Robust multiple imputation based on quantile forests

Kristian Kleinke[1)] & Markus Fritsch[2)]

[1)]*University of Siegen*

[2)]*University of Passau*

Keywords: multiple imputation, random forest, quantile forests, quantile regression

Random Forest (RF) is a machine learning method for classification and regression problems that can be enumerated among the ensemble methods - i.e. the classification decision / prediction is based on an ensemble (forest) of relatively independent statistical models (trees), and among the recursive partitioning methods (the training data are split into several classes so that observational units within one class (leafs) are very ‘similar’ to each other, and ‘different’ from observational units of other classes). In the context of missing data imputation, k bootstrap samples are generated from the remaining observed data, and one tree is grown from each bootstrap sample, using a small group of input variables for finding the best split at each node. Predictions are then made for the incomplete observations regarding which leaf they belong to. Finally, one observation is selected randomly from the observed donors of the matched leafs. This process is repeated M-1 times to obtain the M multiple imputations. Imputation by RF is particularly attractive for large datasets, since no imputation model and auxiliary variables need to be specified, and no functional from needs to be specified, since the underlying functional from is approximated in a data-driven fashion. However, little is yet known about the robustness of RF based imputation. The purpose of the present paper is to elucidate to what extent RF-based multiple imputation is robust, and if imputation based on quantile forests (which focus on the conditional median or other quantiles of interest rather than the mean) might work `better', if the data are skewed and heteroscedastic.