Principal Component Regression Analysis on Automatic Sleep Apnea Detection from ECG Data

Sidik Mulyono, Sani M. Isa, Mohamad Ivan Fanany, Wisnu Jatmiko, and T. Basaruddin

Faculty of Computer Science, Indonesia University

Email: sidik@bppt.go.id, sani@ui.ac.id, {ivan, chan}@cs.ui.ac.id

Abstract—This paper presents regression model to detect apnea sleep disorder by using principal component regression (PCR). We try to model linear correlations between 11 input features, which are statistical values extracted from heart beat intervals in ECG data records, and the AHI (apnea hypoapnea index) that categorize a patient into 3 levels (heavy apnea, middle apnea, and normal). The result shows that the RMSE and correlation value R approximately reach 16,336 and 79.5% respectively. This shows that based on linearity correlation assumption, we can have highly confident prediction for AHI > 30 (heavy apnea). Whereas the prediction of lower index features correspond to middle apnea and healthy people are found as less predictable. We analyze the presence of multicollinearities between features using PCR and found that using only 10 best features would sufficient for the apnea detection.

I. INTRODUCTION

The problem of predicting a real-valued or continuous target feature has been attracting a great deal of interests in machine learning and data mining communities. Multiple linear regression (MLR) is a method to model the linear relationship between a dependent (output) variable and one or more independent (input) variables. After we obtained the model or parameters, which is known as calibration process, we can predict the value of the dependent variable for any values of the independent variable, which is known as prediction or estimation process. MLR is based on least squares, i.e., the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized. Even though its prediction accuracy is not accurate when underlying relationships between input and output variables are nonlinear, MLR is probably the most widely used prediction method due to its simplicity where the result is easy to understand.

In the process of evaluating the relation between dependent and independent variables, especially where the observed independent variables are actually some statistical variables (mean, median, modus, standard deviation, or time series analysis results) obtained from one sensor such as ECG, the problem often encountered is that of multicollinearity, or high correlation among the independent variables in a regression equation. Such multicollinearity makes the regression coefficient incapable of correctly identifying the true contributors to the target or the dependent variable. PCR combines linear regression and principal component analysis. It establishes the relationship between the dependent variable and the selected principal components of the independent variables, which allows the transformation of a set of correlated variables into an equal number of uncorrelated variables.

In this paper, we perform principal regression analysis on the relation between statistical features of heart beat intervals as independent variables to the apnea hypoapnea index (AHI) as dependent variable. The AHI is categorize each ECG records obtained minute by minute during sleep into 3 classes (heavy apnea, middle apnea, or normal). Using PCR we can detect multicollinear variables and select only features which truly contribute linearly to the AHI value prediction.

II. OBSTRUCTIVE SLEEP APNEA (OSA)

Many researchers have done the detection of OSA based on only ECG (Electro Cardiogram) signal relies on the abnormalities in the heart rate variability, which is related to the periodic cycles of breathing cessation and restoration. [1] analyzed ECGs features related to 5 minutes duration to estimate auto-correlation function (ACF), partial auto-correlation function (PACF), and power spectral density (PSD) between normal sleep and OSA. [2] presented the method to diagnose OSA by employing the Hilbert transformation of the sinus inter-beat interval time series from single-lead ECG signal. [3] introduced the automatic processing of the ECG to detect OSA, and used classifier which is based on linear and quadratic discriminants.

Terminology of apnea takes from Greek word, which means “without breath” [1]. Sleep apnea affects the way breath when sleeping. In untreated sleep apnea,
breathing is briefly interrupted or becomes very shallow during sleep. These breathing pauses typically last between 10 to 20 seconds and can occur up to hundreds of times a night. As airflow stops during a sleep apnea episode, the oxygen level in blood drops. And brain responds by briefly disturbing sleep enough to kickstart breathing—which often resumes with a gasp or a choking sound.

Apnea consists of 3 types, namely: Obstructive sleep apnea, Central sleep apnea, Complex sleep apnea. Of these, Obstructive sleep apnea (OSA), the periodic cessation of breathing during sleep due to intermittent airway obstruction, is most common, and occurs in approximately 2 percent of women and 4 percent of men over the age of 35. It is also a frequently undiagnosed condition affecting million of people in the worldwide, and is associated with increased morbidity and mortality.

The best way to diagnose OSA is by using polysomnographic, which requires overnight monitoring of the patient in a specially equipped sleep laboratory. Because of the expense and inconvenience, and time consuming procedure, an effective and low cost, and simple technique for detection of OSA is highly desirable.

The AHI is the parameter recommended by the American Academy of Sleep Medicine (AASM Task Force, 1999). AHI was calculated based on the frequency of occurrence of sleep disordered respiration (apnea or hypoapnea) within one hour. AHI values> 15 are categorized as heavy sleep apnea, AHI value range 5-15 included in the category of borderline or mild apnea, while AHI <5 categorized as normal or control [5].

III. DATABASE AND FEATURES SET

A. Database

The database of ECG signals is taken from PhysioNet. The database was used in Computers in Cardiology Conference Challenge. It consists of 70 recordings, containing a single ECG signal digitized at 100 Hz with 12-bit resolution, continuously for approximately 8 hours (individual recordings vary in length from less than 7 hours to nearly 10 hours). Each recording includes a set of reference annotations, one for each minute of the recording, indicate the presence or absence of apnea during that minute. These reference annotations were made by human experts on the basis of simultaneously recorded respiration signals.

The subjects of these recordings are men and women between 27 and 63 years of age, with weights between 53 and 135 kg (BMI between 20.3 and 42.1); AHI ranges from 0 to 93.5. For classification purposes we divide 70 record into two groups, training data (A01-A35) and testing data (X01-X35), each consist of 35 records with 20 records with AHI> 15 (Category A: patients with sleep apnea), 5 records with 5 ≤ AHI ≤ 15 (category B: patients with borderline / mild apnea), and 10 records with AHI <5 (category N: normal).

B. Feature Sets

In this study we analyze the collinearity of the following features:

- mean and standard deviation RR-interval
- the NN50 measure (variant 1), defined as the number of pairs of adjacent RR-intervals where the 1st RR-interval exceeds the second by more than 50 ms
- the NN50 measure (variant 2), defined as the number of pairs of adjacent RR-intervals where the second RR-interval exceeds the 1st by more than 50 ms
- two pNN50 measures, defined as each NN50 measure divided by the total number of RR-intervals
- the SDSD measures, defined as the standard deviation of the differences between adjacent RR-intervals
- the RMSSD measure, defined as the square root of the mean of the squares of differences between adjacent RR-intervals
- median of RR-intervals
- inter-quartile range, defined as difference between 75th and 25th percentiles of the RR-interval value distribution
- mean absolute deviation values, defined as mean of absolute values obtained by the subtraction of the mean RR-interval values from all the RR-interval values in an epoch.

IV. REGRESSION MODEL

The purpose of regression models, also known as explanatory models, is to identify a functional relationship between the target variable and a subset of the remaining attributes contained in a dataset. Thus, their goal is twofold. On one hand, regression models serve to highlight and interpret the dependency of the target variable on the other variables. On the other hand, they are used to predict the future value of the target attribute, based upon the functional relationship identified and the future value of the explanatory attributes. Therefore, the development of a regression model allows one to acquire a deeper understanding of the phenomenon analyzed and to evaluate the effects determined on the target variable by different combinations of values assigned to the remaining attributes. This subject is of great interest, particularly for analyzing those attributes that are control levers available to effective decision.
Suppose we are given a dataset \( D \) composed of \( m \) observations and \( n+1 \) attributes, among which we distinguish a target variable and \( n \) other variables that may play an explanatory role with respect to the target. The target attribute is also called the dependent variable, response or output, while the explanatory variables are also termed independent variables or predictors. The independent variables of each observation may be represented by a vector \( x_i \), \( i \in M \), in the \( n \)-dimensional space \( \mathbb{R}^n \), while the target attribute is denoted by \( y_i \). For the sake of conciseness, we will write the \( m \) vectors of observations as a matrix \( X \) having dimension \( m \times n \), and the corresponding \( m \)-dimensional vector associated with the target variable as \( Y = [y_1, y_2, \ldots, y_m]^T \). Finally, let \( Y \) be the random variable that represents the target attribute and \( X_{ij}, j \in \mathbb{Z} \), the random variables associated with the explanatory attributes.

Regression models conjecture the existence of a function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) that expresses the relationship between the dependent variable \( Y \) and the \( n \) explanatory variables \( X_i \):

\[
Y = f(X_1, X_2, \ldots, X_n).
\]

In general, the process of identifying the function \( f \), called a hypothesis, can be divided into two sequential phases. First, a choice is made of an adequate class \( F \) of hypotheses which must fulfill two conflicting requirements clearly and rigorously defined within statistical learning theory. On the one hand, the class must be broad enough to allow the identification of an accurate relationship between the target and the independent attributes, in order to guarantee small errors in explaining past data. On the other hand, it must be narrow enough to guarantee good generalization capability in predicting the target variable of new future observations. Considering these conflicting needs, the most popular classes of hypotheses consist of simple and parametric functional relationships of linear, quadratic, logarithmic and exponential nature. In the second phase, once the class of hypotheses \( F \) has been established, the value of the parameters defining the specific function \( f \) within the class \( F \) is determined through the solution of an optimization problem appropriately formulated.

Linear regression models represent the most widely known family of regression models and are based on a class of hypotheses consisting of linear functions. A consequence, the function relation of (1) reduces to

\[
Y = w^T X + b
\]

For multiple linear regression models, the \( n+1 \) parameters \( w \) and \( b \) that identify the hyperplane expressing the linear relationship in the space \( \mathbb{R}^n \) can be derived by the least squares principle, through the minimization of the sum of squared errors. If \( e = [e_1, e_2, \ldots, e_m]^T \) denotes the vector of the residuals associated with the \( m \) pairs of observations \( x_i, y_i \), the \( n \) equalities

\[
y_i = w^T x_i + e_i, \quad i = 1, 2, \ldots, m.
\]

must hold. Assume that the matrix \( X \) associated with the original dataset has been modified by placing to the left an \( m \)-dimensional column vector with all components equal to 1, and denote by \( w = [b, w_1, w_2, \ldots, w_m]^T \) the vector of the slope regression coefficients, extended in turn to the left with the intercept \( b \). We can therefore reformulate equalities (3) in matrix notation as

\[
y = X w^T e.
\]

Hence, the sum of squared errors can be expressed as

\[
SSE = \sum_{i=1}^{m} e_i^2 = \| e \|^2 = \sum_{i=1}^{m} (y_i - w^T x_i)^2 = \sum_{i=1}^{m} (y_i - X w^T y_i - X w)^2
\]

the SSE is a convex quadratic function possessing a unique minimum point that can be calculated analytically by setting equal to zero the partial derivatives evaluated with respect to the components of the vector \( w \):

\[
\frac{\partial SSE}{\partial w} = -2 X^T y + 2 X^T X w = 0
\]

We obtain the normal equation, expressed in matrix form

\[
X^T X w = X^T y
\]

where the solution is uniquely determines the value of the \( n+1 \) regression coefficients

\[
w = (X^T X)^{-1} X^T y
\]

provided that the matrix \( (X^T X)^{-1} \) is invertible.
easily derive the values of response variable \( Y \) predicted by the model, also termed fitted values, as

\[
y = Xw = [X'X]^{-1}X'y = Hy.
\]  

(9)

where \( H = [X'X]^{-1}X' \) is called the hat matrix, which has the properties of being symmetric and idempotent. The matrix \( H \) allows the residuals to be expressed as

\[
e = y - \hat{y} = I - Hy.
\]  

(10)

In most applications, the development of a multiple regression model is a complex matter, involving the identification of the most appropriate predictors to be actually included into the model, starting from the whole set of available explanatory variables.

Two important problems often plague the linear regression analysis: (1) Linear assumption when the actual relationship between dependent and independent variables is non-linear; (2) Multicollinearity between the explanatory variables which are supposed as independent between one another. Principal Components Analysis (PCA) is one solution to avoid the multicollinearity problem. Basically, this analysis is intended to simplify the observed variables or features in a way to shrink (reduce) the dimension or the number of 'important' features. This is done by eliminating the correlation between independent variables through transformation of the independent variables into a set of new variables that do not correlate at all which is commonly referred to as the main components (principal components). After a few principal components are obtained, then these components become a set of new independent variable that will be regressed or analyzed in predicting the value of the dependent variable \( Y \) using regression analysis [7].

There are several ways of finding the principal components of \( A = X'X \) matrix. One possibility is to apply the SVD method to \( A \), writing the reduced form of SVD as follows

\[
A = UDP^T
\]  

(11)

\[
T = UD
\]  

(12)

where:

\( A \) : Centered \( n \times k \) data matrix

\( U \) : Unitary matrix \( n \times n \)

\( D \) : Diagonal matrix correspond to singular values \( \tilde{A} \), \( n \times m \)

\( T \) : Score matrix

\( P \) : Loading matrix consists of columns of normalized eigen vectors \( m \times k \)

The basic idea in Principal Components Regression (PCR) is that after choosing a suitable value for \( g \), the important features of \( X \) have been retained by \( T_g \) [6]. Then perform the multiple linear regression (MLR) with \( T_g \) in place of \( X \) for an \( n \times m \) calibration data matrix \( Y \)

\[
Y = T_gC\varepsilon.
\]  

(13)

Where in this paper \( Y \) is related to the vector of apnea hypopnea Index (AHI). The least squares method then gives coefficient of regression \( C \) as follow

\[
C = (T_g'T_g)^{-1}T_g'Y
\]  

(14)

So in the development of prediction models using principal component regression, the most decisive factors are the score matrix \( T_g \) and coefficient of regression \( C \), while loading matrix \( P_g \) in fact no significant effect on models. But loading vector that is formed to present an anomaly have a strong influence on observed features.

Consider a new sample \( z \) and a predicted value \( y' \) (both uncentered), and let \( X \) and \( Y \) be the calibration sample averages. Then the prediction takes the form

\[
y' = y'z = X'P_gC
\]  

(15)

The matrix \( P_gC \) is called the regression matrix. To determine how many eigen values to be used in the model, then we need to observe all eigen values and calculate the minimum of RMSE. Then we select only the highest values of eigen values that cause little effect on the RMSE.

V. RESULTS AND DISCUSSION

Figure 1 shows the best eight of the training set validated using 6,545 random subsampling or bootstrapping cross validation, to find the smallest RMSE value by simulating different number of principal components (PC). It is shown in the figure that the lowest RMSE value was reached if we use 10 of 11 PCs (with one value reached approximately 99.00%). Consequently, we use 10 of PCs to develop prediction models for AHI using PCR.

Figure 2 shows the calibration model using PCR to search the lowest RMSE, and the highest correlation value R. From this model calibration we perform a regression of coefficient and loading matrix associated with 10 number of PC, that is shown in Table 1. Furthermore, the best loading matrix and coefficients of PCR is used to validate the model for all available samples. We use the loading matrix to transform the
sample data matrix into a score matrix without having to employ SVD matrix factorization.

Figure 3 shows loading vectors associated with 3 PCs, and illustrates the anomaly of some features (NN50v1, NN50v2, and pNN50v), which can be interpreted that these features have a significant role in determining the value of AHI.

Figure 4 shows that when using all samples, the RMSE = 16.336 and R = 0.795. This figure shows that the model developed using PCR (with 10 PC) can effectively predict with good value those features which belongs to severa apnea category (AHI > 30). Whereas, the category of moderate (AHI = 15-30) and mild (AHI <5) are relatively unpredictable, because this model still has a RMSE value or a tolerance of +/- 16.336.

VI. CONCLUSION

This paper presents principal component regression (PCR) analysis on automatic apnea detection from ECG signals. The PCR analyzes the co-linearity and correlation among features. Such analysis is very beneficial for a better understanding of different sources of variation. It is important to find principal components that contain most of the information. Unfortunately, the extracted principal components may be highly sensitive to anomalous observations. Therefore, data reduction based on classical PCR becomes unreliable if outliers are present in the data. In the future, we plan to increase the prediction accuracy for middle and healthy persons by including more data from those categories. We also plan to extend this work into a robust PCR to obtain principal components that are not influenced much by outliers.

REFERENCES

The best loading matrix and coefficients of PCR associated with 10 of PC number

<table>
<thead>
<tr>
<th>PC</th>
<th>Loading Value</th>
<th>Validation Model Using All Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td><img src="image" alt="Fig 3. Loading value" /></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td><img src="image" alt="Fig 4. Validation model using all samples" /></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1**

<table>
<thead>
<tr>
<th>PC Number</th>
<th>Loading Value</th>
<th>Validation Model Using All Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0139</td>
<td><img src="image" alt="Fig 3. Loading value" /></td>
</tr>
<tr>
<td>2</td>
<td>0.08634</td>
<td><img src="image" alt="Fig 4. Validation model using all samples" /></td>
</tr>
<tr>
<td>3</td>
<td>-0.74237</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.62763</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.09007</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.09202</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.08303</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.08347</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.01757</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.09011</td>
<td></td>
</tr>
</tbody>
</table>

- RMSE = 17.650, R² = 0.836
- RMSE = 18.036, R² = 0.625
- RMSE = 16.550, R² = 0.709
- RMSE = 16.474, R² = 0.782
- RMSE = 16.336, R² = 0.795
- RMSE = 16.336, R² = 0.795
- RMSE = 16.336, R² = 0.795
- RMSE = 16.336, R² = 0.795
- RMSE = 16.336, R² = 0.795
- RMSE = 16.336, R² = 0.795