Version dated: June 30, 2016

IMPUTING CONTINUOUS DATA USING PHYLOGENETIC INFORMATION

Patrik Drhlik[1,2] and Simon P. Blomberg[1]

[1]*School of Biological Sciences, University of Queensland, St. Lucia, Queensland, 4072, Australia.*

E-mail: s.blomberg1@uq.edu.au

[2]*Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, 461 17 Liberec, Czech Republic.*

E-mail: patrik.drhlik@tul.cz

Abstract.— Biological databases are often sparse which makes their analyses more difficult. One of the most plausible solutions of dealing with sparse databases is using multiple imputation (MI). MI has already been implemented and is usually a part of most standard statistical softwares. These methods do not use the potential of phylogenies though and do not preserve phylogenetic signal in traits. In this paper, we present new MI methods that preserve phylogenetic signal after the imputation process. These methods were developed in R and based on methods used in R mice package. These methods have been previously suggested as a possible way of improving MI but have not been implemented yet.

(Keywords: Multiple imputation, mice, missing data, phylogenetic signals, phylogenetic imputation, R statistical software)

Possible titles

IMPUTING CONTINUOUS DATA USING PHYLOGENETIC INFORMATION

IMPUTING CONTINUOUS DATA INCORPORATING PHYLOGENETIC INFORMATION

IMPUTING CONTINUOUS DATA INCORPORATING PHYLOGENIES

PHYLOGENETIC IMPUTATION IN R

Notes on observations

phnorm - lambda

tree size x trait size - more or less the same

tree size x tree shape - it seems that pure birth trees give the best results, one sided trees are the worst

tree size x lambda transform - the bigger the lambda before imp, the bigger after, seems to be working slightly better for larger trees

tree size x miss data mech - slightly better when mcar

tree size x miss data level - it works better when the tree is larger and has more missing data (higher the level, more missing data there is)

trait size x tree size - more or less the same

trait size x tree shape - pure birth the best

trait size x lambda transorm - decreasing trend while more traits

trait size x miss data mech - not enough measures in mcar

trait size x miss data level - better when more missing data, gets worse with more traits

tree shape - all variable share the same trend: pure birth best, balanced worse, one sided worst

Introduction

Biological and ecological databases contain vast amounts of data. Unfortunately, these databases are very sparse for various reasons. Even the largest plant trait database has a really sparse data matrix Kattge *et al.* (2011). All of this makes analyses more difficult. There are several methods that can handle the missing data problem. They span from the easiest like casewise deletion Allison (2001) to complex procedures like multiple imputation Allison (2001); Rubin (1996). Methods like a multivariate imputation by chained equations (MICE) have been already successfully implemented van Buuren & Groothuis-Oudshoorn (2011) and are part of a standard software nowadays. None of these methods incorporates the use of phylogenetic information to preserve the phylogenetic signal Blomberg *et al.* (2003). It has been suggested Swenson (2014); Garland & Ives (2000) but not yet implemented. In this paper, we present an extension to R mice package van Buuren & Groothuis-Oudshoorn (2011) that provides a multiple imputation method where imputed data do not lose the phylogenetic information.

Why is missing data a problem?

Everybody who is interested in missing data knows this quote from R.A. Fisher – "The best solution to handle missing data is to have none.". This solution would be a nice way to handle the problem but it is almost impossible to be achieved in real world problems. Any kind of a survey is very likely to have missing data because of the human nature – either we don't want to respond to some questions or we simply forget that there is a second page.

Several problems arise when a data set has missing data. If we use a simple approach where we delete all observations that contain missing data (casewise deletion or complete case analysis), we abandon a lot of data. Imagine a data set that has a hundred observations of ten traits. Thirty of these observations only lack one of the traits and you would have to exclude whole observations because of that even though the other nine traits were fully observed. That leads to a bias in the analysis and decreases a statistical power of the model.

Missing data assumptions

There are several missing data assumptions Allison (2001); Donders *et al.* (2006) that need to be taken in account before handling the data. For a well explaining graphical example about the differences between these mechanisms, see Nakagawa & Freckleton (2008). The strongest assumption is that the data are *missing completely at random* (MCAR). That means that the missingness is not related neither to dependent nor independent variables. If we can assume that this mechanism is present then even a casewise deletion method can produce unbiased estimates.

A weaker but still strong assumption is *missing at random* (MAR). This mechanism says that the missingness of the data may depend on the input variable. In other words, there might be a threshold above which the probability of missing data is very high. That might for example be the case when people do not fill in the amount of money they make in a survey. They might not want to share that information if they make too much.

The weakest assumption is *missing not at random* (MNAR). This mechanism assumes that the missing data in a response variable depends on the response variable itself. Which means that it is something that is not observed. This assumption is said to be *non-ignorable* as opposed to MCAR and MAR which are

ignorable Allison (2001). When the mechanism is identified as MNAR, the analysis is usually very difficult and often requires a unique approach van Buuren & Groothuis-Oudshoorn (2011).

Missing data approaches

We are going to describe a few methods (but not all) that are used in the missing data problems.

The easiest way to handle missing data is a method called casewise deletion (sometimes also complete case analysis or listwise deletion) Allison (2001); Donders *et al.* (2006). You need to delete all observations with missing data and then start your analysis. When it comes to a standard software, this is often a default option how to handle missing data. As mentioned before, this method would not produce biased estimates under the MCAR assumption. That is usually not the case and you should therefore consider different methods.

There is a whole group of method that we call *imputations*. Those are either single or multiple. When using a single imputation method, we impute values to fill the gaps in data sets. After this process, the data set looks like it never had missing values in the first place. We do not try to guess the actual missing values, just to replace them to make the analysis possible. We can either replace them by *constant values* (imputing a mean value), *random values* (randomly choosing one value from observed values) or we can *nonrandomly derive* them (conditional mean imputation, regression imputation). All of these methods are discussed in Donders *et al.* (2006).

Multiple imputation (MI) is the most sophisticated and praised method for handling missing data Allison (2001); Schafer (1997). Single imputation creates one imputed data set whereas MI creates more (usually 3–10, Donders *et al.* (2006)). We then do our statistical analysis on each of the newly created imputed data sets and combine all the results together. The combining process is called *pooling* van Buuren & Groothuis-Oudshoorn (2011). This produces statistically powerful and unbiased results, although some more work needs to be done when the missingness is under the MNAR assumption van Buuren & Groothuis-Oudshoorn (2011). Imputations may sometimes create implausible values (e.g. negative body mass) but can be easily handled by a method suggested by Hussain *et al.* (2010).

Advantages of MI

- generality
- ease of implementation
- incorporation of imputation error

Previous MI implementations

R already has several several MI packages that can be used to deal with missing data. *mice* van Buuren & Groothuis-Oudshoorn (2011) and *Amelia* Honaker *et al.* (2011) packages are simple to use than *mi* package Gelman & Hill (2011). *mi* on the other hand has more thorough MI diagnostics. The main reason we chose to extend *mice* was that it is very easily extensible. It already has a lot of MI methods that a user can use and to add a new one is as easy as creating a function and naming `mice.impute.***`, where `***` stands for the new method name.

Problem

The problem of existing *mice* methods is that all of them treat data as independent. That means that we

do not take in account possible similarities among individual observations in the imputation process. Because of that, traits tend to lose their phylogenetic signal Blomberg *et al.* (2003) which makes the imputation less accurate. Therefore we implemented and present new MI methods which preserve phylogenetic signal that extend the already mentioned R *mice* package.

Methods

Our methods are based on the idea of incorporating phylogenetic variance–covariance matrices that can be obtained from phylogenetic trees. There are several possible ways to implement these methods as suggested by Swenson (2014) and the one we have chosen is to make use of a *generalized linear model* Garland & Ives (2000). A different solution would be to implement it using *phylogenetic eigenvectors* but it has been suggested that *generalized linear models* perform better in the imputation Rohlf (2001); Freckleton *et al.* (2002). We have created a new R package called *phylomice* that contains our phylogenetic imputation methods and we propose that it should serve as a base package for other phylogenetic imputation methods to come in the future.

phnorm

The method called *phnorm* stands for *phylogenetic norm*. It is based on the *mice.impute.norm* (*norm*) function from the *mice* package which imputes data using *Bayesian linear regression analysis*. van Buuren & Groothuis-Oudshoorn (2011) notes that using *norm* for all columns is similar to the *NORM* method by Schafer (1997).

The regression problem can be written as

$$Y = \beta X + \epsilon, \tag{1}$$

as noted by Martins & Hansen (1997).

norm method uses a simple linear model to draw β parameter estimates that can be described as

$$\hat{\beta} = (X^T X)^{-1} X^T y, \tag{2}$$

whereas *phnorm* uses a generalized linear model to do that

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y, \tag{3}$$

where Σ is the phylogenetic variance–covariance matrix. Imputed values are estimated as

$$\hat{Y} = \hat{\beta} X + \epsilon, \tag{4}$$

where the error term has a following distribution

$$\epsilon \sim N(\mu, \sigma^2 c_h). \tag{5}$$

As Garland & Ives (2000) describes in the appendix B, if c_{ih} denotes a shared branch length between species i and h then Σ_{ih} is a vector of values c_{ih} for all species i other than h. c_{hh} stands for the total branch length of species h. Following equations then describe the mean of the error distribution and a part of a variance c_h

$$\mu = \Sigma_{ih}^{'} \Sigma^{-1} (X - \bar{x}), \tag{6}$$

$$c_h = c_{hh} - \Sigma_{ih}^{'} \Sigma^{-1} \Sigma_{ih}. \tag{7}$$

phpp

This method is based on a phylogenetic permutation model described by Lapointe & Garland (2001). The model creates a transition probabilities matrix from the phylogenetic tree and imputes already existing values based on other species relation.

Simulation

Variable name	1st val	2nd val	3rd val	4th val	5th val
Tree size	32	64	128	256	512
Trait size	2	5	10	15	-
Tree shape	balanced	left	pure birth	-	-
λ transform	0	0.5	0.8	1	-
Missing data mechanism	MAR	MCAR	-	-	-
Missing data level	1	2	3	-	-
Methods	norm	phnorm	phpp	-	-

Table 1: List of simulation parameters and their values. We tested every combination of parameters in this table.

We simulated our proposed methods by setting different combinations of parameters. Specifically, we observed how the methods behave when we change a *tree size* (which equals the number of species), *trait size*, *tree shape*, λ *transformation* parameter and a *missing data mechanism* with a certain *amount of missing data*. We tested our *phnorm* and *phpp* methods along with the *norm* method from the *mice* package for comparison.

The simulations were done for each combination of our input variables (described in table 1) and consisted of the following steps:

1. Set simulation parameters
2. Generate tree
3. Generate data
4. Analyse data
5. Generate missing values
6. Impute missing values
7. Analyse data after imputation

First thing after setting the parameters was to generate a tree. Trees were specified by a number of tips and a shape. Examples of simulated tree shapes are shown in figure 1. Balanced and left trees were generated using the *ape* package Paradis *et al.* (2004), while the pure-birth trees were generated with the *phytools* package Revell (2012).

In the second step, we generated our data. Data dimensions were defined by the *tree and trait* size. To ensure that the data contains any phylogenetic signal, we modified the tree we generated in the first step using the specified λ parameter. With this new tree, we created each trait under the Brownian motion Revell (2012).

We analysed our data and saved the following information for each trait: mean, median, mode, standard deviation, λ value and K value. The next step was to create some gaps in the data. We decided to only test missing data in one of the traits. We have simulated two missing data mechanism - MCAR and MAR, both of them in three levels. Levels of MCAR were simply corresponding to the percentage of missing data, where level 1 was 10%, level 2 was 50% and level 3 was 80%. These values were randomly sampled which means that there is no relationship with other traits. Missing data in the MAR mechanism were related to the values of the first trait with a probability $e^{-k|x_1|}$, where x_1 was a value of the first trait and k the parameter govern-

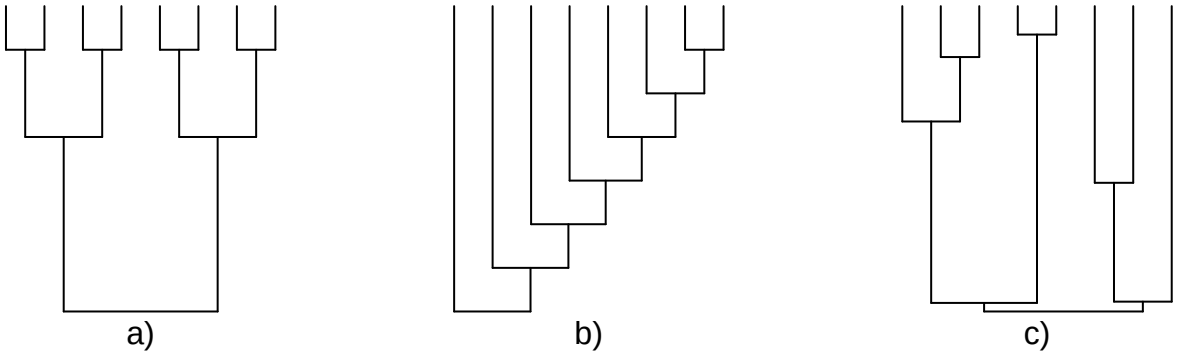

Figure 1: Tree shapes used in simulations – a) balanced tree, b) one sided tree (left), c) pure-birth tree.

ing the strength of the relationship between the two traits. k was set to 0.3, 0.8 and 1.3 for levels 1, 2 and 3, respectively.

We proceeded to the imputation after that. We created 5 imputed data sets in each simulation and we set the `maxit` parameter of the *mice* function to 100. We also specified the currently simulated method and filled the `psi` parameter for *phnorm* and *phpp* and also the `psiinv` parameter for *phnorm*. `psi` is the variance–covariance matrix obtained from the phylogenetic tree in *phnorm* and a matrix with cophenetic distances (Sneath *et al.* (1973)) in *phpp*. `psiinv` is just an inverse of `psi` needed only for the *phnorm* method. It needs to be calculated and passed to `mice` before, because it can be computationally expensive and it would have been calculated several times if we inversed `psi` inside `phnorm`.

Last step after the imputations finished was to analyse the data again and record the same parameters as we did before we made gaps in the data.

Output variables time imputation time *following were measured before and after imputation* column mean column mode column median column sd λ value K value

Note on performance

Other methods extension proposal

Results

Discussion

Funding

This research was supported by the University of Queensland, Australian Research Council grant, Erasmus Mundus NESSIE program and Student Grant Competition of Technical University of Liberec.

Acknowledgments

Thanks to S Rathnayake and ???? for insightful comments on previous drafts of this manuscript.

*

References

Allison, P.D. (2001) *Missing data*, volume 136. Sage publications.

Blomberg, S.P., Garland, T. & Ives, A.R. (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.

Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T. & Moons, K.G.M. (2006) Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, **59**, 1087–91.

Freckleton, R.P., Harvey, P.H. & Pagel, M. (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, **160**, 712–726.

Garland, T. & Ives, A. (2000) Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist*, **155**, 346–364.

Gelman, A. & Hill, J. (2011) Opening windows to the black box. *Journal of Statistical Software*, **40**.

Honaker, J., King, G. & Blackwell, M. (2011) Amelia II: A program for missing data. *Journal of Statistical Software*, **45**, 1–47.

Hussain, S., Mohammed, M.A., Haque, M.S., Holder, R., Macleod, J. & Hobbs, R. (2010) A Simple Method to Ensure Plausible Multiple Imputation for Continuous Multivariate Data. *Communications in Statistics-Simulation and Computation*, **39**, 1779–1784.

Kattge, J., Diaz, S., Lavorel, S., Prentice, I., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B., Wright, I. *et al.* (2011) Try–a global database of plant traits. *Global change biology*, **17**, 2905–2935.

Lapointe, F.J. & Garland, Jr, T. (2001) A generalized permutation model for the analysis of cross-species data. *Journal of Classification*, **18**, 109–127.

Martins, E.P. & Hansen, T.F. (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, pp. 646–667.

Nakagawa, S. & Freckleton, R.P. (2008) Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, **23**, 592–596.

Paradis, E., Claude, J. & Strimmer, K. (2004) Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, **20**, 289–290.

Revell, L.J. (2012) phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

Rohlf, F.J. (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution*, **55**, 2143–2160.

Rubin, D.B. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473–489.

Schafer, J. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Sneath, P.H., Sokal, R.R. *et al.* (1973) *Numerical taxonomy. The principles and practice of numerical classification.*

Swenson, N.G. (2014) Phylogenetic imputation of plant functional trait databases. *Ecography*, **37**, 105–110.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**, 1–67.

Appendix 1