Bayesian Mixture Models for Multiple Imputation

Jerry Reiter

Department of Statistical Science, Duke University

October 14, 2014

General Approaches for MI

- Sequential modeling
  - Estimate a sequence of conditional models
  - Impute from each model, sometimes via Bayesian draws and other times ad hoc (e.g., predictive mean matching)
  - Software: MICE, MI, IVEWARE
- Joint modeling
  - Posit multivariate model (e.g., multivariate normal, loglinear model) for all data
  - Estimate model, usually with Bayesian MCMC methods
  - Impute from conditionals of missing values implied by joint model
  - Software: proc MI, AMELIA II, NORM, CAT

Challenges for Existing Methods

- Sequential modeling
  - Difficult to specify and fit parametric models with high dimensions and complex dependencies (interactions)
  - Not necessarily from coherent joint distribution
- Joint modeling
  - Difficult to specify and fit with high dimensions and complex dependencies (interactions)
  - Typical joint models have restrictive assumptions

Mixture Models as Imputation Engines

Mixture models are widely used in Bayesian (and other types of) inference as flexible models for multivariate data

- Can detect complex structure automatically
- Can scale to large datasets
- Require little tuning by analyst

Two examples discussed in this talk:

- Latent class models for imputation of categorical data (Si and Reiter, 2013; Manrique-Vallier and Reiter, 2013)
- Editing faulty data via mixtures of normal distributions

Categorical Data Imputation

We have n individuals with p variables subject to item nonresponse.
Let $Z_{ij} \in \{1, \ldots, d_j\}$ be value of variable j for individual i.

- Assume each individual i belongs to exactly one of $H < \infty$ latent classes.
- For $i = 1, \ldots, N$, let $s_i \in \{1, \ldots, H\}$ indicate the class of individual i, and let $\pi_h = \Pr(s_i = h)$. $\pi = (\pi_1, \ldots, \pi_H)$ the same for all individuals.
- Within any class, each of the p variables independently follows a class-specific multinomial distribution. For any $z_j \in \{1, \ldots, d_j\}$, let $\psi_{hc_j}^{(j)} = \Pr(Z_{ij} = z_j | s_i = h)$.

Bayesian Latent Class Model

The finite mixture model can be expressed as

$$Z_{ij} \mid s_i, \psi \overset{ind}{\sim} \text{Multinomial}(\psi^{(j)}_{s_i 1}, \ldots, \psi^{(j)}_{s_i d_j}) \text{ for all } i, j \tag{1}$$

$$s_i \mid \pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_H) \text{ for all } i. \tag{2}$$

For prior distributions on π and ψ, we have

$$\pi_h = V_h \prod_{l<h} (1 - V_l) \text{ for } h = 1, \ldots, H \tag{3}$$

$$V_h \overset{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } h = 1, \ldots, H-1, \ V_H = 1 \tag{4}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \tag{5}$$

$$(\psi^{(j)}_{h1}, \ldots, \psi^{(j)}_{hd_j}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}). \tag{6}$$

We set $a_{j1} = \cdots = a_{jd_j} = 1$ for all j, and $(a_\alpha = .25, b_\alpha = .25)$.

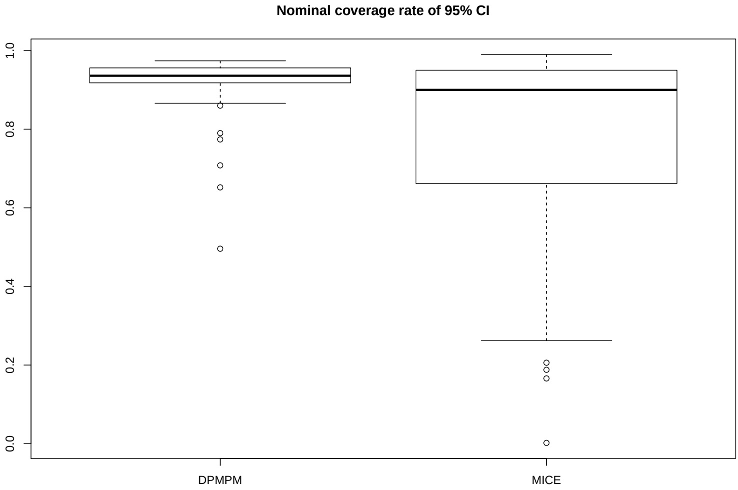
Imputation Algorithm

- Given completed data, sample parameters from full conditionals (all Dirichlet or categorical).
- Given parameter draws, create completed datasets:
  - Draw latent class indicator for each individual from full conditional
  - Draw each missing Z_{ij} from class-specific, independent categorical distributions.
- Computationally efficient since using independent multinomial draws.
- Can enforce structural zeros using ideas of Manrique-Vallier and Reiter (forthcoming).

Some Evidence from Simulation Studies

- Si and Reiter (2013) run repeated sampling simulation studies with $n = 5000$ and $p = 7$ (among others).
- $Z_1, \ldots, Z_5$ generated from loglinear model with all two-way and five three-way interactions.
- Z_6 and Z_7 from logistic regressions with several two-way and three-way interactions.
- (Z_1, Z_2, Z_7) all missing at random via various mechanisms.
- Use latent class model (LC) and MICE with main effects only (a default application) to create $m = 5$ completed datasets for each of 500 runs.
- Estimands: coefficients in log-linear model and logistic regressions (excluding a few 3-way interactions due to sample size issues).

Simulated Coverage Rates of 95% MI Intervals



Average MSE of MI point estimates: .08 for LC model and .13 for MICE.

Editing Faulty Data

Often reported survey data have errors that agency wants to correct before dissemination.

- Categorical data
  - Pregnant males
  - Married eight year olds
- Continuous data
  - Work experience > Age
  - Total salary / Number employees > $1 billion
- Edit and imputation for records with faulty data
  - *Error localization step*: identify set of fields that have errors
  - *Imputation step*: blank and replace these fields with values that satisfy all edit constraints

Edit Rules for Continuous Data

"Edit rule (or shortly edit) is a logical condition to the value of a data field (or variable) which must be met if the data is to be considered correct"[†]

Given observed values of a record $x_{\text{obs}} = \{x_1, \ldots, x_p\}$,

- Range restriction

  e.g, $L_1 \leq x_1 \leq U_1$

- Ratio edit

  e.g, $L_{12} \leq x_1/x_2 \leq U_{12}$

- Balance edit

  e.g, $x_1 = x_2 + x_3$

[†] United Nations Economic Commission for Europe (2000)

How to do edit-imputation?

- Most agencies use variant of Fellegi-Holt (F-H) algorithm:
  - Using optimization techniques, find the minimum number of fields to change to satisfy constraints.
  - Blank and impute, usually via hot deck.
- F-H does not use information about relationships to decide what to replace. Example: if age is 65, replace pregnant rather than male.
- Difficult to find minimum number of fields with balance edits.
- Does not reflect uncertainty in error localization and imputation steps.

Bayesian Data Editing

1. Use a Bayesian approach comprising models for
   1. latent error-free values
   2. latent locations of errors
   3. reported values given error-free values and error locations.
2. Mixture model for the error-free values with support over feasible region
3. Bernoulli distributions for the error locations
4. Measurement error model for reported values
5. Fit model via MCMC to create multiple imputations (or do posterior inference)

Error-Free Value Model $f(x_i|\theta)$

Model for error-free values given inequality constraints $\mathcal{X}$ and n_{bal} balance edits

$$f(x_i|\theta) = f(x_{i,C}|\theta) \cdot \prod_{k=1}^{n_{bal}} I\left[\sum_{j \in C_k} x_{ij} = x_{iT_k}\right] \cdot I[x_i \in \mathcal{X}]$$

1. $x_{i,C} \stackrel{\text{def}}{=} \{x_{ij} : j \in C_k, k = 1, \ldots, n_{bal}\}$: **component variables** modeled by $(p - n_{bal})$-dimensional multivariate distribution
2. $\{x_{iT_k} : k = 1, \ldots, n_{bal}\}$: **sum variables** calculated by balance edits
3. $\mathcal{X}$: the set of convex region with the inequality constraints which all x_i must satisfy

Error-Free Value Model $f(x_i|\theta)$

$$f(x_i|\theta) = f(x_{i,C}|\theta) \cdot \prod_{k=1}^{n_{bal}} I\left[\sum_{j \in C_k} x_{ij} = x_{iT_k}\right] \cdot I[x_i \in \mathcal{X}]$$

The component variables $x_{i,C}$ fit to a mixture of normals with a large number of mixture components:

$$f(x_{i,C}|\theta) \propto \sum_{m=1}^{M} \pi_m N(x_{i,C}; \mu_m, \Sigma_m)$$

Prior for the mixture component weights is

$$\pi_m \sim \text{DirichletProcess}, \quad m = 1, \ldots, M$$

Model for error localizations

For any record i, let $s_i = (s_{i1}, \ldots, s_{ip})$ where

- $s_{ij} = 1$ if variable j is in error and will be blanked and imputed,
- $s_{ij} = 0$ if variable j is not in error and will be released without alteration.
- Example: $s_i = (0, 1, 0)$ means field two is in error and will be replaced.

Model for s_i for all i:

$$
\begin{aligned}
s_{ij} &\sim \text{Bernoulli}(r_j) \\
r_j &\sim \text{Beta}(\alpha_j, \beta_j)
\end{aligned}
$$

where (α_j, β_j) reflects *a priori* knowledge about reliability of variable j. In MCMC check if proposed s_i offers a feasible solution via linear programming.

Measurement Error Model $f(x_{\text{obs},i}|x_i, s)$

Given $x_i = (x_{i1}, \ldots, x_{ip})$ and feasible $s_i = (s_{i1}, \ldots, s_{ip})$, model reported values $x_{\text{obs},i} = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})$ with

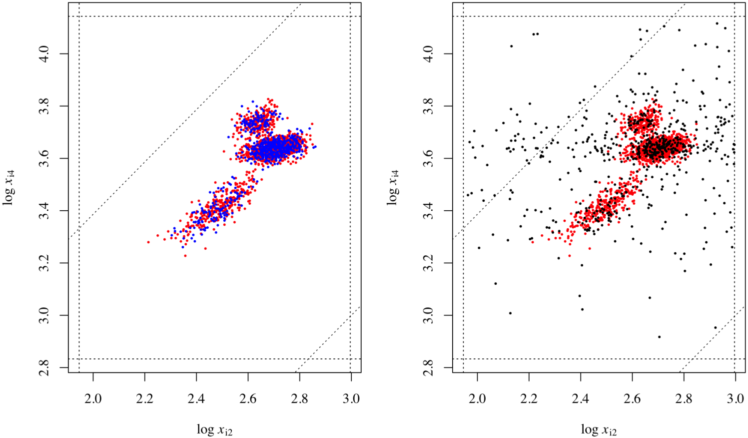
$$f(x_{\text{obs},i}|x_i, s_i) = f\left(x^1_{\text{obs},i}|x_i\right) \prod_{\{j:s_{ij}=0\}} I\left[\tilde{x}_{ij} = x_{ij}\right]$$

1. $x^1_{\text{obs},i} \stackrel{\text{def}}{=} \{\tilde{x}_{ij} : s_{ij} = 1, j = 1, \ldots, p\}$: erroneous variables
2. $f\left(x^1_{\text{obs},i}|x_i\right)$: $(\sum_j s_{ij})$-dimensional density for erroneous variables reflecting the measurement error generating process

Simulation Study

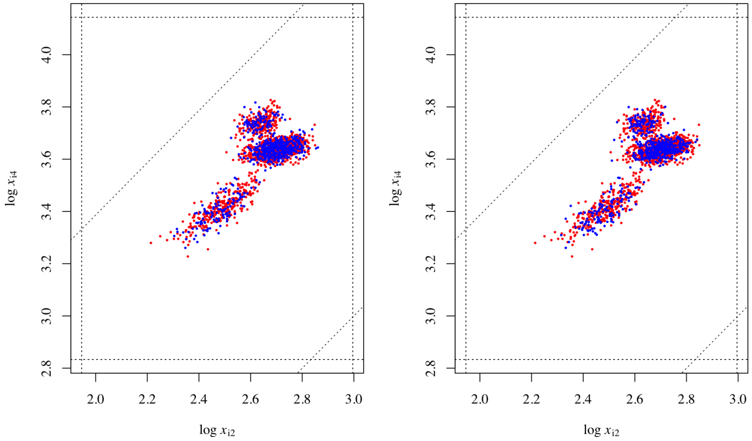
- We introduce edits:
  - range restrictions for each variable, e.g., $L_1 \leq X_1 \leq U_1$
  - ratio edits for some pairs of variables, e.g., $L_{12} \leq X_1/X_2 \leq U_{12}$
  - $q = 2$ balance edits: $X_4 = X_1 + X_2 + X_3$ and $X_7 = X_5 + X_6$
- Generate $n = 2000$ error-free values of $x_i = (x_{i1}, \ldots, x_{i8})$ from
  - mixture of normals for component variables $\{x_{i1}, x_{i2}, x_{i3}, x_{i5}, x_{i6}, x_{i8}\}$
  - balance edits for sum variables $\{x_{i4}, x_{i7}\}$
- For 1000 out of 2000 records, introduce edit-failing records $x_{\text{obs},i} (\neq x_i)$ which are uniformly distributed over a compact region
- We compare
  1. Bayesian editing method
  2. Multivariate imputation method with the minimal changes criterion

Simulated Error-Free values x_i and Observed Values $x_{\mathrm{obs},i}$



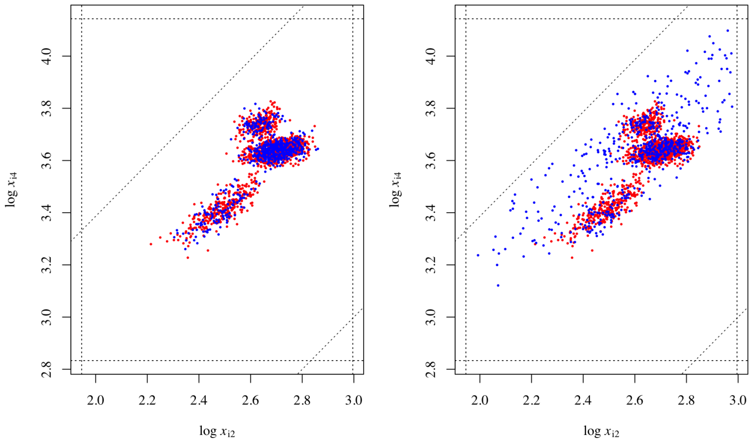
- Left panel: error-free values, x_i
- Right panel: observed edit-failing records (black) and observed edit-passing records (red)

1. Bayesian Editing Method



- Left panel: error-free values, x_i
- Right panel: imputed values (blue) and unchanged values (red)

2. Multivar. Imputation Under Minimal Changes Criterion



- Left panel: error-free values x_i
- Right panel: imputed values (blue) and unchanged values (red)

Summary

- Mixture models can offer flexible approaches to generating multiple imputations from coherent joint distributions.
- My experience: the more variables you have, the more data you need to capture finer features of the joint distribution.
- Some promising research directions:
  - Joint modeling of continuous and categorical data
  - Dealing with high-dimensional continuous data
- It would be very informative to run a bake off between a joint modeling approach and a flexible sequential imputation routine, like sequential CART, on genuine data.