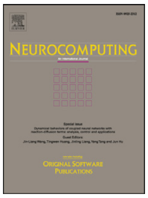
Contents lists available at ScienceDirect

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.

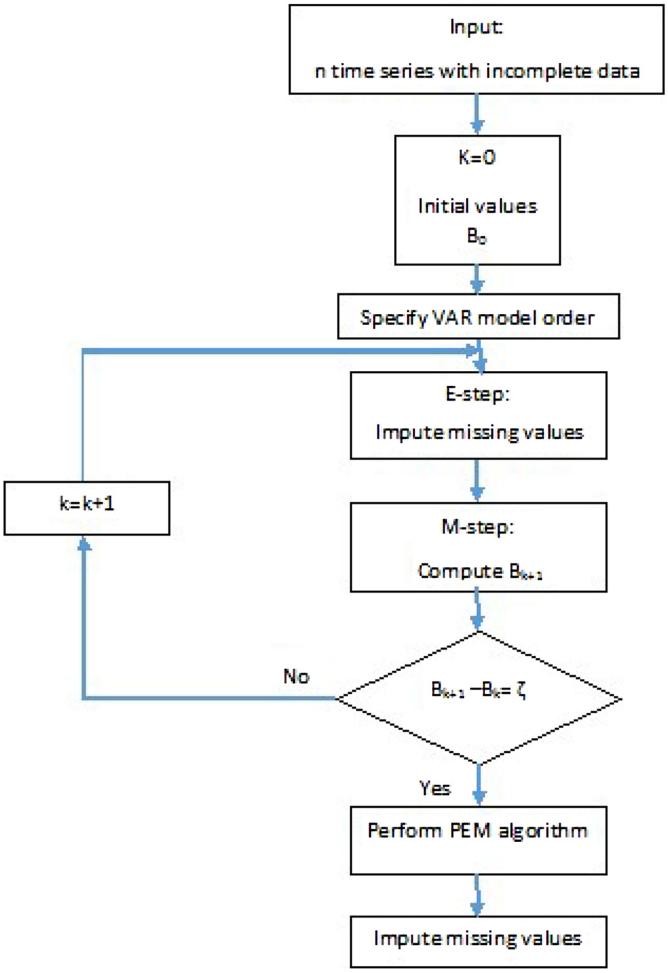


Fig. 1. The flow chart of the VAR-MI algorithm.

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.

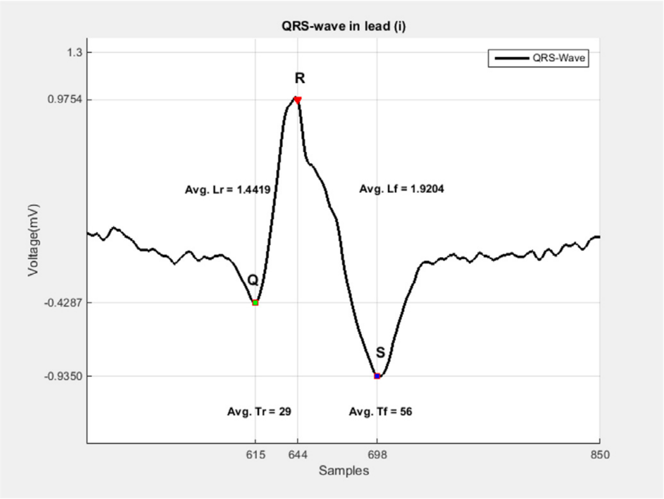


Fig. 2. QRS wave properties for complete data.

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:

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

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

Please cite this article as: F. Bashir, H.-L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.03.097

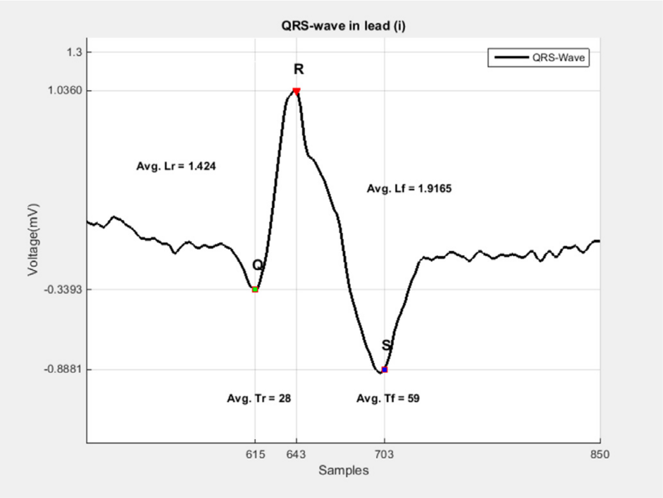


Fig. 3. QRS wave properties VAR-IM imputed data (10%*MCAR*).

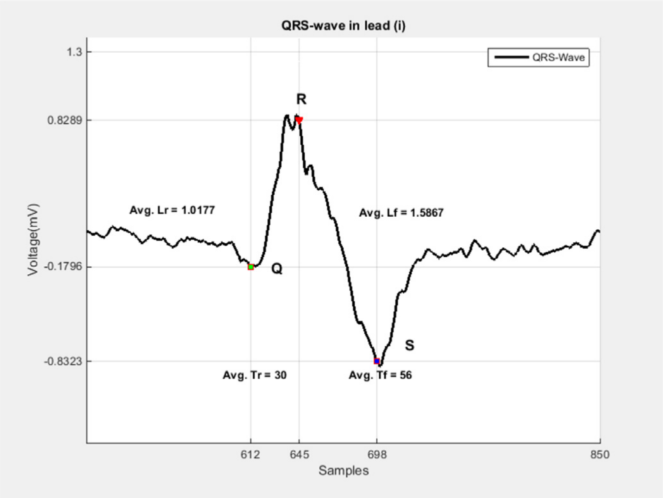


Fig. 4. QRS wave properties for linear-regression imputed data (10%*MCAR*).

drawback of this method is that it handles the available data as static, thus eliminating the property of autoregression.

2.3. Multivariate Auto-Regressive State-Space (MARSS) Model

The Multivariate Auto-Regressive State Space (MARSS) model was introduced in 2012 as the first complete package for handling missing data in multivariate time series data [11]. MARSS incorporates an expectation-maximization (EM) algorithm. It is an R package employing a special formula of vector autoregressive state-space models to fit multivariate time series with missing data via an EM algorithm. A MARSS model has the following matrix structure:

$$\begin{cases} x_t = A_t x_{t-1} + B_t b_t + \varepsilon_t \\ \\ y_t = C_t x_{t-1} + D_t d_t + \mu_t \end{cases} \tag{1}$$

where $\varepsilon_t \sim MVN(0,\ Q_t)$, $\mu_t \sim MVN(0,\ R_t)$ and $x_1 \sim MVN(\pi,\ \Lambda)$ or $x_0 \sim MVN(0,\ \Lambda)$

The state vector is represented by x_t and the measured value is designated by y_t. Driven by data, the model evolves but it is possible that some value may be missing when measuring y. The variables b_t and d_t are inputs representing for example some indicators or exogenous variables. A_t, B_t, C_t, and D_t are system matrices, ε_t and μ_t are process and non-process error respectively, Q_t and R_t are $m \times m$ and $n \times n$ variance–covariance matrices, respectively, where m is number of states and n the number of time series. Compared with the traditional approaches, MARSS can generate better results especially for multivariate time series modelling [12].

Table 1
VAR model order selection.

Model order	AIC	SC	LR	HQ	FPE
1	4.8001	6.4028	21.4948	5.4479	0.0002
2	2.1151	3.2691	17.8260	2.5816	0.0001
3	5.1866	9.3537	91.4041	6.8710	0.0002
4	5.5062	10.956	53.7518	7.7089	0.0002

2.4. EM algorithm

The expectation-maximization (EM) algorithm is an iterative algorithm for parameter estimation using maximum likelihood parameter values when the information (e.g. measurements) of some variables are incomplete [13–15]. The EM algorithm is achieved through two basic steps: estimation step (aka E-step) which replaces missing values by estimated values, and the maximization step (aka M-step) which estimates the parameters. These two steps alternately iterate until convergence [16,17]. The conditional expectations of missing data in observed series and estimates of model parameters in the E-step are calculated by

$$Q(B_n|B_{n+1}) = E_{(x_m|X_0),B_{n+1}}[logL(B; X_0, X_m)] \tag{2}$$

where, $L(B;\ X_0,\ X_m)$ is the likelihood function, B is the parameter vector, B_{n+1} is the estimate of the model parameters, X_0 is observed data, X_m is the missing data. In the M-step, the model parameters can be calculated using (2) to maximize complete data log likelihood function from the E-step:

$$B_{n+1} = arg_B max Q(B|B_n) \tag{3}$$

3. Overview of stationary multivariate time series

A time-series is a sequence of measured values arranged by their sequential time order. The time-series may be in either discrete or continuous time units. Multivariate time series processes are of considerable interest in a variety of fields of engineering, sciences, and medicine. By studying many related variables together rather than a single variable a better understanding of the observed process is often obtained. Nowadays, improved data collection methods permit large amounts of time series multivariate data to be collected from various application domains. For n time series random variables $y_{1t}, y_{2t}, \ldots, y_{nt}$, let Y_t denote a multivariate time series for an n-dimensional time series vector, where each y_{it} time series represents ith raw of Y_t vector, that is, for any time t, $Y_t = (y_{1t}, y_{2t}, \ldots, y_{nt})^T$ One of the fundamental objectives of multivariate time series analysis of Y_t is to fit the data to a model and demonstrate the dynamic relationships among univariate time series. The selection of each time series model, included in Y_t depends on the dynamic interrelationships between these time series variables which are affected directly by time lags between the data points for each time series. The multivariate time series data set Y_t is stationary time series if at arbitrary time intervals $t_1, t_2, \ldots, t_k$ the probability distributions of the component time series variables $y_{t1}, y_{t2}, \ldots, y_{tk}$ and $y_{t1-p}, y_{t2-p}, \ldots, y_{tk-p}$ are the same, where k is the number of the measured values p represents the lag. That means cross time intervals $t_1, t_2, \ldots, t_k$ throughout the stationary multivariate time series, has a random probability distribution of the observed data points with respect to the time lags. Consequently, any stationary multivariate time series should have the

Please cite this article as: F. Bashir, H.-L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.03.097

Table 2
A comparison of different methods for the heart-rate data (10% MCAR).

Method	The conventional 12 leads											
	i	ii	iii	avr	avl	avf	V1	V2	V3	V4	V5	V6
Complete-data	78.12	65.87	73.02	58.58	56.05	65.48	34.96	42.67	52.95	75.82	75.58	74.38
Missing-data	73.9	63.8	66.78	47.8	49.95	60.91	30.48	37.82	45.78	66.23	66.35	64.85
VAR-IM	79	67.08	70.13	54.58	55.08	65.48	37.73	43.58	50.48	70.97	73.51	72.13
List-wise	87.79	74.05	76.07	49.82	58.84	72.17	34.71	43.92	52.03	74.40	74.96	72.85
Linear-reg	76.82	67.80	70.28	50.07	56.57	100.40	76.32	52.65	57.55	85.47	83.57	69.87
MARSS	73.98	63.8	66.83	47.87	49.98	60.93	30.51	37.92	45.82	66.32	66.38	64.9
EM	75.37	64.05	67.33	49.3	51.27	61.38	31.31	38.6	48.95	70.57	66.36	69.6

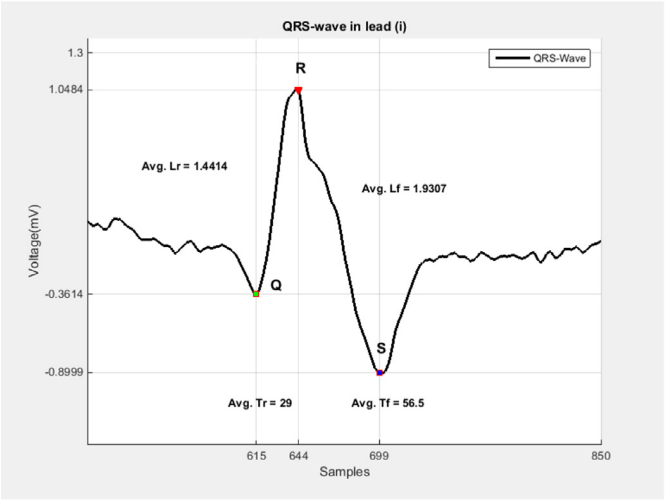


Fig. 5. QRS wave properties for EM imputed data (10%*MCAR*).

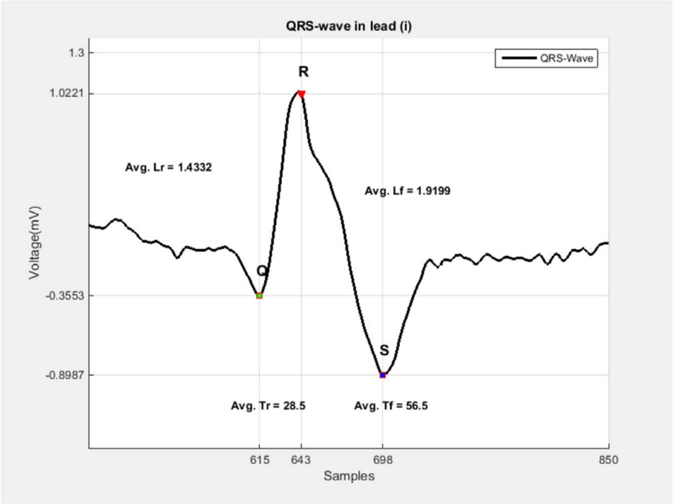


Fig. 6. QRS wave properties for MARSS imputed data (10%*MCAR*) .

same mean value M at any time intervals where:

$$M = E(Y_t) = \begin{bmatrix} m_1 \\ m_2 \\ .. \\ m_n \end{bmatrix} \tag{4}$$

In addition, the covariance matrix, Σ_Y, of a stationary time series Y_t is a constant matrix [18]:

$$\sum_Y = E[(Y_t - M)(Y_t - M)^T].$$

4. Filtering of multivariate time series

A multivariate linear (time-invariant) filter relating an l dimensional input series U_t to n-dimensional output series Y_t often formulated as

$$Y_t = \sum_{N=-\infty}^{\infty} U_{t-N} B_N \tag{5}$$

where B_N are $n \times 1$ matrices. The filter is physically realizable or causal if $B_N = 0$ for $N < 0$ leading to $Y_t = \sum_{N=0}^{\infty} U_{t-N} B_N$, which means that Y_t can be characterized by past values of the input U_t. The filter is said to be stable, if $Y_t = \sum_{N=-\infty}^{\infty} \|B_N\| < \infty$.Under the stability condition, together with an assumption that the input random vectors U_t have uniformly bounded second moments, the output random vector Y_t defined by (5), exists uniquely and represents the limit:

$$\lim_{r \longrightarrow \infty} \sum_{N=-r}^{r} U_{t-N} B_N \tag{6}$$

such that as $r \longrightarrow \infty$

$$Y_t = E\left[\left(Y_t - \sum_{N=-r}^{r} U_{t-N} B_N \right) \left(Y_t - \sum_{N=-r}^{r} U_{t-N} B_N \right)^T \right].$$

When the filter is stable and the input series U_t is stationary with cross-covariance matrices $\Gamma_U(p)$, Eq. (5) is a stationary process [19]. The cross-covariance matrices of the stationary process Y_t are then given by

$$\Gamma_U(p) = Cov(Y_t, Y_{t-p}) = \sum_{i=-\infty}^{i=\infty} \sum_{j=-\infty}^{j=\infty} B_i \Gamma_U(p + i - j) B_j^T \tag{7}$$

5. Vector autoregressive model (VAR)

The vector autoregressive model (VAR) is commonly used model for the analysis of multivariate time series. In many applications where the variables of interest are linearly each related to each other the VAR model has shown to be a good choice for representing and predicting the behaviour of dynamic multivariate time series [20]. It primarily provides good forecasts as compared to models from univariate time series and many other models. Because the VAR model can make conditions on the prediction paths of specified time series within the model itself, the forecasts from this model are relatively easy to derive [20]. In addition to time series analysis and prediction, the VAR model is additionally utilized for causality inference and strategy investigation of the multiple time series. In causality analysis, specific hypotheses of the causality of the time series under analysis are assumed, and the subsequent causal effects of each time series are outlined. This chapter concentrates on the use of the VAR model to analyse stationary multiple time series datasets with missing data.

Table 3
A comparison of different methods for the heart-rate data (20% MCAR).

Method	The conventional 12 leads											
	i	ii	iii	avr	avl	avf	V1	V2	V3	V4	V5	V6
Complete-data	78.12	65.87	73.02	58.58	56.05	65.48	34.96	42.67	52.95	75.82	75.58	74.38
Missing-data	73.90	63.80	66.78	47.80	49.95	60.91	30.48	37.82	45.78	66.23	66.35	64.85
VAR-IM	80.87	68.20	70.55	52.05	56.78	68.15	38.72	44.43	50.53	69.82	73.30	72.22
List-wise	71.63	62.55	64.85	42.27	48.17	58.17	28.43	35.63	44.35	63.57	63.35	61.85
Linear-reg	74.48	66.47	68.35	44.23	57.317	101.32	73.97	56.28	59.45	84.73	82.92	67.03
MARSS	71.67	62.55	64.93	42.32	48.22	58.15	28.50	35.68	44.42	63.63	63.38	61.90
EM	74.80	63.05	65.77	44.28	51.70	59.13	29.68	36.87	49.83	69.58	64.33	67.95

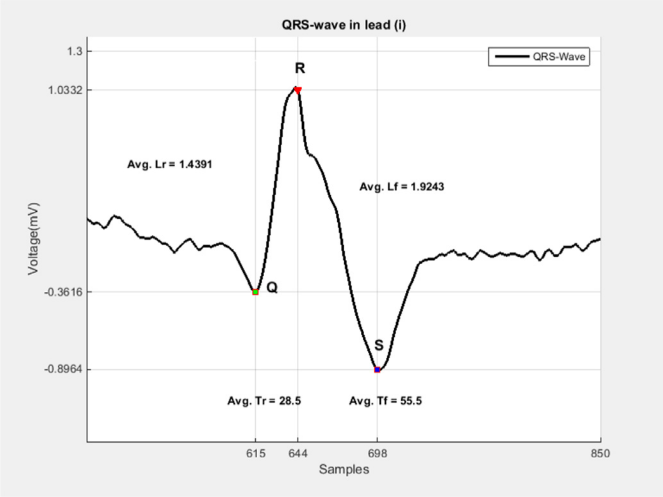


Fig. 7. QRS wave properties for VAR-IM imputed data (20% MCAR).

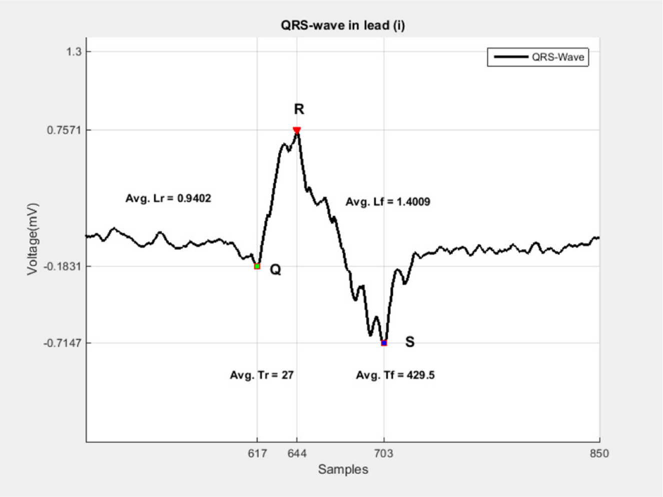


Fig. 8. QRS wave properties for linear-regression imputed data (20%*MCAR*).

5.1. Vector autoregressive state space model

State-space models are models that use state variables to describe a system by a set of differential or difference equations. State variables can be reconstructed from the measured input–output data, but the variables themselves are not measured during an experiment. A state-space models can be estimated using in either time and frequency domains. In this paper, the discrete-time state-space model is used to present the multivariate time series data set, having the following structure [21]:

$$X(t+Ts) = AX(t) + BU(t) + E(t) \tag{8}$$

$$Y(t) = CX(t) + DU(t) + e(t) \tag{9}$$

where $x(t)$ is the vector of state values, A is the state matrix, B is the input matrix, C is the output matrix, D is the feedforward matrix, Y and U are the input and output vectors, respectively, and $\varepsilon(t)$ are state errors as specified with the matrix q. Matrix r contains the output errors, $e(t)$.

5.2. VAR model for stationary time series

Let $Y_t = (y_{1t}), (y_{2t}, \ldots, y_{mt})^T$ be an $(m \times 1)$ time series vector. A $VAR(p)$ model for the multiple time series can be represented by

$$Y_t = A_0 + \sum_{i=1}^{p} A_i Y_{t-i} + \varepsilon(t) \tag{10}$$

$$Y_t = A_0 + A_1Y_{t-1} + A_2Y_{t-2} + \cdots + A_pY_{t-p}\varepsilon(t) \tag{11}$$

where $t = 1, \ldots T$, A_i are $M\times$ coefficient matrices and $\varepsilon(t) \in (0, \Sigma)$ denotes an $M \times 1$ vector of white noise.

Eq. (11) can be written in lagged notation:

$$A_p(L)Y_t = A_0 + \varepsilon(t) \tag{12}$$

where

$$A_p = I_m - A_1L - \cdots - A_pL^p \tag{13}$$

The stability of the VAR model is depended on the roots of (14)

$$|I_m - A_1z - \cdots - A_pz^p| = 0 \tag{14}$$

6. VAR-IM algorithm

The proposed algorithm for imputing missing data into a multivariate time series dataset is to use a vector autoregressive-imputation (VAR-IM) method combined with an EM algorithm together with a prediction error minimization (PEM) algorithm. The method based on a combination of these algorithms can significantly improve the imputation performance for dealing with missing data problem. Specifically, in the first step, the traditional linear interpolation estimate is made for an initial guess of the missing data. Then a VAR(p) model is estimated by selecting the best lag value p. Finally, the parameters of the VAR(p) model are estimated by alternatively using EM and PEM algorithms resulting in an improved value for the data imputation. Basically, the alternation of the two algorithms between imputing missing data and estimating models, improves the model performance by applying PEM algorithm in a way similar to the EM algorithm. The flow chart for the proposed VAR-IM algorithm is shown in Fig. 1.

The VAR-IM method formalizes an intuitive idea for identifying a best VAR model for imputing missing data:

- Calculate the initial values to start the algorithm.

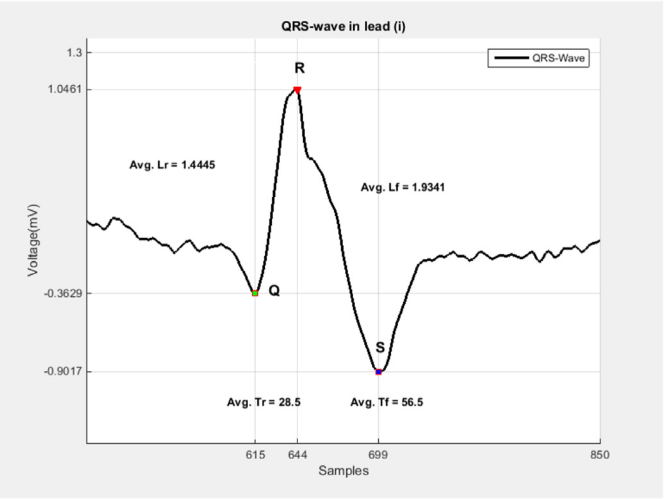


Fig. 9. QRS wave from the imputed data by the EM algorithm (20% MCAR).

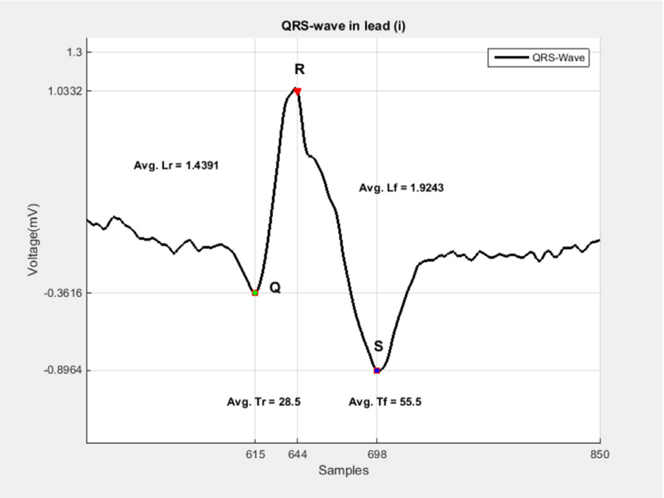


Fig. 10. QRS wave from the imputed data by the MARSS algorithm (20% MCAR).

- Check the causality of time series.
- Select the order of the identified VAR^* Model.
- Impute the missing values by using VAR^*.
- Identify the new VAR model.
- If no convergence, go to step III, otherwise, go to step VII.
- Perform PEM algorithm to update the missing values.
- Impute the missing values.

For more details, assume that X_t represents a multivariate data set and a set of VAR models can simulate X_t with different lags $p = 1, 2, 3,$ and parameters A_p. If there is no missing values, it is of interest to calculate the least squares estimate of A_p based on:

$$X_t = \phi A + E \tag{15}$$

For dynamic systems the auto-regression process depends on the past values of the targeted data point, if the time series includes missing values, means that there is past values missed and the auto-regression cannot be applied in (15). In this case the traditional approaches such as list wise will not work, because ignoring of missing values will effect on the property of dynamic system. To start up the estimation process correctly, the initial values are required; the simple way to determine these initial values is using a simple traditional method such as linear interpolation. We will denote this by expressing X_t as (X_{tmiss}, X_{tobs}), where X_{tmiss} denotes the multivariate data set with missing values, and X_{tobs} represents the multivariate data set with replacing missing values by initial values (imputed by interpolation technique).

Consequently, Eq. (15) becomes [22]:

$$\hat{X_t} = \phi_k A_{pk} + E \Longrightarrow \hat{X_t} = \phi_0 A_{p0} + E \tag{16}$$

$$A_{pk} = (\phi_k^T \phi_k)^{-1} \phi_k^T X_k \Longrightarrow A_{p0} = (\phi_0^T \phi_0)^{-1} \phi_0^T X_0 \tag{17}$$

where ϕ_0 is the initial regression matrix, $k = 0, 1, 2,$ and A_{p0} is the initial coefficients matrix of the select $VAR(p_k)$ model.

The order of the model p_k is updated until the difference$A_{pk} - A_{p(k+N)}$ is less than ξ, where ξ is a prescribed small value [22].

6.1. Model order selection

The model selection for the $VAR(p)$ model is usually specified utilizing model selection criteria. The basic idea is to identify models with different lags values $p = 0, 1, 2, \ldots, p_{max}$ and select the p lag value that can minimizes the model selection criteria [23]. Model order selection formula is represented by

$$IC(p) = ln|\sum(p)| + S_T.\varphi(m, p) \tag{18}$$

where $|\Sigma(p)|$ is the covariance matrix of the residual error S_T are the indexed values sequence $(1, \ldots, T)$, and the penalty function $\varphi(m,\ p)$ which impedes the large models order. The term $|\Sigma(p)|$ is non-growing function whereas the function $\varphi(m,\ p)$ increases with the order P. The basic idea of the model order selection depends on balancing these two functions. There are five techniques for model order selection in the applied $VAR(P)$ model literature generally broadly utilized:

- Akaikes information criterion (AIC) [24].

$$AIC(p) = ln|\sum(p)| + \frac{2}{T}pm^2 \tag{19}$$

Where the penalizing function $\varphi(m, p) = pm^2$ and $S_T = \frac{2}{T}$

- Schwarz criterion (SC) [25].

$$SC(p) = ln|\sum(p)| + \frac{lnT}{T}pm^2 \tag{20}$$

Where the penalizing function $\varphi(m, p) = pm^2$ and $S_T = \frac{lnT}{T}$

- Hannan–Quinn criterion (HQ) [26].

$$SC(p) = ln|\sum(p)| + \frac{2ln(lnT)}{T}pm^2 \tag{21}$$

For which the penalizing function $\varphi(m, p) = pm^2$ and $S_T = \frac{2ln(lnT)}{T}$

Note that for all the three criteria, the penalty function $\varphi(m,\ p)$ has the same formula.

- Final Prediction Error (FPE) [27].

$$FPE(p) = [\frac{T + mp + 1}{T - mp + 1}]^m |\sum(p)| \tag{22}$$

- Likelihood ratio test ($LRtest$) [28].

Where $j = 1, 2, (p - 1)$ Other techniques do exist. They were not included in this study because they are not widely used in the applied VAR model literature.

Table 4
Q–R–S wave properties in case of 10% MCAR.

	Data		Imputed data 10% MCAR missing				
	Complete data	Missing-data	VAR-IM	List-wise	Linear-reg	MARSS	EM
MeanError_Qwave	−0.004	NAN	−0.0109	–	0.1420	−0.0417	−0.006
MeanError_Rwave	0.021	NAN	0.0243	–	0.0277	0.0212	−0.0068
MeanError_Swave	−0.0155	NAN	−0.342	–	−0.0576	−0.0147	−0.018
avg_riseTime	29	NAN	28	–	30	28.5	28
avg_fallTime	56	NAN	59	–	56	56.5	57
avg_riseLevel	1.4419	NAN	1.424	–	1.0171	1.4332	1.4312
avg_fallLevel	1.9204	NAN	1.9165	–	1.5867	1.9199	1.9296

Table 5
Q–R–S wave properties in case of 20% MCAR.

	Data		Imputed data 20% MCAR missing				
	Complete data	Missing -data	VAR-IM	List-wise	Linear-reg	MARSS	EM
MeanError_Qwave	−0.004	NAN	−0.0067	–	0.0131	0.0016	0.0014
MeanError_Rwave	0.021	NAN	0.0191	–	0.0630	0.022	0.0163
MeanError_Swave	−0.0155	NAN	−0.018	–	−0.0802	−0.0096	−0.0191
avg_riseTime	29	NAN	29	–	27	28.5	28.5
avg_fallTime	56	NAN	56.5	–	429.5	55.5	56.5
avg_riseLevel	1.4419	NAN	1.4414	–	0.9402	1.4391	1.4445
avg_fallLevel	1.9204	NAN	1.9307	–	1.4009	1.9243	1.9341

7. Case study

The significance of a good data imputation process is especially important in the field of medicine where discovery and imputation of missing values can help to identify abnormal conditions and reduce incorrect diagnosis [22]. Hence, interest has risen considerably in this and associated fields where it is important to effectively model and analyze multivariate time series data. Therefore to examine the performance of the proposed algorithm for handling a real-world missing data problem, a case study involving Electro-Cardio Gram (ECG) signals was accomplished. An ECG dataset without missing values was obtained from the Physionet website (http://www.physionet.org/physiobank/database/ptbdb). Then two datasets were created by randomly removing data elements. A 10% missing completely at random (MCAR) and a 20% MCAR dataset was created. The initial Physionet dataset included 290 patients with 549 records (aged between 17 and 87, mean 57.2; 209 men, mean age 55.5, and 81 women, mean age 61.6; ages were not recorded for 1 female and 14 male subjects). Each subject is represented by one to five records. There are no subjects numbered 124, 132, 134, or 161. Each record includes 15 simultaneously measured signals: the conventional 12 leads (*i, ii, iii, avr, avl, avf, v*1, *v*2, *v*3, *v*4, *v*5, *v*6) together with three Frank lead ECGs (*vx, vy, vz*). Each signal is digitized at 1000 samples per second, with 16-bit resolution over a range of 16.384 mV. On special request to the contributors of the database, recordings may be available at sampling rates up to 10 KHz. More detailed discussion can be found [29,30]. The diagnostic classes of the patients are divided into nine types; this case study considered a 12-lead ECG signals for two diagnostic classes: myocardial infarction and healthy control for two patients. Two cases of MCAR missing data mechanism with two different percentages 10% and 20% were generated. Table 1 shows the values of the four model order selection criteria. As can be seen each test indicates that the model with lag two has the highest priority. Tables 2 and 3 show the recovering accuracy for the missing data in the heart rate signal using different imputation methods, under two cases of missing data mechanisms, i.e., 10% and 20% MCAR, respectively. In both cases, the proposed method VAR-IM gives better results comparing with the other methods.

7.1. QRS waves

The ventricular depolarization of the heart can be represented by three nodes on the heart electrical wave displayed within ECG signal. These are the Q, R and S nodes, known as the QRS complex. The amplitude of QRS represents the polarization and depolarization of the ventricular. The QRS duration is the required time for the signal to pass through the ventricular myocardium [31]. The normality of QRS complex is measured by its duration (time interval). A normal duration of the QRS complex is between 0.08 and 0.10 s. An intermediate QRS complex has an interval between 0.10 and 0.12 s. While an abnormal QRS time interval is more than 0.12 s. Important QRS properties include rise level (*Lr*), fall level (*Lf*), rise duration (*Tr*), and fall duration (*Tf*). These factors represent the quality of a QRS wave in terms of specifying the ventricular depolarization. The rise and fall levels represent length of edges of R peak on the right and left hand side, respectively, the rise and fall durations are the required time to move from the Q peak to R peak and from R peak to S peak, respectively [32].

Lr = *Amplitude R peak* – *Amplitude Q peak*
Lf = *Amplitude S peak* – *Amplitude R peak*
Tr = *Time point R peak* – *Time point Q peak*
Tf = *Time point R peak* – *Time point Q peak*
Mean Error = *mean (noisy-ECG (QRS locations)* – *((filtered (QRS locations))*

7.2. QRS waves and missing values

The performance of the VAR-IM method is evaluated by comparing the effectiveness of missing data imputation on QRS wave properties, in both cases of missing data (10% and 20% MCAR) and complete data. Furthermore, the efficiency of missing data imputation is considered in the filtering processing. Fig. 2 shows the QRS complex rise level, fall level, rise time and fall time in the case of complete data. In comparison, Figs. 3– 5 show various results with respect to the case of 10% MCAR. The four proposed techniques with VAR-IM methods were applied to solve the missing data problem here. Clearly, most approaches generated obviously different results: for example, the list wise deletion could not achieve any improvement in all features; it gave results similar to the case of missing data. on the other hand, all the other methods gave acceptable results, for some features such as *Lr, Lf, Tr* and *Tf*, but the VAR-IM methods still has the highest priority to be the best methods to recover the missing values, which is similar to the real data especially the QRS peaks locations. Table 4 summarizes the results of the effectiveness of missing data imputation of the four methods for the QRS wave properties.

When the percentage of the missing data increases from 10% to 20%, the proposed VAR-IM method gives the best results among the five methods. Table 5 summarizes the results generated from the recovered data using the five methods, as well as a comparison with that generated from the complete data. As can be noted in both cases of missing data (MCAR 10% and 20%) the MARSS package and EM algorithm have similar results. The reason may be that the MARSS depends mainly on EM algorithm in estimating an MARSS model (Fig. 6–10).

8. Conclusion

It is extremely important to effectively handle multivariate data anomalies that contain missing values. This is especially true for medical data, which could involve great number of critical health diagnostic variables. The proposed VAR-IM method provides improvements to speed and accuracy for imputing missing values of multivariate time series datasets. It outperforms the commonly used methods such as list wise deletion, linear regression imputation, MARSS and EM algorithms. From the results of the case study, the VAR-IM method provides an effective alternative for the imputation of missing values in multivariate time series. While the other proposed traditional and modern methods become less effective with the increase of the proportion of missing data, VAR-IM shows less deterioration in performance with increasing percentages of missing entries. In addition, the VAR-IM method is more robust than the other proposed techniques when applied to the data types discussed in the case study, and performed better on static and noisy data. There are some limitations of the proposed method. Firstly, this study only considered the scenario in which data was missing completely at random (MCAR), that is, the cause of the missing data was independent of both the observed and missing values. A less stringent assumption of missing data mechanism missing at random (MAR) may be more realistic in practice. MAR refers to the case in which missingness is related to the observed values, but not to the missing values themselves. Secondly, the validity of VAR-IM approach requires that the time series should be stationary. Finally, the percentage of missing data has significant impact on most missing data analysis methods, VAR-IM does not have the priority to be used if the percentage of missing data is quite low (say less 10%). Despite these limitations, VAR-IM provides an important alternative to existing methods for handling missing data in multivariate time series. Further extension of the method to include other types of methods will be considered in other future work.

References

[1] J.W. Graham, Missing Data: Analysis and Design, Springer Science & Business Media, 2012.
[2] J. Schafer, Analysis of Incomplete Multivariate Data, Chapman Hall, New York, 1997.
[3] B. Efron, Missing data, imputation, and the bootstrap, J. Am. Stat. Assoc. 89 (426) (1994) 463–475.
[4] M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation, J. Am. stat. Assoc. 82 (398) (1987) 528–540.
[5] D. Rubin, Multiple Imputation for Nonresponse in Surveys, John Willey & Sons, New York, USA, 1987.
[6] C.K. Enders, D.L. Bandalos, The relative performance of full information maximum likelihood estimation for missing data in structural equation models, Struct. Equ. Model. 8 (3) (2001) 430–457.
[7] J.W. Graham, Missing data analysis: making it work in the real world, Ann. Rev. Psychol. 60 (2009) 549–576.
[8] S. Liu, P.C. Molenaar, IVAR: a program for imputing missing data in multivariate time series using vector autoregressive models, Behav. Res. Methods 46 (4) (2014) 1138–1148.
[9] L. Ljung, Prediction error estimation methods, Circ. Syst. Signal Process. 21 (1) (2002) 11–21.
[10] A.J. Isaksson, Identification of ARX-models subject to missing data, IEEE Trans. Autom. Control 38 (5) (1993) 813–819.
[11] E.E. Holmes, E.J. Ward, K. Wills, Marss: multivariate autoregressive state-space models for analyzing time-series data, R J. 4 (1) (2012) 11–19.
[12] E. Holmes, E. Ward, M. Scheuerell, Analysis of multivariate time-series using the marss package, User guide: http://cran.r-project.org/web/packages/MARSS/vignettes/UserGuide.pdf. (2014).
[13] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B Methodol. (1977) 1–38.
[14] R.H. Shumway, D.S. Stoffer, An approach to time series smoothing and forecasting using the EM algorithm, J. Time Ser. Anal. 3 (4) (1982) 253–264.
[15] J.C. Agüero, W. Tang, J.I. Yuz, R. Delgado, G.C. Goodwin, Dual time–frequency domain system identification, Automatica 48 (12) (2012) 3031–3041.
[16] T. Schneider, Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values, J. Clim. 14 (5) (2001) 853–871.
[17] R. Gopaluni, A particle filter approach to identification of nonlinear processes under missing observations, Canad. J. Chem. Eng. 86 (6) (2008) 1081–1092.
[18] R.H. Shumway, D.S. Stoffer, Time series analysis and its applications, Stud. Inf. Control 9 (4) (2000) 375–376.
[19] G.C. Reinsel, Elements of Multivariate Time Series Analysis, Springer Science & Business Media, 2003.
[20] E. Zivot, J. Wang, Vector autoregressive models for multivariate time series (Chapter 11), in: Modeling Financial Time Series with S-PLUS, 2006, pp. 385–429.
[21] W.J. Tsay, Maximum likelihood estimation of stationary multivariate arfima processes, J. Stat. Comput. Simul. 80 (7) (2010) 729–745.
[22] W.L. Wang, Multivariate *t* linear mixed models for irregularly observed multiple repeated measures with missing outcomes, Biometr. J. 55 (4) (2013) 554–571.
[23] H. Lütkepohl, New Introduction to Multiple Time Series Analysis, Springer Science & Business Media, 2005.
[24] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.
[25] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464. https://doi.org/10.1214/aos/1176344136.
[26] E.J. Hannan, B.G. Quinn, The determination of the order of an autoregression, J. R. Stat. Soc. Ser. B Methodol. (1979) 190–195.
[27] H. Akaike, Fitting autoregressive models for prediction, Ann. Inst. Stat. Math. 21 (1) (1969) 243–247.
[28] S. Johansen, Statistical analysis of cointegration vectors, J. Econ. Dyn. Control 12 (2–3) (1988) 231–254.
[29] R. Bousseljot, D. Kreiseler, A. Schnabel, Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet, Biomed. Tech. Biomed. Eng. 40 (s1) (1995) 317–318.
[30] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.
[31] N. Burns, Cardiovascular physiology, 2013. Retrieved from: School of Medicine, Trinity College, Dublin. http://www.medicine.tcd.ie/physiology/assets/docs12_13/lecturenotes/NBurns%20CVS%20lecture.
[32] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. (3) (1985) 230–236.

Faraj Bashir is a Ph.D. candidate at the university of Sheffield, in Automatic Control and Systems Engineering Department, United Kingdom. He received his B.Sc. and M.Sc. degrees from the University of Tripoli, Faculty of Engineering, Automatic Control and Systems Engineering Department, in 2003 and 2008, respectively. His research interests concern application of missing data analysis, generalised linear models, clustering, pattern recognition, regression, Gaussian processes, system identification; polynomial models, VAR modelling, model selection; kernel density estimation and EM algorithm.

Hua-Liang Wei (B.Sc., M.Sc., Ph.D.) is a Senior Lecturer in the Department of Automatic Control and Systems Engineering, University of Sheffield. His recent research interests include system identification; data mining and data analytics; Narmax methodology and its applications; statistical digital signal processing; machine learning; spatio-temporal system modelling; wavelets and neural networks; non-stationary(time varying) process modelling; forecasting of stochastic and dynamical processes; generalized regression analysis; linear and nonlinear optimisation; multidisciplinary applications in medicine and biomedicine, medical informatics, space weather, environmental, economical and financial systems, and social sciences.

Please cite this article as: F. Bashir, H.-L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.03.097