ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Semester Paper Fall 2015

David Pham

Missing Data: Empirical Comparison between Imputation and Nearest Neighbors Algorithms

Submission Date: February 15th 2016

Adviser: Dr. Martin Maechler

To the *R* community and *ESS* developers for their contribution.

Abstract

Incomplete or missing data are common in scientific works, and the most common solution to cope with them is the simply to discard them. However, this might lead to bias in the conclusion. This semester paper summaries modern methods and offers an empirical comparison of the packages *amelia*, *imputeknn*, *mi*, *mice*, *softimpute* with the statistical software R.

Contents

Notation **ix**

1 Theoretical Background **1**
1.1 Mechanism of missingness 1
1.2 Statistical completion 2
1.3 Algorithmic completion 3

Bibliography **5**

List of Figures

List of Tables

Notation

Explain your symbols and abbreviations.

Chapter 1

Theoretical Background

This chapter provides an overview and an intuition on the field of missing data. It follows mainly Schafer and Graham (2002), Little and Rubin (2002), Van Buuren (2012), with some impute from Wikipedia (2015), Matloff (2015), Gelman and Hill (2006), Troyanskaya, Cantor, Sherlock, Brown, Hastie, Tibshirani, Botstein, and Altman (2001).

This chapter begins with a short description on the nature missingness, then describes several procedures in order to handle missing data.

1.1 Mechanism of missingness

Van Buuren (2012) describes two concepts helping us to understand how to solve the problem of missing data: intentional and unintentional missingness, as well as unit and item missingness. The experimenter can decide to not measures all possible variable in an experiment and encode his decisions as missing observations. This is a reasonable decision if the cost of measuring variables is material and unnecessary for some experimental case, such as in medical experimentation. However, it might also happen that the experimenter could not measure some variable, e.g. when a respondent to a survey refuse to answer to some question. In this case, the missingness is named unintentional. The second concept is often missingness is about unit and items: one says a unit is missing when none of the variables of interest could be measured, whereas item refers to some variable missing.

In order to complete missing data, assumptions need to be taken about the underlying mechanism creating missing observations: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Notation Let $Y \in \mathbb{R}^{n \times p}$ be the data matrix containing missing data for n observations with p variables, $R = (r_{ij})_{i=1,j=1}^{n,p} \in \{0,1\}^{n \times p}$ denotes the response y_{ij} (i.e. $r_{ij} = 1$ is y_{ij} is observed, and is 0 otherwise). Y_{obs} and Y_{mis} denote the observation that is observed, respectively, missing, such that $Y = (Y_{obs}, Y_{mis})$. Note that we always observed R whereas we usually do not have Y_{mis}.

MCAR The data are said to be *MCAR* if

$$P(R = 0|Y_{obs}, Y_{mis}) = P(R = 0),$$

or equivalently

$$P(Y = y|R = i) = P(Y = y),\ i \in \{0, 1\},\ y \in \mathbb{R}^{n \times p}.$$

It means the probability of being missing depends does not depends on the actual value of Y.

MAR For multiple imputation, one requires only $R \perp Y_{mis}$, that is

$$P(R = 0|Y_{obs}, Y_{mis}) = P(R = 0|Y_{obs}),$$

that is other observed variables impact of the probability of missingness but the missing mechanism only depends on the observed variables and not the actual missing value. In this case, we say the data Y are *MAR*.

MNAR The data are MNAR if

$$P(R = 0|Y_{obs}, Y_{mis})$$

can not be simplified. It essentially means that the rate of response depends on the actual value of the missing observations. The standard example is the survey about salary where people with extremely high salary tends to hide their earnings.

Modern statistical technique can handle MNAR and MAR cases, whereas simple technique only MCAR, which is quite restrictive.

1.2 Statistical completion

Complete case analysis Unfortunately, One of the most used technique to cope with missing data: the researcher only keeps observation that are complete. This might lead to valid analysis, as the method does not introduce any bias if the missing values are uniformly distributed. Nevertheless, this methodology can not work in modern settings where the probability of one missing variable is quite high: Many data points would be discarded.

Pairwise deletion This methods improve from the previous one by deleting observations only if the variable which is missing must be used in the model. This is typically relevant for computing correlation for example, although some care must be taken in this case.

Single imputation The data matrix is sorted according to some order, *last observation carried forward* is the method of replacing the missing value with last valid value. The missing value can also be replaced with the mean of the other observations, however, correlations are attenuated. Regression imputation use the other variables as predictors to replace the missing value, although precision is misleadingly augmented, hence does not reflect the statistical errors of the missing data. This problem is partially solved by multiple imputation.

Multiple imputation Under the MAR assumption, the multiple imputation (MI) is similar to bootstrapping method: the distribution of each variable conditional and the others is fitted, then in case of missing value, a sample is drawn from this distribution. The desired statistics are averaged except for the standard error which is constructed by adding the variance of the imputed data and the within variance of each data set. The last step solves the problem of understating uncertainty. Standard errors reflect missing-data uncertainty and finite-sample variation.

More precisely, in the one-dimensional case, if the sample is large enough so that the estimator Q follows a Gaussian distribution, then the estimate $\hat{Q}$ and the standard error T can be computed from the estimates of $(Q^j, U^j)_{j=1}^m$, Q^j, respectively, U^j being the fitted value of Q, respectively the standard error, for data sets j:

$$\begin{aligned}
\hat{Q} &= m^{-1} \sum_{j=1}^{m} Q^j, \\
\hat{U} &= m^{-1} \sum_{j=1}^{m} U^j, \\
B &= (m-1)^{-1} \sum_{j=1}^{m} (Q^j - \hat{Q})^2, \\
T &= \hat{U} + (1 + m^{-1})B.
\end{aligned}$$

For confidence interval, the Student's t approximation can be used with the degree of freedom given by

$$\nu = (m-1)\Big[1 + \frac{\hat{U}}{(1+m^{-1})B}\Big]^2.$$

The estimated rate of missing information for Q is approximately $\tau/(\tau+1)$ where $\tau = (1+m^{-1})B/\hat{U}$, the relative increase in variance due to non-response. See Schafer (1997) for more cases.

An advantage of MI is the number of need imputation: the efficiency based on m samples relative o an infinite number is $(1+\lambda/m)^{-1}$, where λ is the rate of missing information, which measures the increase in the large-sample variance of a parameter estimate due to missing values. $m = 20$ is often good in practice.

Obviously, the missing values problem is dealt before the analysis with MI, in contrast with maximum likelihood estimation. The danger from MI is the ability to use different models for imputation and analysis, which might lead to inconsistency.

1.3 Algorithmic completion

Singular value decomposition

Soft-impute completion

K-nearest neighbors completion Finally, one should not forget why these technique exists:

> With or without missing data, the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest – not to estimate, predict, or recover missing observations nor to obtain the same results that we would have seen with complete data. Schafer and Graham (2002)

Bibliography

Gelman, A. and J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Little, R. and D. Rubin (2002). Statistical analysis with missing data.

Matloff, N. (2015). New r software/methodology for handling missing data.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. and J. W. Graham (2002). Missing data: our view of the state of the art. *Psychological methods 7*(2), 147.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman (2001). Missing value estimation methods for dna microarrays. *Bioinformatics 17*(6), 520–525.

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

Wikipedia (2015). Imputation (statistics).

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

. . .

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Muster	Student

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics.*

Place, date:

Zurich August 19th 2009

Signature(s):

bla

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.