Bayesian Methods for Incomplete Data

Source: Chapter 5, Handbook of Missing Data Methodology
Authors: Michael J. Daniels, Joseph W. Hogan

Presentor: Suchit Mehrotra (smehrot@ncsu.edu)

April 24, 2015

Quick Overview of Bayesian Inference

- In Bayesian inference, the parameter $\boldsymbol{\theta}$ is considered a random variable.
- This means that it can be described via a distribution.
- The primary method for inference in the Bayesian paradigm is the posterior distribution of $\boldsymbol{\theta}$ conditioned on the data $\boldsymbol{z}$.
- $p(\boldsymbol{\theta})$ is the prior distribution of parameter.
    - This represents the information about the parameter we bring into the problem.
- The full posterior distribution, using Bayes rule, is written as:

$$p(\boldsymbol{\theta}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{z})}$$

Notation

- Let $Z_i : i = 1, \ldots, N$ be a univariate response.
- Suppose that only a subset of the N responses is observed.
- Let R_i be a random variable which denotes whether or not the i^{th} response is observed.
    - $R_i = 1$ when Z_i is observed and $R_i = 0$ when it is not.
- Then the full data are $(\boldsymbol{Z}, \boldsymbol{R})$.
- Denote the observed part of $\boldsymbol{Z}$ as $\boldsymbol{z}_{(r)}$ and the unobserved as $\boldsymbol{z}_{(\bar{r})}$

An Example - Set Up

- Let $Z_i | R_i = r \sim N(\mu_r, \sigma^2)$
- Let $E(Z) = \mu$ which can be written as:

$$\mu = \mu_0 \, \theta + \mu_1 \, (1 - \theta)$$

- Where:

$$\begin{aligned} \mu_0 &= E(Z|R=0) \\ \mu_1 &= E(Z|R=1) \\ \theta &= P(R=0) \end{aligned}$$

An Example - Importance of Priors

- Let the priors for μ_0 and μ_1 be:

$$p(\mu_0) \sim \mathcal{N}(a_0, \tau_0)$$
$$p(\mu_1) \sim \mathcal{N}(a_1, \tau_1)$$

- Then the posterior of μ is:

$$p(\mu | \boldsymbol{z}_{(r)}, \boldsymbol{r}) = \theta a_0 + (1 - \theta)(B \bar{z}_1 + (1 - B) a_1)$$

- Where:

$$\bar{z}_1 = N_1^{-1} \sum_{i:R_i=1} z_i$$
$$B = \frac{\tau_1}{\tau_1 + \sigma^2 / N_1}$$
$$N_1 = \sum_{i=1}^{N} R_i$$

An Example - Importance of Priors

$$p(\mu|\boldsymbol{z}_{(r)}, \boldsymbol{r}) = \theta a_0 + (1-\theta)(B\bar{z}_1 + (1-B)a_1)$$

- As $N \to \infty$, the impact of $p(\mu_1) \to 0$ but $p(\mu_0)$ has a non-zero weight $\theta = P(R=0) > 0$.
- Additionally:

$$Var(\mu|\boldsymbol{z}_{(r)}, \boldsymbol{r}) = \theta^2 \tau_0 + (1-\theta)^2 \left(\frac{1}{\tau_1} + \frac{N_1}{\sigma^2}\right)^{-1}$$

- Therefore, as $N \to \infty$, $Var(\mu|\boldsymbol{z}_{(r)}, \boldsymbol{r}) \to \theta^2 \tau_0$.

An Example - Importance of Priors

$$Var(\mu|\mathbf{z}_{(r)}, \mathbf{r}) = \theta^2\tau_0 + (1-\theta)^2\left(\frac{1}{\tau_1} + \frac{N_1}{\sigma^2}\right)^{-1}$$

- This means that there is a lower bound on the standard deviation of μ, $\theta\sqrt{\tau_0}$
- The MAR assumption is equivalent to the prior:

$$p(\mu_0|\mu_1) = I\{\mu_0 = \mu_1\}$$

- Therefore, strong prior information is required for inference with missing data.

Types of Missingness Patterns

- Missing Completely at Random (MCAR):

$$pr(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{Z}) = pr(\boldsymbol{R} = \boldsymbol{r}) \iff \boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Z}$$
$$pr(\boldsymbol{R} = \boldsymbol{r}) = \pi(\boldsymbol{r}) \text{ is a constant.}$$

- Missing at Random (MAR):

$$pr(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{Z}) = pr(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{Z}_{(r)}) = \pi(\boldsymbol{Z}_{(r)}, \boldsymbol{r})$$
$$\pi(\boldsymbol{Z}_{(r)}, \boldsymbol{r}) \text{ depends only on the observed data.}$$

- Missing Not at Random (MNAR):

$$pr(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{Z}) = \text{depends on unobserved data.}$$

- These assumptions are not verifiable from the data.

Naive Methods

- Complete Cases
  - Drop all of the observations where all of the variables have not been observed.
- Available Cases
  - In longitudanal settings, keep patients with some missing values and include only those points in the model.
- Last Observation Carry Forward (LOCF)
  - Assume that the last observed point is the next observed point and continue this until real data is observed.
- Single Imputation
  - Plug in the missing values one time using regressions of variables with missing data onto observed variables.

Full Data Joint Densities

- The ideal full data are $(\boldsymbol{R}, \boldsymbol{Z})$, where $\boldsymbol{Z} = (\boldsymbol{Z}_{(r)}, \boldsymbol{Z}_{(\bar{r})})$
- The joint density of these two variables can be factored in multiple ways:
  - Selection Model Factorization:

$$p(\boldsymbol{r}, \boldsymbol{z}) = p(\boldsymbol{r}|\boldsymbol{z})\, p(\boldsymbol{z})$$

  - Pattern Mixture Factorization:

$$p(\boldsymbol{r}, \boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{r})\, p(\boldsymbol{r})$$

  - Extrapolation Factorization:

$$p(\boldsymbol{r}, \boldsymbol{z}) = p(\boldsymbol{z}_{(r)}, \boldsymbol{r})\, p(\boldsymbol{z}_{(\bar{r})}|\boldsymbol{z}_{(r)}, \boldsymbol{r})$$

Bayesian Ignorability

- Using parameters $\boldsymbol{\theta}$, the selection model factorization is:

$$p(\boldsymbol{r}, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{x}, \psi(\boldsymbol{\theta}))\, p(\boldsymbol{z}|\boldsymbol{x}, \gamma(\boldsymbol{\theta}))$$

- In the Bayesian paradigm, ignorability holds under three conditions:
  - When data are MAR.
  - The parameters in $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}$ are distinct.
  - A priori independence: $p(\boldsymbol{\theta}) = p(\gamma, \boldsymbol{\psi}) = p(\gamma)\, p(\boldsymbol{\psi})$.
- If ignorability holds, the posterior distribution of $\boldsymbol{\theta}$ is:

$$p(\boldsymbol{\theta}|\boldsymbol{z}_{(r)}, \boldsymbol{r}, \boldsymbol{x}) = p(\gamma|\boldsymbol{z}_{(r)}, \boldsymbol{x})\, p(\boldsymbol{\psi}|\boldsymbol{z}_{(r)}, \boldsymbol{r}, \boldsymbol{x})$$

- This means that the posterior distribution of $\boldsymbol{\gamma}$ only depends on the observed data.

$$p(\gamma|\boldsymbol{z}_{(r)}, \boldsymbol{x}) \propto p(\boldsymbol{z}_{(r)}|\gamma, \boldsymbol{x})\, p(\gamma)$$

Justification for Multiple Imputation

- Under ignorability, the extrapolation distribution, $p(\boldsymbol{z}_{(\bar{r})}|\boldsymbol{z}_{(r)}, \boldsymbol{x}, \boldsymbol{r})$ simplifies to:

$$p(\boldsymbol{z}_{(\bar{r})}|\boldsymbol{z}_{(r)}, \boldsymbol{x})$$

- This means that the missing data can be imputed from the extrapolation distribution, and a full data analysis can be conducted.
- The classical way to impute the data set is via Bayesian proper imputation (Rubin, 1987).
- Another method that is frequently used is Multiple Imputation via Chained Equations.

Multiple Imputation

- Suppose the full data is $(\boldsymbol{R}, \boldsymbol{Z}_{(r)})$ and a we have posited a likelihood for the full data, $p(\boldsymbol{z})$.
- Donald Rubin outlines multiple imputation as three basic "tasks":
  - Each missing value, is filled in (imputed) M times to create M different data sets.
  - Analyze each data set as if all the observations were observed.
  - The results for the M different analyses are combined into a single analysis by taking into account the variation in imputation.
- The imputations are conducted by first drawing $\theta^{(m)}$ from $p(\theta|\boldsymbol{r}, \boldsymbol{z}_{(r)})$.
- Then draw the missing part of the data from the posterior predictive distribution given the parameter previously drawn, $p(\boldsymbol{z}_{(\bar{r})}|\boldsymbol{z}_{(r)}, \boldsymbol{x}, \theta^{(m)})$
- The estimator for θ is: $\hat{\theta}^* = M^{-1} \sum_{m=1}^{M} \hat{\theta}^{(m)}$

Multiple Imputation by Chained Equations

- One of the ways that missing data are imputed in practice is to assume that everything is normal, and impute based on that.
- An alternative approach is multiple imputation by chained equations (van Buuren, 2007). Let the data set be $\boldsymbol{Z} = \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$.
  - Initialize the algorithm by randomly sampling from the observed data to fill in missing data. If Y_{ik} is missing, impute it with a randomly sampled $\boldsymbol{Y}_k$ from the observed data.
  - Pick any variable with missing values, say $\boldsymbol{Y}_1$, and regress on all other variables using only the observations where $\boldsymbol{Y}_1$ is observed.
  - Then regress $\boldsymbol{Y}_2$, on all other values including $\boldsymbol{Y}_1$.
  - Cycling through all variables is considered one cycle, and repeat for 10-20 cycles to get a single imputed data set.
  - Repeat this procedure M times to create M imputed data sets.
- It should be noted that order in which the variables are imputed is important.

Software for Multiple Imputation

- Packages and procedures exist in both R and SAS to conduct these types of analyses.
- Refer to the examples on Dr. Marie Davidian's webpage for ST 790 for both SAS and R. They are extremely clear and easy to follow.
- *proc mi* and *proc mianalyze* in SAS can conduct these analyses relatively quickly.
- van Buuren (2011) in the Journal of Statistical Software outlines the *mice* package in R to do multiple imputation by chained equations.
- *mice* allows the option to use a variety of regression methods for imputation such as regression trees, random forests, LDA, etc.

Bayesian Non-Ignorability

- If ignorability does not hold, then we have *non-ignorable* missingness.
- In this situation, instead of using $p(\boldsymbol{z}_{(r)}|\gamma, \boldsymbol{x})$, the full data model, $p(\boldsymbol{z}, \boldsymbol{r}|\boldsymbol{\theta})$ has to be specified.
- The extrapolation distribution, $p(\boldsymbol{z}_{(\bar{r})}|\boldsymbol{z}_{(r)}, \boldsymbol{x})$ is unidentifiable without strong assumptions:
  - Assumptions about the missingness that cannot be verified.
  - Modelling assumptions.
  - Informative priors on the parameters of the extrapolation distribution, $\boldsymbol{\theta}_E$.

Mixture Models

- Under the pattern mixture factorization the joint distribution of the full data can be factorized as:

$$p(\boldsymbol{z}, \boldsymbol{r}) = p(\boldsymbol{z}_{(r)}, \boldsymbol{z}_{(\bar{r})} | \boldsymbol{r})\, p(\boldsymbol{r}) = p(\boldsymbol{z}_{(\bar{r})} | \boldsymbol{z}_{(r)}, \boldsymbol{r})\, p(\boldsymbol{z}_{(r)} | \boldsymbol{r})\, p(\boldsymbol{r})$$

- $p(\boldsymbol{z}_{(r)} | \boldsymbol{r})\, p(\boldsymbol{r})$ is identified from the data.
- Assumptions have to be made about $p(\boldsymbol{z}_{(\bar{r})} | \boldsymbol{z}_{(r)}, \boldsymbol{r})$.
- A possible assumption made is of complete case missing values.
  - The density for the missing part of $\boldsymbol{Z}$ is the same as the density for the observed data.
- Other assumptions are available case and neighboring case missing values.
- Unfortunately, none of the assumptions that are made are verifiable from the data.

Summary

- Missing data cannot be ignored in an analysis.
- Strong, unverifiable, assumptions are required to conduct analysis with missing data.
- Priors clearly incorporate these assumptions as part of the model.
- If ignorability can be assumed, the analysis can be done with only the observed data.
- Multiple Imputation methods can be used to fill in the data, but they must be properly analyzed by taking into account the variance between the estimates.
- Different methods for dealing with missing data can lead to different conclusions.