Un estudio sobre el análisis arquetípico

**Resumen: El análisis arquetípico (AA) fue propuesto originalmente en 1994 por Adele Cutler y Leo Breiman como un procedimiento computacional para extraer los aspectos distintivos denominados arquetipos en las observaciones, aproximando cada registro observacional como una mezcla (es decir, una combinación convexa) de estos arquetipos. El AA proporciona así representaciones sencillas, interpretables y explicables para la extracción de características y la reducción de la dimensionalidad, lo que facilita la comprensión de la estructura de los datos de alta dimensionalidad con amplias aplicaciones en todas las ciencias. Sin embargo, el AA también se enfrenta a retos, sobre todo porque el problema de optimización asociado no es convexo. Este estudio ofrece a los investigadores y profesionales de la minería de datos una visión general de las metodologías y oportunidades que ofrece el AA, analizando las numerosas aplicaciones del AA en campos científicos dispares, así como las mejores prácticas para modelar datos utilizando el AA y sus limitaciones. El estudio concluye explicando importantes direcciones de investigación futuras relacionadas con el AA. Términos del índice: análisis arquetípico, agrupamiento, envolvente convexa, ciencia de datos, factorización de matrices, estudio, aprendizaje no supervisado.**

**Términos del índice: análisis arquetípico, agrupamiento, envolvente convexa, ciencia de datos, factorización de matrices, encuesta, aprendizaje no supervisado.**

I. INTRODUCCIÓN

El ANÁLISIS ARQUETÍPICO (AA), propuesto originalmente en [1], es un procedimiento computacional que nos permite extraer estos aspectos distintivos, así como caracterizar las observaciones como combinaciones convexas de dichos aspectos. El AA asume que los arquetipos (es decir, los aspectos distintivos o formas puras) pueden definirse en términos de combinaciones convexas de la representación colectiva definida por todos los puntos de datos. El AA proporciona un marco computacional sencillo pero interesante para la reducción de la dimensionalidad y la extracción de características en el aprendizaje automático con representaciones explicables que son fácilmente interpretables, ya que pueden considerarse formas puras idealizadas de observaciones de datos. Los arquetipos trascienden la filosofía, la psicología y las ciencias en general. La palabra arquetipo, que deriva del sustantivo griego ἀρχέτυπον, se introdujo en el idioma inglés en la década de 1540 con el significado de «modelo, primera forma, patrón original a partir del cual se hacen copias», siendo el adjetivo arquetípico «perteneciente o relativo a un arquetipo». En psicología, Carl Gustav Jung utilizó este concepto para definir los arquetipos como «formas o imágenes de naturaleza colectiva»2. Los arquetipos están, por lo tanto, estrechamente relacionados con las formas puras de Platón, utilizadas para definir la verdadera naturaleza de las cosas3.

AA surgió de los retos asociados con la simulación de la producción de ozono en la atmósfera inferior, un proyecto respaldado por la Agencia de Protección Ambiental de los Estados Unidos.

Aleix Alcacer and Irene Epifanio are with Jaume I University, Spain; Sebastian Mair is with Linköping University, Sweden; and Morten Mørup is with Technical University of Denmark, Denmark. The author list is alphabetically sorted and all authors contributed equally. Corresponding author: Irene Epifanio, [epifanio@mat.uji.es1https://www.etymonline.com/search?q=archetype](mailto:epifanio@mat.uji.es1https://www.etymonline.com/search?q=archetype))

2https://www.etymonline.com/word/archetypal#etymonline\_v\_264923https://marcomasi.substack.com/p/platonic-forms-ideas-and-archetypes

Los complejos modelos informáticos utilizados para este fin contenían cientos de ecuaciones químicas y, por lo general, requerían un tiempo de cálculo considerable, a menudo 24 horas para simular solo 24 horas en tiempo real. Como resultado, los investigadores solo podían modelar unos pocos días de datos en cada proyecto. Esta limitación puso de manifiesto la necesidad de seleccionar datos que representaran unos pocos días «prototípicos», lo que finalmente inspiró el concepto de arquetipos como medio para capturar y analizar eficazmente los patrones clave dentro del conjunto de datos. Este problema fundamental fue analizado en el artículo pionero de [1].

Desde entonces, el AA se ha utilizado en una amplia variedad de campos diferentes, motivado por su capacidad para caracterizar propiedades relevantes y destacadas observadas en diferentes ámbitos. Así, en los sistemas biológicos, el AA puede estar motivado por compensaciones evolutivas y la optimalidad de Pareto, que se ha descubierto que hacen que los datos residan dentro de un politopo que define las compensaciones entre organismos definidos de forma óptima (similares a formas puras) para diversas tareas adecuadas para la supervivencia [2]. En química, el AA puede motivarse por la observación de que las mediciones de muestras de un espacio confinado definen fracciones de concentración en términos de los componentes [3], y en geociencias, que las imágenes hiperespectrales pueden representarse en términos de espectros distintos (los llamados miembros finales), de modo que el espectro registrado puede definirse como combinaciones convexas de estas formas espectrales puras[4]. Para extraer estos componentes puros, las metodologías de extracción de miembros extremos más destacadas se han centrado en intentar minimizar el volumen de la representación de los datos [5]. A diferencia de los enfoques convencionales de minimización de volumen, que se esfuerzan por encapsular la variedad de datos desde el exterior, AA puede considerarse un enfoque que encapsula la variedad de datos de manera similar desde el interior. Por último, en la ciencia de datos en general, la agrupación se utiliza ampliamente para identificar prototipos, pero también es interesante la identificación de características distintivas en los datos y cómo cada observación puede describirse en términos de estas características de manera fácil de explicar [6]. El AA también se relaciona con los enfoques de agrupamiento difuso, en los que las observaciones no se asignan de forma rígida, sino que pertenecen a múltiples grupos de forma flexible [7]-[9], lo que proporciona una representación flexible de la pertenencia a los grupos en términos de un continuo.

La figura 1 muestra un ejemplo de AA calculado sobre el subconjunto del conjunto de datos de dígitos escritos a mano MNIST [10] restringido al dígito 9 con tres arquetipos. La parte izquierda muestra las ponderaciones de mezcla de los 9, mientras que la parte derecha muestra los dígitos manuscritos reales. Se puede observar que los arquetipos extraídos de los 9 manuscritos son «estrechos y rectos», «estrechos e inclinados» y «anchos y rectos». Matemáticamente, las imágenes de 28×28 píxeles que representan los dígitos se encuentran en un espacio de 784 dimensiones en el que encajamos un triángulo (bidimensional) utilizando AA. Las esquinas del triángulo son los arquetipos, y todos los 9 se proyectan en este triángulo. En la parte izquierda se puede observar que tanto los arquetipos como el lado entre ellos atraen más proyecciones que el centro. Además, se proyectan menos puntos en el lado entre los arquetipos «estrecho e inclinado» y «ancho y recto».

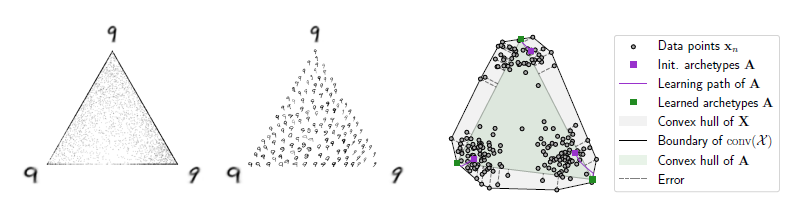


Fig. 1. AA calculado sobre el subconjunto del conjunto de datos de dígitos escritos a mano MNIST que muestra únicamente el dígito 9 con tres arquetipos. La parte izquierda muestra las ponderaciones de mezcla de los nueves, mientras que la parte derecha muestra los dígitos escritos a mano reales.

El AA puede facilitar así la comprensión de la estructura de los datos de alta dimensión en general. Sin embargo, el AA también se enfrenta a retos en sus aplicaciones prácticas. En primer lugar, el problema de optimización asociado no es convexo y no hay garantías de identificar correctamente las formas puras, incluso si están presentes como combinaciones convexas de los datos. Además, no está claro cómo seleccionar el número de arquetipos y, en las aplicaciones prácticas, los valores atípicos pueden influir de manera indeseable en las representaciones aprendidas. Además, es posible que los arquetipos solo sean adecuados dentro de una variedad no lineal, o que los datos se desvíen de la suposición de normalidad gaussiana que se suele asumir en los enfoques convencionales de AA [11].

El objetivo de este estudio es proporcionar a los investigadores y profesionales de la minería de datos en general una visión general de los métodos y oportunidades que ofrece el AA, incluyendo una visión general de los numerosos y dispares campos en los que se ha aplicado el AA. Además, este estudio destacará las mejores prácticas en el modelado de datos utilizando el AA, así como las limitaciones importantes del AA y las futuras direcciones de la investigación para avanzar en el AA.

En aras de la transparencia, la reproducibilidad y la exhaustividad, y con el fin de reducir el sesgo en la selección de artículos, realizamos una búsqueda en la base de datos SCOPUS, la mayor colección en línea de trabajos académicos revisados por pares. Seleccionamos aquellos trabajos que citan el artículo seminal de [1] y que, al mismo tiempo, tienen como palabra clave «análisis arquetípico». Hasta el 15 de abril de 2025 se encontraron un total de 152 documentos. Estos se incorporaron a las diferentes secciones de este manuscrito, junto con otros registros identificados por su cita en dichos documentos.

II. DESCRIPCIÓN TEÓRICO/TÉCNICA:

Análisis arquetípico (AA). Sea un conjunto de datos compuesto por N puntos de datos en M dimensiones y K el número de arquetipos. Esos puntos de datos se apilan como vectores de fila en una matriz de diseño . El objetivo es una descomposición de en un producto de una matriz de pesos y una matriz factorial . La idea del AA es representar cada punto de datos como una combinación convexa de K arquetipos , es decir,

Tenga en cuenta que los vectores contienen los pesos de las combinaciones convexas

Fig. 2. Ejemplo de AA con K = 3 arquetipos en datos bidimensionales ficticios. Tras inicializar los arquetipos (cuadrados morados), el procedimiento de optimización empuja los arquetipos (cuadrados verdes) para que se sitúen en el límite del casco convexo de los datos (véase el Teorema 1). El objetivo general es minimizar la suma de los errores de proyección (es decir, la suma de todas las líneas discontinuas).

que se apilan por filas en S y que A es la matriz de diseño de los arquetipos . Los arquetipos también se representan como combinaciones convexas, pero esta vez de puntos de datos, es decir,

donde, al igual que antes, los vectores se apilan por filas en y contienen los pesos de las combinaciones convexas. Por lo tanto, tenemos . Debido a las restricciones de convexidad, ambas matrices de pesos S y C son estocásticas por filas. En la figura 2 se muestra un ejemplo bidimensional de AA con .

**Aprendizaje.** Las matrices de peso se pueden encontrar minimizando la denominada suma residual de cuadrados (RSS) como función objetivo, que viene dada por

Aquí, denota la norma de Frobenius al cuadrado. Obsérvese que la optimización del RSS para encontrar las matrices de peso óptimas S y C da lugar a un problema de optimización no convexo. Sin embargo, el problema de optimización es convexo en S para un C fijo y viceversa. Esto da lugar al enfoque de optimización alternante estándar, tal y como se describe en el algoritmo 1. La figura 2 muestra la ruta de optimización (línea púrpura) desde los arquetipos inicializados (cuadrados púrpuras) hasta los arquetipos optimizados (cuadrados verdes).

**Propiedades.** Se pueden mostrar varias propiedades de AA. El primer resultado caracteriza la ubicación de los arquetipos: siempre se encuentran en el límite del casco convexo de los datos para , como se muestra en la figura 3.

**Teorema 1** ( [1]). *Sea* *un conjunto de datos discretos,*  *su envolvente convexa y* *la media de* X. *Además, sea* *el número de arquetipos y*  *el contorno de* *con*  *puntos en el contorno. Entonces, se cumple lo siguiente.*

(i) *Si* K = 1, *elegir* a1 = μ *minimiza el* RSS;

Algoritmo 1 Optimización alterna del análisis arquetípico.

Input: matriz de datos , número de arquetipos .

Output: matrices de factores , donde

A ← inicialización de los arquetipos A {Véase la sección IV-A}

**while not** converged do

**for** n = 1, 2, ..., N do

**end for**

**for** k = 1, 2, ..., K do

​.

**end for**

**end while**

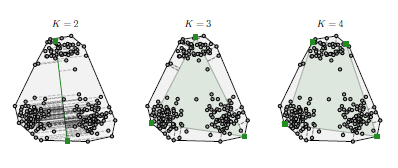


Fig. 3. Ejemplo de AA en dos dimensiones para varios números (K = 2, 3, 4) de arquetipos {a1, . . . , aK}. Los arquetipos siempre se encuentran en el límite del casco convexo de los datos.

(ii) si , *existe un conjunto de arquetipos* *en el límite de conv*(X) *que minimiza el* RSS;

(iii) si , *al elegir* *se obtiene un RSS igual a cero.*

El segundo resultado afirma que las transformaciones afines y el escalado de los datos no influyen en las matrices de peso.

**Lema 1** ( [3]). *Las matrices de ponderación S y C son invariantes ante una transformación afín y un escalado de los datos X.*

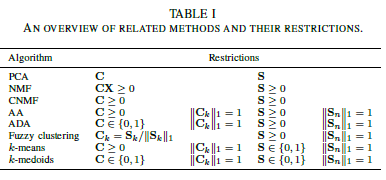
Por lo tanto, la representación AA es invariante al centrar los datos restando la media de cada característica, es decir, aplicando el operador de centrado para formar los datos centrado .

Además, la solución de AA no adolece de ambigüedad rotacional.

**Teorema 2** ([3]). Supongamos que

*and*

*Entonces, la solución de AA es única y no adolece de ambigüedad rotacional, es decir,* X ≈ SCX = SP−1PCX = ˜S ˜CX*, de modo que tanto* S,C *como* ˜ S, ˜C *son soluciones equivalentes, donde P es una matriz de permutación*.



**Perspectiva geométrica**. La minimización del RSS también se puede ver desde una perspectiva geométrica, que también se muestra en la Figura 2. Los arquetipos se eligen para abarcar un gran casco convexo (área sombreada en verde) de modo que las proyecciones por punto (líneas grises) sean mínimas. Esto se puede expresar matemáticamente como

Por lo tanto, se puede considerar que AA encuentra una aproximación (envoltura convexa verde) de la envoltura convexa del conjunto de datos (envoltura convexa gris) con un número determinado de vértices (K).

**Perspectiva volumétrica.** La perspectiva geométrica sugiere que AA consiste en encontrar un politopo con K vértices dentro de la envolvente convexa de los datos. Para reducir la suma de las proyecciones totales, este politopo debe maximizar su volumen. Esto también se consigue mediante un enfoque relacionado denominado maximización del volumen simplex (SiVM) [12]. La idea es elegir secuencialmente puntos como arquetipos de tal manera que se maximice el volumen del casco convexo que inducen. A diferencia del AA estándar, los arquetipos elegidos son puntos de datos reales y no mezclas de los mismos, como en el procedimiento de análisis arquetípico (ADA) que se describe más adelante.

Otros métodos relacionados con la factorización de matrices pueden considerarse desde la perspectiva de la minimización del volumen, por ejemplo, la factorización de matrices no negativas (NMF) [13]. En general, la mayoría de los métodos de factorización de matrices pueden considerarse como la búsqueda de una descomposición X ≈ SA de la matriz de diseño X en un producto de una matriz de peso S y una matriz de factores A. Estos métodos suelen diferir en sus supuestos sobre los datos, las restricciones que imponen a la factorización (véase la tabla I) y la función de pérdida que se utiliza. La NMF asume que la matriz de diseño es no negativa y restringe las matrices de peso y de factores para que también sean no negativas. En la figura 4 se muestra un ejemplo en el que se comparan AA y NMF. Podemos ver que los factores de NMF abren un cono en el que se pueden representar los datos y que el ángulo entre los dos factores podría ser más amplio, lo que alejaría los factores de los datos y reduciría así su interpretabilidad. Por lo tanto, el NMF se regulariza a menudo para minimizar el volumen del cono que abarcan los factores [14], [15]. Otro enfoque para el problema de la interpretabilidad consiste en restringir los factores A para que sean combinaciones convexas de datos, es decir, A = CX como en la ecuación (1), lo que da como resultado un NMF convexo (CNMF) [16].

**Perspectiva de agrupamiento**. También se pueden utilizar diversos métodos de factorización de matrices, como NMF, para el agrupamiento. Del mismo modo, el AA puede considerarse una técnica de agrupamiento en la que los arquetipos denotan representantes de clústeres y las combinaciones convexas (pesos) asociadas a los puntos de datos pueden considerarse probabilidades de pertenencia a clústeres. A diferencia del agrupamiento k-means [17], esto da lugar a una asignación suave. Sin embargo, el agrupamiento k-means está relacionado con el AA de muchas maneras. Consideremos, por ejemplo, su función objetivo.

y compárelo con la perspectiva geométrica de AA tal y como se establece en la ecuación (3). La diferencia radica en la ubicación de la proyección: hacia el envolvente convexo de los factores en AA frente al centro del clúster más cercano en k-means. Por lo tanto, la función objetivo de k-means establece un límite superior a la función objetivo de AA [18], que también se utilizó con fines de inicialización [19]. Tenga en cuenta que k-means también se puede expresar como un problema de factorización de matrices , donde C denota combinaciones convexas para modelar los centros de los clústeres y S contiene un vector base estándar por punto de datos que modela la asignación de clústeres. Al igual que AA, los centros/factores de los

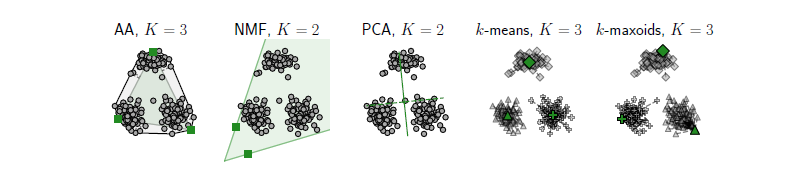


Fig. 4. AA con K = 3 arquetipos en comparación con una factorización de matriz no negativa (NMF) con K = 2 componentes, un análisis de componentes principales (PCA) con K = 2 componentes (el primero es continuo, el segundo discontinuo), agrupamiento k-means con K = 3 clústeres y agrupamiento k-maxoids con K = 3 clústeres.

clústeres son mezclas de puntos de datos y no puntos de datos en sí mismos, lo que dificulta su interpretabilidad. Para k-means, esto se soluciona con k-medoids, mientras que, para AA, se soluciona con el análisis arquetípico (ADA) [20]. Sin embargo, a diferencia de AA, k-means coloca los centros de los clústeres en el interior del conjunto de datos. Esto se evita en la agrupación k-maxoids [21], donde los representantes de los clústeres se colocan en el límite del conjunto de datos, de forma similar al procedimiento de inicialización FurthestSum propuesto para AA [3]. La figura 4 muestra cómo se compara AA con la agrupación k-means y k-maxoids.

**Perspectiva de cambio de base.** AA también puede considerarse como un cambio de representación o un cambio de base de los puntos de datos. En lugar de utilizar combinaciones lineales de la base estándar de un espacio vectorial real, AA utiliza combinaciones convexas de arquetipos, representando así los datos como en el simplex , por ejemplo, como en la parte izquierda de la figura 1. Esto muestra otra relación de AA con enfoques que realizan factorización de matrices y un cambio de base, como el análisis de componentes independientes (ICA) [22] y el análisis de componentes principales (PCA).

En ejemplos pedagógicos como los de este estudio, AA se suele representar para el caso de , donde se muestran cuatro arquetipos (K = 4) en un espacio bidimensional (M = 2). Esto no es habitual en situaciones reales, donde normalmente se trabaja con datos de alta dimensión y un número relativamente bajo de arquetipos. El caso de es problemático, ya que solo se necesitan M + 1 arquetipos para definir un punto utilizando combinaciones convexas y, a menudo, la elección de los arquetipos que se utilizan depende de la elección específica del optimizador. La figura 5 ilustra este problema. Para AA con cuatro arquetipos en un espacio bidimensional, el punto del medio puede utilizar dos subconjuntos diferentes de tamaño tres para definir el punto del medio. Por lo tanto, si se necesita consistencia representativa en el caso de K > M + 1, las combinaciones convexas deben regularizarse [23].

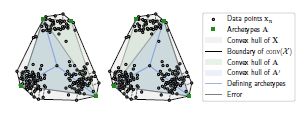


Fig. 5. Solo se necesitan M + 1 arquetipos A′ para definir un punto de datos. Por lo tanto, en el caso de K > M + 1, las ponderaciones no son consistentes. Para el punto de datos resaltado en el centro del conjunto de datos, se pueden utilizar dos subconjuntos diferentes

A′ (área sombreada en azul) de M + 1 arquetipos para definir el punto. Por lo tanto, la representación no es coherente. Aquí, M = 2 y K = 4.

III. AVANCES EN EL ANÁLISIS ARQUETÍPICO

El AA se ha adaptado para alcanzar diferentes objetivos o analizar datos de diferentes tipos. Tipo y ubicación de los arquetipos. Una ventaja del AA radica en el requisito de que los arquetipos deben ser combinaciones convexas de los datos reales, lo que garantiza que la representación se mantenga cercana a los datos observados. Sin embargo, es posible que los arquetipos reales no siempre se representen con precisión como combinaciones convexas de los datos observados.

En este contexto, [3] propuso una relajación del marco de AA para permitir la posibilidad de que los arquetipos puedan existir fuera de la envolvente convexa de los datos, cambiando la restricción original de la ecuación (1) por para algún pequeño . En algunos casos, es esencial que los arquetipos sean observaciones reales, lo que significa que son casos de la muestra X. Para abordar esta cuestión, [20] presentó un nuevo concepto arquetípico, el análisis arquetípico (ADA). Los arquetipoides son puntos de datos reales en el conjunto de datos que mejor representan los tipos puros. En este caso, la restricción original de la ecuación (1) se modifica a . Como resultado, el problema de optimización continua original se transforma en un problema de optimización mixta-entera. En la misma línea, [24] trabajó en AA separable.

**Espacios de representación no lineales.** Cuando el espacio de características se genera de forma no lineal mediante la combinación de arquetipos, se puede considerar una formulación AA no lineal. Entre los trabajos más destacados en este ámbito se encuentra [3], que amplió el AA a datos caracterizados por relaciones por pares, lo que condujo a la definición del AA kernel. En particular, el AA kernel explora cómo la función objetivo de mínimos cuadrados del AA en (2) puede reescribirse trivialmente en términos que dependen del kernel lineal equivalente , es decir, . Este enfoque tiene una amplia gama de aplicaciones prácticas [25], [26]. Además, [27] aplicó AA a las representaciones obtenidas de una capa oculta de una red neuronal de clasificación de imágenes para definir diferentes estilos de imagen. Si bien estos métodos amplían AA a espacios de características no lineales, ambos utilizan una transformación fija en el espacio de datos. Por el contrario, [28] reformuló el problema con el objetivo de aprender (a través de redes neuronales profundas) una transformación no lineal de los datos en un espacio arquetípico latente. En particular, la función objetivo original se cambia en AAnet a , donde la optimización no solo incluye los arquetipos A, sino también la función f, que es aproximadamente invertible. De forma similar a AAnet, la propuesta DeepAA presentada por [29], [30] se basa en modelos generativos probabilísticos (modelo de autoencoder variacional (VAE), cuello de botella de información variacional profunda (DVIB)), en lugar de basarse en autoencoders estándar, es decir, no variacionales. Otra diferencia es la incorporación de información secundaria por parte de [30]. Para tener en cuenta tanto la no linealidad como la estructura de recuento presente en los datos scRNA-seq, [31] desarrolló scAAnet, un método basado en autoencodificadores para realizar AA no lineal. Por último, cuando no se dispone de características y solo se proporcionan disimilitudes, [20] propuso calcular los arquetipos a partir de la matriz producida por un método de escalado multidimensional.

**Representaciones dispersas.** En ciertas aplicaciones del mundo real, como la separación hiperespectral, las soluciones dispersas son esenciales. Esta necesidad ha dado lugar a varias propuestas en este ámbito. [32] propuso un algoritmo AA con restricción de dispersión L1, que incluye también la relajación de AA a arquetipos fuera del casco convexo, tal y como se propone en [3]. La relajación también es empleada por [33] junto con un regularizador L2,1 para hacer que los resultados desmezclados sean más dispersos y un regularizador múltiple que promueve la similitud entre observaciones vecinas dentro de una representación de superpíxeles. Otro enfoque presentado por [34] consiste en transformar la formulación AA en un entorno semisupervisado, donde la función objetivo original se cambia a , donde D son los miembros finales proporcionados por una biblioteca espectral.

**Generalizaciones del análisis arquetípico.** AA solo encuentra arquetipos de observaciones. Una extensión reciente es el análisis biarquetípico (biAA), desarrollado por [35]. BiAA identifica simultáneamente arquetipos tanto de observaciones como de características, que se expresan como mezclas de los biarquetipos. El biAA es al biclustering difuso lo que el AA es al clustering difuso; y el biAA es al AA lo que el biclustering difuso es al clustering difuso. En el biAA, el problema de optimización original se sustituye por , donde los biarquetipos son , *K* arquetipos para filas y *J* para columnas, con , para , J y , para . Los biarquetipos se construyen como mezclas de observaciones y características ponderadas por C y R, respectivamente. Los coeficientes D indican el grado en que cada biarquetipo contribuye a la aproximación de cada característica. Se ha demostrado que el análisis de biarquetipos ofrece ventajas considerables con respecto a los métodos de biclustering, especialmente en lo que se refiere a la interpretabilidad [35].

**Valores faltantes.** Al igual que la mayoría de los métodos estadísticos, el análisis arquetípico asume que los datos están completos. Sin embargo, en las aplicaciones del mundo real es frecuente encontrar datos incompletos (caracterizados por valores faltantes). La primera propuesta que tuvo en cuenta la información faltante fue presentada por [3], quien modificó la función objetivo original. En [36], se asignan pesos variables a los valores no faltantes y faltantes para abordar este problema. En ambos casos, los arquetipos podrían ubicarse fuera del casco convexo. Por esa razón, [37] propuso varias alternativas. En una de ellas, se calcula una matriz de disimilitud basada en todas las disimilitudes por pares entre los puntos de datos y se proyecta para incrustar los casos dentro de un espacio euclídeo donde se lleva a cabo el AA. En otra alternativa denominada AAII, se realizan imputaciones internas durante las actualizaciones de los parámetros de AA. Siguiendo esa línea, el enfoque de [38] consiste en la aplicación del método AA estándar, mejorado con un paso adicional para imputar las entradas faltantes basado en el método de mínimos cuadrados ponderados.

**Otros tipos de datos.** El AA se definió originalmente para datos continuos; sin embargo, los avances recientes han permitido su aplicación a diferentes tipos de datos. [11] propuso una extensión probabilística del marco AA (PAA). La idea es crear la envolvente convexa dentro del espacio de parámetros y obtener perfiles arquetípicos que se encuentren en el espacio de parámetros. [11] abordó específicamente los casos de las distribuciones de probabilidad de Bernoulli, Poisson y multinomial. Cuando el modelo de observación sigue una distribución normal multivariante con una covarianza de identidad, esta formulación es equivalente a resolver el problema AA original. [39] amplió ese marco para abarcar el caso general de las observaciones nominales. Recientemente, [40] exploró cómo las expansiones de Taylor de segundo orden de las funciones de verosimilitud admiten el uso de solucionadores de mínimos cuadrados de última generación para PAA, como se ejemplifica utilizando la distribución de Bernoulli para datos binarios. [41] consideró otro enfoque y empleó un método basado en cópulas para garantizar que AA no se viera afectado por transformaciones estrictamente monótonas de los datos de entrada. La razón detrás de esto es que, por lo general, dichas transformaciones no deberían afectar a la identificación de puntos como arquetipos. Este enfoque permite que las observaciones sean continuas y/o discretas y tengan valores perdidos. [42] propuso un método alternativo para datos binarios, que identificaba arquetipoides en el espacio de observación. Como resultado, los patrones arquetípicos derivados de ADA representan soluciones viables. Este enfoque elimina la necesidad de binarización posterior con PAA y permite una recuperación más precisa de la información arquetípica. En cuanto a los datos ordinales, [43] desarrolló un método de dos pasos para aplicar ADA a respuestas ordinales basado en el modelo de estereotipos ordenados, mientras que [44] propuso un marco de optimización directa para AA para datos ordinales.

AA también se ha extendido a otros objetos de datos. En concreto, AA se ha definido para datos simbólicos [45]. En particular, [46] y [47] han generalizado los arquetipos a datos con valores de intervalo, y [48] a datos con valores de histograma. Además, AA y ADA también se han definido para datos funcionales densos y datos funcionales dispersos en [49] y [50], respectivamente. Anteriormente, [51] presentó una variación de AA diseñada para monitorizar estructuras dinámicas, como ondas viajeras o solitones, y [52] y [53] aplicaron AA a sistemas dinámicos. Además, [54] determinó redes arquetípicas cuando las redes son los objetos de datos, mientras que [55] propuso AA para datos relacionales definidos por redes ponderadas con signo. Por otra parte, [36] estudió las formas arquetípicas con puntos de referencia, mientras que [56] se centró en las formas arquetípicas de las curvas abiertas. Además, [57] consideró el AA direccional para datos axialmente simétricos.

**Análisis arquetípico ponderado.** En el problema AA original, cada observación —y, por consiguiente, cada residuo— contribuye por igual a la solución. Sin embargo, en el problema arquetípico ponderado definido por [58], la función de optimización original se sustituye por , donde W es una matriz N × N de pesos. El AA ponderado permite incorporar información adicional del conjunto de datos, como la importancia de las observaciones o la correlación entre ellas. Además, también se necesita una variante del AA ponderado cuando se utilizan conjuntos básicos. Los conjuntos básicos son subconjuntos ponderados del conjunto de datos completo en los que los modelos funcionan de forma demostrablemente competitiva en comparación con las operaciones en todos los datos, lo que permite aprender el mismo modelo AA más rápidamente y con menos datos [18].

**Robustez.** [58] propuso una forma de robustecer el AA ponderando los residuos y las observaciones respectivamente, considerando un tipo de estimadores M para datos multivariantes de valor real (Mvariate), donde el dominio de su función de pérdida es . En el análisis robusto, el dominio de la función de pérdida suele ser y esto es utilizado por [59] y [60] con la familia Huber de funciones de pérdida. En [61] se utilizó además la familia de funciones de pérdida biweight o bisquare de Tukey, que puede hacer frente mejor a los valores atípicos extremos, para el AA multivariante y funcional robusto.

**Otras variantes de AA.** Se han desarrollado otras variantes de AA para resolver problemas específicos. Por ejemplo, la factorización de matrices no negativas con envolvente convexa jerárquica, diseñada para la agrupación, fue propuesta por [62], [63]. Además, [64] introdujo un enfoque basado en datos para identificar prototipos que integra AA con el análisis de datos composicionales. Por otra parte, [65] amplió AA para dar cabida al modelado de datos multiconjunto teniendo en cuenta datos de resonancia magnética funcional (fMRI) multisujeto, lo que se conoce como AA multisujeto (MS-AA). Ese modelo impone una matriz generadora de arquetipos compartida entre los sujetos, al tiempo que permite arquetipos específicos de cada sujeto y matrices mixtas. A nivel teórico, diversas contribuciones han abordado diferentes cuestiones. [66] estableció la consistencia y convergencia de AA bajo supuestos de soporte acotado. [67] consideró una formulación alternativa de AA basada en la métrica de Wasserstein y obtuvo la existencia de soluciones en diferentes situaciones. [68] propuso una formulación de factorización de matrices no negativas que equilibra el error de reconstrucción con la distancia entre los arquetipos y la envolvente convexa de los datos, en la que un caso especial de esta formulación corresponde a AA. Las propiedades computacionales y de robustez de esta formulación son estudiadas más a fondo por [69]. Otra formulación que tiene AA como caso especial y desarrollada por [70] es AA casi convexa. Aquí, un parámetro regula la distancia máxima entre los vectores base y la envolvente convexa de los puntos de datos.

IV. IMPLEMENTACIÓN EN ANÁLISIS ARQUETÍPICO

En [3] se propone un enfoque de gradiente proyectado alternativo para AA. Dado que , los gradientes con respecto a S y C son y . Por lo tanto, AA se puede calcular actualizando alternativamente y , donde y son parámetros de tamaño de paso. Sin embargo, dado que estos pasos de gradiente pueden mover las actualizaciones fuera de sus conjuntos de restricciones, las filas de S y C deben proyectarse de nuevo en sus regiones factibles, que son el K-simplex y el N-simplex estándar, respectivamente. Alternativamente, se puede aprovechar la optimización en enfoques de variedades [71].

En [72] también se utiliza la optimización eficiente mediante el procedimiento de Frank-Wolfe [73] para evitar la compleja programación cuadrática al actualizar alternativamente C y S. Este procedimiento aprovecha la eficiencia del cálculo de los gradientes y . Sin embargo, a diferencia del método del gradiente proyectado, este enfoque evita la necesidad del paso de retroproyección sobre las restricciones del simplex. Esto se logra realizando actualizaciones de subgradientes a lo largo de direcciones afines y dentro de los simplices.

Los problemas de programación cuadrática definidos mediante la resolución alterna de S y C también pueden resolverse imponiendo directamente la restricción lineal mediante la regularización cuadrática, tal y como se propone en [74]. En consecuencia, al resolver una fila de S dada por minimizar , en la que los elementos de son no negativos y λ se establece muy grande para garantizar que se satisfaga la regularización hacia el simplex. En particular, esto se puede resolver directamente utilizando el procedimiento de conjunto activo de mínimos cuadrados no negativos descrito en [75], [76]; para más detalles, véase también [40]. Cabe destacar que, en [59], la dispersión inherente de la solución se puede utilizar dentro del procedimiento de conjunto activo para acelerar significativamente la convergencia. Concretamente, el algoritmo de conjunto activo actualiza iterativamente un subconjunto S de variables activas (distintas de cero). Dada una estimación actual en una determinada iteración, el algoritmo define un subconjunto y busca una dirección resolviendo el problema reducido: y , , donde representa el complemento de S en el conjunto de índices [K]. A continuación, la estimación se actualiza como , con γ ajustado para mantener dentro del simplex factible . Este enfoque de conjunto activo continúa refinando el conjunto activo S hasta que se alcanza una solución óptima. De forma análoga, este método puede aplicarse para optimizar con respecto a C.

En [40] se observó que, al estimar S considerando un número limitado de arquetipos, un procedimiento de inferencia muy eficiente que impone la restricción de no negatividad y suma a uno es el procedimiento de optimización mínima secuencial (SMO), utilizado originalmente para resolver el problema de optimización cuadrática de la máquina de vectores de soporte (SVM) [77]. La SMO considera el subproblema cuadrático mínimo teniendo en cuenta dos elementos en a la vez y resuelve la redistribución de la masa de estos dos elementos (es decir, la suma de los dos elementos) entre ellos, lo que da como resultado un polinomio simple de segundo orden univariante con solución de forma cerrada; véase también [40] para más detalles.

En [78], un algoritmo basado en el aprendizaje de diccionarios en línea introduce un enfoque novedoso para la AA al desacoplar los arquetipos tal y como se definen directamente en términos de combinaciones convexas de las observaciones de datos. Esto se consigue mediante un algoritmo de AA con restricciones esféricas. Inicialmente, los puntos de datos se proyectan sobre la esfera unitaria, ya sea mediante una simple normalización o, en casos más complejos, mediante proyecciones estereográficas. Tras la proyección, todos los puntos se encuentran en la esfera unitaria. Aquí, los arquetipos se definen como una matriz que se encuentra dentro de la esfera unitaria. Para calcular estos arquetipos, se resuelve el problema de optimización sujeto a. Dado que los arquetipos están restringidos a residir dentro de la esfera unitaria, sus combinaciones convexas forman un casco convexo dentro de la esfera. Dado que los puntos de datos también se normalizan para situarse en la esfera unitaria, se garantiza que quedarán fuera de este casco convexo. En la fase de aprendizaje, el algoritmo ajusta iterativamente las superficies del casco convexo para minimizar el error de representación global en todo el conjunto de datos, acercando eficazmente las superficies a los datos. La resolución de este problema de optimización produce arquetipos que se corresponden aproximadamente con los puntos extremos del conjunto de datos.

Un algoritmo propuesto en [79] aprovecha la geometría subyacente y los patrones de dispersión de las representaciones convexas para identificar arquetipos en las envolturas convexas de los datos. En lugar de optimizar S y C directamente con respecto a toda la matriz de datos X, los autores definen una función objetivo alternativa que permite un aprendizaje eficiente e independiente de C sin necesidad de interactuar directamente con X. Esto se consigue aprendiendo C en el espacio de coeficientes en lugar de en el espacio de señales. El método comienza resolviendo el siguiente problema de optimización sujeto a eliminando un grado de libertad de modo que . Una vez obtenido Z, las matrices de coeficientes S y C pueden derivarse alternativamente resolviendo sujeto a y .

Recientemente, [80] presentó una técnica de compresión de datos basada en un método aleatorio de subespacio de Krylov para reducir la dimensionalidad de los datos. Este método minimiza la necesidad de realizar consultas frecuentes a conjuntos de datos de alta dimensionalidad y representa un enfoque novedoso en el contexto del AA. Además, el estudio propone el uso de proyecciones aleatorias para aproximar la envolvente convexa de los datos, donde las distancias se conservan aproximadamente, reduciendo así la cardinalidad del diccionario para la representación de arquetipos. Ambas técnicas mejoran la eficiencia del algoritmo AA al reducir el tamaño y la dimensionalidad del conjunto de datos. Una idea relacionada es aplicar un PCA a la matriz de datos X para reducir la dimensionalidad antes de ejecutar AA [81] [83]. Este enfoque se denominó análisis arquetípico de espacio reducido (RSAA) y se demostró que produce resultados similares a los del AA básico [84].

Por último, otras aproximaciones AA utilizan algoritmos de recocido cuántico [85] o de descenso entrópico [86].

Además, los modelos sustitutos proporcionan un enfoque eficaz para gestionar las elevadas exigencias computacionales del AA, especialmente con grandes conjuntos de datos. Al aproximar el comportamiento del modelo, se reduce el tiempo necesario para el análisis, lo que permite una exploración rápida y asequible de los patrones de datos.

*A. Estrategias para la inicialización de arquetipos*

Una consideración importante al ajustar los modelos AA es la inicialización de los arquetipos. Una inicialización adecuada es crucial para mejorar la velocidad de convergencia del algoritmo y garantizar la fiabilidad de los arquetipos resultantes.

En [1] se propone un método de inicialización aleatoria. En este enfoque, los arquetipos se seleccionan extrayéndolos del conjunto de datos según una distribución uniforme. Concretamente, cada índice se elige aleatoriamente, donde [N] denota el conjunto de índices del conjunto de datos.

Un enfoque alternativo es el algoritmo FurthestSum, introducido por [3], [87], que modifica el algoritmo FurthestFirst [88] diseñado para k-means a un enfoque de inicialización adecuado para AA. En este método, el primer arquetipo se selecciona aleatoriamente, al igual que en el enfoque FurthestFirst original. Sin embargo, los arquetipos posteriores se eligen en función de la distancia más lejana con respecto a los arquetipos ya seleccionados. El siguiente índice de arquetipo se selecciona según la regla , donde A representa el conjunto de arquetipos ya seleccionados y es un punto de datos del conjunto de datos. Este método garantiza que cada nuevo arquetipo se elija lo más lejos posible de los ya existentes, lo que mejora la diversidad y la representatividad de los arquetipos en todo el conjunto de datos.

En [19] se propone un procedimiento de inicialización diferente, inspirado en el procedimiento de inicialización kmeans++ [89], en el que el primer arquetipo también se selecciona aleatoriamente, pero los arquetipos posteriores se eligen en función de una distribución de probabilidad , que es proporcional a la distancia entre los puntos de datos y la envolvente convexa de los arquetipos ya seleccionados. Concretamente, la distribución de probabilidad viene dada por . Por lo tanto, el siguiente arquetipo se selecciona en función de esta distribución de probabilidad: . Este método garantiza que los puntos más alejados de los arquetipos existentes tengan una mayor probabilidad de ser seleccionados, lo que también favorece una mejor cobertura del conjunto de datos.

Otra línea de trabajo explora directamente el uso de técnicas de inicialización diseñadas originalmente para la agrupación en clústeres. Por ejemplo, estrategias como k-means++ [89], FurthestFirst [88] y el algoritmo Incremental Anomalous Pattern (AP) [8] pueden adaptarse para AA. Aunque estos métodos no están diseñados específicamente para AA como los anteriores, pueden inicializar eficazmente los prototipos. En particular, el algoritmo AP ofrece la ventaja añadida de determinar el número de prototipos que se van a utilizar, lo que lo hace especialmente valioso en escenarios en los que se desconoce el número óptimo de arquetipos.

Es importante señalar que todos estos métodos de inicialización podrían dar lugar a la introducción de arquetipos redundantes. Sin embargo, en [90] se presenta un algoritmo para reciclar los arquetipos eliminando los arquetipos redundantes.

En la figura 6 se puede ver un ejemplo de algunas de estas inicializaciones, y en [19] se puede encontrar una comparación reciente y detallada de muchos enfoques de inicialización.

*B. Técnicas de reducción del tamaño de los conjuntos de datos*

Reducir el tamaño de los conjuntos de datos es fundamental para mejorar la eficiencia computacional en AA, especialmente cuando se trabaja con conjuntos de datos de gran tamaño.

Mediante el empleo de técnicas que reducen el tamaño del conjunto de datos sin alterar su estructura e información fundamentales, es posible realizar análisis con mayor rapidez sin que ello afecte significativamente a la calidad de los resultados.

Un enfoque común consiste en calcular un subconjunto, , del conjunto de datos y utilizarlo para calcular los arquetipos. Dado que es más pequeño que el conjunto de datos original, el algoritmo converge más rápidamente. Por ejemplo, en [91], [92], los arquetipos se minimizan a partir de una submuestra de puntos en la envolvente convexa de X, que se construye tomando la unión de los puntos encontrados en las envolventes convexas de diferentes proyecciones 2D de X.

Alternativamente, en [93], se establece como el subconjunto del conjunto de datos que reside en el límite del casco convexo del conjunto de datos, el denominado marco. Tenga en cuenta que el marco también se puede calcular cuando se utilizan núcleos [94], lo que permite su uso para el AA de núcleo.

En otro enfoque, presentado en [18], se genera un subconjunto representativo mediante el muestreo de puntos de una distribución , donde μ es la media de los puntos de datos. Este subconjunto es un coreset y se pueden establecer garantías teóricas sobre el error de aproximación. Cabe destacar que los puntos del subconjunto tienen pesos y que, al aprender el modelo, se debe utilizar una versión ponderada de AA.

Por otro lado, existe una técnica para reducir el tamaño del conjunto de datos que se aplica durante cada paso de optimización, en lugar de antes de iniciar el algoritmo. Este método, introducido en [92], se basa en la observación de que los puntos de datos dentro de la envolvente convexa de los arquetipos no contribuyen al residuo minimizado por el algoritmo de arquetipos. Como resultado, en cada iteración, el conjunto de datos se divide en dos conjuntos: X− y X+. El conjunto X− contiene puntos que pueden representarse exactamente como combinaciones convexas de arquetipos, mientras que X+ contiene puntos que solo pueden aproximarse. Al centrarse únicamente en X+, que es más pequeño que el conjunto de datos completo, el algoritmo converge más rápidamente. Además, a medida que las estimaciones arquetípicas mejoran con las iteraciones, el número de puntos fuera de la envolvente convexa disminuye, lo que reduce aún más el tamaño del problema de optimización y acelera el algoritmo.

*C. Evaluación de la solidez del modelo*

Evaluar la solidez del modelo es esencial para garantizar que los arquetipos y los conocimientos que proporcionan sean fiables y significativos. Se pueden utilizar varias técnicas para evaluar la estabilidad y la coherencia del modelo, cada una de las cuales aborda diferentes aspectos del análisis.

Una consideración importante en AA es el uso de inicializaciones aleatorias y barras de error en la pérdida. AA suele implicar un proceso de optimización que comienza desde puntos iniciales aleatorios en el espacio de parámetros. La aleatoriedad de estas

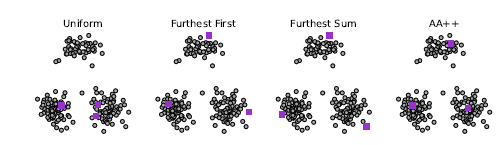


Fig. 6. Comparación de las siguientes técnicas de inicialización de AA para K = 3: Uniform, FurthestFirst, FurthestSum y AA++.

inicializaciones puede dar lugar a diferentes óptimos locales, lo que podría afectar a los arquetipos resultantes. Para evaluar la estabilidad de los resultados, se pueden realizar múltiples ejecuciones de AA con diferentes inicializaciones y se puede supervisar la pérdida (por ejemplo, el error de reconstrucción) a lo largo de estas ejecuciones. Al visualizar las barras de error en los valores de pérdida, podemos medir la variabilidad de los resultados. Las barras de error más pequeñas indican que el modelo es estable, mientras que las barras de error más grandes sugieren una mayor sensibilidad a las condiciones iniciales.

Sin embargo, los resultados pueden seguir produciendo estimaciones fiables del error, mientras que los componentes extraídos subyacentes cambian sustancialmente. Para evaluar la consistencia de las matrices S extraídas, en [65] se propuso la información mutua normalizada entre dos matrices S y S′ estimadas, explorando que forma una distribución conjunta con elementos para los que la información mutua puede cuantificarse mediante y normalizada a una puntuación de similitud entre 0 y 1, es decir,

.

Para la consistencia de los arquetipos para dos matrices A y A′ que emparejan de forma codiciosa los componentes según su proximidad entre sí y promedian estas distancias hasta obtener una distancia cuadrática media, se puede utilizar para definir una puntuación de similitud como

,

donde es la varianza media entre las M características de X.

El bootstrapping es otra técnica utilizada para evaluar la solidez de los arquetipos mediante la generación de múltiples remuestreos de los datos originales con reemplazo [95]. Se aplica AA a cada uno de estos conjuntos de datos remuestreados y se comparan los arquetipos resultantes para medir su variabilidad. El bootstrapping ayuda a determinar si los arquetipos observados son consistentes en diferentes subconjuntos de datos y cómo podrían generalizarse a datos nuevos o desconocidos.

Mediante las técnicas anteriores, podemos evaluar de forma sólida la fiabilidad de los arquetipos generados por AA. Esto garantiza que las conclusiones extraídas del análisis sean significativas y reproducibles, lo que hace que el modelo sea más fiable para su uso en diversas aplicaciones.

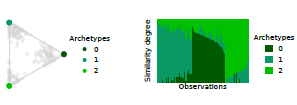
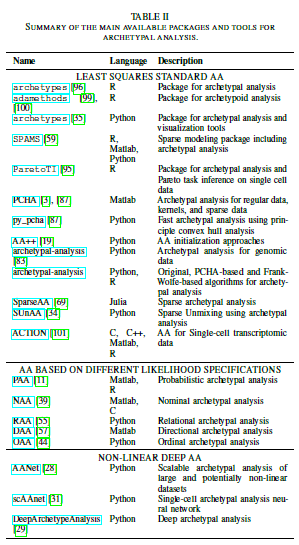


Fig. 7. Panel izquierdo: Gráfico simplex que representa las puntuaciones, donde cada observación es un punto y los arquetipos se sitúan en los vértices coloreados. La ubicación de cada punto indica su contribución relativa a partir de los arquetipos. Panel derecho: Gráfico de barras apiladas de las puntuaciones, donde cada barra representa una observación y los segmentos coloreados corresponden a las contribuciones de cada arquetipo.



*D. Validación del número de arquetipos seleccionados*

Las estrategias para seleccionar el número de arquetipos, es decir, K, se han basado tradicionalmente en la inspección visual. Los métodos de validación comunes incluyen el gráfico de scree o los criterios de codo, que analizan el comportamiento monótono de una función objetiva [96] o una medida de la variación explicada por diferentes modelos [1], [3]. Más recientemente, [97] introdujo un criterio de teoría de la información específicamente adaptado para evaluar la bondad del ajuste en AA. Esta métrica, análoga a la medida similar al AIC discutida en [98], se define como

donde y representan las matrices de covarianza de y , respectivamente.

*E. Herramientas y software accesibles*

Las herramientas de software para AA son esenciales para los investigadores que desean aplicar estos métodos de manera eficaz. El software accesible hace que la AA sea práctica para más campos, lo que favorece la investigación reproducible y fomenta una adopción más amplia en diversas disciplinas. En la Tabla II resumimos algunas de las herramientas de software más destacadas para AA. Estas herramientas están disponibles en varios lenguajes de programación, como R, Python y Matlab, y ofrecen diferentes funcionalidades en función de las necesidades del usuario. Por ejemplo, el paquete archetypes en R está diseñado para AA básica, mientras que adamethods se centra en ADA, una extensión de AA. Otras herramientas, como SPAMS, proporcionan capacidades de modelado disperso, incluyendo AA, mientras que el otro software explicado en[3] incluye opciones avanzadas paratrabajar con datos regulares, kernelizados o dispersos. En particular, el paquete Python archetypes ofrece también excelentes herramientas de visualización para mostrar los resultados de AA, lo que lo hace especialmente útil para los usuarios que buscan representaciones gráficas intuitivas de sus análisis.

*F. Visualización de resultados*

En cualquier tipo de análisis de aprendizaje automático, interpretar los resultados es un paso crucial. En el caso de AA, esto se puede abordar principalmente de dos maneras: analizando las puntuaciones, que miden la similitud de cada observación con cada arquetipo, o examinando directamente los propios arquetipos. Cuando nos centramos en los arquetipos, podemos evaluar sus valores en términos de percentiles dentro de la distribución de cada variable. Varios tipos de visualizaciones, en las que los arquetipos se representan de forma coherente mediante colores, pueden ayudar en este proceso. Estas visualizaciones se pueden generar fácilmente utilizando el paquete de Python archetypes descrito anteriormente, que incluye herramientas prácticas para crear y personalizar gráficos.

La figura 7 ilustra visualizaciones de S que definen cómo cada observación puede visualizarse como una mezcla de los arquetipos. El panel izquierdo de la figura ilustra un gráfico simplex, en el que cada punto corresponde a una observación. La posición de cada punto refleja la contribución relativa de los arquetipos, representados por vértices de colores. Este tipo de gráfico es especialmente eficaz para identificar grupos de observaciones que comparten contribuciones similares de los arquetipos. El gráfico simplex funciona bien con hasta tres arquetipos cuando se visualiza el simplex en 2D, pero se vuelve redundante en la forma en que se reconstruyen los puntos cuando se incluyen más arquetipos.

Como alternativa al gráfico simplex, el panel derecho de la

Figura 7 muestra un gráfico de barras apiladas con las puntuaciones. Cada barra corresponde a una observación, segmentada en secciones de colores que indican la contribución de cada arquetipo. Este gráfico ofrece una forma sencilla de comparar cómo se combinan los diferentes arquetipos en las observaciones y representa con precisión S para un número arbitrario de arquetipos.

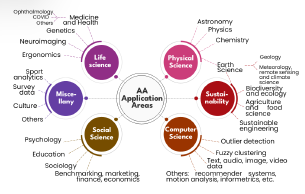


Fig. 8. Mapa de burbujas con las principales áreas de aplicación de la AA.

En general, la mejor forma de visualizar los resultados de AA suele depender mucho del problema concreto. Por lo tanto, cuando se trabaja con datos geoespaciales o formas 3D, suelen ser más adecuados otros tipos de visualizaciones adaptadas a estas estructuras de datos. Por ejemplo, los mapas de calor o los mapas coropléticos pueden ser útiles para representar arquetipos en conjuntos de datos espaciales, mientras que los diagramas de dispersión 3D o las visualizaciones de malla pueden emplearse para analizar formas [102].

V. APLICACIONES DESTACADAS DEL AA

A continuación, describimos las aplicaciones más destacadas de la AA en distintos campos de la ciencia. Cabe señalar que el volumen de investigaciones que aplican el AA es muy amplio, por lo que aquí destacamos algunas de las aplicaciones más destacadas del AA en la literatura. La figura 8 resume el área de aplicación del AA. En las figuras 9-13 destacamos además cinco ejemplos de análisis de AA aplicados respectivamente a datos biológicos sobre las propiedades de los pinzones, datos químicos basados en mediciones de resonancia magnética nuclear (RMN), una imagen hiperespectral de teledetección, dos problemas de detección de valores atípicos, así como un conjunto de datos de ciencias sociales basado en los registros de votación del Congreso. En los análisis, los datos se han centrado restando la media.

*A. Ciencias de la vida*

El AA ha encontrado amplias aplicaciones dentro de las ciencias de la vida. Como tal, la extracción de politopos se ha aplicado a datos biológicos para caracterizar compensaciones evolutivas [2] en las que los arquetipos definen los rasgos óptimos para diversas tareas, teniendo en cuenta también datos sobre pinzones similares al ejemplo que se muestra en la figura 9. Inspirada en dicha extracción de rasgos, el AA se ha utilizado en el ámbito de la genética [95], [103]-[107], incluso para la caracterización de perfiles genéticos destacados a nivel poblacional [83]. El AA se ha aplicado además a datos de secuenciación de ácido ribonucleico unicelular (scRNAseq) para caracterizar la diversidad celular [101] teniendo en cuenta también el núcleo AA basado en similitudes celulares RNAseq [26], así como representaciones no lineales que consideran un enfoque AA profundo [31]. En [108], los datos de microarrays de biopsias de trasplantes renales se puntuaron en términos de aspectos de rechazo de órganos y estas puntuaciones se caracterizaron mediante AA para caracterizar los fenotipos moleculares relacionados con el rechazo de trasplantes renales. En [109], las biopsias de trasplantes hepáticos con pruebas de anticuerpos específicos del donante se analizaron mediante AA para la caracterización de fenotipos asociados con distintas propiedades hepáticas y la supervivencia del trasplante.

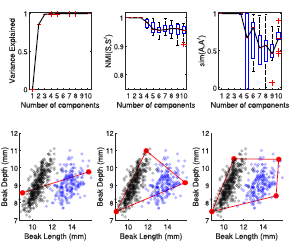


Fig. 9. Análisis de datos sobre pinzones5 en los que se miden dos características basadas en la longitud y profundidad del pico para los pinzones scandens (puntos azules) y fortis (puntos negros) ilustrados para un modelo AA K = 2, K = 3 y K = 4. Cuando se espera la varianza explicada, observamos mejoras sustanciales hasta K = 3, y en cuanto a la reproducibilidad, medida utilizando NMI(S, S′) y sim(A,A′), observamos que un modelo K = 3 también recupera de forma fiable S y A. Mientras que el modelo AA K = 2 discrimina principalmente entre los dos tipos de pinzones, el K = 3 caracteriza en particular la variabilidad en la clase de pinzones fortis, mientras que el K = 4 subdivide aún más la clase de pinzones scandis.

En el ámbito de la neuroimagen, el AA se ha utilizado para caracterizar regiones de baja unión, alta unión y sin unión de trazadores en estudios de tomografía por emisión de positrones (PET) [3]. Además, el AA se ha utilizado para caracterizar distintos patrones de activación funcional en datos de resonancia magnética funcional (fMRI) [65], [110], incluidas las respuestas a estímulos visuales [65], [111], y para la caracterización de la variabilidad de un solo ensayo en datos de magnetoencefalografía (MEG) [112]. Se ha avanzado aún más para dar cabida a la integración multimodal de las respuestas evocadas por estímulos de electroencefalografía y magnetoencefalografía (EEG y MEG) [57], [113], proporcionando patrones topográficos distintos como alternativa a los enfoques convencionales de análisis de microestados [57]. En [114] se aplicó el AA a la imagenología por fluorescencia de calcio para la detección de grupos de neuronas que se coactivan sistemáticamente.

El AA se ha aplicado además en el contexto de la medicina y la salud. Así, se utilizó para caracterizar la dinámica espacio-temporal de los brotes de gripe [115] y COVID-19 [116] en Montana, EE. UU. En [117] se aplicó el ADA para caracterizar los efectos del COVID-19 en la salud en toda Europa, mientras que los datos epidemiológicos de los países europeos sobre el COVID-19 en dos momentos distintos fueron analizados mediante AA en [118]. En [119] se examinó la progresión de la enfermedad en pacientes con esclerosis lateral amiotrófica mediante su caracterización dinámica en términos de arquetipos que describen diversos aspectos de las deficiencias. Cabe destacar que la pérdida del campo visual (CV) se ha caracterizado ampliamente mediante AA [120]-[122] y se ha relacionado con características clínicas establecidas [123], y específicamente para la caracterización de la neuritis óptica en [124] y el glaucoma en [125]-[127]. En este contexto, en [128] se consideró un enfoque de AA basado en el aprendizaje profundo y en [126], [129] se utilizó un enfoque recursivo jerárquico en el que las combinaciones convexas extraídas para la reconstrucción de la observación se descomponían recursivamente mediante AA. En [130] se utilizó AA para caracterizar las redundancias dentro de un catálogo de mutaciones somáticas en el cáncer y en [131] para identificar tipos clínicos extremos de pacientes con esclerosis lateral amiotrófica (ELA). Además, AA se ha utilizado para caracterizar estrategias de recompensa en ratones [132].

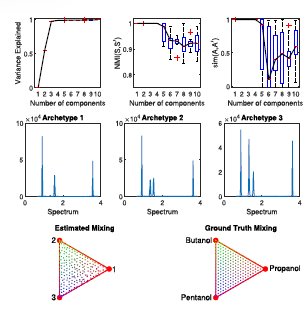


Fig. 10. Análisis de datos sobre un experimento de RMN diseñado con mezclas que contienen fracciones sistemáticamente diferentes de propanol, butanol y pentanol7. Cuando esperamos la varianza explicada, observamos mejoras sustanciales hasta K = 3 y, en cuanto a la reproducibilidad medida con NMI(S, S′) y sim(A,A′), observamos que un K = 3 recupera de forma fiable S y A. Este modelo de tres componentes se corresponde bien con los tres compuestos, mientras que las fracciones de concentración estimadas (indicadas en la parte inferior) se corresponden bien con los niveles de concentración reales de los tres compuestos en cada muestra.

El AA también se ha utilizado para analizar datos de encuestas sobre la actividad física de las mujeres embarazadas [133]. Por último, el AA también se ha aplicado en el estudio de la ergonomía para caracterizar las formas de los pies [134], las mediciones antropométricas relevantes para el diseño de cabinas de aviones [102] y exoesqueletos [135], así como para identificar casos límite para diseños ergonómicos de estaciones de trabajo definidos en términos de los vecinos más cercanos a los arquetipos identificados [136].

*B. Física y química*

El AA se ha aplicado a diferentes problemas en física y química. Estos abarcan desde el análisis de las entidades más grandes del universo, como la astrofísica, hasta las más pequeñas, como la nanotecnología. El AA fue empleado por [137] para examinar conjuntos de datos que consistían en espectros galácticos, ya que se cree que cada espectro es una superposición de emisiones de diversas poblaciones estelares, emisiones nebulares y actividad nuclear dentro de la galaxia. Cada una de estas fuentes de emisión representa un arquetipo potencial para todo el conjunto de datos. Además, los datos del reactor de fusión Tokamak fueron analizados con AA por [1]. El AA también se ha utilizado para analizar datos de sistemas dinámicos [51]-[53], [138], en un experimento de pulso químico [51], una simulación numérica de la ecuación de Kuramoto-Sivashinsky [52] y datos de llamas celulares [53], [138]. Además, el AA ha sido aplicado por [139] en el registro incorrecto de obleas.

También encontramos aplicaciones de AA en química, como el análisis de espectros 1H NMR de mezclas de propanol, butanol y pentanol por [3] (véase también la figura 10) y la biblioteca espectral de infrarrojo medio de muestras de suelo por [140]. Además, se analizan muestras virtuales de nanopartículas de diamante y nanoescamas de grafeno utilizando AA por [141] y nanoescamas de óxido de grafeno por [142], mientras que las nanoestructuras arquetípicas de cinco conjuntos virtuales de puntos cuánticos de silicio (SiQD) se describen en [143]. Por último, el AA se ha aplicado al análisis químico de vinos [144].

*C. Ciencia climática y sostenibilidad*

El AA se ha utilizado en aplicaciones relacionadas con la ciencia climática. Por ejemplo, [145] estudió las precipitaciones diarias y las exportaciones de humedad tropical en el este de Estados Unidos; [71] analizó el monzón asiático de verano utilizando la presión a nivel del mar, junto con la temperatura mensual de la superficie del mar, tal y como hicieron [84], [146] y [147]; [82] examina los fenómenos de flujo de larga duración en el hemisferio sur; [81] identifica los periodos de alta confianza en las previsiones subestacionales; [148] investigó la variabilidad espacio-temporal de las precipitaciones extremas estacionales; [149] estudió las series climatográficas históricas de temperaturas y precipitaciones normales; por último, [150] estudió los arquetipos de lluvias extremas en Australia y publicó los conjuntos de datos relacionados. Los enfoques de minería de datos, incluido el análisis arquetípico para la ciencia meteorológica y climática, también se resumieron y explicaron en [151].

Además de los datos climáticos, el AA se ha utilizado para otros problemas computacionales relacionados con la sostenibilidad. Así, [149] analizó el consumo mundial de electricidad, mientras que la contaminación atmosférica se estudió en el problema seminal que dio origen al AA [1]; [152] analizó la calidad y el origen de los alimentos de los hogares en Togo; [153] predijo la biomasa de caña de azúcar; [154] estudió los arquetipos florísticos de las estepas patagónicas; [155] analizó los modelos agrícolas que sustentan las prácticas de los ganaderos flamencos desde una perspectiva agroecológica; [156] examinó la biogeografía global de los grupos funcionales dentro de las poblaciones de lagartos; [157] estudió los eventos volcánicos y sísmicos del volcán Tungurahua en Ecuador; [158] exploró la composición de las comunidades de peces en los mares europeos; [159] analizó los datos de producción de Olea europaea; [160] analizó las propiedades relacionales dentro de las redes de conocimiento de los bioclústeres europeos; por último, [161] trabajó en la detección de flujos anómalos en las redes de agua urbanas.

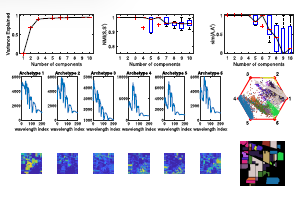


Fig. 11. Análisis de datos de imágenes hiperespectrales de teledetección9 en los que se mide una imagen de un paisaje en Indiana, EE. UU., en 200 longitudes de onda (eliminando las bandas correspondientes a la absorción del agua) en una imagen de 145 × 145. La imagen se convierte en una matriz de 200 características de longitud de onda por 1452 observaciones de píxeles. Cuando se espera la varianza explicada y la reproducibilidad medidas utilizando NMI(S, S′) y sim(A,A′), observamos que un modelo de K = 6 componentes es el máximo que se puede extraer al tiempo que se obtienen soluciones fiables. Al inspeccionar la representación AA aprendida para K = 6, observamos que los arquetipos proporcionan perfiles espectrales distintos con sus combinaciones convexas.

El AA también se ha utilizado en el contexto de la teledetección para el análisis de imágenes satelitales hiperespectrales. Para ver un ejemplo de dicha aplicación del AA a datos hiperespectrales, véase también la figura 11. En [149] se analizaron imágenes hiperespectrales teniendo en cuenta un procedimiento de minimización de volumen simplex, así como el AA, y se descubrió que dichos procedimientos proporcionan representaciones fácilmente interpretables de los perfiles espectrales subyacentes. En [162] se descubrió que este enfoque proporcionaba características muy adecuadas para la clasificación de la cobertura del suelo, mientras que el enfoque se amplió al AA kernel en [163], [164] utilizando un kernel de función de base radial y a arquetipos que representan casos reales (es decir, arquetipoides) en [60]. En [165] se utilizó la probabilidad de Poisson para AA con el fin de tener en cuenta las intensidades de píxeles con valores enteros, imponiendo un marco ponderado basado en las distancias de movimiento de tierra (EMD) entre las observaciones al formar los arquetipos, mientras que en [166] y [167] se utilizó AA para la selección de bandas hiperespectrales adecuadas para la clasificación. En [168] se utilizó un procedimiento de inferencia reversible de Markov Chain Monte Carlo combinado con recocido simulado para identificar representaciones globalmente robustas de arquetipos para imágenes hiperespectrales basadas en representaciones AA candidatas. Además, en [86] se exploró el descenso del gradiente entrópico para proporcionar actualizaciones explícitas eficientes que impusieran las restricciones de mezcla en AA para el análisis de imágenes hiperespectrales. En [34] se definieron los arquetipos en términos de combinación convexa de miembros finales obtenidos de una biblioteca de espectros predefinidos y en [33] se relajaron las restricciones convexas sobre C utilizando el enfoque de relajación propuesto en [3] en combinación con una regularización que impone una reconstrucción similar de los píxeles pertenecientes al mismo segmento de superpíxeles. En [169], el índice de vegetación de diferencia normalizada (NDVI), definido por la diferencia normalizada entre las bandas del infrarrojo cercano y el rojo, se analizó transformando los datos espacio-temporales en espectros de potencia utilizados como entrada para AA basándose en una transformada wavelet continua. En [170], se utilizó AA para la detección de anomalías en imágenes hiperespectrales.

*D. Informática y ciencia de datos*

Al ser un método utilizado con frecuencia en la ciencia de datos, el AA se ha empleado para diversas tareas de ciencia de datos y aprendizaje automático, no solo como herramienta independiente, sino también como extensión de otras herramientas.

Una de las propiedades fundamentales del AA es su capacidad para identificar prototipos extremos que se encuentran en los límites del conjunto de datos. Esto convierte al AA en un candidato perfecto para la detección de anomalías y valores atípicos. Por ejemplo, el AA se utilizó para la verificación de firmas manuscritas fuera de línea [173] y para apoyar a DevOps en su trabajo. [174] construyó un clasificador de una sola clase basado en AA para sistemas ciberfísicos, mientras que [161] utiliza el análisis de datos funcionales con AA y ADA para la detección de anomalías en redes de agua. ADA también fue utilizado por [100], por ejemplo, para datos de series temporales como los datos de ECG. [171] propone combinar AA y un algoritmo de vecino más cercano para obtener un nuevo enfoque de detección de valores atípicos en subespacios, y [172] amplió su enfoque. Para ver ejemplos del uso de AA para la detección de valores atípicos, véase también la figura 12. Por último, [175] utiliza AA para la detección de anomalías en el análisis de juegos electrónicos.

Como se ha señalado anteriormente, el AA también puede considerarse desde una perspectiva de agrupamiento, para lo cual [7]-[9] evalúan el AA para el agrupamiento difuso. Además, [176] utilizan el AA como enfoque de agrupamiento para agrupar a los jugadores de videojuegos según su comportamiento, basándose en representaciones obtenidas a partir de un autoencoder recurrente.

El AA se utiliza con frecuencia para el análisis de datos que se presentan en diferentes modalidades. Para el análisis de datos de texto, [3], [87] utilizan el AA como modelo temático en la línea de la asignación latente de Dirichlet (LDA) [177]. La idea es extraer una serie de temas de una colección de textos en la que cada tema está representado por palabras arquetípicas. En [178] se propone una extensión de la LDA basada en el análisis. Otra línea de investigación en este ámbito se ocupa del resumen de múltiples documentos, en el que se utilizan diversas variantes de AA [179]-[182].

En el caso de los datos de audio, [183] y [184] utilizan AA como herramienta para el aprendizaje de representaciones y la estimación de fuentes. Para la clasificación de audio, [185] combina arquetipos como representaciones de límites y centros de modelos de mezcla gaussiana para obtener mejores incrustaciones para un clasificador de bosque aleatorio, y [186] propone un núcleo de coincidencia intermedio basado en AA para entrenar una máquina de vectores de soporte (SVM) para la clasificación de datos bioacústicos.

Otra modalidad son los datos de imagen. [187] extrae características de las imágenes y calcula un AA sobre esas características para obtener imágenes arquetípicas. [3], [87] ejecuta un AA directamente sobre datos de imagen que muestran rostros humanos. [59] realiza una clasificación basada en AA de imágenes que muestran dígitos. Para otra tarea, [59] divide las imágenes en parches, calcula las características SIFT por parche y calcula palabras visuales utilizando AA para proporcionar una representación para los clasificadores. La idea de dividir las imágenes en parches también fue utilizada por [72]. En lugar de calcular las características por parche, utilizaron los parches tal cual para calcular los arquetipos de los parches. A continuación, los parches de las imágenes se aproximaron mediante combinaciones convexas de parches arquetípicos, lo que dio lugar a una especie de autoencoder. Por último, [188] calcula arquetipos para tareas de reconocimiento facial, no exactamente mediante AA, sino mediante un enfoque relacionado denominado maximización del volumen simplex [12].

Como última modalidad, consideramos los datos de vídeo para los que [189] utilizan AA para resumir vídeos basándose en un conjunto de características, así como en la información textual del título, y [190] realizan resúmenes multivídeo mediante AA ponderado multimodal.

Otro caso de uso destacado de AA son los videojuegos. En este ámbito, [191] extrae primitivas de movimiento utilizando AA para su uso en la creación de bots para videojuegos, [192] propone un sistema de recomendación basado en AA para videojuegos, [193] aprovecha AA para comprender grupos de jugadores en videojuegos, [175] realiza análisis de videojuegos, [176] se utiliza AA como enfoque de agrupación para clasificar a los jugadores de videojuegos según su comportamiento, basándose en representaciones obtenidas de un autoencoder recurrente, y [194] se emplea AA para encontrar arquetipos de estilos de juego que se utilizan para entrenar a un agente por arquetipo que coopera de forma óptima con un humano.

El AA también se utiliza para crear sistemas de recomendación. [3], [87] utilizan el AA para el filtrado colaborativo en los datos de Movielens. En este caso, la tarea consiste en inferir los valores que faltan en las valoraciones de los usuarios sobre las películas. [191] propuso un sistema de recomendación basado en el AA

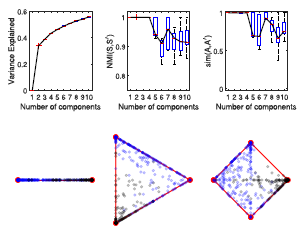


Fig. 13. Análisis de los datos de los registros de votación del Congreso11 teniendo en cuenta 16 proyectos de ley con votos faltantes de algunos miembros del Congreso en algunos proyectos de ley. Aquí se utiliza el procedimiento Kernel AA teniendo en cuenta la similitud de Jaccard entre los miembros del Congreso en términos de los proyectos de ley por los que ambos votaron, ignorando los proyectos de ley que tenían valores faltantes, es decir, KJac. ij = P m:(xim∈{0,1},xjm∈{0,1}) ximxjm P m′:(x im′∈{0,1},x jm′∈{0,1}) 1−(1−xim′ )(1−xjm′ ) . El núcleo se ha centrado dos veces, es decir, TKJac.T⊤ antes de los análisis, lo que corresponde a eliminar la media. Los puntos negros corresponden a los republicanos, mientras que los puntos azules corresponden a los demócratas. Los modelos aprendidos se ilustran para un modelo AA K = 2, K = 3 y K = 4. Cuando se espera la varianza explicada y la reproducibilidad medidas utilizando NMI(S, S′) y sim(A,A′), observamos que hasta K = 4 los modelos AA pueden inferirse de forma fiable, mientras que la varianza explicada mejora constantemente a medida que aumenta el número de componentes. Al inspeccionar el modelo para K = 2, K = 3 y K = 4, observamos que los modelos separan bien a los demócratas (puntos azules) de los republicanos (puntos negros).

para videojuegos, y [195] propuso un sistema de recomendación que combina la clasificación y el filtrado colaborativo basado en los usuarios utilizando el AA para la industria de la moda.

El AA se ha tenido más en cuenta en los campos del análisis del movimiento y el diseño de mapas. [191] utiliza el AA para extraer primitivas de movimiento en videojuegos que se utilizan para la creación de bots de juego. [196] propone un AA temporal para la segmentación de acciones en series temporales multimodales que describen el movimiento humano. [197] y [198] utilizan AA para evaluar rutas en interiores y [199] lo utiliza para evaluar la usabilidad del diseño de mapas.

Dentro de la informetría, [200] utiliza AA para comprender la interacción entre la ciencia y la sociedad y, en especial, el aspecto de la creación de valor. [201] extrae seis arquetipos de economistas basándose, por ejemplo, en indicadores bibliográficos con la ayuda de AA. Este estudio fue ampliado por [202] a ADA. [203] también aprovecha ADA para extraer tres instituciones y facultades arquetípicas dentro del campo de la economía. Por último, [25] utiliza kernel AA para agrupar la dinámica de la atención colectiva en las redes sociales.

Por último, el AA se utilizó en ingeniería de software, por ejemplo, para identificar modelos prototípicos para la tarea de estimación del esfuerzo de software [204], [205], para analizar la deuda técnica [206] y para el análisis de los tipos de contribución de los desarrolladores de software de código abierto [207].

*E. Ciencias sociales y miscelánea*

AA ha encontrado amplias aplicaciones en las ciencias sociales. En el campo de la psicología, [208] examinó los arquetipos del liderazgo, [209] investigó las personalidades arquetípicas y [210] delineó los rasgos de personalidad, mientras que [211], [212] analizaron las identidades sociales y los arquetipos de comportamiento, así como [213], que estudió los perfiles de comportamiento en los juegos digitales. Además, [64] utilizó el AA para estudiar la formación de la identidad en la adolescencia. [64] también consideró el AA en otro campo de aplicación, como el benchmarking, donde se encontraron los mejores resultados entre las unidades operativas de una empresa de telecomunicaciones. [214] también propuso el uso del AA para identificar universidades de referencia. Del mismo modo, [215] aplicó el AA para descubrir clasificadores de benchmarking para la predicción de impagos de préstamos. En un campo relacionado, como el marketing, el AA también se ha utilizado para analizar las operaciones de empresas multinacionales por [216], [217] y para segmentar mercados por [218]. Además, el AA se ha aplicado en los campos de las finanzas y la economía. Así, [61], [219] describieron las empresas del S&P 500 representadas por dos series temporales de cotizaciones bursátiles y [220] las empresas por varios indicadores financieros. Como ejemplo de la aplicación del AA a datos de ciencias sociales, la figura 13 aplica el AA kernel a los datos de los registros de votación del Congreso.

También hay aplicaciones en Sociología. Por ejemplo, [221]

examinó las identidades en los países europeos para estudiar el conflicto social. Otro ejemplo es el análisis del uso del tiempo de los académicos realizado por [222].

Además, el AA se ha aplicado en el ámbito educativo. Por ejemplo, [223] analizó el uso de los cursos y los archivos de registro de los cursos en línea y su evaluación [224], mientras que [225] y [226] analizaron la carga de trabajo de los estudiantes en casa y las respuestas de los estudiantes a las encuestas de los profesores, respectivamente. Por último, [42] analizó cuestionarios binarios para identificar los perfiles de habilidades de los estudiantes y explicar los elementos de un examen de matemáticas.

El análisis deportivo es otro campo de aplicación del AA. Por un lado, [227] aplicó el AA al análisis de varias variables de jugadores de baloncesto y fútbol. Por otro lado, [50] utilizó el ADA, no solo para el análisis de varias variables de jugadores de baloncesto, sino también para estudiar las trayectorias profesionales de los jugadores con funciones y el rendimiento del equipo a través de la información de relaciones asimétricas. Además, [228] utilizó el ADA para funciones dispersas en la predicción del rendimiento de los jugadores de baloncesto.

También hay aplicaciones en otros campos, como la evaluación del habla [229], los modelos culturales y sus implicaciones [230]– [232], el análisis del estilo artístico [27], el análisis de actos terroristas [233], las pruebas y evaluaciones térmicas no destructivas [234] y el diseño arquitectónico [235].

Además, los AA y sus variantes se han aplicado para analizar datos de encuestas muy diversos, como datos sensoriales codificados por intervalos [236], datos sobre conocimientos financieros [237], datos sobre el apego a la comunidad [238], datos sobre acciones de igualdad de género en la universidad [239], datos de la Encuesta Social Europea [44] y datos sobre la calidad de vida de pacientes afectadas por cáncer de mama y la encuesta de satisfacción de los estudiantes [43].

VI. ORIENTACIONES FUTURAS Y PROBLEMAS PENDIENTES

En resumen, el AA ha encontrado amplias aplicaciones en distintos ámbitos de la ciencia para caracterizar conjuntos de datos en términos de propiedades distintivas. Si bien estas caracterizaciones han resultado útiles para muchos fines diferentes, aún quedan muchas vías por explorar en la investigación sobre el AA.

Es importante destacar, como se ha señalado anteriormente, que el modelo AA no es convexo y, por lo tanto, es propenso no solo a mínimos locales, sino también a una posible falta de fiabilidad. Si bien esto suele mitigarse considerando procedimientos de inicialización adecuados, así como múltiples análisis que tienen en cuenta diferentes inicializaciones del modelo, en futuros trabajos se deberían investigar los límites de la convexidad en el AA basándose en ideas derivadas de las propiedades geométricas de las envolturas convexas [240], así como en relajaciones de los enfoques AA similares a las que se han explorado para la agrupación mediante relajaciones espectrales como puntos de partida para los análisis [241], [242].

Otro problema pendiente es la estimación fiable del orden del modelo; en otras palabras, el número de arquetipos K. Aunque esto suele depender del problema y requiere conocimientos del dominio, sería beneficioso disponer de una forma automática de determinar K. Un problema pendiente relacionado es la transferencia de conocimientos de un orden de modelo a otro, es decir, añadir o eliminar un arquetipo sin volver a calcular la factorización desde cero.

Algunos avances en AA se derivan de la exploración de aplicaciones de agrupación y sus extensiones [35]. Siguiendo esta tendencia, AA podría extenderse a otros tipos de datos y configuraciones, como datos censurados, AA justo o AA con privacidad diferencial.

Si bien el AA se ha utilizado para investigar la dinámica temporal, aún quedan importantes líneas de investigación por desarrollar en cuanto a metodologías para que el AA cambie de forma fluida a lo largo del tiempo. Por lo tanto, las futuras líneas de investigación deberían estudiar cómo se pueden combinar las ideas del modelado de sistemas dinámicos con los procedimientos de AA para caracterizar la evolución temporal de los politopos. Esto incluye imponer dependencias temporales en la forma en que los arquetipos (A) pueden evolucionar en el tiempo, así como en la forma en que la reconstrucción temporal de las observaciones puede depender de combinaciones convexas de puntos temporales anteriores, etc.

Una limitación importante del AA es que se basa en la presencia de formas puras en los datos o en que dichas formas puras puedan derivarse como combinaciones convexas de las observaciones. En muchas situaciones, el modelo AA no puede recuperar dichas observaciones puras y es necesario flexibilizar los supuestos del modelo AA para caracterizar las formas puras. Un enfoque que se ha debatido consiste en relajar la suma a una restricción sobre por δ. Sin embargo, esto requiere un ajuste adecuado de δ, mientras que el valor de δ puede depender del arquetipo específico considerado. En futuros trabajos se debería investigar cómo se pueden combinar los procedimientos de AA con enfoques basados en el volumen mínimo y cómo extraer de manera óptima y eficiente los politopos cuando las formas puras no están contenidas dentro de la envolvente convexa de los datos.

El AA convencional asume que los datos residen dentro de un politopo en el espacio de datos considerado. Sin embargo, normalmente los datos residen en una variedad, por lo que se debe tener en cuenta la estructura de la variedad al formar los arquetipos. En la actualidad, el AA de kernel solo puede remediar esto de forma limitada, imponiendo una estructura de kernel preespecificada para tener en cuenta las relaciones por pares y la estructura de vecindad en la variedad de datos. Los procedimientos de AA profunda pueden aprender representaciones latentes adecuadas en las que la variedad de datos se puede desentrañar a través de las representaciones no lineales aprendidas. Por lo tanto, el trabajo futuro debería centrarse en combinar de manera óptima los enfoques de aprendizaje de variedades con el AA.

VII. CONCLUSIÓN

Este estudio presentó el marco del análisis arquetípico (AA), así como sus extensiones y aplicaciones, incluyendo los numerosos y dispares campos de la ciencia en los que el AA se ha aplicado con éxito para caracterizar y obtener información sobre la estructura de los datos de alta dimensión. Esperamos que este estudio también sirva de punto de partida para que los investigadores que no estén familiarizados con esta metodología de modelización adopten este procedimiento como herramienta estándar en sus análisis de datos, con el fin de obtener una comprensión más profunda de los distintos aspectos que caracterizan los datos, así como de cómo las observaciones pueden describirse como combinaciones convexas de los mismos.