$$= \sum_{i=1}^{n} y_i^2 - \left(\hat{\alpha}_0 \sum_{i=1}^{n} P_0(x_i) y_i\right) - \sum_{j=1}^{k} \hat{\alpha}_j \left(\sum_{i=1}^{n} P_j(x_i) y_i\right)$$

$$\left(\because \sum_{i=1}^{n} P_0(x_i) y_i = n\bar{y}\right)$$

$$SSE = \sum y_i^2 - n\bar{y}^2 - \sum_{j=1}^{k} \hat{\alpha}_j \left(\sum_{i=1}^{n} P_j(x_i) y_i\right)$$

$$= \sum_{i=1} (y_i - \bar{y})^2 - \sum_{j=1}^{k} \hat{\alpha}_j \left(\sum_{i=1}^{n} P_j(x_i) y_i\right)$$

$$SSE = SS_{Total} - SS_{Model}$$

Multicollinearity

$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$; $\underset{\sim}{\beta} \in \mathbb{R}^{k+1}$

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

If $|X^T X| = 0$ or $|X^T X| \approx 0$ ← illed-condition

$$\hat{\underset{\sim}{\beta}} = \frac{Adj(X^T X)\, X^T Y}{|X^T X|}$$

$\Rightarrow$ $Var(\hat{\beta}_j)$ may be unbounded.

$\hat{y}_0 = (1 . x_0^T)\hat{\beta}$ will also have large variance.

Why should this problem arise at all?

→ one regressor is a linear combination of the rest.

→ improper collection of the data may lead to multicollinearity.

→ Model building depending on the magnitude

of the regressor.

Consider the expected sq. error for $\underset{\sim}{\beta}$.

$$E\{(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})^T(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})\}$$
$$= E\sum_{j=0}^{k}(\hat{\beta_j}-\beta_j)^2$$
$$= \sum_{j=0}^{k} E(\hat{\beta_j}-\beta_j)^2 = \sum_{j=0}^{k} \mathrm{Var}(\hat{\beta_j})$$
$$= \mathrm{tr}\{\sigma^2(X^TX)^{-1}\}$$

$$\left[\because \hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, \sigma^2(X^TX)^{-1})\right]$$

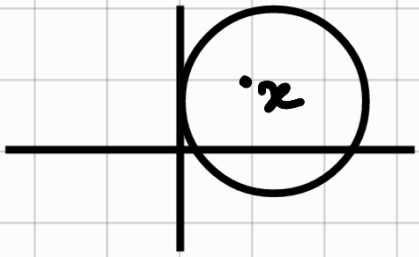
Note.

1) X^TX is a symmetric matrix.
2) $(X^TX)^{-1}$ is also a symmetric matrix, if exists.
3) X^TX is positive semidefinite matrix.
4) X^TX has non-negative evs.
$\lambda_0 > \lambda_1 > \lambda_2 > \dots \geq 0.$
5) $$\lambda_{max} = \max_{\|z\|\neq 0} \frac{\underset{\sim}{z}^T(X^TX)\underset{\sim}{z}}{\underset{\sim}{z}^T\underset{\sim}{z}}$$
$$\lambda_{min} = \min_{\|z\|\neq 0} \frac{\underset{\sim}{z}^T(X^TX)\underset{\sim}{z}}{\underset{\sim}{z}^T\underset{\sim}{z}}$$
6) $\mathrm{tr}(X^TX) = \sum \lambda_i$
7) $\mathrm{tr}((X^TX)^{-1}) = \sum 1/\lambda_i$
8) $|X^TX| = \prod_i \lambda_i$
9) $|(X^TX)^{-1}| = \prod_i 1/\lambda_i$ $\uparrow\infty$ as some $\lambda_i \downarrow 0$

Hence,
$$E\{(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})^T(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})\} = \mathrm{tr}\{\sigma^2(X^TX)^{-1}\}$$
$$= \sigma^2 \sum_i (1/\lambda_i)$$
$$\hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, \sigma^2(X^TX)^{-1})$$

Possible Solutions

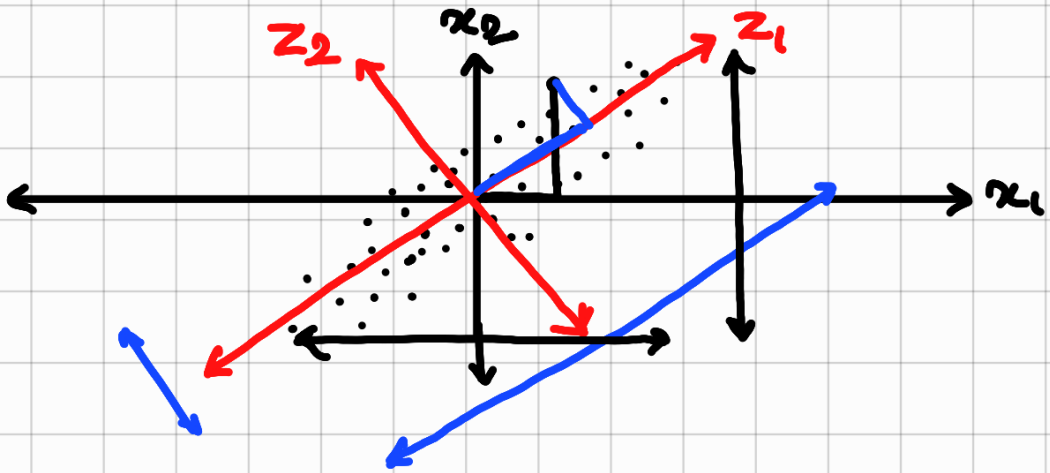
→ Principal Component Regression
→ restrict variance in euclidean norm / l^2-norm



L_1 → LASSO
L_2 → RIDGE

→ Principal Component Regression

SVD — Math
PCA — Statistics



→ one method of dimension reduction

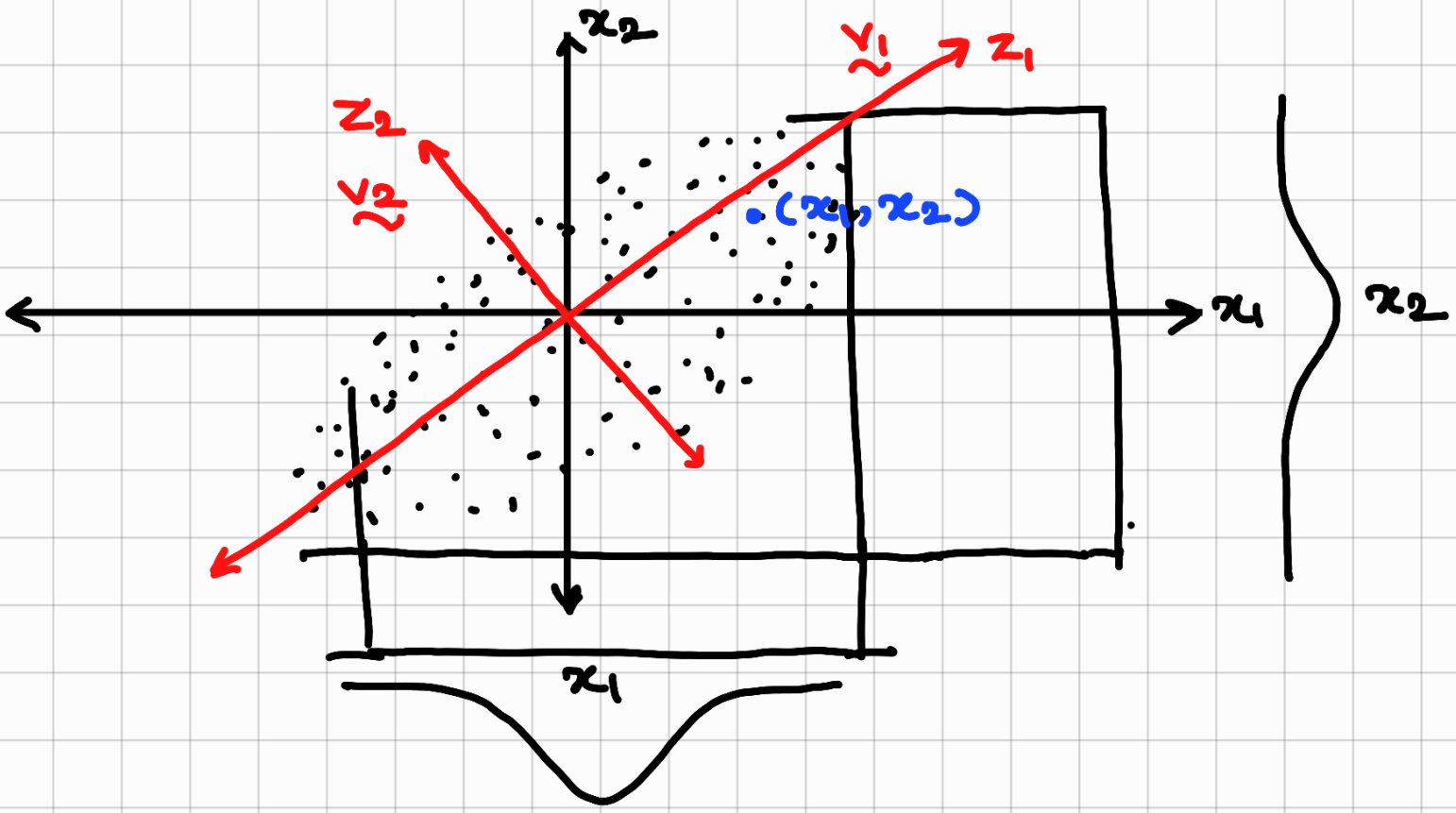
Tentatively, Midsem syllabus is upto and including Polynomial Regression.

Principal Component Regression

$|x^T x| \approx 0$

$x^T x$ p.s.d

$\lambda_0 > \lambda_1 > \lambda_2 \dots \lambda_k \geq 0$



$$e_1 = \begin{pmatrix}1\\0\end{pmatrix}, e_2 = \begin{pmatrix}0\\1\end{pmatrix} \qquad \begin{pmatrix}x_1\\x_2\end{pmatrix} = x_1 \cdot \begin{pmatrix}1\\0\end{pmatrix} + x_2 \cdot \begin{pmatrix}0\\1\end{pmatrix}$$

$$\langle \underset{\sim}{x}, \underset{\sim}{e_1} \rangle \cdot \begin{pmatrix}1\\0\end{pmatrix} + \langle \underset{\sim}{x_2}, \underset{\sim}{e_2} \rangle \cdot \begin{pmatrix}0\\1\end{pmatrix} = \underset{\sim}{x}$$

$$\begin{pmatrix}x_1\\x_2\end{pmatrix} = \langle x, \underset{\sim}{v_1} \rangle \underset{\sim}{v_1} + \langle x, \underset{\sim}{v_2} \rangle \underset{\sim}{v_2}$$
$$= z_1 \underset{\sim}{v_1} + z_2 \underset{\sim}{v_2}$$

$$\lambda_0 > \lambda_1 > \lambda_2 \ldots\ldots > \lambda_k \geq 0$$
$$\begin{bmatrix} \underset{\sim}{v_0} & \underset{\sim}{v_1} & \underset{\sim}{v_2} & \ldots\ldots & \underset{\sim}{v_k} \end{bmatrix} = P$$

$X^T X \longrightarrow (\lambda_i, \underset{\sim}{v_j})_{i=0,1,2,\ldots,k}$

$X^T X = P D P^T$

P ↳ orthogonal matrix

D ↳ diagonal matrix

$P^T P = P P^T = I$.

Example of orthogonal matrix: orth. matrix for reflection → $\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$

$$\underset{\sim}{Y} = X\beta + \underset{\sim}{\epsilon}$$

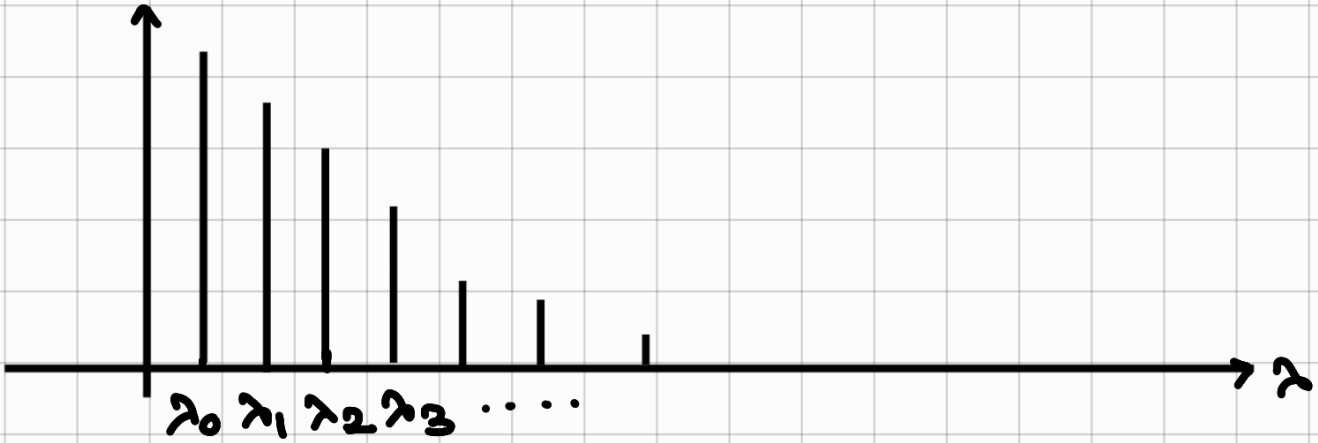
$$\underset{\sim}{Y} = XPP^T\underset{\sim}{\beta} + \underset{\sim}{\epsilon} \Longrightarrow \underset{\sim}{Y} = Z\underset{\sim}{\alpha} + \underset{\sim}{\epsilon}$$

$$\Rightarrow \hat{\alpha} = (Z^TZ)^{-1}Z^T\underset{\sim}{Y}$$

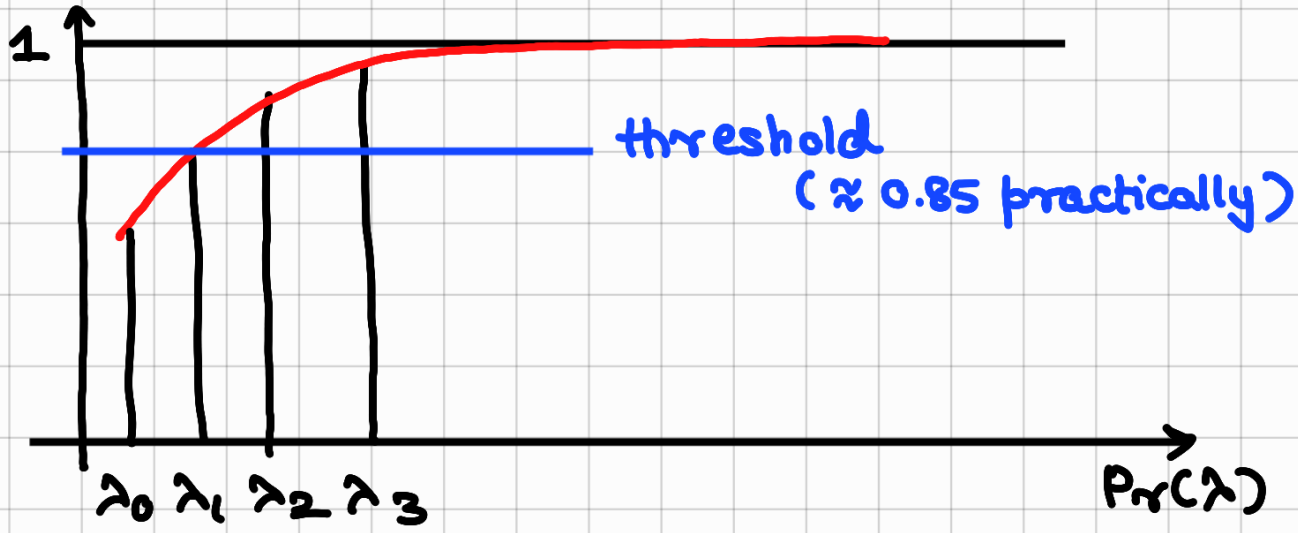
As a process,

$$\underset{\sim}{\alpha}_{(r)} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_r \end{pmatrix}$$

(reducing the dimensions)



$$P_r(\lambda) = \sum_{i=0}^{t} \lambda_i \Big/ \sum_{i=0}^{k} \lambda_i$$



$Z_{(r)}$ is the generated matrix from X and

$[\underset{\sim}{V_0}\ \underset{\sim}{V_1}\ \dots\ \underset{\sim}{V_r}]$

$$Z_{(r)} = \underset{n\times(k+1)}{X}\ [\underset{\sim}{V_0}\ \underset{\sim}{V_1}\ \dots\ \underset{\sim}{V_r}]$$

$$\hat{\alpha}_{(r)} = (Z_{(r)}^T Z_{(r)})^{-1} Z_{(r)}^T Y$$

Transform $\hat{\alpha}_{(r)}$ to β,

$\beta = I\beta$

$\Rightarrow \beta = PP^T\beta$

$\Rightarrow \beta = P.\alpha$

$$\Rightarrow \hat{\beta}_{PC_{(k+1)}} = P.\begin{pmatrix} \hat{\alpha}_{(r)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Tentatively, Midsem syllabus is upto and including Polynomial Regression.

$|X^TX| \approx 0$

$\longrightarrow$ PCR

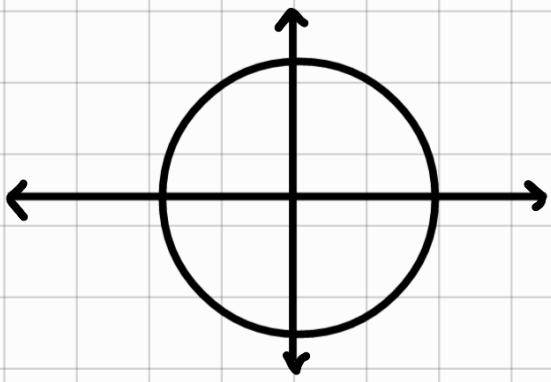
$\longrightarrow$ Shrinkage estimation or regularization

Shrinkage estimation / regularization

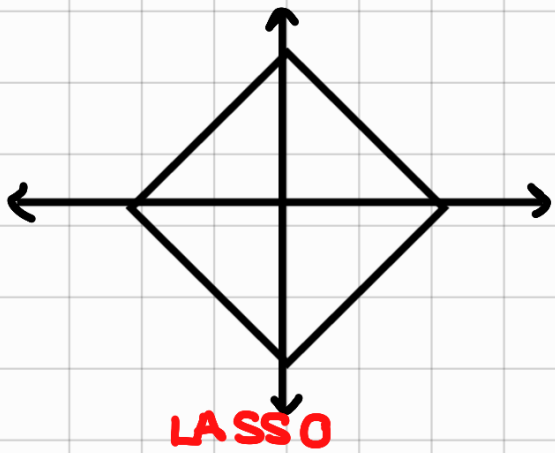
As $|X^TX| \approx 0$, the variance of $\hat{\beta}_{LS}$ is large. Hence, we want some bound on the norm of $\hat{\beta}$.

L_2 : $\|\hat{\beta}\|_2^2 \leq c_2$ or L_1 : $\|\hat{\beta}\|_1 \leq c_1$

$\Leftrightarrow \sum_{i=0}^{k} |\beta_i|^2 \leq c_2$ $\Leftrightarrow \sum_{i=0}^{k} |\beta_i| < c_1$



RIDGE



LASSO

Minimization of LS condition

$$S(\beta) = (y - X\beta)^T (y - X\beta) \text{ w.r.t. } \sum_{i=0}^{k} |\beta_i|^2 < c_2$$

Using Lagrangian multiplier

$$\hat{\beta}_R = \text{argmin } (S(\beta) + \lambda(\beta^T\beta - c_2))$$

$$\hat{\beta}_R = (X^TX + \lambda I_{k+1})^{-1} X^T y$$

Hoerl, Kennard, Baldan (1975)

$$\lambda = (k+1)\hat{\sigma}_{LS} / \hat{\beta}_{LS}^T \hat{\beta}_{LS}$$

These $\hat{\sigma}_{LS}$, $\hat{\beta}_{LS}$ are generated from original model with/without using PCR depending on $(X^TX)^{-1}$.

Note:

1) $\lambda \to \infty \Rightarrow \hat{\beta}_R = 0$
2) $\lambda \to 0 \Rightarrow \hat{\beta}_R = \hat{\beta}_{LS}$

Is $\hat{\beta}_R$ an unbiased estimator of β?

$$\hat{\beta}_R = (X^TX + \lambda I_{k+1})^{-1}(X^TX)(X^TX)^{-1}X^Ty$$
$$= \underbrace{[(X^TX + \lambda I_{k+1})^{-1}(X^TX)]}_{\text{not an identity matrix}} \hat{\beta}_{LS}$$

$\hat{\beta}_R$ is not an unbiased estimator.

Notations:

$W = (X^TX + \lambda I)^{-1}$

$S = (X^TX)$

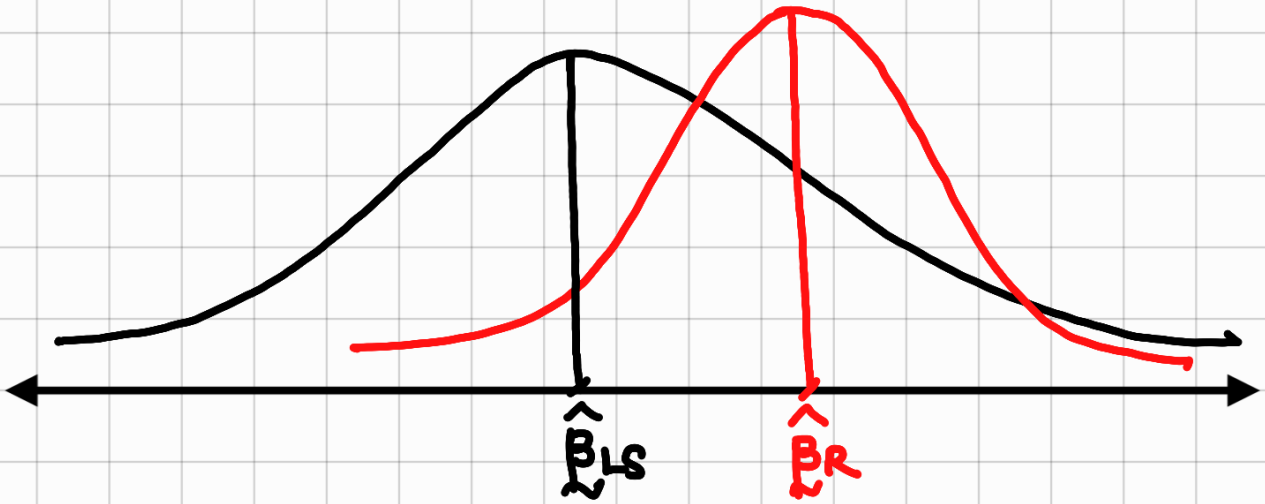
$\therefore \hat{\beta}_R = WS\hat{\beta}_{LS}$

Does RIDGE estimate have less variance?

$$D(\hat{\beta}_{LS}) - D(\hat{\beta}_R) = \sigma^2(S^{-1} - WSS^{-1}SW)$$
$$= \sigma^2(S^{-1} - WSW)$$
$$= \sigma^2 W(W^{-1}S^{-1}W^{-1} - S)W$$
$$= \sigma^2 W((S+\lambda I)S^{-1}(S+\lambda I) - S)W$$
$$= \sigma^2 W(2\lambda I + \lambda^2 S^{-1})W$$

psd psd psd

Hence, $D(\hat{\beta}_{LS}) - D(\hat{\beta}_R)$ is a psd matrix.
This implies RIDGE estimate has less variance.

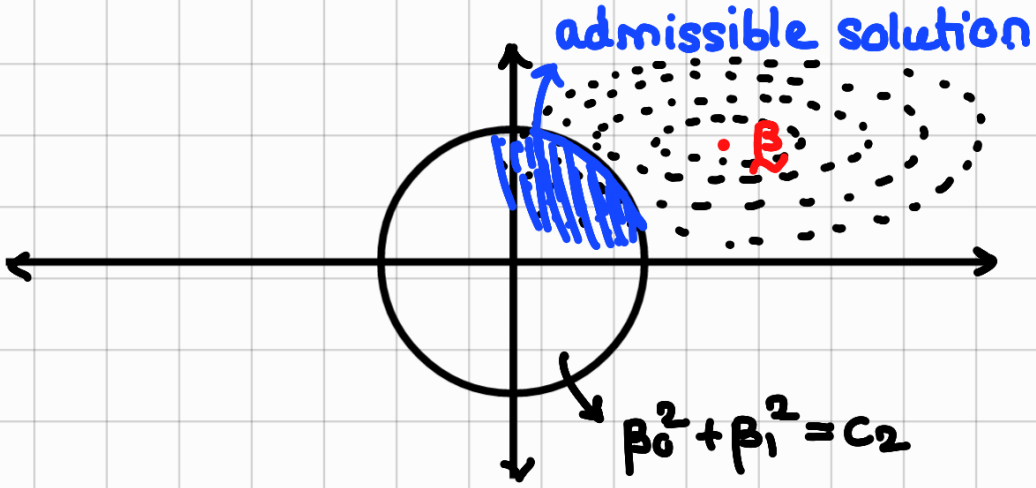


Bias $E(\hat{\beta}_R) - \hat{\beta} = (WS - I)\beta$

$$MSE(\hat{\beta}_R) = tr[D(\hat{\beta}_R)] + \beta^T W^2 \beta \lambda^2$$
$$= \sigma^2 \sum_{i=0}^{k} \lambda_i/(\lambda_i+\lambda)^2 + \lambda^2 \sum_{i=0}^{k} \beta_i^2/(\lambda_i+\lambda)^2$$
$$= \sigma^2 \sum_{i=0}^{k} \lambda_i/(\lambda_i+\lambda)^2 + \sum_{i=0}^{k} \beta_i^2/(1+\lambda_i/\lambda)^2$$

λ_i's are eigenvalues of X^TX.

Solution space for RIDGE Geometrical Intuition



Similar steps for LASSO (differentiation is a bit complex)

Variable selection

$$Y = X\beta + \epsilon$$
$$= [X_p | X_r] \begin{pmatrix} \beta_p \\ \hline \beta_r \end{pmatrix} + \epsilon$$

$(p+r=k+1)$

Data we received : (Y, X_p)
We need r more features to fit data into true model

True Model $Y = X\beta + \epsilon$
Reduced Model $Y = X_p\beta_p + \epsilon$ reduced according to available data

Q1. Is $\hat{\beta}_p$ an unbiased estimator ?
Q2. Does $\hat{\beta}_p$ have less variation compared to the true one ?
Q3. Is σ_p^2 (estimate value of σ^2 in reduced model) an unbiased estimator of σ^2 ?

Note: $\Sigma = \left(\begin{array}{c|c} \Sigma_{pp} & \Sigma_{rp} \\ \hline \Sigma_{pr} & \Sigma_{rr} \end{array}\right)$

reduced model, Σ_{pp} is to be used.

In reduced model,

$$\hat{\underset{\sim}{\beta}}_p = (X_p^T X_p)^{-1} X_p^T \underset{\sim}{Y}$$

$$\begin{aligned} E(\hat{\underset{\sim}{\beta}}_p) &= (X_p^T X_p)^{-1} X_p^T E(\underset{\sim}{Y}) \\ &= (X_p^T X_p)^{-1} X_p^T [X_p \underset{\sim}{\beta}_p + X_r \underset{\sim}{\beta}_r] \\ &= \underset{\sim}{\beta}_p + (X_p^T X_p)^{-1}(X_p^T X_r)\underset{\sim}{\beta}_r \end{aligned}$$

Hence, $\hat{\underset{\sim}{\beta}}_p$ is an unbiased estimator if $X_p^T X_r = 0$
This would happen in case of PCA and orthogonal polynomials.

A1. Not true in general

$$\sigma_p^2 = \frac{\underset{\sim}{Y}^T (I_n - X_p (X_p^T X_p)^{-1} X_p^T)\underset{\sim}{Y}}{n-p}$$

$$\begin{aligned} E(\sigma_p^2) &= [\sigma^2(n-p) + \sigma^2(ncp)]/(n-p) \\ &= \sigma^2 + \sigma^2(ncp)/(n-p) \end{aligned}$$

$$ncp = \frac{1}{\sigma^2}(X\beta)^T (I_n - X_p (X_p^T X_p)^{-1} X_p^T) X\underset{\sim}{\beta}$$

$$= 0+0+0+\frac{1}{\sigma^2}(\underset{\sim}{\beta}_r^T X_r^T (I_n - P_{X_p}) X_r \underset{\sim}{\beta}_r)$$

{ expand $X\underset{\sim}{\beta} = X_p\underset{\sim}{\beta}_p + X_r\underset{\sim}{\beta}_r$
all X_p terms get cancelled }

$$ncp = \frac{1}{\sigma^2}(\underset{\sim}{\beta}_r^T X_r^T (I_n - P_{X_p}) X_r \underset{\sim}{\beta}_r)$$

A3. No (it is unbiased only when we capture all features)

Note: $\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$. Then,

$\Sigma^{-1} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}$, and

$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$

given in prof's slides

Project (Credits : Anubhab Mandal)

20/30 marks

Group size → 4, 5 or 6

Report → 5/6 pages ①
+ code (analysis of dataset using things taught in class) ②

Justification of maths + dataset description

+ PDF of executed code (Python or R) ③

30 marks → Video recording (10 mins) ④ of project (Teams)

Upload all the stuff in a drive folder

All group members must submit project folder by the same name so that checked together

Submission → 15th April 2024

Generalized Linear Model

In linear model : $y_i = \beta^T x_i + \epsilon_i$, $i = 1, 2, \ldots, n$

$E(\epsilon_i) = 0$ $V(\epsilon_i) = \sigma^2$

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$E(y_i | x_i) = \beta^T x_i = I(\beta^T x_i)$ ($I(x) = x$.)

For categorical variables as output, we need

$E(y_i | x_i) = g(\beta^T x_i)$

$x_i \in \mathbb{R}^{k+1} \xrightarrow{\text{linear combination}} x_i^T \beta \in \mathbb{R} \xrightarrow{g} (0,1) \longrightarrow \{0,1\}$ or $\{A, B, C, D\}$

GLM — classification problem

$$g(x) = \begin{cases} \dfrac{e^z}{1+e^z} & z \in \mathbb{R} \quad \text{logit-model} \\ \text{any cdf of a cont. random variable} & \\ \Phi(z) & z \in \mathbb{R} \quad \text{probit-model} \end{cases}$$

Logistic Regression

$$P(y_i = 1) = 1 - P(y_i = 0) = \pi_i = E(y_i | x_i) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

$$\iff \log_e\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T \beta \in \mathbb{R}$$

$\pi_i/(1-\pi_i)$ is known as 'odd' of π_i

Two treatment problem; often defined as $\log_e\left(\frac{p_A/1-p_A}{p_B/1-p_B}\right)$

→ log of odds ratio.

≈ asymptotically follow normal distribution.

$$\psi = \log_e\left(\frac{p_A/1-p_A}{p_B/1-p_B}\right) \; ; \quad \hat{\psi} \sim N(0, \sigma_y^2) \text{ for large } n.$$

Let $y \overset{iid}{\sim}$ Bernoulli(p)

$f(y) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$ joint pdf

$$= p^{\sum_i^n y_i}(1-p)^{n-\sum_i y_i} = \left(\frac{p}{1-p}\right)^{\sum_i^n y_i}(1-p)^n$$

$$\log_e f(\underset{\sim}{y}) = \left(\sum_{i=1}^{n} y_i\right)\left(\log_e\left(p/1-p\right)\right) + n\log(1-p) + 0$$

$$= D_1(\underset{\sim}{y})\, D_2(p) + D_3(p) + D_4(\underset{\sim}{y})$$

For any distribution (almost any), we are able to decompose $\log_e f(\underset{\sim}{y})$ in this form.

$D_1(\underset{\sim}{y})$ → sufficient statistic (all info. of data)

$D_2(p)$ → natural parameter structure

$D_3(p)$ → efficient parameter estimation from data

$D_4(\underset{\sim}{y})$ → carries certain info. about ancillary part depending on family of distribution. does not depend upon parameter

Due to this natural structure, logistic model is preferred over other models.

$$\left(\begin{array}{c} \theta = \log p/1-p \\ p = e^{\theta}/1+e^{\theta} \end{array}\right)$$

Now,

$$E(y_i|x_i) = \pi_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

Likelihood fn. of (β)

$$L(\beta|data) = \pi_{i=1}^{n} \{ \pi_i^{y_i} (1-\pi_i)^{1-y_i} \}$$

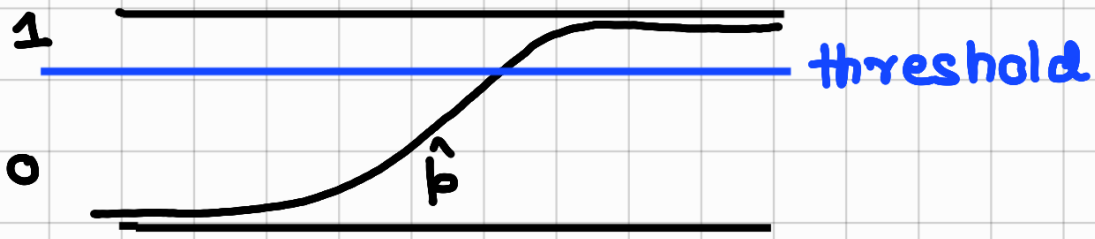
$$\log(L(\beta|y)) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) \right] + \sum_{i=1}^{n} \log(1-\pi_i)$$

$$L(\beta) = \sum_{i=1}^{n} [y_i (x_i^T \beta)] - \sum_{i=1}^{n} \log_e (1 + e^{x_i^T \beta})$$

$$\frac{\partial L(\beta)}{\partial \beta} = 0$$

we get $\hat{\beta}$ as an iterative solution

$$\hat{p} = \frac{e^{\hat{\beta} x}}{1 + e^{\hat{\beta} x}}$$



Error Analysis of MLR

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ assumption

$y_i = x_i^T \beta + \epsilon_i$ data $\qquad H \equiv P_x$

$e_i = y_i - \hat{y}_i$ estimated error

$$e = (Y - \hat{Y}) \sim N(0, \sigma^2 (I_n - P_x))$$

$$\hat{\sigma}^2 = \frac{e^T e}{n-k-1}$$ unbiased estimate of σ^2

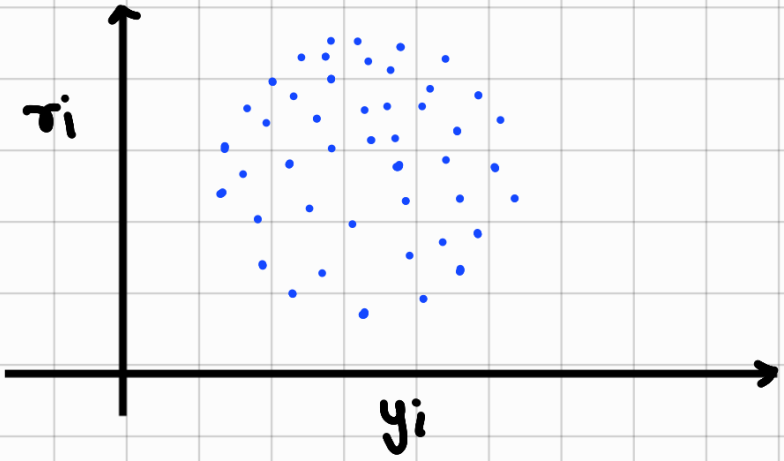
$$cov(e_i, e_j) = \begin{cases} \sigma^2 (1 - h_{ii}) & i = j \\ -\sigma^2 h_{ij} & i \neq j \end{cases}$$

$$MSError = \frac{SSError}{n-k-1}$$

Standardized Residual $d_i = e_i / \sqrt{MSError}$

Studentized Residual $r_i = e_i / \sqrt{MSError(1-h_{ii})}$

If all the analysis is correct, the pattern resembles:



Homework: Show that

$$r_i = \frac{e_i}{\left(MSError\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]\right)^{1/2}} \text{ for SLR.}$$

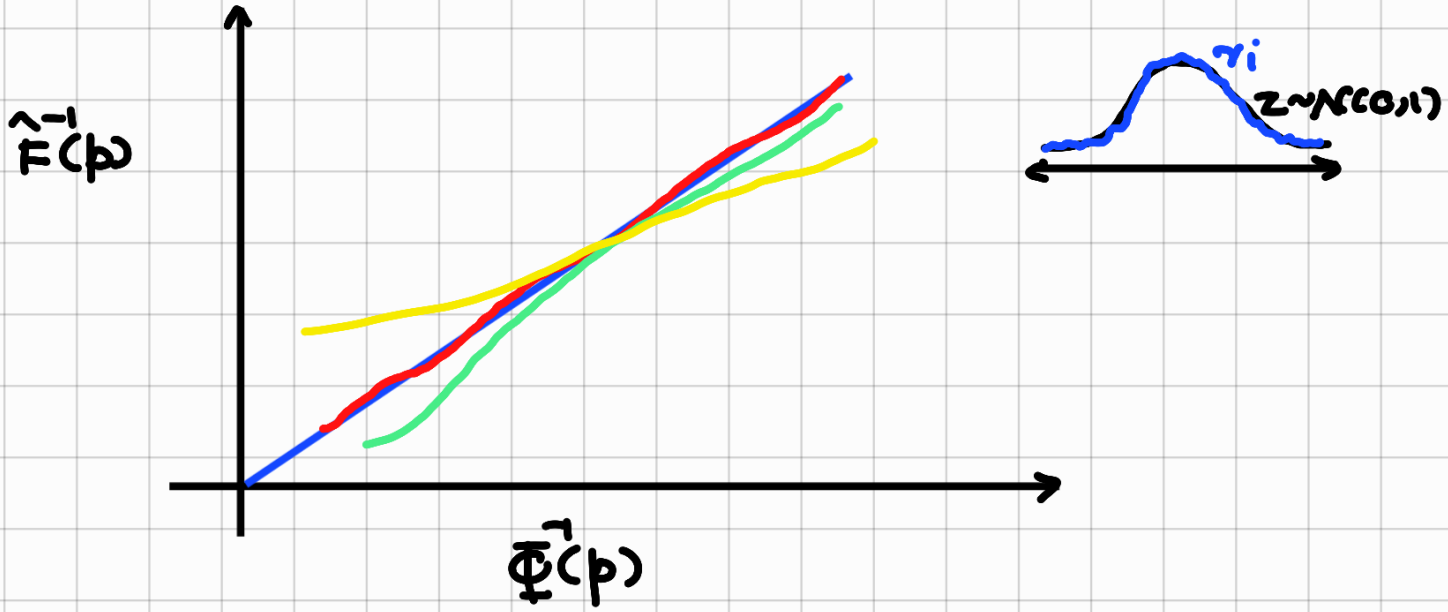
If $n \uparrow$ to ∞ and $k < \infty$, $\hat{\sigma}^2 = MSError$ will converge to σ^2 with probability 1.

$\Rightarrow r_i \overset{a}{\sim} N(0,1)$

To validate this,

- → Q-Q plot
- → χ^2-test
- → KS test

Q-Q Plot

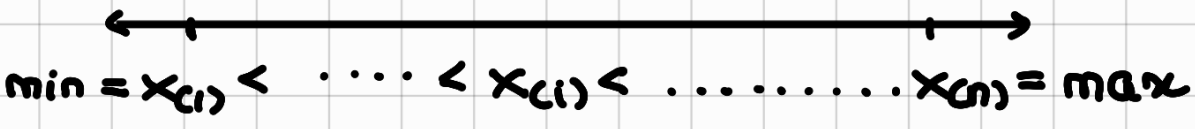


Tail part can most likely fluctuate. The middle part more or less follows the plot

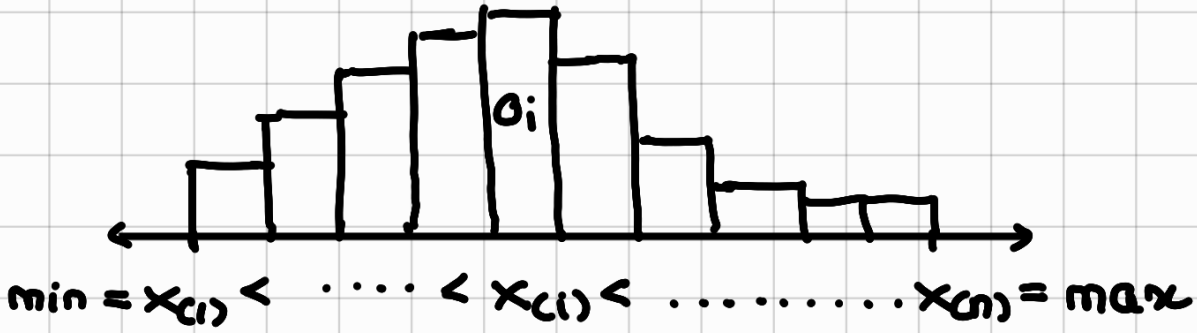
$\Phi^{-1}(p)$ cdf of $N(0,1)$
$\hat{F}$ is the empirical cdf of plot

χ^2-test (goodness of fit)

pdf based test



→ arrange the data in an increasing order
→ divide the range $(x_{(1)}, x_{(n)})$ into k-parts, such that each subinterval has more than 5 data



→ Let there are O_i many observations in the ith interval

→ E_i = expected no. of observations within ith interval

= $n \times p_i$

= (total obs) × prob. of ith interval

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

df = k-1 when parameters are known

df gets decremented with increasing no. of unknown parameters.

If we use $\hat{E_i} = n \times \hat{p_i}$, $\hat{p_i}$ is the probability after estimating parameters from the data, then it will follow $\chi^2_{(k-1)-(\text{no. of parameters estimated})}$

KS Test

Assume that the target distribution specification is completely known.

$$Z_i \sim F(\text{completely known})$$

Z_i are continuous r.v.

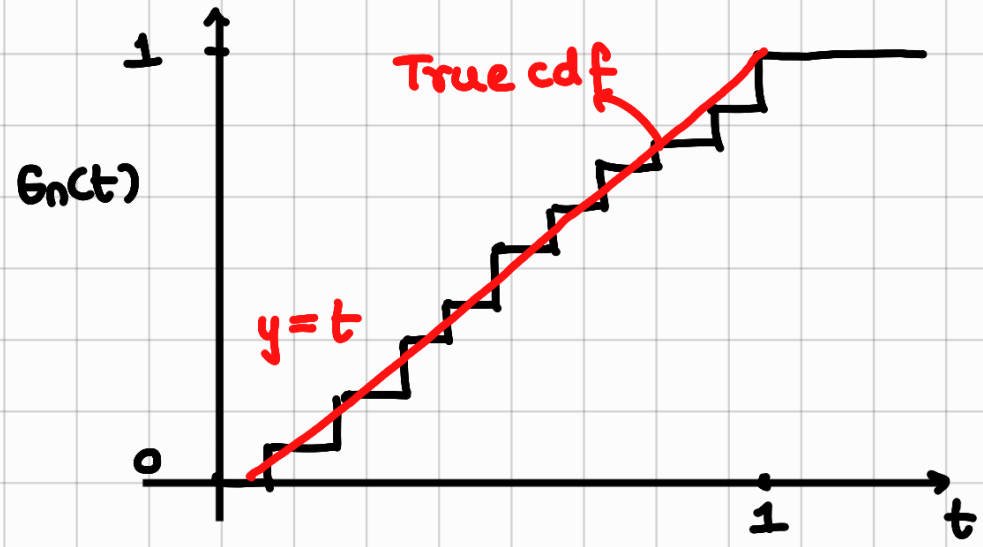
→ test only applicable to continuous r.v.

$W_i = F(Z_i) \overset{iid}{\sim} U(0,1)$

Empirical cdf of W is given as

$G_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{w_i \leq t\}}$

Plot of $G_n(t)$ v/s t looks like :



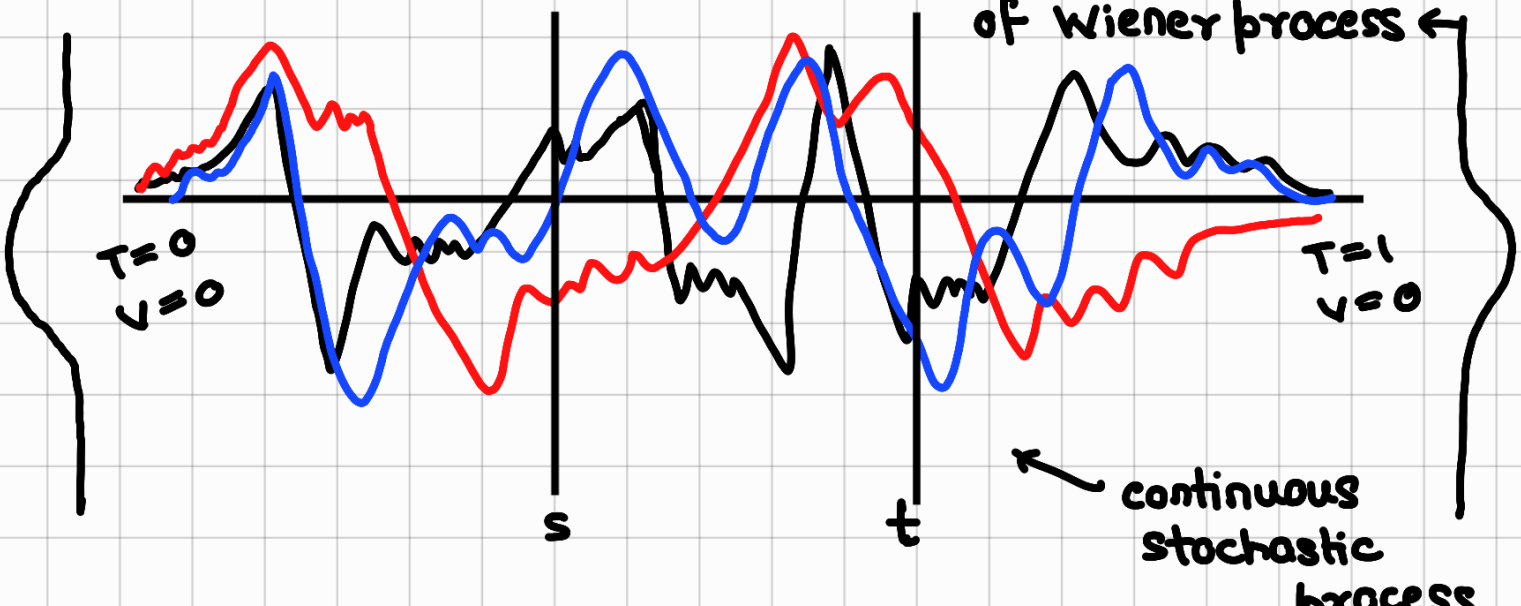
Test statistic: $\sup_{0 \le t \le 1} \sqrt{n}\,|G_n(t)-t| = D$

D is said to follow Kolmogrove distribution as $n \uparrow \infty$.

If $Z_i \sim F$ then $|G_n(t)-t| \longrightarrow 0$ a.s. (almost surely)
{Measure Theory PTSD}

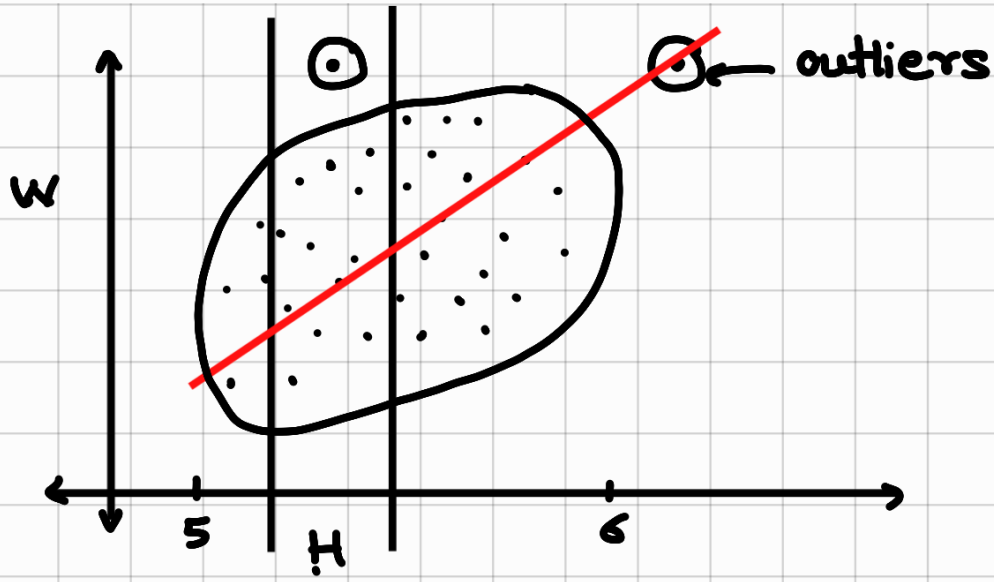
Kolmogorov Distribution:

Brownian Bridge:



$$\begin{pmatrix} B_0(s) \\ B_0(t) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s(1-s) & s(1-t) \\ s(1-t) & t(1-t) \end{pmatrix} \right)$$

$K = \sup_{0 \le t \le 1} |B_0(t)|$ ← r.v. of Kolmogorov distn.



$e_i = y_i - \hat{y}_i$

$= y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}$

we used all $(y_i, \underset{\sim}{x}_i)$ to get $\hat{\underset{\sim}{\beta}}$.

$n-1$ {
~~$(y_1, \underset{\sim}{x}_1)$~~
$(y_2, \underset{\sim}{x}_2)$
:
$(y_n, \underset{\sim}{x}_n)$

get $\hat{\underset{\sim}{\beta}}_{(1)}$

$\hat{y}_{(1)} = \underset{\sim}{x}_1^T \hat{\underset{\sim}{\beta}}_{(1)}$

$e_{(1)} = (y_1 - \hat{y}_{(1)}) = (y_1 - \underset{\sim}{x}_1^T \hat{\underset{\sim}{\beta}}_{(1)})$

$n-1$ {
$(y_1, \underset{\sim}{x}_1)$
~~$(y_2, \underset{\sim}{x}_2)$~~
:
$(y_n, \underset{\sim}{x}_n)$

get $\hat{\underset{\sim}{\beta}}_{(2)}$

$\hat{y}_{(2)} = \underset{\sim}{x}_2^T \hat{\underset{\sim}{\beta}}_{(2)}$

$e_{(2)} = (y_2 - \hat{y}_{(2)}) = (y_2 - \underset{\sim}{x}_2^T \hat{\underset{\sim}{\beta}}_{(2)})$

Jack knife method

for i=1 to n

1) remove the observation
2) do the analysis for (n-1) data
3) Compute the estimated error for ith data

$$e_{(i)} = y_i - \hat{y}_{(i)} = y_i - \hat{\beta}_{(i)}^T x_i$$

Prediction Residual sum of square (PRESS)

$$\sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} (y_{(i)} - y_i)^2$$

It can be shown that $\frac{e_{(i)}}{\sqrt{Var(e_{(i)})}} = \frac{e_i}{\sqrt{V(e_i)}}$

Hence, fitting the whole n observations serves the purpose.

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \sim N\left(0, \frac{\sigma^2}{1-h_{ii}}\right)$$

$$\frac{e_{(i)}}{\sqrt{Var\ e_{(i)}}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}} = r_i$$

$\hat{\sigma}^2$ = MSError may not be a good estimator when the ith observation has been removed

We can use

$$\hat{\sigma}^2 = S_{(i)}^2 = \frac{(n-k-1)MSError - (e_i^2/1-h_{ii})}{n-k-2}$$

$$= \frac{\sum_{i=1}^{n} e_i^2 - e_i^2/1-h_{ii}}{n-k-2}$$

$$T = \frac{e_{(i)}}{S_{(i)}^2/(1-h_{ii})} \sim t_{n-k-2} \text{ under } H_0$$

H_0: ith observation is NOT an outlier

H_1: ith observation is an outlier

$$[X_{(i)}^T X_{(i)}]^{-1} = [X^T X - x_j x_j^T]^{-1}$$

$(k+1)\times(n-1)\quad(n-1)\times(k+1)$ $\qquad$ $(k+1)\times n\quad n\times(k+1)\quad(k+1)\times 1\quad 1\times(k+1)$

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

For related proofs, refer to Appendix C.7 and C.8 from the reference book { The prof is not accountable for the proof :) }

Regression syllabus is completed.

Time-Series

Relation b/w exponential and geometric distribution:

$X \sim \exp(\lambda) \longrightarrow$ continuous

$F(x) = 1 - e^{-\lambda x}$

Consider $Y = [X]$

$$\begin{aligned} P(Y = r) &= P([X] = r) \\ &= P(r \le X < r+1) \\ &= P(X < r+1) - P(X < r) \\ &= (1 - e^{-\lambda(r+1)}) - (1 - e^{-\lambda r}) \\ &= (e^{-\lambda})^r (1 - e^{-\lambda}) \\ &\sim q^r p \longleftarrow \text{geometric}(p = 1 - e^{-\lambda}) \end{aligned}$$

$\longrightarrow$ discrete

This implies that the adjectives of time (discrete, continuous) depend upon the context of approximation of time, and not the data.

Hence, discrete/continuous are adjectives of time and not the data.

Example of time-series

eg 1 White Noise

A time series $\{W_t\}$ is said to follow white noise if

- $E(W_t) = 0$
- $V(W_t) = \sigma_\omega^2$
- and they are UNCORRELATED

$W_t \sim WN(0, \sigma_\omega^2)$ (WN stands for white noise)

- $x_i \overset{iid}{\sim} N(0, \sigma^2)$
- $x_i \sim \begin{cases} N(0,1) & \text{when } i \text{ is even} \\ \exp(1) - 1 & \text{when } i \text{ is odd} \end{cases}$

Homework: Construct uncorrelated but dependent WN for both the cases.

(i) WN need not be normally distributed

(ii) WN need not be iid

(iii) iid sequence with zero mean and finite variance are always WN.

(iv) WN is weakly stationary.

(v) WN with normal distn. are strongly stationary.

eg.2 Binary process

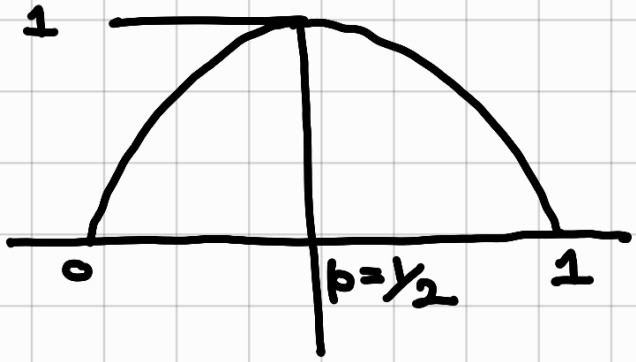
$$x_t = \begin{cases} +1 & \text{with prob. } p = 1/2 \\ -1 & \text{with prob. } 1-p = 1/2 \end{cases}$$

$E(x_t) = 2p - 1$

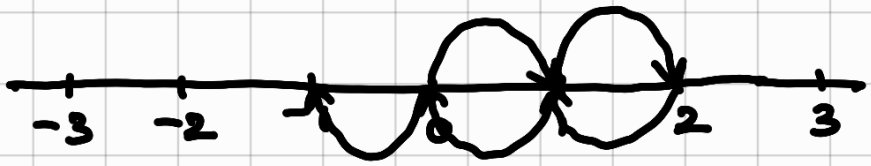
$Var(x_t) = 4p(1-p)$

$Z_t \overset{iid}{\sim} Bernoulli(p)$

$x_t = 2Z_t - 1$



e.g 3 Random walk on $\mathbb{Z}$



$x_0 = 0$

$x_t = x_0 + \sum_{t=1}^{T} W_t$

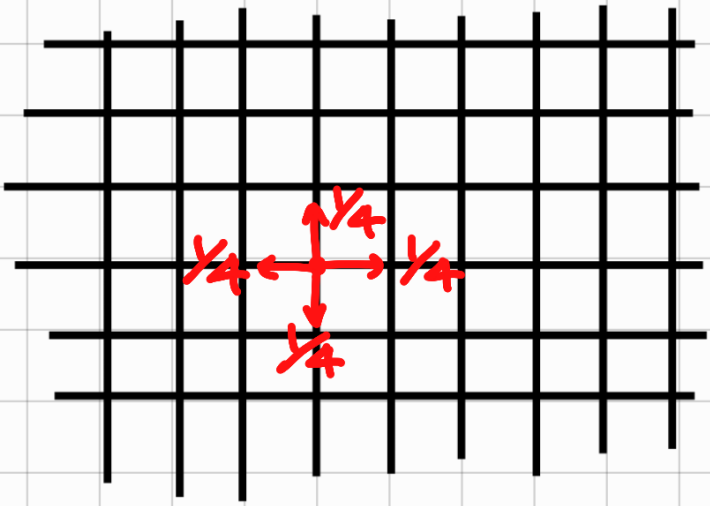
where W_t is a binary process with prob. (p).

$\binom{2n}{n} p^n (1-p)^n$: probability that it returns to 0 in $2n$ steps

Apply stirlings approximation and check for $n \to \infty$.

(i) If $(p = 1/2)$ then sequence will eventually return to zero

(ii) Random walk on $\mathbb{Z}^2$



In $\mathbb{Z}^2$ with equal probability $1/4$ the sequence will eventually return to $(0,0)$.

Reference for proofs of above results;

Markov Chains by Norris

e.g.4. Random walk width drift

$W_t \overset{iid}{\sim}$ $E(W_t) = \delta$

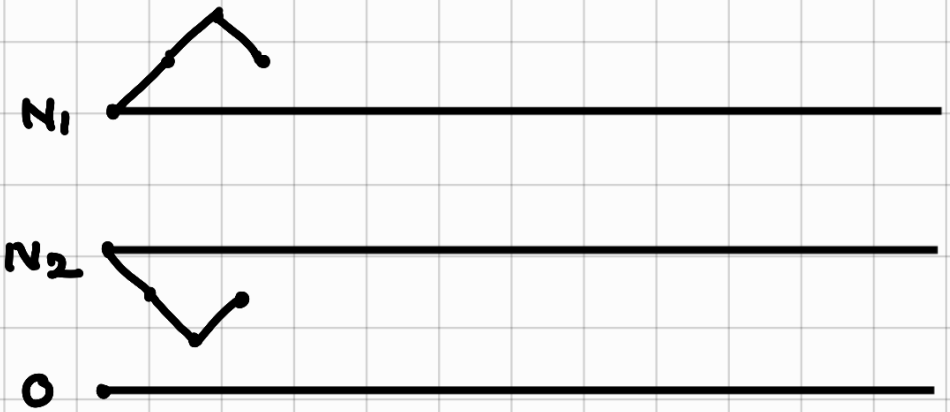
$V(W_t) = \sigma_w^2$

$x_t = x_0 + \sum_{i=1}^{t} W_i$

$E(x_t) = \delta t$

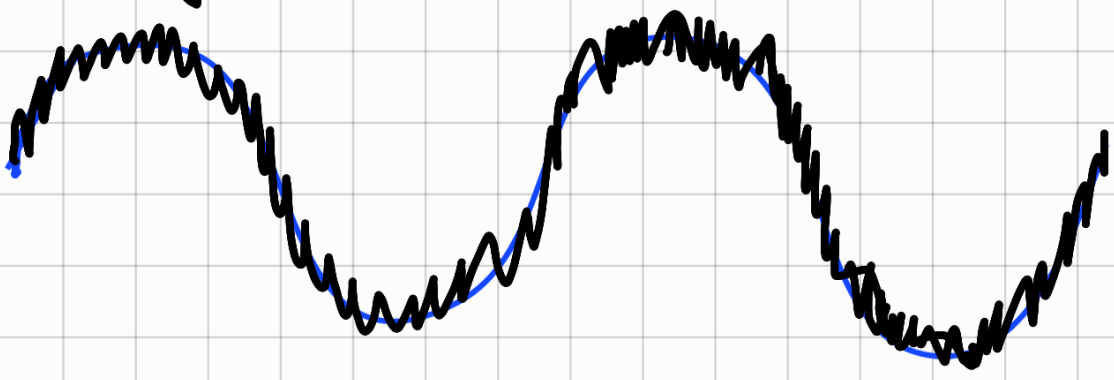
$V(x_t) = \sigma^2 t$

Gambler's ruin problem



First player which reaches 0 coins loses.

e.g.5 Signal with noise



$x_t = A\sin(2\pi f + \phi) + W_t$

$W_t \sim WN(0, \sigma_w^2)$

e.g.6 Moving average prices (order one)

MA(1) process

Let $W_t \sim WN(0, \sigma^2)$

$x_t = 1W_t + \theta W_{t-1}$

e.g. 7 Autoregressive process (order one)

AR(1)

$x_t = \phi x_{t-1} + w_t \quad |\phi| < 1, \ \phi \neq 0$

→ analogous / similar to the regression problem:

$Y_i = \beta^T x_i + \epsilon_i$

Wiener process (Brownian motion)

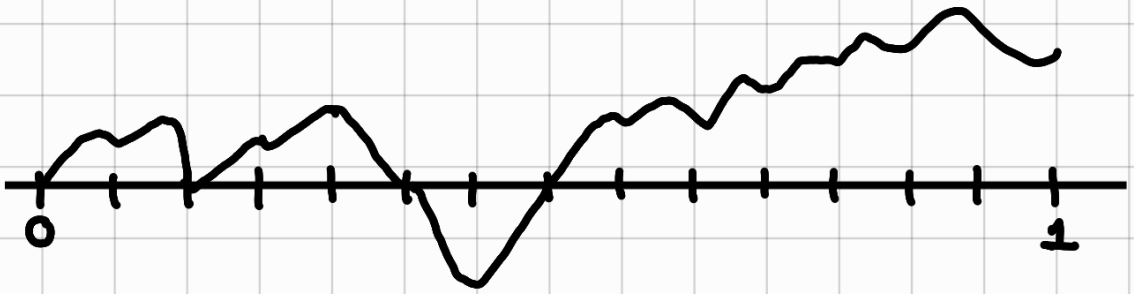
For definition, refer to the slides.

Generating data from Brownian motion (#)

$T_1 = \sup_{0<t<1} W_t$

$T_2 = \int_0^1 w_t^2 \, dt$

→ statistics used to measure qualities of brownian motion



$W_t \sim$ Brownian motion

$y = x^2$ on $(0,1)$

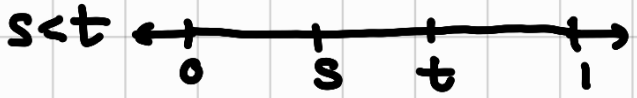
Remark: Let $\{W_i\}$ be a sequence of iid rvs with $E(W_i) = 0$ and $Var(W_i) = 1$ then $Y_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} W_k$ converges to Wiener process X_t on $[0,1]$ for large n.

$$Y_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} W_k$$

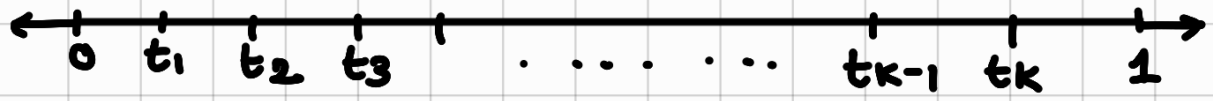
$E(Y_n(t)) = 0$

$$\lim_{n\uparrow\infty}(\mathrm{Var}(Y_n(t)) = \sqrt{\frac{[nt]}{n}}\left(\frac{1}{\sqrt{[nt]}}\sum_{k=1}^{[nt]} W_k\right) \xrightarrow{d} \sqrt{t}\,N(0,1) \equiv N(0,t)$$

Brownian Bridge:



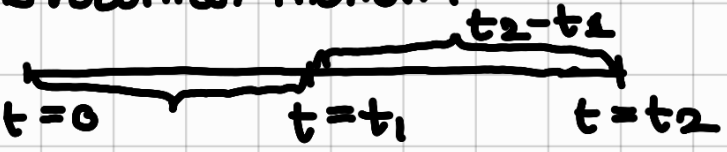
$$\begin{pmatrix} Y_n(s) \\ Y_n(t) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s & \min(s,t) \\ \min(s,t) & t \end{pmatrix}\right)$$



$$\begin{pmatrix} Y_n(t_1) \\ Y_n(t_2) \\ \vdots \\ Y_n(t_k) \end{pmatrix} \sim N\left(\underset{\sim}{0}, \begin{pmatrix} t_1 & \cdots & \min(t_1,t_j) & \cdots \\ \vdots & & \vdots & \vdots \\ \min(t_k,t_j) & \cdots & \cdots & t_k \end{pmatrix}\right)$$

Description of Brownian motion:

$t = 0, X_0 = 0$



$t = t_1, X_{t_1} = X_0 + Z_1$ where $Z_1 \sim N(0, t_1)$

$t = t_2, X_{t_2} = X_{t_1} + Z_2$ where $Z_2 \sim N(0, t_2 - t_1)$

$\mathrm{Var}(X_{t_2}) = 0 + t_1 + t_2 - t_1 = t_2$

If we fix the initial and final time and we simulate a Brownian motion in b/w these, it forms a Brownian bridge.

Brownian Bridge ($B_0(t)$):

Let $B(t)$ be a brownian motion

$B(t_1) = a$ and $B(t_2) = b$, $t_1 < t_2$

$$\text{mean} = a + \frac{t-t_1}{t_2-t_1}(b-a) \equiv \text{function of } t$$

$$\text{Var} = \frac{(t_2-t)(t-t_1)}{(t_2-t_1)}$$

$$\text{cov} = \frac{(t_2-t)(s-t_1)}{(t_2-t_1)} \quad \text{between } B_0(t) \text{ and } B_0(s)$$

$$B_0(t) = B(t) - \frac{t}{T}B(T)$$

Standard Brownian bridge on $[0,1]$

$(t=0, B_0(t)=0)$ and $(t=1, B_0(t)=0)$ $\begin{cases} B_0(t) = B(t) - tB(1) \\ 0 \le t \le 1 \end{cases}$

$$E(B_0(t)) = 0 + \frac{t-0}{1}(0-0) = 0$$

$$\text{Var}(B_0(t)) = \frac{(1-t)(t-0)}{1} = t(1-t)$$

$$\text{Cov}(B_0(s), B_0(t)) = \frac{(1-t)(s)}{1} = s(1-t) \quad \underline{s<t}$$

Brownian Bridge ($B_0(t)$)

→ KS Test for goodness of fit

→ Change point detection in temporal data

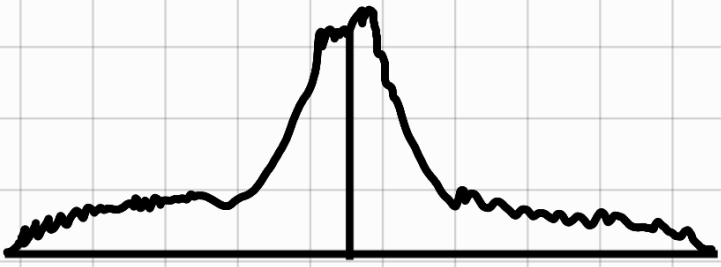
$$S_t := \frac{1}{\sqrt{n}} \sum_{i=1}^{[tn]} \frac{(x_i - \bar{x})}{\hat{\sigma}} \; ; 0 \le t \le 1$$ **CUSUM statistic**

↳ estimated process variance

$S_t \Rightarrow B_0(t)$

$|S(t)|$

$T = \sup_{0 \le t < 1} |S(t)|$



⟶ these statistics are such that software cannot directly estimate.

⟶ we need to simulate the data to find the cut-off values.

$$T = \sup_{0 \le t < 1} |S(t)| \longrightarrow \sup_{0 \le t \le 1} |B_0(t)|$$

Hypothesis Testing:

$H_0: E(X_1) = E(X_2) = \dots = E(X_n)$

$H_1: E(X_1) = E(X_2) = \dots = E(X_k) \neq E(X_{k+1}) \dots = E(X_n)$

Formulation of variance and covariance:

$X_1, X_2, \dots, X_n \overset{iid}{\sim} F$ continuous distribution

$Y = F(X) \sim U(0,1)$

$E\left(\sum_{i=1}^{n} I_{(Y_i < t)}\right) = nt, \quad 0 \le t \le 1$

$V\left(\sum_{i=1}^{n} I_{(Y_i < t)}\right) = nt(1-t)$

$\text{Var}\left(\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} I_{(Y_i < t)} - \sqrt{n}t\right)\right) = t(1-t)$

$$H_n(t) = \sqrt{n}\left(\frac{1}{n}\sum I_{(Y_i < t)} - t\right)$$

$\text{Cov}(H_n(s), H_n(t)) \quad \underline{s < t}$

$= \text{cov}\left(\frac{1}{\sqrt{n}}\sum_{i}^{n} I_{(Y_i < t)}, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{(Y_i < s)}\right)$

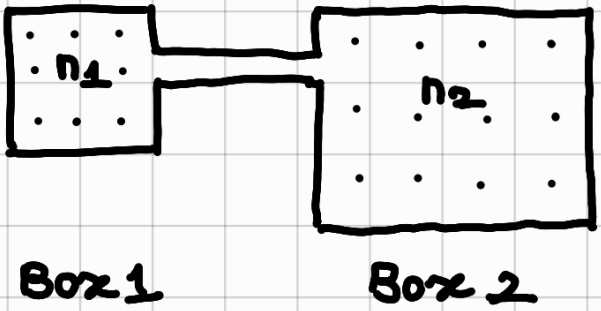
$= \frac{1}{n}\sum_{i}\sum_{j} \text{cov}\left[\left(I_{(Y_i < t)}\right), I_{(Y_j < s)}\right]$

$$\text{Cov} = \min(s,t) - st$$
$$= s(1-t)$$
$$\text{Cov} = \frac{1}{n}.n\{\min(s,t)-st\} = s(1-t)$$

Stationary Timeseries

→ strong stationarity
→ weak stationarity

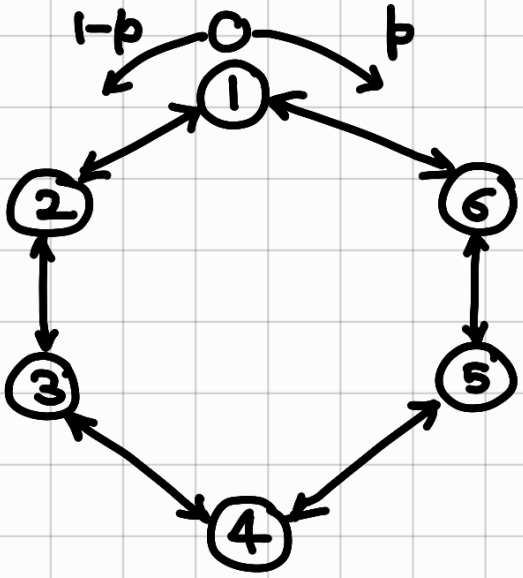
e.g.1 $x_t \sim \text{bin}(n_1+n_2, 1/2)$ as $t \to \infty$



$x_0 = n_1$

x_t = # balls on box 1 at time t

e.g.2



Limiting distribution of $\{x_t\}$ will follow uniform on $\{1,2,3,4,5,6\}$ $P(x_t=1) = 1/6$.

$x_t :=$ location of the ball at time t.

Transition probability matrix :

	1	2	3	4	5	6
1	0	q	0	0	0	p
2	p	0	q	0	0	0
3	0	p	0	q	0	0
4	0	0	p	0	q	0
5	0	0	0	p	0	q
6	q	0	0	0	p	0

Strongly stationary timeseries

Let $\{X_t\}$ be a time series with a joint distribution fn. of $(X_1, X_2, \ldots, X_n)$ as

$P(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n) = F_n(a_1, a_2, \ldots, a_n)$

Now, $\forall n \in \mathbb{N}$, $\forall k \in \mathbb{Z}$, $\forall h \in \mathbb{Z}$ and $\forall a_i \in \mathbb{R}$,

If $P(X_{k+1} \leq a_1, X_{k+2} \leq a_2, \ldots, X_{k+n} \leq a_n)$

$= P(X_{k+h+1} \leq a_1, X_{k+h+2} \leq a_2, \ldots, X_{k+h+n} \leq a_n)$

$= F(a_1, a_2, \ldots, a_n)$,

then $\{X_t\}$ is strongly stationary.

Weakly stationary timeseries

A time series $\{X_t\}$ is said to be weakly stationary if

i) $\mu_t = E(X_t)$ is free from t.

ii) $\text{Cov}(X_t, X_{t+h})$ is free from 't' but can be a function of 'h'.

→ this is similar to saying that this is upto second moment condition.

If a TS is strongly stationary with finite 2nd order moment, then it is weakly stationary.

Now, suppose $\{X_t\}$ is atleast weakly stationary.

$$\gamma_x(h) = \text{Cov}(X_t, X_{t+h})$$
$$= E\{(X_t - \mu)(X_{t+h} - \mu)\}$$

... $t+h = s$ (say)

$$= E\{(X_{s-h} - \mu)(X_s - \mu)\}$$
$$= \gamma_x(-h)$$

'h' is the lag.

Autocorrelation fn.

Auto-correlation coefficient or an 'atleast' weakly stationary timeseries is defined as

$$\rho_x(h) = \gamma_x(h) / \gamma_x(0).$$

Significance/Formulation :

correlation coeff. $\rho_x(h) = \dfrac{\text{cov}(X_{t+h}, X_t)}{\sqrt{\text{Var}(X_{t+h})\,\text{Var}(X_t)}}$

$$= \gamma_x(h) / \sqrt{\gamma_x(0)\,\gamma_x(0)}$$

$$= \gamma_x(h) / \gamma_x(0).$$

e.g.1 $\{X_t\} \overset{\text{WN or iid}}{\sim} E(X_t) = 0,\ \text{Var}(X_t) = \sigma^2$

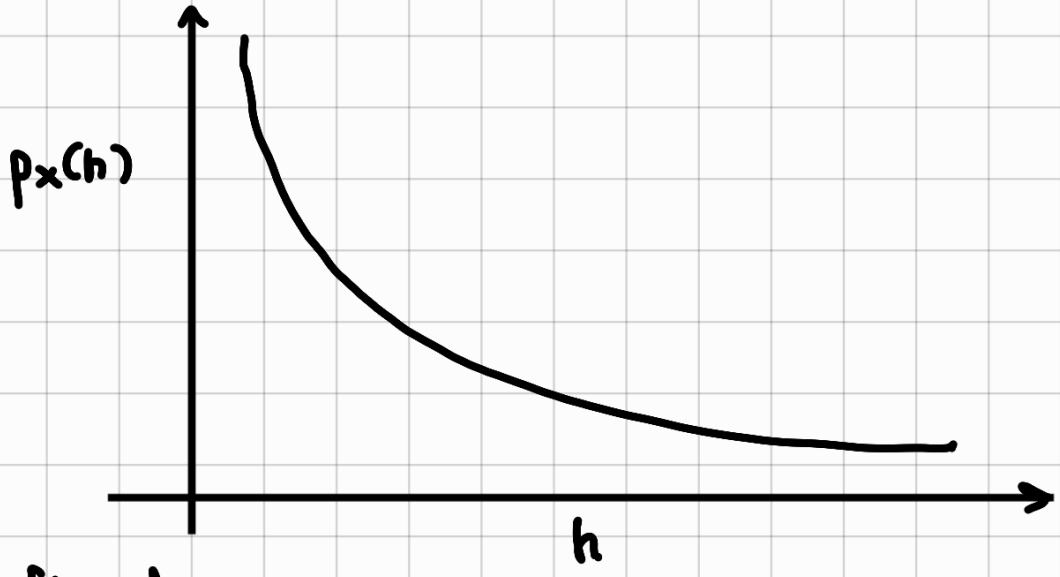
$$\gamma_x(h) = \begin{cases} \sigma^2 & h = 0 \\ 0 & h \neq 0 \end{cases}$$

e.g.2 $S_t = \sum_{i=1}^{t} W_i \quad W_i \overset{iid}{\sim} N(0, \sigma^2)$

Random walk $\gamma_x(h) = \begin{cases} t\sigma^2 & h = 0 \\ t\sigma^2 & h > 0 \end{cases}$

$$\rho_x(h) = \frac{\operatorname{cov}(S_t, S_{t+h})}{\sqrt{V(S_t)V(S_{t+h})}} = \frac{t\sigma^2}{\sigma^2\sqrt{t(t+h)}}$$

$$\rho_x(h) = \sqrt{t/_{t+h}}$$



t is fixed.

e.g.3 MA(1) Moving averages order one

$Z_t = W_t + \theta W_{t-1}$ where $W_t \overset{iid}{\sim} N(0, \sigma^2)$

$E(Z_t) = 0$

$V(Z_t) = (1+\theta^2)\sigma^2$

$$\gamma_Z(h) = \begin{cases} (1+\theta^2)\sigma^2 & h = 0 \\ \theta\sigma^2 & h = \pm 1 \\ 0 & |h| > 1 \end{cases}$$

e.g.4 Autoregressive process of order one AR(1)

$Z_t = \varphi Z_{t-1} + W_t$

- $W_t \overset{iid}{\sim} N(0, \sigma^2)$
- $|\varphi| < 1$
- $\varphi = 0$
- Z_t is weakly stationary

$E(Z_t) = \varphi E(Z_{t-1}) + E(W_t)$

$$E(Z_t) = \varphi E(Z_{t-1}) + 0$$
$$\Rightarrow E(Z_t) = 0.$$

$$V(Z_t) = E(Z_t^2) = E(\varphi^2 Z_{t-1}^2 + W_t^2 + 2\varphi Z_{t-1} W_t)$$
$$= \varphi^2 E(Z_{t-1}^2) + E(W_t^2) + 0$$
$$= \varphi^2 E(Z_{t-1}^2) + \sigma^2$$
$$E(Z_t^2) = \sigma^2/(1-\varphi^2)$$
$$V(Z_t) = \sigma^2/(1-\varphi^2)$$

$$E(Z_t) = 0$$
$$V(Z_t) = \sigma^2/(1-\varphi^2) > \sigma^2$$

→ this is similar to BVN case.

$(X, Y) \sim BVN$

$V(Y|X) = \sigma_y^2(1-\rho^2)$

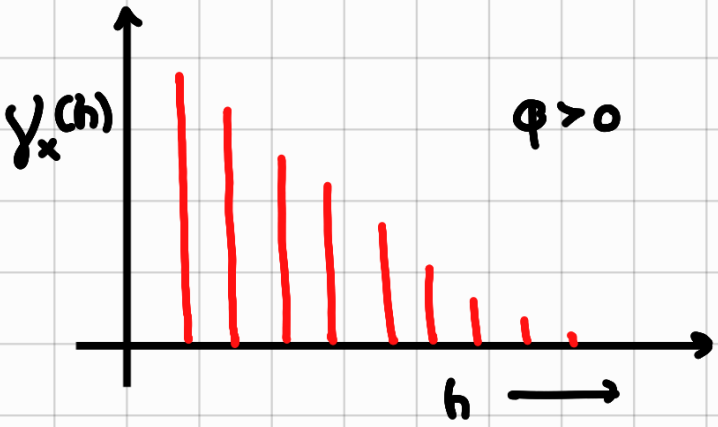
$V(Y) = \sigma_y^2$

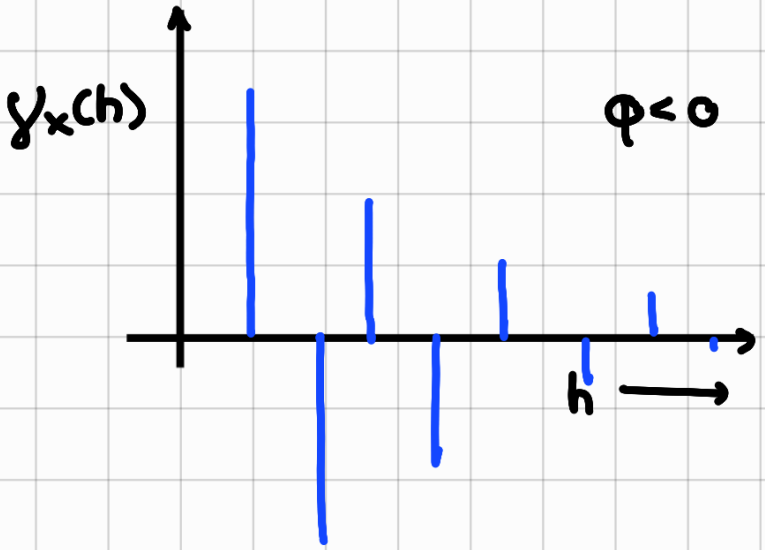
$$\gamma_Z(h) = \begin{cases} (\sigma^2/(1-\varphi^2)) & h = 0 \\ (\sigma^2/(1-\varphi^2))\varphi^h & h \neq 0 \end{cases}$$

$$\gamma_Z(h) = \text{cov}(Z_{t+h}, Z_t)$$
$$= \text{cov}(\varphi Z_{t+h-1} + W_{t+h}, Z_t)$$
$$= \varphi \,\text{cov}(Z_{t+h-1}, Z_t)$$
$$= \varphi^2 \,\text{cov}(Z_{t+h-2}, Z_t)$$
$$\cdots$$
$$= \varphi^h \,\text{cov}(Z_t, Z_t)$$
$$= \varphi^h \left(\frac{\sigma^2}{1-\varphi^2}\right) = \gamma_Z(-h)$$

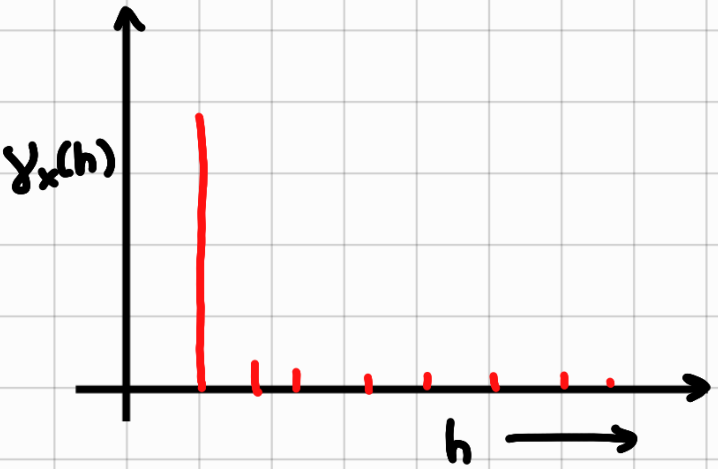
AR(1)

$$\gamma_X(h) = \gamma_X(-h) = \varphi^{|h|}(\sigma^2/(1-\varphi^2))$$





MA(1)



Auto-correlation fn.

Notations of operators

1. Backshift operator (B)

$$B^h x_t \equiv x_{t-h}$$

2. Difference operator (∇)

$$\nabla x_t \equiv x_t - x_{t-1} \equiv (I-B)x_t$$

$$\boxed{\nabla = I - B}$$

$$\nabla^h x_t = (I-B)^h x_t = \sum_{k=0}^{h} \binom{h}{k} (-1)^{h-k} (B^{h-k} x_t)$$

3. Seasonal Operator

$$\nabla_s = (I - B^s)$$

$$\nabla_s x_t = x_t - x_{t-s}$$

Consider $f(x) = a_0 + a_1 x + a_2 x^2 = 1 + 2x + 3x^2$

$x = 1, f(1) = 6$

$x = 2, f(2) = 17$

$x = 3, f(3) = 34$

$x = 4, f(4) = 57$

	∇	∇^2	∇^3
6			
17	11		
34	17	6	0
57	23	6	0

Linear process

A time series $\{X_t\}$ is a linear process if it can be represented as :

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} \quad \text{where } t \in \mathbb{Z},$$

$$= \mu + \sum_{j=-\infty}^{\infty} \psi_j B^j Z_t \qquad Z_t \sim WN(0, \sigma^2),$$

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

$$= \mu + \left(\sum_{j=-\infty}^{\infty} \psi_j B^j\right) Z_t$$

$$= \mu + \Psi(B) Z_t$$

$$\Psi(B) = \sum_{-\infty}^{\infty} \psi_j B^j$$

$$X_t = \{\Psi(B) Z_t\} + \mu$$

Note. $E(X)$ exists if $E|X| < \infty$, i.e.,
$\sum_i |x_i| f(x_i) < \infty$ (or, $\int |x| f(x) dx < \infty$)

Note. Absolutely summable series is always summable

Ex 1. $X_t \sim WN(0, \sigma^2)$

$$\psi_j = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{o.w.} \end{cases}$$

Ex 2. $X_t \sim MA(1)$

$$\psi_j = \begin{cases} 1 & \text{if } j = 0 \\ \theta & \text{if } j = 1 \\ 0 & \text{o.w.} \end{cases}$$

Ex 3. $X_t \sim AR(1)$

$$X_t = \phi X_{t-1} + Z_t$$

$$\psi_j = \begin{cases} \phi^j & \text{if } j \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

MA(q) process $q \in \mathbb{N}$

MA(q) process is a linear process where

$$\begin{cases} \psi_0 = 1 \\ \psi_j = \begin{cases} \theta_j & 1 \leq j \leq q \\ 0 & \text{o.w.} \end{cases} \end{cases}$$

$$\begin{aligned} X_t &= Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j} \\ &= Z_t + \left(\sum_{j=1}^{N} \theta_j B^j\right) Z_t \\ &= \left(I + \sum_{j=1}^{q} \theta_j B^j\right) Z_t \equiv \Theta_q(B) Z_t. \end{aligned}$$

$Z_t \sim WN(0, \sigma^2)$

$$\text{cov}(X_t, X_s) = \begin{cases} 0 & \text{lag } (|t-s| > q) \\ \text{can be non-zero} & \text{o.w.} \end{cases}$$

If $Z_t \overset{iid}{\sim}$ then, (X_t, X_j) are independent if $|t-s| > q$.

Autoregressive Process of order (p) $\equiv$ AR(p) $p \in \mathbb{N}$

A process is said to be an AR(p) process if it can be represented as

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + Z_t$$

where $Z_t \sim WN(0, \sigma^2)$.

$\Rightarrow (x_t) - (\sum_{j=1}^{p} \varphi_j B^j) x_t = Z_t.$

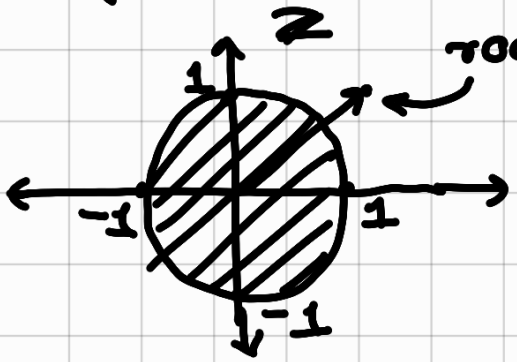
$\Rightarrow (I - \sum_{j=1}^{p} \varphi_j B^j) x_t = Z_t.$

Let $\Phi_p(u) = (1 - \sum_{j=1}^{p} \varphi_j u^j)$, then

$\Phi_p(B) x_t = Z_t$

$\Rightarrow x_t = \left(\frac{1}{\Phi_p(B)}\right) Z_t. \longrightarrow (*)$

$\frac{1}{\Phi_p(B)} = \frac{1}{1 - \sum_{j=1}^{p} \varphi_j B^j}$. This should satisfy unit root condition for $(*)$ to be feasible, that is, for eqn. $1 - gu \equiv 0$, root $u = 1/g > 1$.
(for our case, $gu \equiv \sum_{j=1}^{p} \varphi_j B^j$)
Also, $|g| < 1$, $g \neq 0$.



AR(1) process can be thought as MA(∞) process.

$AR(1) = \lim_{q \to \infty} MA(q)$.

We can represent AR(1) process as MA(∞) process

$x_t = \varphi x_{t-1} + Z_t \quad |\varphi| < 1 \quad \varphi \neq 0$

$\rightarrow (I - \varphi B) x_t = Z_t$

$\rightarrow x_t = \left(\frac{1}{1 - \varphi B}\right) Z_t$

$\Rightarrow x_t = (\sum_{j=0}^{\infty} (\varphi B)^j) Z_t$

$\Rightarrow x_t = \sum_{j=0}^{\infty} \varphi^j Z_{t-j}$

$$x_t = \left(\sum_{j=0}^{k} \phi^j z_{t-j}\right) + \phi^{k+1} x_{t-(k+1)}$$

$$E\left(x_t - \sum_{j=0}^{k} \phi^j z_{t-j}\right)^2 = E\left\{\phi^{k+1} x_{t-(k+1)}\right\}^2$$

$$\lim_{k\to\infty} E\left(x_t - \sum_{j=0}^{k} \phi^j z_{t-j}\right) = \lim_{k\to\infty} E\left\{\phi^{k+1} x_{t-(k+1)}\right\}^2$$

$$= \lim_{k\to\infty} \left(\phi^{2k+2}\right) E\left(x_{t-(k+1)}^2\right)$$

$\phi^{2k+2} \to 0$; $E(x_{t-(k+1)}^2)$ bounded

$$\left(\lim_{n\to\infty} E|x_n - y|^2 \to 0 \Rightarrow x_n \xrightarrow{d} y.\right)$$

Hence, $x_t \overset{d}{=} \sum_{j=0}^{\infty} \phi^j z_{t-j}$

$$\lim_{n\uparrow\infty} \left(x_t - \sum_{j=0}^{k} \phi^j z_{t-j}\right)^2 = \lim_{k\uparrow\infty} \left\{\phi^{2k} E(x_{t-k}^2)\right\}$$

x_t: AR(1) ; $\sum_{j=0}^{k} \phi^j z_{t-j}$: MA(k) ; $\phi^{2k} \to 0$; $E(x_{t-k}^2)$: bounded

$Y_k = \sum_{j=0}^{k} \phi^j z_{t-j}$ such that $E|x_t - Y_k(t)|^2 \to 0$

$\Rightarrow Y_k(t) \xrightarrow{d} x_t$ as $k \uparrow \infty$.

Note. $x_1, x_2, \ldots, x_n$ iid $E(x) = \mu$ $V(x) = \sigma^2$

$T_k = \sqrt{k}(\bar{x} - \mu)/\sigma$

$Z \sim N(0,1)$

Then $E(T_k - Z)^2 \to 0 \Rightarrow T_k \xrightarrow{d} Z$

Covariance of a Linear Process

Let $\{x_t\}$ be a weakly stationary process with $E(x_t) = 0$ and $\gamma_x(h) = \text{cov}(x_t, x_{t+h})$ exists.

If $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then defining $Y_t = \sum_{j=-\infty}^{\infty} \psi_j x_{t-j}$, we have

$\to E(Y_t) = 0$

$\to \gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_x(h+k-j)$

Proof.

$$E|Y_t| = E\left|\sum_{j=-\infty}^{\infty}\psi_j X_{t-j}\right| \le \sum_{j=-\infty}^{\infty}|\psi_j| E|X_{t-j}|$$
$$\le M_1 \sum_{j=-\infty}^{\infty}|\psi_j| < \infty$$

where $E|X_t| < M_1$.

Homework. Let Z be a random variable. Prove that

$$E(|Z|^r) \le \infty \implies E(|Z|^s) < \infty, \quad r \ge s$$

Hence, $E|Y_t|$ exists.

$$E(Y_t) = E\left(\sum_{j=-\infty}^{\infty}\psi_j X_{t-j}\right) = \sum_{j=-\infty}^{\infty}\psi_j E(X_{t-j}) = 0.$$

$$E|Y_{t+h}Y_t| = E\left|\sum_j \psi_j X_{t+h-j} \sum_k \psi_k X_{t-k}\right|$$

$$= E\left|\sum_j\sum_k \psi_j\psi_k X_{t+h-j}X_{t-k}\right|$$
$$\le \sum_j\sum_k |\psi_j||\psi_k| E(|X_{t+h-j}||X_{t-k}|)$$
$$\le \sum_j\sum_k |\psi_j||\psi_k| M_2$$

where $\max_{t,h\in\mathbb{Z}} E|X_{t+h-j}X_{t-k}| < M_2$.

So,

$$E|Y_{t+h}Y_t| \le \sum_j\sum_k |\psi_j||\psi_k| M_2 = M_2\left(\sum_j|\psi_j|\sum_k|\psi_k|\right) < \infty.$$

$$\gamma_Y(h) = \text{cov}(Y_{t+h}Y_t)$$
$$= \text{cov}\left(\sum_j \psi_j X_{t+h-j} \sum_k \psi_k X_{t-k}\right)$$
$$= E\left(\sum_j\sum_k \psi_j\psi_k X_{t+h-j}X_{t-k}\right)$$
$$= \sum_j\sum_k \psi_j\psi_k E(X_{t+h-j}X_{t-k})$$
$$= \sum_j\sum_k \psi_j\psi_k \text{cov}(X_{t+h-j}, X_{t-k})$$
$$= \sum_j\sum_j \psi_j\psi_k\gamma_X(h-j)$$

Homework. $X_t \sim WN(0,\sigma^2)$ and Y_t is a linear process.

Then, $\gamma_Y(h) = \sigma^2\sum_{j=-\infty}^{\infty}\psi_j\psi_{j+h}$.

Homework. $Y_t \sim AR(1)$. $\gamma_Y(h) = \phi^h \sigma^2/(1-\phi^2)$.

ARMA process

ARMA (p,q)

A time series $\{x_t\}$ is said to follow ARMA (p,q) process if we have the representation

$$(x_t - \Sigma_{j=1}^{p} \phi_j x_{t-j}) = z_t + \Sigma_{j=1}^{q} \theta_j z_{t-j}$$
$$\forall\ t \in \mathbb{Z},\ Z \sim WN.$$

$$\rightarrow (I - \Sigma_{j=1}^{p} \phi_j B^j) x_t = (I + \Sigma_{j=1}^{q} \theta_j B^j) z_t$$
$$\rightarrow \Phi_p(B) x_t = \Theta_q(B) z_t \text{ where}$$

$$\Phi_p(x) = 1 - \sum_{j=1}^{p} \phi_j x^j \text{ and } \Theta_q(x) = 1 + \sum_{j=1}^{q} \theta_j x^j$$

$\Phi_p(x)$ and $\Theta_q(x)$ should not have any common root.

For example, ARMA(1,1)

$$x_t - \phi x_{t-1} = z_t + \theta z_{t-1}$$
$$\Rightarrow x_t = \left(\frac{I+\theta B}{I-\phi B}\right) z_t$$

$$\begin{cases} |\phi| < 1,\ \phi \neq 0 \\ \theta + \phi \neq 0 \end{cases}$$

$$x_t = \{(1+\theta B)(\Sigma_{j=0}^{\infty} (\phi B)^j)\} z_t = \Psi(B) z_t$$
$$= (\psi_0 + \Sigma_{j=1}^{\infty} \psi_j B^j) z_t$$

$$\psi_0 = 1$$
$$\psi_j = (\theta+\phi)\phi^{j-1} \quad \forall\ j \geq 1$$

Linear process representation

$$x_t = z_t + (\theta+\phi) \Sigma_{j=1}^{\infty} \phi^{j-1} z_{t-j}$$
$$\gamma_x(0) = \sigma^2 \Sigma_{j=0}^{\infty} \psi_j^2 = \sigma^2 [1 + (\theta+\phi)^2 \Sigma_{j=1}^{\infty} \phi^{j-2}]$$
$$= \sigma^2 \left[1 + \frac{(\theta+\phi)^2}{1-\phi^2}\right]$$

$$\gamma_x(1) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1} = \sigma^2 \left[(\theta+\phi)^2 + \frac{(\theta+\phi)^2 \phi}{1-\phi^2} \right]$$

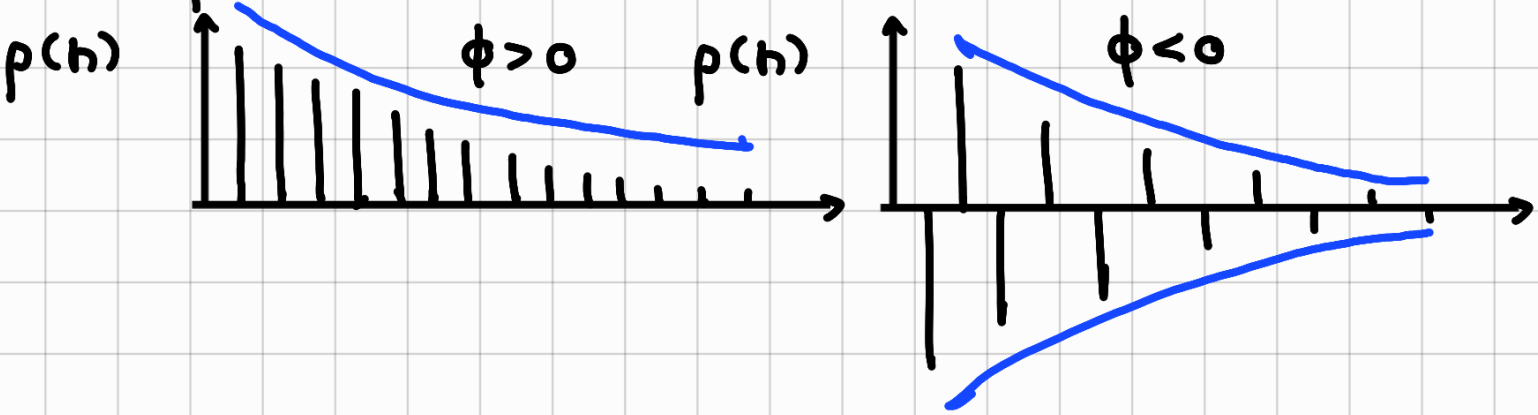
$\gamma_x(h) = \phi^{h-1} \gamma_x(1)$. Homework.

Autocorrelation coefficient of weakly stationary process.

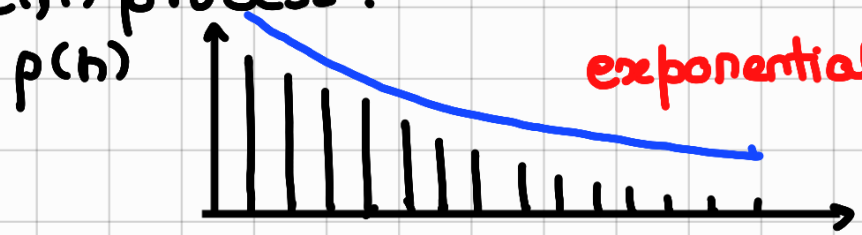
$$\rho(h) = \frac{\gamma_x(h)}{\gamma_x(0)}$$

MA(q) process:

AR(1) process:



ARMA(1,1) process:



ACF can be used to distinguish b/w MA and (AR/ARMA)
But it cannot distinguish b/w AR & ARMA.
To do so, we need partial auto correlation coefficient. (PACF)

Partial correlation Coefficient

given, $\begin{pmatrix} Y \\ Z \\ \underset{\sim}{X} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_Z \\ \underset{\sim}{\mu_X} \end{pmatrix}, \left(\begin{array}{c|c} \Sigma_{YZ} & \Sigma_{YZ\underset{\sim}{X}}^T \\ \hline \Sigma_{YZ\underset{\sim}{X}} & \Sigma_{\underset{\sim}{X}} \end{array}\right) \right)$ $\quad Y \in \mathbb{R}^1,\ Z \in \mathbb{R}^1,\ \underset{\sim}{X} \in \mathbb{R}^p$

$E(Y) = \mu_Y \quad E(Z) = \mu_Z \quad E(\underset{\sim}{X}) = \underset{\sim}{\mu_X}$

$D\begin{pmatrix} Y \\ Z \end{pmatrix} = \Sigma_{YZ} = \begin{pmatrix} \sigma_{YY} & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_{ZZ} \end{pmatrix}_{2\times 2} \quad D(\underset{\sim}{X}) = \Sigma_{\underset{\sim}{X}} = \left((\sigma_{X_iX_j})\right)_{p\times p}$

$$\sigma_{Y\underset{\sim}{X}} = \operatorname{cov}(Y, \underset{\sim}{X}) = \begin{pmatrix} \sigma_{YX_1} \\ \sigma_{YX_2} \\ \vdots \\ \sigma_{YX_p} \end{pmatrix}$$

$$\sigma_{Z\underset{\sim}{X}} = \operatorname{cov}(Z, \underset{\sim}{X}) = \begin{pmatrix} \sigma_{ZX_1} \\ \sigma_{ZX_2} \\ \vdots \\ \sigma_{ZX_p} \end{pmatrix}$$

$$\Sigma_{YZ\underset{\sim}{X}} = \left(\underset{\sim}{\sigma}_{Y\underset{\sim}{X}}, \underset{\sim}{\sigma}_{Z\underset{\sim}{X}} \right)_{p\times 2} \qquad \Sigma_{YZ\underset{\sim}{X}}^T = \begin{pmatrix} \underset{\sim}{\sigma}_{Y\underset{\sim}{X}}^T \\ \underset{\sim}{\sigma}_{Z\underset{\sim}{X}}^T \end{pmatrix}_{2\times p}$$

Partial correlation coefficient is the correlation b/w Y and Z after removing the effect of $\underset{\sim}{X}$ and is denoted as $\rho_{YZ.\underset{\sim}{X}}$

$$\rho_{YZ.\underset{\sim}{X}} = \frac{\operatorname{cov}\left(Y - E(Y|X),\ Z - E(Z|X)\right)}{\sqrt{\operatorname{Var}(Y - E(Y|X))\operatorname{Var}(Z - E(Z|X))}}$$

$$\Rightarrow \rho_{YZ.\underset{\sim}{X}} = \operatorname{cov}(e_{Y|X}, e_{Z|X}) \Big/ \sqrt{\operatorname{Var}(e_{Y|X})\operatorname{Var}(e_{Z|X})}$$

$$= \sigma_{YZ.X} / \sqrt{\sigma_{YY.X}\,\sigma_{ZZ.X}}$$

When we use it for time series we call it PACF.

conditional distribution of $\left(\begin{smallmatrix} Y \\ Z \end{smallmatrix}\right) \Big| X = \underset{\sim}{x}$.

This also follow normal distribution,

$$N\left(\underset{2\times 1}{\begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}} + \underset{2\times p}{\Sigma_{YZX}^{T}} \underset{p\times p}{\Sigma_X^{-1}} \underset{p\times 1}{(\underset{\sim}{x} - \underset{\sim}{\mu_X})}, \underset{2\times 2}{\Sigma_{YZ}} - \underset{2\times p}{\Sigma_{YZX}^{T}} \underset{p\times p}{\Sigma_X^{-1}} \underset{p\times 2}{\Sigma_{YZX}}\right)$$

Case : For Bivariate Normal,

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim BVN\left(\begin{pmatrix} \mu_U \\ \mu_V \end{pmatrix}, \begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix}\right)$$

$$V|U=x \sim N\left(\mu_V + \rho\frac{\sigma_V}{\sigma_U}(x-\mu_U), (1-\rho^2)\sigma_U^2\right)$$

$$E(V|U=u) = \mu_V + \rho\frac{\sigma_U\sigma_V}{\sigma_U^2}(x-\mu_U)$$

$$= \mu_V + \frac{cov(U,V)}{V(U)}(x-\mu_U)$$

$$= \mu_V + cov(U,V)(V(U))^{-1}(x-\mu_U)$$

$$V(V|U=u) = \sigma_V^2 - cov(U,V)(V(U))^{-1}cov(U,V)$$
$$= \sigma_V^2 - \rho\sigma_U\sigma_V(\sigma_U^2)^{-1}\rho\sigma_U\sigma_V$$
$$= \sigma_V^2 - \rho^2\sigma_V^2$$
$$= (1-\rho^2)\sigma_V^2$$

Mean and the variance-covariance structure will remain same, if we do best linear prediction of $\left(\begin{smallmatrix} Y \\ Z \end{smallmatrix}\right)$ based on

$\underset{\sim}{x}$, even when $\begin{pmatrix} y \\ \underset{\sim}{x} \end{pmatrix}$ are not jointly normally distributed

but 2nd order moments exists.