Applied Statistics Notes

Applied statistics (Politecnico di Milano)

Generalities of statistical learnings

★ Dataframes

	X_1	X_2	X_3 ·····	X_p	Y
Unit 1	x_{11}	x_{12}	x_{13} ····	x_{1p}	y_1
Unit 2	x_{21}	x_{22}	x_{23} ···	x_{2p}	y_2
Unit 3	x_{31}	x_{32}	x_{33} ····	x_{3p}	y_3
⋮	⋮				⋮
Unit n	x_{n1}	x_{n2}	x_{n3} ····	x_{np}	y_n

Here $\underline{X} = (X_1, X_2 \cdots X_p) \in R^p$

$Y \in R$

Goal: To explain the variability of y in terms of X

Basic model: $\boxed{Y = f(\underline{X}) + \varepsilon}$

Best solution: $f = \min E[|y - f(x)|^2]$

(min. expected square errors)

Answer to the optimal model

If g is a function such that $g: R^p \to R$

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

Best f ⇒ $f(\underline{X}) = E[Y|\underline{X}]$

(conditional mean of y given X; also called regression function)

To find the best f, we try to approximate $E[Y|\underline{X}]$ by means of a member of a family.

$\{f_\theta \mid f_\theta : R^p \to R$ known when $\theta \in \underline{\theta} \leq R^k$ is known$\}$

θ is a multi-dimensional point in space

Problem

To find $\hat{\theta}$ such that

$\hat{\theta} = \min E[(Y - f_\theta(X))^2]$

& once it is found then

⇒ $f_{\hat{\theta}}$ should be taken as approx of $E[Y|\underline{X}]$

★ Nomenclature in $y = f(x) + \varepsilon$

- $X \in X_1 \cdots X_p$ → features / predictors / regressors / inputs
- $Y \in R$ → response, target, output
- ε → errors, residuals
- using data to find best f → model fitting
- $f_{\hat{\theta}}$ → fitted model with best fit & parameters $\hat{\theta}$

★ Linear Models

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}_{E[Y|\underline{X}]} + \varepsilon$$

(residual error term)

$\forall\ X = (X_1 \cdots X_p) \in R^p$

★ Classical Linear Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad \star$$

$i \in (1, N)$

here ε_i follows a normal distribution

$\varepsilon \in N(0, \sigma^2)$ $\forall i = 1 \cdots N$ (same variability)

The above model can be mentioned in the form of a design matrix of shape $n \times p$ such that

$$\bar{Y} = X\bar{\beta} + \bar{\varepsilon}$$

where $\bar{Y} = (y_1, y_2 \cdots y_n)'$

$\bar{\beta} = (\beta_1, \beta_2 \cdots \beta_n)'$

$\bar{\varepsilon} = (\varepsilon_1, \varepsilon_2 \cdots \varepsilon_n)'$

Data is in the format ($n \times p$ design matrix)

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Remark

Here $\beta_0 = \beta_1$ in this notation

recover ★ by setting $x_{i1} = 1$

Estimating the parameters

- **Method 1: Ordinary Least squares**

Find β that makes the values closest to $\bar{Y} = X\bar{\beta} + \bar{\varepsilon}$

$$\hat{\beta} = \min_{\bar{\beta} \to R^p} \|\bar{Y} - X\bar{\beta}\|^2 = \min_{\bar{\beta}} \sum_{i=1}^{n} (y_i - X_i'\bar{\beta})^2$$

(X_i': transpose; represents i^{th} row of β matrix)

where $X_i'\beta = \beta_1 X_{i1} + \cdots \beta_p X_{ip}$

Geometry of OLS

① $X\underline{\beta}$

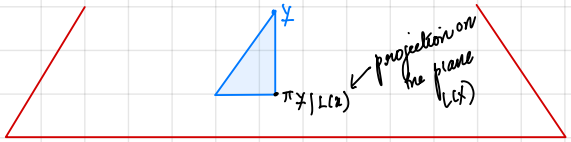
$X = [X^1 \cdots X^p]$

$X^{(j)} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$ where $j = 1 \cdots p$

$\therefore X\underline{\beta} = \beta_1 X^{(1)} + \beta_2 X^{(2)} + \cdots \beta_p X^{(p)}$

If $L(X)$ is the linear space generated by the design matrix

$L(X) = \text{span}(X^{(1)}, \cdots X^{(p)})$



$\pi_{Y|L(X)}$ is an element of the space $L(X)$

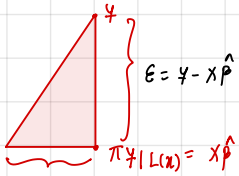
since $\pi_{Y|L(X)} \to L(X)$

$$\pi_{Y|L(X)} = \hat{\beta}\underline{X}^1 + \cdots \hat{\beta}\underline{X}^p = X\hat{\beta}$$

so

$\hat{\beta}$ is a vector of coefficient of $\pi_{Y|L(X)}$

In a space, Y can be described as the sum of 2 vectors



$$\therefore Y = X\hat{\beta} + \hat{\varepsilon}$$
$$= \hat{Y} + \hat{\varepsilon}$$

fitted values — residuals

Assume that $\underline{X}^{(1)} \cdots \underline{X}^{(p)}$ are linearly independent

$\dim(L(X)) = p$ (dimension space of the model)

$\dim(L^{\perp}(X)) = n-p$ (dimension of space of the residuals)

Prop: If X is full rank (means that the rank is p) $n \times p$ such that $(n \times p)$

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

if collinear $X'X$ will be $= 0$

$$\hat{Y} = X(X'X)^{-1}X'Y = HY$$

H: hat matrix

such that $H = X(X'X)^{-1}X'$

$$\underline{\hat{\varepsilon}} = Y - X\hat{\beta} = Y - \hat{Y} = Y - HY = (I-H)Y$$

while estimating the variability for ε

the variability $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$

because $\hat{\varepsilon}$ lies in a plane $(n \times p)$

Properties of $\hat{\beta}_{OLS}$ & $\hat{\sigma}^2_{OLS}$

① They are both unbiased such that

$$E[\hat{\beta}] = \beta$$
$$E[\hat{\sigma}^2] = \sigma^2$$

Maximum likelihood Estimator (MLE) for β & σ^2

$$Y = X\beta + \underline{\varepsilon}$$

$$\varepsilon \sim N(0, \sigma_1 I)$$

Under this assumption

$$Y \sim N(X\beta, \sigma_1 I)$$ (gaussian)

If we look at the component $Y_i \sim N_i(\underline{X}_i'\beta, \sigma^2)$ $i = 1 \cdots n$ (independent)

Then the density of the gaussian distribution

$$f_{\beta,\sigma^2}(y_1 \cdots y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - x_i'\beta)^2}{\sigma^2}}$$

Likelihood

$$L(\beta, \sigma^2 | y_1 \cdots y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - x_i'\beta)^2}{\sigma^2}}$$

MLE for $\hat{\beta}$ & σ^2

$$(\hat{\beta}, \hat{\sigma_1^2}) = \max_{\beta,\sigma^2} L(\beta, \sigma^2 | Y)$$

$$= \max_{\beta,\sigma^2} \log L = \max_{\beta,\sigma} \ell(\beta, \sigma^2 | Y)$$

$$= \max\left[-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \underline{x}_i\beta)^2\right]$$

Proposition

If X is full rank

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'Y \Rightarrow E[\hat{\beta}_{MLE}] = \beta$$

$$\hat{\sigma^2}_{MLE} = \frac{1}{n}\sum\hat{\varepsilon_i^2} \Rightarrow E[\hat{\sigma}^2_{MLE}] = \frac{n-p}{n}\beta$$

Restricted MLE for σ^2

Notice: $Y \sim N_p(X\beta, \sigma^2 I)$

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N_p(\beta, \sigma^2(X'X)^{-1})$$

Recall: If we have a variable W

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

& we take a linear transformation A $(q \times n)$

$$AW \sim N_q(A\mu, A\Sigma A')$$

(if you linearly transform gaussian, you get gaussian)

if $\hat{\beta} = (X'X)^{-1}X'Y \sim N_p(\beta, \sigma^2(X'X)^{-1})$

then

$$\hat{\varepsilon} = (I-H)Y \sim N_n(\underline{0}, \sigma^2(I-H))$$

$\therefore$ RMLE mean

$$\hat{\sigma^2} = \max_{\sigma^2} L(\sigma^2 | \underline{\hat{\varepsilon}})$$

$$\Rightarrow \boxed{\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}$$

Inferences

- $\mathrm{Var}(\hat{\beta}) = \mathrm{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$

  $\mathrm{Var}(\hat{\beta}) = \hat{\sigma}^2_{RMLE}(X'X)^{-1}$

★ **Methods to run test on linear combination of models**

If L is a $q \times p$ matrix

Inference: $L\beta$

$$F = \frac{(L\hat{\beta} - L\beta)'(L\,\mathrm{var}(\hat{\beta})L')^{-1}(L\hat{\beta} - L\beta)}{q}$$

This follows a F distribution $\sim F(q, n-p)$

Testing : $H_0 : L\beta = c_0$
vs
$H_1 : L\beta \neq c_0$

Example

$H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$

If $L = [0 \; \dots \; 1 \; \dots \; 0]$

$q = 1$

then

$$F = \frac{(\hat{\beta_i} - \beta_i)^2}{Var(\hat{\beta_i})}$$

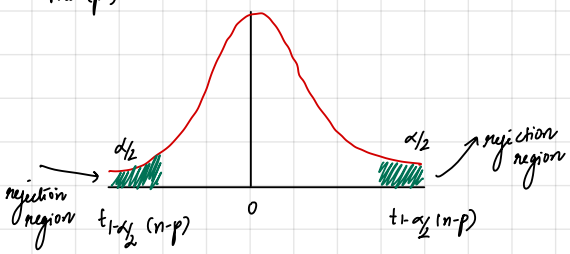
$H_0 : \beta_i = c_0$ vs $H_1 : \beta_i \neq c_0$

F should be rejected if $\frac{(\hat{\beta_i} - c_0)^2}{var(\hat{\beta_i})} > F_{1-\alpha}(q, n-p)$

which is equivalent to

$$\frac{|\hat{\beta_i} - c_0|}{\sqrt{Var(\hat{\beta_i})}} > \sqrt{F_{1-\alpha}(1, n-p)}$$ since $q=1$

$$\Rightarrow \frac{\hat{\beta_i} - c_0}{\sqrt{Var(\hat{\beta_i})}} \sim t(n-p) \text{ (we know)}$$



$$\frac{\hat{\beta_i} - \beta_i}{\sqrt{Var(\hat{\beta_i})}} \sim t(n-p)$$

$$\Rightarrow \left[\hat{\beta_i} \pm t_{1-\frac{\alpha}{2}}(n-p)\sqrt{Var(\hat{\beta_i})}\right] = C_{1-\alpha}(\beta_i)$$

★ Model selection

If there are 2 models for the same data

M_0, M_a

Example:

M_0 : $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

M_A : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$ } Nested models

Here we need to identify if we can eliminate the coefficients β_2 & β_3 such that M_A becomes M_0

Null hypothesis $H_0 : \beta_2 = \beta_3 = 0$

Alternative hypo $H_1 : \beta_2 = \beta_3 \neq 0$

- If the models are nested, we always use the F-test
- If the models are not nested ?

let

$L(\beta, \sigma^2 | data)$ → likelihood

For example: If $\varepsilon_1, \dots \varepsilon_n$ (iid) $\sim N(0, \sigma^2)$

then

$$L(\beta, \sigma^2 | data) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{-\frac{1}{2}\left(\frac{y_i - x_i\beta)^2}{\sigma^2}\right.}$$

then log (likelihood) will be

$$\log L(\hat{\beta}, \hat{\sigma^2} | data)$$

Assume that

$\hat{\beta_A}, \hat{\sigma_A^2}$ are estimates of β, σ^2 under model M_A

$\hat{\beta_0}, \hat{\sigma_0^2}$ are estimates of β, σ^2 for model M_0

We will move from Model M_0 to M_A

if

$$l(\hat{\beta_A}, \hat{\sigma_A^2}) > l(\hat{\beta_0}, \hat{\sigma_0^2})$$

ie. if $l(\hat{\beta_A}, \hat{\sigma_A^2}) - l(\hat{\beta_0}, \hat{\sigma_0^2})$ is large

ie. if $\frac{L(\hat{\beta_A}, \hat{\sigma_A^2})}{L(\hat{\beta_0}, \hat{\sigma_0^2})}$ is large

★ Akaike Information Criterion (AIC)

It is an index used for model selection

$$AIC(\hat{\beta}, \hat{\sigma^2}) = 2p - 2l(\hat{\beta}, \hat{\sigma^2})$$

We should choose the model such that

we choose the model M_A if

$$AIC(\hat{\beta_A}, \hat{\sigma_A^2}) < AIC(\hat{\beta_0}, \hat{\sigma_0^2})$$

★ Bayesian Information Criterion (BIC)

$$BIC(\hat{\beta}, \hat{\sigma^2}) = (\log n) \cdot p - 2l(\beta, \sigma^2)$$

★ Diagnostics

- Residual Analysis
- Influential cases (follows the same concept as Cook's distance)

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})' (var(\hat{\beta}))^{-1} (\hat{\beta} - \hat{\beta}_{(-i)})}{p}$$

$i = 1, \dots n$

$\hat{\beta}_{(-i)}$ after removing 1 item from dataset

Here D_i ∈ Cook's distance

Influential cases are those for which D_i is big

D_i can be re-written as

$$D_i = \frac{1}{\hat{\sigma^2}_{REML}} \left(\frac{\hat{\varepsilon_i}}{\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}}$$

standardized residual

$H = X(X'X)^{-1}X'$ Here h_{ii} is the measure of leverage

Here h_{ii} is the i^{th} element on the diagonal of H

In this $Var(\hat{\varepsilon_i}) \propto 1 - h_{ii}$

Likelihood displacement

$LD_i = 2\left[\ell(\hat{\beta}, \hat{\sigma^2}) - \ell(\hat{\beta}_{(-i)}, \hat{\sigma}^2_{(-i)})\right]$

difference in likelihood before and after removing an influential element

i is influential if LD_i is very large

☆ Linear Models with heterogeneous variances

Model for the phenomenon

$Y = \underbrace{X\beta}_{E[Y|X]} + \varepsilon \rightsquigarrow \cup N(0, \sigma^2)$

Model for observations

$y_i = x_i'\beta + \varepsilon_i \quad i = 1 \cdots n$

$\varepsilon_i \cup N(0, \sigma_i^2)$

parameters: $\beta, \sigma_1^2 \cdots \sigma_n^2$

$n+p$ parameters

Rephrasing the model into vector form

- $Y \in R^n$
- X $n \times p$ design matrix
- $\beta \in R^p$
- $\underline{\varepsilon} \in R^n \quad \varepsilon \cup N_n(0, \sigma^2 R)$
- $R = \Lambda\Lambda'$ where $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$

(Λ: Lambda (diagonal matrix))

ie $\sigma_i^2 = \sigma^2 \lambda_i^2$ (λ_i^2: scaling parameter)

- $\lambda_i = \lambda_i(\underline{\delta}, \mu_i, \underline{v}_i)$ → variance function
- $\underline{\delta} \in R^h$ vector of parameters common to all cases

$\mu_i = E[Y_i | \underline{x}_i] \quad i = 1 \cdots n$

$= \underline{x}_i \beta$

$\underline{v}_i \in R^k$ vector of known covariates

Classification of variance functions

Variance functions can be classified into 4 groups (Pinheiro & Bates 2000 book)

Group 1: known variance function

$\lambda_i = \lambda_i(\underline{v}_i)$ Variance function depends only on known covariates

- Example: Recall

$y_i = x_i'\beta + \varepsilon_i$

$\text{var}(Y_i) = \text{var}(\varepsilon_i) = \sigma^2 \lambda_i$

Suppose that y_i is a mean

$y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}$ if z_{ij} iid with $\text{var}(z_{ij}) = \sigma^2$

$\text{var}(y_i) = \sigma^2 / n_i$

$\lambda_i = 1/n_i$

- Example:

$y_i = \sum_{j=1}^{n_i} z_{ij}$

z_{ij} iid with $\text{var}(z_{ij}) = \sigma^2$

in this case

$\text{var}(y_i) = n_i \sigma^2$

$\therefore \lambda_i = n_i$

In group 1

$\lambda_i = \lambda_i(\underline{v}_i) = \text{var Fixed}(y_i)$

considering $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$

$Y = X\beta + \underline{\varepsilon}$

$\varepsilon \cup N(0, \sigma^2 R)$

$R = \Lambda\Lambda'$

Consider $\Lambda^{-1} = \begin{bmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_n \end{bmatrix}$

$\underbrace{\Lambda^{-1}Y}_{\| \; \tilde{Y}} = \underbrace{\Lambda^{-1}X\beta}_{\| \; \tilde{X}} + \underbrace{\Lambda^{-1}\varepsilon}_{\| \; \tilde{\varepsilon}}$

Now

$\tilde{Y} = \tilde{X}\beta + \tilde{\underline{\varepsilon}}$

where $\tilde{\underline{\varepsilon}} \cup N_n(0, \sigma^2 \underbrace{\Lambda^{-1} R \Lambda^{-1}}_{\Lambda^{-1}\Lambda^{-1} \text{ can be written as } \Lambda^{-1}\Lambda\Lambda^{-1}\Lambda = I})$

$\therefore \tilde{\underline{\varepsilon}} \cup N_n(0, \sigma^2 I)$

Using OLS or ML

$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (X'\Lambda^{-1}\Lambda^{-1}X)^{-1}X'\Lambda^{-1}\Lambda^{-1}Y$

$= (X'R^{-1}X)^{-1}X'R^{-1}Y$

(covariance matrix enters the picture)

$\hat{\underline{\varepsilon}} = \tilde{Y} - \tilde{X}\hat{\beta} = \Lambda^{-1}(Y - X\hat{\beta})$

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{1}{n-p}(Y - X\hat{\beta})'\Lambda^{-1}\Lambda^{-1}(Y - X\hat{\beta})$$
$$= \frac{1}{n-p}(Y - X\hat{\beta})'R^{-1}(Y - X\hat{\beta})$$

★ Weighted Least squares (WLS)

$$\hat{\beta} = \min_{\beta} |\tilde{Y} - \tilde{X}\beta|^2$$
$$= \min_{\beta} (\tilde{Y} - \tilde{X}\beta)'(\tilde{Y} - \tilde{X}\beta)$$
$$= \min_{\beta} (Y - X\beta)'\Lambda^{-1}\Lambda^{-1}(Y - X\beta)$$

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{n} \frac{(Y_i - x_i'\beta)^2}{\lambda_i^2}$$

This is called the weighted least squares (WLS)

so

$$\hat{\beta} = \underbrace{(X'R^{-1}X)^{-1}X'R^{-1}}_{A}Y$$

$\therefore \hat{\beta} = AY$

so $Var(\hat{\beta}) = A \cdot Var(Y) A'$

$Var(Y) = \sigma^2 R$

$$= \sigma^2 (X'R^{-1}X)^{-1}$$

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X'R^{-1}X)^{-1})$$

Inference:

$H_0: L\beta = c_0$ vs $H_1: L\beta \neq c_0$

$L \to q \times p$

then

Test statistic

$$F = \frac{(L\hat{\beta} - c_0)'(L\,Var(\hat{\beta})\,L')^{-1}(L\hat{\beta} - c_0)}{q}$$

so if H_0 is true $\Rightarrow$ $F \sim F(q, n-p)$

& confidence interval for $\beta_i = \left[\hat{\beta}_i \pm t_{1-\frac{\alpha}{2}}(n-p)\sqrt{\widehat{Var}(\hat{\beta}_i)}\right]$

$$\hat{\beta} = (X'R^{-1}X)^{-1}X'R^{-1}Y$$

★ Group 2: $<\delta>$ group

① $\lambda_i^2 = \lambda_i(\underline{\delta}, v_i)$

$\delta \in$ unknown parameters

$v_i \in$ known covariates

Here the units are clustered in $S \geq 1$ groups

Example:

i) $\lambda_i(\delta, v_i) = |v_i|^{\delta_{s_i}}$ (δ_{s_i} is unknown) $\quad \delta_i \in \{1, 2, \cdots S\}$

$= Var\,power(\underline{\delta}; v_i, s_i)$

② $\lambda_i(\underline{\delta}; v_i) = \exp(v_i \delta_{s_i})$ (δ_{s_i} unknown) $\quad i = 1, \cdots n$

$= Var\,exp(\underline{\delta}; v_i, s_i)$

③ $\lambda_i^2(\underline{\delta}; v_i) = \delta_{1,s_i} + |v_i|^{\delta_{2,s_i}}$

$= Var$ const. power $(\underline{\delta}; v_i, s_i)$

④ $\lambda_i(\underline{\delta}; v_i) = \delta_{s_i}$

unknown: $\delta_{s_1}, \delta_{s_2} \cdots \delta_{s_n}$

usually $\delta_{s_1} = 1$ for identifiability

★ Estimation of the parameters

Maximum likelihood

$$L(\beta, \underline{\delta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2\lambda_i^2}} \exp\left[-\frac{1}{2}\frac{(y_i - x_i'\beta)^2}{\sigma^2\lambda_i}\right]$$

(OR)

$$l(\beta, \underline{\delta}, \sigma^2) = \log L(\beta, \underline{\delta}, \sigma^2)$$

$$(\hat{\beta}_{ML}, \hat{\delta}_{ML}, \hat{\sigma}^2_{ML}) = \max_{(\beta, \underline{\delta}, \sigma^2)} l(\beta, \underline{\delta}, \sigma^2)$$

★ optimizing profile likelihood

For any fixed $\delta \Rightarrow \hat{\beta} = \hat{\beta}(\delta)$ obtained by using the diamond formula

$$l^*(\underline{\delta}, \sigma^2) = l(\hat{\beta}(\underline{\delta}), \underline{\delta}, \sigma^2)$$

objective would be to maximize

$$\max_{\delta, \sigma^2} l^*(\underline{\delta}, \sigma^2) = (\hat{\underline{\delta}}, \hat{\sigma}^2)$$

There are chances that can get a biased estimator so instead of writing the likelihood for observations, we should find the likelihood for the residuals. ($\hat{\varepsilon}$)

This gives us the:

$\Rightarrow \hat{\delta}_{REML}, \hat{\sigma}^2_{REML}$

$\Rightarrow \hat{\beta} = \hat{\beta}(\hat{\underline{\delta}}_{REML})$

Inference

$$Var(\hat{\beta}) = \sigma^2(X'R^{-1}X)^{-1}$$

$\Lambda\Lambda^{-1}$ $\quad \Lambda = \begin{bmatrix} \lambda(\delta_1, v_1) & & \\ & \ddots & \\ & & \lambda_n(\delta_n, v_m) \end{bmatrix}$

conduct testing using F statistic estimating $Var(\hat{\beta})$ with $\hat{\sigma}^2(X'\hat{R}^{-1}X)^{-1}$

$H_0: L\beta = c_0$ vs $H_1: L\beta \neq c_0$

distribution of test statistic will be

$F \sim F(q, n-p)$

but if n is large (χ^2 chi-squared distribution)

$qF \sim \chi^2(q)$

$F \sim \frac{1}{q}\chi^2(q)$

Special Important case:

$\lambda_i = \lambda_i(\delta, \nu_i) = \delta_{s_i}$

Consider the LR test based :

$$2\left[l_1(\hat{\beta}_1, \hat{\delta}_1, \hat{\sigma_1^2}) - l_0(\hat{\beta_0}, \hat{\sigma_0^2}) \right] > \chi^2_{1-\alpha}(c-1)$$

if ↗ is satisfied then reject the H_0

Group 3: $<\mu>$

$\lambda_i = \lambda_1(\mu_i, \nu_i)$

$\mu_i = \underline{x_i}' \underline{\beta}$

Notable example:

$\lambda_i = \mu_i$

$Var(Y_i) = \sigma^2 \mu_i^2$

$CV(Y_i) = \frac{std(Y_i)}{\mu_i} = \sigma$

Examples :

① $\lambda_i = |\mu_i|^{\delta_{s_i}}$ (δ_{s_i}: this is known) $\quad s_i \in \{1, \dots S\}$ Group membership

$= Var\,power(\mu_i, \nu_i, s_i)$

② $\lambda_i = exp(\mu_i \delta_{s_i})$

③ $\lambda_i = \delta_{1,s_i} + |\mu_i|^{\delta_{2,s_i}}$ (those are known)

Group 4: $<\mu, \delta>$

$\lambda_i = \lambda_i(\underline{\delta}, \mu_i, \nu_i)$

example would be same as group 3 ($<\mu>$) but with unknown $\underline{\delta}$

Note: In the $<\mu>$ group and $<\mu, \delta>$ group $\underline{\beta}$ (unknown) enters the mean & the variance

1. Try to maximize $L(\underline{\beta}, \underline{\delta}, \sigma^2)$
2. Iterative procedure for estimating $\underline{\beta}, \underline{\delta}, \sigma^2$

Step 0 : Initial guess for $\hat{\underline{\beta}}$: $\hat{\underline{\beta}}_0$ (for instance use OLS)

For $K \geq 1$: repeat "until convergence"

Step 1 : Get $(\hat{\underline{\delta}}^{(K)}, \hat{\sigma}^{2(K)}) = \max_{\underline{\delta}, \sigma^2} l(\hat{\underline{\beta}}^{(K-1)}, \underline{\delta}, \sigma^2)$

ie. assume that $\mu_i = \underline{x_i}' \hat{\underline{\beta}}^{(K-1)}$

Step 2 : Estimate $\hat{\underline{\beta}}^{(K)}$ using $(\diamondsuit)$ and $\hat{R}$ by plugging $\hat{\underline{\delta}}^{(K)}$, $\hat{\sigma}^{2(K)}$

Stop when $\hat{\underline{\beta}}^{(K)}$ and $\hat{\underline{\beta}}^{(K-1)}$ are not very different w.r.t to some metric-

for instance :

$$\| \hat{\underline{\beta}}^{(K)} - \hat{\underline{\beta}}^{(K-1)} \| < \varepsilon \quad \text{(Euclidean distance)}$$

or

$$\sqrt{\left(\hat{\underline{\beta}}^{(K)} - \hat{\underline{\beta}}^{(K-1)}\right) \left(Var\left(\hat{\underline{\beta}}^{(K-1)}\right)\right) \left(\underline{\beta}^{(K)} - \underline{\beta}^{(K-1)}\right)} \leq \varepsilon$$

(Mahalanobis distance)

After this the question becomes what are the distributions of $\hat{\underline{\beta}}$ and F.

★ Linear Models with fixed effects & Correlated errors

$R \to Y = \underline{x}' \underline{\beta} + \varepsilon$

$R^p \to \underline{x}$, $\underline{\beta} \in R^p$, $\varepsilon \in N(0, \sigma^2)$ } Model for phenomenon

Units $i \in \{1, 2, \dots N\}$

each unit i is observed n_i times

$(n = n_1 + n_2 + \dots n_N)$

For a unit i :

$R^{n_i} \to Y_i = X_i \underline{\beta} + \underline{\varepsilon_i}$

$X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{bmatrix}$ $\quad \underline{x_{ij}} \in R^p$ $\quad (n_i \times p$ matrix)

$\varepsilon_i \sim N_{n_i}(0, \sigma^2 R_i)$

R_i need not be diagonal

Moreover: different units, say i and $i' \in \{1 \dots N\}$ are independent

$R \to Y_{ij}$ observation for unit $i \in \{1, \dots, N\}$ at instance $j \in \{1, \dots n_i\}$

- $E[Y_{ij} | x_{ij}] = \underline{x_{ij}}' \underline{\beta}$
- $Var[Y_{ij}] = Var[\varepsilon_{ij}] = \sigma^2 R_i(j,j)$
- $cov[Y_{ij}, Y_{ij'}] = \sigma^2 R_i(j,j')$ such that $j, j' \in \{1, \dots n_i\}$

$\Rightarrow cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$

Recall :

$$corr.(Z, W) = \frac{cov(Z, W)}{\sqrt{var(Z) \cdot var(W)}}$$

$$cov(Z, W) = \sqrt{var(Z)}\ corr(Z, W) \sqrt{var(W)}$$

$$\therefore \sigma^2 R_i(j,j') = cov(Y_{ij}, Y_{ij'}) = cov(\varepsilon_{ij}, \varepsilon_{ij'})$$
$$= \sqrt{var(Y_{ij})}\ corr(Y_{ij}, Y_{ij'}) \sqrt{var(Y_{ij'})}$$
$$= \lambda_{ij}\ c_i(j,j')\ \lambda_{ij'}$$

For • λ_{ij} → variance function already introduced $\langle \delta \rangle, \langle u \rangle, \langle \delta, u \rangle$

• C_{ij} → Correlation structures

Let $\Lambda_i = \begin{bmatrix} \lambda_{i_1} & & 0 \\ & \ddots & \\ 0 & & \lambda_{in_i} \end{bmatrix} = diag[\lambda_{i_1} \cdots \lambda_{in_i}]$ $(n_i \times n_i)$

$C_i = \begin{bmatrix} 1 & C_i(j,j') \\ C_i(j',j) & 1 \end{bmatrix}$ → $= C_i(j,j')$ because of symmetry

then

$R^{n_i} \to \varepsilon_i \sim N_{n_i}(\underline{0}, \sigma^2 R_i)$

$R_i = \Lambda_i C_i \Lambda_i'$

&

$var(y_{ij}) = \sigma^2 \lambda_{ij}^2$

✯ Correlation structures

Most of the times we consider structure of this form:

$Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = C_i(j,j') = h(d(t_{ij}, t_{ij'}), f$

where

d is a metric

t_{ij} is a index in (time, space)

identifying the instance where Y_{ij} has been observed

Assumptions on h:

- continuos (regularity)
- $h(0, f) = 1$

Examples

- $Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho_i \quad , |\rho_i| < 1$ — it is corr. comp. symm in R-studio
- $Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho_i^{|j-j'|}$ = Cor AR(1)

These are very commonly used and are called Serial autocorrelation function

- $Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = e^{-\frac{d(t_{ij}, t_{ij'})}{\rho_i}}, \rho_i > 0$

t_{ij} : location in space where Y_{ij} has been observed

- $Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = e^{-\frac{d^2(t_{ij}, t_{ij'})}{\rho_i^2}}$ = corr. Gauss
- $Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \left[1 - \frac{d(t_{ij}, t_{ij'})}{\rho_i}\right]$ (this is an indicator of $[d < \rho_i]$) = corr. Lin

$\mathbb{1}[d < \rho_i] = \begin{cases} 0 \text{ if } d > \rho_i \\ 1 \text{ if } d \le \rho_i \end{cases}$ (indicator symbol)

- $Corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \left[1 - \frac{3}{2}\frac{d(t_{ij}, t_{ij'})}{\rho_i} + \frac{1}{2}\frac{d^3(t_{ij}, t_{ij'})}{\rho_i^3}\right] \mathbb{1}[d < \rho_i]$

= Corr. spher.

✯ Excurses on stationarity and ARMA models

$\{X_t\}_{t \to T}$ stochastic process

$\{X_t\}_{t \to T}$ is said to be strongly stationary if: $\forall n$ and $(t_1, \cdots t_n)$ and $\tau > 0$

$F(X_{t_1} \cdots X_{t_n}) = F(X_{t_1+\tau}, \cdots X_{t_n+\tau})$

$\Rightarrow F(X_{t_1}) = F(X_{t_2})$ — distributions of X_{t_1} & X_{t_2} are the same

If we have the stationarity then

① $E[X_t] = \mu$

② $Cov(X_{t_1}, X_{t_2}) = h(|t_1 - t_2|)$

If we have the above 2 conditions satisfied then $\{X_t\}$ is said to be weakly stationary

Note:

$Var(X_t) = h(0)$

Consider this process $\{X_t\}$ weakly stationary

- $E[X_t] = 0 \qquad t = -2, -1, 0, 1, 2 \cdots$
- $X_t = \phi X_{t-1} + \varepsilon_t$ ε_t's are independent of X_t's, $\varepsilon_t \sim N(0, \sigma^2)$

$Var(X_t) = Var(\phi X_{t-1} + \varepsilon_t)$

$= \phi^2 Var(X_{t-1}) + \sigma^2$

$Var(X_t) = \phi^2 Var(X_t) + \sigma^2$

$\therefore Var(X_t) = \frac{\sigma^2}{1-\phi^2} \quad |\phi| < 1$

$Corr(X_{t+1}, X_t) = \phi \frac{\sigma^2}{1-\phi^2}$

$Cov(X_{t+1}, X_t) = Cov(\phi X_t + \varepsilon_t, X_t)$

$= \phi \frac{\sigma^2}{1-\phi^2} + 0$

Similarly $Corr(X_{t+2}, X_t) = \phi^2$

$Corr(X_{t+k}, X_t) = \phi^k \quad |\phi| < 1$

In an ARMA model

$X_t = \underbrace{\sum_{i=1}^{p} \phi_i X_{t-i}}_{AR(p)} + \underbrace{\sum_{i=1}^{q} \theta_i \varepsilon_{t-i}}_{MA(q)} + \varepsilon_t$

✯ Estimating β

$Y = X\beta + \varepsilon$

If R is known:

$R^{-1/2} Y = R^{-1/2} X \beta + R^{-1/2} \varepsilon$

such that $R^{-1/2}\varepsilon \sim N_n(0, \sigma^2 I)$

$\hat{\beta} = (X' R^{-1} X)^{-1} X' R^{-1} Y$

Take variance functions in $<\delta>$ group

$\mu = X\beta$

$\hat{\beta} = \hat{\beta}_1(\underline{\delta}, \rho)$

To estimate $\underline{\delta}$ and ρ, estimate the profile likelihood

$l^*lik(\sigma^2, \underline{\delta}, \rho) = llik(\hat{\beta}(\underline{\delta}, \rho), \sigma^2, \underline{\delta}, \rho)$

$(\hat{\sigma^2}, \hat{\underline{\delta}}, \hat{\rho}) = \max_{\sigma^2, \underline{\delta}, \rho} l^*lik(\hat{\beta}, (\underline{\delta}, \rho), \sigma^2, \underline{\delta}, \rho)$

$\Rightarrow \hat{\beta} = \hat{\beta}(\hat{\underline{\delta}}, \hat{\rho}) = (X'\hat{R}^{-1}X)^{-1}X'\hat{R}^{-1}Y$

- For unbiasedness of $\hat{\sigma^2}$, consider REML:

  $l^*lik(\sigma^2, \underline{\delta}, \rho)$ the llik for

  $\hat{\underline{\varepsilon}} = Y - X\hat{\beta}$

  $\Rightarrow \hat{\sigma^2} = \frac{1}{n-p}(Y - X\hat{\beta})'R^{-1}(Y - X\hat{\beta})$

Profile again

$llik(\underline{\delta}, \rho) = llik(\hat{\beta}(\underline{\delta}, \rho), \sigma^2, \underline{\delta}, \rho)$

$(\hat{\delta}, \hat{\rho}) = \max llik(\underline{\delta}, \rho)$

$\therefore \hat{\beta} = \hat{\beta}(\hat{\underline{\delta}}, \hat{\rho})$

$\hat{\sigma^2} = \hat{\sigma^2}(\underline{\delta}, \rho)$

- For $<\mu>$ and $<\mu, \delta>$ group
  → use iterative procedure introduced in previous chapters

The problem with these procedures is that the R matrix is a covariance

$\hat{R} = \begin{bmatrix} \hat{R_1} & & \\ & \ddots & \\ & & \hat{R_N} \end{bmatrix}$ and we should make sure that $\hat{R}$ is a covariance matrix

Covariance Σ of a vector w:

For a vator to be covariant

- Σ $n \times n$ should be symmetric
- $w \in R^n$

  Linear combinations of components of w

  $Var(a_1w_1 + a_2w_2 + \cdots + a_nw_n)$ if $a \in R$

  $= \underline{a}'\Sigma\underline{a} > 0$

★ Linear Mixed Models

Model for the phenomenon

$R \to Y = X'\beta + Z'\underline{b} + \varepsilon$ (random vector of coefficients)

$X \in R^p$ vector of known regressors

$Z \in R^q$ " " " "

$\beta \in R^p$ fixed but unknown

$\underline{b} \sim N_q(0, \sigma^2 D)$ here D is a $q \times q$ matrix

$\varepsilon \sim N(0, \sigma^2)$

Model for the data

Suppose that there is 1 layer of clustering with clusters $\{1, 2, \cdots N\}$

So in cluster i, there contains n_i units

$R^{n_i} \to Y_i = X_i\underline{\beta} + Z_i\underline{b_i} + \underline{\varepsilon_i}$ (random effects; fixed effects)

where X_i : $n_i \times p$ design matrix

Z_i : $n_i \times q$ design matrix

$\underline{b} \sim N_q(0, \sigma^2 D)$

$\beta \in R^p$ fixed coeff (unknown)

$\underline{\varepsilon} \sim N_{n_i}(0, \sigma^2 R_i)$

★ Layers for hierarchy

- 1 layer with N groups $\{1, 2, \cdots N\}$
- 2 layer : in group $i \in \{1, 2 \cdots N\}$
  → n_i groups

  $\Rightarrow$ for $i \in \{1, \cdots N\}$ & $j \in \{1, \cdots n\}$
  there are n_{ij} observations

$R^{n_{ij}} \to Y_{ij} = X_{ij}\underline{\beta} + Z_{1,ij}\underline{b_i} + Z_{2,ij}\underline{b_{ij}} + \underline{\varepsilon_{ij}}$ ($n_{ij} \times p$)

here

X_{ij} : $n_{ij} \times p$ design matrix
$Z_{1,ij}$: $n_{ij} \times q_1$ design matrix
$Z_{2,ij}$: $n_{ij} \times q_2$ design matrix
} These are all known

$b_i \sim N_{q_1}(0, \sigma^2 D_1)$
$b_{ij} \sim N_{q_2}(0, \sigma^2 D_2)$
$\varepsilon_{ij} \sim N_{n_{ij}}(0, \sigma^2 R_{ij})$
} These are all independent of each other

★ Overall model with 1 layer

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_N \end{bmatrix} \in R^{n_1+n_2+\cdots n_N = n}$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \quad n \times p$$

$$Z = \begin{bmatrix} Z_1 & & 0 \\ & Z_2 & \\ & & \ddots \\ 0 & & Z_N \end{bmatrix} \quad n \times Nq$$

$$\underline{b} = N_{Nq}\left[0, \sigma^2 \begin{bmatrix} D & & 0 \\ & \ddots & \\ 0 & & D \end{bmatrix}\right]$$

$$\underline{\varepsilon} \sim N_n\left(0, \sigma^2 \begin{bmatrix} R_1 & & \\ & \ddots & \\ & & R_N \end{bmatrix}\right)$$

Model :

$$\underline{Y} = X\underline{\beta} + Z\underline{b} + \underline{\varepsilon}$$

$\underline{\varepsilon} \perp \underline{b}$ (they are independent)

★ conditional distribution of $\underline{Y}$ given $\underline{b}$

for $i \in \{1, 2, \cdots N\}$

$\underline{Y}_i | b_i \sim N_{n_i}(X_i\underline{\beta} + Z_i\underline{b}_i, \sigma^2 R_i)$

$\underline{Y}_i | \underline{b}_i$ and $\underline{Y}_{i'} | \underline{b}_{i'}$ are independent

★ Marginal distribution of $\underline{Y}$

$\underline{Y}_i | b_i \sim N_{n_i}(X_i\underline{\beta} + Z_i\underline{b}_i, \sigma^2 R_i)$

$\underline{b}_i \sim N_q(0, \sigma^2 D)$

$\Rightarrow \underline{Y}_i \sim N_{n_i}(X_i\underline{\beta}, \sigma^2 Z_i D Z_i' + \sigma^2 R_i)$

$Z_i D Z_i' + R_i = V_i$ (substitute here)

$\therefore \underline{Y}_i \sim N_{n_i}(X_i\underline{\beta} + \sigma^2 V_i)$

where $V_i = \begin{bmatrix} V_1 & & \\ & \ddots & \\ & & V_N \end{bmatrix}$

★ structuring D & R_i

$R_i = \Lambda_i C_i \Lambda_i'$

where

$$\Lambda_i = \begin{bmatrix} \lambda_{i1} & & 0 \\ & \ddots & \\ 0 & & \lambda_{in_i} \end{bmatrix}$$

where λ_{ini} = variance function $\in$ <S> group

C_i : $n_i \times n_i$ correlation $\in$ serial / spatial corr. structure

when q is small $D = \begin{bmatrix} d_{11} & d_{12} & d_{1q} \\ & & d_{qq} \end{bmatrix}$

D → d_{ij} free : $\frac{q(q+1)}{2}$ free parameters

D → when q is large : $D = \begin{bmatrix} q_{11} & 0 \\ 0 & q_{NN} \end{bmatrix}$

★ Estimating β

If V was known (ie. D & R_i's, R_N is known)

$$\hat{\underline{\beta}} = (X^{-1}V^{-1}X)^{-1}X'V^{-1}\underline{Y}$$

use $\hat{\underline{\beta}}$ to profile log likelihood and then maximize w.r.t the other parameters.

$$\hat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{N}(\underline{Y}_i - X_i\hat{\underline{\beta}})V_i^{-1}(\underline{Y}_i - X_i\hat{\underline{\beta}})$$

★ Posterior distribution of $\underline{b}_i$ given $\underline{Y}_i$

$\underline{Y}_i | \underline{b}_i \sim N_{n_i}(X_i\underline{\beta} + Z_i\underline{b}_i, \sigma^2 R_i)$

$\underline{b}_i \sim N_q(0, \sigma^2 D)$

$\underline{Y}_i \sim N_{n_i}(X_i\underline{\beta}, \sigma^2 Z_i D Z_i' + \sigma^2 R_i)$

→ $\underline{b}_i | \underline{Y}_i$ is gaussian in R^q

$\therefore E[\underline{b}_i | \underline{Y}_i] = \text{mean}(\underline{b}_i | \underline{Y}_i)$

$= D Z_i' V_i^{-1}(\underline{Y}_i - X_i\underline{\beta})$

Predictor of $\underline{b}_i$:

$$\hat{\underline{b}_i} = \hat{D} Z_i' \hat{V}_i^{-1}(\underline{Y}_i - X_i\hat{\underline{\beta}})$$

(b_i would be random & different for different observations)

Var $(\hat{\underline{b}_i})$ see formula (13.21) in GLLB

Note: $\forall \underline{c} \in R^q$

$\text{var}(\underline{c}'\hat{\underline{b}_i}) \leq \sigma^2 \underline{c}' D \underline{c} = \text{var}(\underline{c}'\underline{b}_i)$

★ Spatial Statistics

• Tobler's first law of Geography

Everything is related with everything else, but near things are more related than distant things.

• Process

$\{Z_s, s \in D\}$

D: spatial domain

Z_s: observation at location $s \in D$

Spatial locations: $s_1, \dots s_n \in D$

Data: $Z_{s_1}, \dots Z_{s_n}$

n data objects: real numbers, labels, vectors, functions.

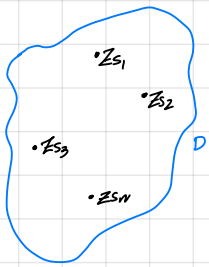
★ Types of spatial data

• Geostatistical Data

In this $D \subseteq R^d$ $d = 1,2,3,\dots$ (D is a subset of R^d)

and D is a rectangle

Z_s is defined for every $s \in D$

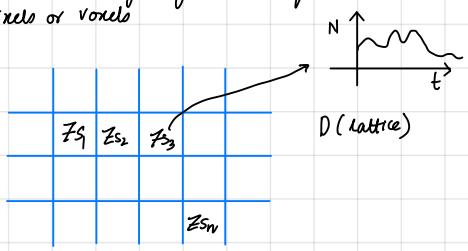


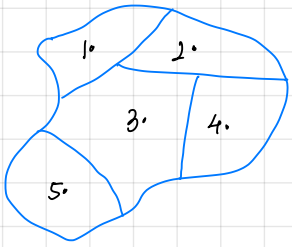
Goal: • To estimate the spatial dependence of variables (how variables are related due to their positions in D)

- Structured spatial covariances → estimate parameters
- Prediction at an unsampled location ($s_0 \in D$)

• Lattice Data

D is a countable (often finite) set of sites (administrative areas) or pixels or voxels





let's suppose, D is countable

$D = \{1, 2, \dots N\}$

Neighbouring structure:

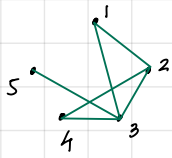
Given $i \in D$

$N_i = \{s \in D$ in the neighbourhood of $i\}$

After the neighbouring structure is specified, D becomes a graph

Ex: i & j are the same neighbour if they share a border

Nodes of the graphs are the sites



⇒ 1 is connected to 2 & 3
2 is connected to 1, 3 & 4
and so on

Formally:

W is a $n \times n$ matrix where $n \in$ no. of sites pixels, elements of lattice

$W = [w_{ij}]_{n \times n}$

$w_{ij} = \begin{cases} \neq 0 & \text{if } j \text{ is a neighbour of } i \\ 0 & \text{otherwise} \end{cases}$

$W =$

	1	2	3	4	5
1	0	1	1	0	0
2	1	0	1	1	0
3			0		
4				0	
5					0

Standardizing the weight matrix

Standardizing by rows

$\tilde{w}_{ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}} \Rightarrow \sum_{j=1}^{n} \tilde{w}_{ij} = 1$

$\sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{w}_{ij} = n$

Suppose $d: D \times D \to [0, \infty)$ is a metric

$d(i,j)$ is the distance b/w i & j in D

Example 1: Fix $\delta > 0$

$w_{ij} = \begin{cases} 1 & \text{if } d(i,j) < \delta \\ 0 & \text{otherwise} \end{cases}$

Example 2: K-nearest neighbour

$w_{ij} = \begin{cases} 1 & \text{if } j \text{ is among the } k\text{- nearest sites to } i \\ 0 & \text{otherwise} \end{cases}$

Example 3: $w_{ij} = f(d(i,j), \theta)$

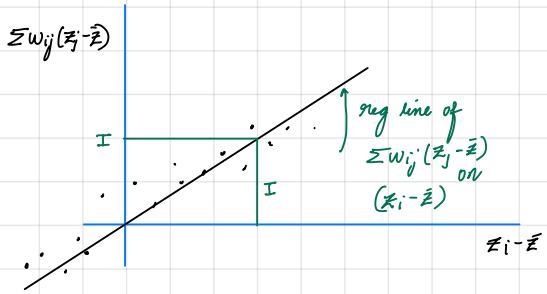
- $w_{ij} = \frac{1}{d^{\alpha}(i,j)} \quad \alpha > 0$
- $w_{ij} = \exp[-\beta d(i,j)], \ \beta > 0$

★ Spatial Auto correlation

Moran's Index (I)

(a) $$I = \frac{\sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum (z_i - \bar{z})^2}$$ if weights are standardized

(b) $$I = \frac{\sum\sum w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum\sum w_{ij}} \cdot \frac{1}{\frac{\sum (z_i - \bar{z})^2}{n}}$$ if weights are not standardized



★ Checking for auto correlation

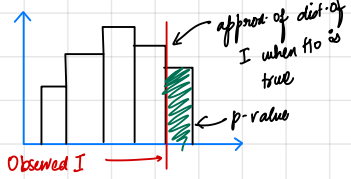
H_0 : no spatial auto corr.

H_1 : spatial auto corr.

we use the permutation test

for $k = 1, \dots, M$ (large)

- permute $(z_1 \cdots z_n)$ (sites/locations) we get $z_{\pi(1)} \cdots z_{\pi(n)}$ where π is a permutation of $\{1, \dots, n\}$
- compute $I : I^k$
- compute the histogram of $I^{(1)}, \dots I^{(k)}$



LISA

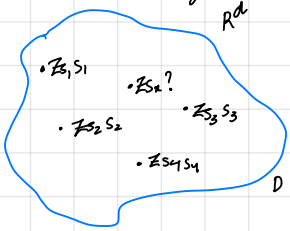
$$I_i = \frac{(z_i - \bar{z}) \sum w_{ij}(z_j - \bar{z})}{\sum (z_i - \bar{z})^2}$$

★ Geostatistical Data

We observe a variable

$R \in Z_s$ where $s \in D \subseteq R^d$, $d = 1, 2, 3, \dots$

D should contain a rectangle in R^d



If data: fix $s_1, \dots s_n$

$Z_{s_1}, \dots Z_{s_n}$

used to predict the value of Z_s in a specific location

General Model

$\forall s \in D$

we assume that

$Z_s = m_s + \delta(s)$ ($\delta(s)$ → random; m_s → deterministic part)

where

$m_s \rightarrow$ drift / trend of the process (captures large scale variation)

$\delta(s) \rightarrow$ residual (captures the local scale variation)

★ Identifying process stationarity

$Z = \{Z_s, s \in D\}$ is a random field

The Law of Z is known if

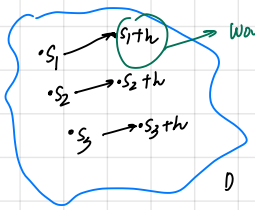
$\forall n \geq 1$ and $\forall s_1, \dots s_n \in D$

$$F_{s_1, \dots s_n}(z_1 \cdots z_n) = P[Z_{s_1} \leq z_1, \dots Z_{s_n} \leq z_n] \quad \forall z_1, \dots z_n \in R$$

These distributions should follow some consistency conditions (Kolmogorov's condition)

Definition : (Strong stationarity)

$\{Z_s, s \in D\}$ is said to be strongly stationary if for every $n \geq 1$ and $s_1, \dots s_n$ the distribution of $Z_{s_1}, \dots Z_{s_n}$ is the same as the distribution of $Z_{s_1+h}, \dots Z_{s_n+h}$ $\forall h \in R^d$



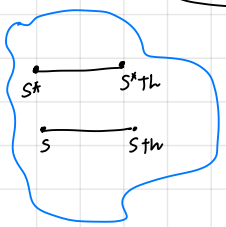
Definition : (weak stationarity)

$\{Z_s, s \in D\}$ is said to be weakly stationary if

(1) $E[Z_s] = m \quad \forall s \in D$

(2) $cov(Z_s, Z_{s+h}) = C(h) \quad \forall s \in D, \forall h \in \mathbb{R}^d$ → C is a function of h

this happens if $E[Z_s^2] < \infty \quad \forall s \in D$



Proposition

Strong stationarity implies weak stationarity but not vice versa

The function

$C : \mathbb{R}^d \to \mathbb{R}$ such that $C(h) = cov(Z_s, Z_{s+h})$

is called a covariogram.

Properties of C :

(1) $\forall n \geq 1$, and any $s_1, \cdots s_n \in D$ define

$c_{ij} = cov(Z_{s_i}, Z_{s_j}) = C(s_i - s_j)$

$C = [c_{ij}]$ n×n matrix

If $(\lambda_1, \cdots \lambda_n) \in \mathbb{R} = \lambda'$

$\leq var\left(\sum_{i=1}^n \lambda_i Z_{s_i}\right) = \lambda' C \lambda$

$= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j c_{ij}$

$= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(s_i - s_j)$

∴ C is a positive definite function

(2) $C(h) = C(-h)$

$\Rightarrow cov(Z_s, Z_{s+h}) = cov(Z_{s-h}, Z_s)$

(3) $|cov(Z_s, Z_{s+h})| \quad \forall h \in \mathbb{R}^d$

$= \frac{|cov(Z_s, Z_{s+h})|}{\sqrt{var(Z_s) \cdot var(Z_{s+h})}}$ → $var(Z_s) = cov(Z_s, Z_s) = C(0)$ since $h=0$

$= \frac{|C(h)|}{\sqrt{C(0) \cdot C(0)}} = \frac{|C(h)|}{C(0)}$

∴ $|C(h)| \leq C(0)$

Definition : Intrinsic stationarity

$\{Z_s, s \in D\}$ is said to be intrinsically stationary if

(1) $E[Z_s] = m \quad \forall s \in D$

(2) $var(Z_s - Z_{s+h}) = 2 \cdot \gamma(h) \quad \forall s \in D, \forall h \in \mathbb{R}^d$

where

$2\gamma : \mathbb{R}^d \to \mathbb{R}$ is called variogram

$\gamma : \mathbb{R}^d \to \mathbb{R}$ is called semivariogram

Proposition: If a field is weakly stationary, it is also intrinsically stationary

Let $\{Z_s, s \in D\}$ is weakly stationary

$var(Z_s - Z_{s+h}) = var(Z_s) + var(Z_{s+h}) - 2 \cdot cov(Z_s, Z_{s+h})$

$= C(0) + C(0) - 2 \cdot C(h)$

$= E[(Z_s - Z_{s+h})^2] \quad = 2C(0) - 2C(h)$

$= 2 \cdot \gamma(h)$

∴ for weakly stationary process

$\gamma(h) = C(0) - C(h) \quad \forall h \in \mathbb{R}^d$

Properties of γ

(1) $\forall n \geq 1, \; s_1, \cdots s_n \in D$

and for $\lambda_1, \cdots \lambda_n$ such that $\sum \lambda_i = 0$

$\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(s_i - s_j) \leq 0$

(negative definition)

(2) $\gamma(h) = \gamma(-h) \quad \forall h \in \mathbb{R}^d$

(3) $\gamma(h) \geq 0 \quad \forall h \in \mathbb{R}^d$

(4) $\gamma(0) = var(Z_s - Z_s) = 0$

(5) $\lim_{\|h\| \to \infty} \frac{\gamma(h)}{\|h\|^2} = 0$ (sub quadratic)

$\|h\|$ → distance

Definition

If $\gamma(h)$ is indeed a function of $\|h\|$ then $\{Z_s, s \in D\}$ intrinsically stationary is also isotropic.

ie.

$var(Z_s - Z_{s+h}) = 2\gamma(\|h\|) \quad \forall s \in D, h \in D$

WkT

① $\gamma(0) = 0$ since $\gamma(0) = \text{Var}(Z_s - Z_s) = \text{Var}(0)$

$\lim_{\|h\| \to 0} \gamma(h) = \tau^2 \geq 0$ if $\tau^2 > 0 \Rightarrow \gamma(h)$ is not continuos in 0

τ^2 is called nugget

(measures instantaneous variability)

② $\lim_{\|h\| \to \infty} \gamma(h) = \gamma^2 + \tau^2$ if it is finite

$= \text{Var}[Z_s - Z_{s+h}) = 2[C(0) - C(h)]$

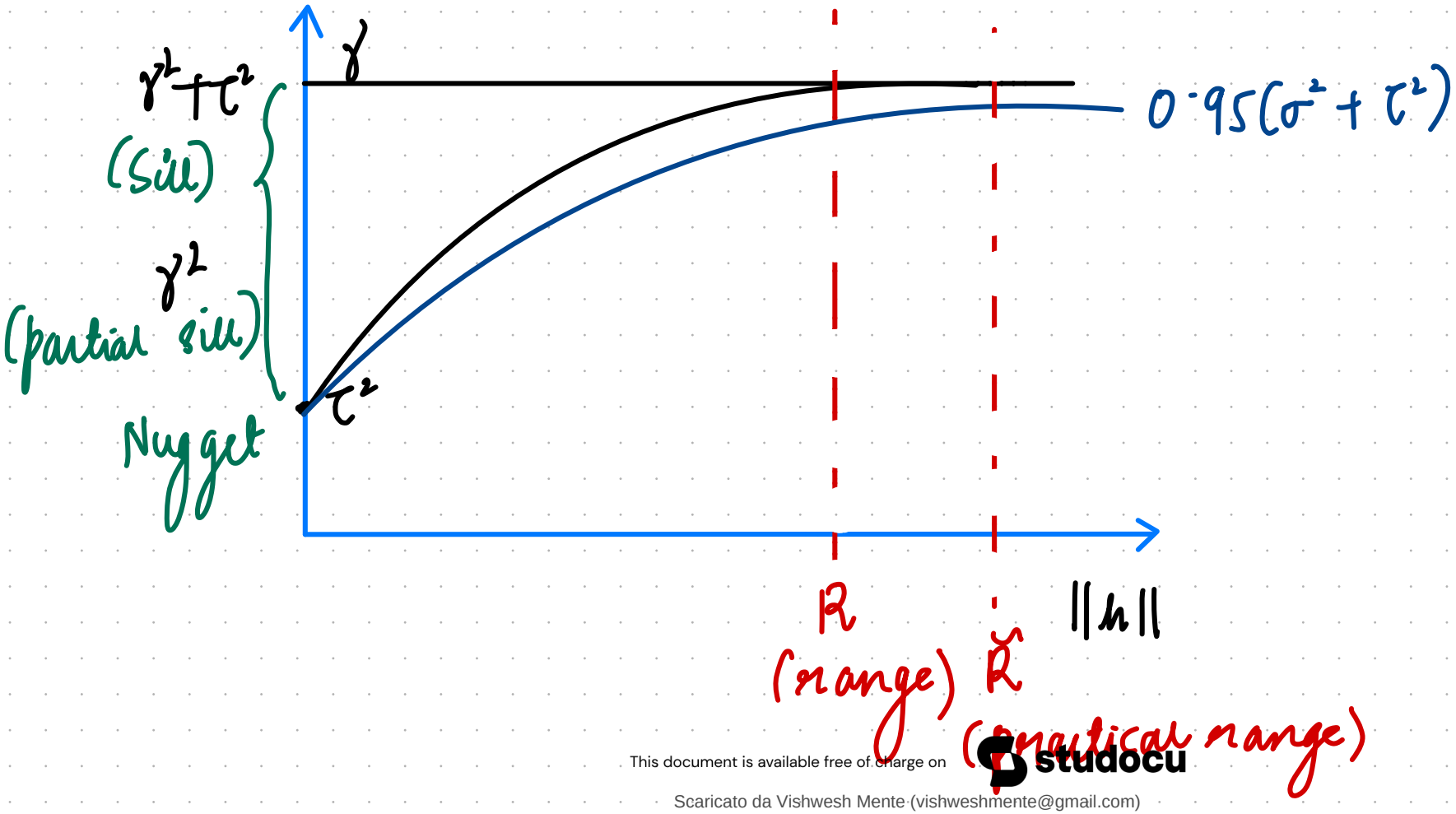
when $h \to \infty$, $C(h) \to 0$

$\therefore = 2 \times C(0)$

If $\gamma^2 + \tau^2 = \lim_{\|h\| \to \infty} \gamma(h)$ is finite then

$\{Z_s, s \in D\}$ is weakly stationary

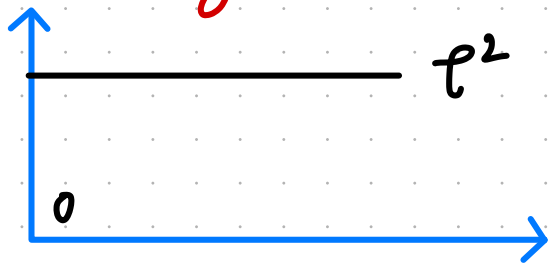
$\gamma^2 + \tau^2$ is called sille



Parametric Models for semi-variogram

- **Pure Nugget**

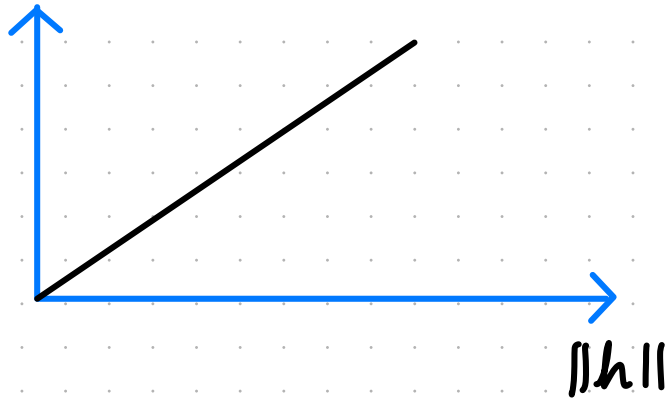
$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \tau^2 & h \neq 0 \end{cases}$$



- **Linear Model (Isotropic)**

$$\gamma(h) = a^2 \|h\|$$

(no sill, no nugget)



- **Exponential Model**

$$\gamma(h) = \sigma^2 \left(1 - \exp\left[-\frac{\|h\|}{a^2}\right]\right) \quad a, \sigma \neq 0$$

No nugget, sill: σ^2, practical range: $3a^2$

- **Gaussian Model**

$$\gamma(h) = \sigma^2 \left[1 - \exp\left[-\frac{\|h\|^2}{a^2}\right]\right) \quad a, \sigma \neq 0$$

No nugget, sill $= \sigma^2$, practical range = ?

- **Spherical Model**

$$\gamma(h) = \begin{cases} \sigma^2 \left\{\frac{3}{2}\frac{\|h\|}{a^2} - \frac{1}{2}\left(\frac{\|h\|}{a^2}\right)^3\right\} & \text{for } 0 \leq \|h\| \leq a^2 \\ \sigma^2 & \text{for } \|h\| \geq a^2 \end{cases}$$

No nugget, sill : σ^2, range $= a^2$

Proposition

If γ_1 & γ_2 are valid semi-variograms then

(1) $\gamma_1 + \gamma_2$ is a valid semi-variogram.

(2) $b^2\gamma_1$ is a valid semi-variogram.

Estimating the semi-variogram

$$\gamma(h) = C(0) - C(h)$$

$$\gamma(h) = E\left[(Z_s - Z_{s+h})^2\right] \quad h \in R^d$$

$$\hat{\gamma}(h) = \frac{1}{2\,|N(h)|} \sum_{(s_i, s_j) \in N(h)} \left[Z_{s_i} - Z_{s_j}\right]^2$$

} Empirical estimate of γ denoted by $\hat{\gamma}(h)$

$$N(h) = \{ (s_i, s_j) : \|s_i - s_j\| = \|h\| \}$$

There are only a finite no. of $(s_i - s_j)$ ie only a finite number of h's for which $\hat{\gamma}$ is computed

To deal with the variability, we adjust the formula such that

$$\hat{\gamma}(h_i) = \frac{1}{2|N(h)|} \sum_{(s_i - s_j) \in N(h)} [z_{s_i} - z_{s_j}]^2$$

$$N(h) = \{(s_i, s_j) : \|s_i - s_j\| \in T_i\}$$

where h_i is the center of T_i

$i = 1, \ldots K$

$K \in$ no. of bins

2-step procedure of estimating $\hat{\gamma}$

(1) Estimate empirical semi-variogram $\hat{\gamma}$ by binning data

$\Rightarrow \hat{\gamma}(h_1), \dots \hat{\gamma}(h_K) \in K$ bins

(2) Fit a parametric model to $(\hat{\gamma}(h_1), \dots \hat{\gamma}(h_K))$

ie. $\gamma(h;\theta)$ parametric model (spherical, gauss., lin. etc)

and then

$$\min_{\theta} \sum_{i=1}^{n} \left[\hat{\gamma}(h_i) - \gamma(h_i;\theta)\right]^2$$

Better approach to ↗

$$\min_{\theta} \sum_{i=1}^{K} w_i \left[\hat{\gamma}(h_i) - \gamma(h_i;\theta)\right]^2$$

w_i → weight prop. to size of bin i

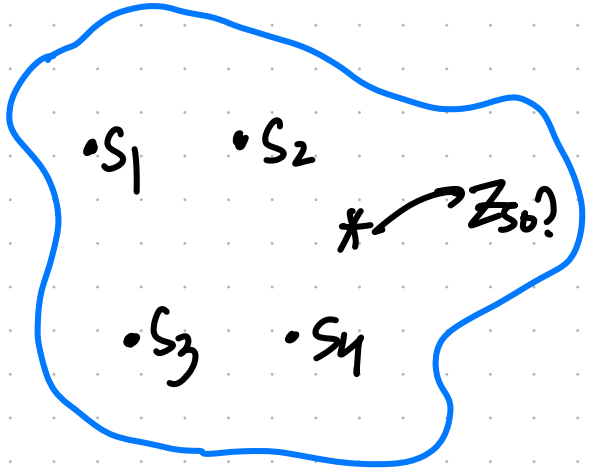
Even better approach

$$\hat{\gamma}(h) = (\hat{\gamma}(h_1) \dots \hat{\gamma}(h_K))$$

(Generalised Least squares - GLS)

$$\min_{\theta} \left[\underline{\hat{\gamma}(h)} - \underline{\gamma(h;\theta)}\right]' \text{cov}^{-1} \left[\underline{\hat{\gamma}(h)} - \underline{\gamma(h;\theta)}\right]$$

★ Prediction (Kriging)



Given the position of Z_s

at locations $Z_{s_1} \dots Z_{s_n}$

$\Rightarrow$ predict the Z_s at location Z_{s_0}

$\underline{Z} = (Z_{s_1} \dots Z_{s_n})$ gives a prediction of Z_{s_0}

$f(\underline{z}) = Z_{s_0}^*$ (predicted) → this is called kriging

Goal is to minimize

$$E\left[(Z_{s_0} - Z_{s_0}^*)^2\right] = E\left[(Z_{s_0} - f(\underline{z}))^2\right]$$

$$Z_{s_0}^* = E\left[Z_{s_0} \mid Z_{s_1}, \dots Z_{s_n}\right]$$

Properties: ① $E\left[Z_{s_0}^*\right] = E\left[Z_{s_0}\right]$ unbiased

② $\mathrm{cov}\left(Z_{s_0}^*, Z_{s_0} - Z_{s_0}^*\right) = 0$

③ If $s_0 \in \{s_1, \dots s_n\}$ then $E\left[Z_{s_0} \mid Z_{s_1}, \dots Z_{s_n}\right] = Z_{s_0}$

(interpolation)

④ If $\{Z_s, s \in D\}$ is gaussian then $E\left[Z_{s_0} \mid Z_{s_1} \dots Z_{s_n}\right] = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z_{s_i}$ (linear)

Problems in Kriging

Find $\hat{\lambda_0}, \hat{\lambda_1}, \dots \hat{\lambda_n}$ such that $Z_{s0}^* = \hat{\lambda_0} + \sum_{i=1}^{n} \lambda_i Z_{si}$

Satisfies:

1. $E[Z_{s0}^*] = E[Z_{s0}]$
2. $(\hat{\lambda_0}, \dots \hat{\lambda_n}) = \min_{(\lambda_0, \dots \lambda_n)} E\left[\left(Z_{s0} - (\lambda_0 + \sum_{i=1}^{n} \lambda_i Z_{si})\right)^2\right]$

Only if Z_{s0}^* exist, then Z_{s0}^* is the BLUP for Z_{s0}

(Best linear unbiased predictor)

$$Z_{s0}^* = \hat{\lambda_0} + \sum_{i=1}^{n} \hat{\lambda_i} Z_{si}$$

Simple kriging

Assume that $E[Z_s] = m_s$ is known for $s \in \{s_0, s_1, \dots s_n\}$

then unbiasedness requires that

$$E\left[\lambda_0 + \sum \lambda_i Z_{s_i}\right] = \lambda_0 + \sum \lambda_i m_{s_i}$$

$$\| $$

$$E[Z_{s_0}] = m_{s_0}$$

$$\boxed{\hat{\lambda_0} = m_{s_0} - \sum_{i=1}^{n} \hat{\lambda_i} m_{s_i}}$$

Now we need to

$$\min_{(\lambda_1 \dots \lambda_n)} E\left[Z_{s_0} - \left(\sum_{i=1}^{n} \lambda_i Z_{s_i} + \left(m_{s_0} - \sum_{i=1}^{n} \lambda_i m_{s_i}\right)\right)^2\right]$$

$$\therefore \min E\left[\left((Z_{S0} - m_{S0}) - \left[\sum_{i=1}^{n} \lambda_i Z_{Si} - \sum_{i=1}^{n} \lambda_i m_{Si}\right)\right)^2\right]$$

$$= \min \quad Var(Z_{S0}) + Var\left(\sum \lambda_i Z_{Si}\right) - 2 \cdot Cov\left(Z_{S0}, \sum \lambda_i Z_{Si}\right)$$

$$= \min \quad var(Z_{S0}) + var\left(\sum \lambda_i Z_{Si}\right) - 2 \cdot \sum \lambda_i \, Cov\left(Z_{S0}, Z_{Si}\right)$$

Notations

$\sigma_{00} = Var(Z_{S0})$

$\sigma_{0i} = Var\left[Z_{S0}, Z_{Si}\right)$

$\underline{c}_0 = (\sigma_0, \cdots \sigma_{0n})$

$\sigma_{ij} = Cov(Z_{Si}, Z_{Sj}) \quad i,j = 1 \cdots n$

$\Sigma = [\sigma_{ij}]$ $n \times n$ cov matrix of $(Z_{S1}, \cdots Z_{Sn})$

$\underline{\lambda} = (\lambda_1 \cdots \lambda_n)'$

$\therefore$ $\boxed{\min_{\underline{\lambda}} \underline{\lambda}' \Sigma \underline{\lambda} - 2\underline{\lambda}' \underline{c}_0 + \sigma_{00}}$

$\Rightarrow \Sigma \underline{\lambda} = \underline{c}_0 \Rightarrow \hat{\underline{\lambda}} = \Sigma^{-1} \underline{c}_0$ (if Σ is inverted)

$$Z_{s0}^* = \left(m_{s0} - \sum \hat{\lambda}_i m_{si} \right) + \sum_{i=1}^{n} \hat{\lambda}_i Z_{si}$$

so we need to know Σ and $\underline{c}_0$

★ Ordinary Kriging

$E[Z_s] = m \quad \forall s \in D$

where m is unknown

Unbiasedness:

$$E[\lambda_0 + \sum \lambda_i Z_{si}] = E[Z_{s0}] \quad \therefore \hat{\lambda_0} + \sum \hat{\lambda_i} m = m$$

$$\forall (\lambda_1, \ldots \hat{\lambda_n}, \hat{\lambda_0})$$

$$\left.\begin{aligned} &\hat{\lambda}_0 = 0 \\ &\sum_{i=1}^{n} \hat{\lambda}_i = 1 \end{aligned}\right\} \text{ Constants}$$

$$= \begin{cases} \min\limits_{(\lambda_1, \cdots \lambda_n)} E\left[(Z_{S0} - \sum \lambda_i Z_{Si})^2\right] \\ \sum \lambda_i = 1 \end{cases}$$

$$= \begin{cases} \min\limits_{\underline{\lambda}} \underline{\lambda}' \Sigma \underline{\lambda} - 2 \underline{\lambda} c_0 + \sigma_{00} \\ \lambda' = 1 \end{cases}$$

$$= \min_{\underline{\lambda}, \xi} \lambda' \Sigma \underline{\lambda} - 2\lambda c_0 + \sigma_{00} + \xi(\lambda' - 1)$$

$$= \begin{bmatrix} \Sigma & 1 \\ 1' & 0 \end{bmatrix} \begin{bmatrix} \underline{\lambda} \\ \xi \end{bmatrix} = \begin{bmatrix} c_0 \\ 1 \end{bmatrix} = \hat{\lambda}$$

Universal Kriging

$$E[Z_s] = a_0 + a_1 f_1(s) + \cdots a_L f_L(s)$$

where

$a_0, a_1, \cdots a_L$ are coefficients - unknown

$f_1(s), \cdots f_L(s)$ are covariates - known $\forall s \in D$

Model Z_s: $m_s + \delta(s)$ and $E[\delta(s)] = 0$

Unbiasedness: $E\left[\hat{\lambda_0} + \sum \hat{\lambda_i} Z_{s_i}\right] = E[Z_{s_0}]$

$$\hat{\lambda_0} + \sum_{i=1}^{n} \hat{\lambda_i}\left(a_0 + a_1 f_1(s_1) + \cdots a_L f_L(s_i)\right) = a_0 + a_1 f_1(s_0) + \cdots a_L f_L(s_0)$$

$\Rightarrow \hat{\lambda_0} = 0$ then $\sum \hat{\lambda_i} f_i(s_i) = f_i(s_0)$

$\sum \hat{\lambda_i} = 1$ then $\sum \hat{\lambda_i} f_j(s_i) = f_j(s_0) \quad j = 1 \cdots L$

Hence

$$\begin{cases} \min_{\lambda} \ \underline{\lambda}'\Sigma\underline{\lambda} - 2\underline{\lambda}'\underline{c}_0 + \sigma_{00} \\ \sum \lambda_i = 1 \\ \sum \lambda_i f_j(s_i) = f_j(s_0) \quad j = 1, \dots, L \end{cases}$$

$$\Rightarrow \begin{bmatrix} \Sigma & F \\ F' & 0 \end{bmatrix} \begin{bmatrix} \underline{\lambda} \\ \xi \end{bmatrix} = \begin{bmatrix} c_0 \\ f_0 \end{bmatrix}$$

where $F = \begin{bmatrix} 1 & f_1(s_1) & \dots & f_L(s_1) \\ 1 & f_1(s_2) & \dots & f_L(s_2) \\ 1 & f_1(s_n) & \dots & f_L(s_n) \end{bmatrix}$ (design matrix)

and where

$$f_0 = \begin{pmatrix} f_1(s_0) \\ \vdots \\ f_L(s_0) \end{pmatrix}$$

$$\therefore \quad Z_{s_0}^{*} = \sum \hat{\lambda}_i Z_{s_i}$$

To solve simple, ordinary & universal kriging equations

we need to know Σ & $\underline{c}_0$

Assumptions:

for simple kriging: $\{Z_s - m_s, s \in D\}$ is weakly stationary

for ordinary kriging: $\{Z_s, s \in D\}$ is weakly stationary

for universal kriging: $\{\delta(s), s \in D\}$ is weakly stationary

with these assumptions

- Use data $(Z_{s_1}, \dots Z_{s_n})$ + covariates to estimate $\hat{\gamma}$

$$\gamma(h) = C(0) - C(h) \quad \text{where} \quad C(0) = \text{Var}(Z_s)$$
$$C(h) = \text{cov}(Z_s, Z_{s+h})$$

$\therefore \hat{C}(h) = \hat{C}(0) - \hat{\gamma}(h)$ here $\hat{C}(0)$ is the sill

$\text{Cov}(Z_{s_i}, Z_{s_j}) = \hat{\sigma}_{ij} = \hat{C}(0) - \hat{\gamma}(s_i - s_j)$

In universal kriging, for estimating δ, you need to know δ

$Z_s = a_0 + a_1 f_1(s) + \cdots a_L f_L(s) + \delta(s)$

$\underline{Z} = (Z_{s_1}, \cdots Z_{s_n})'$ $\underline{a} = (a_0, a_1, \cdots a_L)$

$$\boxed{\underline{Z} = F\underline{a} + \underline{\delta}}$$ linear model $E(\underline{\delta}) = 0$

$\underline{\delta} = (\delta(s_1), \cdots \delta(s_n))$ $\text{Cov}(\underline{\delta}) = \Sigma$

- Estimate $\underline{a}$ with $\hat{\underline{a}} \longrightarrow \quad \hat{\underline{a}} = (F' \Sigma F)^{-1} F' \Sigma^{-1} z$
- Estimate $\delta = z - F\hat{a}$
- Estimate γ

Iterative Process

Step 0: $\hat{a}^{(0)} = (F'F)^{-1} F' z \quad$ (OLS)

Step i: $\underline{\delta} = \underline{z} - F\hat{\underline{a}}^{(K-1)}$

$\Rightarrow$ Estimate $\gamma^{(i)} = \hat{\Sigma}^{(i)} = \hat{\underline{a}}^{(i)} = \left(F' \Sigma^{(i)} F\right)^{-1} F' \hat{\Sigma}^{(i)-1} z$

Variance of Kriging predictors

- $\text{Var}\left(Z_0^*(s_k)\right) = E\left[\left(Z_{s0} - Z_{s0}^*(s_k)\right)^2\right]$

$$= C(0) - \sum_{i=1}^{n} \lambda_i C(s_0 - s_i) = \sigma^2_{sk}$$

- $\text{Var}\left(\hat{Z}_{s0}^*(OK)\right) = C(0) - \sum \hat{\lambda}_i C(s_0 - s_i) - \hat{\xi} = \sigma^2_{OK}$
- $\text{Var}\left(Z_{s0}^*(UK)\right) = C(0) - \sum \hat{\lambda}_i C(s_0 - s_i) - \sum \hat{\xi}_j f_j(s_0) = \sigma^2_{UK}$

Markov Processes (Chains)

We have a process $\{X_t, t \in [0, \infty)\}$ with $X_t \in R$ is called Markov if, given X_t, the distribution of X_s for all $s > t$ does not depend of X_u for $u < t$.

Equivalently.

given X_t,

X_u with $u < t$
X_s with $s > t$
} are independent

S is called state space

We will consider discrete time process so $t = n = 0, 1, 2, 3, \ldots$

moreover $X_n \in S$ discrete, finite, countable $\{X_n\} \rightarrow$ markov chains

Markov property

$$P[X_{n+1}=j \mid X_0=i_0, X_1=i_1, \cdots X_n=i_{n+1}] \quad \forall\ i_0, i_1, \cdots i_n \in S$$

this probability will be = $P[X_{n+1}=j \mid X_n=i_n]$

↳ only dependent on previous element

Notation : $\forall i,j \in S$

$$P[X_{n+1}=j \mid X_n=i] = P_{ij} \quad \text{(transition probability)}$$

- If $P_{ij}^{(n,n+1)}$ $\forall\ i,j \in S$ and $\forall n$ are stationary, then $P_{ij}^{(n,n+1)} = P_{ij}$ and the chain is said to be homogeneous.

- For $i \in S$, let

$$p_i = P[X_0 = i] \quad \text{for initial periods i.e. } p_i \geq 0 \text{ and } \sum p_i = 1$$

Notice that, because of the markov property,

$$\forall\ i_0, i_1, \cdots i_n \in S$$

$$P[X_0 = i_0, X_1 = i, X_2 = i_2, \cdots X_n = i_n] = p_i\, p_{i_0 i}\, p_{i i_2} \cdots p_{i_{n-1} i_n}$$

Introducing a matrix (transition matrix)

$$P = [p_{ij}] \quad \text{of order } |S| \times |S|$$

$$\therefore \sum_{j \in S} p_{ij} = 1 \quad , \forall p_{ij} \geq 0$$

Central to analysis of Markov chains

$$p_{ij}^{(n)} = P[X_n = j \mid X_0 = i] = P[X_{m+n} = j \mid X_m = i]$$

$$P^{(n)} = [p_{ij}^{(n)}] \quad \text{(n-step transition matrix)}$$

$$\text{Set } p_{ij}^{(0)} = \begin{cases} 0 & \forall\, i \neq j \\ 1 & \forall\, i = j \end{cases}$$

Proposition: For all $i, j \in S$:

$$p_{ij}^{(n)} = \sum_{k \in S} p_{ik}\, p_{kj}^{(n-1)} \quad \text{ie.} \quad \boxed{P^{(n)} = P \cdot P^{(n-1)}}$$

$$P^{(n)} = P \cdot P^{(n-1)} = \underbrace{P \cdot P \cdots P}_{n\text{-times}} = P^n$$

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

(This means that for the probability at n^{th} observation ($P^{(n)}$) we need to multiply P, n times ie. P^n (which = $P^{\wedge n}$)

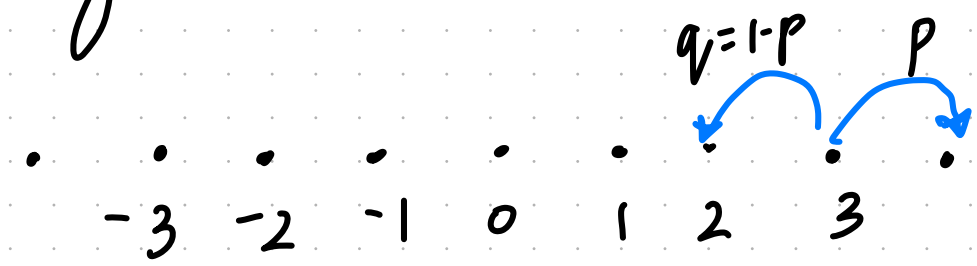
If $0 \leq K \leq n$

then $P^n = P^K \cdot P^{n-K} = P^{(K)} \cdot P^{(n-K)}$

Chapman-Kolmogorov's equations

Examples (1)

Considering a 1-D random walk (T)



then the transition matrix

$$P = i \begin{bmatrix} \cdots & \cdots & q & 0 & p & 0 & \cdots \\ & & q & 0 & p & & \\ & & 0 & q & & & \end{bmatrix}$$

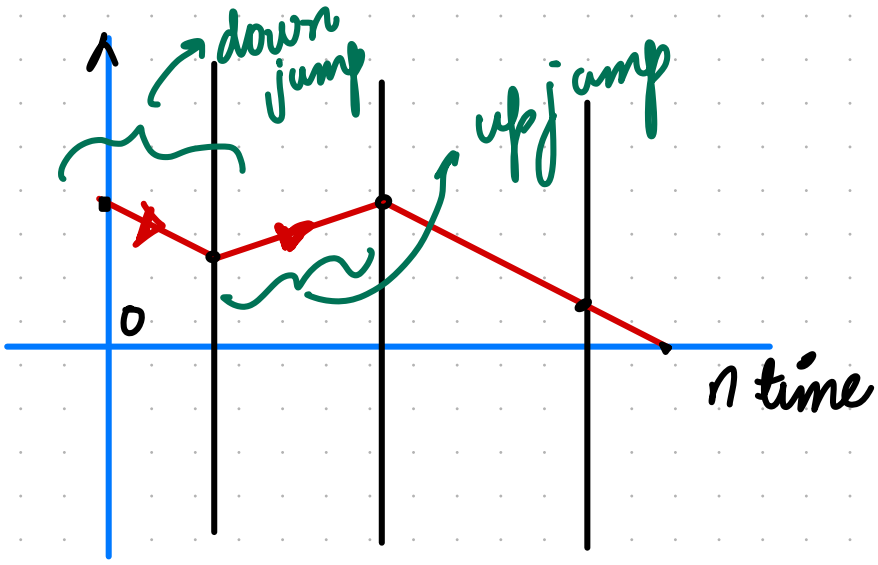
$i-1 \qquad i+1$

$\therefore p_{ij}^{(n)} = P[X_n = j \mid X_0 = i]$

Let $j \geq i$

and u: sum of up jumps

d: sum of down jumps



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

where $Y_1, Y_2 \cdots Y_n$ = iid

$$Y_i = \begin{cases} +1 & p \\ -1 & 1-p \end{cases}$$

$$\therefore u = \sum_{k=1}^{n} \mathbb{1}[Y_k = +1] \quad \text{(jump-up)}$$

$$d = \sum_{k=1}^{n} \mathbb{1}[Y_k = -1] \quad \text{(jump-down)}$$

here $u + d = n$ &
$u - d = j - i$ } adding them

$\therefore n+j-i$ must be even

$2u = n + j - 1$

$$\boxed{u = \tfrac{1}{2}(n+j-1)}$$

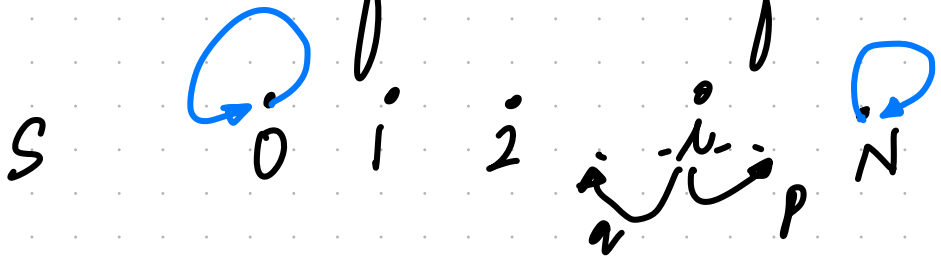
$$P_{ij}^{(n)} = \begin{cases} 0 & \text{if } n+j-i \text{ is odd} \\ \binom{n}{\frac{1}{2}(n+j-i)} p^{\frac{1}{2}(n+j-i)} \times (1-p)^{n-\frac{1}{2}(n+j-i)} & \text{if } n \text{ is even} \end{cases}$$

Eg: $P_{00}^{(2n)} = \binom{2n}{n} p^n (1-p)^n$

probability of coming back from where you started

② Gambler's ruin

Here S is a finite no. of integers



If you reach 0, you stay at 0

If you reach N, you stay at N

only if you are at $i > 0$ and $i < N$, you can jump up or down

where $i \in$ current gambler fortune

$N - i \in$ fortune of adversary

Transition matrix

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & & & & N \end{array} \\ \begin{bmatrix} 1 & 0 & & & & & 0 \\ q & 0 & p & & & & \\ 0 & q & 0 & p & & & 0 \\ \vdots & & & & & & \\ 0 & & & & 0 & q & 0 & p \\ 0 & & & & & & 1 \end{bmatrix} \end{array}$$

$p_{ij} \rightarrow$ compute $P^{(n)}$ with R

if $T = \inf\{n : X_n = 0 \text{ or } X_n = N\}$

$P[X_T = 0 \mid X_0 = i] \quad i = 0, 1, \dots N$

probability of gambler going broke

$\boxed{\lambda_0 = 1}$ (probability of going broke with 0 euros)

$\lambda_1 = q + p \cdot \lambda_2$

$\lambda_2 = q \cdot \lambda_1 + p \cdot \lambda_3$ $\quad \therefore \lambda_{N-1} = q \cdot \lambda_{N-2} + p \cdot 0$

(p of going broke with N euros)

$\boxed{\lambda_N = 0}$

Exercise: when $N=3$, prove that $\lambda_1 = \frac{q}{1-pq}$ and $\lambda_2 = \frac{q^2}{1-pq}$

if $p=q=\frac{1}{2}$, $\lambda_1 = \frac{2}{3}$, $\lambda_2 = \frac{1}{3}$

For general N

$$\lambda_i = \begin{cases} \frac{N-i}{N} & \text{if } p=q=\frac{1}{2} \\ \frac{(q/p)^i - (q/p)^N}{1-(q/p)^N} & \text{if } p \neq q \end{cases}$$

Now if $N \to \infty$

$\lim_{N\to\infty} \lambda_i = 1$ if $p=q$

$\lim_{N\to\infty} \lambda_i = 1$ if $q>p$

$\lim_{N\to\infty} \lambda_i = (q/p)^i$ $p>q$

③ Inventory problem

Consider a warehouse of c stock units

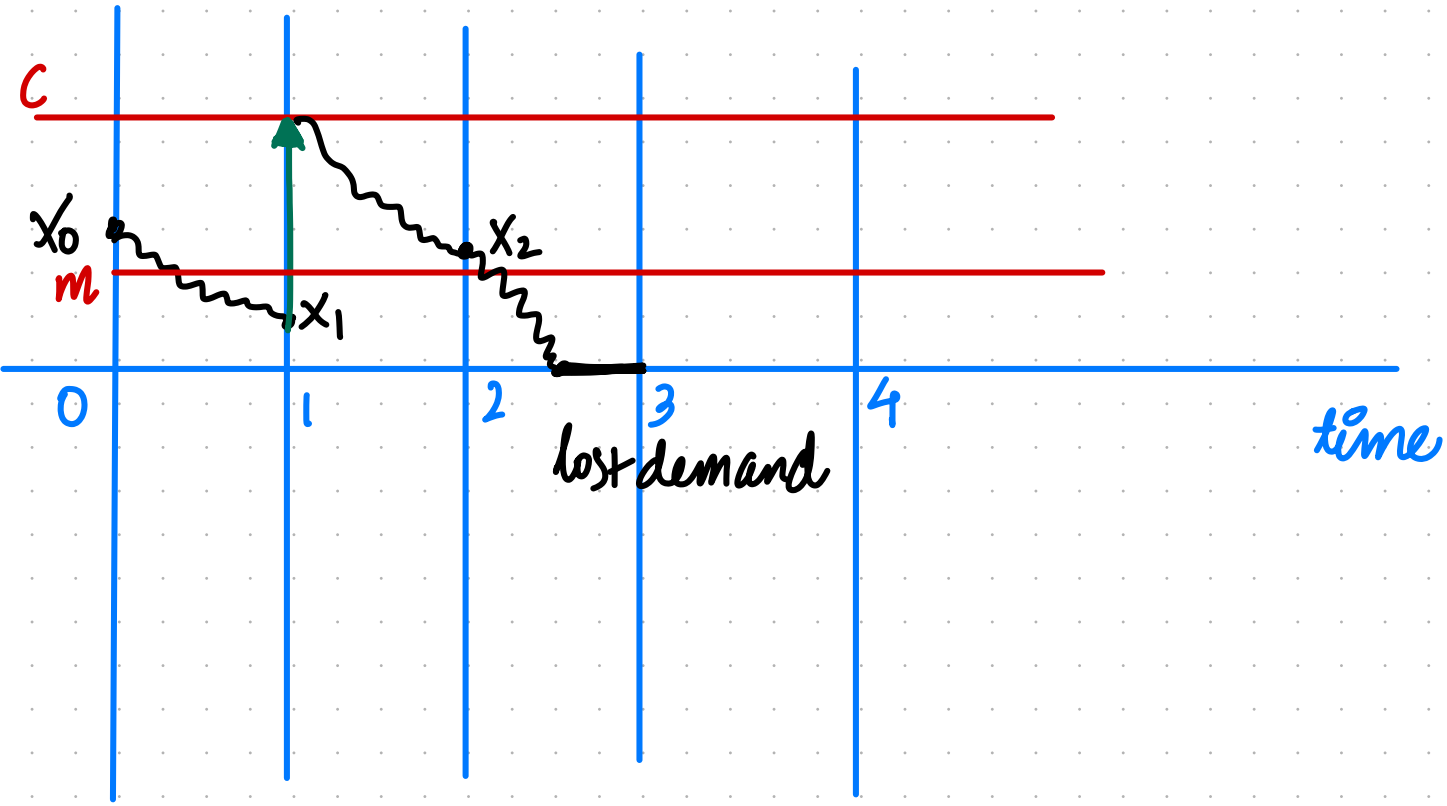
Every time period n: $D_n \in \{0, 1, 2, \dots\}$ demand met, if possible

X_n: residual stock in warehouse at end of period n.

Assume $D_1, D_2, \dots D_n$ iid

Restock strategy: let $0 \leq m \leq c-1$

- if $X_n > m$ don't do anything
- if $X_n \leq m$ restock to full capacity c



$$X_{n+1} = \begin{cases} C - D_{n+1} & \text{if } 0 \le X_n \le m \\ (X_n - D_{n+1}) & \text{if } m < X_n \le C \end{cases}$$

is not dependent upon C and instead on previous obs. ∴ it is a markov chain

Relevant Questions

① $\lim_{n \to \infty} p_{c_0}^{(n)}$

(prob. of finding warehouse empty at EOD)

② $\lim_{n \to \infty} \sum_j p_{cj}^{(n)}$

③ Long run avg. demand that is not met $\left(\lim_{n \to \infty} \sum_{i=0}^{c} u_i p_{ci}^{(n)} \right)$

$D_1, D_2, \ldots D_n$ iid

$X_n = i \quad \{0, 1, 2, \ldots c\}$

$$u_i = \begin{cases} |D - c| & 0 \le i \le m \\ |D - i| & m \le i \le c \end{cases}$$ where $u_i \in$ Expected unmet demand at end of period $n+1$

Long run behaviour of Markov chains

Consider a markov chain $\{X_n\}$ on $S = \{1, 2, \cdots N\}$

Assume $K \geq 1$ such that $\underline{P}^{(K)}$ has all non-negative entries

ie. $\forall\ i, j \in S$

$$p_{ij}^{(K)} > 0$$

Such a markov chain is called regular.

Proposition: If $\{X_n\}$ is regular, for all $i, j \in S$

$$\lim_{n \to \infty} p_{ij}^{(n)} = \pi_j > 0$$

does not depend on i

Moreover

$$\sum_{j \in S} \pi_j = 1$$

$\pi = (\pi_1, \dots \pi_N)'$ is a probability distribution on S and is the solution of this eigen equation.

$$\boxed{\pi' = \pi' P}$$

ie. π' is a left eigen vector of P corresponding to eigen vector 1.

Interpretation

- π_j is the long term probability of finding the process $\{X_n\}$ in state j.
- π_j is the long run mean fraction of time that the chain spends in state j.

Consider: $E\left[\frac{1}{m}\sum_{n=0}^{m}\pi[X_n=j]\right] = \frac{1}{m}\sum_{n=0}^{m}p_{ij}^{(n)}$

WKT when $m\to\infty$; $p_{ij}^{(n)} \to \pi_j$

$\therefore E\left[\frac{1}{m}\sum_{n=0}^{m}\pi[X_n=j]\right] = \pi_j$

when $m\to\infty$

Example : Inventory problem

$$X_{n+1} = \begin{cases} (C - D_n) & \text{if } 0 \le X_n \le m \\ (X_n - D_n) & \text{if } m < X_n \le C \end{cases}$$

Now if $D \sim D_1, D_2 \cdots D_n$ iid is such that

$$P[D=k] > 0 \text{ for } k = 0, 1, 2, \cdots M \text{ with } M \geq C$$

$$S = \{0, 1, \cdots C\}$$

Let $\pi = (\pi_0, \cdots \pi_C)'$ such that $\pi' = \pi' P$

then

- π_0 (long run freq. of residual stock) is 0
- $\sum_{j \in S} j\pi_j$ (long run average no. of items in stock)

- Expected un-met demand

For $i \in S$

$$u_i = \begin{cases} E[D-C] & \text{if } 0 \le i \le m \\ E[(D-i)] & \text{if } m < i \le C \end{cases}$$

- Long run average of un-met demand

$$u(m) = \sum_{i \in S} u_i \pi_i \quad \text{here } u(m)\,(\downarrow) \text{ as } m\,(\uparrow)$$

Suppose if a is the cost of restocking and b is the profit for each item sold, then

Long run average of restocking

$$g(m) = \sum_{j=0}^{m} \pi_j \quad \text{if } m(\uparrow) \text{ then } g(m)[\uparrow]$$

$$\text{cost}: a \sum_{j=0}^{m} \pi_j + b \cdot \sum_{i \in S} u_i \pi_i$$

Here the optimization problem would be to find m^* such that $m^* = \min \text{cost}(m)$

Exercise: Find the optimal 'm' when $C=3$ (possible values for m are 2, 1, 0) ; $D \in \{0, 1, 2, 3 \cdots\}$ and $P[D=k] = \frac{1}{2^{k+1}}$ $k = 0, 1, \cdots$

Solution:

$$m = \begin{cases} 2 & \text{if } a \leq 1/4 \\ 1 & \text{if } 1/4 < a \leq 1 \\ 0 & \text{if } a \geq 1 \end{cases}$$

Definitions Let $i, j \in S$

j is reachable from i, if $n \geq 1$ such that $P_{ij}^{(n)} > 0$

If j is reachable from i and i is reachable from j } i & j communicate $(i \leftrightarrow j)$

Proposition

$\leftrightarrow$ is an equivalence relation for states in S

ie - reflexive
- transitive
- symmetric

Hence since it is an equivalence relation

$$S = C_0 \cup C_1 \cup C_2 \cdots \cup C_n \cup \ldots$$

C_i equiv classes induced by $\leftrightarrow$

ie $C_i \neq \emptyset \ \forall i$, and $C_i \cap C_j = \emptyset \ \forall i \neq j$

• Starting in one class C_i, the process can move to another class C_j but it cannot come back

In case of Gambler's ruin, classes would be $\{0\}, \{1 \cdots N-1\} \{N\}$

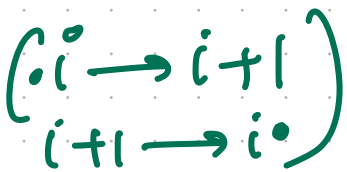
Definition: If $S = C$ (one equivalence class)

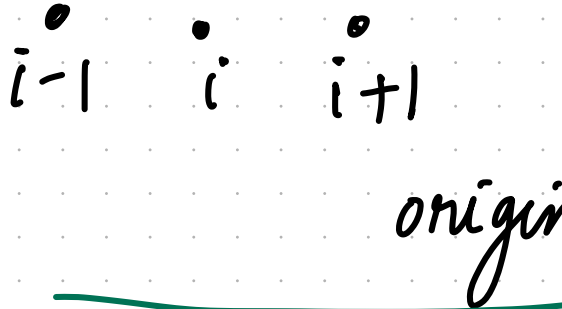
then Markov chain is said to be <u>reducible</u>

Definition: Period of a state $i \in S$

Period is a max. common divisor of n such that $P_{ii}^{(n)} > 0$

Example: In a simple random walk





the only way to come back to the original pos $\Rightarrow$ 2 steps $\therefore$ Period = 2

Proposition: All the states in the same class have the same period.

Definition: If $i \in S$ is said to be recurrent if

$$P[\text{returning to } i \mid X_0 = i] = 1$$

(belongs to a single state)

If $P[\text{returning to } i \mid X_0 = i] < 1$, it is called transient

(In case of transient state, after some runs, the probability of returning back becomes zero)

Proposition: If $i \leftrightarrow j$ then either i & j are recurrent or i & j are transient.

Example: Gambler's ruin

$\{0\}$ $\{1, 2, \cdots N-1\}$ $\{N\}$

$0 \in$ recurrent $\quad \{1, 2, \cdots N-1\} \in$ transient

$N \in$ recurrent (because eventually you either reach 0 or N)

Random walk

If $p = q = 1/2$ then all states are recurrent

$p \neq q$ then states are all transient

Let

$T_i = \inf\{n : X_n = i \mid X_0 = i\}$ time of first return

$E[T_i \mid X_0 = i] = \mu_i$

If i is transient, $\mu_i = \infty$

If i is recurrent, $\mu_i = \infty$ (null recurrent)

$\mu_i < \infty$ (positive recurrent)

* All states in the same class are only one & one of this:

1. Transient
2. positive recurrent
3. Negative recurrent

Theorem: If a Markov chain is reducible and the states are positive recurrent then $\pi = (\pi_0, \pi_1 \cdots)$ distribution on S which is an unique solution of $\pi' = \pi' P$ and $\pi_i = 1/\mu_i$

Here π_i is called the stationary distribution of chain.

Theorem: If a Markov chain is irreducible and aperiodic then

$$\pi_i = \lim_{n \to \infty} p_{ij}^{(n)} = \frac{1}{\mu_i}$$

✯ Markov Decision Processes

we are in a state space S

$\{X_n\}$ is a process such that $X_n \in S$ for $n = 1, 2, 3, \ldots$

and A is an action set such that $A = \{a_1, a_2 \cdots a_m\}$

$q : S \times A \rightarrow$ prob. distribution on S

ie.

$$P[X_{n+1} = j \mid X_n = i, \sigma_n = a] = q_{(S,a)}(j) \quad \text{[Law of motion]}$$

(σ_n: action chosen at time n)

$\sigma_{\text{strategy}} = (\sigma_0, \sigma_1, \sigma_2, \ldots)$ where $\sigma_n : h_n$ (history)

$$h_n = (S_0, a_0), (S_1, a_1), \ldots (S_{n-1}, a_{n-1}), S_n$$

(S_n: state at time $(n+1)$)

$u : (S \times A)(S \times A) \cdots \to R$

(utility)

ie. if $h = (s_0, a_0), (s_1, a_1), \cdots \cdots)$

then $u(h)$ is the gain.

Goal: Find σ^* such that $\max_{\sigma} E[u(h)]$

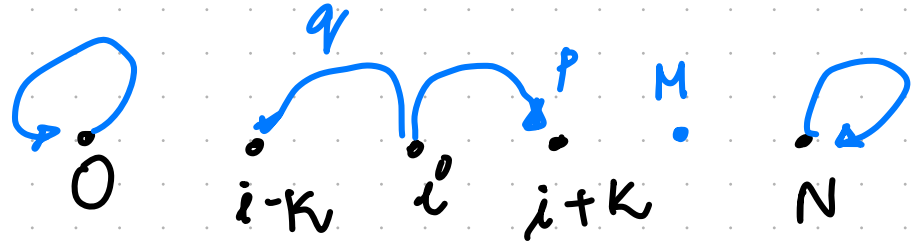
Example:

$$u(h) = \sum_{K=1}^{\infty} \beta^{K-1} u(s_K)$$

$$u : S \to R$$

Example: Extension of Gambler's ruin

→ Red & Black



Bids: (action) $K = X$ euro $\leq i$ if X_n is in $?$

$$u := \begin{cases} 1 & \text{if } h \text{ is s.t } X_n \text{ reached } M \leq N \\ 0 & \text{otherwise} \end{cases}$$

Optimal strategy:

$\Rightarrow$ Bold play: Play everything but do not overshoot the target (maximize probability of reaching M)

★ Hidden Markov Models (HMM)

A process made of 2 random variables $\{(X_n, Y_n)\}_{n=0,1,2\ldots}$

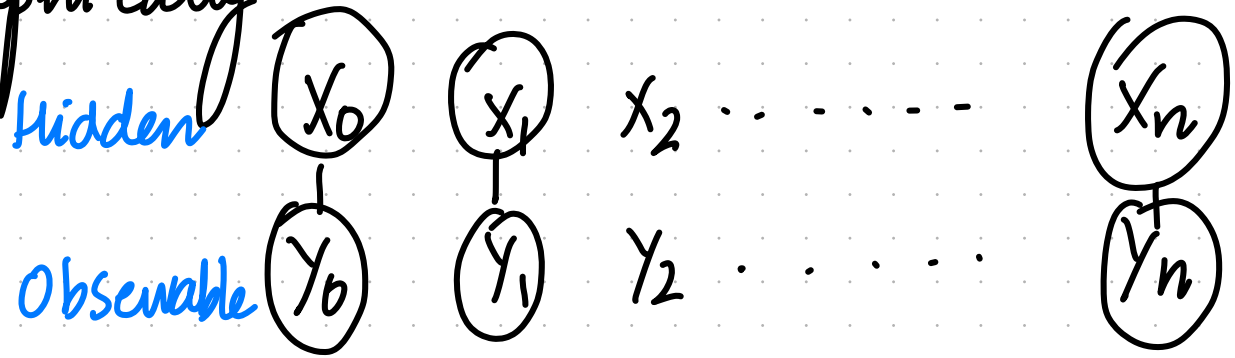
$\{X_n\}$ is a MC on S - state space finite

Conditionally on $\{X_n\}$

$Y_0, Y_1, \dots Y_n$ are independent (random variables with values in $\mathbb{R}$)

and $P(Y_n \in A \mid X_0, X_1 \dots X_{n-1}, \dots) = P(Y_n \in A \mid X_n)$

Graphically



Ingredients:

→ S state space of chain $\{X_n\}$, $S = \{s_1, s_2 \dots s_N\}$

→ finite

→ Observables: finite $\mathcal{O} = \{v_1 \dots v_m\}$ infinite $R: \mathbb{Z}, \mathbb{N}$

- P : transition matrix on S

- Initial distribution : λ is a dist. on S

$$\lambda(i) = P[X_0 = i] \quad i \in S$$

- Emission Matrix: if $\mathcal{O}$ is finite : $b_j(v_k) = P[Y_n = v_k \mid X_n = j]$

if $\mathcal{O}$ is $\mathbb{R}$: b_j is the dist. of $Y_n \mid X_n = j$

Eg: $Y_n \mid X_n = j \sim N(\mu_j, \sigma_j^2)$ is gaussian

we will have a matrix B (N rows $\times$ M col.) if $\mathcal{O}$ is finite

$$B = \begin{bmatrix} b_1(v_1) & b_2(v_2) & \cdots & b_n(v_n) \\ \vdots & & & \\ b_N(v_1) & b_N(v_2) & \cdots & b_N(v_N) \end{bmatrix}$$

- if $\mathcal{O}$ is R & $Y_n \mid X_n = j \sim N(\mu_j, \sigma_j^2)$ then

$$B = \begin{bmatrix} N(\mu_1, \sigma_1^2) \\ N(\mu_2, \sigma_2^2 \\ \vdots \\ N(\mu_n, \sigma_N^2) \end{bmatrix}$$

Example:

We have a signal data: 1001110010 ....

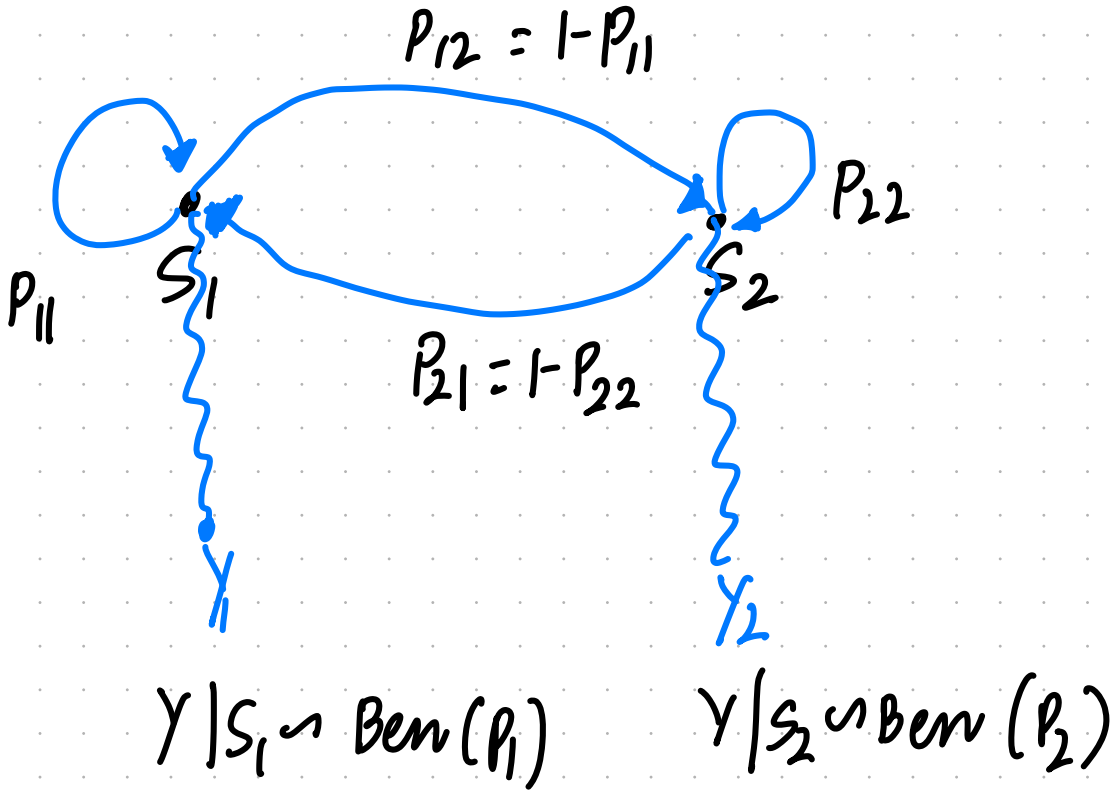
Here X_n iid $\sim$ Bernoulli (P)

this is called HMM_0 (order 0)

If $\lambda = \begin{Bmatrix} S_1 & \pi_1 \\ S_2 & 1-\pi_1 \end{Bmatrix}$ then $Y \mid S_1 \sim$ Bernoulli(P_1) this is HMM_1

$Y \mid S_2 \sim$ Bernoulli (P_2)

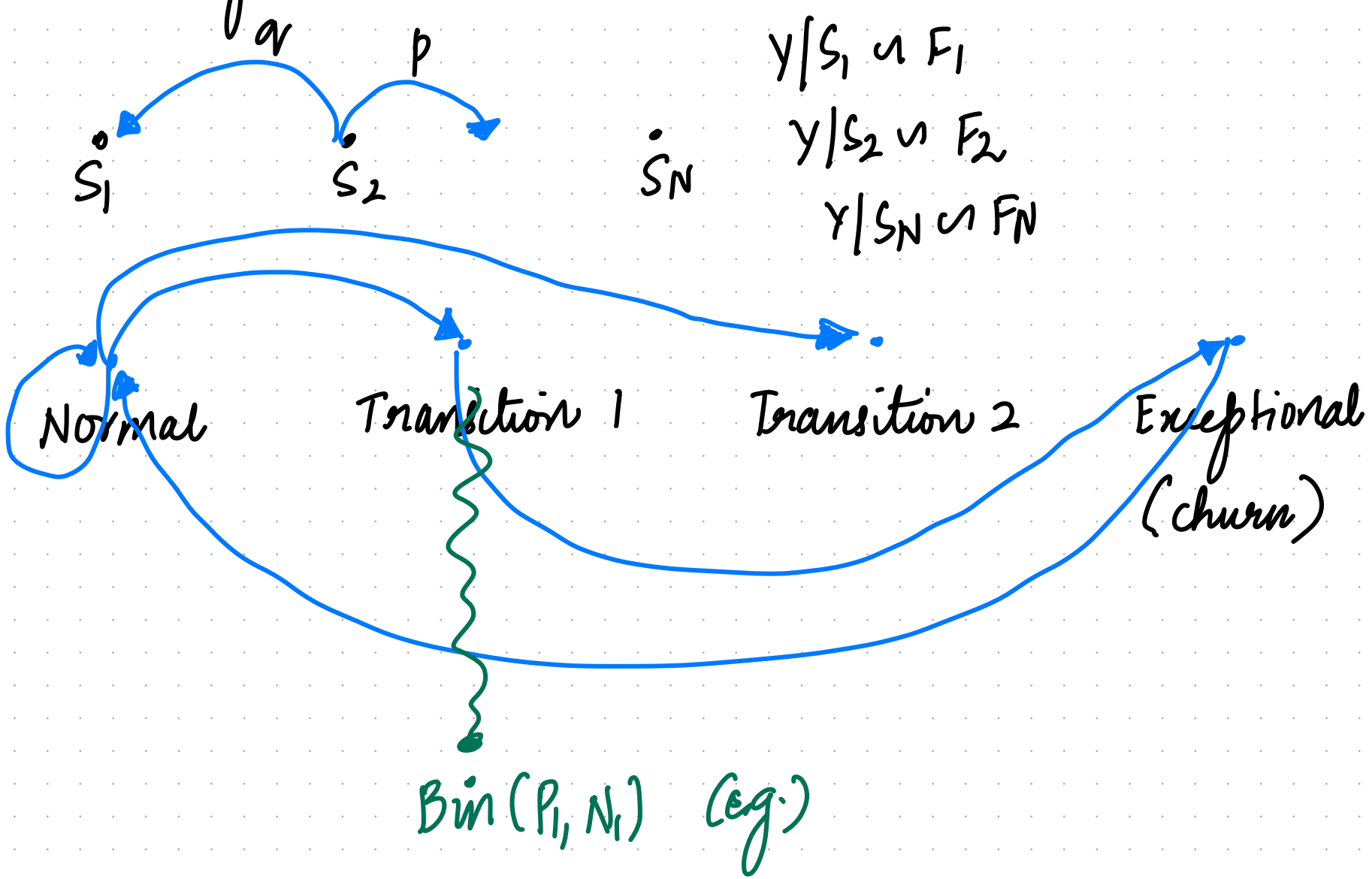
- HMM_2



$Y|S_1 \sim Ber(P_1)$ $Y|S_2 \sim Ber(P_2)$

A general example

We have a state space S



at each state

$y|s_1 \sim F_1$

$y|s_2 \sim F_2$

$y|s_N \sim F_N$

Problems

- Given a model $m = (\lambda, P, B)$

  compute the probability

  $$P(y_0 = y_0, \dots y_T = y_T) \text{ for all } T = 0, 1, 2, \dots$$

  Baum's Algorithm

- Estimate optimally the history of process $\{X_n\}$ upto time T, $T \geq 1$ ie. $X_0 = x_0, \dots X_T = x_T$ (smoothing problem)

  Viterbi's Algorithm

  Given a model, $m = (\lambda, P, B)$

  (Incomplete)

- Estimate m (computationally heavy)

  (Baum-welch Algorithm), estimates of parameters in m through ML.

★ State space Models and Kalman Filters

Linear Models- Interaction terms

Consider the linear model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

$m_1 \leftarrow lm(y \sim x_1 + x_2)$ $\quad [\hat{y} = b_0 + b_1 x_1 + b_2 x_2]$

$m_2 \leftarrow lm(y \sim x_1 * x_2)$ $\quad [\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2]$

AIC & BIC

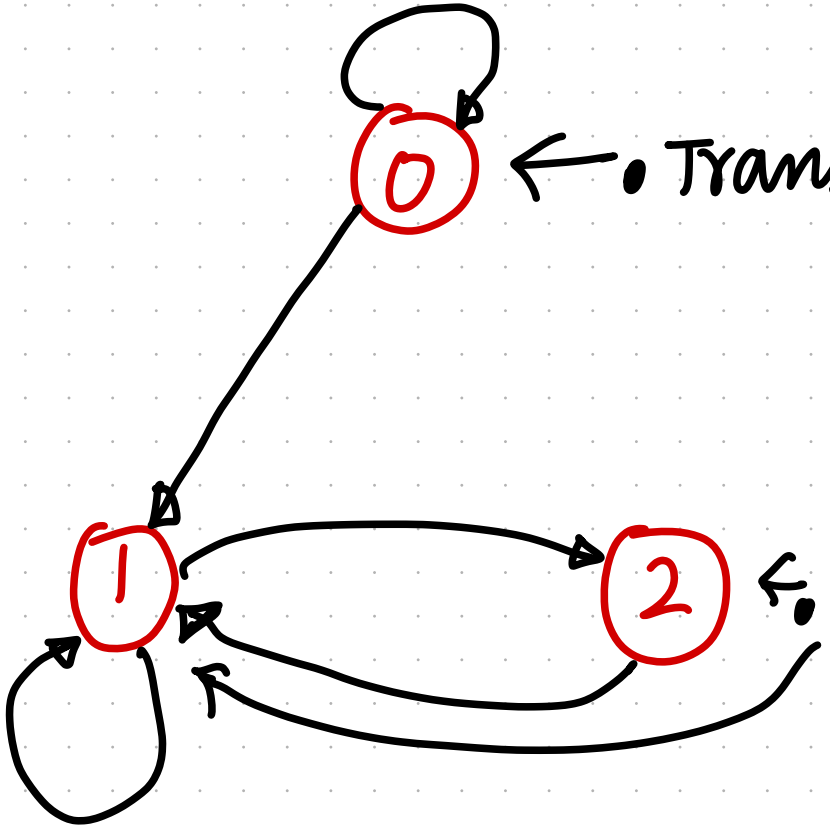
AIC = 2 x no. of parameters − 2 x L-Ratio

BIC = no. of parameters x log(sample size) − 2 x L-ratio

$\log(\text{sample size}) \in [0, \infty)$

If sample size > 1; BIC > AIC

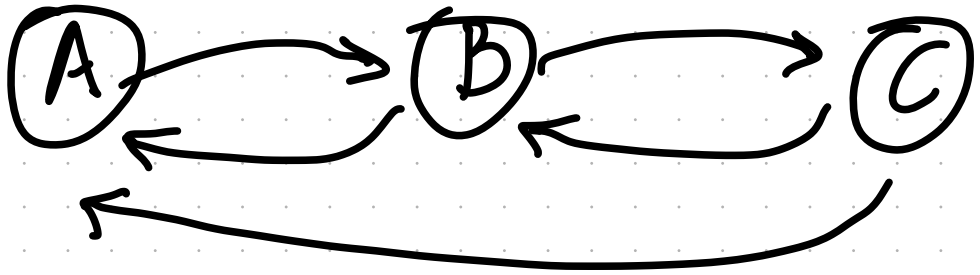
Markov chains & Markov Models



- Transient state: [If left from the state then no chance of coming back]
- Recurrent state: [There is a probability of coming back in a random walk]

- Reducibility: If some states are unreachable from others
- Irreducibility: If all states can be left & returned back
- Absorbing: once entered, cannot be left

- If there is 1 self loop, it makes the whole chain aperiodic. (period = 1)



Possible paths:

$A \to B \to A$ – 2

$A \to B \to C - A$ – (3)

$A \to B \to C - B - A$ – (4)

Period = GCD (2, 3, 4) = 1

If period = 1; (aperiodic)