STAT 530: Bayesian treatment of missing data

Mar 3, 2010

Recall: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N_p(\theta, \Sigma)$

Binary indicator R_{ij} of whether Y_{ij} is observed or not.

Useful to write $Y_i = (Y_i^{obs}, Y_i^{mis})$.

Bayesian inference via $(\theta, \Sigma | y^{obs})$

Usual latent variable idea, can obtain this by marginalizing $(\theta, \Sigma, y^{mis} | y^{obs})$, where

$$p(\theta, \Sigma, y^{mis} | y^{obs}) \quad \propto \quad p(y^{mis}, y^{obs} | \mu, \Sigma) p(\mu, \Sigma)$$

Gibbs sampling straightforward (Hoff's code on web)

$(\theta | \Sigma, y^{mis}, y^{obs})$ as before

$(\Sigma | \theta, y^{mis}, y^{obs})$ as before

$(y^{mis} | \theta, \Sigma, y^{obs}) \sim MVN()$ (independently across subjects)

First version of the data ($n = 200$, $p = 4$)

```
      glu        bp      skin       bmi
 Min.   : 56.0   38.00    7.00   18.20
 1st Qu.:100.0   64.00   20.75   27.57
 Median :120.5   70.00   29.00   32.80
 Mean   :124.0   71.26   29.21   32.31
 3rd Qu.:144.0   78.00   36.00   36.50
 Max.   :199.0  110.00   99.00   47.90
```

```
      glu   bp skin  bmi
glu  1.00 0.27 0.22 0.22
bp   0.27 1.00 0.26 0.24
skin 0.22 0.26 1.00 0.66
bmi  0.22 0.24 0.66 1.00
```

First version (no missing data!), continued

```
### Hoff's choices for hyperparameters
mu0<-c(120,64,26,26)
sd0<-(mu0/2)
L0<-matrix(.1,p,p) ; diag(L0)<-1 ; L0<-L0*outer(sd0,sd0)
nu0<-p+2 ; S0<-L0
```

Posterior mean and SD for $(\theta_1, \ldots, \theta_4)$

	glu	bp	skin	bmi
mean	124.0	71.27	29.20	32.29
SD	2.3	0.82	0.81	0.41

Second version of the data

```
      glu        bp    skin      bmi
Min.   : 56.0    38.00    7.0   18.20
1st Qu.:100.0    64.00   20.5   27.60
Median :120.0    70.00   29.0   32.80
Mean   :123.6    70.84   29.1   32.21
3rd Qu.:142.0    78.00   36.0   36.48
Max.   :199.0   110.00   99.0   47.90
NA's   : 15.0    23.00   25.0   22.00
```

Posterior mean and SD

glu	bp	skin	bmi
123.5	71.03	29.34	32.18
2.4	0.83	0.91	0.47

Third version of the data

```
      glu        bp    skin      bmi
 Min.   : 56.0    38.00    7.00   18.20
 1st Qu.:100.0    64.00   20.75   31.90
 Median :120.5    70.00   29.00   34.30
 Mean   :124.0    71.26   29.21   34.65
 3rd Qu.:144.0    78.00   36.00   38.25
 Max.   :199.0   110.00   99.00   47.90
NA's    :                        76.00
```

Posterior mean and SD

glu	bp	skin	bmi
124.0	71.22	29.21	33.01
2.3	0.81	0.82	0.56

Fourth version of the data

```
      glu        bp    skin      bmi
Min.   :  56.0    38.00    7.00   22.50
1st Qu.: 100.0    64.00   20.75   33.10
Median : 120.5    70.00   29.00   35.35
Mean   : 124.0    71.26   29.21   35.79
3rd Qu.: 144.0    78.00   36.00   38.25
Max.   : 199.0   110.00   99.00   47.90
       :                        76.00
```

Posterior mean and SD

glu	bp	skin	bmi
123.99	71.22	29.21	34.93
2.26	0.81	0.82	0.42

How was missingness actually introduced (in these phoney examples)?

Scenario 2: $Pr(R_{ij} = 1|Y_i, R_{i,-j}) = 0.8$

Scenario 3: $\text{logit}Pr(R_{i4} = 1|Y_i) = \alpha_0 + \alpha_1 Y_{i3}$

Scenario 4: $\text{logit}Pr(R_{i4} = 1|Y_i) = \alpha_0 + \alpha_1 Y_{i4}$

Postulate that Scenario 4 is fundamentally different than 1, 2, 3, and causes the method to fail. Does the math support this?

Missingness model (generically...)

Model of interest for Y, paraemterized by ϕ

Missingness model for $R|Y$, parameterized by α

Posterior for unobservables given observeables

$$p(\phi, \alpha, y^{mis}|y^{obs}) \quad \propto \quad p(\phi)p(\alpha)p(y^{mis}, y^{obs}|\phi)p(r|y^{obs}, y^{mis}, \alpha)$$

Simplifies if the distribution of R doesn't depend on y^{mis}....

Taxonomy of missing data

MCAR	R doesn't depend on Y at all
MAR	R depends on Y only through Y^{obs}
nonignorable	R depends on Y^{mis}

In last case, forced to:

- commit to modeling $R|Y$
- jointly infer $(\lambda, \alpha, Y^{mis}|Y^{obs})$

Lots of issues surrounding when MAR is a justified assumption, how to model $R|Y$ when not...