

Time Series Project

Jacob Clinton George Smith-Kolff

2024-10-06

Contents

1	Introduction	2
1.1	What is missing data	2
1.2	Issues caused by missing data	2
1.3	Structure of the report	2
2	Theory	3
2.1	Time series	3
2.2	Missing data	3
2.3	Missing data imputation	4
3	Illustration of the method application	7
3.1	The data	7
3.2	The methods	7
4	Conclusion	8
5	Appendix	9

Chapter 1

Introduction

(Probably work on this section last)

1.1 What is missing data

1.2 Issues caused by missing data

1.3 Structure of the report

Chapter 2

Theory

2.1 Time series

Missing at random. This basically means that the missingness in the data is not informative

We begin by providing definitions for both uni variate and multivariate time series:

A univariate time series $X = \{x_1, x_2, \dots, x_t\} \in \mathbb{R}^t$ is a sequence t observations on a single variable.

This can be extended to multivariate time series:

$X = \{x_1, x_2, \dots, x_t\} \in \mathbb{R}^{t \times d}$ where each x_i is a d dimensional vector.

2.2 Missing data

remove the subsection numbers later

Data missingness is a common occurrence that is common in working with real world data. Missing data can arise from device failures such as measuring equipment failing, or from data censoring (such as from governments) [1]. Missing data typically falls into one of three categories:

2.2.1 Missing Completely at Random (MCAR)

Data is said to be missing completely at random if the distribution that describes how missingness occurs is independent of both the observed and unobserved values in the time series.

2.2.2 Missing at Random (MAR)

Missing at random is when missingness is related to the observed data, but is independent of the unobserved data. This means that there is some external factor that is causing the missingness. An example given in [2] is that data from a sensor is more likely to be missing on weekends since on some weekends the sensor is shutdown.

2.2.3 Not Missing at Random (NMAR)

Not missing at random means that the missingness is related to the value of the observation itself. An example of this is a sensor that will return a missing value if the recorded value is above 100°.

Come back and include probability notation

2.3 Missing data imputation

Most missing data imputation methods require MCAR or MAR, this is because these systems of missingness are not informative towards the missing values themselves.

Can extend this writeup later.

Idea is to start with the most basic techniques, then more the more advanced ones

2.3.1 Multivariate data imputation techniques

2.3.1.1 Simple (naive?) methods

- Mean/Median/Mode
- Forward/backward fill
- Linear interpolation
- MA

2.3.1.2 K-nearest neighbours

K-Nearest neighbors is a simple non-parametric machine learning algorithm, it can be used for either classification or regression. It can be applied in a time series data imputation context to estimate missing values in a time series.

$$\frac{1}{K} \sum_{j=1}^K Y_j$$

Consider a missing value x_i in a time series, k-nearest neighbors will impute a value for the missing value by calculating the mean of the datapoints in the neighborhood determined by a distance metric such as Euclidean distance. The algorithm is simple and is shown to be quite effective [3]. K-nearest neighbors algorithm has a hyper-parameter K, the number of nearest (non missing) data

points to consider. A choice of K that is too large may lead to smoothing over temporal patterns (under fitting) this has the effect of ignoring potentially important seasonal patterns. On the other hand a value of K that is too small could lead to being overly sensitive to the noise. The value for K needs to be carefully chosen, typically by cross validation, but in a time series context it could be chosen using sliding windows and forward chaining. The curse of dimensionality is an issue for K -nearest neighbors. With higher dimension it becomes harder to find data points that are closer, this can lead to an increased sensitivity to noise and misleading imputations. The computational cost also increases rapidly with more dimensions in the data making the algorithm less efficient. Another issue related to computational cost, K -nearest neighbors is a lazy learning algorithm so the complete dataset ends up being stored for large datasets this can make the algorithm slow or even infeasible

Come back
to this sec-
tion later

2.3.1.3 Multivariate Imputation by Chained Equations (MICE)

2.3.1.4 General Adversarial Networks (GAN)

2.3.2 Univariate data imputation

2.3.2.1 The struggle with univariate data imputation

Unlike with a multivariate time series, a univariate time series is only one sequence of observations $\{y_1, y_2, \dots, y_n\}$. This simpler form actually leads to increased difficulty when it comes to imputing a missing value of y , since we can no longer exploit any dependencies between the predictor variables in the series. With univariate imputation we can only rely on the previous (or future) values for y along with the implicit time variable ¹.

2.3.2.2 Last Observed Carried Forward (LOCF) Next Observation Carried Backward (NOCB)

²

2.3.2.3 Mean/Median imputation

2.3.2.4 ARIMA-based Imputation

Missing data can be imputed using an ARIMA model

$$\left(1 + \sum_{i=1}^q \beta_i B^i\right) y(t) W(t)$$

Where p is the AR part, d is the degree of differencing, q is order of the moving average part. In ARIMA based imputation the model will use the rest of the series to impute the missing value, if there are several missing values then each

¹moritz2015comparison

²ahn2022comparison

missing value is calculated one at a time and the model will use the previously imputed values as well as the data in the series to predict the next missing values
Kalman filtering

Chapter 3

Illustration of the method application

Using the methods of missing data in practice.

3.1 The data

3.2 The methods

Chapter 4

Conclusion

- [1] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.
- [2] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, and J. Stork, “Comparison of different methods for univariate time series imputation in r,” *arXiv preprint arXiv:1510.03924*, 2015.
- [3] H. Ahn, K. Sun, K. P. Kim, *et al.*, “Comparison of missing data imputation methods in time series forecasting,” *Computers, Materials & Continua*, vol. 70, no. 1, pp. 767–779, 2022.

Chapter 5

Appendix