1$

$D_{KL}(X \| Y) = 0.3511$ when $\sigma = 2$

$D_{KL}(X \| Y) = 0.1779$ when $\sigma = 3$

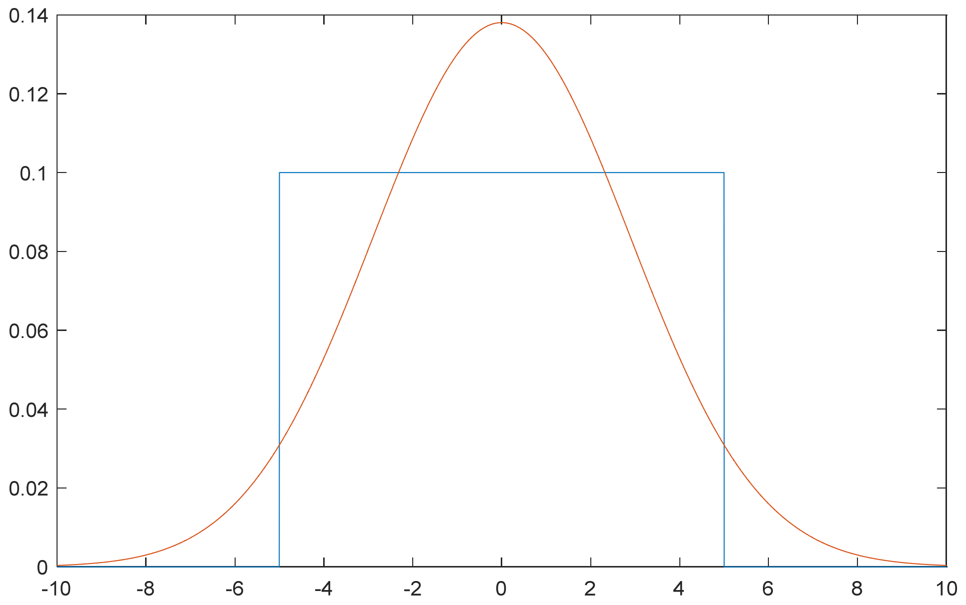
$D_{KL}(X \| Y) = 0.2630$ when $\sigma = 4$

$$D_{KL}\left(X \,\|\, Y\right) = 0.2092 \qquad \text{when } \sigma = 3.5$$

$$D_{KL}\left(X \,\|\, Y\right) = 0.1992 \qquad \text{when } \sigma = 2.5$$

$$\vdots$$

$$D_{KL}\left(X \,\|\, Y\right) = 0.176417 \qquad \text{when } \sigma = 2.89$$



9.4.2 Cross Entropy

[Cross Entropy]

$$H(X,Y) = D_{KL}(X \| Y) + H(X)$$

Since $D_{KL}(X \| Y) = \sum_n P_X(n) \ln \dfrac{P_X(n)}{P_Y(n)}$

$$H(X) = -\sum_n P_X(n) \ln\left(P_X(n)\right) \qquad \text{if } P_X(n) \neq 0$$

$$H(X,Y) = \sum_n P_X(n) \left[\ln\left[\frac{P_X(n)}{P_Y(n)}\right] - \ln\left(P_X(n)\right) \right]$$

$$H(X,Y) = -\sum_n P_X(n)\left[\ln P_Y(n)\right] \qquad \text{(discrete case)}$$

$$H(X,Y) = -\int P_X(x)\left[\ln P_Y(x)\right] dx \qquad \text{(continuous case)}$$

Note:

(1) $H(X,Y) \neq H(Y,X)$

(2) $H(X,Y) \geq H(X)$

(3) If it happen that $P_Y(x) = 0$ but $P_X(x) \neq 0$ for some x, then

$$H(X,Y) \to \infty$$

[Second Definition of the Cross Entropy]

Suppose that both X and Y are both Bernoulli distribution

$$P_X(0) = 1 - q, \quad P_X(1) = q$$

$$P_{Y|X}(0|0) = 1 - p_1, \quad P_{Y|X}(1|0) = p_1$$

$$P_{Y|X}(0|1) = 1 - p_2, \quad P_{Y|X}(1|1) = p_2$$

Then the cross-entropy of X and Y is

$$H(X,Y) = -P_X(0)\log\left[P_{Y|X}(0,0)\right] - P_X(1)\log\left[P_{Y|X}(1,1)\right]$$

$$= -(1-q)\log(1-p_1) - q\log(p_2)$$

In general,

$$H(X,Y) = -\sum_n P_X(n)\log\left[P_{Y|X}(n,n)\right]$$

This definition is often used in machine learning and classification, but it is different from the standard one.

Note: When applying the second definition of the cross entropy,

(1) $H(X,Y) \neq H(Y,X)$

(2) $H(X,Y) \geq 0$

(3) If $P_{Y|X}(n,n) = 1 \quad \textit{for all } n$, i.e.,

$P_{Y|X}(m,n) = 0 \quad \textit{for } m \neq n$

then $H(X,Y) = 0$

(4) If $P_{Y|X}(n,n) = 0 \quad \textit{for some } n$

then $H(X,Y) \to \infty$

[Example 3] Determine the cross entropy of X and Y if

$P_{X,Y}(n, m)$	X=0	X=1
Y=0	0.3	0.1
Y=1	0.2	0.4

(Solution): When using the definition on page 767, since

$$P_X(0) = P_X(1) = 0.5, \qquad P_Y(0) = 0.4, \quad P_Y(1) = 0.6$$

we have

$$H(X,Y) = -0.5\ln(0.4) - 0.5\ln(0.6) = 0.7136$$

When using the definition on page 769, since

$$P_{Y|X}(0|0) = 0.6 \quad \frac{0.3}{0.3+0.2} \qquad P_{Y|X}(1|1) = 0.8 \quad \frac{0.4}{0.1+0.4}$$

$$H(X,Y) = -0.5\ln(0.6) - 0.5\ln(0.8) = 0.3670$$

Section 9.5 Basic Concepts of Random Process

9.5.1 Definition

In the case where the data is not known explicitly (e.g., the noise, the future information, an uncertain data, or a non-fixed signal), then it is a random process.

If the data has a fixed form, then it is a deterministic signal.

For example,

$$X(t)=1 \qquad X(t)=\cos t \qquad X(t)=t^{3/4}$$

are all deterministic signals, and

1st measurement $X(t,1)=x_1(t)$

2nd measurement $X(t,2)=x_2(t)$

$\vdots \qquad \vdots$

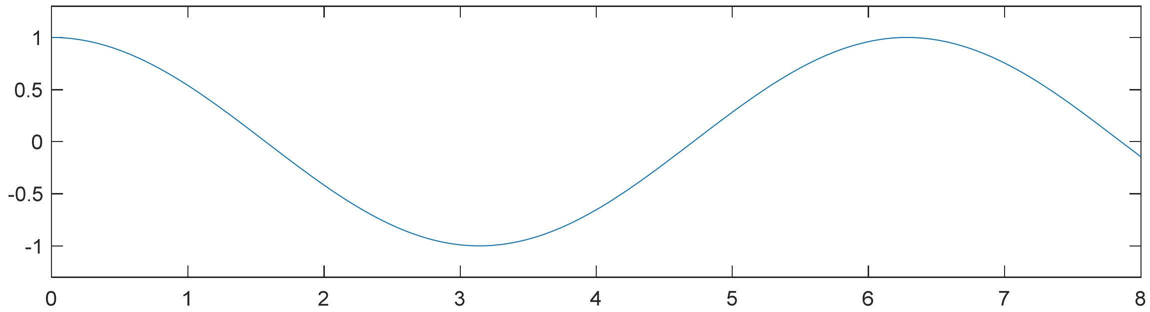
Mth measurement $X(t,M)=x_M(t)$

where $X(t, m)$ is the mth measuring result for $X(t)$

is a random process.

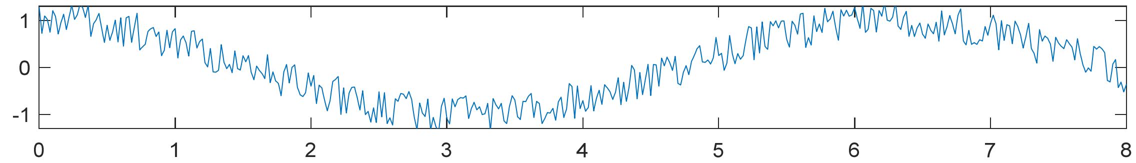
deterministic signal

$\cos(t)$

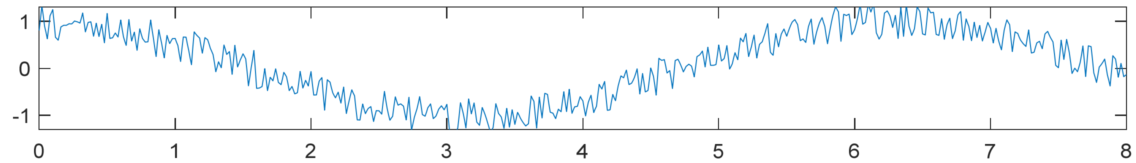


random process

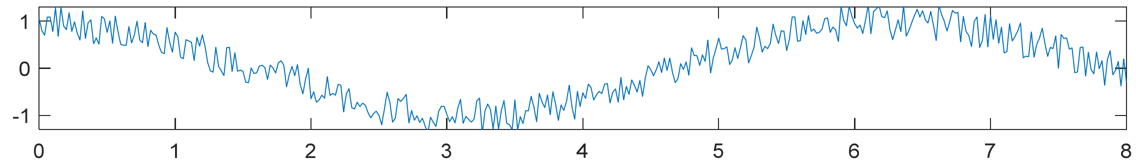
1st measurement



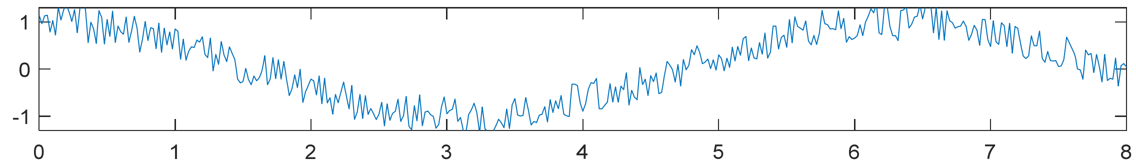
2nd measurement



3rd measurement



4th measurement



One cannot use a function to express a random process explicitly. Instead, one often use some metrics related to probability to express the random process.

(1) Mean $\mu_X(t) = mean[X(t)] = \frac{1}{M}\sum_{m=1}^{M} X(t,m)$

m: the m^{th} time of measurement

Note: (i) On page 773, the random process has the mean of $\mu_X(t) = \cos(t)$.

(ii) For a pure noise, it is usually assumed that $\mu_X(t) = 0$ for all t.

(2) Variance $\sigma_X^2(t) = \text{var}[X(t)] = \frac{1}{M}\sum_{m=1}^{M} \left(X(t,m) - \mu_X(t)\right)^2$

(3) Standard Deviation: $\sigma_X(t) = \sqrt{\sigma_X^2(t)} = \sqrt{\frac{1}{M}\sum_{m=1}^{M} \left(X(t,m) - \mu_X(t)\right)^2}$

(4) Auto-Covariance:

$$cov_X(t,t_1) = cov[X(t), X(t_1)]$$
$$= \frac{1}{M}\sum_{m=1}^{M}\left(X(t,m)-\mu_X(t)\right)\left(X(t_1,m)-\mu_X(t_1)\right)$$

(5) Auto-Correlation:

In many literature, the auto-variance is also called the auto-correlation.

However, its standard definition should be

$$corr_X(t,t_1) = corr[X(t), X(t_1)] = \frac{cov[X(t), X(t_1)]}{\sigma_X(t)\sigma_X(t_1)}$$

(6) Joint Probability

$$P_X\left(t_1,t_2,\cdots,t_N;c_1,c_2,\cdots,c_N\right)=Prob\left\{X\left(t_1\right)=c_1,\ X\left(t_2\right)=c_2,.....,X\left(t_N\right)=c_N\right\}$$

$$P_X\left(t;c\right)=Prob\left\{X\left(t\right)=c\right\}$$

9.5.2 Stationary Random Process

Suppose that

$$P_X(t_1,t_2,\cdots,t_N;c_1,c_2,\cdots,c_N) = Prob\{X(t_1)=c_1,\ X(t_2)=c_2,\ .....,\ X(t_N)=c_N\}$$

(joint probability for the random process)

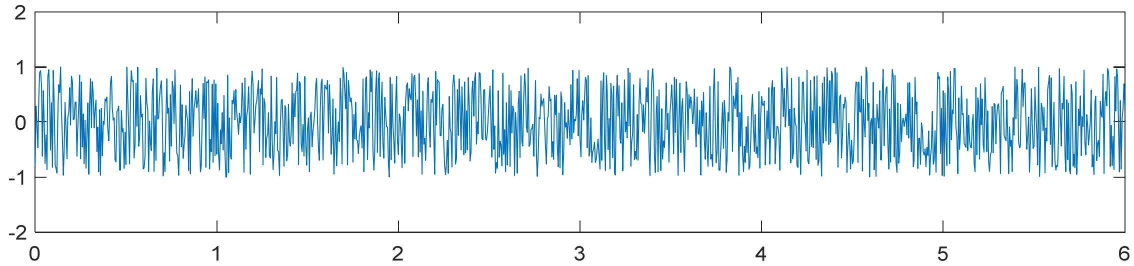
If

$$P_X(t_1,t_2,\cdots,t_N;c_1,c_2,\cdots,c_N) = P_X(t_1+\tau,t_2+\tau,\cdots,t_N+\tau;c_1,c_2,\cdots,c_N)$$

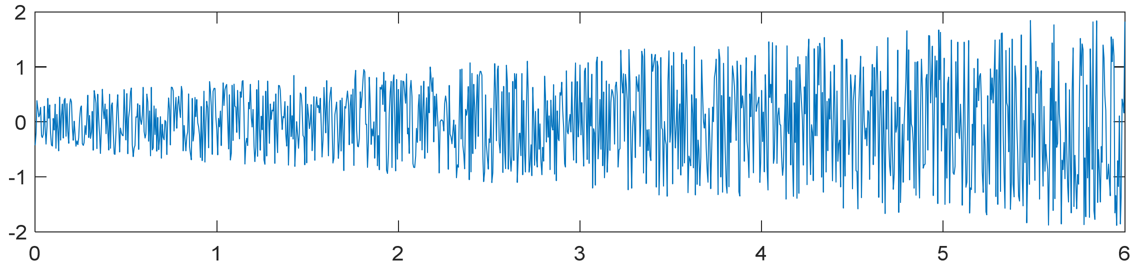
for all possible $\tau,t_1,t_2,\cdots,t_N;c_1,c_2,\cdots,c_N,N$

then the random process is called a strict-sense stationary random process (also called the strictly stationary random process or the strongly stationary random process).

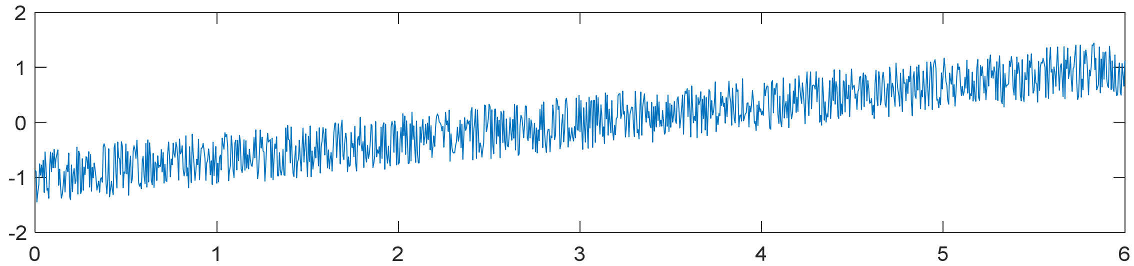
stationary



non-stationary



non-stationary



K^{th} Order Stationary

If

$$P_X\left(t_1,t_2,\cdots,t_N;c_1,c_2,\cdots,c_N\right)=P_X\left(t_1+\tau,t_2+\tau,\cdots,t_N+\tau;c_1,c_2,\cdots,c_N\right)$$

for all possible $\tau,t_1,t_2,\cdots,t_N;c_1,c_2,\cdots,c_N$ and $N = 1, 2, \ldots, K$,

then X is a K^{th} order stationary random process.

2nd Order Stationary (WSS)

Specially, when $\boldsymbol{K=2}$, then X is a 2nd order stationary random process.

The definition of the 2nd order stationary random process is similar to that of the wide sense stationary random process.

Wide-Sense Stationary (WSS) Random Process

(1) $\mu_X(t), \sigma_X(t), \sigma_X^2(t)$ are all constants and can be denoted by

$$\mu_X, \sigma_X, \sigma_X^2$$

(2) $$cov_X(t, t+\tau) = cov_X(t_1, t_1+\tau)$$

In fact, we can replace $cov_X(t_a, t_b)$ by $cov_X(t_b - t_a)$

$$cov_X(t_a, t_b) = R_X(t_b - t_a)$$

since the covariance is only dependent on the difference of t_a and t_b.

(3) $E\left\{|X(t)|^2\right\} < \infty$ for all t.

The wide-sense stationary random process is also called the weak-sense stationary or covariance stationary random process.

Most random processes can be expressed by the addition (or multiplication) of a stationary random process with a deterministic signal.

For example, for the random process on page 773,

$$X(t) = \cos t + n(t)$$

stationary random process

For the 2nd random process on page 778,

$$X(t) = (1 + t/2)\,n(t)$$

For the 3rd random process on page 779,

$$X(t) = -1 + t/3 + n(t)$$

9.5.3 White Noise

Power Spectral Density (PSD)

$$S_X(f,t) = \mathcal{F}_{\tau \to f}\left[cov_X(t,t+\tau)\right] = \int_{-\infty}^{\infty} cov_X(t,t+\tau) e^{-j2\pi f\tau} d\tau$$

If X is a WSS random process and

$$R_X(\tau) = cov_X(t,t+\tau) = E\Big(\big(X(t) - \mu_X(t)\big)\big(X(t+\tau) - \mu_X(t+\tau)\big)\Big)$$

$$= \frac{1}{M}\sum_{m=1}^{M}\big(X(t,m) - \mu_X(t)\big)\big(X(t+\tau,m) - \mu_X(t+\tau)\big)$$

then the Power Spectral Density (PSD) of X is

$$S_X(f) = \mathcal{F}_{\tau \to f}\left[R_X(\tau)\right] = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau} d\tau$$

White Noise

The white noise is a WSS random process where

$$R_X(\tau) = c\delta(\tau) \qquad c = E\left(X^2(t)\right)$$

$$\mu_X = 0$$

In other words,

$$cov_X(t, t+\tau) = E\left[X(t)X(t+\tau)\right] = 0$$

when $\tau \neq 0$. It means that the noises at different time are uncorrelated.

The PSD of a white noise is

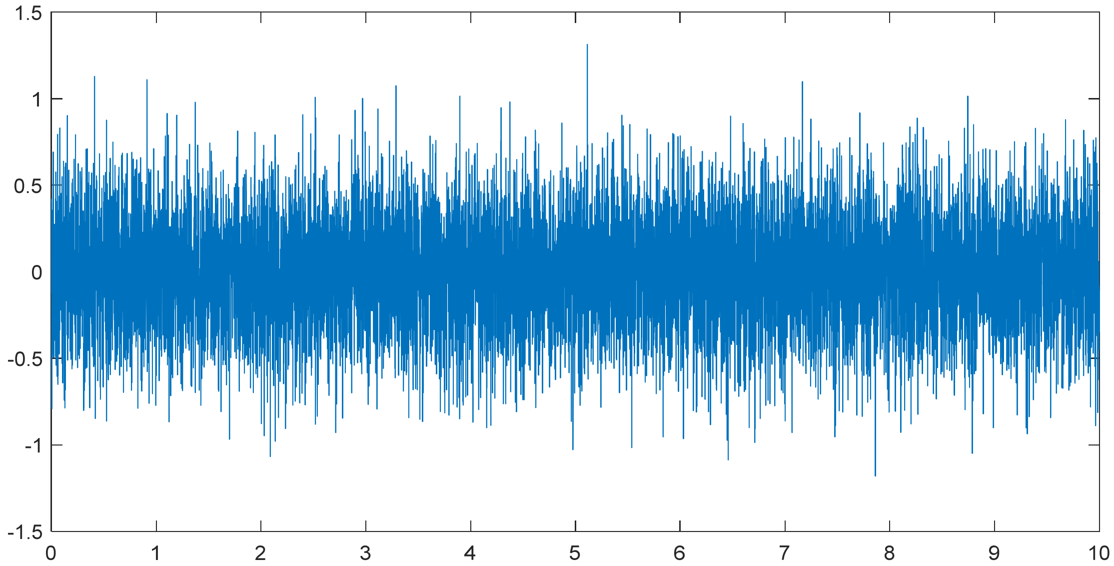
$$S_X(f) = c$$

Additive White Gaussian Noise (AWGN)

For a white noise X, if it is additive and its probability density function (PDF) is a Gaussian distribution with zero mean:

$$P_X(x) = \text{Prob}(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

then X is an additive white Gaussian noise (AWGN).



an AWGN with $\sigma = \pi/10$

Section 9.6 Independent Component Analysis

Suppose that $s_1, s_2, \ldots, s_N$ are sources and $y_1, y_2, \ldots, y_N$ are outputs. If

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix}$$

If $a_{1,1}, a_{1,2}, \ldots, a_{N,N}$ are all known, then one can recover the sources $s_1, s_2, \ldots, s_N$ from the received signals $y_1, y_2, \ldots, y_N$:

$$\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N} \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Q: However, how do we recover the sources s_1, s_2, …, s_N if $a_{1,1}$, $a_{1,2}$, …, $a_{N,N}$ are all unknown?

We can make an assumption that the sources s_1, s_2, …, s_N are independent and have zero correlations:

$$cov(s_n, s_k) = \frac{1}{M}\sum_{m=1}^{M}(s_n[m]-\mu_n)(s_k[m]-\mu_k) = 0 \qquad \text{if } n \neq k$$

where

$$\mu_n = \frac{1}{M}\sum_{m=1}^{M} s_n[m], \quad \mu_k = \frac{1}{M}\sum_{m=1}^{M} s_k[m]$$

$s_n[m]$, $s_k[m]$ means the values of s_n, s_k for the m^{th} measurement

Based on the above assumption, we perform independent components analysis (ICA).

$$cov(s_n, s_k) = E\left[(s_n - \mu_n)(s_k - \mu_k)\right] = 0 \qquad \text{if } n \neq k$$

Therefore,

$$E\left[(\mathbf{s} - \mathbf{s_0})(\mathbf{s} - \mathbf{s_0})^T\right] = \mathbf{D}$$

where

$$\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \cdots & s_N \end{bmatrix}^T \qquad \mathbf{s_0} = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_N \end{bmatrix}^T \qquad \mu_n = E(s_n)$$

$$\mathbf{D} = E\left[(\mathbf{s} - \mathbf{s_0})(\mathbf{s} - \mathbf{s_0})^T\right]$$

$$= \begin{bmatrix} E[(s_1-\mu_1)(s_1-\mu_1)] & E[(s_1-\mu_1)(s_2-\mu_2)] & \cdots & E[(s_1-\mu_1)(s_N-\mu_N)] \\ E[(s_2-\mu_2)(s_1-\mu_1)] & E[(s_2-\mu_2)(s_2-\mu_2)] & \cdots & E[(s_2-\mu_2)(s_N-\mu_N)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(s_N-\mu_N)(s_1-\mu_1)] & E[(s_N-\mu_N)(s_2-\mu_2)] & \cdots & E[(s_N-\mu_N)(s_N-\mu_N)] \end{bmatrix}$$

D is a diagonal matrix:

$$D[n,k] = cov(s_n, s_k)$$

$$= E\left[(s_n - \mu_n)(s_k - \mu_k)\right] = 0 \qquad \text{if } n \neq k$$

$$D[n,n] = cov(s_n, s_n)$$

$$= E\left[(s_n - \mu_n)^2\right]$$

If

$$\mathbf{y} = \mathbf{A}\mathbf{s}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^T \qquad \mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N} \end{bmatrix}$$

then

$$E[\mathbf{y}] = E[\mathbf{A}\mathbf{s}] = \mathbf{A}E[\mathbf{s}] = \mathbf{A}\mathbf{s}_\mathbf{0}$$

where

$$E(\mathbf{y}) = \begin{bmatrix} E(y_1) & E(y_2) & \cdots & E(y_N) \end{bmatrix}^T$$

If we set $\mathbf{Y} = E\left[(\mathbf{y}-\mathbf{y_0})(\mathbf{y}-\mathbf{y_0})^T\right]$ $\qquad \mathbf{y_0} = E[\mathbf{y}]$

then

$$\mathbf{Y} = E\left[(\mathbf{A}\mathbf{s}-\mathbf{A}\mathbf{s}_\mathbf{0})(\mathbf{A}\mathbf{s}-\mathbf{A}\mathbf{s}_\mathbf{0})^T\right]$$

$$= \mathbf{A}E\left[(\mathbf{s}-\mathbf{s}_\mathbf{0})(\mathbf{s}-\mathbf{s}_\mathbf{0})^T\right]\mathbf{A}^T = \mathbf{A}\mathbf{D}\mathbf{A}^T$$

Therefore,

$$\mathbf{Y} = E\left[(\mathbf{y}-\mathbf{y_0})(\mathbf{y}-\mathbf{y_0})^T\right] = \mathbf{ADA}^T$$

where $\mathbf{D} = E\left[(\mathbf{s}-\mathbf{s_0})(\mathbf{s}-\mathbf{s_0})^T\right]$

A **D**

It is similar to eigenvector-eigenvalue decomposition.

Note that since $E\left[(\mathbf{y}-\mathbf{y_0})(\mathbf{y}-\mathbf{y_0})^T\right]$ is a symmetric matrix,

the eigenvectors of $E\left[(\mathbf{y}-\mathbf{y_0})(\mathbf{y}-\mathbf{y_0})^T\right]$ can form a

complete and orthonormal set.

$$\mathbf{A}^{-1} = \mathbf{A}^T$$

Process of Independent Component Analysis

Suppose that we have measured the outputs $\mathbf{y} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}}$ many times.

(1) Construct an NxN covariance matrix $\mathbf{Y}$ where

$$\mathbf{Y} = E\left[(\mathbf{y}-\mathbf{y_0})(\mathbf{y}-\mathbf{y_0})^T\right]$$

i.e., $\mathbf{Y}[n,k] = E[(y_n-\rho_n)(y_k-\rho_k)]$ $\qquad \rho_n = E[y_n]$

(2) Perform eigenvector-eigenvalue decomposition for $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{A_1 D_1 A_1^T}$$

where each column of $\mathbf{A_1}$ is an eigenvector of $\mathbf{Y}$ (with normalization) and the diagonal entries of the diagonal matrix $\mathbf{D_1}$ are the corresponding eigenvalues.

(3) The independent components can be reconstructed from $\mathbf{c} = \mathbf{A_1^T y}$

(**c** may not be the same as **s**)

[Example 1] Suppose that

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

Suppose that s_1 and s_2 are two independent AGWNs with $\sigma = 0.3409$ and 0.5504, respectively (simulated by Matlab programs).

Also suppose that we measure y_1 and y_2 300 times and obtain

$$\mathbf{Y} = E\left[(\mathbf{y} - \mathbf{y_0})(\mathbf{y} - \mathbf{y_0})^T\right] = \begin{bmatrix} 0.7465 & -0.6472 \\ -0.6472 & 2.8785 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 0.9630 & -0.2694 \\ 0.2694 & 0.9630 \end{bmatrix} \begin{bmatrix} 0.5655 & 0 \\ 0 & 3.0596 \end{bmatrix} \begin{bmatrix} 0.9630 & -0.2694 \\ 0.2694 & 0.9630 \end{bmatrix}^T$$

Therefore, $\mathbf{A_1} = \begin{bmatrix} 0.9630 & -0.2694 \\ 0.2694 & 0.9630 \end{bmatrix}$

Note: $\mathbf{A_1} = \begin{bmatrix} 0.9630 & -0.2694 \\ 0.2694 & 0.9630 \end{bmatrix}$ is unequal to $\begin{bmatrix} 2 & -1 \\ 1 & 3 \end{bmatrix}$

Then, the independent components can be reconstructed from

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathbf{A_1^T y}$$

Note that although c_1 and c_2 are not equal to s_1 and s_2 but

$$cov(c_1, c_2) = E\left[\left(c_1 - E(c_1)\right)\left(c_2 - E(c_2)\right)\right] = 0$$

Note that

(1) The ICA is very similar to the PCA.
(compared to page 697)
The difference is that ICA is to find independent components.
PCA is to select principle components.

(2) The independent component set obtained by the above ICA process is not the only solution for independent component separation.

Happy Summer Vacation!

祝各位學習與研究順利，暑假愉快！