Computational Statistics and Data Analysis 95 (2016) 24–38

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models

Min Cherng Lee [1], Robin Mitra *

Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 8 October 2014
Received in revised form 24 July 2015
Accepted 6 August 2015
Available online 9 September 2015

Keywords:
Data augmentation
Latent variable
Missing data
Multiple imputation

ABSTRACT

Multiple imputation is a commonly used approach to deal with missing values. In this approach, an imputer repeatedly imputes the missing values by taking draws from the posterior predictive distribution for the missing values conditional on the observed values, and releases these completed data sets to analysts. With each completed data set the analyst performs the analysis of interest, treating the data as if it were fully observed. These analyses are then combined with standard combining rules, allowing the analyst to make appropriate inferences which take into account the uncertainty present due to the missing data. In order to preserve the statistical properties present in the data, the imputer must use a plausible distribution to generate the imputed values. In data sets containing variables with different measurement scales, e.g. some categorical and some continuous variables, this is a challenging problem. A method is proposed to multiply impute missing values in such data sets by modelling the joint distribution of the variables in the data through a sequence of generalised linear models, and data augmentation methods are used to draw imputations from a proper posterior distribution using Markov Chain Monte Carlo (MCMC). The performance of the proposed method is illustrated using simulation studies and on a data set taken from a breast feeding study.

1. Introduction

Multiple imputation is a widely used and appealing approach to deal with missing values (Rubin, 1996; Gelman et al., 2005; Wallace et al., 2010; Bernhardt et al., 2014; Hapfelmeier and Ulm, 2014; Xu and Zhang, 2010; Consentino and Claeskens, 2010). In this approach the imputer models the joint distribution of the variables in the data set and, using this model, imputes missing values from their posterior predictive distribution conditional on the observed data. This is done m times to yield m multiply imputed data sets which are then given to the analysts. Analysts can then perform the same standard analysis that would have been performed on fully observed data on each of these completed data sets, and combine the analyses across data sets using simple combining rules, thereby incorporating the additional uncertainty due to the missing values. An appealing feature of multiple imputation is that the imputer and analyst can be different individuals, as this allows trained statistical modellers to deal appropriately with the often complex missing data problem, and reduces the subsequent burden on the analyst, who is often not a proficient modeller, of dealing with the missing values.

* Corresponding author.
E-mail address: R.Mitra@soton.ac.uk (R. Mitra).
[1] Present address: Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, Malaysia.

http://dx.doi.org/10.1016/j.csda.2015.08.004
0167-9473/

It is crucial to draw imputations from an appropriate distribution to preserve relationships present in the data set. There are often complications in drawing imputations from their posterior predictive distribution: (1) the pattern of missing values in the data set may not be monotone, and (2) the variables in the data set may be continuous or categorical. One way to simplify this problem is to model the joint distribution of variables in the data set through a sequence of generalised linear models (GLMs) where the link function of the GLM is determined by the measurement scale of the variable being modelled; missing values can then be imputed on a variable by variable basis. This approach has some advantages over multivariate imputation models (Schafer, 1997), as it is easier to detect model inadequacies in each GLM than having to assess the fit over a joint model for the whole data set. Still, to draw imputations with a non-monotone pattern of missing values from their joint posterior predictive distribution requires Markov Chain Monte Carlo (MCMC) methods (Little and Rubin, 2002). In this article we use the decomposition above and data augmentation methods (Albert and Chib, 1993) to draw missing values from their posterior predictive distribution. In data sets with continuous, binary and ordinal variables such an approach allows imputations to be drawn within a Gibbs sampler. When data sets include nominal variables, we develop a Metropolis within Gibbs sampler to impute missing values with an innovative Metropolis–Hastings proposal distribution.

The decomposition above has been previously proposed in the literature to deal with missing values (Ibrahim et al., 2005, 1999, 2002; Chen and Ibrahim, 2001). While these approaches use the same decomposition, they do not present specific models to use for each conditional regression, instead presenting the strategy as a general framework. Some also focus on methods to obtain maximum likelihood estimates from the data through the construction of E–M algorithms (Ibrahim et al., 1999; Chen and Ibrahim, 2001). For those that proceed with Bayesian modelling using this decomposition (Ibrahim et al., 2002, 2005), the computations are also typically performed in a different way to those described in this article, making use of adaptive rejection sampling (Gilks and Wild, 1992). The data augmentation methods used in this article to model variables in the data have also been considered in the literature. De Leon and Carrièger (2007) use similar models proposed in this article to jointly model categorical and continuous variables, but are primarily concerned with likelihood methods to estimate parameters of the distribution. Unlike this article, De Leon and Carrièger (2007) also do not explicitly embed the imputation of missing values within a full Bayesian framework. Another article by Goldstein et al. (2009) proposes a related imputation modelling strategy for multilevel models, that uses latent normal random variables. However, our approach differs fundamentally with Goldstein et al. (2009) in the modelling of nominal variables, and thus also in the development of the posterior computations. To our knowledge no article has comprehensively reviewed the proposed imputation approach through extensive simulation studies, and the development of the innovative Metropolis–Hastings algorithm in this article is an original contribution within multiple imputation for missing data.

We illustrate the performance of this imputation strategy on simulated data as well as on a breast feeding study. We also compare the performance of this strategy with the approach of multiple imputation via chained equations (MICE) in these scenarios. MICE is a commonly used method for imputing missing values in these types of situations (Raghunathan et al., 2001; Van Buuren, 2011) and so it would be interesting to see if there were any advantages gained by using the more formal modelling strategy proposed.

The remainder of the article is structured as follows; Section 2 briefly reviews the approach of multiple imputation to deal with the problem of missing values. Section 3 describes the modelling strategy we propose for imputing missing values, and briefly describes the approach of MICE to impute missing values, Section 4 illustrates the performance of both approaches in a simulation study, Section 5 illustrates the performances of both approaches in a breast feeding study, finally Section 6 presents some concluding remarks.

2. Missing data and multiple imputation

In this section we briefly describe the missing data framework and how the multiple imputation procedure is used for inference. We suppose that we have a $n \times p$ data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})'$, $j = 1, \ldots, p$ is the jth variable in $\mathbf{X}$. Also denote a $n \times p$ missing data indicator matrix $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_p)$, where $\mathbf{m}_j = (m_{1,j}, \ldots, m_{n,j})'$ and $m_{i,j} = 1$ indicates $x_{i,j}$ is missing and $m_{i,j} = 0$ indicates $x_{i,j}$ is observed, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We can then denote the observed and missing portions of $\mathbf{X}$ by $\mathbf{X}_{obs} = \{x_{i,j} : m_{i,j} = 0\}$ and $\mathbf{X}_{mis} = \{x_{i,j} : m_{i,j} = 1\}$ respectively.

If the conditional distribution of $p(\mathbf{M}|\mathbf{X}, \boldsymbol{\phi}) = p(\mathbf{M}|\mathbf{X}_{obs}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ are the parameters of the distribution, so missingness depends only on the observed components in $\boldsymbol{X}$, then the data are said to be missing at random (MAR) (Little and Rubin, 2002). In this article we focus on covariates that are MAR, so the modelling strategy proposed in the next section makes this assumption and the simulations in Section 4 utilise MAR mechanisms.

We assume that an analyst wishes to use the data to infer about some population quantity Q, this might be for example the mean of jth variable or it could be the coefficient from a regression model for one of the variables on some other subset of variables. To do this they obtain a point estimate, q, for Q, and an estimate of its variance u. With the presence of missing data, analysts may no longer be able to obtain these estimates.

In multiple imputation, the missing values are imputed from their posterior predictive distribution $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$, this is done m times to generate m completed data sets $\mathbf{X}_{com}^{(1)}, \ldots, \mathbf{X}_{com}^{(m)}$. Analysts can then treat each completed data set as a fully observed data set to obtain m sets of point and variances estimates (q_k, u_k) from each $\mathbf{X}_{com}^{(k)}$, $k = 1, \ldots, m$. Appropriate inference for Q can then be made following simple combining rules (Rubin, 1987). Specifically the analyst computes the