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 powerful new 3D scanning techniques has enabled the generation of large up-to-date anthropometric databases which provide highly valued data to improve the ergonomic design of products adapted to the user population. As a consequence, Ergonomics and Anthropometry are two increasingly quantitative fields, so advanced statistical methodologies and modern software tools are required to get the maximum benefit from anthropometric data.

This paper presents a new R package, called **Anthropometry**, which is available on the Comprehensive R Archive Network. It brings together some statistical methodologies concerning clustering, statistical shape analysis, statistical archetypal analysis and the statistical concept of data depth, which have been especially developed to deal with anthropometric data. They are proposed with the aim of providing effective solutions to some common anthropometric problems, such as clothing design or workstation design (focusing on the particular case of aircraft cockpits). The utility of the package is shown by analyzing the anthropometric data obtained from a survey of the Spanish female population performed in 2006 and from the 1967 United States Air Force survey.

This manuscript is contained in **Anthropometry** as a vignette.

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

1. Introduction

Ergonomics is the science that investigates the interactions between human beings and the elements of a system. The application of ergonomic knowledge in multiple areas such as clothing and footwear design or both working and household environments is required to achieve the best possible match between the product and its users. To that end, it is fundamental to know the anthropometric dimensions of the target population. Anthropometry refers to the study of the measurements and dimensions of the human body and is considered a very important branch of Ergonomics because of its significant influence on the ergonomic design of products (Pheasant 2003).

A major issue when developing new patterns and products that fit the target population well is the lack of up-to-date anthropometric data. Improvements in health care, nutrition and living conditions and the transition to a sedentary life style have changed the body dimensions of people over recent decades. Anthropometric databases must therefore be updated regularly. Traditionally, human physical characteristics and measurements have been manually taken

using rudimentary methods like calipers, rulers or measuring tapes (Simmons and Istook 2003; Lu and Wang 2008; Shu, Wuhrer, and Xi 2011). These procedures are simple (user-friendly), non-invasive and no particularly expensive. However, measuring a statistically useful sample of thousands of people by hand is time-consuming and error-prone: the set of measurements obtained, and therefore the shape information, is usually imprecise and inaccurate.

In recent years, the development of new three-dimensional (3D) body scanner measurement systems has represented a huge step forward in the way anthropometric data are collected and updated. This technology provides highly detailed, accurate and reproducible anthropometric data from which 3D shape images of the people being measured can be obtained (Istook and Hwang 2001; Lerch, MacGillivray, and Domina 2007; Wang, Wu, Lin, Yang, and Lu 2007; D'Apuzzo 2009). The great potential of 3D body scanning techniques constitutes a true breakthrough in realistically characterizing people and they have made it possible to conduct new large-scale anthropometric surveys in different countries (for instance, in the USA, the UK, France, Germany and Australia). Within this context, the Spanish Ministry of Health sponsored a 3D anthropometric study of the Spanish female population in 2006 (Alemany, González, Nácher, Soriano, Arnáiz, and Heras 2010). A sample of 10,415 Spanish females from 12 to 70 years old, randomly selected from the official Postcode Address File, was measured. Associated software provided by the scanner manufacturers made a triangulation based on the 3D spatial location of a large number of points on the body surface. A 3D binary image of the trunk of each woman (white pixel if it belongs to the body, otherwise black) is produced from the collection of points located on the surface of each woman scanned, as explained in Ibáñez, Simó, Domingo, Durá, Ayala, Alemany, Vinué, and Solves (2012a). The two main goals of this study, which was conducted by the Biomechanics Institute of Valencia, were as follows: firstly, to characterize the morphology of females in Spain in order to develop a standard sizing system for the garment industry and, secondly, to encourage an image of healthy beauty in society by means of mannequins that are representative of the population. In order to tackle both these objectives, Statistics plays an essential role.

In every methodological and practical anthropometric problem, body size variability within the user population is characterized by means of a limited number of anthropometric cases. This is what is called *a user-centered design process*. An anthropometric case represents the set of body measurements the product evaluator plans to accommodate in design (HFES 300 Committee 2004). A case may be a particular human being or a combination of measurements. Depending on the features and needs of the product being designed, three types of cases can be distinguished: central, boundary and distributed. If the product being designed is a one-size product (one-size to accommodate people within a predetermined portion of the population), as may be the case in working environment design, the cases are selected on an accommodation boundary. However, if we focus on a multiple-size product (n sizes to fit n groups of people within a predetermined portion of the population), clothing design being the most apparent example, central cases are selected. The statistical methodologies that we have developed seek to define central and boundary cases to tackle the clothing sizing system design problem and the workplace design problem (focusing on the particular case of an aircraft cockpit).

Clothing sizing systems divide a population into homogeneous subgroups based on some key anthropometric dimensions (size groups), in such a way that all individuals in a size group can wear the same garment (Ashdown 2007; Chung, Lin, and Wang 2007). An efficient and optimal sizing system must accommodate as large a percentage of the population as possible, in as few sizes as possible, that best describes the shape variability of the population. In

addition, the garment fit for accommodated individuals must be as good as possible. Each clothing size is defined from a person who is near the center for the dimensions considered in the analysis. This central individual, which is considered as the size representative (the size prototype), becomes the basic pattern from which the clothing line in the same size is designed. Once a particular garment has been designed, fashion designers and clothing manufacturers hire fit models to test and assess the size specifications of their clothing before the production phase. Fit models have the appropriate body dimensions selected by each company to define the proportional relationships needed to achieve the fit the company has determined (Ashdown 2005; Workman and Lentz 2000; Workman 1991). Fit models are usually people with central measurements in each body dimension. The definition of an efficient sizing system depends to a large extent on the accuracy and representativeness of the fit models.

Clustering is the statistical tool that classifies a set of individuals into groups (clusters), in such a way that subjects in the same cluster are more similar (in some way) to each other than to those in other clusters (Kaufman, L. and Rousseeuw, P. J. 1990). In addition, clusters are represented by means of a representative central observation. Therefore, clustering comes up naturally as a useful statistical approach to try to define an efficient sizing system and to elicit prototypes and fit models. Specifically, five of the methodologies that we have developed are based on different clustering methods. Four of them are aimed at segmenting the population into optimal size groups and obtaining size prototypes. The first one, hereafter referred to as *trimowa*, has been published in Ibáñez, Vinué, Alemany, Simó, Epifanio, Domingo, and Ayala (2012b). It is based on using a especial distance function that mathematically captures the idea of garment fit. The second and third ones (called *CCbiclustAnthropo* and *TDDclust*) belong to a paper in progress (Vinué and Ibáñez 2014). The current version of this report can be accessed on the author's website, http://www.uv.es/vivigui/docs/biclustDepth. The *CCbiclustAnthropo* methodology adapts a particular clustering algorithm mostly used for the analysis of gene expression data to the field of Anthropometry. *TDDclust* uses the statistical concept of data depth (Liu, Parelius, and Singh 1999) to group observations according to the most central (deep) one in each cluster. As mentioned, traditional sizing systems are based on using a suitable set of key body dimensions, so clustering must be carried out in the Euclidean space. In the three previous procedures, we have always worked in this way. Instead, in the fourth and last one, hereinafter called as *kmeansProcrustes*, a clustering procedure is developed for grouping women according to their 3D body shape, represented by a configuration matrix of anatomical markers (landmarks). To that end, the statistical shape analysis (Dryden and Mardia 1998) will be fundamental. This approach has been accepted for publication (Vinué, Simó, and Alemany 2014c). The preprint version is available on the author's website, http://www.uv.es/vivigui/docs/kmeansProcADAC.pdf. Lastly, the fifth clustering proposal is presented with the goal of identifying accurate fit models and is again used in the Euclidean space. It is based on another clustering method originally developed for biological data analysis. This method, called *hipamAnthropom*, has been published in Vinué, León, Alemany, and Ayala (2014b). Well-defined fit models and prototypes can be used to develop representative and precise mannequins of the population.

A sizing system is intended only to cover what is known as the "standard" population, leaving out the individuals who might be considered outliers with respect to a set of measurements. In this case, outliers are called disaccommodated individuals. Clothing industries usually design garments for the standard sizes in order to optimize market share. The four aforementioned methods concerned with apparel sizing system design (*trimowa*, *CCbiclustAnthropo*, *TDDclust*

and *kmeansProcrustes*) take into account this fact. In addition, because *hipamAnthropom* is based on hierarchical features, it is capable of discovering and returning true outliers.

Unlike clothing design, where representative cases correspond to central individuals, in designing a one-size product, such as working environments or the passenger compartment of any vehicle, including aircraft cockpits, the most common approach is to search for boundary cases. In these situations, the variability of human shape is described by extreme individuals, which are those that have the smallest or largest values (or extreme combinations) in the dimensions considered in the study. These design problems fall into a more general category: the accommodation problem. The supposition is that the accommodation of boundaries will facilitate the accommodation of interior points (with less-extreme dimensions) (Bertilsson, Högberg, and Hanson 2012; Parkinson, Reed, Kokkolaras, and Papalambros 2006; HFES 300 Committee 2004). For instance, a garage entrance must be designed for a maximum case, while for reaching things such as a brake pedal, the individual minimum must be obtained. In order to tackle the accommodation problem, two methodological contributions based on statistical archetypal analysis are put forward. An archetype in Statistics is an extreme observation that is obtained as a convex combination of other subjects of the sample (Cutler and Breiman 1994). The first of these methodologies was published in Epifanio, Vinué, and Alemany (2013), and the second has been submitted for publication (Vinué, Epifanio, and Alemany 2014a). The preprint version is available on the author's website, http://www.uv.es/vivigui/docs/archetypoidsCSDAr1_DEF.pdf.

As far as we know, there is currently no reference in the literature related on Anthropometry or Ergonomics that provides the programming of the proposed algorithms. In addition, to the best of our knowledge, with the exception of modern human modelling tools like Jack and Ramsis, which are two of the most widely used tools by a broad range of industries (Blanchonette 2010), there are no other general software applications or statistical packages available on the Internet to tackle the definition of an efficient sizing system or the accommodation problem. Within this context, this paper introduces a new R package (R Development Core Team 2013) called **Anthropometry**, which brings together the algorithms associated with all the above-mentioned methodologies. All of them were applied to the anthropometric study of the Spanish female population and to the 1967 United States Air Force (USAF) survey. **Anthropometry** includes several data files related to both anthropometric databases. All the statistical methodologies, anthropometric databases and this R package were announced in the author's PhD thesis (Vinué 2014), which is freely available in a Spanish institutional open archive. The latest version of **Anthropometry** is always available from the Comprehensive R Archive Network at http://cran.r-project.org/package=Anthropometry.

The outline of the paper is as follows: Section 2 describes all the data files included in **Anthropometry**. Section 3 gives a brief explanation of each statistical technique developed and Section 4 presents how they are implemented in this package. In Section 5 some examples of their application are shown, pointing out at the same time the consequences of choosing different argument values. Section 6 is intended to guide users in their choice of the different methods presented. Finally, concluding remarks are given in Section 7.

2. Data

2.1. Spanish anthropometric survey

The Spanish National Institute of Consumer Affairs (INC according to its Spanish acronym), under the Spanish Ministry of Health and Consumer Affairs, commissioned a 3D anthropometric study of the Spanish female population in 2006, after signing a commitment with the most important Spanish companies in the apparel industry. The Spanish National Research Council (CSIC in Spanish) planned and developed the design of experiments, the Complutense University of Madrid was responsible for providing advice on Anthropometry and the study itself was conducted by the Biomechanics Institute of Valencia (Alemany *et al.* 2010). The target sample was made up of 10,415 women grouped into 10 age groups ranging from 12 to 70 years, randomly chosen from the official Postcode Address File.

As illustrative data of the whole Spanish survey, **Anthropometry** contains a database called `dataDemo`, made up of a sample of 600 Spanish women and their measurements for five anthropometric variables: bust, chest, waist and hip circumferences and neck to ground length. These variables are chosen for three main reasons: they are recommended by experts, they are commonly used in the literature and they appear in the European standard on sizing systems. Size designation of clothes. Part 2: Primary and secondary dimensions (European Committee for Standardization 2002).

This data set will be used by *trimowa*, *TDDclust* and *hipamAnthropom*. As mentioned above, the women's shape is represented by a set of landmarks, specifically 66 points. A data file called `landmarks` contains the configuration matrix of landmarks for each of the 600 women. The *kmeansProcrustes* methodology will need this data file.

As also noted above, a 3D binary image of each woman's trunk is available. Hence, the dissimilarity between trunk forms can be computed and a distance matrix between women can be built. The distance matrix used in Vinué *et al.* (2014a) is included in **Anthropometry** and is called `cMDSwomen`.

2.2. USAF survey

This database contains the information provided by the 1967 United States Air Force (USAF) survey. It can be downloaded from http://www.dtic.mil/dtic/. This survey was conducted in 1967 by the anthropology branch of the Aerospace Medical Research Laboratory (Ohio). A sample of 2420 subjects of the Air Force personnel, between 21 and 50 years of age, was measured at 17 Air Force bases across the United States of America. A total of 202 variables were collected. The dataset associated with the USAF survey is available on `dataUSAF`. In the methodologies related to archetypal analysis, six anthropometric variables from the total of 202 will be selected. They are the same as those selected in Zehner, Meindl, and Hudson (1993) and are called cockpit dimensions because they are critical in order for designing an aircraft cockpit.

2.3. Geometric figures

In the *kmeansProcrustes* approach, a numerical simulation with controlled data is performed to show the utility of our methodology. The controlled data are two geometric figures, a cube and a parallelepiped, made up of 8 and 34 landmarks. These configurations are saved in four files called `cube8`, `cube34`, `parallelepiped8` and `parallelepiped34`, respectively.

3. Statistical methodologies

In Section 3.1, the *trimowa*, *CCbiclustAnthropo*, *TDDclust* and *hipamAnthropom* methodologies are described. Section 3.2 focuses on the *kmeansProcrustes* methodology. Section 3.3 provides an explanation of the methodologies based on archetypal analysis.

For practical guidance, the method used for the clustering-based approaches is as follows: the data matrix is segmented using a primary control dimension (bust circumference in the case of *trimowa*, *hipamAnthropom*, *kmeansProcrustes* and *TDDclust*, and waist circumference in the case of *CCbiclustAnthropo*, according to the classes suggested in the European standard on sizing systems. Size designation of clothes. Part 3: Measurements and intervals (European Committee for Standardization 2005)). Then, a further segmentation is carried out using other secondary control anthropometric variables. In this way, the first segmentation provides a first easy input to choose the size, while the resulting clusters (subgroups) for each bust (or waist) and other anthropometric measurements optimize the sizing.

Regarding the methodologies using archetypal analysis, the steps are as follows: first, depending on the problem, the data may or may not be standardized. Then, an accommodation subsample is selected to obtain the archetypal individuals as the third and last step.

3.1. Anthropometric dimensions-based clustering

The *trimowa* methodology

The aim of a sizing system is to divide a varied population into groups using certain key body dimensions (Ashdown 2007; Chung *et al.* 2007). Three types of approaches can be distinguished for creating a sizing system: traditional step-wise sizing, multivariate methods and optimization methods. Traditional methods are not useful because they use bivariate distributions to define a sizing chart and do not consider the variability of other relevant anthropometric dimensions. Recently, more sophisticated statistical methods have been developed, especially using principal component analysis (PCA) and clustering (Gupta and Gangadhar 2004; Hsu 2009b; Luximon, Zhang, Luximon, and Xiao 2012; Hsu 2009a; Chung *et al.* 2007; Zheng, Yu, and Fan 2007; Bagherzadeh, Latifi, and Faramarzi 2010). Peter Tryfos was the first to suggest an optimization method (Tryfos 1986). Later, McCulloch et al. (McCulloch, Paal, and Ashdown 1998) modified Tryfos' approach.

The first clustering methodology proposed, called *trimowa*, is closed to the one developed in McCulloch *et al.* (1998). However, there are two main differences. First, when searching for k prototypes, a more statistical approach is assumed. To be specific, a trimmed version of the partitioning around medoids (PAM or k-medoids) clustering algorithm is used. The trimming procedure allows us to remove outlier observations (García-Escudero, Gordaliza, Matrán, and Mayo-Iscar 2008; García-Escudero, Gordaliza, and Matrán 2003). Second, the dissimilarity measure defined in McCulloch *et al.* (1998) is modified using an OWA (ordered weighted average) operator to consider the user morphology. Our approach allows us to obtain more realistic prototypes (medoids) because they correspond to real women from the database and the selection of individual discommodities. In addition, the use of OWA operators has resulted in a more realistic dissimilarity measure between individuals and prototypes. This approach was published in Ibáñez *et al.* (2012b) and it is implemented in the `trimowa` function.

We learned from this situation that there is an ongoing search for advanced statistical approaches that can deliver practical solutions to the definition of central people and optimal

size groups. Consequently, we have come across two different statistical strategies in the literature and have aimed to discuss their potential usefulness in the definition of an efficient clothing sizing system. These approaches are based on biclustering and data depth and will be summarized below.

The *CCbiclustAnthropo* methodology

In the analysis of gene expression data, conventional clustering is limited to finding local expression patterns. Gene data are organized in a data matrix where rows correspond to genes and columns to experimental samples (conditions). The goal is to find submatrices, i.e., subgroups of genes and subgroups of conditions, where the genes exhibit a high degree of correlation for every condition (Madeira and Oliveira 2004). Biclustering is a novel clustering approach that accomplishes this goal. This technique consists of simultaneously partitioning the set of rows and the set of columns into subsets.

In a traditional row cluster, each row is defined using all the columns of the data matrix. Something similar would occur with a column cluster. However, with biclustering, each row in a bicluster is defined using only a subset of columns and vice versa. Therefore, clustering provides a global model but biclustering defines a local one. This interesting property made us think that biclustering could perhaps be useful for obtaining efficient size groups, since they would only be defined for the most relevant anthropometric dimensions that describe a body in the detail necessary to design a well-fitting garment.

Recently, a large number of biclustering methods have been developed. Some of them are implemented in different sources, including R. Currently, the most complete R package for biclustering is **biclust** (Kaiser and Leisch 2008; Kaiser, Santamaria, Khamiakova, Sill, Theron, Quintales, and Leisch 2013). The usefulness of the approaches included in **biclust** for dealing with anthropometric data was investigated in Vinué (2012). Among the conclusions reached, the most important was concerned with the possibility of considering the Cheng & Church biclustering algorithm (Cheng and Church 2000) (referred to below as CC) as a potential statistical approach to be used for defining size groups. Specifically, in Vinué (2012) an algorithm to find size groups (biclusters) and disaccommodated women with CC was set out. This methodology is called *CCbiclustAnthropo* and it is implemented in the `CCbiclustAnthropo` function.

Designing lower body garments depends not only on the waist circumference (the principal dimension in this case), but also on other secondary control dimensions (for upper body garments only the bust circumference is usually needed). Biclustering produces subgroups of objects that are similar in one subgroup of variables and different in the remaining variables. Therefore, it seems more interesting to use a biclustering algorithm with a set of lower body dimensions. For that purpose, all the body variables related to the lower body in the Spanish anthropometric survey were chosen (there were 36). An efficient partition into different biclusters was obtained with promising results. All individuals in the same bicluster can wear a garment designed for the specific body dimensions (waist and other variables) which were the most relevant for defining the group. Each group is represented by the median woman. The CC algorithm is nonexhaustive, i.e., some rows (and columns) do not belong to any bicluster. This property can be used to fix a proportion of non-accommodated sample.

The main interest of this approach was descriptive and exploratory and the important point to note here is that `CCbiclustAnthropo` cannot be used with `dataDemo`, since this data file

does not contain variables related to the lower body in addition to waist and hip. However, this function is included in the package in the hope that it could be helpful or useful for other researchers. All theoretical and practical details are given in Vinué and Ibáñez (2014), Vinué (2014) and Vinué (2012).

The *TDDclust* methodology

The statistical concept of data depth is another general framework for descriptive and inferential analysis of numerical data in a certain number of dimensions. In essence, the notion of data depth is a generalization of standard univariate rank methods in higher dimensions. A depth function measures the degree of centrality of a point regarding a probability distribution or a data set. The highest depth values correspond to central points and the lowest depth values correspond to tail points (Liu *et al.* 1999; Zuo and Serfling 2000). Therefore, the depth paradigm is another very interesting strategy for identifying central prototypes.

The development of clustering and classification methods using data depth measures has received increasing attention in recent years (Dutta and Ghosh 2012; Lange, Mosler, and Mozharovskyi 2012; López and Romo 2010; Ding, Dang, Peng, and Wilkins 2007). The most relevant contribution to this field has been made by Rebecka Jörnsten in Jörnsten (2004) (see Jörnsten, Vardi, and Zhang 2002; Pan, Jörnsten, and Hart 2004, for more details). She introduced two clustering and classification methods (*DDclust* and *DDclass*, respectively) based on L_1 data depth (see Vardi and Zhang (2000)). The *DDclust* method is proposed to solve the problem of minimizing the sum of L_1-distances from the observations to the nearest cluster representatives. The L_1 data depth is the amount of probability mass needed at a point z to make z the multivariate L_1-median (a robust representative) of the data cluster.

An extension of *DDclust* is introduced which incorporates a trimmed procedure, aimed at segmenting the data into efficient size groups using central (the deepest) people. This methodology will be referred to below as *TDDclust* and it can be used within **Anthropometry** by using a function with the same name. All details about *TDDclust* are described in Vinué and Ibáñez (2014) and Vinué (2014).

The *HipamAnthropom* methodology

Representative fit models are important for defining a meaningful sizing system. However, there is no agreement among apparel manufacturers and almost every company employs a different fit model. Companies try to improve the quality of garment fit by scanning their fit models and deriving dress forms from the scans (Ashdown 2007; Song and Ashdown 2010). A fit model's measurements correspond to the commercial specifications established by each company to achieve the company's fit (Loker, Ashdown, and Schoenfelder 2005; Workman and Lentz 2000; Workman 1991). Beyond merely wearing the garment for inspection, a fit model provides objective feedback about the fit, movement or comfort of a garment in place of the consumer.

The *hipamAnthropom* methodology is proposed in order to provide new insights about this problem. It consists of two classification algorithms based on the hierarchical partitioning around medoids (HIPAM) clustering method presented in Wit and McClure (2004), which has been modified to deal with anthropometric data. This procedure was published in Vinué *et al.* (2014b). The dissimilarity measure defined in McCulloch *et al.* (1998) and a different method for obtaining a classification tree (Irigoien and Arenas 2008) were incorporated. One

algorithm was called $HIPAM_{MO}$ and the other, $HIPAM_{IMO}$. The outputs of both include a set of central representative subjects or medoids taken from the original data set, which constitute our fit models. They can also detect outliers. This methodology is available in the `hipamAnthropom` function.

3.2. Statistical shape analysis

The *kmeansProcrustes* methodology

The clustering methodologies explained in Section 3.1 use a set of control anthropometric variables as the basis for a different type of sizing system in which people are grouped in a size group based on a full range of measurements. Consequently, clustering is done in the Euclidean space. The shape of the women recruited into the Spanish anthropometric survey is represented by a set of correspondence points called landmarks. Taking advantage of this fact, we have adapted the k-means clustering algorithm to the field of statistical shape analysis, to define size groups of women according to their body shapes. The representative of each size group is the average woman. This approach has been accepted for publication (Vinué *et al.* 2014c). We have adapted both the Hartigan-Wong (H-W) and original Lloyd versions of k-means to the field of shape analysis and we have demonstrated, by means of a simulation study, that the Lloyd version is more efficient for clustering shapes than the H-W version.

The function that uses the Lloyd version of k-means adapted to shape analysis (what we called *kmeansProcrustes*) is `LloydShapes`. The function that uses the H-W version of k-means adapted to shape analysis is `HartiganShapes`. A trimmed version of *kmeansProcrustes* can be also executed with `trimmedLloydShapes`.

3.3. Archetypal analysis

In ergonomic-related problems, where the goal is to create more efficient people-machine interfaces, a small set of extreme cases (boundary cases), called human models, is sought. Designing for extreme individuals is appropriate where some limiting factor can define either a minimum or maximum value which will accommodate the population. The basic principle is that accommodating boundary cases will be sufficient to accommodate the whole population.

For too long, the conventional solution for selecting this small group of boundary models was based on the use of percentiles. However, percentiles are a kind of univariate descriptive statistic, so they are suitable only for univariate accommodation and should not be used in designs that involve two or more dimensions. Furthermore, they are not additive (Zehner *et al.* 1993; Robinette and McConville 1981; Moroney and Smith 1972). Today, the alternative commonly used for the multivariate accommodation problem is based on PCA (Friess and Bradtmiller 2003; Hudson, Zehner, and Meindl 1998; Robinson, Robinette, and Zehner 1992; Bittner, Glenn, Harris, Iavecchia, and Wherry 1987). However, it is known that the PCA approach presents some drawbacks (Friess 2005). In Epifanio *et al.* (2013), a different statistical approach for determining multivariate limits was put forward: archetypal analysis (Cutler and Breiman 1994), and its advantages regarding over PCA were demonstrated. The function that allows us to reproduce the results discussed in Epifanio *et al.* (2013) is `archetypesBoundary`.

Archetypes computed by archetypal analysis are a convex combination of the sampled individuals, but they are not necessarily real observations. In some problems, it is crucial that

the archetypes are real subjects, observations of the sample, and not fictitious. To that end, we have proposed a new archetypal concept: the archetypoid, which corresponds to specific individuals and each observation of the data set can be represented as a mixture of these archetypoids (Vinué *et al.* 2014a). We have developed an efficient computational algorithm based on PAM to compute archetypoids (called archetypoid algorithm), we have analyzed some of their theoretical properties, we have explained how they can be obtained when only dissimilarities between observations are known (features are unavailable) and we have demonstrated some of their advantages regarding over classical archetypes. The `stepArchetypoids` function calls the `archetypoids` function to run the archetypoid algorithm repeatedly.

The archetypoid algorithm has two phases: a BUILD phase and a SWAP phase, like PAM. In the BUILD step, an initial set of archetypoids is determined, made up of the nearest individuals to the archetypes returned by the **archetypes** R package. This set can be defined in two different ways: on the one hand, as mentioned in Epifanio *et al.* (2013) (set *nearest*) and on the other hand, as used in Eugster (2012) and Seiler and Wohlrabe (2013) (set *which*). Accordingly, the initial set of archetypoids is either *nearest* or *which*. The aim of the SWAP phase of the archetypoid algorithm is the same as that of the SWAP phase of PAM, but the objective function changes (see Vinué *et al.* (2014a); Vinué (2014)).

4. The Anthropometry R package

In this section we will look more closely at the package functions associated with each of the methodologies introduced in Section 3.

4.1. Anthropometric dimensions-based clustering

A key element of two of the aforementioned clustering methodologies -*trimowa* and *hipamAnthropom*- is the global dissimilarity function used. It is the same dissimilarity measure proposed in McCulloch *et al.* (1998) but incorporates a set of OWA weights to highlight high dissimilarities. The global dissimilarity described in McCulloch *et al.* (1998) is defined as a sum of squared discrepancies over each of the p anthropometric measurements considered. In this way, the different discrepancies are aggregated and an OWA operator can be used. An OWA operator allows us to adjust the importance of each one of the p discrepancies by assigning a particular weight to each of them. The largest discrepancy is assigned the largest weight, the second largest discrepancy is assigned the second largest weight and so on for the p variables. Because the OWA operators are bounded between the max and min operators, a measure called orness is needed. See Vinué (2014, p. 22-24) for a detailed explanation of the OWA operators.

The code for computing the global dissimilarity with **Anthropometry** is written in C and is exported from the NAMESPACE file. *Trimowa* and *hipamAnthropom* incorporate the calculus of the dissimilarity matrix within their main functions (`trimowa` and `hipamAnthropom`, respectively).

We will now give a comprehensive description of the arguments of the `trimowa`, `TDDclust` and `hipamAnthropom` functions. The `CCbiclustAnthropo` function is not detailed because it

will not be used in Section 5.

The `trimowa` function

```
trimowa(x, w, K, alpha, niter, Ksteps, ahVect = c(23, 28, 20, 25, 25))
```

Its arguments are as follows:

- x: Data frame. In our approach, this is each of the subframes originated after segmenting the whole anthropometric Spanish survey into twelve bust segments, according to the European standard on sizing systems. Size designation of clothes. Part 3: Measurements and intervals. Each row corresponds to an observation, and each column corresponds to a variable. All variables are numeric.
- w: The aggregation weights of the OWA operator. They are computed with the `WeightsMixtureUB` function.
- K: Number of clusters.
- alpha: Proportion of trimmed sample.
- niter: Number of random initializations.
- Ksteps: Steps per initialization.
- ahVect: Constants that define the *ah* slopes of the distance function in `GetDistMatrix`. Given the five variables considered, this vector is c(23,28,20,25,25). This vector would be different according to the variables considered.

The `TDDclust` function

```
TDDclust(x, K, lambda, Th, A, T0, alpha, Trimm, data1)
```

Its arguments are as follows:

- x: Data frame. Each row corresponds to an observation, and each column corresponds to a variable. All variables must be numeric.
- K: Number of clusters.
- lambda: Tuning parameter that controls the influence the data depth has over the clustering, see Jörnsten (2004).
- Th: Threshold for observations to be relocated, usually set to 0.
- A: Number of iterations.
- T0: Simulated annealing parameter. It is the current temperature in the simulated annealing procedure.

- alpha: Simulated annealing parameter. It is the decay rate, default 0.9.
- Trimm: Proportion of non-accommodated sample.
- data1: The same data frame as x, used to incorporate the trimmed observations into the rest of them for the next iteration.

The `hipamAnthropom` function

```
hipamAnthropom(x, asw.tol = 0, maxsplit = 5, local.const = NULL,
               orness = 0.7, type, ahVect = c(23, 28, 20, 25, 25), ...)
```

Its arguments are as follows:

- x: Data frame. In our approach, this is each of the subframes originated after segmenting the whole anthropometric Spanish survey into twelve bust segments, according to the European standard on sizing systems. Size designation of clothes. Part 3: Measurements and intervals. Each row corresponds to an observation, and each column corresponds to a variable. All variables are numeric.
- asw.tol: If this value is given, a tolerance or penalty can be introduced (asw.tol > 0 or asw.tol < 0, respectively) in the branch splitting procedure. Default value (0) is maintained. See Wit and McClure (2004, p.154) for more details.
- maxsplit: The maximum number of clusters that any cluster can be divided into when searching for the best clustering.
- local.const: If this value is given (meaningful values are those between -1 and 1), a proposed partition is accepted only if the associated asw is greater than this constant. Default option for this argument is maintained, that is to say, this value is ignored. See Wit and McClure (2004, p.154) for more details.
- orness: Quantity to measure the degree to which the aggregation is like a min or max operation. See `WeightsMixtureUB` and `GetDistMatrix`.
- type: Type of HIPAM algorithm to be used. The possible options are 'MO' (for $HIPAM_{MO}$) and 'IMO' (for $HIPAM_{IMO}$).
- ahVect: Constants that define the ah slopes of the distance function in `GetDistMatrix`. Given the five variables considered, this vector is c(23,28,20,25,25). This vector would be different according to the variables considered.
- ...: Other arguments that may be supplied to the internal functions of the HIPAM algorithms.

4.2. Statistical shape analysis

The `LloydShapes`, `HartiganShapes` and `trimmedLloydShapes` functions are examined.

The `LloydShapes` function

```
LloydShapes(dg, K, Nsteps = 10, niter = 10, stopCr = 0.0001, simul, print)
```

Its arguments are as follows:

- dg: Array with the 3D landmarks of the sample objects. Each row corresponds to an observation, and each column corresponds to a dimension (x,y,z).
- K: Number of clusters.
- Nsteps: Number of steps per initialization. Default value is 10.
- niter: Number of random initializations. Default value is 10.
- stopCr: Relative stopping criteria. Default value is 0.0001.
- simul: Logical value. If TRUE, this function is used for a simulation study.
- print: Logical value. If TRUE, some messages associated with the running process are displayed.

The `HartiganShapes` function

```
HartiganShapes(dg, K, Nsteps = 10, niter = 10, stopCr = 0.0001, simul,
               initLl, initials, print)
```

Its arguments are as follows:

- dg: Array with the 3D landmarks of the sample objects. Each row corresponds to an observation, and each column corresponds to a dimension (x,y,z).
- K: Number of clusters.
- Nsteps: Number of steps per initialization. Default value is 10.
- niter: Number of random initializations. Default value is 10.
- stopCr: Relative stopping criteria. Default value is 0.0001.
- simul: Logical value. If TRUE, this function is used for a simulation study.
- initLl: Logical value. If TRUE, see next argument *initials*. If FALSE, they are new random initial values.
- initials: If `initLl`=TRUE, they are the same random initial values used in each iteration of `LloydShapes`. If `initLl`=FALSE this argument must be passed simply as an empty vector.
- print: Logical value. If TRUE, some messages associated with the running process are displayed.

The `trimmedLloydShapes` function

```
trimmedLloydShapes(dg, n, alpha, K, Nsteps = 10, niter = 10,
                   stopCr = 0.0001, print)
```

Its arguments are as follows:

- dg: Array with the 3D landmarks of the sample objects. Each row corresponds to an observation, and each column corresponds to a dimension (x,y,z).
- n: Number of individuals.
- alpha: Proportion of trimmed sample.
- K: Number of clusters.
- Nsteps: Number of steps per initialization. Default value is 10.
- niter: Number of random initializations. Default value is 10.
- stopCr: Relative stopping criteria. Default value is 0.0001.
- print: Logical value. If TRUE, some messages associated with the running process are displayed.

4.3. Archetypal analysis

Finally, this section provides a detailed explanation of the `archetypesBoundary`, `archetypoids`, `stepArchetypoids` and `stepArchetypesMod` functions.

The `archetypesBoundary` function

```
archetypesBoundary(data, numArchet, verbose, nrep)
```

Its arguments are as follows:

- data: USAF 1967 database (see `dataUSAF`). Each row corresponds to an observation, and each column corresponds to a variable. All variables are numeric.
- numArchet: Number of archetypes.
- verbose: Logical value. If TRUE, some details of the execution progress are shown (this is the same argument as that of the `stepArchetypes` function of the **archetypes** R package (Eugster and Leisch 2009)).
- nrep: For each archetype run `archetypes` `nrep` times (this is the same argument as that of the `stepArchetypes` function of **archetypes**).

The `archetypoids` function

```
archetypoids(i, data, huge = 200, step, init, ArchObj, nearest, sequ, aux)
```

Its arguments are as follows:

- i: Number of archetypoids.
- data: Data matrix. Each row corresponds to an observation and each column corresponds to an anthropometric variable. All variables are numeric.
- huge: This is a penalization added to solve the convex least squares problems regarding the minimization problem to estimate archetypoids, see Eugster and Leisch (2009). Default value is 200.
- step: Logical value. If TRUE, the archetypoid algorithm is executed repeatedly within `stepArchetypoids`. Therefore, this function requires the next argument `init` (but neither the `ArchObj` nor the `nearest` arguments) that specifies the initial vector of archetypoids, which has already been computed within `stepArchetypoids`. If FALSE, the archetypoid algorithm is executed once. In this case, the `ArchObj` and `nearest` arguments are required to compute the initial vector of archetypoids.
- init: Initial vector of archetypoids for the BUILD phase of the archetypoid algorithm. It is computed within `stepArchetypoids`. See `nearest` argument below for an explanation of how this vector is calculated.
- ArchObj: The list returned by the `stepArchetypesMod` function. This function is a slight modification of the original `stepArchetypes` function of **archetypes** to apply the archetype algorithm to raw data. The `stepArchetypes` function standardizes the data by default and this option is not always desired. This list is needed to compute the nearest individuals to archetypes. Required when `step`=FALSE.
- nearest: Initial vector of archetypoids for the BUILD phase of the archetypoid algorithm. Required when `step`=FALSE. This argument is a logical value: if TRUE (FALSE), the *nearest* (*which*) vector is calculated. Both vectors contain the nearest individuals to the archetypes returned by the `archetypes` function of **archetypes** (In Vinué *et al.* (2014a), archetypes are computed after running the archetype algorithm twenty times). The *nearest* vector is calculated by computing the Euclidean distance between the archetypes and the individuals and choosing the nearest. It is used in Epifanio *et al.* (2013). The *which* vector is calculated by consecutively identifying the individual with the maximum value of alpha for each archetype, until the defined number of archetypes is reached. It is used in Eugster (2012).
- sequ: Logical value. It indicates whether a sequence of archetypoids (TRUE) or only a single number of them (FALSE) is computed. It is determined by the number of archetypes computed by means of `stepArchetypesMod`.

- aux: If `sequ`=FALSE, this value is equal to `i`-1 since for a single number of archetypoids, the list associated with the archetype object only has one element.

The `stepArchetypoids` function

```
stepArchetypoids(i, nearest, data, ArchObj)
```

Its arguments are as follows:

- i: Number of archetypoids.
- nearest: Initial vector of archetypoids for the BUILD phase of the archetypoid algorithm. This argument is a logical value: if TRUE (FALSE), the *nearest* (*which*) vector is calculated. Both vectors contain the nearest individuals to the archetypes returned by the `archetypes` function of **archetypes** (In Vinué *et al.* (2014a), archetypes are computed after running the archetype algorithm twenty times). The *nearest* vector is calculated by computing the Euclidean distance between the archetypes and the individuals and choosing the nearest. It is used in Epifanio *et al.* (2013). The *which* vector is calculated by consecutively identifying the individual with the maximum value of alpha for each archetype, until the defined number of archetypes is reached. It is used in Eugster (2012).
- data: Data matrix. Each row corresponds to an observation and each column corresponds to an anthropometric variable. All variables are numeric.
- ArchObj: The list returned by the `stepArchetypesMod` function. This function is a slight modification of the original `stepArchetypes` function of **archetypes** to apply the archetype algorithm to raw data. The `stepArchetypes` function standardizes the data by default and this option is not always desired. This list is needed to compute the nearest individuals to archetypes.

The `stepArchetypesMod` function

```
stepArchetypesMod(data, k, nrep = 3, verbose = TRUE)
```

Its arguments are as follows:

- data: Data to obtain archetypes.
- k: Number of archetypes to compute, from 1 to `k`.
- nrep: For each `k`, run `archetypes` `nrep` times.
- verbose: If TRUE, the progress during execution is shown.

5. Examples

This section presents a detailed explanation of the numerical and graphical outcome provided by each method by means of several examples. In addition, some relevant comments are given about the consequences of choosing different argument values in each case.

First of all, **Anthropometry** must be loaded into R:

```
library("Anthropometry")
```

5.1. Anthropometric dimensions-based clustering

The following code executes the *trimowa* methodology. A similar code was used to obtain the results described in Ibáñez *et al.* (2012b). We use `dataDemo` and its five anthropometric variables. The bust circumference is used as the primary control dimension. Twelve bust sizes (from 74 cm to 131 cm) are defined according to the European standard on sizing systems. Size designation of clothes. Part 3: Measurements and intervals (European Committee for Standardization 2005)).

```
dataDef <- dataDemo
num.variables <- dim(dataDef)[2]
bust <- dataDef$bust
bustCirc_4 <- seq(74, 102, 4)
bustCirc_6 <- seq(107, 131, 6)
bustCirc <- c(bustCirc_4, bustCirc_6)
nsizes <- length(bustCirc)
```

The aggregation weights of the OWA operator are computed. They are used to calculate the global dissimilarity between the individuals and the prototypes. We give orness a value of 0.7 in order to highlight the largest aggregated values, that is to say, the largest discrepancies between the women's body measurements and those of the prototype. An orness value close to 1 gives more importance to the worst fit, whilst an orness value close to 0 gives more importance to the best fit (see Vinué (2014, p. 27-31) for details).

```
orness <- 0.7
w <- WeightsMixtureUB(orness, num.variables)
```

Next the `trimowa` algorithm is used within each bust class. Three size groups (clusters, argument `K`) are calculated per bust segment. This number of groups is quite well aligned with the strategy used by companies to design sizes. A larger `K` will result in many sizes being designed, increasing the production a lot. A smaller `K` corresponds to too few sizes being designed and having a poor accommodation index.

The trimmed proportion, `alpha`, is prefixed to 0.01 per segment (therefore, the accommodation rate in each bust size will be 99%). This selection allows us to accommodate a very large percentage of the population in the sizing system. A larger trimmed proportion would result in a smaller amount of accommodated people. The number of random initializations is 10 (`niter`), with seven steps per initialization (`Ksteps`). These values are small in the interests

of a fast execution. The more random repetitions, the more accurate the prototypes and the more representative of the size group. In Ibáñez *et al.* (2012b), the number of random initializations was 600.

In addition, a vector of five constants (one per variable) is needed to define the dissimilarity. The numbers collected in the `ahVect` argument are related to the particular five variables selected in `dataDemo`. Different body variables would require different constants (see McCulloch *et al.* 1998; Vinué 2014, for further details).

To reproduce results, a seed for randomness is fixed.

```
K <- 3 ; alpha <- 0.01 ; niter <- 10 ; Ksteps <- 7
ahVect <- c(23, 28, 20, 25, 25)

set.seed(2014)
res_trimowa <- list()
for (i in 1 : (nsizes - 1)){
 data = dataDef[(bust >= bustCirc[i]) & (bust < bustCirc[i + 1]), ]
 res_trimowa[[i]] <- trimowa(data, w, K, alpha, niter,
                             Ksteps, ahVect = ahVect)
}
```

The prototypes are the clustering medoids.

```
medoids <- list()
for (i in 1 : (nsizes - 1)){
medoids[[i]] <- res_trimowa[[i]]$meds
}
```

Figure 1 shows the scatter plots of bust circumference against neck to ground with the three medoids obtained for each bust class without (left) and with (right) the prototypes defined by the European standard. The medoids color and the plot title must be provided.

```
bustVariable <- "bust"
xlim <- c(70, 150)
color <- c("black", "red", "green", "blue", "cyan", "brown", "gray",
           "deeppink3", "orange", "springgreen4", "khaki3", "steelblue1")

variable <- "necktoground"
ylim <- c(110, 160)
title <- "Medoids \n bust vs neck to ground"

plotMedoids(dataDef, medoids, nsizes, bustVariable, variable, color,
            xlim, ylim, title, FALSE)
plotMedoids(dataDef, medoids, nsizes, bustVariable, variable, color,
            xlim, ylim, title, TRUE)
```

The following sentences illustrate how to use the *hipamAnthropom* methodology. The same twelve bust segments as in `trimowa` are used.

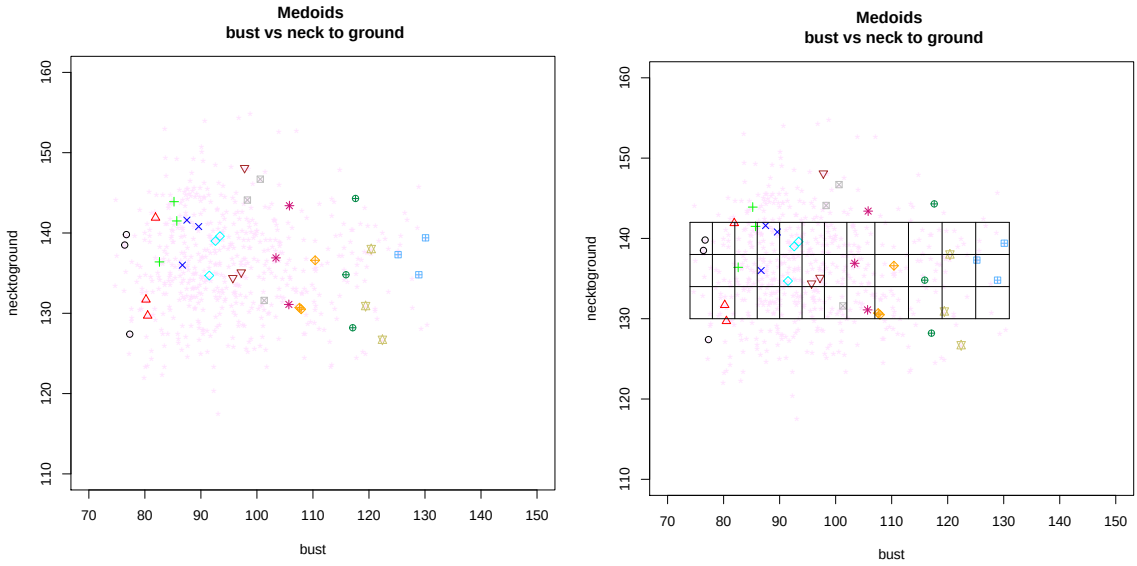

Figure 1: Bust vs. neck to ground, jointly with our medoids (left) and the prototypes defined by the European standard (right).

```
dataDef <- dataDemo
bust <- dataDef$bust
bustCirc_4 <- seq(74, 102, 4)
bustCirc_6 <- seq(107, 131, 6)
bustCirc <- c(bustCirc_4, bustCirc_6)
nsizes <- length(bustCirc)
```

The $HIPAM_{IMO}$ algorithm is used. It was verified in Vinué *et al.* (2014b) that $HIPAM_{IMO}$ showed better performance for finding representative prototypes. The maximum number of clusters that any cluster can be divided into is fixed to five (`maxsplit`). In the HIPAM algorithm the number of sub-clusters that any cluster is potentially divided into is between 2 and `maxsplit`. A larger `maxsplit` than five could result in too many clusters, which is not interesting from the point of view of the strategy used by companies to design sizes.

The same orness and vector of constants as in `trimowa` are used. To reproduce results, a seed for randomness is fixed.

```
type <- "IMO"
maxsplit <- 5 ; orness <- 0.7
ahVect <- c(23, 28, 20, 25, 25)

set.seed(2013)
hip <- list()
for(i in 1 : (nsizes - 1)){
data =  dataDef[(bust >= bustCirc[i]) & (bust < bustCirc[i + 1]), ]
d <- as.matrix(data)
```

```
hip[[i]] <- hipamAnthropom(d, maxsplit = maxsplit, orness = orness,
                           type = type, ahVect = ahVect)
}
```

The `hipamBigGroups` function is a function of **Anthropometry** that returns the medoids of the clusters with more than 2 elements. These medoids constitute our fit models. The `outlierHipam` function is another function of **Anthropometry** that returns the individuals of the clusters with 1 or 2 elements (outliers).

```
list.meds <- lapply(1:(nsizes - 1), FUN = hipamBigGroups, hip)
list_outl1_2 <- sapply(1 : (nsizes - 1), FUN = outlierHipam, hip)
```

Figure 2 displays the medoids (left) and the outlier women (right) corresponding to each bust size. The medoids color and the plot title must be provided. The important point to note here is the fact that each bust segment has a small sample size. This might explain the fact that this algorithm (and also $HIPAM_{MO}$) does not find large homogeneous clusters and therefore identifies a lot of women as outliers in each class for this database. One of the features of the HIPAM algorithm is that it is a very sensitive algorithm for identifying outliers. A broad discussion, analysis and thoughts on the anthropometric meaning of these outliers is given in Vinué *et al.* (2014b) (including the supplementary material).

```
bustVariable <- "bust"
xlim <- c(70, 150)
color <- c("black", "red", "green", "blue", "cyan", "brown", "gray",
           "deeppink3", "orange", "springgreen4", "khaki3", "steelblue1")

variable <- "hip"
ylim <- c(80, 160)
title <- "Medoids HIPAM_IMO \n bust vs hip"
title_outl <- "Outlier women HIPAM_IMO \n bust vs hip"

plotMedoids(dataDef, list.meds, nsizes, bustVariable, variable, color,
            xlim, ylim, title, FALSE)
plotTrimmOutl(dataDef, list_outl1_2, nsizes, bustVariable, variable, color,
              xlim, ylim, title_outl)
```

To conclude this section a basic example of the *TDDclust* methodology is shown. Computing data depth is very demanding. As an illustration, only 25 individuals are selected. In addition, the neck to ground, waist and bust variables are selected.

```
dataDef <- dataDemo[1 : 25, c(2, 3, 5)]
data1 <- dataDemo[1 : 25, c(2, 3, 5)]
```

In line with `trimowa`, three size groups are calculated (`K`) and a trimmed proportion is fixed to 0.01 (`percTrimm`). The `lambda` controls the influence the data depth has over the clustering. If `lambda` is 1, the clustering criterion is equivalent to the average silhouette width. On the

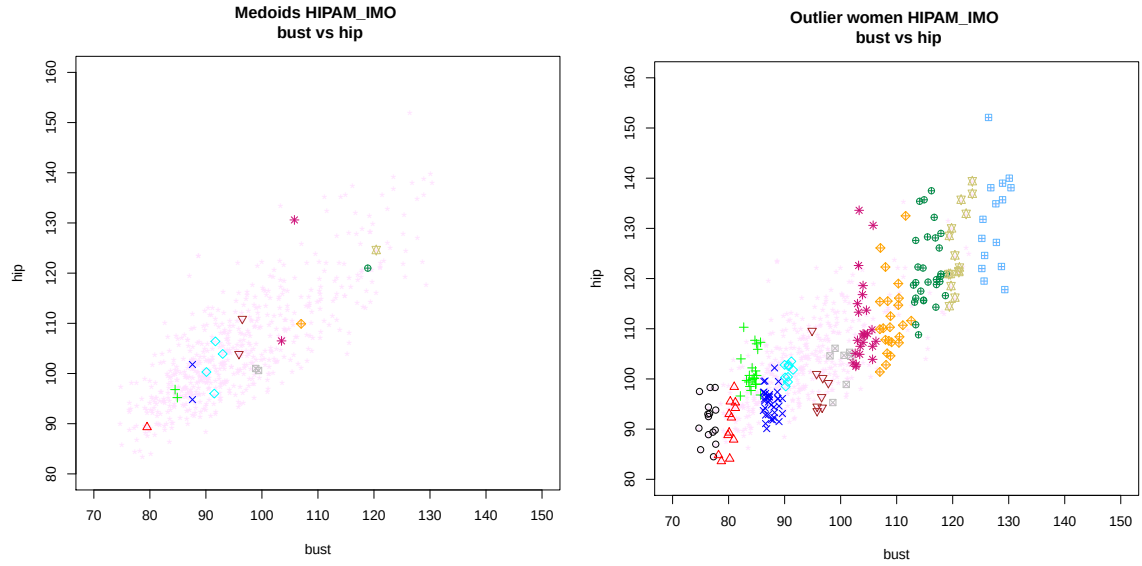

Figure 2: Bust vs. hip with the medoids (left) and with the outliers (right) obtained using $HIPAM_{IMO}$.

contrary, if `lambda` is 0, it is given by the average relative data depth. Fixing `lambda` to 0.5 is an intermediate and suggested scenario. A detailed explanation of the consequences of different `lambda` values is given in Jörnsten (2004). Because the depth computation is costly, we only run the algorithm for five iterations (`A`).

The other arguments are given by default. A different value for `Th` may result in the optimum clustering not being found (see Jörnsten 2004, page 75). A different simulated annealing parameter (`T0` and `alpha`) may change the clustering results obtained. To reproduce results, a seed for randomness is fixed.

```
K <- 3 ; percTrimm <- 0.01 ; lambda <- 0.5 ; A <- 5
Th <- 0 ; T0 <- 0 ; alpha <- 0.9

set.seed(2014)
Dout <- TDDclust(x = dataDef, K = K, lambda = lambda, Th = Th, A = A,
                 T0 = T0, alpha = alpha, Trimm = percTrimm, data1 = data1)
```

The following sentences allow us to analyze the clustering results, the final value of the optimal partition, the iteration in which the optimal partition was found and the trimmed observations, respectively.

```
table(Dout$NN[1,])
Dout$Cost
Dout$klBest
Dout$indivTrimmed
```

5.2. Statistical shape analysis

In this section, the use of the *kmeansProcrustes* methodology is illustrated. For the sake of simplicity of the computation involved only a small sample (the first 50 individuals) is selected. When there are missing values, they are removed.

```
landmarks1 <- na.exclude(landmarks)
num.points <- (dim(landmarks1)[2]) / 3
landmarks2 <- landmarks1[1 : 50, ]
n <- dim(landmarks2)[1]
```

We have to define an array with the 3D landmarks of the sample objects.

```
dg <- array(0, dim = c(num.points, 3, n))
for(k in 1 : n){
 for(l in 1 : 3){
  dg[, l, k] <- as.matrix(as.vector(landmarks2[k, ][seq(l, dim(landmarks2)[2]
                                                        + (l - 1), by = 3)]),
                          ncol = 1, byrow = T)
 }
}
```

Again, three size groups are calculated (`K`) and a trimmed proportion is fixed to 0.01 (`alpha`).

The `trimmedLloydShapes` algorithm is used with only five iterations and five steps per initialization in the interests of a fast execution. A larger number of repetitions is suggested to obtain more optimal results. The default relative stopping criteria is 0.0001. Using this small value ensures that the algorithm stops when the decrease in the objective function is hardly visible. A larger stopping value could prematurely stop the algorithm (but the decrease in the objective function should have been taken into account).

To reproduce results, a seed for randomness is fixed.

```
K <- 3 ; alpha <- 0.01 ; Nsteps <- 5 ; niter <- 5 ; stopCr <- 0.0001
set.seed(2013)
res <- trimmedLloydShapes(dg, n, alpha, K, Nsteps, niter, stopCr, TRUE)
```

The clustering results and the optimal centers are obtained in the following way:

```
asig <- res$asig
table(asig)
copt <- res$copt
```

The trimmed individuals of the optimal iteration can be also identified:

```
iter_opt <- res$trimmsIter[length(res$trimmsIter)]
trimm <- res$trimmWomen[[iter_opt]][[res$betterNstep]]
```

In order to examine the differences between clusters for some key anthropometric dimensions, their boxplots can be represented. To do this, we need to identify the first 50 individuals in `dataDemo` and to remove the trimmed ones. Figure 3 (left) displays the boxplots for neck to ground measurement for the three clusters calculated.

```
data <- dataDemo[1 : 50, ]
data <- data[-trimm, ]
boxplot(data$necktoground ~ as.factor(asig), main = "Neck to ground")
```

In addition, Figure 3 (right) displays the projection on the xy plane of the recorded points and mean shape for cluster 1. To that end, we first need to carry out a generalized Procrustes analysis in each cluster to obtain the full Procrustes rotated data.

```
out_proc <- list()
for(h in 1 : K){
 out_proc[[h]] = shapes::procGPA(dg[, , asig == h], distances = T,
                                 pcaoutput = T)
}

shapes::plotshapes(out_proc[[1]]$rotated)
points(copt[, , 1], col = 2)
legend("topleft", c("Registrated data", "Mean shape"), pch = 1,
       col = 1:2, text.col = 1:2)
title("Procrustes registrated data for cluster 1 \n
      with its mean shape superimposed", sub = "Plane xy")
```

5.3. Archetypal analysis

We focus on the cockpit design problem. The accommodation of boundaries (our archetypoids) ensures the accommodation of interior points in the cockpit. We use the `dataUSAF` database. Again, as an illustrative example only the first 50 individuals are chosen. From the total variables, the six so-called cockpit dimensions are selected. We convert the variables from mm into inches in order to compare our results with those discussed in Zehner *et al.* (1993) (see Epifanio *et al.* (2013)). Then, before computing archetypes and archetypoids, the data must be preprocessed. This is done with the following `accommodation` function.

This step includes a possible standardization of the variables and fixing a percentage of the population to accommodate. In this case, the variables are standardized as they measure different dimensions (first `TRUE` in `accommodation`) and the accommodation percentage is fixed to 0.95. When designing a workspace, it has typically been a requirement that between 90 and 95 percent of the relevant population are accommodated. Finally, the second `TRUE` indicates that the Mahalanobis distance will be used to remove the more extreme 5% data. If `FALSE`, a depth procedure is used (see Epifanio *et al.* 2013, section 2.2.2 for more details).

```
m <- dataUSAF[1 : 50, ]
sel <- c(48, 40, 39, 33, 34, 36)
```

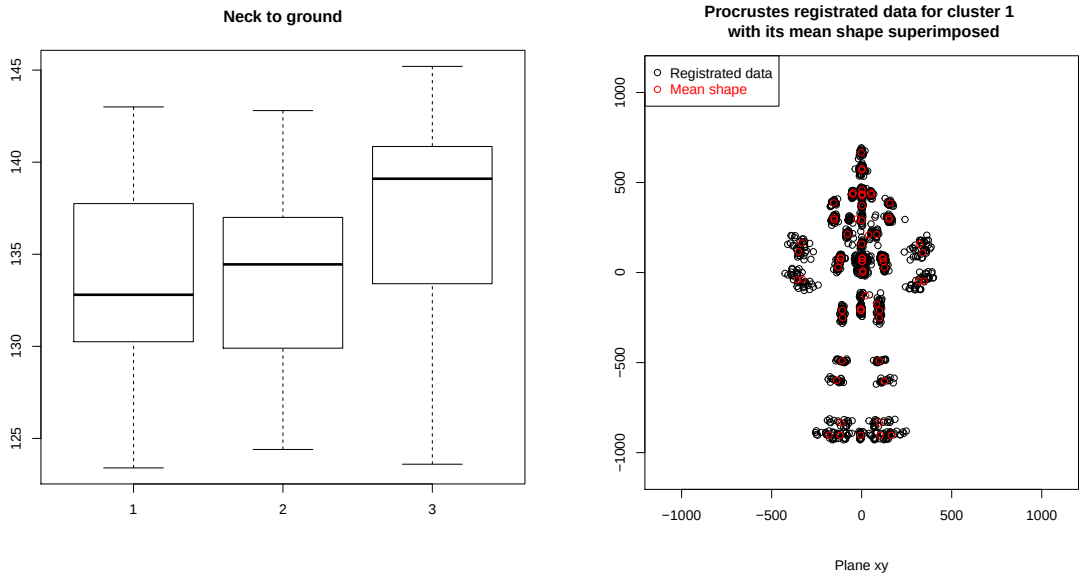

Figure 3: Boxplots for the neck to ground measurement for three clusters (left) and projection on the xy plane of the recorded points and mean shape for cluster 1 (right). Results provided by trimmed kmeansProcrustes.

```
mpulg <- m[,sel] / (10 * 2.54)
preproc <- accommodation(mpulg, TRUE, 0.95, TRUE)
```

Next the archetype algorithm is run repeatedly from 1 to `numArch` archetypes. The user can decide how many archetypes are to be considered. We chose `numArch` equal to 10 because a larger number of boundary cases may be may overwhelm the designer and therefore be counterproductive. The argument `nrep` specifies the number of repetitions of the algorithm. Choosing twenty repetitions ensures that the best possible archetypes are obtained. To reproduce results, a seed for randomness is fixed.

```
set.seed(2010)
numArch <- 10 ; nrep <- 20
lass <- stepArchetypesMod(data = preproc$data, k = 1 : numArch,
                          verbose = FALSE, nrep = nrep)
screeplot(lass)
```

According to the screeplot and following the elbow criterion, we compute three archetypoids (beginning from *nearest* and *which* sets).

```
i <- 3
res <- archetypoids(i, preproc$data, huge = 200, step = FALSE,
                    ArchObj = lass, nearest = TRUE, sequ = TRUE)
res_which <- archetypoids(i, preproc$data, huge = 200, step = FALSE,
                          ArchObj = lass, nearest = FALSE, sequ = TRUE)
```

```
aux <- res$archet
aux_wh <- res_which$archet
```

In this case, the nearest and which archetypoids match (although the nearest and which archetypes do not), so it is enough to represent a single percentile plot. To that end, the `compPerc` computes the percentiles of the archetypoids for every column of the data frame.

```
percs <- list()
for(j in 1 : length(aux)){
 percs[[j]] <- sapply(1 : dim(preproc$data)[2], compPerc, aux[j],
                      preproc$data, 0)
}
m <- matrix(unlist(percs), nrow = 6, ncol = length(percs), byrow = F)
```

Figure 4 shows the percentiles of three archetypoids, beginning from *nearest* (left) and with *which* (right).

```
barplot(m, beside = TRUE, main = paste(i, " archetypoids", sep = ""),
        ylim = c(0, 100), ylab = "Percentile")
```

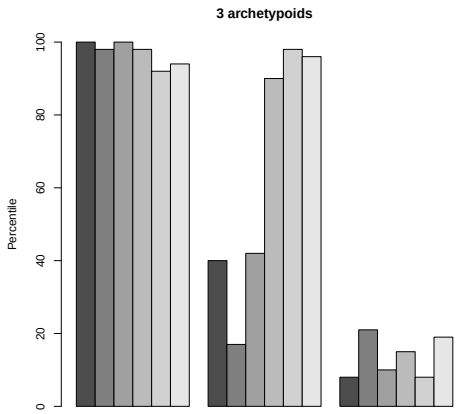

Figure 4: Percentiles of three archetypoids, beginning from the *nearest* and *which* sets for `dataUSAF`. In this case, the nearest and which archetypoids coincide.

6. Comparison of the clustering methods: Guidance for users

In the **Anthropometry** R package five clustering methods are available (*trimowa*, *CCbiclustAnthropo*, *TDDclust*, *HipamAnthropom* and *kmeansProcrustes*), each offering a different theoretical foundation and practical benefits. The purpose of this section is to provide users with insights that can enable them to make a suitable selection of the proposed methods.

The main difference between them is their practical objective. This is the first key to finding out which method is right for the user. If the goal of the practitioner is to obtain representative fit models for apparel sizing, the *HipamAnthropom* algorithm must be used. Otherwise, if the goal is to create clothing size groups and size prototypes, the other four methods are suitable for this task. If the user wanted to design lower body garments, *CCbiclustAnthropo* should be chosen. Otherwise, *trimowa*, *TDDclust* and *kmeansProcrustes* are suitable for designing upper body garments. Finally, choosing one of the latter three methods depends on the kind of data being collected. If the database contains a set of 3D landmarks representing the shape of women, the *kmeansProcrustes* method must be applied. On the other hand, *trimowa* and *TDDclust* can be used when the data are 1D body measurements.

For illustrative purposes, Figure 5 shows a decision tree that helps the user to decide which clustering approach is best suited.

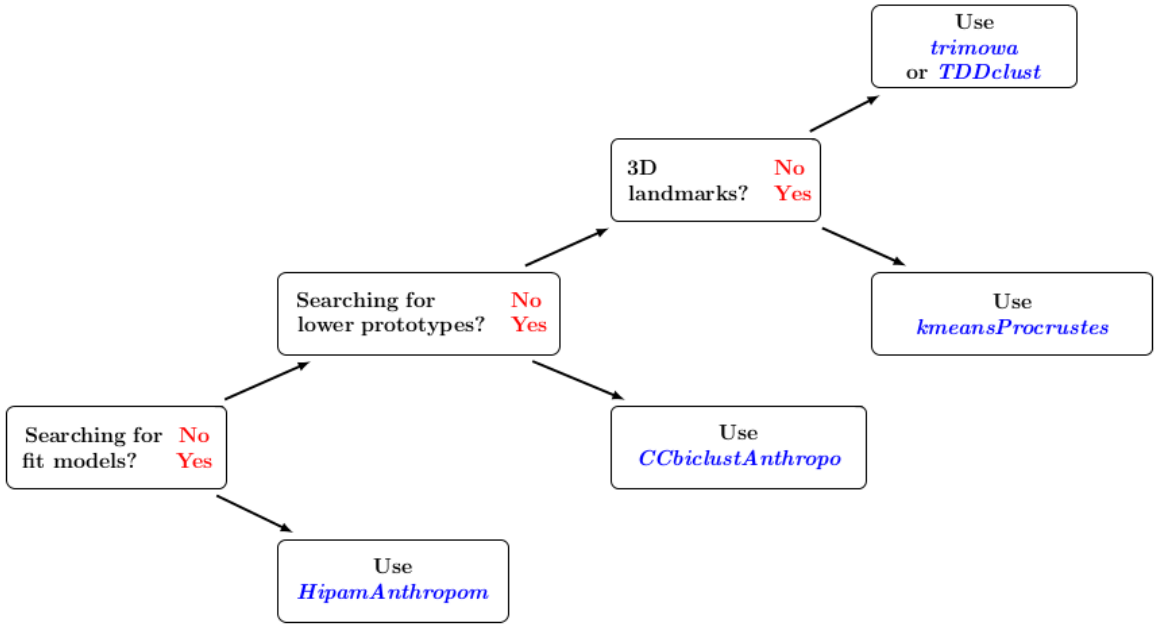

Figure 5: Decision tree as user guidance for choosing which of the different clustering methods to apply.

As a conclusion to this discussion, an illustrative comparison of the outcomes of using *trimowa* and *TDDclust* on a random sample subset is given below. We restrict our attention to these two methods because both of them have the same intention.

We run both algorithms for twenty randomly selected women. To reproduce results, a seed for randomness is fixed.

```
set.seed(1900)
rand <- sample(1:600,20)

dataDef <- dataDemo[rand, c(2, 3, 5)]
```

```
data1 <- dataDemo[rand, c(2, 3, 5)]

K <- 3 ; lambda <- 0.5 ; Th <- 0
A <- 5 ; TO <- 0 ; alpha <- 0.9
percTrimm <- 0.01 ;ahVect <- c(28, 25, 25)
orness <- 0.7 ; niter <- 10 ; Ksteps <- 7

#TDDclust:
Dout <- TDDclust(x = dataDef, K = K, lambda = lambda, Th = Th, A = A,
                 TO = TO, alpha = alpha, Trimm = percTrimm,
                 data1 = data1)
Dout$Y

#Trimowa:
num.variables <- dim(dataDef)[2]
w <- WeightsMixtureUB(orness, num.variables)
res_trimowa <- trimowa(dataDef, w, K, percTrimm, niter, Ksteps,
                       ahVect = ahVect)
dataDemo[res_trimowa$meds,]
```

Table 1 shows, in blue and with a frame box, the upper prototypes obtained with *TDDclust* and with *trimowa*, respectively. In this case, two of the three prototypes match. However, it is worth pointing out that in another case it is possible that none of them would match. This is because of the different statistical foundation of each approach. At this point, it would be recommendable to use the *trimowa* methodology because it has been developed further than *TDDclust*, returns outcomes with a significantly lower computational time, regardless of the sample size, and is endorsed by a scientific publication.

Label women	neck to ground	waist	bust
92	134.3	71.1	82.7
340	136.3	85.9	95.9
480	133.1	96.8	106.5
396	136.6	90.2	100.2

Table 1: Upper size prototypes obtained by *TDDclust* (in blue) and by *trimowa* (frame box).

6.1. Additional remark: selecting anthropometric cases

Clustering methodologies have been developed to obtain central cases. On the other hand, methods based on archetype and archetypoid analysis aim to identify boundary cases. Having explained the differences between the clustering methods, it is also of great importance to remember when each approach is best suited to obtain representative central or boundary cases. Fig. 6 shows a decision tree providing guidance in this question.

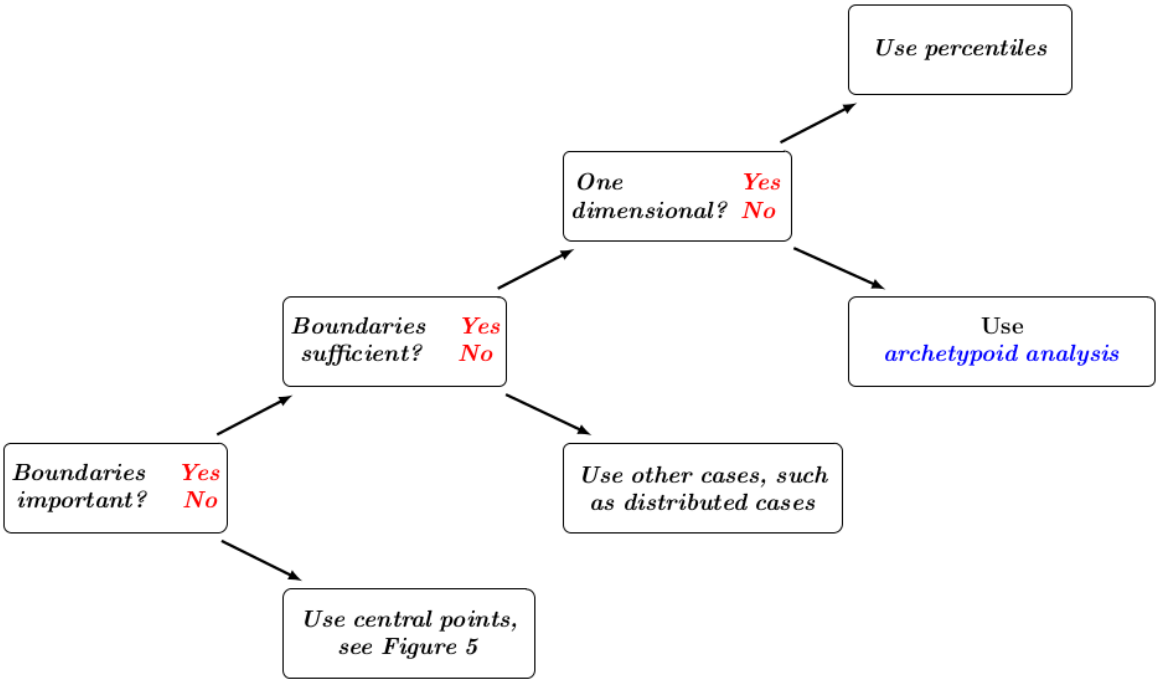

Figure 6: Decision tree for case selection methods.

7. Conclusions

New three-dimensional whole-body scanners have drastically reduced the cost and duration of the measurement process. These types of systems, in which the human body is digitally scanned and the resulting data converted into exact measurements, make it possible to obtain accurate, reproducible and up-to-date anthropometric data. These databases constitute very valuable information to effectively design better-fitting clothing and workstations, to understand the body shape of the population and to reduce the design process cycle. Therefore, rigorous statistical methodologies and software applications must be developed to make the most of them.

This paper introduces a new R package called **Anthropometry** that brings together different statistical methodologies concerning clustering, the statistical concept of data depth, statistical shape analysis and archetypal analysis, which have been especially developed to deal with anthropometric data. The data used have been obtained from a 3D anthropometric survey of the Spanish female population and from the USAF survey. Procedures related to clustering, data depth and shape analysis are aimed at defining optimal clothing size groups and both central prototypes and fit models. The two approaches based on archetypal analysis are useful for determining boundary human models which could be useful for improving industry practice in workspace design.

The **Anthropometry** R package is a positive contribution to help tackle some statistical problems related to Ergonomics and Anthropometry. It provides a useful software tool for engineers and researchers in these fields so that they can analyze their anthropometric data in a comprehensive way.

Acknowledgments

The author gratefully acknowledges the many helpful suggestions of I. Epifanio and G. Ayala. The author would also like to thank the Biomechanics Institute of Valencia for providing us with the Spanish anthropometric data set and the Spanish Ministry of Health and Consumer Affairs for having commissioned and coordinated the "Anthropometric Study of the Female Population in Spain". This paper has been partially supported by the following grants: TIN2009-14392-C02-01, TIN2009-14392-C02-02. The author would also like to thank the referees for their very constructive suggestions, which led to a great improvement of this paper.

References

Alemany S, González JC, Nácher B, Soriano C, Arnáiz C, Heras H (2010). "Anthropometric Survey of the Spanish Female Population Aimed at the Apparel Industry." In *Proceedings of the 2010 International Conference on 3D Body Scanning Technologies*. Lugano, Switzerland.

Ashdown S & Loker S (2005). "Improved Apparel Sizing: Fit and Anthropometric 3D Scan Data." *Technical report*, National Textile Center Annual Report.

Ashdown SP (2007). *Sizing in Clothing: Developing Effective Sizing Systems for Ready-To-Wear Clothing*. Woodhead Publishing in Textiles.

Bagherzadeh R, Latifi M, Faramarzi AR (2010). "Employing a Three-Stage Data Mining Procedure to Develop Sizing System." *World Applied Sciences Journal*, **8**(8), 923–929.

Bertilsson E, Högberg D, Hanson L (2012). "Using Experimental Design to Define Boundary Manikins." *Work: A Journal of Prevention, Assessment and Rehabilitation*, **41**(Supplement 1), 4598–4605.

Bittner AC, Glenn FA, Harris RM, Iavecchia HP, Wherry RJ (1987). "CADRE: A Family of Manikins for Workstation Design." In *Asfour, S.S. (ed.) Trends in Ergonomics/Human Factors IV. North Holland*, pp. 733–740.

Blanchonette P (2010). "Jack Human Modelling Tool: A Review." *Technical Report DSTO-TR-2364*, Defence Science and Technology Organisation (Australia). Air Operations Division.

Cheng Y, Church GM (2000). "Biclustering of Expression Data." *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, **8**, 93–103.

Chung MJ, Lin HF, Wang MJJ (2007). "The Development of Sizing Systems for Taiwanese Elementary- and High-School Students." *International Journal of Industrial Ergonomics*, **37**, 707–716.

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

D'Apuzzo N (2009). *Recent Advances in 3D Full Body Scanning with Applications to Fashion and Apparel*. Gruen, A., Kahmen, H. (eds.). Optical 3-D Measurement Techniques IX.

Ding Y, Dang X, Peng H, Wilkins D (2007). "Robust Clustering in High Dimensional Data Using Statistical Depths." *BMC Bioinformatics*, **8**(Suppl 7:S8), 1–16.

Dryden IE, Mardia KV (1998). *Statistical Shape Analysis*. John Wiley & Sons.

Dutta D, Ghosh A (2012). "On Robust Classification Using Projection Depth." *Annals of the Institute of Statistical Mathematics*, **64**(3), 657–676.

Epifanio I, Vinué G, Alemany S (2013). "Archetypal Analysis: Contributions for Estimating Boundary Cases in Multivariate Accommodation Problem." *Computers & Industrial Engineering*, **64**, 757–765.

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

Eugster MJA (2012). "Performance Profiles Based on Archetypal Athletes." *International Journal of Performance Analysis in Sport*, **12**(1), 166–187.

European Committee for Standardization (2002). "Size Designation of Clothes. Part 2: Primary and Secondary Dimensions."

European Committee for Standardization (2005). "Size Designation of Clothes. Part 3: Measurements and Intervals."

Friess M (2005). "Multivariate Accommodation Models Using Traditional and 3D Anthropometry." *Technical report*, SAE.

Friess M, Bradtmiller B (2003). "3D Head Models for Protective Helmet Development." *Technical report*, SAE.

García-Escudero LA, Gordaliza A, Matrán C (2003). "Trimming Tools in Exploratory Data Analysis." *Journal of Computational and Graphical Statistics*, **12**(2), 434–449.

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008). "A General Trimming Approach to Robust Cluster Analysis." *The Annals of Statistics*, **36**, 1324–1345.

Gupta D, Gangadhar BR (2004). "A Statistical Model for Developing Body Size Charts for Garments." *International Journal of Clothing Science and Technology*, **16**(5), 458–469.

HFES 300 Committee (2004). *Guidelines for Using Anthropometric Data in Product Design*. Human Factors and Ergonomics Society.

Hsu CH (2009a). "Data Mining to Improve Industrial Standards and Enhance Production and Marketing: An Empirical Study in Apparel Industry." *Expert Systems with Applications*, **36**, 4185–4191.

Hsu CH (2009b). "Developing Accurate Industrial Standards to Facilitate Production in Apparel Manufacturing Based on Anthropometric Data." *Human Factors and Ergonomics in Manufacturing*, **19**(3), 199–211.

Hudson JA, Zehner GF, Meindl RD (1998). "The USAF Multivariate Accommodation Method." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **42**(10), 722–726.

Ibáñez MV, Simó A, Domingo J, Durá E, Ayala G, Alemany S, Vinué G, Solves C (2012a). "A Statistical Approach to Build 3D Prototypes from a 3D Anthropometric Survey of the Spanish Female Population." In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, volume 1, pp. 370–374. Vilamoura, Algarve, Portugal.

Ibáñez MV, Vinué G, Alemany S, Simó A, Epifanio I, Domingo J, Ayala G (2012b). "Apparel Sizing Using Trimmed PAM and OWA Operators." *Expert Systems with Applications*, **39**, 10512–10520.

Irigoien I, Arenas C (2008). "INCA: New Statistic for Estimating the Number of Clusters and Identifying Atypical Units." *Statistics in Medicine*, **27**, 2948–2973.

Istook CL, Hwang SJ (2001). "3D Body Scanning Systems with Application to the Apparel Industry." *Journal of Fashion Marketing and Management*, **5**, 120–132.

Jörnsten R (2004). "Clustering and Classification Based on the L_1 Data Depth." *Journal of Multivariate Analysis*, **90**, 67–89.

Jörnsten R, Vardi Y, Zhang C (2002). "A Robust Clustering Method and Visualization Tool Based on Data Depth." In *Statistical Data Analysis Based on the L1-norm and Related Methods (Neuchâtel, 2002), Statistics for Industry and Technology*, pp. 353–366.

Kaiser S, Leisch F (2008). "A Toolbox for Bicluster Analysis in R." *Technical report*, Department of Statistics (University of Munich).

Kaiser S, Santamaria R, Khamiakova T, Sill M, Theron R, Quintales L, Leisch F (2013). ***biclust**: BiCluster Algorithms*. R package version 1.0.2, URL http://CRAN.R-project.org/package=biclust.

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

Lange T, Mosler K, Mozharovskyi P (2012). "Fast Nonparametric Classification Based on Data Depth." *Statistical Papers*, **5**, 1–22.

Lerch T, MacGillivray M, Domina T (2007). "3D Laser Scanning: A Model of Multidisciplinary Research." *Journal of Textile and Apparel, Technology and Management*, **5**, 1–22.

Liu RY, Parelius JM, Singh K (1999). "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference." *The Annals of Statistics*, **27**(3), 783–858.

Loker S, Ashdown S, Schoenfelder K (2005). "Size-Specific Analysis of Body Scan Data to Improve Apparel Fit." *Journal of Textile and Apparel, Technology and Management*, **4**(3), 1–15.

López A, Romo J (2010). "Simplicial Similarity and Its Application to Hierarchical Clustering." *Technical report*, Universidad Carlos III de Madrid. Departamento de Estadística. Working papers. Statistics and Econometrics.

Lu JM, Wang MJJ (2008). "Automated Anthropometric Data Collection Using 3D Whole Body Scanners." *Expert Systems with Applications*, **35**, 407–414.

Luximon A, Zhang Y, Luximon Y, Xiao M (2012). "Sizing and Grading for Wearable Products." *Computer-Aided Design*, **44**, 77–84.

Madeira SC, Oliveira AL (2004). "Biclustering Algorithms for Biological Data Analysis: A Survey." *IEEE Transactions on Computational Biology and Bioinformatics*, **1**, 24–45.

McCulloch CE, Paal B, Ashdown SP (1998). "An Optimization Approach to Apparel Sizing." *Journal of the Operational Research Society*, **49**, 492–499.

Moroney WF, Smith MJ (1972). "Empirical Reduction in Potential User Population As the Result of Imposed Multivariate Anthropometric Limits." *Technical report*, Naval Aerospace Medical Research Laboratory.

Pan JZ, Jörnsten R, Hart RP (2004). "Screening Anti-Inflammatory Compounds in Injured Spinal Cord with Microarrays: A Comparison of Bioinformatics Analysis Approaches." *Physiological Genomics*, **17**, 201–214.

Parkinson MB, Reed MP, Kokkolaras M, Papalambros PY (2006). "Optimizing Truck Cab Layout for Driver Accommodation." *Journal of Mechanical Design*, **129**(11), 1110–1117.

Pheasant S (2003). *Bodyspace: Anthropometry, Ergonomics and the Design of Work*. Taylor & Francis, Ltd.

R Development Core Team (2013). *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/.

Robinette KM, McConville JT (1981). "Alternative to Percentile Models." *Technical report*, SAE.

Robinson JC, Robinette KM, Zehner GF (1992). "User's Guide to the Anthropometric Database at the Computerized Anthropometric Research and Design (CARD) Laboratory (U)." *Technical report*, Systems Research Laboratories.

Seiler C, Wohlrabe K (2013). "Archetypal Scientists." *Journal of Informetrics*, **7**, 345–356.

Shu C, Wuhrer S, Xi P (2011). "Geometric and Statistical Methods for Processing 3D Anthropometric Data." In *International Symposium on Digital Human Modeling*.

Simmons KP, Istook CL (2003). "Body Measurement Techniques: Comparing 3D Body-Scanning and Anthropometric Methods for Apparel Applications." *Journal of Fashion Marketing and Management*, **7**(3), 306–332.

Song HK, Ashdown SP (2010). "An Exploratory Study of the Validity of Visual Fit Assessment From Three-Dimensional Scans." *Clothing and Textiles Research Journal*, **28**(4), 263–278.

Tryfos P (1986). "An Integer Programming Approach to the Apparel Sizing Problem." *The Journal of the Operational Research Society*, **37**(10), 1001–1006.

Vardi Y, Zhang CH (2000). "The Multivariate L_1-Median and Associated Data Depth." *Proceedings of the National Academy of Sciences*, **97**, 1423–1426.

Vinué G (2012). *Métodos Biclustering Aplicados a Datos Antropométricos: Exploración de Su Posible Aplicación en el Diseño de Indumentaria*. Master's thesis, School of Mathematics, University of Valencia (Spain). In Spanish.

Vinué G (2014). *Development of Statistical Methodologies Applied to Anthropometric Data Oriented Towards the Ergonomic Design of Products*. Ph.D. thesis, Faculty of Mathematics. University of Valencia, Spain, http://hdl.handle.net/10550/35907.

Vinué G, Epifanio I, Alemany S (2014a). "Archetypoids: A New Approach to Define Representative Archetypal Data." Submitted.

Vinué G, Ibáñez MV (2014). "Data Depth and Biclustering Applied to Anthropometric Data: Exploring Their Utility in Apparel Design." In progress.

Vinué G, León T, Alemany S, Ayala G (2014b). "Looking for Representative Fit Models for Apparel Sizing." *Decision Support Systems*, **57**, 22–33.

Vinué G, Simó A, Alemany S (2014c). "The k-Means Algorithm for 3D Shapes with an Application to Apparel Design." Accepted for publication in Advances in Data Analysis and Classification.

Wang MJJ, Wu WY, Lin KC, Yang SN, Lu JM (2007). "Automated Anthropometric Data Collection from Three-Dimensional Digital Human Models." *The International Journal of Advanced Manufacturing Technology*, **32**, 109–115.

Wit E, McClure J (2004). *Statistics for Microarrays: Design, Analysis and Inference*. John Wiley & Sons.

Workman J (1991). "Body Measurement Specifications for Fit Models As a Factor in Clothing Size Variation." *Clothing and Textiles Research Journal*, **10**(1), 31–36.

Workman JE, Lentz ES (2000). "Measurement Specifications for Manufacturers' Prototype Bodies." *Clothing and Textiles Research Journal*, **18**(4), 251–259.

Zehner GF, Meindl RS, Hudson JA (1993). "A Multivariate Anthropometric Method for Crew Station Design: Abridged." *Technical report*, Human Engineering Division, Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio.

Zheng R, Yu W, Fan J (2007). "Development of a New Chinese Bra Sizing System Based on Breast Anthropometric Measurements." *International Journal of Industrial Ergonomics*, **37**, 697–705.

Zuo Y, Serfling R (2000). "General Notions of Statistical Depth Function." *The Annals of Statistics*, **28**(2), 461–482.

Affiliation:

Guillermo Vinué
Department of Statistics and Operations Research
Faculty of Mathematics

University of Valencia
46100 Burjassot, Spain
E-mail: Guillermo.Vinue@uv.es
URL: http://www.uv.es/vivigui