# *Scoring* y análisis del sueño con técnicas basadas en la proyección sobre estructuras latentes y aprendizaje automático

A. González Cebrián[1], A.J. Ferrer Riquelme[1]

[1] Grupo en Ingeniería Estadística Multivariante (GIEM), Dpto. de Estatística e I.O. Aplicadas y Calidad, Universitat Politècnica de València, Valencia, España, algonceb@upv.es, aferrer@eio.upv.es

## Resumen

*En este trabajo se pretende caracterizar el sueño en base a criterios cuantitativos con el fin de llegar a una metodología para el análisis del sueño y el scoring que proporcione resultados repetibles y reproducibles.*
*Se propone la aplicación de técnicas de proyección sobre estructuras latentes y de aprendizaje automático, con el fin de obtener clasificadores que asignen una fase del sueño a cada tramo de 30 segundos (epoch) empleando información sobre Polisomnografías de pacientes distintos. Para lograrlo se ha realizado previamente una extracción de características basadas en una transformación al espacio tiempo-frecuencia mediante la Transformada Wavelet Discreta (TWD), de cada tramo de 30 segundos (epoch).*
*Tomando las fases asignadas por un profesional médico como referencia, se han construido y evaluado clasificadores basados en las diferentes técnicas estudiadas, estableciendo una comparación final entre los resultados.*

## 1. Motivación

Las alteraciones del hábito de sueño poseen una alta prevalencia y acarrean disfunciones en procesos fisiológicos básicos. Uno de los pasos críticos para la monitorización y diagnóstico del sueño es su *scoring*, un proceso de análisis mediante el que se asigna una fase del ciclo a cada tramo de 30 segundos (epoch) de la Polisomnografía (PSG). La realización manual basada en un análisis visual de las señales, sin un protocolo cuantitativo que sirva como estándar, conlleva variabilidad inter e intra-profesional en el diagnóstico, además de un gran consumo de tiempo [1], [2].

La integración de sistemas que permitan un *screening* rápido de los registros, o incluso la automatización del análisis visual mediante algoritmos, es uno de los grandes hitos planteados actualmente [3], [4]. Este cambio de paradigma permitiría mejorar la eficacia y eficiencia en la gestión de los trastornos del sueño. Además, esto supondría el abaratamiento del coste y, por tanto, del acceso a estos servicios [5]. Es, por tanto, necesario desarrollar protocolos que contemplen la inclusión de nuevas tecnologías con el fin de aumentar la eficiencia, repetibilidad y reproducibilidad con las que cuenta actualmente.

## 2. Introducción

En los últimos años se han desarrollado varias metodologías *data-driven* para un análisis rápido y robusto de los registros polisomnográficos.

Una vez registradas las señales, pueden extraerse características de distinta naturaleza en base a las PSGs:

- **Características en el dominio temporal.** Pueden establecerse dos grandes tipos de parámetros: parámetros sobre la morfología de la señal (desde estadísticos clásicos a otros menos comunes [6]) y parámetros sobre la información de la señal (parámetros no lineales o métodos fractales basados en la comparación con señales de nivel de ruido conocido).
- **Características en el dominio espectral.** Explotando la relación conocida entre bandas de frecuencia e información fisiológica, suele emplearse la *Fast Fourier Transform* (*FFT*) y, posteriormente, pueden calcularse parámetros o gráficos.
- **Características en el dominio tiempo-frecuencia.** Permiten, además de una descomposición en bandas de frecuencia, mantener la información temporal. Algunas opciones son la *Short Time Fourier Transform* (*STFT)* o la Transformada *Wavelet* (TW). Sobre las señales transformadas, parámetros que calculen la energía de cada banda de frecuencias en las ventanas temporales son útiles para conocer qué zona del espectro tiene más información de actividad que el resto.

Tras haber calculado las características, pueden emplearse distintas técnicas para la explotación de la información obtenida:

- **Técnicas de aprendizaje automático**. Suelen ofrecer soluciones rápidas en problemas de clasificación, como en este caso es el proceso de asignación de fases del sueño. Sin embargo, dada la naturaleza de estas técnicas (Máquinas de Soporte Vectorial (*SVM*), *K*-Vecinos Más Próximos o Árboles de Clasificación, entre otras), la comprensión de las variables que afectan al modelo, así como las relaciones entre estas y entre las clases, debe inferirse de los propios resultados dado que no se obtiene una formulación del modelo como tal. Es común la selección de variables en base a su importancia dentro del modelo.
- **Técnicas estadísticas clásicas**. En el otro extremo se encuentran herramientas clásicas como los Modelos de Regresión Logística. Sin embargo, su aplicación en contextos como este se ve prácticamente imposibilitada por la propia naturaleza y estructura de los datos, siendo habitual encontrar datos faltantes o

un gran número de variables explicativas con una compleja estructura de correlación, lo que resulta en el cálculo de matrices singulares o mal condicionadas para obtener el modelo. Por estas razones estas técnicas clásicas muestran una capacidad mucho menor que otras técnicas más modernas para modelar procesos o fenómenos complejos.

- **Técnicas basadas en proyección sobre estructuras latentes**. Estas técnicas ofrecen la posibilidad de desentrañar las complejas relaciones existentes dentro de una estructura de datos mediante la identificación de las variables latentes subyacentes. Entre este tipo de técnicas se encuentran el Análisis de Componentes Principales (*PCA*), y el Análisis Discriminante basado en Mínimos Cuadrados Parciales (*PLS-DA*). Estas técnicas pueden ser usadas en situaciones con datos faltantes y matrices mal condicionadas (o incluso singulares), a la vez que facilitan la comprensión de los sistemas estudiados, gracias a la reducción del número de variables a considerar (variables latentes).

## 3. Materiales y métodos

Entre las variantes mencionadas para lograr el objetivo de este trabajo, se ha propuesto el uso de la TW Discreta (TWD) para la **extracción de características** basada en un análisis tiempo-frecuencia. Posteriormente, respondiendo a las necesidades y retos clínicos, se han propuesto **tres metodologías *data-driven*** para la explotación de las características extraídas.

### 3.1. Estructura de la Base de Datos

El punto de partida son polisomnografías de 103 pacientes con información de Electroencefalograma (EEG), Electrooculograma (EOG) y Electromiograma (EMG). Esta información se encuentra en forma de ficheros en formato *edf* (*European Data Format*), un estándar dentro de la comunidad médica para esta clase de archivos. La asignación de las fases del sueño se encuentra en otros ficheros (Figura 1).

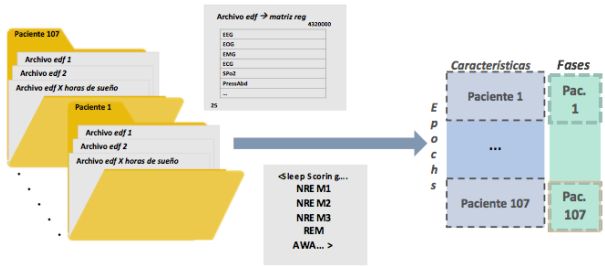


**Figura 1.** *Estructura de la información en su estado inicial, en distintos ficheros y carpetas.*

Toda esta información desagregada en distintos ficheros y carpetas debe organizarse en una matriz de observaciones por características o variables explicativas, sobre la cual se aplicarán diversas técnicas de clasificación y análisis.

### 3.2. Extracción de características basadas en la TWD

Al tratar de aunar en una sola matriz todos los registros surgen limitaciones de memoria y velocidad de procesado debido a que un solo electrodo registra al cabo de la PSG, del orden de millones de datos. En base a la información en el EEG, EOG y EMG, se ha representado cada *epoch* mediante un vector fila de características, sumando la fase del sueño asignada (Figura 2).

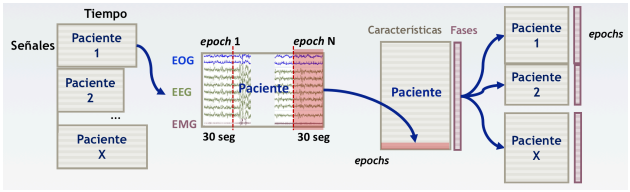


**Figura 2.** *Esquema con los pasos para la extracción de características.*

Para esta extracción de características se ha optado por la TWD, una herramienta matemática que, partiendo de una señal temporal, obtiene unas series de coeficientes: L con información de bajas frecuencias y H de altas frecuencias.

Un parámetro a configurar es la elección de la *wavelet* madre ($\Psi$), una función con cierto comportamiento oscilatorio que es dilatada y trasladada. La elección de las *wavelets* se ha realizado en base a información encontrada en la bibliografía [7]–[9] y el número de niveles de descomposición en base a las frecuencias con información de interés clínico.

El número de elementos de las series de coeficientes iguala la longitud original de la *epoch*, que es del orden de $10^4$ datos. Para hacer manejable esta cantidad de información, se ha recurrido al cálculo de la varianza de los coeficientes y de cada una de las *epochs* (Figura 3).

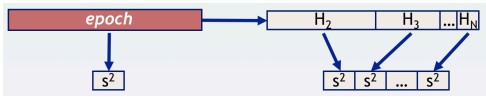


**Figura 3.** *Extracción de características de cada epoch.*

Tras aplicar este proceso a cada *epoch* de cada uno de los electrodos para cada paciente, se tiene una base de datos con 78 variables: 77 variables explicativas y una variable respuesta correspondiente a la fase del sueño asignada por un profesional médico. El número final de pacientes es de 38, con un promedio de 900 *epochs* por cada paciente.

### 3.3. Metodologías propuestas para el análisis del sueño y el *scoring*

Las soluciones propuestas basadas en técnicas de análisis de datos han sido tres:

- **Análisis preliminar de la base de datos.** Es aconsejable comprobar si hay observaciones atípicas o extremas en la base de datos que no deban considerarse. Teniendo en cuenta que se trata de un entorno en el que hay una gran cantidad de datos, se ha considerado el uso de PCA[10] y PLS-DA[11] para realizar un análisis preliminar de los datos con **fines** básicamente **exploratorios** e **inferenciales**.

Este apartado se ha realizado con el software MVA-GIEM.

- ***Scoring* automático del sueño.** Con el fin de **reducir** el **tiempo** del *scoring*, se han comparado los siguientes clasificadores: Árbol de Clasificación, Árbol de Clasificación con selección de variables más importantes tras un Bosque Aleatorio, *K*-Vecinos Más Próximos y Máquina de Soporte Vectorial. Este clasificador se ha denominado **global** debido a que se entrena con datos de unos pacientes para clasificar datos de pacientes totalmente distintos. Los resultados se han evaluado mediante el Área Bajo la Curva *ROC* [12], pudiendo discriminar así el acierto y fallo según fases.
- ***Scoring* semi-automático del sueño.** Tratando de mejorar los resultados obtenidos con la aproximación previa, este enfoque **individual** es especialmente recomendado cuando la **variabilidad entre pacientes** es elevada. Este método usa los datos del *scoring* manual de cierta parte del sueño de cada paciente para el entrenamiento del clasificador, realizando el *scoring* automático del resto del registro de cada individuo. Para cuantificar cuánto afectaría la cantidad de conocimiento incorporado *a priori* por el profesional médico, se ha considerado entrenar con un 70% o un 50% de las *epochs*. Los resultados se han evaluado calculando la tasa de acierto.

# 4. Resultados

## 4.1. Análisis preliminar

Tras obtener los modelos PCA y PLS-DA, se han evaluado las posibles observaciones atípicas o extremas con los gráficos de los estadísticos Error Cuadrático de Predicción y $T^2$ de Hotelling, respectivamente.

En el caso del modelo PCA, con cuatro Componentes Principales (CPs) se explica casi el 80% de la variabilidad en los datos. Para saber con qué variables originales están relacionadas estas componentes, se muestran los siguientes gráficos de pesos o *loadings* (Figura 4).

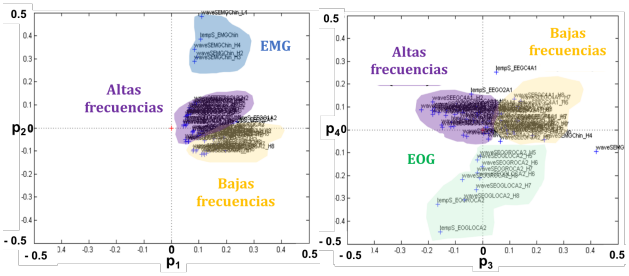


**Figura 4.** *Gráficos de loadings para las CP 1 y 2 (izda.) y CP 3 y 4 (dcha.).*

En ellos se aprecia relación entre CPs e información médica, teniendo la actividad muscular mayor peso en la segunda CP, la actividad ocular mayor peso en la cuarta, y en base a la tercera CP se distinguen las frecuencias altas y bajas (Figura 4).

En cuando al modelo PLS-DA, su capacidad de predicción de las variables respuesta es globalmente baja (30% con 6 variables latentes), siendo un poco mejor para para las fases NREM3 (fase III sin movimientos oculares rápidos), REM (fase de movimientos oculares rápidos) y AWA (despierto). Analizando los coeficientes *Jackniffe* pueden detectarse las variables estadísticamente significativas para las tres fases del sueño mejor ajustadas por el modelo.

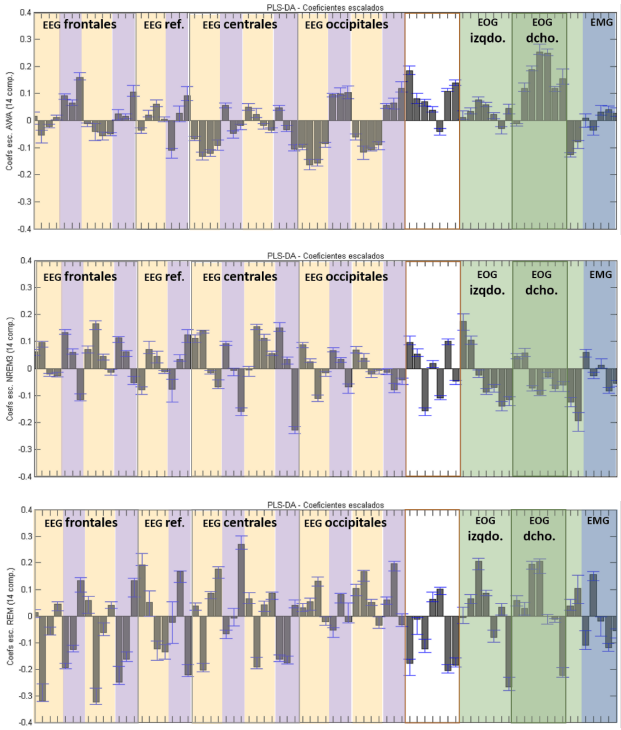


**Figura 5.** *Coeficientes Jackniffe con la importancia de las variables explicativas para las fases AWA (sup.), NREM3 (med.) y REM(inf.).*

En el caso de la fase AWA (superior) tienen peso mayoritario las variables asociadas a la actividad ocular y variables referentes a actividad cerebral registrada en la región occipital, clínicamente relacionada a estados de relajación previos al sueño. Para la fase NREM3 (medio) hay un cambio en la importancia de las variables, especialmente con información de bajas frecuencias (amarillo), clínicamente relacionadas con las fases de sueño profundo. Finalmente, en la fase REM (inferior) vuelve a sobresalir la actividad del EOG y aparecen coeficientes elevados para la actividad cerebral de alta frecuencia (morado), característica de esta fase del sueño.

## 4.2. Clasificadores globales

Tras realizar un ANOVA sobre el *AUC* teniendo en cuenta los factores Clasificador, Fase del sueño e Iteración (considerado como factor de bloqueo para aumentar la potencia estadística del test), todos ellos resultan estadísticamente significativos con un riesgo de primera especie del 5%. Tras validar el modelo comprobando la homocedasticidad y normalidad de los residuos, se analizan los intervalos LSD para la predicción del *AUC* según los clasificadores y fases (Figura 6).

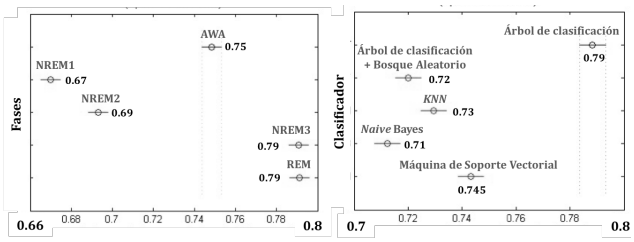


**Figura 6.** *Intervalos LSD del AUC medio predicho para los factores Fase (izda.) y Clasificador (dcha.).*

Los mejores resultados han sido entorno al 80% para las fases NREM3, REM y los Árboles de Clasificación, seguidos por un 75% para la fase AWA y las Máquinas de Soporte Vectorial.

### 4.3. Clasificadores individuales

De la misma forma que se ha explicado para los modelos globales, tras realizar un ANOVA para los factores Clasificador, Paciente y % de Entrenamiento, los tres resultan estadísticamente significativos. Pese a constatarse significativa la importancia del factor Paciente, al analizar los intervalos LSD para la predicción de la tasa de acierto media, se observa, que según el sujeto, esta fluctúa entre un 50% y un 75%. En este caso los mejores resultados son los obtenidos por la Máquina de Soporte Vectorial.

Por otro lado, la diferencia entre emplear un 50% o un 70% de los datos para entrenar supone una diferencia del 1% en la tasa de acierto, mientras que, a nivel de reducción del tiempo empleado, realizar el *scoring* manual de un 50% supondría una reducción del 25% del tiempo empleado.

## 5. Conclusiones y líneas futuras

Tras haber realizado el trabajo, se ha comprobado que el uso de distintas aproximaciones *data-driven* podría ser beneficioso para mejorar la eficacia y la eficiencia en los trastornos del sueño.

En cuanto al Análisis Preliminar de los datos, se propone PLS-DA como método para un análisis exploratorio e inferencial que permita estudiar la existencia de observaciones atípicas o extremas que no debiesen tenerse en cuenta para la construcción de los clasificadores posteriores.

Para una clasificación que permita obtener resultados repetibles y reproducibles en el *scoring*, se ha comprobado que un clasificador global ofrece tasas de acierto similares al individual, pero reduciendo el tiempo de forma mucho más drástica, pasando de unas 3 horas por análisis de paciente, a solo unos pocos minutos. Los Árboles de Clasificación o Máquinas de Soporte Vectorial han mostrado los mejores resultados.

Por otro lado, quedan aspectos abiertos al estudio en un futuro. Sería también de gran interés el estudio de la influencia del paciente en la clasificación, analizando una posible relación entre patologías y el ajuste de distintos modelos. Se podría considerar el uso de estadísticos robustos para la detección de observaciones atípicas o extremas, así como probar más clasificadores o características extraídas. Por ejemplo, *kernel*-PLS [13]. permitiría tanto un análisis preliminar exploratorio e inferencial, como la obtención de modelos que reflejen no linealidades en los datos.

## Referencias


[1] R. Boostani, F. Karimzadeh, y M. Nami, «A comparative review on sleep stage classification methods in patients and healthy individuals», *Comput. Methods Programs Biomed.*, vol. 140, pp. 77-91, 2017.

[2] J. E. Russo, R. R. McCool, y L. Davies, «VA Telemedicine: An Analysis of Cost and Time Savings», *Telemed. e-Health*, vol. 22, n.º 3, pp. 209-215, mar. 2016.

[3] N. A. Collop *et al.*, «Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients. Portable Monitoring Task Force of the American Academy of Sleep Medicine.», *J. Clin. Sleep Med.*, vol. 3, n.º 7, pp. 737-47, 2007.

[4] G. Singh, J. D. Miller, F. H. Lee, D. Pettitt, y M. W. Russell, «Prevalence of cardiovascular disease risk factors among US adults with self-reported osteoarthritis: data from the Third National Health and Nutrition Examination Survey.», *Am. J. Manag. Care*, vol. 8, n.º 15 Suppl, pp. S383-91, oct. 2002.

[5] N. F. Watson, «Health care savings: The economic value of diagnostic and therapeutic care for obstructive sleep apnea», *J. Clin. Sleep Med.*, vol. 12, n.º 8, pp. 1075-1077, 2016.

[6] B. Hjorth, «EEG analysis based on time domain properties.», *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, n.º 3, pp. 306-10, sep. 1970.

[7] P. S. Addison, «Wavelet transforms and the ECG: a review», *Physiol. Meas.*, vol. 26, n.º 5, pp. R155-R199, 2005.

[8] M. Akin, «Comparison of Wavelet Transform and FFT Methods in the Analysis of EEG Signals», *J. Med. Syst.*, vol. 26, n.º 3, pp. 241-247, 2002.

[9] V. Gerla, K. Paul, L. Lhotska, y V. Krajca, «Multivariate analysis of full-term neonatal polysomnographic data», *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, n.º 1, pp. 104-110, 2009.

[10] S. Wold, K. Esbensen, y P. Geladi, «Principal component analysis», *Chemom. Intell. Lab. Syst.*, vol. 2, n.º 1-3, pp. 37-52, ago. 1987.

[11] P. Geladi y B. R. Kowalski, «Partial Least-Squares Regression: a Tutorial», *Anal. Chim. Acta Elsevier Sci. Publ. B.V*, vol. 185, pp. 1-17, 1986.

[12] T. Sing, O. Sander, N. Beerenwinkel, y T. Lengauer, «ROCR: visualizing classifier performance in R», vol. 21, n.º 20, pp. 3940-3941, 2005.

[13] R. Vitale, O. E. de Noord, y A. Ferrer, «A kernel-based approach for fault diagnosis in batch processes», *J. Chemom.*, vol. 28, n.º 8, pp. 697-707, 2014.