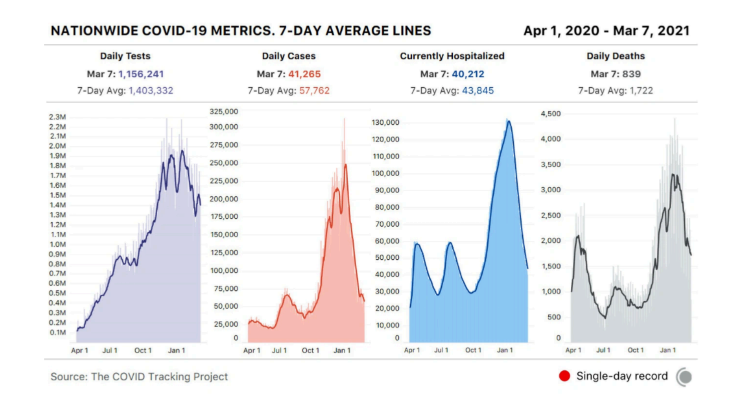
Mathematical Statistics II

STA2212H S LEC9101

Week 10

March2 24 2021

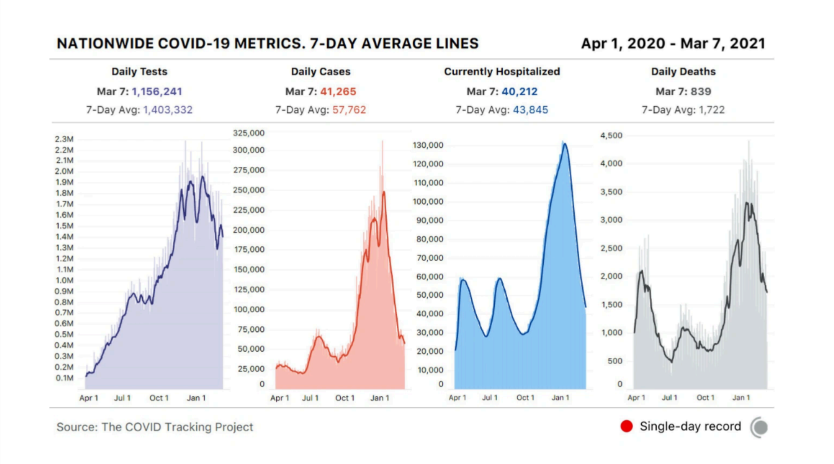
Start recording!



link

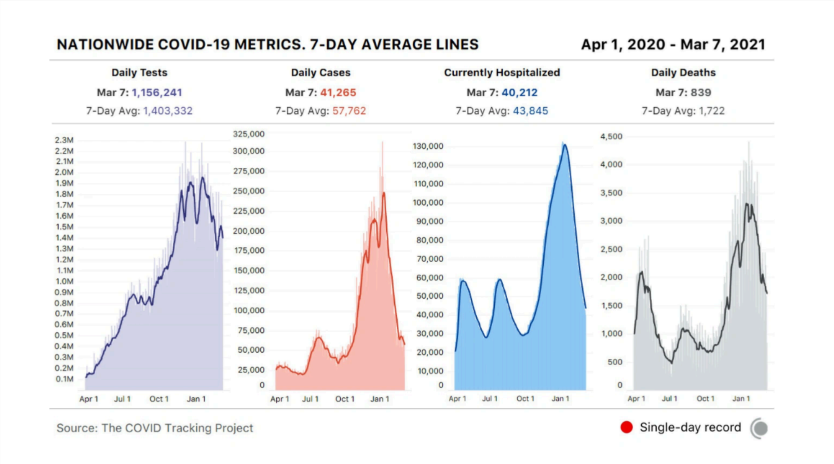
Why the Pandemic Experts Failed

The Atlantic, Mar 15, Meyer & Madrigal



Why the Pandemic Experts Failed

The Atlantic, Mar 15, Meyer & Madrigal



"*Data* might seem like an overly technical obsession, an oddly nerdy scapegoat on which to hang the deaths of half a million Americans. But data are how our leaders apprehend reality. In some sense, data are the federal government's reality. As a gap opened between the data that the leaders imagined should exist and the data that actually did exist, it swallowed the country's pandemic response."

"Before March 2020, the country had no shortage of pandemic preparation plans. Many stressed the importance of data-driven decision making. Yet these plans largely assumed that detailed and reliable data would simply … exist. They were less concerned with how those data would actually be made."

In Canada

Replying to @CTVNews and @CitImmCanada

Last year, Canada hit a record high in the number of deaths recorded in a single year, reporting over 300,000 deaths in 2020.

According to the Public Health Agency of Canada, COVID-19 accounted for approximately one death out of every 20 deaths in Canada last year.

Calling Bullshit @callin_bull · Mar 19

This is a train wreck of a data graphic from the Public Health Agency of Canada.

It implies a minimum of a 5% infection fatality rate, and that only if everyone in Canada had been infected.

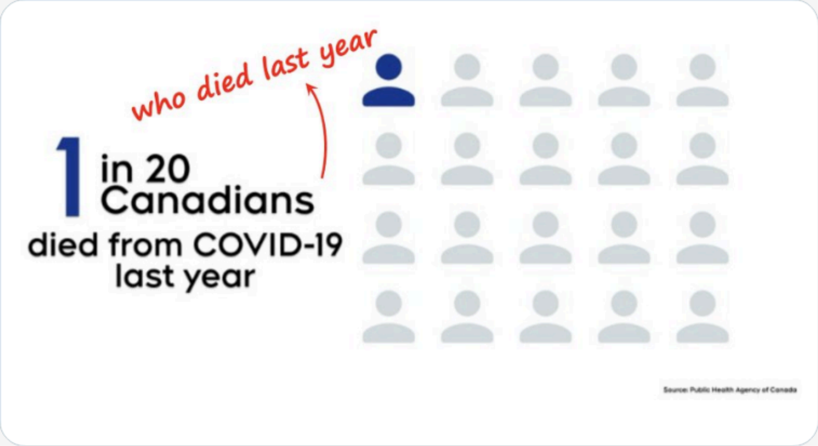
Can you spot went wrong?

Answer in the next post.

14 23 96

Calling Bullshit @callin_bull · Mar 19

What it should say is that one in twenty deaths in Canada last year were due to Covid, or that one in twenty Canadians *who died last year* died of Covid.



2 10 62

link

Recap: shrinkage estimation

- Bayesian hierarchical models lead to shrinkage estimators

$$x_i \mid \theta_i \sim N(\theta_i, \sigma^2); \theta_i \mid \mu \sim N(\mu, \tau^2); \mu \sim N(\mu_0, b^2)$$

- shrinks towards mean of hyper-prior $\tilde{\theta}_{HB,i} = w_i x_i + (1 - w_i)\widetilde{\mathrm{E}}(\mu \mid x)$ Mar 17 Slide 3

Recap: shrinkage estimation

- Bayesian hierarchical models lead to shrinkage estimators

$x_i \mid \theta_i \sim N(\theta_i, \sigma^2); \theta_i \mid \mu \sim N(\mu, \tau^2); \mu \sim N(\mu_0, b^2)$

- shrinks towards mean of hyper-prior $\tilde{\theta}_{HB,i} = w_i x_i + (1 - w_i)\widetilde{\mathrm{E}}(\mu \mid x)$ Mar 17 Slide 3

- empirical Bayes replaces hyper-prior with estimates $\hat{\mu}, \hat{\tau}^2$ (parametric EB)

using $m(\boldsymbol{x}; \mu, \tau^2) = \int L(\theta; \boldsymbol{x})\pi(\theta \mid \mu, \tau^2)d\theta$

- leads to shrinkage estimators $\tilde{\theta}_{EB,i} = w_i x_i + (1 - w_i)\hat{\mu}$
- in some models it may be possible to estimate prior nonparametrically

e.g. Poisson; EH §6.1

- SM, Example 11.30 "Shakespeare's vocabulary data"; also EH §6.2

"How many words did Shakespeare know?" (Efron & Thisted 1976)

"Did Shakespeare write a newly discovered poem?" (Thisted & Efron, 1987)

- Loss function $L(\theta, \hat{\theta})$
- Risk function – average loss over model $R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x)) f(x|\theta) dx$ $f(x \mid \theta)$

E under model

Recap: decision theory of estimation

- Loss function
- Risk function – average loss over model
- Bayes risk – average Risk function over prior
- "Bayes estimator" – minimizes Bayes risk

$L(\theta, \hat{\theta})$

$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}) f(x|\theta) dx$ $\quad f(x \mid \theta)$

$\int R(\theta, \hat{\theta}) \pi(\theta) d\theta$ $\quad \pi(\theta)$

$r_\pi(\hat{\theta})$ or $r(f, \hat{\theta})$ AoS

- Loss function
- Risk function – average loss over model $f(x \mid \theta)$

- Bayes risk – average Risk function over prior $\pi(\theta)$
- Bayes estimator – minimizes Bayes risk

- minimax estimator – minimizes maximum risk $\longleftarrow$ $\min_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$

- Loss function
- Risk function – average loss over model $f(x \mid \theta)$

$\pi(\theta) \propto 1$

- Bayes risk – average Risk function over prior $r_\pi(\theta) = \int\int L(\theta, \hat{\theta}(x)) f(x|\theta)\, dx\, d\theta$ $\pi(\theta)$
- Bayes estimator – minimizes Bayes risk

- minimax estimator – minimizes maximum risk

$R(\theta, \hat{\theta})$

x Bayes

- admissible estimator – no other estimator has smaller Risk function for all $\theta \in \Theta$
- Bayes' estimators with proper priors are admissible, ... with some caveats

✓ Bayes

ass$^=$ on model $X \sim N(\theta, 1)$

AoS Ch. 12 illustrates a useless admissible est. $\hat{\theta} = 3$

Recap: decision theory of estimation

- Loss function
- Risk function – average loss over model $f(x \mid \theta)$

- Bayes risk – average Risk function over prior $\pi(\theta)$
- Bayes estimator – minimizes Bayes risk

- minimax estimator – minimizes maximum risk

- admissible estimator – no other estimator has smaller Risk function for all $\theta \in \Theta$
- Bayes' estimators with proper priors are admissible

- Stein's paradox – when estimating many parameters, shrinkage estimators beat non-shrinkage (mle) estimators

$\hat{\theta}_{i,JS} = \left(1 - \frac{n-2}{\sum X_i^2}\right) X_i$ HW 9 no prior needed!

$\hat{\theta}_{i,JS} = \left(1 - \frac{n-3}{S^2}\right) X_i \pm$??

Today

1. Calendar, HW, Final HW ← 9th out 19th in
2. Friday (lecture)
3. Stein's paradox – baseball example
4. multivariate distributions (Ch 14 AoS, SM 6.3)

- Mar 29 3.00 – 4.00 pm EDT
  Data Science ARES
  John Aston, University of Cambridge
  "Functional Data in Constrained Spaces"
- Professor of Statistics in Public Life;
  formerly Chief Scientific Adviser
  to the Home Office, UK Government.

+ Day Week Month Year Eastern Time

April 2021

Sun	Mon	Tue	Wed	Thu	Fri
21	22 • Data Science ARES 3 PM • Office Hours 7 PM	23	Mar 24 • STA2212 10 AM	25 • Office Hours 11 AM	Lecture ~~26~~ • Grad Student ev... 8:45 AM • STA2212 10 AM • Office Hours 11 AM HW 10
28	29 • Office Hours 7 PM	30	31 AMS Notices review • STA2212 10 AM ✓ causality	Apr 1 Yanbo's nomination due • Office Hours 11 AM	2 Good Friday x HW 11
4 Easter	5 Easter Monday • Office Hours 7 PM	6	7 • STA2212 10 AM 7th	8 • Office Hours 11 AM	9 • STA2212 10 AM • Office Hours 11 AM
11	12 • STA2212 10 AM • Office Hours 7 PM	13	14	15 • Office Hours 11 AM	16 • Office Hours 11 AM
18 ~~Final due~~	19 Final due 9:00 am	20 Grader due	21	22 RS Finance • DEPARTMENTAL... 10 AM	23 Grad student research d...

graphical models

HW 12 + Final ?

Example J-S Paradox

EH §7.2

Table 7.1 *Eighteen baseball players;* `MLE` *is batting average in first 90 at bats;* `TRUTH` *is average in remainder of 1970 season; James–Stein estimator* `JS` *is based on arcsin transformation of MLEs. Sum of squared errors for predicting* `TRUTH`*:* `MLE` *.0425,* `JS` *.0218.*

data at end

Player	MLE	JS	TRUTH	x
1	.345	.283	.298	11.96
2	.333	.279	.346	11.74
3	.322	.276	.222	11.51
4	.311	.272	.276	11.29
5	.289	.265	.263	10.83
6	.289	.264	.273	10.83
7	.278	.261	.303	10.60
8	.255	.253	.270	10.13
9	.244	.249	.230	9.88
10	.233	.245	.264	9.64
11	.233	.245	.264	9.64
12	.222	.242	.210	9.40
13	.222	.241	.256	9.39
14	.222	.241	.269	9.39
15	.211	.238	.316	9.14
16	.211	.238	.226	9.14
17	.200	.234	.285	8.88
18	.145	.212	.200	7.50

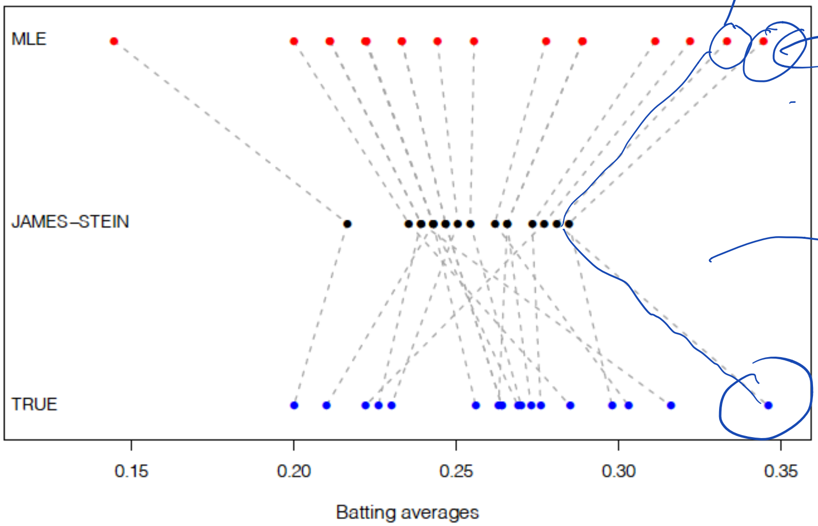
data early in season

$\hat{\theta}_{JS,i}$ = shrinks to $\hat{\mu}$

batting average

$\dfrac{X}{n}$ ← # hits / ← # of bats

Example J-S Paradox



#2

data after 90 at bats

Roberto Clemente

data at end of season

Figure 7.1 Eighteen baseball players; top line MLE, middle James–Stein, bottom true values. Only 13 points are visible, since there are ties.

vector random variables

column

$$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \qquad E(\underset{\sim}{X}) = \underset{\sim}{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$$

cov need $E(X_i X_j)$ $i,j \in \{1,\dots,k\}$

3rd moment needs $E(X_i X_j X_h)$

3-way array

$$\underset{k\times k}{\Sigma} = \mathrm{cov}(\underset{\sim}{X}) = E\{(\underset{k\times 1}{\underset{\sim}{X} - \underset{\sim}{\mu}})(\underset{1\times k}{\underset{\sim}{X} - \underset{\sim}{\mu}})^T\}$$

$\sigma_{ii} = \mathrm{var}(X_i)$

$\sigma_{ij} = \mathrm{cov}(X_i, X_j)$

expected value, variance, precision

μ $\quad \Sigma \quad \Sigma^{-1}$

$E(\underset{\sim}{a}^T \underset{\sim}{X}) = \underset{\sim}{a}^T \underset{\sim}{\mu}$ $\qquad \mathrm{cov}(a^T X) = \underset{1\times p}{a^T} \underset{p\times p}{\Sigma} a = a^T \Sigma a$ $\qquad \mathrm{var}(aX_i) = a^2 \sigma_i^2$

$E(AX) = A\mu$ $\qquad A_{hk}$ $\qquad \boxed{\mathrm{cov}(AX) = A \Sigma A^T}$

i.i.d. samples $\underset{k\times 1}{\underset{\sim}{X}_1}, \ldots, \underset{\sim}{X}_n$ $\underset{\sim}{X}_1 = \begin{pmatrix} X_{11} \\ \vdots \\ X_{1k} \end{pmatrix}$ $\underset{\sim}{X}_2 = \begin{pmatrix} X_{21} \\ \vdots \\ X_{2k} \end{pmatrix} \cdots \underset{\sim}{X}_n \begin{pmatrix} X_{n1} \\ \vdots \\ X_{nk} \end{pmatrix}$

(x_i, z_i)

$\uparrow$

$f(x_i \mid z_i)$

$$\underset{\sim}{\bar{X}} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_k \end{pmatrix} \qquad E(\underset{\sim}{\bar{X}}) = \mu \qquad \operatorname{cov}(\bar{X}) = E\{(\bar{X} - \mu)(\bar{X} - \mu)^T\}$$

$$= \underset{k\times k \text{ matrix}}{\Sigma} / n$$

unbiased estimators of μ & Σ AoS Thm 14.1

$$\hat{\mu} = \underset{\sim}{\bar{X}} \qquad \underset{k\times k}{\tilde{\Sigma}} = \frac{1}{n-1} S^2 \longleftarrow ?$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (\underset{\sim}{X}_i - \underset{\sim}{\bar{X}})(\underset{\sim}{X}_i - \underset{\sim}{\bar{X}})^T$$

$$E\tilde{\Sigma} = \Sigma$$

(nfbc)

confidence intervals for correlations

AoS §14.2

$$\tilde{\Sigma} = \frac{1}{n-1} S^2$$

$$\tilde{\sigma}_{ab} = \widetilde{cor}(X_a, X_b) = \frac{1}{n-1} = \sum_{i=1}^{n} (X_{ai} - \bar{X}_a)(X_{bi} - \bar{X}_b) \cdot \frac{1}{n-1}$$

$$\tilde{\rho}_{ab} = \underset{(\text{in } N)}{\hat{\rho}_{mle}} = \frac{\tilde{\sigma}_{ab}}{(\tilde{\sigma}_{aa}\ \tilde{\sigma}_{bb})^{1/2}}$$

estimates $\rho_{ab} = \dfrac{\sigma_{ab}}{(\sigma_{aa}\sigma_{bb})^{1/2}}$

$\in (-1, 1)$

It turns out $\hat{z} = \log\left(\dfrac{1+\hat{\rho}}{1-\hat{\rho}}\right)$ has $var(z) \simeq \dfrac{1}{n-3}$ Fisher 1915

"var. stabilizing"

doesn't dep on ρ

$\hat{z} \pm 1.96 \frac{1}{\sqrt{n-3}}$ approx 95% C.I. for z based on normal approx =

$(\hat{z}_L \quad , \hat{z}_U)$

$\longrightarrow (\hat{\rho}_L, \; \hat{\rho}_U)$ using inv. transf.

$\hat{\rho} = \frac{e^{2\hat{z}} - 1}{e^{2\hat{z}} + 1}$ is the inverse multivariate normal distribution $\underset{\sim}{X} \sim N_k(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$

AoS Thm 14.2

$$f(\underset{\sim}{x}; \underset{\sim}{\mu}, \underset{\sim}{\Sigma}) = \frac{1}{(\sqrt{2\pi})^k} \cdot \frac{1}{|\Sigma|^{1/2}} \exp - \frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})$$

need Σ full rank
i.e. +ve def.

if k is extremely large, Σ^{-1} (SM Ω^{-1}) could be difficult to compute $O(k^3)$ operⁿ.

sampling and likelihood function

$\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ iid $N_k(\mu, \Omega)$

$$L(\mu, \Omega; \underset{\sim}{x}_1 \ldots \underset{\sim}{x}_n) \propto e^{\left\{-\frac{n}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^n (\underset{\sim}{X}_i - \underset{\sim}{\mu})^T \Omega^{-1} (\underset{\sim}{X}_i - \underset{\sim}{\mu})\right\}}$$

$$\ell(\mu,\Omega) = -\frac{n}{2}\ln|\Omega| - \frac{1}{2}\sum_{i=1}^{n}(\underset{\sim}{X}_i-\underset{\sim}{\mu})^T\Omega^{-1}(\underset{\sim}{X}_i-\underset{\sim}{\mu})^T$$

$$= -\frac{n}{2}\ln|\Omega| - \frac{1}{2}\sum(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}} + \bar{\underset{\sim}{X}} - \mu)^T\Omega^{-1}(\qquad\qquad)$$

$$= -\frac{n}{2}\ln|\Omega| - \frac{1}{2}\sum(\underset{\sim}{X}_i-\bar{\underset{\sim}{X}})^T\Omega^{-1}(\underset{\sim}{X}_i-\bar{\underset{\sim}{X}}) \leftarrow \sum(X_i-\bar{X})^2$$

$$-\frac{1}{2}n(\bar{\underset{\sim}{X}}-\underset{\sim}{\mu})^T\Omega^{-1}(\bar{\underset{\sim}{X}}-\underset{\sim}{\mu}) \quad \sqrt{n}(\bar{X}-\mu)^2$$

$\max_{\mu,\Omega} \ell(\mu,\Omega)$ to define mle

$\hat{\mu}$ only appears $\qquad \hat{\mu} = \bar{X}$ maximizes $\ell(\mu,\Omega)$ over μ

$$\ell_{profile}(\Omega) = -\frac{n}{2}\ln|\Omega| - \frac{1}{2}\mathrm{tr}\sum(X_i-\bar{X})^T\Omega^{-1}(X_i-\bar{X})$$

$\hat{\Omega}_{mle}$ $\qquad \left.\frac{\partial \ell_{prof} \leftarrow \text{scalar}}{\partial\Omega \leftarrow \text{matrix}}\right|_{\Omega=\hat{\Omega}} = 0$

$$\frac{\partial \ell_{pr}(\Omega)}{\partial\sigma_{ab}} = \frac{\partial\log|\Omega|}{\partial\sigma_{ab}} - \frac{1}{2}\frac{\partial}{\partial\sigma_{ab}}\left[\Omega^{-1}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T\right]$$

SM Ch6 ...

Exercise 6.3.1.

$$\hat{\Omega} = \frac{1}{n}\sum(\underset{\sim}{X}_i-\bar{\underset{\sim}{X}})(\underset{\sim}{X}_i-\bar{\underset{\sim}{X}})^T$$

$$\hat{\underset{\sim}{\mu}} = \bar{\underset{\sim}{X}} \qquad \text{m.l.est.}$$

$$\frac{\partial\log|A|}{\partial A} = \qquad \frac{\partial A^{-1}}{\partial A} \overset{?}{=} A^{-1}\partial A\, A^{-1\,T}$$

Hotelling's T^2 statistic

$$T^2 = n(\bar{\underline{x}} - \underline{\mu}_0)^T S^{-1} (\bar{\underline{x}} - \underline{\mu}_0)$$

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$$

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

under $N(\mu_0, \Omega)$,

$$T^2 \sim \frac{k}{n-k} \cdot (n-1) \, F_{k, n-k}$$

provides a test of $H_0 : \underline{\mu} = \mu_0$

"omnibus test"

partial and marginal correlation

$\underset{\sim}{X}_i \sim N(\mu, \Omega)$

$(i = 1, \dots, n \text{ iid})$

$\Omega = ((\sigma_{ab}))$

$a, b \in \{1, \dots, k\}$

$\sigma_{ab} = \mathrm{cov}(X_a, \cancel{X}_b) \rightarrow \frac{\sigma_{ab}}{(\sigma_{aa}\sigma_{bb})^{1/2}} = \rho_{ab}$

$= \mathrm{corr}^{n}(X_a, X_b)$

$\underline{\underline{\Omega}} \rightarrow R = \begin{pmatrix} 1 & \rho_{ab} & \\ & \ddots & \\ \rho_{ab} & & 1 \end{pmatrix}$

$R^{-1}_{(ab)} = \underline{\underline{\text{partial}}}\ \underline{\underline{\text{corr}}}^{n}$

of (X_a, X_b), given $\underline{X}_{(a,b)}$

graphical models

Partial corr$^{=}$ and marginal corr$^{=}$ in

MVN $\quad X \sim N(\mu, \Omega)$

$$\rightarrow \underbrace{E(X_a \mid \underset{\sim}{X}_{(-a)})}_{\text{scalar}} = \cdots = \underbrace{\mu_a}_{\text{scalar}} + \underline{\underline{\Omega^T_{a,-a}\Omega^{-1}_{-a,-a}(\underset{\sim}{X}_{(-a)} - \underset{\sim}{\mu}_{(-a)})}}$$

$$\begin{pmatrix} X_1 \\ \underset{\sim}{X}_{-1,1} \\ \vdots \\ X_{-1\,k-1} \end{pmatrix}$$

$$\Omega = \left(\begin{array}{c|c} \sigma_{aa} & \sigma_{a2} \cdots \sigma_{a,k-1} \\ \hline \updownarrow & \Omega_{-a,-a} \end{array}\right)$$

$$\operatorname{var}(X_a \mid \underset{\sim}{X}_{(-a)}) = \sigma_{aa} - \underset{1\times k-1}{\Omega^T_{a,-a}}\ \underset{k-1\times k-1}{\Omega^{-1}_{-a,-a}}\ \underset{k-1\times 1}{\Omega_{-a,a}}$$

"If the cond'l variance of X_a given the others is much less than marginal variance$^{\#}$, then $X_{(-a)}$ carry information about dist$^{=}$ of X_a"

If $\boxed{\sigma_{ab} = 0 \text{ then } X_a \perp\!\!\!\perp X_b}$

(indep't) (marg'l ind)

o.w. $X_a \not\!\perp\!\!\!\perp X_b$

if $\underset{\sim}{X} \sim N_k(\underset{\sim}{\mu}, \Omega)$

text "$\ell\ell\ell$"

(reminder independent $\neq$ uncorrelated unless MVN)

$\Omega \quad \Omega^{-1}$

$((\sigma_{ab})) \quad ((\sigma^{ab}))$

ntbf.

$$\sigma^{ab} = \left(\underline{\sigma_{ab} - \Omega_{a,-(a,b)}\Omega^{-1}_{-ab,-ab}\Omega_{-(a,b),a}}\right)^{-1}$$

$*$ $\overline{\sigma^{ab} = 0} \implies X_a \perp\!\!\!\perp X_b$ given $X_{-(a,b)}$

264 $\Omega_{5\times5}$ $\hat{\Omega}_{5\times5}$ *6 · Stochastic Models*

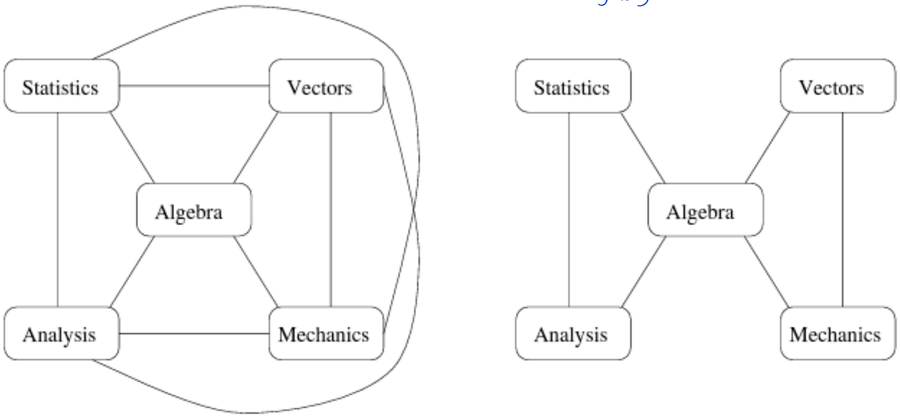


Figure 6.10 Graphs for the full model (left) and a reduced model (right) for the maths marks data. The interpretation of the reduced model is that given the result for algebra, results for vectors and mechanics are independent of those for analysis and statistics.

each variable $a = 1, \dots 5$ is a node on a graph

lines between show "association" bet. 2 or more of them

Multinomial distribution