Principal Component Analysis Based Method for Reconstruction of Fragments of Corrupted or Lost Signal in Multilead Data Reflecting Electrical Heart Activity and Hemodynamics

R Petrolis, R Simoliuniene, A Krisciukaitis

Kaunas University of Medicine, Kaunas, Lithuania

Abstract

Multilead signals reflecting electrical activity of the heart and hemodynamics give comprehensive but usually redundant representation of the processes. Therefore fragments of transient corruption or loss of the data in one or more leads can be restored substituting them by the signals reconstructed using information carried by the other leads. Principal component analysis used for reduction of dimensionality in representation of the physiological processes. It concentrates essential information represented by multilead signal into few principal components which could form a set of basis functions for optimal representation and reconstruction of the original signals. The idea of the method is to calculate these components from normal fragment of the multilead signal and use them for reconstruction of missing lead in the corrupted fragment of it.

1. Introduction

Multivariate analysis methods, particularly Principal Component Analysis (PCA) or Karhunen Loeve Transform, are successfully used for optimal representation of quasiperiodic biomedical signals as ECG or EEG [1]. The aim of this procedure often is to reduce dimensionality of data representation. The PCA transforms the original data set into a new set of vectors (the principal components) which are uncorrelated and each of them involve information represented by several interrelated variables in the original set. According to the experience of various authors only few first principal components are enough for optimal representation of information carried by, for example, 12 lead ECG recording [2]. We found that first 5 principal components were representing 99.54% of variation in 12 lead ECG recordings containing T-wave alternans episodes [3]. Moreover, diagnostically informative variation in these recordings was represented only by 2 or 3 principal components. It shows significant redundancy of the original representation, so it is highly expected that corrupted or lost signal in one of several ECG leads could be restored with proper accuracy using data from other leads.

Other then ECG heart activity reflecting signals (e.g. plethysmographic or chest impedance signals), registered together also have the same origin – contraction of the heart muscle. However there is not much data published so far about interdependency of their shape changes. Some suggestions could be made from our previous investigations on chest impedance signal shape [4]. Anyway, we can expect such interdependency and it is worth to try to develop this idea.

2. Methods

We used PhysioNet database (http://physionet.org/physiobank/database/) signals for processing. Signal preprocessing was started with detection of fiducial point of each cardiocycle – R-wave. After preliminary detection using filtered derivative of the ECG signal we maximized cross-correlation of the sliding in time R-wave template with the ECG signal. R-wave template was constructed from first 10 cardiocycles of the recording and updated after every processed cardiocycle. Fixed length exerpts of 640 ms, or 80 samples surrounding R-wave were concerned as cardiocycles. The length and position in regard to the fiducial point of excerpts was defined during preliminary tests. In case of pletismography, arterial blood preasure or impedance signals, the excerpt was taken after fiducial point. The excerpts from all leads of one heartbeat were concatenated one by one into one array. Such concatenated arrays of all uncorrupted cardiocycles formed a matrix of samples $\boldsymbol{X}$, which was giving a redundant but comprehensive representation of the shape of cardiocycles from the recording considered for analysis:

$$\mathbf{X} = \begin{matrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,1} & \dots & x_{2,n} \\ \dots & \dots & x_{i,j} & \dots \\ x_{p,1} & x_{p,2} & \dots & x_{p,n} \end{matrix}, \qquad (1)$$

where $x_{i,j}$ is the i^{th} sample of the j^{th} cardiocycle. The calculated principal components from this matrix were

used to perform the optimal representation of all cardiocycles in the recording (including the corrupted or lost fragments in some leads). The calculated principal components are ordered so that the very first of them retain most of the variation present in all the original variables. Thus it is possible to perform a truncated expansion cardiocycle representing vectors by using only the first several principal components. Every vector $\boldsymbol{x}_i$ representing ordinary cardiocycle is then represented by linear combination of the principal components $\boldsymbol{\phi}_k$ multiplied by the coefficients $w_{i,k}$:

$$\mathbf{x}_i = \sum_{k=1}^{p} w_{i,k}\varphi_k \ . \qquad (2)$$

Every coefficient of k-th principal component of j-th cardiocycle $w_{i,k}$ is calculated as following:

$$w_{i,k} = \varphi_k{}^T x_i \ . \qquad (3)$$

So it is done only for the uncorrupted period of the recordings. In corrupted periods the samples of lost data lead were excluded from both: registered data array x_i and principal components ϕ. Shortened arrays used in calculations obviously caused reduction in value of coefficients $w_{i,k}$. Correction of this reduction was performed multiplying the coefficients by some constants. The values of them were calculated using average ratios of the two coefficients calculated for the same cardiocycle using full arrays of data and principal components and the shortened ones from uncorrupted periods of the signals, giving maximal agreement between restored signals. The agreement was estimated by the sum of the squares of the residuals obtained sample by sample subtracting these two signals.

Minimal yet sufficient number of principal components to be used for representation of every analyzed recording we determined according to our experience described in details in [5]. For determination of this number we used cross-validation criterion based on the parameter called PRESS (**PRE**diction **S**um of **S**quares) proposed by Wold [6]:

$$PRESS(m) = \sum_{i=1}^{n}\sum_{j=1}^{p}({}_m\hat{x}_{ij} - x_{ij})^2 \ , \ (4)$$

where ${}_m\hat{x}_{ij}$ is the estimate of the original data set based not on all but the first m basis functions, x_{ij} - the original data set.

3. Results

The example of extracted cardiocycles from the recording containing period of corrupted data is presented on fig. 1. First ten principle components calculated from concatenated arrays of uncorrupted cardiocycles of the recording are presented on fig.2.

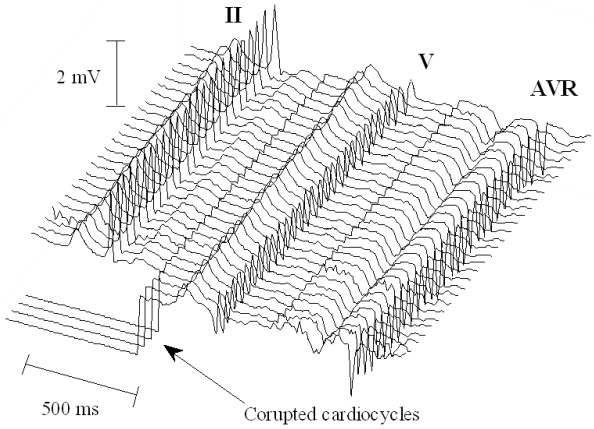

Figure 1. Extracted cardiocycles from the recordings.

Percentage of the variation of uncorrupted cardiocycles in this recording represented by principal components is presented in fig.3 together with Wold's cross-validatory estimation criteria PR used to determine the minimal yet sufficient number of components. As one can see, first 4 principle components were representing 99.96 % of variation, however according to our experience in determining minimal, yet sufficient, number of principal components we decided to use first 2 principal components (the curve of Wold's criteria has characteristic discontinuity at the 2th principal component). About the same situation we observed in the majority of the recordings.

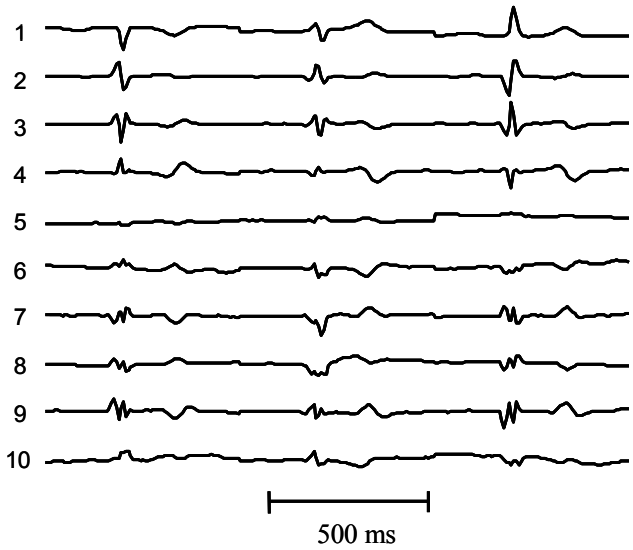

Figure 2. First ten principle components of uncorrupted cardiocycles.

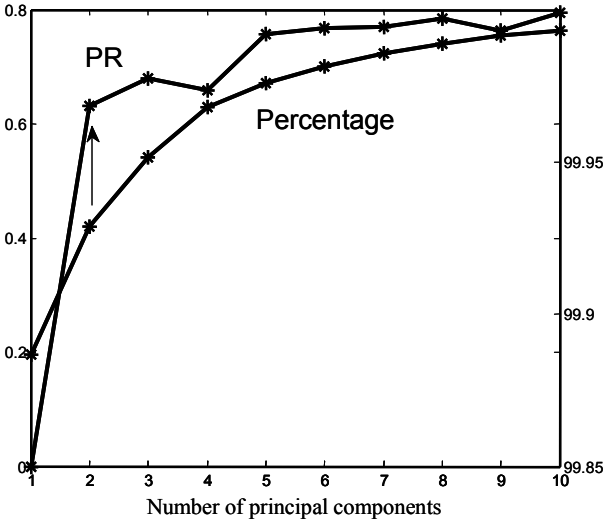

Figure 3. Variation of principal components together with Wold's cross-validatory estimation.

The corrupted period of the recording presented on fig.1 contained zeros in II lead of ECG. Example of the lost samples of one cardiocycle in lead II is presented on fig.4, by dashed line. Reconstructed samples of this lead are shown by solid line.

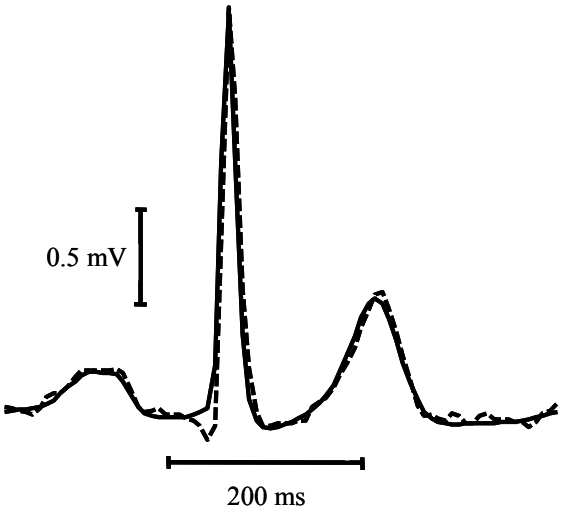

Figure 4. Reconstructed and control cardiocycle data.

The average ratio between coefficients of first 2 principal components calculated using full concatenated arrays of cardiocycles and truncated ones excluding corrupted leads was 1.3798 and varied between 0.7 and 1.7 in all processed recordings.

4. Discussion

Number of minimal yet sufficent number of principal components for representation of concatenated arrays of cardiocycles in all tested recordings was significantly less then number of leads in the recordings. That fact confirmed our idea that all informtive variation in the signals is redundantly represented by multiple leads. Therefore it should be possible to restore lost data in one of the leads using information carried by the others. In majority of tested recordings the clinical sitution was stable and only few principal components were needed to represent all informative variation. Tests on recordings containing some trends in clinical situation could reveal the boundaries of usage of the method.

Acknowledgements

The work is supported by Research Council of Lithuania (Grant: MIP-68/2010).

References

[1] Sörnmo L., Laguna P. Bioelectrical Signal Processing in Cardiac and Neurological Applications. Academic Press; 1 edition (June 15, 2005) ISBN-10: 0124375529.

[2] Castells F., Laguna P., Sornmo L., Bollmann A., Roig J.M. Principal Component Analysis in ECG Signal Processing. EURASIP Journal on Advances in Signal Processing, Volume 2007, Article ID 74580, 21 pages doi:10.1155/2007/74580

[3] Simoliuniene R, Krisciukaitis A, Macas A, Baksyte G, Saferis V, Zaliunas R. Principal Component Analysis Based Method for Detection and Evaluation of ECG T-Wave Alternans. Computers in Cardiology 2008;35:757−760.

[4] Tamošiūnas M, Macas A, Bakšytė G, Kriščiukaitis A, Bražd̄žionytė J. Monitoring of cardiac output by means of chest impedance signal morphology analysis. Proceedings of the 6th Nordic Conference on eHealth & Telemedicine "From tools to services" : NCeHT2006, 2006 Helsinki, Finland / Editor: Persephone Doupi [et al]. Helsinki, 2006. p. 257-258.

[5] Kriščiukaitis A, Tamošiūnas M, Jakuška P, Veteikis R, Lekas R, Šaferis V, Benetis R. Evaluation of ischemic injury of the cardiac tissue by using the principal component analysis of an epicardial electrogram. Computer methods and programs in biomedicine. 2006, vol. 82, no. 2. p. 121-129.

[6] Wold S., Cross-validatory estimation of the number of components in factor and pricipal component models. Technometrics, (1978) 20 pp. 397-405.

Address for correspondence.

Robertas Petrolis
Institute for Biomedical Research,
Kaunas University of Medicine
Eiveniu str. 4, LT50009 Kaunas, Lithuania.
E-mail address robertas.petrolis@bmti.kmu.lt