Missing Data

Ex: Pima Indians ; Precision Medicine

We record:

glu : blood glucose concentration (mg/deciliter)

bp : diastolic blood pressure (bottom #; pressure in arteries when heart rests btwn beats)

skin : skin fold thickness (mm)

bmi : body mass index (kg/m^2)

Questions: - How do these measurements compare (in Pima popn) to national average?

- How do these measurements relate (covary) w/in the Pima popn?

▷ What kind of model could help answer these questions?

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

Y_i: 4×1 vector of measurements for i^{th} individual

Complication: some data are missing.

How can we handle this?

- throw out missing data? No! Lose a lot of info.
- impute (w/ mean, w/ nearest neighbour)
  - issue: we have uncertainty about the value & throw that out.
  - another issue: we might hurt cov. structure estimation.

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

Let $\underline{O}_i = \begin{bmatrix} O_{i1} \\ \vdots \\ O_{i4} \end{bmatrix}$ be a observation indicator vector.

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

$O_{ij} = 0$ if Y_{ij} missing

Let $\underline{\underline{Y}} = \underline{Y}_1, \ldots, \underline{Y}_n$

$\underline{\underline{O}} = \underline{O}_1, \ldots, \underline{O}_n$

This notation lets us write the

COMPLETE DATA Likelihood:

$$p(\underline{\underline{Y}}, \underline{\underline{O}} \mid \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\Phi})$$

Let $\underline{\underline{Y}} = [\underline{\underline{Y}}_{obs}, \underline{\underline{Y}}_{mis}]$

$\underline{\underline{Y}}_{obs} = \underline{\underline{Y}}[\underline{\underline{O}} == 1]$

$\underline{\underline{Y}}_{mis} = \underline{\underline{Y}}[\underline{\underline{O}} == 0]$

observed data likelihood:

$$p(\underline{\underline{Y}}_{obs}, \underline{\underline{O}} \mid \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\Phi}) = \int p(\underline{\underline{Y}}_{obs}, \underline{\underline{Y}}_{mis}, \underline{\underline{O}} \mid \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\Phi}) \, d\underline{\underline{Y}}_{mis}$$

3 WAYS TO CONSIDER MISSING DATA MECHANISM:

1) MCAR : Missing Completely At RANDOM
   - no dependence on obs or missing data whatsoever.

2) MAR : Missing at random
   - missingness only depends on obs data (but not missing data).
   i.e. $\underline{\theta}_i \perp \underline{Y}_{mis,i} \mid \underline{Y}_{obs,i}$

3) MNAR : missing not at random
   - missingness may depend on missing data.

Assumption we'll make: Data are MAR

Implication:

complete data likelihood:

$$p(\underline{Y}, \underline{\theta} \mid \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\phi}) = \underbrace{p(\underline{\theta} \mid \underline{Y}_{obs}, \underline{Y}_{mis}, \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\phi})}_{p(\underline{\theta} \mid \underline{Y}_{obs}, \underline{\phi})} p(\underline{Y} \mid \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\phi})$$

---

we are interested in, as Bayesians,

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

$$p(\underline{\theta}, \underline{\underline{\Sigma}}, \underline{Y}_{mis} \mid \underline{Y}_{obs}, \underline{\theta}) \propto p(\underline{\theta}, \underline{\underline{\Sigma}}, \underline{Y}_{obs}, \underline{Y}_{mis}, \underline{\theta})$$

$$\propto \underbrace{p(\underline{Y}_{mis}, \underline{Y}_{obs} \mid \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\theta})}_{\text{complete data likelihood}} \cdot \underbrace{p(\underline{\theta}, \underline{\theta}, \underline{\underline{\Sigma}})}_{p(\underline{\theta}, \underline{\underline{\Sigma}} \mid \underline{\theta}) \cdot p(\underline{\theta})}$$

I want to approximate this posterior.

What priors could enable Gibbs sampling?

$\underline{\theta} \sim MVN(\underline{\mu}_0, \underline{\underline{V}}_0)$

$\underline{\underline{\Sigma}} \sim \text{inv-Wishart}(\eta_0, \underline{\underline{S}}_0)$

assumption:

$p(\underline{\theta}, \underline{\underline{\Sigma}} \mid \underline{\theta}) = p(\underline{\theta})\, p(\underline{\underline{\Sigma}})$

Gibbs sampler:

$p(\underline{\theta} \mid \cdot) = dMVN(\underline{\mu}_n, \underline{\underline{V}}_n)$

↓ ↓ functions of complete data $\underline{\underline{Y}}$.

$p(\underline{\underline{\Sigma}} \mid \cdot) = d\text{inv-Wishart}(\eta_n, \underline{\underline{S}}_n)$

$p(\underline{Y}_{mis} \mid \cdot) =$ conditional Normal

↓ can see this easily from data generative model or write out

$p(\underline{Y}_{mis} \mid \cdot) \propto \underbrace{p(\underline{Y}_{mis}, \underline{Y}_{obs} \mid \underline{\theta}, \underline{\underline{\Sigma}}, \underline{\theta})}_{MVN}$