Joint Analysis and Imputation of Incomplete Data in R: A Fully Bayesian Approach for Handling Missing Values in Complex Settings

Nicole Erler

Abstract

Missing values are a challenge regularly encountered in the analysis of real-world data. One of the most popular approaches to handle missing values is multiple imputation, usually using a full conditional specification (FCS). In multiple imputation, each missing values is imputed multiple times, generating multiple completed datasets, which can then be analysed using standard methods. FCS is based on the idea of the Gibbs sampler, and values are imputed from univariate models that have all other variables in their linear predictor. The separation of imputation and analysis and the specification of full-conditional models are attractive features since they allow multiple analyses to be performed on the same set of imputed data and facilitate a straightforward specification of imputation models for variables of mixed types. In more complex settings such as multi-level data, time-to-event analyses or settings involving non-linear associations, however, important assumptions of the FCS approach are likely violated, which may lead to biased results. These violations occur, for example, when it is not possible to include the response variable(s) into the linear predictor of the imputation models without simplification or summarising it, or when the shape of the association between response and incomplete covariate implied by the imputation model is misspecified.

An alternative for analysing incomplete data in such setting is a fully Bayesian approach in which the parameters of interest are estimated jointly with the missing values. The joint distribution of response variable(s), incomplete variables and parameters can be conveniently split into a sequence of (univariate) conditional distributions, allowing the choice of appropriate distributions for mixed-type variables. When the model(s) for the response variable(s) are included in this sequence, it is not necessary to have the response variables into the linear predictors of the models for covariates. This makes the approach suitable for highly complex substantive models, such as multivariate joint models of longitudinal and survival data. Moreover, any complex association structures specified in the substantive model(s) are automatically taken into account during imputation, ensuring compatibility between all sub-models involved. This fully Bayesian approach is implemented in the R package JointAI, which provides functionality to fit models of various types using syntax that users are familiar with from standard R functions.

Besides giving an introduction to the theoretical background of the fully Bayesian approach, the basic use of JointAI and some of its extended functionality will be demonstrated.