# *Developing statistical methodologies for Anthropometry*

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

Jointly with Guillermo Ayala Gallego, Juan Domingo Esteve, Esther Durá Martínez (University of Valencia), Amelia Simó Vidal, María Victoria Ibañez Gual, Irene Epifanio López (University Jaume I of Castellón) and Sandra Alemany Mut (Biomechanics Institute of Valencia)

18th European Young Statisticians Meeting, 26-30 August 2013, Osijek, Croatia

# Outline

# Motivation and problematic

- Clothing development process and human modelling require updated anthropometric data to develop new patterns.
- Body measurements have been traditionally taken by using rudimentary methods.
  - Advantages:
    - Procedures are very easy to use.
    - No particularly expensive.
  - Drawbacks:
    - The shape information is imprecise and inaccuracy.
    - Interaction with real subjects.
- New 3D body scanner technologies constitute a step forward in the way of collecting anthropometric data.
- Anthropometric surveys in different countries (Spain, 2006).

# Anthropometric database

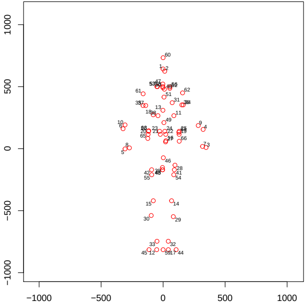
- A national 3D anthropometric survey of the female population was conducted in Spain in 2006 by the Spanish Ministry of Health.

- Aim: To generate anthropometric data from the female population addressed to the clothing industry.

- Database: Sample of 10.415 Spanish women randomly selected:
  - From 12 to 70 years old.
  - 95 anthropometric measurements.
  - 66 points representing their shape.
  - Socio-demographic survey.

# Landmarks

- The shape of all the women of the database is represented by landmarks.
- Landmark: Point $(x, y, z)$ of correspondence on each individual that matches between and within populations.
- The configuration is the set of landmarks $\Rightarrow X \in \mathcal{M}_{66\times 3}(\mathbb{R})$

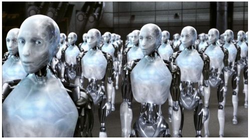



| Landmark | Description |
|---|---|
| 1. Head back | Most prominent point of the head in the sagital plane |
| 2. Head front | Glabela (most prominінet point of the forehead) |
| 3. Forearm wrist left | Maximum girth of the left forearm |
| 4. Forearm girth left | Maximum girth of the left forearm just under the left elbow |
| 5. Forearm wrist right | Maximum girth of the right forearm |
| ..... | ..... |
| 66. Left iliac crest | Physical marker on the left of the iliac crest |

# Statistical approaches

- Clothing development process

- Human modelling



- R package

- Clustering:
  - * Trimmed PAM.
  - * Hierarchical PAM.
- Statistical shape analysis:
  - * k-means adapted.
- Archetypal analysis:
  - * Archetypes & Archetypoids.

## *Clothing development process*

# Clothing development process (I)

- Clothing development process: To define a sizing system that fits good.
- Current sizing systems don't cover all morphologies.
- Causes:
  - Old size charts.
  - Clothing manufacturers work by trial and error.
  - Sizing systems are not standardized.
- The final evaluation of fit needs fit models.
- A good fit model is the basis for defining an accurate sizing system.
- They define a single fit model for the whole target market.
- There is not much information to help choose a fit model.

# Clothing development process (II)

- Consequences: Lack of fitting of the sizing systems.
  * Large amount of unsold garments (company competitivity loss).
  * High index of returned garments (customer dissatisfaction).
- Proposals in the literature to define sizing systems:
  * Multivariate approaches (PCA, Clustering).
  * Optimization algorithms (McCulloch et al. (1998)).
- Proposals in the literature to define fit models:
  * There are not any statistical method aimed at defining fit models.
- Our proposals:
  * Sizing system:
    - Trimmed PAM.
    - *k*-means in the shape space.
  * Fit models: Hierarchical PAM.

# Clustering: Trimmed PAM

- Our proposal is an improvement of the work of McCulloch et al. (1998):
  - $k$-medoids clustering method → prototypes will be real persons.
  - Trimmed $k$-medoids to leave out outlier individuals.
  - Extension of the dissimilarity used by McCulloch et al. (1998), $d_{MO}$.
  - OWA operators to take into account to the user.
- Data set (6013 women and 5 dimensions):
  - Not pregnant women. ; Not breast feeding at the time of the survey.
  - No cosmetic surgery. ; Between 20 and 65 years.
  - Bust, chest, waist and hip circumference and neck to ground length.
- Procedure:
  - The data set is segmented in 12 bust classes (EN 13402-2).
  - The trimmed $k$-medoids is applied to each segment with $k = 3$.
- Conclusions of this work:
  - Assignment of individuals to size.
  - Derivation of realistic prototypes.
  - Selection of individual discommodities.

Expert Systems with Applications 39 (2012) 10512–10520

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

# Apparel sizing using trimmed PAM and OWA operators

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

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

A R T I C L E I N F O

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

A B S T R A C T

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

# Clustering: Hierarchical PAM (HIPAM)

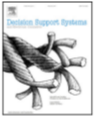
- HIPAM algorithm: divisive hierarchical clustering method based on PAM (Wit et al. (2004)).
- Two HIPAM algorithms adapted to anthropometric data:
  - * $HIPAM_{MO}$: HIPAM + dissimilarity $d_{MO}$ + asw.
  - * $HIPAM_{IMO}$: HIPAM + dissimilarity $d_{MO}$ + INCA (Irigoien et al. (2008)).
- The INCA criterion is used to divide the clusters and as a stopping rule.
- Data set and procedure:
  - * Same data set used with by the trimmed PAM procedure.
  - * The data set is also segmented in 12 bust classes (EN 13402-2).
  - * Both HIPAM algorithms are applied to each segment.
- Conclusions of this work:
  - * Robust statistical tool to determine accurate fit models.
  - * Automatic detection of outliers.

Decision Support Systems xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

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



# Looking for representative fit models for apparel sizing

G. Vinué [a,*], T. León [a], S. Alemany [b], G. Ayala [a]

[a] Department of Statistics and O.R., Avda, Vicent Andrés Estellés, 1. 46100-Burjasot, University of Valencia, Valencia, Spain
[b] Biomechanics Institute of Valencia, Universidad Politécnica de Valencia, Valencia, Spain

ARTICLE INFO

*Article history:*
Received 15 February 2013
Received in revised form 23 July 2013
Accepted 23 July 2013
Available online xxxx

*Keywords:*
HIPAM
Hierarchical tree
Partitioning around medoids
Fit models
Mean split silhouette
INCA statistic

ABSTRACT

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

We propose two algorithms based on a new hierarchical partitioning around medoids clustering method originally developed for gene expression data. We are concerned with a different application; therefore, the dissimilarity between the objects has to be different and must be designed to deal with anthropometric features. Furthermore, one of the algorithms incorporates a different rule to split the clusters, which, in our case, provides better results. Our procedures not only make it possible to obtain optimal prototypes, but also to detect outliers. These outliers should be removed before defining prototypes so that the companies' market share can be optimized.

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

# Statistical shape analysis

- Aim: To use $k$-means to divide the population according to their shapes.
- Fundamental concepts of the statistical shape analysis:
  - * Pre-shape of an object: It is what is left after allowing for the effects of translation and scale.
  - * Shape of an object: It is what is left after allowing for the effects of translation, scale, and rotation.
  - * Procrustes distance, $\rho$: Closest great circle distance between pre-shapes on the pre-shape sphere.
  - * Procrustes mean: The shape that has the least summed squared Procrustes distance to all the configurations of a sample.
- $k$-means has been usually applied using a set of anthropometric variables as input.
- $k$-means: The sample mean is the value that minimizes the Euclidean distance from each point, to the centroid of the cluster to which it belongs.

# K-means algorithm in the Shape Space

- By integrating the Procrustes mean and the Procrustes distance into $k$-means, we can use it in the shape analysis context.
- $k$-means algorithm to $X_1, \ldots, X_n$ configuration matrices:
  - (i) Given $Z = ([Z_1], \ldots, [Z_k])$ $[Z_i] \in \Sigma_3^{66}$ $i = 1, \ldots, k$, we minimize with respect to $\mathcal{C} = (C_1, \ldots, C_k)$ assigning each shape $([X_1], \ldots, [X_n])$ to the class whose centroid has minimum Procrustes distance to it.
  - (ii) Given $\mathcal{C}$, we minimize with respect to $Z$, taking $Z = ([\widehat{\mu_1}], \ldots, [\widehat{\mu_k}])$, being $[\widehat{\mu}_i]$ $i = 1, \ldots, k$ the Procrustes mean of shapes in $C_i$.
- Steps (i) and (ii) are repeated until convergence of the algorithm.
- We adapt the Lloyd $k$-means unlike Amaral et al. (2010) that use Hartigan-Wong $k$-means.

# K-means algorithm in the Shape Space

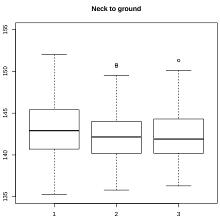
- Computational statistical tool: R package *shapes*.
- Data set and procedure:
  - Same data set used with the trimmed PAM procedure.
  - We segment our data set using bust and height measurements.
  - In this way, the size effect is filtered out in an easy way.
  - We apply the $k$-means algorithm to each segment ($k = 3$).

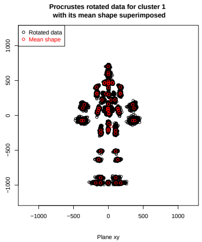
| Bust | Height1 $\leq 162$ cm | Height2 $[162 - 174[$ cm |
|---|---|---|
| $[74 - 82[$ cm | 240 | 97 |
| $[82 - 90[$ cm | 1052 | 694 |
| $[90 - 98[$ cm | 1079 | 671 |
| $[98 - 106[$ cm | 772 | 311 |
| $[106 - 118[$ cm | 446 | 170 |

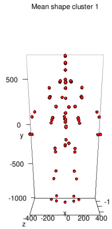
- Conclusions of this work:
  - $k$-means adapted in the context of shape analysis to cluster human body shapes based on landmarks.
  - Lloyd version has better performance than Hartigan-Wong version.

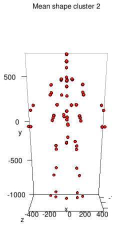
# Clustering results

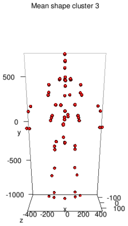
| Bust $\in$ [90-98[ ; Height $\in$ [162-174[ | | |
|---|---|---|
| 671 women | | |
| Cluster 1 | Cluster 2 | Cluster 3 |
| 153 | 206 | 312 |



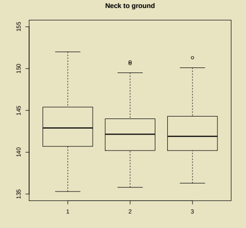



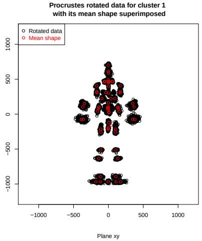



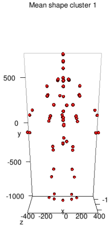



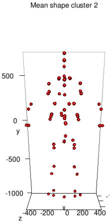
# Clustering results

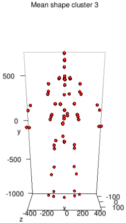
| Bust ∈ [90-98[ ; Height ∈ [162-174[ | | |
|---|---|---|
| 671 women | | |
| Cluster 1 | Cluster 2 | Cluster 3 |
| 153 | 206 | 312 |



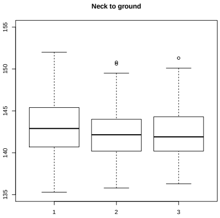



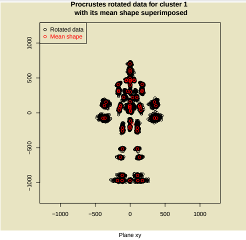



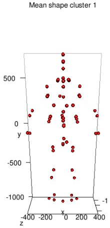



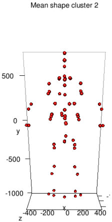
# Clustering results

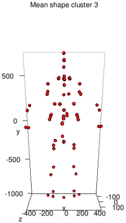
| Bust ∈ [90-98[ ; Height ∈ [162-174[ | | |
|---|---|---|
| 671 women | | |
| Cluster 1 | Cluster 2 | Cluster 3 |
| 153 | 206 | 312 |




SECOND REVISION

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

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

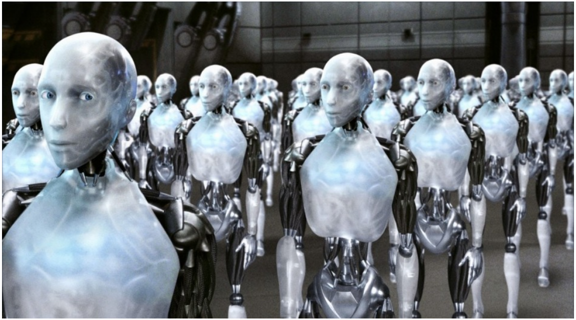
(1) *Department of Statistics and O.R., University of Valencia, C/ Dr. Moliner, 50. 46100-Burjassot, Valencia, Spain.*
(2) *Department of Mathematics-IMAC. University Jaume I. Avda. del Riu Sec s/n. 12071-Castellón, Spain.*
(3) *Biomechanics Institute of Valencia, Universidad Politécnica de Valencia,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

January 2, 2013

**Abstract**

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

# *Human modelling*



# Human modelling

- Products intended to *fit* the users must be designed considering their size and shape → Generation of several representative human models.
- The human models represent the anthropometric variability of the target population.
- The appropriate selection of this small group is critical.
- If the *hard to fit* extreme individuals are previously identified, the time and cost of the design process is reduced.
- Usual approaches:
    - Percentile analysis.
    - Regression.
    - PCA.
- Our proposal: ARCHETYPAL ANALYSIS (AA) (Cutler et al. (1994)).

# Archetypal analysis

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

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

under the constraints

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

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

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

# Archetypal analysis vs PCA

- Aircraft pilots database from the United States Air Force (USAF) Survey.
  - From the total of variables, we select six (the cockpit dimensions).
- Conclusions of this work:
  - * The goal of AA is to obtain extreme individuals.
  - * The level of accommodation is reached with AA.
  - * Archetypes cannot be obtained with PCA.
  - * Number of archetypes is decided either by the user or by a criterion.

| Dimension | 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. |

Computers & Industrial Engineering 64 (2013) 757–765

Contents lists available at SciVerse ScienceDirect

Computers & Industrial Engineering

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

ELSEVIER

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

CrossMark

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

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

ARTICLE INFO

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

ABSTRACT

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

# Archetypoid analysis

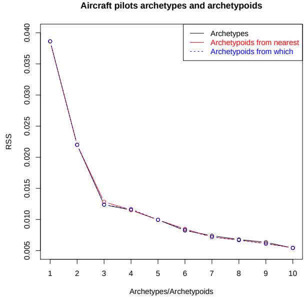
- The archetypes do not correspond to observed individuals:

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

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

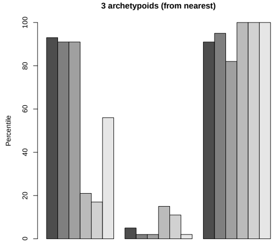
$$\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l \quad \text{with} \quad \sum_{l=1}^{n} \beta_{jl} = 1 \quad \text{and} \quad \boxed{\beta_{jl} \in \{0, 1\}}$$

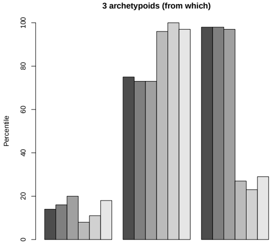
- An archetypoid is a real observed individual.
- The archetypoids might not be the same as the *nearest*/*which*.
- The archetypoids always exist even when the features are unavailable.




- Aircraft pilots database.

| | RSS |
|---|---|
| 3 archetypes | 0.012380 |
| 3 archetypoids from which (1632,1822,52) | 0.012385 |
| 3 archetypoids from nearest (2177,2240,1691) | 0.0128 |
| which (1421,314,1691) | 0.0195 |
| nearest (511,314,1691) | 0.0182 |




*Submitted to the Annals of Applied Statistics*
arXiv: arXiv:0000.0000

# ARCHETYPOIDS: A NEW APPROACH TO DEFINE REPRESENTATIVE ARCHETYPAL DATA*

By Guillermo Vinué, Irene Epifanio and Sandra Alemany

*University of Valencia, Universitat Jaume I and Biomechanics Institute of Valencia*

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

# *R package*

# R package

- A new R package called **Anthropometry** gathers together all these methodologies.
- Hopefully soon available from CRAN.

*Journal of Statistical Software*
*MMMMMM YYYY, Volume VV, Issue II.* *http://www.jstatsoft.org/*

**Anthropometry: An R Package for Analysis of Anthropometric Data**

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

**Abstract**

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

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

# Conclusions

- Updated anthropometric data constitute valuable information...
  - To optimize sizing systems.
  - To understand the body shape of the population.
  - To reduce the design process cycle.
- Rigorous statistical methodologies and software tools must be developed.
- Our research group has proposed several statistical methods concerning clustering, the statistical shape analysis and the archetypal analysis.
- Clustering methodologies and the shape analysis allow...
  - To define an optimal clothing sizing system.
  - To identify the fit models of each size.
  - To cluster individuals according to their shape.
- Methodologies based on the archetypal analysis look for representative subjects, which help to define coherent human models and mockups.
- A new R package has been introduced that joints all these methodologies.

# Basic references I

Alemany, S., González, J. C., 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 Intl. Conference on 3D Body scanning Technologies. pp. 307–315.

Amaral, G. J. A., Dore, L. H., Lessa, R. P., Stosic, B., 2010. k-Means Algorithm in Statistical Shape Analysis. Communications in Statistics - Simulation and Computation 39 (5), 1016–1026.

Claude, J., 2008. Morphometrics with R. Use R! Springer.

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

Dryden, I. E., Mardia, K. V., 1998. Statistical Shape Analysis. John Wiley & Sons.

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

European Committee for Standardization, 2002. European Standard EN 13402-2: Size system of clothing. Primary and secondary dimensions.

García-Escudero, L. A., Gordaliza, A., Matrán, C., Jun. 2003. Trimming tools in exploratory data analysis. Journal of Computational and Graphical Statistics 12 (2), 434–449.

Georgescu, V., 20th-24th July 2009. Clustering of Fuzzy Shapes by Integrating Procrustean Metrics and Full Mean Shape Estimation into K-Means Algorithm. In: IFSA-EUSFLAT Conference.

# Basic references II

Irigoien, I., Arenas, C., 2008. INCA: New statistic for estimating the number of clusters and identifying atypical units. Statistics in Medicine 27, 2948–2973.

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

León, T., Zuccarello, P., Ayala, G., de Ves, E., Domingo, J., 2007. Applying logistic regression to relevance feedback in image retrieval systems. Pattern Recognition 40, 2621–2632.

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

McCulloch, C., Paal, B., Ashdown, S., 1998. An optimization approach to apparel sizing. Journal of the Operational Research Society 49, 492–499.

Pollard, K. S., van der Laan, M. J., 2002. A method to identify significant clusters in gene expression data. Vol. II of SCI2002 Proceedings. pp. 318–325.

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