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

I. Epifanio(1)(*), G. Vinué (2), S. Alemany (3)

(1). Ph. 34 964728390, fax 34 964728429, epifanio@uji.es. Dept. Matemàtiques. Universitat Jaume I. Castelló. Spain (*) Corresponding author.

(2) Department of Statistics and O.R., University of Valencia, Valencia, Spain.

(3) Biomechanics Institute of Valencia, Universidad Politécnica de Valencia, Valencia, Spain.

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 or leave the selection by a criterion. Unlike PCA, the objective of the archetypal analysis is obtaining extreme individuals, so it is the appropriate statistical technique for solving this type of problem. Archetypes cannot be obtained with PCA even if we consider all the components, as we show in the application.

Keywords: representative case; archetype; percentile; anthropometry; representative human model generation

ACKNOWLEDGEMENTS

This work has been partially supported by Grants CICYT TIN2009-14392- C02-01, CICYT TIN2009-14392- C02-02, MTM2009-14500-C02-02, GV/2011/004 and Bancaixa-UJI P11A2009- 02.

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

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 or leave the selection by a criterion. Unlike PCA, the objective of the archetypal analysis is obtaining extreme individuals, so it is the appropriate statistical technique for solving this type of problem. Archetypes cannot be obtained with PCA even if we consider all the components, as we show in the application.

Keywords: representative case; archetype; percentile; anthropometry; representative human model generation

1. Introduction

Products intended to "fit" their users must be designed with careful consideration of the size and shape of the user population. In ergonomic design and evaluation, a small group of human models which represents the anthropometric variability of the target population is commonly used. Use of a small group of human models provides designers an efficient way to develop and evaluate a product design. In the multivariate accommodation problem, a set of representative cases (human models) are searched in order to cover a certain percentage of the user population. The appropriate selection of this small group is critical if we want to accommodate a certain percentage of the population.

Two strategies can be considered in searching the human models according to the characteristics of product being designed: searching on a boundary or a set of grids. If the product being designed is a a one-size product (one-size to accommodate people within a designated percentage of the population) such as a bus operator's workstation or a helicopter cockpit, the cases are selected on an accommodation boundary. However, if we are designating a multiple-size product (n sizes to fit n groups of people within a designated percentage of the population), being clothing the most apparent example, the cases are selected over a set of grids formed in the distribution of anthropometric dimensions [12]. In this work, we center on the first situation: one-size product.

It has long been demonstrated that the use of percentiles is not appropriate, due to the fact that, with the exception of $50th$ percentiles, percentile values are not additive [17, 26, 23]. Different alternatives have been proposed using different statistical techniques such as regression [23, 7, 16] or cluster analysis [14]. However, the most common approach is based on the use of principal component analysis (PCA) [26, 1, 10, 9, 11, 24]. The idea of this approximation consists in considering the first principal components and selecting several extreme points in an ellipse (or in a circle if they are standardized) which covers a certain percentage of the data (95%, for example). If a workspace is designed to enable all these cases to operate efficiently, then all other less extreme body types and sizes in the target population (within the circle) should also be well accommodated.

Friess in [8] makes an excellent analysis of the PCA-approach, where his comparison reveals that PCA approach have many limits: 1) in its simplest variant it can lead to enormous portions of the population (nearly 50%) being left out; 2) an improved version of it requires the use of a great number of components (if not all) and the contribution of octant points to the determination of multivariate boundaries remains unclear. Still, even this version did not achieve the level of accommodation it set out to reach.

Note that the PCA-approach followed for example in [26, 11, 24] has several drawbacks. As it only chooses the first components, part of the data variation is removed (according to the variation explained by the first components. In addition, not considered variation may represent cases difficult to accommodate). Therefore, when building the ellipse, the true covered percentage is not the 95%. Furthermore, with two and three components the selected cases are respectively, eight and fourteen, so the number of cases would increase if we would want to represent more than three components in order to consider more variation. It may not be practical to select too many cases. Moreover, if we restrict ourselves to the chosen components, there might be combinations of variables which were not collected by the principal components (even considering all the possible components) and which correspond to extreme data, since the goal of PCA is not the calculation of extreme data. This final consideration will be shown in Section 3.

Therefore, an alternative to the previous methodology, is proposed: the archetypal analysis [3]. We propose a methodology with which we can assure the covering of a certain proportion of the population. Archetypal analysis assumes that there are several "pure" individuals who are on the "edges" of the data, and all others individuals are considered to be mixtures of these pure types. Archetypal analysis (AA) estimates the convex hull of a data set, as such AA favors features that constitute representative "corners" of the data, i.e. archetypes. Archetypes are almost always easy to interpret as they represent extreme combinations of features. In the original paper on AA [3] the method was demonstrated useful in the analysis of air pollution and head shape and later also for tracking spatio-temporal dynamics. Recently, AA has found use in benchmarking and market research [15] and in particular, for identifying typically extreme practices, rather than just good practices [21], as well as in the analysis of astronomy spectra [2] as an approach for the end-member extraction problem [20]. Ref. [6] is another interesting contribution in which archetypal athletes are determined for American basketball and European soccer, according the data from their most representative leagues. AA has been shown to be relevant also for a large variety of machine learning problems and for high-dimensional data arising from video-taped images [18, 19, 25]. A recent application of AA for comparing different species of bats is found in [4].

Archetypes can be computed easily by means of a library of free software R [5, 22]. The code developed to calculate them from our data is freely available and it can be seen in Appendix A. The outline of the paper is as follows: section 2 describes the data set and the methodology used in this paper. The application of our procedure is given in Section 3. Conclusions and possible further developments conclude the paper in Section 4.

2. Materials and Methods

2.1. Data

Our data set comes from the 1967 United States Air Force (USAF) Survey (available from *http://www.dtic.mil/dtic/*, and as supplemental material for be readded with our code). The 1967 USAF Survey was conducted during the first three months of 1967 under the direction of the Anthropology Branch of the Aerospace Medical Research Laboratory, located in Ohio. Subjects were measured at 17 Air Force bases across the United States of America. A total of 202 variables (including body dimensions and background variables) were taken on 2420 Air Force personnel between 21 and 50 years of age. From the total of variables, we select in particular six anthropometric measurements, the same selected in [26]. These six dimensions are the so-called cockpit dimensions because they are the most important dimensions in order to design aircraft cockpits. The summary statistics of these variables can be seen in Table 1. Table 2 gives the description of each one of them, regarding [13].

We have chosen this well-known problem and database in order to highlight the contribution of our methodology.

Table 1: Summary statistics for the six variables considered.

Measurement (inches)	Mean	Standard Deviation
Thumb Tip Reach	31.618	1.567
Buttock-Knee Length	23.781	1.064
Popliteal Height Sitting	17.206	0.885
Sitting Height	36.687	1.251
Eye Height Sitting	31.870	1.188
Shoulder Height Sitting	24.037	1.126

Table 2: Description of the six variables considered.

Measurement	Description
Thumb Tip Reach	Measure the distance from the wall to the tip of the thumb.
Buttock-Knee Length	Measure the horizontal distance from the rearmost surface of the right buttock to the forward surface of the right kneecap.
Popliteal Height Sitting	Measure the vertical distance from the footrest surface to the superior margin of the right kneecap.
Sitting Height	Measure the vertical distance from the sitting surface to the top of the head.
Eye Height Sitting	Measure the vertical distance from the sitting surface to the right external canthus (outer “corner” of eye).
Shoulder Height Sitting	Measure the vertical distance from the sitting surface to the right Acromion - the bony landmark at the tip of the shoulder.

Fig. 1 shows a common skeleton of an aircraft pilot with explanations of the six selected measurements.

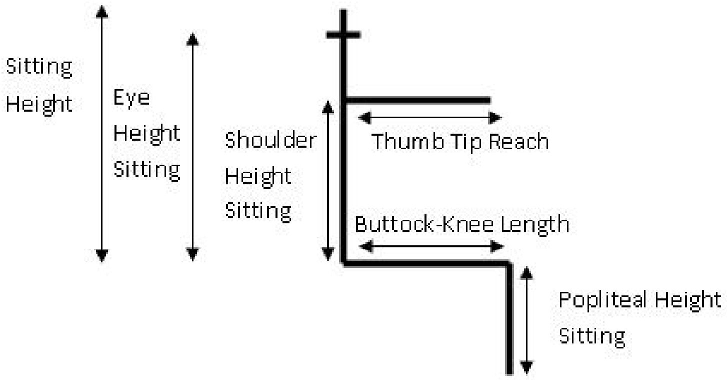


Figure 1: Generic skeleton for an aircraft pilot.

2.2. Methodology

2.2.1. Archetypal analysis

Consider an $n \times m$ matrix $\mathbf{X}$ representing a multivariate data set with n observations and m variables. The goal of the archetypal analysis is to find a $k \times m$ matrix $\mathbf{Z}$ that characterize the archetypal patterns in the data, such that data can be represented as mixtures of those archetypes. More precisely, the archetypal analysis aims at obtaining the two $n \times k$ coefficient matrices α and β which minimize the residual sum of squares

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

under the constraints

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

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

Constraint 1) tell us that predictors of $\mathbf{x_i}$ are finite mixtures of archetypes, $\mathbf{x_i} = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z_j}$, while constraint 2) implies that archetypes $\mathbf{z_j}$ are convex combinations of the data points, $\mathbf{z_j} = \sum_{l=1}^{n} \beta_{jl}\mathbf{x_l}$.

2.2.2. Calculation of archetypes with a 95% *accommodated*

The procedure that we propose is as follows. First, we standardize the variables, as [26]. Since we are looking for archetypes for the 95% of the sample, we have to remove the more extreme 5% data. We can do this in two ways. If we assume that the data comes from a multivariate (m-variate) normal (as we are dealing with anthropometric measurements, we can assume this), then we can use the fact that the Mahalanobis distance from a observation to the mean $D^2 = (\mathbf{x} - \hat{\mu})'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x} - \hat{\mu})$, where $\hat{\mu}$ is the estimated mean and $\hat{\mathbf{\Sigma}}$ is the estimated covariance matrix, is distributed according to the Chi-square distribution with m degrees of freedom. Therefore, those observations more far away from the 95th percentile of the Chi-square distribution, can be removed from the analysis. We use this procedure because with the PCA approach, the normality is assumed when drawing the circle. Anyway, if the normality hypothesis is not acceptable, a non parametric alternative might be employed. For example, the depth of each data can be calculated. Then a removing procedure from the less to the deepest data can be applied until getting the desired 95%. However, this approach has the disadvantage that the desired percentage is not under control of the analyst as with the former procedure. For instance, there was almost a 7% of less deep data in the USAF Survey ($169/2420 = 0.0698$), each one of them with the same depth. We checked that there was pretty agreement using the Mahalanobis distance and depth procedure with the USAF database. That's why we only consider the former approach using the Mahalanobis distance.

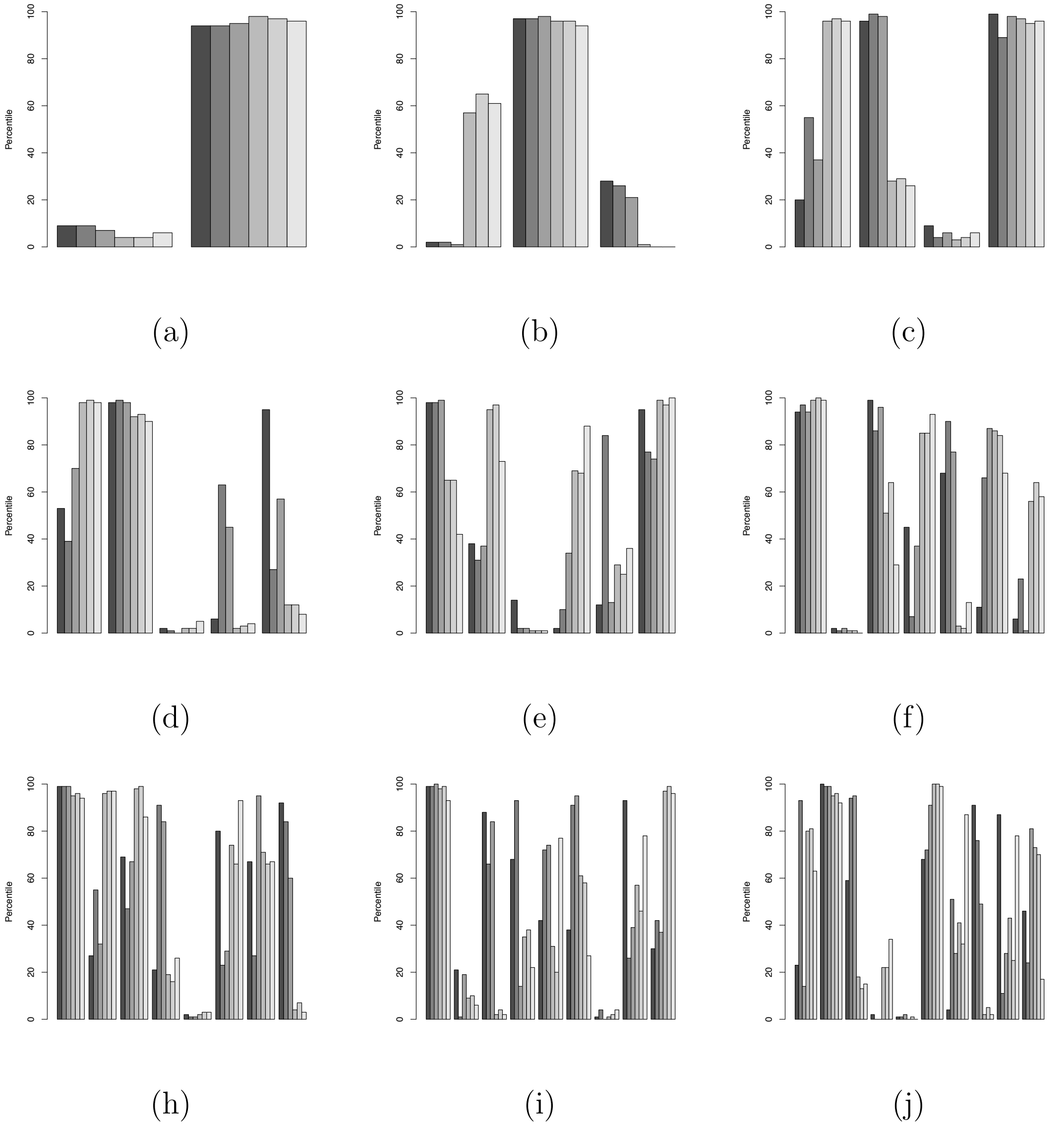
After removing the more extreme 5% data, we apply archetypal analysis to get the archetypes. The number of them is decided by our own criterion or by an external criterion, as explained in Section 3.2. We would like to point out that the archetypes are not nested. For instance, if we first calculate three archetypes and then we calculate four archetypes, there is no reason so that these four include those first three obtained, as the existing ones can change to better capture the shape of the data set. The archetypes fall on the convex hull of the data, except when $k = 1$, where the archetype obtained is the sample mean. Once obtained them (they are measurements for each dimension), we will be able to calculate to which percentile corresponds each one of the variables used.

In summary, the steps are the following. First, depending on the problem, to standardize the data or not (to use standardization depends on one's sense about the data, but in this case the variables should be standardized as they measure different dimensions). Second, to use Mahalanobis distance and Chi-square distribution to select the subsample for obtaining the archetypes as the third and last step.

3. Results

3.1. Archetypes for 1967 USAF

We have computed the archetypes from $k = 1$ to $k = 10$ (remember that for $k = 1$ the mean of each variable is obtained). Fig. 2 displays the percentile value of each variable for each archetype, from $k = 2$ (a) to $k = 10$ (j). The percentiles of each archetype are represented by each set of bars, where a bar represents a different variable, from dark gray (Thumb Tip Reach) to light gray (Shoulder Height Sitting). For example, in Fig. 2 (a), the first archetype is low in all variables, whereas the second archetype is high in the six variables. In Fig. 2 (b), the percentiles for each one of the three archetypes are shown. The first archetype has small percentiles for the first three variables (corresponding to limb dimensions), while has average measures for the last three variables (corresponding to torso dimensions). The second archetype represents individuals which are huge in all measurements, and the third archetype represents individuals which are small, although for the first three variables not very small, around the 25nd percentile. As said before, archetypes are not nested. As more archetypes are found, the existing ones can change to better capture the information of the data set. So, we have to determine which is the number of archetypes to be considered.



(a) (b) (c)

(d) (e) (f)

(h) (i) (j)

Figure 2: Percentiles of archetypes from $k = 2$ to $k = 10$.

3.2. Choosing the number of archetypes

The user can decide how many archetypes wants to consider. However, in case that you are not sure about which is the best number, the residual function can orientate you. As in many cases there is no rule for the correct number of archetypes k. A simple method the determine the value of k is to run the algorithm for different numbers of k and use the elbow criterion on the residual sum of squares, RSS, where a flattening of the curve indicates the correct value of k. This method is very common in statistics. The RSS from $k = 2$ to 15 is graphed in Fig. 3.

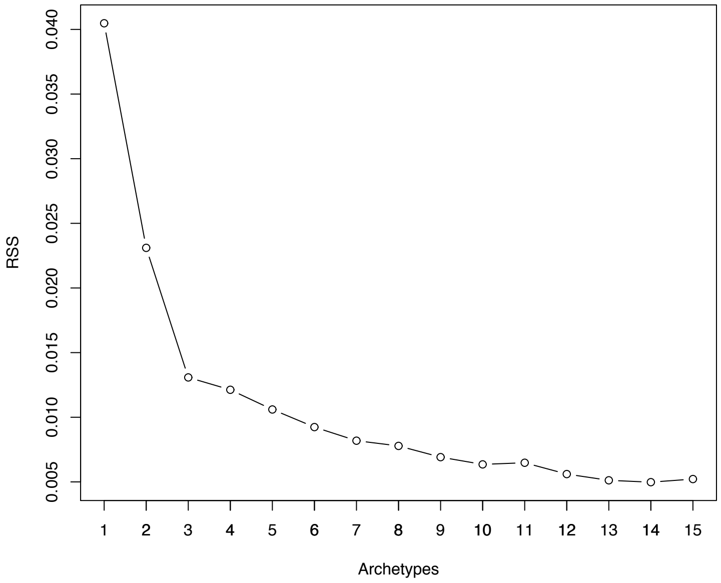


Figure 3: Screeplot of the residual sum of squares.

According to Fig. 3, an elbow occurs at $k = 3$. There is another one at k= 7 (there is a flat zone in the plot from $k = 7$ to $k = 8$), and maybe at $k = 10$. Corresponding to Occam's razor three and seven archetypes are considered as the best numbers of archetypes (the law of parsimony is considered since a large numbers of representative cases may overwhelm the designer and thus, be counterproductive, although if he/she is interested in more archetypes, they can be computed). Results for three archetypes where commented in subsection 3.1. We focus on seven archetypes, whose percentiles were represented in Fig. 2 (f). The first archetype has high percentiles for all variables, the opposite to the second archetype with low percentiles in all variables. The third archetype has high percentiles in the first three variables (those related with limb dimensions), whereas has middle percentiles for the last three variables (those related with torso dimensions), just the opposite for the fourth archetype (middle percentiles for the first three variables and high percentiles for the last three variables). In the fifth archetype, the percentiles are middle-high for the first three variables and low for the last three, the opposite of the seventh archetype, with low percentiles for the first three variables and middle-high for the last ones. The sixth archetype shows a different profile, with high percentiles for all variables except the first one, the only one related with arms (a man which is huge in all measurements, but with short arms).

If the body size variability exhibited by these archetypes is accommodated into a new aircraft design, then the target percentage of the total population will. This assumes, that the seat, rudder, and other adjustable components can be adjusted in sufficiently small increments. Without such adjustability, it may be necessary to pick more representative cases.

3.3. Comparison with PCA results

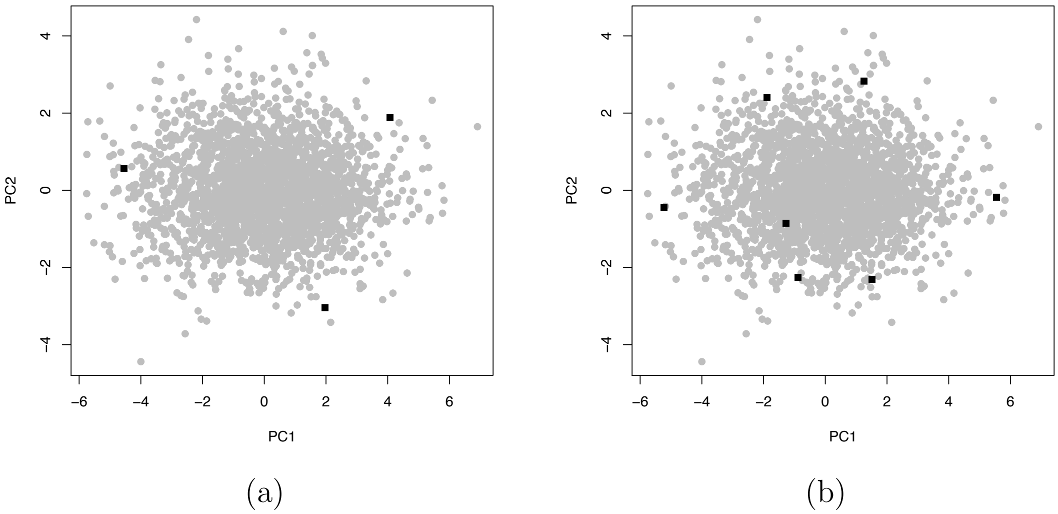
In order to compare the archetypes obtained with our methodology with that obtained with PCA as in [26, 24], we have computed PCA for the six standardized variables and all the individuals. Table 3 shows the coefficients for the six principal components, the percentage of variance explained for each component, and the cumulative percentage.

Table 3: PCA coefficients and percentage of explained variance

	PC1	PC2	PC3	PC4	PC5	PC6
Thumb Tip Reach	-0.364	0.453	0.697	0.418	0.04	-0.001
Buttock-Knee Length	-0.36	0.464	-0.716	0.374	0.036	-0.043
Popliteal Height Sitting	-0.39	0.408	0.025	-0.809	-0.144	0.077
Sitting Height	-0.46	-0.353	0.02	-0.082	0.305	-0.751
Eye Height Sitting	-0.449	-0.367	-0.025	0.004	0.494	0.648
Shoulder Height Sitting	-0.416	-0.392	-0.01	0.155	-0.8	0.098
% Explained Variance	61.5	21.0	6.59	5.69	4.08	1.07
Cumulative %	61.5	82.6	89.15	94.84	98.93	100

Note that the first two components capture the 82.6% of variability (89.15% with the first three components). If only the first two components are considered, some variability (maybe important) is discarded. The first component can be interpreted as the overall size of the individuals. The second component contrasts (the sign is different) the limb dimensions (the first three) and the torso dimensions (the last three). The third and fourth components show a contrast inside the limbs (Thumb Tip Reach versus Buttock-Knee Length for the third, and Thumb Tip Reach and Buttock-Knee Length versus Popliteal Height Sitting for the fourth). The torso dimensions are contrasted instead by the fifth and sixth components (Sitting Height and Eye Height Sitting versus Shoulder Height Sitting for the fifth, and Sitting Height versus Eye Height Sitting for the sixth).

Fig. 4 shows the scores for the two first principal components of all individuals in gray, with the scores for the three archetypes (a), and seven archetypes (b) in black squares.



(a)

(b)

Figure 4: PC scores for three (a) and seven (b) archetypes

The archetypes obtained with $k = 3$ are similar to those that can be obtained with the two first principal components (see the results for three archetypes in subsection 3.1, the second archetype corresponds to an extreme of PC1, and the first and third archetype correspond to a combination of extremes of PC1 and PC2, octants). In the case that $k = 7$ archetypes, all except the sixth archetype correspond to extreme combinations of PC1 and PC2 (they form a circle). However, the sixth archetype (the one with scores -1.28 and -0.86 for PC1 and PC2 respectively) cannot be extracted as a combination of the two first PC. In fact, it cannot be obtained with any combination of the PCs.

Table 4 shows the percentile values for the 8 cases extracted with the classical PCA approach for 95% accommodation (see [24] for details).

Table 4: Percentile values for two principal component representative cases

	A	B	C	D	W	X	Y	Z
Thumb Tip Reach	98	38	2	62	96	90	4	10
Buttock-Knee Length	98	37	2	63	96	90	4	10
Popliteal Height Sitting	98	31	2	69	97	87	3	13
Sitting Height	80	1	20	99	99	16	1	84
Eye Height Sitting	78	1	22	99	98	16	2	84
Shoulder Height Sitting	74	2	26	98	98	14	2	86

Percentiles for the archetypes with $k = 7$ appear in table 5.

Table 5: Percentile values for seven archetypes

	1	2	3	4	5	6	7
Thumb Tip Reach	94	2	99	44	68	10	6
Buttock-Knee Length	97	1	86	7	89	66	22
Popliteal Height Sitting	94	2	96	37	77	87	0
Sitting Height	99	0	51	86	3	86	57
Eye Height Sitting	99	1	64	85	2	84	64
Shoulder Height Sitting	99	1	29	93	13	68	57

The case W corresponds with archetype 1, the Y with archetype 2. The case A is in the middle between archetype 1 and 3. There is no case with PCA for archetype 3. The same occurs with case B, which is in the middle between archetype 2 and 5. The case X is the nearest to archetype 5, although it does not correspond exactly. The case C is in the middle between archetype 2 and 7. The case D could be seen as a kind of combination of archetypes 1, 4 and 6, although there is not case for archetypes 4 and 6. The case Z is in the middle between archetypes 4 and 7. As we have seen, except in two cases there is no clear coincidence between the cases for PCA and archetypes.

In table 6 and 7 the corresponding values for each variable are displayed.

Table 6: Variable values for two principal component representative cases

	A	B	C	D	W	X	Y	Z
Thumb Tip Reach	34.93	31.14	28.31	32.14	34.3	33.61	28.94	29.62
Buttock-Knee Length	26.02	23.44	21.55	24.13	25.60	25.12	21.96	22.44
Popliteal Height Sitting	19.07	16.77	15.35	17.64	18.83	18.21	15.58	16.20
Sitting Height	37.74	33.89	35.63	39.48	39.41	35.46	33.96	37.92
Eye Height Sitting	32.8	29.24	30.98	34.5	34.39	30.67	29.35	33.08
Shoulder Height Sitting	24.77	21.6	23.3	26.48	26.28	22.83	21.8	25.24

Table 7: Variable values for seven archetypes

	1	2	3	4	5	6	7
Thumb Tip Reach	34.18	28.51	35.34	31.34	32.33	29.69	29.24
Buttock-Knee Length	25.85	21.23	24.94	22.27	25.09	24.18	22.97
Popliteal Height Sitting	18.65	15.39	18.79	16.89	17.84	18.22	14.99
Sitting Height	39.66	33.57	36.7	38	34.46	38.07	36.88
Eye Height Sitting	35.05	29.24	32.28	33.08	29.58	33.04	32.28
Shoulder Height Sitting	26.73	21.26	23.41	25.8	22.82	24.56	24.22

Fig. 5 shows the skeletons for each one of these seven archetypes.

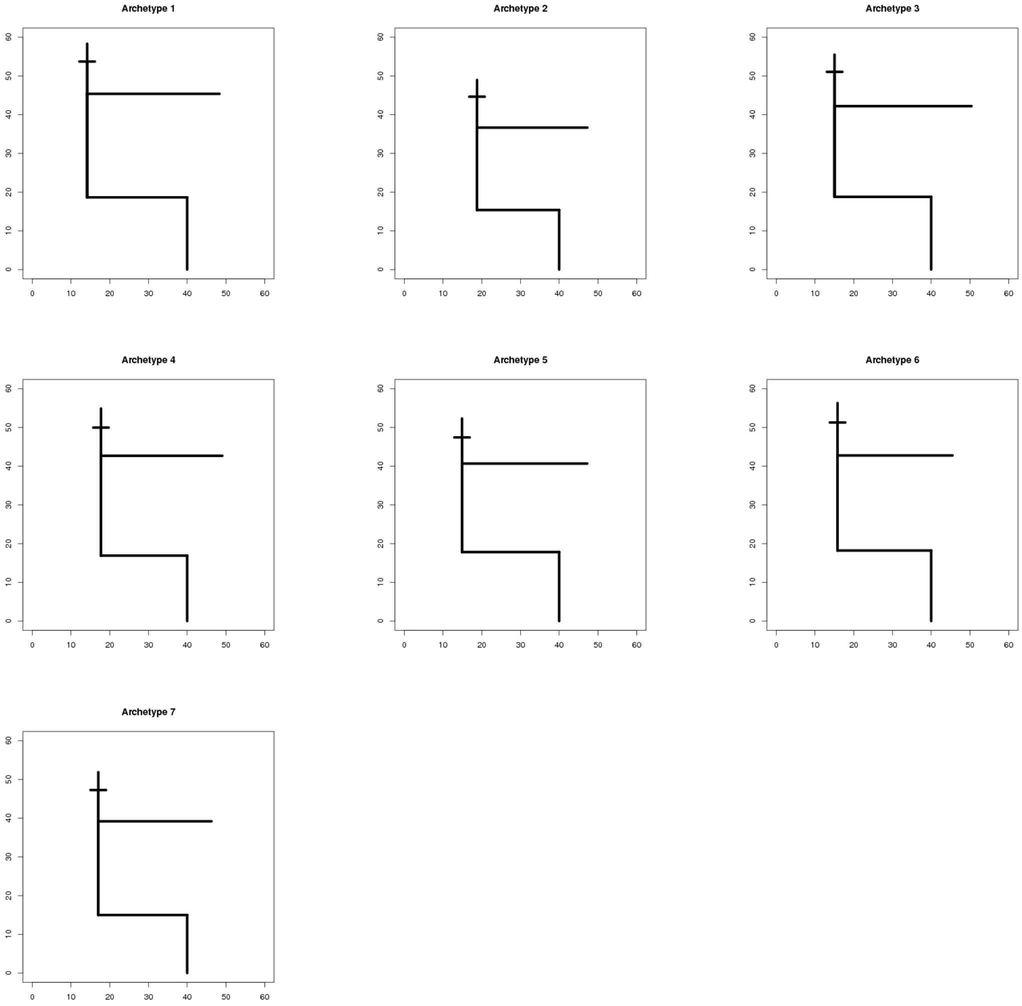


Figure 5: Skeleton plots visualizing the seven archetypes.

The nearest individual to each archetype can be obtained by simply computing the distance between the archetypes and the individuals and choosing the nearest. The code can be seen in Appendix A.

4. Conclusions

We have proposed an alternative to determine test cases based on archetypal analysis. This technique effectively considers a certain percentage of the population for accommodation, not as the classical PCA where the percentage of accommodation is determined without consider all the variability, and therefore it does not consider effectively the accommodation percentage desired previously. We have applied the technique to a classical database and we have compared it with the methodology based on PCA, which is the most common for obtaining the boundary cases. With our methodology, we obtained seven archetypes, one of them could not be extracted by any principal component. Friess [8] indicates that the contribution of octant points to the determination of multivariate boundaries with PCA remains unclear, and he suggests that caution must be used when relying on PCA outcome, as it does not achieve the level of accommodation it set out to reach. He also recommends that a PCA derived boundary model be systematically tested against the sample from which it was calculated to allow for possible adjustments. As the objective of the archetypal analysis is obtaining extreme individuals, just as the objective of obtaining boundary cases, these adjustments are not necessary with our methodology. We have shown how to select the number of the archetypes, based on the elbow criterion for the RSS and the Occam's razor, since large numbers of representative cases may overwhelm the designer and thus, be counterproductive. The archetypes can be obtained easily as the code is free and open.

The use of archetypal analysis with 3D body scanner is our future study.

References

[1] Bittner, A., Glenn, F., Harris, R., Iavecchia, H., Wherry, R., 1987. CADRE: A family of mannikins for workstation design. In: Asfour, S.S. (ed.) Trends in Ergonomics/Human Factors IV. North Holland. pp. 733–740.

[2] Chan, B., Mitchell, D., Cram, L., 2003. Archetypal analysis of galaxy spectra. Monthly Notices of the Royal Astronomical Society 338.

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

[4] D'Esposito, M. R., Palumbo, F., Ragozini, G., 2012. Interval Archetypes: A New Tool for Interval Data Analysis. Statistical Analysis and Data Mining.
URL `DOI:10.1002/sam.11140`

[5] Eugster, M. J., Leisch, F., April 2009. From Spider-Man to Hero - Archetypal Analysis in R. Journal of Statistical Software 30 (8), 1–23. URL http://www.jstatsoft.org/

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

[7] Flannagan, C. A., Manary, M. A., Schneider, L. W., Reed, M. P., 1998. An Improved Seating Accommodation Model with Application to Different User Populations. In: Human Factors in Driving, Vehicle Seating, and Rear Vision. SAE SP 1358.

[8] Friess, M., 2005. Multivariate Accommodation Models using Traditional and 3D Anthropometry. In: SAE Technical Paper.

[9] Friess, M., Bradtmiller, B., 2003. 3D Head Models for Protective Helmet Development. In: Proceedings of the SAE 2003.

[10] Gordon, C. C., Churchill, T., Clauser, C. E., Bradtmiller, B., McConville, J. T., Tebbetts, I., Walker, R. A., March 1989. 1988 Anthropometric Survey of U.S. Army personnel: Summary statistics interim report. Tech. rep., US Army Natick Research, Development and Engineering Center.

[11] Hudson, J. A., Zehner, G. F., Meindl, R. D., October 1998. The USAF Multivariate Accommodation Method. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 42 (10), 722–726.

[12] Jung, K., Kwon, O., You, H., 2010. Evaluation of the multivariate accommodation performance of the grid method. Applied Ergonomics 42 (1), 156–61.

[13] Kennedy, K. W., June 2001. Anthropometric accommodation in aircraft cockpits.
URL http://www.humanics-es.com/anthro/

[14] Kim, K., Kim, H., Lee, J., Lee, E., Kim, D., 2004. Development of a New 3D Test Panel for Half-Mask Respirators by 3D Shape Analysis for Korean Faces. Journal of the International Society for Respiratory Protection 21, 125–134.

[15] Li, S., Wang, P., Louviere, J., Carson, R., December 2003. Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals. In: ANZMAC 2003 Conference Proceedings.

[16] Manary, M. A., Flannagan, C. A., Reed, M. P., Schneider, L. W., 1998. Development of an Improved Driver Eye Position Model. In: Human Factors in Driving, Vehicle Seating, and Rear Vision. SAE SP 1358.

[17] Moroney, L. W. F., MSC, USN, Smith, M. J., September 1972. Empirical reduction in potential user population as the result of imposed multivariate anthropometric limits. Tech. rep., Naval Aerospace Medical Research Laboratory.

[18] Mørup, M., Hansen, L. K., 2010. Archetypal Analysis for Machine Learning. In: IEEE International Workshop on Machine Learning for Signal Processing.

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

[20] Plaza, A., Martínez, P., Pérez, R., Plaza, J., 2004. A Quantitative and Comparative Analysis of Endmember Extraction Algorithms From Hyperspectral Data. IEEE Transactions on Geoscience and remote sensing 42 (3).

[21] 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.

[22] R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

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

[24] Robinson, J. C., Robinette, K. M., Zehner, G. F., February 1992. User's guide to the anthropometric database at the computerized anthropometric research and design (card) laboratory (U). Tech. rep., Systems Research Laboratories Inc.

[25] Stone, E., Olson, B., 1999. Archetypal analysis of cellular flame data. Tech. rep., Dept Mathematics and Statistics, Utah State University.

[26] 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 Laboratory, Wright-Patterson Air Force Base, Ohio.

Appendix A: Archetypal Analysis in R with our data

The following commands (also available at *http://www3.uji.es/∼epifanio/RESEARCH/archboundary.rar*) show how to reproduce the results from this article in R. Package archetypes is freely available on *http://cran.R-project.org*.

First, we read and preprocess the database USAF 1967.

```
> m <- read.table(med1967.dat, sep =" ")
> #Variable selection:
> sel <- c(48, 40 ,39, 33, 34, 36)
> #Changing to inches:
> mpulg <- m[,sel] / (10 * 2.54)
> #Standarizing (normalizing):
> smpulg <- scale(mpulg, center = mean(mpulg[,]), scale = sd(mpulg[,]))
```

Next, we remove the more extreme 5% data using the Mahalanobis distance and we check that calculating the depth of each data, we get very similar results.

```
> Sx <- cov(smpulg)
> D2 <- mahalanobis(smpulg, colMeans(smpulg), Sx)
> #Number of individuals not considered:
> sum(D2 > qchisq(0.95, df = 6))
> #Which individuals are not considered:
> qchi <- which(D2 > qchisq(0.95, df = 6))

> library(depth)
> dt = c()
> for(i in 1 : nrow(smpulg)){
 dt[i] <- depth(smpulg[i,], smpulg)
 }
> #Number of individuals not considered:
> sum(dt == min(dt))
> #Which individuals are not considered:
> qd <- which(dt == min(dt))

> #Agreement between Mahalanobis and depth:
> intersect(qd, qchi)
```

Results shown in Section 3 are obtained as follows.

```
> #Individuals considered for covering 95%:
> qlchi <- which(D2 <= qchisq(0.95, df = 6))
> #Database with the 6 variables and the selected individuals
> lsmpulg <- smpulg[qlchi,]

> library("archetypes")
> #For reproducing results, seed for randomness:
> set.seed(2010)
> #Run archetypes algorithm repeatedly from 1 to 15 archetypes:
> lass15 <- stepArchetypes(data = lsmpulg, k = 1:15,
 verbose = FALSE, nrep = 3)
> #Plot from 1 to 15 archetypes:
> screeplot(lass15)
> #3 archetypes:
> a3 <- bestModel(lass15[[3]])
> parameters(a3)
>  #7 archetypes:
> a7 <- bestModel(lass15[[7]])
> parameters(a7)
> #Plotting the percentiles of each archetype:
> barplot(a3, smpulg, percentiles = T, which = "beside")
> barplot(a7, smpulg, percentiles = T, which = "beside")
```

We get the results in Section 3.3 related with PCA.

```
> pznueva <- prcomp(smpulg, scale = T, retx = T)
> summary(pznueva)
> #PCA scores for 3 archetypes:
```

```
> p3 <- predict(pznueva, parameters(a3))
> #PCA scores for 7 archetypes:
> p7 <- predict(pznueva, parameters(a7))
> #Representing the scores:
> xyplot.pca(p3[,1:2], pznueva$x[,1:2], data.col = gray(0.7), atypes.col =
1, atypes.pch = 15)
> xyplot.pca(p7[,1:2], pznueva$x[,1:2], data.col = gray(0.7), atypes.col =
1, atypes.pch = 15)
```

At last, we show how the empirical percentiles and the nearest individuals to each one of archetypes can be obtained.

```
> #Function for computing empirical percentiles:
> .perc <- function(x, data, digits = 0) {
   Fn <- ecdf(data)
   round(Fn(x) * 100, digits = digits)
 }
> #Percentiles for 3 archetypes:
> .perc(parameters(a3),smpulg)
> #Percentiles for 7 archetypes:
> .perc(parameters(a7), smpulg)
```

```
> #Which is the nearest individual to archetypes?.
> #Example for three archetypes:
> i = 3
> ai <- bestModel(lass15[[i]])
> ras <- rbind( parameters(ai),smpulg)
> dras <- dist(ras, method = "euclidean", diag =F, upper = T, p = 2)
> mdras <- as.matrix(dras)
> diag(mdras) = 1e+11
> wh1 <- which.min(mdras[1,]) - (i)
> min(mdras[1,])
> wh2 <- which.min(mdras[2,]) - (i)
> min(mdras[2,])
> wh3 <- which.min(mdras[3,]) - (i)
> min(mdras[3,])
> pa <- parameters(ai)
> dist(rbind(pa[1,],smpulg[wh1,]))
> smpulg[wh1,]
```

In addition, we can turn the standarized values to the original variables.

```
> parameters(a7)
> p <- parameters(a7)
> m <- mean(mpulg[,])
> s <- sd(mpulg[,])
> d <- p
> for(i in 1 : 6){
  d[,i] = p[,i] * s[i] + m[i]
 }
> t(d)
```

The code of function *xyplot.pca* is the following:

```
> xyplot.pca <- function(x, y, data.col = 1, data.pch = 19, data.bg = NULL,
  atypes.col = 2, atypes.pch = 19, ahull.show = F, ahull.col = atypes.col,
  chull = NULL, chull.col = gray(0.7), chull.pch = 19, adata.show = FALSE,
  adata.col = 3, adata.pch = 13, link.col = data.col, link.lty = 1, ...){

  zs <- x
  data <- y

  plot(data, col = data.col, pch = data.pch, bg = data.bg, ...)
  points(zs, col = atypes.col, pch = atypes.pch,  ...)

   if(!is.null(chull)){
    points(data[chull, ], col = chull.col, pch = chull.pch, ...)
    lines(data[c(chull, chull[1]), ], col = chull.col, ...)
   }

     if(ahull.show)
      lines(ahull(zs), col = ahull.col)
     if(adata.show){
      adata <- fitted(zs)
      link.col <- rep(link.col, length = nrow(adata))
      link.lty <- rep(link.lty, length = nrow(adata))
      points(adata, col = adata.col, pch = adata.pch, ...)
       for (i in seq_len(nrow(data)))
```

```
        lines(rbind(data[i, ], adata[i, ]), col = link.col[i],
              lty = link.lty[i], ...)
    }
    invisible(NULL)
}
```