Clustering human body shapes using k-means algorithm

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 Jaime I of Castellón) and Sandra Alemany Mut (Biomechanics Institute of Valencia)

20th International Conference on COMPUTATIONAL STATISTICS (COMPSTAT 2012), 27-31 August 2012, Limassol, Cyprus

Outline

Abstract

- *k*-means algorithm: Minimizes $\sum_{i=1}^{k} \sum_{j \in C_i} d_E(x_j, \bar{x}_{C_i})^2$, where $\bar{x}_{C_i}$ is the sample mean of each group $C_1, \dots, C_k$ and d_E is the Euclidean distance.
- Idea: To integrate Procrustes mean and Procrustes distance into *k*-means.
- Several attempts in that sense (Amaral et al. (2010), Georgescu (2009)):
  - Amaral et al. $\Rightarrow$ Hartigan-Wong k-means algorithm.
  - Georgescu $\Rightarrow$ k-means algorithm similar to Lloyds algorithm.
- We will compare the performance of Hartigan-Wong and Lloyds versions of k-means in the field of Statistical Shape Analysis (SSA).
- Both algorithms will be applied to a recently 3D anthropometric female Spanish data base.

Motivation

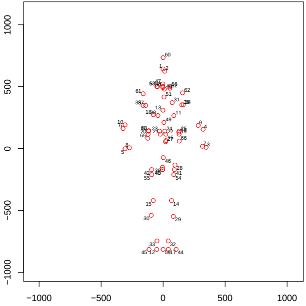
- Our application: the apparel sizing system design.
- Apparel development process $\Rightarrow$ To define a sizing system that fits good.
- Current sizing systems don't cover all morphologies.
- Causes:
  - Old size charts.
  - Apparel manufacturers work by trial and error.
  - Sizing systems are not standardized.
- Consequences: Lack of fitting of the sizing systems.
  - Large amount of unsold garments (company competitivity loss).
  - High index of returned garments (customer dissatisfaction).
- Clothing fit is a problem for both customer and apparel industry.
- Anthropometric surveys in different countries (Spain, 2006).

Anthropometric dataset

- 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 measures.
  - 66 points representing their shape.
  - Socio-demographic survey.

Landmarks

- The shape of all the women of our data base 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 promininet 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

Objectives

- Sizing system: Divides a population into homogeneous subgroups.
- Multivariate approaches proposed to develop optimal sizing systems:
  - Clustering ⇒ *k*-means using anthropometric variables as inputs.
- Our case: Clustering objects whose shapes are based on landmarks.
- Main objectives:
  1. To show how *k*-means can be adapted to cluster objects based on their shape in order to build optimal sizes.
  2. To compare Hartigan and Lloyds versions in SSA.
  3. To analyze the shape variability using PCA.
  4. To add a trimmed procedure into the Lloyds algorithm.

Shape Space and shape distances

- Pre-shape of an object: It is what is left after allowing for the effects of translation and scale.
- Pre-shape space: Set of all possible pre-shapes.
- Pre-shape space: Hypersphere of unit radius in (k-1)m real dimensions.
- Shape of an object: It is what is left after allowing for the effects of translation, scale, and rotation.
- Shape space Σ_3^{66}: Set of all possible shapes.
- Full Procrustes distance, $d_F(X_1, X_2)$: Square root of the sum of squared differences between the positions of the landmarks in two optimally superimposed configurations.
- Procrustes distance, ρ: Closest great circle distance between pre-shapes on the pre-shape sphere. $\Rightarrow d_F = sin(\rho)$.
- Procrustes mean: The shape that has the least summed squared Procrustes distance to all the configurations of a sample.

K-means algorithm in the Shape Space

- We apply the k-means algorithm to $X_1, \ldots, X_n$ configuration matrices, by using the Procrustes distance and Procrustes mean.
  (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 use Procrustes distance, ρ, because a computational time reason.

Experimental results

- Data set (6013 women):
  - Not pregnant women. ; Not breast feeding at the time of the survey.
  - No cosmetic surgery. ; Between 20 and 65 years.
- Computational statistical tool: R package *shapes*.
- Procedure:
  - We segment our data set using the European Normative.
  - We apply the k-means algorithm to each segment ($k = 3$).

Bust	Height1 ≤ 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

Lloyds and Hartigan algorithms comparison

- Three different sample sizes.
- Same random initial values for both algorithms.

Bust in [74-82[cm and height in [162-174[cm. (97 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	30	47	20	$\approx$ 7 min.	0.008931727
Hartigan version	31	50	16	$\approx$ 20 min.	0.008931948
Bust in [106-118[cm and height $\leq$ 162 cm. (446 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	183	113	150	$\approx$ 30 min.	0.006525749
Hartigan version	175	117	154	$\approx$ 3 h.	0.006522669
Bust in [82-90[cm and height $\leq$ 162 cm. (1052 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	195	539	318	$\approx$ 1 h.	0.004637619
Hartigan version	295	531	226	$\approx$ 15 h.	0.004604781

- Clustering results (groups and objective function) are very similar.
- Computational time increases dramatically for Hartigan version with big samples.
- Lloyds algorithm is more appropriate in SSA.

Lloyds and Hartigan algorithms comparison

- Three different sample sizes.
- Same random initial values for both algorithms.

Bust in [74-82[cm and height in [162-174[cm. (97 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	30	47	20	≈ 7 min.	0.008931727
Hartigan version	31	50	16	≈ 20 min.	0.008931948
Bust in [106-118[cm and height $\leq$ 162 cm. (446 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	183	113	150	≈ 30 min.	0.006525749
Hartigan version	175	117	154	≈ 3 h.	0.006522669
Bust in [82-90[cm and height $\leq$ 162 cm. (1052 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	195	539	318	≈ 1 h.	0.004637619
Hartigan version	295	531	226	≈ 15 h.	0.004604781

- Clustering results (groups and objective function) are very similar.
- Computational time increases dramatically for Hartigan version with big samples.
- Lloyds algorithm is more appropriate in SSA.

Lloyds and Hartigan algorithms comparison

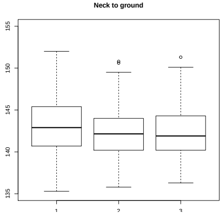
- Three different sample sizes.
- Same random initial values for both algorithms.

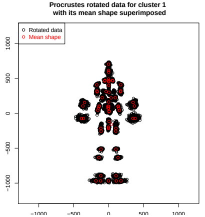
Bust in [74-82[cm and height in [162-174[cm. (97 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	30	47	20	≈ 7 min.	0.008931727
Hartigan version	31	50	16	≈ 20 min.	0.008931948
Bust in [106-118[cm and height ≤ 162 cm. (446 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	183	113	150	≈ 30 min.	0.006525749
Hartigan version	175	117	154	≈ 3 h.	0.006522669
Bust in [82-90[cm and height ≤ 162 cm. (1052 women)					
	Cluster 1	Cluster 2	Cluster 3	Computational time	Obj. function
Lloyds version	195	539	318	≈ 1 h.	0.004637619
Hartigan version	295	531	226	≈ 15 h.	0.004604781

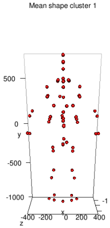
- Clustering results (groups and objective function) are very similar.
- Computational time increases dramatically for Hartigan version with big samples.
- Lloyds algorithm is more appropriate in SSA.

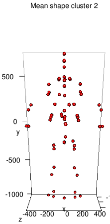
Clustering results

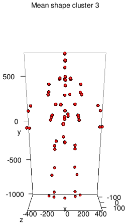
Bust ∈ [90-98[; Height ∈ [162-174[		
671 women		
Cluster 1	Cluster 2	Cluster 3
153	206	312





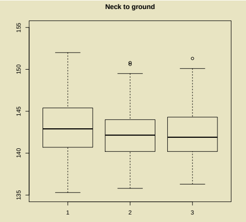


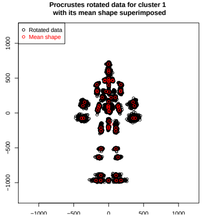


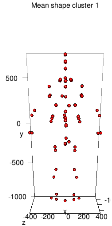


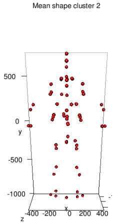
Clustering results

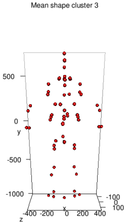
Bust $\in$ [90-98[; Height $\in$ [162-174[		
671 women		
Cluster 1	Cluster 2	Cluster 3
153	206	312





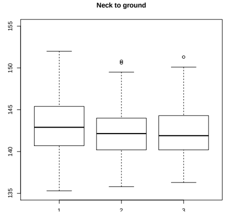


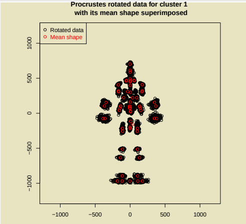


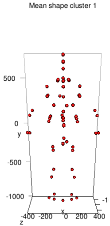


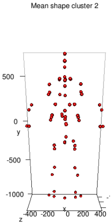
Clustering results

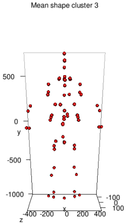
Bust ∈ [90-98[; Height ∈ [162-174[		
671 women		
Cluster 1	Cluster 2	Cluster 3
153	206	312











Analysis of shape variability I

- We have calculated the mean shape in each cluster through Procrustes superimposition.
- We want to describe now the variability in shape in each cluster $\Rightarrow$ PCA:
  - * Shows similarities and differences as simple scatter plots.
  - * Returns new variables for further statistical analysis.
- Analysis for previous cluster 1 (analogous for the other two clusters).
- Dryden and Mardia (1998) propose to evaluate:

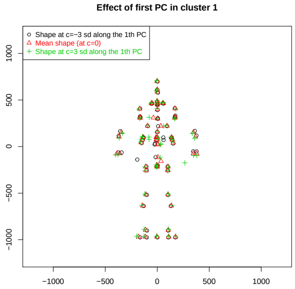
$$v(c, j) = \bar{v} + c\lambda_j^{1/2}\gamma_j, \quad j = 1, \ldots, p$$

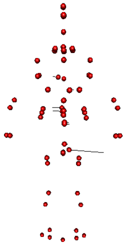
for a range of values of the standardized PC score c.

  - * v: Data in tangent space.
  - * $\bar{v}$: Mean shape.
  - * γ_j: PC of the matrix covariance of Procrustes residuals.
  - * λ_j: Corresponding eigenvalues of γ_j.

Analysis of shape variability II

- There are several ways to visualize the effect of each PC.
- We plot an icon projected in the xy plane for the values $c \in \{-3, 0, 3\}$.

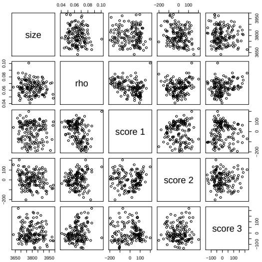




- The first PC shows variability at the belly and the iliac crest.

Analysis of shape variability III

- Pairwise plots of $(s_i, \rho_i, c_{i1}, c_{i2}, c_{i3})$, $i = 1, \ldots, 153$.
  - * s_i are the centroid sizes of the configuration.
  - * ρ_i are the Riemannian distances to the mean shape.
  - * c_{i1}, c_{i2}, c_{i3} are the first three standardized PC scores.



There appears to be one woman more far away than the rest: The Procrustes distance serves to find outliers.

- $RMS(d_F) = 0.07 \Rightarrow$ shape variability in cluster 1 is quite small.

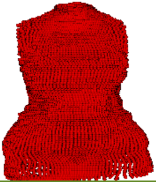
Trimmed k-means

- Results of k-means can be influenced by outliers.
- Garcia et al. (1999) proposed a way of robustify k-means $\Rightarrow$ trimmed procedure:
  * A proportion α ($\alpha \in [0, 1]$) of the total observations n is removed.
- An apparel sizing system is intended to cover only the standard population $\Rightarrow$ trimmed version of Lloyds k-means.
  * The $n\alpha$ shapes with largest distances are removed.
  * The $n(1 - \alpha)$ left are assigned to the class whose centroid has Procrustes minimum distance to it.
- Example: Group with bust in [74-82[cm and height in [162-174[cm.

	Cluster 1	Cluster 2	Cluster 3
Lloyds version (original)	30	47	20
Lloyds version (trimmed)	29	47	20

Conclusions and future work

- It has been shown how *k*-means can be adapted to SSA.
- We have applied it to the Anthropometric data base of Spanish women.
- It has been demonstrated that Lloyds version works better than Hartigan in SSA.
- We have used it to define a sizing system.
  - We have analyzed the shape variability of the clustering results.
- We have added a trimmed procedure to Lloyds algorithm.



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. Lugano, Switzerland.
- 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.
- Chunga, M., Lina, H., , Wang, M.-J. J., 2007. The development of sizing systems for taiwanese elementary- and high-school students. International Journal of Industrial Ergonomics 37, 707–716.
- Claude, J., 2008. Morphometrics with R. Use R! Springer.

Basic references II

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

García-Escudero, L. A., Gordaliza, A., 1999. Robustness properties of k-means and trimmed k-means. Journal of the American Statistical Association 94 (447), 956–969.

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.

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