Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Estadística

Métodos para ordenar observaciones de señales oscilatorias. Aplicación a la ordenación de expresiones de genes

Autor: *Víctor López Iglesias*

Tutor/es: *Cristina Rueda Sabater, Yolanda Larriba González*

MÉTODOS PARA ORDENAR OBSERVACIONES DE SEÑALES OSCILATORIAS. APLICACIÓN A LA ORDENACIÓN DE EXPRESIONES DE GENES

TFG ESTADÍSTICA

Víctor López Iglesias
Autor del trabajo
Estudiante de INdat*
Universidad de Valladolid
victor.lopez@alumnos.uva.es

Cristina Rueda Sabater
Tutora del trabajo
Departamento de Estadística e IO
Universidad de Valladolid
cristina.rueda@uva.es

Yolanda Larriba González
Tutora del trabajo
Departamento de Estadística e IO
Universidad de Valladolid
yolanda.larriba@uva.es

ABSTRACT

El presente trabajo tiene como objetivo realizar una introducción al tema de agregación de órdenes circulares y a como resuelve un problema de interés en el ámbito de la cronobiología. Además se pretende proponer algunas alternativas metodológicas y validar las mismas.

The objective of this document is to introduct the circular order aggregation topic and his relevance in chronobiology. Moreover, we proppose some methodological alternatives and we validate them with real data.

Keywords Agregación de órdenes circulares · Circular order aggregation · Señales oscilatorias · Oscillatory signals · Cronobiología · Chronobiology

*Ingeniería Informática y Estadística

1. Introducción

En este trabajo, se busca proponer y resolver un problema de interés en el ámbito de la cronobiología, que busca refinar las técnicas existentes y realizar una nueva propuesta metodológica no presentada hasta la fecha en esta disciplina. Por ello, antes de comenzar con detalles más técnicos es importante conocer un poco más esta disciplina.

La **cronobiología** es una rama de la biología que estudia los sucesos de carácter periódico que acontecen en los seres vivos. En esta disciplina se busca secuenciar de manera temporal las alteraciones que se producen en los organismos de los seres vivos, así como descubrir qué mecanismos regulan estos cambios. Son de gran importancia sus aplicaciones en la ciencia del sueño y en la comprensión del comportamiento de los seres vivos.

La gran mayoría de los procesos estudiados en cronobiología están regulados por el ciclo luz-oscuridad, conocido como ciclo circadiano. De forma que gran parte de las variables biológicas se ven alteradas de manera rítmica por este fenómeno cíclico de carácter diario. Este ciclo circadiano es fundamental para una correcta comprensión del funcionamiento de los procesos biológicos que ocurren en los seres vivos.

En concreto, la gran mayoría de procesos fisiológicos están regulados por las proteínas que son sintetizadas en los genes. La sintetización de estas proteínas es el resultado de un proceso de transformación de la información codficada en el ADN y el ARN (ácidos nucleicos), que tiene lugar en los seres vivos y les permite realizar sus funciones vitales y desarrollarse correctamente. Este proceso se conoce como **expresión génica**.

La expresión génica suele presentar un carácter rítmico, con un patrón oscilatorio *up-down-up* a lo largo del tiempo. El momento de máxima expresión del gen coincide con la hora del día en la que se lleva a cabo la función biológica asociada a dicho gen. Sin embargo, esta ritmicidad no está presente en todos los genes y, por tanto, existe una diferenciación entre genes rítmicos y no rítmicos.

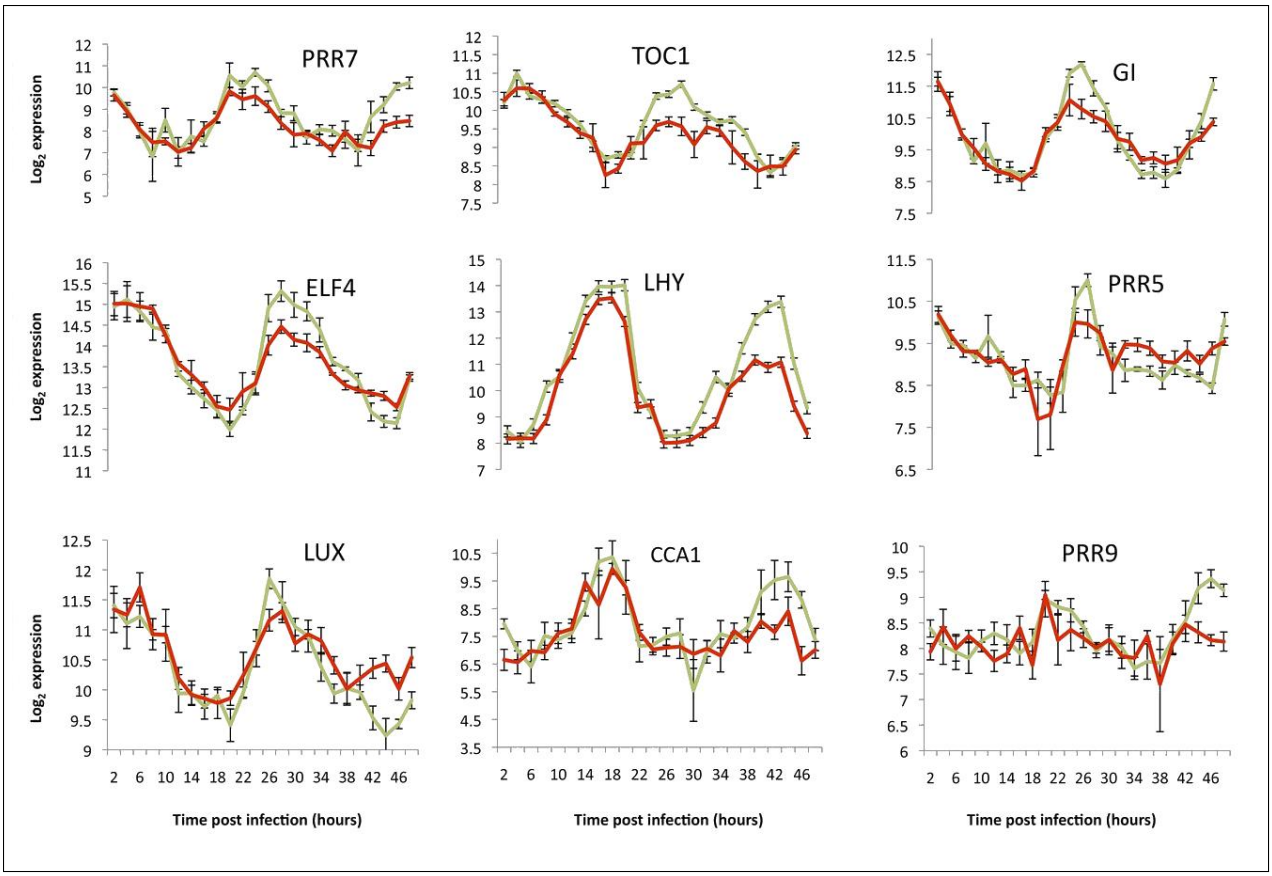

Figura 1: Expresión de distintos genes de una especie de planta, en rojo individuos infectados por un virus, en verde individuos control

Es de especial interés, en consecuencia, la elaboración de técnicas para la detección de la ritmicidad de un gen a partir de su expresión génica, con el fin de conocer qué fenómenos biológicos tienen una fuerte componente rítmica y qué influencia tiene este ciclo en ellas. Así mismo, es muy importante conocer el ciclo exacto que sigue esta expresión génica (su ordenación) para conocer la relación entre el ciclo ambiental y las funciones biológicas.

Ambos son problemas matemáticos, y más en concreto estadísticos, muy interesantes y complejos cuya resolución es de gran importancia para la consecución de avances en esta disciplina de la biología. La detección de señales rítmicas es un tema tratado en [2]. En este documento se proporciona una metodología novedosa, aplicando la teoría ORI, del inglés Order Restricted Inference [3]. En él, también se compara con un método aplicado habitualmente en la cronobiología para detectar patrones rítmicos, conocido como JTK, véase [4].

Uno de los objetivos del trabajo es realizar una introducción a las técnicas de estimación de un orden circular desconocido que subyace en unos datos de expresión génica, los cuáles están formados por varios genes con distintos valores de expresión, tomados cada uno de ellos en individuos diferentes y en un momento distinto.

La secuencia temporal real de estas muestras es lo que es desconocido y, en consecuencia, se busca estimar dicho orden, como paso previo a determinar si el patrón de expresión de los genes de los que disponemos (una vez ordenado) es cíclico o no.

Para estimar este orden, puesto que se dispone de varios genes diferentes, se busca un orden común, construido a partir de la información de todos los genes disponibles, que es lo que conocemos como orden agregado.

Para ayudar a entender esto, la diferencia principal al disponer de varios genes con respecto de cuando disponemos de un solo gen, es como si en una competición, los atletas, en lugar de realizar una única carrera (con sus puestos correspondientes), realizasen varias y la clasificación final fuese el resultado de ponderar de alguna manera cada una de las carreras (órdenes de los distintos genes) para formar una clasificación final (orden agregado). Salvando las distancias, puesto que la agregación de órdenes en este ejemplo sería en la línea mientras que nosotros vamos a trabajar en la geometría circular.

La estimación de este orden es de gran importancia puesto que no siempre se dispone del instante de tiempo en el que se han tomado las muestras o bien este es erróneo, como sucede con el TOD. El TOD es el momento en que se produce la muerte del individuo sobre el que se toma la muestra de expresión génica (*Time of Death*), el cual, en muchas ocasiones se toma como referencia temporal cuando esta muestra se realiza en la autopsia del individuo.

Sin embargo, el TOD da lugar a órdenes erróneos, entre otras razones, debido a que el instante en el que se produce lo que se conoce como muerte clínica (paro cardiaco) no supone un paro inmediato de todas las funciones de los órganos (muerte biológica), con lo cual este momento no coincide con el momento real de expresión de los genes en cuestión.

La reconstrucción de este orden real, pese a ser un problema de gran importancia, no ha sido muy tratado en la literatura y es que, hasta la aparición del artículo [1], no se había dado solución al problema de agregación de órdenes circulares.

En dicho artículo, se proponen dos alternativas al problema: el TSP y un método de tripletas basado en la teoría de Hodge, de las cuáles se hablará más en profundidad en la metodología. Estos métodos utilizan datos circulares para realizar la estimación del orden agregado.

No obstante, en ocasiones, los datos de los que se dispone son euclídeos, como es el caso de los utilizados en este trabajo. Para los cuáles, en el artículo [5] publicado posteriormente, se propone otro método basado en las componentes principales, el CPCA, del que también se hablará en la sección de la metodología.

De esta diferencia entre datos euclídeos y circulares, nace el segundo objetivo del trabajo, que es realizar una propuesta novedosa, en la cual se busca poder aplicar la metodología ya desarrollada para datos circulares a los datos euclídeos, realizando una transformación de estos a circulares a través del CPCA. Esta novedosa técnica se explicará en mayor pofundidad en la sección 2.4.2.

Para probar esta nueva técnica trabajaremos con datos de diversos tejidos, en formato euclídeo, como ya hemos comentado, de los cuales se tiene evidencia de ritmicidad. Estos datos proceden del portal GTEx, como se comenta en la sección de resultados, en la cual, se proporciona más información acerca de este conjunto de datos.

Sobre estos datos, se buscará comparar la estimación del orden que produce el CPCA (aplicable sobre los datos sin ninguna transformación) con la que producen diferentes alternativas de la metodología de Hodge sobre los datos transformados en datos circulares (vía CPCA).

En las siguientes secciones se busca realizar una introducción a la metodología utilizada hasta el momento para resolver el problema y presentar la técnica aplicada en este proyecto (sección 2), presentar los resultados resultantes de aplicar la metodología propuesta sobre unos datos euclídeos (sección 3) y realizar una reflexión sobre estos resultados y las conclusiones extraídas de los mismos (sección 4).

2. Metodología

En esta sección, se pretenden detallar las técnicas de las que se dispone actualmente para la agregación de órdenes circulares, parte de las cuáles se utilizarán con los datos disponibles. Antes de ello, plantearemos una notación adecuada.

2.1. Notación

Disponemos de un conjunto de datos con p genes y n instantes de tiempo diferentes.

Tenemos que $\boldsymbol{x_j} = (x_{1j}, \ldots, x_{nj})'$ con $j = 1 \ldots p$ es el vector de expresiones génicas del gen j para los instantes de tiempo identificados con el subíndice $i = 1 \ldots n$ para los datos euclídeos. El conjunto de los datos para los p genes será identificado como $\boldsymbol{X}$.

En el caso de los datos circulares, tenemos $\boldsymbol{\theta_j} = (\theta_{1j}, \ldots, \theta_{nj})'$ con $j = 1 \ldots p$ como vector de datos de la expresión génica en formato angular (entre 0 y 2π) del gen j para los instantes de tiempo identificados con el subíndice $i = 1 \ldots n$. El conjunto de los p vectores será identificado como $\boldsymbol{\Theta}$.

En el caso del vector de ordenación de los datos, tenemos $\boldsymbol{\tau_j} = (\tau_{1j}, \ldots, \tau_{nj})$ como vector de posiciones para cada una de las observaciones del gen *j*. Esta notación será únicamente utilizada para la definición de una de las distancias utilizadas en el método de Hodge, no será necesaria para la metodología.

Dado que algunos métodos utilizan matrices de preferencias, tenemos que definir la notación utilizada para las mismas. En el caso de las preferencias entre pares (bidimensionales), se utilizará Y_{ih}^{j} para denotar el grado de preferencia del elemento i sobre el h para el gen j.

Para el caso de las hipermatrices, usadas por ejemplo en el método de Hodge para preferencias entre tripletas, Ψ_{ihk}^{j} indica el grado de preferencia del orden circular $i \leq h \leq k \leq i$ sobre el orden $i \leq k \leq h \leq i$.

Establecida esta notación, únicamente falta por definir la notación utilizada en lo referente a órdenes.

Denotamos que un vector angular $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ sigue un órden $\boldsymbol{O} \in \mathcal{O}$ como $\boldsymbol{\alpha} \sim \boldsymbol{O}$, siendo $\mathcal{O}$ el espacio de todos los órdenes.

Por último, el vector que denota el CIRE, es decir, el vector que verifica el orden $\boldsymbol{O}$ más cercano al vector de observaciones angulares respecto a la suma de errores circulares (SCE) es $\tilde{\boldsymbol{\theta}}_{\boldsymbol{j}}^{(\boldsymbol{O})} = (\tilde{\theta}_{1j}^{(O)}, \ldots, \tilde{\theta}_{nj}^{(O)})'$. Lo que es lo mismo:

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{j}}^{(\boldsymbol{O})} = \underset{\boldsymbol{\alpha} \sim \boldsymbol{O}}{arg\ min}\ SCE(\theta, \alpha) = \underset{\boldsymbol{\alpha} \sim \boldsymbol{O}}{arg\ min} \sum_{i=1}^{n} (1 - cos(\theta_{ij} - \alpha_i)) \tag{1}$$

Esta definición la utilizaremos más adelante, en el apartado 2.3.1, para conocer el problema matemático al que nos enfrentamos.

2.2. Agregación de órdenes en la línea

Existe una contrapartida al problema presentado en este trabajo que es la agregación de órdenes en la línea. Parte de los métodos desarrollados para este problema, han sido utilizados como punto de partida para algunos de los métodos existentes en la agregación de órdenes circulares, por lo que es conveniente ponerlos en contexto.

Como ejemplo para entender el problema en el caso de la línea, tenemos que se dispone de **p** jueces que opinan sobre **n** productos diferentes. Para ello cada juez ordena los productos de menor a mayor, por lo que dispondremos de **p** rankings distintos que pueden diferir unos de otros. Lo que buscamos es encontrar un ranking global de los productos, o lo que es lo mismo, un orden común de preferencia de un producto respecto a otro.

Existe gran variedad de métodos para resolver este problema que se agrupan en dos familias dependiendo de la forma de abordarlo, de manera que, aparece una familia de métodos que elaboran rankings basados en la información que proporciona cada

elemento de manera individual y otra familia cuyos métodos utilizan las preferencias de un elemento frente a otro para estimar el orden global.

Uno de los métodos destacados para resolver este problema son el de Borda, que es el más simple pero uno de los más utilizados, consistente en sumar para cada elemento su posición obtenida en cada ranking. Estas sumas se ordenan produciendo el orden común estimado. De este método, deriva el método de Borda circular, como apliación de este a la geometría circular.

Otro método relevante es *HodgeRank*, que está basado en la teoría de Hodge para pares. La extensión a tripletas (orden mínimo en el círculo) de este es de donde nace el método de Hodge que se presenta en la sección 2.3.3.

Además de estos, existen innumerables métodos dado que este problema ha sido mucho más tratado en la literatura que el problema en la geometría circular. Por citar algunos, tenemos los basados en Montecarlo, los basados en cadenas de Markov, el método de Footrule ...

2.3. Agregación de órdenes circulares con información angular

2.3.1. Problema de minimización

Como ya hemos comentado, el problema al que nos enfrentamos es el de agregar órdenes circulares. Por una parte, dado un orden concreto, para cada gen tendremos un CIRE $\tilde{\boldsymbol{\theta}}_{\boldsymbol{j}}^{(\boldsymbol{O})}$ (el cual se explicó en la subsección 2.1), es decir, un vector que minimice la SCE para los datos de expresión génica de ese gen en concreto.

Con ello, podemos definir una distancia entre un vector de observaciones angulares y un orden concreto:

$$d(\boldsymbol{\theta_j}, \boldsymbol{O}) = MSCE(\boldsymbol{\theta_j}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{j}}^{(\boldsymbol{O})}) = \frac{1}{n}\sum_{i=1}^{n}(1 - cos(\theta_{ij} - \tilde{\theta}_{ij}^{(O)})) \tag{2}$$

Como buscamos un orden global, definimos en función de la distancia expuesta anteriormente (MSCE), la cual se define individualmente para cada uno de los genes, una segunda distancia sobre el conjunto de genes, es decir, teniendo en cuenta todos los vectores angulares de expresión asociados a los genes de los que se dispone:

$$d^*(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}^{(\boldsymbol{O})}) = MSCE(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}^{(\boldsymbol{O})}) = \sum_{j=1}^{p} \omega_j d(\boldsymbol{\theta_j}, \boldsymbol{O}) \tag{3}$$

En esta distancia, los ω_j son una medida de la variabilidad de cada uno de los p genes, que sirven para ponderar las diferencias de variabilidad entre los diferentes experimentos.

Esta distancia, da lugar a un orden óptimo que es el que la minimiza dentro del espacio de todos los órdenes posibles ($\boldsymbol{O^*} \in \mathcal{O}$), el cual, se define más formalmente como:

$$\boldsymbol{O^*} = \underset{\boldsymbol{O} \in \mathcal{O}}{arg\ min}\ d^*(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}^{(\boldsymbol{O})}) \tag{4}$$

2.3.2. TSP

El TSP (*Travelling Salesman Problem*) es un famosísimo problema que ha protagonizado incluso películas como *Travelling Salesman*, en la que un grupo de matemáticos trata de resolver el debate *P vs NP* tratando de solucionar el TSP en tiempo polinomial.

Es un problema NP-Hard[2] ampliamente abordado en la rama de la Investigación Operativa y las Ciencias de la Computación que resuelve el supuesto de tener que trazar una ruta a través de n ciudades, separadas cada una de ellas por una distancia dada, en la cual se pase una única vez por cada ciudad terminando el recorrido en la ciudad de origen. Todo ello, tratando de minimizar la distancia total recorrida.

[2]Un problema que es al menos tan complejo como un problema NP, los cuales se caracterizan por no ser resolubles en tiempo polinomial con respecto al tamaño de la entrada (n)

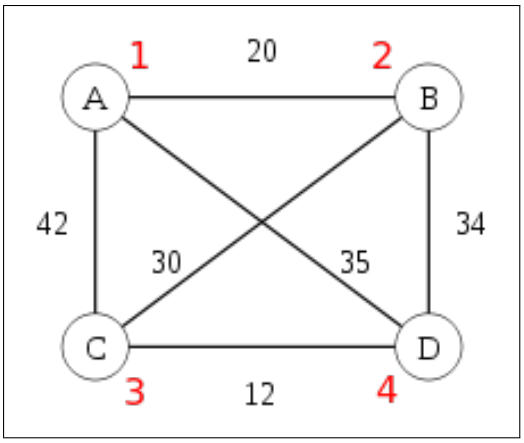

Figura 2: Ejemplo del *Travelling Salesman Problem*, solución en rojo

Como podemos ver es un problema cuya formulación se traduce a un grafo con aristas ponderadas entre los distintos nodos, en el que buscamos la ruta de menor peso total.

Este supuesto se puede extrapolar a una infinitud de problemas, como podemos imaginar, desde problemas en el ámbito de la logística (el campo de aplicación más evidente) pasando por el testeo de chips (el escáner recorre el circuito pasando por los distintos componentes minimizando la longitud del recorrido por razones de tiempo y ahorro energético) hasta llegar a la ordenación de señales oscilatorias.

La aplicación del **TSP** en la agregación de órdenes circulares requiere de un ejercicio de abstracción, pues en un primer vistazo las distintas componentes del problema pueden no ser tan evidentes.

En este ámbito tenemos como nodos los **n** instantes de tiempo en los que se han tomado valores de la expresión génica y los pesos de las aristas serían las distancias (la medida escogida para esta distancia es flexible) en el círculo entre cada par de instantes, buscando por tanto el recorrido (orden) que minimice la suma de distancias en el círculo entre los distintos nodos que conforman el *tour* (separación entre los mismos) dando lugar al orden más çoherente"de los datos.

Como disponemos de diferentes genes el peso de las aristas es un valor agregado de las distancias entre instantes para cada gen en el círculo. Comenzaremos definiendo estas distancias asociadas a cada gen j representadas como E^j_{hk}.

Los E_{hk}^{j}, como ya hemos comentado, representan las distancias en el círculo entre las distintas tomas de expresión génica (en momentos diferentes) para el gen j (distancia entre θ_{hj} y θ_{kj}). Esta distancia, en términos generales, será:

$$E_{hk}^{j} = d_{\alpha}(\theta_{hj}, \theta_{kj}) = min(d_R(\theta_{hj}, \theta_{kj}), \alpha \cdot d_C(\theta_{hj}, \theta_{kj})) \tag{5}$$

En esta definición entran en juego varios elementos clave: d_R, d_C y α. Hay que tener en cuenta que, al tratarse de un problema en la geometría circular, entre dos elementos existen dos distancias, una en una de las direcciones de rotación (d_R) y la otra, en la contraria (d_C). Con esto en mente, α es una constante de penalización para la dirección contraria.

Con respecto a α, su dominio de valores plausibles es $\alpha \in [1, \infty)$. Cuando $\alpha = 1$ estamos tratando con distancias no dirigidas, mientras que cuando $\alpha = \infty$ nos movemos sólo en la dirección de rotación. Por tanto, los valores intermedios medirán cuán indirigida es esta distancia.

Se puede utilizar cualquier tipo de métrica para las distancias entre elementos, siempre y cuando cumplan que sean:

1. Continuas
2. Acotadas
3. Positivas
4. Verifiquen una desigualdad triangular relajada (dados $\theta_i, \theta_h, \theta_k \in [0, 2\pi]$ y $\alpha \geq 1$):
$$d_{\alpha}(\theta_h, \theta_k) \leq 2(d_{\alpha}(\theta_h, \theta_i) + d_{\alpha}(\theta_i, \theta_k)) \tag{6}$$

Con la matriz de aristas asociadas a cada gen definida, queda conocer como se construye la matriz de aristas agregadas E. La construcción es la siguiente.

$$E_{hk} = \sum_{j=1}^{p} \omega_j E_{hk}^{j} \tag{7}$$

Los ω_j son las medidas de la variabilidad de cada gen de las que hablamos a la hora de definir la función objetivo del problema en el apartado 2.3.1.

Conocido cómo se forma la matriz de pesos procederemos a observar más detenidamente la formulación del problema.

El problema se formula como un problema de minimización de esta ruta que busca la matriz binaria, donde 1 en el elemento (i, j) se traduce a que en el *tour* circular se pasa del elemento i al j. Así del conjunto de tours posibles representado por el conjunto de todas las matrices binarias $n \times n$ ($\mathcal{X}$) buscamos $\hat{X}$, tal que:

$$\hat{X} = \underset{X \in \mathcal{X}}{arg\ min} \sum_{hk} X_{hk} E_{hk} \tag{8}$$

Es decir encontrar la matriz binaria que minimice la suma de pesos de las aristas correspondientes. Todo ello, por supuesto, sujeto a las restricciones que dan forma al problema (sino sería la matriz de ceros):

1. Se llega una única vez a cada nodo desde otro nodo.

$$\sum_{h=1}^{n} X_{hk} \quad \forall k = 1, \ldots, n \tag{9}$$

2. Se sale una única vez desde cada nodo a otro nodo.

$$\sum_{k=1}^{n} X_{hk} \quad \forall h = 1, \ldots, n \tag{10}$$

3. Siendo $V = \{1, \ldots, n\}$ el conjunto de nodos en el problema (el conjunto de tomas para la expresión génica), la suma de la matriz binaria de un subconjunto de nodos con más de un elemento será:

$$\sum_{h,k \in S} X_{hk} \leq |S| - 1 \quad \forall S \subset V, |S| > 1 \tag{11}$$

Esta formulación es la que sigue el TSP clásico, la cual, nos permite sacar ventaja de las heurísticas existentes, en la extensa literatura desarrollada al respecto, para aproximar la solución óptima.

2.3.3. Hodge

La teoría de Hodge fue desarrollada por W.V.D. Hodge en la década de los 30. Actualmente, constituye una herramienta muy importante en la rama de la topología y la geometría, concretamente en el cálculo de formas diferenciales en un variabilidad diferenciable M.

Mediante el uso de esta técnica se ha podido desarrollar este novedoso procedimiento para la ordenación de señales oscilatorias. Este método utiliza hipermatrices de preferencia para el orden circular definido por cada tripleta.

El método permite transformar un problema de agregación de órdenes en la geometría circular en un problema de mínimos cuadrados que simplifica el cálculo del orden común y permite calcular una solución aproximada de manera eficiente. El problema a resolver es de la forma:

$$\hat{\Psi} = \underset{\Psi \in \mathcal{H}_C}{arg\ min}\ ||\overline{\Psi} - \Psi||^2 \tag{12}$$

$\overline{\Psi}$ es la hipermatriz de preferencias agregada mientras que $\hat{\Psi}$ es la solución al problema (el orden circular más próximo al agregado en el sentido de mínimos cuadrados). $\mathcal{H}_C$ es el conjunto de hipermatrices que forman el espacio de soluciones al orden agregado buscado. Es un subconjunto restringido del conjunto de hipermatrices $n \times n \times n$ pero su definición no se corresponde con los objetivos de este trabajo.

Por tanto, partiremos de una hipermatriz de preferencias de tripletas por cada uno de los genes de los que disponemos, es decir, un total de p hipermatrices. Así, para cada gen j tendremos una Ψ^j asociada en la que cada uno de los elementos Ψ^j_{ihk} mide el grado de preferencia del orden circular $i \leq h \leq k \leq i$ sobre el orden $h \leq i \leq k \leq h$.

Estas hipermatrices verifican la propiedad de que entre sus elementos $\Psi^j_{ihk} = \Psi^j_{kih} = \Psi^j_{hki} = -\Psi^j_{hik} = -\Psi^j_{khi} = -\Psi^j_{ikh}$. Esto viene a representar que elementos que definen la preferencia de un mismo orden de tres elementos son iguales (evidentemente) y que, elementos con ordenes contrarios tienen el mismo valor absoluto pero signo opuesto.

Así, todos los Ψ^j_{ihk} siguen una norma general de construcción que tiene la forma:

$$\Psi^j_{ihk} = sign^j(i,h,l) \cdot \lambda^j_{ihk} \tag{13}$$

En esta regla $sign^j(i,h,k) = sign^j(\theta_{hj} - \theta_{ij}) + sign^j(\theta_{kj} - \theta_{hj}) + sign^j(\theta_{ij} - \theta_{kj})$ y λ^j_{ihk} es el parámetro que mide la intensidad de la preferencia en la tripleta y que se puede definir de diferentes formas, dandole flexibilidad y versatilidad al método.

Cada uno de los Ψ^j define un orden circular para el gen j inducido por los órdenes circulares entre cada una de las tripletas. Así pues, $\hat{\Psi}$ define un orden circular global resultante de la agregación de los órdenes de los p genes.

Una vez definida la construcción de las hipermatrices Ψ^j, el cálculo de la hipermatriz agregada $\overline{\Psi}$ resulta de usar en ellas la media aritmética o circular.

Antes de pasar a la obtención de la solución, debemos conocer los operadores que define la teoría de Hodge, permitiendo el paso de hipermatrices antisimétricas a matrices antisimétricas (δ_1^*) o viceversa(δ_1) y de matrices antisimétricas a vectores(δ_0^*) o viceversa (δ_0).

$$\Psi \overset{\delta_1^*}{\Longrightarrow} Y \overset{\delta_0^*}{\Longrightarrow} s$$
$$s \overset{\delta_0}{\Longrightarrow} Y \overset{\delta_1}{\Longrightarrow} \Psi$$

$$\delta_1^*(\Psi) = Y \qquad Y_{ih} = \sum_k \Psi_{ihk} \qquad ; \qquad \delta_1(Y) = \Psi \qquad \Psi_{ihk} = Y_{ih} + Y_{hk} + Y_{ki}$$

$$\delta_0^*(Y) = s \qquad s_i = \sum_h Y_{ih} \qquad ; \qquad \delta_0(s) = Y \qquad Y_{ih} = s_i - s_h$$

Finalmente, teniendo construidas las Ψ_{ihk}^j y habiendo obtenido la hipermatriz agregada $\overline{\Psi}$, buscamos hallar la hipermatriz solución al problema de agregación de órdenes circulares $\hat{\Psi}$.

Para ello debemos de pasar primero de la hipermatriz agregada $\hat{\Psi}$ a la matriz agregada $\hat{Y}$ vía δ_1^* y obtenemos $\boldsymbol{s}$:

$$s_i = -\frac{1}{n-1} \sum_{h \neq l_0} \overline{Y}_{ih} \qquad \forall_{i \neq l_0}, \qquad l_0 = arg\ max_l \sum_h \overline{Y}_{lh}^2 \tag{14}$$

Realizando el paso de la hipermatriz agregada $\overline{\Psi}$ a la matriz agregada $\overline{Y}$ y posteriormente de esta a los scores $\boldsymbol{s}$, podemos obtener el orden estimado por el método como la ordenación de estos scores.

La metodología de Hodge ha sido escogida para probar la nueva propuesta. Con los datos circulares transformados, se han aplicado diferentes variantes del método para estimar el orden agregado:

Alternativa	$\lambda_{ihk}^j \quad \forall i,h,k \in V$
POS	$\lvert F_{ih}^j - D_{ih}^j \rvert + \lvert F_{hk}^j - D_{hk}^j \rvert + \lvert F_{ki}^j - D_{ki}^j \rvert$
COS	$(1 + cos(\theta_{hj} - \theta_{ij})) + (1 + cos(\theta_{kj} - \theta_{hj})) + (1 + cos(\theta_{ij} - \theta_{kj}))$
AVE	$Ave(arc(\theta_{ij}, \theta_{hj}), arc(\theta_{hj}, \theta_{kj}), arc(\theta_{kj}, \theta_{ij}))$
E3	$d_3(\theta_{ij}, \theta_{hj}) + d_3(\theta_{ij}, \theta_{kj}) + d_3(\theta_{hj}, \theta_{kj}) + d_3(\theta_{hj}, \theta_{ij}) + d_3(\theta_{kj}, \theta_{ij}) + d_3(\theta_{kj}, \theta_{hj})$

Tabla 1: Alternativas utilizadas para el método de Hodge

Para la alternativa **POS** se utilizan F_{ih}^{j} y D_{ih}^{j} que se definen como:

$$D_{hk}^{j} = (\tau_{kj} - \tau_{h-1j})(mod\ n) \tag{15}$$

$$F_{hk}^{j} = (\tau_{hj} - \tau_{k-1j})(mod\ n) \tag{16}$$

Esta medida por tanto mide la diferencia entre el número de elementos entre h y k en el sentido de rotación y el número de elementos entre h y k en contrarrotación para cada par de elementos del orden definido por la tripleta. Son medidas usadas en métodos basados en cadenas de Markov.

La alternativa **COS** busca medir las distancias en el círculo entre cada par del orden de elementos definido por la tripleta, utilizando la medida más utilizada en la geometría circular para medir distancias, la cual resulta invariante frente a la orientación de los elementos.

En la variante **AVE** se utiliza la media circular de los arcos mínimos entre los pares de la tripleta. La media circular dado $\boldsymbol{\theta}$ se define como:

$$\overline{\theta} = Ave(\boldsymbol{\theta}) = atan2\left(\frac{1}{n}\sum_{i=1}^{n} sin(\theta_i), \frac{1}{n}\sum_{i=1}^{n} cos(\theta_i)\right) \tag{17}$$

Por último la alternativa **E3** utiliza la suma de unas distancias definidas incialmente para el problema del TSP. Recordemos de la sección 2.3.2 que estas distancias tienen varias componentes: una distancia de rotación (d_R), una distancia de contrarrotación (d_C) y un α que mide la penalización de la contrarrotación (en este caso, $\alpha = 3$):

$$d_R(\theta_{kj}, \theta_{hj}) = \begin{cases} 1 - cos(\theta_{kj} - \theta_{hj}) & \text{si } 0 \leq \theta_{kj} - \theta_{hj} \leq \pi \\ 3 - cos(\theta_{kj} - \theta_{hj} - \pi) & \text{si } \pi < \theta_{kj} - \theta_{hj} < 2\pi \end{cases} \tag{18}$$

$$d_C(\theta_{kj}, \theta_{hj}) = \begin{cases} 3 - cos(\theta_{kj} - \theta_{hj} - \pi) & \text{si } 0 \leq \theta_{kj} - \theta_{hj} \leq \pi \\ 1 - cos(\theta_{kj} - \theta_{hj}) & \text{si } \pi < \theta_{kj} - \theta_{hj} < 2\pi \end{cases} \tag{19}$$

2.4. Agregación de órdenes con datos euclídeos

2.4.1. CPCA

El método de análisis de componentes principales circulares (CPCA) está fuertemente basado en el análisis de componentes principales tradicional (PCA).

El PCA es un procedimiento estadístico para datos multivariantes que permite reducir la dimensionalidad (número de variables) del conjunto de datos minimizando la pérdida de información y reduciendo la interdependencia entre las distintas variables.

Este procedimiento data de 1901 y el autor del documento original es Karl Pearson, aunque, debido a la exigencia computacional del método este no se popularizó hasta la década de los 60, con la aparición de computadoras lo suficientemente potentes para implentarlo.

El método se basa en construir nuevas variables a partir de combinaciones lineales de las anteriores que expliquen la mayor parte de la variabilidad posible (recogiendo la mayor información).

Siendo $X_{n\times p}$ la matriz de datos, las componentes principales se obtienen utilizando la combinación lineal dada por los autovectores de $X'X$ cuya matriz denominamos U. Las nuevas variables (componentes principales) se construyen como XU (la combinación lineal de la que hablábamos).

Estos autovectores representan los ejes ortogonales que recogen la mayor variabilidad de la nube de puntos. Los autovalores sirven como medida de la variabilidad explicada por la componente principal correspondiente con respecto a la variabilidad total del conjunto de datos.

De esta idea nace el método desarrollado en [5]. La idea es realizar un PCA sobre los datos de manera que obtengamos autovectores para los genes (*eigengenes*).

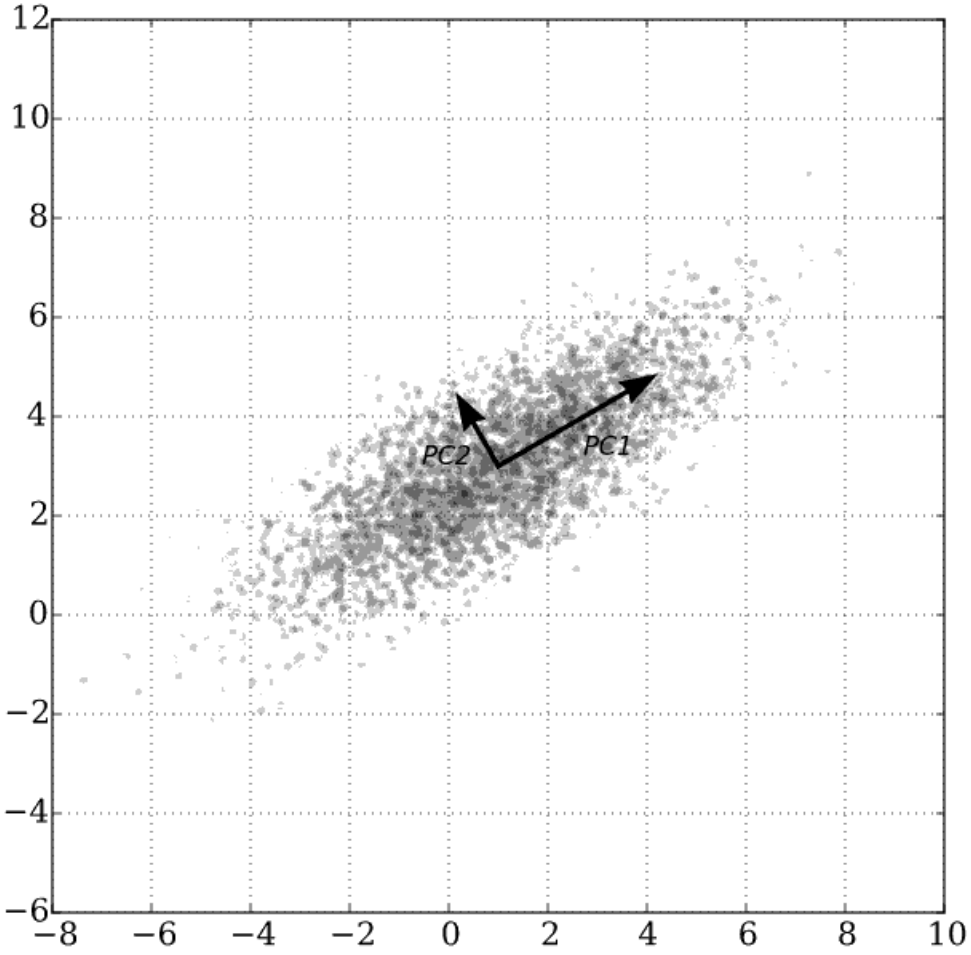

Figura 3: Ejemplo del análisis en componentes principales (PCA)

Del PCA resultante, nos interesan los dos primeros autovectores, los cuales, al ser representados en dos dimensiones, revelan una estructura circular subyacente debida a la ritmicidad de los genes (partimos de que lo son).

Con estos dos primeros eigengenes realizamos un paso de coordenadas cartesianas a coordenadas polares, utilizando la arcotangente, y obteniendo así, el ángulo en el que se encuentra cada observación, del cual, podremos inferir el orden circular agregado subyacente.

Si $\boldsymbol{u_1}$ y $\boldsymbol{u_2}$ son, respectivamente, el primer y segundo eigengen, entonces para obtener las proyecciones de los individuos

$$\boldsymbol{x} = \sqrt{\frac{\boldsymbol{u_1}}{\boldsymbol{u_1^2} + \boldsymbol{u_2^2}}} \qquad \boldsymbol{y} = \sqrt{\frac{\boldsymbol{u_2}}{\boldsymbol{u_1^2} + \boldsymbol{u_2^2}}} \tag{20}$$

Por tanto los ángulos resultantes los obtendremos como:

$$\boldsymbol{\theta} = arc\ tan(\boldsymbol{y}/\boldsymbol{x}) \tag{21}$$

Estas observaciones angulares generan un orden, al ser ordenadas de 0 a 2π, que es el que proporciona esta metodología.

Este método será comparado en el presente documento con la nueva propuesta puesto que es el que se ha venido utilizando hasta ahora para estimar el orden circular agregado en conjuntos de datos euclídeos.

2.4.2. Transformación de datos euclídeos en circulares

La nueva propuesta, presentada por primera vez en este trabajo, se basa en transformar los datos de expresión génica disponibles en formato euclídeo en datos circulares. Para ello se hace uso del CPCA y la equivalencia de ambos espacios cuando existe un orden circular. Esto nos permite aplicar posteriormente los métodos existentes cuando se dispone de información angular para estimar el orden circular agregado.

Como paso previo a la aplicación del método, debemos de realizar un escalado para cada $\boldsymbol{x_j}$ al intervalo [-1, 1]. Esto se hace para que la función arcoseno tome para los valores extremos de expresión los valores 0 y π correspondientemente.

Al realizar el CPCA sobre los datos euclídeos disponibles, obtenemos $\boldsymbol{\theta}$ como se explica en la sección 2.4.1. De la ordenación de $\boldsymbol{\theta}$ obtenemos $\hat{\boldsymbol{O}}$, que es el orden estimado por el CPCA.

Una vez obtenido este orden, debemos conocer para los datos de expresión génica de cada gen $\boldsymbol{x_j}$ en qué posición se encuentra L, punto más bajo de la señal oscilatoria (tomaremos el valor mínimo de expresión génica) y en cuál se encuentra U (tomaremos el valor máximo) que se corresponde con el punto más alto de la señal.

Conocidos L y U, utilizaremos $\hat{\boldsymbol{O}}$ para saber dentro de los datos de cada gen, en qué posición se encuentra cada instante de tiempo respecto a L y U, es decir, si se encuentran entre L y U o en el conjunto $\{1 \dots L\} \cup \{U \dots n\}$. En términos referentes a la figura 4 nos interesaría conocer qué observaciones se encuentran en el intervalo marcado con el color azul y cuales en los intervalos marcados en color rojo. Todo ello asumiendo el caso en el que $L < U$.

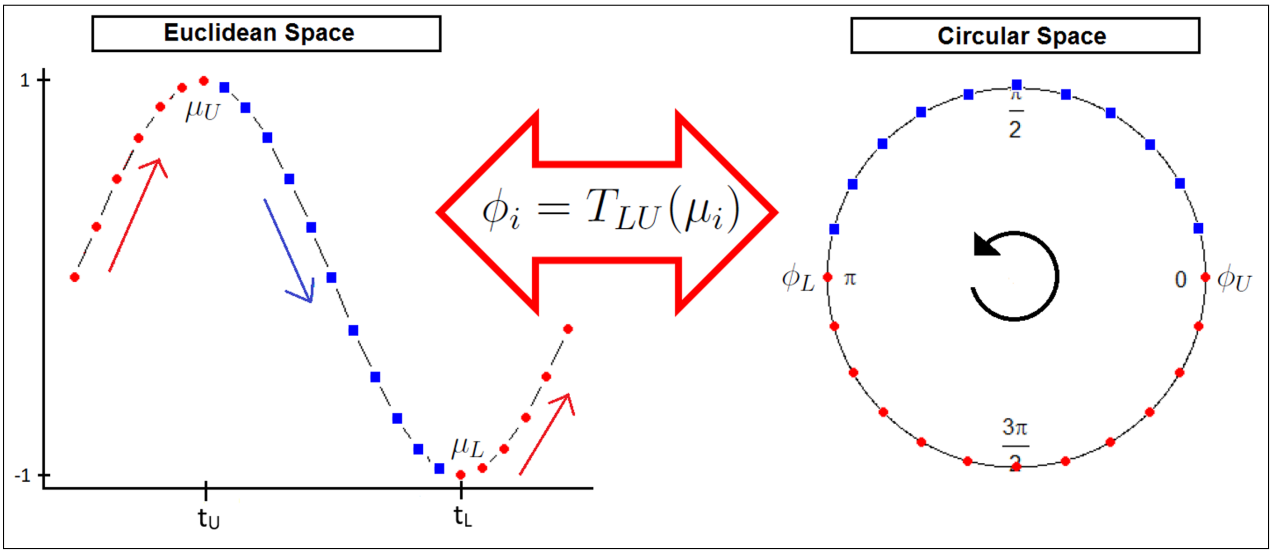

Figura 4: Equivalencia onda-circunferencia

Como podemos ver en la figura 4, L y U se corresponde en el espacio circular con 0 y π, respectivamente. Así, también podemos ver la correspondencia existente entre el espacio circular y el eculídeo (la onda).

La función que nos permite pasar del espacio euclídeo al circular es clave. En primer lugar, hay que tener en consideración la posición de L con respecto de U para cada gen. Habrá genes para los que $L < U$, mientras que para otros, $L > U$. Esta distinción es importante pues la función a aplicar es diferente.

Para diferenciar los L y U de cada gen denotaremos estos como L_j y U_j de manera que estos serán los asociados al gen j.

Para los casos en los que $L_j > U_j$ tenemos que la función que nos proporciona este paso es:

$$\theta_{ij} = T_{LU}(x_{ij}) = \begin{cases} arcsin(x_{ij}) - \frac{\pi}{2} & i \in \{1, \dots, U_j\} \cup \{L_j, \dots, n\} \\ \frac{\pi}{2} - arcsin(x_{ij}) & - \end{cases} \tag{22}$$

En cambio, para el caso contrario ($U_j > L_j$) tenemos:

$$\theta_{ij} = T_{LU}(x_{ij}) = \begin{cases} arcsin(x_{ij}) - \frac{\pi}{2} & L_j \leq i \leq U_j \\ \frac{\pi}{2} - arcsin(x_i) & - \end{cases} \tag{23}$$

El resultado de aplicar esta transformación para cada $\boldsymbol{x_j}$ dará lugar a los $\boldsymbol{\theta_j}$ que necesitamos, por ejemplo, para estimar un orden mediante el método de Hodge.

2.5. Medida de validación de los órdenes estimados

En el estudio realizado se comparan varios órdenes estimados para ver cuál de ellos resulta más coherente a los datos de los que disponemos. Lo ideal sería disponer del orden real de los datos y conocer cuál de ellos es más próximo a este orden. No obstante, como es común en este campo, no disponemos de dicho orden en los datos utilizados (como ya se explicó en la Introducción).

Por todo ello, se ha utilizado una medida basada en el ajuste del modelo cosinor, véase [6].

2.5.1. Modelo COSINOR

El modelo cosinor es un modelo de regresión desarrollado para ser utilizado en el análisis de series temporales y la detección de ritmicidad. Es un modelo muy utilizado en la cronobiología, aunque recientemente han surgido otros modelos más adecuados que representan patrones rítmicos más próximos a los de los genes.

La definición del modelo cosinor simple es la que se muestra a continuación, donde t es la variable que representa el tiempo:

$$X(t) = M + A \cdot cos\left(\frac{2\pi t}{\tau} + \phi\right) + e(t) \tag{24}$$

Siendo M el intercept, A la amplitud de la onda, ϕ la acrofase (el tiempo que tarda en aparecer el pico de la onda), τ el periodo de la onda y e(t) el término del error, que sigue una $N(0, \sigma)$ y es independiente. En este modelo, se asume que τ es conocido. Estos parámetros miden diferentes características de las ondas, tal y como se ilustra en la figura 5.

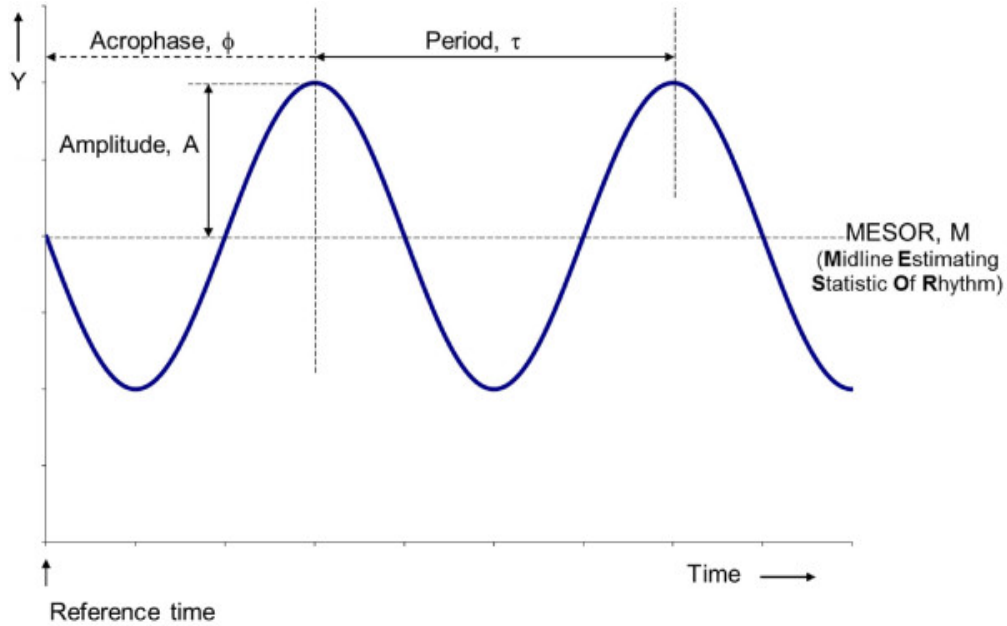

Figura 5: Parámetros del modelo cosinor

Para facilitar la tarea de encontrar los estimadores para este modelo, es necesario realizar una transformación de los parámetros aplicando identidades trigonométricas:

$$\beta = A \cdot cos\phi \qquad \gamma = A \cdot sin\phi$$
$$y = cos\left(\frac{2\pi t}{\tau}\right) \quad z = sin\left(\frac{2\pi t}{\tau}\right)$$

Con esta reparametrización, el modelo resulta en:

$$Y(t) = M + \beta y + \gamma z + e(t) \tag{25}$$

Con este modelo se realizan las estimaciones mediante el método de mínimos cuadrados. Con los parámetros estimados se pueden obtener los parámetros iniciales de la siguiente manera:

$$\hat{A} = \sqrt{\hat{\gamma}^2 + \hat{\beta}^2} \qquad \hat{\phi} = atan2(\hat{\gamma}, \hat{\beta}) + K\pi$$

Este modelo cosinor se ha ajustado a los datos euclídeos ordenados según el orden agregado estimado y se ha obtenido el error cuadrático medio individual para cada uno de estos ajustes (MSE_j). Con los MSE_j obtenidos se ha realizado el promedio para proporcionar una medida global del comportamiento del método para los datos.

$$MSE_j = \frac{\sum_{i=1}^{n}(\hat{x_{ij}}^{(O)} - x_{ij}^{(O)})^2}{n} \tag{26}$$
$$MSE = \frac{\sum_{j=1}^{p} MSE_j}{p} \tag{27}$$

Donde $\hat{x_{ij}}^{(O)}$ es el valor ajustado por el modelo cosinor.

3. Resultados

Los datos utilizados en trabajo para obtener los resultados que se muestran en esta sección proceden del portal de datos GTEx (`www.gtexportal.org`). Este portal proporciona de forma pública datos postmortem de expresión de gen para 54 tejidos humanos.

Este trabajo se centra en el análisis de los tejidos pulmonar, sanguíneo y muscular. Esta decisión se debe, entre otras razones, a que son tejidos con un alto porcentaje de genes rítmicos. Además, estos tejidos han sido utilizados en investigaciones previas y han producido resultados satisfactorios. Otra de las razones de la elección de estos tejidos, es que presentan un número de muestras suficiente, con un total de 97 individuos para el caso del tejido pulmonar, 98 individuos para el tejido sanguíneo y 111 individuos para el tejido muscular. Tejidos como la pituitaria, conocidos por presentar genes rítmicos y utilizados en otros estudios han sido descartados por no disponerse de un número de muestras satisfactorio.

En concreto para cada tejido, se han tomado datos de expresión génica para un total de 26831 genes del tejido pulmonar, 21044 genes del tejido sanguíneo y 22584 genes del tejido muscular.

De este conjunto de genes se ha hecho una selección de genes típicamente rítmicos conocidos como *genes core clock*, [7]. Estos genes presentan una fuerte componente rítmica y y revelan un patrón rítmico incluso en situaciones de mucha variabilidad, como es el caso de los datos de GTEx. El papel de estos genes es clave en este trabajo, puesto que la base de datos GTEx contiene tanto genes genes rítmicos como genes no rítmicos que actúan como ruido a la hora de estimar el orden circular.

La selección de *genes core clock* para este estudio se compone de 12 genes, que aparecen con frecuencia en estudios del campo de la cronobiología, véase [7], [8] y [9]. En concreto hemos seleccionado los genes *PER1*, *PER2*, *PER3*, *CRY1*, *CRY2*, *ARNTL*, *CLOCK*, *NR1D1*, *RORA*, *DBP*, *TEF* y *STAT3*.

Para cada uno de los tejidos se ha abordado el problema de agregación de ordenes circulares de manera independiente. En primer lugar, se ha realizado una normaliza-

ción de los datos como paso previo a la obtención del CPCA. Después se ha estimado el orden agregado del CPCA con los datos euclídeos, como se explicó en la sección 2.4.1, con el orden resultante se ha realizado la transformación de los datos de partida (euclídeos) en datos circulares, tal y como se indica en la sección 2.4.2. Por último, con estos datos circulares se ha estimado el orden agregado con las cuatro variantes expuestas en la Tabla 1 para el método de Hodge.

Se han elaborado gráficos para ilustrar el rendimiento de cada uno de los métodos evaluados para los diferentes tejidos y genes. En ellos se representan los valores euclídeos de expresión génica ordenados según el orden agregado estimado para cada método. También se proporcionan los tres primeros eigengenes, ordenados respecto al orden correspondiente en cada caso. Estos eigengenes son patrones de expresión característicos, que abarcan los perfiles de expresión global observados en los datos [10].

En estos gráficos se ha ajustado un modelo cosinor (sección 2.5.1) y se proporcionan los valores del error cuadrático medio individuales para cada gen (MSE_j) y el promedio global (MSE), estos últimos ilustrados en la Tabla 2 serán de utilidad para las subsecciones 3.1, 3.2 y 3.3.

Método Agregación	*MSE (Pulmón)*	*MSE (Sangre)*	*MSE (Músculo)*
CPCA	0.084	0.063	0.0747
Hodge POS	0.1078	0.0858	0.0993
Hodge COS	0.1069	0.0756	0.0914
Hodge AVE	0.0964	0.0893	0.1176
Hodge E3	0.0967	0.086	0.1176

Tabla 2: *MSE* para cada método y tejido

3.1. Tejido pulmonar

En las figuras 6 a 10 se muestra el ajuste realizado para los *core clock genes* de acuerdo al orden estimado por cada uno de los método de agregación estudiados para el tejido del pulmón. En la Tabla 2 podemos observar como en este tejido el mejor ajuste global es el que se obtiene con el CPCA, seguido por las variantes AVE y E3, cuyos resultados son similares.

Comparando los resultados en los distintos órdenes estimados, se observa que en el tejido pulmonar, los genes PER2 y CRY2 son los presentan un patrón rítmico más marcado. En ambos genes CPCA es el método que proporciona valores más bajos del *MSE*, destacando la de CRY2 (con un *MSE* por debajo de 0.05). La ritmicidad de estos dos genes también se observa de forma clara cuando la estimación del orden se obtiene a partir de los métodos E3 y AVE, siendo las variantes POS y COS, en las que el patrón rítmico subyacente no es tan evidente. En general, si nos fijamos en los eigengenes obtenidos para el tejido pulmonar de acuerdo a los distintos órdenes estimados, se observa que el método COS es el que presenta una mayor variabilidad, siendo CPCA el método en el que la estructura circular subyacente es más evidente en base al patrón de expresión de los dos primeros eigengenes.

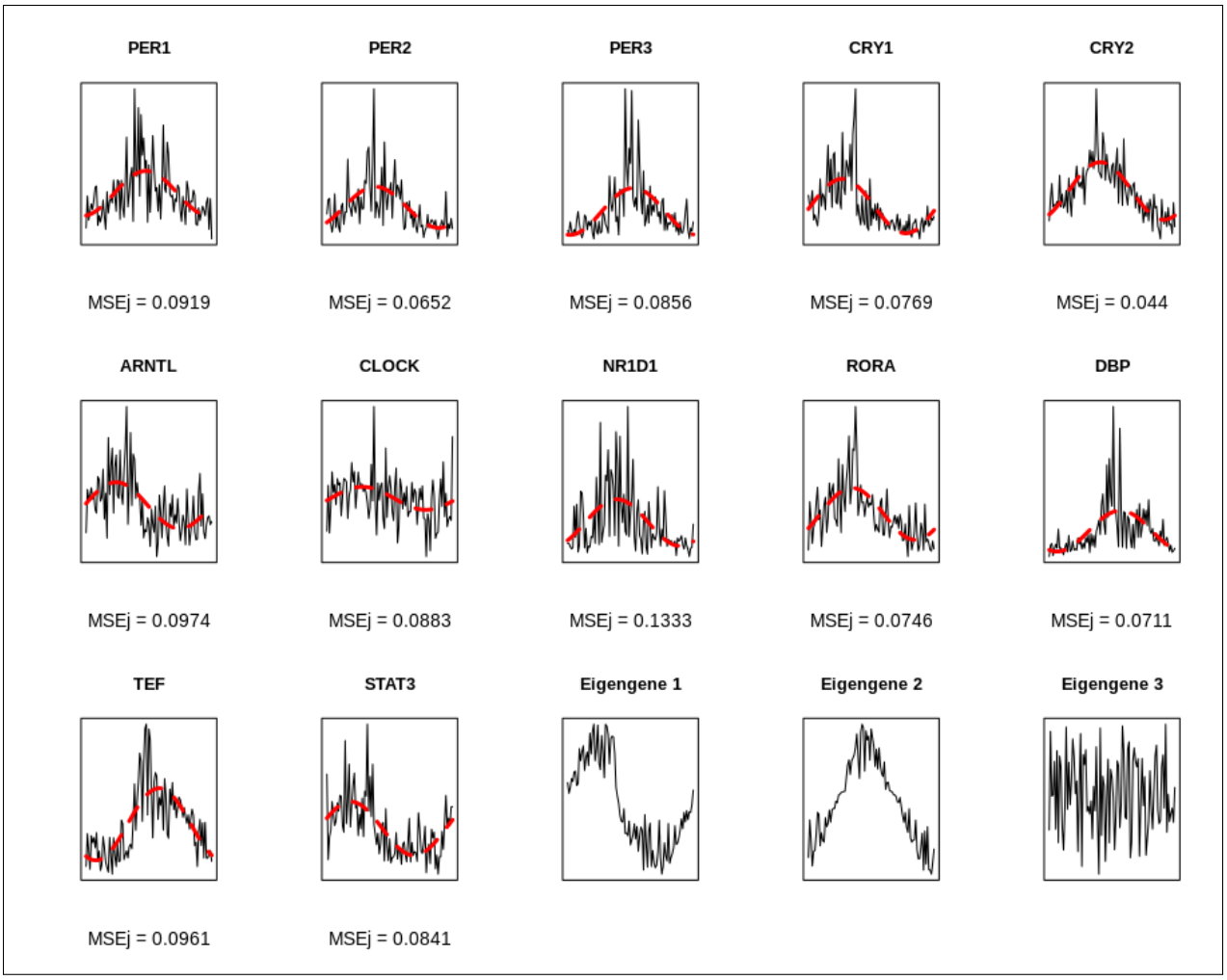

Figura 6: Resultados para el orden estimado con CPCA para el tejido pulmonar

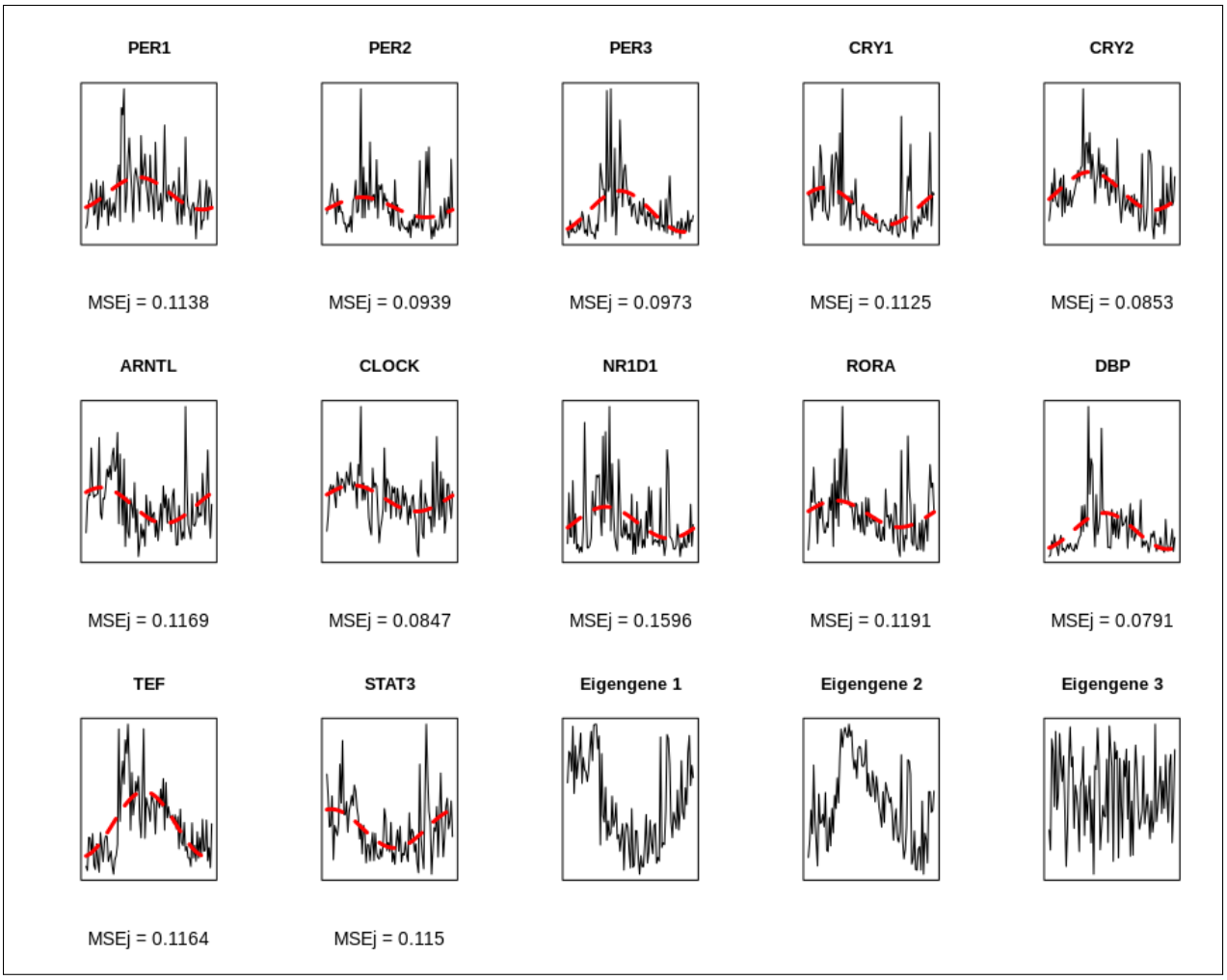

Figura 7: Resultados para el orden estimado con HODGE POS para el tejido pulmonar

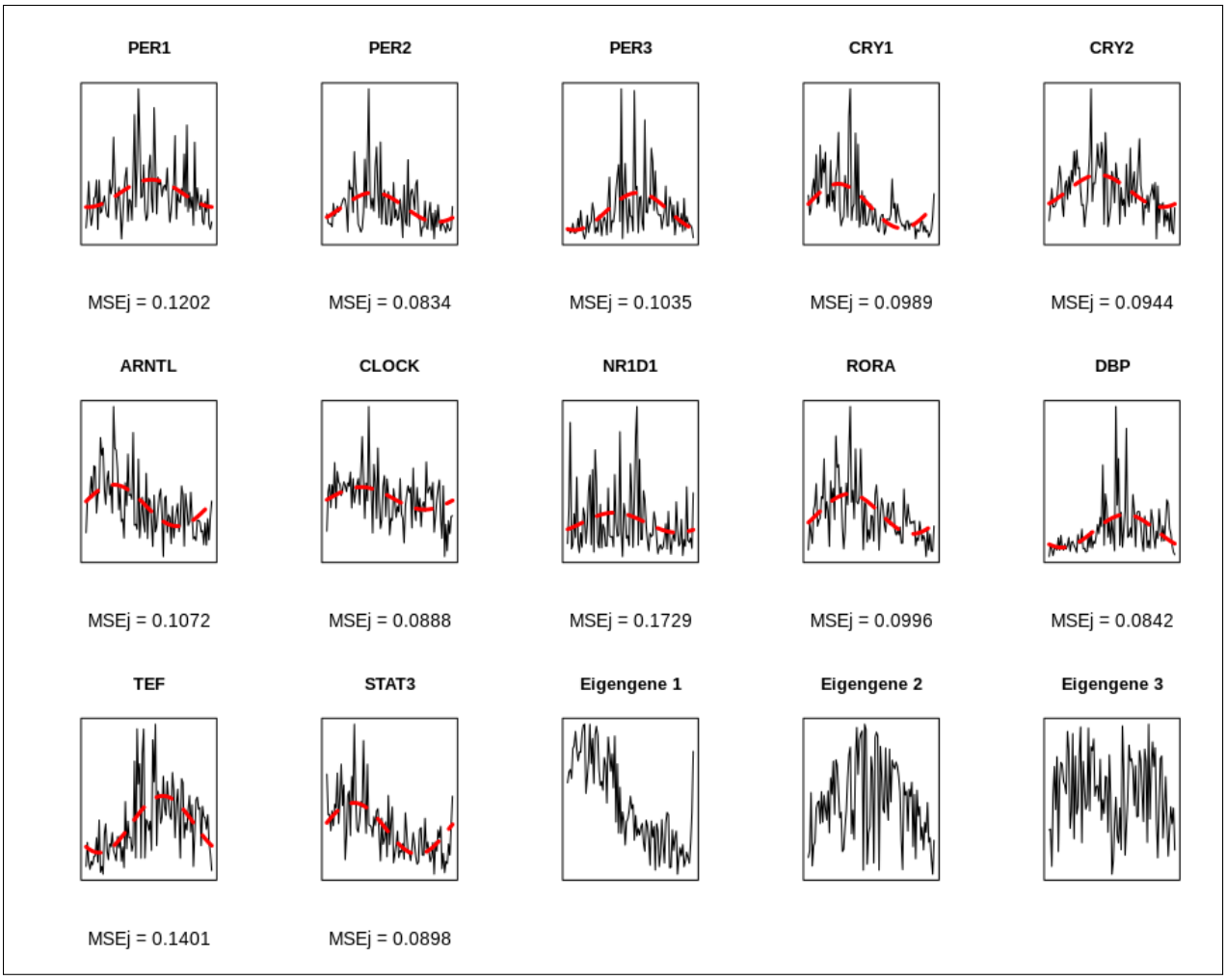

Figura 8: Resultados para el orden estimado con HODGE COS para el tejido pulmonar

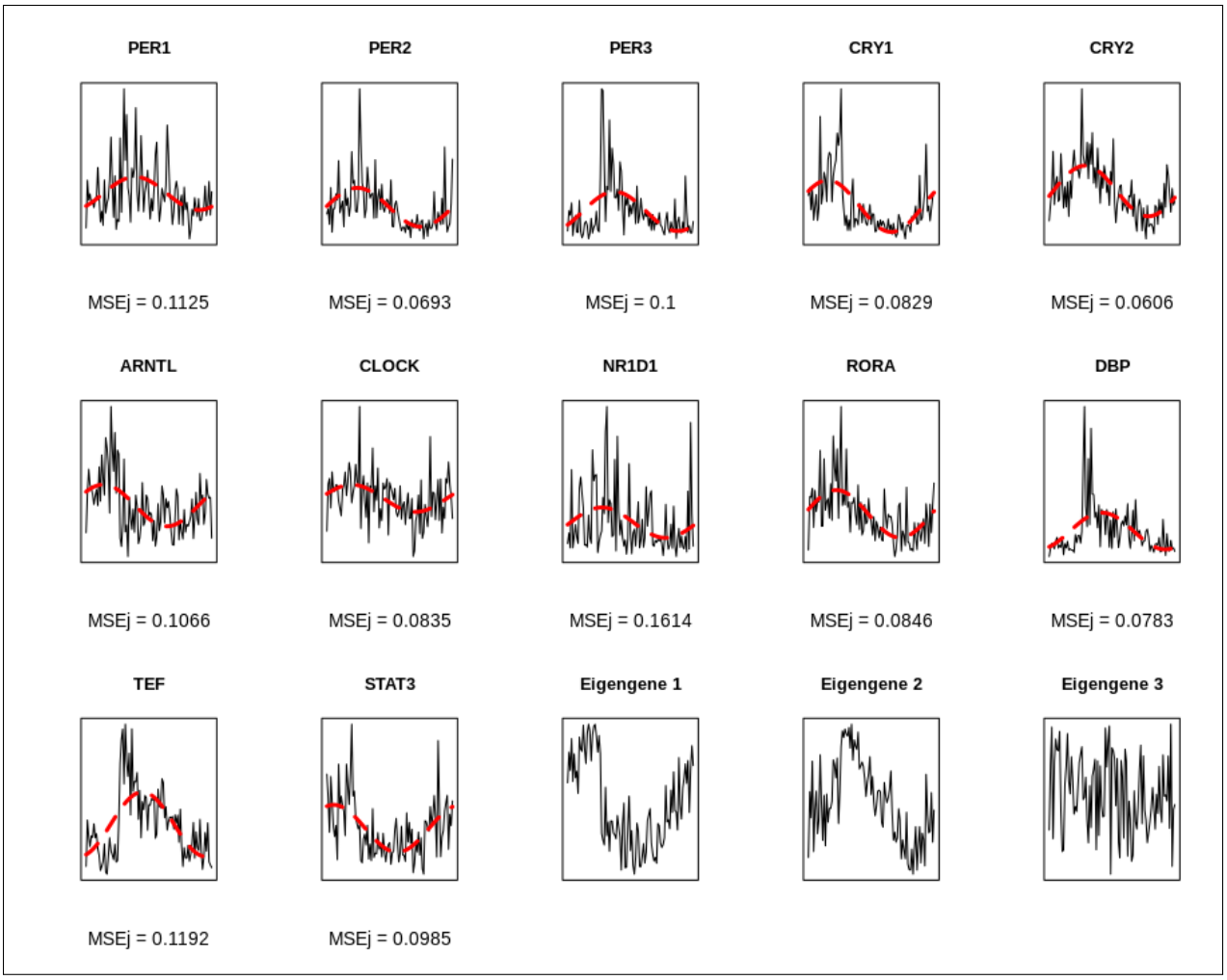

Figura 9: Resultados para el orden estimado con HODGE AVE para el tejido pulmonar

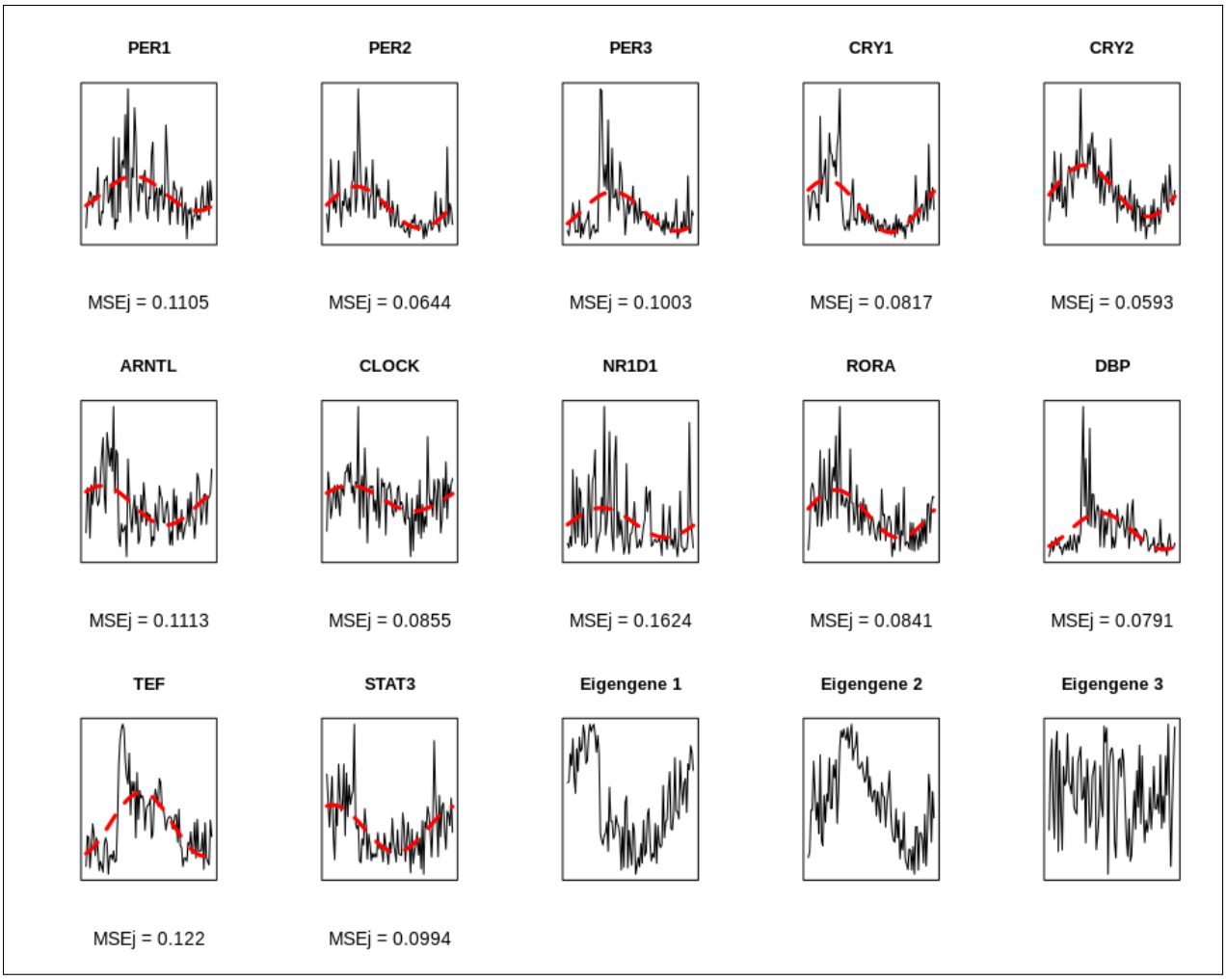

Figura 10: Resultados para el orden estimado con HODGE E3 para el tejido pulmonar

3.2. Tejido sanguíneo

En las figuras 11 a 15 se muestra el ajuste realizado para los *genes core clock* de acuerdo al estimado por cada uno de los método de agregación estudiados para el tejido sanguíneo. En base a las medidas establecidas, se observa que para este tejido el método más eficiente es el CPCA, destacando entre las alternativas del método de Hodge estudiadas la variante COS, propocionando mejores resultados, en términos de *MSE* que el resto de alternativas, véase Tabla 2.

En este tejido, por lo general, parece que la expresión génica de los *genes core clock* estudiados tiene una componente rítmica más fuerte que el tejido pulmonar estudiado en 3.1. Podemos destacar la ritmicidad del gen RORA, del gen NR1D1 y del gen CRY2. En el CPCA, de nuevo, es donde el patrón rítmico se aprecia con mayor claridad, con un *MSE* de en torno al 0.05. Para estos genes, salvo la alternativa de Hodge COS, el resto no resultan muy razonables en este tejido. La alternativa COS, sin embargo, obtiene ajustes satisfactorios rivalizando, en general, con las del CPCA. Estos comentarios se reflejan de forma general en los patrones de los dos primeros eigengenes para los métodos de agregación de órdenes propuestos.

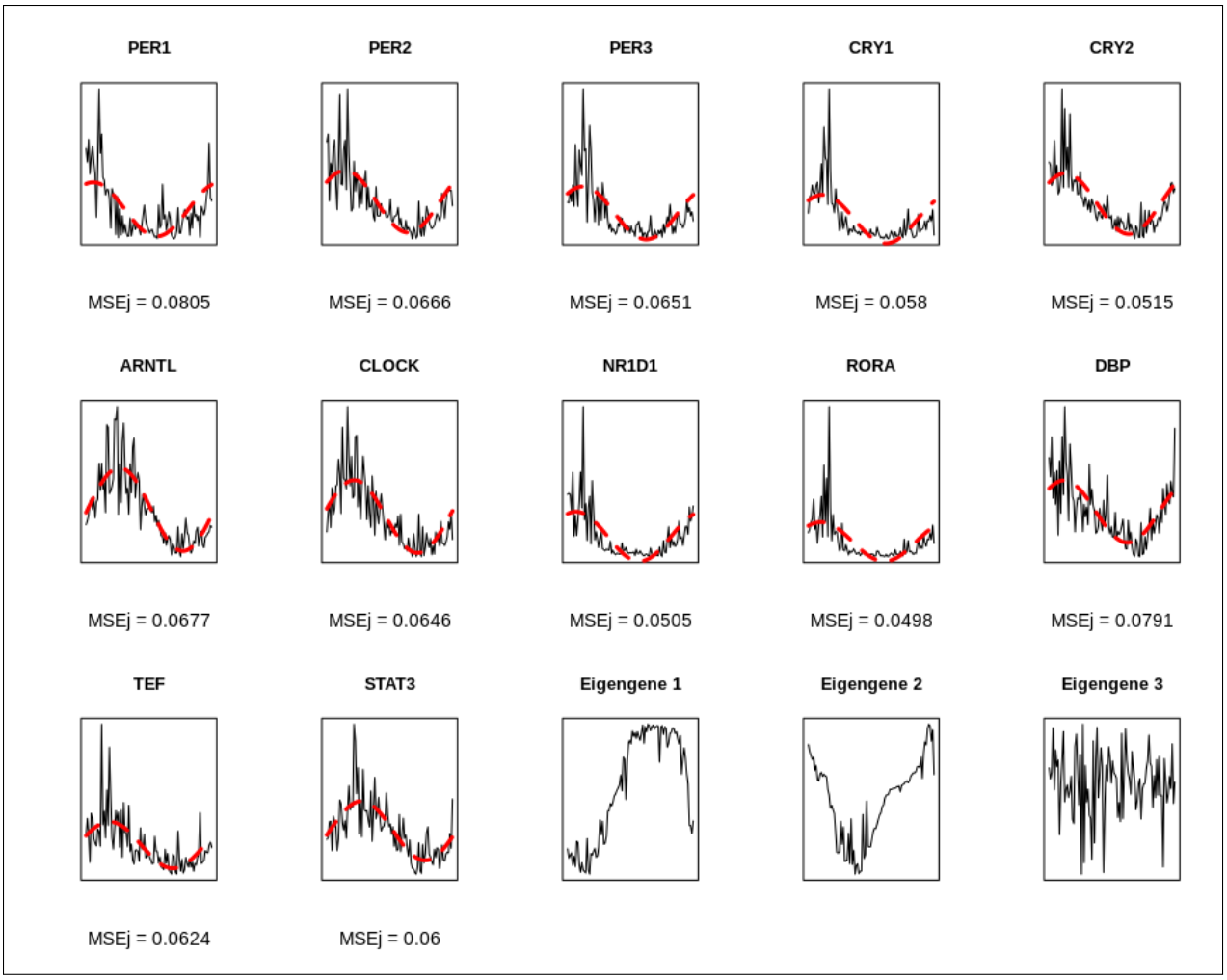

Figura 11: Resultados para el orden estimado con CPCA para el tejido sanguíneo

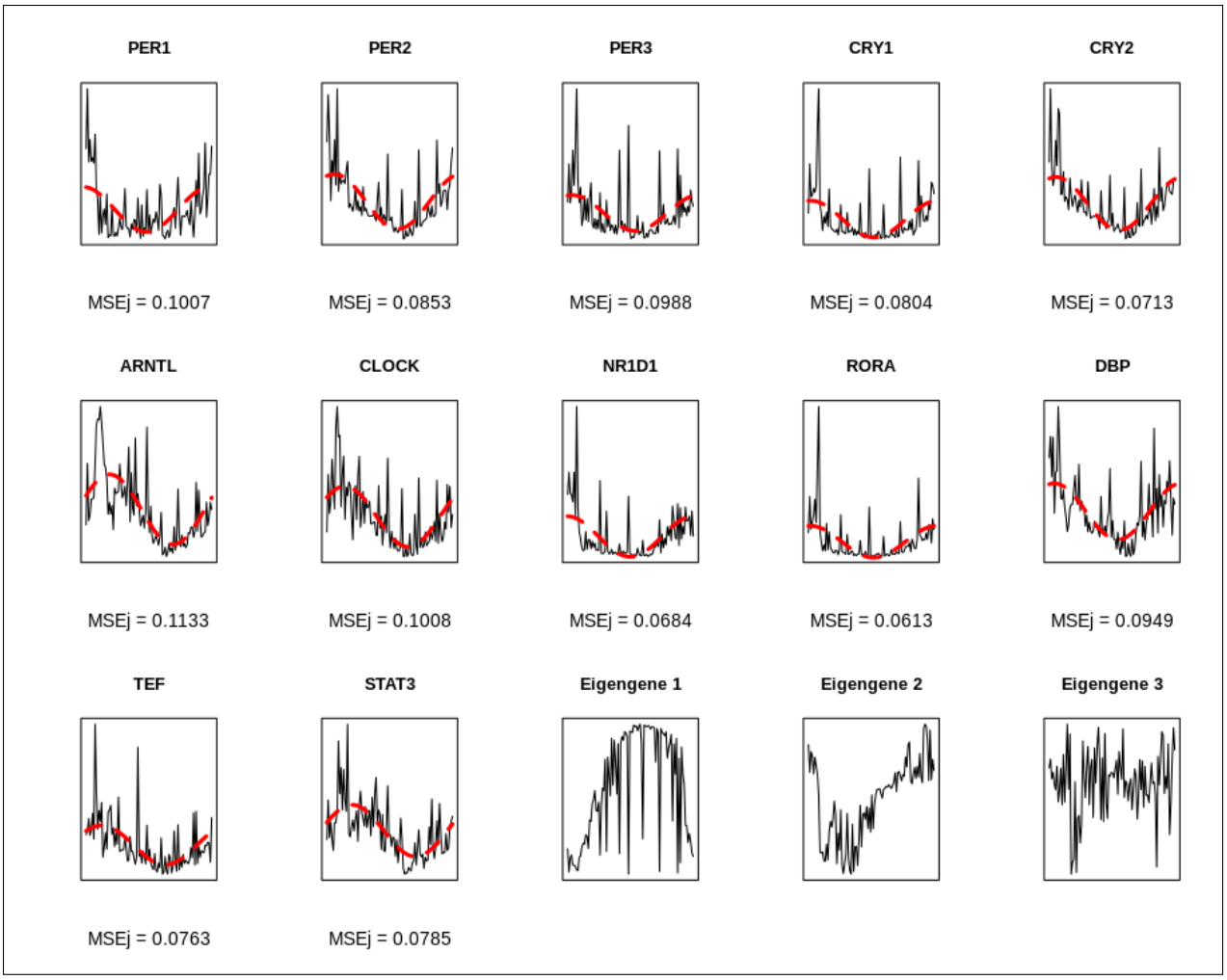

Figura 12: Resultados para el orden estimado con HODGE POS para el tejido sanguíneo

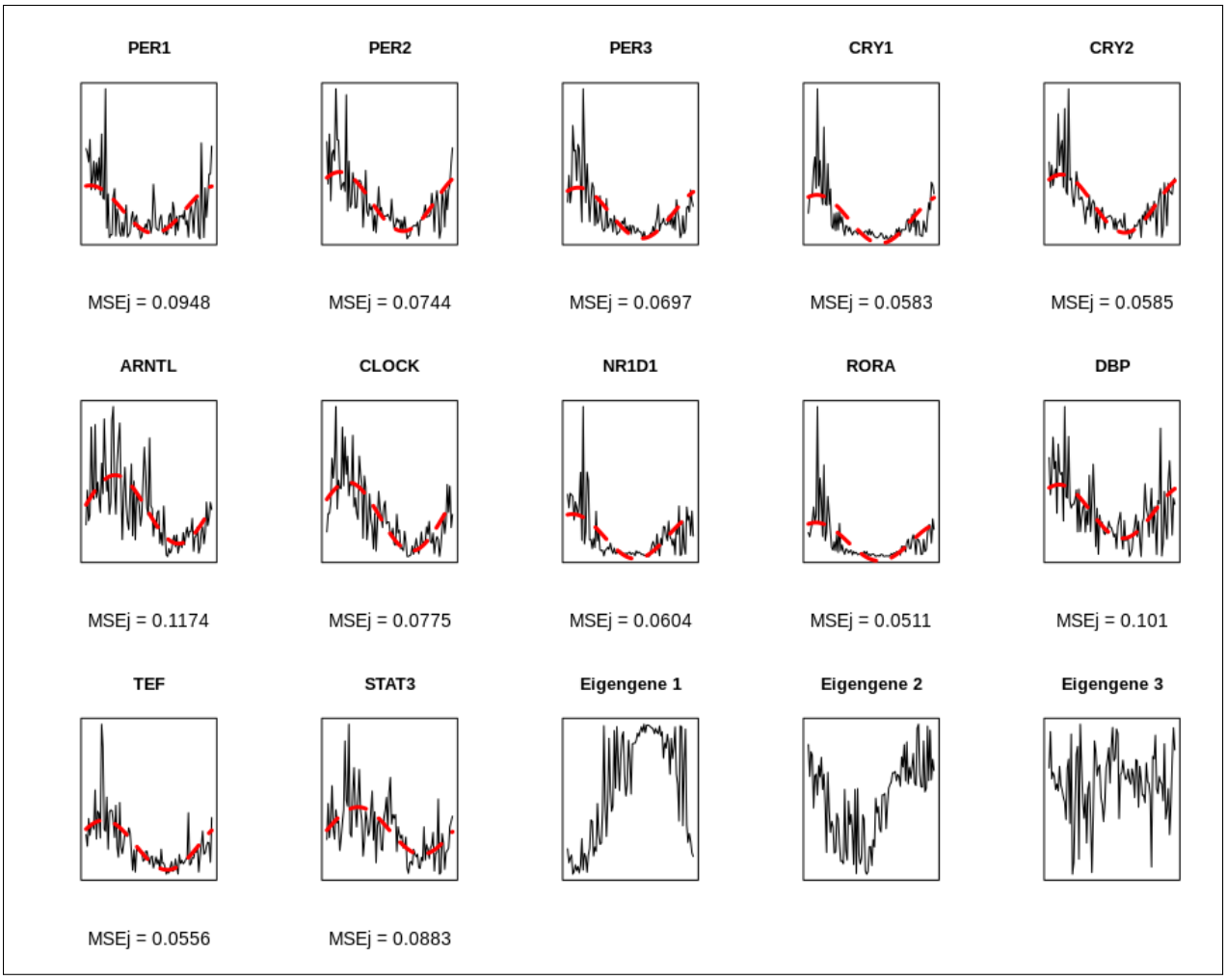

Figura 13: Resultados para el orden estimado con HODGE COS para el tejido sanguíneo

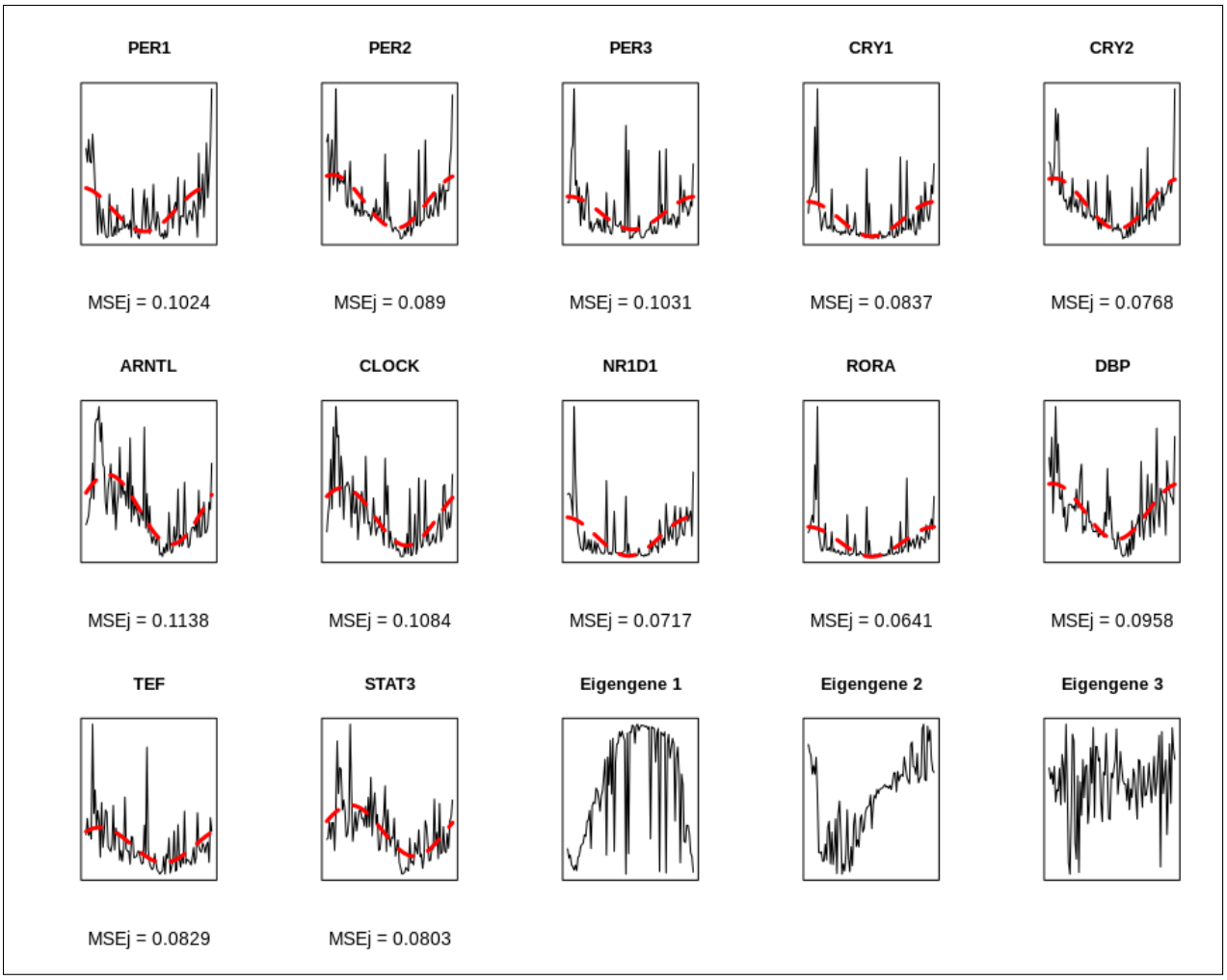

Figura 14: Resultados para el orden estimado con HODGE AVE para el tejido sanguíneo

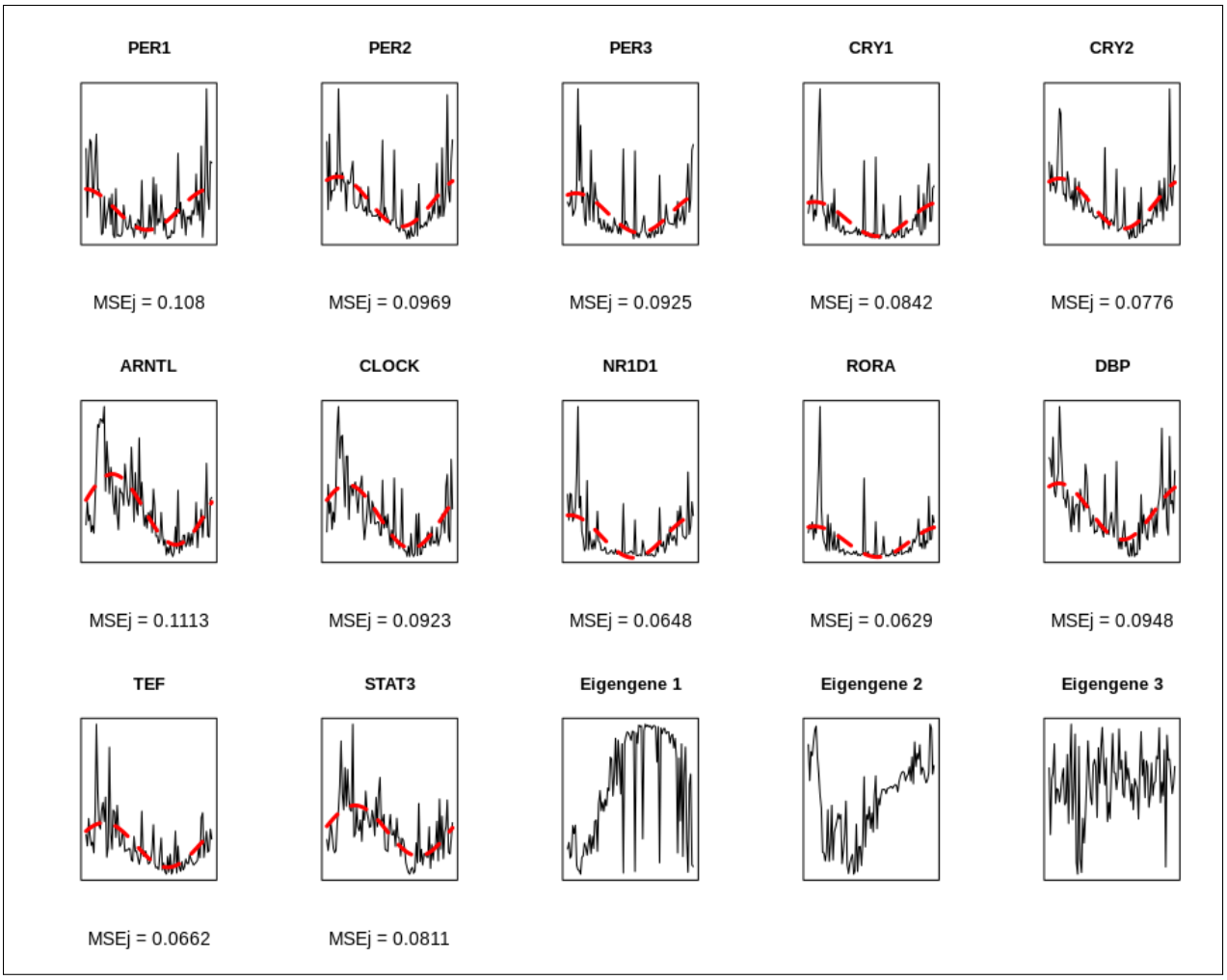

Figura 15: Resultados para el orden estimado con HODGE E3 para el tejido sanguíneo

3.3. Tejido muscular

En las figuras 16 a 20 se muestra el ajuste realizado para los *core clock genes* de acuerdo al estimado por cada uno de los método de agregación estudiados para el tejido muscular. En este tejido destaca por su rendimiento con las medidas propuestas, de nuevo, el CPCA, como podemos ver en la Tabla 2. Le siguen las variantes Hodge COS y POS, mientras que las alternativas AVE y E3 presentan un rendimiento significativamente inferior, con un *MSE* cercano al 0.12.

En el tejido muscular destacan a lo largo de los diferentes métodos los genes PER3 y TEF por su ritmicidad. El *MSE* para estos genes logrado por el CPCA se encuentra por debajo del 0.06, le sigue la alternativa COS cuyos valores del *MSE* para PER3 y TEF se encuentran en torno al 0.09. Este tejido nos pone en evidencia cómo varía el comportamiento particular de los genes en las distintas alternativas. Por ejemplo, las alternativas AVE y E3 pese a presentar un comportamiento global peor que la variante POS funcionan mejor para estos genes. Las alternativas AVE y E3 tienen un *MSE* de en torno al 0.10 para estos genes. La variante POS, en cambio, tiene un *MSE* bastante elevado para el caso particular del gen PER3 (0.1347) y un MSE, aunque más razonable, también peor que el de las variantes AVE y E3 para el gen TEF (0.1041).

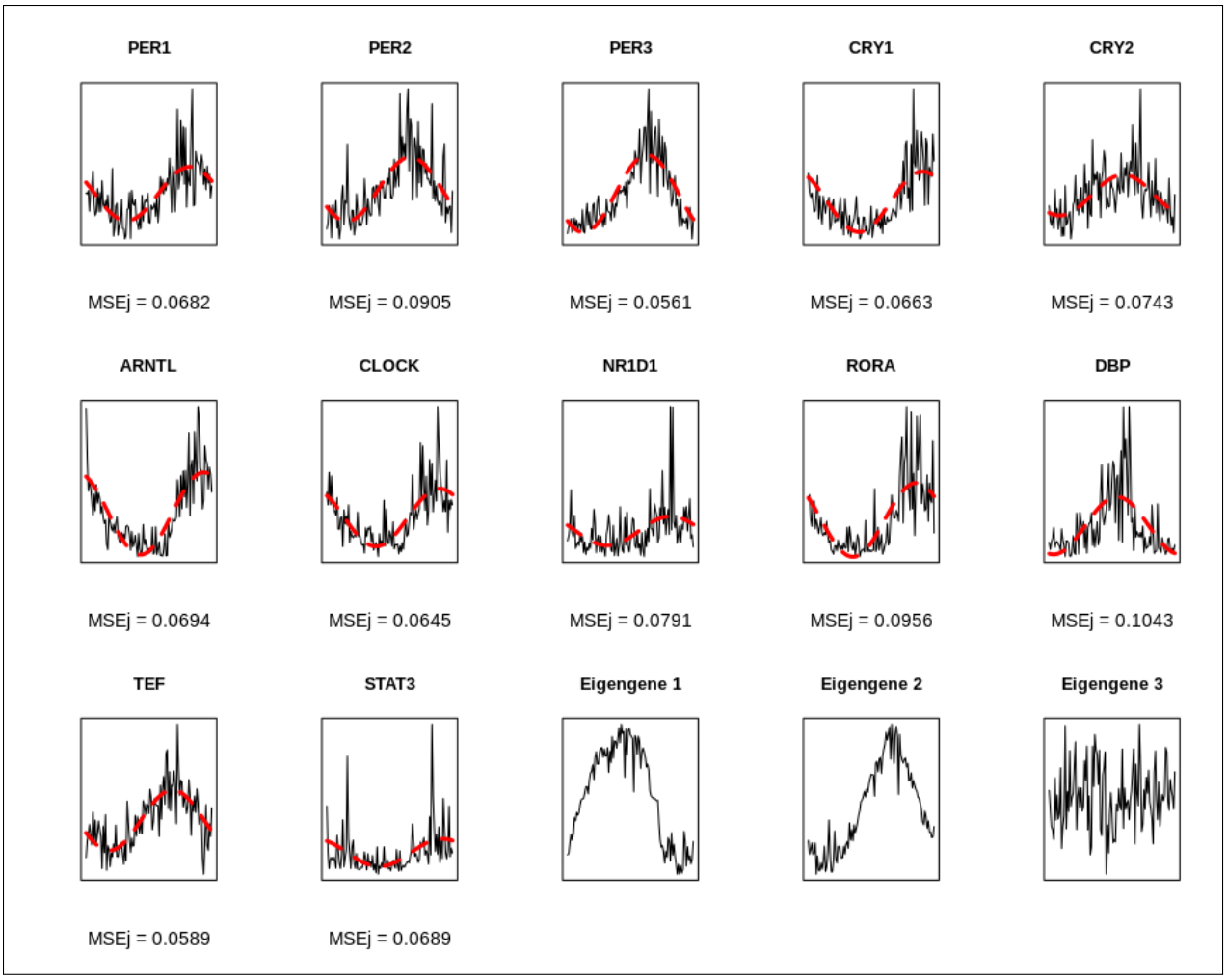

Figura 16: Resultados para el orden estimado con CPCA para el tejido muscular

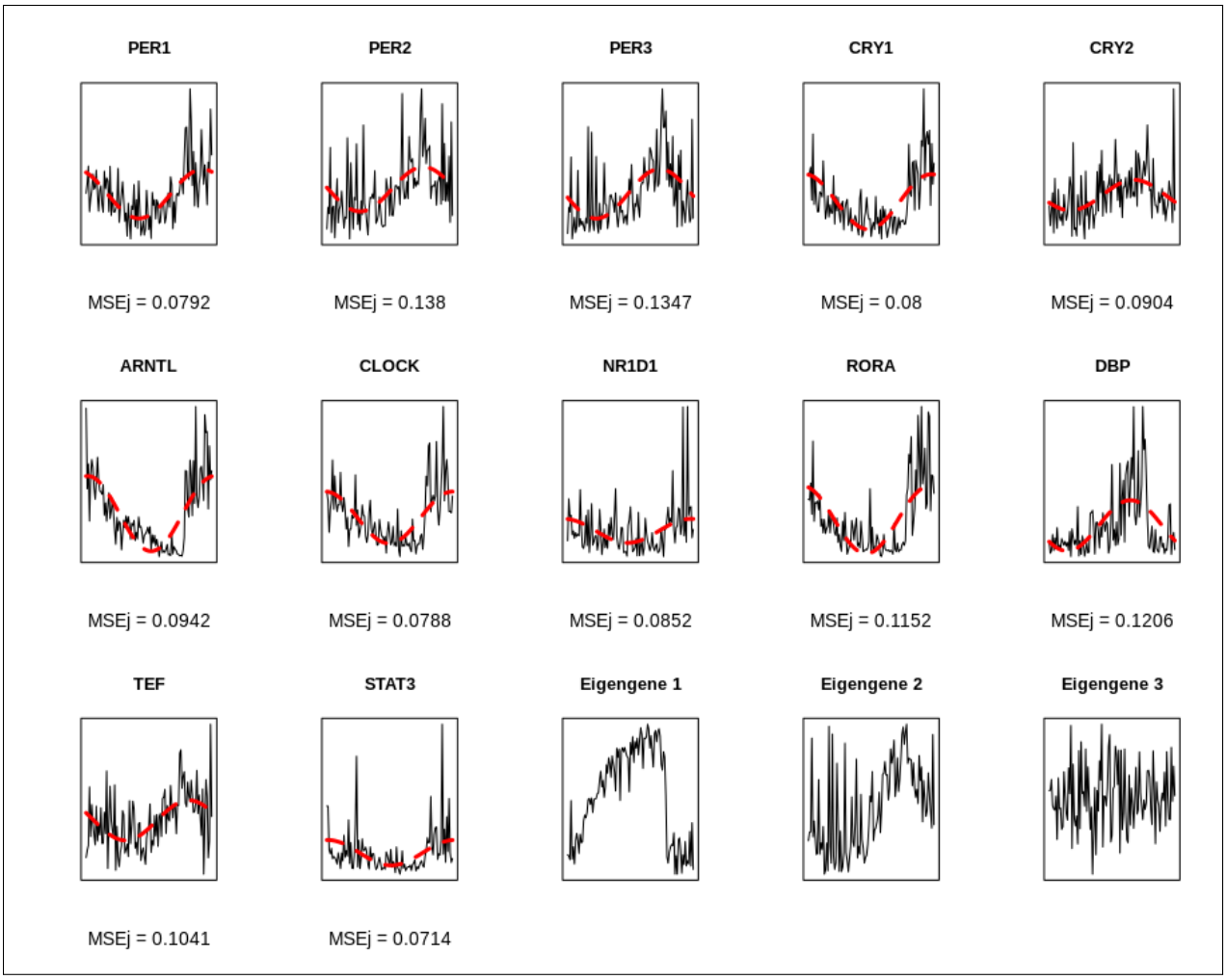

Figura 17: Resultados para el orden estimado con HODGE POS para el tejido muscular

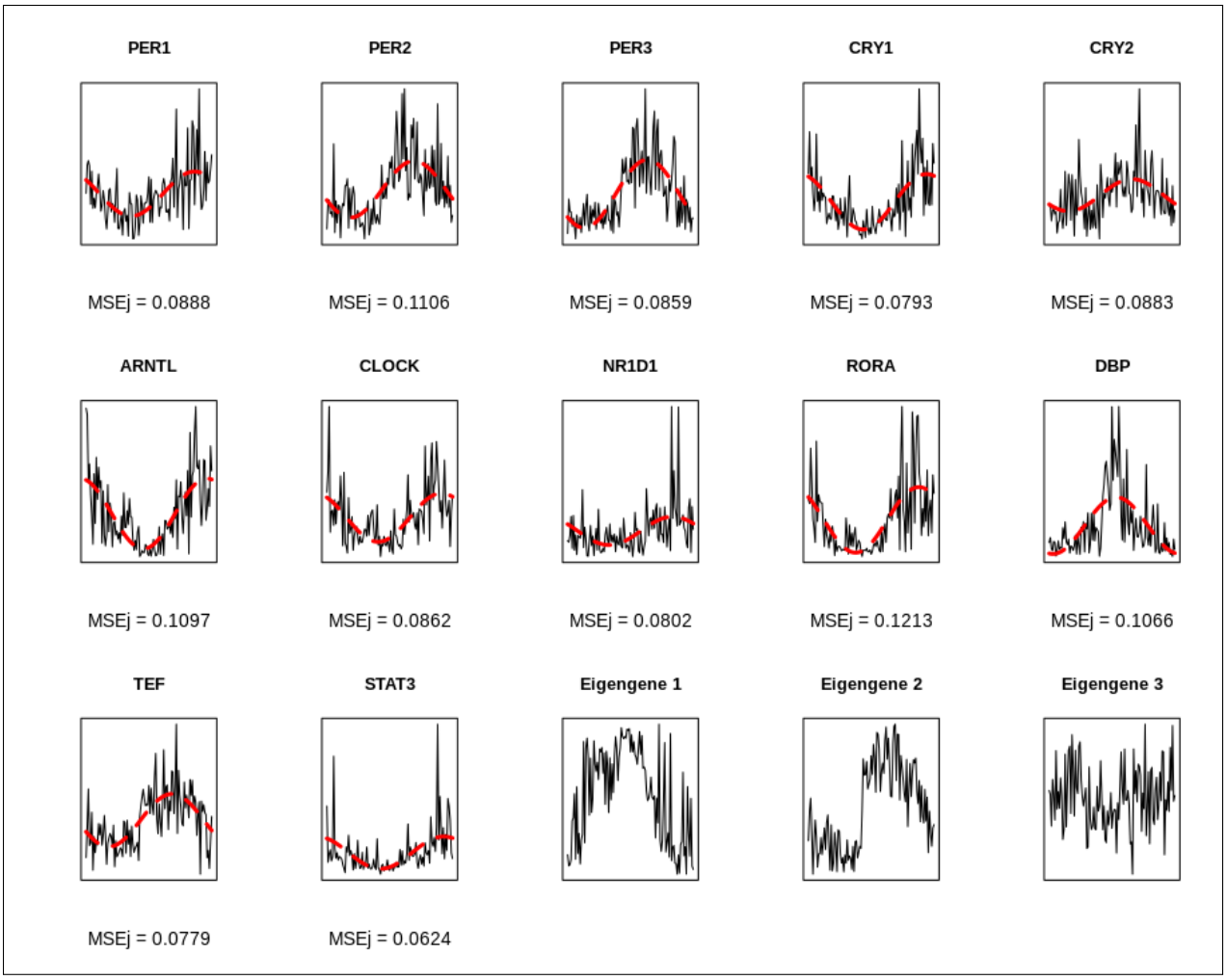

Figura 18: Resultados para el orden estimado con HODGE COS para el tejido muscular

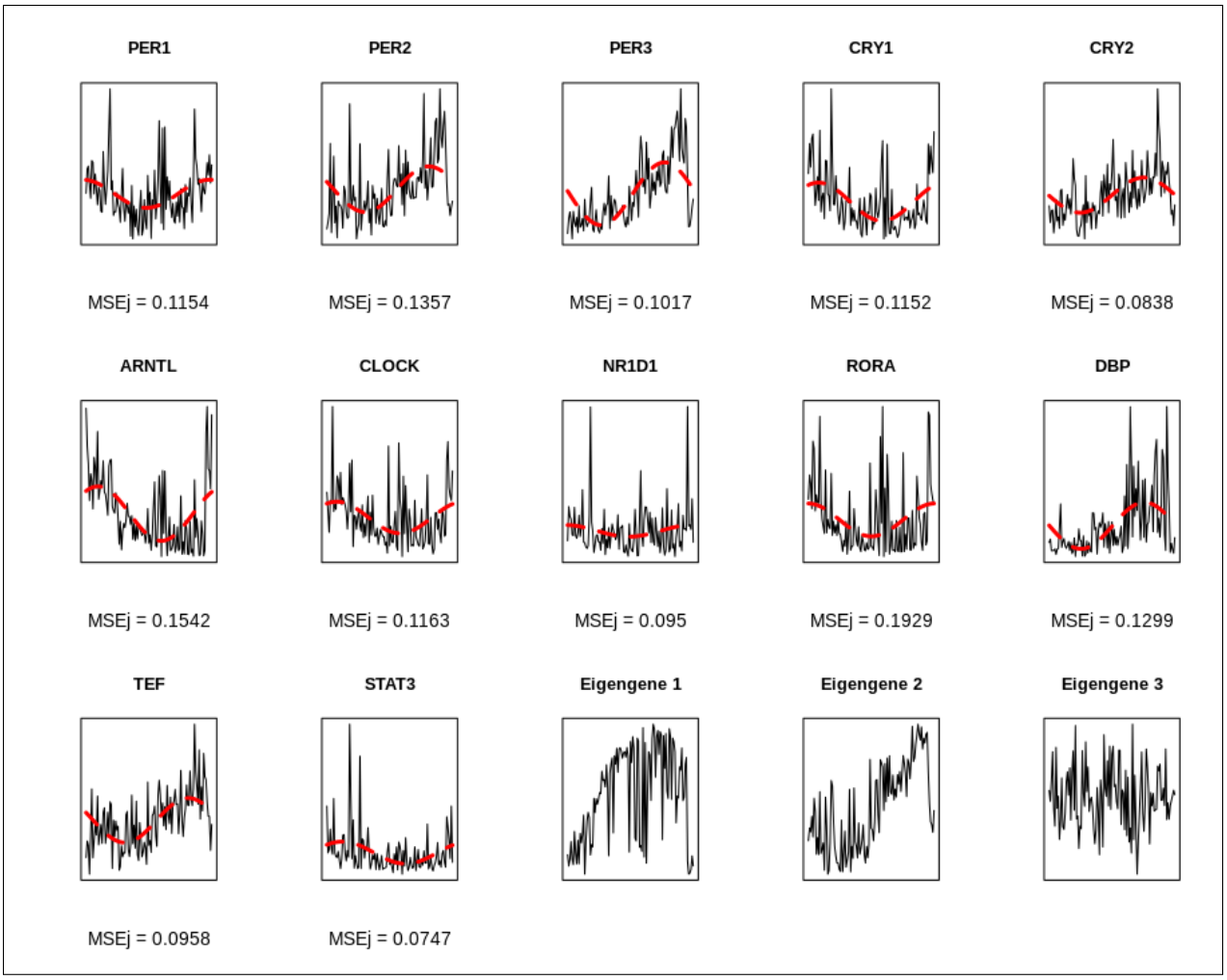

Figura 19: Resultados para el orden estimado con HODGE AVE para el tejido muscular

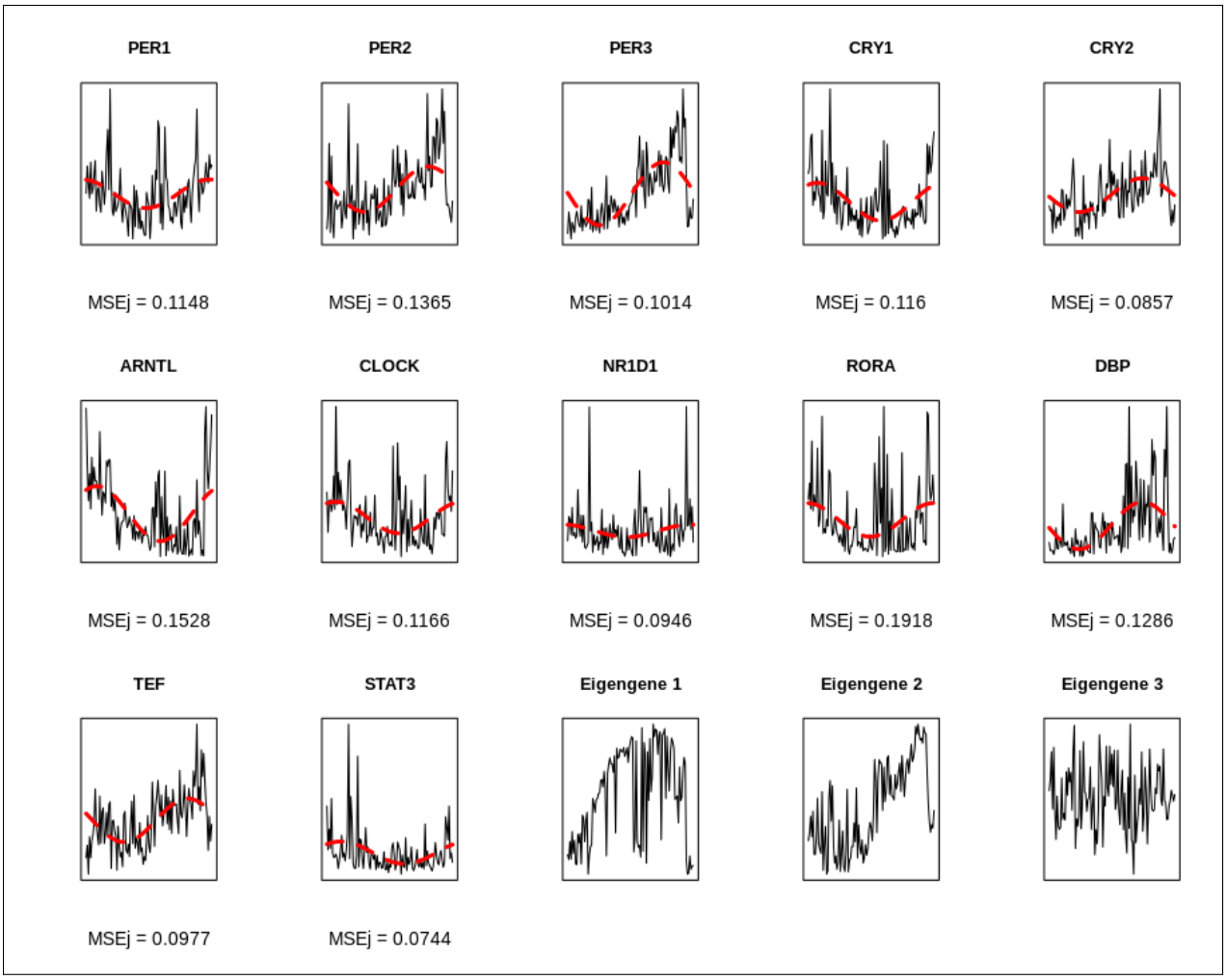

Figura 20: Resultados para el orden estimado con HODGE E3 para el tejido muscular

4. Conclusión

Tras haber analizado los resultados obtenidos, parece evidente que para este conjunto de datos y en todos los tejidos evaluados se comporta mejor el orden estimado por el CPCA, produciendo ordenaciones de los genes con una mayor ritmicidad.

Entre las alternativas para la definición de los λ_{ihk}^{j}, la que mejor parece funcionar es COS, aunque para el tejido pulmonar tanto AVE como E3 eran significativamente mejores. Las diferencias, en todo caso, no son tan destacadas como lo es la existente entre CPCA y el resto.

La transformación de los datos euclídeos a circulares y la posterior estimación del orden agregado con el método de Hodge parece producir órdenes circulares agregados razonables, en los que se puede percibir una ritmicidad evidente.

No obstante, la metodología de Hodge es muy flexible y hay diferentes elementos que pueden alterarse para obtener procedimientos mas eficaces. En particular, las hipermatrices se pueden definir a partir de la información de los datos euclídeos directamente usando otro tipo de transformación o medidas de distancia entre tripletas de observaciones.

Una cuestión muy importante que dejamos también para futuras investigaciones es la robustez de las diferentes propuestas, una cuestión especialmente relevante en los datos analizados en este estudio.

Referencias

[1] Barragán S, Rueda C, Fernández MA. Circular Order Aggregation and its Application to Cell-cycle Genes Expressions, 2016.

[2] Larriba Y, Rueda C, Fernández MA, Peddada SD. Order restricted inference in chronobiology, 2019.

[3] T. Robertson, F.Wright, and R. Dykstra. Order Restricted Statistical Inference. John Wiley & Sons, 1988.

[4] M. Hughes, J. Hogenesch, and K. Kornacker. JTK CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. Journal of Biological Rhythms, 25(5):372-380, 2010. doi: 10.1177/0748730410379711.

[5] Larriba Y, Scheer F, Saxena R, Rueda C, Mason IC. Human circadian atlas from GTEx data base, 2019.

[6] Cornelissen Germaine. Cosinor-based rhythmometry, 2014

[7] Zhang, R., Lahens, N., Ballance, H., Hughes, M. & Hogenesch, J. A circadian gene expression atlas in mammals: Implications for biology and medicine. Proceedings of the National Academy of Sciences of the United States of America 111, (2014).

[8] Mure, L. S. et al. Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. Science 359 (2018).

[9] Ruben, M. D. et al. A database of tissue-specific rhythmically expressed human genes (2018).

[10] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA. 2000;97:10101–10106.