Método de Análisis de Ritmos Fisiológicos Basado en Group Lasso

R. Goya-Esteban[1], O. Barquero-Pérez[1], R. García Carretero[2], A. García-Alberola[3], J.L. Rojo-Álvarez[1,4]

[1] Departamento Teoría de la Señal y Comunicaciones, Universidad Rey Juan Carlos, Fuenlabrada, España, {rebeca.goyaesteban,oscar.barquero, joseluis.rojo}@urjc.es

[2] Servicio de Medicina Interna, Hospital Universitario de Móstoles, España, rgcarretero@salud.madrid.org

[3] Unidad de Arritmias, Hospital Universitario Virgen de la Arrixaca, Murcia, España, arcadi@secardiologia.es

[4] Center for Computacional Simulation, Universidad Politécnica de Madrid, España

Resumen

Los ritmos fisiológicos surgen de interacciones no lineales entre mecanismos biológicos y las condiciones ambientales. Habitualmente, se utilizan modelos matemáticos simplificados para caracterizar la dinámica de estos ritmos. Uno de los aspectos más importantes en estos modelos es determinar la significancia estadística de los ritmos presentes en las series temporales.

El objetivo de este trabajo es proponer un método de análisis ritmométrico automático basado en la regresión lineal con regularización norma l_1 (lasso), utilizando como características del modelo de regresión los ritmos fisiológicos. Este tipo de modelos regularizados con norma l_1 poseen soluciones dispersas, lo que permite identificar los ritmos más relevantes. Un ritmo puro de un determinado periodo está compuesto por la suma de un coseno y un seno, por lo que estos forman una estructura grupal natural. Para respetar esta estructura grupal se propone utilizar Group Lasso. Para la selección del parámetro de regularización proponemos utilizar un esquema de validación cruzada que respeta la estructura temporal de los datos. Para probar el método propuesto se utilizaron un conjunto de señales sintéticas compuestas por un conjunto de ritmos sinusoidales y ruido gaussiano añadido. Así mismo se exploró el funcionamiento del método en un conjunto de señales reales de ritmo cardíaco (RC) y tensión arterial (TA).

El método propuesto detectó correctamente el 98% de los patrones de ritmo en los datos sintéticos. Así mismo, fue capaz de identificar los ritmos significativos de las señales de RC y TA. Dado que Lasso es el método convexo más cercano al problema de selección de subconjuntos de características, el método propuesto permite la identificación óptima de ritmos presentes en las señales fisiológicas.

1. Introducción

Tal y como ha sido demostrado en muchos estudios, la mayoría de los parámetros biológicos varían ampliamente a diferentes escalas temporales, tanto en condiciones de salud como en enfermedad [1]. Uno de los ejemplos más ampliamente estudiados es el ritmo cardíaco (RC) que oscila con diferentes períodos [1, 2]. Este tipo de ritmos en señales fisiológicas son de interés clínico dado que se encuentran en la literatura estudios que identifican dinámicas diferentes entre individuos sanos y pacientes. Algunos estudios han referido alteraciones de los ritmos circadianos de diferentes variables fisiológicas en pacientes [3–5]. Una posible hipótesis es que los mecanismos fisiológicos de los individuos sanos permiten una mejor adaptación a las variaciones ambientales que los equivalentes en pacientes enfermos y, quizá, con un agravamiento progresivo en la capacidad de adaptación que podría estar relacionada con la severidad de la patología.

Un posible enfoque para estudiar este problema es por medio de modelos matemáticos simples de los sistemas fisiológicos [6]. Un aspecto fundamental al desarrollar estos modelos es cómo determinar la significancia estadística de los ritmos presentes en las series temporales. En [7], presentamos un análisis basado en *lasso path* para identificar el orden de activación de los ritmos en señales de RC, donde el orden de activación representa la importancia del ritmo. En el presente trabajo, desarrollamos dicho modelo para proponer un método de análisis automático de ritmos basado en modelos de regresión lineal con regularización de norma l_1 (lasso), con los ritmos fisiológicos como características del modelo. Este tipo de modelos lasso poseen soluciones dispersas, es decir, muchos de los coeficientes del modelo son cero. La dispersión en la solución contribuye a identificar los ritmos relevantes. Dado que las componentes coseno y seno de un ritmo puro constituyen una estructura grupal natural, proponemos utilizar modelos *group lasso* [8]. La selección del parámetro libre de regularización se realiza utilizando un método de validación cruzada adecuado, con el objetivo de preservar la estructura temporal de las señales temporales. El método propuesto se evalúa utilizando un conjunto de señales sintéticas que combinan diferentes ritmos sinusoidales con ruido gaussiano añadido. El método también se prueba en señales reales de RC y tensión arterial (TA).

La estructura del trabajo es como sigue. En la Sección 2 se describen las diferentes partes del método propuesto. En la Sección 3 se detallan los experimentos realizados. En la Sección 4 se muestran los resultados. Finalmente, las conclusiones se incluyen en la Sección 5.

2. Métodos

Existe una amplia evidencia de que casi todas las variables biológicas exhiben un comportamiento más o menos periódico. A menudo es útil considerar las series de medidas compuestas por una parte determinista, que puede contener tanto componentes periódicas como

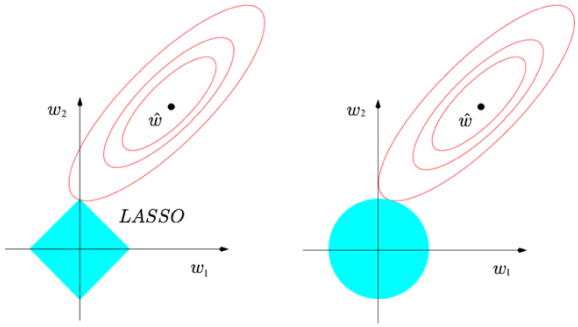

Figura 1 *Comparación de las restricciones para la estimación de los pesos entre lasso, norma* $\boldsymbol{l_1}$*, (izquierda) y ridge regression, norma* $\boldsymbol{l_2}$*, (derecha). El punto* $\hat{\boldsymbol{\omega}}$ *representa la estimación LS de los pesos sin regularización. Las elipses rojas representan las líneas de contorno de la función de coste y las regiones continuas azules son las regiones asociadas a cada una de las restricciones. Adaptado de [10].*

aperiodicas, y una parte aleatoria (considerada ruido) [1]. Proponemos, en este trabajo, un método de análisis ritmométrico basado en modelos de regresión lineal con regularización de norma l_1, cuyas características sean los ritmos fisiológicos. Proponemos un modelo conocido como group lasso, para mantener la estructura grupal de las componentes coseno y seno de los ritmos puros. Asimismo, se propone elegir el parámetro de regularización mediante un esquema de validación cruzada que preserva la estructura temporal de la serie temporal.

2.1. Análisis ritmométrico con lasso y group lasso

La caracterización conjunta de diferentes ritmos se puede realizar mediante el siguiente modelo de componentes múltiples [1],

$$y_n = M + \sum_i A_i \cos(2\pi f_i t_n + \phi_i) + e_n \qquad (1)$$

$$\text{para n} = 1, \dots, \text{N}$$

donde M denota la media del ritmo ajustado o MESOR (midline estimating statistic of rhythm), f_i, A_i y ϕ_i representan la frecuencia, la amplitud y la acrofase (el retardo entre un punto de referencia temporal y el máximo de la señal sinusoidal ajustada a los datos) para cada ritmo considerado y, finalmente, N es la longitud de la señal. La variable aleatoria e_n corresponde a la diferencia entre la muestra real observada (y_n) y el valor estimado ($\hat{y}_n$). La estimación de los parámetros del modelo se puede realizar con el método de mínimos cuadrados (least squares, LS) Sin embargo, típicamente los coeficientes obtenidos con LS serán todos diferentes de cero, lo que complica la interpretación del modelo resultante.

En cambio, podemos restringir el proceso de estimación utilizando regularización de norma L1. Como consecuencia algunos coeficientes serán cero, dando lugar a modelos dispersos [9]. Podemos reescribir la Eq. 1 para expresar el modelo de forma lineal,

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

donde $\alpha_i = A_i \cos(\phi_i)$ y $\beta_i = -A_i \sin(\phi_i)$. Por lo tanto,

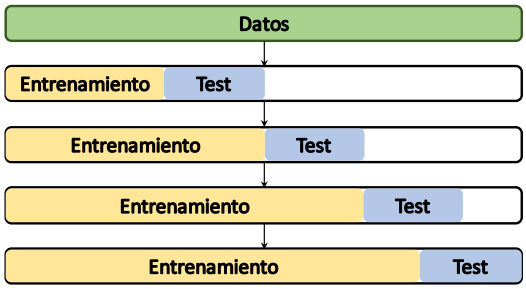

Figura 2 Representación del procedimiento de validación cruzada para series temporales

los senos y cosenos de frecuencias f_i son las características del modelo. Si agrupamos todos los coeficientes de las características en un vector de pesos $\boldsymbol{\omega} = [1, \alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k]$, donde k es el número de ritmos, y agrupando todas las características y el MESOR en una matriz $\boldsymbol{X}$, el modelo ritmométrico puede ser formulado de forma matricial como,

$$\boldsymbol{y} = \boldsymbol{X\omega} + \boldsymbol{e} \qquad (3)$$

Los pesos del modelo, $\boldsymbol{\omega}$, pueden ser estimados utilizando LS incluyendo un término de regularización,

$$\hat{\boldsymbol{\omega}} = argmin\|\boldsymbol{y} - \boldsymbol{X\omega}\|_2^2 + \lambda\|\boldsymbol{\omega}\|_1 \qquad (4)$$

donde $\|\boldsymbol{\omega}\|_1 = \sum_{p=1}^{2k+1}|\omega_p|$ es la norma l_1 de ω, y λ es el parámetro de regularización que debe ser especificado por el usuario [10].

La regularización lasso permite controlar el número de pesos activos ($\omega_p \neq 0$), de forma que, si λ es suficientemente grande, algunos de los pesos serán exactamente cero. Esto no se cumple para las normas l_q con $q > 1$ [8], ver Fig. 1. Por lo tanto, podemos utilizar modelos lasso para extraer las características relevantes de las señales [11].

Podemos encontrar modelos de regresión en los cuales las características poseen una estructura grupal natural, en el análisis de ritmos, las componentes seno y coseno de un determinado periodo constituyen un grupo. En estos casos es razonable que todas las características pertenecientes a un mismo grupo se encuentren activas (o inactivas) simultáneamente, esto es, que sus respectivos pesos sean diferentes de cero (o cero) [8].

Si consideramos un modelo de regresión lineal con J grupos de características, donde para $j = 1, \dots, J$, el vector $\boldsymbol{Z}_j$ agrupa las características del grupo j, y $\boldsymbol{\theta}_j$ representa el conjunto de coeficientes de regresión para el grupo j. Reuniendo todos los grupos en una matriz $\boldsymbol{Z}$, el modelo puede ser formulado de forma matricial,

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

Group lasso resuelve el siguiente problema convexo,

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

donde $\|\boldsymbol{\theta}\|_2 = \sum_{j=1}^{J}\|\boldsymbol{\theta}_j\|_2$ es la norma l_2 de $\boldsymbol{\theta}$. Una consideración importante es que este criterio conlleva la suma de normas l_2, en lugar de normas l_2 al cuadrado (*ridge regression*). De forma que esta formulación

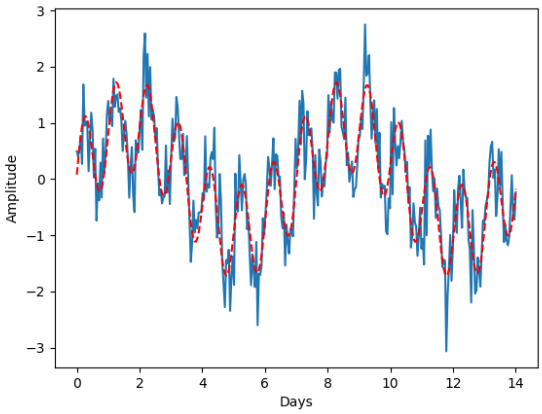

Figura 3. *Ejemplo de señal sintética (línea continua azul) creada como combinación de un ritmo de periodo 24 horas, un ritmo de periodo 7 días y ruido gaussiano. La señal resultante posee una SNR de 6 dB. El modelo ritmométrico resultante obtenido con el método propuesto (línea roja discontinua).*

impone una restricción global $l_1/\ l_2$ sobre el conjunto total de coeficientes. La consecuencia de esta restricción de grupo es que se seleccionan todas las características de un grupo para estar (o no) en el modelo resultante. Así mismo, dentro de un grupo de características (si es seleccionado) los coeficientes se obtienen mediante estimación LS.

2.2 Validación Cruzada para Series Temporales

Con el objetivo de seleccionar el parámetro de regularización λ, implementamos un esquema de validación cruzada capaz de preservar la estructura temporal de las señales.

Las series temporales se caracterizan por la correlación entre observaciones que se encuentran cercanas en tiempo. Las técnicas habituales de validación cruzada asumen que las observaciones son independientes e idénticamente distribuidas, por lo tanto, aplicar estas técnicas a series temporales da lugar a una correlación entre los conjuntos de entrenamiento y test que deriva en una pobre capacidad de generalización de los modelos.

Una aproximación razonable consiste en evaluar el modelo en observaciones futuras de la serie temporal. En este trabajo utilizamos una variación de la técnica de validación cruzada *k-fold,* donde las primeras *k-fold* son utilizadas como conjunto de entrenamiento y la *k+1 fold* como conjunto de test [12]. En este estudio hemos considerado cada *fold* igual a una muestra de la señal temporal. De esta forma, en la primera iteración del algoritmo (ver Fig. 2), las *l* primeras muestras de la señal conforman el conjunto de entrenamiento y la muestra *l+1* el de test. En la iteración *m*, las primeras *l+m-1* muestras pasan a ser el conjunto de entrenamiento y la muestra *l+m* el de test. Las iteraciones finalizan cuando *l+m = N*. La elección del valor de *l* depende de aspectos computacionales del algoritmo, así como de la longitud y frecuencia de muestreo de la señal.

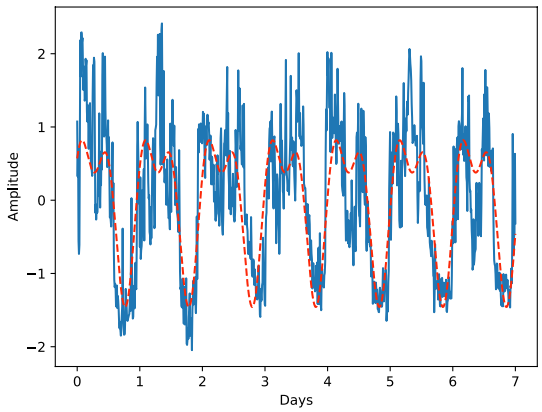

Figura 4. *Ejemplo de una señal real de RC (línea continua azul). El modelo ritmométrico resultante obtenido con el método propuesto (línea roja discontinua).*

El proceso se repite para un conjunto de valores de λ, seleccionando finalmente el valor del parámetro de regularización que proporciona el mínimo error cuadrático medio.

3. Experimentos

En este trabajo se han considerado los ritmos con periodos de 8, 12 y 24 horas, así como los ritmos de periodos 3.5 y 7 días, como características del modelo. No obstante, podría considerarse cualquier conjunto de ritmos. Con el objetivo de comprobar el funcionamiento del método se crearon 200 señales sintéticas. Cada una de las señales estaba formada por una combinación aleatoria de diferentes componentes sinusoidales (ritmos puros) y ruido gaussiano añadido. Se obtuvieron señales con relaciones señal a ruido (*signal to noise ratio*, SNR) entre 3 y 8 dB. Estas señales sintéticas simulaban la evolución temporal de una determinada variable durante 14 días, con un periodo de muestreo de 1 hora.

La Figura 3 muestra un ejemplo de una señal sintética creada como la combinación de una componente sinusoidal de periodo 24 horas, una componente sinusoidal de periodo 7 días y ruido gaussiano. La señal resultante posee una SNR de 6 dB (línea continua azul).

Adicionalmente el método se evalúo cualitativamente en dos conjuntos de señales cardíacas. Un conjunto de cuatro señales de RC obtenidas a partir de registros Holter de 7 días recogidos en la Unidad de Arritmias del Hospital Virgen de la Arrixaca de Murcia [2]. Las señales utilizadas en este trabajo fueron obtenidas como el RC medio de cada 10 minutos a lo largo de 7 días. La Figura 4 presenta un ejemplo de una señal real de RC (línea continua azul). Se utilizó también un segundo conjunto de cuatro señales reales, recogidas en la Unidad de Medicina Interna del Hospital Universitario de Móstoles. Estas señales presentan la evolución de la TA durante 24 horas. La Figura 5 presenta un ejemplo de una señal real de TA (línea continua azul).

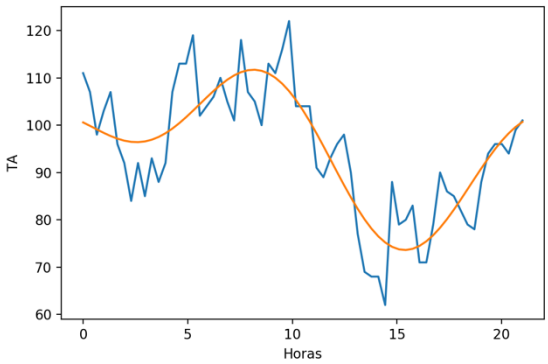

Figura 5. Ejemplo de una señal real de TA (línea continua azul). El modelo ritmométrico resultante obtenido con el método propuesto (línea roja discontinua).

Para seleccionar el parámetro de regularización (ver Sección 2.2), después de una primera inspección en un rango amplio de valores, concentramos la búsqueda de λ en el rango [0.001,1], probando 30 valores logarítmicamente espaciados para cada señal. Finalmente seleccionando aquel valor de λ que proporcionó el mínimo error cuadrático medio. El valor de *l,* fue fijado a 30 en el caso de las señales sintéticas y las señales reales de TA, y a 140 para las señales reales de RC.

4. Resultados

El método propuesto detectó correctamente el 98% de los patrones de ritmo en las señales sintéticas. Considerando un acierto únicamente cuando hay una coincidencia exacta entre el conjunto completo de ritmos seleccionado por el modelo y los presentes en la señal sintética. La Figura 3 muestra un ejemplo del modelo ritmométrico resultante al aplicar el método propuesto (línea discontinua roja) a la señal sintética (línea continua azul). El método detectó correctamente los dos ritmos (con periodos de 24 horas y 7 días) presentes en la señal ruidosa.

En las señales reales podemos observar cualitativamente que el método es capaz de extraer los ritmos presentes. La Figura 4 muestra el modelo ritmométrico resultante al aplicar el método propuesto (línea discontinua roja) a la señal de RC (línea continua azul), el método detectó ritmos de periodos 24 y 12 horas. La Figura 5 muestra el modelo ritmométrico resultante al aplicar el método propuesto (línea discontinua roja) a la señal de TA (línea continua azul), el método detectó ritmos de periodos 12 y 8 horas.

5. Conclusiones

En este trabajo proponemos un método automático de análisis de ritmos en series temporales basado en lasso o regularización con norma l_1. Aplicamos el método al análisis de señales cardíacas, pero es fácilmente adaptable para el análisis de cualquier tipo de series temporales. Dado que Lasso es el método convexo más cercano al problema de selección de subconjuntos de características, el método propuesto permite la identificación óptima de ritmos presentes en las señales.

Agradecimientos

Este trabajo ha sido financiado parcialmente por los proyectos TEC2013-48439-C4-1-R, TEC2016-75161-C2-1-R, TEC2013- 46067-R, TEC2016-81900-REDT (Ministerio de Economía y Competitividad, Gobierno de España).

Referencias

[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-Álvarez 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 O, Alzueta J, et al. Amulticentric 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 withSparsity: The Lasso and Generalizations. *CRC Press*, 2015.

[9] James G, Witten D, Hastie T, Tibshirani R. 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 learningin Python. *Journal of Machine Learning Research* 2011;12:2825–28.