Statistics of multiple variables

lecture 4

suppose you have two random variables

X and Y

X = sea surface temperature

Y = evaporation

The statistics of X and Y is defined by the JOINT DISTRIBUTION FUNCTION

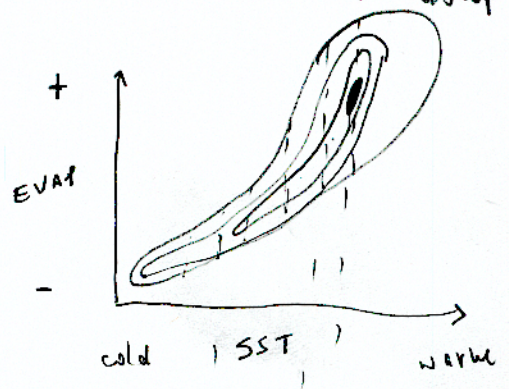
$$P_{xy}(X, Y) = \langle \delta(X - x)\, \delta(Y - y) \rangle$$

more likely higher temp. with enhanced evaporation

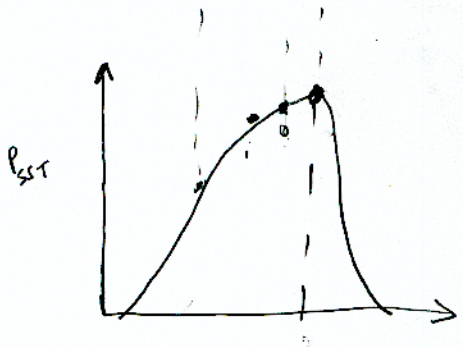
again

$$\int_{-\infty}^{\infty}\int P_{xy}\, dX\, dY = 1$$

$$P_x = \int_{-\infty}^{\infty} P_{xy}\, dY$$



If we want to know how probable SST is given a certain measure of EVAP (or vice versa)



$$P_{x|y}(X|Y) = \frac{P_{xy}}{P_y}$$

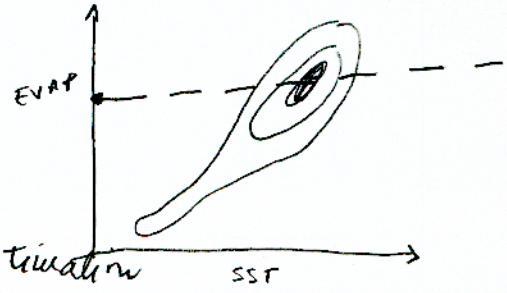
PDF of x conditioned on Y → conditional PDF

note that this equation is the basis for Bayes Rule

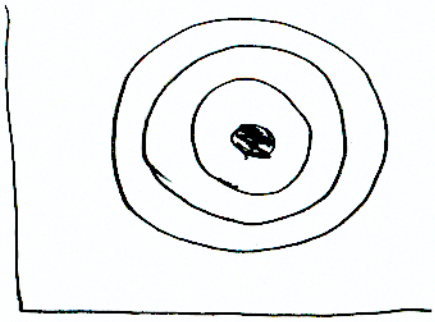
$$P_y\, P_{x|y}(X|Y) = P_{xy}$$

$$P_x\, P_{y|x}(Y|X) = P_{xy}$$

→ Bayesian statistics / estimation



suppose now to pick tropical SST and & air temp. over Atlanta in summer



$P_{x|y} = P_x$

meaning that Atlanta temperature carries no information on tropical SST.
(this may not be true if we evaluate the JPDF of $SST(t)$ and $T_{air}(t + 6 \text{ months later})$. This would be a set of different random variable!)

Assume
If X and Y are independent

$$P_{xy} = P_x P_y \qquad \text{e.g. } P_x \sim e^{-x^2} \text{ and } P_y \sim e^{-y^2}$$

$$P_{xy} = P_x P_y \sim e^{-x^2-y^2}$$

The expected value of

$$\langle xy \rangle = \int_{-\infty}^{\infty}\int \bar{X}\bar{Y} P_{xy}\, dX\, dY \qquad \text{covariance}$$

if $P_{xy} = P_x P_y$

$$\langle xy \rangle = \int_{-\infty}^{\infty} \bar{X} P_x\, dX \int_{-\infty}^{\infty} \bar{Y} P_y\, dY = \langle x \rangle \langle y \rangle = \emptyset \checkmark$$

- x and y are independent
- $\langle x \rangle$ and $\langle y \rangle = \emptyset$
- assume normal POF

$$P_x \sim e^{-\frac{x^2}{\sigma_x^2}} \qquad P_y \sim e^{-\frac{y^2}{\sigma_y^2}}$$

$$P_{xy} = P_x P_y = e^{-x^2/\sigma_x^2} e^{-y^2/\sigma_y^2} = \exp\left[-\frac{x^2}{\sigma_x^2} - \frac{y^2}{\sigma_y^2}\right]$$

- Do coordinate transformation
- change of variable

$$\hat{x} = a_{11} x + a_{12} y$$

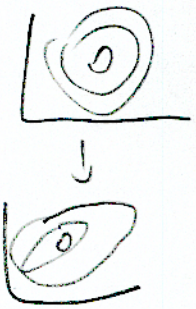
$$\hat{y} = a_{21} x + a_{22} y$$

you can already note that this transformation is building correlation in the new variable set.

$$\langle \hat{x} \rangle = 0$$

$$\langle \hat{y} \rangle = 0$$

$$\begin{cases} \langle \hat{x}^2 \rangle = a_{11}^2 \langle x^2 \rangle + a_{12}^2 \langle y^2 \rangle \\ \sigma_{\hat{x}}^2 = a_{11}^2 \sigma_x^2 + a_{12}^2 \sigma_y^2 \\ \langle \hat{y}^2 \rangle = \text{- - - - -} \end{cases}$$



$$\langle \hat{x}\hat{y} \rangle = a_{11} a_{21} \langle x^2 \rangle + a_{12} a_{21} \langle y^2 \rangle \neq 0$$

(page 37)

$$P_{\hat{x}\hat{y}} = P_{xy} \frac{\partial(x,y)}{\partial(\hat{x},\hat{y})} \simeq \exp\left[-\frac{\hat{x}^2}{\sigma_{\hat{x}}^2} - \underbrace{\frac{2\varrho\, \hat{x}\hat{y}}{\sigma_{\hat{x}}\sigma_{\hat{y}}}}_{\text{correlation term}} - \frac{\hat{y}^2}{\sigma_{\hat{y}}^2}\right]$$

Let us go back to the case of SST and EVAP so that $p_{xy} \neq p_x p_y$ and x and y are dependent.

This implies that if I have a measurement x and y I can define a linear relationship (or non linear)

① $y = \alpha x + n$ our model

② compute expected value and subtract from ①

$$\langle y \rangle = \alpha \langle x \rangle + \langle n \rangle$$

$$y' = \alpha x' + n'$$

③ now multiply by x' and take expected value again

$$\langle x'y' \rangle = \alpha \langle x'x' \rangle + \cancel{\langle x'n' \rangle}^{=0}$$

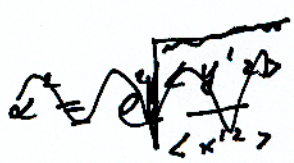
$$\alpha = \frac{\langle x'y' \rangle}{\langle x'x' \rangle}$$

an estimate of α based on the statistics of x and y

④ now let us compute the variance $\langle y'^2 \rangle$

$$\langle y'^2 \rangle = \alpha^2 \langle x'^2 \rangle + \langle n'^2 \rangle$$

define $\rho \equiv \dfrac{\langle x'y' \rangle}{\left[\langle x'^2 \rangle \langle y'^2 \rangle\right]^{1/2}}$ Pearson



$$\alpha^2 = \rho^2 \frac{\langle y'^2 \rangle}{\langle x'^2 \rangle}$$

$$\langle y'^2 \rangle = e^2 \langle y'^2 \rangle + \langle n'^2 \rangle$$

$$(1 - e^2) \langle y'^2 \rangle = \langle n'^2 \rangle$$

↑ correlation coef. is the fraction of variance in $\langle y'^2 \rangle$ that can be explained by x', the stronger the smaller is $\langle n'^2 \rangle$

$\hat{y}' = \alpha x'$ → this minimizes $\langle n'^2 \rangle$ ↗ of α

Note that this estimate is a least square estimate!

$$y' = \alpha x' + n'$$

it is equivalent of minimizing $\langle n'^2 \rangle$ with respect of α.

$J = \langle n'^2 \rangle$ the sum of the squares of the model error, the inability of x to predict y

$$J = \langle (y' - \alpha x')^2 \rangle$$

$$\frac{\partial J}{\partial \alpha} = \langle 2(y' - \alpha x')(-x') \rangle = 0$$

$$= 2\langle (-y'x' + \alpha x'x') \rangle = 2\left[-\langle y'x' \rangle + \alpha \langle x'x' \rangle \right]$$

$$\alpha = \frac{\langle y'x' \rangle}{\langle x'x' \rangle}$$ ✓

Relationship between least squares and probability distribution function

Want to show that $\hat{y}' = \alpha x'$ is equivalent to maximizing

$$P_{y'|x'}(Y' \mid X' = x')$$

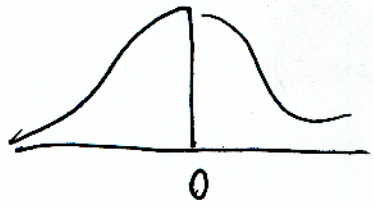
$$\hat{y}' = \int_{-\infty}^{\infty} Y' \, P_{y'|x'} \, dy$$

How to show this?

when we estimate α we minimized

$$J = \langle n'^2 \rangle$$

assume that the model errors are normally distributed



$$P_N \sim e^{-N^2/2} = e^{-(y' - \alpha x')^2} =$$

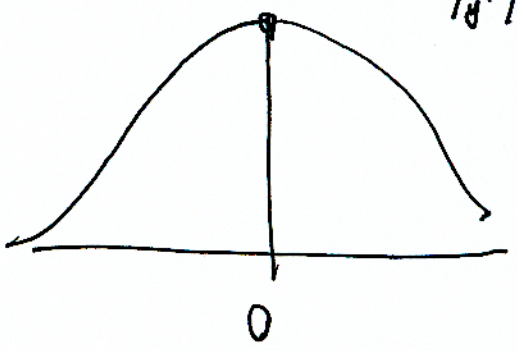
$$\exp\left[-y'^2 - \underbrace{\alpha x' y'}_{\text{correlation term}} - x'^2 \right] = P_{yx}$$

if we fix $X' = x'$

$$P_{y'|x'}(Y' \mid X' = x') = e^{-(y' - \alpha x')^2}$$

The maximum value of this pdf is

for $\begin{cases} y' = \alpha x' \\ \hat{y}' = \alpha x' \end{cases}$



this can also be seen by computing the expected value

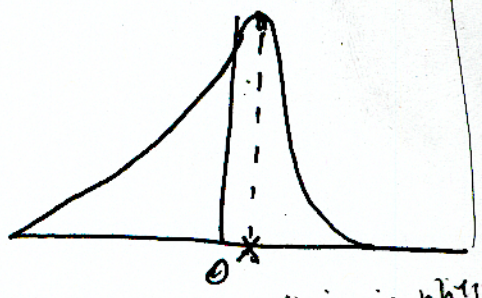
$$\hat{y}' = \int_{-\infty}^{\infty} y' \, p_{y'|x'} \, dy = \alpha x'$$

$\Rightarrow$ least square which minimizes $\langle n'^2 \rangle = \langle (y' - \alpha x')^2 \rangle$ is equivalent of maximizing the probability

$\begin{cases} p_{n'} & \text{conditioned on } x \\ p_{y'|x'} \end{cases}$

the big assumption however is that

n' is Gaussian



this implies better estimate

e.g. say you had a log normal distribution

$$p_{xy} = \exp\left[- \ln y'^2 - \ln \alpha x' y' - \ln x' \right]$$

knowing $p_{y'|x'}$ is the best.

STATIONARITY

Markov chain Process

suppose that

$$X_{n+1} = X_n$$

Why is $P_{x|y}$ a powerfull tool ?

Assume

$$\frac{dz}{dt} = \beta z$$

MARKOV PROCESS of 1st order also called MARKOVIAN

$$z_{t+1} = z_t + \beta z_t \Delta t = (1 + \beta \Delta t) z_t$$

if you know

$$P_{z_{t+1} | z_t} \left(z_{t+1} \,/\, z_t = z_t \right)$$

then

$$z_{t+1} = \int_{-\infty}^{\infty} P_{z_{t+1} | z_t} \, z_t \, dz$$

if in a random process variable future value is only determined by knoledge of the previous value so that $P_{z_{t+1}}$ is only conditional on z_t

→ MARKOV PROPERTIES

Assume now

$$\frac{dz}{dt} = F(z)$$

$$\hat{F}(z) = \int_{-\infty}^{\infty} p_{\dot{z}|z, \frac{dz}{dt}|z} \frac{dz}{dt} \, dz$$

compute this numerically

Assume now x and y are correlated and you want to make them uncorrelated!

$\langle xy \rangle \neq 0$ to a new $\langle \hat{x}\hat{y} \rangle = 0$

example:

$$\hat{x} = x\cos\phi + y\sin\phi$$

$$\hat{y} = -x\sin\phi + y\cos\phi$$

a rotation of angle ϕ

you can find the ϕ $\quad \langle \hat{x}\hat{y} \rangle = 0$

$$\tan 2\phi = \frac{2\langle xy \rangle}{\langle x^2 \rangle - \langle y^2 \rangle}$$

simply evaluate

$$\langle \hat{x}\hat{y} \rangle = (x\cos\phi + y\sin\phi)\cdot(-x\sin\phi + y\cos\phi)$$

$$-x^2 \cos\sin + xy\cos^2$$

$$-xy\sin^2 + y^2 \cos\sin \ldots$$

$$= -(x^2 - y^2)\cos\sin$$

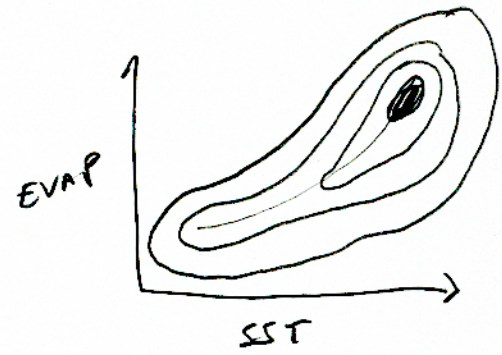
$$xy(\cos^2 - \sin^2)$$

$$= -(x^2 - y^2)\cos\sin$$

$$xy\cos^2 - xy$$

NON GAUSSIAN VARIABLES

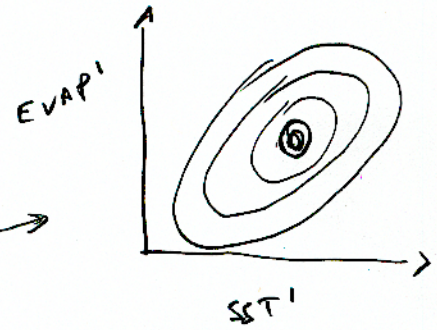
What to do with least square fit, or statistical models developped based on linear correlations?



If the non gaussian effect are related to a "mean", which we know that the anomalies are likely to be gaussian

↓

we can model those.



HURRICANE EXAMPLE

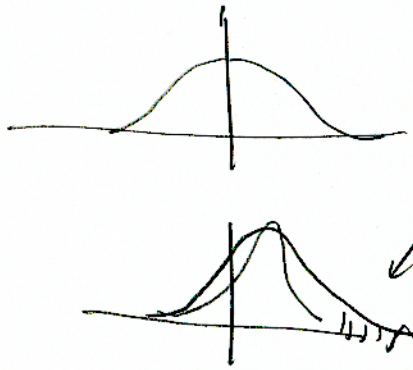
- number of hurricanes m ? $\frac{\partial SST}{\partial y}$ (SST - deep Tropospheric Temp)
- intensity I

$P_{n|i}$ and ask the question, given i, how many n are likely to occur

$$PDI \equiv \int_0^\tau V_{max}^3 \, dt$$

SST ↑ destructive power

1) The PDF of PDI changes from year to year.



PDI

p PDI(year) is not stationary

statistics depends on the time you pick!

1st order $\int p_x \, dx = \text{const}$

2nd order $\int p_x x^2 \, dx$ const. variance, covariance does not chag.

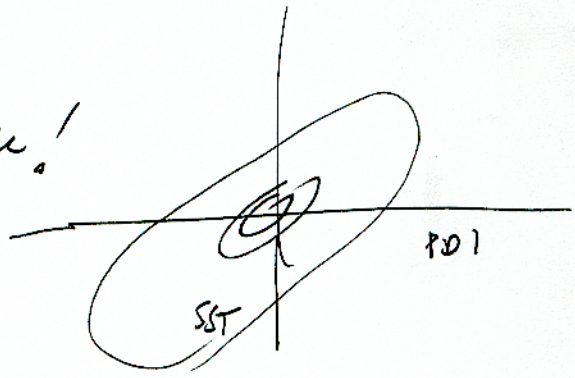
etc

2) Consider now the JPDF, if you new it!

SST' PDI'

likely to not change very much!

STATIONARITY again



3) PDI defined as destructive loss of economy ?