Mathematical Statistics II

STA2212H S LEC9101

Week 11 March

31 2021 Start

recording!

link

Today

Start Recording

1. HW 10 (current) not due until Monday April 5
   HW 11 posted April 2 due April 9
   Take-home posted April 9 due April 19 April 7 if I can
   No class Friday April 2; no office hour Monday April 5

2. Course evaluations available until April 12
3. Bayesian methods for text classification; discriminant analysis
4. Intro to graphical models and causality

Recap

- Multivariate normal distribution

$$f(x; \mu, \Omega) = \frac{1}{(\sqrt{2\pi})^k} |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Omega^{-1}(x-\mu)}$$

- $X \sim N_k(\mu, \Omega)$

$x_1, \ldots, x_n$ iid

$$\hat{\mu} = \bar{x} \quad \hat{\Omega} = \frac{1}{n} S = \ldots$$

- correlation $\rho_{\cancel{ij}\, rs}$

$$= \operatorname{cov}(X_r, X_s) \Big/ \sqrt{\operatorname{var}(X_r)\operatorname{var}(X_s)}$$

$$-1 \le \rho_{rs} \le 1$$

$$\rho_{rs} = 0 \;=\; X_r \perp\!\!\!\perp X_s$$

(Normality)

- partial correlation $\rho_{\cancel{ij}\, rs|-\cancel{(i,j)}\,(r,s)}$

$\rho_{rs|-(r,s)}$ corr$^{=}$ between X_r, X_s, conditional on all other X's

$$\Omega^{rs} = (\Omega^{-1})_{rs} = \operatorname{cov}(X_r, X_s \mid X_{-(r,s)})$$

↑ Mar 24 Ω_{rs} − … N ↑

Ω_{22}

$$f(\underline{x}) = \underline{f(x_{(1)} | x_{(2)})}\, \underbrace{f(x_{(2)})}_{\uparrow N}$$

Multinomial distribution

- $X \sim Mult_k(n; p)$ $j = 1, \ldots, k$ categories

  X_j = number of obs in category j

- $\mathrm{pr}(X_1 = x_1, \ldots, X_k = x_k; p) = \dfrac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$

  $0 \le p_j \le 1 \quad \sum_{j=1}^{k} p_j = 1$

  $\sum x_j = n$

- $\mathrm{E}(X) = \cancel{n} p_k$

- $\mathrm{cov}(X) =$

$$\begin{bmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ \vdots & \ddots & & \\ & & \ddots & \\ -np_1p_k & \cdots & & np_k(1-p_k) \end{bmatrix}$$

AoS Thm 14.4

- $\hat{p} = \dfrac{X}{n}$

- $\mathrm{cov}(\hat{p}) = \dfrac{\mathrm{cov}(X)}{n^2}$

Multinomial distribution

[see Ch 15 of AoS]

X_j = number of obs in category j

- $X \sim Mult_k(n; p)$
- $\mathrm{pr}(X_1 = x_1, \ldots, X_k = x_k; p) =$
- $\mathrm{E}(X) =$
- $\mathrm{cov}(X) =$
- $\hat{p} = \frac{X}{n}$
- $\mathrm{cov}(\hat{p}) =$ ntbc

$p_k = p_k(\theta)$ $\dim \theta < k-1$ Contingency tables

1	$\cdots$	k
$np_1(\theta)$	$\cdots$	$np_k(\theta)$
X_1		X_k
$\hat{np}_1$		$\hat{np}_k$

constrained

AoS Thm 14.4

χ^2 test compares these

(loglinear models)

	$1 \cdots m$
$1 \vdots \ell$	X_{11} $\cdots$ $X_{\ell m}$

$X = (X_{ij})$ $k = m \times \ell$

$i = 1, \ldots, \ell$; $j = 1, \ldots, m$

Overview

- multivariate and multinomial distributions used to study the joint distribution
- analogous to unsupervised learning ← learning relationships among variables on an equal footing
- AoS §15.1,2: 2 binary variables; 2 discrete variables multinomial
- AoS §14.2: pairs of normal variables
- AoS §15.4 one discrete, one continuous variable compare

$$\overset{\text{cont's}}{F_1(\underline{x})} \neq F_2(\underline{x})$$

discrete

- AoS Ch.15 Inference about independence

	T	C
D	—	—
A	—	-

$h(X): \mathcal{X} \to \mathcal{Y}$

$$\hat{h}(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } \hat{r}(x) > \frac{1}{2} \\ 0 & \quad\quad < \frac{1}{2} \end{cases}$$ binary

Bayes rule is

$$\max_k \frac{f_k(\underset{\sim}{x})\pi_k}{\sum_{r=1} f_r(\underset{\sim}{x})\pi_r}$$

($\hat{h}(x = k)$ if we predict $Y|X = k$)

could use parametric models for $\hat{f}_k$ (2-class) $\to$ MVN $\to$ (LDA & QDA)

or nonpar models to get ~~for~~ $\hat{f}_k$

$$\arg\max_k \frac{\hat{f}_k(x)\hat{\pi}_k}{\sum \hat{f}_r(x)\hat{\pi}_r} \qquad \hat{\pi}_r = \frac{\#(Y\text{'s in class } r)}{n}$$

BCE loss f — binary cross-entropy

$$-\frac{1}{n}\sum_{i=1}^{n}\{y_i \log p_i + (1-y_i)\log(1-p_i)\}$$

y_i — 0, 1

$$= -\frac{1}{n}\,\ell(\underset{\sim}{p};\underset{\sim}{y}) \text{ under } Y_i \sim Ber(p_i) \qquad p_i = pr(y_i = 1)$$

"log-likelihood loss" $\hat{p}_i = y_i$

Overview

supervised vs unsupervised   M/L   Stats   categorical vs continuous

some labelled data $y_1, \dots, y_n$

some features $\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n$

use features to predict y

$$p(x) = P_r(y = 1 \mid \underset{\sim}{x}) = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}$$

(e.g.)

text analysis

$$E(y \mid x) = \underset{\sim}{x}^T \beta$$

or

$$= e^{\underset{\sim}{x}^T \beta}$$

or

$$= e^{\underset{\sim}{P}(\underset{\sim}{x})^T \cdot \underset{\sim}{\theta}}$$

map $h : \mathcal{X} \to \mathcal{Y}$ classifier

$$L(h) = P\{h(X) \neq Y\} \qquad \underbrace{\frac{1}{n} \sum_{i=1}^{n} 1\{h(X_i) \neq Y_i\}}_{\hat{L}_n(h)}$$

Models for dependent data

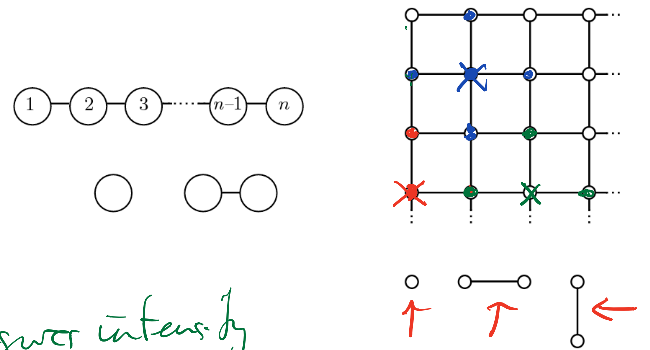
- Markov chains — SM §6.1
- continuous time Markov models — finite state space $\mathcal{S}$ — SM §6.2
- Markov random fields; directed acyclic graphs — SM §6.2
- Multivariate normal — SM §6.3
- Time series — SM §6.4
- Point processes — SM §6.5

- $X_1, \dots, X_n$ a random sequence
- Markov property:

$$\mathrm{pr}(X_{j+1} = x_{j+1} \mid X_j = x_j, \dots, X_1 = x_1) = \Pr(X_{j+1} = x_{j+1} \mid X_j = x_j)$$

- set of sites $\mathcal{J} = \{1, \dots, n\}$ $\quad \Pr(X_j = x_j \mid \cancel{X}_{-j} = x_{-j}) = \Pr(X_j = x_j \mid X_{\mathcal{N}_j} = x_{\mathcal{N}_j})$
- neighbourhood system $\mathcal{N} = \{\mathcal{N}_j, j \in \mathcal{J}\}$

Figure 6.4 Markov random fields. Left: neighbourhood structure for first-order Markov chain and its cliques and their subsets. Right: first-order neighbourhood structure, cliques and their subsets for rectangular grid of sites.



labels $y_1, \dots, y_n$

$\cancel{x}_{i1} \quad \cancel{x}_{i2} \quad \dots \quad \cancel{x}_{i\, m\times m}$

$i = 1, \dots, n$

$\cancel{x} \in \{1, \dots, k\}$

$\cancel{x} \in \mathbb{R}$ measures intensity

e.g.

Directed acyclic graphs

- can be convenient for studying relationships between variables ← "mechanistic"
- through a probability distribution on the graph
- that is specified by a factorization of the joint density

parents

nodes
edges
directed edges

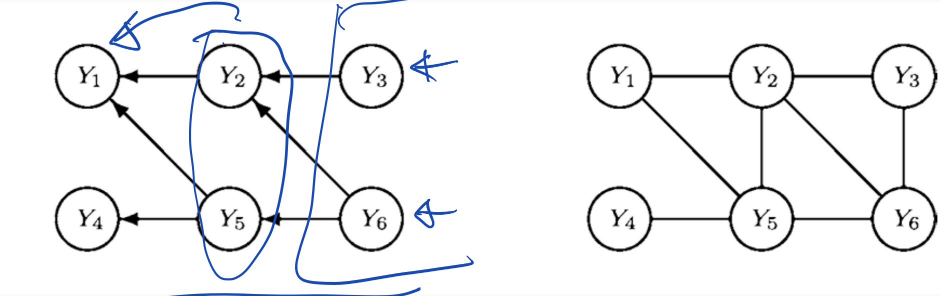


Figure 6.6 Directed acyclic and moral graphs. Left: directed acyclic graph representing (6.17). Right: moral graph, formed by moralizing the directed acyclic graph, that is, 'marrying' parents and dropping arrowheads.

cycles
cliques
colliders

$$f(y_1, \dots, y_6) = f(y_6)\, f(y_5 \mid y_6)\, f(y_2 \mid y_3, y_6)\, f(y_1 \mid y_2, y_5)\, f(y_2 \mid y_3)$$

$$f(y_4 \mid y_5)\, f(y_3)$$

$$\rightarrow f(y) = \prod_{j \in \mathcal{J}} f(y_j \mid \text{parents of } y_j)$$

(check SM)

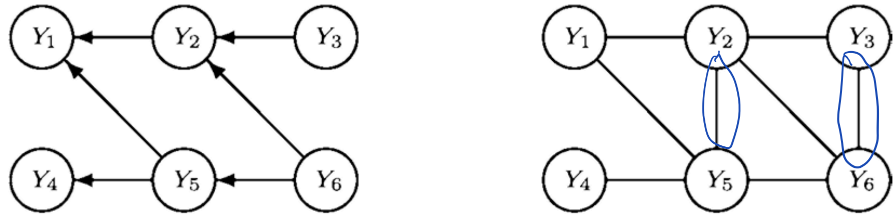


Figure 6.6 Directed acyclic and moral graphs. Left: directed acyclic graph representing (6.17). Right: moral graph, formed by moralizing the directed acyclic graph, that is, 'marrying' parents and dropping arrowheads.

31/2 and p.250

graph ← maybe motivated by an appⁿ

~~st~~ cond'l independence encoded

need a prob. distⁿ on graph ← modelling

pref. MRF

we convert a DAG to a MRF by: removing arrows, joining parents

(6.18)

$$f(y) = \prod_{j \in J} f(y_j \mid \text{parents } y_j)$$

Those models encode structured dependence

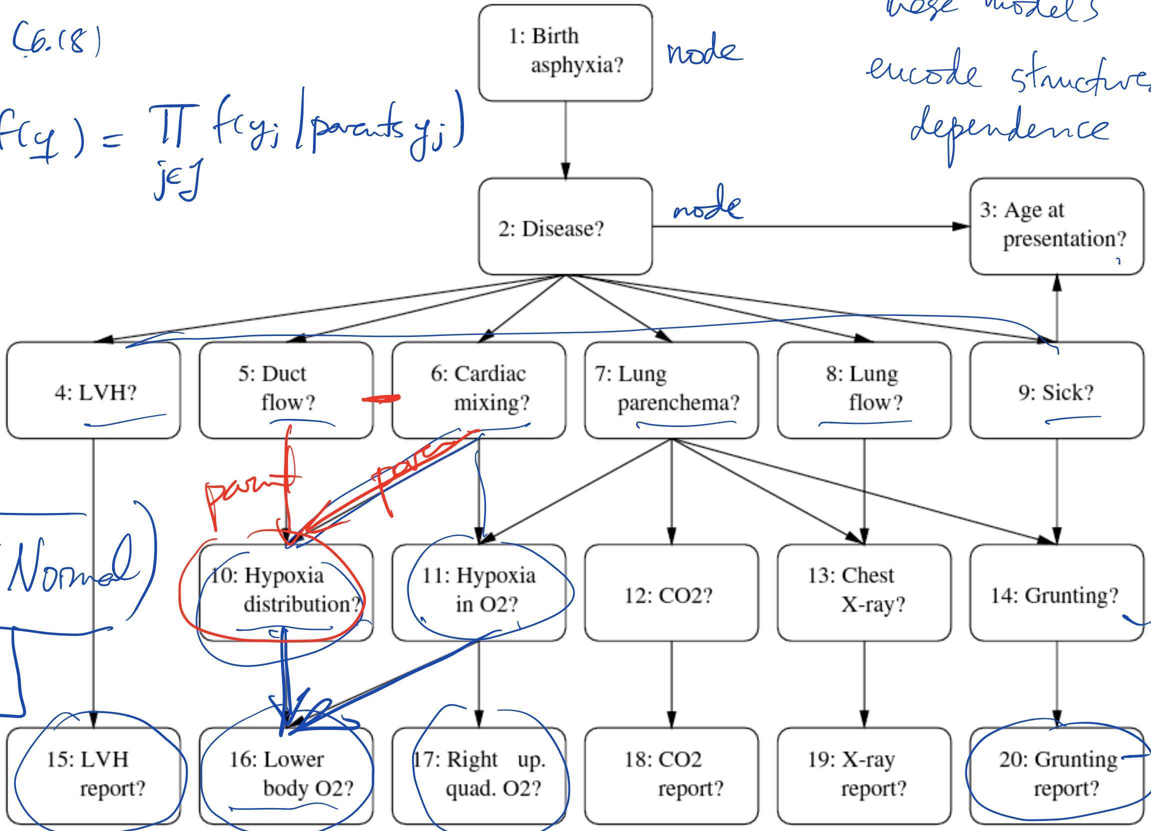


Figure 6.7 Directed acyclic graph representing the incidence and presentation of six possible diseases that would lead to a 'blue' baby (Spiegelhalter *et al.*, 1993). LVH means left ventricular hypertrophy.

$Y_j \quad j = 1, \dots, \#\text{nodes}$

dist= for Y_j

$\text{pr}(Y_j \in \text{class } k)$

Graphical Gaussian models

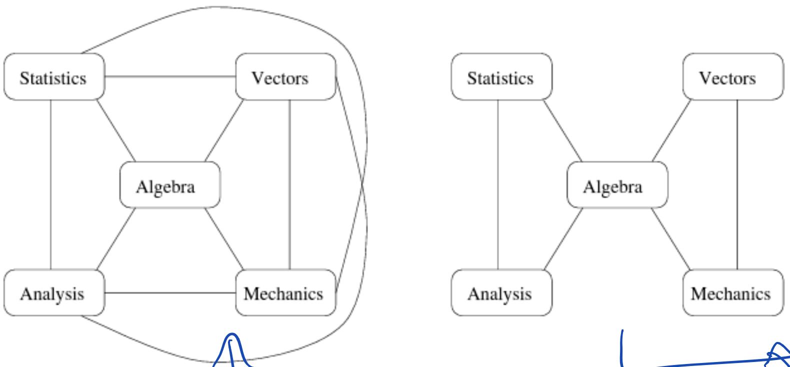


Figure 6.10 Graphs for the full model (left) and a reduced model (right) for the maths marks data. The interpretation of the reduced model is that given the result for algebra, results for vectors and mechanics are independent of those for analysis and statistics.

$\Omega^{rs} \neq \Omega^{-1}_{rs}$

$\|$

$(\Omega^{-1})_{rs} \neq \Omega^{-1}_{rs}$

complete indirected graph

M

simplified graph

If $\Omega^{rs} = 0$ then $X_r \perp\!\!\!\perp X_s \mid X_{-(r,s)}$

$= (\Omega^{-1})_{r,s}$

- Fig 17.2, Example 17.4
- Fig 17.3

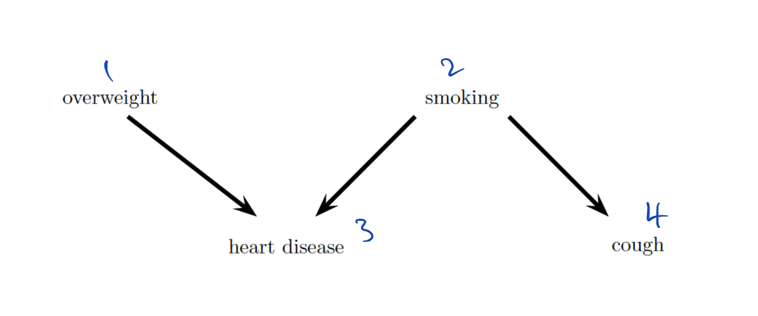


FIGURE 17.2. DAG for Example 17.4.

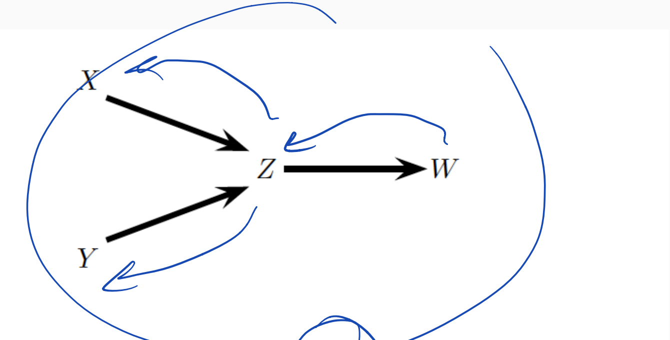


FIGURE 17.3. Another DAG.

$$f(y) = f(y_3 \mid y_2, y_1) \cdot f(y_1) f(y_2) f(y_4 \mid y_2)$$

encodes dep. in graph.

$$f(x,y,z,w) = f(w \mid z) f(z \mid x,y) f(x) f(y)$$

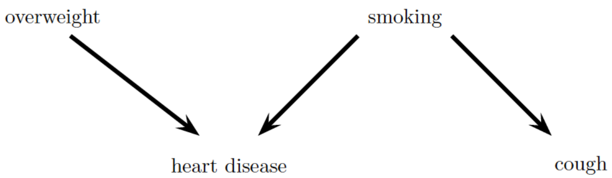


FIGURE 17.2. DAG for Example 17.4.

17.4 Example. Figure 17.2 shows a DAG with four variables. The probability function for this example factors as

$$\begin{aligned} & f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) \\ & \quad = f(\text{overweight}) \times f(\text{smoking}) \\ & \quad \times f(\text{heart disease} \mid \text{overweight}, \text{smoking}) \\ & \quad \times f(\text{cough} \mid \text{smoking}). \quad \blacksquare \end{aligned}$$

17.5 Example. For the DAG in Figure 17.3, $\mathbb{P} \in M(\mathcal{G})$ if and only if its probability function f has the form

$$f(x, y, z, w) = f(x)f(y)f(z \mid x, y)f(w \mid z). \quad \blacksquare$$

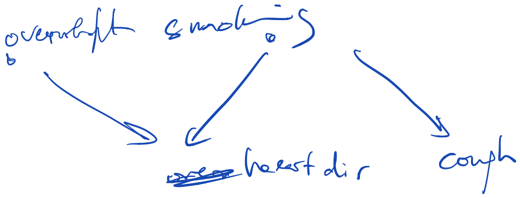
AoS (17.8)
SM 9.1

DAGs are related to causality

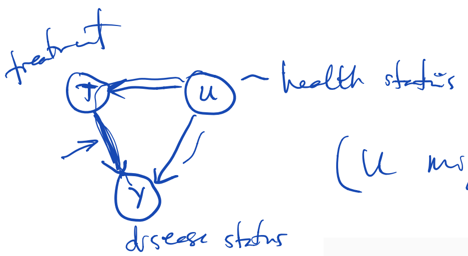
value of y at each parent node is a potential cause of the response at a child node

Causality Club
AoS

"counterfactuals"

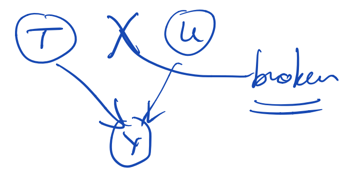


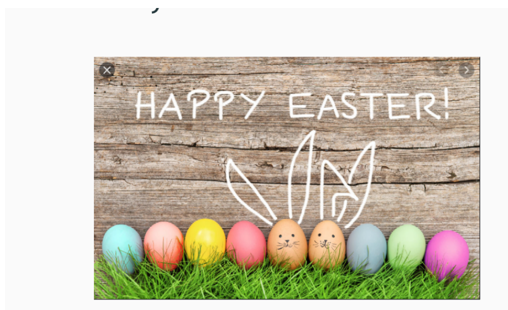
prob. model to encode all possible dep. of interest; see if data supports this, or a simple model



(U might be unseen)

If we randomize treatment





April 7 – causality + (DAG)

9 – visualization

April 12 office hours