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

List of Figures

List of Tables

Notation

Explain your symbols and abbreviations.

Chapter 1

Introduction

Description of the work. Prepare the reader for the following chapters.

You will cite litterature here, typically

Chapter 2

First Chapter

2.1 To include a picture

Figure 2.1: Old Faithful Geyser eruption lengths, $n = 272$; binned data and two (Gaussian) kernel density estimates ($\times 10$) with $h = h^* = .3348$ and $h = .1$ (dotted).

Or also with `includegraphics`:
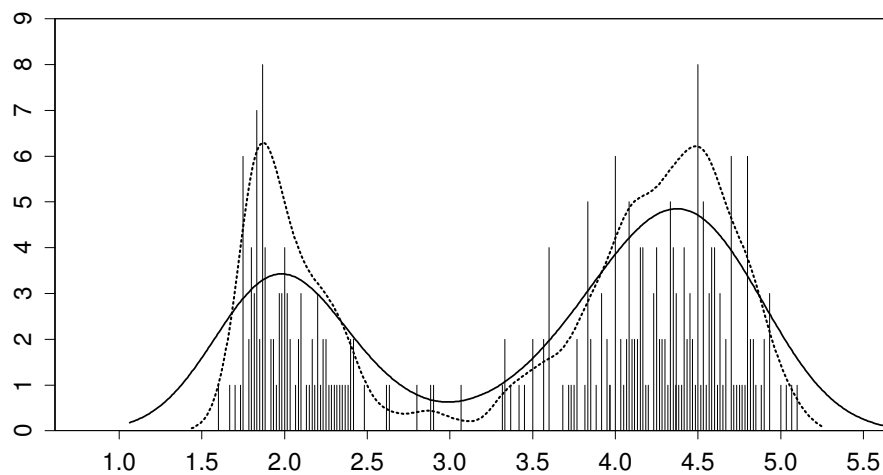

Figure 2.2: Old Faithful Geyser eruption lengths, $n = 272$; binned data and two (Gaussian) kernel density estimates ($\times 10$) with $h = h^* = .3348$ and $h = .1$ (dotted).

2.2   To make a proof

Proof. $1 + 1 = 2$ □

2.3   To include **R** code

See information in Appendix **??**.

2.4   Other information

Put a text between quotes: make sure to use nice quotes, such as "quote".

Cite a document in the bibliography (an example here): **?**. Or mention that **?** (a person) or **?** (two persons) have already done quite a bit work.

Referencing a different part of your work: please refer to Appendix **??**.

Chapter 3

First Chapter

3.1 To include a picture

Figure 3.1: Old Faithful Geyser eruption lengths, $n = 272$; binned data and two (Gaussian) kernel density estimates ($\times 10$) with $h = h^* = .3348$ and $h = .1$ (dotted).
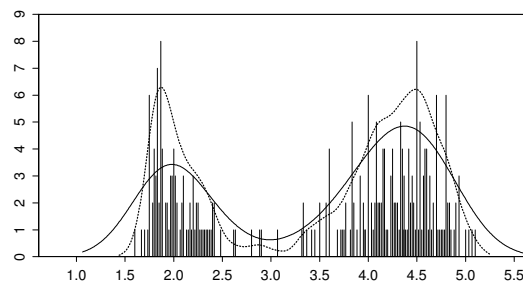
Or also with `includegraphics`:
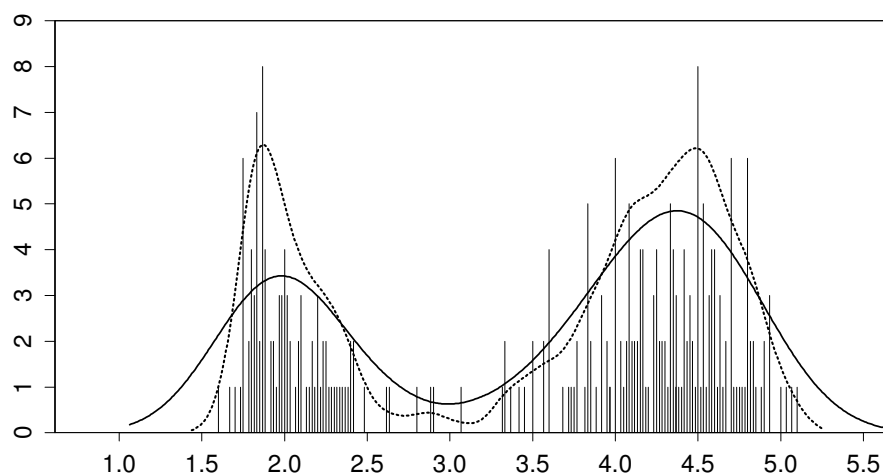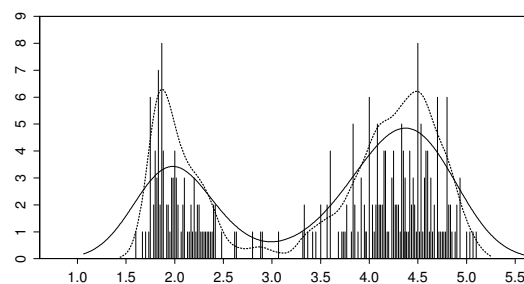

Figure 3.2: Old Faithful Geyser eruption lengths, $n = 272$; binned data and two (Gaussian) kernel density estimates ($\times 10$) with $h = h^* = .3348$ and $h = .1$ (dotted).

3.2   To make a proof

Proof. $1 + 1 = 2$ □

3.3   To include **R** code

See information in Appendix **??**.

3.4   Other information

Put a text between quotes: make sure to use nice quotes, such as "quote".

Cite a document in the bibliography (an example here): **?**. Or mention that **?** (a person) or **?** (two persons) have already done quite a bit work.

Referencing a different part of your work: please refer to Appendix **??**.

Chapter 4

Summary

Summarize the presented work. Why is it useful to the research field or institute?

4.1 Future Work

Possible ways to extend the work.

Bibliography

Author, F. and S. Author (year of publication). Title of the article. *Journal where the article has been published volume of the journal*(issue number), firstpage–lastpage.

Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics 27*(2), 95–107.

Stahel, W. and S. Weisberg (1991). *Directions in Robust Statistics and Diagnostics, 2 vol.* N. Y.: Springer-Verlag.

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):**

Mustern ....   *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: **Signature(s):**

Zurich August 19th 2009   *bla* ....

....

....

....

....

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