We record 4 measurements for 200 women.

- blood plasma glucose concentration — glu — Y_{i1}
- diastolic blood pressure — bp — Y_{i2}
- skin fold thickness — skin — Y_{i3}
- body mass index — bmi — Y_{i4}

Some of the data is missing e.g. I have

$\vec{Y}_2 = [195 \;\; 70 \;\; 33 \;\; NA]^T$.

(glu, bp, skin, bmi)

I want to know $\theta = [\theta_1 \;\; \theta_2 \;\; \theta_3 \;\; \theta_4]$, the +ve pop'n mean ($\mathbb{E}$ glu, $\mathbb{E}$ bp, $\mathbb{E}$ skin, $\mathbb{E}$ bmi)

AND

I want to know $\underset{4\times4}{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ & & \sigma_{34} & \sigma_{34} \\ & \circ & & \sigma_{44} \end{bmatrix}$

that describes how each measurement covaries w/in the pop'n.

AND

I want to guess at the missing data Y_{miss}.

Overall I want:

$p(\vec{\theta}, \Sigma, Y_{miss} \mid Y_{obs})$ ← posterior distr. (target distr.)

(unknown)

This is difficult to compute so I will ....

sample from it

but its difficult to sample from directly ...

so I will Gibbs sample.

Assumption: Missing at random.

I need:

1. $\vec{\theta} \mid \Sigma, y_{obs}, y_{missing}$ ✓
2. $\Sigma \mid \vec{\theta}, y_{obs}, y_{missing}$ ✓
3. $y_{missing} \mid \vec{\theta}, \Sigma, y_{obs}$?

$p(y_{miss} \mid y_{obs}, \theta, \Sigma) \propto \overbrace{P(y_{miss}, y_{obs} \mid \theta, \Sigma)}^{\text{MVN}}$

It turns out, if we partition y into b (missing) & a observed, (index of elements)

$$\vec{y}_{[b]} \mid \vec{y}_{[a]}, \theta, \Sigma \sim \text{MVN}(\vec{\theta}_{b|a}, \Sigma_{b|a}) \quad \text{where}$$

$$\vec{\theta}_{b|a} = \vec{\theta}_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\vec{y}_{[a]} - \vec{\theta}_{[a]})$$

$$\Sigma_{b|a} = \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}$$

$\theta_{[b]}$: elements of θ corresponding to missing obs.

$\Sigma_{[b,a]}$: elements of Σ corresponding missing rows, obs columns.