Statistical method for ECG analysis and diagnostic

Hanen Chaouch*, Khaled Ouni and Lotfi Nabli

National School of Engineers of Monastir,
Ibn Aljazzar City,
Monastir 5019, Tunisia
Email: hanenchaochu8@gmail.com
Email: khaled_ouni@yahoo.fr
Email: lotfinabli@yahoo.fr
*Corresponding author

Abstract: In this paper, a statistical method of ECG analysis and diagnostic is proposed. This method is based on three parts: data simplification using multiscaled PCA, faults detection and localisation by introducing classic linear PCA. The studied data is presented as a multivariate matrix. The variables of this matrix are extracted from the ECG waves characteristics: waves amplitudes and segments measurements. The developed approach allows detecting arrhythmias and heart beat troubles. Comparing the results obtained by this approach and the data of the expert, we approve the performance of our study.

Keywords: ECG diagnostic; multiscaled PCA; fault detection; fault isolation; arrhythmia.


Biographical notes: Hanen Chaouch received PhD degree in Electric Engineering from National School of Engineers of Monastir in 2015. In 2009, she obtained her Master’s in Measurement and Instrumentation from INSAT, Tunis. She is currently Assistant in the Superior Institute of Applied Maths and Informatics in Kairouan (ISMAIK). Her research interests include modelling, control, monitoring and command of biologic and industrial non linear Systems.

Khaled Ouni received PhD degree in Electric Engineering from National of Engineers School of Monastir in 2012. In 2007, he obtained his Master’s in Safety and Security of the Industrial Systems from ISSAT, Sousse. He is currently associate professor in the Superior Institute of Applied Sciences and Technology of Gabes (ISSATG). His research interests include modelling, control, monitoring and command of manufactory systems.

Lotfi Nabli obtained Master’s of Sciences and DEA from ENSET, Tunis in 1989 and 1991 respectively. In 2000, he obtained his doctorate degree in Industrial Automation: Automatic and Industrial Computing from the University of Sciences and Technologies of Lille France. He is currently professor of Electrical Engineering at the National School of Engineers of Monastir (ENIM), Tunisia and a Master of conference candidate. His research interests include modelling and identification.
1 Introduction

Recently, several studies have introduced the ECG signal in their work. Indeed, analysis of this signal has two objectives, waves segmentation and diseases classification (Ponrani et al., in press; Chatterjee et al., 2015; Mahajan and Bansal, 2015; Dai et al., 2015; Kumari and Kumar, 2016; Bhattacharjee et al., 2015; Chattopadhyay et al., 2013; Tsau et al., 2011; Sram et al., 2012; Heden et al., 1993; Joseph et al., 1999; Azar, 2011; Bahoura et al., 1997; Clavier and Boucher, 1996; Dokur et al., 1997). This work comes in this context, in which a method of ECG diagnostic method is proposed using the multiscaled principal components analysis (MSPCA). This method enables to combine modelling through PCA of every wavelet coefficient issued of multiscaled decomposition. The data will be reconstructed on the scales of resolution regarding more information than others and in which the principal components are determined. The projection of the re-built data on the linear PCA model is then applied (Nomikos et al., 1994). The PCA method is based on three parts. For the principal components number determining, several techniques have been applied in the literature (Kaiser, 1960; Giancarlo and Chiara, 2002; Xu and Kailath, 1994; Kaiser, 1960 and Giancarlo and Chiara, 2002). Indeed, the choice of method has an impact on the results of the analysis steps of the PCA approach. In this paper, the method of determining the number of principal components made is based on the minimisation criterion error (PVC) (Nabli and Ouni, 2008). This step allows identifying the components that carry important information relative to the other (residues). Only the principal components obtained are introduced later in the approach. In the second part, the faults detection of the studied data is based on the principle of comparing the behaviour of the process to normal (healthy heart). To do this, we introduce two methods, the SPE method (statistical prediction error) and Hotteling $T^2$ statistic (Nabli et al., 2008). Faults are detected using the thresholds of controls for each statistic; each of these limits described the occurrence of a fault. The last step is the faults location, which allows the determination of the defective variable. The used method for fault isolation is based on the calculation of contributions (Zumoffen and Basualdo, 2007). Indeed, variables that have high contributions are considered defective. The set of defected parameters on the ECG, and with the intervention of the expert, allowing giving the existing pathology in the subject matter. This approach to diagnosis of the ECG has given reliable results compared to expert data, which proves its effectiveness for the supervision of the cardiovascular system.

2 Contribution of developed approach

Any abnormality in the ECG signal expresses problems in electrical conduction of infraction and thus contraction. Interpretation of ECG signal falling within the area of the practitioner uses the analysis or modelling of the ECG signals while taking into account the physiology of the myocardium.

This work has begun in the context of surveillance of the cardiovascular system through the diagnostic of the ECG taken from MI/BIH database by integrating automated
and computer tools. The diagnostic of such signal includes ECG waves and segments study using modelling tool. Our contribution allows introducing a new statistic tool, the principal components analysis combined with wavelet transform. This approach has been introduced in industrial process modelling and has shown an efficient result. We propose integrate this method on a biological process “the cardiovascular system”.

3 Wavelet transform and multiscaled analysis

The wavelet transform is a mathematic tool which allows decomposing signal into frequencies by maintaining a space location. The original signal is projected on asset of basic functions that vary in frequency and space. These basic functions adapt to the frequencies of the signal which will be analysed. The wavelet transform allows having a location in time and frequency. The wavelet analysis adapts to a prototype function of wavelets called “mother wavelet”. This wavelet generates by recursive translation and dilatation a set of basic function called “daughter wavelet”. The principal of the mother wavelet is given by the following equation:

\[ \Psi(\tau, s) = \frac{1}{\sqrt{s}} \Psi\left(\frac{t-\tau}{s}\right) \]  

with \( \tau \) and \( s \) are respectively the translation factor and the dilatation factor. We can find infinity of values for these two factor, we can vary it continuously: this is the principle of the continuous wavelet transform (CWT). This transform as defined is redundant because it gives more wavelet coefficient that it is necessary to describe the signal exhaustively. Practically, we are often deal with discrete signal. That is why we are interested to discretise the values of \( \tau \) and \( s \), we are talking here about the discrete wavelet transform (DWT). This transform is easier then the CWT in implementation. In this paper, we introduce the DWT in the multiscaled analysis. The discrete wavelet transform allows translating and dilating the wavelet within discrete values.

The coefficients are dicretised as follow:

\( s = s_0^j \) and \( \tau = k\tau_0 \) with \( s_0 > 1 \) and \( \tau_0 > 1 \) fixed and belong to \( \mathbb{Z} \).

The DWT is given by the following equation:

\[ DWT(\tau_0, s_0) = \frac{1}{\sqrt{s_0^j}} \int_{-\infty}^{\infty} x(t) \Psi\left(\frac{t-k\tau_0}{s_0}\right) dt \]  

with \( s_0^j \) is the scale factor, \( \tau_0 \) is the translation factor, \( k \) and \( j \) are integer. \( \Phi(t) \) and \( \Psi(t) \) are the respectively the scale function and the wavelet function defined as follow:

\[ \Phi(t) = 2^{-j/2} \Phi(2^{-j}t-k) \]  

\[ \Psi(t) = 2^{-j/2} \Psi(2^{-j}t-k) \]  

with \( k, j \in \mathbb{Z} \).
The projection on the scale function gives the scale coefficients or approximation coefficients $a_j$. While, the projection on the wavelet function is known as the wavelet coefficient $d_j$ and called the signal detail passing from one scale to another larger.

$$a(j) = \sum_{k=-\infty}^{\infty} x(k) g(2j-k)$$  \quad (5)

$$a(j) = \sum_{k=-\infty}^{\infty} x(k) h(2j-k)$$  \quad (6)

with $h$ and $g$ are respectively the scale function coefficients (low-pass filter) and the wavelet function coefficients (high-pass filter).

## 4 Linear Principal Components Analysis (LPCA)

The method of fault detection and isolation has a great contribution in biological and industrial signal monitoring. This approach rests on three steps: number of principal component determining, detection and isolation of faults. Determining the number of the principal components $\ell$ is based on the minimisation of the variance of the reconstruction error. This method was proposed by Rinses (2001). A process in its normal operating condition is presented by a data matrix with $n$ measurements and $m$ variables. The following equation allows to calculate the variance of the reconstruction error of the $n^{th}$ component of:

$$x(k) = [x_1, \ldots, x_m]^T \in \mathbb{R}^n$$  \quad (7)

$$\rho_i = \text{var} \{ \zeta_i^T (x(k) - x(k))^T \}$$  \quad (8)

With $\zeta_i$ is the $i^{th}$ column of the identity matrix and $X(k)$ is the measurement vector.

The $i^{th}$ component of $X(k)$ has been reconstructed as follow:

$$x(k)_i = \left[ C_i^T C_i \right]^{-1} \frac{C_i^T C_i}{1-C_i} x(k)$$  \quad (9)

With $C = [C_1^T, C_2^T, \ldots, C_m^T]$, $C_i$ is the $i^{th}$ column of the matrix $C$. The vectors made up by the first $(i-1)$ and the last $(m-i)$ elements are presented by the signs $(+i)$ and $(-i)$.

The determining of the number of principal components is based on the minimisation criterion explained as follow:

$$J(\ell) = \frac{\rho_i}{\text{var} \{ \zeta_i^T x_k \}} = \sum_{i=1}^{m} \frac{\rho_i}{\zeta_i^T \zeta_i}$$  \quad (10)
The second part of the PCA approach is the fault detection. Two methods are introduced at this step, the Hotelling statistic $T^2$ and the SPE method (Squared Prediction Error) which are defined as follow:

$$T^2(k) = \sum_{i=1}^{\ell} \lambda_i$$

$$SPE(k) = \sum_{j=1}^{m} (e_j(k))^2$$

with $e_j(k)$ is the $j^{th}$ residue.

The fault detection is based on these following conditions (Qin, 1998):

$$SPE < \delta_a$$

$$T^2 < \delta_a$$

Where $\delta_a$ and $\delta_a$ are respectively the thresholds of $T^2$ and SPE statistics.

For determining the defected variables we introduce the localisation method of the PCA approach. We use here the principle of calculating contributions of variables at the principal component. In fact, every principal component is expressed as follow:

$$t_i = \rho_i^T x = \sum_{j=1}^{m} \rho_{ij} x_j$$

With $\rho_i$ is the eigen vector corresponding to the value $\lambda_i$.

The total contribution of the variable $x_i$ on the $q$ highest components is given by (Jackson and Mudeholkar, 1998):

$$Cont_i = \sum_{i=1}^{q} Cont_{ij}$$

$$Cont_{ij} = \frac{t_i}{\lambda_i} \rho_{ij} x_j$$

$\rho_{ij}$ is $j^{th}$ component of the proper vector $p_i$.

5 Multiscaled linear PCA

The first part consists of applying a multi-scale analysis based on discrete wavelet transform (DWT); it allows decomposing the data matrix into wavelet coefficients on five levels of multi-scale resolution. Then, we apply the principal component analysis PCA wholes on the coefficients obtained to determine the number of principal components to retain and which provide more information. Data matrix is reconstructed by referring only scales of resolution whose main components are found. In the second part, we introduce the resulting data for detecting and localising faults using PCA by introducing two statistics: SPE and $T^2$. Variables are determined defective by the method of calculating contributions by the same previous statistics. The developed approach based on multiscaled linear PCA is described in Figure 1.
6 ECG analysis using multiscaled PCA

6.1 Pretreatment of data matrix

In this paper, the data matrix is composed from 9 variables and 500 measurements of an ECG signal. These variables are determined from the characteristics of the ECG waves which are the amplitude of P, R, Q, T and S waves. Moreover, the segments QS, QP, RR and ST are the rest of variables. The data matrix is centred and reduced as shown in Figure 2.

Figure 1 Multiscaled PCA algorithm (see online version for colours)

Figure 2 Centred and reduced variables (see online version for colours)
6.2 ECG signal simplification

The first step in the ECG diagnostic approach is the application of the MSPCA to simplify the studied data. The result of the simplification step is shown in Figure 4. The studied data matrix is a multivariate one, composed by 9 variables as shown in Figure 3. These variables are determined from the ECG wave’s amplitudes and segments. The application of the MSPCA allows simplifying the multivariate data. In fact, after decomposing the signal into wavelet coefficient using the discrete wavelet transform (DWT) as shown in Figure 3, we apply the linear PCA approach on each variable to find the number of principal component. These variables are reconstructed using only the retained principal components. The reconstructed signal is called simplified as shown in Figure 4. This signal will be used in the rest of the proposed study.

![Figure 3](image1) Original signal for 9 variables (see online version for colours)

![Figure 4](image2) Simplified signal for 9 variables (see online version for colours)
6.3 Principal component number determining

Table 1 shows the results of principal component number determining using the minimisation criterion error. We note that the used criterion is minimal for 3 variables (6, 7 and 9), so the number of PC is equal to 3.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( \ell = 1 )</th>
<th>( \ell = 2 )</th>
<th>( \ell = 3 )</th>
<th>( \ell = 4 )</th>
<th>( \ell = 5 )</th>
<th>( \ell = 6 )</th>
<th>( \ell = 7 )</th>
<th>( \ell = 8 )</th>
<th>( \ell = 9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>2.17</td>
<td>1.49</td>
<td>1.17</td>
<td>0.98</td>
<td>0.97</td>
<td>0.91</td>
<td>0.66</td>
<td>0.51</td>
<td>0.11</td>
</tr>
</tbody>
</table>

6.4 Fault detection and location

- Case 1: Normal ECG

In the detection part, we introduce the method of squared predictive error SPE to identify faults on the studied data. Figure 5 shows that there is no detected faults using the SPE method. That can be explained by the normality of the studied ECG. In fact, we have to generate some faults on the data matrix for effective assessment of our study. We increase a dozen of measures using the same amplitude; this fault is generated on the variable RA. After detecting defect in Figure 6, the next step is the isolation by the method of calculating contributions. In fact, the variable which has the greatest contributions is the defected one. Figure 7 shows that the variable RA is defected because it has the greatest contributions. We note that this result is consistent with the fault generation hypothesis. In fact, we locate the fault at the same variable that it was generated.

- Case 2: Tachycardia disease

In this case the ECG present a tachycardia which is characterised by an irregular heart beat (RR segment). Figure 8 shows the fault detection using the statistic SPE. Figure 9 shows the fault isolation, we note that the variable RR is defected; this result is efficient with comparison with the expert data.
Statistical method for ECG analysis and diagnostic

Figure 6  Fault detection with $T^2$ method (see online version for colours)

![Figure 6](image1)

Figure 7  Fault localisation with SPE method (see online version for colours)

![Figure 7](image2)

Figure 8  Fault detection (case of ECG with tachycardia) (see online version for colours)

![Figure 8](image3)
7 Discussion and comparative study

Several studies have addressed this problem by introducing different methods. Tsipouras et al. (2007) has established an approach for the automatic generation of a fuzzy expert system (FES). This work has shown more or less interesting results in the classification of cardiac arrhythmias unless they are limited because of the use of fuzzy logic. Chazal et al. (2004) propose a method for automatic processing of the ECG from the database MIT/BIH for pulsation classification. The independent evaluation of the performance of this configuration resulted in a sensitivity of 75.9%, a positive prediction of 38.5%, and false positive rate of 4.7% for arrhythmia detection (ESV, E...). These results appear very limited compared to the expert needs.

Formant et al. (2006) discussed the multi-source learning problem using Inductive Logic Programming (ILP), which characterises cardiac arrhythmias from various heterogeneous data sources such as the different routes of an ECG or the measurement of blood pressure. This approach showed significant efficacy in a small volume of data. At this level, our work does not have limitations in studied data volume thanks to the introduced statistical tools (PCA). In conclusion, the developed diagnostic approach has several advantages in comparison to previous works.

8 Conclusion

In this paper, an ECG signal diagnostic method is proposed. This method is based on the multiscaled principal components analysis which combine the discrete wavelet transform (DWT) and the linear PCA. First, the multivariate data matrix is decomposed using the DWT on wavelet coefficient at five scales. Then, the linear PCA is applied on each coefficient and the number of principal components is calculated. After that, we reconstruct data using only the retained principal components. The simplified data is then introduced in the linear PCA steps to detect faults. While, we know that the studied ECG is a normal one that is why we should create some faults to precede our study. The
proposed method is efficient in detection and location of faults. In fact, we detect and locate the fault in the same variable that it was created (variable RA). In the second case, the studied ECG presents a tachycardia disease. The application of this approach shows that the variable RR is defected, which means that the heart beat are irregular.

In other words, this method is limited due to fact that the contributions of variables are very close and that other methods are considered to better isolate faults as the reconstruction principle. In summary, we find that we can apply this method based on multiscaled principal component analysis for electrocardiogram signal monitoring. Arrhythmias identification is based by the set of defective variables founded in the ECG. By comparison with other techniques that introduce fuzzy logic and neural street-network, our proposed approach is more robust in terms of running time and results.

References


