Computational Statistics and Data Analysis 92 (2015) 84–96

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Using mixtures of t densities to make inferences in the presence of missing data with a small number of multiply imputed data sets

S. Rashid[a,*], R. Mitra[a], R.J. Steele[b]

[a] University of Southampton, UK
[b] McGill University, Canada

ARTICLE INFO

Article history:
Received 25 September 2014
Received in revised form 29 May 2015
Accepted 29 May 2015
Available online 19 June 2015

Keywords:
Missing data
Multiple imputation
Bayesian statistics
Disclosure avoidance
Mixture distribution
Monte Carlo

ABSTRACT

Strategies for making inference in the presence of missing data after conducting a Multiple Imputation (MI) procedure are considered. An approach which approximates the posterior distribution for parameters using a mixture of t-distributions is proposed. Simulated experiments show this approach improves inferences in some aspects, making them more stable over repeated analysis and creating narrower bounds for certain common statistics of interest. Extensions to the existing literature have been executed that provide further stability to inferences and also a strong potential to identify ways to make the analysis procedure more flexible. The competing methods have been first compared using simulated data sets and then a real data set concerning analysis of the effect of breastfeeding duration on children's cognitive ability. R code to implement the methods used is available as online supplementary material.

1. Introduction

Multiple Imputation, MI (Rubin, 1987) is one of the most popular approaches to address missing data problems (van Buuren, 2012). In this approach an imputer imputes missing values from their posterior predictive distribution conditional on the observed data. The imputer repeats this process m times, thus generating m imputed data sets. These m data sets can then be released to analysts. Analysts process each imputed data set as if it were fully observed, obtaining point and variance estimates in the same way they would have proceeded faced with the original data. Simple rules proposed by Rubin (1987) allow analysts to combine their estimates across imputations and obtain appropriate point and variance estimates that take into account the additional uncertainty due to the missing data.

Rubin's combining rules in general perform quite well when m is large, and for a wide variety of estimates such as means and regression coefficients (Rubin, 1996), essentially where a normal approximation for the sampling distribution of the estimator (if frequentist) or for the posterior distribution of the estimand (if Bayesian) is a reasonable assumption. We will evaluate the performance of the combining rules assuming that we are taking a Bayesian inference approach, although similar conclusions could be expected under a frequentist approach, provided non-informative priors are used.

However, some evidence suggests that the distributional approximations underlying the MI approach to inferences may not be suitable in all cases (Marshall et al., 2009). Concerns regarding the variability of inferences obtained from multiply-imputed data (Graham et al., 2007; Bodner, 2008) and improving computation power both motivate the need to avoid

[*] Corresponding author.
E-mail address: S.Rashid@soton.ac.uk (S. Rashid).

http://dx.doi.org/10.1016/j.csda.2015.05.009
0167-9473/© 2015 Elsevier B.V. All rights reserved.

such approximations. In particular, fitting Rubin's proposed t distribution to approximate the posterior distribution of the estimand may not be necessary or even suitable in some cases. Recent work has focused on sampling based approaches that utilise the imputed data sets to approximate the full posterior distribution of the estimand. Zhou and Reiter (2010) combined posterior draws of the estimands from each imputed data set to approximate the posterior distribution. Steele et al. (2010) used the fact that for a large enough sample size, posterior distributions conditional on an imputed data set can be well approximated with a normal distribution and suggested modelling the posterior distribution using a mixture of normals approach, where each normal distribution receives equal weight, $1/m$. This approach was shown to be beneficial in creating more stable confidence intervals than the usual method of moments t distribution approximation. However, one drawback with this approach was that it resulted in confidence intervals that tended to be too narrow with a lower than nominal coverage rate. The purpose of this paper is to extend this methodology to consider approximating the posterior distribution with mixture of t distributions. By doing so, we hope to correct for the problem of under coverage noted by Steele et al. (2010) while still retaining the benefits of improved stability in the confidence intervals created with such an approach.

The rest of the paper is organised as follows. In Section 2 we briefly review how one can use multiply imputed data sets to infer about an estimand in the population and illustrate the drawbacks of existing inferential techniques. In Section 3 we propose the inferential procedure based on a mixture of t-distributions. In Section 4 we illustrate the procedure on a breast-feeding study. Finally in Section 5 we end with some concluding remarks.

2. Multiple imputation inference

Suppose we have a $n \times p$ covariate data set $X = (x_1, \ldots, x_p)$, where x_{ij} is the i^{th} unit's value for the j^{th} covariate, $i = 1, \ldots, n, j = 1, \ldots, p$. For each x_{ij} we also define a corresponding missing data indicator m_{ij} where $m_{ij} = 1$ implies x_{ij} is missing and $m_{ij} = 0$ implies x_{ij} is observed. We can then decompose X into its missing and observed parts with $X_{mis} = \{x_{ij} : m_{ij} = 1\}$ and $X_{obs} = \{x_{ij} : m_{ij} = 0\}$ respectively.

In multiple imputation we fill in missing values with draws from their predictive distribution conditional on the observed values i.e. from $p(X_{mis}|X_{obs})$. We take m sets of draws to create m complete data sets; denote these data sets by $X^{(k)}_{com}$, $k = 1, \ldots, m$. We assume that an analyst of the data is interested in inferring about some estimand in the population, Q, and that the posterior distribution of the estimand, i.e. $p(Q|X)$, can be well approximated with a normal distribution. For large values of n and for most common estimands, such as the sample mean of a variable in the data or coefficients from a regression model, the normal will be a reasonable approximation. The analyst then proceeds to obtain m sets of point and variances estimates, q_k and u_k, for Q in each of the imputed data sets $X^{(k)}_{com}$. Typically an analyst will then combine the estimates in the following way:

$$\bar{q} = \frac{\sum_{k=1}^{m} q_k}{m}, \tag{1}$$

$$T_m = (1 + 1/m)\frac{\sum_{k=1}^{m}(q_k - \bar{q})^2}{m-1} + \frac{\sum_{k=1}^{m} u_k}{m} = (1 + 1/m)B + \bar{U}. \tag{2}$$

Inferences can then be made using $\bar{q}$ as a point estimate for Q and T_m as an estimate of the variance of $\bar{q}$. The variance clearly reflects the additional uncertainty present due to the missing data. In the absence of missing data, the average of individual variances $\bar{U}$ would be sufficient as a final variance estimate; the variance of individual point estimates, B adds to this estimate and the $1/m$ correction provides further variability considering a finite number of imputations, m. Confidence intervals can be constructed using a t-distribution with degrees of freedom given by $\nu = (m-1)(1 + r_m^{-1})^2$, where, the relative increase in variance due to missing information, $r_m = (1 + 1/m)\bar{B}/\bar{U}$. This expression is obtained through an approximation that matches the first two posterior moments of T_m to a chi-squared distribution. For small m, an analyst can use the adjusted degrees of freedom, as proposed by Barnard and Rubin (1999), $\nu* = 1/(1/\nu + 1/\nu_0)$, where $\nu_0 = (a+1) * a/(a+3) * \bar{U}/(\bar{U} + (1+1/m)\bar{B})$ and a is the complete data degrees of freedom.

While this procedure is straightforward and relatively easy to apply in standard software packages, some of its limitations have been noted. More specifically, there have been concerns about the reproducibility of inferences over repeated application of the MI procedure with Rubin's combining rules. In particular, inferences gathered from a small number of imputations, such as length of confidence intervals and p-values, have been shown to have considerable variability over repeated MI analysis. Graham et al. (2007) demonstrate the power fall-off of a hypothesis test when the number of imputations falls from 100 to 3. Bodner (2008) and White et al. (2011) further support this observation demonstrating concern over the variability of inferences. Nevertheless, the confidence intervals obtained from Rubin's combining rules provide an appropriate probability coverage on average (Rubin and Schenker, 1986).

The current literature recommends increasing the number of imputations, m, to be able to produce stable inferences, which, given the processing power of modern computers, is a practical solution. Nevertheless, there are several reasons why the use of small m is still desirable. Practices in many academic disciplines and the industry are often driven by convention