A Short Note on the Multiple Imputation Framework

Nandana Sengupta

October 16, 2015

1 Overview of Technique

Let the complete dataset, the observed data and the missing data be denoted by Y, Y_{obs} and Y_{miss} respectively. The key assumptions in this literature is that the data is 'Missing At Random' (MAR) i.e the missingness depends only on the observed data Y_{obs} and is independent of the missing data Y_{miss}. The true value of the statistic of interest is Q and its estimate from the complete dataset is denoted by $\widehat{Q}$.

The Multiple Imputation technique (Rubin (1987)) aims to represent statistical uncertainty involved in missing data imputation. The technique involves three steps:

- **Imputation:** 'm' complete datasets are created by filling in missing values in the incomplete datasets m times. Typically the value of m is between 5 and 10.
- **Analysis:** The m complete datasets are used to perform the main analysis of interest leading to m estimates $\widehat{Q}^{(l)}$ with estimated variance $U^{(l)}$, $\quad l = \{1, 2, \cdots, m\}$.
- **Pooling:** The m estimates are averaged to get the final estimate $\overline{Q}_m$:

$$\overline{Q}_m = \frac{1}{m}\sum_{i=1}^{m} \widehat{Q}^{(l)}.$$

2 Pooling Imputations

The total estimated variance of $\overline{Q}_m$ is

$$T_m = \frac{B_m}{1+m} + U_m,$$

which is composed of the 'within-imputation' and 'between-imputation' variances:

$$\overline{U_m} = \frac{1}{m}\sum_{l=1}^{m} U^{(l)}, \qquad B_m = \frac{1}{m}\sum_{l=1}^{m}\left(\widehat{Q}^{(l)} - \overline{Q}_m\right)^2.$$

Starting with the standard normal assumption on the complete data estimate $\widehat{Q}$ with estimated variance U

$$(\widehat{Q} - Q)/\sqrt{U} \sim N(0,1),$$

Rubin (1987) shows that if the imputation technique is 'proper' (discussed in the next section) then the tests and confidence intervals for the pooled estimate $\overline{Q}_m$ follow a t-distribution:

$$(\overline{Q}_m - Q)/\sqrt{T_m} \sim t_\nu$$

with degrees of freedom

$$\nu = (m-1)\left[1 + \frac{\overline{U_m}}{(1+m^{-1})B_m}\right]^2.$$

3 'Proper' Imputations

Rubin (1996) lays out a number of technical conditions under which imputations are 'proper'. One of the more easily interpretable versions of the conditions is

$$(\widehat{Q} - \overline{Q}_{\infty})/\sqrt{B_{\infty}} \sim N(0,1).$$

In other words, the pooled estimate from the imputed datasets when the number of imputations appraches ∞ should approach the estimate from the complete dataset.

Although Rubin argues that pooling of proper imputations is a principled way to account for statistical uncertaintly in imputations, in practice the only methods he proposes to obtain such imputations is via Bayesian models of the data: 1) a parameteric model of the complete data is assumed (typically $Y \sim N(\mu, \Sigma)$), 2) a prior is applied to the unknown parameters and updated till convergence, 3) finally m independent draws from the distribution of Y_{miss} conditional on Y_{obs} make up the m imputed datasets.

4 Criticisms, Response and Current State of Literature

Given the popularity of Multiple Imputation in the social sciences and the dependence of Multiple Imputation on Bayesian models, the missing value literature is mainly restricted to Bayesian models of the complete data. There have been a number of authors who have criticized the framework and have suggested competiting imputation techniques (notably Rao and Shao (1992), Fay (1991, 1996), Nielsen (2003)).

However, Rubin (1996) argues strongly against most criticisms of the technique. Other influential authors such as Schafer (1997), Meng (1994) and King (2012) also lend support to the framework. The current state (and implementation) of the literature on missing values in the social sciences therefore remains firmy centered around the Multiple Imputation framework, and as a result around Bayesian models of the data.

In terms of software packages the two most popular ones are 'Amelia' (Blackwell, King, Honaker (2012)) which uses Expectation-Maximization and 'MICE' (Stef Van Burren (2011)) which uses Chained Equations in order to estimate the parameters of the Likelihood Function. These packages are both available in *R*.

5 Discussion/Section in our current paper

So far in our analysis, we have compared single imputations from 'Amelia', 'MICE' and 'LowRankModels' by assuming that some proportion of the observed datapoints is missing and comparing the error rates for these datapoints across the different techniques. We have found that 'LowRankModels' imputations dominate in terms of cross-validation error.

However, in order to get our results closer to the mainstream in social science, we need to consider the analysis stage (after the completed dataset is obtained). In particular it will be useful to include a discussion on:

- Extending the 'LowRankModels' techniques to the Multiple Imputation framework.
- Including a discussion on variance estimation of the parameters of interest in the final analysis.