# MISSING DATA

EX: PIMA INDIANS ; PRECISION MEDICINE

We record:

<u>glu</u> : blood glucose concentration

<u>bp</u> : diastolic blood pressure

<u>skin</u> : skin fold thickness

<u>bmi</u> : body mass index

<u>Question</u>: How do these measurements compare in PIMA pop'n to national avg.?

How do these measurements covary in PIMA pop'n?

$$\underset{4\times1}{Y_i} = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} \sim MVN\left(\underset{4\times1}{\theta}, \underset{4\times4}{\Sigma}\right)$$

vector of measurements for $i^{th}$ individual

Complication: some data are missing.

how to handle?

- throw out missing data? No! Lose a lot of info.
- impute w/ mean of a column? No! Lose all cov-structure

To handle this in a principled Bayesian way, I need to account for the missingness

Let $O_i = \begin{bmatrix} O_{i1} \\ \vdots \\ O_{i4} \end{bmatrix}$ be an observation indicator vector.

$$O_{ij} = 1 \quad \text{if } Y_{ij} \text{ obs.}$$
$$= 0 \quad \text{if } Y_{ij} \text{ missing}$$

Let $Y = Y_1, \dots, Y_n$
$O = O_1, \dots, O_n$

<u>COMPLETE</u> DATA LIKELIHOOD:

$$p(Y, O \mid \theta, \Sigma, \phi)$$
$$= p(O \mid Y, \theta, \Sigma, \phi) \cdot p(Y \mid \theta, \Sigma)$$

Let $Y = [Y_{OBS}, Y_{MIS}]$
$Y_{OBS} = Y[O == 1]$
$Y_{MIS} = Y[O == 0]$

<u>OBSERVED</u> DATA LIKELIHOOD

$$p(Y_{OBS}, O \mid \theta, \Sigma, \phi) = \int p(Y_{OBS}, Y_{MIS}, O \mid \theta, \Sigma, \phi) \, dY_{MIS}$$

ASSUMPTION : DATA are MAR
"missing at random"

$$p(O \mid y, \theta, \Sigma, \phi) = p(O \mid \phi)$$

where $\phi$ does not depend on $\theta, \Sigma$

$$O \perp y$$

---

We are interested in, as Bayesians,

$p(\text{unknowns} \mid \text{knowns})$

$$p(\theta, \Sigma, Y_{mis} \mid O, Y_{OBS})$$

$$\propto \; p(\theta, \Sigma, Y_{mis}, Y_{OBS}, O)$$

$$\propto \; \underbrace{p(Y_{mis}, Y_{OBS} \mid \theta, \Sigma, O)}_{\text{complete data likelihood}} \cdot \underbrace{p(O, \theta, \Sigma)}_{p(\theta, \Sigma \mid O) \cdot p(O)}$$

I want to approx. this posterior. What priors would enable Gibbs sampling?

$\theta \sim MVN(\mu_0; \tau_0^2)$

$\Sigma \sim$ inverse-Wishart$(n_0, S_0)$

Assumption : $p(\theta, \Sigma \mid O) = p(\theta)\, p(\Sigma)$

Gibbs sampling proceeds for each unknown:

$$p(\theta \mid \cdot) = \overbrace{dMVN(\mu_n, \tau_n^2)}^{\text{full cond'l posterior}}$$

$\mu_n, \tau_n^2$ → function of $\underline{\text{data}}$ (complete: $y_{mis}$, $y_{obs}$) and $\mu_0, \tau_0^2$

$$p(\Sigma \mid \cdot) = d\text{inv-wishart}(\nu_n, S_n)$$

$\nu_n, S_n$ → function of complete data & $\nu_0, S_0$

"$\cdot$" means everything

$$p(Y_{mis} \mid \cdot) \propto p(Y_{OBS}, Y_{mis} \mid \theta, \Sigma, O)$$

$$\propto \underbrace{p(Y_{mis} \mid Y_{OBS}, \theta, \Sigma, O)}_{\text{conditional normal}}$$

my full cond'l posteriors are all proportional to the joint.