Imputation of multivariate continuous data with non-ignorable missingness

Thais Paiva    Jerry Reiter

Department of Statistical Science
Duke University

NCRN Meeting Spring 2014

May 23, 2014
Outline

1. Introduction
2. Methodology
3. Simulation Study
4. Real Data application
5. Conclusions
Adaptive Design  ➔ In an ongoing survey, decide to:
1. stop the data collection or
2. invest on collecting more data.
If decide to stop, impute the missing data based on the observed data.
If decide to continue, collect an extra wave and impute the remaining.

$$N = n_R + n_{FUS} + n_{NR}$$
If decide to continue, collect an extra wave and impute the remaining.

\[ N = n_R + n_{FUS} + n_{NR} \]
How to decide to stop or not?

**Information measure**
How different is the non-respondents distribution from the respondents?

**Cost measure**
How much does it cost to collect more data and what is the budget?
We need to consider the hypothesis that the non-respondents are **Missing Not At Random**.
Assume that the non-respondents have a different distribution than the respondents.

For now, we are considering only *unit non-response*, but the method could be adapted to deal with item non-response as well.
Model for the observed data

- Continuous multivariate data
- The variables are likely correlated and with heavily skewed distributions
- The model has to be flexible to capture any distributional features from the data
Model for the observed data

- Mixture of multivariate normal distributions
- Dirichlet Process prior to allow for more flexibility and better density estimation (Ishwaran and James, 2001)
Dirichlet Process Mixture Model

\( Y_n = y_1, \ldots, y_n \) \( n \) complete \( p \)-dimensional observations. Assume each variable is standardized.

\( z_i \in 1, \ldots, K \) component indicator of \( i \)-th observation, with probability \( \pi_k = P(z_i = k) \)

Each component \( k \) follows a MVN distribution \( N(\mu_k, \Sigma_k) \)

**Mixture model:**

\[
\begin{align*}
    y_i | z_i, \mu, \Sigma & \sim N(y_i | \mu_{z_i}, \Sigma_{z_i}) \\
    z_i | \pi & \sim \text{Multinomial}(\pi_1, \ldots, \pi_K)
\end{align*}
\]

Marginal mixture model: \( p(y_i | \mu, \Sigma, \pi) = \sum_{k=1}^{K} \pi_k \ N(y_i | \mu_k, \Sigma_k) \)
Prior specification

With **conjugate priors** (Kim et al., 2014), the posterior samples can be obtained using a Gibbs sampler (Ishwaran and James, 2001).

**Components:**

\[ \mu_k | \Sigma_k \sim N(\mu_0, h^{-1} \Sigma_k) \]
\[ \Sigma_k \sim IW(f, \Phi) \]

\[ \Phi = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_p \end{bmatrix} \text{ with } \phi_j \sim \text{Gamma}(a_\phi, b_\phi) \]
\[ a_\phi = b_\phi = 0.25 \quad \mu_0 = 0 \]
\[ df: f = p + 1 \quad h = 1 \]

**Stick-breaking representation for the weights:**

\[ \pi_k = v_k \prod_{g < k} (1 - v_g) \text{ for } k = 1, \ldots, K \]
\[ v_k \sim \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K - 1; \ v_K = 1 \]
\[ \alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \]

\[ a_\alpha = b_\alpha = 0.25 \]
Imputation under MNAR

**MAR**

**Generate impute data from the posterior predictive distribution**

Respondents: $D_R$

Non-respondents: $D_{NR}$

**Imputation with non-ignorable missingness**
Imputation under MNAR

MNAR

Generate impute data from the altered posterior predictive distribution

Respondents

$D_R$

Non-respondents

$D_{NR}$

mixture model

reflect a hypothesis for the non-respondents pattern

$\mu$

$\Sigma$

$\pi^*$
Imputation under MNAR

MNAR

Generate impute data from the altered posterior predictive distribution

Respondents $D_R$

Non-respondents $D_{NR}$

mixture model

Imputation

$\mu$, $\Sigma$, $\pi^*$

Reflect a hypothesis for the non-respondent pattern.
If MNAR is being considered, it is likely that the missing data will have more extreme values than the observed.

We need to leverage the weights of the clusters on the tails.

Rank the components based on the distance to the origin $\mu' \mu$

- post-simulation
- only non-empty components are considered
Changing the mixture weights

Many ways to choose the new weights $\pi^*$:
- set to fixed values;
- rescale based on the posterior samples;
- sampled from a random distribution;
- incorporate information from auxiliary variables, etc.

With a moderate number of components:
- fix the new values for $\pi^*$

As the number of components increases, it becomes harder:
- choose a subset of components
Selecting posterior samples

**Multiple Imputation:** Select $m$ samples from the MCMC iterations

- If the cluster allocations are similar across the $m$ samples, specify overall probabilities and proceed with standard MI methods.

- Otherwise, summarize the samples by selecting the sample that has the largest posterior value (Fraley and Raftery, 2007).
Simulation Study

**Toy example:** The true complete data distribution can be recovered if the missing data mechanism is known.

Repeat 500 times:
1. Generate complete data (observed and missing)
2. Fit the mixture model to the observed data
3. Set $\pi^*$ to the true missing proportions
4. Generate $m = 5$ imputed data sets under MAR and MNAR
Toy example

Inference on:

- **complete** complete original data sets (no missing data)
- **observed** original data sets with just the observed responses
- **MNAR** observed + multiple imputed data sets under MNAR (combining rules from Reiter (2003))
- **MAR** observed + multiple imputed data sets under MNAR (combining rules from Reiter (2003))
Inference on:

- Marginal means
- Linear regression coefficients

<table>
<thead>
<tr>
<th>Coverage rates:</th>
<th>$\bar{y}_1$</th>
<th>$\bar{y}_2$</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.92</td>
</tr>
<tr>
<td>MNAR</td>
<td>0.99</td>
<td>0.99</td>
<td>0.96</td>
<td>0.89</td>
</tr>
<tr>
<td>observed</td>
<td>0.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>MAR</td>
<td>0.00</td>
<td>0.00</td>
<td>0.93</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Thais Paiva, Jerry Reiter

Imputation with non-ignorable missingness

May 23, 2014 20 / 29
**Colombian Annual Manufacturing Survey in 1991 (N=6609)**

- **Variables:** RVA (real value-added), RMU (real material used in products) and CAP (capital in real terms).

- **Missing data indicator:**

  \[ R_i \sim Bern(\theta_i) \quad \text{where} \quad \theta_i = \logit^{-1}(\beta_0 + \beta_1 Y_i) \]

  \( \beta_0 \) and \( \beta_1 \) are fixed such that the plants with larger quantities are more likely to not respond.
The values are log transformed and positively standardized.
Clusters from the iteration with maximum posterior with default priors

RVA

RMU

CAP

Clusters

- 1 (\(\pi = 0.25\))
- 2 (\(\pi = 0.05\))
- 3 (\(\pi = 0.07\))
- 4 (\(\pi = 0.4\))
- 5 (\(\pi = 0.01\))
- 6 (\(\pi = 0.22\))
Imputed data from the top cluster only
Results with default prior are not flexible enough

Change prior (fix covariance matrices to enforce smaller clusters)
Sensitivity Analysis

Cluster Probabilities:

- $\pi_1^*: 0.044$
- $\pi_2^*: 0.096$
- $\pi_3^*: 0.176$
- $\pi_4^*: 0.228$
- $\pi_5^*: 0.012$
- $\pi_6^*: 0.124$
- $\pi_7^*: 0.192$
- $\pi_8^*: 0.1$
- $\pi_9^*: 0.012$
- $\pi_{10}^*: 0.016$
- $\pi_{11}^*: 0.0$

RVA

RMU

CAP

Clusters:
- 1 ($\pi^* = 0.04$)
- 7 ($\pi^* = 0.19$)
- 2 ($\pi^* = 0.1$)
- 8 ($\pi^* = 0.1$)
- 3 ($\pi^* = 0.18$)
- 9 ($\pi^* = 0.01$)
- 4 ($\pi^* = 0.23$)
- 10 ($\pi^* = 0.02$)
- 5 ($\pi^* = 0.01$)
- 11 ($\pi^* = 0$)

Legend:
- observed
- imputed
**Conclusions**

**Imputation under MNAR:**

- Flexible model that is able to capture different features of the data
- Under MNAR, the missing data distribution is unknown. The method works for different levels of prior information
- Interface to facilitate Sensitivity Analysis

**Next steps: Adaptive Design**

- *Information measure*: based on propensity scores to compare data sets imputed under different scenarios
- *Cost function* and *Stopping rule*
Thank you!

tvp@stat.duke.edu

