University of Oxford

MSc in Statistical Science

Final thesis

# Missing data imputation for Haemorrhagic shock prediction

*Author:*
Antoine Ogier

*Supervisor:*
Pr. Julie Josse
(École polytechnique)
Pr. Geoff Nicholls
(University of Oxford)

September 2018

**Abstract**

Lorem ipsum dolot sit amet nunc cui Brexit.

# Acknowledgements

Lorem ipsum dolot sit amet

# Contents

# Introduction

# Chapter 1

# Goal and data

## 1.1 The problem of haemorrhagic shock

Prediction is very hard, as described in [1]

## 1.2 The Traumabase data

## 1.3 Exploratory data analysis

### 1.3.1 Variables

### 1.3.2 Missing data

## 1.4 The problem of imputation

# Chapter 2

# Imputation methods

## 2.1 Main types of imputation

### 2.1.1 Joint parametric specification

### 2.1.2 Fully conditional specification: the MICE algorithm

### 2.1.3 Low-rank approximation for imputation

### 2.1.4 ML-based

## 2.2 Multiple imputation

### 2.2.1 Principle

### 2.2.2 Rubin's rule and prediction aggregation

## 2.3 Normality hypothesis: transforming the data

# Chapter 3

# Methodology: imputation and the validation split

The task we are trying to solve is quite particular: we are trying to impute the missing values in the data, not to perform a statistical analysis, but to select and train a prediction model. Our final benchmark of performance is not parameter estimation but predictive loss. As the next two chapters will show, although this seems like a minor difference, this actually leads to some major changes.

It is interesting to note in that regard how the communities of statistics and machine learning seem to lack any point of convergence on the subject. Missing data have been an active field of research in statistics for a long time [2], and many complex methods have been developed and proven for statistical inference [3] (such as those described in Chapter 2).

On the other hand, there is almost no research on these methods when applied to prediction. Even recent manuals for machine learning [4] generally make only a quick mention of missing data, and in practice it is extremely rare for anything else than imputation by the mean to be used. *Scikit-learn* [5], by far the most-used machine learning package in the community, only proposes implementation by the mean as of now.

A few research papers [6] [7] try to assess the performance of more modern imputation methods when used in a predictive context. However, they do not propose any framework or theory on this endeavour. They take it for granted that they can impute the whole dataset before performing the subsequent analysis. However, when performing prediction there is one significant difference with statistical inference, which we describe further in this chapter: the data is split into one dataset to learn the model, and another one to validate its performance. This raises many questions that we discuss here and in Chapter 4.

After laying out the general framework of Empirical Risk Minimization [8], which is the general paradigm used for prediction, we adapt it to fit the context of missing data. We notice that current implementations of modern imputation methods are incompatible with this framework, and try to devise solutions for this issue.

## 3.1 Empirical risk minimization (ERM): classical context

We first describe Empirical Risk Minimization (ERM) without missing data, as described in [8].

### 3.1.1 Context and notations

We are provided with a matrix $X$ of size $n \times p$ and response vector $y$ of size $n$. Our goal is to learn a model in the form $\hat{y} = f(x, \psi)$, where $\psi$ is some parameter to choose, $\hat{y}$ is a predicted value for $y$ and $f$ is a fixed parametric predictive function (usually corresponding to a choice of function in a given class).

The quality of a prediction is evaluated with the loss $L(y, \hat{y})$. The end goal in this context is to find the parameter $\hat{\psi}$ which minimizes the risk: $R = \mathbb{E}(L(y, f(X)))$. However, we do not have access to the real expectation of the risk, so we must use a proxy for this value. We define the empirical risk:

$$R_{\text{emp}}(y, f(X, \psi)) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(X_i, \psi))$$

Empirical Risk Minimization corresponds to choosing $\psi$ minimizing $R_{emp}$ for our $X$ and $y$. However, this is not enough. We do not just want the optimal $\psi$, we also want a measure of how well the final model would perform on new data. This is important because this is what we will take into account to make our choice of $f$. But the empirical risk gives us no measure of how well our model generalizes whatsoever, only of how closely it can fit known data.

To address the issue of model selection, the standard practice is to measure the error on some data that were not used to learn the model: it is the principle of cross-validation.

### 3.1.2 Cross-validation

To perform Cross-validation, we divide the data in two datasets: first we choose $n_A < n$ entries in the dataset to be used to loearn $\psi$: this is the training dataset $X_A$ and response $y_A$. We denote $I_A = (i_1, \dots, i_{n_A})$ the set of indices chosen for the training data. The rest of the observations are noted $X_V$ and $y_V$ and called the validation dataset.

Once this is done, ERM is performed as before, using only the training data. The obtained parameter $\hat{\psi}$ can then be evaluated with the validation error:

$$R_V = \frac{1}{n_V} \sum_{i=1, i \notin I_A}^{n} L(y_i, f(X_i, \hat{\psi}))$$

It is this value that we can compare to choose our our model class $f$.

## 3.2 ERM with missing data: the problem of current methodologies

We now place ourselves in the same context as before, except some values are missing from $X$, both in the training and the validation data. The context is almost the same as before: choosing a parametric model that takes as input the observed data and outputs a prediction for $y$.

### 3.2.1 Imputation seen as an ERM

Remember that the purpose of this work is to impute the data independently of the predictive model used afterwards. This does not change the framework of ERM but it does mean that we cannot use any function we like to go from $X$ to $\hat{y}$. The prediction is the composition of two steps.

**Imputation step** First we choose an imputation model $X^{\text{complete}} = g(X, \phi)$ where $X^{\text{complete}}$ is the completed dataset and $\phi$ some parameter. This is similar to predicting $y$ as we did previously with one caveat: we do not know the true data, even on the training dataset. Thus we choose $\hat{\phi}$ to minimize some unsupervised loss

$$L'(g(X, \phi), \phi)$$

measuring the fit of the model to the data. Generally, this means that we choose a parameter that maximizes the likelihood of the observed data

according to some generative model (though it is not always the case). Once this is done, we obtain a completed dataset $\hat{X}$.

**Prediction step** Once the imputation is done, we can perform as before to choose a parameter $\hat{\psi}$ that minimizes the empirical risk when using the completed data:

$$R_{\text{emp}}(y, f(\hat{X}_i, \psi)) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\hat{X}_i, \psi))$$

Putting it all together, we can define

$$h(X, (\psi, \phi)) = f(X^{imp}, \psi) = f(g(X, \phi), \psi)$$

the combined model that takes the observed data as input and outputs a predicted $y$. This is optimized as

$$\hat{\phi} = \arg\min_{\phi} L'(g(X, \phi, \phi)$$
$$\hat{\psi} = \arg\min_{\psi} R_{\text{emp}}(y, h(X, (\psi, \hat{\phi})))$$

We choose to use this notation to illustrate our point that imputation is an integral part of the ERM, not a separate, preliminary process. In particular, it means that in theory its parameters *must* be subjected to cross-validation just like those of the prediction. That is $X_A$ the training data are used to estimate $(\hat{\psi}, \hat{\phi})$ as shown just above. Then, we can compute a prediction $\hat{y}_V = h(X_V, (\hat{\psi}, \hat{\phi}))$ which can be compared to $y_V$ to evaluate the choice of model.

The bottom line is that just like for the prediction, the imputation parameter $\phi$ should be estimated only on the training data and then used on the validation data. As we will see, this raises an issue with the way current imputation methods are implemented.

### 3.2.2 Unsuitability of current methods

Over the years, many imputation have been proposed, and we describe some in Chapter 2. Mention sklearn

## 3.3 Possible solutions

### 3.3.1 Using current implementations

### 3.3.2 A new variant: Multivariate Normal Mode with reserved data

### 3.3.3 Comparison on simulated data

# Chapter 4

# Error sources and best imputation: the case of linear regression with missing data

## 4.1 Problem set-up

### 4.1.1 Notations

### 4.1.2 Objective

## 4.2 Partial resolution

### 4.2.1 General loss

### 4.2.2 When the validation set is fully observed

Strong consistency of the least square estimator [9]

### 4.2.3 When the data is large and the training data is fully observed

## 4.3 Consequences

### 4.3.1 Theoretical implications for our data

### 4.3.2 Verification with simulated data

# Chapter 5

# Analysis: imputing the Traumabase data for prediction

## 5.1 Criteria for evaluation

## 5.2 Single imputation

## 5.3 Multiple imputation

# Chapter 6

# Results

# Conclusion

# Bibliography


[1] Matthew J Pommerening, Goodman, et al. Clinical gestalt and the prediction of massive transfusion after trauma. *Injury*, 46(5):807–813, 2015.

[2] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[3] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

[4] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Yvonne Vergouwe, Patrick Royston, Karel GM Moons, and Douglas G Altman. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of clinical epidemiology*, 63(2):205–214, 2010.

[7] José M Jerez, Ignacio Molina, Pedro J García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115, 2010.

[8] Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[9] TW Anderson, John B Taylor, et al. Strong consistency of least squares estimates in normal linear regression. *The Annals of Statistics*, 4(4): 788–790, 1976.