Biostat 250C

Class Notes

WEEK 1
Lec 1

Biostat 250C

Spring 2021 TA: Mr. Rogers

Introduction to Graphical Modelling by David Edwards (springer)

Intro. to multivariate statistical analysis by T. W. Anderson

Univariate $\leftarrow X \sim F(\cdot)$ $\mathbb{P}(X \in A)$ or $\mathbb{P}(X \le x)$

$\underbrace{X_1, \cdots, X_n}_{\text{Realization from } F(\cdot)} \overset{iid}{\sim} F(\cdot)$

(Ex) $X \sim \mathcal{N}(\mu, \sigma^2)$ $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$

You can not model using iid realizations

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim F(\cdot) \quad \text{(multivariate)}$$

$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n)$ ← How to Construct it ?

Elements of X are dependant ! Model dependency.

$n=2$: $\mathbb{P}(X, Y) = \mathbb{P}(X) \cdot \mathbb{P}(Y|X)$

Notations:

$[X, Y] = [X][Y|X]$

$[X] = \int [X, Y]$; $[Y] = \int [X, Y]$

$[X][Y|X] = [X, Y] = [Y][X|Y]$

$\Rightarrow [X|Y] = \dfrac{[X][Y|X]}{[Y]}$

$\Rightarrow \underbrace{[X|Y] \propto [X][Y|X] = [X, Y]}_{\text{Bayes' Theorem}}$

Q: Suppose I give you $[X|Y]$ & $[Y|X]$

Let us suppose $[X, Y]$ exists.

$\underbrace{\int\int [X, Y] = 1}$

- Is it possible to find $[X, Y]$ from $[X|Y]$ & $[Y|X]$?

$$\frac{[Y|X]}{[X|Y]} = \frac{[Y]}{[X]} \Rightarrow \int \frac{[Y|X]}{[X|Y]} = \frac{1}{[X]}$$

$$\Rightarrow [X] = \left(\int \frac{[Y|X]}{[X|Y]} \right)^{-1}$$

$$\Rightarrow [X, Y] = [X][Y|X]$$

$$= [Y|X] \left(\int \frac{[Y|X]}{[X|Y]} \right)^{-1}$$

What is the key ?

$$\frac{[X]}{[Y]} = \frac{[X|Y]}{[Y|X]} \Rightarrow [X] = \frac{[Y][X|Y]}{[Y|X]}$$

or $[X] \propto \dfrac{[X|Y]}{[Y|X]}$.

I can evaluate $[X]$.

DO NOT NEED INTEGRATION EXPLICITLY!

Conditional Independence

$X \perp Y$ or $X \perp\!\!\!\perp Y$ iff

$P(X,Y) = P(X) \cdot P(Y)$ or

$[X,Y] = [X] \cdot [Y]$

$\Leftrightarrow \underbrace{\frac{[X \cdot Y]}{[Y]}} = [X]$

$\Leftrightarrow [X|Y] = [X]$

$\Leftrightarrow [Y|X] = [Y]$

Intro. a 3rd variable Z

$X \perp\!\!\!\perp Y \mid Z \overset{\text{def}}{\Leftrightarrow} [X,Y|Z] = [X|Z][Y|Z]$

or $P(X,Y|Z) = P(X|Z) \cdot P(Y|Z)$

what about $[X|Y,Z]$?

$[X|Y,Z] = [X|Z]$

Because

$[X,Y|Z] = [Y|Z]\,[X|Y,Z]$

$\xrightarrow{\text{C.I.}} = [Y|Z]\,[X|Z]$

Decomposition property of Cond. Indept.

$X \perp\!\!\!\perp (Y,W) \mid Z \Rightarrow$

$X \perp\!\!\!\perp Y \mid Z \wedge X \perp\!\!\!\perp W \mid Z$

Proof:

$[X,Y,W|Z] =$

$\overset{\text{def}}{=} [X|Z]\,[Y|X,Z]\,[W|X,Y,Z]$

$= [X|Z]\,[Y,W|Z]$

Integrate / marginalize out W from both sides.

LHS: $\int [X,Y,W|Z] = [X,Y|Z]$

RHS: $\int [X|Z]\,[Y,W|Z] = [X|Z]\,[Y|Z]$

Done!

HW: Given $X \perp\!\!\!\perp (Y,W) \mid Z$, prove $X \perp\!\!\!\perp W \mid Z$ using factorization & integration.

HW: $X \perp\!\!\!\perp (Y,W) \mid Z$, prove $X \perp\!\!\!\perp Y \mid W,Z$

HW: Contraction.

$(X \perp\!\!\!\perp W \mid Z,Y) \wedge (X \perp\!\!\!\perp Y \mid Z) \Rightarrow$

$X \perp\!\!\!\perp (Y,W) \mid Z$

Important HW rule:

Any HW given in a week is due on or before the beginning of class of the Wed. of the following week.

Comment: If $[X,Y]$ exists $[X|Y]$ & $[Y|X]$ defined uniquely

But what if $[X,Y]$ does NOT $\exists$?

$p(x,y) \propto e^{-\frac{1}{2}(x+y)^2}$

$\iint p(x,y)\,dx\,dy = +\infty$.

$[X|Y] \sim \mathcal{N}(y,1)$, $[Y|X] \sim \mathcal{N}(x,1)$

WEEK 5 Lec 9 04-26

Determinant version of SWM

Recall

① $\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & D \end{bmatrix}$

② ← ② + X①

$[C:D] = [C:D] + X[A:B]$

Row(i) of $[C:D]$ ←

Row(i) of $[C:D]$ + row(i) of X $[A:B]$

- $C = AB$ (Digression)

row(i) of C = (row i of A) · B

Matrix multiplication:

$C = [C_{*1} : C_{*2} : \cdots : C_{*p}]$ ← column p of C

$= \begin{bmatrix} C_{1*}^T \\ C_{2*}^T \\ \vdots \\ C_{m*}^T \end{bmatrix}$ ← row 2 of C

$C^T = [C_{1*} \; C_{2*} \cdots C_{m*}] = \begin{bmatrix} C_{*1}^T \\ \vdots \\ C_{*p}^T \end{bmatrix}$

$C = AB \Rightarrow$

$C_{ij} = (\text{row } i \text{ of } A) \cdot (\text{col } j \text{ of } B)$

$= a_{i*}^T b_{*j}$

$C_{i*}^T = a_{i*}^T [b_{*1}; b_{*2}; \cdots; b_{*p}]$

$= a_{i*}^T B$

$C_{*j} = A b_{*j}$ Similarly.

- $\begin{bmatrix} a_{1*}^T x_{*j} \\ \vdots \\ a_{p*}^T x_{*j} \end{bmatrix} = \begin{bmatrix} a_{1*}^T \\ \vdots \\ a_{p*}^T \end{bmatrix} x_{*j}$ NEED A PROOF!

(Another Digression)

$x^T A = (x_1, \cdots, x_m)\begin{pmatrix} a_{1*}^T \\ \vdots \\ a_{m*}^T \end{pmatrix}$

$1\times m \quad m\times n$

$= \sum_{i=1}^{m} x_i (\text{row } i \text{ of } A)$

$x^T A$: linear comb. of rows of A

Ax : linear comb. of cols of A.

Back to Block Elimination.

$\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ XA+C & XB+D \end{bmatrix}$

$[XA + C : XB + D] = X[A;B] + [C;D]$

So As a transformation,

$[C:D] \leftarrow [C:D] + X[A:B]$

row (i) of $[C:D]$ ←

row (i) of $[C:D]$ + (row (i) of X) $[A:B]$

linear comb. of rows of $[A:B]$

$\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}$: execute a series of type-I row operations as a block.

$\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}$: execute a series of type-I row operations as a block.

- Type I op:

$\text{row}(i) \leftarrow \text{row}(i) + \alpha\,\text{row}(j)$

$\text{row}(i) \leftarrow \text{row}(i) + \sum_{k=1}^{r} \alpha_k \text{row}(j_k)$

This is a series of elem. row-ops.

$\text{row}(i) \leftarrow \text{row}(i) + \alpha_1 \text{row}(j_1)$

$\text{row}(i) \leftarrow \text{row}(i) + \alpha_2 \text{row}(j_2)$

& so on.

- For (k in 1:r) {

  row(i) ← row(i) + α_k row(j_k)

  }

- $\begin{bmatrix} A & B \\ C & D \end{bmatrix} \to \begin{bmatrix} \alpha A & \alpha B \\ \beta C & \beta D \end{bmatrix}$

Is this a seq./block of elementary ops? If so, how/why?

Yes: $\begin{bmatrix} \alpha I & 0 \\ 0 & \beta I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix}$

$\begin{bmatrix} I & 0 \\ 0 & \beta I \end{bmatrix} \begin{bmatrix} \alpha I & 0 \\ 0 & I \end{bmatrix}$

- Back to Determinant SWM.

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & D-CA^{-1}B \end{bmatrix}$$

Block elementary operation

Recall elementary row operation of the form

$\text{row}(i) \leftarrow \text{row}(i) + \alpha\,\text{row}(j)$

$\Rightarrow$

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det \begin{pmatrix} A & B \\ 0 & D-CA^{-1}B \end{pmatrix}$$

(*) Most mature def (algebraic)

$$\det_{n\times n}(A) = \sum_{\pi} (-1)^{\text{sign}(\pi)} a_{1\pi_1} a_{2\pi_2} \cdots a_{n\pi_n}$$

(Ex) $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$

π	term	sign
(1,2)	$a_{11}a_{22}$	+1
(2,1)	$a_{12}a_{21}$	-1

- $a_{11}a_{22} - a_{12}a_{21} = \det|A|$

$$\Rightarrow \det \begin{pmatrix} A & B \\ C & 0 \end{pmatrix} = \det(A)\det(D-CA^{-1}B)$$

Let B = 0:

$$\begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A-BD^{-1}C & 0 \\ C & D \end{bmatrix}$$

block elem. ops. mat

$$\Rightarrow \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(D)\det(A-BD^{-1}C)$$

- $|A||D-CA^{-1}B| = |D||A-BD^{-1}C|$

HW WEEK 5

① Let A be 3×3. Fill out a table of all permutations of $(1,2,3)$

π	Term	Sign

Find $\det(A)$.

② A is $p\times p$, D is $n\times n$

$|A|\neq0$, $|D|\neq0$. Prove:

$$\det\begin{pmatrix} A & B \\ 0 & D \end{pmatrix} = \det A \det D.$$

③ what is the SWM form for

$|U+VWX|$, where U is $n\times n$, W is $p\times p$, V is $n\times p$, X is $p\times n$ & $|U|\neq0, |W|\neq0$

Application of SWM.

Linear mixed models.

$$\underset{n\times1}{Y} = \underset{n\times p}{X}\underset{p\times1}{\beta} + \underset{n\times r}{Z}\underset{r\times1}{u} + \underset{n\times1}{\epsilon}$$

$\epsilon \sim \mathcal{N}(0, D)$, $u \sim \mathcal{N}(0, D)$

Unknown: β, u & maybe para in D & R

Posterior: $\mathcal{N}(y|X\beta+Zu, D) \times$
$\mathcal{N}(u|0,R) \times$
$\mathcal{N}(\beta|0,V_\beta)$, $V_\beta^{-1}\to0$.
$\mathcal{N}(\beta|0,V_\beta) \to c$. (improper)

So posterior (or likelihood)

$$\mathcal{N}(u|0,R)\times\mathcal{N}(y|X\beta+Zu,D)$$

Dim of unknown {u & β}:

Integrate out the "u".

Stats: marginalize.

Machine Learning: Collapsing.

$$\int \mathcal{N}(u|0,R)\times\mathcal{N}(y|X\beta+Zu,D)\,du.$$

WEEK 5 Lec10 04-28

Sequential Bayesian Learning

Idea: Single data point $Y_1 \sim p(y|\theta)$

$\theta \sim p(\theta)$

Posterior: $p(\theta|Y_1) \propto p(\theta)\, p(Y_1|\theta)$

Second data point: $Y_2 \sim p(y|\theta)$

$Y_2 \perp Y_1 | \theta$

Update:

$$p(\theta|Y_1,Y_2) \propto p(\theta)\, p(Y_1,Y_2|\theta)$$
$$\propto \underbrace{p(\theta)\, p(Y_1|\theta)}_{\propto p(\theta|Y_1)}\, p(Y_2|\theta)$$
$$\propto p(\theta|Y_1)\, p(Y_2|\theta)$$

Prior on $\theta \xrightarrow{\text{Updated}} p(\theta|Y_1) \xrightarrow{\text{Updated}} p(\theta|Y_1,Y_2)$

$\rightarrow \cdots \cdots \xrightarrow{\text{Updated}} p(\theta|Y_1, Y_2, \cdots, Y_n)$

Ex $Y_i|\theta \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$, σ^2 known

$i = 1, \cdots, n$, n: sample size, n_0: prior sample size

$\theta \sim \mathcal{N}(\mu, \frac{\sigma^2}{n_0})$

n = Sample size

n_0 = "prior sample size"

(larger $n_0 \Rightarrow$ more precise priors)

$$p(\theta|Y_1) \propto \mathcal{N}(\theta|\mu, \tfrac{\sigma^2}{n_0}) \times \mathcal{N}(Y_1|\theta,\sigma^2)$$
$$\propto e^{-\frac{n_0}{2\sigma^2}(\theta-\mu)^2} \times e^{-\frac{1}{2\sigma^2}(Y_1-\theta)^2}$$
$$\propto e^{-\frac{1}{2\sigma^2}\{n_0(\theta-\mu)^2 + (Y_1-\theta)^2\}}$$

$$n_0(\theta-\mu)^2 + (Y_1-\theta)^2 = (n_0+1)\theta^2 - 2(\mu n_0 + Y_1)\theta + Y_1^2 + n_0\mu^2$$
$$= (n_0+1)\left\{\theta - \frac{n_0\mu + Y_1}{n_0+1}\right\}^2 + \text{Const.}$$

$$\therefore p(\theta|Y_1) = \mathcal{N}\left(\theta \middle| \frac{n_0\mu + Y_1}{n_0+1}, \frac{\sigma^2}{n_0+1}\right)$$

mean is weighted average:

$$\underbrace{\frac{n_0}{n_0+1}}_{\text{prior weight}}\mu + \underbrace{\frac{1}{n_0+1}}_{\text{data weight}} Y_1$$

Precision enhanced: $\frac{\sigma^2}{n_0} \rightarrow \frac{\sigma^2}{n_0+1}$

$$p(\theta|Y_1,Y_2) \propto p(\theta|Y_1)\, p(Y_2|\theta)$$
$$\propto e^{-\frac{(n_0+1)}{2\sigma^2}(\theta-\theta_*)^2} \times e^{-\frac{1}{2\sigma^2}(Y_2-\theta)^2}$$

$\theta_* = \frac{n_0\mu + Y_1}{n_0+1} \Rightarrow$

$$p(\theta|Y_1,Y_2) = \mathcal{N}\left(\theta \middle| \underbrace{\frac{(n_0+1)\theta_* + Y_2}{(n_0+1)+1}}_{(\Delta)}, \frac{\sigma^2}{(n_0+1)+1}\right)$$

$$(n_0+1)\theta_* + Y_2 = (n_0+1)\left\{\frac{n_0\mu + Y_1}{n_0+1}\right\} + Y_2 = n_0\mu + (Y_1+Y_2)$$

$$(\Delta) = \frac{n_0\mu + 2\bar{Y}}{n_0+2}$$

Carry on:

$$p(\theta|Y_1,Y_2) = \mathcal{N}\left(\theta \middle| \frac{n_0\mu + 2\bar{Y}}{n_0+2}, \frac{\sigma^2}{n_0+2}\right)$$

$\cdots\cdots$

$$p(\theta|Y_1,\cdots,Y_{n-1}) = \mathcal{N}\left(\theta \middle| \underbrace{\frac{n_0\mu + (n-1)\bar{Y}_{n-1}}{n_0+(n-1)}}_{\theta_*}, \frac{\sigma^2}{n_0+n-1}\right)$$

$$p(\theta|Y_1,\cdots,Y_n) \propto \mathcal{N}(\theta|\theta_*, \tfrac{\sigma^2}{n_0+n-1}) \times \mathcal{N}(Y_n|\theta,\sigma^2)$$
$$= \mathcal{N}\left(\theta \middle| \frac{(n_0+n-1)\theta_* + Y_n}{(n_0+n-1)+1}, \frac{\sigma^2}{n_0+n}\right)$$

$$\theta_* = \frac{n_0\mu + (n-1)\bar{y}_{n-1}}{n_0+n-1} \Rightarrow$$

$$(n_0+n-1)\theta_* = n_0\mu + (n-1)\bar{y}_{n-1} + y_n$$

$$\Rightarrow p(\theta|y_1,\dots,y_n) =$$

$$\mathcal{N}\left(\theta \,\middle|\, \frac{n_0\mu + \sum_{i=1}^{n-1} y_i + y_n}{n_0+n}, \frac{\sigma^2}{n_0+n}\right)$$

$$= \mathcal{N}\left(\theta \,\middle|\, \frac{n_0\mu + n\bar{y}}{n_0+n}, \frac{\sigma^2}{n_0+n}\right)$$

- **Much simpler to use** Sufficient statistic (Le2AB)!

$$y_1,\dots,y_n|\theta \overset{iid}{\sim} \mathcal{N}(\theta,\sigma^2)$$

$$\Rightarrow \bar{y} \sim \mathcal{N}(\theta, \tfrac{\sigma^2}{n})$$

$$p(\theta|y_1,\dots,y_n) = p(\theta|\bar{y})$$

$$\propto \mathcal{N}(\theta|\mu, \tfrac{\sigma^2}{n_0}) \times \mathcal{N}(\bar{y}|\theta, \tfrac{\sigma^2}{n})$$

$$= \mathcal{N}\left(\theta \,\middle|\, \frac{n_0\mu + n\bar{y}}{n_0+n}, \frac{\sigma^2}{n_0+n}\right)$$

Return to HW:

$$\begin{cases} y = X\beta + e_y; & e_y \sim \mathcal{N}(0,\sigma^2 I_n) \\ \beta = \mu_\beta + e_\beta; & e_\beta \sim \mathcal{N}(0,\sigma^2 V_\beta) \\ \tilde{y} = \tilde{X}\beta + \tilde{e}; & \tilde{e} \sim \mathcal{N}(0,\sigma^2 I_m) \end{cases}$$

- Assume σ^2 known (since we want to find $p(\tilde{y}|\sigma^2, y)$)

- $p(\beta,\tilde{y}|\sigma^2,y) \propto p(\beta|\sigma^2,y) \times p(\tilde{y}|\beta,\sigma^2)$

Since $\tilde{y} \perp y|\beta,\sigma^2$

$$\tilde{y} = \tilde{X}\beta + \tilde{e}; \quad \tilde{e} \sim \mathcal{N}(0,\sigma^2 I_m)$$

- β's distribution has been updated to

$$\beta|y \sim \mathcal{N}(M m, \sigma^2 M)$$

$$-M^{-1} = V_\beta^{-1} + X^T X$$

$$-m = V_\beta^{-1}\mu_\beta + X^T y$$

- Writing as a linear model

$$\beta = Mm + e_{\beta|y}, e_{\beta|y} \sim \mathcal{N}(0,\sigma^2 M)$$

To find $p(\tilde{y}|\sigma^2,y)$ we substitute the LM for β (posterior) into the LM for $\tilde{y}$:

$$\tilde{y} = \tilde{X}(Mm + e_{\beta|y}) + \tilde{e}$$

$$= \tilde{X}Mm + \underbrace{\tilde{X}e_{\beta|y} + \tilde{e}}_{e_*}$$

$$= \tilde{X}Mm + e_*$$

$$e_* \sim \mathcal{N}(0, \sigma^2(I_m + \tilde{X}M\tilde{X}^T))$$

$$\Rightarrow \tilde{y}|y,\sigma^2 \sim$$

$$\mathcal{N}(\tilde{X}Mm, \sigma^2(I_m + \tilde{X}M\tilde{X}^T))$$

Application of SWM to Linear Mixed Models

$$\underset{n\times 1}{y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times r}{Z}\underset{r\times 1}{u} + \underset{n\times 1}{e};$$

$$e \sim \mathcal{N}(0, D)$$

$$u \sim \mathcal{N}(0, R)$$

Dimension Reduction: Choose a design matrix $Z_{n\times r}$ & $r \ll n$ Not: R is $r\times r$

Collapse: $y \sim \mathcal{N}(X\beta, ZRZ^T + D)$

$$y|\beta \sim \mathcal{N}(X\beta, ZRZ^T + D)$$

Compute likelihood: to sol mle of β

$(\underbrace{ZRZ^T + D}_{n\times n})^{-1}$ is required

What if (n) is BIG?

- SWM to the rescue:

$$(D + ZRZ^T)^{-1} =$$

$$D^{-1} - D^{-1}Z(R^{-1} + Z^TD^{-1}Z)^{-1}Z^TD^{-1}$$

Advantage: D is diagonal/sparse s.t. D^{-1} is cheap.

- Z is $n\times r$ with $r \ll n$. Modeler chooses r & Z. R is $r\times r$, R^{-1} is cheap.

Prediction when $\tilde{Y} \not\perp y|\beta, \sigma^2$

$$\begin{bmatrix} y \\ \tilde{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} X\beta \\ \tilde{X}\beta \end{bmatrix}, \sigma^2\begin{bmatrix} V_y & V_{y\tilde{Y}} \\ V_{\tilde{Y}y} & V_{\tilde{Y}} \end{bmatrix}\right)$$

- $p(\tilde{y}|y, \sigma^2, \beta)$

$$= \mathcal{N}(\tilde{X}\beta + V_{\tilde{Y}y}V_y^{-1}(y - X\beta), \sigma^2(V_{\tilde{Y}} - V_{\tilde{Y}y}V_y^{-1}V_{y\tilde{Y}}))$$

- $p(\beta|y, \sigma^2) = \mathcal{N}(\beta|M m, \sigma^2 M)$
- $p(\sigma^2|y) = IG(\sigma^2|a^*, b^*)$

- Stat: multivariate Bayesian stat.
- ML: probabilistic machine learning.

WEEK6 Lec 11 May 03.2024

Recall: $\mathcal{N}(\beta|\mu_\beta, \sigma^2 V_\beta) \times \mathcal{N}(y|X\beta, \sigma^2 V_y)$

$$p(y|\sigma^2) = \int \underbrace{p(\beta, y|\sigma^2)}_{p(\beta|\sigma^2)p(y|\beta,\sigma^2)} d\beta$$

$$= \int \mathcal{N}(\beta|\mu_\beta, \sigma^2 V_\beta) \times \mathcal{N}(y|X\beta, \sigma^2 V_y)\, d\beta$$

Easy to avoid integration:

$$[y|\beta,\sigma^2] \rightarrow y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 V_y)$$

$$[\beta|\sigma^2] \rightarrow \beta = \mu_\beta + \epsilon_\beta;\ \epsilon_\beta \sim \mathcal{N}(0, \sigma^2 V_\beta)$$

$$\Rightarrow [y|\sigma^2] \rightarrow y = X(\mu_\beta + \epsilon_\beta) + \epsilon$$

$$= X\mu_\beta + \underbrace{X\epsilon_\beta + \epsilon}_{\mathcal{N}(0,\ \sigma^2(XV_\beta X^T + V_y))}$$

[75-year-old prof.]

Similar for predictions:

$$p(\tilde{y}|y,\sigma^2) = \int p(\tilde{y}|\beta,\sigma^2,y)\cdot p(\beta|\sigma^2,y)\, d\beta$$

$$[\tilde{y}|\beta,\sigma^2,y] \rightarrow \tilde{y} = \tilde{x}\beta + \tilde{e};$$

assume $\tilde{y} \perp\!\!\!\perp y\,|\,\beta,\sigma^2$ $\quad \tilde{e} \sim \mathcal{N}(0, \sigma^2 \tilde{V})$

$$[\beta|\sigma^2, y] \rightarrow \beta = M m + e_\beta;$$

$$e_\beta \sim \mathcal{N}(0, \sigma^2 M)$$

$$[\tilde{y}|\sigma^2, y] \rightarrow \tilde{y} = \tilde{x}(Mm + e_\beta) + \tilde{e}$$

$$= \tilde{x}Mm + \underbrace{\tilde{x}e_\beta + \tilde{e}}_{\mathcal{N}(0,\ \sigma^2(\tilde{x}M\tilde{x}^T + \tilde{V}))}$$

Q Why $e \perp\!\!\!\perp e_\beta$?

Key point: Look @ the integration.

Always remember:

$$[\text{Unknown}|\text{known}] \propto [\text{unknown}, \text{known}]$$

$$[\beta, \sigma^2, \tilde{y}|y] \propto [\beta, \sigma^2, \tilde{y}, y]$$

$$\propto \underbrace{[\sigma^2][\beta|\sigma^2][\tilde{y}|\beta,\sigma^2][y|\beta,\sigma^2]}_{\propto [\beta,\sigma^2|y]} \times$$

$$\propto [\beta,\sigma^2|y][\tilde{y}|\beta,\sigma^2]$$

Prediction vs. Model-fitting

$$y_i|\beta,\sigma^2 \overset{ind}{\sim} \mathcal{N}(x_i^T\beta, \sigma^2)$$

$$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$$

$$\sigma^2 \sim IG(a,b)$$

$$(\tilde{y}, \tilde{x}) \rightarrow \tilde{y}|\beta,\sigma^2 \sim \mathcal{N}(\tilde{x}^T\beta, \sigma^2)$$

$$\Rightarrow y = X\beta + \epsilon;\ \beta = \mu_\beta + \epsilon_\beta;$$

$$\tilde{y} = \tilde{x}\beta + \tilde{\epsilon}$$

$$\tilde{y}|y,\sigma^2 \sim \mathcal{N}(\tilde{x}M m,\ \sigma^2(I_m + \underbrace{\tilde{x}M\tilde{x}^T}_{m\times p\ p\times p\ p\times m}))$$

Prediction: New (but given) $\tilde{x}$

Predict $\tilde{y}$ for that new $\tilde{x}$.

Model-fitting: $\tilde{y}$ but with X, not $\tilde{x}$!

$\Rightarrow$ replicate

$$\tilde{y}|y,\sigma^2 \sim \mathcal{N}(XMm, \sigma^2(I_m + XMX^T))$$

Beneath the above derivation, we still assume $\tilde{y} \perp\!\!\!\perp y \mid \beta, \sigma^2$

$p(\tilde{y} \mid \sigma^2, y)$ ← model's ability to replicate the data y!

Ex $V_\beta^{-1} = 0$, $a = -\frac{p}{2}$, $b = 0$

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\right)^{a+\frac{p}{2}+1}$$

Then $\left\{ \begin{array}{l} M = (X^TX)^{-1} \\ m = X^Ty \end{array} \right\} \rightarrow Mm = \hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty$

model-fitting

$$\tilde{y} \mid y, \sigma^2 \sim \mathcal{N}(X\hat{\beta}_{OLS}, \sigma^2(I_m + P_X))$$

Prediction

$$\tilde{y} \mid y, \sigma^2 \sim \mathcal{N}(\tilde{X}\hat{\beta}_{OLS}, \sigma^2(I_m + \tilde{X}(X^TX)^{-1}\tilde{X}^T))$$

To compare models:

most model comparison matrices use:

$p(\tilde{y} \mid y)$ → pointwise

We want to investigate behavior or performance in terms of

$p(\tilde{y}_i \mid y)$ for $i = 1, 2, \cdots, n$

Ex $\tilde{y} \mid y, \sigma^2 \sim \mathcal{N}(X\hat{\beta}_{OLS}, \sigma^2(I + P_X))$

$\tilde{y}_i \mid y, \sigma^2 \sim \mathcal{N}(x_i^T\hat{\beta}_{OLS}, \sigma^2(I + P_X)_{ii})$

$P_X = X(X^TX)^{-1}X^T$; $[P_X]_{ii} = x_i^T(X^TX)^{-1}x_i$

where

x_i = row i of X or col i of X^T

⇒ Assume replicates are indep. across i:

$$y_i \mid \sigma^2, y \sim \mathcal{N}(x_i^T\hat{\beta}_{OLS}, \sigma^2(1 + x_i^T(X^TX)^{-1}x_i))$$

Common model comparison metric:

$\sum_{i=1}^{n} \log p(\tilde{y}_i \mid y)$ → leads to WAIC (after penalty) — widely

General:

$$p(\tilde{y}_i \mid y) = \int p(\tilde{y}_i \mid \theta) p(\theta \mid y)\, d\theta$$

$$= \mathbb{E}_{\theta \mid y}[p(\tilde{y}_i \mid \theta)]$$

$$= \int p(\tilde{y}_i \mid \theta) \underbrace{P_{post}(\theta)}_{p(\theta \mid y)}\, d\theta$$

$$P_{post}(\tilde{y}_i) = p(\tilde{y}_i \mid y)$$

Suppose we have generated samples

$\tilde{y}_i^{(1)}, \cdots, \tilde{y}_i^{(n)} \sim p(\tilde{y}_i \mid y)$

$$\tilde{\mu}_i = \mathbb{E}(\tilde{y}_i \mid y) \approx \frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}} \tilde{y}_{(i)}$$

Goodness of fit measure:

$$G = \sum_{i=1}^{n} (y_i - \tilde{\mu}_i)^2$$

Also need to penalize models.

$$P = \sum_{i=1}^{n} Var(\tilde{Y}_i | Y),$$

$$Var(\tilde{Y}_i | Y) = \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} (\tilde{Y}_i^{(j)} - \tilde{\mu}_i)^2$$

model comparison score:

$$D = \underbrace{G}_{\text{goodness of fit}} + \underbrace{P}_{\text{penalty}}.$$

Marginalize σ^2

$$\tilde{Y} | Y, \sigma^2 \sim \mathcal{N}(\tilde{X}\mu_m, \sigma^2(I_m + \tilde{X}V\tilde{X}^T))$$

$$[\tilde{Y}|Y] = \int [\tilde{Y}|Y,\sigma^2][\sigma^2|Y]$$

$$= \int \mathcal{N}(\tilde{Y}|\tilde{X}\mu_m, \sigma^2(I_m+\tilde{X}V\tilde{X}^T)) \times IG(\sigma^2|a_*, b_*)\, d\sigma^2$$

$$= \int \mathcal{N}IG(\tilde{Y}, \sigma^2 | \underbrace{\tilde{X}\mu_m}_{\mu}, \underbrace{I_m+\tilde{X}V\tilde{X}^T}_{V}, \underbrace{a_*}_{a}, \underbrace{b_*}_{b})\, d\sigma^2$$

↓

Multivariate (non-central) student's t distribution

Another look @ normal densities & Linear models.

$$\underset{n\times 1}{y} \sim \mathcal{N}(\mu, \Sigma)$$

$$y = (Y_1, Y_2, \cdots, Y_n)^T$$

- $p(y) = p(y_1)\, p(y_2|y_1)\, p(y_3|y_1,y_2) \cdots p(y_n|y_1,\cdots,y_{n-1})$

$[Y_1] \rightarrow Y_1 = \mu_1 + \epsilon_1;\quad \epsilon_1 \sim \mathcal{N}(0, d_1)$

$[Y_2|Y_1] \rightarrow Y_2 - \mu_2 = a_{21}(Y_1-\mu_1) + \epsilon_2;$

$\epsilon_2 \sim \mathcal{N}(0, d_2)$

$[Y_3|Y_2,Y_1] \rightarrow Y_3 - \mu_3 = a_{31}(Y_1-\mu_1) +$

$a_{32}(Y_2-\mu_2) + \epsilon_3;$

$\epsilon_3 \sim \mathcal{N}(0, d_3)$

$\vdots$

$[Y_i | Y_{i-1}, \cdots, Y_1] \rightarrow$

$$y_i - \mu_i = \sum_{j=1}^{i-1} a_{ij}(y_j - \mu_j) + \epsilon_i;\ \epsilon_i \sim \mathcal{N}(0, d_i)$$

$i = 2, \cdots, n$

$\Rightarrow$

$$\begin{bmatrix} Y_1-\mu_1 \\ Y_2-\mu_2 \\ \vdots \\ Y_n-\mu_n \end{bmatrix} =$$

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & \cdots & 0 & 0 \\ a_{31} & \ddots & \ddots & \ddots & \vdots \\ a_{(n-1)1} & \cdots & a_{(n-1)(n-2)} & 0 & \\ a_{n1} & a_{n2} & \cdots & a_{n(n-1)} & 0 \end{bmatrix} \begin{bmatrix} Y_1-\mu_1 \\ Y_2-\mu_2 \\ \vdots \\ Y_n-\mu_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- $y - \mu = A(y-\mu) + \epsilon$

$\epsilon \sim \mathcal{N}(0, D),\ D = Diag(d_i)$

$\Rightarrow (I-A)(Y-\mu) = \epsilon$

$(I-A)x = 0 \Rightarrow$

$$\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ a_{n1} & & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow x_1 = x_2 = \cdots = x_n = 0!$$

$\Rightarrow I-A$ is non-singular!

$\Rightarrow (I-A)^{-1}$ exists!

So $(I-A)(Y-\mu) = \epsilon$;

$(I-A)\,\mathrm{Var}(Y-\mu)(I-A)^T = \mathrm{Var}(\epsilon)$

or

$(I-A)\,\Sigma\,(I-A)^T = D$

$\Rightarrow$

$$\Sigma = \underbrace{(I-A)^{-1} D (I-A)^{-T}}_{\text{Cholesky}}$$

- Notation:

$$(I-A)^{-T} = \left[(I-A)^{-1}\right]^T = \left[(I-A)^T\right]^{-1}$$
$$= \left[\underbrace{(I-A^T)}_{\text{upper triangular}}\right]^{-1}$$

$$\Sigma^{-1} = (I-A)^T D^{-1} (I-A)$$

Cholesky: $\Sigma = L D L^T$

L is unit lower triangular, D has diagonal element.

$D = D^{\frac{1}{2}} D^{\frac{1}{2}}$;

$$L D L^T = L D^{\frac{1}{2}} D^{\frac{1}{2}} L^T$$
$$= \underbrace{(L D^{\frac{1}{2}})}_{\tilde{L}} \underbrace{(L D^{\frac{1}{2}})^T}_{\tilde{L}^T}$$

We have derived the Cholesky decomp. using a probability factorization.

Let K be a p.d. matrix.

Spectral decomp. $K = P \Lambda P^T$

- $z \sim \mathcal{N}(0, I)$
- $y = \mu + P \Lambda^{\frac{1}{2}} z$

$\Rightarrow y \sim \mathcal{N}(\mu, K)$

So $\forall$ p.d. K leads to

$$Y \sim \mathcal{N}(\mu, K)$$

So every P.d. K has a Cholesky decomp.

WEEK 7 LEC 13
May 10. 2021

Recap of Midterm

1) Brook's Lemma
2) Bayesian LR

$y = (y_1^T, y_2^T, \cdots, y_K^T)^T$

$X = \begin{bmatrix} X_1 \\ \vdots \\ X_K \end{bmatrix}$ $\quad D_k = \{y_k, X_k\}$

- $p(\beta, \sigma^2 | D_1, D_2, \cdots, D_K)$

Step I:

$p(\beta, \sigma^2 | D_1) \propto p(\beta, \sigma^2) \times p(D_1 | \beta, \sigma^2)$

$\propto NIG(\beta, \sigma^2 | \mu_\beta, V_\beta, a, b) \times N(y_1 | X_1\beta_1, \sigma^2 I_{m_1})$

$\propto NIG(\beta, \sigma^2 | M_1 m_1, M_1, a_1, b_1)$

where

$$\begin{cases} m_1 = V_\beta^{-1}\mu_\beta + X_1^T y_1 \\ M_1^{-1} = V_\beta^{-1} + X_1^T X_1 \end{cases} \quad \begin{cases} a_1 = a + \frac{m_1}{2} \\ b_1 = b + \frac{1}{2}\{y_1^T y_1 + \mu_\beta^T V_\beta^{-1}\mu_\beta - m_1^T M_1 m_1\} \end{cases}$$

Step II:

$p(\beta, \sigma^2 | D_1, D_2) \propto p(\beta, \sigma^2) p(D_1, D_2 | \beta, \sigma^2)$

$\propto \underbrace{p(\beta, \sigma^2)\, p(D_1 | \beta, \sigma^2)}_{p(\beta, \sigma^2 | D_1)}\, p(D_2 | \beta, \sigma^2)$

$\propto NIG(\beta, \sigma^2 | M_1 m_1, M_1, a_1, b_1) \cdot N(y_2 | X_2\beta, \sigma^2 I_{m_2})$

$\propto NIG(\beta, \sigma^2 | M_2 m_2, M_2, a_2, b_2)$

where

$$\begin{cases} m_2 = m_1 + X_2^T y_2 \\ M_2^{-1} = M_1^{-1} + X_2^T X_2 \end{cases} \quad \begin{cases} a_2 = a_1 + \frac{m_2}{2} \\ b_2 = b_1 + \frac{1}{2}\{y_2^T y_2 + m_1^T M_1 m_1 - m_2^T M_2 m_2\} \end{cases}$$

We can go for **K** steps:

- $m_K = m_{K-1} + X_K^T y_K$
  $= V_\beta^{-1}\mu_\beta + \sum_{i=1}^{K-1} X_i^T y_i + X_K^T y_K$
  $= V_\beta^{-1}\mu_\beta + \sum_{i=1}^{K} X_i^T y_i$
- $M_K^{-1} = M_{K-1}^{-1} + X_K^T X_K$
  $= V_\beta^{-1} + \sum_{i=1}^{K-1} X_i^T X_i + X_K^T X_K$
  $= V_\beta^{-1} + \sum_{i=1}^{K} X_K^T X_K$
- $a_* = a_K = a + \frac{1}{2}\sum_{k=1}^{K} m_k = a + \frac{n}{2}$
- $b_* = b_K = b + \frac{1}{2}\{\sum_{k=1}^{K} y_i^T y_i + \mu_\beta^T V_\beta^{-1}\mu_\beta - m_K^T M_K^{-1} m_K\}$

So two approaches give the same result.

Homework week 7

3) Augmented linear model

4) Density function we have:

$X = (X_1, \dots, X_n)^T$

$p(x) = \prod_{i=1}^{n} f(x_i|\theta)$ and

$X_i|\theta \overset{iid}{\sim} f(\cdot|\theta)$

$\int p(\theta)d\theta = +\infty$ and

$\int p(\theta) f(x_1|\theta)\, dx_1 < +\infty$

What can we say about

$$\int p(\theta) \prod_{i=1}^{n} f(x_i|\theta) d\theta = ?$$

$p(\theta) f(x_1|\theta) \propto p(\theta|x_1)$

$\int p(\theta|x_1) d\theta = 1$

NOTE $p(\theta|x_1) = C \cdot p(\theta) \cdot f(x_1|\theta)$

If the prior × likelihood integrates to finite number, the the posterior is well-defined.

$$\int p(\theta|x_1) d\theta = \int \frac{p(\theta) f(x_1|\theta)}{[\int p(\theta) f(x_1|\theta) d\theta]} d\theta$$

$$= 1.$$

Also if $\int p(\theta|x) d\theta = c \Rightarrow$

$$\frac{1}{c} \int p(\theta|x) d\theta = 1$$

So $\frac{1}{c} p(\theta|x)$ is a proper prior.

Look @ $p(\theta|x_1, x_2)$

$p(\theta|x_1, x_2) \propto p(\theta) \cdot f(x_1, x_2|\theta)$

$\propto p(\theta) f(x_1|\theta) f(x_2|\theta)$

$\propto p(\theta|x_1)$

$\propto p(\theta|x_1) f(x_2|\theta)$

proper prior proper density

$\Rightarrow p(\theta|x_1, x_2)$ must be well-defined.

$\Rightarrow \int p(\theta) f(x_1, x_2|\theta) d\theta$ is well-defined.

□

WEEK 7 LEC 14 · May 12

Recall joint prob. model:

$$p(y_1)\,p(y_2|y_1)\,p(y_3|y_1,y_2) \cdots \cdots p(y_n|y_1,\cdots,y_{n-1})$$

We saw the linear model representation

$$y_1 = \mu_1 + \epsilon;\ \epsilon \sim \mathcal{N}(0, d_1)$$

$$y_i = \sum_{j=1}^{i-1} a_{ij} y_j + \epsilon_i;\ \epsilon_i \sim \mathcal{N}(0, d_2)$$
$$i = 2, \cdots, n.$$

$$\Rightarrow \quad y = Ay + \epsilon,$$

$A = \{a_{ij}\}$ strictly lower-Δ

$$\Rightarrow \quad \underbrace{(I-A)}_{\text{unit lower-}\Delta}\, y = \epsilon$$

$$y \sim \mathcal{N}\left(0, (I-A)^{-1} D (I-A^T)^{-1}\right)$$

If $y \sim \mathcal{N}(0, \Sigma)$, then

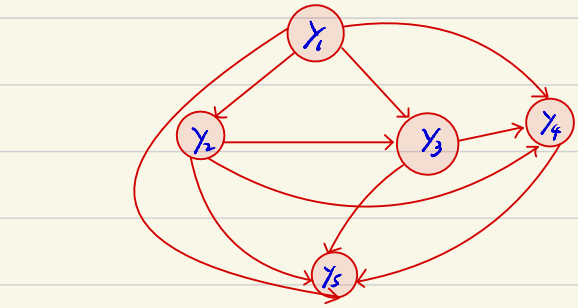
$$\underbrace{\Sigma}_{\text{p.d.}} = \underbrace{(I-A)^{-1}}_{L}\, D\, \underbrace{(I-A^T)^{-1}}_{L^T}$$

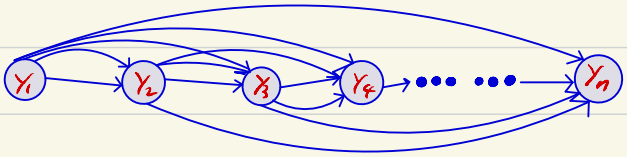
Cholesky Decomp.

DAGS.

$$p(y_1)\,p(y_2|y_1) \cdots \cdots p(y_n|y_1,\ldots,y_{n-1})$$

Directed Acyclic Graphical Model





DAGs : Bayesian Networks

All Bayesian models are DAGs.

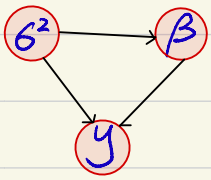
- DAGs because they (& the core) specify joint distributions.

Ex $y_i|\beta,\sigma^2 \sim \mathcal{N}(x_i^T\beta, \sigma^2)$

$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$

$\sigma^2 \sim IG(a, b)$

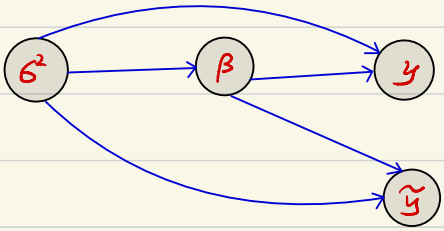
How to represent it as a DAG?



Suppose:

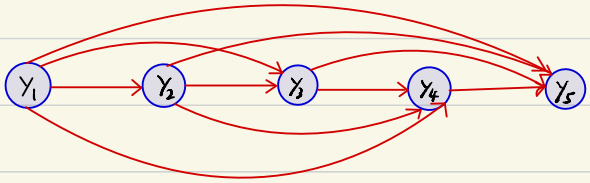
$$y|\beta,\sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$
$$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$$
$$\sigma^2 \sim IG(a, b)$$
$$\tilde{y}|\beta,\sigma^2 \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 I_n)$$

What is the DAG?



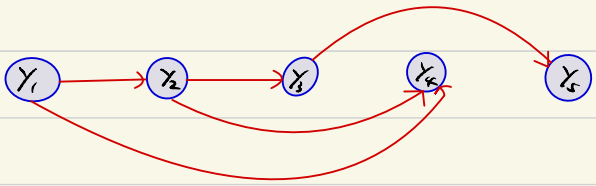
DAGs are easily modified to incorporate conditional independence (and hence sparsity).

Consider $n=5$, $y_1, y_2, \dots, y_5$, complete DAG:



$$p(y_1)\,p(y_2|y_1)\,p(y_3|y_1,y_2)\,p(y_4|y_1,y_2,y_3)\,p(y_5|y_1,\dots,y_4)$$

Drop edges:



$$p(y_1)\,p(y_2|y_1)\,p(y_3|y_2)\,p(y_4|y_1,y_2)\,p(y_5|y_3)$$

$$\Rightarrow \quad Y_3 \perp\!\!\!\perp Y_1 \mid Y_2;\ Y_4 \perp\!\!\!\perp Y_3 \mid Y_1, Y_2$$
$$Y_5 \perp\!\!\!\perp (Y_1, Y_2, Y_4) \mid Y_3$$

Conditional indep $\longrightarrow$ Sparsity

Return to Linear model representation

Suppose $\mu = 0$,

$$y_1 = 0 + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0, d_1)$$
$$y_2 = a_{21} y_1 + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, d_2)$$
$$y_3 = a_{31} y_1 + a_{32} y_2 + \epsilon_3, \quad \epsilon_3 \sim \mathcal{N}(0, d_3)$$

$$Y_3 \perp\!\!\!\perp Y_1 \mid Y_2 \Rightarrow a_{31} = 0$$

$$y_4 = a_{41} y_1 + a_{42} y_2 + a_{43} y_3 + \epsilon_4;$$
$$\epsilon_4 \sim \mathcal{N}(0, d_4)$$

$$Y_4 \perp\!\!\!\perp Y_3 \mid Y_1, Y_2 \Rightarrow a_{43} = 0$$

$$y_5 = a_{51} y_1 + a_{52} y_2 + \dots + a_{54} y_4 + \epsilon_5;$$

$$Y_5 \perp\!\!\!\perp Y_1, Y_2, Y_4 \mid Y_3 \Rightarrow a_{51} = a_{52} = a_{54} = 0$$

$$p(y_5|y_1, y_2, y_3, y_4) = p(y_5|y_3)$$
$$p(y_4|y_1, y_2, y_3) = p(y_4|y_1, y_2)$$
$$p(y_3|y_1, y_2) = p(y_4|y_3)$$

Parents shrinking $\Leftrightarrow$ a_{ij}'s $= 0$

$$\Sigma = (I-A)^{-1} D\, (I-A^T)^{-1}$$
$$\Sigma^{-1} = (I-A^T) D^{-1} (I-A)$$

Sparse $A \longleftrightarrow$ sparsity in Σ^{-1}.

$$y^T \Sigma^{-1} y = \underbrace{y^T (I-A^T)}_{z^T} D^{-1} \underbrace{(I-A)\, y}_{z}$$

$z = (I-A)\, y$.

Dense $y \xrightarrow{?}$ sparse $\underline{z}$.

Statistical interpretation comes from

$\underbrace{\Sigma^{-1}}_{\text{precision matrix}} \longleftrightarrow A$

General DAG model

$$\prod_{i=1}^{n} p(y_i \mid y_{Pa[i]})$$

$Pa[1] = \{\emptyset\}$ empty set so:

$p(y_1 \mid y_{Pa[1]}) = p(y_1)$

Full conditionals from a DAG

(Ex) $\tilde{p}(y_1, \cdots, y_n) = p(y_1)p(y_2|y_1)p(y_3|y_2)$
$p(y_4|y_1, y_2)p(y_5|y_3)$

- $\tilde{p}(y_1|\cdot) = \tilde{p}(y_1|y_2, y_3, y_4, y_5)$

$\propto p(y_1)p(y_2|y_1)p(y_4|y_1, y_2)$

$\propto p(y_1|y_{Pa[1]}) \prod_{\{j: y_j \in ch(y_1)\}} p(y_j|y_{Pa[j]})$

$\tilde{p}(y_3|\cdot) \propto p(y_3|y_{Pa[3]}) \cdot \underbrace{p(y_5|y_3)}_{\prod_{\{j: y_j \in ch(y_3)\}} p(y_j|y_{Pa(j)})}$

In general,

$$\tilde{p}(y_i|\cdot) = p(y_i \mid Pa[i]) \prod_{\{j \in ch(i)\}} p(y_j \mid y_{Pa[j]})$$

Suppose $y = (y_1, \cdots, y_n)^T \sim \mathcal{N}(0, Q^{-1})$

$Q = \Sigma^{-1}$ is the precision mat.

$y_i \perp\!\!\!\perp y_j \mid \underbrace{y_{-(i,j)}}_{\text{everything except } i \& j}$

$p(y_i|y_{(-i)})$ does not depend on y_j.

i.e. $p(y_i|y_{(-i)}) = p(y_i|y_{-(i,j)})$?? when?

Can we get a clue in terms of Q?

Keep an eye here!

$p(y) \propto e^{-\frac{1}{2} y^T Q y}$

$$y^T Q y = \sum_{k=1}^{n} \sum_{l=1}^{n} q_{kl} y_k y_l$$

$$= \sum_{k=1}^{n} q_{kk} y_k^2 + \sum_{k \neq l} q_{kl} y_k y_l$$

$p(y_i|\cdot)$ for a fixed i:

need only terms involving i:

$$y^T Q y \Rightarrow q_{ii} y_i^2 + y_i \left(\sum_{l \neq i} q_{il} y_l\right)$$

$$= q_{ii} y_i^2 + y_i \left(2 q_{ij} y_j + \sum_{l \neq \{i,j\}} q_{il} y_l\right)$$

$$= q_{ii} y_i^2 + 2 q_{ij} y_i y_j + 2 y_i f(y_{(-i,-j)})$$

$\Rightarrow$

$y^T Q y \Rightarrow a_{ii} y_i^2 + 2 a_{ij} y_i y_j +$
$2 y_i$ (terms free of y_j)

So $p(y_i | y_{(-i)})$ will be free of y_i if & only if $q_{ij} = 0$

So $p(y_i | y_{(-i)}) = p(y_i | y_{(-(ij))})$

iff $q_{ij} = 0$.

$Y_i \perp\!\!\!\perp Y_j \mid Y_{(-(ij))} \iff Q_{ij} = 0$

WEEK 8 • LEC 15 • May 17

Recall precision matrix

$Y_i \perp\!\!\!\perp Y_j \mid Y_{(-(ij))}$ iff $Q_{ij} = 0$

Closer look @ multivariate normal with a precision matrix Q

$$X = \begin{matrix} n_A \{ \\ n_B \{ \end{matrix} \overbrace{\begin{bmatrix} x_A \\ x_B \end{bmatrix}}^{1} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}}_{\mu}, \underbrace{\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}}_{\Sigma} \right)$$

$n_A + n_B = n$.

μ_A $n_A \times 1$, μ_B $n_B \times 1$, Σ_{AA} $n_A \times n_A$

Σ_{AB} $n_A \times n_B$, Σ_{BA} $n_B \times n_A$, Σ_{BB} $n_B \times n_B$

$\Sigma_{AB} = \Sigma_{BA}^T$

$$Q = \Sigma^{-1} = \begin{bmatrix} Q_{AA} & Q_{AB} \\ Q_{BA} & Q_{BB} \end{bmatrix}$$

$$p(x) \propto e^{-\frac{1}{2}(x-\mu)^T Q (x-\mu)}$$

Closer look @ $(x-\mu)^T Q (x-\mu)$

$$\begin{bmatrix} (x_A-\mu_A)^T & (x_B-\mu_B)^T \end{bmatrix} \begin{bmatrix} Q_{AA} & Q_{AB} \\ Q_{BA} & Q_{BB} \end{bmatrix} \begin{bmatrix} x_A-\mu_A \\ x_B-\mu_B \end{bmatrix}$$

$$= (x_A-\mu_A)^T Q_{AA} (x_A-\mu_A) + (x_B-\mu_B)^T Q_{BB} (x_B-\mu_B)$$

$$+ \underbrace{(x_A-\mu_A)^T Q_{AB} (x_B-\mu_B)}_{\text{Scalars: transpose of each other}} + \underbrace{(x_B-\mu_B)^T Q_{BA} (x_A-\mu_A)}_{\Rightarrow \text{ same or equal.}}$$

$$= (x_A-\mu_A)^T Q_{AA} (x_A-\mu_A) + (x_B-\mu_B)^T Q_{BB} (x_B-\mu_B)$$

$$+ 2 (x_A-\mu_A)^T Q_{AB} (x_B-\mu_B)$$

Mn-formula

We know $p(x) = p(x_B)p(x_A|x_B)$

Goal: identify $p(x_B)$ and $p(x_A|x_B)$.

- Factorise $p(x) \propto f(x_B)\, g(x_A, x_B)$ s.t. $\int f(x_B) = C_1 < \infty$, $\int g(x_A, x_B)\, dx_A = C_2 < \infty$

$\Rightarrow$ $f(x_B) \propto p(x_B)$, $g(x_A, x_B) \propto p(x_A|x_B)$

Pf: $p(x) = p(x_B)p(x_A|x_B) = \tilde{p}(x_B)\tilde{p}(x_A|x_B)$

$\Rightarrow p(x_B) = \tilde{p}(x_B)$ Because

$$p(x_B) = \int p(x_B)p(x_A|x_B)dx_A = \int \tilde{p}_1(x_B)\tilde{p}_2(x_A|x_B)dx_A$$
$$= \tilde{p}_1(x_B)\underbrace{\int \tilde{p}_2(x_A|x_B)dx_A}_{1}$$
$$= \tilde{p}_1(x_B)$$

Hence, $p(x_B)p(x_A|x_B) = \tilde{p}_1(x_A)\tilde{p}_2(x_A|x_B)$

$\Rightarrow p(x_A|x_B) = \tilde{p}_2(x_A|x_B)$ $\forall$ $x_B \in \{x : p_{x_B}(x) > 0\}$ $\square$.

Return to our case:

$$p(x) \propto e^{-\frac{1}{2}x^T Q x}$$

$$x^T Q x = (x_A-\mu_A)^T Q_{AA}(x_A-\mu_A) + (x_B-\mu_B)^T Q_{BB}(x_B-\mu_B) + 2(x_A-\mu_A)^T Q_{AB}(x_B-\mu_B)$$
$$= x_A^T Q_{AA} x_A - 2x_A^T Q_{AA}\mu_A + 2x_A^T Q_{AB}(x_B-\mu_B) - 2x_B^T Q_{BA}\mu_A + x_B^T Q_{BB} x_B - 2x_B^T Q_{BB}\mu_B + \text{Const.}$$
$$= G(x_A, x_B) + F(x_B) + \text{const.}$$

Where

$$G(x_A, x_B) = x_A^T Q_{AA} x_A - 2x_A^T Q_{AA}\mu_A + 2x_A^T Q_{AB}(x_B-\mu_B)$$
$$F(x_B) = x_B^T Q_B x_B - 2x_B^T(Q_{BA}\mu_B + Q_{BA}\mu_A)$$

$$\therefore p(x) \propto e^{-\frac{1}{2}G(x_A, x_B)} e^{-\frac{1}{2}F(x_B)}$$

($\propto$: luxury)

$$G(x_A, x_B) = x_A^T M^{-1} x_A - 2x_A^T m$$

$$\begin{cases} M^{-1} = Q_{AA} \\ m = Q_{AA}\mu_A - Q_{AB}(x_B-\mu_B) \end{cases},$$

$\Rightarrow$

$$G(x_A, x_B) = (x_A - Mm)^T M^{-1}(x_A - Mm)$$

$p(x_A|x_B) = \mathcal{N}(x_A | Mm, M)$

$\mathrm{Var}(x_A|x_B) = Q_{AA}^{-1} = M$

$\mathbb{E}(x_A|x_B) = Mm = Q_{AA}^{-1}(Q_{AA}\mu_A - Q_{AB}(x_B-\mu_B))$
$= \mu_A + (-Q_{AA}^{-1}Q_{AB}(x_B-\mu_B))$

$$G(x_A, x_B) + F(x_B)$$
$$= (x_A - Mm)^T M^{-1}(x_A - Mm) - m^T Mm + F(x_B)$$

HW: Show $p(x_B) = \mathcal{N}(\mu_B, \mathcal{K}^{-1})$

$$\mathcal{K} = \Sigma_{BB}^{-1}$$

$$x_B = \mu_B + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma_{BB})$$

$$x_A = \mu_A + (-Q_{AA}^{-1} Q_{AB})(x_B - \mu_B)$$

Also from standard normal:

$$\mathbb{E}[x_A | x_B] = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B)$$

$$\Sigma_{AB}\Sigma_{BB}^{-1} = -Q_{AA}^{-1} Q_{AB}$$

Recall $\Sigma^{-1} = Q$, **HW:**

without explicitly solving a linear system,

Can you use the factorization

$$p(x_A, x_B) = p(x_B)\, p(x_A | x_B)$$

to express:

Q_{AA}, Q_{AB}, Q_{BB} in terms of

$\Sigma_{AA}, \Sigma_{AB}, \Sigma_{BB}$.

- $x_A \perp\!\!\!\perp x_B$ iff $Q_{AB} = 0$

  $p(x_A | x_B) = p(x_A)$

$m^T \mathcal{M} m = (Q_{AA}\mu_A - Q_{AB}(x_B - \mu_B))^T \mathcal{M} m$

$= (Q_{AA}\mu_A - Q_{AB}(x_B - \mu_B))^T$

$(\mu_A - Q_{AA}^{-1} Q_{AB}(x_B - \mu_B))$

$= \mu_A^T Q_{AA} \mu_A - (x_B - \mu_B)^T Q_{BA} \mu_A -$

$\mu_A Q_{AB}(x_B - \mu_B) + (x_B - \mu_B)^T Q_{BA} Q_{AA}^{-1} Q_{AB}(x_B - \mu_B)$

$= (x_B - \mu_B)^T Q_{BA} Q_{AA}^{-1} Q_{AB}(x_B - \mu_B) -$

$2(x_B - \mu_B)^T Q_{BA}\mu_A + \text{const.}$

$F(x_B) - m^T \mathcal{M} m$

$= (x_B - \mu_B)^T (Q_{BB} - Q_{BA} Q_{AA}^{-1} Q_{AB})(x_B - \mu_B) -$

$2(x_B - \mu_B)^T (Q_{BA}\mu_A - Q_{BA}\mu_A)$

$= (x_B - \mu_B)^T [Q_{BB} - Q_{BA} Q_{AA}^{-1} Q_{AB}](x_B - \mu_B)$

$\Rightarrow p(x_B) \propto e^{-\frac{1}{2}\{F(x_B) - m^T \mathcal{M} m\}}$

$= \mathcal{N}(x_B | \mu_B, [Q_{BB} - Q_{BA} Q_{AA}^{-1} Q_{AB}])$

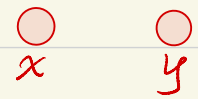
So $\Sigma_{BB}^{-1} = Q_{BB} - Q_{BA} Q_{AA}^{-1} Q_{AB}$.

HW is DONE!

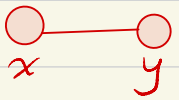
WEEK 9 · LEC 16
May 24 , 2021

Recall Directed Acyclic graphical models $\Leftrightarrow$ joint probability models

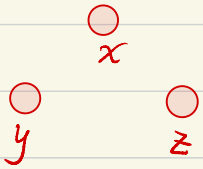
Another class of directed models arise from undirected graphical models.



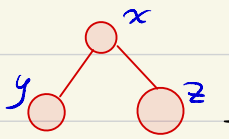
$x \perp\!\!\!\perp y$



$p(x,y) \neq p(x)p(y)$



$p(x,y,z) = p(x)p(y)p(z)$



$p(x,y|z) \neq p(x|z)p(y|z)$

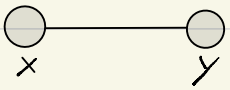
or $x \not\perp\!\!\!\perp y \mid z$ & $x \not\perp\!\!\!\perp z \mid y$

But: $y \perp\!\!\!\perp z \mid x$.

In general, $G = \{V, E\}$ where

V = vertices & RVs

E = edges (conditional indept. relations)



$p(x|y) = \mathcal{N}(x|y,1)$

$p(y|x) = \mathcal{N}(y|x,1)$

$$\Rightarrow \frac{p(x,y)}{p(x_0,y_0)} = \frac{p(y|x)}{p(y_0|x)} \frac{p(x|y_0)}{p(x_0|y_0)}$$

Brook's Lemma

$$\Rightarrow p(x,y) \propto e^{-\frac{1}{2}(x-y)^2}$$

$$(x-y)^2 = x^2 - 2xy + y^2$$

$$= (x \; y) \underbrace{\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}}_{D-W} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Diagonal matrix with # of neighbors along diagonal

Adjacency matrix

Generalize:

$$x_i | x_{-i} \sim \mathcal{N}\left(\rho \sum_{j=1}^{n} \frac{w_{ij} x_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\sigma^2}{\sum_{j=1}^{n} w_{ij}}\right)$$

$w_{ii} = 0 \quad \forall \; i = 1, 2, \cdots, n$

$$w_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$$

$\sum_{j=1}^{n} w_{ij} = m_i$ = # of neighbors of i.

(w_{i+} often used)

Homework

Apply Brook's Lemma to the above framework, where

$$x_i \mid x_{(-i)} \sim \mathcal{N}\left(\rho \frac{\sum_{j=1}^{n} w_{ij} x_j}{w_{i+}}, \frac{\sigma^2}{w_{i+}}\right)$$

to show that

$$p(x_1, x_2, \cdots, x_n) \propto e^{-\frac{1}{2\sigma^2} x^T (D - \rho W) x}$$

where $x = (x_1, \cdots, x_n)^T$

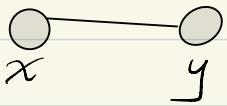
$$D = \text{Diag}(w_{1+}, w_{2+}, \cdots, w_{n+})$$

and W is the binary adjacency matrix.

Hint:

$$\frac{p(x)}{p(0)} = \prod_{i=1}^{n} p(x_i \mid \cdot) \Big/ \prod_{i=1}^{n} p(0 \mid \cdot)$$

As we saw for



the joint distribution $p(x, y)$ need not be well defined.

So: when will

$$p(x) \propto e^{-\frac{1}{2\sigma^2} x^T (D - \rho W) x}$$

be well-defined?

ANS: If $D - \rho W$ is p.d. i.e. all eigen-values are > 0.

So our question:

when is $D - \rho W \succ 0$?

Fact: $D - \rho W$ is symmetric because graph is undirected (Relation defined by the graph is a sym relation)

So $w_{ij} = w_{ji}$.

So W is symmetric & D is diagonal

Can we analytically access the eigen values of $D - \rho W$?

$$D - \rho W = D^{\frac{1}{2}} \left(I - \rho \underbrace{D^{-\frac{1}{2}} W D^{-\frac{1}{2}}}_{\widetilde{W}}\right) D^{\frac{1}{2}}$$

McGreevy's Decomposition

$$\Rightarrow D - \rho W = D^{\frac{1}{2}} (I - \rho \widetilde{W}) D^{\frac{1}{2}}$$

$$\widetilde{W} = \underbrace{D^{-\frac{1}{2}} W D^{-\frac{1}{2}}}_{\text{symmetric}}$$

$\Rightarrow \widetilde{W}$ has real eigen-values.

Spectral Decomposition for real symmetric matrices:

$$\widetilde{W} = \underset{\text{ortho.}}{P} \underset{\text{diag.}}{\Lambda} \underset{\text{ortho.}}{P^T}$$

We wish to find when

$$x^T (D - \rho W) x > 0 \quad \forall x \neq 0$$

$$\Leftrightarrow \underbrace{x^T D^{\frac{1}{2}}}_{y^T} (I - \rho \widetilde{W}) \underbrace{D^{\frac{1}{2}} x}_{y} > 0 \quad \forall x \neq 0$$

$$\Leftrightarrow y^T (I - \rho \widetilde{W}) y > 0 \quad \forall y \neq 0$$

Suffices to see when

$$I - \rho\widetilde{W} \succ 0.$$

$$\downarrow$$

$$I - \rho\widetilde{W} = PP^T - \rho P\Lambda P^T$$

$$= \underbrace{P(I - \rho\Lambda)P^T}$$

Spectral decomp. for $I - \rho\widetilde{W}$

$\Rightarrow \{1-\rho\lambda_1, 1-\rho\lambda_2, \cdots, 1-\rho\lambda_n\}$ are the eigen-vals of $I - \rho\widetilde{W}$

$I - \rho\widetilde{W}$ is p.d. $\Leftrightarrow$

$1 - \rho\lambda_i > 0 \quad \forall\ i = 1, \cdots, n$

Can we find an interval for ρ s.t. $I - \rho\widetilde{W}$ is p.d.?

Take the ordered eigen-vals:

$\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$: λ_i: eigen-vals of $\widetilde{W} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$

$1 - \rho\lambda_{(1)} > 0 \quad \& \quad 1 - \rho\lambda_{(n)} > 0$

Can we say that $\lambda_{(1)} < 0$ and $\lambda_{(n)} > 0$? If yes, then

$$\underbrace{1 - \rho\lambda_{(1)} > 0 \text{ iff } 1 + \rho|\lambda_{(1)}| > 0}$$

$$\Rightarrow \rho > -\frac{1}{|\lambda_{(1)}|} = \frac{1}{\lambda_{(1)}}$$

Also, $\rho < \frac{1}{\lambda_{(n)}}$

$$\Rightarrow \boxed{(I - \rho\widetilde{W}) \succ 0 \text{ iff } \rho \in \left(\frac{1}{\lambda_{(n)}}, \frac{1}{\lambda_{(1)}}\right)}$$

So is $\lambda_{(1)} < 0$ & $\lambda_{(n)} > 0$?

- $\widetilde{W} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$

$$\mathrm{Tr}(\widetilde{W}) = \mathrm{Tr}(D^{-\frac{1}{2}} W D^{-\frac{1}{2}})$$

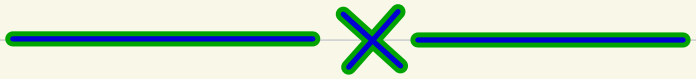
$$= \mathrm{Tr}(D^{-1} W) \quad \{W = (\circ \cdots \circ),\ D = (\circ \cdots \circ)\}$$

$$= 0 \text{ (verify easy)}$$

But $\mathrm{Tr}(\widetilde{W}) = \sum_{i=1}^n \lambda_i = \sum_{i=1}^n \lambda_{(i)}$

$\Rightarrow$ At least one of the $\lambda_{(i)}$'s is < 0 & at least one of the $\lambda_{(i)}$'s is > 0.

$\Rightarrow \lambda_{(1)} < 0 \quad \& \quad \lambda_{(n)} > 0$



Suppose Prof. Horvath gives KM a graph representing relationships in a gene network. undirected graph.

He asks KM to construct a probability model as this network.

Response: y_i, $i=1,2,\cdots,n$

Covariates: $\underset{p\times 1}{x_i}$, $i=1,2,\cdots,n$

$$y_i = x_i^T\beta + \phi_i + \epsilon_i \overset{=}{\sim} \mathcal{N}(0,\tau^2)$$

(could as well be a glm)

$$\phi_i \sim \mathcal{N}(0, \Sigma_\phi) \hookrightarrow \mathrm{Var}(\phi_i)$$

$$\phi_i = \sum_{j=1}^{n} w_{ij}\phi_j / w_{i+} + \eta_i,\ \eta_i \sim \mathcal{N}(0, \tfrac{\sigma^2}{w_{i+}})$$

$$p(\phi) \propto e^{-\frac{1}{2\sigma^2}\phi^T(D-W)\phi}$$

$D-W$ is singular! Because

$$(D-W)\mathbf{1} = \begin{bmatrix} w_{1+}-w_{1+} \\ \vdots \\ w_{n+}-w_{n+} \end{bmatrix} = \mathbf{0}$$

or $1 \in \mathcal{N}(D-W)$

$\Rightarrow$ rank$(D-W) < n$

modify: $\phi_i = \rho\sum_{j=1}^{n}\frac{w_{ij}\phi_j}{w_{i+}} + \eta_i$

$\Rightarrow p(\phi_i) \propto e^{-\frac{1}{2\sigma^2}(D-\rho W)}$

$D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ ← eigen analysis

get $\lambda_{(1)}, \lambda_{(n)}$, $\rho \sim \mathrm{Unif}(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}})$

$\phi_i \sim \mathcal{N}(0, \sigma^2(D-\rho W)^{-1})$

$D-\rho W$ is called the

Laplacian of the graph.

precision (inverse of the covariance)

WEEK9·LEC17·May 26·2021

Central to modern multivariate statistics & machine learning.

① Simple Linear Reg.

one out come Y

one explanatory variable X

$y = \beta_0 + \beta_1 X + \text{Error}$

② multiple Linear Regression

one outcome: Y

$p+1$ explanatory variables: $X_1, \cdots, X_p$

$y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \text{Error}$

$\Rightarrow$ with n measurements:

$$\underset{n\times 1}{y} = \underset{n\times(p+1)}{X}\underset{(p+1)\times 1}{\beta} + \underset{n\times 1}{e}$$

③ multivariate (matrix-valued) linear Regression

m outcomes: $Y_1, \cdots, Y_m$

common design matrix X

$$[\underbrace{Y_1 \cdots Y_m}_{\text{correlated}}] = X[\beta_1 \cdots \beta_m] + [\underbrace{E_1 \cdots E_m}_{\text{correlated}}]$$

Take a closer look:

m = # of outcomes: $Y_1, \cdots, Y_m$

n = # of observations for each outcome

Let y_{ij} = i^{th} observation on outcome j,

let x_i be a $p\times 1$ vector of explanatory variables corresponding to observation i.

$$y_{ij} = x_i^T \beta_j + e_{ij}, \quad \begin{matrix} i=1,\cdots,n \\ j=1,\cdots,m \end{matrix}$$

$$\Rightarrow \underset{n\times 1}{y_j} = X\beta_j + e_j, \quad j=1,\cdots,m$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

y_j's are NOT indept.

So how do we model e_j's jointly?

If e_j's were indept., then

$$\text{Cov}(e_j, e_{j'}) = \begin{cases} 0 & \forall j \neq j' \\ \underset{n\times n}{\sigma_j^2 V} & \forall j = j' \end{cases}$$

Dependence can be introduced as follows:

$$\text{Cov}(e_j, e_{j'}) = \underset{\text{scalar}}{u_{jj'}}\underset{n\times n}{V} \quad \forall j, j'$$

$$\underset{nm\times 1}{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} X & & \\ & \ddots & \\ & & X \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}$$

- $$\underset{mn\times 1}{y} = \underbrace{(\underset{m\times m}{I} \otimes \underset{n\times p}{X})}_{mn\times mp}\underset{mp\times 1}{\beta} + \underset{nm\times 1}{e}$$

$$\text{Cov}(e) = \text{Cov}\left(\begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix}\right)$$

$$= [\text{Cov}(e_i, e_j)]_{\substack{i=1,\cdots,m \\ j=1,\cdots,m}}$$

$$= [u_{ij} V]$$

(Ex) $m=2$:

$$\text{Cov}\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{bmatrix} \text{Var}\, e_1 & \text{Cov}(e_1, e_2) \\ \text{Cov}(e_2, e_1) & \text{Var}(e_2) \end{bmatrix}$$

$$= \begin{bmatrix} u_{11}V & u_{12}V \\ u_{21}V & u_{22}V \end{bmatrix}$$

$$\overset{\text{def}}{=} u \otimes V, \quad u = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

In general,

$$\text{Cov}(e) = \underset{m\times m}{U} \otimes \underset{n\times n}{V}$$

(m×m: outcomes; n×n: measurements)

$$y \sim \mathcal{N}((I\otimes X)\beta, U\otimes V)$$

✕ Digression ✕

Kronecker Products

Def: $A\otimes B = [a_{ij}B]$ ($(ij)^{th}$ element of A)

$$= \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1p}B \\ \vdots & & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mp}B \end{bmatrix}$$

$mn \times pq$

Start simple:

$$(\underset{1\times p}{x^T}\otimes \underset{n\times q}{B}) = [x_1B : x_2B : \cdots : x_pB]$$

($n \times pq$)

$$\begin{bmatrix} x^T \\ y^T \end{bmatrix} \otimes B = \begin{bmatrix} x_1B & \cdots & x_pB \\ y_1B & \cdots & y_pB \end{bmatrix} = \begin{bmatrix} x^T\otimes B \\ y^T \otimes B \end{bmatrix}$$

$$A\otimes B = \begin{bmatrix} a_{1*}^T \\ \vdots \\ a_{m*}^T \end{bmatrix} \otimes B$$

$$= \begin{bmatrix} a_{1*}^T \otimes B \\ \vdots \\ a_{m*}^T \otimes B \end{bmatrix}$$

- Understanding products of Kronecker (tensor) products.

$$(\underset{m\times p}{A}\otimes \underset{n\times q}{B})\cdot(\underset{p\times r}{C}\otimes \underset{q\times s}{D}) = ?$$

Start simple: $m=1$, $r=1$,

$$(\underset{1\times p}{x^T}\otimes \underset{n\times q}{B})\cdot(\underset{p\times 1}{y}\otimes \underset{q\times s}{D})$$

$$= [x_1B : x_2B : \cdots : x_pB] \begin{bmatrix} y_1D \\ y_2D \\ \vdots \\ y_pD \end{bmatrix}$$

$$= \sum_{i=1}^{p} x_iy_iBD = \left(\sum_{i=1}^{p} x_iy_i\right) BD$$

$$= x^Ty \otimes BD$$

Return to general case:

$$A\otimes B = \begin{bmatrix} a_{1*}^T\otimes B \\ \vdots \\ a_{m*}^T \otimes B \end{bmatrix}$$

$$C\otimes D = [C_{*1}\otimes D : \cdots : C_{*r}\otimes D]$$

$$(A\otimes B)\cdot(C\otimes D)$$

$$= \begin{bmatrix} a_{1*}^T\otimes B \\ \vdots \\ a_{1*}^T \otimes B \end{bmatrix} [C_{*1}\otimes D : C_{*2}\otimes D : \cdots : C_{*r}\otimes D]$$

$$= \left[(a_{i*}^T\otimes B)\cdot(C_{*j}\otimes D)\right]_{i,j=1,\cdots,m}$$

$$= \left[\underbrace{(a_{i*}^T C_{*j})}_{\text{scalar}\ (AC)_{ij}} \underset{n\times q}{B}\underset{q\times s}{D}\right]_{i,j=1,2,\cdots,m}$$

$$= AC\otimes BD$$

Thus,

$$(A\otimes B)\cdot(C\otimes D) = AC\otimes BD$$

If A, B are non-singular (square & invertible), then

$$(A\otimes B)\cdot(A^{-1}\otimes B^{-1}) = (AA^{-1})\otimes(BB^{-1})$$
$$= I_m \otimes I_n$$

$\Rightarrow A\otimes B$ is non-singular with

$$(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$$

Return to $Cov(y) = Cov(e) = U\otimes V$

$$y \sim \mathcal{N}[(I\otimes X)\beta, U\otimes V]$$

$$Z = y - (I\otimes X)\beta$$

$$p(z) = \frac{1}{(2\pi)^{\frac{mn}{2}}}\frac{1}{|U\otimes V|^{\frac{1}{2}}} e^{-\frac{1}{2}Z^T(U^{-1}\otimes V^{-1})Z}$$

- U and V are 2 p.d. matrices.

Then is $U\otimes V$ p.d. ?

Ans: Yes.

Proof: $U = P_U\Lambda_U P_U^T$ (spectral)

$V = P_V\Lambda_V P_V^T$ (spectral)

$P_U P_U^T = I_m$, $P_V P_V^T = I_n$,

$$U\otimes V = (P_U\Lambda_U P_U^T)\otimes(P_V\Lambda_V P_V^T)$$
$$\overset{\text{by Lemma}}{=} (P_U\Lambda_U\otimes P_V\Lambda_V)\cdot(P_U^T\otimes P_V^T)$$
$$= (P_U\otimes P_V)\cdot(\Lambda_U\otimes\Lambda_V)\cdot(P_U^T\otimes P_V^T)$$

Quick result:

$$(A\otimes B)^T = [a_{ij}B]^T = [a_{ji}B^T]$$
$$= A^T\otimes B^T$$

$\Rightarrow$

$$U\otimes V = \underbrace{(P_U\otimes P_V)}_{\widetilde{P}}\underbrace{(\Lambda_U\otimes\Lambda_V)}_{\widetilde{\Lambda}}\underbrace{(P_U\otimes P_V)^T}_{\widetilde{P}^T}$$

$\widetilde{\Lambda} = \text{diag}(\lambda_U^{(i)}\lambda_V^{(j)})$

Is $\widetilde{P}\widetilde{\Lambda}\widetilde{P}^T = U\otimes V$ a spectral decomposition?

Ans: only if $\widetilde{P}$ is orthogonal.

- $(P_U\otimes P_V)\cdot(P_U^T\otimes P_V^T)$
$= P_U P_U^T\otimes P_V P_V^T$
$= I_{mn}$

So $U\otimes V = \widetilde{P}\widetilde{\Lambda}\widetilde{P}^T$, where $\widetilde{\Lambda}$'s diagonal elements are the eigen-vals of $U\otimes V$.

$\therefore$ the eigen-vals of $U\otimes V$ are the set

$$\{\lambda_U^{(i)}\lambda_V^{(j)} : \begin{matrix} i=1,\dots,m \\ j=1,\dots,n\end{matrix}\}$$

$\Rightarrow$ all >0 since $\begin{cases}\lambda_U^{(i)}>0\\ \lambda_V^{(j)}>0\end{cases}$

$\Rightarrow U\otimes V$ is positive definite.

$\therefore \mathcal{N}((I\otimes X)\beta, U\otimes V)$ is a valid distribution!

WEEK 10 • LEC 18 • June 02, 2024

- $\det(U \otimes V) = \det(\tilde{P} \tilde{\Lambda} \tilde{P}^T)$

$$= \det(\tilde{\Lambda})$$

$$= \det(\Lambda_U \otimes \Lambda_V)$$

$$= \det \begin{pmatrix} \lambda_U^{(1)} \Lambda_V & & 0 \\ & \ddots & \\ 0 & & \lambda_U^{(m)} \Lambda_V \end{pmatrix}$$

$$= \prod_{i=1}^{m} \det(\lambda_U^{(i)} \Lambda_V)$$

$$= \prod_{i=1}^{m} \left\{ [\lambda_U^{(i)}]^n \prod_{j=1}^{n} \lambda_V^{(j)} \right\}$$

$$= (\lambda_U^{(1)} \cdots \lambda_U^{(m)})^n [\det(\Lambda_V)]^m$$

$$= (\det(P_U \Lambda_U P_U^T))^n (\det(P_V \Lambda_V P_V^T))^m$$

$$= (\det \Lambda_U)^n (\det \Lambda_V)^m$$

- $Z = Y - (I \otimes X)\beta$ has density

$$p(z) = \frac{1}{(2\pi)^{\frac{mn}{2}}} \frac{1}{|U|^{\frac{n}{2}} |V|^{\frac{m}{2}}} e^{-\frac{1}{2} z^T (U^{-1} \otimes V^{-1}) z}$$

□.

Next step: Construct a density for the $m \times n$ matrix

$$\underset{n \times m}{Y} = [y_{ij}]$$

$$= [y_1, \cdots, y_m]$$

Recall: Matrix-variate regression

$$y_{ij} = x_i^T \beta_j + e_{ij}; \quad \begin{matrix} i = 1, \cdots, n \\ j = 1, \cdots, m \end{matrix}$$

$$y_j = X\beta_j + e_j; \quad j = 1, 2, \cdots, m$$

$$\underbrace{[y_1 : y_2 : \cdots : y_m]}_{\underset{n \times m}{Y}} = X \underbrace{[\beta_1 : \beta_2 : \cdots : \beta_m]}_{\underset{p \times m}{B}} + \underbrace{[e_1 ; e_2 ; \cdots ; e_m]}_{\underset{n \times m}{E}}$$

Matrix-variate regression:

$$\underset{n \times m}{Y} = \underset{n \times p}{X} \underset{p \times m}{B} + \underset{n \times m}{E}$$

Find a distribution for E, hence for Y.

Recall from last lecture:

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \sim N(0, \underset{m \times m}{U} \otimes \underset{n \times n}{V})$$

- $U \otimes V$ is p.d. as long as U and V are p.d.
- $(U \otimes V)^{-1} = U^{-1} \otimes V^{-1}$
- $\det(U \otimes V) = (\det U)^n (\det V)^m$
- $(A \otimes B)(C \otimes D) = (AC \otimes BD)$

Define: vec(·) operator

$$\text{vec}(\cdot): \underbrace{\mathcal{M}_{n\times m}}_{\text{matrices}} \longmapsto \mathbb{R}^{nm\times 1}$$

Let A be an $n\times m$ matrix

$$A = [a_{*1} : a_{*2} : \cdots : a_{*m}]$$

$$\text{vec}(A) = \begin{bmatrix} a_{*1} \\ a_{*2} \\ \vdots \\ a_{*m} \end{bmatrix} \in \mathbb{R}^{nm\times 1}$$

Back to matrix-variate linear model:

$$Y = XB + E$$

$$\text{vec}(E) = \begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix}$$

$$\text{vec}(E) \sim \mathcal{N}(0, U\otimes V)$$

$p(\text{vec}(E))$ is

$$\frac{1}{(2\pi)^{\frac{mn}{2}}} \frac{1}{|U\otimes V|^{\frac{1}{2}}} e^{-\frac{1}{2}\text{vec}(E)^T (U\otimes V)^{-1}\text{vec}(E)}$$

$$\Rightarrow \frac{1}{(2\pi)^{\frac{mn}{2}}} \frac{1}{|U|^{\frac{n}{2}}|V|^{\frac{m}{2}}} e^{-\frac{1}{2}\text{vec}(E)^T (U^{-1}\otimes V^{-1})\text{vec}(E)}$$

↑ want to replace with matrix E

Goal: write the above density as a function of E (not vec(E)) with parameters U and V, no $\otimes$.

Need to simplify:

$$\underbrace{\text{vec}(E)^T}_{\text{long vec}} \underbrace{(U^{-1}\otimes V^{-1})}_{\text{long vec}} \text{vec}(E)$$

Digression vec algebra

Suppose $Y = [y_1 : \cdots : y_m]$ $n\times m$

$X = [x_1 : \cdots : x_m]$ $n\times m$

$$\text{vec}(Y)^T \text{vec}(X) = \text{Tr}(Y^T X) = \text{Tr}(X Y^T)$$

Next question:

$$(U^{-1}\otimes V^{-1})\text{vec}(E) = \text{vec}(?)$$

- Second digression

$$\underbrace{(A}_{m\times n} \otimes \underbrace{B}_{n\times p}) \overbrace{\text{vec}(X)}^{np\times 1}_{p\times n} = \text{vec}(?)$$

$$X = \underset{p\times n}{[x_1 : x_2 : \cdots : x_n]}$$

$$(A\otimes B)\text{vec}(X) = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ \vdots & & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{n} a_{1j} B x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} B x_j \end{bmatrix} \begin{matrix} \} n\times 1 \\ \\ \} n\times 1 \end{matrix}$$

($np\times 1$, $p\times 1$; $mn\times 1$)

$$= B\left([x_1 : x_2 : \cdots : x_n] \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{in} \end{bmatrix} \right)$$

$$\Rightarrow B(X a_{i*})$$

$\Rightarrow i^{th}$ row(block) of $(A\otimes B)\text{vec}(X)$

Hence

$$(A \otimes B)\,vec(X)$$

$$= \begin{bmatrix} BXa_{1*} \\ BXa_{2*} \\ \vdots \\ BXa_{m*} \end{bmatrix}$$

$$= vec(BXA^T)$$

(Since a_{i*} is the i^{th} row of A)

$\Rightarrow$

$$\boxed{(A \otimes B)\,vec(X) = vec(BXA^T)}$$

——— Digression Ends ———

Back to our question:

$$(U^{-1} \otimes V^{-1})\,vec(E) = vec(V^{-1}EU^{-1})$$

note: $(U^{-1})^T = U^{-1}$ since U sym.

So using first digression:

$$vec(E)^T vec(V^{-1}EU^{-1})$$

$$= Tr(E^T V^{-1} E U^{-1})$$

$$= Tr(V^{-1}EU^{-1}E^T)$$

$E : n \times m$

Thus,

$$p(E) = \frac{1}{(2\pi)^{\frac{nm}{2}}} \times \frac{1}{|U|^{\frac{n}{2}}|V|^{\frac{m}{2}}} \times e^{-\frac{1}{2} vec(E)^T (U^{-1} \otimes V^{-1}) vec(E)}$$

$$\Rightarrow \frac{1}{(2\pi)^{\frac{nm}{2}}} |U|^{-\frac{n}{2}} |V|^{-\frac{m}{2}} e^{-\frac{1}{2} Tr(E^T V^{-1} E U^{-1})}$$

$$= MN(E \mid 0, U, V)$$

$\Rightarrow E = Y - XB$ leads to

$$p(Y) = (2\pi)^{-\frac{nm}{2}} |U|^{-\frac{n}{2}} |V|^{-\frac{m}{2}} e^{-\frac{1}{2} Tr[(Y-XB)^T V^{-1} (Y-XB) U^{-1}]}$$

$$= MN(\underset{n \times m}{Y} \mid XB, \underset{m \times m}{U}, \underset{n \times n}{V})$$

matrix normal.

$U \leftarrow$ Covariance matrix among **columns** of Y.

$V \leftarrow$ Covariance matrix among **rows** of Y.

Homework:

Q1: ① $Vec(AB) = (?)\,vec(B)$

$= (?)\,vec(A)$

② $vec(ABC) = (?)\,vec(C)$

$= (?)\,vec(A)$

Q2: Generating from a Matrix-normal.

- Devise an algorithm to generate

$$Y \sim MN(M, U, V)$$

WITHOUT using kronecker product.

- Matrix Normal – Inverse Wishart

MNIW

Inverse Wishart: Family of $\mathbb{P}$ & p on positive definite (covariance) matrices.

- Extends inverse-Gamma to multivariate.
  Σ is $m \times m$ covariance matrix

$$p(\Sigma \mid \nu, S) = IW(\Sigma \mid \nu, S) \propto |\Sigma|^{-\frac{\nu+m+1}{2}} e^{-\frac{1}{2}\mathrm{Tr}(S\Sigma^{-1})}$$

$$MNIW(B, \Sigma \mid C, U, \nu, S) = \underbrace{IW(\Sigma \mid \nu, S)}_{p(\Sigma)} \times \underbrace{MN(B \mid C, U, \Sigma)}_{p(B \mid \Sigma)}$$

(B: $p \times m$; U: $p \times p$; Σ: $m \times m$)

General Framework:

$$Y_{n \times m} \mid B, \Sigma \sim MN(XB, \Sigma, I_n) \quad (XB: n \times p \text{ prm})$$

$$B \mid \Sigma \sim MN(C, U, \Sigma)$$

$$\Sigma \sim IW(\nu, S)$$

- Question: $p(B, \Sigma \mid Y)$??

  It turns out that $p(B, \Sigma \mid Y) = MNIW(B, \Sigma \mid M, V, \nu, S)$

- $p(B, \Sigma \mid Y) \propto MNIW(B, \Sigma \mid C, U, \nu, S) \times MN(Y \mid XB, \Sigma, I_n)$

$$\propto IW(\Sigma \mid \nu, S) \times MN(B \mid C, U, \Sigma) \times MN(Y \mid XB, \Sigma, I_n)$$

$$\propto |\Sigma|^{-\frac{\nu+m+1}{2}} e^{-\frac{1}{2}\mathrm{Tr}(S\Sigma^{-1})} \times |U|^{-\frac{m}{2}} |\Sigma|^{-\frac{p}{2}} \times e^{-\frac{1}{2}\mathrm{Tr}[(B-C)^T \Sigma^{-1}(B-C)U^{-1}]} \times |\Sigma|^{-\frac{n}{2}} |I_n|^{-\frac{m}{2}} e^{-\frac{1}{2}\mathrm{Tr}[(Y-XB)^T(Y-XB)\Sigma^{-1}]}$$

Keeping track of this!

$$\propto IW(\Sigma \mid \nu^*, S^*) \times MN(B \mid M, V, \Sigma)$$

Find M & V!