Lecture Notes - 18

24-Sep-09

Suppose you know that on an average 500 customers visit a supermarket everyday. Also suppose it is known that every customer spends Rs. 200 on an average in each of his visit. Now can we compute the average money received by the supermarket in a day (i.e. average money spent by all customers in a day)?

Any school kid will say the answer is Rs. $200 \times 500 = 100000$

Now let us try to answer this question using the theory (r.v.) we learnt till now:

Suppose X_i is the amount of money spent by customer 'i' in a visit.

We are given that $E[X_i] = 200 \quad \forall i$. ~~Now consider~~

~~$X = X_1 + X_2 + \cdots + X_N$~~

Let N be the no. of customers visiting in a day. We are given that $E[N] = 500$. Now consider a new r.v.:

$$X = X_1 + X_2 + \cdots + X_N$$ (sum of random number of random variables!)

In words, X is the total money received by supermarket in a day. We want to compute $E[X]$.

i.e. $E\left[\sum_{i=1}^{N} X_i\right]$.

This indeed seems to be a very difficult problem. How do you find exp. of sum of random number of random variables? Moreover too we are not given the distributions of X_i & N !!

But somehow a school kid knows:

$$E\{X\} = E\left\{\sum_{i=1}^{N} X_i\right\} \overset{\text{??}}{\underset{\text{How?}}{=}} E\{N\}\,E\{X_i\} = 500 \times 200 = 100000$$

(Though we are, seemingly not as clever as the kid, let us try to prove his statement).

With a little thought we would realize that if it were exp sum of fixed number of r.v's then we know the answer is sum of expectations of r.v. So can we exploit this? In particular can we (say) condition on N (i.e. assume $N = n$ say)?

i.e. Consider the random variable $X/_{N=n}$.

we know $X/_{N=n}$ is nothing but $\sum_{i=1}^{n} X_i$

So $E\{X/_{N=n}\} = E\left[\sum_{i=1}^{n} X_i\right] = n\,E\{X_i\} = 200n$

Now expectation seems to be something dependent on the value taken by the conditioned rv. N !!

In other words, we can view $E[X/N=n]$ as a function of the random variable N!

How? →

Consider $g(n) = 200n$

and $g(N) = 200N$

This is exactly the quantity $E[X/N=n]$

→ We denote this random variable as $E[X/N]$

In summary,

$E[X/N]$ is a random variable which takes on values $E[X/N=n]$ $\forall$ n (n is the values taken by N).

$E[X/N]$ is called the conditional expectation of X given N

Now $E[X/N]$ is a random variable, in fact a function of r.v. N! So we know how to compute its expectation!

$$E[E[X/N]] = E[g(N)] = \sum_n g(n) p_N(n).$$

$E[200N] = 200\,E[N]$

$= 200 \times 500$

$= 100000$

~~We claim that~~

~~$E[E[X/N]] = E[X]$~~

Finally we got the answer the kid got but we still need to prove the kid's claim that

$$E[E[X/N]] = E[X] \; !$$

We will see more general results in the following:

(a) Consider a two RV's X, Y (joint dist. exists) and three cases we were allways considering: (a) X, Y both discrete (b) X, Y jointly conts (c) $X/_{Y=y}$ conts. & Y is discrete (same as) $\Rightarrow$ X is conts.

(d) $Y/_{X=x}$ discrete & X is conts $\Rightarrow$ Y is discrete

(recall that in each of these 4 cases we placed total Prob. Rule & Baye's rule!)

~~We can summarize these cases as two cases:~~

(i) ~~X is conts i.e X and $X/_{Y=y}$ are conts. i.e f_X, $f_{X/Y=y}$ are pdf's~~

(ii) ~~X and $X/_{Y=y}$ are discrete i.e f_X, $f_{X/Y=y}$ are pmf's~~

~~(irrespective of nature of X)~~

Lets now $E\left[E\left[g(X)/Y\right]\right] = E\left[g(X)\right]$ for all the four general cases.

Now suppose X, $X/Y=y$ are conts.

Proof

$$E\left[g(X)/Y=y\right] = \int_{-\infty}^{\infty} g(x)\, f_{X/Y}(x/y)\, dx$$

This is a function of y! Let it be $h(y)$

$$h(y) = \int_{-\infty}^{\infty} g(x)\, f_{X/Y}(x/y)\, dx$$

$E\left[g(X)/Y\right]$ is the random variable $h(Y)$.

(BTW you can show that $h(Y)$ is a valid r.v. if $g(X)$ is a r.v. & $E[g(X)]<\infty$ (not in this class))

Now, $E\left[E\left[g(X)/Y\right]\right] = E\left[h(Y)\right]$

$$= \begin{cases} \int_{-\infty}^{\infty} h(y)\, f_Y(y)\, dy & \text{if } Y \text{ is conts.} \\ \sum_{\forall y} h(y)\, f_Y(y) & \text{if } Y \text{ is discrete} \end{cases}$$

$$= \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\, f_{X/Y}(x/y)\, dx\, f_Y(y)\, dy = \int_{-\infty}^{\infty} g(x) \underbrace{\left[\int_{-\infty}^{\infty} f_{X/Y}(x/y)\, f_Y(y)\, dy\right]}_{f_X(x)} dx \\ \qquad = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx = E\left[g(X)\right] \quad \text{(if } Y \text{ is conts)} \\ \sum_{\forall y}\int_{-\infty}^{\infty} g(x)\, f_{X/Y}(x/y)\, dx\, f_Y(y) = \int_{-\infty}^{\infty} g(x) \underbrace{\left[\sum_{\forall y} f_{X/Y}(x/y)\, f_Y(y)\right]}_{f_X(x)} dx = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx \\ \qquad = E\left[g(X)\right] \quad \text{(if } Y \text{ is discrete)} \end{cases}$$

So proof is done for two cases.

Next case is say X, $X/_{Y=y}$ are discrete rv's.

Then,

$$E[g(X)/_{Y=y}] = \sum_{\forall x} g(x) f_{X/Y}(x/y)$$

This is a function of y! Let it be $h(y)$

$$h(y) = \sum_{\forall x} g(x) f_{X/Y}(x/y)$$

$\Rightarrow E[g(X)/Y]$ is ~~the~~ random variable $h(Y)$

(Again if $g(x)$ is a rv & $E[g(x)] < \infty$, then $h(Y)$ is a rv! / valid)

$$E[E[g(X)/Y]] = E[h(Y)]$$

$$= \begin{cases} \sum_{\forall y} h(y) f_Y(y) = \sum_{\forall y} \sum_{\forall x} g(x) f_{X/Y}(x/y) f_Y(y) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} h(y) f_Y(y)\, dy = \int_{-\infty}^{\infty} \sum_{\forall x} g(x) f_{X/Y}(x/y) f_Y(y)\, dy & \text{if } Y \text{ is conts} \end{cases}$$

$$= \begin{cases} \sum_{\forall x} g(x) \underbrace{\left[\sum_{\forall y} f_{X/Y}(x/y) f_Y(y)\right]}_{f_X(x)} = \sum_{\forall x} g(x) f_X(x) = E[X] & \text{if } Y \text{ is discrete} \\ \sum_{\forall x} g(x) \underbrace{\left[\int_{-\infty}^{\infty} f_{X/Y}(x/y) f_Y(y)\, dy\right]}_{f_X(x)} = \sum_{\forall x} g(x) f_X(x) = E[X] & \text{if } Y \text{ is conts.} \end{cases}$$

So all cases are placed. Hence now we are as smart as a school kid!! :D

Now we can easily verify the following properties of conditional expectation:

(i) linearity: $E[a_1 h_1(X) + a_2 h_2(X) / Y] = a_1 E[h_1(X)/Y] + a_2 E[h_2(X)/Y]$

of cond. exp.

(ii) if $h(X) \geq 0$, then $E[h(X)/Y] \geq 0$

(with prob. 1) of almost surely $\Leftrightarrow$ as a result (with prob 1) (almost surely)

if $X_1 \geq X_2$, then $E[X_1/Y] \geq E[X_2/Y]$

(almost surely) (almost surely)

(iii) $E[c/Y] = c$ (c is const.) (iv) $E[h(Y)/Y] = h(Y)$

also if X, Y are independent then

$E[g(X)/Y] = E[g(X)]$ etc.

also if X, Y are independent then

$E[g(X)h(Y)/Y] = h(Y) E[g(X)]$

(*Realize where we used this in supermarket problem)

Lecture Notes-19

25-Sep-09

Consider series of coin flips with usual assumptions. Suppose we are interested in average no. flips for seeing the first 'HT'. Here's a way to do it using the concept of conditional expectation:

Let X_1 : rv. representing no. flips for seeing first 'H' (geometric rv)

X_2 : " " " " " 'HT'

Lets looks at the following rv which reads as "no. flips for seeing first HT given first head was at x_1^{th} toss":

$$X_2/X_1 = x_1 = \begin{cases} x_1+1 & 1-p \\ x_1+2 & p(1-p) \\ x_1+3 & p^2(1-p) \\ \vdots & \vdots \end{cases}$$

(I)

→ note that this is indeed a valid pmf and indeed $X_2/X_1=x_1$ is a valid rv.

$$\text{Now } E[X_2/X_1=x_1] = \sum_{i=1}^{\infty} (x_1+i)\, p^{i-1}(1-p)$$

$$= x_1 \sum_{i=1}^{\infty} p^{i-1}(1-p) + \sum_{i=1}^{\infty} i\, p^{i-1}(1-p)$$

(exp. of geometric rv with parameter $1-p$!)

$$= x_1 + \frac{1}{1-p}$$

$$\Rightarrow E[X_2/X_1] = X_1 + \frac{1}{1-p}$$ } the cond. exp. rv.

$$\Rightarrow E\{X_2\} = E\{E\{X_2/X_1\}\} = E\left\{X_1 + \frac{1}{1-p}\right\} = E\{X_1\} + \frac{1}{1-p} = \frac{1}{p} + \frac{1}{1-p}$$

(* Put $p = 1/2$ and realize what is happening)

Now suppose we want "avg. no. flips for seeing the first HH".

Again let X_1 : no. flips for seeing first head (geometric rv)

X_2 : " " " " HH

Again look at:

$$Y = X_2/X_1 = x_1 = \begin{cases} x_1+1 & p \\ x_1+1+X_2 & 1-p \end{cases} \quad \begin{matrix} \to \text{again valid pmf} \\ \text{(II)} \end{matrix}$$

But contrast this defn. and (I). In fact we have never seen such a defn! The values taken by Y must be numbers. Here it looks like another rv! (function)

With a little thought you can convince yourself that this is indeed a valid defn. What we mean is:

To see this let X_2 r.v take values in set E & $P\{X_2 = \gamma_i\} = p_i$ $\forall \gamma_i \in E$

(of course $\sum_{i=1}^{\infty} p_i = 1, p_i \geq 0$) we can re-write (II) as:

$$X_2/X_1 = x_1 = \begin{cases} x_1+1 & p \\ x_1+1+\gamma_i & (1-p)p_i \end{cases} \quad \to \text{again valid pmf and indeed a "usual" way of defining a r.v.!}$$

$\longrightarrow$ Now,

$$E\{X_2/X_1 = x_1\} = (x_1+1)p + \sum_{\forall \gamma_i} (x_1+1+\gamma_i)(1-p)p_i = x_1+1+(1-p)\sum_{\forall \gamma_i} \gamma_i p_i$$

$$= x_1+1+(1-p)E[X_2]$$

$$\Rightarrow E\{X_2/X_1\} = X_1+1+(1-p)E[X_2]$$

$$\Rightarrow E\{X_2\} = E\{E\{X_2/X_1\}\} = E\{X_1\}+1+(1-p)E\{X_2\}$$

$\Rightarrow \quad pE\{X_2\} = 1 + E\{X_1\} \Rightarrow E\{X_2\} = \dfrac{1 + \frac{1}{p}}{p} = \dfrac{p+1}{p^2}$

(* put $p = 1/2$ compare with prev. result. Explain intuitively!)

~~How can you think about n consecutive heads here?~~
~~(you need to consider $E\{X_2\}$, $E\{X_3\}$, ... $E\{X_n\}$)~~
~~known~~ ~~from~~ known

~~(easy to see that $E\{X_n\} = \frac{E\{X_{n-1}\}+1}{p}$ this is recursive relation & we know boundaries $E\{X_2\}, E\{X_1\}$)~~

⟶ Another eg. where cond. exp. is used:

Consider a quicksort algorithm whose pivot is always chosen as the first number. Lets do amortized analysis of computational cost. In other words assume the algorithm gets all possible ordering of the same 'n' numbers with equal probability and we want "to find the average no. comparisons need to be done to sort the n numbers".

Let C_n be the r.v. representing no. comparisons for sorting 'n' numbers. It is a random variable because it depends on the order in which numbers are given!

Let Y_n be the correct position of pivot (obtained after doing n-1 comparisons)

Consider the r.v. $C_n / Y_n = y$

(By the nature of quicksort alg. we have)

$$C_n / Y_n = y = C_{y-1} + C_{n-y} + n-1$$

$$\Rightarrow E\{C_n / Y_n = y\} = E[C_{y-1}] + E\{C_{n-y}\} + n-1$$

$$= \bar{C}_{y-1} + \bar{C}_{n-y} + n-1$$

(let us denote $E[C_n] = \bar{C}_n$).

Now we can consider the random variable $E[C_n / Y_n]$ which takes values $E\{C_n / Y_n = y\}$ for $y = 1$ to n and of course the dist. of Y_n is uniform dist. (no ~~Prob~~ $P[Y_n = y] = \frac{1}{n}$ $\forall y = 1$ to n).

$$\Rightarrow \bar{C}_n = E[C_n] = E\{E[C_n / Y_n]\} = \sum_{y=1}^{n} \left(\bar{C}_{y-1} + \bar{C}_{n-y} + n-1\right) \frac{1}{n}$$

$$\Rightarrow n\bar{C}_n = n(n-1) + 2\sum_{y=1}^{n-1} \bar{C}_y$$

solving recursion we get, $\bar{C}_n \sim 2(n+1)\log(n+1)$

which is a well-known result.

⟶ another eg. for use of cond. exp.:

Suppose X, Y are jointly conts. then,

$$E\{g(X,Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\, dx\, dy.$$

If X, Y are discrete, then $E[g(X,Y)] = \sum_{\forall x} \sum_{\forall y} g(x,y) f_{XY}(x,y)$.

Now can we still compute $E[g(X,Y)]$ in the ~~cases~~ core:

~~*~~ ~~(a)~~ X/y is conts., Y is discrete $\Leftrightarrow$ X is conts. & Y/x is discrete

~~(in the words, if X is conts, Y is discrete succeeded)~~

In this case we noted many times that F_{XY} obviously exists but there is no notion of "f_{XY}". (f_{XY} does not make sense of for this case)

Using notion of conditional exp. we can still compute $E[g(X,Y)]$:

Consider $E[g(X,Y)/Y=y] = E[g(X,y)/Y=y]$

$$= \int_{-\infty}^{\infty} g(x,y)\, f_{X/Y}(x/y)\, dx$$

exp. of function of X wrt. $f_{X/Y}$

Now again we can ~~no~~ consider the random variable which takes values (at $Y=y$). Now: $E[g(X,Y)/Y]$

$$E[g(X,Y)] = E\big[\, E[g(X,Y)/Y] \,\big]$$

exp. of function of Y wrt. f_Y

$$= \sum_{\forall y} E[g(X,y)/Y=y]\, f_Y(y)$$

$$= \sum_{\forall y} \int_{-\infty}^{\infty} g(x,y)\, f_{X/Y}(x/y)\, dx\, f_Y(y)$$

$$\Rightarrow E[g(X,Y)] = \int_{-\infty}^{\infty} \left(\sum_{\forall y} g(x,y) f_{X/Y}(x/y) f_Y(y) \right) dx$$

(this is a useful expression.)

also by Baye's rule we have:

$$\Rightarrow E\{g(X,Y)\} = \int_{-\infty}^{\infty} \sum_{\forall y} g(x,y) f_{Y/X}(y/x) f_X(x) \, dx$$

$$= \sum_{\forall y} \left(\int_{-\infty}^{\infty} g(x,y) f_{Y/X}(y/x) f_X(x) \, dx \right)$$

(this is also a useful expression.)

In lecture we saw how this can be used as starting step for Bayesian Decision Theory.

Conditional Variance

Similar to notion of conditional exp. we can define conditional variance:

$$\text{var}(X/Y) = E[X^2/Y] - \left(E\{X/Y\}\right)^2 \quad - (I)$$

↓ (new defn.) conditional variance

again a r.v. and infact a function of Y.

Now we also know through notion of var(x) and notion of conditional exp. ~~we have~~ the following:

$$\text{var}(X) = E\{X^2\} - \left(E\{X\}\right)^2 \qquad \text{('defn. of var(x))}$$

$$= E\left[E\{X^2/Y\}\right] - \left(E\{E\{X/Y\}\right)^2 \qquad \text{(result we proved in case of cond. exp.)}$$

(II)

using (I) & (II) we can prove (re assign.) the following useful result:

$$\text{var}(X) = \text{var}\left(E\{X/Y\}\right) + E\left[\text{var}(X/Y)\right] \quad \text{(III)}$$

Lets look at an application of this:

~~Q~~ Let us compute the variance of $X = \sum_{i=1}^{N} X_i$

(This is supermarket eg. assume again X_i & N are independent)

$$\text{var}(X/N=n) = \text{var}\left(\sum_{i=1}^{N} X_i / N=n\right) = \text{var}\left(\sum_{i=1}^{n} X_i / N=n\right)$$

$$= \text{var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= n\,\text{var}(X_i) \qquad \text{if } X_i \text{ are iid.}$$

$$\Rightarrow \text{var}(X/N) = N\,\text{var}(X_i)$$

(note that this is indeed a function of N)

Using (III):

$$\text{var}(X) = \text{var}\left(E\{X/N\}\right) + E\left[N\,\text{var}(X_i)\right]$$

$$= \text{var}\left(N E\{X_i\}\right) + E\{N\}\,\text{var}(X_i) = \left(E\{X_i\}\right)^2 \text{var}(N) + E\{N\}\,\text{var}(X_i)$$

(we will later use these results in C.L.T. Central limit theorem.)

Lecture Notes-20

29-Sep-09

mgf of a mrv

In case of a r.v. X we have defined mgf as: $M_X(s) = E[e^{sX}]$.

Now for a mr.v. X we define mgf analogously: $M_X(\underline{s}) = E[e^{\underline{s}^T X}]$

If X is the collection of the r.v.s $X_1, X_2, \ldots, X_n$, then is also known as the joint mgf of $X_1, X_2, \ldots, X_n$ and is represented as follows:

$$M_{X_1,X_2,\ldots,X_n}(\underbrace{s_1, s_2, \ldots, s_n}_{\underline{s}}) = M_X(\underline{s}) = E[e^{\underline{s}^T X}] = E[e^{s_1X_1 + s_2X_2 + \cdots + s_nX_n}]$$

Now if $X_1, X_2, \ldots, X_n$ are independent, then we already know that expectations factorize hence we have:

$$M_{X_1,\ldots,X_n}(s_1,\ldots,s_n) = E[e^{s_1X_1+\cdots+s_nX_n}] = E[e^{s_1X_1} e^{s_2X_2} \ldots e^{s_nX_n}]$$

$$= E[e^{s_1X_1}]E[e^{s_2X_2}]\ldots E[e^{s_n^T X_n}] \quad \left.\right\} X_1, X_2 \ldots X_n \text{ are independent}$$

$$= M_{X_1}(s_1) M_{X_2}(s_2) \ldots M_{X_n}(s_n).$$

Hence if $X_1, \ldots, X_n$ are independent then the joint mgf factorizes!

Also it is easy to see that:

$$\left.\frac{\partial^{m_1+m_2+\cdots+m_n} M_X(\underline{s})}{\partial s_1^{m_1} \partial s_2^{m_2} \ldots \partial s_n^{m_n}}\right|_{\underline{s}=0} = E[X_1^{m_1} X_2^{m_2} \ldots X_n^{m_n}]$$

This shows that all moments can be "generated" from the mgf of X!

One can also show that (not in this course) given the joint mgf, the joint dist. function of $X_1, \ldots, X_n$ is determined and vice-versa.

Hence joint mgf is indeed a useful function for characterizing the collection of r.v.s.

Another look at mean vector & covariance matrix of m.r.v.

Given a m.r.v. $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, we have defined:

$$E[X] \equiv \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}$$

(mean vector of X) (vector of means)

&

$Cov(X) \equiv$ matrix with $(i,j)^{th}$ element as $Cov(X_i, X_j)$.

(covariance matrix of X)

Now lets look at a linearly related m.r.v: $Y = \underline{A}X + \underline{b}$. Note that $\underline{A}$ is a $n \times n$ matrix & $\underline{b}_{n \times 1}$ vector. Y is another m.r.v. which is a linear (affine) function of X.

It is an exercise to show that: (atleast take $n = 2, 3$ and convince yourself)

$$E[Y] = \underline{A}\, E[X] + \underline{b}$$

$$Cov(Y) = \underline{A}\, Cov(X)\, \underline{A}^T \quad (\star)$$

(in fact these are merely alternate ways of stating the linearity prop. of E)

Comparing the diagonal terms in the matrix equality (I) we have:

$$\text{var}(Y_i) = \underline{a}_i^T \text{Cov}(X) \underline{a}_i \quad \text{(why?)}$$

(Here $\underline{a}_i$ is the i^{th} column in the matrix $\underline{A}$)

Now $\text{var}(Y_i) \geq 0 \Rightarrow \underline{a}_i^T \text{Cov}(X) \underline{a}_i \geq 0$ irrespective of what is $\underline{a}_i$

Also by the very construction of $\text{Cov}(X)$, the $(i,j)^{th}$ element $= \text{Cov}(X_i, X_j)$

$(j,i)^{th}$ element $= \text{Cov}(X_j, X_i)$

$= \text{Cov}(X_i, X_j)$

Here $\text{Cov}(X)$ is a matrix which satisfies two properties:

(i) it is symmetric

(ii) $\forall \underline{a} \in \mathbb{R}^n$, $\underline{a}^T \text{Cov}(X) \underline{a} \geq 0$

Matrices satisfying the above two properties are well-studied and are known as positive semi-definite (psd) matrices!

In other words, covariance matrix of any m.r.v. is psd.

On a similar note there are matrices which are positive definite (pd)

They satisfy the below two prop:

(i) Symmetric

(ii) $\forall \underline{a} \in \mathbb{R}^n, \underline{a} \neq \underline{0}$ $\underline{a}^T \text{Cov}(X) \underline{a} > 0$

(non-zero vector)

Of course all pd matrices are psd but converse is not true.

One interesting property of pd matrices is worth noting:

If M is pd $\iff$ $M = L D L^T$

where L is an orthonormal matrix

D is a diagonal matrix with pos entries.

(we will later on see that such a decomposition of matrix is known as the eigen-value decomposition of M!)

Rather than attempting to prove this (we will do it later) we will not understand the relation.

An orthonormal matrix is a matrix whose column (l_i) vectors are

(i) Unit length $\|l_i\| = 1$ ($\Rightarrow l_i^T l_i = 1$)

(ii) orthogonal $\Rightarrow l_i^T l_j = 0 \quad (i \neq j)$.

With this it is easy to see that $L^{-1} = L^T$. $\left(\because \begin{matrix} L^T L = I, \\ L L^T = I \end{matrix} \right)$.

Also $|\det(L)|$ is nothing but the (n-dim) area of the hyper-rectangle formed by $l_1, \ldots, l_n$ vectors ; ~~for already true~~ which is 1 since each of l_i is a unit vector.

This says that $\det M = \det L \det D \det L^T = \det D = \prod_{i=1}^{n} d_i$

= ~~product of~~

(assuming $D = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix}$)

Also, $M^2 = LDL^T LDL^T = LD^2L^T$ ~~by~~ $M^p = LD^pL^T$!

(~~also~~ atleast for rational $p \geq 0$)

~~Now~~ Note

Consider $N = LD^{-1}L^T$. It is again easy to see that $MN = I$, $NM = I$.

$\Rightarrow M^{-1} = LD^{-1}L^T$ ~~by~~ $M^n = LD^nL^T$!

(atleast for all rational $n < 0$)

In other words computing any rational powers of M is trivial if given the eigen-value decomposition of M!

In the following we will use another important feature of this decomposition.

⟶ Multivariate Normal r.v.

In one of prev. lectures we defined a multivariate std. Normal r.v. Now we define $\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ to be multivariate Normal or multivariate Gaussian r.v iff

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2} |\underline{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x}-\underline{\mu})} \quad \forall \underline{x} \in \mathbb{R}^n$$

(where $\underline{\Sigma}$ is a pd matrix).

Also, if this holds, $X_1, X_2, \ldots, X_n$ are said to be jointly Normal or jointly Gaussian.

Let us first prove that $f_{\underline{X}}$ is a valid pdf. Indeed it is ≥ 0. Need to check if area under it is 1:

i.e. TST $\int \cdots \int e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x}-\underline{\mu})} d\underline{x} = (2\pi)^{n/2} |\underline{\Sigma}|^{1/2}$

i.e. TST $\int \cdots \int e^{-\frac{1}{2} \underline{y}^T \underline{\Sigma}^{-1} \underline{y}} d\underline{y} = (2\pi)^{n/2} |\underline{\Sigma}|^{1/2}$

(Put $\underline{y} = \underline{x} - \underline{\mu}$)

i.e. TST $\int\cdots\int e^{-\frac{1}{2}y^T \Sigma^{-1} y}\, dy = (2\pi)^{n/2} \prod_{i=1}^{n} d_i^{1/2}$

(we $\because \Sigma = L D L^T \Rightarrow \Sigma^{-1} = L D^{-1} L^T$ and $\det \Sigma = \prod_{i=1}^{n} d_i$)

(exists because Σ is pd) $\rightarrow \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix}$

i.e. TST $\int\cdots\int e^{-\frac{1}{2}x^T D^{-1} x}\, dx = (2\pi)^{n/2} \prod_{i=1}^{n} d_i^{1/2}$

(Put $x = L^T y \Rightarrow y = L x \Rightarrow |J| = |\det L| = 1$ (magnification factor / Jacobian))

Now LHS = i.e. TST $\int\cdots\int e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2/d_i}\, dx_1 \ldots dx_n$

$= \prod_{i=1}^{n} d_i^{1/2} \int\cdots\int e^{-\frac{1}{2}\sum_{i=1}^{n} t_i^2}\, dt_1 \ldots dt_n$ (Put $\frac{x_i}{\sqrt{d_i}} = t_i$)

$= \prod_{i=1}^{n} d_i^{1/2} \prod_{i=1}^{n} \int e^{-\frac{1}{2}t_i^2}\, dt_i = (2\pi)^{n/2} \prod_{i=1}^{n} d_i^{1/2}$ = RHS. Hence Proved.

Also in the process we have proved that the transformation

$$x = y + \mu = L z + \mu = L D^{1/2} t + \mu$$

takes the initial multiple integral and completely factorizes it.

$\Rightarrow X = L D^{1/2} T + \mu$ is a transformation which takes the multivariate Normal r.v X to r.v T which is nothing but a multivariate std Normal r.v!

Lecture Notes - 21

1-Oct-09

Continuing the discussion in the prev. lecture, we saw that the following rules of change of variables has done the job of factoring the multiple integral into product of 1-d integrals:

$$\underline{X} = \underline{Y} + \underline{\mu} = L\underline{Z} + \underline{\mu} = LD^{1/2}\underline{T} + \underline{\mu}.$$

Now look at the corresponding transformed rvs:

$$X = Y + \underline{\mu} = LZ + \underline{\mu} = LD^{1/2}T + \underline{\mu}$$

Let us introduce one notation: "$X \sim N(\underline{\mu}, \underline{\Sigma})$" means that X is a multivariate Normal rv with parameters $\underline{\mu}, \underline{\Sigma}$.

It is easy to see the following from the derivation in the prev. lecture:

$$Y \sim N(\underline{0}, \underline{\Sigma}), \quad Z \sim N(\underline{0}, D), \quad T \sim N(\underline{0}, I)$$

& in other words T is a multivariate std Normal rv!

& in other words it is collection of n independent r.v. $T_1, T_2, \ldots, T_n$ all of which are std. Normal.

Hence X is nothing but an affine transformation (which simply rotates & translates axes) of the rv. T ⟶

Also

$$E[X] = LD^{1/2}E[T] + \underline{\mu} = \underline{\mu} \quad (E[T] = \underline{0})$$

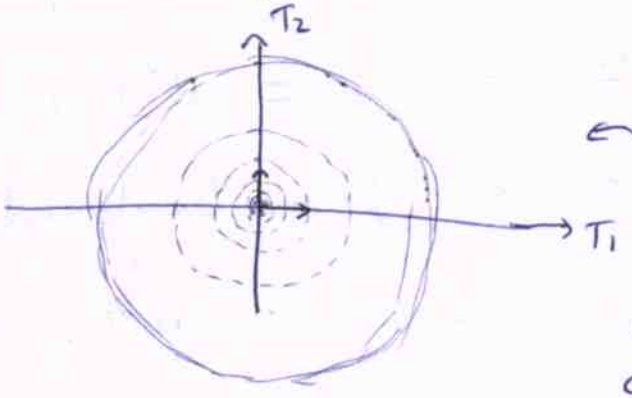
$$\mathrm{Cov}(X) = LD^{1/2}\,\mathrm{Cov}(T)\,D^{1/2}L^T = LDL^T = \underline{\Sigma} \quad (\mathrm{Cov}(T) = I)$$

Here the parameters $\underline{\mu}, \underline{\Sigma}$ are nothing but the mean vector & cov. matrix!

In other words, given the pdf expression we can just read out the mean vector & covariance matrix for a (multivariate) Normal r.v.

Here how the data generated from each of r.v's T, Z, Y, X would look like (the eigen vectors & values are also indicated in each case):
(2-d example) of Σ

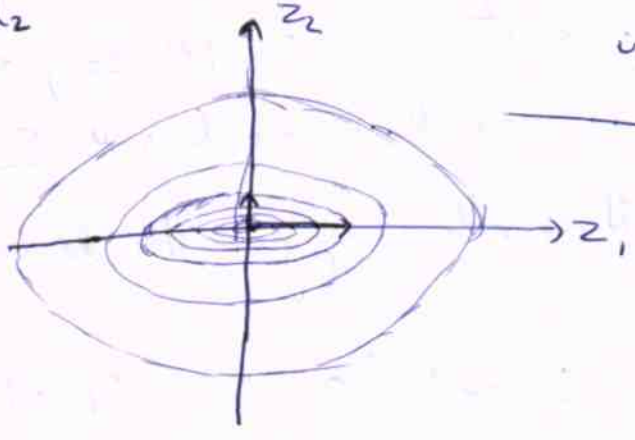
$T \sim N(\underline{0}, I)$



(imagine a bell shape curve replacing the prob. density sitting on this)

$Z \sim N(\underline{0}, D)$ $D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$ let $d_1 > d_2$

(in other words $Z = DT$
$Z_1 = d_1 T_1$
$Z_2 = d_2 T_2$)
(transformation representing scaling of axes)

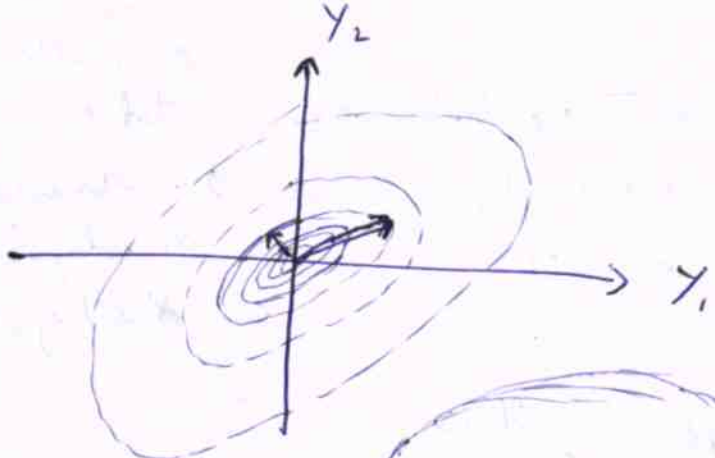


Here spacing between concentric circles is prop. to prob. density.

imagine a skewed bell curve sitting on this plot representing prob. density.

$Y \sim N(\underline{0}, \underline{\Sigma})$

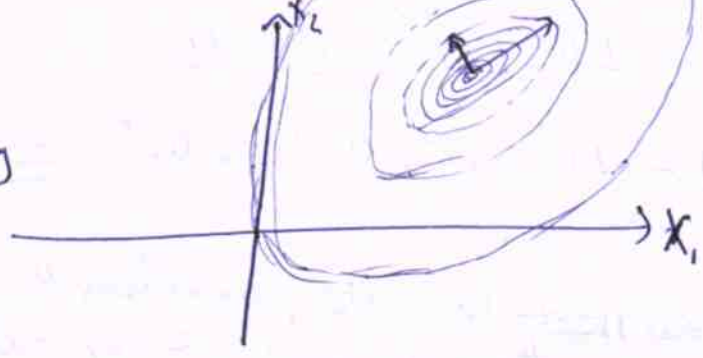
$Y = LZ$
(transformation representing rotation of axes)



A point sitting is the ellipse will be || to axis if cov. matrix is diagonal

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

$X = Y + \underline{\mu}$
(transformation representing shifting of axis)



In the above figures the ellipses/circles are nothing but the cross-sections of the bell-shaped pdf at equal intervals (heights).

In words, any multivariate Normal r.v. can be produced by shifting, rotating and scaling of a multi. std. Normal r.v. (which is nothing but a collection of independent std. Normals).

→ What is mgf of $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$?

Ans $M_X(\underline{\delta}) = E[e^{\underline{\delta}^T \underline{X}}] = \frac{1}{(\sqrt{2\pi})^n |\underline{\Sigma}|^{1/2}} \int\cdots\int e^{\underline{\delta}^T \underline{x}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1}(\underline{x}-\underline{\mu})} d\underline{x}$

Again do the change of variables: $\underline{x} = \underline{L}\underline{D}^{1/2}\underline{t} + \underline{\mu}$, we get:

$$M_X(\underline{\delta}) = \frac{1}{(\sqrt{2\pi})^n} \int\cdots\int e^{\underline{\delta}^T(\underline{L}\underline{D}^{1/2}\underline{t} + \underline{\mu})} e^{-\frac{1}{2}\underline{t}^T\underline{t}} d\underline{t}$$

Put $\underline{\delta}^T\underline{L}\underline{D}^{1/2} = \underline{a}^T$

$$= e^{\underline{\delta}^T\underline{\mu}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{a_i t_i - \frac{1}{2}t_i^2} dt_i$$

$\int_{-\infty}^{\infty} e^{a_i t_i - \frac{1}{2}t_i^2} dt_i$ is nothing but $M_{T_i}(a_i)$ → T_i is std. Normal r.v

$$= e^{\underline{\delta}^T\underline{\mu}} \prod_{i=1}^{n} e^{a_i^2/2}$$

$$= e^{\underline{\delta}^T\underline{\mu} + \frac{1}{2}\underline{a}^T\underline{a}} = e^{\underline{\delta}^T\underline{\mu} + \frac{1}{2}\underline{\delta}^T\underline{L}\underline{D}\underline{D}^{1/2}\underline{L}^T\underline{\delta}} = e^{\underline{\delta}^T\underline{\mu} + \frac{1}{2}\underline{\delta}^T\underline{\Sigma}\underline{\delta}}$$

$$\Rightarrow \boxed{M_X(\underline{\delta}) = e^{\underline{\delta}^T\underline{\mu} + \frac{1}{2}\underline{\delta}^T\underline{\Sigma}\underline{\delta}}}$$

This is also worth remembering.

Following things abt $M_X(\Delta)$ are notable:

(i) It is again a Gaussian function! (exp. of a quadratic term)

(ii) It shows that all moments (and in fact the pdf) depend only on $\underline{\mu}$, $\underline{\underline{\Sigma}}$ & in other words the entire distribution is completely specified by the first two moments!

(iii) From the joint mgf $M_X(\Delta)$ we can easily write down the mgf of X_i:

choose $\Delta = \begin{bmatrix} 0 \\ \vdots \\ \delta_i \\ \vdots \\ 0 \end{bmatrix} \to \delta_i$

Using the mgf expression of $M_X(\underline{\Delta})$ we have

$$M_X(\underline{\Delta}) = E[e^{\underline{\Delta}^T \underline{X}}] = e^{\delta_i \mu_i + \frac{1}{2}\delta_i^2 \Sigma_{ii}}$$

also equal to $M_{X_i}(\delta_i)$

μ_i : i^{th} component of $\underline{\mu}$

Σ_{ii} : i^{th} diagonal entry in $\underline{\underline{\Sigma}}$

This shows that X_i is indeed a Normal RV with mean μ_i & variance Σ_{ii}.

In other words, if $X_1, X_2, \ldots, X_n$ are jointly Normal, then each of X_i is Normal.

⟶ Another property:

- $X_1, X_2, \ldots, X_n$ are jointly Normal
  & $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ is multivariate Normal
  $\iff$ $\underline{a}^T X$ is Normal RV $\forall\, \underline{a} \neq \underline{0}$ (II)

Proof for $\Rightarrow$ is given below. (Proof for $\Leftarrow$ is an assignment)

Proof: We need to show that $Y = \underline{a}^T X$ is Normal. The idea is to show that mgf of Y looks like that of a Normal rv:

$$M_Y(s) = E[e^{sY}] = E[e^{s\underline{a}^T X}] \qquad (\text{Put } s\underline{a}^T = \underline{c}^T)$$

$$= E[e^{\underline{c}^T X}] \longrightarrow \text{mgf of } X!$$

$$= e^{\underline{c}^T \underline{\mu} + \frac{1}{2}\underline{c}^T \underline{\Sigma}\, \underline{c}}$$

$$= e^{s(\underline{a}^T \underline{\mu}) + \frac{1}{2}s^2(\underline{a}^T \underline{\Sigma}\, \underline{a})}$$

$\downarrow$

Indeed looks like mgf of Normal rv. with mean as $\underline{a}^T \underline{\mu}$ & variance as $\underline{a}^T \underline{\Sigma}\, \underline{a}$!

Hence Proved.

$\Longrightarrow$ One consequence of this is that if $X_1, X_2, \ldots, X_n$ are jointly Normal then $X_1 + X_2 + \ldots + X_n$ (& any partial sum) is a Normal rv! (In other words, convoluting Gaussian functions is again Gaussian!)

$\longrightarrow$ Another property:

If $X_{n\times 1}$ is a n-dim multivariate Normal rv. & $A_{m\times n}$ is a $m \times n$ matrix such that all rows of A are linearly independent then $Y_{m\times 1} = A_{m\times n} X_{n\times 1}$ is also a multivariate Normal rv!

Let the rows of $\underline{A}$ be $\underline{a}_i^T$ $(i=1 \text{ to } m)$ then:

$$Y_i = \underline{a}_i^T X$$

we know that each of $\underline{a}_i \neq \underline{0}$ (since rows are lin. ind.)

By II ($\Rightarrow$ part) we get that each of Y_i are Normal rvs.

Now we want to show that all (non-trivial) linear combinations of Y_i are Normal then by II ($\Leftarrow$ part) we again can say that $Y_1, Y_2, \ldots, Y_m$ are jointly Normal.

↳ lets show this:

Consider $\underline{b}^T Y = \sum_{i=1}^{n} b_i Y_i = \left(\sum_{i=1}^{m} b_i \underline{a}_i^T\right) X = \underline{c}^T X$ $(\underline{c} \neq \underline{0})$

- $\underline{b}^T Y$: If $\underline{b}$ is arbitrary $\underline{b} \neq \underline{0}$ then this is arbitrary lin. comb. of Y_i's.
- $\sum_{i=1}^{m} b_i \underline{a}_i^T$: nothing but linear comb. of rows of $\underline{A}$ which is guaranteed to be a non-zero row vector $\underline{c}^T$. ($\because$ rows are lin. ind.)
- $\underline{c}^T X$ is indeed Normal by II ($\Rightarrow$) again.

Hence Proved.

Now lets look at the case when $X_1, X_2, \ldots, X_n$ are jointly Normal. We have already seen that if the covariance matrix of $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ is diagonal then the pdf factorizes and hence each of $X_1, \ldots, X_n$ are independent Normal rvs.

However covariance matrix is diagonal $\Leftrightarrow$ all covariance terms $Cov(X_i, X_j) = 0$ $(i \neq j)$. In other words for Normal rvs

$X_1, X_2, \ldots, X_n$ are independent $\Leftrightarrow$ $X_1, X_2, \ldots, X_n$ are uncorrelated.

$\longrightarrow$ To summarize:

For $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, a multivariate Normal rv. with parameters $\underline{\mu}, \underline{\Sigma}$ we have,
(or in other words $X_1, \ldots X_n$ are jointly Normal)

(1) $\underline{\mu}$ is mean vector of X & $\underline{\Sigma}$ is covariance matrix of X.

(2) Each of $X_1, \ldots, X_n$ are individually Normal rvs with mean
$E[X_i] = \mu_i$, $Var(X_i) = \underline{\Sigma}_{ii}$

(3) $X_1, X_2, \ldots, X_n$ are independent $\Leftrightarrow$ $X_1, X_2, \ldots, X_n$ are uncorrelated
(i.e. $\underline{\Sigma}$ is diagonal)

(4) There exists an affine transformation (i.e. $\exists$ $A_{n \times n}$ such that):

$$X = AT + \underline{\mu}$$

where T is a multivariate std. Normal rv (i.e. T is collection of iid std. Normals)

In fact A is nothing but $L D^{1/2}$ where $\underline{\Sigma} = L D L^T$ (L is orthogonal, D is diagonal)

(In other words, data from X can be generated by scaling, rotating and shifting data for data generated by T)

(5) $X_1, X_2, \ldots X_n$ are jointly Normal $\Leftrightarrow$ $\underline{a}^T X$ is Normal $\forall$ $\underline{a} \neq 0$

(6) X is multivariate Normal $\Rightarrow$ $Y = AX$ is also multivariate Normal for A full row ranked A.

Intuitive meaning of Conditional Expectation

In one of prev. lectures we ~~say~~ saw that $\operatorname{argmin}_c E[(X-c)^2] = E[X]$

i.e. $E[X]$ is the "best" constant value that approximates rv X.

Now we will show that the "best" approximation of a rv Y as a function of another rv X is nothing but $E[Y/X]$

(recall that this is indeed a rv and a function of X)

i.e.

TST $\quad \operatorname{argmin}_{g(x)} E[(Y-g(x))^2] = E[Y/X]$

Proof $\quad E[(Y-g(x))^2] = E[((Y-E[Y/X]) + (E[Y/X]-g(x)))^2]$

(add and subtract $E[Y/X]$)

$$= E[(Y-E[Y/X])^2] + E[(E[Y/X]-g(x))^2]$$

$$+ 2E[(Y-E[Y/X])(E[Y/X]-g(x))]$$

$(Y-E[Y/X])$: fn of X,Y, let it be $f(X,Y)$; $(E[Y/X]-g(x))$: only a fn of X, let it be $h(X)$

~~Let's write this term done as follows~~

~~$E[(Y-E[Y/X])(E[Y/X]-g(x))]$~~

Now this expectation is:

$$E[f(X,Y)h(X)] = E[E[f(X,Y)h(X)/X]]$$

$$= E[h(X)\,E[f(X,Y)/X]]$$

$$= E[h(X)\,\underbrace{E[(Y-E[Y/X])/X]}_{\text{is zero}}] = 0$$

$$\therefore E[(Y-g(x))^2] = E[(Y-E[Y/X])^2] + E[(E[Y/X]-g(x))^2]$$

$$\geq E[(Y-E[Y/X])^2]$$

In the above inequality, an equality is achieved iff $E[Y/X] = g(X)$.

Hence $\underset{g(x)}{\text{Min}}\, E\{(Y-g(X))^2\} = E\{Y/X\}$. Hence Proved.

5-Oct-09

Lecture Notes-22

Sequences of RVs

Eg Let us recall an example from one of the prev. lectures.

Consider $M_n = \max(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots, X_n$ are iid rvs, suppose $F_{X_i} = F$, $f_{X_i} = f$ (denote their common dist. by F)

We showed that $F_{M_n}(x) = (F(x))^n \quad \forall x$ (I)

~~Previously we considered the special case $F(x) = \ldots$ for $x \ldots$~~

Now let each of X_i be upper bounded by some "a". i.e.

$$F(x) = \begin{cases} \text{some monotonic fn} < 1 & \text{for } x < a \\ 1 & x \geq a \end{cases} \quad \text{(II)}$$

(Recall that in prev. lecture we restricted X_i to be a uniform rv.) Here we don't.)

Now, consider this sequence of rvs:

$$M_1, M_2, \ldots, M_n, \ldots$$

Intuitively, $\lim_{n\to\infty} M_n$ (the limiting rv) will be the degenerate discrete rv taking value "a" with prob 1. This is because if we consider "infinite" iid rvs X_i, some where we must hit "a".

~~Hence~~ Now lets look at

$$\lim_{n\to\infty} F_{M_n}(x) \quad \forall x$$

It is easy to see that $= \begin{cases} 0 & x < a \\ 1 & x \geq a \end{cases}$ (by (I), (II))

(At "a" the limit doesn't exist. Lets choose it to be 1.)

Note that this limiting dist. (i.e. $\lim_{n\to\infty} F_{M_n}(x)$) is indeed a distribution and infact it corresponds to that of the ~~yet~~ degenerate discrete rv taking value "a" with prob. 1!

This motivates for the following defn. of convergence of rvs:

Convergence in Law

The sequence of rvs $\{X_n\}$ is said to converge to a rv X "in law" or "in distribution" (denoted by $\{X_n\} \xrightarrow{\mathcal{L}} X$ or $\{X_n\} \xrightarrow{\mathcal{D}} X$) if

$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ ($\forall x$ where F_X is conts.)

* Note that even in prev. eg, $x=a$ was point of discontinuity of F_X and infact the $\lim_{n\to\infty} F_{X_n}(a)$ doesn't exist. We can choose $F_X(a)$ conventionally to make F_X a right conts. (i.e. a valid distribution fn.)

* The issue of convergence of rvs is posed as a problem of point-wise convergence of dist. functions! (which we are very familiar with). Here this form of convergence is simplest to understand and is useful in practice.

* If we consider arbitrary sequences of rvs, then their dist. functions may not converge (point-wise) or even if they do, the limiting function may not be a dist. function!

eg: Consider $X_n \sim U[0,n]$ (Intuitively $\lim_{n\to\infty} X_n$ should not exist and this is what $\xrightarrow{\mathcal{D}}$ shows)

It is easy to see that $\lim_{n\to\infty} F_{X_n}(x) = 0 \quad \forall x$ which is not a valid dist. fn!

* In the M_n eg. $\{M_n\} \xrightarrow{\mathcal{L}} a$ this is a degenerate discrete rv.

(there are conts. rv.)

Here sequences of conts rv may $\xrightarrow{\mathcal{L}}$ to a discrete rv etc. So in general it may not make sense to talk abt convergence in pdf/pmf etc.

* Though useful, the notation of $\xrightarrow{d}$ is ~~terrible~~ loose in the sense that values the rvs take cannot be determined by the dist fns. alone (~~if~~ if $X \sim U[0,1]$ then $1-X$ is also $\sim U[0,1]$!)

Here we can easily generate toy eg. where the values rvs take are diff. but dist. fns. is same. So we define convergence notion that more how says $\{X_n\}$ as $n \to \infty$ is very close to X.

Convergence in Probability

A sequence of rvs $\{X_n\}$ is said to converge to X in probability (denoted by $\{X_n\} \xrightarrow{p} X$) if $\lim_{n\to\infty} P[|X_n - X| > \varepsilon] = 0 \quad \forall \varepsilon > 0$.

* We can equivalently write down the δ-neighbourhood version of this defn: Given $\delta > 0$, $\exists N \in \mathbb{N} \ni \forall n \geq N \quad P[|X_n - X| > \varepsilon] < \delta$. (and this happens for $\forall \varepsilon > 0$).

* Taking some eg of M_n, let us see if $\lim_{n\to\infty} P[|M_n - a| > \varepsilon] = 0$?

$$M_n \leq a \Rightarrow P[|M_n - a| > \varepsilon] = P[a - M_n > \varepsilon] = P[M_n < a - \varepsilon]$$

$$= \text{some monotonic}\ (g(a-\varepsilon))^n$$

(g is some monotonic fnc. < 1)

$\Rightarrow$ indeed $\lim_{n\to\infty} P[|M_n - a| > \varepsilon] = \lim_{n\to\infty} (g(a-\varepsilon))^n = 0$!

Here $\{M_n\} \xrightarrow{p} a$.

* In fact, one can show that: (not proof out of scope)

$$\{X_n\} \xrightarrow{p} X \Rightarrow \{X_n\} \xrightarrow{d} X$$

Here convergence in Prob. is strictly stronger notion of convergence than convergence in dist.

Now we have seen a notion of convergence which only looks at dist. functions (i.e. probabilities) and a notion which actually looks at values rv's take. Now lets look at a notion of convergence of rvs which uses notion of moments (i.e. convergence in averages!)

Convergence in n^{th} Moment

A sequence of rvs $\{X_n\}$ is said to converge to X in the n^{th} moment (denoted by $\{X_n\} \xrightarrow{n} X$) if $\lim_{n\to\infty} E[|X_n - X|^n] = 0$

* Lets again work out the $\{M_n\}$ example:

we have $F_{M_n}(x) = (F(x))^n \Rightarrow f_{M_n}(x) = n(F(x))^{n-1} f(x)$

↓ non-zero for $x \le a$

Lets take the case F, f represent uniform rv between $[0, a]$.

$$\Rightarrow F_{M_n}(x) = \begin{cases} 0 & x<0 \\ (x/a)^n & 0 \le x < a \\ 1 & x \ge a \end{cases} \Rightarrow f_{M_n}(x) = \begin{cases} \frac{n x^{n-1}}{a^n} & 0 \le x < a \\ 0 & \text{otherwise} \end{cases}$$

Also consider $n=2$ (convergence in mean squared sense)

$$E[|M_n - a|^2] = \int_0^a (x-a)^2 \frac{n x^{n-1}}{a^n} = \frac{n}{a^n}\left[\int_0^a (x-a)^2 x^{n-1} dx\right]$$

$$= \frac{n}{a^n}\left[\frac{a^{n+2}}{n+2} - \frac{2a^{n+2}}{n+1} + \frac{a^{n+2}}{n}\right]$$

$$= \frac{2a^2}{(n+1)(n+2)}$$

$$\Rightarrow \{M_n\} \xrightarrow{n=2} a.$$

Now what is the relation between $\{X_n\} \xrightarrow{n_1} X$ and $\{X_n\} \xrightarrow{n_2} X$??

(see assign. problem)

* It turns out that:

$$\{X_n\} \xrightarrow{r} X \implies \{X_n\} \xrightarrow{p} X \left(\xRightarrow{\text{and hence}} \{X_n\} \xrightarrow{d} X\right)$$

(for some r)

$$\{X_n\} \xrightarrow{p} X \not\Longrightarrow \{X_n\} \xrightarrow{r} X$$

(however! converse is not true)

Proof of $\Rightarrow$ is by what is known as Markov inequality which we will see shortly. Here is counter eg for the converse:

eg Consider $P\{X_n = n\} = \frac{1}{n}$

$$P\{X_n = 0\} = 1 - \frac{1}{n}$$

It is easy to see that $\{X_n\} \xrightarrow{p} 0$. Here is why:

$$P\{|X_n - 0| \geq \varepsilon\} = P\{X_n = n\} = \frac{1}{n} \Rightarrow \lim_{n\to\infty} P\{|X_n - 0| > \varepsilon\} = 0$$

However $E\{|X_n - 0|^r\} = \frac{1}{n} n^r = n^{r-1}$ (of course $r \geq 1$)

$$\Rightarrow \lim_{n\to\infty} E\{|X_n - 0|^r\} \neq 0 \Rightarrow \{X_n\} \not\xrightarrow{r} 0.$$

Markov Inequality

Let $Y = g(X)$ be a non-negative r.v. Then

$$P\{g(X) > \varepsilon\} \leq \frac{E\{g(X)\}}{\varepsilon} \quad \forall \varepsilon > 0$$

very imp. inequality.

(as X is cont.)

Proof: $E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx = \int_{x: g(x) \leq \varepsilon} g(x) f_X(x)\,dx + \int_{x: g(x) > \varepsilon} g(x) f_X(x)\,dx$

$$\geq \int_{x: g(x) > \varepsilon} g(x) f_X(x)\,dx$$

($\because g(x) \geq 0$ this is OK)

$$\Rightarrow E\{g(x)\} \geq \int_{x:g(x)\geq\varepsilon} \varepsilon f_X(x)dx$$

$$= \varepsilon P[g(x)>\varepsilon]$$

$$\Rightarrow P[g(x)>\varepsilon] \leq \frac{E[g(x)]}{\varepsilon}$$. Here Proved

Now put $g(x)=|X|^n$, $\varepsilon=\delta^n$

~~$P[|X|^n \geq \delta^n] \leq P[|X|>\delta]$~~

$$P[|X|>\delta] = P[|X|^n > \delta^n] \leq \frac{E[|X|^n]}{\delta^n}$$

(inequality relating absmoments & the corresponding Prob.)

$\longrightarrow$ Now put $g(x)=|X_n-X|^n$, $\varepsilon=\delta^n$ ~~we get~~ we get:

$$P[|X_n-X|>\delta] \leq \frac{E[|X_n-X|^n]}{\delta^n}$$

If $\{X_n\} \xrightarrow{n} X$, then RHS $\to 0$ as $n\to\infty$ $\Rightarrow \lim_{n\to\infty} LHS = 0 \Rightarrow \{X_n\} \xrightarrow{P} X$

$\longrightarrow$ Now put $g(x)=(X-E[X])^2$, $\varepsilon=\delta^2$, we get:

$$P[|X-E[X]|>\delta] = P[(X-E[X])^2 > \delta^2] \leq \frac{E[(X-E[X])^2]}{\delta^2} = \frac{\text{var}(x)}{\delta^2}$$

$$\Rightarrow P[|X-E[X]|>\delta] \leq \frac{\text{var}(x)}{\delta^2}$$ (very imp. Chebyshev's inequality)

Now put $\delta = k\sqrt{\text{var}(X)}$, then:

$$P\left[\,|X - E[X]| > k\sqrt{\text{var}(X)}\,\right] \leq \frac{1}{k^2}$$

↓ (independent of X !)

Says that prob. that <u>any</u> rv deviates from its mean by k times std. div, is <u>always</u> $\leq \frac{1}{k^2}$.

Though it looks like we have ~~one more~~ neglected some terms to arrive at the inequality, it happens that ~~one~~ "given only two moments of a rv, one cannot beat/improve ~~the~~ what the Chebyshev's inequality gives!

In other words I can come with distributions satisfying the inequality exactly (~~to be~~ i.e. with equality).

Lecture Notes-23

06-Oct-09

In the previous lecture we saw three modes of convergence of rvs:

(i) $\{X_n\} \xrightarrow{d} X$ iff $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ ($\forall x$ where F_X is conts.)
(Convergence in Distribution or Law)

(ii) $\{X_n\} \xrightarrow{p} X$ iff $\lim_{n\to\infty} P[|X_n - X| > \varepsilon] = 0 \quad \forall \varepsilon$
(Convergence in Probability)

(iii) $\{X_n\} \xrightarrow{r} X$ iff $\lim_{n\to\infty} E[|X_n - X|^r] = 0$ (r is given, $r \in \mathbb{N}$)
(Convergence in r^{th} moment)

Convergence in 1st moment is also known as convergence in mean
" 2nd " " " " mean square

Convergence in distribution, though is the weakest form, is also in some sense the most useful form of convergence. One reason is this: Usually we need to guess what "X" is going to be. However the notion of conv. in dist. gives us an easy way out. Given the marginal distribution functions F_{X_n} we can always see if the pointwise limit of these functions exists and if so is it a valid dist. function. This immediately gives F_X! So this is an easy way to guess what "X" is.

On the other hand, unless X is known, (a priori) the other notions of convergence are hard to verify. In fact, unless the X is a constant number, to verify (ii), (iii) itself is not straightforward because we will need the joint dist. of X_n & X!

(Note that $X_n \forall n$, X need to be rvs defined on the same probability space $\Pi = (\Omega, \mathcal{F}, P)$)

However in cases where X turns out to be a constant, ~~in many cases~~ most of the times, (iii) is the easiest to verify (we will see eg. of "Law of Large no." later)

So the ans to the question "which is the apt mode of convergence" depends on ~~the~~ need and convinience of the problem in hand. w.r.t.

Now in case mgf's of X_n & X all exist, then it is easy to see that convergence of mgf of X_n to mgf X indeed (pointwise) implies that $\{X_n\} \xrightarrow{\mathcal{D}} X$. This is because, if mgf exists, mgf and dist. fnc. uniquely determine each other.

Ill by if pdf of $X_n \xrightarrow[\text{converges}]{\text{pointwise}}$ pdf of $X \Rightarrow \{X_n\} \xrightarrow{\mathcal{D}} X$

Ill by pmf of $X_n \longrightarrow$ pmf of $X \Rightarrow \{X_n\} \xrightarrow{\mathcal{D}} X$

In summary, convergence (pointwise) of mgf's, pdf's & pmf's all imply convergence in distribution. However ~~the~~ converse is not true the reasons are simple:

(i) mgf may not always exists! So if $\{X_n\} \xrightarrow{\mathcal{D}} X$ we may not even have mgf existing so what is the question of their convergence?

(ii) We saw in prev. class that a sequence of conts. rv can convergence to a discrete rv (in dist.) and in fact we can also ~~show that~~ give examples where sequence of discrete rvs converge (in dist.) to a conts. rv!

So $\{X_n\} \xrightarrow{\mathcal{D}} X \not\Rightarrow$ convergence in pdf/pmf.

However if the mgf/pdf/pmf exist for all X_n & X then convergence in dist and convergence in mgf/pdf/pmf are equivalent.

All said and done, a rv is nothing but a function from $\mathbb{R} \to \mathbb{R}$! So we can as well talk abt notions of pointwise convergence of these functions:

Convergence Everywhere & Sure Convergence

$\{X_n\} \xrightarrow{} X$ if $\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \forall \omega \in \mathbb{R}$.

($\{X_n\}$ is said to surely converge to X) (pointwise convergence)

$\{X_n\} \xrightarrow{e} X$

It is easy to see that this is the "strongest" notion of convergence. But rarely used in practice (because of difficulty in verification)

Now we can also relax above condition and say that the set at which pointwise convergence does not happen must be of negligible (i.e. prob. of those set of outcomes is zero). This leads to the following relaxed notion of convergence:

Almost sure convergence (Convergence with prob. 1) or almost everywhere convergence

$\{X_n\} \xrightarrow{a.s.} X$ if $P(\{\omega \in \mathbb{R} / X_n(\omega) \not\to X(\omega) \text{ as } n\to\infty\}) = 0$

($\{X_n\}$ is said to almost surely converge to X)

or

$P(\{\omega \in \mathbb{R} / X_n(\omega) \to X(\omega) \text{ as } n\to\infty\}) = 1$

It turns out that there are convenient ways of verifying a.s convergence but we will not venture into them. However, note that, this notion of convergence is strong and is also not very difficult to verify in many cases.

Finally we have following comparison of modes of convergence:

$$\{X_n\} \xrightarrow{a} X \Rightarrow \{X_n\} \xrightarrow{a.s.} X \Rightarrow \{X_n\} \xrightarrow{p} X \Rightarrow \{X_n\} \xrightarrow{D} X$$

$$\{X_n\} \xrightarrow{r} X \nearrow$$

In general, none of the converse implications are true. However if X is a constant one can show that convergence in prob. & in dist. are equivalent. Also, in general, convergence in a.s. and r are not comparable. (one may not imply the other)

Law of Large Numbers

Suppose you are conducting an experiment and say you have measured a certain quantity. Since experimental procedures are prone to errors, you repeat the expt. (independently) many times and take average of the readings. Then we justify saying that this average for large number of readings is close to the true value.

Lets justify this using the notion of convergence of r.vs.

Let μ be the true quantity to be estimated. Let $X_1, X_2, \ldots, X_n, \ldots$ be observed values of this true quantity.

↓ each are i.i.d r.vs. represent the true value + noise

In other words, $E[X_i] = \mu$; however due to some noise X_i are different from $E[X_i]$ i.e. Let the variance $\text{var}(X_i) = \sigma^2$ be non-zero.

Our intuition says $\lim_{n\to\infty} \bar{S}_n = \left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \mu \; (E[X_i])$ i.e.

→ average of n readings

↓ n is large

Let $\sum_{i=1}^{n} X_i = S_n$, $\frac{S_n}{n} = \bar{S}_n$

Consider the sequence of rvs:

$$\bar{S}_1, \bar{S}_2, \ldots, \bar{S}_n, \ldots$$

we know $E[\bar{S}_n] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{n\mu}{n} = \mu$

Also $\text{var}(\bar{S}_n) = E\left[\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)^2\right]$

$$= E\left[\frac{\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2}{n^2}\right] = \frac{\sum_{i=1}^{n} E\{(X_i-\mu)^2\}}{n^2}$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Now, we claim that $\{\bar{S}_n\} \xrightarrow{n=2} \mu$.

Proof

$$\lim_{n\to\infty} E\left[|\bar{S}_n - \mu|^2\right] = \lim_{n\to\infty} \text{var}(\bar{S}_n) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

$\therefore \{\bar{S}_n\}$ converges to μ in mean squared sense and hence in probability i.e. $\{\bar{S}_n\} \xrightarrow{p} \mu$

This statement is known as the "Weak law of large numbers".

There is also the "strong" version of this: $\{\bar{S}_n\} \xrightarrow{a.s.} \mu$ (we won't prove this here)

→ The laws of large numbers also helps us to link our axiomatic defn. of probability to classical notion:

Let X_i be indicator rv of an event E. $(\forall i \in \mathbb{N})$

$\Rightarrow E[X_i] = P(E)$.

Now $\bar{S}_n \xrightarrow{a.s.} E\{X_i\} = P(E)$

$\hookrightarrow = \frac{\sum_{i=1}^{n} X_i}{n}$

So if we want to estimate prob. of any event in a random expt. then we can repeat the expt. (independently & identically) a large no. of times & avg. no. times event occurs is $P(E)$!

Lecture Notes-24

08-Oct-09

Recall the (strong) law of large numbers: it says that if $X_1, X_2, \ldots, X_n, \ldots$ are a sequence of iid rvs with $E[X_i] = \mu$, then

$$\overline{S}_n \xrightarrow{a.s.} \mu \quad \left(\text{where } \overline{S}_n = \frac{\sum_{i=1}^{n} X_i}{n}\right).$$

Let X be a random quantity in a random experiment E. You can think about $X_1, X_2, \ldots, X_n, \ldots$ as being the values the of taken by this random quatity in independent trials of the same random experiment E. Once this view is clear, the power of law of large numbers becomes evident: $\overline{S}_n$ is a quatity easily computable ~~but~~ by just finding mean of the values the random quality takes. What law of large no. says is that $\overline{S}_n \xrightarrow{a.s.} \mu$. Note that, μ, on the other hand is a quantity which cannot be observed through repeats of the random ext. E! Infact knowing μ amounts to knowing partial information regarding the distribution of X whereas $\overline{S}_n$ is simply a quantity easily computable in terms of observed values.

~~The law of large no. can also be employed to estimate quantity other than~~ This discussion (and that in prev. class) clearly show that this result can be employed to estimate mean of an unknown distribution.

In fact, we can try estimating other quantities of (parameters) of an unknown distribution by appropriately (& cleverly) choosing the r.v. X such that $E[X] = \theta$.

Here are some examples:

Suppose we are given a coin and asked to estimate the prob. of getting heads with it:

Let X be the (Bernoulli rv) indication of obtaining heads with the coin. $P\{X=1\} = p \longrightarrow$ to be determined.

$P\{X=0\} = 1-p$. Also $E[X] = p$.

Now consider repeats (independent) of coin tosses. Suppose $X_1, X_2, \ldots, X_n, \ldots$ represent the indicators of obtaining heads. Obviously they are iid and have the same distribution as X.

So applying law of large no. we get:

$$\overline{S}_n \xrightarrow{a.s.} E[X] = p.$$

By defining iid Bernoulli rvs in a similar fashion allows us to estimate $P(E)$ in any random expt.

prob. of some event.

$\longrightarrow$ Another eg.

Suppose the true mean $E[X] = \mu$ of a rv X is known. However $\mathrm{var}(X)$ is to be estimated. Again from school days knowledge we would say: estimate of $\mathrm{var}(X) = \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$. ($X_i$ are iid)

How is this justified?

Consider rv $Y = (X-\mu)^2$ then $E[Y] = \mathrm{var}(X)$.

Hence $\overline{S}_n = \dfrac{\sum_{i=1}^{n} Y_i}{n} \xrightarrow{a.s.} E[Y] = \mathrm{var}(X)$ (by law of large no.).

($\sum_{i=1}^{n}(X_i-\mu)^2 / n$; again are iid)

In general, one may want to estimate a parameter θ in an unknown distribution (of course some partial information regarding the unknown dist. may be available for eg. in the prev. situation we know the true mean of the dist.)

The idea is to come up with $\hat{\theta}_n$ ($\to$ r.v. which is a function of n iid samples from the unknown distribution.) Note that $\hat{\theta}_n$ can be computed using the observable values whereas θ is something abstract.), such that $\hat{\theta}_n \longrightarrow \theta$.

(Convergence in some sense)

Such a r.v $\hat{\theta}_n$ is known as an estimator of θ.

The following are two desirable properties of an estimator:

(i) Unbiased estimator : $E[\hat{\theta}_n] = \theta \quad \forall n$

(ii) Minimum variance & Unbiased estimator $\hat{\theta}_n^*$: $\text{var}(\hat{\theta}_n^*) \leq \text{var}(\hat{\theta}_n)$

optimal min var., unbiased estimator

$\forall n$

any unbiased estimator

(i) condition says that with any no. samples expected value of the estimator is the true value of the estimated quantity.

(Note that the estimators for mean, P(E), variance discussed above are all unbiased estimators. Refer to assign. prb. for eg. of an estimator which is not unbiased).

(ii) condition says that at with any no. samples, the deviation from true value θ is lower than with any other unbiased estimator.

This is the starting of Estimation Theory.

Central limit theorem

Theorem: Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of iid rvs.

Then $\tilde{S}_n \xrightarrow{\mathcal{D}} N(0,1)$ (std. Normal) where $\tilde{S}_n = \frac{S_n - E[S_n]}{\sqrt{\text{var}(S_n)}}$.

(recall that $S_n = \sum_{i=1}^{n} X_i$)

$\longrightarrow$

Note that $E[\tilde{S}_n] = 0$, $\text{var}(\tilde{S}_n) = 1$ and is an affine transformation. In other words $\tilde{S}_n$ is nothing but sum of n iid rvs normalized to zero mean & unit variance.

(In fact for any X, $Y = \frac{X - E[X]}{\sqrt{\text{var}(X)}}$ is the usual way of normalizing to zero mean & unit variance.)

Let $E[X_i] = \mu$, $\text{var}(X_i) = \sigma^2$.

$$\tilde{S}_n = \frac{S_n - E[S_n]}{\sqrt{\text{var}(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

$$= \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

$$= \frac{\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)}{\sqrt{n}} = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}$$

Y_i's are nothing but normalized versions of X_i (mean 0, variance 1). Also if X_i are iid then, Y_i are also iid.

$Y_i = \frac{X_i - \mu}{\sigma}$

Hence the theorem above is proved if we show that:

$$\tilde{S}_n = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1)$$ for any $Y_1, Y_2, \ldots, Y_n, \ldots$ iid and mean 0, variance 1.

Here is the proof:

Proof Suppose mgf of Y_i exists ⟶ (In the generic proof we need not assume this. However, for the scope of this class, the assumption is OK).

Now,

$$M_{Y_i}(\lambda) = E[e^{\lambda Y_i}]$$

$$\Rightarrow \prod_{i=1}^{n} M_{Y_i}(\lambda) = \prod_{i=1}^{n} E[e^{\lambda Y_i}] = E\left[\prod_{i=1}^{n} e^{\lambda Y_i}\right] = E\left[e^{\lambda \sum_{i=1}^{n} Y_i}\right] \quad (I)$$

$\left(M_X(\lambda)\right)^n$ ($\because Y_i$ are identically distributed) ($\because Y_i$ are independent)

Here mgf of $\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}$ exists and is infact $E\left[e^{\frac{\lambda}{\sqrt{n}} \sum_{i=1}^{n} Y_i}\right] = \left(M_{Y_i}\left(\frac{\lambda}{\sqrt{n}}\right)\right)^n$

Our idea is to show that the sequence of mgf:

$$M_{S_1}(\lambda), M_{S_2}(\lambda), \ldots, M_{S_n}(\lambda), \ldots \xrightarrow{n\to\infty} M_X(\lambda) = e^{\lambda^2/2}$$

X is a std Normal

↳ If we show this $\forall\lambda$ then we are done.

in other words T.S.T. $\lim_{n\to\infty} M_{S_n}(\lambda) = e^{\lambda^2/2}$. $(\forall \lambda)$

~~$\lim_{n\to\infty} M_{S_n}(\lambda) \in \lim_{n\to\infty} \left(M_{Y_i}\left(\frac{\lambda}{\sqrt{n}}\right)\right)^n$~~ i.e. TST $\lim_{n\to\infty} \left(M_{Y_i}\left(\frac{\lambda}{\sqrt{n}}\right)\right)^n = e^{\lambda^2/2}$ $(\forall \lambda)$

Lets expand $M_{Y_i}(\lambda/\sqrt{n})$ using Maclaurin series:

$$M_{Y_i}\left(\frac{\lambda}{\sqrt{n}}\right) = 1 + \frac{(\lambda/\sqrt{n})}{1!} E[Y_i] + \frac{(\lambda/\sqrt{n})^2}{2!} E[Y_i^2] + \underbrace{\frac{(\lambda/\sqrt{n})^3}{3!} E[Y_i^3] + \ldots}_{O\left(\left(\frac{\lambda}{\sqrt{n}}\right)^3\right)}$$

($E[Y_i] \to 0$, $E[Y_i^2] \to 1$)

$$= 1 + \frac{\lambda^2}{2n} + O\left(\left(\frac{\lambda}{\sqrt{n}}\right)^3\right)$$

Now $\lim_{n\to\infty} \left(1 + \frac{\lambda}{n}\right)^n = e^{\lambda}$ and $O\left(\left(\frac{\lambda}{\sqrt{n}}\right)^3\right)$ go faster to zero than $\frac{1}{n}$, we have that

$$\lim_{n\to\infty} \left(1 + \frac{\lambda^2}{2n} + O\left(\left(\frac{\lambda}{\sqrt{n}}\right)^3\right)\right)^n = e^{\lambda^2/2}.$$ Hence Proved.

The key advantage of the central limit theorem (for us) is to approximate the distributions of S_n and $\bar{S}_n$!

S_n ↓ sum of n iid rvs X_i ; $\bar{S}_n$ → mean of n iid rvs. X_i

[Note that though the pmf/pdf of S_n is the convolution of pmf/pdf of X_i, repeated n times, this convolution may not be easily computable. However the approximation provided by CLT for distribution is "easy" to compute and also is a good approximation for moderately large n (20-30 in practice!)]

CLT says that $\tilde{S}_n \longrightarrow N(0,1)$

$$\Rightarrow P[\tilde{S}_n \le x] \approx \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt .$$

(Φ: dist. func. of Std. Normal.)

$$\Rightarrow P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right] \approx \Phi(x)$$

$$P[S_n \le x\sigma\sqrt{n} + n\mu] \approx \Phi(x)$$

$$\Rightarrow P[S_n \le y] \approx \Phi\left(\frac{y-n\mu}{\sigma\sqrt{n}}\right)$$

$$\Rightarrow F_{S_n}(y) \approx \Phi\left(\frac{y-n\mu}{\sigma\sqrt{n}}\right)$$

$$P\left[\frac{S_n/n - \mu}{\sigma/\sqrt{n}} \le x\right] \approx \Phi(x)$$

$$\Rightarrow P\left[\frac{\sqrt{n}(\bar{S}_n - \mu)}{\sigma} \le x\right] \approx \Phi(x)$$

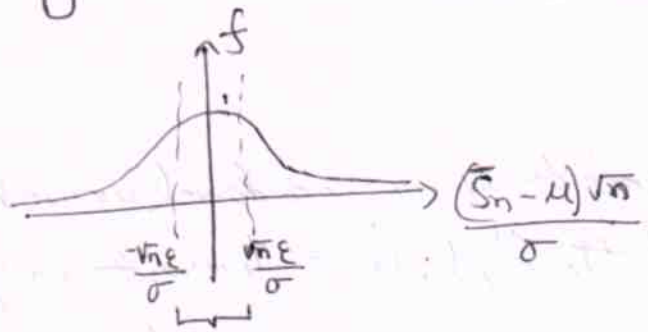
$$\Rightarrow P\left[\bar{S}_n \le \frac{\sigma x}{\sqrt{n}} + \mu\right] \approx \Phi(x).$$

$$\Rightarrow F_{\bar{S}_n}(y) \approx \Phi\left(\frac{(y-\mu)\sqrt{n}}{\sigma}\right)$$

(dist. of both)

S_n, $\bar{S}_n$ can be approx. in terms of that of std. Normal rv!

What this says alternatively (for large n):



Area of this region $\geq 1-\delta$

$1-\delta$ is usually known as the confidence

δ is " " the significance.

$\rightarrow$ nothing but $1-2\left(1-\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)\right)$

Many times we would wish that

$|\bar{S}_n-\mu| < \varepsilon$ (tolerance)

(deviation from true mean)

$\Rightarrow \frac{\sqrt{n}|\bar{S}_n-\mu|}{\sigma} < \frac{\sqrt{n}\varepsilon}{\sigma}$

but of course $\bar{S}_n$ is r.v. so we can only satisfy this with high probability.

$P\left[\frac{\sqrt{n}|\bar{S}_n-\mu|}{\sigma} < \frac{\sqrt{n}\varepsilon}{\sigma}\right] \geq 1-\delta$

(δ is small ≈ 0.05)

Here we want $1-2\left(1-\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)\right) \geq 1-\delta$

$\Rightarrow 1-\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) \leq \delta/2 \Rightarrow n \geq \left(\frac{\sigma\Phi^{-1}(1-\delta/2)}{\varepsilon}\right)^2$ (II)

(inequality is maintained since Φ is monotonic)

This shows that if we need to ensure that with high confidence ($1-\delta \approx 0.95$), the $\bar{S}_n$ (sample mean) is within ε (≈ 0.1) tolerance of the true mean, then we need to have atleast these many samples.

Here CLT can also give an idea about how many repeats of experiments are to be performed during estimation of a parameter

The same expression in (II) can be used to answered the following question:

Suppose δ is fixed (confidence/significance level we work with is fixed). Also say we have repeated the exp. for n times & we get $\bar{S}_n$. Now we can find critical value of ε below which the value is acceptable as an approximation of μ. i.e.

$$\varepsilon \geq \frac{\sigma \Phi^{-1}(1-\delta/2)}{\sqrt{n}} \Rightarrow \varepsilon_{\text{critical}} = \frac{\sigma \Phi^{-1}(1-\delta/2)}{\sqrt{n}}.$$

$\therefore$ If $|\bar{S}_n - \mu| < \varepsilon_{\text{critical}}$ we accept $\bar{S}_n$ as a valid approx. of μ and otherwise reject it!

This notion leads to the theory of Hypothesis Testing.

Lecture Notes-25 & 26

10-Oct-09

Random (or) Stochastic Processes

A collection of RVs (from the same prob. space $(\Omega, \mathcal{F}, P)$), $\{X_t, t \in T\}$, (Here T is some countable & uncountable index set) is called as a stochastic or random process.

The case of $T = \mathbb{N}$ gives back the case of sequences of rvs, which we have been analysing since few lectures (in a limited sense)

Usually the set T has an interpretation of "time". Hence the name random process. Also till now we were concerned only with the case $T = \mathbb{N}$ and mostly looking at how the "limiting" RV looks like. Now we are going to look at the collections of rvs in their entirety. i.e. specify all orders of joint-dist. functions:

i.e. specify $F_{X_{t_1} X_{t_2} \cdots X_{t_k}}$ $\forall\ t_1, t_2, \ldots t_k$ and $\forall k$.

eg: The simplest eg. of stochastic process is Bernoulli process
→ nothing but collection of independent Bernoulli rvs.

for instance: flipping of a coin is a Bernoulli process.

Note that it is easy to write down all orders of joint dist. fnc. in this case

$$f_{X_{t_1} \cdots X_{t_k}} = f^k \quad \forall\ t_1, t_2, \ldots t_k \text{ and } k$$ where f is pmf of the Bernoulli rv.

Note that for the Bernoulli process, we can consider the "sub-process" $\{X_t, t \in J, t \geq t_1\}$ where t_1 is some $t_1 \in J$.

↳ This is again the same Bernoulli process $\{X_t, t \in J\}$

↓

in the sense that all orders of joint distributions are the same

In other words, it does not matter since when the process originated, the distribution (laws) is the same.

Such processes are known as Stationary processes. We formally define stationary process as:

joint dist of $X_{t_1}, \ldots, X_{t_n}$ is same as that of $X_{t_1+s}, X_{t_2+s}, \ldots, X_{t_n+s}$

and this happens for all s and $t_1, t_2, \ldots, t_n$ and all n

Some implications are:

(take $n=1$) $F_{X_{t_1}} = F_{X_{t_2}}$ $\forall t_1, t_2$. (Marginal distributions are same) (I)

(take $n=2$) $F_{X_{t_1}X_{t_2}} = F_{X_{t_1+s}X_{t_2+s}}$ (Pairwise distributions only depend on $|t_1 - t_2|$)

... and so on.

⟶ classification of rps (random processes)

(a) Depending on whether J is countable or uncountable:

If J is countably infinite ⟶ discrete time rp

J is uncountable ⟶ continuous time rp

(b) Depending on whether the rv take on discrete or conts. values.

If X_t is discrete : discrete-state rp

X_t is conts rv : conts.-state rp

Actually, the set of values the rvs take (in context of rp) is known as state-space. Hence these names.

rp's with all $2 \times 2 = 4$ combinations are possible.

One can have further classification in discrete-state : finite or infinite state space.

Two rp we will be studying are (i) Markov chains
(ii) Poisson process

(i) is a finite-~~state~~ state discrete time rp

(ii) is an infinite-~~state~~ state & continuous time rp.

Another rp which is easy to describe is Gaussian Process (GP)

(analogous with defn of Bernoulli process and relation with Bernoulli rv)

~~For now we are concerned with discrete time GP.~~

A rp is called GP if any order joint dist is (multivariate) Gaussian distribution.

i.e. $X_{t_1}, X_{t_2}, \ldots, X_{t_k}$ are jointly Gaussian & jointly Normal. $\forall\, t_1, t_2, \ldots, t_k$ and $\forall\, k$

there is none as saying the matrix $\begin{bmatrix} \mathrm{Cov}(X_{t_1}, X_{t_1}) & \mathrm{Cov}(X_{t_1}, X_{t_2}) & \cdots & \mathrm{Cov}(X_{t_1}, X_{t_k}) \\ \vdots & & & \\ \mathrm{Cov}(X_{t_k}, X_{t_1}) & - & - & \mathrm{Cov}(X_{t_k}, X_{t_k}) \end{bmatrix}$

is pd !

(II)

So in order to specify a GP we just need to specify

$E\{X_t\}\ \forall t$ $\xrightarrow{\text{known as}}$ mean function } (imagine discrete and cont. I)

& $cov(X_{t_1}, X_{t_2})\ \forall t_1, t_2$ $\xrightarrow{\text{known as}}$ auto covariance function. }

In GP ~~terminology~~ notation, mean function is denoted by $\eta_x(t) = E\{X_t\}$

auto covariance function by $C_x(t_1, t_2) = cov(X_{t_1}, X_{t_2})$

We also know that $cov(X_{t_1}, X_{t_2}) = E\{X_{t_1} X_{t_2}\} - E\{X_{t_1}\} E\{X_{t_2}\}$

↓

~~correlation~~ called auto correlation

auto correlation function is denoted by $R_x(t_1, t_2) = E\{X_{t_1} X_{t_2}\}$.

With this terminology we can ~~say~~ define GP with its mean function and auto covariance function (provided it is positive definite)

i.e. condition (II)

We can ask when is GP stationary?

Ans (Refer (I)) first condition is Marginals same $\Rightarrow E\{X_{t_1}\} = E\{X_{t_2}\}\ \forall t_1, t_2$

and

$var(X_{t_1}) = var(X_{t_2})\ \forall t_1, t_2$

In other words $\eta_x(t)$ is a constant (mean function) and $C_x(t_1, t_1) = C_x(t_2, t_2)$

second condition is pairwise dist. only depend on $|t_1 - t_2|$

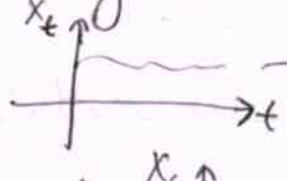
i.e. $C_x(t_1, t_2)$ is function of $|t_1 - t_2|\ \forall t_1, t_2 \in J$.

In fact these are necessary & sufficient conds. for GP to be stationary

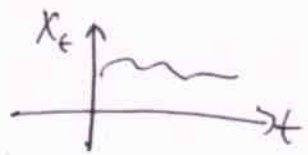
(why?)

Before going into the next example of Poisson Process, let us understand that we can view a rp in a slightly different way:

Alternative View of rp

Suppose I conduct the random experiment of ~~coin~~ toss repeated flip and observe the corresponding Bernoulli process. What I would register is a sequence (infinitely long) of 0 & 1. Now imagine a conts. time, conts.-state rp. Then if I plot the values taken by rp then we will get some 

→ curve like this

Now when we repeat the expt we may get 

and so on.

So we can also view a rp to be that which outputs a function of time for every outcome in a rand. expt. In other words:

$$X : \Omega \times J \longrightarrow \mathbb{R} \quad \text{is a rp}$$

(compare with defn. of r.v as $X : \Omega \to \mathbb{R}$)

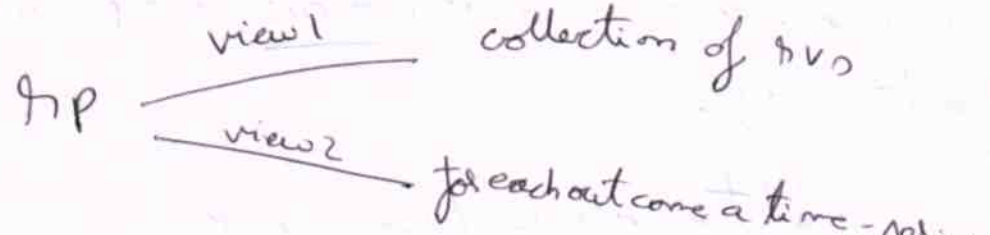
i.e. $X(\omega, t_1)$ is some number $\forall\, \omega \in \Omega$ (sample space of the rand. expt.), $t_1 \in J$ (time domain).

$X(\cdot, t_1)$ is nothing but the rv X_{t_1} i.e. $X_{t_1} : \Omega \to \mathbb{R}$

$X(\omega, \cdot)$ is nothing but a curve 

i.e. $X(\omega, \cdot) : J \to \mathbb{R}$

To summarize



⟶ **POISSON PROCESS**

Consider the D-Mart supermarket & consider a rv N_t : no. people who arrived at ~~the~~ it before and till t (↓ some time)

~~Consider~~ Consider the continuous time rp $\{N_t\}$.

Note that since values N_t can take is discrete (possibly countably infinite) the state space is countably infinite.

Such a rp which represent counts of no. events (eg: no. of people arrivals) occured since a reference '0' are known as counting processes. Here is formal defn.

Let us choose some reference time '0' as the origin & $N_0 = 0$ $t \geq 0$.

(i) $T = \mathbb{R}^+$ (set of positive reals)

(ii) $N_0 = 0$ (iii) N_t takes on values as Whole numbers.

(iv) If $t_1 \leq t_2$, then $N_{t_1} \leq N_{t_2}$

(v) Of course for a counting process no. events occured between $(t_1, t_2] = N_{t_2} - N_{t_1}$.

Often two kinds of assumptions are made in a counting process:

(a) No. events in disjoint intervals of time are independent
Called the independent increment assumption.

(Looks like for DMart example it is fair to assume this unless people look at crowd inside and decide not to visit :D)

e.g. $P[N_{t_1} \le x, N_{t_2} - N_{t_1} \le y] = P[N_{t_1} \le x] P[N_{t_2} - N_{t_1} \le y]$

(b) $N_{t_0+t} - N_{t_0}$ is independent of t_0. $\forall t_1 \le t_2$

(b) No. events occuring in a duration of time is dependent (if at all) only on the duration rather than the exact start & end times.

i.e. $$P[N_{t+s} - N_t \le x] = P[N_s \le x] \quad \forall t, s$$

This is called the stationary increment assumption

(This may not be correct assumption for DMart as there are peak hours, market closed etc. :P
However we might consider analysis a duration in which this kind of up & down behaviour is not there for e.g. analyze the peak hour itself !:P

So there are real-world eg. where these assumptions make sense.)

Now a Poisson process is a special counting process which makes both assumptions i.e. stationary as well as independent increment.

One definition of Poisson process is here: (other equivalent defns. are possible refer Ross's Book)

A counting process satisfying ~~independent increments~~ :

(i) independent increments cond.

(ii) $P\{N_{t+s} - N_s = n\} = \dfrac{e^{-\lambda t}(\lambda t)^n}{n!} \quad \forall\ t, s \geq 0.$

↓ Note that ↓ says that Poisson process satisfies assumption of stationary increments. Next, No. events occurring in duration t is simply given by Poisson r.v with parameter λt.

Now take $s=0 \Rightarrow P\{N_t = n\} = \dfrac{e^{-\lambda t}(\lambda t)^n}{n!}$

$\Rightarrow E[N_t] = \lambda t \longrightarrow$ nothing but the mean function.

↓

Since mean function is not a constant, the Poisson process is NOT a stationary process!

Note that $\lambda = E[N_t]$. In words, λ is the average rate of occurrence of events!

(Intuitively this is obviously true because it indeed matters from where I start observing the process)

~~Note however that X_{t_0} is indeed stationary~~

(only incrementally it is stationary and that is exactly stationary inc. defn.)

Now lets derivette auto ~~covariance~~ correlation fuction of N:

$$R_N(t_1, t_2) = E\{N_{t_1} N_{t_2}\} = E\{N_{t_1}(N_{t_2} - N_{t_1}) + N_{t_1}^2\}$$

(Suppose $t_1 \le t_2$)

$$= E\{N_{t_1}\}E\{N_{t_2} - N_{t_1}\} + E\{N_{t_1}^2\}$$

($\because$ No. events in disjoint intervals $(0, t]$ & $(t_1, t_2]$ are independent)

$$= \lambda t_1 (\lambda(t_2 - t_1)) + \lambda t_1 + \lambda^2 t_1^2$$

- λt_1 → mean of Poisson r.v with parameter as λt_1
- $\lambda(t_2 - t_1)$ → mean of Poisson r.v with parameter $\lambda(t_2 - t_1)$
- $\lambda t_1 + \lambda^2 t_1^2$ → second moment of Poisson r.v with param. as λt_1

$$= \lambda^2 t_1 t_2 + \lambda t_1$$

Similarly in case $t_1 > t_2$ we will get $\lambda^2 t_1 t_2 + \lambda t_2$.

~~$C_N(t_1, t_2) = R_N(t_1, t_2) - \lambda^2 t_1 t_2$~~

$$\Rightarrow R_N(t_1, t_2) = \lambda \min(t_1, t_2) + \lambda^2 t_1 t_2$$

$$\Rightarrow C_N(t_1, t_2) = R_N(t_1, t_2) - m_N(t_1) m_N(t_2)$$

$$= \lambda \min(t_1, t_2).$$

Note that ✓ is also not a fuction of $|t_1 - t_2|$ again confirming that Poisson process is not stationary.

In assignment you will prove that:

T_i is an exponential random variable with parameter λ, where T_i is waiting time for i^{th} event to occur from the instant of $i-1^{th}$ event occurring.

It is easy to see that ~~T_1 are~~ $T_1, T_2, T_3, \ldots, T_n, \ldots$ is a sequence of iid exponential rvs with parameter λ. (same as Poisson Process's parameter)

$\longrightarrow$

One can show that $W_i = \sum_{j=1}^{i} T_j$ → nothing but waiting time of event i (from the origin) is gamma distributed.
(Sum of ~~independent~~ iid exponential rvs is ↑ gamma)

$\longrightarrow$

Markov chains (MC)

To motivate (MC) let us look at the changeover problem again:

1st toss, 2nd toss, 3rd toss, 4th toss

$X_0, X_1, X_2, X_3, \ldots, X_n, \ldots$

Consider the rp X which is collection of $\{X_n\}$ $\forall n \in W$ → set of whole numbers.
where each X_n is indicator of changeover at $(n+1)^{th}$ toss.

Note that $P\{X_0 = 0\} = 1$

$P\{X_n = 1\} = 2p(1-p)$ $\forall n$, $P\{X_n = 0\} = 1 - 2p(1-p)$ $\forall n$.

Also, $P[X_2 = 1, X_{1} = 1] = p^2(1-p) + p(1-p)^2 = p(1-p)$

(probabilities as HTH, THT) → Illy $P[X_n = 1, X_{n-1} = 1] = p(1-p)$

~~adapted~~ $\Rightarrow$ X_n and X_{n-1} are ~~a~~ NOT independent

Illy we can show that X_n & X_m are independent if $|n-m| > 1$
A generalization of this notion that neighbouring rvs are not independent & far off rvs are independent is what is known as the Markov Property.

Lecture Notes-27

13-Oct-09

In these couple of lectures (left in Prob. theory) we will study Markov chains (in fact overview is a better word than study). Before venturing into that, we will introduce few useful notions regarding independence of rvs:

Suppose we are concerned with Discrete rvs only.

We know,

X, Y are independent $\iff$ $f_{XY}(x,y) = f_X(x) f_Y(y)$ $\forall x, y$

(joint-pmf factorizes)

Now, in some situations it may so happen that two rvs are not independent but given information abt another rv, they behave like independent rvs!

eg Let X, Y be marks obtained (a performance) by two students in IPL class. Initially one might be tempted to say that X, Y are independent. But a bit more thought will convince us that performance of students in general also depends on the quality of the instructor.

For instance, if the instructor is good (like me), then the performance of students is more likely to be good and vice-versa. I can represent this relation:

However, given the quality of instructor, then the performance are pretty much independent!

X, Y are linked thru Z and when Z is given, they are no more linked

In such cases, we usually talk abt the notion of ~~condition~~ Conditional Independence:

We say X is cond. ind. of Y given Z iff:

(I) $\forall x,y,z \quad f_{XY/Z}(x,y/z) = f_{X/Z}(x/z)\, f_{Y/Z}(y/z)$ (conditional pmfs factorize)

Note that X, Y are ind $\Longrightarrow$ X, Y are cond. ind. given Z $\forall z$

However X, Y are cond. ind. given Z $\not\Longrightarrow$ X, Y are indepdt.

↳ (we already gave eg. of student's performance).

Now (I) gives:

$$\frac{f_{XY/Z}(x,y/z)}{f_{X/Z}(x/z)} = \frac{f_{XYZ}(x,y,z)/f_{Z}(z)}{f_{XZ}(x,z)/f_{Z}(z)} = \frac{f_{Y/XZ}(y/x,z)}{f_{Y/X}(y/x)} \overset{\text{by (I)}}{=} f_{Y/Z}(y/z)$$

(

$$\text{Illy} \quad \frac{f_{XY/Z}(x,y/z)}{f_{Y/Z}(x/z)} = \frac{f_{XYZ}(x,y,z)/f_{Z}(z)}{f_{YZ}(y,z)/f_{Z}(z)} = f_{X/YZ}(x/y,z) \overset{\text{by (I)}}{=} f_{X/Z}(x/z)$$

In other words, if X, Y are cond. ind. given Z, then:

$$f_{Y/XZ} = f_{Y/Z} \quad ; \quad f_{X/YZ} = f_{X/Z}$$

X provides no more information given Z about Y

Y provides no more information given Z about X.

We are just dropping the arguments of cond. pmf to ease notation

Also, both these conditions imply each other.

My notion can be extended to 4 rvs:

X is cond ind. of Y, Z given W means:

$$f_{X/YZW} = f_{X/W} \iff f_{YZ/XW} = f_{YZ/W}$$

In words, Y and Z is cond. ind of X given W

Beware that this does not mean that X, Y, Z are cond. ind. given W.

$$\Updownarrow$$

$$f_{XYZ/W} = f_{X/W} f_{Y/W} f_{Z/W}$$

$$\Updownarrow$$

$$f_{X/YZW} = f_{X/W} \text{ and } f_{Y/ZW} = f_{Y/W}$$

III My notion can be extended to n rvs like this:

~~X_n is cond. ind. of X~~

$$f_{X_n/X_{n-1}X_{n-2}\ldots X_2X_1} = f_{X_n/X_{n-1}} \quad \text{(II)}$$

↳ This in words is same as saying X_n is cond. ind. of $X_{n-2}X_{n-3}\ldots X_1$ given X_{n-1}.

In fact, this property is known as Markov property and leads us to the defn of Markov chain.

MARKOV CHAINS

A Markov Chain (MC) is a discrete time, discrete-state random process where the Markov property (II) holds for all n

i.e.

$$X_0\ X_1\ X_2 \ldots\ldots X_{n-1}\ X_n\ X_{n+1} \ldots\ldots$$

↳ This collection (countable collection) of rvs is called a MC iff:

(i) all the rvs ($X_n\ \forall n$) take on discrete values i.e. values from a discrete set $S = \{s_1, s_2, \ldots\}$

state-space; $s_1, s_2, \ldots$ states

(ii) Markov property (II) holds $\forall\ n\ (\geq 1)$

i.e. for eg $f_{X_3/X_2X_1X_0} = f_{X_3/X_2}$; $f_{X_2/X_1X_0} = f_{X_2/X_1}$ so on.

Now,

$$f_{X_0X_1\cdots X_n} = f_{X_0}\, f_{X_1/X_0}\, f_{X_2/X_1X_0}\, f_{X_3/X_2X_1X_0} \cdots f_{X_{n-1}/X_{n-2}X_{n-3}\cdots X_1X_0}\, f_{X_n/X_{n-1}X_{n-2}\cdots X_1X_0}$$

$f_{X_0X_1}$

$f_{X_0X_1X_2}$

$f_{X_0X_1X_2X_3}$

$f_{X_0X_1\cdots X_{n-2}X_{n-1}}$

(Note that this relation holds for any rvs $X_0,\ldots,X_n$ and NOT particular to MC).

However in case of MC; the above relation boils down to:

$$f_{X_0X_1\cdots X_{n-1}X_n} = f_{X_0}\, f_{X_1/X_0}\, f_{X_2/X_1}\, f_{X_3/X_2} \cdots f_{X_{n-2}/X_{n-3}}\, f_{X_{n-1}/X_{n-2}}\, f_{X_n/X_{n-1}} \quad \text{(III)}$$

(You can view this as first step towards forgoing assumption of independence of rvs (; as for ind rvs we would have $f_{X_0\cdots X_n} = f_{X_0} f_{X_1}\cdots f_{X_n}$) and of course the weaker notion of independence i.e. the cond. ind. helped us in doing this)

Note that (III) implies that just by specifying one marginal (i.e. f_{X_0}) and specifying pairwise cond. for neighbours ($f_{X_n/X_{n-1}}\ \forall n\in\mathbb{N}$), one can specify the MC (as all orders of joint pmfs would have then been specified).

Here is some notation:

Usually we denote:

$f_{X_n}(x)$ as $\Pi_n(x)$ where $x \in S$

and $f_{X_n/X_{n-1}}(x/y)$ as $P(n, y, x)$ where $x, y \in S$ & $n \in \mathbb{N}$

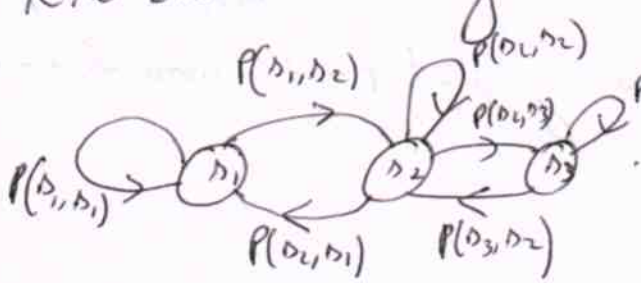
In other words, to specify a MC we just need:

$\Pi_0(x)$ $\forall x \in S$ i.e. the initial state prob. dist.

$P(n, y, x)$ $\forall x, y \in S, \forall n \in \mathbb{N}$ i.e. the state transition probabilities for all pairs of states and for all stages n

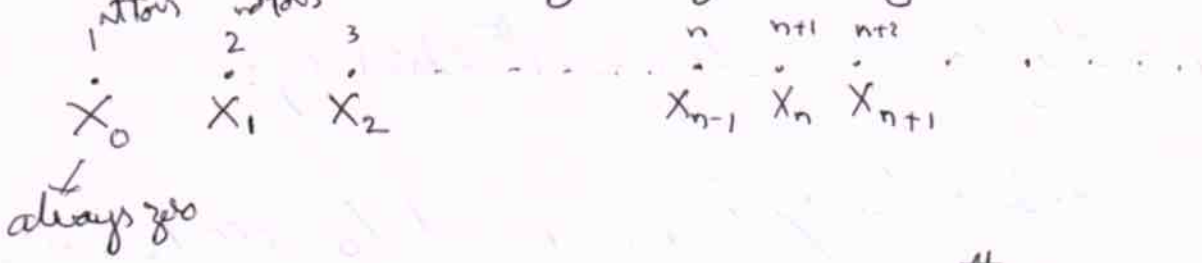
If it so happens that $P(n, y, x)$ is independent of n, i.e. if $P(n, y, x) = P(y, x)$, then these MC's are known as homogeneous MC.

One can also view a MC as a state transition process, where initial state prob. dist. is given and the law of transitions (i.e. state transition prob.) are given. And hence MC can also be represented using a state transition diagram:



$\Pi(S_1)$, $\Pi(S_2)$ ... also also given.

Eg 1 Consider the (countably infinite) collection of rvs which represent "indicator of seeing a changeover at ~~the~~ toss".



Each X_n indicator of a changeover at $(i+1)^{th}$ toss.

In last class we argued that

X_n & X_m are { independent if $|n-m|>1$; dependent if $|n-m|\leq 1$

Hence the Markov property is indeed satisfied. ~~Hence~~ Hence ~~i.e. X_n~~ this rp is a MC.

However note that, MC only requires that:

X_n is cond. ind. of $X_{n-2}, X_{n-3}, \ldots X_1 X_0$ given X_{n-1}

but here a stronger condition than this i.e.:

X_n is ind. of $X_{n-2}, X_{n-3}, \ldots, X_1 X_0$ is being satisfied.

So this rp is a MC, but some thing "more" than a MC.

Now one can also consider the (problem in Mid-Sem) collection of rvs which indicate presence of HHH, in that case it will turn out that X_n & X_m are { ind. if $|n-m|>2$; not ind. otherwise.

Here is the motivate us to define 2nd order Markov property and so on... And in general we can write down a k^{th} order Markov property (and hence define a k^{th} order MC):

X_n is cond. ind. of $X_{n-k-1} X_{n-k-2} \cdots X_1 X_0$ given $X_{n-1}, \ldots, X_{n-k}$

---

eg 2 Consider the gambler's ruin problem (refer prob. 4a of first assignment)

~~It is easy to see that even in this case, the Markov property is satisfied.~~ Let X_n be the amount of money Player A has when he plays the n^{th} round of gambling.

$X_1 = n_1$ → given he starts with n_1 rupees.

$$X_2 = \begin{cases} n_1 + 1 & \text{with prob } p_1 \\ n_1 - 1 & \text{" } 1 - p_1 \end{cases}$$

$$X_3 = \begin{cases} n_1 + 2 & \text{with prob } p_1^2 \\ n_1 \bullet & p_1(1-p_1) + (1-p_1)p_1 \\ n_1 - 2 & \text{with prob } (1-p_1)^2 \end{cases}$$

⋮

so on.

Note that given $X_{n-1} = x$, then $P\{X_n = x+\} = p_1$, $P\{X_n = x-\} = 1-p_1$. In other words distribution of X_n is determined completely given X_{n-1} alone (and additional knowledge of $X_{n-2} X_{n-3} \ldots X_1$ are not required)

Hence $\{X_n \; n \in \mathbb{N}\}$ is indeed a MC.

Also note that X_n & X_m are not independent for any n, m !! (why?)

So the Markov property is "tightly" satisfied (unlike eg 1).

Now, a little thought shows that this MC is not a homogeneous MC:

$$P(2, n_1, n_1+1) = p_1 \quad P(2, n_1, n_1-1) = 1-p_1$$

and $P(2, y, x) = 0 \;\forall\; y \neq n_1$ or $x \neq n_1+1$ or n_1-1

However $P(3, n_1+1, n_1+2) = p_1 \quad P(3, n_1+1, n_1) = 1-p_1$

$$P(3, n_1-1, n_1) = p_1 \quad P(3, n_1-1, n_1-2) = 1-p_1$$

and $P(3, y, x) = 0 \;\forall\; y \neq n_1+1$ or n_1-1 or $x \neq n_1+2$ or n_1 or n_1-2

In particular $P(2, n_1, n_1+1) = p_1$, whereas $P(3, n_1, n_1+1) = 0$

Hence this MC is not homogeneous.

In next lecture we will analyze a homogeneous MC of simple nature & provide some results on MCs.

)

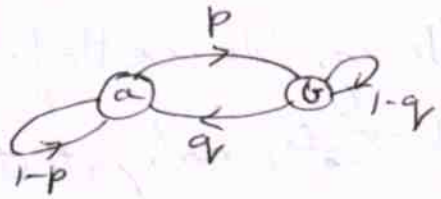
Lecture - 28

15-Oct-09

Rev.

In ~~last~~ lecture we learnt that MC can also be represented using a state-transition diagram. Consider the following eg:

eg!



(assume $p, q \neq 0$ & $p, q \neq 1$)
(consider non-trivial cases)

Now, $\Pi_n(a) = \Pi_{n-1}(a) P(a,a) + \Pi_{n-1}(b) P(b,a)$ — (I)

Illy, $\Pi_n(b) = \Pi_{n-1}(a) P(a,b) + \Pi_{n-1}(b) P(b,b)$

Since there are two states in this MC we get two eqns and in each two terms

If there were m states, we would end-up with m equations and in each eqn. we would have m terms (when added give LHS)

Some times it is convinient to represent eqns. like (I) (which relate $\Pi_n(\cdot)$ with $\Pi_{n-1}(\cdot)$ in a vectorial form:

define $\Pi_n = \begin{bmatrix} \Pi_n(S_1) \\ \Pi_n(S_2) \\ \vdots \\ \Pi_n(S_m) \end{bmatrix}$ & $P = \begin{bmatrix} P(S_1,S_1) & P(S_1,S_2) & \cdots & P(S_1,S_m) \\ P(S_2,S_1) & P(S_2,S_2) & \cdots & P(S_2,S_m) \\ \vdots & & & \\ P(S_m,S_1) & P(S_m,S_2) & \cdots & P(S_m,S_m) \end{bmatrix}_{m \times m}$

Π_n: column vector of size $m \times 1$ (m is no. states)

P: state transition prob. matrix size $m \times m$

It is easy to see that: $\Pi_n^T = \Pi_{n-1}^T P$ — (II)

(this is nothing but vectorial representation of (I)).

P is known as stochastic matrix as eqn. Matrix shortly

Of course, when S is countably-infinite (i.e. no. states is not finite) then form (II) is not convinient representation and we can have this relation:

(reading that in each row of P, the row sum is 1. Such matrices are known as Stochastic Matrices.)

$$\Pi_n(x) = \sum_{\forall y \in S} \Pi_{n-1}(y)\, P(y,x) \qquad \forall x \in S \quad \text{(III)}$$

~~(Is this the case)~~ →

Nevertheless, lets come back to the eg. We have the following by writing $P(a,b) = p$ & $P(b,a) = q$ and ~~$\Pi_{n+1}(a)$~~ realizing that $\Pi_n(a) + \Pi_n(b) = 1 \;\forall n$:

$$\Pi_n(a) = \Pi_{n-1}(a)(1-p) + (1-\Pi_{n-1}(a))q$$

$$= \Pi_{n-1}(a)(1-p-q) + q \qquad \text{(recursive relation)}$$

IIIly $\Pi_{n-1}(a) = \Pi_{n-2}(a)(1-p-q) + q$

$$\Rightarrow \Pi_n(a) = \left(\Pi_{n-2}(a)(1-p-q)+q\right)(1-p-q) + q$$

$$= \vdots$$

$$= \Pi_0(a)(1-p-q)^n + q(1-p-q)^{n-1} + q(1-p-q)^{n-2} + \cdots + q(1-p-q) + q$$

$$= \Pi_0(a)(1-p-q)^n + \frac{q\left(1-(1-p-q)^n\right)}{p+q}$$

$$\Rightarrow \Pi_n(a) = (1-p-q)^n\left(\Pi_0(a) - \frac{q}{p+q}\right) + \frac{q}{p+q} \quad \text{(IV)}$$

IIIly we get $\Pi_n(b) = (1-p-q)^n\left(\Pi_0(b) - \frac{p}{p+q}\right) + \frac{p}{q+p}$

Note that (IV) can also be obtained from the recursive relation (II):

$$\Pi_n^T = \Pi_{n-1}^T P$$

$$= \left(\Pi_{n-2}^T P\right)P = \Pi_{n-2}^T P^2$$

$$= \vdots$$

$$\Rightarrow \Pi_n^T = \Pi_0^T P^n \quad \text{(V)}$$

Now, a closer look at (IV) reveals many important things regarding the MC in the example, and thereby opens up many questions regarding generic Markov chains.

→

The first observation is that if $\Pi_0(a)$ was chosen as $\frac{q}{p+q}$
& $\Pi_0(b)$ " " $\frac{p}{p+q}$.

(Note that $\Pi_0(a)+\Pi_0(b)=1$, so OK)

then, $\Pi_n(a) = \Pi_0(a) = \frac{q}{p+q}$
$\Pi_n(b) = \Pi_0(b) = \frac{p}{p+q}$ $\forall n$ (from (IV))

This says that all marginals $f_{X_0}, f_{X_1}, \dots f_{X_n}, \dots$ are the same.

In fact, for a HMC (homogeneous Markov chain) $\Pi_n(x)$ is independent of n $\forall n \in S$
is same as saying that the HMC is a stationary proc.

Here is why:
We know for HMC:

$$f_{X_0 X_1 \cdots X_{n-1} X_n}(x_0, x_1, \dots, x_{n-1}, x_n) = f_{X_0}(x_0)\, f_{X_1/X_0}(x_1/x_0) \cdots f_{X_n/X_{n-1}}(x_n/x_{n-1})$$
$$= \Pi_0(x_0)\, P(x_0, x_1) \cdots P(x_{n-1}, x_n) \quad \text{(VI)}$$

Now lets shift time by $+s$ and consider:

$$f_{X_0 X_{0+1} \cdots X_{s+n-1} X_{n+0}}(x_0, x_1, \dots, x_{n-1}, x_n) = \Pi_s(x_0)\, P(x_0, x_1) \cdots P(x_{n-1}, x_n).$$

By the very defn. of stationary process:

$$f_{X_0 \cdots X_n} = f_{X_s X_{s+1} \cdots X_{n+s}} \quad \forall s$$

$\Rightarrow$ $\Updownarrow$ by (VI)

$$\Pi_s(x_0) = \Pi_0(x_0) \quad \forall s \quad (\text{and for all } x_0)$$

In otherwords iff $\Pi_n(x) = \Pi_0(x) \quad \forall x \in S$ then the HMC is stationary

In terms of (V) : $\Pi_0^T = \Pi_n^T = \Pi_0^T P^n \quad \forall n \iff \Pi_0^T = \Pi_0^T P$

In terms of (III) : $\Pi_0(x) = \sum_{\forall y \in S} \Pi_0(y) P(y,x)$ (VII)

Such a distribution Π which satisfies i.e.

$$\Pi(x) = \sum_{\forall y \in S} \Pi(y) P(y,x) \quad \forall x$$

is defined as the stationary dist. of the HMC. And if $\Pi_0 = \Pi$ then the HMC is stationary.

In case of eg!, Π exists and in fact $\Pi = \begin{bmatrix} \frac{q}{p+q} \\ \frac{p}{p+q} \end{bmatrix}$

This observation obviously raises the following question:

① Will a HMC always have Π? If not characterize HMC based on existence of Π.

In general, the answer to question (1) is not all HMC will have a Π.

But it is easy to see that if HMC has finite states then Π will exist!

Proof: In case of finite states Π is that vector which satisfies:

$$\Pi^T = \Pi^T P \qquad \text{(by VII)}$$

(Π^T: $1\times m$, Π^T: $1\times m$, P: $m\times m$)

$$\Pi^T = \Pi^T P \iff \Pi^T(P-I) = 0 \iff \exists \text{ a linear comb. of rows of } P-I \text{ giving row of zeros}$$

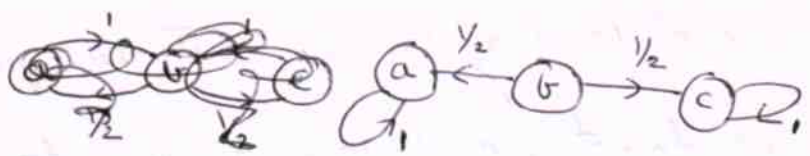
$$\iff \det(P-I) = 0$$

$\det(P-I)$ is indeed zero for any stochastic matrix P.

(because, if we add all columns of $P-I$ we will get zero as each row in P sums to 1)

$\therefore$ In case of finite HMC, Π always exists. Lets look at another example:

eg 2



By inspection of this HMC we can immediately tell that

$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ both satisfy (VII)

Because $\begin{Bmatrix}1\\0\\0\end{Bmatrix}$ says we start in (a) and if so we are bound to be in (a) because $P(a,a)=1$.

Illy $\begin{Bmatrix}0\\0\\1\end{Bmatrix}$ says we start in (c) & we are bound to be in (c) as $P(c,c)=1$.

Now lets solve $\pi^T = \pi^T P$ and see if $\begin{Bmatrix}1\\0\\0\end{Bmatrix}$ and $\begin{Bmatrix}0\\0\\1\end{Bmatrix}$ are the only two possible solutions or are there more?

$$\begin{Bmatrix}\pi(a) & \pi(b) & \pi(c)\end{Bmatrix} = \begin{Bmatrix}\pi(a) & \pi(b) & \pi(c)\end{Bmatrix}\begin{bmatrix}1 & 0 & 0\\ 1/2 & 0 & 1/2\\ 0 & 0 & 1\end{bmatrix}$$

also $\pi(a)+\pi(b)+\pi(c)=1 \longrightarrow$ (it must be valid dist. after all)

You will realize the only condition required is $\pi(b)=0$

$\pi(a), \pi(c)$ can be chosen anything such that they add to 1.

Hence $\pi = \begin{bmatrix}p\\0\\1-p\end{bmatrix}$ $\forall\ 0\le p\le 1$ are all stationary distributions.

So this says the simple HMC in eg2 has uncountable number of stationary distributions! This raises an important question:

(2) For what kind of HMC's is π unique?

Again we do not provide here a complete answer, however eg2 shows that if π is not unique then it can be realized as convex combination of few fundamental stationary distributions! Here is what we mean:

Realize that $\begin{bmatrix} p \\ 0 \\ 1-p \end{bmatrix} = p\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 1-p\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

↓ generic solⁿ. for π

↓ fundamental stationary dist.

Convex comb. meaning ~~[illegible]~~ combined with ~~[illegible]~~ non-negative weights adding to 1.

This in fact turns out to be the generic nature of HMCs! i.e. Everytime a HMC has more than one π, then it can be realized as convex comb. of few fundamental π's.

→ Next from (IV) of (eq 1) we will also find that:

$$\lim_{n\to\infty} \pi_n(a) = \pi(a) = \frac{q}{p+q} \qquad \left(\text{since } \lim_{n\to\infty}(1-p-q)^n = 0 \text{ as } 1-p-q \text{ is in } (0,1)\right)$$

$$\lim_{n\to\infty} \pi_n(b) = \pi(b) = \frac{p}{p+q}$$

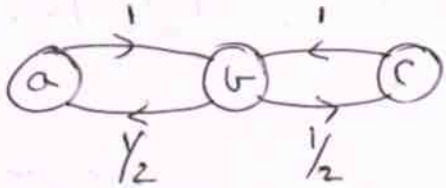
and this happens irrespective of whether $\pi_0 = \pi$!

In other words if we run this π_0 (in eq ①) then after a long time, it will attain/achieve stationarity! This raises an important question:

③ ~~With~~ Suppose π exists and moreover is unique. Then can we always say $\lim_{n\to\infty} \pi_n = \pi$? (convergence in pointwise norm)

In general, it turns out that the answer to above question is NO. Here is example:

eg 3



$$P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}.$$

The π turns out to be $\begin{bmatrix} 1/4 \\ 1/2 \\ 1/4 \end{bmatrix}$ in this case. (π exists & is unique)

However it turns out that $P^n = \begin{cases} P & n \text{ is odd} \\ P^T & n \text{ is even} \end{cases}$

~~$\pi_n = \pi_0 P$ if n is odd~~

$$\Rightarrow \pi_n^T = \pi_0^T P^n = \begin{cases} \pi_0^T P & \text{if } n \text{ is odd} \\ \pi_0^T P^T & \text{if } n \text{ is even} \end{cases}$$

You can also convince yourself that unless $\pi_0 = \pi$, in general, $\pi_0^T P \neq \pi_0^T P^T$. In other words, π_n oscillates between two distinct vectors. Hence $\lim_{n\to\infty} \pi_n$ (pointwise) does not even exist !!

It turns out that such things happen because of periodic nature of HMCs. for instance in (eg 3) each state can return to itself only in multiples of 2 steps! For ~~peri~~ aperiodic HMC (where this kind of patterns are not there) where π is unique, it does happen that $\lim_{n\to\infty} \pi_n = \pi$ (pointwise)