Missing Data

Ex Pima Indians ; Precision Medicine

We record:

glu : blood glucose concentration

bp : diastolic blood pressure

skin : skin fold thickness

bmi : body mass index

Questions: How do these measurements in the Pima popn compare to national averages?
How do these measurements covary in Pima popn?

$$\underset{4\times 1}{Y_i} = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} \sim MVN\left(\underset{4\times 1}{\theta}, \underset{4\times 4}{\Sigma}\right)$$

↑ vector of measurements for ith indiv.

Complication: some data are missing

How to handle?

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

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

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

$O_{ij} = 1$ if Y_{ij} obs.

$= 0$ if Y_{ij} missing

COMPLETE DATA LIKELIHOOD: Let $Y = Y_1, \ldots, Y_n$

$O = O_1, \ldots, O_n$

$$P(Y, O \mid \theta, \Sigma, \phi)$$

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

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

$Y_{OBS} = Y[O == 1]$

$Y_{MIS} = Y[O == 0]$

obs. matrix

OBSERVED 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 ϕ does not depend on θ, Σ, or y_{mis}.

$$O \perp Y_{mis}$$

MAR: missingness does not depend on missing data whatsoever, but may depend on observed data.

MCAR: "missing completely at random"
$= p(O \mid \phi)$ & ϕ does not even depend on Y_{obs}.

missingness has no dependency on data.

MNAR: missing NOT at random
missingness depends on missing data.

We are interested in, as Bayesians

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

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

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

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

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

$$\theta \sim MVN(M_0, \overline{\Sigma_0})$$

$$\Sigma \sim \text{inv-Wishart}(n_0, S_0)$$

assumption:

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

Gibbs sampling proceeds for each unknown:

$$p(\theta \mid \cdot) = \text{MVN}(\mu_n, \Psi_n)$$

(μ_n, Ψ_n ↓ function of complete data ymis & yobs and μ_0, Σ_0)

$$p(\Sigma \mid \cdot) = \text{Inv-Wishart}(n_n, S_n)$$

(n_n, S_n ↓ function of complete data & n_0, S_0)

$\therefore$ new everything else ↓

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

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

condit normal