UNIVERSITAT
JAUME I

Programa de Doctorat en Ciències

Escola de Doctorat de la Universitat Jaume I

Some contributions to archetypal analysis with applications

Memòria presentada per Ismael Cabero Fayos per a optar al grau de doctor per la Universitat Jaume I

Doctorand:
Ismael Cabero Fayos

Directora:
Irene Epifanio López

Castelló de la Plana, setembre 2019

Aquest treball està recolzat per les subvencions següents:

DPI2017-87333-R del Ministeri de Ciència, Innovació i Universitats (AEI / FEDER, UE) i UJIB2017-13 de la Universitat Jaume I.

Agraïments

Sempre m'he considerat una persona afortunada, però en aquest viatge acadèmic, anomenat doctorat, l'adjectiu hauria d'escriure's en forma superlativa.

He tingut la grandíssima sort de tindre al meu costat, com a Directora, a Irene Epifanio que no ha escatimat a l'hora de compartir la seua experiència, saviesa, paciència,... sempre amb un somriure (fins i tot digitalment :-)) i una dedicació insuperable. Sense tot el seu suport acadèmic i moral hauria sigut impossible fer ni una ínfima part del que hem aconseguit. Li estic profundament agraït.

A més, des del començament del trajecte (crec que sol ser un recorregut prou solitari), he anat agafat de la mà del meu amic Marc Pallarès, que m'ha guiat, m'ha aconsellat, m'ha ajudat... incansablement i de forma altruista, cada dia, sense fallar-ne ni un, més enllà del que hom podria pensar. Sóc un privilegiat!

Un altre puntal ha sigut la meua família, sempre ha confiat en mi i ha respectat la meua dedicació, esperem que el resultat revertisca en ells. Doneu sentit a tot el que faig.

Vull agrair també al professor Guillermo Ayala per la seua col·laboració desinteressada i per la seua eminència.

Així mateix no em puc deixar a un amic que m'ha ajudat a arribar fins al doctorat, al meu company de màster Aitor Alfonso, un altre bàcul en el qual recolzar-me i del que sempre he rebut ajuda.

Ha sigut un plaer poder col·laborar amb Ana Piérola i Alfredo Ballester.

A tots i totes elles, GRÀCIES per fer-ho possible.

Índex

Capítol 1

Introducció

1.1 Plantejament de la investigació

En les darreres dècades la millora en la velocitat de computació i la capacitat d'arreplegar una gran quantitat de dades han fet que s'incorporaren noves tècniques i algorismes per poder predir el comportament de les dades i extraure'n informació. L'aprenentatge automàtic (Machine Learning) és el camp de la intel·ligència artificial encarregat de crear aquests algorismes i l'aprenentatge estadístic (Statistical Learning) és la branca de l'estadística que dona resposta a l'aprenentatge automàtic emfatitzant els models estadístics i l'avaluació de la incertesa.

Tot aquest bullici de nous camps i nous aprenentatges estan sota el mant de l'anomenada Ciència de les dades (Data science) que fa de connexió entre totes les disciplines involucrades, dona ferramentes per millorar la presa de decisions i evita utilitzar la intuició (Sund; 2019). Dins d'aquest entramat, algunes de les tasques que estan reservades a les persones estadístiques són assegurar la fiabilitat de la informació extreta de les dades, treballar amb les dades no disponibles i altres fonts d'error no mostral, quantificar la incertesa de prediccions, previsions i models, descriure l'estructura de les dades, els paràmetres a trobar, crear mètodes per a inferència causal i efectivitat comparativa, combinar informació de múltiples fonts, i la determinació de tècniques efectives de visualització de dades... Com indica Weihs and Ickstadt (2018), l'estadística és una de les disciplines més importants per proporcionar eines i mètodes per trobar l'estructura i donar una visió més profunda de les dades, i la disciplina més important per analitzar i quantificar la incertesa.

Per ubicar-se i saber on s'inscriu aquesta tesi és important entendre que l'aprenentatge estadístic es pot classificar com un aprenentatge supervisat i no supervisat i depenent del tipus de dades de què disposem utilitzarem unes tècniques o unes altres; en el nostre cas ens centrarem en les no supervisades.

En l'aprenentatge supervisat tenim unes dades expressades en parells d'objectes, unes variables d'entrada (input) que estan relacionades amb una o unes variables d'eixida (output) i l'objectiu d'aquest aprenentatge és crear funcions o algorismes que prediguen resultats per a aquelles entrades en les quals desconeguem les eixides.

En l'aprenentatge no supervisat tan sols tenim un conjunt de dades entrada, no existeixen dades etiquetades i l'objectiu és trobar associacions i patrons entre aquest conjunt de dades d'entrada.

A partir dels noranta fou quan Cutler and Breiman (1994) presentaren una anàlisis estadística no supervisada novedosa, l'Anàlisi d'Arquetipus (AA). Anàlisi que cerca uns representants extrems que permeten explicar tots els elements de la mostra com a combinació convexa dels mateixos. D'aquesta manera, s'ha anat consolidant aquest tipus d'anàlisi gràcies a la seua validesa i a la seua capacitat per descodificar la informació oculta en les dades.

Les tècniques d'anàlisi arquetípiques es troben entre dues tècniques estadístiques no supervisades molt conegudes: Anàlisi de components principals (PCA) i anàlisi de clúster (CLA). En les tècniques de descomposició de dades per trobar les components latents, un conjunt de dades es considera com una combinació lineal de diversos factors. Diferents eines d'anàlisi prototípiques sorgeixen depenent de les restriccions sobre els factors i la seua combinació (Mørup and Hansen; 2012; Vinué et al.; 2015). Els factors amb menys restriccions són els produïts per PCA, ja que són combinacions lineals de variables. Un dels avantatges és que aquest fet ajuda a explicar la variabilitat de les dades; no obstant això, la interpretabilitat dels factors està compromesa. Al contrari, les restriccions més grans es troben en les eines de clúster, com ara k-means o k-medoids. Els seus factors són fàcilment interpretats perquè són centroides (mitjanes de grups de dades) o medoids (observacions concretes) en el cas de k-means i k-medoids, respectivament. El preu que les eines de clusterització paguen per la interpretabilitat és la pèrdua de flexibilitat de modelització a causa de l'assignació binària de dades als clústers. Les eines arquetípiques, d'altra banda, gaudeixen d'una flexibilitat de modelització més elevada que les eines de clúster, però sense perdre la interpretabilitat dels seus factors.

Encara que no és tan àmpliament aplicada com altres mètodes, AA és una poderosa eina per a la mineria de dades, reconeixement de patrons o classificació, ja que proporciona resultats fàcilment interpretables en l'anàlisi de components latents, reducció de dimensionalitat i agrupació.

Tal com ens mostra Davis and Love (2010) els humans interpretem amb més facilitat les dades més extremes pel principi dels oposats (Thurau et al.; 2012). Això és d'una gran rellevància per a l'AA, ja que els arquetipus que ens torna són models extrems que ens faciliten la interpretació i el descobriment de perfils que ens puguen descriure tots els elements.

AA s'ha aplicat a un ampli espectre de camps, com la biologia (D'Esposito et al.; 2012), astrofísica (Chan et al.; 2003), climatologia (Steinschneider and Lall; 2015), la psicologia del desenvolupament (Ragozini et al.; 2017), didàctica (Cabero; 2018), enginyeria (Epifanio et al.; 2013, 2018; Millán-Roures et al.; 2018), finances (Moliner and Epifanio; 2019), genètica (Thøgersen et al.; 2013), desenvolupament global (Epifanio; 2016; Epifanio et al.; 2019), problemes d'aprenentatge automàtic (Mørup and Hansen; 2012), investigació de mercat (Li et al.; 2003; Porzio et al.; 2008; Midgley and Venaik; 2013), resum de diversos documents (Canhasi and Kononenko; 2013, 2014), neurociències (Tsanousa et al.; 2015; Hinrich et al.; 2016) i esports (Eugster; 2012; Vinué and Epifanio; 2017) ente d'altres.

AA es va definir originalment per dades multivariants contínues. En aquesta tesi analitzarem quina és la millor opció si volem trobar perfils arquetípics quan les dades són binàries. A més, mostrarem els avantatges que aporta l'AA per resoldre diferents problemes, com són la segmentació de textures i la detecció d'outliers. La fonamentació d'aquest treball i la millora que representa aquesta anàlisi queda reflexada en tres investigacions independents amb diferents enfocaments.

En el primer article, fruit d'aquestes investigacions, treballem amb dades binàries i funcionals. La restricció que obliga l'AA a que els arquetipus siguen una combinació convexa dels elements de la base de dades, no és una condició necessaria perquè formen part d'ella i això pot ser un inconvenient en alguns camps, encara més si parlem de persones en les quals els arquetipus no garanteixen que hi haja persones reals que s'apropen a les característiques d'aquests representants (Seiler and Wohlrabe; 2013). L'any 2015, Vinué et al. (2015) varen proposar una variant a AA, l'anàlisi d'arquetipoids (ADA), una tècnica d'aprenentatge estadística no supervisada, en la qual els representants que es cerquen ja no poden ser teòrics, és necessari que formen part de les dades. Aquest fet facilita la comprensió del conjunt de dades en mantindre el fet de ser elements extrems i individus reals.

En el cas de la utilització de dades binàries, els arquetipus són una barreja de dades i no serien necessàriament vectors binaris, i com a conseqüència no serien interpretables. Però utilitzant ADA, els factors (arquetipoids) són casos reals, de manera que ADA es pot aplicar a dades binàries sense perdre la interpretabilitat dels factors. Mitjançant un estudi de simulació, compararem les diverses opcions.

Aquest treball l'hem ampliat amb un estudi amb dades funcionals (les dades binàries són convertides en dades funcionals i treballem en Anàlisi de dades funcionals (FDA)). En el context funcional, les funcions del conjunt de dades són aproximades per barreges de funcions arquetípiques, aplicant així l'Anàlisi Funcional d'Arquetipoids (FADA).

Presentem dues aplicacions reals i comparem les solucions ADA amb les d'altres tècniques no supervisades establertes per il·lustrar els avantatges de l'ADA i FADA en les ciències educatives i de comportament.

Amb ADA, obtindrem els perfils d'habilitats de l'alumnat. D'aquesta manera, els estudiants es poden agrupar per la similitud en el domini d'habilitats. Una forma clàssica d'agrupar els perfils establerts per l'habilitat de l'estudiantat és mitjançant un mètode de clúster, tal com ho fan Dean and Nugent (2013), però en termes d'interpretabilitat humana, els punts centrals retornats per les eines de clúster no semblen tan favorables com els punts extrems retornats per ADA. Per exemple, una anàlisi clúster podria diferenciar l'alumnat en tres grups, els que trauen bona nota, els que en trauen dolentes i els que tenen notes al voltant de la mitjana, però ADA els ha agrupat cercant també diferències qualitatives, en aquest cas revisant si l'alumnat necessita més suport d'anàlisi o d'àlgebra. Amb la utilització de les dades funcionals aconseguirem un marc de referència que no ens pot proporcionar PCA, ja que l'objectiu de PCA no és recuperar patrons extrems. De fet, les corbes amb puntuacions extremes de PCA no corresponen necessàriament a observacions arquetípiques. Això es discuteix a Cutler and Breiman (1994) i es mostra a Epifanio et al. (2013) mitjançant un exemple on els arquetips no es podrien recuperar amb PCA, fins i tot considerant totes les components.

En el segon article ens centrem en la segmentació d'imatges, i en particular en la segmentació de textures, una de les tasques més importants i difícils en el processament d'imatges, que de manera succinta consisteix a separar les diferents textures que es presenten a la imatge.

Treballem la segmentació no supervisada de textures, és a dir, quan no hi ha informació prèvia sobre les textures de la imatge. Hi ha dos enfocaments generals per a dur-la a terme, són els enfocaments basats en la regió o els enfocaments basats en fronteres. Considerem un enfocament basat en la regió que consisteix en el càlcul de característiques de textures en finestres petites centrades en cada píxel de la imatge o en una mostra de píxels, i després utilitzar una anàlisi clúster de les característiques d'elles (Soille; 2003).

Aquest procediment retorna una mostra de característiques que provenen de diferents tipus de textures pures (quan les finestres contenen un tipus de textura única) i també una mostra de característiques procedents d'una barreja de textures. Per aquest motiu, es proposa una tècnica d'aprenentatge estadístic no supervisada com a alternativa a l'anàlisi clúster: l'AA. L'objectiu d'AA és precisament extreure els arquetips, que són perfils purs en un conjunt de dades, i expressar les dades com a barreges d'aquests arquetips. Els arquetips són una barreja d'observacions del conjunt de dades. Per tant, amb AA, podem obtenir el conjunt de textures pures presents a la imatge i trobar les fronteres entre textures pures, com aquelles regions compostes per barreges de diversos arquetips. Es testarà la validesa d'aquesta metodologia en una comparativa amb altres alternatives en una aplicació en teledetecció.

En el tercer article presentem un nou algoritme per a la detecció no supervisada de valors estranys o atípics (outliers) en dades multivariants contínues basada en l'ús d'AA.

Un valor atípic és una observació que no sembla correspondre's amb la resta d'observacions del conjunt de dades objecte d'estudi. La detecció d'aquest tipus de dades és un dels problemes interessants que es presenten en l'anàlisi de dades. La relativa subjectivitat de l'anàlisi d'aquestes dades, planteja el dubte de si aqueixa observació, que no sembla correspondre's amb la resta d'observacions, és realment un valor atípic. En general, el tractament de la informació es realitza utilitzant les dades atípiques i es comparen les conclusions amb altres tractaments que no els utilitzen. Si aquestes són diferents, s'ha de realitzar una anàlisi més detallada d'aquests.

Realitzem una comparació experimental amb un gran conjunt de bases de dades conegudes i algoritmes estàndard. Aquesta avaluació proporciona resultats molt favorables a la nostra nova proposta. A més, apliquem el nou algorisme a un conjunt de dades originals de mesures del peu, que s'utilitza en un problema d'enginyeria que presentem aquí. La detecció d'outliers en Antropometria només s'ha utilitzat com a tècnica de neteja, per corregir o eliminar els valors molt inferiors o superiors abans d'analitzar dades (Kouchi; 2014; Kuehnapfel et al.; 2016). No obstant això, els outliers aporten informació molt valuosa en el procés de disseny del calçat, ja que poden indicar quins tipus de peus són més diferents de la resta i, per tant, quins individus poden tenir problemes en el calçat si el disseny no és apropiat.

1.2 Organització de la investigació

La distribució de la tesi continua amb el capítol 2 en el qual mostrem les bases teòriques que utilitzem en AA, ADA, FDA, en textures i en morfologia matemàtica. Es prossegueix amb el capítol 3 on es descriuen les aportacions, els objectius de la tesi i els tres estudis realitzats mostrant les dades utilitzades, la metodologia i els resultats. Per finalitzar, els capítols 4 i 5 presenten les conclusions d'aquesta tesi i els possibles treballs futurs de la investigació, respectivament.

Capítol 2

Fonaments

En aquest capítol pretenem fer un repàs als conceptes i resultats fonamentals que després utilitzarem al llarg de la memòria. En concret presentarem AA i ADA, tot seguit revisarem FDA, després l'ànalis de textures i per últim conceptes bàsics de morfologia matemàtica.

2.1 Anàlisi d'Arquetipus

Siga X una matriu $n \times m$ que representa una base de dades amb n observacions i m variables. El nostre propòsit serà trobar una matriu Z de dimensions $k \times m$ que caracteritze els patrons arquetípics de les dades de manera que cada dada es puga representar com una barreja d'aquests arquetipus. Específicament, com ens indica Vinué et al. (2015), l'anàlisi d'arquetipus cerca obtindre les dues matrius $n \times k$ dels coeficients α i β que minimitzen la suma residual dels quadrats que sorgeix de l'equació que mostra x_i com una aproximació d'una combinació convexa dels arquetipus z_j i les equacions que mostren z_j com a combinació convexa de les dades.

$$RSS = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2,$$

amb condicions

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ amb $\alpha_{ij} \geq 0$ i $i = 1, \dots, n$.

2)$\sum_{l=1}^{n} \beta_{jl} = 1$ amb $\beta_{jl} \geq 0$ i $j = 1, \dots, k$.

Aleshores per 1) tenim que les aproximacions de x_i són una barreja finita d'arquetipus $\hat{x}_i = \sum_{j=1}^{k} \alpha_{ij} z_j$ i les α_{ij} ens indicaran el pes que té cada arquetipus j per a l'individu i. Per un altre costat la restricció 2) ens indica que els arquetipus z_j són combinacions convexes de les dades, $z_j = \sum_{l=1}^{n} \beta_{jl} x_l$, i això fa que els arquetipus puguen no ser elements de la mostra.

2.2 Anàlisi d'Arquetipoids

Basant-se en Vinué et al. (2015), el problema d'optimització en l'anàlisi d'arquetipoids es transforma a minimitzar:

$$RSS = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2$$

sota les condicions

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ amb $\alpha_{ij} \geq 0$ i $i = 1, \ldots, n$

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ amb $\beta_{jl} \in \{0, 1\}$ i $j = 1, \ldots, k$ i.e. $\beta_{jl} = 1$ per a tan sols una l i la resta de $\beta_{jl} = 0$.

I ara sí que és absolutament necessari que els arquetipoids z_j siguen un element de les dades.

Si continuem amb l'explicació de Vinué et al. (2015) tenim que, a causa de les seves definicions, els arquetipus i els arquetipoids són representants extrems de les dades. Si k >1, els arquetipus se situen en la envoltura convexa de les dades (vegeu Cutler and Breiman (1994)), mentre que això no necessàriament té lloc per als arquetipoids (vegeu Vinué et al. (2015)). Si k = 1, la mitjana de les dades és l'arquetipus, i l'arquetipoid és el medoid de les dades (Kaufman and Rousseeuw; 1990).

Es va proposar un algorisme alternant de minimització per Cutler and Breiman (1994) per resoldre el problema d'AA. S'alterna entre estimar el millor α per arquetipus donats Z i els millors arquetipus Z per α determinats. Els problemes de mínims quadrats convexos es van resoldre amb una versió penalitzada de l'algorisme de quadrats mínims no negatius proposada per Lawson and Hanson (1974).

Eugster and Leisch (2009) implementa aquest algorisme a la biblioteca R **archetypes**, on les dades estan estandarditzades per defecte. Hem basat el nostre algorisme en aquest, però les dades no estan estandarditzades i es considera la norma Frobenius, tal com s'indica a l'equació (2.1), enlloc de la norma espectral utilitzada per Eugster and Leisch (2009).

En el cas d'ADA, per resoldre el problema , es va proposar en Vinué et al. (2015) un algorisme basat en la idea de l'algorisme de clúster de Particionar al voltant de Medoids (PAM)(Kaufman and Rousseeuw; 1990).

L'algorisme té dues fases:

(i) En la primera fase, BUILD, es selecciona un conjunt d'arquetipoids com a conjunt inicial.

(ii) En la segona fase, SWAP, es tracta de millorar la qualitat dels arquetipoids intercanviant-los per dades que redueixen l'RSS.

A partir d'un conjunt inicial d'arquetipoids computats en el pas BUILD, la fase SWAP millora aquest conjunt mitjançant l'intercanvi de casos seleccionats per observacions no seleccionades i comprovant si aquests reemplaçaments redueixen l'RSS. Aquest algorisme es va implementar a la biblioteca R **Anthropometry** by Vinué et al. (2017). Es consideren tres alternatives per a la fase BUILD en la implementació R.

Els primers candidats són els veïns més propers (utilitzant la distància euclidiana) als arquetipus, l'anomenat conjunt $cand_{ns}$.

Els segons candidats inicials, anomenats el conjunt $cand_{\alpha}$, són els casos amb el valor màxim de α per cada arquetipus j, és a dir, els casos amb la proporció relativa més gran per als respectius arquetipus.

El tercer conjunt de candidats, el conjunt $cand_{\beta}$, consisteix en les observacions amb el valor màxim de β per cada arquetipus j, és a dir, els principals contribuents en la generació dels arquetips. A partir d'aquests tres conjunts inicials, després de la fase SWAP, l'ADA torna tres conjunts d'arquetipoids. El conjunt amb RSS més baix (sovint el mateix conjunt s'obté de les tres inicialitzacions) és la solució ADA retornada.

Una pregunta oberta és el número k d'arquetipus o arquetipoids per calcular. Tingueu en compte que ni arquetipus ni arquetipoids estan necessàriament niats. Ho pot decidir l'usuari o bé el criteri de colze es podria utilitzar tal com ho va fer Cutler and Breiman (1994); Eugster and Leisch (2009); Vinué et al. (2015) (el valor k és seleccionat com el punt on es troba el colze sobre la representació RSS per a una sèrie de diferents valors de k).

2.3 Anàlisi funcional de dades

2.3.1 Introducció

Malgrat que les dades recollides dins de l'àmbit científic solen recollir-se de manera discreta, és a dir en certs instants de temps, moltes vegades ens interessa poder treballar aquesta informació com una funció contínua al llarg del temps i poder interpretar-la com una unitat, permetent-nos així tindre una visió global, conèixer valors d'aqueixa variable en qualsevol instant i així proveir al futur anàlisi de major integritat i certesa.

Les dades funcionals, és com es coneixen aquests tipus de dades, s'inscriuen dins de l'anàlisi de dades funcionals (FDA) de creació actual i sol utilitzar-se en casos com: l'anàlisi de la variació d'una magnitud al llarg dels anys (quantitat de pardals vistos, variació de l'alçada dels humans, . . .), indicadors esportius i econòmics, estudis meteorològics, etc.

El principi essencial de FDA rau a utilitzar una successió d'observacions individuals com tan sols una observació, oferint-nos els avantatges abans esmenats. El terme funcional fa referència a l'estructura interna de les dades en lloc de la seua forma explícita, és a dir, s'assumeix l'existència d'una funció que dóna lloc a les dades observades.

Realment, cada dada funcional es recull directament com a p parells (t_j, y_j) on y_j és el valor observat en el temps t_j, i anomenem soroll als errors observacionals que poden afectar els valors.

Si un conjunt de dades funcionals és una mostra, tindrem que cada individu de la mostra és una funció x_i constituïda per p_i parells de la forma (t_{ij}, y_{ij}), $j = 1, ..., p_i$. Com realment estem agrupant en una funció diferents valors discrets, no és necessari que els valors t_{ij} siguen iguals per a cada observació, el mateix passa amb l'interval T en el qual es recullen totes les dades.

Habitualment la variable contínua sobre la qual es registren les dades funcionals sol ser el temps, però també es poden registrar sobre altres variables com poden ser la posició, la freqüència, etc.

És molt important que les funcions siguen suaus, fet que implicarà que dos valors adjacents y_j i y_{j+1} siguen prou similars, perquè permetrà a les funcions que tinguen una o més derivades. Els gràfics de les derivades primera i segona com a funcions de t, o gràfics dels valors de la segona derivada com a funcions dels primers valors derivats, poden revelar aspectes importants dels processos que generen les dades. Si no poguérem tindre derivades en lloc de tractar les dades com a funcions les podríem tractar simplement com a dades multivariants.

No obstant això, les dades observades poden no ser del tot suaus a causa de l'error observacional. Cada observació y_{ij} la podem expressar com:

$$y_{ij} = x_i(t_{ij}) + \epsilon_{ij}$$

on $x_i(t_{ij})$ és el valor real de la funció x_i en el temps t_{ij} i ϵ_{ij} és l'error de mesurament. És absolutament bàsic i necessari en el moment d'utilitzar dades discretes com a funcionals filtrar aqueix soroll, que pot distorsionar els resultats fàcilment. Amb aquest objectiu, s'apliquen tècniques de suavitzat, és a dir, s'ajusten les dades a una base.

2.3.2 Exemple amb dades funcionals

Si revisem el nombre de persones individuals que han visitat el Centre Cultural la Beneficència de València en els darrers anys, obtenim la següent taula 2.1:

	Mesos											
	G	F	M	A	M	J	J	A	S	O	N	D
2017	3102	4439	5237	5669	4536	4171	4605	4757	4740	5204	4587	4343
2016	3392	3189	3817	3480	3745	3893	3287	3496	3744	5536	4072	3298
2015	2989	3506	3022	2858	2225	2399	2646	3109	2795	3655	3087	3598
2014	3729	3585	4565	3844	3659	3209	3420	3567	4401	5010	5611	3234

Taula 2.1: Nombre de visitants individuals per mesos al Centre Cultural La Beneficència de València des del 2014 fins el 2017. *FONT. Ajuntament de València* (2019).

Tanmateix, malgrat que la dada d'afluència de visitants és única per a cada mes, és interessant interpretar la mitjana de visites que s'han fet al cap de l'any com una funció al llarg del temps, és més fàcil distingir certes tendències que si ho fem com un vector de dades amb la mitjana de cada mes. Si transformem aquesta taula en dades funcionals obtenim la Fig. 2.1.

L'objectiu és obtindre funcions que ens ajuden a interpretar de manera més adient la distribució de les dades. En aquest cas ens mostren que la tendència és que març i octubre són els mesos amb més afluència, mentre que en maig-juny i desembre-gener són mesos amb afluència baixa. Evidentment estudiant períodes de temps més llargs podríem consolidar aquestes afirmacions i veure la tendència evolutiva dels visitants al museu.

S'ha realitzat una interpolació utilitzant B-splines cúbics amb 12 nodes, que serà explicada en la secció 2.3.7.

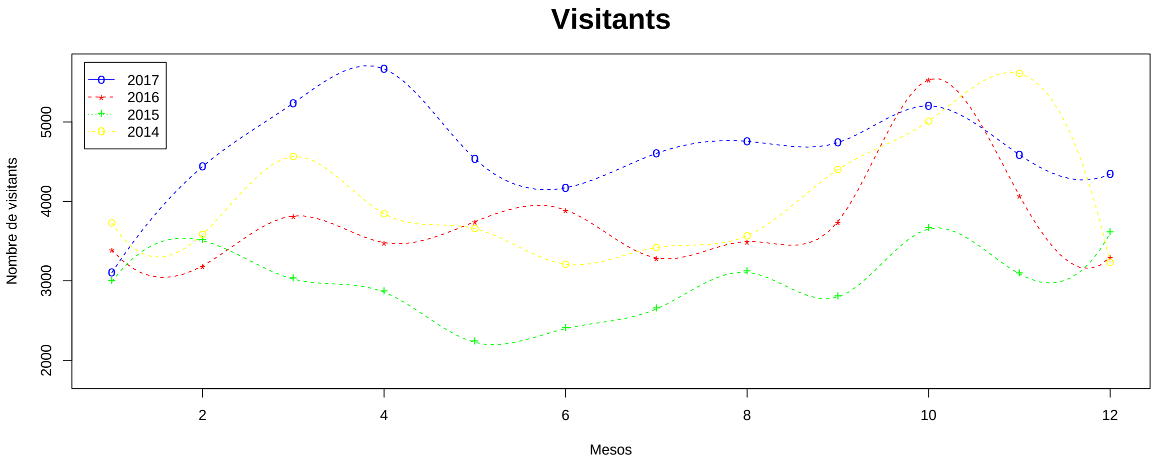


Figura 2.1: Afluència de visitants al Centre cultural La Beneficència de València des del 2014 fins el 2017.

2.3.3 Objectius de la FDA

L'anàlisi de la FDA, com a part de l'estadística comparteix els seus objectius generals des d'una perspectiva pròpia que millora i facilita l'anàlisi en certs aspectes. Aquests objectius els podríem resumir en els següents, entre d'altres:

- Representar les dades de manera que faciliten una anàlisi posterior.
- Mostrar les dades per a ressaltar diverses característiques.
- Descriu la naturalesa de les dades a analitzar.
- Estudiar patrons en les dades.
- Explorar la variació d'un conjunt de dades funcionals (components principals).
- Explicar el valor d'una variable dependent en funció de diverses variables independents, funcionals o no.
- Formar grups d'observacions similars (aprenentage estadístic no supervisat).
- Classificar una observació en una classe, on les classes estan predefinides a priori (aprenentage estadístic supervisat).
- Detectar dades atípiques (outliers).

Estadístic	**Estimació**
Mitjana	$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$
Variància	$var_x(t) = \frac{1}{n-1}\sum_{i=1}^{n}\left[x_i(t) - \bar{x}(t)\right]^2$
Desviació típica	$sd_x(t) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[x_i(t) - \bar{x}(t)\right]^2}$
Covariància	$cov_x(t_1, t_2) = \frac{1}{n-1}\sum_{i=1}^{n}\left[x_i(t_1) - \bar{x}(t_1)\right]\left[x_i(t_2) - \bar{x}(t_2)\right]$
Correlació	$corr_x(t_1, t_2) = \frac{cov_x(t_1,t_2)}{\sqrt{var_x(t_1)var_x(t_2)}}$

Taula 2.2: Estadístics per a dades funcionals.

En aquesta tesi es treballen alguns d'aquests objectius però específicament, amb les dades funcionals, anem a estudiar patrons de les dades i crear grups d'observacions similars (classificació no supervisada) amb l'objectiu d'extraure informació oculta.

2.3.4 Estadístics per a dades funcionals

Siga $x_i(t), i = 1, \ldots, n$, una mostra de dades funcionals. Per tal de definir els estadístics per dades funcionals, es generalitzen les definicions dels estadístics per a mostres multivariants, i es poden calcular punt a punt, tal i com s'ha recollit a la taula 2.2.

2.3.5 Representar funcions amb bases de funcions

Una vegada tenim les dades discretes, el primer que hem de fer és transformar-les en funcions que representaran cada individu, és a dir, a partir dels parells de valors (t_{ij}, y_{ij}) corresponents a cada individu de la mostra x_i amb $i = 1, ..., n$ on n és el nombre d'individus de la mostra, s'ha de representar cada element respecte a un sistema de funcions base.

Un sistema de funcions base és un conjunt de funcions conegudes ϕ_k independents unes de les altres. Permeten representar una funció $x(t)$ com una combinació lineal de funcions base de la següent manera:

$$x(t) \approx \sum_{k=1}^{K} c_k \phi_k(t)$$

on K és el nombre de funcions base ϕ_k utilitzades. Per tant, les funcions base tenen la propietat de poder aproximar arbitràriament bé qualsevol funció si es pren un número gran K d'aqueixes funcions i es trien adequadament els pesos c_k de la combinació lineal.

Si c és el vector de longitud K on els seus coeficients són les c_k utilitzades per a representar $x(t)$ i ϕ és el vector format per les funcions ϕ_k, podem expressar matricialment l'equació anterior utilitzant la transposada, que denotarem mitjançant una t:

$$x = \mathbf{c}^{\mathbf{t}}\phi = \phi^{\mathbf{t}}\mathbf{c}$$

Quan $K = p$ es duu a terme una interpolació lineal en el sentit que es trien els coeficients c_k per tal que $x(t_j) = y_j$ per a cada j. Per tant, el grau en el qual les dades y_j es suavitzen en oposició a la interpolació ve donada pel nombre de funcions base K que es consideren.

La similitud de característiques de les funcions base amb les propietats de les funcions que es volen estimar afavoreix i redueix el nombre K de funcions base necessàries per a tindre una aproximació satisfactòria i a més a més aqueixa reducció en el nombre de bases repercuteix en un menor cost computacional.

Malgrat que hi ha molts sistemes de bases; exponencials, bases de potències, bases polinòmiques, ondetes, etc. les més utilitzades són les bases de Fourier i les bases B-spline. Les primeres s'utilitzen per a descriure dades periòdiques, mentre que les segones es poden utilitzar quan les dades no siguen periòdiques.

2.3.6 Base de Fourier

Les sèries de Fourier és un dels sistemes de bases de funcions més utilitzats. Les funcions base són:

$$1, sin\omega t, cos\omega t, sin2\omega t, cos2\omega t, \ldots$$

Que les podríem generalitzar de la forma:

$$\begin{aligned}\phi_0(t) &= 1,\\ \phi_{2r-1}(t) &= sinr\omega t.\\ \phi_{2r}(t) &= cosr\omega t.\end{aligned}$$

Per tant, podem aproximar la funció $x(t)$ mitjançant $\hat{x}(t)$ com s'indica seguidament:

$$\hat{x}(t) = c_0 + c_1 sin\omega t + c_2 cos\omega t + c_3 sin2\omega t + c_4 cos2\omega t + \dots$$

Notem que en ser una base formada per funcions trigonomètriques sinus i cosinus, és una base periòdica. El paràmetre ω es coneix com la freqüència del senyal, indica el nombre de vegades que el senyal passa per un punt en la unitat de la variable independent del senyal. Segons si es mesura en hertzs ($Hz =$ voltes/segon) o en radians per segon, el període serà $\frac{1}{r\omega}$ o $\frac{2\pi}{r\omega}$.

La nostra base serà ortonormal si els instants de temps, t_j, en els quals han sigut registrats els valors de la mostra, estan equiespaiats en l'interval de temps amb el qual es treballa, T, i el període és igual a la longitud de l'interval T, dividint per les constants adequades: $\sqrt{p}$ per a j = 0 i $\sqrt{\frac{p}{2}}$ per a la resta de valors de j, és a dir, si es defineix Φ la matriu que conté els valors de les K funcions base, $\Phi^t\Phi = \mathbf{I}$ on $\mathbf{I}$ és la matriu identitat.

Si p és una potència de 2 i els instants en els quals s'obtenen les observacions de la mostra estan igualment espaiats, podem utilitzar l'algorisme de la Transformada Ràpida de Fourier (FFT) per a calcular els coeficients c_k de la sèrie de Fourier. Utilitzem aquesta casuística tan especial perquè ens permet reduir significativament el cost computacional. El nombre d'operacions necessàries per a calcular tots els coeficients c_k i les p dades suavitzades de cada observació $x(t)$ avaluada en els valors t_j és d'$O(plogp)$ en lloc d'$O(p^2)$ que es requeriria per avaluació directa de la fórmula de la transformada discreta de Fourier. Si considerem el nombre de mostres $p = 1024$ aquesta simplificació redueix el nombre d'operacions d'una mica més d'un milió a tan sols uns milers.

Si derivem les bases:

$$\begin{aligned}&\frac{d}{dt}(sinr\omega t) = r\omega cos(r\omega t),\\ &\frac{d}{dt}(cosr\omega t) = -r\omega sin(r\omega t).\end{aligned}$$

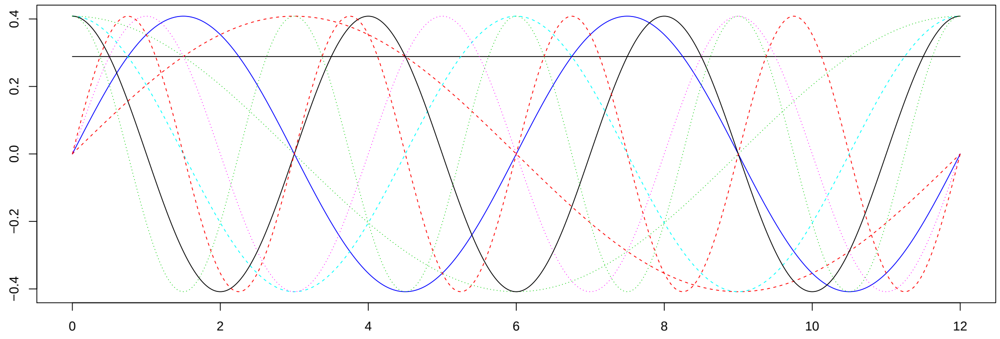


Figura 2.2: Bases de Fourier per a K=9.

els coeficients de la derivada de la base de Fourier venen donats per

$$(0, \omega c_1, -\omega c_2, 2\omega c_3, -2\omega c_4, \ldots)$$

La sèrie de Fourier és realment vàlida per a aproximar qualsevol funció, sobretot si són funcions estables sense característiques locals fortes, i computacionalment és molt eficient. S'utilitza en la majoria de camps de les ciències i té un pes rellevant en totes les enginyeries. No obstant això, no són adequades per a dades que presenten discontinuïtats bé en la funció o en les seues primeres derivades. Tampoc seria adeqüada si la funció no és periòdica, per això vorem a la secció següent altra base més addient per a eixes situacions.

En la Figura 2.2 es mostra la base de Fourier per a 9 funcions base.

2.3.7 Base de B-Splines

Les funcions spline són les funcions polinòmiques més utilitzades per a aproximar funcions no periòdiques. Gràcies al fet que ofereixen una major flexibilitat i al seu reduït cost computacional han aconseguit substituir pràcticament a les bases de polinomis. S'han desenvolupat sistemes de bases per a les funcions spline, el cost computacional de les quals és $O(p)$, la qual cosa resulta essencial perquè moltes aplicacions treballen amb grans quantitats d'observacions.

Introduirem l'estructura d'una funció spline i així posteriorment podrem explicar el sistema B-spline, que és el sistema de bases que es sol utilitzar per a construir splines.

Una vegada tenim l'interval sobre el qual es volen aproximar les funcions, haurem de dividir-lo en L subintervals separats pels valors o nodes τ_l on $l = 1, . . . , L$-1. El conjunt format per tots el nodes, rep el nom de vector nu.

Un spline serà un polinomi d'un grau específic m, que aproximarà a la funció en cada interval. Els splines de grau 0 són funcions constants, els de grau 1 són rectes, els de grau dos paràboles, i així successivament. L'ordre del polinomi és el nombre de constants necessàries per a definir-lo que equival al seu grau més 1. Els diferents splines adjacents aniran unint-se de forma suau proporcionant continuïtat de tipus C^{m-2} en cadascun dels nodes.

Es pot reduir la continuïtat en un node repetint el valor d'aqueix node. Si es té un spline d'ordre m amb un node de multiplicitat s (el valor del node es repeteix s vegades), la continuïtat en el node es redueix de C^{m-2} a C^{m-s-1} .

A mesura que augmenta l'ordre dels polinomis, l'aproximació millora, sempre i quan la posició dels nodes es mantinga. Tanmateix, s'ha de tindre en compte que el que fa guanyar felixibilitat és el nombre de nodes. Habitualment s'utilitzen nodes equiespaits però és més adequat utilitzar més nodes en els intervals on la funció tinga més variabilitat, això sí, cada interval hauria de contindre un valor inicial o dada. Aleshores podem afirmar que una funció spline està determinada per dos elements: a) L'ordre dels segments polinomials; b) El vector nu donat per la successió $\tau_0, ..., \tau_j$ amb $l = 1, . . . , L$ - 1.

El nombre de paràmetres per a definir una funció spline, en el cas en que cada node tinga tan sols multiplicitat u, ve donat per l'ordre més el nombre de nodes interiors, $m + L - 1$.

Bases B-spline per a les funcions spline

Malgrat que ja tenim definida el que és una funció spline, per a construir-la necessitarem un sistema de funcions base $\phi_k(t)$ que han de verificar les següents propietats:

- Cada funció base $\phi_k(t)$ ha de ser una funció spline definida amb un ordre m i un vector nu τ.
- Qualsevol combinació de les funcions base ha de ser una funció spline, ja que la suma i el producte de funcions spline és també una funció spline.
- Cada funció spline definida amb un ordre m i un vector nu τ s'ha de poder expressar com una combinació lineal d'aqueixes funcions base.

La manera més coneguda de construir funcions spline és utilitzant el sistema de bases B-splines que va ser desenvolupat per De Boor et al. (1978), que recursivament, a partir de B_0,

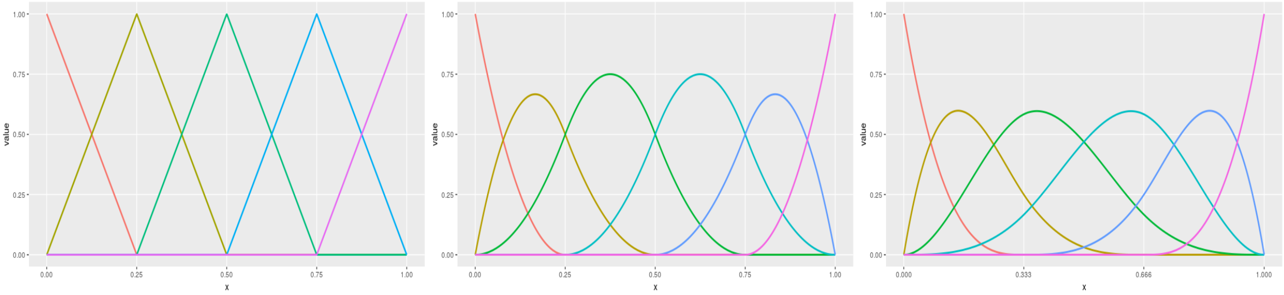


Figura 2.3: B-splines de grau 2, 3 i 4 per a un vector nu uniforme (amb nodes equiespaiats).

crea totes les funcions B-spline d'ordre superior. Una funció spline $S(t)$ amb nodes discrets interiors es defineix de la següent manera:

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau),$$

on $B_k(t, \tau)$ denota el valor del k-èssim B-spline en el punt t definit pel vector nu τ.

La figura 2.3 representa els B-splines lineals, quadràtics i cúbics amb tres nodes respectivament.

Si a un ajustament utilitzant B-splines se li augmenta el nombre de nodes o se li augmenta l'ordre dels B-splines deixant el vector nu intacte, els resultats tan sols poden millorar, però és notori tindre en compte que l'augment del número de B-splines no sempre produeix un millor ajust. És possible que la posició dels nodes faça que un sistema B-spline de menor dimensió millore l'ajustament que un sistema de dimensions superiors.

Això és pel fet que l'espai de funcions definit per K B-splines (amb la posició de nodes distinta) no sempre està contingut en el definit per $K + 1$ B-splines.

Com a exemple de la funcionabilitat dels splines, en la gràfica 2.4 mostrem una aproximació a la funció $y = cos(2x)$ en l'interval $[0, 2\pi]$.

Els ordres dels splines pels quals s'aproximen a la funció són de dalt a baix 2, 3 i 4 . Per a mostrar la importància del nombre de nodes, comparem l'aproximació utilitzant tres punts de tall o cinc punts de tall:

- En el primer cas l'interval queda subdividit en quatre subintervals. Si s'inclouen també 0 i 2π com a punts de tall, llavors es numeren com τ_0,...,τ_L on $L = 4$.

- En el segon cas l'interval queda subdividit en sis subintervals. Si s'inclouen també 0 i 2π com a punts de tall, llavors es numeren com τ_0,...,τ_L on $L = 6$.

2.3.8 Mètodes de suavitzat

Les dades funcionals, com a un conjunt de p parelles d'observacions de la forma (t_j, y_j) on y_j és el valor observat en el temps t_j, són recollides de forma discreta i per a transformar-les a una funció s'utilitza una tècnica que es coneix com suavitzat i que consisteix a ajustar les dades a una base de funcions, permetent també eliminar el soroll registrat en obtindre les observacions.

Existeixen diversos mètodes de suavitzat entre els quals destaquen el suavitzat per mínims quadrats, el suavitzat mitjançant kernels o el suavitzat per regularització. Nosaltres tan sols usarem el primer mètode, per a un estudi detallat d'altres mètodes es pot consultar Ramsay (2006).

2.3.9 Suavitzar dades funcionals per mínims quadrats

L'objectiu és ajustar una corba $x(t)$ a les observacions discretes y_j , j = 1, ..., p (on $y_j = x(t_j) + \epsilon_j$ sent $x(t_j)$ el valor de la funció $x(t)$ en l'instant t_j i sent ϵ_j l'error observacional corresponent a l'observació y_j) usant una combinació lineal de funcions base per a $x(t)$ de la forma

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c^t} \phi$$

Els vectors c i ϕ són de longitud K i contenen els coeficients c_k i les funcions base ϕ_k respectivament.

Definim Φ com la matriu que conté els valors de les K funcions base per als p punts de la mostra, és a dir, és una matriu $p \times K$ que conté els valors $\phi_k(t_j)$.

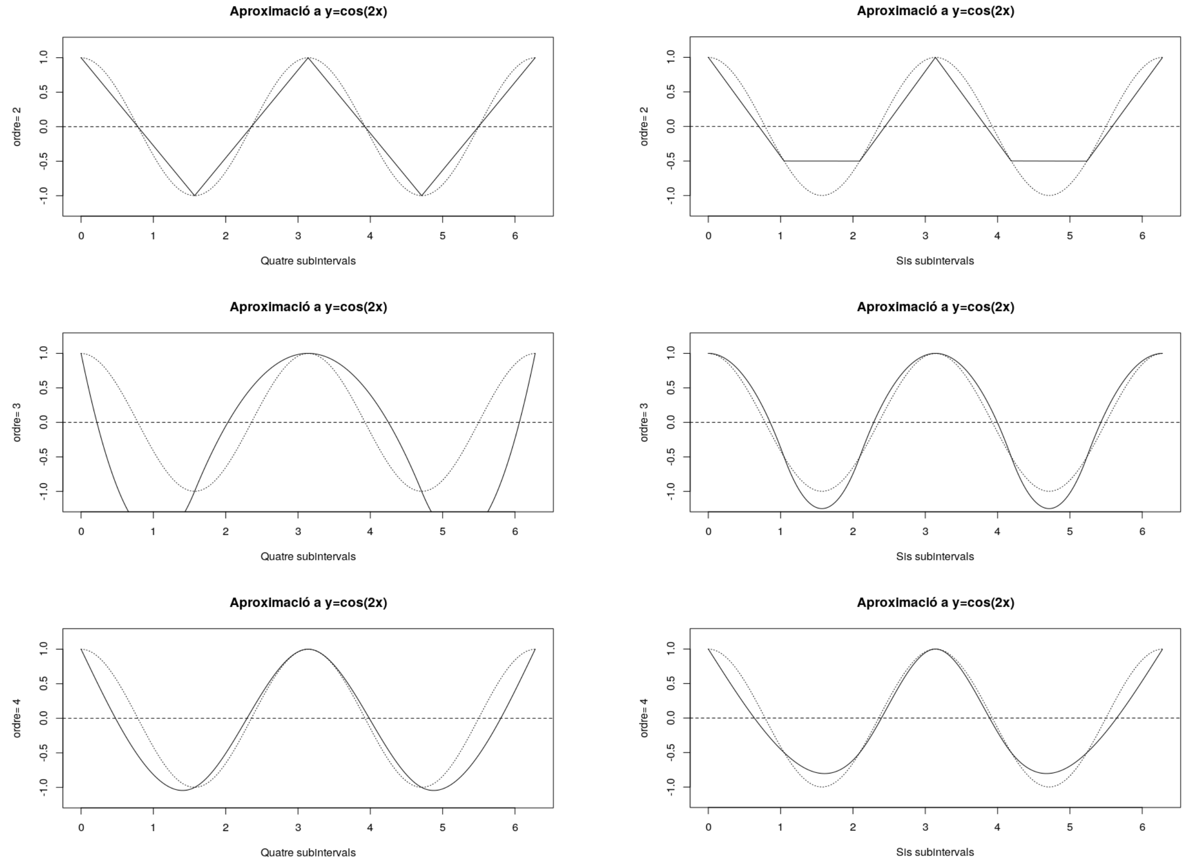


Figura 2.4: Aproximació mitjançant funcions splines a la funció $cos(2x)$ sobre l'interval $[0,\, 2\pi]$ amb ordres 2, 3 i 4 respectivament, i amb 4 i 6 subintervals. Les línies puntejades representen la funció que s'ha d'aproximar i les línies negres les aproximacions realitzades mitjançant les funcions spline.

Ajust per mínims quadrats no ponderats

Els coeficients c_k es determinen pel criteri de mínims quadrats

$$SMSSE(y \mid c) = \sum_{j=1}^{p} \left[y_i - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2 .$$

Expressat en forma matricial s'obté

$$SMSSE(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})^{\mathbf{t}} (\mathbf{y} - \mathbf{\Phi c}) = \left(\mathbf{y^t} - (\mathbf{\Phi c})^{\mathbf{t}}\right)(\mathbf{y} - \mathbf{\Phi c}) =$$
$$\left(\mathbf{y^t} - \mathbf{c^t \Phi^t}\right)(\mathbf{y} - \mathbf{\Phi c}) = \mathbf{y^t y} - \mathbf{y^t \Phi c} - \mathbf{c^t \Phi^t y} + \mathbf{c^t \Phi^t \Phi c}.$$

Prenent la derivada de $SMSSE(\mathbf{y} \mid \mathbf{c})$ respecte a c, obtenim l'equació.

$$2\Phi^t \Phi \mathbf{c} - 2\Phi^t \mathbf{y} = \mathbf{0}$$

i aillant $\mathbf{c}$ s'obté l'estimador $\hat{\mathbf{c}}$ que minimitza $SMSSE(\mathbf{y} \mid \mathbf{c})$

$$\hat{c} = \left(\Phi^t \Phi\right)^{-1} \Phi^t \mathbf{y}$$

La corba ajustada és

$$\hat{\mathbf{y}} = \Phi \hat{\mathbf{c}} = \Phi(\Phi^t \Phi)^{-1} \Phi^t \mathbf{y}.$$

L'aproximació per mínims quadrats és adequada quan els residus ϵ_j són independents i idènticament distribuïts amb mitjana zero i variància constant.

Ajust per mínims quadrats ponderats

Quan les variàncies dels errors no són constants o els errors no estan idènticament distribuïts, s'ha d'aportar un pes diferent als dels diferents residus. Per a això, s'estén el criteri de mínims quadrats de la forma

$$SMSSE(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})^{\mathbf{t}}\, \mathbf{W}(\mathbf{y} - \mathbf{\Phi c})$$

on $\mathbf{W}$ és una matriu simètrica i definida positiva (per a tots els vectors $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v^t W v} > \mathbf{0}$). Si la matriu de variàncies i covariàncies Σ_e per als residus ϵ_j és coneguda llavors

$$\mathbf{W} = \sum\nolimits_{\mathbf{e}}^{\mathbf{-1}},$$

perquè tota matriu definida positiva és invertible (el seu determinant és positiu), i la seua inversa és definida positiva.

En aplicacions on no és factible estimar Σ_e, s'assumeix que les covariàncies entre els errors són zero i en aqueix cas $\mathbf{W}$ és diagonal amb la variància de l'error associat als y_j's en la diagonal.

D'aquesta manera, derivant l'expressió $SMSSE(\mathbf{y} \mid \mathbf{c})$ i aillant $\mathbf{c}$, l'estimador de mínims quadrats ponderats per als coeficients c_k és

$$\hat{\mathbf{c}} = (\Phi^t \mathbf{W \Phi})^{\mathbf{-1}} \mathbf{\Phi^t W y}.$$

2.4 Anàlisi de Textures

Malgrat que el concepte de textura no té una definició exacta en el processament d'imatges, la idea subjacent és que són aquelles característiques que es repeteixen com a patrons locals.

L'aproximació estadística i la sintàctica (Haralick; 1979) són les més destacades per a la descripció de textures. En el primer cas la textura es descriu mitjançant un vector de propietats que representa la textura com un punt en un espai de característiques multidimensional (Sonka et al.; 1993). En la segona aproximació, es tracta de determinar les primitives que constitueixen la textura i les seues regles de col·locació.

Per a enfrontar-se al problema d'anàlisi de textures, malgrat que no ho fem de manera exhaustiva (tampoc és la pretensió d'aquesta tesi), podem classificar els mètodes en quatre grans grups (Sonka et al.; 1993; Tuceryan and Jain; 1993): mètodes estadístics, geomètrics, basats en models i de processament del senyal.

2.4.1 Mètodes estadístics

Els mètodes estadístics treballen amb la distribució espacial dels nivells de gris i entre els més coneguts cal citar: les matrius de co-ocurrència i la funció d'autocorrelació.

2.4.2 Mètodes geomètrics

Els mètodes geomètrics consideren la textura composta de 'elements' o primitives. Una vegada identificats aquests elements, la textura pot analitzar-se sota dues perspectives: computar propietats estadístiques dels elements extrets o bé extraure la regla de col·locació que descriga la textura, aquesta última aproximació pot comportar mètodes geomètrics o sintàctics.

2.4.3 Mètodes basats en models

Aquests mètodes es basen en la construcció d'un model, els paràmetres estimats del qual descriurien les qualitats de la textura. Els mètodes basats en els camps aleatoris de Markov o els fractals, es considerarien en aquesta metodologia.

2.4.4 Mètodes de processament del senyal

Les tècniques que hem vist fins ara operen sobre el senyal definit en el domini espacial. Tal vegada podria ser interessant representar la imatge en un altre domini, de manera que en aquest nou domini ens facilite l'extracció de determinades característiques. Moltes de les tècniques d'aquest apartat consisteixen a calcular certes característiques de les imatges després d'haver sigut filtrades, a més algunes es recolzen en el fet que en una primera etapa, el cervell humà duu a terme una anàlisi freqüencial de la imatge. Dividirem aquesta secció en tres apartats.

1. **Filtres en el domini espacial**
   Dins d'aquesta categoria podem incloure des de mètodes que mesuren la densitat de vores a través de la magnitud de les respostes de màscares de Robert i laplaciana (Gibson; 1950; Laws; 1980), fins a mètodes basats en moments espacials (Laws; 1980) i altres obtinguts mitjançant filtres espacials i operadors no lineals (Unser and Eden; 1990).

2. **Filtres en el domini de Fourier**
   Seguint els resultats psicovisuals, s'han desenvolupat sistemes d'anàlisis de textures de filtrat en el domini de Fourier per a obtindre diferents característiques. Cada filtre és selectiu tant en freqüència com en orientació (Coggins and Jain; 1985).

3. **Filtres Gabor i onetes (wavelets)**
   La transformada de Fourier proporciona una anàlisi de la freqüència global. Hi ha vegades que és més apropiat que l'anàlisi estiga localitzada en el domini espacial.

Una forma clàssica d'aconseguir-ho, és mitjançant la transformada de Fourier amb finestra. Quan la funció finestra és una gaussiana, la transformació és una funció de Gabor (Gabor; 1946). Una funció de Gabor de dues dimensions (Daugman; 1985) definida en el domini espacial ($\mathbf{x} = (x, y)$), s'assoleix mitjançant el producte de dues funcions:

- Una funció harmònica bidimensional, l'oscil·lació de la qual s'estén perpendicularment al vector $\mathbf{f_0}$ =(f_{ox}, f_{oy}) amb freqüència $|\mathbf{f_0}|$, amb origen en un determinat $\mathbf{x_0} = (x_0, y_0)$ i fase ϕ.
- Una gaussiana de base el·líptica centrada en el punt $\mathbf{x_0}$ i l'eix major de la qual forma un angle θ amb l'eix X.

$$G(\mathbf{x}, \mathbf{x_0}, \mathbf{f_0}, \sigma_{\mathbf{x}}, \sigma_{\mathbf{y}}, \theta, \phi) = \mathbf{e}^{-\pi\left(\frac{\mathbf{x}'^{\mathbf{2}}}{\sigma_{\mathbf{x}}^{\mathbf{2}}}+\frac{\mathbf{y}'^{\mathbf{2}}}{\sigma_{\mathbf{y}}^{\mathbf{2}}}\right)}\mathbf{e}^{\mathbf{i}(\mathbf{2}\pi\mathbf{f_0}(\mathbf{x}-\mathbf{x_0})+\phi)},$$

on $\mathbf{x}' = (x', y')$ apareix d'una traslació i un gir d'eixos, és a dir,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} cos(-\theta) & sen(-\theta) \\ -sen(-\theta) & cos(-\theta) \end{pmatrix} \cdot \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}.$$

Específicament, podem construir diversos filtres de Gabor variant la longitud d'ona (λ), l'orientació (θ), la fase (ϕ) i les desviacions típiques de la gaussiana (σ), usant:

$$G(x, y|\lambda, \theta, \phi, x_0, y_0) = \exp\left(\frac{(x - x_0)^2 + (y - y_0)^2}{-2\sigma^2}\right) \sin\left(\frac{2\pi}{\lambda}(x\cos\theta - y\sin\theta) + \phi\right)$$

Varen ser Turner (1986) i Clark and Bovik (1987) els que proposaren emprar filtres de Gabor en l'anàlisi de textures. També s'han utilitzat onetes (wavelets) (Mallat; 1989) de manera exitosa en l'anàlisi de textures. Podem veure exemples de l'ús d'ambdues transformades en Jain and Farrokhnia (1991); Chang and Kuo (1993); Unser (1995) o Portilla and Simoncelli (2000).

És ben sabut que hi ha una gran varietat d'aplicacions de les textures en l'àrea del procés d'imatge, des de problemes de classificació, segmentació, síntesi, compressió, anàlisi de forma (per a extraure informació de la superfície), recuperació d'imatges en una base de dades, etc. (Karu et al.; 1996). El ventall on posar en pràctica l'aplicació de les textures és immens, des d'aplicacions industrials fins i tot en el camp mèdic. Existeixen aplicacions en inspecció de la qualitat en indústries tèxtils i automobilístiques (Dewaele et al. (1988); Chetverikov (1988); Chen and Jain (1988); Conners et al. (1983); Siew et al. (1988); Jain et al. (1990); Wood (1990)), multitud d'aplicacions mèdiques (Sutton and Hall (1972); Harms et al. (1986); Landeweerd and Gelsema (1978); Insana et al. (1986); Chen et al. (1989); Lundervold (1992); Miller and Astley (1992); Toulson and Boyce (1992)), en procés de documents (Wang and Srihari (1989); Wahl et al. (1982); Fletcher and Kasturi (1988); Taxt et al. (1989); Jain and Bhattacharjee (1992); Jain and Farrokhnia (1991)), recuperació d'imatges d'una base d'imatges (Puzicha et al. (1997)) i per descomptat en teledetecció (Haralick et al. (1973); Rignot and Kwok (1990); Schistad and Jain (1992); Du (1990); Lee and Philpot (1990); Monjoux and Rudant (1991)), entre altres.

2.5 Morfologia matemàtica

Hi ha moltes maneres d'estudiar una imatge, una d'elles és la Morfologia Matemàtica, que ens permet transformar la nostra imatge en una altra i, triant adequadament la transformació, destacar algun tret de la nostra primitiva imatge que resultarà senzill de mesurar. El procés lògic serà doncs:

Imatge	Transformació Morfològica	Paràmetre
X	$T(X)$	$\mu(T(x))$

2.5.1 Transformacions en tot o res

Si analitzem una imatge binària X i la transformem mitjançant l'estudi de les relacions que lliguen als punts que la formen i que constitueixen la seua estructura, aquesta transformació rep el nom de *tot o res*, que seqüencialment realitza els següents passos:

1. Prenem un conjunt B de geometria coneguda que conté a l'origen i al qual anomenem **element estructurant.**

2. Desplacem B de manera que l'origen passa per tots els punts de l'espai; B_x serà el resultat d'aquest desplaçament en cada cas.

3. En cadascuna d'aquestes posicions ens preguntem si els conjunts $B_x \cap X$, $B_x \cap X^c$ o $B_x \cup X$ verifiquen certa condició.

4. Aquells punts x en els quals la resposta és *afirmativa* passen a formar part de $T(X)$.

Erosió (ϵ): Es defineix l'*erosió* de X mitjançant l'element estructurant B com el conjunt de punts x tals que $B_x \subset X$. Com a transformació en tot o res, suposa preguntar-se en cada desplaçament si $B_x \cap X^c = \emptyset$. Designem per $\epsilon^B(X)$ al resultat de la transformació, també conegut com l'*erosionat* de X. Aleshores,

$$\epsilon^B(X) = \cap_{y \in \check{B}} X_y = X \ominus \check{B}.$$

on $\check{B}$ és el simètric de B respecte a l'origen, $\check{B} = \{-y; y \in B\}$.

Dilatació (δ): La *dilatació* de X mitjançant B, $\delta^B(X)$, es defineix com el conjunt dels x tals que B_x toca a X, és a dir, $B_x \cap X \neq \emptyset$.

$$\delta^B(X) = \{x; B_x \cap X \neq \emptyset\} = X \oplus \check{B}.$$

Obertura (γ): Es defineix l'*obertura* de X mitjançant B, $\gamma^B(X)$, com la dilatació de l'erosionat, és a dir,

$$\gamma^B(X) = (X \ominus \check{B}) \oplus B = \delta^{\check{B}}(\epsilon^B(X)).$$

2.5.2 Morfologia per a imatges a nivells de gris

Les tècniques morfològiques van ser desenvolupades originalment per a conjunts o, la qual cosa és equivalent, per a imatges binàries. Va ser a mitjans dels anys setanta quan es van estendre aquests conceptes a imatges amb nivells de gris.

Per a entendre les transformacions morfològiques de funcions cal definir el concepte de subgraf. Siga $\{f(x) : x \in \mathbb{R}^2\}$ una imatge a nivells de gris (és a dir, amb $f(x) \geq 0$), aleshores

$$U_f = \{(x, t) : t \leq f(x)\},$$

és el subgraf de f. Conèixer la imatge f és equivalent a conèixer U_f.

2.5.3 Erosió i dilatació per elements estructurants plans

B és un element estructurant pla (com pot ser un cercle, segment, quadrat, $\ldots$). Erosionem el conjunt U_f mitjançant el conjunt B i obtenim el conjunt

$$\epsilon^B U_f = U_f \ominus \check{B}.$$

Es comprova que aquest conjunt és el subgraf de la funció

$$\epsilon^B f(x) = f(x) \ominus \check{B} = inf\{f(u) : u \in B_x\}$$

que anomenem funció erosionada de f mitjançant B.

Aleshores, l'erosió de f per B és $[\epsilon^B(f)](x) = \min_{b\in B} f(x+B)$ i la dilatació de f per B és $[\delta^B(f)](x) = \max_{b\in B} f(x+b)$.

2.5.4 Obertura

Combinant erosió i dilatació podem definir l'obertura d'una funció a nivells de gris. L'obertura de f mitjançant B, es defineix com

$$\gamma^B f(x) = \delta^{\check{B}} \epsilon^B f(x).$$

Capítol 3

Aportacions

3.1 Objectius de la investigació

L'objectiu general d'aquesta tesi és demostrar la validesa de l'Anàlisi d'Arquetipus en diferents vessants del coneixement en els que encara no s'ha implementat i contribuir així de manera original al seu desenvolupament. Aquest objectiu comú l'hem desglossat en tres investigacions en les que hem fet que aquest algorisme treballe amb diverses dades i amb diferents propòsits:

- Comprovar la validesa d'ADA (una variant de l'AA) amb dades binàries i demostrar la intuïtivitat dels seus resultats.
- Comprovar la validesa de FADA (la mateixa variant de l'AA, però ara utilitzant funcions com a dades).
- Demostrar la transcendència de la utilització d'AA en la segmentació de textures.
- Demostrar que en la detecció mutivariant d'outliers, l'AA és una ferramenta útil.
- Utilitzar amb bases de dades reals l'anàlisi AA i demostrar la seua vàlua tant amb dades binàries, funcionals i contínues.

Tots aquests objectius han quedat testimoniats amb investigacions quantitatives que han donat pas a la redacció de tres articles.

1r article Les dades d'enquestes binàries són de gran importància en les ciències socials. Moltes dades en brut d'exàmens, enquestes d'opinió, qüestionaris d'actitud, etc. es presenten en forma de matriu de dades binàries, és a dir, les respostes de l'examinant estan codificades com 0/1 (1 si contesta correctament, en cas contrari 0). La matriu binària es pot veure des de dos punts de vista. En el primer, l'interès resideix a les files, és a dir, a les persones; mentre que, en la segona, l'interès resideix en les columnes que contenen els elements o variables. En ambdós casos, l'anàlisi de dades exploratòries (EDA) té com a objectiu trobar informació en dades i generar idees (Unwin; 2010). Per ser útil com a eina per a EDA en conjunts de dades, l'eina ha de ser senzilla i fàcil d'usar, amb pocs paràmetres i que revelen les característiques més destacades de les dades de manera que els humans puguen visualitzar-les (Friedman and Tukey; 1974).

Així com hem explicat a la Secció 1.1, en la primera investigació, per primera vegada, proposem l'ús de l'Anàlisi d'Arquetipoids (ADA) per a aquest tipus de dades, per tal d'entendre, descriure, visualitzar i extreure informació que siga fàcilment interpretable, fins i tot per no experts. La nostra base dades es basa en un qüestionari matemàtic passat a estudiants novells de la UJI (Universitat Jaume I de Castelló), on les correccions aportaven dades binàries (0 resposta incorrecta, 1 resposta correcta). El nombre total d'alumnes dels que tenim dades és de 690 i el qüestionari conté 21 preguntes que es poden classificar sobretot dins de les branques d'anàlisi i àlgebra. A més a més hem fet un estudi comparatiu de l'anàlisi d'arquetipoids amb altres anàlisis estadístiques que serveixen per a treballar dades booleanes amb diferents objectius, però sempre de forma no supervisada, és a dir, sense que hi haja una variable resposta. Concretament les anàlisis treballades a banda d'ADA són; Anàlisi Clúster (PAM), HOMALS (acrònim de homogeneity analysis by means of alternating least squares (anàlisi d'homogeneïtat mitjançant mínims quadrats alternants)) i PAA (Probabilistic archetypal analysis).

A la segona aplicació, que correspon al segon punt de vista de la matriu binaria, utilitzem les mateixes dades i enfocaments seguits per Ramsay and Silverman (2002) i Rossi et al. (2002). Les dades utilitzades són les respostes 0/1 (incorrectes / correctes) de 2115 homes de l'administració d'una versió del Test de Matemàtiques de 60 ítems de l'American College Testing Program. Encara que aquesta matriu binària no sembla curvilínia a primera vista, fent el supòsit simplificador que les probabilitats P_{ih} (la probabilitat que l'examinant h responga l'element i correctament) varien d'una manera unidimensional entre els examinats, podem estimar la curva d'espai de capacitat. Aleshores, podem treballar amb les funcions de resposta de l'element (IRFs) $P_i(\theta)$ com a dades funcionals (Ramsay and Silverman; 2005), on θ és la variable que mesura les posicions al llarg de la corba d'espai de capacitat. O millor, podem treballar amb les funcions de registre odds-ratio $W_i(\theta)$, ja que aquestes transformacions de les funcions de resposta de l'element tenen la variació sense restriccions que solem veure en les corbes observades habitualment. Ramsay and Silverman (2002) i Rossi et al. (2002) utilitzen PCA funcional (FPCA) per estudiar variacions entre aquestes funcions. En canvi, es proposa utilitzar ADA funcional (FADA), que revela patrons molt interessants que no es van descobrir quan FPCA es va usar. AA i ADA es van ampliar a dades funcionals per Epifanio (2016).

En el context funcional, les funcions del conjunt de dades s'aproximen amb barreges (mixtures) de funcions arquetípiques.

Podem utilitzar diversos enfocaments per calcular IRFs. Un enfocament comú és estimar els IRFs de forma paramètrica, és a dir, les corbes són modelades per un conjunt de paràmetres que s'estimen, com en el model logístic de tres paràmetres (3PL). Un altre enfocament important és calcular els IRFs de forma no paramètrica, per exemple, per suavitzat kernel (Ramsay; 1991). Ramsay (1997) i Rossi et al. (2002) defensen la flexibilitat dels mètodes no paramètrics contra les restriccions dels paràmetres. En qualsevol cas, podem aplicar ADA als IRFs estimats per qualsevol mètode seleccionat per l'investigador.

Comparem els resultats amb aquests dos mètodes d'estimació també. La finalitat d'ambdues aplicacions, així com és objectiu comú de l'aprenentatge estadístic no supervisat, seria descobrir el millor mètode per a descriure l'organització interna de les dades.

2n article Les imatges són font d'informació i la capacitat d'extreure-hi informació és una àrea bàsica d'aplicació en la tecnologia de la imatge digital. Un dels primers passos per entendre la imatge és la segmentació de la imatge, en la que podem dividir la imatge en algunes àrees que tenen les mateixes característiques. En aquest treball, ens centrem en la segmentació no supervisada de textures, és a dir, quan no hi ha informació prèvia sobre les textures de la imatge.

Les regions d'una imatge són un grup de píxels connectats amb propietats similars. En els enfocaments basats en la regió, cada píxel s'assigna a un objecte o regió concret, és a dir particionen una imatge en regions que són properes a un conjunt de criteris predefinits (Janowczyk et al.; 2011).

Per a calcular aquestes regions, és habitual emprar una anàlisi de clúster de les característiques obteses en finestres xicotetes centrades en cada píxel i agrupant-les en grups amb característiques similars. La nostra proposta serà enlloc d'utilitzar clústers utilitzar AA ja que una vegada tingam definits els arquetipus (textures extremes i aleshores "pures"), quan les finestres siguen una barreja d'aquestes textures, tindrem dues avantatges:

- per un costat separarem les textures diferents que correspondran amb cada arquetipus.
- per un altre l'AA podrà contrastar exactament quin pes de cada textura pura té, ja que podrà definir cada finestra com una combinació convexa de les característiques de les textures pures que formen la barreja.

3r article Ara que estem en l'era de l'anomenada "Big Data" és important realitzar una anàlisi de qualitat previ a la ingent quantitat de dades que tenim per evitar prendre decisions equivocades. Una de les possibles causes d'aquestes decisions són els "outliers" o dades atípiques que ja hem presentat a la introducció 1.1. En molts casos els outliers es consideren els valors més extrems com a sorolls o errors, però sovint incorporen informació vital. Però és clar que el soroll per a algú pot ser un focus d'interès per a un altre (Johnson et al.; 1998), i depenent del que estigueu estudiant (per exemple, en la detecció de frau en targetes de crèdit) és un fet molt important.

Trobar els outliers en dades univariants és relativament senzill perquè és fàcil trobar els casos extrems. Tanmateix, en el cas multivariant, la detecció de valors extrems és més difícil, ja que els outliers multidimensionals són observacions que es consideren estranyes, no pel valor que prenen en una determinada variable, sinó pel valor dins del conjunt (Gnanadesikan and Kettenring; 1972; Campbell; 1978).

S'han proposat moltes tècniques per a la detecció d'outliers al llarg del temps (vegeu Aggarwal (2017) per obtenir una explicació detallada de molts d'ells per a diferents tipus de dades). En el cas de dades multivariants, Goldstein and Uchida (2016) van revisar i comparar molts dels algorismes de detecció d'anomalies no supervisats més estàndard en un conjunt de dades de referència. Goldstein and Uchida (2016) van proposar una taxonomia d'algoritmes de detecció d'anomalies no supervisats, que es divideixen en quatre categories: (1) tècniques basades en veïns més propers (NN), (2) mètodes basats en clustering, (3) algorismes estadístics i (4) tècniques de projecció en subespai.

Proposem un nou algorisme sense supervisió que detecte outliers en dades multivariants contínues. Es pot classificar en diverses d'aquestes categories, principalment (1) i (4), ja que utilitza una tècnica d'aprenentatge no supervisada (una tècnica de projecció en un subespai), que també es pot veure com una tècnica de cluster (Epifanio et al.; 2019), i també utilitza tècniques basades en NN. Tingueu en compte que les tècniques basades en distàncies són molt populars a causa dels seus bons resultats, la seva simplicitat conceptual i la seva interpretabilitat. Tanmateix, quan el nombre de variables és elevat, aquestes tècniques poden fallar a causa de la complexitat de la multidimensionalitat. Un punt clau per resoldre aquest problema seria reduir les dimensions i utilitzar subespais, on els outliers es puguen revelar fàcilment. Aquesta és la idea de l'algorisme proposat: primer en projectar dades en subespais rellevants i, a continuació, utilitzar tècniques basades en la proximitat per detectar outliers en aquests subespais.

L'algorisme proposat es basa en l'Anàlisi d'Arquetips (AA). Els arquetips es troben a la frontera de l'envoltura convexa de les dades, per tant són perfils extrems i això fa que siguen sensibles als outliers. AA no és una tècnica paramètrica, és un mètode basat en dades, de manera que no hem de fer cap presumpció sobre la distribució de dades. Per tant, la combinació d'AA juntament amb els mètodes de proximitat dóna com a resultat un mètode no paramètric amb un alt nivell d'interpretabilitat, que és molt important en moltes aplicacions.

3.2 Aportació I: Finding archetypal patterns for binary questionnaires [1]

[1] Actualment Cabero and Epifanio (2019c) està sotmés i a més es presentà a la XVII Conferència Espanyola i VII Trobada Iberoamericana de Biometria CEB-EIB 2019 (Cabero and Epifanio; 2019b).

Finding archetypal patterns for binary questionnaires

Ismael Cabero[1], Irene Epifanio[2]

Abstract

Archetypal analysis is an exploratory tool that explains a set of observations as mixtures of pure (extreme) patterns.If the patterns are actual observations of the sample, we refer to them as archetypoids. We propose to use archetypoid analysis for binary observations. This tool can contribute to the understanding of a binary data set, as in the multivariate case. We illustrate the advantages of the proposed methodology in a simulation study and two applications, one exploring objects (rows) and the other exploring items (columns). One is related to determining student skill set profiles and the other to describing item response functions.

MSC:

62H99, 62P25

Keywords:

dichotomous item test, archetypal analysis, functional data analysis, item response theory, skill profile

[1] Department de Didàctica de les Matemàtiques, Universitat de València, Spain

[2] Departament de Matemàtiques-IMAC, Universitat Jaume I, Castelló 12071, Spain. Email: epifanio@uji.es)

1. Introduction

Mining binary survey data is of utmost importance in social sciences. Many raw data from exams, opinion surveys, attitude questionnaires, etc. come in the form of a binary data matrix, i.e. examinees' responses are coded as 0/1 (1 if examinee i answers item h correctly, otherwise 0). The binary matrix can be viewed from two points of views. In the first, the interest lies in the rows, i.e. in the people, while in the second, the interest lies in the columns that contain the items or variables. In both cases, exploratory data analysis (EDA) aims to find information in data and generate ideas (Unwin, 2010). In order to be useful as a tool for EDA on data sets, the tool should be simple and easy to use, with few parameters, and reveal the salient features of the data in such a way that humans can visualize them (Friedman and Tukey, 1974).

For the first time, we propose the use of the exploratory tool Archetypoid Analysis (ADA) for this kind of data in order to understand, describe, visualize and extract information that is easily interpretable, even by non-experts. ADA is an unsupervised statistical learning technique (see Hastie et al. (2009, Chapter 14) for a complete review of unsupervised learning techniques). Its objective is to approximate sample data by a convex combination (a mixture) of k pure patterns, the archetypoids, which are extreme representative observations of the sample. Being part of the sample makes them interpretable, but also being extreme cases facilitates comprehension of the data. Humans understand the data better when the observations are shown through their extreme constituents (Davis and Love, 2010) or when features of one observation are shown as opposed to those of another (Thurau et al., 2012).

ADA was proposed by Vinué et al. (2015) as a derivative methodology of Archetype Analysis (AA). AA was formulated by Cutler and Breiman (1994), and like ADA, it seeks to approximate data through mixtures of archetypes. However, archetypes are not actual cases, but rather a mixture of data points. Recently, Seth and Eugster (2016b) proposed a probabilistic framework of AA (PAA) to accommodate binary observations by working in the parameter space.

AA and ADA have been applied to many different fields, such as astrophysics (Chan et al., 2003), biology (D'Esposito et al., 2012), climate (Steinschneider and Lall, 2015; Su et al., 2017), developmental psychology (Ragozini et al., 2017), e-learning (Theodosiou et al., 2013), finance (Moliner and Epifanio, 2019), genetics (Thøgersen et al., 2013), human development (Epifanio, 2016; Epifanio et al., 2018), industrial engineering (Epifanio et al., 2013, 2018; Millán-Roures et al., 2018), machine learning (Mørup and Hansen, 2012; Seth and Eugster, 2016a,b; Ragozini and D'Esposito, 2015), market research (Li et al., 2003; Porzio et al., 2008; Midgley and Venaik, 2013), multi-document summarization (Canhasi and Kononenko, 2013, 2014), nanotechnology (Fernandez and Barnard, 2015), neuroscience (Tsanousa et al., 2015; Hinrich et al., 2016) and sports (Eugster, 2012; Vinué and Epifanio, 2017).

Archetypal analysis techniques lie somewhere in between two well-known unsupervised statistical techniques: Principal Component Analysis (PCA) and Cluster Analysis (CLA). In data decomposition techniques, a data set is viewed as a linear combination of several factors to find the latent components. Different prototypical analysis tools arise depending on the constraints on the factors and how they are combined (Mørup and Hansen, 2012; Vinué et al., 2015). The factors with the least restrictions are those produced by PCA, since they are linear combinations of variables. One of the advantages is that this helps explain the variability of the data; however, the interpretability of the factors is compromised. Instead, the greatest restrictions are found in cluster tools, such as k-means or k-medoids. Their factors are readily interpreted because they are centroids (means of groups of data) or medoids (concrete observations) in the case of k-means and k-medoids, respectively. The price that clustering tools pay for interpretability is their modeling flexibility due to the binary assignment of data to the clusters. Archetypal tools, on the other hand, enjoy higher modeling flexibility than cluster tools but without losing the interpretability of their factors. A table summarizing the relationship between several unsupervised multivariate techniques is provided by Mørup and Hansen (2012) and Vinué et al. (2015).

1.1. *Illustrative example*

In Figure 1 a toy two-dimensional data set is used to illustrate what archetypoids mean and the differences compared with CLA and PCA, as well as to provide some intuition on what these pure and extreme patterns imply in behavioral sciences. Two numeric variables are considered from the data set personality-1.0 of the R package **neuropsychology** (Makowski, 2016), which contains personality traits data from an online questionnaire: Empathy.Agreeableness and Honesty.Humility. We apply k-means and ADA with $k = 3$, i.e. we find 3 clusters and archetypoids. We also apply PCA. Archetypoids are people with extreme values, which have clear profiles: archetypoid 1 is characterized by a very low Empathy.Agreeableness value together with a high Honesty.Humility value (1, 5.25), archetypoid 2 has the maximum values for both Empathy.Agreeableness and Honesty.Humility (7,7), while the third archetypoid has a very high Empathy.Agreeableness value together with the lowest Honesty.Humility value (6,0). Archetypoids are the purest people. The rest of the individuals are expressed as mixtures (collected in alpha coefficients; this is explained in Section 2) of these ideal people. For example, an individual with values of 6.25 and 0.75 for Empathy.Agreeableness and Honesty.Humility, respectively, is explained by 11% of archetypoid 2 plus 89% of archetypoid 3.

This is compatible with the natural tendency of humans to represent a group of objects by its extreme elements (Davis and Love, 2010). Figure 1 d) shows the partition

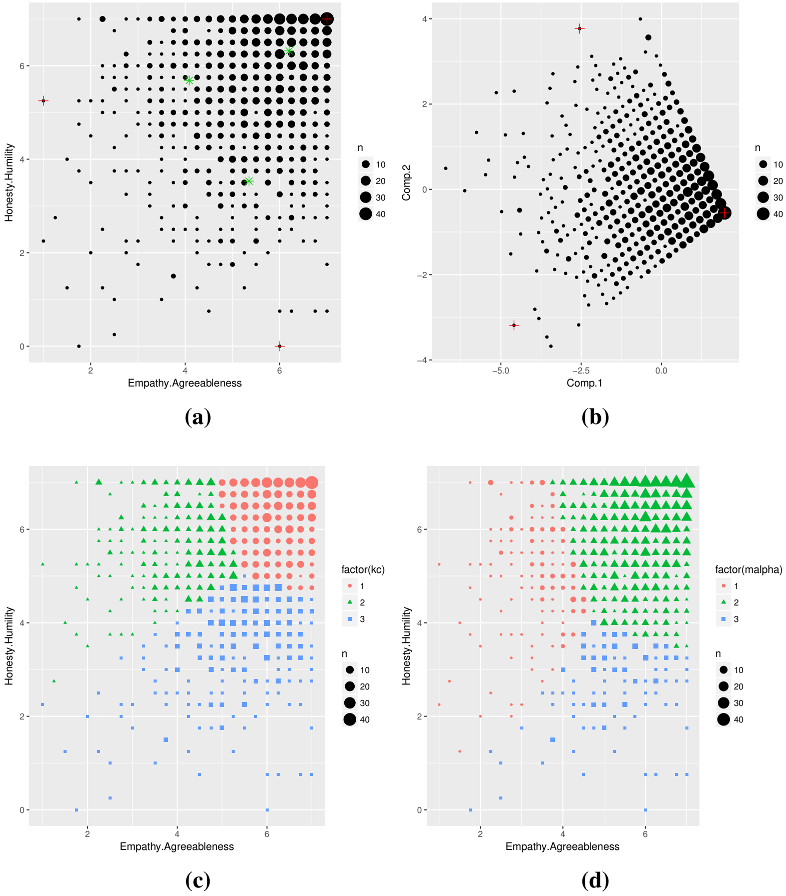


Figure 1. *(a) Plot of the toy example. The size of the points depends on their frequency. The red crosses represent the archetypoids, while the green stars represent the centroids of each cluster; (b) PC scores. Projected archetypoids are represented by red crosses; (c) k-means cluster assignments; (d) ADA assignments by the maximum alpha (see Section 2), i.e. assigned to the archetypoid that best explains the corresponding observation.*

of the set generated by assigning the observations to the archetypoid that best explains each individual. However, when we apply *k*-means to this kind of data set, without differentiated clusters, the centroids are in the middle of the data cloud. Centroid profiles are not as differentiated from each other as archetypoid profiles. This happens because centroids have to cover the set in such a way that the set is partitioned by minimizing the distance with respect to the assigned centroid (see Wu et al. (2016) about the connection between set partitioning and clustering). On the one hand, this means that the set partition generated by *k*-means and ADA would be different (Figures 1 c) and d)). On the other hand, centroids are not the purest, and therefore their profiles are not as clear as those of archetypoids. For example, centroids 2 and 3 have values (4.1, 5.7) and (5.3, 3.5), which are not as intuitively interpretable as archetypoids. If we look again at the individual with values (6.25, 0.75) from the clustering point of view this individual is clearly assigned to cluster 3, with centroid (5.3, 3.5), but clustering does not say anything about the distance of this point with respect to the assigned centroid, or in which direction they are separated. In fact, (6.25, 0.75) is quite far from (5.3, 3.5). This happens because the objective of clustering is to assign the data to groups, not to explain the structure of the data more qualitatively. Finally, note that archetypoids do not coincide with the individuals with the most extreme PC scores (see Figure 1 b)).

In summary, depending on our objective, the appropriate analysis should be selected. The objective of PCA is to reduce data dimension. Although PCA returns the location of the observations in the new dimensions by PC scores, there is no guarantee that the principal components are interpretable. In other words, observations are expressed in a new base, but in general the PCA base is not easily interpretable. However, the objective of CLA is to segment the data, i.e. to make groups of data by finding modes in data. Although the modes can be easily interpretable, CLA does not return an expression about the location of each observation with respect to each mode. On the other hand, finding extreme profiles, which are easily interpretable, is not the objective of PCA or CLA, but that of AA or ADA. These techniques also return the location of the observations as a function of the extreme profiles, in fact as a mixture (a convex combination), which is more easily interpretable than a linear combination. This provides a complete overview of the data set, generally supported by visual methods, i.e. this allows data to tell us more beyond the formal modeling or hypothesis testing task.

1.2. Our approach and applications

AA and ADA were originally thought of for real-valued observations. For AA, as the factors (archetypes) are a mixture of data, they would not necessarily be binary vectors, and as a consequence they would not be interpretable. In ADA though, the factors (archetypoids) are actual cases, so ADA can be applied to binary data without losing the interpretability of the factors. To perform a sanity check and provide insight we analyze the solutions obtained by AA, PAA and ADA through a simulation study. Furthermore, we present two real applications and compare ADA solutions with those of other established unsupervised techniques to illustrate the advantages of ADA in educational and behavioral sciences, when used as another useful tool for data mining in these fields (Slater et al., 2017).

In the first application, which corresponds to the first point of view of the binary matrix (analysis of the rows), we analyze the data set described by Orús and Gregori (2008), which was obtained through the application of a test on the initial mathematics skills of 690 first-year students of the College of Technology and Experimental Sciences at Jaume I University (Spain) at the beginning of the 2003-04 academic year. The test consisted of 17 questions corresponding to 21 single items, the answers to which were coded as 0 (incorrect or unanswered) or 1 (correct). The items of the test were selected in order to ascertain some given didactic hypotheses on the didactic discontinuities between mathematics at pre-university and university levels. The complete description of the questions can be seen in Orús and Gregori (2008). With ADA, we could obtain students' skill set profiles. In this way, students can be grouped by their similar mastery of skills. For instance, students showing consistently high levels of aptitude may be selected for an advanced class or students with similar difficulties could receive extra instruction as a group and also teaching strategies could also be adapted to suit their level. A classical way to group student skill set profiles is by using a clustering method, as carried out by Dean and Nugent (2013), but in terms of human interpretability, the central points returned by clustering tools do not seem as favorable as the extreme points returned by ADA. Results from different exploratory tools are compared.

In the second application, which corresponds to the second point of view of the binary matrix (analysis of the columns), we use the same data and approach followed by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002), although another strategy could be considered (Ramsay and Wiberg, 2017). The data used are the 0/1 (incorrect/correct) responses of 2115 males from administration of a version of the American College Testing Program 60-item Mathematics Test. Although this binary matrix does not seem curvaceous at first sight, by making the simplifying assumption that the probabilities P_{ih} (probability that examinee h gets item i right) vary in a smooth one-dimensional way across examinees, we can estimate the ability space curve that this assumption implies. Then, we can work with item response functions (IRFs) $P_i(\theta)$ as functional data (Ramsay and Silverman, 2005), where θ is the charting variable that measure out positions along the ability space curve. Or rather, we can work with the log odds-ratio functions $W_i(\theta)$, since these transformations of the item response functions have the unconstrained variation that we are used to seeing in directly observed curves. Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) used functional PCA (FPCA) to study variations among these functions. Instead, we propose to use functional ADA (FADA), which reveals very interesting patterns that were not discovered with FPCA. AA and ADA were extended to functional data by Epifanio (2016). In the functional context, functions from the data set are approximated by mixtures of archetypal functions.

As mentioned above, we can use other approaches to estimate IRFs. One common approach is to estimate IRFs parametrically, i.e. curves are modeled by a set of parameters that are estimated, as in the three-parameter logistic (3PL) model. Another important approach is not to assume any mathematical form and to estimate IRFs nonparametrically, for example by kernel smoothing (Ramsay, 1991). Ramsay (1997) and Rossi et al. (2002) defend the flexibility of nonparametric methods compared with the restrictions of parametric methods. In any case, we can apply ADA to the IRFs estimated by any method selected by the researcher. We also compare the results using these two estimation methods. Note that in the literature, we find other terms for IRFs, such as option characteristic curves, category characteristic curves, operating characteristic curves, category response functions, item category response functions or option response functions (Mazza et al., 2014).

The outline of the paper is as follows: In Section 2 we review AA and ADA for real-valued multivariate and functional data and PAA. In Section 2.4 we introduce the analysis for binary multivariate data. In Section 3, a simulation study with binary data compares the different strategies for obtaining archetypal patterns. In Section 4, our proposal is applied to two real data sets and compared to the results of other well-known unsupervised statistical learning techniques. Section 5 contains conclusions and some ideas for future work.

The data sets and code in R (R Development Core Team, 2018) for reproducing the results for both artificial and real data are available at `http://www3.uji.es/~epifanio/RESEARCH/adaedu.rar`.

2. Archetypal analysis

2.1. AA and ADA in the real-valued multivariate case

Let $\mathbf{X}$ be an $n \times m$ real-valued matrix with n observations and m variables. Three matrices are established in AA: a) the k archetypes $\mathbf{z}_j$, which are the rows of a $k \times m$ matrix $\mathbf{Z}$; b) an $n \times k$ matrix $\alpha = (\alpha_{ij})$ with the mixture coefficients that approximate each observation $\mathbf{x}_i$ by a mixture of the archetypes ($\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$); and c) a $k \times n$ matrix $\beta = (\beta_{jl})$ with the mixture coefficients that characterize each archetype ($\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$). To figure out these matrices, we minimize the following residual sum of squares (RSS) with the respective constraints ($\|\cdot\|$ denotes the Euclidean norm for vectors):

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2, \qquad (1)$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

As previously mentioned, archetypes do not necessarily match real observations. Indeed, this will only happen when one and only one β_{jl} is equal to one for each archetype, i.e. when each archetype is defined by only one observation. So, in ADA the previous constraint 2) is substituted by the following one, and as a consequence in ADA a mixed-integer problem is optimized instead of the AA continuous optimization problem:

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k$.

As regards the location of archetypes, they are on the boundary of the convex hull of the data if $k > 1$ (see Cutler and Breiman (1994)), although this does not necessarily happen for archetypoids (see Vinué et al. (2015)). Nonetheless, the archetype is equal to the mean and to the medoid in case of the archetypoid (Kaufman and Rousseeuw (1990)), if $k = 1$.

We want to emphasize that archetypal analysis is an EDA technique based on a geometric formulation (no distribution of data is assumed). It is not an inferential statistical technique, i.e. it is not about fitting models, parameter estimation, or testing hypotheses. Nevertheless, a field to study in the future would be to view archetypal analysis as a feature extraction method (Hastie et al., 2009, Ch. 5), where the raw data are preprocessed and described by α, which can be used as inputs into any learning procedure for compositional data (Pawlowsky-Glahn et al., 2015).

2.1.1. Computation of AA and ADA

The estimation of the matrices in the AA problem can be achieved by means of an alternating minimizing algorithm developed by Cutler and Breiman (1994), where the best α for given archetypes $\mathbf{Z}$ and the best archetypes $\mathbf{Z}$ for a given α are computed by turns. To solve the convex least squares problems, a penalized version of the non-negative least squares algorithm by Lawson and Hanson (1974) is used. Eugster and Leisch (2009) implemented that algorithm in the R package **archetypes**, although with some changes. Specifically, the data are standardized and the spectral norm in equation 1 is used instead of Frobenius norm for matrices. In our R implementation those changes were annulled, i.e. the data are not standardized by default and the objective function to minimize is defined by equation 1.

With respect to the estimation of the matrices in the ADA problem, it can be achieved using the algorithm developed by Vinué et al. (2015). It is composed of two steps: the BUILD step and the SWAP step. The objective of the BUILD step is to determine an initial set of archetypoids that will be upgraded during the following step. The objective of the SWAP step is to improve the primitive set by exchanging the selected instances for unselected observations and checking whether these replacements decrease the RSS. Vinué (2017) implemented that algorithm in the R package **Anthropometry** with three possible original sets in the BUILD step: $cand_{ns}$, $cand_{\alpha}$ and $cand_{\beta}$. These sets correspond to the nearest neighbor observations in Euclidean distance to the k archetypes, the cases with the maximum α value for each archetype j and the observations with the maximum β value for each archetype j, respectively. Then three possible solutions are obtained once these three sets go through the SWAP step, but only the solution with lowest RSS (often the same final set is returned from the three initializations) is chosen as the ADA solution.

One important point is the selection of k, since archetypes are not necessarily nested and neither are archetypoids. If the user has prior knowledge of the structure of the data, the value of k can be chosen based on that information. Otherwise, a simple but effective heuristic (Cutler and Breiman, 1994; Eugster and Leisch, 2009; Vinué et al., 2015; Seth and Eugster, 2016b) such as the elbow criterion can be used. With the elbow criterion, we plot the RSS for different k values and the value of k is selected as the point where the elbow is located.

2.2. Probabilistic archetype analysis

The idea underlying PAA is to work in a parameter space instead of the observation space, since the parameter space is often vectorial even if the sample space is not. The key is to assume that data points come from a certain distribution (from the Bernoulli distribution in the case of binary observations). Then the maximum likelihood estimates of the parameters of the distributions are seen as the parametric profiles that best describe each observation, and archetypal profiles are computed in the parameter space by maximizing the corresponding log-likelihood under the constraints for α and β. In summary, probabilistic archetypes lie in the parameter space, whereas classical archetypes lie in the observation space. Thus, archetypal profiles for binary data are the probability of a positive response. Details can be found in Seth and Eugster (2016b).

2.3. AA and ADA in the functional case

In Functional Data Analysis (FDA) each datum is a function. Therefore, the sample is a set of functions $\{x_1(t), ..., x_n(t)\}$ with $t \in [a,b]$, i.e. the values of the m variables in the standard multivariate context are replaced by function values with a continuous index t. We assume that these functions belong to a Hilbert space, satisfy reasonable smoothness conditions and are square-integrable functions on that interval. Simplistically, the sums are replaced by integrals in the definition of the inner product.

In functional archetype analysis (FAA), we seek k archetype functions that approximate the functional data sample by their mixtures. In other words, the objective of FAA is the same as AA, but now both archetypes ($z_j(t)$) and observations ($x_i(t)$) are functions. As a consequence, RSS is now calculated with a functional norm instead of a vector norm. We consider the L^2-norm, $||f||^2 =< f, f >= \int_a^b f(t)^2 dt$. The interpretation of matrices α and β is the same as in the classical multivariate case.

Analogously, FADA is also a generalization of ADA, where k functional archetypoids, which are functions of the sample, approximate the functions of the sample through the mixtures of these functional archetypoids. Again, vectors are replaced by functions and vector norms by functional norms, and the matrices are interpreted is the same way as before.

To obtain FAA and FADA in a computationally efficient way (Epifanio (2016)), functional data are represented by means of basis functions (see Ramsay and Silverman (2005) for a detailed explanation about smoothing functional data). Let B_h (h = 1, ..., m) be the basis functions and $\mathbf{b}_i$ the vector of coefficients of length m such that $x_i(t) \approx \sum_{h=1}^m b_i^h B_h(t)$. Then, RSS is formulated as (see Epifanio (2016) for details):

$$RSS = \sum_{i=1}^n ||x_i - \sum_{j=1}^k \alpha_{ij} z_j||^2 = \sum_{i=1}^n ||x_i - \sum_{j=1}^k \alpha_{ij} \sum_{l=1}^n \beta_{jl} x_l||^2 = \sum_{i=1}^n \mathbf{a}_i' \mathbf{W} \mathbf{a}_i, \qquad (2)$$

where $\mathbf{a}'_i$ = $\mathbf{b}'_i - \sum_{j=1}^k \alpha_{ij} \sum_{l=1}^n \beta_{jl} \mathbf{b}'_l$ and $\mathbf{W}$ is the order m symmetric matrix with the inner products of the pairs of basis functions w_{m_1,m_2} = $\int B_{m_1} B_{m_2}$. If the basis is orthonormal, for instance the Fourier basis, $\mathbf{W}$ is the order m identity matrix and FAA and FADA can be estimated using standard AA and ADA with the basis coefficients. If not, $\mathbf{W}$ has to be calculated previously one single time by numerical integration.

2.4. Archetypal analysis for binary data

Let $\mathbf{X}$ be an $n \times m$ binary matrix with n observations and m variables. The idea behind archetypal analysis is that we can find a set of archetypal patterns, and that data can be expressed as a mixture of those archetypal patterns. In the case of binary data, on the one hand the archetypal patterns should also be binary data, as the population from which data come. For example, if pregnancy was one of the binary variables, it would not make sense to consider as an archetypal observation a woman who was pregnant 0.7. In other words, archetypal patterns should be binary in order to have a clear meaning and not lose their interpretability, which is the cornerstone of archetypal techniques, i.e. they should not be 'mythological', but rather something that might be observed. On the other hand, in order to describe data as mixtures, we should assume that observations exist in a vector space, i.e. that observations can be multiplied by scalars (in this case in the interval $[0,1]$) and added together.

A solution that meets all these ideas is to apply ADA to $\mathbf{X}$, since the feasible archetypal patterns belong to the observed sample. In fact, ADA was originally created as a response to the problem in which pure non-fictitious patterns were sought (Vinué et al., 2015).

Instead, the archetypes returned by applying AA or PAA do not need to be binary, i.e. they do not need to belong to the feasible set of solutions. In fact, Seth and Eugster (2016b) binarized the archetypes obtained by AA or PAA in experiments. However, using a continuous optimization problem to solve a problem whose feasible solutions are not continuous can fail badly (Fletcher, 2000, Ch. 13). Indeed, there is no guarantee that this approach will provide a good solution, even by examining all the feasible binary solutions in a certain neighborhood of the continuous solution.

Therefore, we propose to use ADA to handle binary observations.

3. Simulation study

We have carried out a simulation study to assess all the alternatives in a controlled scenario. The design of the experiment has been based on simulation studies that appear in Vinué et al. (2015) and Seth and Eugster (2016b). We generate $k = 6$ archetypes, ζ_i, with $m = 10$ binary variables by sampling them from a Bernoulli distribution with a probability of success $p = 0.7$, $\mathbf{A} = [\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5, \zeta_6]$. Given the archetypes, we generate $n = 100$ observations as the binarized version of $\mathbf{x_i} = \mathbf{\tilde{A}_i h_i} + \mathbf{E_i}$, where $\mathbf{\tilde{A}_i}$ contains the archetypes after adding salt and pepper noise to them, $\mathbf{h_i}$ is a random vector sampled from a Dirichlet distribution with $\alpha = (0.8, 0.8, 0.8, 0.8, 0.8, 0.8)$, and $\mathbf{E_i}$ is a 10-dimensional random vector of Gaussian white noise with a mean of zero and standard deviation of 0.1. The binarized versions are obtained by replacing all values above 0.5 with 1 and others with 0. The noise density added to $\mathbf{A}$ is 0.05 (the default value used in MATLAB). With salt and pepper noise, a certain amount of the data is changed to either 0 or 1. To ensure that $\mathbf{\tilde{A}_i}$'s are archetypes, we chose $\alpha = 0.8$, a value near to but less than one.

We compute PAA, AA and ADA. The archetypes returned by PAA and AA are binarized for comparison with the true ones, $\mathbf{A}$. We calculated the Hamming distance (Manhattan distance between binary vectors), which is the same as the misclassification error used with binary images, between each archetypal solution and the true archetypes after permuting the columns of each archetypal solution to match the true archetypes in such a way that the least error with the city block distance is provided.

This was repeated 100 times. The first 10 times are displayed in Figure 2. The solutions returned by all the methods are quite similar to the true archetypes, i.e. the number of errors (a zero in the solution where the true value is 1, or vice versa) is very small. Nevertheless, there are differences between the methods, which are more evident in columns 5 and 6. For columns 5 and 6, the number of errors for PAA is 5 and 5, it is 4 and 2 for AA, but only 2 and 2 for ADA. Table 1 shows a the summary of the misclassifications. The archetypoids returned by ADA match the true archetypes better than those returned by AA or PAA, in this order, i.e. ADA provides the smallest mean misclassification error.

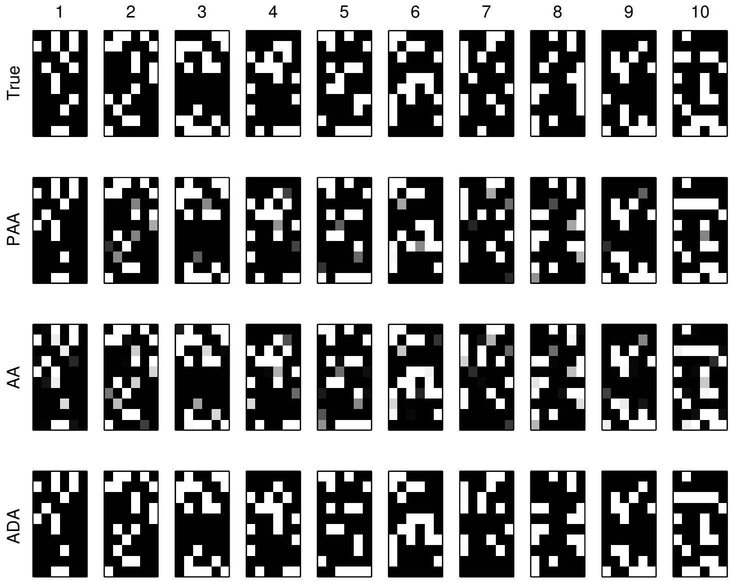


Figure 2. *Comparison between true archetypes and those returned by PAA, AA and ADA, respectively. The 10 columns represent the first 10 repetitions of the simulation. Black represents 0 and white 1.*

Table 1. *Summary of misclassification errors of the archetype profiles for each method over 100 simulations.*

Method	PAA	AA	ADA
Mean (Std. dev.)	4.20 (1.86)	3.59 (1.99)	3.19 (1.88)

4. Applications

4.1. An initial mathematical skills test for first-year university students

We would like to estimate the skill set profiles hidden in the data set described by Orús and Gregori (2008) and introduced in Section 1. In other words, we would like to discover the data structure. Our intuition tells us that skill sets vary continuously across students, i.e. we do not expect there to be clearly differentiated (separate) groups of students with different abilities. Even so, CLA has been used to generate groups of students with similar skill set profiles (Chiu et al., 2009; Dean and Nugent, 2013). Here, we are going to consider the raw binary data and let the data speak for themselves, as ADA is a data-driven method. We compare the ADA solution with others from well-established unsupervised techniques to highlight the information about the quality understanding of data provided by ADA. In particular, besides AA and PAA, we use homogeneity analysis (multiple correspondence analysis) using the R package **homals** (de Leeuw and Mair, 2009) (HOMALS) and CLA by Partitioning Around Medoids (PAM) from the R package **cluster** (Maechler et al., 2018; Kaufman and Rousseeuw, 1990), since it returns representative objects or medoids among the observations of the data set. The pairwise dissimilarities between observations in the data set needed for PAM are computed with the *daisy* function from the R package **cluster** (Maechler et al., 2018), specifically using Gower's coefficient (Gower, 1971) for binary observations. Other popular clustering methods (Flynt and Dean, 2016) are also used in the comparison: latent class analysis (LCA) from the R package **poLCA** (Linzer and Lewis, 2011), which is a finite mixture model clustering for categorical data, and classical k-means clustering (Lloyd, 1982). It is used in the literature (Henry et al., 2015), despite not being recommended for binary data (IBM Support, 2016). For that reason, we also consider PAM, since it is a robustified version of k-means (Steinley, 2006) that can be used with distances other than Euclidean, and observations, rather than centroids, serve as the exemplars for each

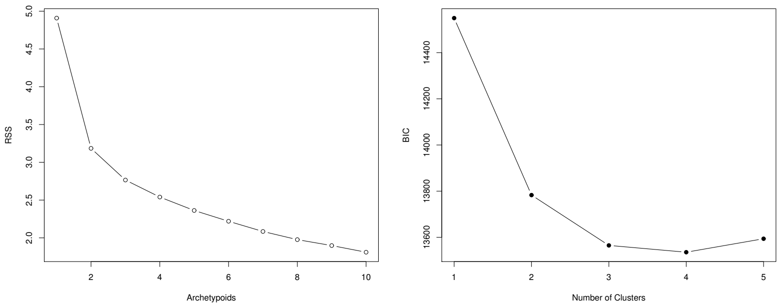


Figure 3. *Initial mathematical skills test data: Screeplot of ADA (left-hand panel); screeplot of LCA (right-hand panel).*

cluster.

For the sake of brevity and as an illustrative example, we examine the results of k = 3. The RSS elbow for ADA and the Bayesian Information Criterion (BIC) elbow for LCA are found at k = 3 (see Figure 3).

The highest value of the silhouette coefficient (a method of interpretation and validation of consistency within clusters of data, see Kaufman and Rousseeuw (1990) for details) is 0.22 (for k = 2 and k = 3 clusters), which means that no substantial cluster structure was found, as we predicted. We perform an h-plot (a multidimensional scaling method that is particularly suited for representing non-Euclidean dissimilarities, see Epifanio (2013) for details) on the dissimilarities used by PAM to graphically summarize the data set and to visualize the obtained clusters by PAM in two dimensions (see Figure 4). Effectively, separate clusters do not seem to exist.

This is also corroborated by Figure 5, where the students' scores from HOMALS are plotted in two dimensions. As regards the interpretation of the dimensions of HOMALS, the loadings are displayed in Figure 6 and Table 2 shows their exact values, together with the number of correct answers. As also happens with PCA, their interpretation is not always easy and immediate. For the first dimension, all the coefficients are positive

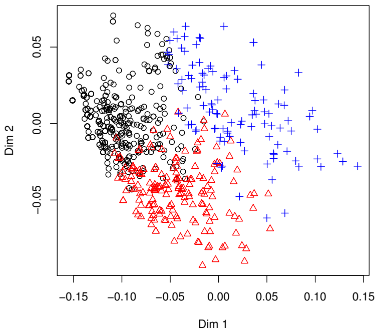


Figure 4. *H-plot of dissimilarities for the initial mathematical skills test data. We perform PAM. The black circles represent data points assigned to the first cluster, the red triangles to the second cluster and the blue crosses to the third cluster.*

(as a measure of size), which can indicate a kind of sum score. The highest coefficients more or less correspond to the last questions of the test, which fewer students answered correctly. The second dimension compares, above all, questions 4, 5, 6a and 6b (with high positive coefficients) with 13a and 13b (with low negative values), while in the third dimension, questions 1, 3, 7, 8 and 10 (with high positive coefficients) are compared with 14a and 14b (with low negative values). However, we do know how the meaning of these contradistinctions is interpreted.

LCA returns the conditional item response probabilities by outcome variable for each class. Table 3 lists these probabilities for correct answer. The predicted classes for each student are shown in Figure 5, since the profiles of cluster 1 and 3 are mainly differentiated in questions 4, 5, 13a and 13b, which are the most relevant variables of dimension 2 of HOMALS.

Table 3 also lists the profiles of the medoids, centroids of k-means and the archetypal profiles for AA, PAA and ADA. For medoids and archetypoids, the code of the corresponding observation is also displayed. To facilitate the analysis we also show the binarized profiles of AA and PAA.

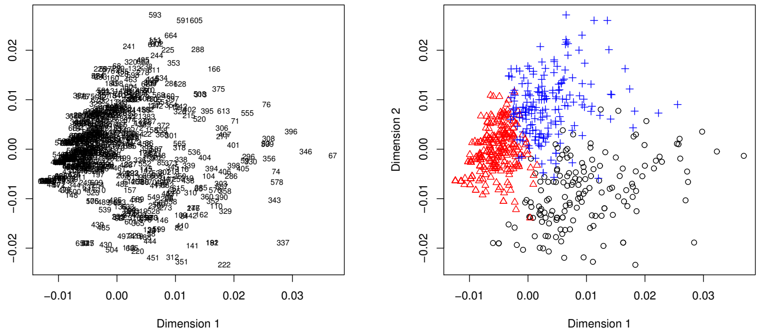


Figure 5. *HOMALS of the initial mathematical skills test data. Plot of students' scores. The numbers indicate the code of each of the 690 students (left-hand panel). We perform LCA. The black circles represent data points assigned to the first cluster, the red triangles to the second cluster and the blue crosses to the third cluster (right-hand panel).*

As a simple summary of the profiles, we compute the percentage of correct answers for each profile. For PAM, the percentages are 9.5%, 33.3% and 57.1%; for binarized LCA, 38.1%, 9.5% and 33.3%; for binarized k-means, 38.1%, 9.5% and 42.9%; for BAA, 9.5%, 47.6% and 61.9%; for BPAA, 9.5%, 42.9% and 57.1%; and for ADA, 57.1%, 52.4% and 9.5%, respectively. Note that the median of the percentage of correct answers in the data set is 28.6% (the minimum is 0, the first quartile is 19.1%, the third quartile is 38.1%, while the maximum is 95.2%).

One profile is repeated in all the methods, a student who only answers questions 1 and 2 correctly, i.e. a student with a serious lack of competence. We therefore concentrate the analysis on the other two profiles for each method.

In contrast with the third archetypoid, i.e. the student with very poor skills, the first and second archetypoids correspond to students with very high percentages of correct answers. In fact, the first archetypoid corresponds to the 92nd percentile of the data set, while the second archetypoid corresponds to the 88th percentile. Nevertheless, both profiles are quite different. In fact, the Hamming distance between archetypoids 1 and 2

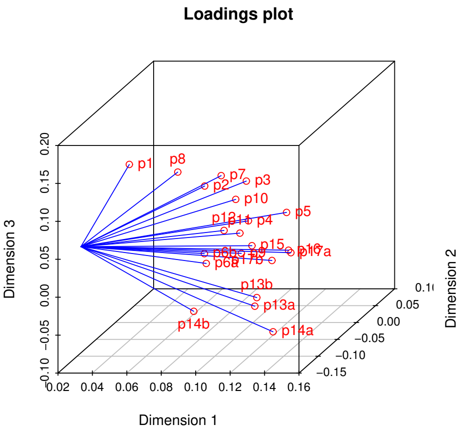


Figure 6. *HOMALS of the initial mathematical skills test data. Loadings plot.*

is 13, which means that although they answered a lot of items correctly, these correctly answered items do not coincide. In other words, archetypoids 1 and 2 are somehow complementary. Both answered items 1, 2, 3, 12 and 16 correctly, which were among the most correctly answered items. Neither of them answered items 11, 14b and 17a correctly, which were among the least correctly answered items. On the one hand, the items that archetypoid 1 answered correctly, but archetypoid 2 did not are 8, 10, 13a, 13b, 14a, 15 and 17b. These items are about nonlinear systems and linear functions. On the other hand, the items that archetypoid 2 answered correctly, but archetypoid 1 did not are 4, 5, 6a, 6b, 7 and 9. These items are about the calculation of derivatives and integrals and algebraic interpretation. The skills of these archetypoids are clear and different to each other.

We can use the alpha values for each of the students to learn about their relationship to the archetypoid profiles. The ternary plot in Figure 7 displays the alpha values that provide further insight into the data structure. Note that the majority of the data is concentrated around archetypoid 1, i.e. the one with very poor skills. If we wanted to form three groups using the alpha values, we could assign each student to the group in

Table 2. *Number of correct answers and loadings of the first three dimensions by HOMALS for the initial mathematical skills test data.*

Question	No. correct answers	D1	D2	D3
1	621	0.02862783	-0.0023240445	0.109355256
2	589	0.05796987	0.0626839294	0.052003283
3	301	0.09098387	0.0229084678	0.076212734
4	233	0.07597107	0.0959592091	-0.008635664
5	253	0.09922374	0.0910173149	0.004650411
6a	231	0.05413846	0.0839718688	-0.059468794
6b	105	0.05230047	0.0873491670	-0.048042860
7	270	0.07408320	0.0332456303	0.078814530
8	140	0.06009601	-0.0172906782	0.105955795
9	109	0.07749049	0.0705908727	-0.040283362
10	202	0.09540734	-0.0246907330	0.073252540
11	71	0.07484609	0.0786972991	-0.017170762
12	329	0.08006423	0.0138046564	0.015179516
13a	177	0.12934508	-0.1274837375	-0.021853917
13b	132	0.12952677	-0.1231243539	-0.012406654
14a	114	0.11951449	-0.0355861709	-0.096428155
14b	22	0.07564891	-0.0454400932	-0.065063675
15	183	0.10748607	-0.0366512703	0.017544829
16	236	0.12062274	-0.0001786048	-0.004647963
17a	47	0.12115852	0.0035393336	-0.009522369
17b	62	0.10884146	0.0095484411	-0.022486862

which their corresponding alpha is the maximum, as we did in Figure 1 (d). In this way, the number of students similar to archetypoid 1 is 113, to archetypoid 2 it is 110 and to archetypoid 3 it is 467.

The profiles of medoids 2 and 3 are not as complementary as the previous archetypoids. In fact, medoid 2 corresponds to the 56th percentile, while medoid 3 corresponds to the 92nd percentile. In this case, the percentage of correct answers for medoid 2 is not high. The Hamming distance between the two medoids is only 7. On the one hand, both answered items 1, 2, 3, 5, 7 and 12 correctly, which are the most correctly answered items. On the other hand, both failed items 6a, 6b, 8, 9, 11, 14b, 17a and 17b, many more items than in the case of ADA. The only item that medoid 2 answered correctly but medoid 3 did not is item 4. The items that medoid 3 answered correctly but medoid 2 did not are 10, 13a, 13b, 14a, 15 and 16. It seems as if the cluster definition was guided by the number of correct answers rather than by the kind of item answered correctly.

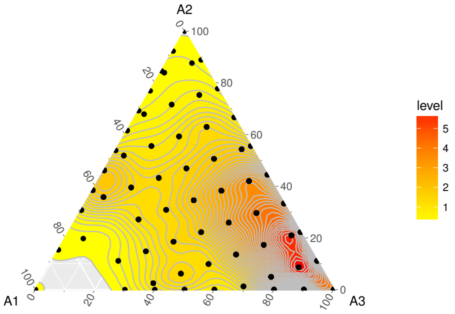


Figure 7. *Ternary plot of αs of ADA together with a plot density estimate for the initial mathematical skills test data.*

This is the reason why PAM selects medoid 2 in the middle of the data cloud. PAM, and usual clustering methods, tries to cover the set in such a way that every point is near to one medoid or one cluster center. The number of students belonging to each cluster is 398, 179 and 113, respectively. Note that the size of the cluster of students with poor skills is smaller than in the case of ADA, because some of those students are assigned to the cluster of medoid 2.

The binarized profile of LCA 1, corresponding to the 75th percentile, is similar to medoid 3, but with a lower number of correct answers (5, 7, 14a and 15), while the binarized profile of LCA 3, corresponding to the 56th percentile, is similar to medoid 2, only differentiated by two items (7 and 16). Therefore, they are even less complementary than the previous medoids. The Hamming distance between both LCA-profiles is only 5. The number of students belonging to each cluster is 155, 352 and 183, respectively. Note that the size of the cluster of students with poor skills is smaller than in the case of PAM.

The binarized profile of the first centroid of *k*-means, corresponding to the 75th

Table 3. *Profiles for the initial mathematical skills test data, for PAM, LCA, k-means (k-M), AA (and binarized, BAA), PAA (and binarized, BPAA) and ADA, with k = 3. The numbers in brackets for PAM and ADA indicate the code of the representative student*

Methods	1	2	3	4	5	6a	6b	7	8	9	10	11	12	13a	13b	14a	14b	15	16	17a	17b
PAM (661)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PAM (586)	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
PAM (162)	1	1	1	0	1	0	0	1	0	0	1	0	1	1	1	1	0	1	1	0	0
LCA 1	0.93	0.88	0.57	0.36	0.48	0.37	0.17	0.49	0.32	0.20	0.51	0.14	0.62	1.00	0.85	0.43	0.10	0.50	0.55	0.18	0.19
LCA 2	0.88	0.79	0.28	0.17	0.15	0.24	0.07	0.30	0.14	0.03	0.15	0.02	0.32	0.05	0	0.03	0	0.13	0.14	0	0
LCA 3	0.91	0.94	0.60	0.62	0.65	0.48	0.27	0.47	0.23	0.35	0.36	0.23	0.63	0.04	0	0.20	0.03	0.31	0.53	0.09	0.17
k-M 1	0.91	0.95	0.64	0.70	0.74	0.43	0.25	0.53	0.24	0.32	0.38	0.22	0.63	0.05	0.00	0.19	0.02	0.33	0.52	0.10	0.15
k-M 2	0.88	0.78	0.26	0.13	0.11	0.27	0.09	0.27	0.13	0.05	0.14	0.02	0.32	0.06	0.01	0.03	0.01	0.12	0.15	0.01	0.01
k-M 3	0.93	0.91	0.59	0.34	0.49	0.37	0.17	0.50	0.33	0.20	0.54	0.14	0.64	1	0.88	0.46	0.10	0.52	0.58	0.18	0.19
AA 1	0.85	0.68	0.02	0	0	0.05	0	0.04	0.05	0	0.01	0	0.16	0	0	0	0	0.01	0	0	0
AA 2	0.90	1	0.87	1	1	1	0.63	0.82	0.19	0.52	0.24	0.38	0.65	0	0	0.16	0	0.07	0.43	0.09	0.15
AA 3	1	1	0.89	0.32	0.53	0.19	0.06	0.71	0.58	0.26	1	0.17	1	1	1	0.67	0.18	1	1	0.36	0.37
BAA 1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BAA 2	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0
BAA 3	1	1	1	0	1	0	0	1	1	0	1	0	1	1	1	1	0	1	1	0	0
PAA 1	0.86	0.72	0.13	0	0	0	0	0.12	0.07	0	0	0	0.23	0	0	0	0	0	0	0	0
PAA 2	0.90	1	0.78	1	1	1	0.61	0.73	0.27	0.40	0.38	0.31	0.66	0	0	0	0	0	0.43	0	0
PAA 3	0.99	1	0.82	0.36	0.57	0.25	0	0.66	0.44	0.27	0.86	0.12	0.85	1	1	0.73	0.15	1	1	0.32	0.42
BPAA 1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BPAA 2	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
BPAA 3	1	1	1	0	1	0	0	1	0	0	1	0	1	1	1	1	0	1	1	0	0
ADA (182)	1	1	1	0	0	0	0	0	1	0	1	0	1	1	1	1	0	1	1	0	1
ADA (274)	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0
ADA (1)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

percentile, is similar to medoid 2, only differentiated by item 16, while the binarized profile of the third centroid, corresponding to the 82nd percentile, is similar to medoid 3, but with a lower number of correct items (5, 7 and 14a). The Hamming distance between both centroids is 7. The level of complementarity between both centroids is similar to that of the medoids of PAM, but the number of correct answers of medoid 3 is higher than binarized centroid 3. The number of students belonging to each cluster is 196, 349 and 145, respectively. Note that the size of the cluster of students with poor skills is smaller than in the case of PAM, but larger than in the case of PAM for cluster 3, which in both clustering methods corresponds to the students with more correct answers.

In the clustering methods, the profiles of each cluster are not as extreme as archetypoids. Archetypoids are also more complementary, which makes it clearer to establish which kinds of features distinguish one group from another.

The profiles of BAA2 and BAA3 and BPAA2 and BPAA3 are quite similar to the profiles of archetypoid 2 and 1, respectively, but with slight differences. The percentiles corresponding to correctly answered items are also high, although for one of the archetypes not as high as for archetypoids. The percentiles are the 82nd and 94th for BAA2 and BAA3, and the 75th and 92nd for BPAA2 and BPAA3, respectively. Therefore, the archetypoids are more extreme than the binarized archetypes of AA and PAA. Although the profiles for BAA and BPAA are also complementary, they are not as complementary as the two archetypoids. The Hamming distance between BAA2 and BAA3 is 11, and 9 between BPAA2 and BPAA3. Archetypoids therefore manage to find more complementary profiles.

4.2. An ACT Mathematics Test

As mentioned in Section 1, we used the same data and approach followed by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) to estimate IRFs, $P_i(\theta)$, and their logit functions, $W_i(\theta) = log(P_i(\theta)/(1-P_i(\theta)))$. In particular, a penalized EM algorithm was used and functions were expanded by terms of 11 B-spline basis functions using equally spaced knots. Figure 8 displays the estimated IRFs, $exp(W_i(\theta))/(1+exp(W_i(\theta)))$, and their log odds-ratio functions $W_i(\theta)$ for the 60 items. As expected, this kind of graphs with superimposed curves is largely uninformative and aesthetically unappealing (Jones and Rice, 1992).

To explore a set of curves Jones and Rice (1992) proposed the use of functions with extreme principal component scores. This could be viewed as finding the archetypoid functions. Nevertheless, the aim of PCA is not to recover extreme patterns. In fact, curves with extreme PCA scores do not necessarily correspond to archetypal observations. This is discussed in Cutler and Breiman (1994) and shown in Epifanio et al.

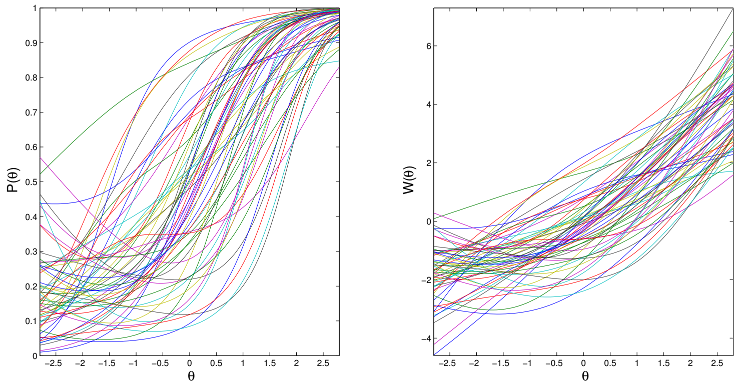


Figure 8. *Estimated IRFs (left-hand panel) and log odds-ratio functions (right-hand panel) for the ACT math exam estimated from the male data.*

(2013) through an example where archetypes could not be restored with PCA, even if all the components had been considered. Not only that, Stone and Cutler (1996) also showed that AA may be more appropriate than PCA when the data do not have elliptical distributions.

In order to show the advantages of ADA over PCA, we compute FPCA and FADA for $W(\theta)$, since they are unconstrained, therefore making them more appropriate for PCA application than the bounded $P_i(\theta)$. This is not a problem with FADA as it works with convex combinations. Figure 9 displays the first four PCs after a varimax rotation having been back-transformed to their probability counterparts, as performed by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002). We base the interpretation of each PC on the detailed description carried out by Ramsay and Silverman (2002, Ch. 9).

The percentage of total variation explained by those four components is nearly 100%, while the percentage explained by each component is reported in Figure 9. The first component concentrates on the middle part of the ability range, in such a way that an item with a high (low) score in that component has a higher (lower) slope than the mean from approximately 0 to 2, i.e. it quantifies a discriminability trade-off between average stu-

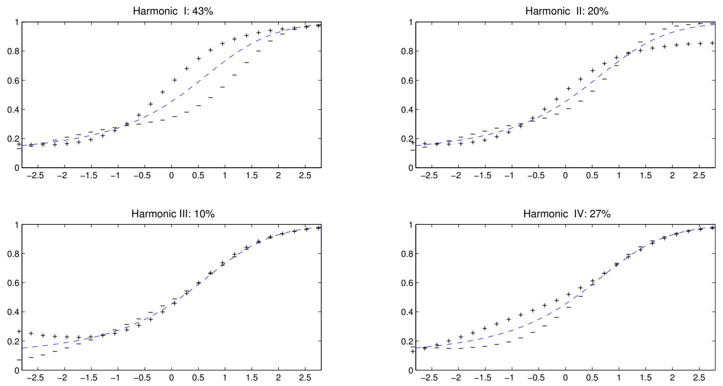


Figure 9. *The first four functional PCs in IRFs after VARIMAX rotation. Plus (negative) signs indicate the effect of adding (subtracting) a multiple of a component to the mean function. The mean function is the dashed line.*

dents and those with rather high abilities. Analogously, the fourth component quantifies a discriminability trade-off between average examinees and those with rather low abilities. On the contrary, the second component concentrates on the upper end of the ability range. As Rossi et al. (2002) explained, the 3PL model is not well suited to modeling this type of variation. An item with a low score on this component is good at sorting out very high ability students from others of moderately high ability, whereas if the score for this item is high, it will discriminate well among most of the population but will be found to be of approximately equal difficulty by all the very good students. Nevertheless, conclusions on the extreme part of the ability range should be made with caution, since the estimation is carried out using a relatively small numbers of students. The third component also accounts for variation in the characteristics of test items in the extreme ability range, but now in low ability ranges. PC scores for these four components can be seen in Figure 10. Note that to evaluate the 4 PC scores simultaneously and combine them to give an idea about each item, it is not easily comprehensible or human-readable.

Figure 11 displays the archetypoids for $k = 4$ explaining 97% of the variability, which is nearly as high as FPCA. The archetypoids are items 2, 18, 28 and 60. These four items

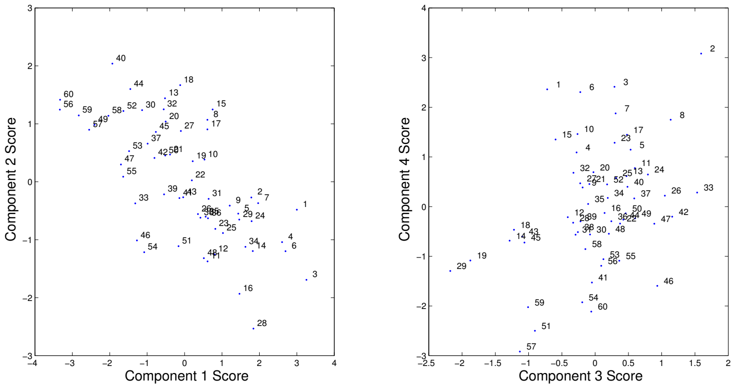


Figure 10. *Bivariate plots of principal component scores of IRFs. PC1 versus PC2 (left-hand panel); PC3 versus PC4 (right-hand panel).*

describe the extreme patterns found in the sample. Item 2 has very high scores in PC 3 and PC 4, high scores in PC 1 and a score of nearly zero for PC 2. Its IRF is quite flat with a very slight slope, it seems to be a very easy item, with high probabilities of success throughout the ability range. The other archetypoids discriminate better between low and high ability students but in very different ways. Item 18 has a very high score for PC 2 and a negative score for PC 3, but nearly zero for PC 1 and PC 4. It is an item that is quite difficult even for the students in the very high ability range. The IRF of item 28 is quite similar to that of item 18 for the low ability range until θ 0, but its slope for the high ability range is higher, and the probabilities of success are higher than 0.9 for θs higher than 1. On the contrary, the probabilities of success of the IRF of item 60 are quite low as far as 1.5, which means that it is a difficult item, but the probabilities of success for the best students are high. In fact, the probabilities of success for item 60 for θ higher than 2 are higher than those of item 18. Item 28 has high score for PC 1 and low score for PC2, while it has a score of nearly zero for PC 3 and PC 4. However, item 60 has low scores for PC 1 and PC 4, a high score for PC 2 and nearly zero for PC 3. In other words, it would have been very difficult to guess the extreme representatives of the sample returned by ADA from an analysis of the scores in Figure 10.

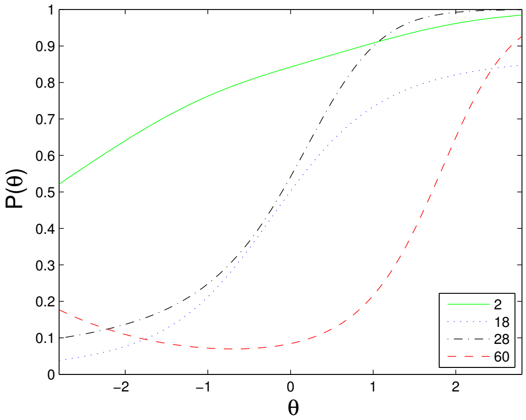


Figure 11. *ACT data: The four IRF archetypoids are items 2, 18, 28 and 60. See the legend inside the plot.*

The alpha values (from 0 to 1) tell us about the contribution of each archetypoid to each item. Remember that they add up to 1. Figure 12 shows star plots of the alpha values for each archetypoid, thus providing a complete human-readable view of the data set. The 4 alpha values in this case are represented starting on the right and going counter-clockwise around the circle. The size of each alpha is shown by the radius of the segment representing it. The items that are similar to the archetypoids can be clearly seen (for example, 7 and 8 are somehow similar to 2; 15 and 19 are somehow similar to 18; 14 and 16 are somehow similar to 28; and 56, 57 and 59 are similar to 60), as can the items that are a mixture of several archetypoids (for example, item 1 is a mixture of mainly item 2, together with items 28 and 18, to a lesser extent). Item 1 was selected by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) as an example of a low difficulty item, although it seems that item 2 would be a better representative of this kind of item. Item 9 was selected by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) as an example of a medium difficulty item, and it is mainly a mixture of items 18 and 28. Finally, item 59 was selected by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) as an example of a hard item. Item 59 was mainly explained by item 60.

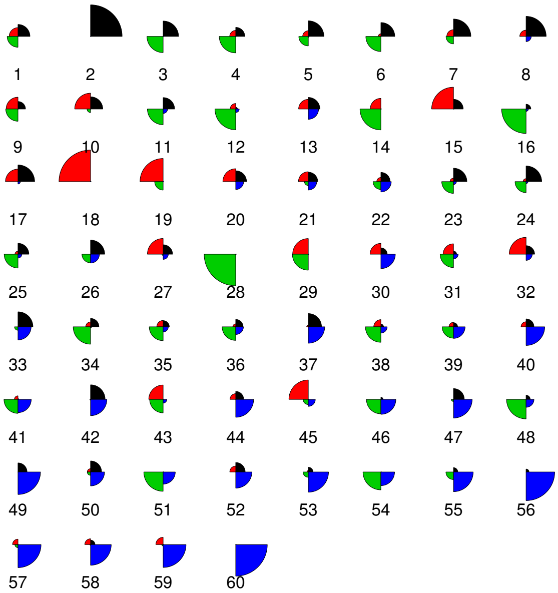


Figure 12. *Star plots of the alphas of each archetypoid for IRFs. The item number appears below each plot. The archetypoids are 2, 18, 28 and 60.*

4.2.1. FADA with kernel and parametric IRF estimates

Figure 13 shows IRFs estimated by kernel smoothing with the R package **KernSmoothIRT** (Mazza et al., 2014) and the 3PL model with the R package **irtoys** (Partchev and Maris, 2017; Rizopoulos, 2006). Note that the estimates are quite different, also if we compare them with those in Figure 8. On the one hand, parametric models are not as flexible as nonparametric methods (remember the previous analysis about the variation in the upper asymptote for the second PC component). The possible shapes of the 3PL model estimates are restricted by the functional form. On the other hand, although kernel smoothing makes it possible to represent the data well, there is too much local curvilinearity, i.e. they are not as smooth as the estimated IRFs in the previous Section. With kernel and 3PL model estimation methods, we have the estimates of IRFs in a series of points and we can apply ADA to obtain the functional archetypoids. Note that in FDA we can also work point-wise with discretized functions to a fine grid. Nevertheless, as the input data are different, we can expect variation in the archetypoids obtained for the different estimation methods.

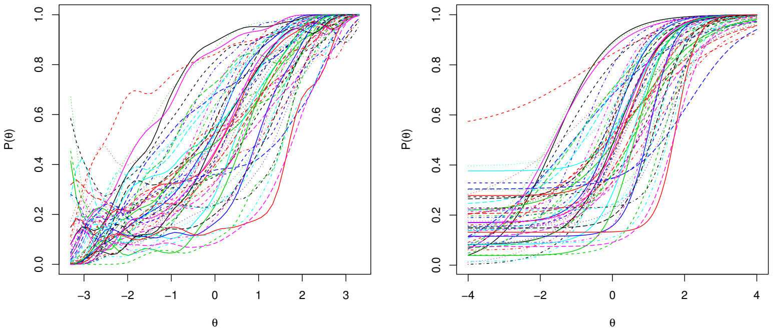


Figure 13. *Estimated IRFs by kernel smoothing (left-hand panel) and the 3PL model (right-hand panel) for the ACT math exam.*

Figure 14 shows the PC scores together with the functions with extreme scores for each component, as computed by the R package **KernSmoothIRT**. Note the strange estimate for item 3 in the lower end of the ability range, and the noisy estimates for the other items. The 4 archetypoids derived from kernel smoothing and 3PL model estimates are shown in Figure 15. Note that none of the archetypoids coincides with the extreme PC scores. Two of the archetypoids coincide in all three estimation methods: item 2 and item 60, which are typical of the easiest and hardest items, respectively. The other two archetypoids in each case show a similar positive slope, but begin at different values of θ.

In summary, we can apply ADA to different ways of estimating IRFs. Obviously, depending on the input, ADA, or any other method, returns different solutions. So we

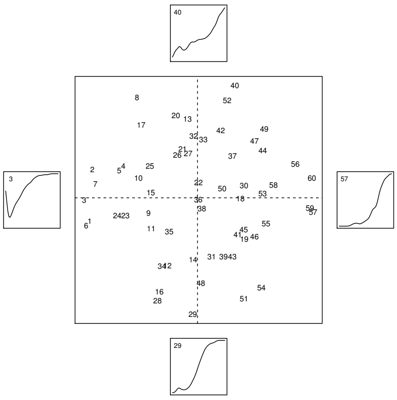


Figure 14. *First two principal components for the ACT math exam with kernel smoothing. In the interior plot, numbers are the identifiers of the items. The small plots show the estimated IRFs for the most extreme items for each principal component.*

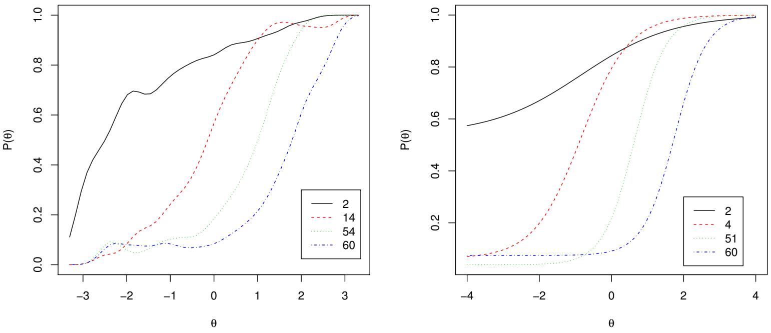


Figure 15. *The four archetypoids for the estimated IRFs by kernel smoothing (left-hand panel) and the 3PL model (right-hand panel) for the ACT math exam. See the legend inside each plot.*

must be cautious with the estimation method we use.

5. Conclusion

We have proposed to find archetypal patterns in binary data using ADA for a better understanding of a data set. A simulation study and results provided in two applications have highlighted the benefits of ADA for binary questionnaires as an alternative that can be used instead of (or in addition to) other established methodologies.

Although, much of statistics is based on the idea that averaging over many elements of a data set is a good thing to do, in this paper we adopt a different perspective. We have selected a small number of representative observations, archetypal observations, and the data composition is explained through mixtures of those extreme observations. We have shown that this can be highly informative and is a useful tool for making a data set more "human-readable", even to non-experts.

As regards future work, throughout the paper all variables share the same weight, but for certain situations some variables could have more weight in RSS. Another direction of future work would be to consider ADA for nominal observations, for example, by converting those variables into dummy variables, i.e. with binary codes. Another not so immediate extension, would be to consider the case of mixed data, with real valued and categorical data, together with missing data.

Acknowledgments:

This work is supported by the following grants: DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities (AEI/FEDER, EU) and UJI-B2017-13 from Universitat Jaume I.

References

Canhasi, E. and I. Kononenko (2013). Multi-document summarization via archetypal analysis of the content-graph joint model. *Knowledge and Information Systems*, 1–22.

Canhasi, E. and I. Kononenko (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications 41*(2), 535 – 543.

Chan, B., D. Mitchell, and L. Cram (2003). Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society 338*(3), 790–795.

Chiu, C.-Y., J. A. Douglas, and X. Li (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika 74*(4), 633.

Cutler, A. and L. Breiman (1994). Archetypal Analysis. *Technometrics 36*(4), 338–347.

Davis, T. and B. Love (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science 21*(2), 234–242.

de Leeuw, J. and P. Mair (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software 31*(4), 1–20.

Dean, N. and R. Nugent (2013). Clustering student skill set profiles in a unit hypercube using mixtures of multivariate betas. *Advances in Data Analysis and Classification 7*(3), 339–357.

D'Esposito, M. R., F. Palumbo, and G. Ragozini (2012). Interval Archetypes: A New Tool for Interval Data Analysis. *Statistical Analysis and Data Mining 5*(4), 322–335.

Epifanio, I. (2013). h-plots for displaying nonmetric dissimilarity matrices. *Statistical Analysis and Data Mining 6*(2), 136–143.

Epifanio, I. (2016). Functional archetype and archetypoid analysis. *Computational Statistics & Data Analysis 104*, 24 – 34.

Epifanio, I., M. V. Ibáñez, and A. Simó (2018). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. *The American Statistician*.

Epifanio, I., M. V. Ibáñez, and A. Simó (2018). Archetypal shapes based on landmarks and extension to handle missing data. *Advances in Data Analysis and Classification 12*(3), 705–735.

Epifanio, I., G. Vinué, and S. Alemany (2013). Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. *Computers & Industrial Engineering 64*(3), 757–765.

Eugster, M. J. and F. Leisch (2009). From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software 30*(8), 1–23.

Eugster, M. J. A. (2012). Performance profiles based on archetypal athletes. *International Journal of Performance Analysis in Sport 12*(1), 166–187.

Fernandez, M. and A. S. Barnard (2015). Identification of nanoparticle prototypes and archetypes. *ACS Nano 9*(12), 11980–11992.

Fletcher, R. (2000). *Practical Methods of Optimization* (Second ed.). John Wiley & Sons.

Flynt, A. and N. Dean (2016). A survey of popular R packages for cluster analysis. *Journal of Educational and Behavioral Statistics 41*(2), 205–225.

Friedman, J. H. and J. W. Tukey (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers C-23*(9), 881–890.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*(4), 857–871.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning. Data mining, inference and prediction*. 2nd ed., Springer-Verlag.

Henry, D., A. B. Dymnicki, N. Mohatt, J. Allen, and J. G. Kelly (2015). Clustering methods with qualitative data: a mixed-methods approach for prevention research with small samples. *Prevention Science 16*(7), 1007–1016.

Hinrich, J. L., S. E. Bardenfleth, R. E. Roge, N. W. Churchill, K. H. Madsen, and M. Mørup (2016). Archetypal analysis for modeling multisubject fMRI data. *IEEE Journal on Selected Topics in Signal Processing 10*(7), 1160–1171.

IBM Support (2016). Clustering binary data with K-Means (should be avoided). http://www-01.ibm.com/support/docview.wss?uid=swg21477401. Accessed: 2018-07-09.

Jones, M. C. and J. A. Rice (1992). Displaying the important features of large collections of similar curves. *The American Statistician 46*(2), 140–145.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley.

Lawson, C. L. and R. J. Hanson (1974). *Solving Least Squares Problems*. Prentice Hall.

Li, S., P. Wang, J. Louviere, and R. Carson (2003). Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals. In *ANZMAC 2003 Conference Proceedings*, pp. 1674–1679.

Linzer, D. A. and J. B. Lewis (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software 42*(10), 1–29.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory 28*, 129–137.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik (2018). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.7-1.

Makowski, D. (2016). *Package 'neuropsychology': An R Toolbox for Psychologists, Neuropsychologists and Neuroscientists*. (0.5.0).

Mazza, A., A. Punzo, and B. McGuire (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software 58*(6), 1–34.

Midgley, D. and S. Venaik (2013). Marketing strategy in MNC subsidiaries: pure versus hybrid archetypes. In *P. McDougall-Covin and T. Kiyak, Proceedings of the 55th Annual Meeting of the Academy of International Business*, pp. 215–216.

Millán-Roures, L., I. Epifanio, and V. Martínez (2018). Detection of anomalies in water networks by functional data analysis. *Mathematical Problems in Engineering 2018*(Article ID 5129735), 13.

Moliner, J. and I. Epifanio (2019). Robust multivariate and functional archetypal analysis with application to financial time series analysis. *Physica A: Statistical Mechanics and its Applications 519*, 195 – 208.

Mørup, M. and L. K. Hansen (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing 80*, 54–63.

Orús, P. and P. Gregori (2008). *Fictitious Pupils and Implicative Analysis: a Case Study*, pp. 321–345. Berlin, Heidelberg: Springer.

Partchev, I. and G. Maris (2017). *irtoys: A Collection of Functions Related to Item Response Theory (IRT).* R package version 0.2.1.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.

Porzio, G. C., G. Ragozini, and D. Vistocco (2008). On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry 24*, 419–437.

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ragozini, G. and M. R. D'Esposito (2015). Archetypal networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, New York, NY, USA, pp. 807–814. ACM.

Ragozini, G., F. Palumbo, and M. R. D'Esposito (2017). Archetypal analysis for data-driven prototype identification. *Statistical Analysis and Data Mining: The ASA Data Science Journal 10*(1), 6–20.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika 56*(4), 611–630.

Ramsay, J. O. (1997). *A Functional Approach to Modeling Test Data*, pp. 381–394. New York, NY: Springer New York.

Ramsay, J. O. and B. W. Silverman (2002). *Applied Functional Data Analysis*. Springer.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed.). Springer.

Ramsay, J. O. and M. Wiberg (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics 42*(3), 282–307.

Rizopoulos, D. (2006). ltm: An r package for latent variable modeling and item response analysis. *Journal of Statistical Software, Articles 17*(5), 1–25.

Rossi, N., X. Wang, and J. O. Ramsay (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics 27*(3), 291–317.

Seth, S. and M. J. A. Eugster (2016a). Archetypal analysis for nominal observations. *IEEE Trans. Pattern Anal. Mach. Intell. 38*(5), 849–861.

Seth, S. and M. J. A. Eugster (2016b). Probabilistic archetypal analysis. *Machine Learning 102*(1), 85–113.

Slater, S., S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics 42*(1), 85–106.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology 59*(1), 1–34.

Steinschneider, S. and U. Lall (2015). Daily precipitation and tropical moisture exports across the Eastern United States: An application of archetypal analysis to identify spatiotemporal structure. *Journal of Climate 28*(21), 8585–8602.

Stone, E. and A. Cutler (1996). Introduction to archetypal analysis of spatio-temporal dynamics. *Physica D: Nonlinear Phenomena 96*(1), 110 – 131.

Su, Z., Z. Hao, F. Yuan, X. Chen, and Q. Cao (2017). Spatiotemporal variability of extreme summer precipitation over the Yangtze river basin and the associations with climate patterns. *Water 9*(11).

Theodosiou, T., I. Kazanidis, S. Valsamidis, and S. Kontogiannis (2013). Courseware usage archetyping. In *Proceedings of the 17th Panhellenic Conference on Informatics*, PCI '13, New York, NY, USA, pp. 243–249. ACM.

Thøgersen, J. C., M. Mørup, S. Damkiær, S. Molin, and L. Jelsbak (2013). Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics 14*, 279.

Thurau, C., K. Kersting, M. Wahabzada, and C. Bauckhage (2012). Descriptive matrix factorization for sustainability: Adopting the principle of opposites. *Data Mining and Knowledge Discovery 24*(2), 325–354.

Tsanousa, A., N. Laskaris, and L. Angelis (2015). A novel single-trial methodology for studying brain response variability based on archetypal analysis. *Expert Systems with Applications 42*(22), 8454 – 8462.

Unwin, A. (2010). Exploratory data analysis. In P. Peterson, E. Baker, and B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)*, pp. 156 – 161. Oxford: Elsevier.

Vinué, G. (2017). Anthropometry: An R package for analysis of anthropometric data. *Journal of Statistical Software 77*(6), 1–39.

Vinué, G. and I. Epifanio (2017). Archetypoid analysis for sports analytics. *Data Mining and Knowledge Discovery 31*(6), 1643–1677.

Vinué, G., I. Epifanio, and S. Alemany (2015). Archetypoids: A new approach to define representative archetypal data. *Computational Statistics & Data Analysis 87*, 102 – 115.

Wu, C., E. Kamar, and E. Horvitz (2016). Clustering for set partitioning with a case study in ridesharing. In *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1384–1388.

3.3 Aportació II: Archetypal Analysis: an alternative to Clustering for unsupervised texture segmentation [2]

[2] L'article Cabero and Epifanio (2019a) està publicat per la revista *Image Analysis and Stereology*.

Image Anal Stereol 2019;38:151-160 doi:10.5566/ias.2052
Original Research Paper

ARCHETYPAL ANALYSIS: AN ALTERNATIVE TO CLUSTERING FOR UNSUPERVISED TEXTURE SEGMENTATION

ISMAEL CABERO[1] AND IRENE EPIFANIO✉,[2]

[1]Department of Didactics of Mathematics, Universitat de València, Avda. Tarongers, 4, 46022 València, Spain;
[2]Department of Mathematics and IMAC, Universitat Jaume I, Campus del Riu Sec, Castelló, 12071, Spain
e-mail: ismael.cabero@uv.es, epifanio@uji.es

(Received November 8, 2018; revised April 26, 2019; accepted April 30, 2019)

ABSTRACT

Texture segmentation is one of the main tasks in image applications, specifically in remote sensing, where the objective is to segment high-resolution images of natural landscapes into different cover types. Often the focus is on the selection of discriminant textural features, and although these are really fundamental, there is another part of the process that is also influential, partitioning different homogeneous textures into groups. A methodology based on archetype analysis (AA) of the local textural measurements is proposed. AA seeks the purest textures in the image and it can find the borders between pure textures, as those regions composed of mixtures of several archetypes. The proposed procedure has been tested on a remote sensing image application with local granulometries, providing promising results.

Keywords: archetype, image segmentation, local granulometries, mathematical morphology, texture analysis.

INTRODUCTION

Image segmentation, and texture segmentation in particular, is one of the most important and difficult tasks in image processing. It consists of separating the different textures presented in the image. In this work, we focus on unsupervised texture segmentation, *i.e.*, when no previous information about the textures in the image is available. Texture segmentation is part of the broader field of texture analysis (Tuceryan and Jain, 1993). Although the concept of texture has no exact definition in image processing, the underlying idea is that it is something where local patterns are repeated.

Two general approaches for carrying out texture segmentation are region-based approaches or boundary-based approaches. We consider a common region-based approach that consists of computing (local) textural features in small windows centered on each pixel of the image or on a sample of pixels and then performing a clustering analysis of them (Soille, 2003, Ch. 11). The reason for using clustering techniques is that feature vectors that have common attributes will form clusters in the feature space (Fletcher and Evans, 2005). However, this rationale does not take into account the cases in which the window is centered on the border between two or more different textures, *i.e.*, if the window contains a mixture of textures.

In summary, that procedure returns a sample of features from different pure types of textures (when the windows contain a single kind of texture), but also a sample of features from a mixture of textures. For that reason, we propose a more appropriate unsupervised statistical learning technique as an alternative to cluster analysis: archetype analysis (AA). The objective of AA is precisely to extract the archetypes, which are pure profiles in a data set, and to express the data as mixtures of those archetypes. The archetypes are themselves a mixture of observations from the data set. Therefore, with AA we can obtain the set of pure textures in the image and find the borders between pure textures, as those regions that are composed of mixtures of several archetypes. Textures are described by features. Obtaining good results will depend on the following premises: using textural features that on the one hand can successfully separate the characteristics of the different pure textures in the image, and on the other hand, in the case of a mixture of textures, its features are a convex combination of the features of the pure textures that form the mixture.

A toy example is used to illustrate what AA means and how it differs from clustering. Let us assume that we have in Fig. 1 a sample of six small windows from a certain image to segment. In this sample, the first and last window contain pure textures; the first one is completely smooth, while the last one is noisy. However, the other windows are a mixture (in different degrees) of the two textures. In order to characterize the windows, we consider the following feature: the percentage of smoothness in the window, measured as the area of smooth zones in the window divided by the total area of the window. So, we can summarize each windows as 100, 80, 60, 40, 20 and 0, respectively. Applying k-means, with $k = 2$, to this sample returns 80 and 20 as centers, corresponding to the two windows in the central column of Fig. 1, which are mixtures. Applying AA (with two archetypes) to this sample instead returns 100 and 0 as archetypes, corresponding to the first and last window, which are pure textures. Clustering returns central points, while AA returns extreme points, which are purer profiles than the central points. Furthermore, in a matrix α AA returns the composition (from 0 to 1 and adding up to 1) of the windows as a function of the archetypes found. In this example, α is [1 0; 0.8 0.2; 0.6 0.4; 0.4 0.6; 0.2 0.8; 0 1], *i.e.*, it returns how archetypes are mixed in each sample. So through α we can know that the first and last window correspond to pure textures, each corresponding to a different archetype, while the other windows are mixtures, to some extent, of those two textures. For example, the second window is a mixture between 80% of the first archetype (the smooth one) and 20% of the second archetype (the noisy one).

AA was proposed by Cutler and Breiman (1994) and it has found applications in diverse fields, such as computer vision (Chen *et al.*, 2014; Bauckhage *et al.*, 2015; Sun *et al.*, 2017a;b; Mair *et al.*, 2017), developmental psychology (Ragozini *et al.*, 2017), engineering (Epifanio *et al.*, 2013; Vinué *et al.*, 2015; Vinué, 2017; Epifanio *et al.*, 2018b; Millán-Roures *et al.*, 2018), finance (Moliner and Epifanio, 2019), genetics (Thøgersen *et al.*, 2013), global development (Epifanio, 2016; Epifanio *et al.*, 2018a), machine learning problems (Mørup and Hansen, 2012), neuroscience (Tsanousa *et al.*, 2015; Hinrich *et al.*, 2016) and sports (Eugster, 2012; Vinué and Epifanio, 2017).

The purpose of this work is to introduce the idea of using AA for texture segmentation and to apply it to a real problem. Section "Material and methods" presents the data and describes the methodology. Section "Results" shows the segmentation results. Finally, conclusions and further developments are discussed in Section "Discussion". The code in R (R Development Core Team, 2018) and data for reproducing the results are available at http://www3.uji.es/~epifanio/RESEARCH/aasegmentation.rar.

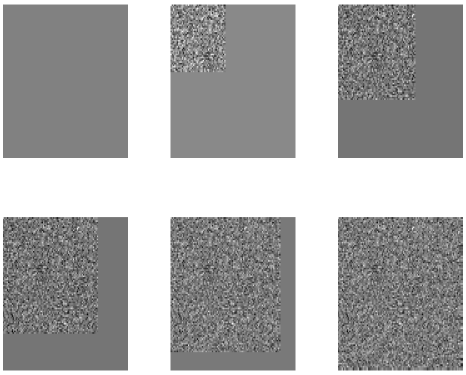


Fig. 1. *Toy example illustrating the difference between clustering and AA results (see text for details).*

MATERIAL AND METHODS

PRELIMINARY DEFINITIONS

Let us review some basic morphological transformations (see Soille, 2003 for an in-depth introduction). Let B be a structuring element, let $\check{B}$ be the reflection of B, and let f be a gray scale image. Then, the erosion of f by B is $[\varepsilon_B(f)](x) = \min_{b \in B} f(x+b)$; the dilation of f by B is $[\delta_B(f)](x) = \max_{b \in B} f(x+b)$, and the opening of f by B: $\gamma_B(f) = \delta_{\check{B}}[\varepsilon_B(f)]$.

Matheron (1975) defined a granulometry on a family $\mathscr{A}$ of sets, as a one-parameter family ψ_λ, with $\lambda \geq 0$, of mappings from $\mathscr{A}$ into itself such that: (i) $\psi_\lambda(A) \subset A$ for any $\lambda > 0$ and $A \in \mathscr{A}$; (ii) if $A, B \in \mathscr{A}$ and $A \subset B$, imply $\psi_\lambda(A) \subset \psi_\lambda(B)$; (iii) $\lambda_1 \geq \lambda_2 > 0$ imply $\psi_{\lambda_1}(A) \subset \psi_{\lambda_2}(A)$ and (iv) $\psi_{\lambda_1} \circ \psi_{\lambda_2} = \psi_{\lambda_2} \circ \psi_{\lambda_1} = \psi_{sup(\lambda_1,\lambda_2)}$. Additionally, it is usual to consider that $\psi_\lambda(A) = A$.

DATA

Compositions of artificial textures are the data sets commonly used for assessing texture segmentation procedures. However, we use several images of forest stands, with natural (real) textures, since textures on natural landscapes are more difficult to process owing to their high natural variation (Epifanio and Soille, 2007).

Many different local textural features can be used, such as classical spatial moments (Tuceryan, 1994) or Gabor filters (Jain and Farrokhnia, 1991), where features are computed in a sample of the pixels for computational efficiency and clustered.

Here, we prefer to use features based on mathematical morphology tools that have been proven to be successful in tackling different problems of geoscience and remote sensing (Epifanio and Ayala, 2002; Soille and Pesaresi, 2002; Plaza *et al.*, 2005; Benediktsson *et al.*, 2005; Fauvel *et al.*, 2008). In particular, we use the well-known local granulometries (Dougherty *et al.*, 1989), specifically granulometries by opening using squares of increasing size, as presented in (Soille, 2003, Ch. 11) for similar satellite images. Openings using squares (S) of increasing size λ obey the above definition. So, in our application we define a granulometric size distribution on f by: $1 - V(\gamma_S(f))/V(f)$, where V stands for the volume, *i.e.*, the sum of the pixel values.

METHODOLOGY

The proposed procedure uses two unsupervised learning procedures: trimmed *k*-means (Cuesta-Albertos *et al.*, 1997) and AA. First, these procedures are reviewed.

Trimmed *k*-means

Trimmed *k*-means is analogous to *k*-means but a proportion δ (between 0 and 1) of observations is discarded by the procedure itself, *i.e.*, the trimmed points are self-determined by the data. Trimmed *k*-means aims to robustify *k*-means, *i.e.*, to determine appropriate clusters when noisy data or outliers are present, which are detected and returned as trimmed observations by the procedure itself.

Let x_1, ..., x_n be n points of dimension p. Let k be the number of clusters. The k-means returns a set of k points, m_1^*, ..., m_k^*, the centroids, verifying ($\|\cdot\|$ denotes the Euclidean norm for vectors)

$$\{m_1^*,...,m_k^*\} = \arg \min_{m_1,...,m_k} \frac{1}{n} \sum_{i=1}^{n} \inf_{1\leq j\leq k} \|x_i - m_j\|^2, \quad (1)$$

and each observation x_i is assigned to its closest centroid m_j^*.

The trimmed k-means, with trimming size δ, returns k points, m_1^*, ..., m_k^* such that

$$\{m_1^*,...,m_k^*\} = \arg \min_{\mathbf{Y},\{m_1,...,m_k\}} \frac{1}{\lceil n(1-\delta)\rceil} \sum_{x_i\in\mathbf{Y}} \inf_{1\leq j\leq k} \|x_i - m_j\|^2, \quad (2)$$

where $\mathbf{Y}$ ranges on subsets of x_1, ..., x_n containing $\lceil n(1-\delta)\rceil$ points ($\lceil\cdot\rceil$ denotes the integer part of a given value). Each non-trimmed point x_i is assigned to its closest centroid m_j. A trimmed k-means algorithm can be found in García-Escudero *et al.* (2003).

Archetype analysis

In AA, three matrices are returned: 1) the k archetypes $\mathbf{z}_j$, which are the rows of a $k \times p$ matrix $\mathbf{Z}$; 2) an $n \times k$ matrix $\alpha = (\alpha_{ij})$ that contains the mixture coefficients that approximate each observation $\mathbf{x}_i$ by a mixture of the archetypes ($\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$); and 3) a $k \times n$ matrix $\beta = (\beta_{jl})$ that contains the mixture coefficients that define each archetype ($\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$). To find these matrices, we minimize the following residual sum of squares (RSS):

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 \\ &= \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2, \end{aligned} \quad (3)$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1,\ldots,n$ and
2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1,\ldots,k$.

To estimate those matrices Cutler and Breiman (1994) developed an alternating minimizing algorithm, which was implemented in R by Eugster and Leisch (2009).

If we know that the number of textures on the image is k, we can use this number for the AA algorithm. Otherwise, a simple but effective heuristic (Cutler and Breiman, 1994; Eugster and Leisch, 2009; Vinué *et al.*, 2015; Seth and Eugster, 2016) such as the elbow criterion can be used. With the elbow criterion, we plot the RSS for different k values and the value of k is selected as the point where the elbow is located.

Proposed procedure

The idea of the procedure is to extract textural features in small windows around each pixel of the image to segment. These textural features describe the windows. Then, we look for the extreme or archetypal features that correspond to the pure textures (AA is used). The textural features of the windows are expressed as a convex combination (the α matrix) of the archetypal features. One of the α values will be high (near one) for the windows with pure or nearly pure textures. However, windows with mixed textures because they are on the border between different textures will have intermediate α values. We can cluster the textural features using the α values to segment the image. If we assume that the percentage of windows with mixed textures because they are on the border between different textures is small, those windows would correspond to the trimmed observations of the trimmed k-means. Let us look at an implementation of this idea based on the description of textures by granulometric curves.

Our texture segmentation algorithm consists of the following steps: (a) Computing the granulometric curves within a small window around each pixel, (b) performing AA on the granulometric curves of a sample of pixels in the image and selecting a certain number k of archetypes by using the elbow criterion, (c) performing a trimmed k-means of the α values and (d) classifying every pixel in the image according to the results of step (c) by computing the α values corresponding to the archetypes obtained in step (b).

The granulometric curves are not part of the methodology and could be substituted for other appropriate textural features. Note that if the scales of new textural features are not comparable, they should be standardized before applying AA.

The selection of the window size depends on the image resolution, which should be high enough to capture the textures.

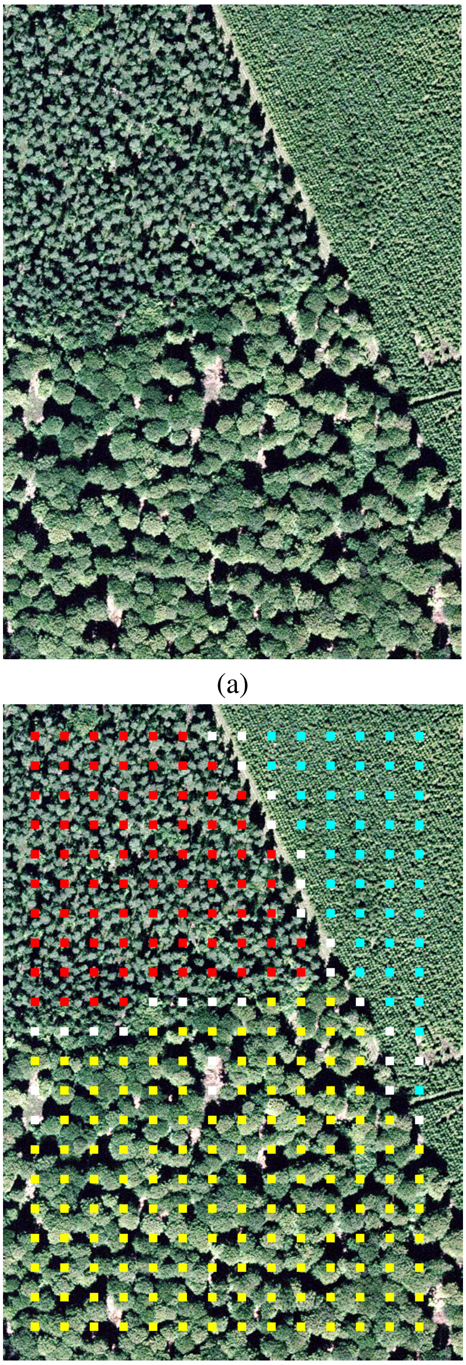
RESULTS

First the proposed procedure is illustrated by an image of several forests and compared with alternative unsupervised learning methods other than AA. In Sect. 3.1 the methodology is applied in a remote sensing problem and compared with other methodologies.

Fig. 2a shows the image to segment. We have computed the granulometric curves by opening using squares of increasing size (from 1 to 50), in windows with a size of 51 × 51 centered on a systematic sample of pixels, as shown in Fig. 2b, where each color indicates a different texture and white indicates that the pixel is on a border between textures. Those labels will act as a ground truth. The 294 granulometries are shown in Fig. 3. The distribution functions for the group colored in cyan move from 0 to 1 more rapidly (*i.e.*, they have a more concentrated probability distribution) than the corresponding ones from the group colored in red, and these move more rapidly than the distribution functions from the class colored in yellow. This occurs because the grain sizes (tree canopy) are small for the cyan class and they increase for the red and yellow classes.

AA is applied to the granulometries from $k = 1$ to 10. Fig. 4 shows the screeplot (RSS versus the number of archetypes). According to the elbow criterion, $k = 3$ is selected, which is the number of true texture classes, although it is determined in an unsupervised way. The archetypes (together with the α values) are computed for $k = 3$. The nearest granulometric curves to each archetype are the curves from the samples with numbers 28, 99 and 190. The windows surrounding those archetypal pixels are displayed in Fig. 5. In this way, we obtain a (pure) representative of each class.

The ternary plot with the α values is shown in Fig. 6. The different colors and numbers represent true classes, as in Fig. 2b, where the black points with the number 0 indicate the pixels on borders between textures. Note that the majority of points are situated around the archetypes, each of which is on one corner of the plot. In the middle of the plot there are some spread points, which are a mixture of the archetypes, and the majority of them are shown in black coded with zero, *i.e.*, they are border pixels. Therefore, our premises are met. We apply the trimmed k-means, with $k = 3$ and the proportion of trimmed sample $\delta = 0.095$, the percentage of pixels coded with zero, which can be considered as outliers. The matching matrix is shown in Table 1. The errors occur in border pixel classes. We have computed the adjusted Rand index (Hubert and Arabie, 1985) to compare the true partition


(a)

(b)

Fig. 2. *(a) Image of several forests. (b) Sampled pixels with their respective true label (see text for details).*

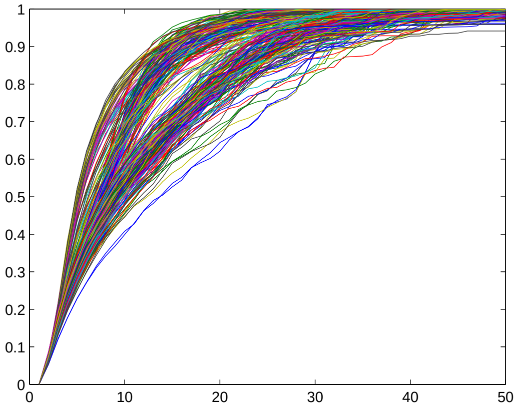


Fig. 3. *Granulometric curves.*

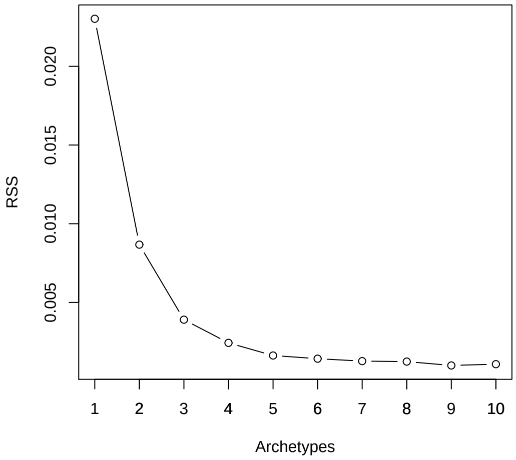


Fig. 4. *Screeplot of the residual sum of squares.*

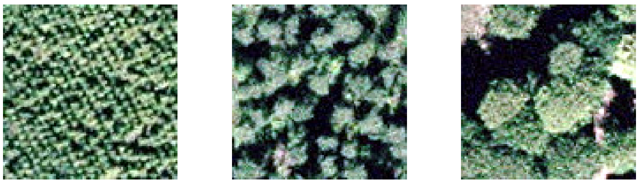


Fig. 5. *Archetypal representative of each group.*

with the one obtained using our methodology. The Adjusted Rand Index (ARI) is 0.87, which indicates a high level of agreement between the partitions (the closer to one, the more similar the partitions are).

We have also computed the ARI using different δ values in order to check the robustness of the results when the exact percentage of border pixels is not known. When $\delta = 0$, the classical k-means is applied. Furthermore, in order to test the improvement achieved by using AA instead of a clustering algorithm in step (b), for comparative purposes we change AA in step (b) for one of two possibilities. The first possibility, referred to as clustering, consists of applying the clustering algorithm (k-means or its trimmed version) directly to the granulometries. The second possibility, referred to as PC clustering, consists of applying the clustering algorithm (k-means or its trimmed version) to the principal component (PC) scores of the granulometries. The first two PCs are considered, since they explain more than 95% of the variability and better results are obtained than if we consider, for example, the first five PCs, which explain more than 99% of the variability. Table 2 shows the ARI values. On the one hand, for all the δ values, the highest ARI is obtained by AA, showing that AA returns the most similar the partition to the ground truth. On the other hand, for all the δ values, the ARI values for AA are high, in the majority of cases above 0.85.

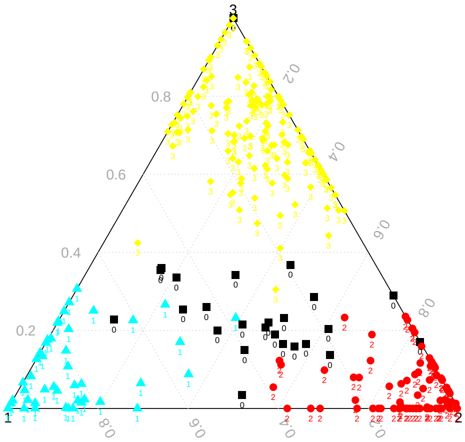


Fig. 6. *Ternary plot (see text for details about the color and number codes).*

Table 1. *Matching matrix on the sampled pixels (true class labels in the first row; labels obtained with out methodology in the first column).*

	0	1	2	3
0	17	3	1	6
1	6	77	0	0
2	4	0	136	0
3	1	0	0	43

APPLICATION

We consider the segmentation of orthophotos. Fig. 7 shows two 50 cm resolution images of the Alps used to assess the accuracy of a method monitoring the increase in woody vegetation from multitemporal Landsat images (Maggi *et al.*, 2007). These orthophotos need to be segmented into three different classes of tree densities: dense, sparse, and empty.

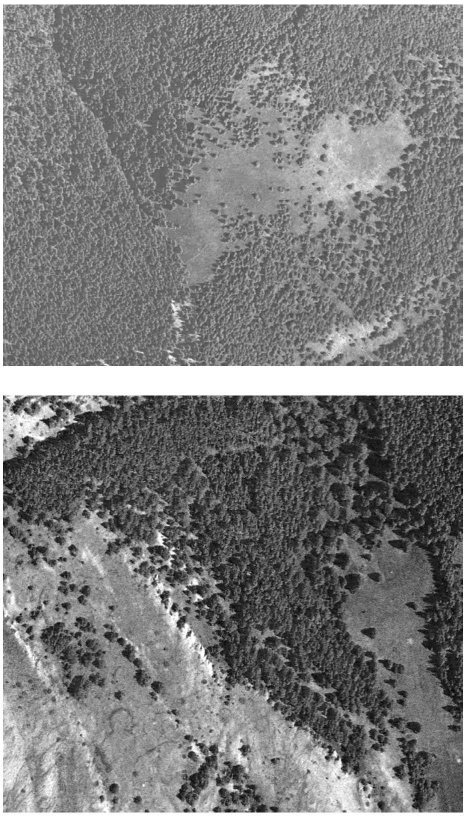


Fig. 7. *Orthophotos with dense, sparse, and empty tree densities.*

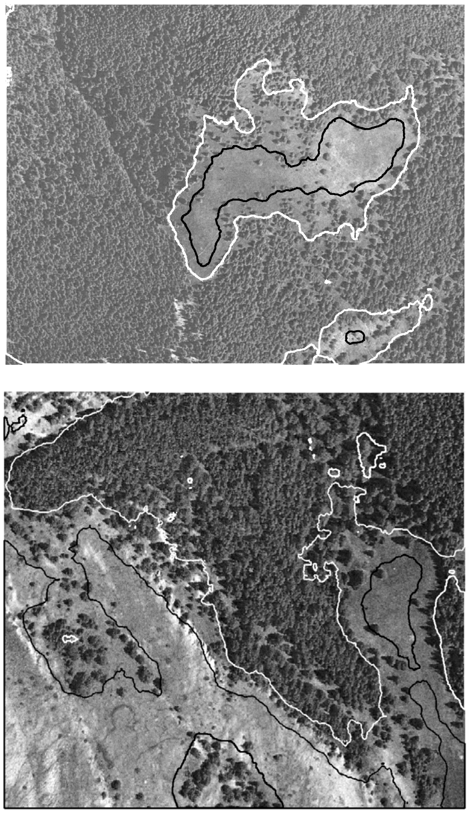


Fig. 8. *Supervised segmentation of orthophotos with the methodology described in Epifanio and Soille (2007).*

These orthophotos were originally used in Epifanio and Soille (2007) with a supervised texture segmentation methodology. This consisted of describing each pixel by several morphological features computed in a centered window of size 81, manually selecting nine prototypes from an orthophoto as the training data (five for dense, three for sparse, and one for the empty class), and using the minimum distance classifier, *i.e.*, the nearest neighbor according to the Euclidean distance. The segmentation results with this supervised methodology are shown in Fig. 8. The boundaries of the dense vegetation zones and the treeless zones are traced in white and black, respectively. The remaining areas correspond to the zones with sparse tree density. In Supplementary Material the labeled regions are visualized.

We now apply the approach described in Jain and Farrokhnia (1991) using Gabor filters to perform unsupervised texture segmentation, as explained in MathWorks (2019). A total of 32 Gabor features and 2 spatial features for each pixel in the input image are clustered with *k*-means. Fig. 9 shows the segmentation. Boundaries traced in white and black for the first orthophoto separate the classes returned. The obtained segmentation does not correspond to the density of trees. Even if we consider two or four groups instead three classes, the segmentation does not correspond to the density of trees (Fig. 10). For the second orthophoto, as before, the boundaries of the dense vegetation zones and the treeless zones are traced in white and black, respectively. There are several clear errors in this segmentation, for example, in the top left part of the image a dense vegetation zone is segmented as sparse; and a zone in the bottom right of the image with vegetation is classified as empty. In Supplementary Material the labeled regions are visualized.

We now apply our proposed procedure with windows of size 81. We apply AA and $\delta = 0$ for the trimmed *k*-means to facilitate comparison with the

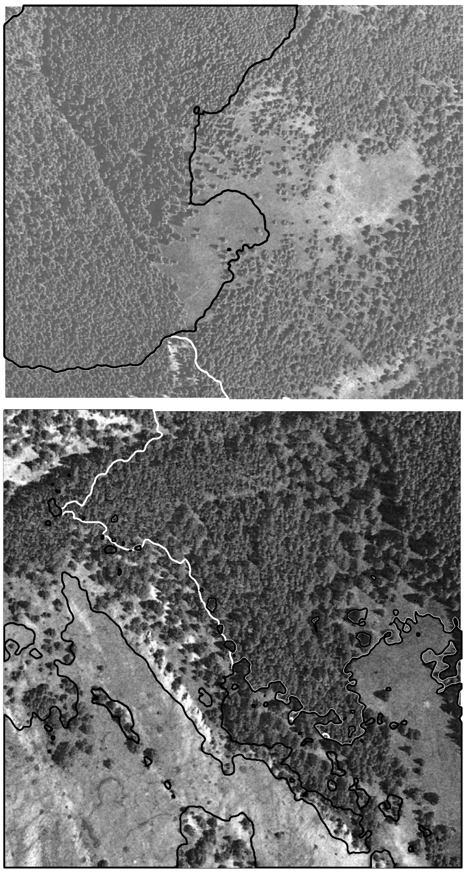


Fig. 9. *Unsupervised segmentation of orthophotos using Gabor filters.*

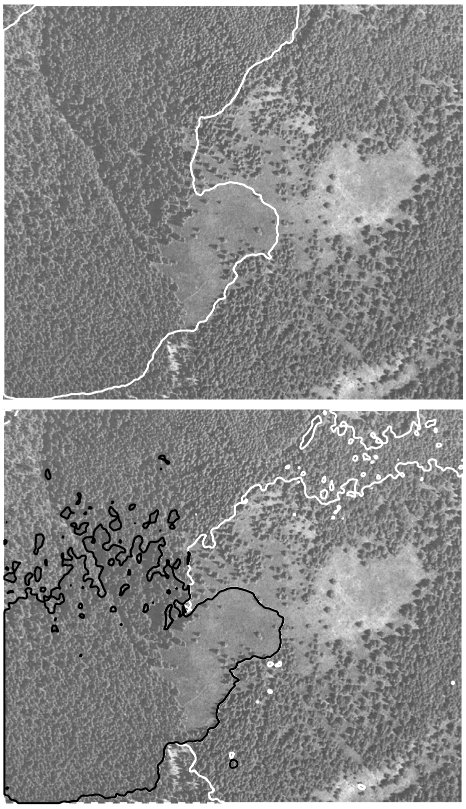


Fig. 10. *Unsupervised segmentation of the first orthophoto using Gabor filters with two and four classes.*

Table 2. *ARI for clustering, PC clustering and AA, using different δ values.*

δ	Clustering	PC clustering	AA
0.00	0.82	0.82	0.85
0.01	0.83	0.83	0.85
0.02	0.83	0.83	0.86
0.03	0.81	0.82	0.88
0.04	0.81	0.81	0.88
0.05	0.80	0.80	0.89
0.06	0.78	0.81	0.88
0.07	0.79	0.81	0.88
0.08	0.78	0.79	0.87
0.09	0.77	0.78	0.86
0.10	0.76	0.78	0.87
0.11	0.75	0.77	0.86
0.12	0.74	0.76	0.83
0.13	0.73	0.74	0.82
0.14	0.72	0.73	0.80
0.15	0.71	0.72	0.79

other methods, *i.e.*, no observation is trimmed. For the first orthophoto the elbow is at $k = 4$, so for the first orthophoto we consider four classes. The boundaries of the dense vegetation zones are traced in white (the areas surrounded by dashed lines have lower density than the solid lines).

The boundaries of the treeless zones are traced in black, and the remaining areas correspond to the zones with sparse tree density. Fig. 11 shows the segmentation results with our proposed procedure. If instead of using AA, the local granulometries are directly clustered by k-means, the segmentation obtained can be seen in Fig. 12. In this last case, some dense vegetation zones are misclassified as sparse tree density zones. In Supplementary Material the labeled regions are visualized.

Our procedure returns very satisfactory segmentation results, comparable with the results of the supervised methodology, despite being completely unsupervised, *i.e.*, no information is provided.

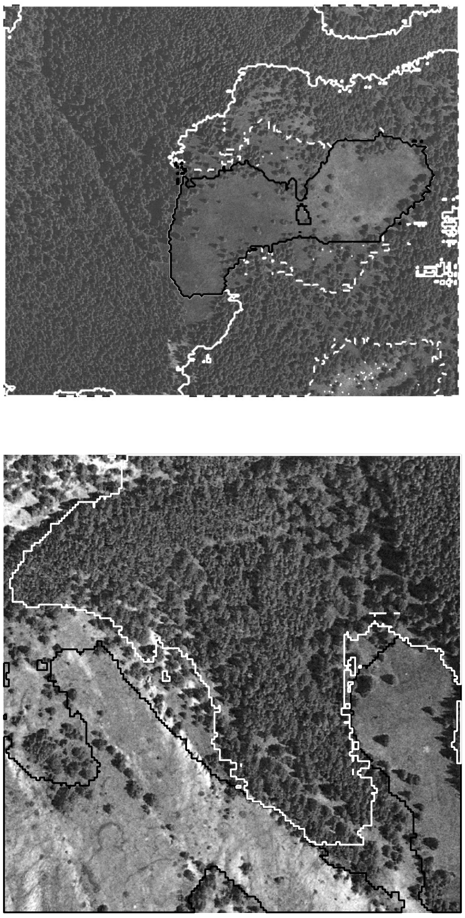


Fig. 11. *Unsupervised segmentation of orthophotos using our procedure.*

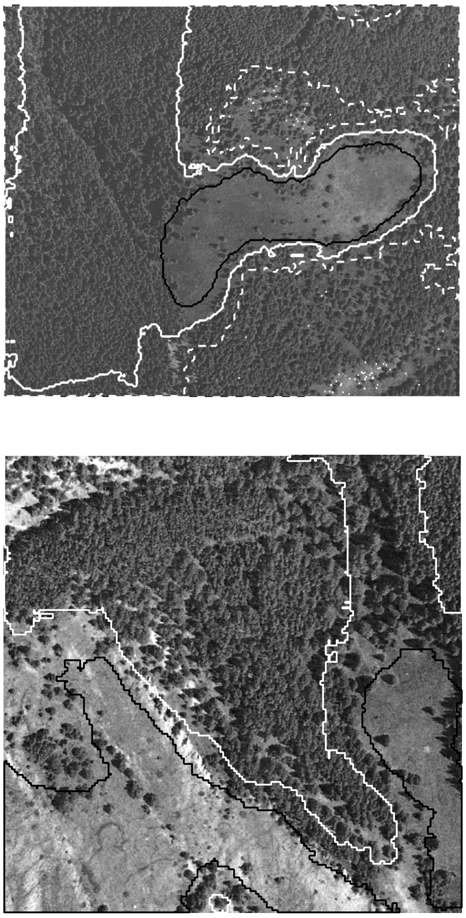


Fig. 12. *Unsupervised segmentation of orthophotos by clustering the local granulometries.*

DISCUSSION

A common problem of texture segmentation is that local windows centered on each pixel from which features are extracted can contain more than one texture. Some attempts to solve this consist of considering windows not centered on the pixel, as in Wang *et al.* (1993) and Epifanio and Soille (2007). However, those procedures are more labor-intensive than the procedure that we propose, where windows are centered on each pixel, and what we change is the clustering phase for an archetypal analysis phase followed by a clustering phase. The preliminary results show the relevance of the proposed approach. Our procedure has been compared with other methods. In spite of its simplicity, the relevance of its results has been shown, especially in a remote sensing application. The results of our procedure are competitive not only with other unsupervised methodologies, but also in comparison with a supervised methodology.

Our procedure can also return archetypal windows of each group. Instead of using the α values in the segmentation, the archetypal windows themselves could be used in the segmentation process by acting as prototypes.

Some additional points could also be studied. As an alternative to AA, we could use archetypoid analysis (ADA) (Vinué *et al.*, 2015), where instead of the archetypal representatives being built as a mixture of observations, they are actual observations. Simmilarly, instead of trimmed *k*-means, we could use trimmed *k*-medoids (Ibáñez *et al.*, 2012). On the other hand, we have treated textural features as multivariate features, since this is the most usual case in this context and we have preferred to do this due to comparison with other well-known techniques. We have not exploited the fact that granulometries are functions, and functional data analysis (Ramsay and Silverman, 2005) techniques could be used; in fact, functional archetypal analysis (Epifanio, 2016) could have been used.

For other applications, alternative textural features could be considered, since the proposed methodology does not depend on the selected textural features, but rather on the fact that they are discriminant between classes and the mixtures of textures are transferred to the vector of characteristics.

ACKNOWLEDGMENTS

This work has been partially supported by Grant DPI2017-87333-R from the Spanish Ministerio de Ciencia, Innovación y Universidades (AEI/FEDER, UE) and UJIB2017-13 from Universitat Jaume I. We would also like to thank Pierre Soille for providing us with the images.

REFERENCES

Bauckhage C, Kersting K, Hoppe F, Thurau C (2015). Archetypal Analysis as an Autoencoder. In: Workshop New Challenges in Neural Computation.

Benediktsson J, Palmason J, Sveinsson J (2005). Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Geosci Remote 43:480–91.

Chen Y, Mairal J, Harchaoui Z (2014). Fast and Robust Archetypal Analysis for Representation Learning. In: Proc 2014 IEEE Conf Comput Vision Pattern Recong (CVPR) 1478–85.

Cuesta-Albertos JA, Gordaliza A, Matrán C (1997). Trimmed *k*-means: an attempt to robustify quantizers. Ann Stat 25:553–76.

Cutler A, Breiman L (1994). Archetypal Analysis. Technometrics 36:338–47.

Dougherty ER, Kraus EJ, Pelz JB (1989). Image segmentation by local morphological granulometries. In: Proc 12th Can Symp Remote Sens Geosci Remote Sens Symp 3:1220–3.

Epifanio I (2016). Functional archetype and archetypoid analysis. Comput Stat Data An 104:24–34.

Epifanio I, Ayala G (2002). A random set view of texture classification. IEEE T Image Process 11:859–67.

Epifanio I, Ibáñez MV, Simó A (2018a). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. Am Stat, in press.

Epifanio I, Ibáñez MV, Simó A (2018b). Archetypal shapes based on landmarks and extension to handle missing data. Adv Data Anal Classi 12:705–35.

Epifanio I, Soille P (2007). Morphological texture features for unsupervised and supervised segmentations of natural landscapes. IEEE Geosci Remote 45:1074–83.

Epifanio I, Vinué G, Alemany S (2013). Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. Comput Ind Eng 64:757–65.

Eugster MJ, Leisch F (2009). From Spider-Man to Hero – Archetypal Analysis in R. J Stat Soft 30:1–23.

Eugster MJA (2012). Performance profiles based on archetypal athletes. Int J Perf Anal Spor 12:166–87.

Fauvel M, Benediktsson J, Chanussot J, Sveinsson J (2008). Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. IEEE Geosci Remote 46:3804–14.

Fletcher ND, Evans AN (2005). Texture segmentation using area morphology local granulometries. In: Ronse C, Najman L, Decencière E, eds., Mathematical Morphology: 40 Years On. Dordrecht: Springer.

García-Escudero LA, Gordaliza A, Matrán C (2003). Trimming tools in exploratory data analysis. J Comput Graph Stat 12:434–49.

Hinrich JL, Bardenfleth SE, Roge RE, Churchill NW, Madsen KH, Mørup M (2016). Archetypal analysis for modeling multisubject fMRI data. IEEE J Sel Top Signa 10:1160–71.

Hubert L, Arabie P (1985). Comparing partitions. J Classif 2:193–218.

Ibáñez MV, Vinué G, Alemany S, Simó A, Epifanio I, Domingo J, Ayala G (2012). Apparel sizing using trimmed PAM and OWA operators. Expert Syst Appl 39:10512–20.

Jain AK, Farrokhnia F (1991). Unsupervised texture segmentation using Gabor filters. Pattern Recogn 24:1167–86.

Maggi M, Estreguil C, Soille P (2007). Woody vegetation increase in alpine areas: a proposal for a classification and validation scheme. Int J Remote Sens 28:143–66.

Mair S, Boubekki A, Brefeld U (2017). Frame-based data factorizations. In: Proc 34th Int Conf Machine Learn (ICML'17) 70:2305–13.

Matheron G (1975). Random sets and integral geometry. New York: Wiley.

MathWorks (2019). Texture segmentation using Gabor filters. Accessed 2019-04-05.

Millán-Roures L, Epifanio I, Martínez V (2018). Detection of anomalies in water networks by functional data analysis. Math Probl Eng 2018:5129735

Moliner J, Epifanio I (2019). Robust multivariate and functional archetypal analysis with application to financial time series analysis. Physica A 519:195–208.

Mørup M, Hansen LK (2012). Archetypal analysis for machine learning and data mining. Neurocomputing 80:54–63.

Plaza A, Martinez P, Plaza J, Perez R (2005). Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. IEEE Geosci Remote 43:466–79.

R Development Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ragozini G, Palumbo F, D'Esposito MR (2017). Archetypal analysis for data-driven prototype identification. Stat Anal Data Min 10:6–20.

Ramsay JO, Silverman BW (2005). Functional data analysis, 2nd ed. New York: Springer.

Seth S, Eugster MJA (2016). Probabilistic archetypal analysis. Mach Learn 102:85–113.

Soille P (2003). Morphological image analysis: Principles and applications, 2nd ed. Berlin, Heidelberg: Springer.

Soille P, Pesaresi M (2002). Advances in mathematical morphology applied to geoscience and remote sensing. IEEE Geosci Remote 40:2042–55.

Sun W, Yang G, Wu K, Li W, Zhang D (2017a). Pure endmember extraction using robust kernel archetypoid analysis for hyperspectral imagery. ISPRS J Photogramm 131:147–59.

Sun W, Zhang D, Xu Y, Tian L, Yang G, Li W (2017b). A probabilistic weighted archetypal analysis method with Earth mover's distance for endmember extraction from hyperspectral imagery. Remote Sensing 9:841.

Thøgersen JC, Mørup M, Damkiær S, Molin S, Jelsbak L (2013). Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. BMC Bioinformatics 14:279.

Tsanousa A, Laskaris N, Angelis L (2015). A novel single-trial methodology for studying brain response variability based on archetypal analysis. Expert Syst Appl 42:8454–62.

Tuceryan M (1994). Moment-based texture segmentation. Pattern Recogn Lett 15:659–68.

Tuceryan M, Jain AK (1993). Texture analysis. In: Chen CH, Pau LF, Wang PSP, eds., Handbook of Pattern Recognition & Computer Vision. River Edge: World Scientific, 235–76.

Vinué G (2017). Anthropometry: An R package for analysis of anthropometric data. J Stat Soft 77:1–39.

Vinué G, Epifanio I (2017). Archetypoid analysis for sports analytics. Data Min Knowl Disc 31:1643–77.

Vinué G, Epifanio I, Alemany S (2015). Archetypoids: A new approach to define representative archetypal data. Comput Stat Data An 87:102–15.

Wang D, Haese-Coat V, Bruno A, Ronsin J (1993). Texture classification and segmentation based on iterative morphological decomposition. J Vis Commun Image R 4:197–214.

Submitted to Image Anal Stereol, 3 pages
Original Research Paper

SUPPLEMENTAL MATERIAL OF “ARCHETYPAL ANALYSIS: AN ALTERNATIVE TO CLUSTERING FOR UNSUPERVISED TEXTURE SEGMENTATION”

ISMAEL CABERO[1] AND IRENE EPIFANIO[2]

[1]Department of Didactics of Mathematics, Universitat de València, Avda. Tarongers, 4, 46022 València, Spain,
[2]Department of Mathematics and IMAC, Universitat Jaume I, Campus del Riu Sec, Castelló, 12071, Spain
e-mail: ismael.cabero@uv.es, epifanio@uji.es
(Submitted)

ABSTRACT

We visualize the segmented regions of the Application Section of “Archetypal analysis: an alternative to clustering for unsupervised texture segmentation”.

Keywords: Archetype, Image segmentation, Local granulometries, Mathematical morphology, Texture analysis.

RESULTS OF APPLICATION

The segmentation results with the supervised methodology are shown in Fig. 1 (it corresponds to the Fig. 8 of the main paper). The black areas correspond to the dense vegetation zones, the white regions to the treeless zones, and the remaining areas to the zones with sparse tree density.

Fig. 2 shows the segmentation with Gabor filters. (It corresponds to the Fig. 9 of the main paper). For the first orthophoto each color represents one of the classes returned. The obtained segmentation does not correspond to the density of trees. Even if we consider two or four groups instead three classes, the segmentation does not correspond to the density of trees (see Fig. 3). (This corresponds to the Fig. 10 of the main paper). For the second orthophoto, as before, the black areas correspond to the dense vegetation zones, while the white regions correspond to the treeless zones.

Fig. 4 shows the segmentation results with our proposed procedure. (It corresponds to the Fig. 11 of the main paper). The black or gray areas correspond to the dense vegetation zones (the gray areas have lower density than the black areas), the white regions to the treeless zones, and the remaining areas to the zones with sparse tree density. If instead of using AA, the local granulometries are directly clustered by *k*-means, the segmentation obtained can be seen in Fig. 5. (This corresponds to the Fig. 12 of the main paper).

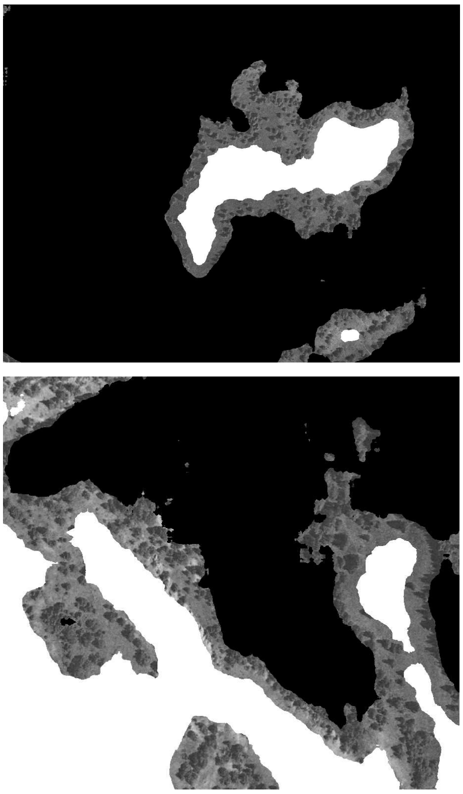


Fig. 1. *Supervised segmentation of orthophotos with the methodology described in Epifanio and Soille (2007).*

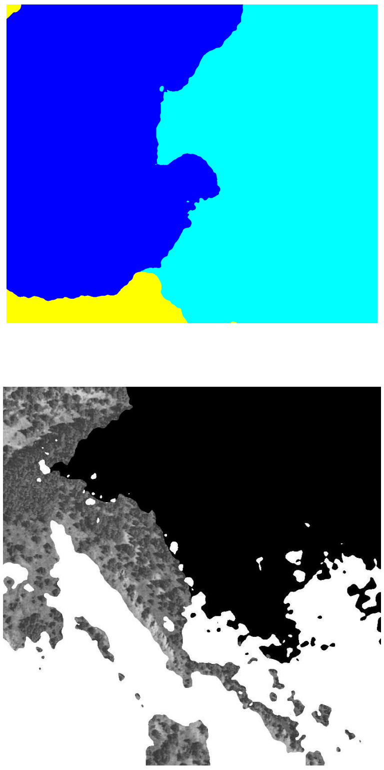


Fig. 2. *Unsupervised segmentation of orthophotos using Gabor filters.*

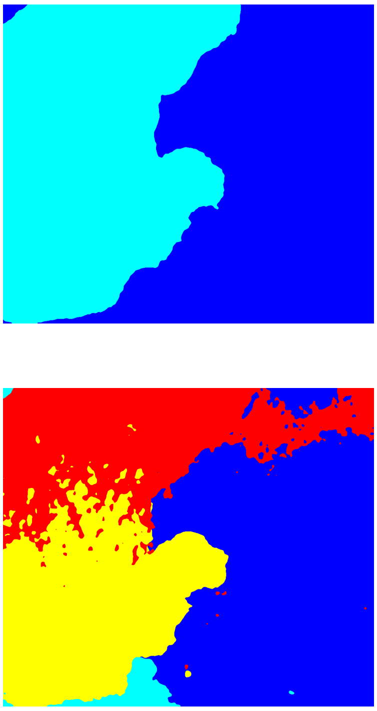


Fig. 3. *Unsupervised segmentation of the first orthophoto using Gabor filters with two and four classes.*

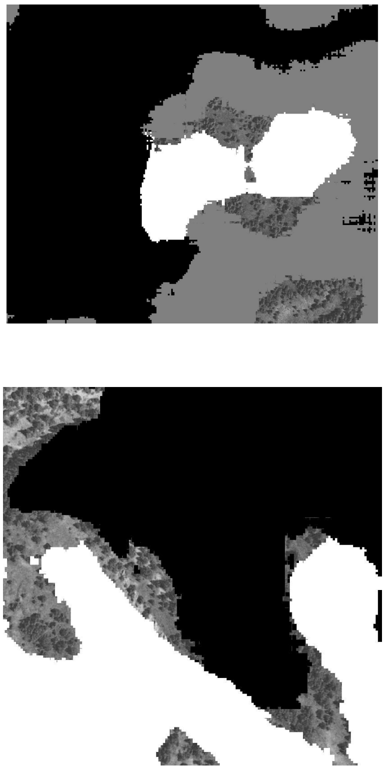


Fig. 4. *Unsupervised segmentation of orthophotos using our procedure.*

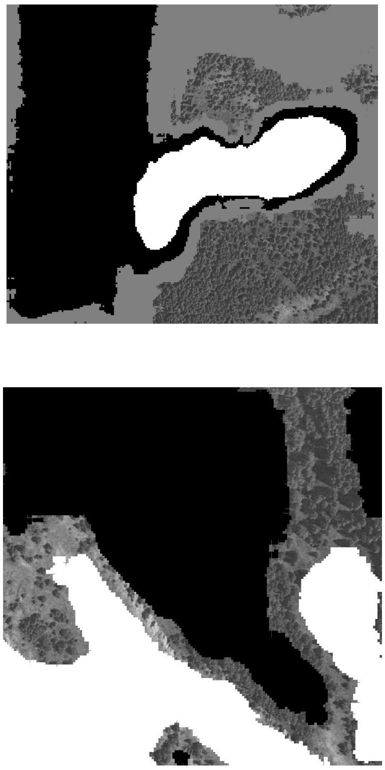


Fig. 5. *Unsupervised segmentation of orthophotos by clustering the local granulometries.*

REFERENCES

Epifanio I, Soille P (2007). Morphological texture features for unsupervised and supervised segmentations of natural landscapes. IEEE Transactions on Geoscience and Remote Sensing 45:1074–83.

3.4 Aportació III: Archetype analysis based algorithm to find outliers [3]

[3] Actualment Cabero, Epifanio, Piérola and Ballester (2019) está sotmés i es presentà a la XVII Conferència Espanyola i VII Trobada Iberoamericana de Biometria CEB-EIB 2019 (Cabero, Epifanio and Piérola; 2019).

Archetypal analysis-based method to find outliers[☆]

Ismael Cabero[a], Irene Epifanio[b,*], Ana Piérola[c], Alfredo Ballester[c]

[a]*Departament de Didàctica de la Matemàtica. Universitat de València, 46022 València, Spain*
[b]*Dept. Matemàtiques and Institut de Matemàtiques i Aplicacions de Castelló. Universitat Jaume I, 12071 Castelló, Spain*
[c]*Institut de Biomecànica de València, 46022 València, Spain*

Abstract

The problem of detecting outliers in multivariate data sets with continuous numerical variables is addressed by a new method. This method combines projections into relevant subspaces by archetype analysis with a nearest neighbor algorithm, through an appropriate ensemble of the results. A procedure for converting the outlier scores returned by the method into binary labels is also proposed. The method is illustrated by several simple examples and assessed with a battery of outlier detection algorithms with several benchmark data sets. The comparison shows that the performance of our proposal is very favorable. Finally, a novel industrial data set is introduced, and an outlier analysis is carried out to improve the fit of footwear, since this kind of analysis has never been fully exploited in the anthropometric field.

Keywords: Archetype Analysis, Unsupervised Anomaly Detection, Nearest neighbors, Ensembles, Multivariate Outlier Detection, Footwear

1. Introduction

Nowadays, we tend to work with enormous amounts of data and variables, which greatly hinders their analysis. It is necessary to perform a quality analysis to avoid making wrong decisions. One of the possible causes of such decisions is outliers. A classic definition of an outlier given by Hawkins (1980), is "an

[☆] The code of our procedure, the artificial examples and the novel industrial data set for reproducing the results are available at `http://www3.uji.es/~epifanio/RESEARCH/AAout.rar`, while the rest of the data sets and code used in the comparison are available at Goldstein and Uchida (2016).

[*] Tel.: +34 964728390; fax: +34 964728429.

Email addresses: `ismael.cabero@uv.es` (Ismael Cabero), `epifanio@uji.es` (Irene Epifanio), `ana.pierola@ibv.org` (Ana Piérola), `alfredo.ballester@ibv.org` (Alfredo Ballester)

observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Outliers can also be defined as observations whose characteristics differ significantly from the normal profile.

Detection of outliers is an early and necessary step in any data analysis application. Although outliers are considered in many cases as noise or errors, they often incorporate vital information. But it is clear that what is noise for one person can be a focus of interest for another (Johnson et al., 1998), and depending on what you are studying, an outlier (e.g. in the detection of credit card fraud) can be of great importance. Failure to look for them and study them can lead to poor specification of the model and an incorrect estimate of its parameters. It is therefore important to identify them prior to modeling and analysis (Williams et al., 2002).

The appearance of outliers can be due to many causes, such as errors, anomalies, inaccuracies generated in the process of coding or transmission of the actual data, incorrect transformations to use the data in statistical analysis or false assumptions in the distribution of the data (Anscombe, 1960). However, they can also occur as a result of fraudulent behavior, or incorrect reports in a random or intentional manner, or they may be conditioned by elements that are external to the population studied (such as environmental conditions). They can also be the result of incorrect sample selection. All these causes of outliers open up a huge range of possibilities for their uses. Some examples to show their repercussion and impact include detection of credit card fraud (Bolton et al., 2001), insurance or health care claim fraud (Chandola et al., 2009), quality control and fault detection (Rousseeuw and Driessen, 1999), web-log anomalies, email spam, anomalies in economic applications (Galeano et al., 2006; Peña and Prieto, 2007), network intrusion (Teng et al., 1990), cellular fraud (Fawcett and Provost, 1997), clinical processes (Penny and Jolliffe, 2001), military surveillance for enemy activities (Chandola et al., 2009), data cleaning (Rousseeuw and Hubert, 2018), etc. (see Aggarwal (2017) for a detailed discussion about applications).

Identyfing outliers in univariate data is relatively simple because it is easy to find the extreme cases. However, in the multivariate case, the detection of outliers is more difficult because multidimensional outliers are observations that are considered strange not because of the value they take in a certain variable, but due to the value in all of them (Gnanadesikan and Kettenring, 1972).

Many techniques have been proposed for the detection of outliers along time (see Aggarwal (2017) for a detail explanation of many of them for different types of data). In the case of multivariate data, Goldstein and Uchida (2016) reviewed and compared many of the most standard unsupervised anomaly detection algorithms in a set of benchmark data sets. A similar study was carried out by Campos et al. (2016) and Domingues et al. (2018). Goldstein and Uchida (2016) proposed a taxonomy of unsupervised anomaly detection algorithms, which are divided into four categories: (1) Nearest-neighbor (NN) based techniques, (2) Clustering-based methods, (3) Statistical algorithms, and (4) Subspace techniques.

We propose a new method for unsupervised (no labels are available) detection of outliers in continuous multivariate data. It can be categorized into several of those categories, mainly (1) and (4), because it uses an unsupervised learning technique (a subspace technique), which can also be used as a clustering technique (Epifanio et al., 2018a), and it also relies on NN-based techniques. Note that techniques based on distances are very popular due to their good results, conceptual simplicity and interpretability. However, when the number of variables is high, these techniques can fail because of the curse of dimensionality. A key point to solve this problem would be to eliminate the dimensions and project the data into subspaces, where outliers can be easily revealed. This is the idea of the proposed method: first to project the data into the relevant subspaces and then to use proximity-based techniques to detect outliers in those subspaces.

The proposed method, which we refer to as AA + k-NN, is based on Archetype Analysis (AA), the objective of which is to represent the observations by means of a mixture of archetypes, which are a mixture of observations. Archetypes lie on the boundary of the convex hull of the data, meaning that they are extreme profiles. This makes AA sensitive to outliers, and we will take advantage of this in order to detect outliers. AA is not a parametric technique, it is a data-driven method, so we do not have to make any assumption about data distribution. Furthermore, the results returned by AA are easily interpretable, even for non-experts. The combination of AA together with proximity-based methods therefore results in a non-parametric method with a high level of interpretability, which is very important in many applications.

AA was defined by Cutler and Breiman (1994) and has been applied in a broad spectrum of fields, such as biology (D'Esposito et al. (2012)), developmental psychology (Ragozini et al. (2017)), didactics (Cabero (2018)), engineering (Epifanio et al. (2013); Vinué et al. (2015); Vinué (2017); Epifanio et al. (2018b); Millán-Roures et al. (2018)), finance (Moliner and Epifanio, 2019), genetics (Thøgersen et al. (2013)), global development (Epifanio (2016)), machine learning problems (Mørup and Hansen (2012)), market research (Porzio et al. (2008)), multi-document summarization (Canhasi and Kononenko (2014)), neuroscience (Tsanousa et al. (2015); Hinrich et al. (2016)) and sports (Eugster (2012); Vinué and Epifanio (2017)).

The main contributions of this paper are as follows: we present a new method for unsupervised detection of outliers in multivariate data. We conduct an experimental evaluation with a large number of well-known data sets and standard algorithms. In this comparison our new proposal provides very favorable results.

Furthermore, we apply the new method to an original data set of foot measurements, which is used in an engineering problem that we introduce here. Outlier detection in Anthropometry has only been used as a cleaning technique for correcting or removing the outliers before analyzing data (Kouchi, 2014; Kuehnapfel et al., 2016). However, outliers report very valuable information in the footwear design process, since they can show which kinds of feet are more different from the rest and may therefore pose fitting problems in footwear if the design is not appropriate.

The rest of this article is organized as follows. Section 2 presents the data sets and standard methods used in the comparison, and AA is also reviewed. In Section 3, we introduce our method. Section 4 presents the results of the comparison. The new methodology is applied to a new data set in an engineering problem in Section 5. Finally, we finish with some conclusions and future prospects for further research in Section 6.

2. Preliminary

2.1. Benchmark data sets

We use all the data sets employed by Goldstein and Uchida (2016) except for two of them, which were excluded because they contain mixed data (numerical and categorical data). Therefore, our benchmark data sets contain numerical data, although some of them contain not only continuous numerical variables, but also discrete numerical variables. Remember that our proposed method is appropriate for continuous numerical data. The data sets have been obtained from multiple sources, such as Dheeru and Karra Taniskidou (2017), and can be found in Goldstein (2015). The data sets contain a variable with labels that indicate whether or not an observation is an outlier. However, we work with an unsupervised anomaly detection method and labels will not be used, except at the end for assessing the results. Details about the construction of the data sets can be found in Goldstein and Uchida (2016), but a brief summary of them is given below.

Breast Cancer Wisconsin (Diagnostic): This data set is composed of 367 individuals with 30 different variables and 10 anomalies, which represent 2.72%. This data set focuses on the diagnosis of breast cancer to discriminate between benign and malignant tumors. It includes a set of features of cell nuclei from a digitized image of a fine needle aspirate (FNA) from a breast mass. The anomalies correspond to malignant instances, while the rest of the data set consists of benign instances.

Pen-Based Recognition of Handwritten Text (global): This set has 16 features and 809 observations, 11.1% of which are anomalies. This data set contains handwritten digits; in particular, the digit 8 is considered as the normal class and a sample of 10 digits from all of the other classes are considered anomalies. Therefore, there is a large normal class, and the anomalies are very different from each other.

Pen-Based Recognition of Handwritten Text (local): This data set also contains 0-9 handwritten digits. Specifically, it has 6724 cases with 16 variables. There are 9 large clusters corresponding to all digits, except the digit 4. For this class, only 10 cases are considered and they are therefore considered outliers, representing 0.15% of the data.

Letter Recognition: The data set has 1600 observations and 32 features, extracted from the 26 letters of the English alphabet. Three letters constitute the normal class, while anomalies come from small samples of the rest of the letters and represent 6.25% of the data.

Speech Accent Data: This data set contains real world data from recorded English language, the features of which are 400-dimensional vectors. This set has 3686 observations with 1.65% anomalies, corresponding to speech segments from speakers who do not have an American accent, which is the normal class.

Landsat Satellite: This data set contains 5100 instances with 1.49% anomalies and 36 features. These features were extracted from satellite observations.

Statlog Shuttle: This data set describes radiator positions in a NASA space shuttle; it contains 46,464 observations with 9 features, including 3.02% anomalies.

Object Images (ALOI): This data set contains 27-dimensional features extracted from images of small objects taken under different light conditions and viewing angles. It has 50,000 observations, 3.02% of which are anomalies.

Before applying any method for detecting anomalies, the data should be preprocessed so that the features have equal weights. Goldstein and Uchida (2016) use classic min-max normalization (the range transformation scales the data to be within $[0, 1]$). With our proposal, besides min-max normalization, we also use standardization (the mean is subtracted and values are divided by the standard deviation), since it is the common preprocessing procedure in AA.

2.2. Unsupervised Anomaly Detection Algorithms

There are a huge number of unsupervised anomaly detection algorithms. Let us take a quick look at the most widely used in practice and those used by Goldstein and Uchida (2016). Furthermore, these algorithms will be used in the comparison.

k-NN Anomaly Detection: This algorithm searches for the nearest k-neighbors for every element in the database and calculates the average distance of the k-neighbors. This procedure returns outlier scores, which depends on the selection of k. In the experiments, we follow the same strategy as in Goldstein and Uchida (2016): values from $k = 10$ to 50 are considered and averaged in order to achieve a fair evaluation when comparing algorithms. It focuses on global outliers.

kth-NN Global Anomaly Detection: As above, but once we have the nearest k-neighbors, only the distance of the k-th nearest neighbor is considered. It also focuses on global outliers.

Local Outlier Factor (LOF) and LOF-upper bound (LOF-UB): This algorithm is designed to find local outliers. It follows these steps: 1) search for the k-NN for each observation; 2) compute the local density for each observation; 3) the LOF score is computed by comparing the local densities of each observation with those of its k neighbors. See Breunig et al. (2000) for details. This algorithm finds local outliers and also global ones, but if we are only interested in global outliers, we will have a lot of false alarms. The choice of k will have a great influence on the results. Therefore, we will follow the same strategy as in Goldstein and Uchida (2016): scores for different ks up to an upper bound are calculated and the maximum of these scores is considered. This strategy is referred to as LOF. However, we can also consider different upper bounds and average the results. This strategy is referred to as LOF-UB.

Connectivity-Based Outlier Factor (COF): This algorithm, proposed by Tang et al. (2002) works like LOF except that instead of using the Euclidean distance, COF uses the "chaining distance"; this distance is the minimum sum of all the distances connecting all the k-neighbors and the case. The objective of changing the distance is to avoid the lack of precision of LOF when the density of the data that is around the observation has some kind of linear correlation, which is not appreciated with the inherent sphericity of the Euclidean distance.

Influenced Outlierness (INFLO): If the data set has close clusters with very different densities, it is possible that an algorithm such as LOF may identify the border points between cluster as outliers. To avoid this, Jin et al. (2006) proposed to work just like LOF but also taking into account the "reverse neighbors". This makes it possible to calculate the outlier scores of that kind of points more precisely.

Local Outlier Probability (LoOP): LoOP (Kriegel et al., 2009) instead of assigning a score to the outlier, it gives the probability that each element is an outlier. LoOP also studies the local density of the neighbors around each element, but it assumes that the distances to the nearest neighbors follow a Gaussian distribution.

Local Correlation Integral (LOCI): This algorithm tries to eliminate the difficulty of choosing the best k (number of neighbors). To do this, instead of looking at the nearest k-neighbors, it takes a circle of radius r around the case and study the density that exists in it. This radius r expands over time and, like LoOP, it also calculates the density using a Gaussian average distribution and compares two neighborhoods of different sizes instead of the ratio of local densities. A parameter α controls the relationship between the different neighborhoods. See Papadimitriou et al. (2003) for details. Computationally, it is a very expensive algorithm and is too slow for large data sets so, as stated by Hofmann and Klinkenberg (2013), it can only be applied to very small data sets (at most 3000 observations).

Approximate Local Correlation Integral (aLOCI): In order to reduce the computational cost of LOCI, this algorithm uses quad trees and some restrictions on α. However, due to these approximations, its performance can sometimes be very poor (Hofmann and Klinkenberg, 2013; Goldstein and Uchida, 2016).

Cluster-Based Local Outlier Factor (CBLOF/ uCBLOF): The algorithm proposed by He et al. (2003) no longer uses the NNs to estimate density, but divides the data into different clusters and determines the density of each cluster. The most commonly used algorithm is k-means due to its small computational cost. Then clusters are classified as large and small. For the large ones, a weighted distance from the center to each element of the cluster is calculated and for the small ones the distance to the nearest large cluster is calculated. However, the weighting strategy used can lead to

an incorrect density estimation (Goldstein and Uchida, 2016). A modified version to solve this problem is uCBLOF, which simply neglects the weighting. Algorithms based on cluster analysis still have a problem similar to those of k-neighbors, because they have to choose the number k of clusters.

Local Density Cluster-based Outlier Factor (LDCOF): LDCOF (Amer and Goldstein, 2012) is analogous to the previous procedure, but now for each cluster the average distance of all cluster members to the centroid is calculated. Then the score is calculated by dividing the distance of an observation to its cluster center by the average distance.

Clustering-based Multivariate Gaussian Outlier Score (CMGOS-Red, Reg and MCD): This algorithm works like the previous ones, finding clusters with the k-means and separating them into large and small ones. For each cluster, the covariance matrix $\mathbf{\Sigma}$ is robustly estimated by three different procedures that give rise to the CMGOS-Red, CMGOS-Reg and CMGOS-MCD algorithms (see Goldstein and Uchida (2016) for details). The outlier score is calculated by dividing the Mahalanobis distance from one observation to its nearest cluster center by the certain percentile of the chi-square distribution.

Histogram-based Outlier Score (HBOS): HBOS (Goldstein and Dengel, 2012) is an algorithm that assumes the independence of the variables. For each variable, a histogram is computed and normalized, and the height of each bin is used to compute the outlier scores. The histogram can be created in different ways, and the number k of bins also influences the results.

Robust Principal Component Analysis (rPCA): Shyu et al. (2003) use robust principal component analysis, and in particular, the major and minor components.

One-Class Support Vector Machine (oc-SVM and ν-oc-SVM): one-class SVM with robust techniques and a modification in the objective function (a ν parameter is included) is trained using the data set and afterwards, each observation is scored by a normalized distance to the determined decision boundary (Amer et al., 2013).

In order to assess the algorithms for detection of outliers with unsupervised data, apart from taking into account the accuracy, the order of the outliers must be considered, especially because outlier scores are available. Therefore, we reproduce the same strategy followed by Goldstein and Uchida (2016), which consists of ranking the outlier scores and iteratively applying a threshold from the first to the last rank. In this way, n tuple values (true positive rate and false positive rate) are obtained, and a single receiver operator characteristic (ROC) is generated. As an assessment measure we use the integral of the ROC, i.e. the area under the curve (AUC). Note that the AUC value can be interpreted as the probability that an outlier detection algorithm will assign a lower score to a randomly chosen normal observation than to a randomly chosen anomalous observation (Fawcett, 2006). On the other hand, note that many algorithms depend on a parameter, e.g. k for all the NN or clustering-based algorithms, which can be critical. In order to ensure fair comparisons, we also follow the same strategy as Goldstein and Uchida (2016) and compute the AUC from $k = 10$ to $k = 50$. Then the AUC results are averaged and the standard deviation is also computed.

2.3. Archetype Analysis

AA is an unsupervised statistical learning technique (Hastie et al., 2009, Chapter 14). AA seeks out extreme profiles called archetypes, which are restricted to being a convex combination of the elements of the database. Also, these archetypes will represent each individual in our database as a convex combination of the archetypes. Expressing the observations as mixtures of extremes profiles facilitates comprehension of the data. Humans understand the data better when the instances are shown through their extreme constituents (Davis and Love, 2010) or when features of one instance are shown as opposed to those of another (Thurau et al., 2012).

AA lies somewhere in between two well-known unsupervised statistical techniques: Principal Component Analysis (PCA) and cluster analysis. Those techniques are also data decomposition techniques, where a data matrix is decomposed as a linear combination of several factors to find the latent components. Depending on the decomposition, different techniques are obtained. A table summarizing the relationship between several unsupervised techniques is provided by Mørup and Hansen (2012) and Vinué et al. (2015). With PCA, factors are linear combinations of features; therefore, they are the least restrictive. On the other hand, PCA bases are the most difficult to interpret, while the factors of clustering techniques, such as the centroids (averages of groups of data) of k-means, have more restrictions in their set-up, but their interpretation is very easy. However, their modeling flexibility is compromised due to the binary assignment of data to the clusters. Nevertheless, AA lies in between PCA and cluster tools, with higher modeling flexibility than cluster techniques but without losing the interpretability of their factors.

Let $\mathbf{X}$ be an $n \times m$ matrix that represents a database with n observations and m variables. Our goal is to find a $p \times m$ matrix $\mathbf{Z}$ that characterizes the archetypal patterns of the data so that each data point can be represented as a mixture of these archetypes. Specifically, AA tries to obtain the two $n \times p$ matrices of the coefficients α and β that minimize the residual sum of the squares (RSS) that arise from the equation that shows x_i as an approximation of a convex combination of archetypes z_j and the equations that show z_j as a convex combination of data:

$$RSS = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} \sum_{l=1}^{n} \beta_{ji} x_i\|^2, \tag{1}$$

with two conditions: 1) $\sum_{j=1}^{p} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, ..., n$, and 2)$\sum_{l=1}^{n} \beta_{ji} = 1$ with $\beta_{ji} \geq 0$ and $j = 1, ..., p$.

Therefore from 1) the approximations of x_i are a finite archetypal mixture $\hat{x}_i = \sum_{j=1}^{k} \alpha_{ij} z_j$ and the α_{ij} will indicate the weight of each archetype z_j for the element x_i. On the other hand, restriction 2) will show that the archetypes z_j are convex combinations of the data, $z_j = \sum_{l=1}^{n} \beta_{jl} x_l$.

AA is an exploratory data analysis (EDA) tool that is based on a geometric formulation (no distribution of data is assumed). Cutler and Breiman (1994) showed that archetypes are on the boundary of the convex hull of the data if $p > 1$ (the archetype coincides with the mean for $p = 1$).

Cutler and Breiman (1994) developed an alternating minimizing algorithm to compute the matrices in the AA problem, where the best α for given archetypes $\mathbf{Z}$ and the best archetypes $\mathbf{Z}$ for a given α are estimated by turns. A penalized version of the non-negative least squares algorithm by Lawson and Hanson (1974) is used to solve the convex least squares problems. That algorithm was implemented in the R package **archetypes** by Eugster and Leisch (2009), although with some modifications. For example, the spectral norm in equation 1 is used instead of the Frobenius norm for matrices. We reverted those modifications in our R implementation, i.e. the objective function to minimize is defined by equation 1. This algorithm is not deterministic, AA is run beginning from 20 random initializations, and the best model is selected for each p.

Archetypes are not necessarily nested or orthogonal to one another, so the selection of p is an important issue. A simple but effective heuristic tool for choosing p, which has been used elsewhere (Cutler and Breiman, 1994; Eugster and Leisch, 2009; Vinué et al., 2015; Seth and Eugster, 2016), is the elbow criterion. With the elbow criterion, we plot the RSS for different p values and the value of p is selected as the point where the elbow is located. Nevertheless, in our case, once the elbow has been identified, the selection of p is not as critical as in problems with merely EDA objectives, where only one p needs to be selected for interpretative purposes and where we want to avoid outliers being selected as archetypes. However, for our purposes, this is not the case; we can consider different p values, since we prefer to better capture the shape of the data set by changing the resulting archetypes and collect the information for different p values. For us, it is not a problem that an archetype is an outlier, since our objective is to detect them and, in fact, this can facilitate the mission.

3. The AA + *k*-NN method for detecting anomalies

Feature extraction is a well-known and powerful method for improving the performance of learning algorithms (Hastie et al., 2009). In the same way, a sensible tactic in the outlier detection field is to identify relevant subspaces where outlier analysis can be honed, i.e. where outliers deviate clearly from the normal observations after projection on the relevant subspaces, and then combine the results from different subspaces in order to create a more robust ensemble (Aggarwal, 2017).

This is the idea of our proposal. An overall picture of our method is to compute AA for a certain p and project the data. We then apply the k-NN method to the

α values. Note that AA actually seeks extreme profiles, so we can take advantage of this fact. If we repeat the procedure for different p values, we will have different explanations of the data, and we can use independent ensembles (the combined procedures are independent) to combine the results.

The outline of the procedure is as follows:

Step 1 Min-max normalize or standardize the data.

Step 2 Compute AA from $p = 1$ to $p = P$, and determine the value e where the elbow is found.

Step 3 Apply k-NN (sum of distance to k nearest neighbors) for a certain k for the α matrices from e to P. Then, the P - e + 1 outlier scores obtained in each subspace are merged by a cumulative-sum approach, which is equivalent to averaging the scores.

Note that the bias-variance trade-off in anomaly detection is almost identical to that in classification (Aggarwal, 2017), so it follows that averaging also reduces variance in anomaly detection.

This procedure returns outlier scores; as usual, the highest score denotes the highest degree of outlierness. A way to establish a binary decision about whether or not to label a point as an outlier, is to use a box-plot with the outlier scores and to consider the points detected as outliers by the box-plot as anomalies. Obviously, this hardening method will work well if the outlier scores corresponding to true outliers are well separated from those of the normal cases.

Let us give the details of each step. In Step 1, we consider both alternatives in the experiments: min-max normalization and standardization. In Step 2, we consider two values in the experiments, $P = 10$ and $P = 15$. In Step 3, we begin with e, since it is expected to be the first value for which archetypes explain the data well. The following values from e to P should also describe the data well, but may give different descriptions. This can be desirable, since diversity and accuracy are two key factors in the success of ensembles. The aggregation ensemble of Step 3 is valid since the scale of the outlier scores is the same, as we use the same k each time. Also note that the α values always add 1, for any p.

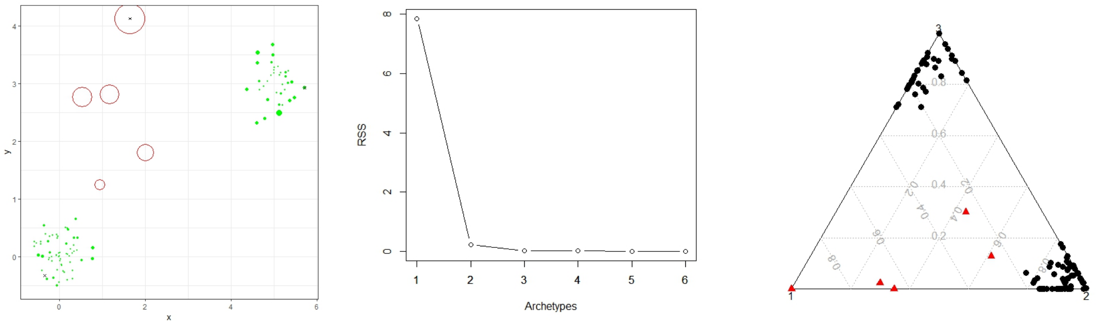
In the experiments, instead of considering a single k, we evaluate the procedure for different k values, from $k = 10$ to 50, as in Goldstein and Uchida (2016). Then the summary, mean and standard deviation for all these AUCs are calculated. If we wanted to report the results in terms of binary labels, i.e. to convert the scores into binary labels, we could use the box plot-based hardening strategy explained above for each k, then the final decision can be given by aggregation and majority voting (Nguyen et al., 2010). In other words, we have the binary labels for each k, from $k = 10$ to 50, and finally, we consider the points that are labeled as anomalies at least 50% of the times as outliers.

The proposed procedure is illustrated with two toy data sets. The first example is shown in Fig. 1 a). The plot consists of two Gaussian clusters, whose data points are represented by solid green circles, plus five uniformly sampled outliers that are represented by red unfilled circles. If we directly apply the k-NN method to these data with k =5 and a box-plot to the outlier scores, the five outliers are detected, but another five points are also falsely labeled as outliers.

We compute the archetypes from 1 to 6 (P) and the screeplot is displayed in Fig. 1 b). The elbow is found at $p = 2$ (e), so we consider the alpha values of AA from $p = 2$ to 6. As an example, we display the α values for $p = 3$ in a ternary plot in Fig. 1 c), where the outliers are represented by red triangles, while the rest of the points are represented by black circles. The archetypes for $p = 3$ are represented as black crosses in Fig. 1 a). We apply the k-NN method to the α values with $k = 5$, from $p = 2$ and $p = 6$, and the outlier scores are the sum of these results. The outlier scores are visualized by the bubble-size of each case in Fig. 1 a). Then we apply a box-plot to the outlier scores: the five outliers are detected and only one point is falsely labeled as an outlier. In summary, our procedure gives only one error, unlike the five errors given by k-NN in this example. AA + k-NN therefore manages to distinguish between anomalies and normal instances better than simply using k-NN.

Note that AA is a very intuitive tool and its results are easily interpretable, much more so than a PCA transformation. For example, in Fig. 1 a) the data are expressed as a mixture (the α values represented in Fig. 1 c)) of the archetypes (the black crosses). For example, the outlier located at (2, 1.8) is expressed as 26% of archetype 1, plus 44% of archetype 2, plus 30% of archetype 3. Interpretability is a valuable factor for the analyst (Aggarwal, 2017). Knowing the reasons why a particular data point is labeled as an outlier, i.e. discovering the intensional knowledge about the outliers (Knorr and Ng, 1999), can be a great help in real applications, as will be shown in Section 5.

In the second example, we show a simple two-dimensional data set, where the features have a linear dependency, except one point that does not follow the linear relationship of the other points. We apply eight different procedures, and the outlier scores are shown in Fig. 2 as above, by the bubble-size of each case, with filled green circles denoting normal cases and an unfilled red circle denoting the outlier. We consider $k = 3$ for the NN algorithms. For our procedure, the elbow is at $p = 2$, and we consider $p = 2$ and 3 for the α computation, since the convex hull of this data set is formed by three vertices. We consider two new procedures here, which were not considered by Goldstein and Uchida (2016). In RPCA + k -NN, we follow a similar procedure as in our proposal, but instead of using AA, robust PCA is considered and k-NN is used with the PC scores. This alternative method is considered to show that AA is more useful than (robust)

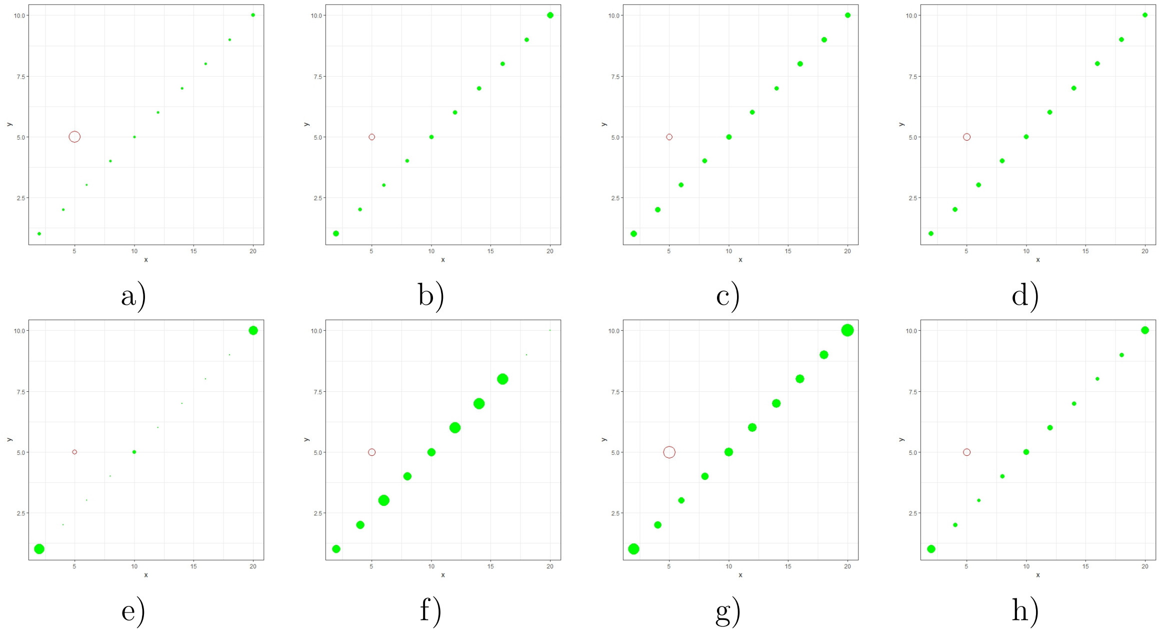


a) b) c)

Figure 1: Example 1: a) plot of the data set (see the text for details); b) Screeplot; c) Ternary plot.

PCA for this situation. In RDOS, the recent procedure proposed by Tang and He (2017) is applied. This procedure uses a density-based outlier detection approach with local kernel density estimation, and instead of using only k nearest neighbors, they also consider reverse nearest neighbors and shared nearest neighbors of a case for density distribution estimation. We use the implementation of the R package **DDoutlier** (Madsen, 2018).

We also compute the AUC values, which are shown in Table 1. For AA + k-NN, the outlier score of the anomalous point is more than double the next highest score. For k-NN the outlier score of the anomalous point is the fourth lowest (remember that the data set is composed of 11 points). For LOF, the outlier score of the anomalous point is the lowest (0.86, below 1, when 1 is supposed to be the score for normal cases). For COF, the outlier scores are the same for all the points. For LoOP, the probability that the anomalous point is an outlier is only 0.17. In fact, the probabilities are higher for three other points. For HBOS, the outlier score of the anomalous point is the third lowest. For RPCA + k-NN, the outlier score of the anomalous point is the fourth lowest. Finally, the outlier score of the anomalous point is the fifth highest for RDOS. In summary, our procedure provides the highest AUC of all the algorithms in this example. Note that although COF was designed to detect this kind of anomaly, it is not able to identify it, because in this case the anomaly is too close to the other points. With the same distribution of points, if the anomaly was $(5, 5 + \epsilon)$, with $\epsilon > 0$, instead of (5,5), COF would be able to detect it. However, our procedure can detect any anomalous point outside the pattern of the linear relationship, since that anomalous point would not belong to the convex hull of the rest of the data, i.e it is a vertex of the convex hull of the whole data set and, therefore, it is used as an archetype in AA for $p = 3$.



a) b) c) d)

e) f) g) h)

Figure 2: Example 2, plot of the anomaly scores for different algorithms (see the text for details): a) AA + k-NN; b) k-NN; c) LOF; d) COF; e) LoOP; f) HBOS; g) RPCA + k-NN; h) RDOS.

AA + k-NN	k-NN	LOF	COF	LoOP	HBOS	RPCA + k-NN	RDOS
1	0.3	0	0.5	0.7	0.2	0.3	0.6

Table 1: The AUC results for example 2.

4. Results and discussion

Table 2 shows the results for all the algorithms and benchmark data sets. Our proposal has been run using min-max normalization as in Goldstein and Uchida (2016), for fair comparison, and $e = 4$ and $P = 10$ in all cases. The best result for each data set is highlighted in bold font. Note that our proposal is the best for the breast cancer and pen local data sets. For the other data sets, the results of AA + k-NN are also very competitive.

Nevertheless, AA was frequently used with standardized data so Table 3 reports the results of our proposal when the two options for Step 1 (normalization or standardization) and Step 2 ($P = 10$ or $P = 15$) are used. Note that the good results for the breast cancer and pen local data sets are improved if the data are standardized rather than normalized.

Method	*breast cancer*	*pen global*	*pen local*	letter	speech	satellite	shuttle	aloi
AA + k-NN	**0.9851** ±0.0030	0.9634 ±0.0030	**0.9915** ±0.0013	0.8605 ±0.0133	0.4351 ±0.0053	0.9073 ±0.0178	0.7317 ±0.02341	0.6003 ±0.0122
k-NN	0.9791 ±0.001	**0.9872** ±0.0055	0.9837 ±0.0018	0.8719 ±0.0176	0.4966 ±0.0101	**0.9701** ±0.0007	0.9424 ±0.0069	0.6502 ±0.0191
kth -NN	0.9807 ±0.0008	0.9778 ±0.0142	0.9757 ±0.0069	0.8268 ±0.0228	0.4784 ±0.0007	0.9681 ±0.0015	0.9434 ±0.0101	0.6177 ±0.0189
LOF	0.9816 ±0.0024	0.8495 ±0.0679	0.9877 ±0.0016	0.8673 ±0.0271	0.5038 ±0.0215	0.8147 ±0.1126	0.5127 ±0.0129	0.7563 ±0.0135
LOF-UB	0.9805 ±0.0020	0.8541 ±0.0777	0.9876 ±0.0013	0.9019 ±0.0030	0.5233 ±0.0134	0.8425 ±0.0839	0.5182 ±0.0124	0.7713 ±0.0045
COF	0.9518 ±0.0054	0.8695 ±0.1261	0.9513 ±0.0134	0.8336 ±0.0228	0.5218 ±0.0287	0.7491 ±0.0952	0.5257 ±0.0086	0.7857 ±0.0118
INFLO	0.9642 ±0.0171	0.7887 ±0.0540	0.9817 ±0.0024	0.8632 ±0.0250	0.5017 ±0.0191	0.8272 ±0.0761	0.4930 ±0.0175	0.7684 ±0.0142
LoOP	0.9725 ±0.0123	0.7684 ±0.0994	0.9851 ±0.0068	**0.9068** ±0.0078	**0.5347** ±0.0343	0.7681 ±0.0433	0.5049 ±0.0035	**0.7899** ±0.0093
LOCI	0.9787	0.8877	-	0.7880	0.4979	-	-	-
aLOCI	0.8105 ±0.0883	0.6889 ±0.0345	0.8011 ±0.0615	0.6208 ±0.0220	0.4992 ±0.0348	0.8324 ±0.0372	0.9474 ±0.0379	0.5855 ±0.0143
CBLOF	0.2983 ±0.1492	0.3190 ±0.1155	0.6995 ±0.1407	0.6792 ±0.0386	0.5021 ±0.0680	0.5539 ±0.0692	0.9037 ±0.1263	0.5393 ±0.0154
uCBLOF	0.9496 ±0.0390	0.8721 ±0.0511	0.9555 ±0.0109	0.8192 ±0.0231	0.4692 ±0.0029	0.9627 ±0.0038	0.9716 ±0.0324	0.5575 ±0.0061
LDCOF	0.7645 ±0.1653	0.5948 ±0.0825	0.9593 ±0.0145	0.8107 ±0.0244	0.4366 ±0.0099	0.9522 ±0.0325	0.8076 ±0.1814	0.5726 ±0.0146
CMGOS-Red	0.9140 ±0.0815	0.5693 ±0.1000	0.9727 ±0.0141	0.7711 ±0.0614	0.5077 ±0.0158	0.9054 ±0.0233	0.5425 ±0.2446	0.5852 ±0.0161
CMGOS-Reg	0.8992 ±0.0643	0.6994 ±0.0681	0.9449 ±0.0510	0.8902 ±0.0200	0.5081 ±0.0161	0.9056 ±0.0233	0.5679 ±0.2402	0.5855 ±0.0161
CMGOS-MCD	0.9196 ±0.0830	0.6265 ±0.0969	0.9038 ±0.0511	0.7848 ±0.0485	-	0.9120 ±0.0520	0.6903 ±0.1670	0.5547 ±0.0160
HBOS	0.9827 ±0.0016	0.7477 ±0.0206	0.6798 ±0.0249	0.6216 ±0.0073	0.4708 ±0.0030	0.9135 ±0.0047	**0.9925** ±0.0039	0.4757 ±0.0010
rPCA	0.9664 ±0.0000	0.9375 ±0.0001	0.7841 ±0.0151	0.8095 ±0.0029	0.5024 ±0.0000	0.9461 ±0.0023	0.9963 ±0.0000	0.5621 ±0.0000
oc-SVM	0.9721 ±0.0102	0.9512 ±0.0436	0.9543 ±0.0130	0.5195 ±0.0382	0.4650 ±0.0021	0.9549 ±0.0021	0.9862 ±0.0002	0.5319 ±0.0021
ν-oc-SVM	0.9581 ±0.0311	0.8993 ±0.0387	0.9236 ±0.0140	0.7298 ±0.1365	0.4649 ±0.0026	0.9430 ±0.0058	0.9848 ±0.0019	0.5221 ±0.0025
Best NN	0.9816 ±0.0024	0.9872 ±0.0055	0.9877 ±0.0016	0.9068 ±0.0078	0.5347 ±0.0343	0.9701 ±0.0007	0.9474 ±0.0379	0.7899 ±0.0093
Best Cluster	0.9496 ±0.0390	0.8721 ±0.0511	0.9727 ±0.0141	0.8902 ±0.0200	0.5081 ±0.0161	0.9627 ±0.0038	0.9716 ±0.0324	0.5855 ±0.0161

Table 2: Average AUC together with the standard deviation for each algorithm and benchmark data set.

Method	*breast cancer*	*pen global*	*pen local*	letter	speech	satellite	shuttle	aloi
AA + k-NN min-max, P=10	0.9851 ±0.0030	0.9634 ±0.0030	0.9915 ±0.0013	**0.8605** ±0.0133	**0.4351** ±0.0053	**0.9073** 0.0178	0.7317 ±0.0234	0.6003 ±0.0122
AA + k-NN standardization, P=10	**0.9862** ±0.0023	0.9812 ±0.0025	0.9944 ±0.0008	0.8315 ±0.0132	0.3591 ±0.0065	0.9060 ±0.0131	**0.7360** ±0.0068	0.5747 ±0.0062
AA + k-NN min-max, P=15	0.9807 ±0.0027	0.9798 ±0.0041	0.9943 ±0.0007	0.8524 ±0.0137	0.4139 ±0.0035	0.8806 ±0.0258	0.6584 ±0.0106	**0.6213** ±0.0153
AA + k-NN standardization, P=15	0.9782 ±0.0024	**0.9818** ±0.0028	**0.9962** ±0.0005	0.8405 ±0.0166	0.3458 ±0.0041	0.8849 ±0.0189	0.7216 ±0.0274	0.5980 ±0.0106

Table 3: Mean AUC and the standard deviation for different option of AA + k-NN.

We also analyze the stability of the results of our proposal if e is changed and also if a different range of k-values are used. The breast-cancer data set is used as an illustration and the results are shown in Table 4. We see that the best results are achieved with $e = 2$ and with small k values. With this option, we can convert the outlier scores into binary levels as explained above. Table 5 shows the confusion matrices with min-max normalization, k from 5 to 15, and different e. Zero indicates a normal case, whereas one indicates an outlier. All the outliers (10) are correctly identified, but some cases are erroneously labeled as outliers. We also show the results of hardening with k-NN, which returns a high number of errors, not only as false positives, but also as false negatives.

e	$Min - max$ k = 5 to 15	$Standardization$ k = 5 to 15	$Min - max$ k = 10 to 50	$Standardization$ k = 10 to 50
e =2	**0.9930** ±0.0016	0.9913 ±0.0007	0.9882 ±0.0018	0.9892 ±0.0012
e =3	0.9928 ±0.0016	0.9913 ±0.0005	0.9873 ±0.0024	0.9884 ±0.0017
$e = 4$	0.9916 ±0.0013	0.9900 ±0.0006	0.9851 ±0.0030	0.9862 ±0.0023
$e = 5$	0.9895 ±0.0017	0.9886 ±0.0003	0.9832 ±0.0030	0.9849 ±0.0026

Table 4: Mean AUC and the standard deviation for different options of AA + k-NN with breast cancer.

Our proposal is composed of two major computational parts. The computational complexity of the algorithm used to compute AA has been analyzed in detail by Eugster and Leisch (2009). It is a computer intensive method. More efficient alternative algorithms for computing AA have been proposed, especially for large data sets, such as the implementation by Mørup and Hansen (2012), Chen et al. (2014), Bauckhage et al. (2015) and Mair et al. (2017). On the other hand, k-NN

True Labels	e=2		e=3		e=4		e=5		k-NN	
Predictions	0	1	0	1	0	1	0	1	0	1
0	341	0	343	0	344	0	343	0	338	2
1	16	10	14	10	13	10	14	10	19	8

Table 5: Confusion matrices with binary labels for different options of AA + k-NN and k-NN with breast cancer (min-max normalization and k from 5 to 15).

may require O(n^2) time to compute all k-nearest neighbor distances (Aggarwal, 2017). Therefore, AA + k-NN is not a computationally efficient method, but this may be compensated for by its ease of interpretability and intuitive analysis, and its mathematical precision (effectiveness).

5. Application

Knowledge of foot shape is of great importance for the appropriate design of footwear. It is a crucial issue for manufacturing shoes, since a proper fit is a key factor in the decision to buy, besides the fact that poorly fitting footwear can cause foot pain and deformity, especially in women. For these reasons, there are a large number of studies on foot shapes, such as Krauss et al. (2008), Delgado-Abellán et al. (2014), Hong et al. (2011), Krauss et al. (2011), Tomassoni et al. (2014), Saghazadeh et al. (2015), etc. In many of these studies, and in anthropometric studies devoted to product design in general, or apparel design in particular, data are studied without carrying out an outlier analysis, as in Jung et al. (2010) or Alemany et al. (2019). However, this is crucial, not only for data cleaning (Ibáñez et al., 2012; Pierola et al., 2016; Markiewicz et al., 2017), which is a classical application of outlier analysis (Aggarwal, 2017), but also to take advantage of the information that outliers can provide with regard to the design of shoes that fit well for a high percentage of the population. For example, in the apparel industry, many brands offer special sizes. However, outlier detection in the field of anthropometry is usually carried out by means of very simple procedures, as is the case in Kouchi (2014) or Kuehnapfel et al. (2016), where they look at extreme values in individual variables, or two-dimensional plots are inspected, which are the recommendations given in ISO 15535:2012 (2012) for cleaning anthropometric databases. Obviously, a somewhat more sophisticated method can be more effective, and better advantage can be taken of the information.

Therefore, the purpose of this Section is to detect the outliers in an anthropometric foot database, before form analysis is carried out. Here, we restrict and focus on the outlier analysis part only. We carry out a separate analysis for men and women, since gender foot shape differences are well-known (Krauss et al., 2011; Saghazadeh et al., 2015). Furthermore, footwear designers usually propose different designs for women and men. We apply our proposal, which also helps us to understand why those points are labeled as outliers.

5.1. Foot database

Foot measurements have been extracted from an anthropometric database of 775 3D right foot scans representing the Spanish adult female and male population, 382 corresponding to women and 393 to men. The data were collected in different regions across Spain at shoe shops and workplaces using an INFOOT laser scanner (I-Ware Laboratory, 2018). The scanning process is carried out while the participant stands upright placing equal weight on each foot, in a specific position and orientation (see Fig. 3). The result is a 3D point cloud representing the complete outer surface of the foot, including the sole of the foot.

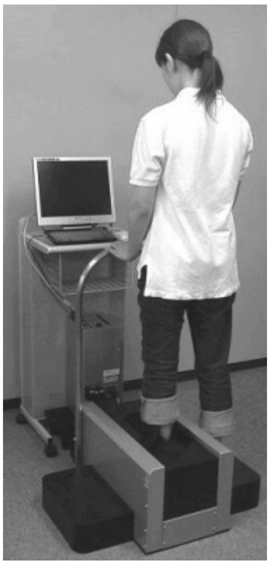


Figure 3: Infoot$^{®}$ scanner.

3D foot shapes were registered using the method described by Allen et al. (2003) with a template made up of 5,000 vertices, with five foot landmarks (i.e. 1st and 2nd toe tips; 1st and 5th metatarsal heads; and pternion; see Fig. 4). This set of landmarks is automatically located on 3D foot scans and allows the extraction of 22 key foot measurements.

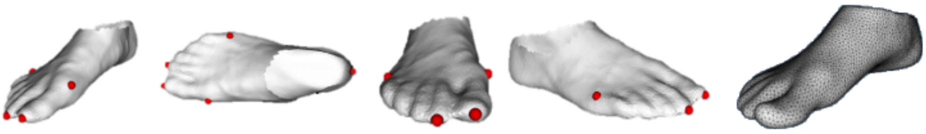


Figure 4: Foot landmarks used in the registration of the database and foot template topology (the last image).

The 22 foot measurements (see Fig. 5) are used in product design and in clinical assessment. All 3D registered feet were digitally measured with the algorithms developed by the IBV (Biomechanics Institute of Valencia). In contrast to body measurements, foot measurements are not standardized. Only Foot Length, Ball Girth and Ball Width are considered in ISO 8559-1:2017 (2017), ASTM D5219-15 (2015) and ISO 7250-1:2008 (2008). The definitions are those used by the Human Shape Lab of the IBV, which comply with standards and are compatible with the accepted definitions found in the literature (Rossi and Tennant, 2013; Ramiro et al., 1995; AIST, Digital Human Research Group, 2018; Goonetilleke, 2012; Luximon, 2013).

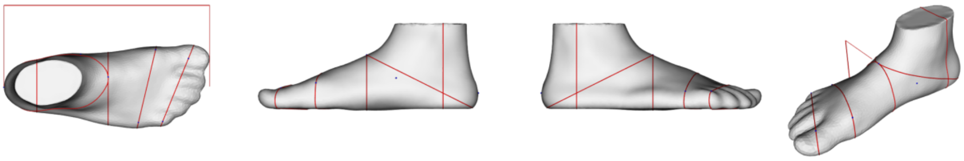


Figure 5: Examples of digital measurements elicited from a 3D registered foot.

5.2. *Outlier analysis*

Instead of the whole set of 22 variables, only the 4 variables that could most influence shoe fitting according to shoe design experts are analyzed. Specifically, these variables are: Foot Length, FL (distance between the rear and foremost point the foot axis); Ball Girth, BG (perimeter of the ball section), Ball Width, BW (maximal distance between the extreme points of the ball section projected onto the ground plane); and Instep Height, IH (maximal height of the instep section, located at 50% of the foot length).

In this application, size and shape is important. According to shoe experts, FL is the variable that best describes the size of the foot; in fact, this variable has great importance for shoe size. Therefore, we consider the size, represented by FL, and the shape, as explained by Dryden and Mardia (2016), separately. Shape corresponds to the geometrical information that remains once the scale is removed. Therefore, to describe the shape, we consider the rest of the variables after removing the scale by dividing each of the variables by FL: BG/FL, BW/FL and IH/FL.

For FL we can use simple box-plots to determine the outlier in size. Fig. 6 shows the different size ranges for women and men, with a different number of outliers. For women, five outliers are detected: one due to a very small size and four due to very large sizes. However, for men, the number of extreme sizes is smaller: there is one outlier corresponding to a very small size and another corresponding to a very large size. The variation in men, in terms of both range and interquartile range, is greater than in women. This could explain the fact that more outliers are found in women.

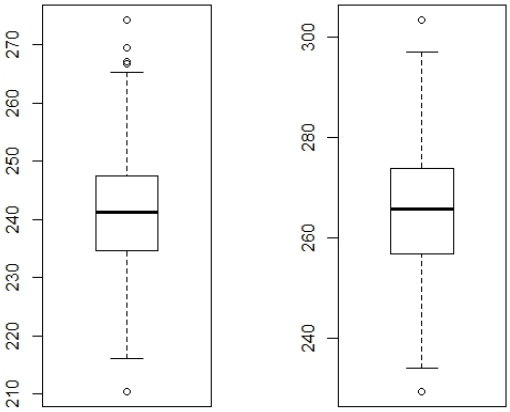


Figure 6: Box-plots for women (left-handed) and men (right-handed), respectively.

We apply AA + k-NN to the previously standardized foot shape variables, with $e = 3$, since the elbow appears at this value for both men and women, and with k from 10 to 50. We convert the outlier scores into binary labels for the shake of brevity in the illustration.

In literature on AA, archetypes are usually displayed by the percentile values of each variable as compared to the data. We consider the same strategy here to interpret the outliers found. Tables 6 and 7 show the percentile profiles of the outliers found in foot shape variables for women and men, respectively. This information is useful not only for cleaning, but also for shoe designers to know which shapes are “not normal”. For that reason, we also include the percentile of

FL for the outliers, although this variable is not used in the outlier detection of shapes.

BG / FL	BW /FL	IH /FL	FL
87	40	94	48
90	80	100	4
96	82	97	28
90	63	89	8
2	3	86	10
99	100	83	35
97	88	94	36
93	62	91	51
99	99	71	55
0	0	44	62
96	95	100	17
99	99	99	2
73	39	96	87
100	100	98	2

Table 6: Percentile profiles of outliers of foot shape variables for women.

For women, a total of 14 outliers are found, more than in the group of men, where 8 outliers are detected. One type of outlier detected in both men and women corresponds to points with very high percentiles in all three shape variables. We refer to these as type 1 outliers. Another type of outlier, type 2, is the points with a high percentile in BG/FL and IH/FL, but a medium percentile in BW/FL. This kind of outlier is mainly found in women. For men, two outliers could also be included in this type, but their BG/FL percentiles are not as high as in the case of women. In women, we find another type of outlier, type 3, with very low percentiles for BG/FL and BW/FL. Only one man is an outlier of this type. In men, we find another two types of outliers that do not appear in women: type 4 are outliers with a very low percentile in BW/FL, but a very high percentile in IH/FL, whereas type 5 are outliers with high percentiles in BG/FL and BW/FL and very low percentiles in IH/FL. Note that the majority of outliers have one or more variables with high percentiles, more so than with low percentiles, so they are due to excess, especially for women.

In summary, for women the outliers are grouped into three sets: one from type 1 (2nd, 3rd, 6th, 7th, 9th, 11th, 12th and 14th), one from type 2 (1st, 4th, 8th and 13th) and one from type 3 (5th and 10th), while for men the outliers are from type 1 (6th, 7th, 8th), a variation of type 2 (5th and 10th), type 3 (1st), type 4

BG /FL	BW/FL	IH/FL	FL
0	0	26	31
31	3	96	29
64	75	3	42
37	5	95	69
70	60	99	39
100	99	100	6
100	100	97	8
84	82	100	53
82	86	1	43
56	31	99	3

Table 7: Percentile profiles of outliers of foot shape variables for men.

(2nd and 4th) and type 5 (3rd and 9th). Type 1 outliers have small-size feet, i.e. their FL percentiles are small, although in the case of women, some of them are not excessively small, and for one man it is medium. Type 2, 3 and 4 outliers are found in feet of all sizes. The only two type 5 outliers correspond to medium-size feet.

6. Conclusions

We have proposed a method to detect outliers based on projection into relevant subspaces by means of AA, applying a k-NN algorithm to these subspaces and combining the results. This method returns outlier scores and we have also proposed a procedure to binarize the scores. We have illustrated their advantages in two simple examples. We have compared our proposal with a large battery of anomaly detection algorithms, 20 in total, in eight benchmark data sets. In terms of averaged AUC, as expected, no algorithm is the best for all data sets, but AA + k returns competitive results. Our proposal obtains the best results in two data sets, k-NN is the best in another two data sets, LoOP in three of the data sets, while HBOS is the best in one of them. The best results of AA + k-NN have been yielded by data sets with numerical continuous variables (breast cancer, pen global and pen local), while the other data sets contain discrete numerical variables. Remember that AA is suited for multivariate continuous numerical data sets. In other words, it worked well with data sets with global and local anomalies, with continuous numerical variables, such as pen global and pen local. We have also seen that changing the normalization procedure for standardization and also the k values used in the second part of our method can improve the results. Nevertheless, we have obtained good results with all the different combinations of parameters. Therefore, it does not seem very sensitive to parameter choice.

As discussed in Section 4, its weak point is its computational inefficiency, but new AA implementations could improve its speed. On the other hand, it has the advantage of its effectiveness (accuracy) and interpretability, which has been shown in the illustrative examples of Section 3 and the application of Section 5. Although our AA implementation is not deterministic, its solutions are stable (Eugster and Leisch, 2009).

We applied AA + k-NN to a novel industrial data set and outliers were detected and interpreted. There are more outliers in women's feet than in men's. For example, in the case of women, there are more outliers due to a very long FL than to a short FL; and as regards their shapes, many of the outliers are due to large dimensions in BG, BW and IH relative to their small FLs. This information can be taken into account in the design process, or also to propose a range of shoes of special lengths or shapes for women. However, for men there are fewer shape outliers but with more different typologies. In summary, detecting the outliers in this kind of data sets can help shoe designers adjust their designs to a larger part of the population and be aware of the characteristics of the users that will make them uncomfortable to wear, whether when considering a range of special sizes or modifying any shoe feature to fit more customers.

We have used AA, but in future work a variant of AA such as archetypoid analysis (Vinué et al., 2015) could be tested. The hardening process could also be improved by changing simple box-plots for other alternatives, such as those proposed by Kriegel et al. (2011). The speed of AA could be improved by using alternative AA algorithms such as those discussed in Section 4. On the other hand, we have considered only complete instances, but in real problems not all the cases are complete. We could easily extend the methodology for data sets with missing data, taking into account the proposal of Epifanio et al. (2018a). Finally, as regards the application, we have focused on the most important variables in shoe design, but a more complete and exhaustive study could be carried out by considering other important variables.

Acknowledgments

The authors would like to thank IMPIVA for having promoted "3D anthropometric and morphological study of the feet of the Spanish population for its application to the design of footwear and components" (IMPRDA/2005/38). This work is supported by the following grants: DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities (AEI/FEDER, EU) and UJI-B2017-13 from Universitat Jaume I.

References

Aggarwal, C. C., 2017. Outlier analysis, 2nd Edition. Springer.

AIST, Digital Human Research Group, 2018. `http://www.dh.aist.go.jp/research/centered/anthropometry/M_foot.html.en`.

Alemany, S., Ballester, A., Parrilla, E., Pierola, A., Uriel, J., Nacher, B., Remon, A., Ruescas, A., Durá, J. V., Piqueras, P., Solves, C., 2019. 3D body modelling and applications. In: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). Springer International Publishing, Cham, pp. 623–636.

Allen, B., Curless, B., Popović, Z., 2003. The space of human body shapes: reconstruction and parameterization from range scans. ACM Transactions on Graphics (TOG) 22 (3), 587–594.

Amer, M., Goldstein, M., 2012. Nearest-neighbor and clustering based anomaly detection algorithms for RapidMiner. In: Proceedings of the 3rd RapidMiner Community Meeting and Conference. pp. 1–12.

Amer, M., Goldstein, M., Abdennadher, S., 2013. Enhancing one-class support vector machines for unsupervised anomaly detection. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description. ODD '13. ACM, New York, NY, USA, pp. 8–15.

Anscombe, F. J., 1960. Rejection of outliers. Technometrics 2 (2), 123–146.

ASTM D5219-15, 2015. Standard terminology relating to body dimensions for apparel sizing. ASTM International, West Conshohocken, PA, 2015.

Bauckhage, C., Kersting, K., Hoppe, F., Thurau, C., 2015. Archetypal analysis as an autoencoder. In: Workshop New Challenges in Neural Computation. pp. 8–15.

Bolton, R. J., Hand, D. J., et al., 2001. Unsupervised profiling methods for fraud detection. Credit Scoring and Credit Control VII, 235–255.

Breunig, M., Kriegel, H.-P., Ng, R. T., Sander, J., 2000. Lof: Identifying density-based local outliers. In: Proceeding of the 2000 ACM Sigmoid internationla conference on management data. pp. 93–104.

Cabero, I., 2018. Aplicació en Didàctica de la Matemàtica de l'Anàlisi d'Arquetipoids. Master's thesis, Universitat Jaume I, Spain.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., Assent, I., Houle, M. E., 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining and Knowledge Discovery 30 (4), 891–927.

Canhasi, E., Kononenko, I., 2014. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. Expert Systems with Applications 41 (2), 535 – 543.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR) 41 (3), 15.

Chen, Y., Mairal, J., Harchaoui, Z., 2014. Fast and Robust Archetypal Analysis for Representation Learning. In: CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition. pp. 1478–1485.

Cutler, A., Breiman, L., 1994. Archetypal Analysis. Technometrics 36 (4), 338–347.

Davis, T., Love, B. C., 2010. Memory for category information is idealized through contrast with competing options. Psychological Science 21 (2), 234–242.

Delgado-Abellán, L., Aguado, X., Jiménez-Ormeño, E., Mecerreyes, L., Alegre, L. M., 2014. Foot morphology in Spanish school children according to sex and age. Ergonomics 57 (5), 787–797.

D'Esposito, M. R., Palumbo, F., Ragozini, G., 2012. Interval Archetypes: A New Tool for Interval Data Analysis. Statistical Analysis and Data Mining 5 (4), 322–335.

Dheeru, D., Karra Taniskidou, E., 2017. UCI machine learning repository.
URL `http://archive.ics.uci.edu/ml`

Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J., 2018. A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition 74, 406–421.

Dryden, I. L., Mardia, K. V., 2016. Statistical Shape Analysis: With Applications in R. John Wiley & Sons, Chichester.

Epifanio, I., 2016. Functional archetype and archetypoid analysis. Computational Statistics & Data Analysis 104, 24 – 34.

Epifanio, I., Ibáñez, M. V., Simó, A., 2018a. Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. The American Statistician.

Epifanio, I., Ibáñez, M. V., Simó, A., 2018b. Archetypal shapes based on landmarks and extension to handle missing data. Advances in Data Analysis and Classification 12 (3), 705–735.

Epifanio, I., Vinué, G., Alemany, S., 2013. Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. Computers & Industrial Engineering 64 (3), 757–765.

Eugster, M. J., Leisch, F., 2009. From Spider-Man to Hero - Archetypal Analysis in R. Journal of Statistical Software 30 (8), 1–23.

Eugster, M. J. A., 2012. Performance profiles based on archetypal athletes. International Journal of Performance Analysis in Sport 12 (1), 166–187.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters 27 (8), 861–874.

Fawcett, T., Provost, F., 1997. Adaptive fraud detection. Data mining and knowledge discovery 1 (3), 291–316.

Galeano, P., Peña, D., Tsay, R. S., 2006. Outlier detection in multivariate time series by projection pursuit. Journal of the American Statistical Association 101 (474), 654–669.

Gnanadesikan, R., Kettenring, J. R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28 (1), 81–124.

Goldstein, M., 2015. Unsupervised anomaly detection benchmark. Harvard Dataverse.
URL https://doi.org/10.7910/DVN/OPQMVF

Goldstein, M., Dengel, A., 2012. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track, 59–63.

Goldstein, M., Uchida, S., 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE 11 (4), e0152173.

Goonetilleke, R. S., 2012. The science of footwear. CRC Press.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data mining, inference and prediction. 2nd ed., Springer-Verlag, New York.

Hawkins, D. M., 1980. Identification of outliers. Vol. 11. Springer.

He, Z., Xu, X., Deng, S., 2003. Discovering cluster-based local outliers. Pattern Recognition Letters 24 (9), 1641 – 1650.

Hinrich, J. L., Bardenfleth, S. E., Roge, R. E., Churchill, N. W., Madsen, K. H., Mørup, M., 2016. Archetypal analysis for modeling multisubject fMRI data. IEEE Journal on Selected Topics in Signal Processing 10 (7), 1160–1171.

Hofmann, M., Klinkenberg, R., 2013. RapidMiner: Data mining use cases and business analytics applications. CRC Press.

Hong, Y., Wang, L., Xu, D. Q., Li, J. X., 2011. Gender differences in foot shape: a study of chinese young adults. Sports Biomechanics 10 (02), 85–97.

I-Ware Laboratory, 2018. `http://www.i-ware.co.jp/`.

Ibáñez, M. V., Vinué, G., Alemany, S., Simó, A., Epifanio, I., Domingo, J., Ayala, G., 2012. Apparel sizing using trimmed PAM and OWA operators. Expert Systems with Applications 39 (12), 10512 – 10520.

ISO 15535:2012, 2012. General requirements for establishing anthropometric databases.

ISO 7250-1:2008, 2008. Basic human body measurements for technological design - part 1.

ISO 8559-1:2017, 2017. Size designation of clothes - Part 1: Anthropometric definitions for body measurement.

Jin, W., Tung, A. K. H., Han, J., Wang, W., 2006. Ranking outliers using symmetric neighborhood relationship. In: Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, pp. 577–593.

Johnson, T., Kwok, I., Ng, R. T., 1998. Fast computation of 2-dimensional depth contours. In: KDD. pp. 224–228.

Jung, K., Kwon, O., You, H., 2010. Evaluation of the multivariate accommodation performance of the grid method. Applied Ergonomics 42 (1), 156 – 161.

Knorr, E. M., Ng, R. T., 1999. Finding intensional knowledge of distance-based outliers. In: Proceedings of the 25th International Conference on Very Large Data Bases. pp. 211–222.

Kouchi, M., 2014. 3 - anthropometric methods for apparel design: body measurement devices and techniques. In: Gupta, D., Zakaria, N. (Eds.), Anthropometry, Apparel Sizing and Design. Woodhead Publishing, pp. 67 – 94.

Krauss, I., Grau, S., Mauch, M., Maiwald, C., Horstmann, T., 2008. Sex-related differences in foot shape. Ergonomics 51 (11), 1693–1709.

Krauss, I., Langbein, C., Horstmann, T., Grau, S., 2011. Sex-related differences in foot shape of adult Caucasians a follow-up study focusing on long and short feet. Ergonomics 54 (3), 294–300.

Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A., 2009. Loop: Local outlier probabilities. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. pp. 1649–1652.

Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A., 2011. Interpreting and unifying outlier scores. In: Proceedings of the SIAM International Conference on Data Mining. pp. 13–24.

Kuehnapfel, A., Ahnert, P., Loeffler, M., Broda, A., Scholz, M., 2016. Reliability of 3D laser-based anthropometry and comparison with classical anthropometry. Scientific reports 6, 26672.

Lawson, C. L., Hanson, R. J., 1974. Solving Least Squares Problems. Prentice Hall, Englewood Cliffs.

Luximon, A., 2013. Handbook of footwear design and manufacture. Elsevier.

Madsen, J. H., 2018. DDoutlier: Distance & Density-Based Outlier Detection. R package version 0.1.0.
URL `https://CRAN.R-project.org/package=DDoutlier`

Mair, S., Boubekki, A., Brefeld, U., 2017. Frame-based data factorizations. In: International Conference on Machine Learning. pp. 2305–2313.

Markiewicz, L., Witkowski, M., Sitnik, R., Mielicka, E., 2017. 3D anthropometric algorithms for the estimation of measurements required for specialized garment design. Expert Systems with Applications 85, 366 – 385.

Millán-Roures, L., Epifanio, I., Martínez, V., 2018. Detection of anomalies in water networks by functional data analysis. Mathematical Problems in Engineering 2018 (Article ID 5129735), 13.

Moliner, J., Epifanio, I., 2019. Robust multivariate and functional archetypal analysis with application to financial time series analysis. Physica A: Statistical Mechanics and its Applications 519, 195 – 208.

Mørup, M., Hansen, L. K., 2012. Archetypal analysis for machine learning and data mining. Neurocomputing 80, 54–63.

Nguyen, H. V., Ang, H. H., Gopalkrishnan, V., 2010. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In: Database Systems for Advanced Applications. Springer Berlin Heidelberg, pp. 368–383.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., Faloutsos, C., 2003. LOCI: Fast outlier detection using the local correlation integral. In: ICDE. pp. 315–326.

Peña, D., Prieto, F. J., 2007. Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. Journal of Computational and Graphical Statistics 16 (1), 228–254.

Penny, K. I., Jolliffe, I. T., 2001. A comparison of multivariate outlier detection methods for clinical laboratory safety data. Journal of the Royal Statistical Society: Series D (The Statistician) 50 (3), 295–307.

Pierola, A., Epifanio, I., Alemany, S., 2016. An ensemble of ordered logistic regression and random forest for child garment size matching. Computers & Industrial Engineering 101, 455 – 465.

Porzio, G. C., Ragozini, G., Vistocco, D., 2008. On the use of archetypes as benchmarks. Applied Stochastic Models in Business and Industry 24, 419–437.

Ragozini, G., Palumbo, F., D'Esposito, M. R., 2017. Archetypal analysis for data-driven prototype identification. Statistical Analysis and Data Mining: The ASA Data Science Journal 10 (1), 6–20.

Ramiro, J., Alcántara, E., Forner, A., Ferrandis, R., García-Belenguer, A., Durá, J., Vera, P., Brizuela, G., Llana, S., 1995. Guía de recomendaciones para el diseño de calzado. Instituto de Biomecánica de Valencia, 135–151.

Rossi, W. A., Tennant, R., 2013. Professional shoe fitting. National Shoe Retailers Association.

Rousseeuw, P. J., Driessen, K. V., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41 (3), 212–223.

Rousseeuw, P. J., Hubert, M., 2018. Anomaly detection by robust statistics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (2), e1236.

Saghazadeh, M., Kitano, N., Okura, T., 2015. Gender differences of foot characteristics in older Japanese adults using a 3D foot scanner. Journal of Foot and Ankle Research 8 (1), 29.

Seth, S., Eugster, M. J. A., 2016. Probabilistic archetypal analysis. Machine Learning 102 (1), 85–113.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., Chang, L., 2003. A novel anomaly detection scheme based on principal component classifier. In: Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop. pp. 171–179.

Tang, B., He, H., 2017. A local density-based approach for outlier detection. Neurocomputing 241, 171–180.

Tang, J., Chen, Z., Fu, A. W.-C., Cheung, D. W., 2002. Enhancing effectiveness of outlier detections for low density patterns. In: Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, pp. 535–548.

Teng, H. S., Chen, K., Lu, S. C., 1990. Adaptive real-time anomaly detection using inductively generated sequential patterns. In: Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy. pp. 278–284.

Thøgersen, J. C., Mørup, M., Damkiær, S., Molin, S., Jelsbak, L., 2013. Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. BMC Bioinformatics 14, 279.

Thurau, C., Kersting, K., Wahabzada, M., Bauckhage, C., 2012. Descriptive matrix factorization for sustainability adopting the principle of opposites. Data Mining and Knowledge Discovery 24 (2), 325–354.

Tomassoni, D., Traini, E., Amenta, F., 2014. Gender and age related differences in foot morphology. Maturitas 79 (4), 421 – 427.

Tsanousa, A., Laskaris, N., Angelis, L., 2015. A novel single-trial methodology for studying brain response variability based on archetypal analysis. Expert Systems with Applications 42 (22), 8454 – 8462.

Vinué, G., 2017. Anthropometry: An R package for analysis of anthropometric data. Journal of Statistical Software 77 (6), 1–39.

Vinué, G., Epifanio, I., 2017. Archetypoid analysis for sports analytics. Data Mining and Knowledge Discovery 31 (6), 1643–1677.

Vinué, G., Epifanio, I., Alemany, S., 2015. Archetypoids: A new approach to define representative archetypal data. Computational Statistics & Data Analysis 87, 102 – 115.

Williams, G., Baxter, R., He, H., Hawkins, S., Gu, L., 2002. A comparative study of RNN for outlier detection in data mining. In: IEEE International Conference on Data Mining. pp. 709–712.

Capítol 4

Conclusions

El propòsit primordial d'aquesta tesi era demostrar la transcendència de l'Anàlisi d'Arquetipus dins de les tècniques estadístiques no supervisades. Ho hem fet utilitzant diferents tipus de dades, algunes d'elles implementades per primera vegada en AA o en alguna de les seues variants (ADA i FADA), com són les dades binàries i també veient la seua rellevància en la resolució de problemes complexos com són la segmentació no supervisada de textures, i també la detecció d'outliers. En totes les aportacions, hem comparat la nostra proposta amb tècniques reconegudes i també hem utilitzat bases de dades reals, testimoniejant la seua qualitat i idoneïtat. Fet que evidencia un pas endavant per a l'aprenentatge estadístic (Statistical Learning), sobretot en les conclusions qualitatives de les dades.

Detalladament, en el primer article, hem proposat trobar patrons arquetípics en dades binàries utilitzant ADA per a una millor comprensió d'un conjunt de dades. Un estudi de simulació i els resultats proporcionats en dues aplicacions han posat en relleu els beneficis de l'ADA per als qüestionaris binaris com a alternativa que es pot utilitzar enlloc (o conjuntament amb altres) de metodologies establertes.

Tot i que gran part de les anàlisis parteixen de la idea del fet que fer una mitjana de molts elements d'un conjunt de dades és positiu, en aquest article adoptem una perspectiva diferent. Hem seleccionat un petit nombre d'observacions "representatives", observacions arquetípiques i la composició de les dades s'explica a través de barreges d'aquestes observacions extremes. Hem demostrat que aquesta interpretació aporta molta informació i és una eina útil per fer que un conjunt de dades siga més "llegible", fins i tot per als no experts.

En el cas dels alumnes nouvinguts a la Universitat, si ens fixem en els resultats obtinguts utilitzant arquetipoids, ens han aparegut tres representants molt complementaris que faciliten la classificació. Si donem una ullada més profunda a les dades, ens adonem que els tres representants es podrien definir com:

-Un alumne amb un grau de coneixements molt baix i amb quasi totes les respostes nul·les.

-Dos alumnes amb un número semblant de respostes positives, però complementàries i amb perfils diferents, un encerta la part dels sistemes no lineals i les funcions lineals (afins) i l'altre encerta la part de derivades, integrals i la interpretació algebraica.

Aquestes tres tipologies d'alumnes formarien tres grups que serien la nostra proposta a l'hora de dividir l'alumnat i incidir-hi en les seues mancances, assignant-los al grup on tinguen el α més gran.

A més a més hem emprat una de les anàlisis més contrastada, l'anàlisi Clúster. Ens torna uns representants que no són tan extrems com els arquetipoids i en els que no hi ha complementarietat en les respostes dels representats triats, fet que fa pensar que la formació dels diferents clústers s'ha basat en la quantitat de respostes encertades i no en la "qualitat" de les pròpies, motiu que diferencia significativament el dos tipus de classificació.

L'anàlisi HOMALS tampoc ens ha permès treure cap conclusió clara, els resultats no mostren una relació clara entre les preguntes.

L'última anàlisi que hem aplicat, l'anàlisi PAA de Seth and Eugster (2016), en la que s'optimitzen de forma contínua dades discretes en l'espai de paràmetres, donen (en el cas k=3) tres arquetipus de característiques similars als arquetipoids però amb els que es mostren un desequilibri en l'àlgebra i l'anàlisi, no hi ha cap arquetipus amb un nivell acceptable en anàlisi. En els arquetipus reals de PAA, sí que trobem una idoneïtat semblant amb els arquetipoids. Tanmateix la dificultat de resoldre un problema de programació discreta simplificant-lo en un problema continu és molt habitual, però tal com reflexiona Fletcher (2000) el fet de jugar a transformar dades discretes (en aquest cas binàries) en contínues, resoldre el problema d'optimització, i després tornar-les a transformar en binàries i cercar una solució la més propera possible no és el més adient i no garanteix una bona solució.

Respecte a l'anàlisi arquetipoid funcional FADA, s'ha mostrat utilitzant-ho com analitzar les qüestions d'un test de resposta binària. Hem seleccionat un petit nombre d'observacions "representatives", observacions arquetípiques i la composició de les dades s'explica a través de les barreges d'aquestes observacions extremes. Hem demostrat que això pot ser molt informatiu i que és una eina útil per fer que el conjunt de dades sigui "fàcilment llegible", fins i tot per als no experts. A més a més hem demostrat que FADA proporciona una informació diferent que no és possible obtindre si s'usa PCA.

En l'estudi de la segmentació de textures, un problema comú que ens trobem és que les finestres locals centrades en cada píxel des de les quals s'extreuen les característiques, poden contenir més d'una textura. Alguns intents per resoldre això consisteixen a considerar les finestres no centrades en el píxel, com a Wang et al. (1993) i Epifanio and Soille (2007). No obstant això, aquests procediments són més intensius computacionalment que el procediment que hem proposat, on les finestres sí que es centren en cada píxel i el que canviem és la fase de clusterització per a una fase d'anàlisi arquetípica seguida d'una fase de clusterització.

Els resultats preliminars mostren la rellevància de l'enfocament proposat. El nostre procediment s'ha comparat amb altres mètodes. Malgrat la seua simplicitat, s'ha demostrat la trascendència dels seus resultats, especialment en una aplicació de teledetecció. Els resultats del nostre procediment són competitius no només amb altres metodologies no supervisades, sinó també en comparació amb una metodologia supervisada.

En el darrer article, hem proposat un algorisme per detectar outliers basat en projectar en subespais mitjançant AA, aplicant un algoritme k-NN en aquests subespais i combinant els resultats. Aquest algorisme retorna els outliers.

Hem comparat la nostra proposta amb una gran bateria d'algoritmes de detecció d'anomalies, 20 en total, en huit bases de dades de referència. En termes de mitjanes d'AUC, com era d'esperar, cap algoritme és el millor per a tots els conjunts de dades, però AA + k-NN retorna resultats molt competitius. La nostra proposta obté els millors resultats en dos conjunts de dades, k-NN és el millor en altres dos conjunts de dades, LoOP en tres dels conjunts de dades, mentre que HBOS en un d'ells. Els millors resultats de AA + k-NN han estat proporcionats per conjunts de dades amb variables contínues numèriques (Breast Cancer, Pen global i Pen local), la resta de conjunts de dades contenen variables numèriques discretes.

També hem vist que canviar el procediment de normalització per estandardització i també els valors k usats en la segona part del nostre algorisme poden millorar els resultats. Però hem obtingut bons resultats amb totes les diferents combinacions de paràmetres. Per tant, no sembla molt sensible a l'elecció de paràmetres.

Com hem comprovat, el seu punt feble és la seua ineficàcia computacional, però les noves implementacions d'AA poden millorar la seua velocitat. Els seus punts forts són la seua efectivitat (exactitud) i la seua interpretabilitat. A més a més hem aplicat AA + k-NN a un nou conjunt de dades industrials.

S'han detectat i interpretat outliers, els quals són més abundants en els peus de les dones que en els homes. La detecció de valors atípics d'aquest tipus de conjunts de dades pot ajudar els dissenyadors de sabates a ajustar els seus dissenys a un nombre més gran de població i ser conscients de les característiques dels usuaris per modificar qualsevol variable del calçat amb l'objectiu de cobrir més clients.

Aleshores, si donem una ullada genèrica a aquests resultats concloem que:

1. Hem aconseguit demostrar la qualitat d'ADA per a dades binàries constatant una profunditat qualitativa, en la capacitat d'agrupar les dades, molt superior a les altres anàlisis.

2. En eixa mateixa línea hem vist com analitzar les preguntes d'un qúestionari binari mitjançant FADA, i hem pogut descubrir patrons que no hagueren estat detectats amb FPCA.

3. Hem obert un nou camí en la segmentació de textures utilitzant AA, amb uns resultats molt competitius i una baixa càrrega operativa.

4. La comparativa dels algorismes no supervisats a la recerca d'outliers mostra la combinació d'AA+k-NN com una de les més competitives, sinó la més quan treballem amb variables contínues numèriques.

5. Tots els exemples reals en els quals hem utilitzat AA addueixen la seua destresa i l'amplitud de camps de coneixement on es pot aplicar.

Capítol 5

Línies d'investigació futures

L'aprenentatge estadístic és una de les branques de la ciència amb més desenvolupament i major projecció en l'actualitat. El fet de treballar amb una tècnica tan avantguardista i obrir-la a nous camps permet que el ventall de possibilitats per a línies d'investigació futura siga molt ampli.

Fruit del primer treball, com a futura investigació en el camp de les aplicacions, es podria aplicar ADA a qualsevol branca on les dades vinguen recollides de forma binària, com marketing, sociologia, psicologia, etc. és a dir, en totes aquelles disciplines on les enquestes juguen un paper fonamental.

En el camp de l'estadística, a escala teòrica es vol desenvolupar un treball de simulació per comparar adequadament totes les alternatives. A més, hi ha treball futur per estudiar en ADA per al cas de tindre dades no disponibles (missing data), o dades de diferents tipus a la vegada: continues i binàries, per exemple.

En el nostre treball el paper de totes les variables és equilibrat, és a dir, comparteixen el mateix pes, però per a certes situacions, algunes variables podrien tindre més pes en RSS, s'hauria de tindre en compte aquesta situació.

Una altra direcció del treball futur seria considerar ADA per a observacions nominals, per exemple, convertint aquestes variables en indicadors Booleans (variables dummy), és a dir, amb codis binaris.

El segon article permet investigar alguns punts addicionals. Podríem plantejar-se en lloc d'utilitzar AA, treballar amb ADA (Vinué et al.; 2015). En la mateixa línia, en lloc del k-medoids, podríem utilitzar els k-medoids retallats (Ibáñez et al.; 2012). D'altra banda, hem tractat les característiques texturals com a característiques multivariants, ja que aquest és el cas més habitual en aquest context i hem preferit fer-ho a causa de la comparació amb altres tècniques conegudes. No hem explotat el fet que les granulometries siguen funcions i es puguen utilitzar tècniques d'anàlisi de dades funcionals (Ramsay and Silverman; 2005), de fet, hauríem pogut utilitzar l'anàlisi arquetípic funcional FADA (Epifanio; 2016).

Per a altres aplicacions, es podrien considerar funcions texturals alternatives perquè la metodologia proposada no depèn de les característiques texturals seleccionades sinò del fet que siguen discriminatòries entre les classes i les barreges entre textures es transfereixen al vector de característiques.

El nostre procediment també pot retornar finestres arquetípiques de cada grup.

En lloc d'utilitzar els valors α en la segmentació, les mateixes finestres arquetípiques es podrien utilitzar en el procés de segmentació actuant com a prototips.

Respecte a la tercera aportació, nosaltres hem proposat un algorisme no supervisat de recerca d'outliers molt vàlid, però només hem considerat bases de dades completes, però en problemes reals no tots els casos estan complets. Es pot ampliar fàcilment la metodologia a conjunts de dades amb dades mancants, tenint en compte la proposta de Epifanio et al. (2019).

Hem utilitzat AA, però com a treball futur es pot provar ADA (Vinué et al.; 2015). El procés de hardening (ficar etiquetes fixes en lloc de puntuacions) també es podria millorar canviant l'opció del diagrama de caixes per altres alternatives, com les proposades per Kriegel et al. (2011). La velocitat d'AA podria millorar-se mitjançant l'ús d'algoritmes alternatius per realitzar AA.

Finalment, pel que fa a l'aplicació real, ens hem centrat en les variables més importants en el disseny de sabates, però es podria fer un estudi exhaustiu més complet tenint en compte altres variables que nosaltres hem arrecerat en el moment de reduir la variabilitat, ja que hem considerat que la majoria de la informació que contenen estarà inclosa en les que per a nosaltres han sigut les variables principals.

Gràcies a la contemporaneïtat del camp de treball i la relativa novetat de l'anàlisi, la quantitat de camins per continuar les nostres investigacions és molt dilatada, la utilització d'AA i les seues variants poden donar pas a noves estratègies que sàpiguen conjugar la qualitat de la classificació de grups d'aquest tipus d'anàlisi i perfeccionar els seus punts febles, i a més puguen ampliar el ventall d'àmbits on aplicar-los.

Bibliografia

Aggarwal, C. C. (2017). *Outlier analysis*, 2 edn, Springer.

Ajuntament de València (2019).
URL: *https://www.valencia.es*

Cabero, I. (2018). Anàlisi d'arquetipoids aplicats a la didàctica de la matemàtica. UJI's Masther Tesis.

Cabero, I. and Epifanio, I. (2019a). Archetypal analysis: An alternative to clustering for unsupervised texture segmentation, *Image Analysis and Stereology* **38**(2): 151–160.
URL: *https://www.ias-iss.org/ojs/IAS/article/view/2052*

Cabero, I. and Epifanio, I. (2019b). Archetypoid analysis applied to the teaching of mathematics, *Llibre de resums de la XVII Conferència Espanyola i VII Trobada Iberoamericana de Biometria CEB-EIB 2019*, p. 110.

Cabero, I. and Epifanio, I. (2019c). Finding archetypal patterns for binary questionnaires, *submitted.*

Cabero, I., Epifanio, I. and Piérola, A. (2019). Detection of anomalies by gender in foot shapes, *Llibre de resums de la XVII Conferència Espanyola i VII Trobada Iberoamericana de Biometria CEB-EIB 2019*, p. 56.

Cabero, I., Epifanio, I., Piérola, A. and Ballester, A. (2019). Archetype analysis based algorithm to find outliers, *submitted.*

Campbell, N. A. (1978). The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics* pp. 251–258.

Canhasi, E. and Kononenko, I. (2013). Multi-document summarization via archetypal analysis of the content-graph joint model, *Knowledge and Information Systems* pp. 1–22.

Canhasi, E. and Kononenko, I. (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization, *Expert Systems with Applications* **41**(2): 535 – 543.

Chan, B., Mitchell, D. and Cram, L. (2003). Archetypal analysis of galaxy spectra, *Monthly Notices of the Royal Astronomical Society* **338**(3): 790–795.

Chang, T. and Kuo, C. (1993). Texture analysis and classification with tree-structured wavelet transform, *IEEE Transactions on Image Processing* **2**: 429–441.

Chen, C., Daponte, J. and Fox, M. (1989). Fractal feature analysis and classification in medical imaging, *IEEE Trans. Medical Imaging* **8**: 133–142.

Chen, J. and Jain, A. (1988). A structural approach to identify defects in textured images, *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, Beijing*, pp. 29–32.

Chetverikov, D. (1988). Detecting defects in texture, *Proc. 9th Int. Conf. on Pattern Recognition, Rome, Italy*, pp. 61– 63.

Clark, M. and Bovik, A. (1987). Texture segmentation using Gabor modulation /demodulation, *Pattern Recognition Letters* **6**: 261–267.

Coggins, J. and Jain, A. (1985). A spatial filtering approach to texture analysis, *Pattern Recognition Letters* **3**: 195–203.

Conners, R., McMillin, C., Lin, K. and Vasquez-Espinosa, R. (1983). Identifying and locating surface defects in wood: Part of an automated lumber processing system, *IEEE Trans. Pattern Anal. Mach. Intell.* **5**: 573–583.

Cutler, A. and Breiman, L. (1994). Archetypal Analysis, *Technometrics* **36**(4): 338–347.

Daugman, J. (1985). Uncertainity relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters, *Journal of the Optical Society of America* **2**(7): 1160–1169.

Davis, T. and Love, B. C. (2010). Memory for category information is idealized through contrast with competing options, *Psychological Science* **21**(2): 234–242.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C. and De Boor, C. (1978). *A practical guide to splines*, Vol. 27, springer-verlag New York.

Dean, N. and Nugent, R. (2013). Clustering student skill set profiles in a unit hypercube using mixtures of multivariate betas, *Advances in Data Analysis and Classification* **7**(3): 339–357.

D'Esposito, M. R., Palumbo, F. and Ragozini, G. (2012). Interval Archetypes: A New Tool for Interval Data Analysis, *Statistical Analysis and Data Mining* **5**(4): 322–335.

Dewaele, P., Van Gool, P. and Oosterlinck, A. (1988). Texture inspection with self-adaptative convolution filters, *Proc. 9th Int. Conf. on Pattern Recognition, Rome, Italy*, pp. 56–60.

Du, L. (1990). Texture segmentation of SAR images using localized spatial filtering, *Proc. Int. Geoscience and Remote Sensing Symp., Washington, DC.*

Epifanio, I. (2016). Functional archetype and archetypoid analysis, *Computational Statistics & Data Analysis* **104**: 24 – 34.

Epifanio, I., Ibáñez, M. V. and Simó, A. (2019). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles, *The American Statistician* .

Epifanio, I., Ibáñez, M. V. and Simó, A. (2018). Archetypal shapes based on landmarks and extension to handle missing data, *Advances in Data Analysis and Classification* **12**: 705–735.

Epifanio, I. and Soille, P. (2007). Morphological texture features for unsupervised and supervised segmentations of natural landscapes, *IEEE Transactions on Geoscience and Remote Sensing* **45**(4): 1074–1083.

Epifanio, I., Vinué, G. and Alemany, S. (2013). Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem, *Computers & Industrial Engineering* **64**(3): 757–765.

Eugster, M. J. A. (2012). Performance profiles based on archetypal athletes, *International Journal of Performance Analysis in Sport* **12**(1): 166–187.

Eugster, M. J. and Leisch, F. (2009). From Spider-Man to Hero - Archetypal Analysis in R, *Journal of Statistical Software* **30**(8): 1–23.

Fletcher, J. and Kasturi, R. (1988). A robust algorithm for text string from mixed text / graphics images, *IEEE Trans. Pattern Anal. Mach. Intell.* **10**: 910–918.

Fletcher, R. (2000). *Practical Methods of Optimization*, second edn, John Wiley & Sons.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on computers* **100**(9): 881–890.

Gabor, D. (1946). Theory of communication, *J. Inst. Elect. Eng.* **93**: 429–457.

Gibson, J. (1950). *The Perception of the Visual World*, Houhton Mifflin, Boston, MA.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics* pp. 81–124.

Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PLOS ONE* **11**(4): e0152173.

Haralick, R. (1979). Statistical and structural approaches to texture, *Proceedings IEEE* **67**(5): 786–804.

Haralick, R., Shanmugam, K. and Dinstein, I. (1973). Textural features for image classification, *IEEE Trans. Syst. Man Cyberm.* **3**: 610–621.

Harms, H., Gunzer, U. and Aus, H. (1986). Combined local color and texture analysis od stained cells, *Comput. Vision Graph. Image Process.* **33**: 364–376.

Hinrich, J. L., Bardenfleth, S. E., Roge, R. E., Churchill, N. W., Madsen, K. H. and Mørup, M. (2016). Archetypal analysis for modeling multisubject fMRI data, *IEEE Journal on Selected Topics in Signal Processing* **10**(7): 1160–1171.

Ibáñez, M. V., Vinué, G., Alemany, S., Simó, A., Epifanio, I., Domingo, J. and Ayala, G. (2012). Apparel sizing using trimmed PAM and OWA operators, *Expert Systems with Applications* **39**(12): 10512 – 10520.

Insana, M., Wagner, R., Garra, B., Brown, D. and Shawker, T. (1986). Analysis of ultrasound image texture via generalized rician statistics, *Opt. Engin.* **25**: 743–748.

Jain, A. and Bhattacharjee, S. (1992). Text segmentation using Gabor filters for automatic document processing, *Mach. Vision and Appl.* **5**: 169–184.

Jain, A. and Farrokhnia, F. (1991). Unsupervised texture segmentation using Gabor filters, *Pattern Recognition* **24**: 1167–1186.

Jain, A., Farrokhnia, F. and Alman, D. (1990). Texture analysis of automotive finishes, *Proc. of SME Machine Vision Applications Conf., Detroit, MI*, pp. 1–16.

Janowczyk, A., Chandran, S., Singh, R., Sasaroli, D., Coukos, G., Feldman, M. D. and Madabhushi, A. (2011). High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts, *IEEE transactions on biomedical engineering* **59**(5): 1240–1252.

Johnson, T., Kwok, I. and Ng, R. T. (1998). Fast computation of 2-dimensional depth contours, *KDD*, pp. 224–228.

Karu, K., Jain, A. and Bolle, R. (1996). Is there any texture in the image?, *Pattern Recognition* **29**(9): 1437–1446.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York.

Kouchi, M. (2014). 3 - anthropometric methods for apparel design: body measurement devices and techniques, *in* D. Gupta and N. Zakaria (eds), *Anthropometry, Apparel Sizing and Design*, Woodhead Publishing Series in Textiles, Woodhead Publishing, pp. 67 – 94.

Kriegel, H.-P., Kröger, P., Schubert, E. and Zimek, A. (2011). Interpreting and unifying outlier scores, *Proceedings of the SIAM International Conference on Data Mining*, pp. 13–24.

Kuehnapfel, A., Ahnert, P., Loeffler, M., Broda, A. and Scholz, M. (2016). Reliability of 3D laser-based anthropometry and comparison with classical anthropometry, *Scientific reports* **6**: 26672.

Landeweerd, G. and Gelsema, E. (1978). The use of nuclear texture parameters in the automatic analysis of leukocytes, *Pattern Recognition* **10**: 57–61.

Laws, K. (1980). *Textured Image Segmentation*, PhD thesis, University of Southern California.

Lawson, C. L. and Hanson, R. J. (1974). *Solving Least Squares Problems*, Prentice Hall.

Lee, J. and Philpot, W. (1990). A spectral textural classifier for digital imagery, *Proc. Int. Geoscience and Remote Sensing Symp., Washington, DC.*

Li, S., Wang, P., Louviere, J. and Carson, R. (2003). Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals, *ANZMAC 2003 Conference Proceedings*, pp. 1674–1679.

Lundervold, A. (1992). Ultrasonic tissue characterization- a pattern recognition approach, *Technical report*, Norwegian Computing Center.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(7): 674–693.

Midgley, D. and Venaik, S. (2013). Marketing strategy in MNC subsidiaries: pure versus hybrid archetypes, *P. McDougall-Covin and T. Kiyak, Proceedings of the 55th Annual Meeting of the Academy of International Business*, pp. 215–216.

Millán-Roures, L., Epifanio, I. and Martínez, V. (2018). Detection of anomalies in water networks by functional data analysis, *Mathematical Problems in Engineering* **2018**(Article ID 5129735).

Miller, P. and Astley, S. (1992). Classification of breast tissue by texture analysis, *Image and Vision Computing* **10**(5): 277–282.

Moliner, J. and Epifanio, I. (2019). Robust multivariate and functional archetypal analysis with application to financial time series analysis, *Physica A: Statistical Mechanics and its Applications* **519**: 195–208.

Monjoux, E. and Rudant, J. (1991). Texture segmentation in aerial images, *Image Processing Algorithms and Techniques II, San Jose.*

Mørup, M. and Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining, *Neurocomputing* **80**: 54–63.

Portilla, J. and Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelets coefficients, *International Journal of Computer Vision* **40**(1): 49–71.

Porzio, G. C., Ragozini, G. and Vistocco, D. (2008). On the use of archetypes as benchmarks, *Applied Stochastic Models in Business and Industry* **24**: 419–437.

Puzicha, J., Hofmann, T. and Buhman, J. (1997). Non-parametric similarity measure for unsupervised texture segmentation and image retrieval, *Computer Vision and Pattern Recognition*, pp. 262–272.

Ragozini, G., Palumbo, F. and D'Esposito, M. R. (2017). Archetypal analysis for data-driven prototype identification, *Statistical Analysis and Data Mining: The ASA Data Science Journal* **10**(1): 6–20.

Ramsay, J. (1997). A functional approach to modeling test data, *Handbook of modern item response theory*, Springer, pp. 381–394.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation, *Psychometrika* **56**(4): 611–630.

Ramsay, J. O. (2006). Functional data analysis, *Encyclopedia of Statistical Sciences* **4**.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*, Springer.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd edn, Springer.

Rignot, E. and Kwok, R. (1990). Extraction of textural features in SAR images: Statistical model and sensitivity, *Proc. Int. Geoscience and Remote Sensing Symp., Washington, DC.*

Rossi, N., Wang, X. and Ramsay, J. O. (2002). Nonparametric item response function estimates with the em algorithm, *Journal of Educational and Behavioral Statistics* **27**(3): 291–317.

Schistad, A. and Jain, A. (1992). Texture analysis in the presence of speckle noise, *Proc. IEEE Geoscience and Remote Sensing Symp., Houston, TX.*

Seiler, C. and Wohlrabe, K. (2013). Archetypal scientists, *Journal of Informetrics* **7**(2): 345–356.

Seth, S. and Eugster, M. J. (2016). Probabilistic archetypal analysis, *Machine learning* **102**(1): 85–113.

Siew, L., Hodgson, R. and Wood, E. (1988). Texture measures for carpet wear assesment, *IEEE Trans. Pattern Anal. Mach. Intell.* **10**: 92–105.

Soille, P. (2003). Texture analysis, *Morphological Image Analysis: Principles and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 317–346.

Sonka, M., Hlavac, V. and Boyle, R. (1993). *Image Processing, Analysis and Machine Vision*, Chapman and Hall.

Steinschneider, S. and Lall, U. (2015). Daily precipitation and tropical moisture exports across the Eastern United States: An application of archetypal analysis to identify spatiotemporal structure, *Journal of Climate* **28**(21): 8585–8602.

Sund, R. (2019). Computer age statistical inference: Algorithms, evidence, and data science bradley efron and trevor hastie institute of mathematical statistics monographs cambridge university press, 2016,(8th printing 2018), xix+ 475 pages, hardcover isbn: 978-1-107-14989-2, *International Statistical Review* **87**(1): 186–188.

Sutton, R. and Hall, E. (1972). Texture measures for automatic classification of pulmonary disease, *IEEE Trans. Comput.* **21**: 667–676.

Taxt, T., Flyn, P. and Jain, A. (1989). Segmentation of document images, *IEEE Trans. Pattern Anal. Mach. Intell.* **11**: 1322–1329.

Thøgersen, J. C., Mørup, M., Damkiær, S., Molin, S. and Jelsbak, L. (2013). Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways, *BMC Bioinformatics* **14**: 279.

Thurau, C., Kersting, K., Wahabzada, M. and Bauckhage, C. (2012). Descriptive matrix factorization for sustainability adopting the principle of opposites, *Data Mining and Knowledge Discovery* **24**(2): 325–354.

Toulson, D. and Boyce, J. (1992). Segmentation of MR images using neural nets, *Image and Vision Computing* **10**(5): 324–328.

Tsanousa, A., Laskaris, N. and Angelis, L. (2015). A novel single-trial methodology for studying brain response variability based on archetypal analysis, *Expert Systems with Applications* **42**(22): 8454 – 8462.

Tuceryan, M. and Jain, A. (1993). *Texture analysis*, Handbook of Pattern Recognition and Computer Vision, World Scientific Publishing Company, chapter 2.1, pp. 235–276.

Turner, M. (1986). Texture discrimination by Gabor functions, *Biol. Cybern.* **55**: 71–82.

Unser, M. (1995). Texture classification and segmentation using wavelet frames, *IEEE Transactions on Image Processing* **4**: 1549–560.

Unser, M. and Eden, M. (1990). Nonlinear operators for improving texture segmentation based on features extracted by spatial filtering, *IEEE Trans. Syst. Man Cybern.* **20**: 804–815.

Unwin, A. (2010). Exploratory data analysis, *in* P. Peterson, E. Baker and B. McGaw (eds), *International Encyclopedia of Education (Third Edition)*, Elsevier, Oxford, pp. 156 – 161.

Vinué, G. and Epifanio, I. (2017). Archetypoid analysis for sports analytics, *Data Mining and Knowledge Discovery* **31**(6): 1643–1677.

Vinué, G., Epifanio, I. and Alemany, S. (2015). Archetypoids: A new approach to define representative archetypal data, *Computational Statistics & Data Analysis* **87**: 102–115.

Vinué, G., Epifanio, I., Simó, A., Ibáñez, M., Domingo, J. and Ayala, G. (2017). *Anthropometry: An R Package for Analysis of Anthropometric Data.* R package version 1.9.

Wahl, F., Wong, K. and Casey, R. (1982). Block segmentation and text extraction in mixed text / image documents, *Comput. Vision Graph. Image Process* **20**: 375–390.

Wang, D., Haese-Coat, V., Bruno, A. and Ronsin, J. (1993). Texture classification and segmentation based on iterative morphological decomposition, *Journal of Visual Communication and Image Representation* **4**(3): 197–214.

Wang, D. and Srihari, S. (1989). Classification of newspaper image blocks using texture analysis, *Comput. Vision Graph. Image Process* **47**: 327–352.

Weihs, C. and Ickstadt, K. (2018). Data science: the impact of statistics, *International Journal of Data Science and Analytics* **6**(3): 189–194.

Wood, E. (1990). Applying Fourier and associated transforms to pattern characterization in textiles, *Textile Research Journal* **60**(4): 212–220.