Statistical Analysis with Missing Data

Module 6

Bayes Inference

University of Michigan

Objectives

- Gibbs' sampler to simulate posterior distribution of parameters
- Bayesian theory of Multiple Imputation under explicit models
- Bayes generates proper multiple imputations – propagates error in estimating parameters

Gibbs sampling for missing-data problems

$y = (y_{(0)}, y_{(1)}), y_{(0)}$ = observed data, $y_{(1)}$=missing data; assume MAR

Model for full data: $f_Y(y \mid \theta)$; prior: $\pi(\theta)$

Objective: draws $\theta^{(d)}$ from posterior distribution of θ, that is:

$p(\theta \mid y_{(0)}) \propto \pi(\theta) f(y_{(0)} \mid \theta)$

Often easier to draw $\theta \sim p(\theta \mid y_{(0)}, y_{(1)})$, the complete-data posterior distribution, rather than $\theta \sim p(\theta \mid y_{(0)})$

Often easier to draw $y_{(1)} \sim p(y_{(1)} \mid y_{(0)}, \theta)$ rather than $y_{(1)} \sim p(y_{(1)} \mid y_{(0)})$

So, we apply the Gibbs' sampler to $(y_{(1)}, \theta)$:

Gibbs sampler for missing-data

Initial draw of $\theta = \theta^{(0)}$; then draw $y_{(1)}^{(0)} \sim f_Y(y_{(1)} \mid y_{(0)}, \theta^{(0)})$

Let $(\theta^{(t)}, y_{(1)}^{(t)})$ be draws at iteration t. Then for iteration $t+1$ draw:

P step: $\theta^{(t+1)} \sim p(\theta \mid y_{(0)}, y_{(1)}^{(t)})$, posterior for θ with $y_{(1)}^{(t)}$ imputed for $y_{(1)}$

I step: $y_{(1)}^{(t+1)} \sim f_Y(y_{(1)} \mid y_{(0)}, \theta^{(t+1)})$, predictive dn of $y_{(1)}$ given $\theta = \theta^{(t+1)}$

(P for "posterior," I for "imputation."

the order of the P and I steps is not important).

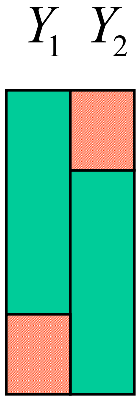
As $t \to \infty, (\theta^{(t)}, y_{(1)}^{(t)})$ converges to a draw from $p(\theta, y_{(1)} \mid y_{(0)})$

After burn-in a, draws $\{\theta^{(a+t)}, t = 1, 2, ...\}$ simulate posterior dn of θ

(Recommended: run 2 or more chains to ensure mixing)

Example: bivariate normal MAR data

- Bivariate normal data with missing data on both variables
- MAR mechanism
- Gibbs' for iteration t consists of an I step and a P step.
- I-Step is like an E step, except that conditional mean is replaced by a draw:



$$\text{missing } y_{i2} : (y_{i2}^{(t+1)} \mid y_{i1}, \theta^{(t)}) \sim_{ind} N(\beta_{20\cdot 1}^{(t)} + \beta_{21\cdot 1}^{(t)} y_{i1}, \sigma_{22\cdot 1}^{(t)})$$

$$\text{missing } y_{i1} : (y_{i1}^{(t+1)} \mid y_{i2}, \theta^{(t)}) \sim_{ind} N(\beta_{10\cdot 2}^{(t)} + \beta_{12\cdot 2}^{(t)} y_{i2}, \sigma_{11\cdot 2}^{(t)})$$

P-Step is like M-Step of EM, with maximization

replaced by draw from complete-data posterior distribution:

$$\Sigma^{(t+1)} \sim \text{Inv}-\text{Wishart}(S^{(t+1)}, n-1)$$

$$\mu^{(t+1)} \mid \Sigma^{(t+1)} \sim N(\overline{x}^{(t+1)}, \Sigma^{(t+1)})$$

Bayes and multiple imputation

Draws $y_{(1)}^{(t)}$ from $p(y_{(1)} \mid y_{(0)})$ can also used to create

multiply-imputed data sets $\left((y_{(0)}, y_{(1)}^{(d)}), d = 1, \ldots D\right)$

E.g. impute missing values $(y_{(1)}^{(a+db)})$ for dth MI dataset,

b chosen so that imputations are roughly uncorrelated

Or run a separate chain for each MI data set.

- The reason is that the MI combining rules are Bayesian: specifically, as I now discuss, they are simulation approximations of the posterior mean and variance under a Bayesian model

MI Inference for a Scalar Estimand

$\theta =$ estimand of interest

$\hat{\theta}_d =$ estimate from dth dataset $(d = 1, \ldots, D)$

The MI estimate of θ is $\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$

$W_d =$ estimate of variance of $\hat{\theta}_d$ from dth dataset

The MI estimate of variance is $T_D = \bar{W}_D + (1 + 1/D)B_D$

$$\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d = \text{Within-Imputation Variance}$$

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta}_D)^2 = \text{Between-Imputation Variance}$$

Bayesian Theory of MI

Model: $f_Y(y \mid \theta) \Rightarrow$ Likelihood $L(\theta \mid y) \propto f_Y(y \mid \theta)$

Prior distribution: $\pi(\theta)$; md mechanism: MAR

$y = \left(y_{(0)}, y_{(1)} \right), y_{(0)}$ = observed data, $y_{(1)}$ = missing data

Complete-data posterior distribution,

if there were no missing values:

$$p(\theta \mid y_{(0)}, y_{(1)}) \propto \pi(\theta)\, f_Y(\, y_{(0)}, y_{(1)} \mid \theta)$$

Posterior distribution given observed data:

$$p(\theta \mid y_{(0)}) \propto \pi(\theta)\, f(y_{(0)} \mid \theta)$$

Theory relates these two distributions ...

Relating the posteriors

- The posterior is related to the complete-data posterior by:

$$p(\theta \mid y_{(0)}) = \int p(\theta \mid y_{(0)}, y_{(1)}) p(y_{(1)} \mid y_{(0)}) dy_{(1)}$$

$$\approx \frac{1}{D} \sum_{d=1}^{D} p(\theta \mid y_{(0)}, y_{(1)}^{(d)}), \text{ where } y_{(1)}^{(d)} \sim p(y_{(1)} \mid y_{(0)})$$

$y_{(1)}^{(d)}$ is a draw from the predictive distribution of the missing values

The accuracy of the approximation increases with D and the fraction of observed data

MI approximation to posterior mean

- Similar approximations for posterior mean and variance yield the MI combining rules given earlier:

$$E(\theta \mid y_{(0)})$$

$$= \int E(\theta \mid y_{(0)}, y_{(1)}) p(y_{(1)} \mid y_{(0)}) dy_{(1)}$$

$$\approx \frac{1}{D} \sum_{d=1}^{D} E(\theta \mid y_{(0)}, y_{(1)}^{(d)}) = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d,$$

where $\hat{\theta}_d =$ is posterior mean from *d*th imputed dataset

MI approximation to posterior variance

$$\text{Var}(\theta \mid y_{(0)}) = E(\theta^2 \mid y_{(0)}) - \left(E(\theta \mid y_{(0)})\right)^2$$

Apply above approx to $E(\theta \mid y_{(0)})$ and $E(\theta^2 \mid y_{(0)})$

Algebra then yields:

$$\text{Var}(\theta \mid y_{(0)}) \approx \bar{V} + B$$

$$\bar{V} = \frac{1}{D}\sum_{d=1}^{D} V_d = \text{ within-imputation variance,}$$

$V_d = \text{Var}(\theta \mid y_{(0)}, y_{(1)}^{(d)})$ is posterior variance from dth dataset

$$B = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta}_D\right)^2 = \text{between-imputation variance}$$

Refinements of MI combining rules for small D

(A): $\mathrm{Var}(\theta \mid y_{(0)}) \approx \bar{V} + \left(1 + 1/D\right) B$

(B) Replace normal reference distribution by t distribution with df

$$\nu = (D-1)\left(1 + \frac{D}{D+1}\frac{\bar{V}}{B}\right)^2$$

(C) For normal sample with variance based on ν_{com} df, replace ν by

$$\nu^* = \left(\nu^{-1} + \hat{\nu}_{\mathrm{obs}}^{-1}\right)^{-1}, \hat{\nu}_{\mathrm{obs}} = (1 - \hat{\gamma}_D)\left(\frac{\nu_{\mathrm{com}} + 1}{\nu_{\mathrm{com}} + 3}\right)\nu_{\mathrm{com}}$$

$$\hat{\gamma}_D = \frac{\left(1 + D^{-1}\right)B}{\bar{V} + \left(1 + D^{-1}\right)B} = \text{estimated fraction of missing information}$$

Logistic regression example revisited

- Imputation Model

  $X_{edi} \sim \text{iid } N(\mu_{ed}, \sigma^2);$

  e=0,1, d=0,1, subject i
- Imputations are draws from the posterior predictive distribution
- Draw σ^2, then μ_{ed} and then missing X_{edi}

Predictive Distributions

- Draw σ^2

$$\sigma^2 \sim \frac{WSS}{\chi^2_{r-4}}$$

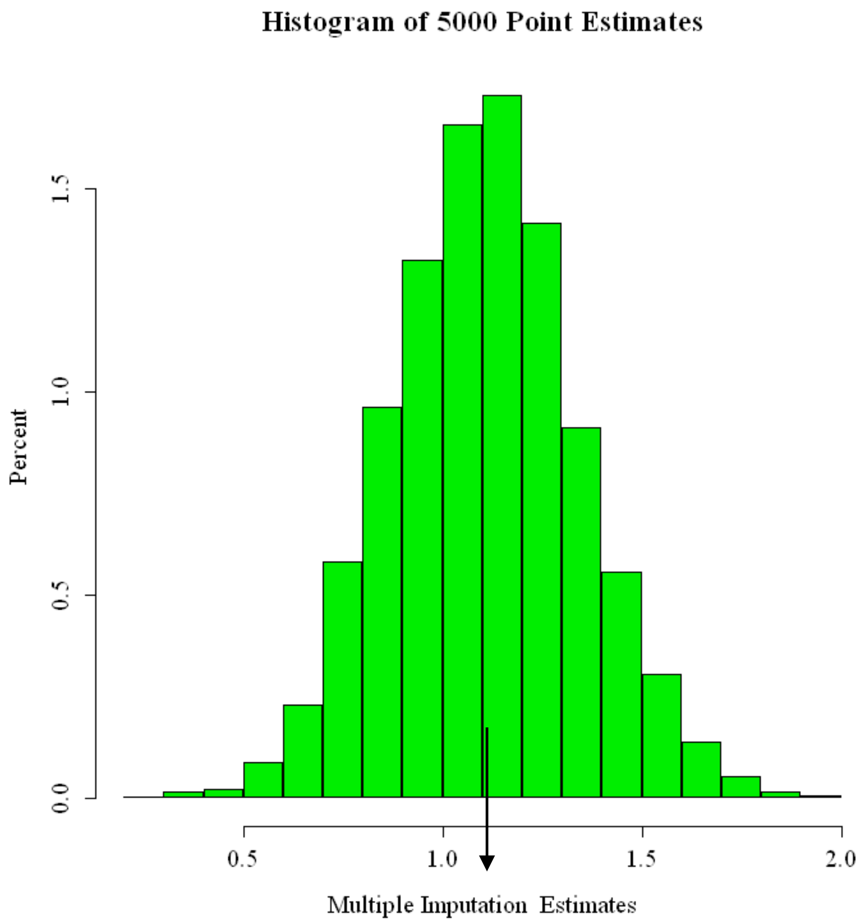
- Draw μ_{ed}

- WSS=Residual sum of squares
- r_{ed}= Number of respondents in cell ed
- $\bar{x}_{ed}$= Mean for cell ed

$$\mu_{ed} | X_{obs}, D, E, \sigma^2 \sim N(\bar{x}_{ed}, \sigma^2 / r_{ed})$$

- Draw $X_{edi} \sim N(\mu_{ed}, \sigma^2)$

Histogram of Multiple Imputation Estimates



- 5 Imputations per missing value
- 5 completed Datasets
- Analyze each separately
- Combine using the formulae given earlier

Coverage and MSE of Various Methods

METHOD	COVERAGE (95% Nominal)	MSE
Before Deletion	*94.68*	*0.0494*
Complete-case	37.86	0.4456
Weighted Complete-case	97.42	0.0538
Hot-Deck Single Imputation	90.28	0.0566
Multiple Imputation	94.56	0.0547

Use of Auxiliary Information in Imputations

- Imputation may involve many more variables though a particular substantive analysis may only use a subset of variables
- Example: Public use data sets or a data set to be used by multiple researchers from different perspectives
- Improve efficiency, reduce bias

Expanded Simulation Study

- Add auxiliary variable: $Z \sim N(0,1)$, Corr(Z, X)=ρ

ρ	Efficiency of MI Using Z compared to Ignoring Z
0.89	1.42
0.71	1.31
0.55	1.21
0.35	1.12
0	0.97

Bayes or MI?

- Gibbs sampler can be used to simulate posterior distribution of parameters under a particular model – no need for MI data sets and combining rules
- However, MI data sets are useful for non-Bayesian analyses, or situations where model from MI differs from analysis model, for example by including variables as predictors that are not in the final model.

Conclusions

- Gibbs sampler useful tool for drawing from the posterior distribution when data are incomplete
- Multiple imputations are a by-product of Gibbs, and can be useful for other analyses
- Other Bayesian simulation methods (SIR, Metropolis-Hastings) can also be useful for handling models where Gibbs is not straightforward