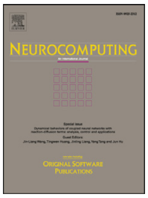
Neurocomputing 000 (2017) 1–8

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm

Faraj Bashir[a], Hua-Liang Wei[b,c,*]

[a] Department of Automatic Control and Systems Engineering, University of Shffield, Mapping Street, S1 4DT, UK
[b] Department of Automatic Control and Systems Engineering, University of Shffield, Mapping Street, Shffield S1 3JD, UK
[c] INSIGNEO Institute for in Silico Medicine, University of Shffield, Mapping Street, Shffield S1 3JD, UK

ARTICLE INFO

Article history:
Received 25 January 2016
Revised 25 January 2017
Accepted 12 March 2017
Available online xxx

Keywords:
Missing data
EM algorithm
VAR model
ECG

ABSTRACT

Imputing missing data from a multivariate time series dataset remains a challenging problem. There is an abundance of research on using various techniques to impute missing, biased, or corrupted values to a dataset. While a great amount of work has been done in this field, most imputing methodologies are centered about a specific application, typically involving static data analysis and simple time series modelling. However, these approaches fall short of desired goals when the data originates from a multivariate time series. The objective of this paper is to introduce a new algorithm for handling missing data from multivariate time series datasets. This new approach is based on a vector autoregressive (VAR) model by combining an expectation and minimization (EM) algorithm with the prediction error minimization (PEM) method. The new algorithm is called a vector autoregressive imputation method (VAR-IM). A description of the algorithm is presented and a case study was accomplished using the VAR-IM. The case study was applied to a real-world data set involving electrocardiogram (ECG) data. The VAR-IM method was compared with both traditional methods list wise deletion and linear regression substitution; and modern methods Multivariate Auto-Regressive State-Space (MARSS) and expectation maximization algorithm (EM). Generally, the VAR-IM method achieved significant improvement of the imputation tasks as compared with the other two methods. Although an improvement, a summary of the limitations and restrictions when using VAR-IM is presented.

1. Introduction

Throughout the literature, many imputation methods for missing data have been proposed. The methods fall primarily into two broad classifications: traditional and modern techniques. Traditional techniques such as simple deletion, averaging, or regression estimation are limited but still used in many cases. On the other hand, modern approaches such as multiple imputation (MI) and maximum likelihood (ML) routines, have proved superior and are have gained favour. In fact modern data imputation algorithms that use these approaches are very prevalent and can be easily administered in standard statistical packages such as Statistical Package for Social Sciences (SPSS) and Multivariate Autoregressive State-Space (MARSS or even standalone applications such as NORM [1,2]. The MI approach first imputes multiple data sets from random samples of the population using techniques such as bootstrapping [3] or data augmentation [4]. Then, using Rubins rules, the results from the imputed data sets are combined [5]. The ML technique for handling missing data is becoming commonplace in microcomputer packages. Specifically, ML algorithms are currently available in many existing software packages (e.g. EM algorithm) [6]. When conducted properly, both ML and MI techniques enable researchers to make valid statistical inferences when data are missing at random [7]. However, these techniques either have limitations or are difficult to carry out for dynamic systems modelling [8]. For example, many dynamic models involve autoregressive variables and the output is normally a linear or nonlinear combination of a lagged variable. The estimation of autoregressive models requires that the data be fully observed. With the existence of missing values, this is not possible, rendering it impossible to estimate the model. Furthermore, these methods often lead to bias in the estimates. In this paper, a new method is proposed for missing data imputation in multivariate time series datasets. The new algorithm utilizes a vector autoregressive model (VAR) to handle missing data by combining the prediction error minimization (PEM) [9] with an EM algo-

[*] Corresponding author at: Department of Automatic Control and Systems Engineering, University of Shffield, Mapping Street, Shffield S1 3JD, UK.

E-mail addresses: faabashir1@sheffield.ac.uk (F. Bashir), w.hualiang@sheffield.ac.uk (H.-L. Wei).

https://doi.org/10.1016/j.neucom.2017.03.097
0925-2312/© 2017 Elsevier B.V. All rights reserved.

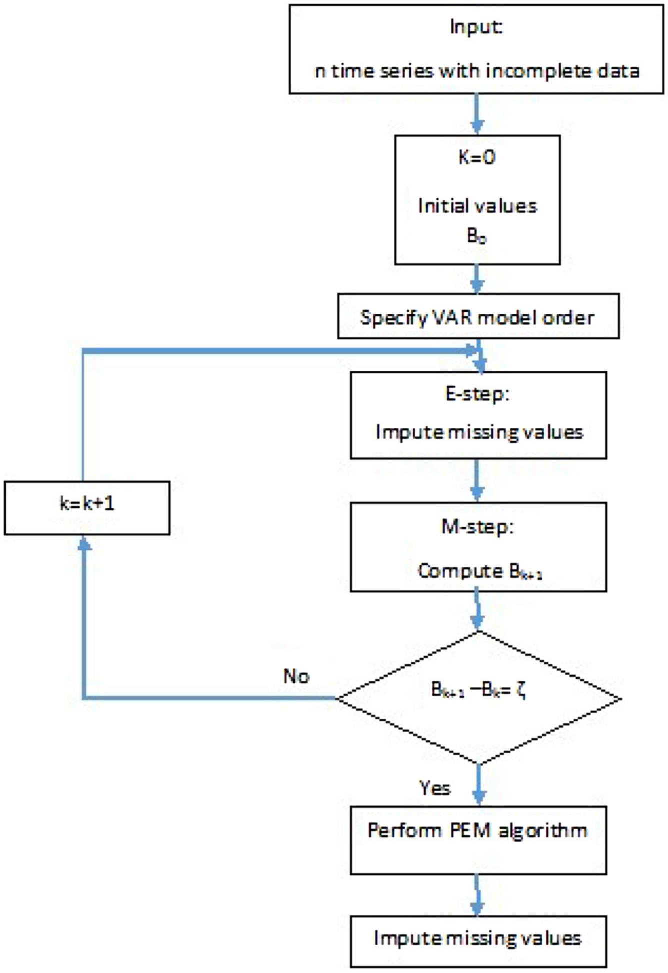


Fig. 1. The flow chart of the VAR-MI algorithm.

rithm. The new algorithm is called a vector autoregressive imputation method (VAR-IM). A description of the algorithm is presented and a case study was accomplished using the VAR-IM. The case study involved electrocardiogram waves that contain multivariate time series data. Also the advantages and limitations of the proposed method are analyzed. Finally, a simulation study of the proposed algorithm is compared to traditional and modern imputation methods.

2. Overview of traditional and modern data imputation techniques

Obtaining good, reliable, and complete data for a research study is often taken for granted, however, without good data; the results of a research project will be incorrect and could lead to significant errors in model development. For various reasons the obtained data may be corrupted with missing, incorrect, or distorted values. These anomalies may occur during or after the data collection process. The problem of how to deal with corrupted data has been a significant problem throughout many research fields for many years. Data imputation is the process of replacing missing, abnormal and distorted values of dataset. Many techniques of imputing missing data have been developed as it constitutes a central part of data mining and analysis [10]. For this study, two of the traditional and modern methods were selected as baseline comparisons to the proposed new algorithm. These are list wise deletion, linear regression imputation, MARSS package and EM algorithm.

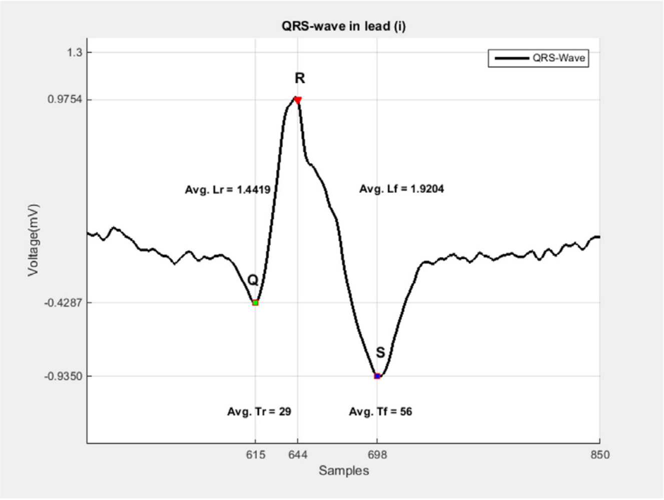


Fig. 2. QRS wave properties for complete data.

2.1. Listwise deletion

List wise deletion is among the simplest techniques for imputing missing data. Specifically, in this technique, all measured values at a specific time point, are ignored if one of the variables has a missing value for that specific measurement. Because this method removes the data with missing values, it decreases the number of variables and the length of sequences resulting in a reduced sample size. In dynamic modelling where all values are important for estimating the current values, the list wise deletion approach can significantly affect the autoregressive model estimation. Although even with these weaknesses, this approach is still being used for missing data analysis due to its simplicity. In some mainstream statistical programming such as R and SAS, this method is the most popular one for dealing with missing values, especially when analysing time series. However, there is no obvious indication that list wise deletion is adequate for handling missing data involving multivariate time series modelling [8].

2.2. Linear regression imputation

Linear regression imputation is a very general technique for dealing with missing values in time series analysis. Linear regression imputation uses the available data (observed data) to estimate the missing values by using a linear model:

$$Y_1 = B_{10} + B_{11}Y_2 + B_{12}Y_3 + \cdots + B_{1n}Y_n + e$$

$$Y_2 = B_{20} + B_{21}Y_1 + B_{22}Y_3 + \cdots + B_{2n}Y_n + e$$

$$Y_n = B_{n0} + B_{n1}Y_1 + B_{n2}Y_2 + \cdots + B_{nn}Y_{n-1} + e$$

$$\{Y_1\} = \{Z_1\}\{B\} + \{e\}$$

where $\{Y_1\}$ contains the imputation data, $\{B\}$ is the parameters of the linear model, $\{e\}$ is the error vector at each data point, and $[Z_1]$ is regression matrix with n time series and m length of observed data:

$$Z_1 = \begin{bmatrix} 1 & Y_{21} & Y_{31} & Y_{1n} \\ 1 & Y_{22} & Y_{32} & Y_{n2} \\ 1 & .. & .. & .. \\ 1 & Y_{2m} & Y_{3m} & B_{nm} \end{bmatrix}$$

The main advantage of this method is that it does not decrease the variation of data as compared to mean substitution. The main