# Análisis de papers

## A survey on archetypal analysis

**Año:** 2025

**Autores:** Aleix Alcacer, Irene Epifanio, Sebastian Mair, Morten Morup

### Principales hallazgos

#### Introducción

**Definición de Análisis Arquetípico (AA)**

* Procedimiento que extrae aspectos distintivos (“formas puras”) de los datos.
* Cada observación se representa como combinación convexa de esos arquetipos.

**Origen del concepto**

* Surge de un problema en química atmosférica (simulación de ozono).
* Había que resumir grandes cantidades de datos en pocos patrones representativos.

**Carácter interpretativo**

* Los arquetipos son explicables y fáciles de interpretar (se entienden como “formas ideales” de los datos).
* Se diferencian de otros métodos de reducción de dimensionalidad porque no pierden la conexión con ejemplos concretos.

**Aplicaciones en diferentes campos**

* Biología: compensaciones evolutivas, optimalidad de Pareto.
* Química: mezclas de concentraciones.
* Geociencias: imágenes hiperespectrales, “miembros extremos”.
* Ciencia de datos: agrupamiento, prototipos, agrupamiento difuso.

**Ejemplo práctico: MNIST dígito 9**

* Se extraen 3 arquetipos de “formas puras” del número 9.
* Los datos de 784 dimensiones se reducen a un triángulo (2D).
* Cada observación se proyecta hacia los vértices = interpretabilidad clara.

**Limitaciones del AA**

* Problema de optimización no convexo (no siempre encuentra los arquetipos “verdaderos”).
* Difícil elegir K (el número de arquetipos).
* Sensibilidad a outliers.
* Puede que los datos estén en estructuras no lineales; AA tradicional no encaja bien.
* No siempre cumple supuestos de normalidad gaussiana.

**Objetivo del paper**

* Ofrecer un estado del arte del AA.
* Mostrar aplicaciones en distintos campos.
* Discutir limitaciones y buenas prácticas.
* Revisar 152 documentos relevantes hasta 2025.

**Resumen**

El Análisis Arquetípico busca encontrar “formas puras” en los datos que representen los casos más extremos o prototípicos. En lugar de describir los datos con promedios (como PCA) o grupos rígidos (como k-means), AA genera puntos “ideales” y cada observación se entiende como una mezcla de esos puntos.

Ejemplo: si analizas estrellas, los arquetipos serían formas puras de curvas de luz (estrella muy variable, estrella poco variable, estrella periódica perfecta, etc.). Cada estrella real sería una mezcla de esas formas puras.

#### Descripción teórico/técnica

**Definición formal del AA**

* **Datos de entrada:**
  + Conjunto de puntos
  + Se apilan en la matriz de diseño
* **Objetivo:** Descomposición aproximada

donde:

* + = pesos de cada observación respecto a los arquetipos
  + = matriz de arquetipos
* Interpretación

Cada dato es una **combinación convexa** de arquetipos .

* + Restricciones:
  + Los arquetipos también son combinaciones convexas de los datos:
  + Por tanto,
* Tanto S como C son matrices estocásticas por filas (filas suman 1, entradas

**Aprendizaje (Optimización)**

* Se minimiza la suma residual de cuadrados (RSS)

Donde es la norma de Frobenius al cuadrado.

* Problema: la optimización es no convexa en conjunto, pero:
  + Convexa en S si C está fijo
  + Convexa en C si S está fijo.
* Solución: usar optimización alternante (algoritmo 1)

**Algoritmo 1**

1. Inicializar arquetipos A
2. Mientras no haya convergencia:
   * Para cada dato , calcular el vector de pesos resolviendo:
   * Actualizar arquetipos:
   * Para cada arquetipo , calcular su combinación de datos:
   * Ajustar

**Propiedades teóricas**

1. **Geométrica:**
   * AA encontrar una envolvente convexa simplificada de los datos.
   * Objetivo: minimizar distancias de proyección.
2. **Volumétrica:**
   * AA busca un politopo de vértices que maximice volumen dentro del casco convexo de los datos.
   * Relación con SiVM (maximización de volumen simplex).
   * Diferencia: SiVM elige datos reales como arquetipos, AA permite combinaciones.
3. **Comparación con NMF (Non-negative Matrix Factorization):**
   * NMF impone no negatividad en todos los factores.
   * AA asegura que los arquetipos sean combinaciones convexas (más interpretables)
   * CNMF (NMF convexos) es una versión intermedia.
4. **Perspectiva de agrupamiento:**
   * AAun clustering suave (cada dato pertenece a varios clusters con distintos pesos)
   * Relación con K-means y K-medoids:
     + K-means ubica centros dentro de los datos menos interpretables.
     + AA los coloca en el borde (mejor interpretabilidad)
5. **Perspectiva de cambio de base:**
   * En vez de base estándar, se usa la base de los arquetipos.
   * Relación con PCA e ICA.
   * Problema si : redundancia, necesidad de regulación.

**Resumen**

El análisis arquetípico (AA) es un método de factorización matricial cuyo objetivo es representar un conjunto de datos como una combinación convexa de un número reducido de arquetipos. Cada dato se expresa como mezcla ponderada de arquetipos, y a su vez, los arquetipos se definen como combinaciones convexas de los propios datos. De esta manera, tanto los datos como los arquetipos quedan vinculados mediante las matrices de peso , que son estocásticas por filas.

El problema de aprendizaje se formula como la minimización de la suma residual de cuadrados (RSS), medida con la norma de Frobenius al cuadrado. Aunque la optimización conjunta es no convexa, resulta convexa si se fija una de las matrices de peso, lo que permite resolver el problema mediante un esquema de optimización alternante. Este procedimiento ajusta iterativamente los pesos y arquetipos hasta la convergencia.

Teóricamente, los arquetipos se ubican siempre en el límite del casco convexo de los datos cuando , y en el caso extremo de tomar como arquetipos todos los puntos del borde (, el error de reconstrucción se anula. Asimismo, se ha demostrado que el AA es invariante ante transformaciones afines y escalado de los datos, y que la solución es única hasta permutaciones, evitando problemas de rotación como en otros métodos de factorización.

El AA puede interpretarse desde diferentes perspectivas: geométrica (minimización de proyecciones hacia un politopo de vértices), volumétrica (maximización del volumen del politopo dentro de la envolvente convexa), y de agrupamiento (asignación suave de pertenencia a clusters). Además, guarda relación con otras técnicas de reducción de dimensionalidad como PCA, ICA o NMF, aunque se distingue por la interpretabilidad que ofrece al definir arquetipos como combinaciones convexas de datos reales.

#### Avances en el análisis arquetípico

**Puntos principales (Adaptaciones del AA)**

* **Tipo y ubicación de los arquetipos**
  + Ventaja: arquetipos = combinaciones convexas de datos reales.
  + Problema: a veces los arquetipos reales no están bien representados como combinaciones convexas.
  + Extensiones:
    - Relajación del AA: permitir arquetipos fuera de la envolvente convexa con parámetros .
    - ADA (Análisis de Arquetipoides): arquetipos deben ser observaciones reales problema se convierte en optimización mixta-entera.
    - AA separable variante para casos especiales.
* **Espacios de representación no lineales**
  + **Kernel AA:** usa kernels (ej: relaciones por pares), reescribe la función objetivo en términos de matrices de kernel amplias aplicaciones.
  + **DeepAA/AAnet:** reformulan el AA con redes neuronales aprenden transformaciones no lineales al espacio arquetípico.
  + **scAAnet:** adaptado a datos de biología (scRNA-seq) usando autoencoders variacionales/probabilísticos.
  + Extensiones cuando no hay características: se usan disimilitudes + espacio multidimensional.
* **Representaciones dispersas**
  + Útiles en aplicaciones como separación hiperespectral
  + Se agregan restricciones de dispersión (L1, L2, 1, regularizadores múltiples).
  + También versiones semisupervisadas con matrices adicionales.
* **Generalizaciones**
  + **BiAA (análisis biarquetípico):** extiende AA a filas y columnas equivalente a biclustering difuso.
  + Permite identificar arquetipos en observaciones y características simultáneamente.
  + Ventaja: mayor interpretabilidad.
* **Valores faltantes**
  + Variantes para trabajar con datos incompletos.
  + Estrategias. Modificar la función objetivo, imputaciones internas, trabajar en espacios de disimilitudes.
* **Otros tipos de datos**
  + **Probabilístico (PAA):** extiende AA a distribuciones (Bernoulli, Poisson, multinomial).
  + Extensiones a datos binarios, ordinales, simbólicos, funcionales, redes, formas.
  + Diferentes adaptaciones según el tipo de variable.
* **AA ponderado**
  + Cada observación puede tener diferente peso en la función objetivo.
  + Útil para incorporar información externa (importancia, correlación).
  + También se aplica en **conjuntos básicos** entrenar AA más rápido con subconjuntos representativos.
* **Robustez**
  + Variantes robustas al ruido y outliers.
  + Uso de funciones de pérdida robustas (Huber, Tukey biweight)
* **Otras variantes**
  + AA jerárquico, multisujeto (MS-AA), casi convexo, formulaciones en métrica de Wasserstein.
  + Convergencia y propiedades teóricas estudiadas en trabajos recientes.

**Resumen**

El análisis arquetípico ha sido objeto de múltiples extensiones y adaptaciones para ampliar su aplicabilidad y superar limitaciones del modelo original. Una primera línea de desarrollo se centra en el tipo y ubicación de los arquetipos. Aunque la definición clásica exige que los arquetipos sean combinaciones convexas de los datos reales, en algunos casos resulta ventajoso relajar esta restricción y permitir que los arquetipos se sitúen fuera de la envolvente convexa, mediante un parámetro de tolerancia δ. Otra alternativa es el Análisis de Arquetipoides (ADA), en el que los arquetipos deben ser observaciones reales del conjunto de datos, lo que convierte el problema en una optimización mixta-entera.

Otra línea de investigación ha extendido el AA hacia espacios de representación no lineales. Ejemplos de ello son el Kernel AA, que reformula la función objetivo en términos de matrices de kernel, y los enfoques basados en redes neuronales profundas, como AAnet y DeepAA, que aprenden transformaciones no lineales hacia un espacio arquetípico latente. Estas propuestas han demostrado su utilidad en aplicaciones como biología computacional (scAAnet) e imagen digital.

En escenarios donde se requiere mayor interpretabilidad, se han propuesto representaciones dispersas, que introducen regularizadores L1 y L2,1 para promover soluciones parciales y localizadas. Asimismo, el análisis biarquetípico (BiAA) generaliza el modelo clásico al identificar arquetipos tanto de observaciones como de características, lo que establece un paralelismo con los métodos de biclustering.

La literatura también ha abordado limitaciones prácticas como la presencia de valores faltantes, proponiendo variantes que integran imputaciones internas, el uso de espacios de disimilitudes o modificaciones directas de la función objetivo. De manera similar, se han diseñado extensiones para otros tipos de datos, incluyendo versiones probabilísticas (PAA) para distribuciones discretas, así como adaptaciones a datos binarios, ordinales, simbólicos, funcionales, redes y formas.

Por otra parte, se ha introducido el AA ponderado, en el que las observaciones contribuyen de forma diferenciada a la solución, permitiendo incorporar información externa o utilizar subconjuntos representativos de los datos. También se han desarrollado versiones robustas, que emplean funciones de pérdida como Huber o Tukey biweight, con el fin de reducir la sensibilidad a valores atípicos. Finalmente, se han propuesto múltiples variantes adicionales, como el AA jerárquico, el AA multisujeto (MS-AA) y el AA casi convexo, junto con formulaciones en métricas alternativas como la de Wasserstein, lo que ha permitido profundizar en el estudio teórico de la consistencia y la convergencia del modelo.

#### Implementación en análisis arquetípico

**Optimización y algoritmos**

* [3]: **Gradiente proyectado alternativo**
  + Actualiza matrices con pasos de gradiente.
  + Requiere proyectar resultados al simplex (factibilidad).
  + Optimización eficiente sin retroproyección.
  + Usa direcciones afines dentro del simplex.
* [74–76]: **Programación cuadrática con regularización**
  + Añade término de regularización grande para forzar simplex.
  + Usa algoritmos de conjunto activo iteraciones más rápidas.
* [59]: La **dispersión inherente** puede acelerar la convergencia en el conjunto activo.
* [40]: **SMO (Sequential Minimal Optimization)**
  + Adaptado de SVM.
  + Actualiza pares de coeficientes en 𝑆 solución cerrada y eficiente.
* [78]: **AA con restricciones esféricas**
  + Datos y arquetipos proyectados en la esfera unitaria.
  + Mejora la geometría de las combinaciones convexas.
* [79]: **AA en el espacio de coeficientes**
  + Optimiza sobre 𝑍 en vez de sobre 𝑋.
  + Reduce complejidad computacional.
* [80–84]: **Compresión y reducción dimensional**
  + Krylov subspace, proyecciones aleatorias, PCA.
  + Variante: RSAA (AA en espacio reducido).
* Otros: recocido cuántico, descenso entrópico, modelos sustitutos para grandes datasets.

1. **Estrategias de inicialización**

* Inicialización aleatoria ([1]).
* FurthestSum ([3], [87]): elige puntos lo más alejados posible diversidad.
* k-means++ adaptado ([19]): selecciona puntos según distribución proporcional a distancia de la envolvente convexa.
* Métodos de clustering adaptados: k-means++, FurthestFirst, AP ([8]).
* Problema de redundancia: [90] propone reciclaje de arquetipos.

1. **Reducción del tamaño de datos**

* Selección de subconjuntos representativos ():
  + Submuestreo de proyecciones 2D [91,92].
  + Uso del marco (puntos en el límite del casco convexo) [93,94].
  + Coresets con pesos, muestreados según distancia a la media [18].
* Reducción dinámica durante la optimización [92]:
  + Dividir en (dentro de la envolvente) y (fuera)
  + Optimizar solo en .

1. **Evaluación de solidez**

* Estabilidad frente a inicializaciones aleatorias: comparar pérdidas con barras de error.
* Consistencia de S: usar información mutua normalizada (NMI).
* Consistencia de A: medir distancia promedio entre arquetipos (sim(A,A’)).
* Bootstrapping: evaluar robustez con remuestreos.

1. **Validación del número de arquetipos**

* Métodos clásicos: gráfico de codo, scree plot, variación explicada.
* [97]: criterio de teoría de la información adaptado a

1. **Herramientas y software**

* R: archetypes, adamethods.
* Python: archetypes (con visualización).
* Matlab: otras implementaciones.
* SPAMS: modelado disperso + AA.

1. **Visualización de resultados**

* Dos enfoques:

1. Arquetipos A analizar en percentiles.
2. Puntuaciones S interpretar mezcla de arquetipos.

* Gráficos comunes:
  + Simplex (bueno hasta 3 arquetipos).
  + Barras apiladas (funciona con más K).
* Otros según el tipo de dato: mapas de calor, coropléticos, dispersión 3D, mallas.

**Resumen**

El desarrollo de algoritmos eficientes ha sido un aspecto clave en la evolución del análisis arquetípico. Inicialmente, se propusieron métodos basados en gradiente proyectado para actualizar iterativamente las matrices de coeficientes, aunque con la necesidad de reproyectar las soluciones dentro de los simplex factibles. Posteriormente, se introdujo el uso del algoritmo de Frank-Wolfe, que evita dicha retroproyección al realizar actualizaciones en direcciones afines, lo que mejora la eficiencia computacional. Otros enfoques recurren a la programación cuadrática con regularización, en combinación con algoritmos de conjunto activo, que permiten optimizar sobre subconjuntos reducidos de variables y acelerar la convergencia. De manera similar, el procedimiento de optimización mínima secuencial (SMO), originalmente diseñado para máquinas de soporte vectorial, se ha adaptado con éxito para el AA.

En paralelo, han surgido variantes que redefinen la geometría del problema, como el AA esférico, donde los datos y los arquetipos se proyectan sobre la esfera unitaria, o los métodos que formulan la optimización en el espacio de coeficientes en lugar del espacio de datos, reduciendo la complejidad del problema. También se han explorado técnicas de reducción dimensional, incluyendo métodos de subespacio de Krylov, proyecciones aleatorias y PCA, lo que dio origen al AA de espacio reducido (RSAA). Más recientemente, se han propuesto aproximaciones innovadoras basadas en recocido cuántico, descenso entrópico y modelos sustitutos, que facilitan la escalabilidad del AA a grandes volúmenes de datos.

Un aspecto determinante para la calidad de los resultados es la inicialización de los arquetipos. La literatura ofrece múltiples estrategias, desde inicializaciones aleatorias hasta procedimientos inspirados en clustering, como FurthestSum o k-means++ adaptado, que favorecen la diversidad y cobertura de los arquetipos. Sin embargo, estas técnicas pueden producir redundancia, lo que ha motivado la creación de algoritmos de reciclaje de arquetipos.

Otra línea de trabajo fundamental es la reducción del tamaño de los conjuntos de datos. Entre las estrategias propuestas destacan el uso de subconjuntos representativos (coresets), la selección de puntos en el marco del casco convexo, y técnicas de reducción dinámica que optimizan solo sobre los datos fuera de la envolvente convexa. Estas aproximaciones logran acelerar significativamente el proceso de ajuste sin sacrificar calidad en los resultados.

Para garantizar la fiabilidad del modelo, se han desarrollado diversas técnicas de evaluación de solidez. Estas incluyen la repetición del AA con múltiples inicializaciones aleatorias y el uso de métricas de estabilidad como la información mutua normalizada (NMI) o la similitud entre arquetipos reconstruidos. Asimismo, el bootstrapping permite evaluar la consistencia de los arquetipos frente a variaciones de los datos.

La selección del número de arquetipos (K) constituye otro desafío metodológico. Aunque los criterios tradicionales se basan en métodos visuales como el gráfico de codo, investigaciones recientes han introducido métricas basadas en teoría de la información, análogas a criterios como el AIC, que proporcionan un marco más formal para esta decisión.

En cuanto a la implementación práctica, existen diversas herramientas de software que facilitan la aplicación de AA en distintos lenguajes de programación, incluyendo R, Python y Matlab. Entre ellas se destacan archetypes, adamethods y SPAMS, con funcionalidades que van desde el AA básico hasta variantes dispersas y kernelizadas.

Finalmente, la visualización de resultados desempeña un papel esencial en la interpretación del AA. Dos enfoques principales son la representación de arquetipos en términos de percentiles y el análisis de las puntuaciones como mezclas convexas. Gráficos como el simplex o las barras apiladas permiten explorar estas relaciones de manera intuitiva, mientras que, en aplicaciones específicas, como datos espaciales o tridimensionales, resultan más adecuados mapas coropléticos, diagramas 3D o visualizaciones de malla.

#### Aplicaciones destacadas del AA

1. **Ciencias de la vida**

* **Evolución y biología:**
  + Extracción de politopos → caracterización de compensaciones evolutivas en especies (ej. pinzones).
  + Genética: perfiles poblacionales, scRNAseq para diversidad celular, enfoques kernel y AA profundo.
  + Microarrays de trasplantes renales y hepáticos → fenotipos moleculares y de supervivencia.
* **Neurociencia:**
  + PET: regiones de baja/alta unión de trazadores.
  + fMRI/MEG/EEG: patrones de activación, variabilidad de ensayo, integración multimodal, microestados.
  + Fluorescencia de calcio: detección de neuronas coactivadas.
* **Medicina y salud:**
  + Dinámica espacio-temporal de epidemias (gripe, COVID-19).
  + Progresión de enfermedades como ELA y glaucoma.
  + Pérdida del campo visual y enfermedades oculares → AA profundo y enfoques jerárquicos.
  + Cáncer: redundancias en mutaciones somáticas.
* **Otras aplicaciones biomédicas y sociales:**
  + Estrategias de recompensa en ratones.
  + Encuestas sobre actividad física en embarazadas.
  + Ergonomía: formas de pies, diseño de cabinas, exoesqueletos y estaciones de trabajo.

1. **Física y química**

* **Astrofísica:**
  + Espectros galácticos → combinación de poblaciones estelares, nebulosas y núcleos activos.
  + Fusión nuclear: análisis de datos del reactor Tokamak.
  + Sistemas dinámicos: experimentos químicos, simulaciones no lineales (Kuramoto-Sivashinsky), llamas celulares.
  + Nanotecnología: nanopartículas de diamante, grafeno, puntos cuánticos de silicio.
* **Química clásica:**
  + Espectros NMR de alcoholes.
  + Espectros infrarrojos de suelos.
  + Análisis de vinos.

1. **Ciencia climática y sostenibilidad**

* **Clima:**
  + Monzones, precipitaciones extremas, lluvias en Australia.
  + Fenómenos de flujo, predicción subestacional.
  + Series climatológicas históricas.
* **Sostenibilidad y medioambiente:**
  + Consumo eléctrico global, contaminación atmosférica.
  + Agricultura y ecología: caña de azúcar, estepas, lagartos, peces, olivo.
  + Eventos volcánicos y sísmicos.
  + Redes de agua urbanas (detección de anomalías).
* **Teledetección (hiperespectral):**
  + Clasificación de suelos, vegetación (NDVI), anomalías en imágenes.
  + Métodos kernel AA, arquetipoides, Poisson-AA, selección de bandas, inferencia bayesiana (MCMC).

1. **Informática y ciencia de datos**

* **Detección de anomalías y outliers:**
  + Firmas manuscritas, sistemas ciberfísicos, ECG, redes de agua, videojuegos.
* **Clustering y agrupamiento:**
  + Aplicaciones en jugadores de videojuegos, agrupamiento difuso.
* **Datos multimodales:**
  + Texto: AA como modelo temático (similar a LDA), resumen de documentos.
  + Audio: representación, clasificación bioacústica, separación de fuentes.
  + Imágenes: rostros, dígitos, parches, autoencoders arquetípicos.
  + Video: resúmenes unimodales y multimodales.
* **Videojuegos:**
  + Bots, recomendadores, clasificación de jugadores, estilos de juego cooperativos.
* **Sistemas de recomendación:**
  + Películas (Movielens), videojuegos, moda.
* **Movimiento y mapas:**
  + Segmentación de acciones, rutas interiores, usabilidad en mapas.
* **Informetría y bibliometría:**
  + Arquetipos de economistas, instituciones, dinámicas sociales.
* **Ingeniería de software:**
  + Estimación de esfuerzo, deuda técnica, contribuciones en proyectos open source.

1. **Ciencias sociales y otras áreas**

* **Psicología y sociología:**
  + Liderazgo, personalidad, identidades sociales, adolescencia.
* **Educación:**
  + Cursos en línea, carga de trabajo, habilidades estudiantiles.
* **Deportes:**
  + Baloncesto y fútbol → rendimiento y trayectorias profesionales.
* **Marketing y economía:**
  + Segmentación de mercados, multinacionales, benchmarking financiero.
  + Empresas del S&P 500, indicadores financieros.
  + Registros de votación del Congreso (AA kernel).
* **Otras aplicaciones:**
  + Evaluación del habla.
  + Modelos culturales.
  + Arte y estilos artísticos.
  + Actos terroristas.
  + Diseño arquitectónico.
  + Encuestas sobre género, calidad de vida, satisfacción estudiantil.

**Resumen**

El análisis arquetípico (AA) se ha consolidado como una herramienta versátil aplicada en un amplio espectro de disciplinas científicas. Su utilidad radica en la capacidad de identificar prototipos extremos que resumen las características fundamentales de un conjunto de datos, lo que lo convierte en un método adecuado tanto para la exploración descriptiva como para tareas analíticas específicas.

En el ámbito de las ciencias de la vida, el AA ha sido utilizado para estudiar procesos evolutivos, diversidad genética y fenotipos moleculares asociados a trasplantes, así como para caracterizar patrones de activación cerebral mediante neuroimagen (PET, fMRI, MEG, EEG). También ha encontrado aplicaciones relevantes en epidemiología, donde se ha empleado para el análisis de epidemias como la gripe y el COVID-19, y en el estudio de enfermedades neurológicas y oculares como la esclerosis lateral amiotrófica o el glaucoma. Más allá del área biomédica, el AA se ha aplicado en estudios de ergonomía y comportamiento animal, mostrando su alcance interdisciplinario.

En física y química, el AA se ha aplicado a escalas tan diversas como el análisis de espectros galácticos en astrofísica y la caracterización de nanopartículas en nanotecnología. En química, se ha empleado en el análisis espectral (NMR, IR) y en estudios aplicados como el análisis de vinos. También se ha aplicado al estudio de sistemas dinámicos complejos, incluyendo simulaciones no lineales y experimentos de combustión.

En el campo de la ciencia climática y la sostenibilidad, el AA ha sido utilizado para caracterizar fenómenos meteorológicos (monzones, precipitaciones extremas, series históricas), así como en estudios de ecología y recursos naturales. Destacan sus aplicaciones en la teledetección mediante imágenes hiperespectrales, donde se ha extendido con variantes como el kernel AA, arquetipoides y aproximaciones probabilísticas.

Dentro de la informática y la ciencia de datos, el AA se emplea en detección de anomalías, agrupamiento, minería de texto, análisis de audio, imagen y vídeo, así como en sistemas de recomendación y videojuegos. Su flexibilidad permite integrarlo en flujos multimodales, combinarlo con técnicas de aprendizaje profundo y utilizarlo como base para tareas específicas como la clasificación, el resumen de documentos o la segmentación de series temporales. Además, ha demostrado ser útil en ingeniería de software y en estudios bibliométricos.

Por último, en el ámbito de las ciencias sociales, el AA se ha aplicado en psicología, sociología, educación, economía y deportes, así como en marketing y finanzas. Ejemplos notables incluyen el análisis de perfiles de personalidad, el estudio de dinámicas sociales en redes, la segmentación de mercados y el análisis del rendimiento deportivo. El AA también se ha empleado en arte, cultura, seguridad y diseño arquitectónico, lo que evidencia su carácter transversal.

En conjunto, esta amplia variedad de aplicaciones demuestra que el análisis arquetípico no solo constituye una herramienta metodológica robusta, sino también un marco flexible adaptable a datos heterogéneos y a problemas científicos de distinta naturaleza.

#### Orientaciones futuras y problemas pendientes

**Limitaciones actuales del AA**

1. **No convexidad del modelo**

* El AA no es convexo → riesgo de mínimos locales y falta de fiabilidad.
* Mitigación actual: inicializaciones múltiples y adecuadas.
* Futuro: explorar límites de convexidad basados en propiedades geométricas y relajaciones espectrales (similar al clustering).

1. **Número de arquetipos (K)**

* Difícil de estimar de forma automática.
* Hoy depende del conocimiento del dominio.
* Futura línea: métodos automáticos para seleccionar K y técnicas para transferir modelos entre diferentes órdenes (añadir/eliminar arquetipos sin recalcular todo).

1. **Extensiones necesarias**

* Aplicar AA a nuevos tipos de datos:
  + Datos censurados.
  + AA justo (fairness).
  + AA con privacidad diferencial.

1. **Dinámica temporal**

* Aunque AA se ha aplicado a series temporales, falta un marco robusto para:
  + Arquetipos que evolucionen en el tiempo.
  + Reconstrucciones que dependan de puntos previos (dependencias temporales).
* Futuro: integración con modelos de sistemas dinámicos.

1. **Dependencia de formas puras**

* Supuesto fuerte: las “formas puras” deben estar presentes o derivarse como combinaciones convexas.
* Problema: no siempre se cumplen esas condiciones.
* Futuro:
  + Relajar la restricción de convexidad (introducir δ en los coeficientes C\_k).
  + Combinar AA con métodos de volumen mínimo para extraer politopos aun si las formas puras no están en la envolvente convexa.

1. **Estructura de la variedad de datos**

* El AA convencional supone que los datos están en un politopo, pero en realidad suelen estar en una variedad.
* Limitaciones del kernel-AA: solo capta relaciones por pares.
* Futuro: AA profundo y combinación con aprendizaje de variedades (representaciones latentes no lineales).

**Resumen**

Si bien el análisis arquetípico (AA) ha demostrado ser una herramienta poderosa y versátil, presenta diversas limitaciones que abren nuevas líneas de investigación. Una de las principales dificultades radica en la no convexidad del modelo, lo que lo hace propenso a caer en mínimos locales y compromete la fiabilidad de los resultados. Este problema se suele mitigar mediante múltiples inicializaciones, pero se requieren avances metodológicos que exploren los límites de convexidad a través de propiedades geométricas o relajaciones espectrales.

Otro aspecto crítico es la determinación del número de arquetipos (K). Actualmente, esta elección depende en gran medida del conocimiento experto y del dominio de aplicación, lo cual limita la automatización. Sería deseable contar con métodos que permitan estimar K de forma objetiva y, además, mecanismos que faciliten la transferencia entre modelos con diferente número de arquetipos sin recalcular completamente la factorización.

En cuanto a las extensiones del modelo, existen oportunidades en la aplicación del AA a nuevos tipos de datos, como los datos censurados, así como en el desarrollo de variantes orientadas a la justicia algorítmica y la privacidad diferencial. De igual forma, aunque el AA se ha empleado en contextos temporales, todavía falta un marco sistemático que permita modelar de manera fluida la dinámica temporal de los arquetipos, integrando dependencias en el tiempo y conceptos de sistemas dinámicos.

Otra limitación significativa es la fuerte dependencia del AA respecto a la existencia de formas puras dentro de los datos o derivables mediante combinaciones convexas. En escenarios donde esta condición no se cumple, el AA puede fallar en su capacidad de representación. Para superar esta restricción, se plantea la necesidad de relajar los supuestos de convexidad, introducir parámetros de tolerancia y combinar el AA con enfoques de mínimo volumen, con el fin de extraer politopos más flexibles.

Finalmente, debe considerarse que el AA clásico asume que los datos residen en un politopo del espacio original, cuando en realidad suelen encontrarse en variedades no lineales. Aunque el kernel-AA constituye un primer paso al imponer estructuras de vecindad, su capacidad es limitada. Los desarrollos más recientes en AA profundo abren la posibilidad de integrar el aprendizaje de variedades con representaciones latentes más expresivas, lo que representa una dirección prometedora para futuras investigaciones.

En síntesis, el avance del análisis arquetípico depende de superar sus limitaciones teóricas y computacionales, ampliando su aplicabilidad a nuevos tipos de datos, mejorando su capacidad para modelar dinámicas temporales y extendiendo sus fundamentos a contextos no lineales y de mayor realismo.