A Group Lasso Based Method for Automatic Physiological Rhythm Analysis

Rebeca Goya-Esteban[1], Óscar Barquero–Pérez[1], Carlos Figuera-Pozuelo [1], Arcadi García-Alberola[3], José Luis Rojo-Álvarez[1,2]

[1] Rey Juan Carlos University, Fuenlabrada, Madrid, Spain
[2] Center for Computational Simulation, Polytechnic University of Madrid, Spain
[3] Virgen de la Arrixaca University Hospital, Murcia, Spain

Abstract

Physiological rhythms arise from nonlinear interactions between biological mechanisms and environmental conditions. A possible approach to study these dynamics is by means of simplified mathematical models. An essential aspect of these models is how to determine the statistical significance of the rhythms present in a temporal series.

The aim of this work is to propose an automatic rhythm analysis method based on lasso or l_1-regularized linear regression, with physiological rhythm components as features. These models have sparse solutions, allowing to identify relevant rhythms. Since the sine and cosine components of a given period constitute a natural group structure, we used a group lasso model. A cross-validation scheme preserving the temporal structure of the signal allowed to select the regularization parameter. Synthetic signals were used to test the method, combining different sinusoidal rhythm components plus gaussian noise. The method was also applied to study the rhythms in heart rate signals (HR).

The method correctly detected 98% of rhythm patterns on the synthetic data. The method was also able to extract significant cardiac rhythms in HR signals. Since lasso is the closest convex relaxation of the best feature subset selection problem, the proposed method is able to optimally identify the rhythms present physiological signals.

1. Introduction

As many studies have shown most biological variables vary greatly along several time scales in health and disease [1]. For example, heart rate (HR) shows oscillations with different periods that have been widely studied [1, 2]. These rhythms are of clinical interest since several studies indicate dynamics with differences between normal individuals and patients. Some studies have shown blunted or altered circadian rhythms of different physiological variables [3–5]. A suitable hypothesis is that physiological mechanisms in healthy subjects are more adaptable to environmental changes than those in pathological subjects, and maybe with a progressive deterioration of this adaptability related to the severity of the pathological condition. A possible approach to study these dynamics is by means of simplified mathematical models of physiological systems [6]. An essential aspect of these models is how to determine the statistical significance of the rhythms present in a temporal series.

In [7] we presented a lasso path approach to analyze the order of activation of the rhythms, representing the importance of each rhythm, in HR signals. In the present work, we further develop the approach proposing an automatic rhythm analysis method based on lasso or l_1-regularized linear regression, with physiological rhythm components as features. These models have sparse solutions, i.e. many estimated weights are zero, allowing to identify relevant rhythms. Since the sine and cosine components of a given period constitute a natural group structure, we propose to use a group lasso model [8]. In order to select the regularization parameter a suitable cross-validation scheme, preserving the temporal structure of the signals, was implemented. Synthetic signals were used to test the method, combining different sinusoidal rhythm components plus gaussian noise. The method was also applied to study the rhythms present in 7-day heart rate signals.

The structure of the paper is as follows. Section 2 describes the the methodological approach. Section 3 depicts the results. Finally, the conclusions are outlined in Section 4.

2. Methods

Extensive sets of time series collected over several decades show that nearly all biological variables show some degree of more or less periodic behaviour. In many cases, it is useful to look upon a measurement series as consisting on a deterministic part, which may have both rhythmic and arrhythmic systematic components, and a noise part [1].

 ISSN: 2325-887X DOI:10.22489/CinC.2017.198-213

We propose an automatic rhythm analysis method based on lasso or l_1-regularized linear regression, with physiological rhythm components as features. Since the sine and cosine components of a given period constitute a natural group structure, we used a group lasso model. A cross-validation scheme preserving the temporal structure of the signal allowed to select the regularization parameter.

2.1. Rhythm Analysis with Lasso and Group Lasso

Joint characterization of a set of rhythms can be performed by a multiple components model [1],

$$\begin{aligned} y_n &= M + \sum_i A_i \cos(2\pi f_i t_n + \phi_i) + e_n \quad (1) \\ n &= 1, ..., N \end{aligned}$$

where M denotes the rhythm-adjusted mean or MESOR (midline estimating statistic of rhythm), f_i, A_i and ϕ_i are the frequency, the amplitude and the acrophase (i.e., the lag from a defined reference time point to the crest time in the cosine curve fitted to the data) corresponding to each considered rhythm, and N the signal length. The random variable e_n corresponds to the difference between the observed sample y_n and the value provided by the estimated regression model $\hat{y}_n$. The least squares (LS) method can be applied to determine the regression parameters. However, typically all of the least-squares estimates from Eq. 1 will be nonzero. This complicates the interpretation of the final model.

Instead we can constrain, or regularize the estimation process. We propose to use l_1-norm regularization which has the effect of forcing some of the coefficient estimates to be exactly zero, yielding to sparse models [9].

In order to use this approach, we need to reformulate the rhythm model as a linear one, rewriting Eq. 1 as

$$y_n = M + \sum_i \alpha_i \cos(2\pi f_i t_n) + \beta_i \sin(2\pi f_i t_n) + e_n, \quad (2)$$

where $\alpha_i = A_i \cos(\phi_i)$ and $\beta_i = -A_i \sin(\phi_i)$. Therefore, the sines and cosines of frequencies f_i are the characteristics (features) of the model. Collecting all the coefficients in a vector of weights $\boldsymbol{w} = [M, \alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k]$, where k is the number of rhythm components, and collecting all the characteristics and the MESOR in a matrix X, the rhythm model can be compactly written as

$$\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{e} \quad (3)$$

Weights of the model, $\boldsymbol{w}$, can be estimated using LS including a regularization term

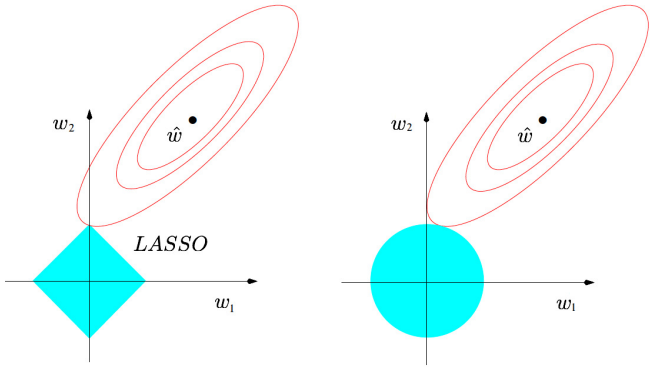

Figure 1. Comparison of estimation weights constraints between lasso (left) and regularized regression (right). Adapted from [10].

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{y} - X\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 \quad (4)$$

where $\|\boldsymbol{w}\|_1 = \sum_{p=1}^{2k+1} |w_p|$ is the l_1 norm of w, and λ is a user specified parameter [10].

The nature of lasso constraint allows to control the number of weights actives ($w_p \neq 0$), so that, making λ sufficiently large will cause some of the weights to be exactly zero. This does not hold for l_2 or l_q with $q > 1$ [8], see Figure 1. Accordingly, it is possible to use lasso models to find the most important variables (features) of the signals in the sense of mean squared error (MSE) [11].

There are regression problems in which the characteristics have a natural group structure, in the present case the sine and cosine components of a given period (rhythm) constitute a natural group structure. In such cases it is desirable to have all coefficients within a group become nonzero (or zero) jointly [8].

Consider a linear regression model involving J groups of characteristics, where for $j = 1, ..., J$, the vector Z_j represents the characteristics in group j, and θ_j represents the set of regression coefficients for group j. Collecting all the groups in a matrix Z, the rhythm model can be compactly written as

$$\boldsymbol{y} = Z\boldsymbol{\theta} + \boldsymbol{e} \quad (5)$$

The group lasso solves the convex problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\boldsymbol{y} - Z\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2 \quad (6)$$

where $\|\boldsymbol{\theta}\|_2 = \sum_{j=1}^{J} \|\theta_j\|_2$ is the l_2 norm of θ. It is important that this criterion involves the sum of the ordinary l_2-norms, as opposed to the squared l_2-norms. In this way, it amounts to imposing a block l_1/l_2 constraint on the overall collection of coefficients. The effect of this group

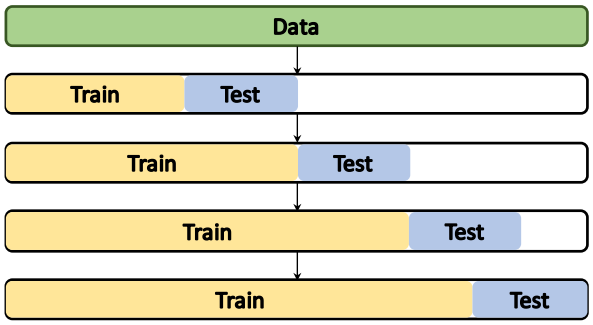

Figure 2. Representation of the cross validation for time series data.

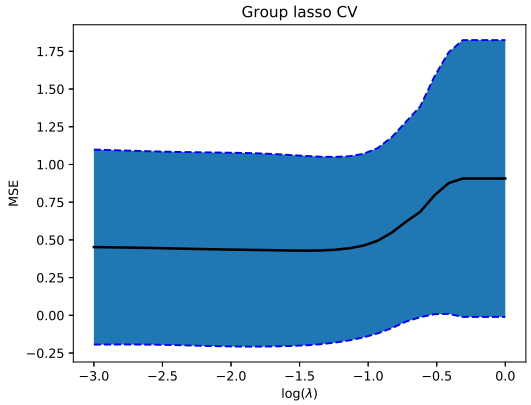

Figure 3. MSE provided by the temporal cross validation as a function of the $log(\lambda)$ values for the HR signal in Fig. 5. Mean (solid black line) and standard deviation (blue dashed lines) of the set of folds.

penalty is to select all the coefficients, of a group of characteristics, to be in or out of the model [8].

2.2. Cross Validation for Time Series Data

In order to select the regularization parameter λ a suitable cross-validation scheme, preserving the temporal structure of the signals, was implemented. Time series data is characterized by the correlation between observations that are near in time. Classical cross-validation techniques assume the samples are independent and identically distributed, therefore applying such techniques would result in unreasonable correlation between training and testing instances yielding poor generalization. A reasonable approach is to evaluate the model for time series data on the future observations. We used a variation of the $k-$fold cross validation technique, where the first k folds are used as train set and the $k+1$ fold as test set, see Fig. 2. Unlike standard cross-validation methods, successive training sets are supersets of those that come before them [12].

2.3. Experiments

In the present study, 8, 12, 24 hour and 7 day period rhythms are considered by the model as characteristics, but any set of rhythms could be considered.

In order to test the proposed methodology we created a set of 200 synthetic signals. The synthetic signals were assembled as combination of different number of sinusoidal rhythm components, and also adding gaussian noise with SNR_s ranging between 3 and 8 dB. The synthetic signals simulated the time evolution of a certain variable during 14 days with a sampling period of 1 hour. The temporal cross validation was implemented with the train and test sets containing the signal samples as follows (see Fig. 2)

- First iteration: Train $\rightarrow$ 1 to 30; test $\rightarrow$ 31.
- Iteration m: Train $\rightarrow$ train + test from iteration $m-1$; test $\rightarrow 31+m-1$. For $m = 2...N-30$.

Figure 4 shows an example of a synthetic signal created as combination of a sinusoidal rhythm of 24 hour period, a sinusoidal rhythm of 7 day period and gaussian noise with 6 dB SNR (blue solid line).

We also analyzed a set of HR signals, obtained from a 7 day Holter database from patients with congestive heart failure, collected in the Arrhythmia Unit of Virgen de la Arrixaca University Hospital (Spain) [2]. The signals were obtained as the mean HR in each 10 minutes window during de 7 days. Figure 5 shows an example of a real HR signal (blue solid line). The temporal cross validation was implemented as for the synthetic signals, but the training set for the first iteration contained 140 samples, which corresponds proximately to the samples in the first 24 hours of the signal.

To select the regularization parameter (see Sec 2.2), after some inspection of a wide range of values for λ, we narrowed the search from 0.001 to 1, testing 30 logarithmically spaced values for each signal. Figure 3 shows the MSE as a function of $log(\lambda)$ values provided by the temporal cross validation applied to a real HR signal.

3. Results

The method correctly detected 98% of rhythm patterns on the synthetic data. Considering a success only when the exact set of rhythms selected by the model matched the set of rhythms in the synthetic signal. Figure 4 shows an example of the resulting rhythm model after applying the proposed method (red dashed line) to the synthetic signal (blue solid line). The method correctly detected the two rhythms present in the noisy signal.

For real HR signals we qualitatively observed that the method was able to extract the underlying cardiac rhythms. Figure 5 shows the resulting rhythm model (red dashed line) for a HR signal (blue solid line), the method detected a 24 hour and a 12 hour period rhythms.

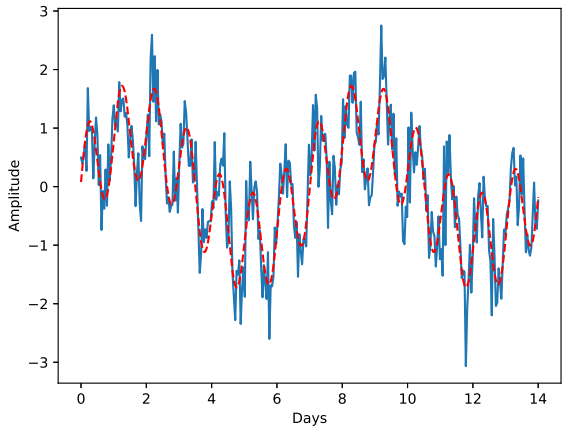

Figure 4. Example of a synthetic signal created as combination of a sinusoidal rhythm of 24 hour period, a sinusoidal rhythm of 7 day period and gaussian noise with 6 dB SNR (blue solid line). The resulting rhythm model obtained by the proposed method (red dashed line)

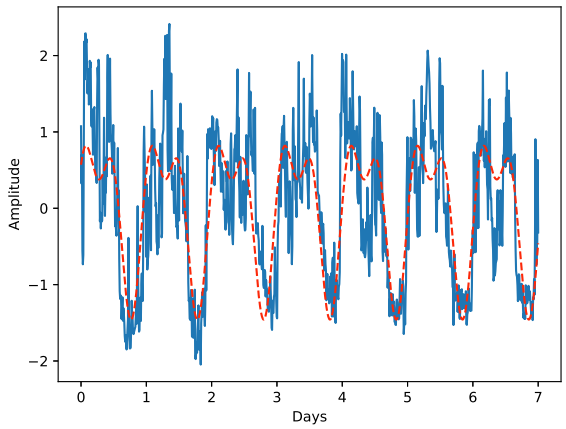

Figure 5. Example of a HR signal obtained from a 7-day Holter recording (blue solid line). The resulting rhythm model obtained by the proposed method (red dashed line)

4. Conclusions

We propose an automatic rhythm analysis method based on lasso or l_1-regularized linear regression. Since lasso is the closest convex relaxation of the best feature subset selection problem, the proposed method is able to optimally identify the rhythms present in physiological signals.

The method can be easily adapted to extract rhythm information from any temporal signal.

Acknowledgements

This work has been partially supported by the Research Projects from the Spanish Ministry of Economy and Competitiveness, TEC2013-48439-C4-1-R, TEC2016-75161-C2-1-R, TEC2013- 46067-R, TEC2016-81900-REDT.

References

[1] Bingham C, Arbogast B, Guillaume GC, Lee J, Halberg F. Inferential statistical methods for estimating and comparing cosinor parameters. Chronobiologia 1982;9(4):397–439.

[2] Goya-Esteban R, Mora-Jiménez I, Rojo-Alvarez JL, Barquero-Pérez O, Pastor-Pérez FJ, Manzano-Fernández S, Pascual-Figal DA, García-Alberola A. Heart rate variability on 7-day holter monitoring using a bootstrap rhythmometric procedure. IEEE Transactions on Biomedical Engineering 2010;57(6):1366–1376.

[3] Guzzetti S, Dassi S, Pecis M, Casat R, Masu AM, Longoni P, Tinelli M, Cerutti S, Pagani M, Malliani A. Altered pattern of circadian neural control of heart period in mild hypertension. Journal of Hypertension 1991;9(9):831–838.

[4] Touitou Y, Bogdan A, Levi F, Benavides M, Auzeby A. Disruption of the circadian patterns of serum cortisol in breast and ovarian cancer patients: relationships with tumour marker antigens. British Journal of Cancer 1996; 74(8):1248.

[5] Burger AJ, Charlamb M, Sherman HB. Circadian patterns of heart rate variability in normals, chronic stable angina and diabetes mellitus. International journal of cardiology 1999;71(1):41–48.

[6] Glass L. Synchronization and rhythmic processes in physiology. Nature 2001;410(8):277–284.

[7] Goya-Esteban R, Barquero-Pérez Ó, Alzueta J, et al. A multicentric study of long-term rhythm patterns in heart rate. In Computing in Cardiology Conference (CinC), 2016. IEEE, 2016; 909–912.

[8] Hastie T, Robert T, Martin W. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015.

[9] James G, Witten D, Hastie T. An Introduction to Statistical Learning: With Applications in R. Springer, 2014.

[10] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 1996;58(1):267–288.

[11] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics 2004;32(2):407–499.

[12] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011; 12:2825–2830.

Address for correspondence:

Rebeca Goya-Esteban
Camino del Molino s/n, Departamental III D202, 28943, Fuenlabrada, Spain
rebeca.goyaesteban@urjc.es