Computational Statistics and Data Analysis 72 (2014) 92–104

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Recursive partitioning for missing data imputation in the presence of interaction effects[☆]

L.L. Doove [a,b,*], S. Van Buuren [c,a], E. Dusseldorp [c,b]

[a] Department of Methodology and Statistics, Faculty of Social Sciences, University of Utrecht, PO Box 80140, 3508 TC Utrecht, The Netherlands

[b] Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102 - bus 3713, Leuven, Belgium

[c] Netherlands Organisation for Applied Scientific Research TNO, PO Box 2215, 2301 CE Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 18 July 2012
Received in revised form 31 October 2013
Accepted 31 October 2013
Available online 13 November 2013

Keywords:
CART
Classification and regression trees
Interaction problem
MICE
Nonlinear relations
Random forests

ABSTRACT

Standard approaches to implement multiple imputation do not automatically incorporate nonlinear relations like interaction effects. This leads to biased parameter estimates when interactions are present in a dataset. With the aim of providing an imputation method which preserves interactions in the data automatically, the use of recursive partitioning as imputation method is examined. Three recursive partitioning techniques are implemented in the multiple imputation by chained equations framework. It is investigated, using simulated data, whether recursive partitioning creates appropriate variability between imputations and unbiased parameter estimates with appropriate confidence intervals. It is concluded that, when interaction effects are present in a dataset, substantial gains are possible by using recursive partitioning for imputation compared to standard applications. In addition, it is shown that the potential of recursive partitioning imputation approaches depends on the relevance of a possible interaction effect, the correlation structure of the data, and the type of possible interaction effect present in the data.

1. Introduction

Today's state of the art solution for handling missing data is multiple imputation. In approaches to implement multiple imputation, different methods are available to use the information from the data at hand (Van Buuren, 2012). The common element in these methods is that they model the relations between variables. Hereby, it is particularly important to reflect the structure of the data since otherwise, parameter estimates under multiple imputation will be biased. Caution is therefore needed when data contain nonlinear structures like a quadratic relation. Approaches to implement multiple imputation, like Multiple Imputation by Chained Equations (MICE; Van Buuren, 2007), do not automatically incorporate nonlinear relations. We focus on a special case of nonlinear relations, namely interaction effects. For the purpose of this study, both cross-products and quadratic terms are denoted by interactions.

MICE is a popular approach for implementing multiple imputation because of its flexibility. In MICE, multivariate missing data are imputed on a variable by variable basis, called fully conditional specification (Van Buuren, 2007). This means that per variable imputations are created, such that for each incomplete variable a specified imputation model is required. In these imputation models, interactions can be modelled in two ways: first, by specifying models including interaction effects manually and second by imputing subgroups of the data separately. For example, one could create distinct imputation

[☆] Supplementary materials related to the implementation of proposed methods are available online (see Appendix C).

[*] Corresponding author at: Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102 - bus 3713, Leuven, Belgium. Tel.: +32 16 3 25977. *E-mail addresses:* lisa.doove@ppw.kuleuven.be (L.L. Doove), stef.vanbuuren@tno.nl (S. Van Buuren), elise.dusseldorp@tno.nl (E. Dusseldorp).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.025

models for males and females. Besides the fact that both approaches are somewhat cumbersome, they are often unusable as the structure of the data is usually unknown before imputation. Therefore, models should preferably be fitted to the data in an automatic fashion without unnecessary user involvement.

A technique that can handle interactions with ease is recursive partitioning (Burgette and Reiter, 2010; Hand, 1997). One of the first implementations of recursive partitioning is called Automatic Interaction Detection (Morgan and Sonquist, 1963). The recursive partitioning technique models the interaction structure in the data by sequentially splitting a dataset into increasingly homogeneous subsets (Breiman et al., 1984). Essentially recursive partitioning finds the split that is most predictive of the response variable by searching through all predictor variables (Merkle and Schaffer, 2011). Within the subgroups created from one predictor variable, the algorithm goes on to partition the data based on other variables or other splits of the same predictor. The resulting series of splits can be represented by a tree structure like Fig. 1, to which we will return in Section 2. Since splits are conditional on previous splits, the variables used may indicate interaction effects. By constructing models in this manner, possible interactions are automatically taken into account.

Others have worked on this idea of combining recursive partitioning with imputation methods, e.g., Burgette and Reiter (2010), Iacus and Porro (2007, 2008), Nonyane and Foulkes (2007), Stekhoven and Bühlmann (2012), and Van Buuren (2012, p. 83). The main shortcoming of most of the proposed methods is that recursive partitioning is combined with single imputation instead of multiple imputation. Therefore, they cannot be used for making appropriate statistical inferences. Another shortcoming is that, except for Burgette and Reiter, the performance of these methods is not investigated on data containing interaction effects. In the current study, we would like to overcome these shortcomings by providing a framework for connecting recursive partitioning techniques with multiple imputation. This type of imputation takes into account the uncertainty associated with the missing data (Rubin, 1996), which results in parameter estimates with appropriate confidence intervals.

The purpose of our study is to gain insight into whether the use of recursive partitioning in multiple imputation (i.e., MICE) is a convenient way to preserve interaction effects. We consider two main questions: which recursive partitioning techniques create appropriate variability between repeated imputations? What are the statistical properties (e.g., bias, coverage, confidence interval width) of estimates of the interaction parameters? In gaining insight into these questions, distinctions will be made between different types of interactions. In addition, the two questions will be considered for both continuous and categorical data. Burgette and Reiter (2010) embarked on the implementation of recursive partitioning in MICE and demonstrated the performance of the method on a single model with continuous predictor and response variables. We want to elaborate on the work of Burgette and Reiter and, to be complete, also consider categorical predictor and response variables. Different results are expected for both types of data since recursive partitioning techniques are known to perform especially well for data with interactions between categorical variables (Dusseldorp et al., 2010).

This paper is organized as follows. In Section 2, MICE will first be elaborated further after which two main recursive partitioning techniques will be considered, namely Classification And Regression Trees (CART; Breiman et al., 1984) and random forests (Breiman, 2001). Subsequently, incorporation of recursive partitioning in the MICE framework will be presented. In Section 3 different interaction types will be discussed, which will be observed in answering the research questions. Then we make the distinction between predictor and response variables either being continuous (Section 4) or categorical (Section 5). In both Sections 4 and 5, a simulation study is described, carried out to investigate which of the discussed methods are convenient to preserve interaction effects, followed by the results of the simulation study. The results from both simulation studies will be discussed in Section 6, at the end of which some final conclusions are given.

2. MICE and recursive partitioning

2.1. Multiple imputation by chained equations

Imagine a set of variables, $y_1, \ldots, y_j$, some or all of which have missing values. Handling these data using MICE comprises three main steps: generating multiple imputation, analyzing the imputed data, and pooling the analysis results (Van Buuren, 2007). The main idea is to impute each incomplete variable using its own imputation model. All missing values are initially filled in at random. The first variable with at least one missing value, say y_1, is then regressed on the remaining variables, $y_2, \ldots, y_j$. This is restricted to individuals with observed values for y_1. The missing values in y_1 are now replaced by simulated draws from the posterior predictive distribution of y_1. The next variable with missing values, say y_2, is then regressed on all the other variables, $y_1, y_3, \ldots, y_j$. This estimation is restricted to individuals with observed y_2 and uses the imputed values of y_1. Again, missing values in y_2 are replaced by draws from the posterior predictive distribution of y_2. This process is repeated for all other variables with missing values in turn. To stabilize the results this cycle is iterated a number of times, producing one imputed dataset. The entire procedure is repeated m times, yielding m imputed datasets. Each complete dataset is analyzed separately by MICE, after which the results are pooled.

2.2. Recursive partitioning

In this study we consider two main recursive partitioning techniques, namely CART and random forests. We will first elaborate on CART and return to random forests later on in this section. Depending on the response variable of interest