Archetypes and Archetypoids: Contributions for estimating boundary cases in Anthropometry and Ergonomics

Guillermo Vinué Visús
PhD student
Faculty of Mathematics
University of Valencia, Spain

Jointly with Irene Epifanio López (University Jaume I of Castellón) and Sandra Alemany Mut (Biomechanics Institute of Valencia)

XIV Conferencia Española de Biometría, 22-24 May 2013, Ciudad Real, Spain

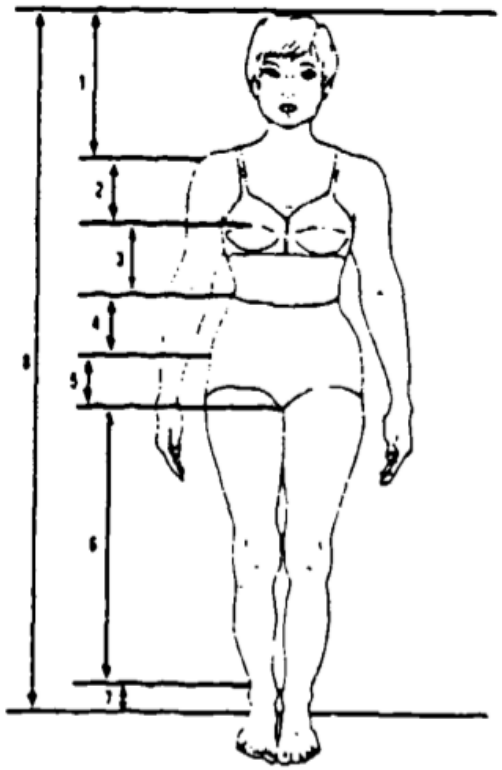
Outline

Accommodation problem in Anthropometry and Ergonomics

- The accommodation is the process of mutual adaptation between persons.
- Products intended to *fit* the users must be designed considering their size and shape $\rightarrow$ Generation of several representative human models.
- The human models represents the anthropometric variability of the target population.
- The appropriate selection of this small group is critical.
- If the *hard to fit* extreme individuals are previously identified, the time and cost of the design process is reduced.

Usual approaches

- **Percentile analysis**: Traditional method used.
  - Drawbacks: Univariate approach. They are not additive.

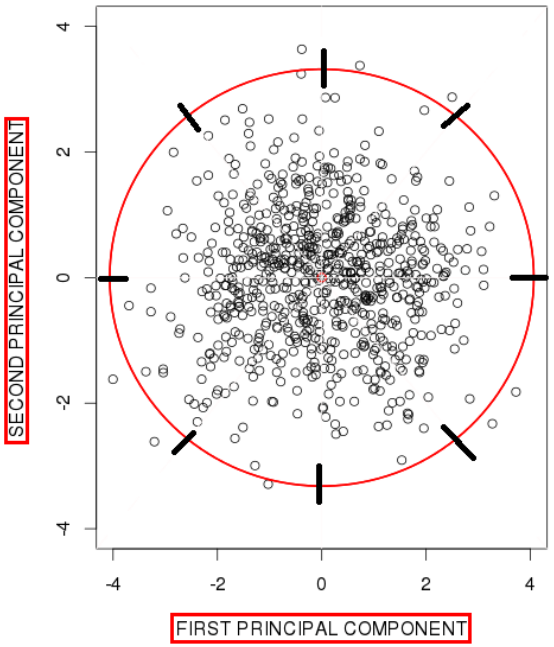


Var. No.	Variable Name	5th%ile	95th%ile
1	Shoulder to Vertex	27.05	32.79
2	Bust to Shoulder	10.79	17.77
3	Waist to Bust	13.42	21.78
4	Buttock to Waist	13.78	21.67
5	Crotch to Buttock	4.78	10.42
6	Ankle to Crotch	57.84	71.09
7	Ankle Height	9.23	13.29
	TOTAL	136.89	188.81
8	Stature	152.50	173.06

Total sample size=3235
All values are in centimeters

(Image taken of Robinette et al. (1981))

- **Regression**: An alternative to approximate a percentile person.
  - The predicted numbers are additive.
  - Drawback: It only provides average values for the predicted measurements.
- **Principal component analysis (PCA)**:



It considers the first PC and selects several extreme points in an ellipse (or circle) which covers a certain percentage of the data.

Drawbacks: Part of the data variation is removed. The number of cases would increase if we wanted to consider more variation.

- Our proposal: ARCHETYPAL ANALYSIS (AA) (Cutler et al. (1994)).

In our everyday language...

Arquetipo de anciano vencido

JAVIER MEMBA

El debut en la gran pantalla de Joan Dalmau fue tardío. Aun así, su actividad cinematográfica se prolongó a lo largo de más de 50 títulos. Ahora bien, tuvieron que pasar varios años para que impusiera su arquetipo de anciano vencido en un pasado, casi siempre concerniente a la Guerra Civil. Tras dar vida a algunos de estos personajes a las órdenes de Montxo Armendáriz, le consagró Miralles, el posible salvador de Rafael Sánchez Mazas (Ramón Fontserè) en *Soldados de Salamina* (2003), la aplaudida adaptación de David Trueba de la novela homónima de Javier Cercas. Aquella creación fue merecedora del

en menor medida también en la pequeña pantalla, cuando Dalmau se estrenó en el cine por primera vez a las órdenes de Gonzalo Herralde en *Últimas tardes con Teresa* (1984). Ya en 1987, Jaime de Arminán le confiaba el personaje del vicealmirante Céspedes en su comedia *Mi general*.

Los hermanos Pastor nos proponen una visión arquetípica del fin del mundo basada en

But in Statistics???

meaningful subsets of the population. An archetype is defined as a centrally located subject in each cluster that is the true representative of the cluster [2]. The use of cluster analysis to identify such archetypes would go a long way to increase our understanding of the data. For

AND:

this goal, the clustering approach should be able to represent each cluster by its archetype [2].

Source: Paquet, E. *Exploring Anthropometric Data Through Cluster Analysis*. Published in Digital Human Modeling for Design and Engineering (DHM). June 15-17, 2004. Oakland University, Rochester, Michigan, USA. NRC 46564.

- An archetype is not a prototype!!!

Statistical definition of archetype

- Let be an $n \times m$ matrix $\mathbf{X}$, multivariate database.
- The AA aims at obtaining the $n \times k$ matrices α and β which minimize:

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

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, \ldots, n \implies \hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$

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

- ARCHETYPE: extreme member of the data set that is a mixture of the actual data points: $\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$
- Archetypes can be computed with the R package *archetypes*.

Archetypal analysis vs PCA

- The goal of AA is to obtain extreme individuals.
- The level of accommodation is reached with AA.
- Archetypes cannot be obtained with PCA.
- The number of archetypes can be decided by the user or by a criterion.

Computers & Industrial Engineering 64 (2013) 757–765

Contents lists available at SciVerse ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Archetypal analysis: Contributions for estimating boundary cases in multivariate accommodation problem

I. Epifanio [a,*], G. Vinué [b], S. Alemany [c]

[a] Dept. de Matemàtiques, Universitat Jaume I, Castelló, Spain
[b] Department of Statistics and O.R., University of Valencia, Valencia, Spain
[c] Biomechanics Institute of Valencia, Universidad Politécnica de Valencia, Valencia, Spain

ARTICLE INFO

Article history:
Received 3 August 2012
Received in revised form 6 December 2012
Accepted 7 December 2012
Available online 29 December 2012

ABSTRACT

The use of archetypal analysis is proposed in order to determine a set of representative cases that entail a certain percentage of the population, in the accommodation problem. A well-known anthropometric database has been used in order to compare our methodology with the common used PCA-approach, showing the advantages of our methodology: the level of accommodation is reached unlike the PCA approach, no more adjustments are necessary, the user can decide the number of archetypes to consider

Archetypoid analysis

- The archetypes do not correspond to observed individuals:

$$\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l \quad \text{with} \quad \sum_{l=1}^{n} \beta_{jl} = 1 \quad \text{and} \quad \boxed{\beta_{jl} \geq 0}$$

- In some cases it is critical that the archetypes are real subjects.
- So far the nearest individuals to archetypes are computed in two ways:
  1. *nearest*: Subjects who have the closest d_E to archetypes.
  2. *which*: Subjects with the greatest α for each archetype.
- The identified archetypes can be artificial: "*no economist in our sample fits this archetype to 100%*" (Seiler et al. (2013)).
- A new archetypal concept is proposed: the ARCHETYPOID:

$$\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l \quad \text{with} \quad \sum_{l=1}^{n} \beta_{jl} = 1 \quad \text{and} \quad \boxed{\beta_{jl} \in \{0, 1\}}$$

- An archetypoid is a real observed individual.
- The archetypoids might not be the same as the *nearest*/*which*.
- The archetypoids always exist even when the features are unavailable.

- An algorithm inspired by PAM is developed to compute archetypoids:
  1. An initial vector of archetypoids is obtained (*nearest*/*which*).
  2. The initial vector of archetypoids is attempted to be improved.
- Cases of study:

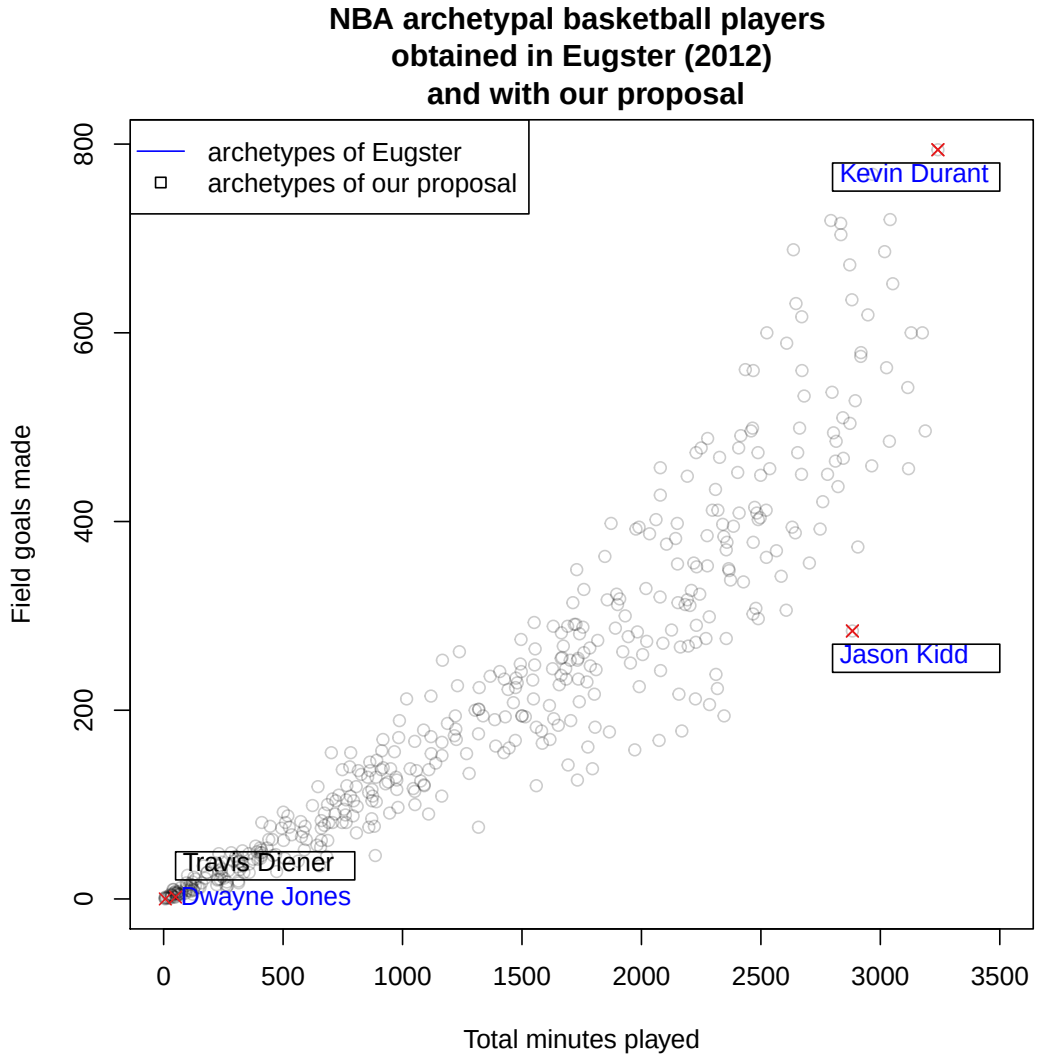
Sportive example

Clothing design problem

1. NBA database: Total minutes played and field goals made of 441 players, season 2009/2010.
2. Spanish 3D anthropometric survey: 10415 women (12-70 years old).

	Archetypal players of Eugster (2012)		
	Name	Total minutes	Field goals
124	Kevin Durant	3241	794
243	Jason Kidd	2883	284
236	Dwayne Jones	7	0

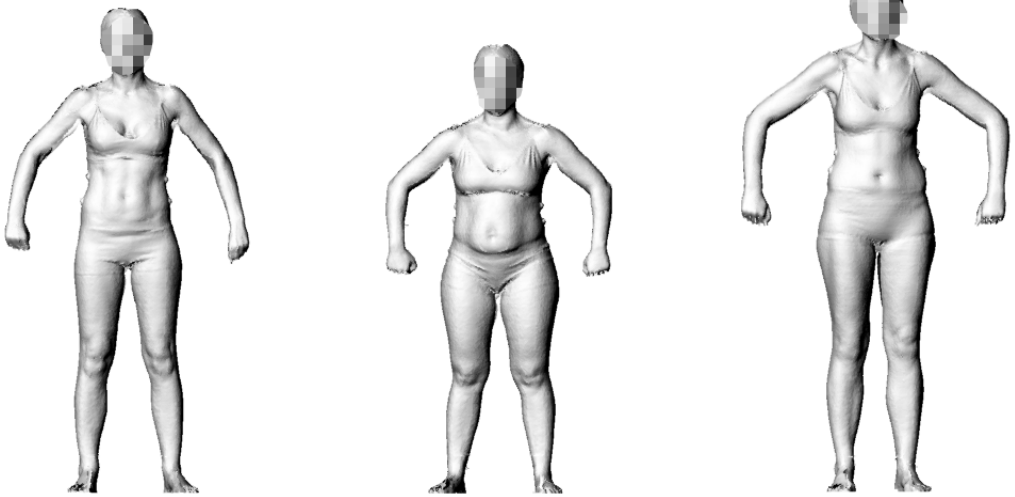
	Archetypoid players of our proposal		
	Name	Total minutes	Field goals
124	Kevin Durant	3241	794
243	Jason Kidd	2883	284
113	Travis Diener	50	2



- Kevin Durant: Good scorer.
- Jason Kidd: Other role.
- Dwayne Jones & Travis Diener: Benchwarmers.

- n=470 women, age $\in$ [25,45] y., bust $\in$ [86,90[cm.
- High variability of body shapes.
- Archetypoids to identify subjects with fittings problems.

- Let $\mathbf{D}$ be a $n \times n$ matrix where d_{ij} is the distance between womens i and j.
- Compute cMDS where m is the dimension of the space which the data are to be represented. We choose $m = 4$ as the first integer for which the good of fit (Mardia et al. (1979),eq. 14.4.8) is greater than 90%.
- Compute the archetypoids of the resulting $n \times m$ matrix $\mathbf{X}$.



Archetypoids: a new approach to define representative archetypal data

Abstract

A new concept is introduced: archetypoids. Archetypoid analysis represents each observation in a data set as a mixture of actual observations in the data set, which are pure type or archetypoids. Unlike archetype analysis, archetypoids are real observations, not a mixture of observations. This is relevant when existing archetypal observations are needed, and not fictitious. An algorithm is also proposed to find them. Some of their theoretical properties are introduced. We also show how to obtain them when only dissimilarities between observations are known (features are unavailable). Archetypoid analysis is illustrated in three different problems and several examples, comparing them with the archetypes and the nearest observations to them.

Keywords: Archetype; Convex hull; Unsupervised learning; Extremal point

Conclusions

- A new proposal for the accommodation problem.
- Advantages of the archetypal analysis regarding PCA.
- New archetypal concept: the archetypoid.
- Algorithm to obtain archetypoids.
- Advantages of archetypoids regarding archetypes.
- All the R code is freely available from my website: http://www.uv.es/vivigui/software.html
- It also belongs to an R package hopefully soon available from CRAN.

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Anthropometry: An R Package for Analysis of Anthropometric Data

Guillermo Vinué
Department of Statistics and O.R., University of Valencia, Valencia, Spain.

Abstract

The development of new powerful 3D scanning techniques has enabled the generation of updated anthropometric data, very representative of the target population. In order to get full benefit from them, they must be comprehensively analyzed by means of rigorous statistical methodologies. This paper presents a new R package called **Anthropometry** that joints together some statistical methodologies concerning clustering, the statistical shape analysis and the archetypal analysis, specially developed to deal with anthropometric data. The utility of the package is shown by analyzing anthropometric data obtained from a survey of the Spanish female population and from the 1967 United States Air Force Survey.

Keywords: R, anthropometric data, clustering, statistical shape analysis, archetypal analysis.

Expert Systems with Applications 39 (2012) 10512–10520

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Apparel sizing using trimmed PAM and OWA operators

M.V. Ibáñez [a,*], G. Vinué [b], S. Alemany [c], A. Simó [a], I. Epifanio [a], J. Domingo [d], G. Ayala [b]

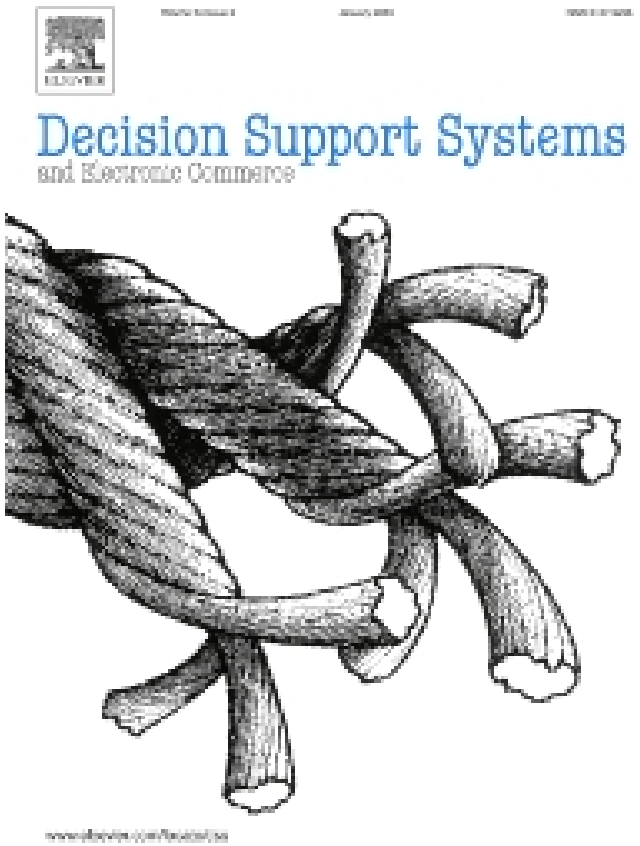
[a] Department of Mathematics, Universitat Jaume I, Castellón, Spain
[b] Department of Statistics and O.R., University of Valencia, Valencia, Spain
[c] Biomechanics Institute of Valencia, Universidad Politécnica de Valencia, Valencia, Spain
[d] Department of Informatics, University of Valencia, Valencia, Spain

ARTICLE INFO

Keywords:
Anthropometric data
Sizing systems
Trimmed k-medoids
OWA operators

ABSTRACT

This paper is concerned with apparel sizing system design. One of the most important issues in the apparel development process is to define a sizing system that provides a good fit to the majority of the population. A sizing system classifies a specific population into homogeneous subgroups based on some key body dimensions. Standard sizing systems range linearly from very small to very large. However, anthropometric measures do not grow linearly with size, so they can not accommodate all body types. It is important to determine each class in the sizing system based on a real prototype that is as representative as possible of each class. In this paper we propose a methodology to develop an efficient apparel sizing system based on clustering techniques jointly with OWA operators. Our approach is a natural extension and improvement of the methodology proposed by McCulloch, Paal, and Ashdown (1998), and we apply it to the anthropometric database obtained from a anthropometric survey of the Spanish female population, performed during 2006.



SECOND REVISION

Looking for representative fit models for apparel sizing

G. Vinué[(1)], T. León[(1)], S. Alemany[(2)] and G. Ayala[(1)]

(1) Department of Statistics and O.R., Avda. Vicent Andrés Estellés, 1. 46100-Burjasot. University of Valencia, Valencia, Spain.
(2) Biomechanics Institute of Valencia, Universidad Politécnica de Valencia, Valencia, Spain.

Abstract

This paper is concerned with the generation of optimal fit models oriented to apparel design. Representative fit models or prototypes are important for defining efficient sizes. However, there is no agreement among apparel manufacturers and each one has their own prototypes and size charts. As a consequence, there is a lack of standard sizes in garments from different apparel manufacturers.

We propose a methodology based on a new hierarchical partitioning around medoids clustering method that, in a previous version, has been used in gene clustering. We are going to consider a different dissimilarity measure and a different method in order to obtain a classification tree. We will obtain optimal prototypes and outliers. These outliers have to be removed before defining prototypes so that the companies' market share can be optimized.

All the analyses are performed over the anthropometric database obtained from a survey of the Spanish female population.

Keywords: HIPAM, Hierarchical tree, Partitioning around medoids, Fit models, Mean Split Silhouette, INCA statistic.

The k-means algorithm for 3D shapes with an application to apparel design

G. Vinué[(1)], A. Simó[(2)], and S. Alemany[(3)]

[(1)] *Department of Statistics and O.R., University of Valencia, C/ Dr. Moliner, 50. 46100-Burjassot, Valencia, Spain.*
[(2)] *Department of Mathematics-IMAC. University Jaume I. Avda. del Riu Sec s/n. 12071-Castellón, Spain.*
[(3)] *Biomechanics Institute of Valencia, Universidad Politécnica de Valencia, Camino de Vera s/n. 46022-Valencia, Spain.*

January 2, 2013

Abstract

Since the basic foundation of the k-means algorithm is to use the fact that the sample mean is the value that minimizes the Euclidean distance from each point to the centroid of the cluster to which it belongs, the idea of integrating the Procrustes distance and Procrustes mean into the k-means algorithm to adapt it to the shape analysis context arises in a natural way. There have been several attempts in that way, each one adapting a different version of the k-means algorithm. In this paper we propose to adapt the Lloyd version of the k-means algorithm to the field of statistical shape analysis, focusing on the three dimensional case. We present a study

Basic references


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


Eugster, M. J., Leisch, F., April 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.


Mardia, K., Kent, J., Bibby, J., 1979. Multivariate Analysis. Academic Press.


Robinette, K., McConville, J., 1981. An Alternative to Percentile Models. In: SAE Technical Paper.


Seiler, C., Wohlrabe, K., 2013. Archetypal scientists. Journal of Informetrics 7, 345–356.


Zehner, G. F., Meindl, R. S., Hudson, J. A., April 1993. A Multivariate Anthropometric Method For Crew Station Design: Abridged. Tech. rep., Human Engineering Division, Armstrong Lab., Wright-Patterson Air Force Base, Ohio.

Example of a *which* vector of archetypes

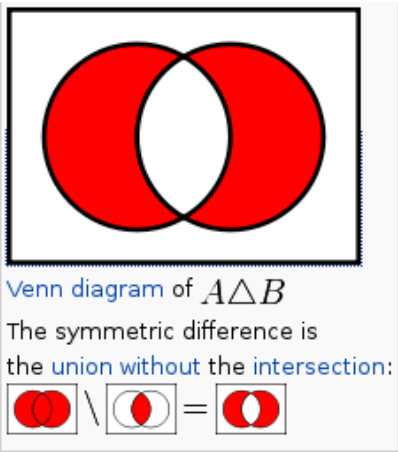
	Name	Archetype 1	Archetype 2	Archetype 3
124	Kevin Durant	1.00000	0.00000	0.0000
243	Jason Kidd	0.06203	0.00000	0.93788
236	Dwayne Jones	0.00000	0.99997	1.5589e-14

- Each α_{ij} is the weight of the archetype j for the individual i, that is to say, the α coefficients represent how much each archetype contributes to the approximation of each individual (as convex combination).
- This allows the assignment of the observations to their nearest archetypes and, consequently, the identification of the most archetypal observation(s).

Misclassification error

- Let X be a image raster. A binary image b is identified with a subset $B \subseteq X$, $B = \{x \in X : b(x) = 1\}$.
- Let A and B be two binary images associated to the trunk of two women and defined in a lattice Λ.
- There are several metrics for measuring the differences between A and B.
- We use the simplest one, which is the misclassification error:

$$d(A, B) = \frac{nu(A \Delta B)}{nu(\Lambda)}$$, where Δ is the set symmetric difference and nu counts the number of pixels in that set, that is to say, the volume of the set.



Set symmetric difference: Set of elements which are in either A or B and not in their intersection.