Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

Multiple imputation of multivariate missingness

- Last week we learned about multiple imputation for univariate missingness.
- In practice, missing data may occur in several variables and the goal of this week's lecture is to learn about how to deal with this case in a multiple imputation context.
- As we have seen back in week 4, when we superficially covered multivariate missingness in a single imputation context, there are two main routes:
  - Joint model imputation.
  - Multiple imputation by chained equations/fully conditional specification.

Multiple imputation of multivariate missingness

Joint model approach

- The joint model approach is the original approach for multiple imputation of multivariate missingness and it rests on specifying a multivariate (regression) model for all the variables that contain missing values.

- The main problem with this approach is that it is not always easy to set up a reasonable multivariate regression model, and this may be particularly complicated if the variables containing missing values are of mixed type (e.g., continuous and binary/categorical variables).

- However, software exists to fit such models (e.g., `norm`, `cat`, `mix`, and `jomo`) and there are still new techniques within this line of multiple imputation being developed.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- In this course we will cover in detail the approach based on multiple imputation by chained equations (MICE), also known as fully conditional specification, which is nowadays often considered as the gold standard approach to multivariate multiple imputation.

- As we can guess by the name, this is the approach implemented by the `R` package `mice` when several variables have missing values.

- The MICE approach imputes multivariate missing data on a variable by variable basis.

- The method requires the specification of a regression imputation model for each incomplete variable and creates imputations per variable in an iterative fashion.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

↪ The following scheme illustrates the main idea behind the MICE approach.

Y_1	Y_2	Y_3	Y_4	...
✓	✓	NA	NA	...
✓	NA	NA	✓	...
NA	✓	✓	NA	...

↪ $Y_1 \sim Y_2 + Y_3 + Y_4 + \ldots$

↪ $Y_2 \sim Y_1 + Y_3 + Y_4 + \ldots$

↪ $Y_3 \sim Y_1 + Y_2 + Y_4 + \ldots$

↪ $Y_4 \sim Y_1 + Y_2 + Y_3 + \ldots$

↪ Under this imputation scheme, we can easily deal with variables of mixed types and we reduce the problem of multivariate imputation to several univariate imputation problems.

↪ The options that we have available in `mice` for each univariate regression model depends on the nature of the variable being imputed and we have already covered them last week.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- Let us use the toy example from the previous slide to describe how MICE works. The iterative procedure is as follows:
  - Start with initial guesses for the missing values (e.g., random draws from the observed data of the respective variable).
  - Update/impute Y_1 based on Y_2, Y_3, Y_4, and the other fully observed variables (in case they exist). Note that at this stage Y_2, Y_3, and Y_4 still have the missing values set to the initial guesses.
  - Update/impute Y_2 based on Y_1, Y_3, Y_4, and the other fully observed variables. Note that now for the missing values on Y_1 we are using the updated values from the previous step, while for Y_3 and Y_4 the missing values are still set to the initial guesses.
  - ...
  - Update/impute Y_1 again based on updated Y_2, Y_3, and Y_4 and on the values of the fully observed variables.
  - Repeat until convergence.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- ↪ The imputed values for Y_1, Y_2, Y_3, Y_4 from the last iteration of the algorithm are then used to replace the missing data in the original dataset.
- ↪ Note that one run through the algorithm until convergence leads to one imputed dataset.
- ↪ We need to repeat the procedure M times to impute values in all M copies of the dataset that we have created.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- ↪ MICE is a **Markov chain Monte Carlo** (MCMC) method. More specifically, it is based on the idea of **Gibbs sampling**.
- ↪ MCMC? Gibbs sampling?
- ↪ MCMC is a technique that allows drawing samples from a complex distribution.
- ↪ MCMC works by creating a chain of random variables (a Markov chain)! The distribution that each element in the chain is sampled from depends on the values of the previous element.
- ↪ Under certain conditions, the chain eventually stabilizes (i.e., converges to the stationary distribution).
- ↪ It can be proved that samples from the chain are then samples from our complex distribution of interest.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- ↪ The Gibbs sampler is a MCMC method that allows drawing samples from a univariate distribution by splitting it into a set of univariate full conditional distributions.

- ↪ It can be shown that a sample from the multivariate distribution of interest can be obtained by repeatedly sampling from each of the univariate conditional distributions.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- The sequence of imputed values for each missing value is called a (Markov!) chain.
- Each run through the MICE algorithm produces one chain per missing value.
- The reason for why we need to iterate is because the imputed values in one variable depend on the imputed values of the other variables and to start the procedure we use random draws from the observed data for each particular variable.
- It is not unreasonable to think that these draws are far from the actual distribution and so the first draws may not be draws from the distribution of interest.

Multiple imputation of multivariate missingness

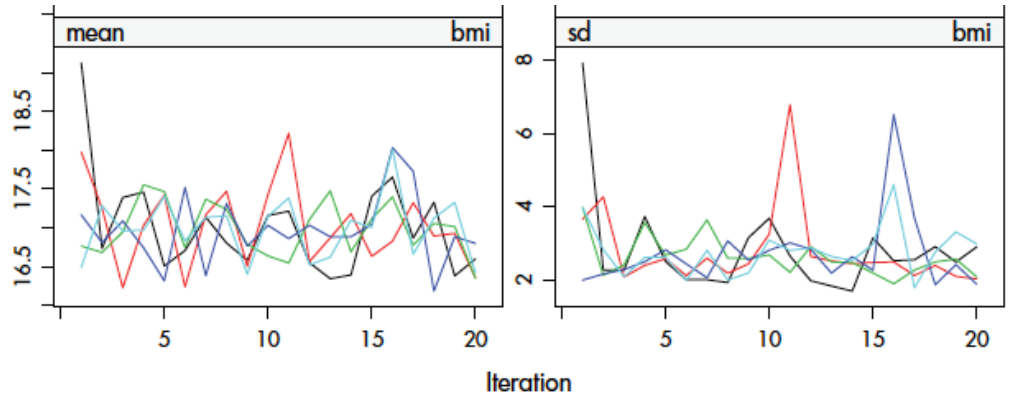
Multiple imputation by chained equations

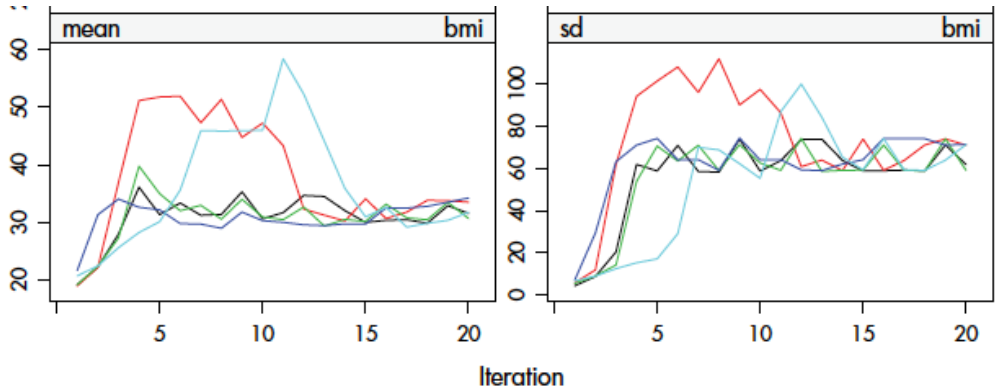
- ↪ Note that we have several variables with missing values and each of these may contain several missing values, and therefore at the end of the iterative procedure we have a potential large number of chains of imputed values.
- ↪ Then, because we need to repeat the procedure for all M copies of the dataset, we have a potential very large number of chains of imputed values multiplied by M.
- ↪ To monitor convergence is thus infeasible to look at all produced chains.
- ↪ Alternatively, as a way to reduce the number of chains we have to look at, and as implemented in `mice`, we could look at the chains of the mean and/or standard deviation (or other summary statistic) of the imputed values per variable.
- ↪ Fortunately, as pointed by van Buuren, the number of iterations to reach convergence is typically much lower than that required in typical modern MCMC applications. The default in `mice` is 5, but it is quite difficult to judge convergence based on a so low number of iterations and, in practice, 20 or 30 iterations are often recommend.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

↪ Figures from van Buuren (2018, p. 188 and 189). The top row shows the `mice` output for a healthy convergence case, while the second row shows a traceplot that indicates non convergence.





Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- ↪ It is interesting to note that Gibbs sampling exploits the fact that a multivariate joint distribution is fully determined by its full univariate conditional distributions.
- ↪ But MICE imputation models work the other way around. The univariate imputation models (= univariate full conditional distributions) are specified directly and there is no guarantee that a corresponding joint distribution exists.
- ↪ This leads us to the concept of **compatibility**.
- ↪ To conditional densities $f(Y_1 \mid Y_2)$ and $f(Y_2 \mid Y_1)$ are said to be compatible if a joint distribution $f(Y_1, Y_2)$ exists that has $f(Y_1 \mid Y_2)$ and $f(Y_2 \mid Y_1)$ as its conditional densities.
- ↪ So, the main problem of MICE is that there is no guarantee that the distributions implied by the different univariate imputation regression models are compatible.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- As stated by van Buuren (2018, p. 123): "*though MICE is only guaranteed to work if the conditional are compatible, these simulations suggest the results may be robust against violations of compatibility*."

- van Buuren also concluded the discussion on this topic with the following: "*Apart from potential feedback problem, it appears that incompatibility seems like a relatively minor problem in practice, especially if the missing data rate is modest and the imputation models fit the data well*."

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- We have already discussed the concept of **congeniality** and how often it is confused with the compatibility concept.
- Remember that compatibility refers to the property that the conditionally specified models together specify some joint distribution from which imputations are to be drawn. It is a theoretical requirement of the Gibbs sampler.
- Congeniality refers to the relation between the substantive model (step 2) and the imputation model.
- It is widely accepted that the imputation model should be more general than the substantive model.
- Further, interactions, quadratic terms, etc, appearing in the substantive model should be preserved in the imputation model.

Multiple imputation of multivariate missingness

Multiple imputation by chained equations

- To finish we should keep in mind the assumptions made by MICE:
  - The missing mechanism is MAR (or MCAR).
  - All associations between variables are linear.
  - Compatibility and congeniality.