Universidad de Valencia

Tesis Doctoral

Development of statistical methodologies applied to anthropometric data oriented towards the ergonomic design of products

---

Autor:
Guillermo Vinué Visús

Directoras:
Mª Teresa León Mendoza
Amelia Simó Vidal
Irene Epifanio López

Programa de Doctorado en Estadística y Optimización

Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas, Universidad de Valencia
Burjassot, España

Universidad de Valencia

Tesis Doctoral

Development of statistical methodologies applied to anthropometric data oriented towards the ergonomic design of products

---

Autor:
Guillermo Vinué Visús

Directoras:
Mª Teresa León Mendoza
Amelia Simó Vidal
Irene Epifanio López

Programa de Doctorado en Estadística y Optimización

Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas, Universidad de Valencia
Burjassot, España

A mis padres y mi familia.

Agradecimientos

En primer lugar, quiero agradecer a Guillermo Ayala la confianza que depositó en mí al concederme la beca FPI con la que he realizado esta tesis doctoral, la cual ha supuesto un paso adelante muy importante en mi carrera profesional. Le agradezco mucho también la flexibilidad y libertad que me ha dado para trabajar y su gran calidad humana para atenderme siempre que se me ha presentado un problema, ya fuera matemático o personal. Muchas gracias por todo, Guillermo.

A continuación, quisiera dar un agradecimiento especial a mis directoras de tesis, Irene Epifanio, Amelia Simó y Maite León, por todo lo que me han enseñado y ayudado estos años en todos los trabajos y actividades que hemos realizado y, a nivel personal, por su apoyo, confianza y calidez.

De igual forma, quiero expresar mi gratitud a Toia Ibañez por su cercanía y por su ayuda y sugerencias de mejora en el trabajo de la profundidad.

Muchas gracias también a Sandra Alemany, por habernos ayudado a interpretar todos los resultados que hemos obtenido, por habernos enviado tanta bibliografía, así como por haberme introducido en el IBV, en el que siempre que he estado, me he sentido como uno más.

Estoy muy agradecido del mismo modo a Juan Domingo, por haberme ayudado siempre que lo he necesitado en todas las dudas y cuestiones relacionadas con el software que hemos utilizado, sobre todo con Linux, su consola y con nuestro servidor johnford. Juan, muchas gracias.

Por último, quiero mencionar a tod@s mis compañer@s y amig@s del departamento, facultad e IBV, por los buenos momentos vividos y compartidos.

Esta tesis doctoral ha sido financiada mediante una beca FPI (Formación de Personal Investigador) correspondiente a la convocatoria del año 2010, con referencia BES-2010-035842 y asociada al proyecto "Aplicación de técnicas morfométricas al diseño y evaluación funcional de indumentaria", con código TIN2009-14392-C02-01 y TIN2009-14392-C02-02.

Preface

Ergonomics is the scientific discipline that studies the interactions between human beings and the elements of a system and presents multiple applications in areas such as clothing and footwear design or both working and household environments. In each of these sectors, knowing the anthropometric dimensions of the current target population is fundamental to ensure that products suit as well as possible most of the users who make up the population. Anthropometry refers to the study of the measurements and dimensions of the human body and it is considered a very important branch of Ergonomics because its considerable influence on the ergonomic design of products [162].

Human body measurements have usually been taken using rules, calipers or measuring tapes. These procedures are simple and cheap to carry out. However, they have one major drawback: the body measurements obtained and consequently, the human shape information, is imprecise and inaccurate. Furthermore, they always require interaction with real subjects, which increases the measure time and data collecting. The development of new three-dimensional (3D) scanning techniques has represented a huge step forward in the way of obtaining anthropometric data. This technology allows 3D images of human shape to be captured and at the same time, generates highly detailed and reproducible anthropometric measurements.

The great potential of these new scanning systems for the digitalization of human body has contributed to promoting new anthropometric studies in several countries, such as United Kingdom, Australia, Germany, France or USA, in order to acquire accurate anthropometric data of their current population. In this context, in 2006 the Spanish Ministry of Health commissioned a 3D anthropometric survey of the Spanish female population, following the agreement signed by the Ministry itself with the Spanish associations and companies of manufacturing, distribution, fashion design and knitted sec-

tors. A sample of 10415 Spanish females from 12 to 70 years old, randomly selected from the official Postcode Address File, was measured. The two main objectives of this study, which was conducted by the Biomechanics Institute of Valencia, were the following: on the one hand, to characterize the shape and body dimensions of the current Spanish women population to develop a standard sizing system that could be used by all clothing designers. On the other hand, to promote a healthy image of beauty through the representation of suited mannequins [3]. In order to tackle both objectives, Statistics plays an essential role. Thus, the statistical methodologies presented in this PhD work have been applied to the database obtained from the Spanish anthropometric study.

Clothing sizing systems classify the population into homogeneous groups (size groups) based on some key anthropometric dimensions. All members of the same group are similar in body shape and size, so they can wear the same garment. In addition, members of different groups are very different with respect to their body dimensions. An efficient and optimal sizing system aims at accommodating as large a percentage of the population as possible, in the optimum number of size groups that better describes the shape variability of the population. Besides, the garment fit for the accommodated individuals must be as good as possible. A very valuable reference related to sizing systems is the book *Sizing in clothing: Developing effective sizing systems for ready-to-wear clothing*, by Susan Ashdown [7]. Each clothing size is defined from a person whose body measurements are located toward the central value for each of the dimensions considered in the analysis. The central person, 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.

Clustering is the statistical tool that divides a set of individuals in groups (clusters), in such a way that subjects of the same cluster are more similar to each other than to those in other groups [115]. In addition, clustering defines each group by means of a representative individual. Therefore, it arises in a natural way the idea of using clustering to try to define an efficient sizing system. Specifically, four of the methodologies presented in this PhD thesis aimed at segmenting the population into optimal sizes, use different clustering methods. The first one, called *trimowa*, has been published in *Expert Systems with Applications* [102]. It is based on using an especially defined distance to examine differences between women regarding their body measurements. The second and third ones (called *biclustAnthropom* and *TDDclust*, respectively) will soon be submitted in the same paper [214]. *BiclustAnthropom* adapts

to the field of Anthropometry a clustering method addressed in the specific case of gene expression data [140]. Moreover, *TDDclust* uses the concept of statistical depth [132] for grouping according to the most central (deep) observation in each size. As mentioned, current sizing systems are based on using an appropriate set of anthropometric dimensions, so clustering is carried out in the Euclidean space. In the three previous proposals, we have always worked in this way. Instead, in the fourth and last approach, called *kmeansProcrustes*, a clustering procedure is proposed for grouping taking into account the women shape, which is represented by a set of anatomical markers (landmarks). For this purpose, the statistical shape analysis [47] will be fundamental. This contribution has been submitted for publication [216].

A sizing system is intended to cover the so-called "standard" population, discarding the individuals with extreme sizes (both large and small). In mathematical language, these individuals can be considered outliers. An outlier is an observation point that is distant from other observations. In our case, a person with extreme anthopometric measurements would be considered as a statistical outlier. Clothing companies usually design garments for the standard sizes so that their market share is "optimal". Nevertheless, with their foreign expansion, a lot of brands are spreading their collection and they already have a special sizes section. In last years, Internet shopping has been an alternative for consumers with extreme sizes looking for clothes that follow trends. The custom-made fabrication is other possibility with the advantage of making garments according to the customers' preferences. The four aforementioned methodologies (*trimowa*, *biclustAnthropom*, *TDDclust* and *kmeansProcrustes*) have been adapted to only accommodate the "standard" population.

Once a particular garment has been designed, the assessing and analysis of fit is performed using one or more fit models. The fit model represents the body dimensions selected by each company to define the proportional relationships needed to achieve the fit the company has determined. The definition of an efficient sizing system relies heavily on the accuracy and representativeness of the fit models regarding the population to which it is addressed. In this PhD work, a statistical approach is proposed to identify representative fit models. It is based on another clustering method originally developed for grouping gene expression data. This method, called *hipamAnthropom*, has been published in *Decision Support Systems* [215]. From well-defined fit models and prototypes, representative and accurate mannequins

of the population can be made.

Unlike clothing design, where representative cases correspond with central individuals, in the design of working and household environments, the variability of human shape is described by extreme individuals, which are those that have the largest or smallest values (or extreme combinations) in the dimensions involved in the study. This is often referred to as the accommodation problem. A very interesting reference in this area is the book entitled *Guidelines for Using Anthropometric Data in Product Design*, published by The Human Factors and Ergonomics Society [93]. The idea behind this way of proceeding is that if a product fits extreme observations, it will also fit the others (less extreme). To that end, in this PhD thesis we propose two methodological contributions based on the statistical archetypal analysis. An archetype in Statistics is an extreme individual that is obtained as a convex combination of other subjects of the sample [37]. The first of these methodologies has been published in *Computers & Industrial Engineering* [54], whereas the second one has been submitted for publication [213].

The outline of this PhD report is as follows:

Chapter 1 reviews the state of the art of Ergonomics and Anthropometry and introduces the anthropometric survey of the Spanish female population.

Chapter 2 presents the *trimowa*, *biclustAnthropom* and *hipamAnthropom* methodologies.

In Chapter 3 the *kmeansProcrustes* proposal is detailed.

The *TDDclust* methodology is explained in Chapter 4.

Chapter 5 presents the two methodologies related to the archetypal analysis.

Since all these contributions have been programmed in the statistical software R [165], Chapter 6 presents the **Anthropometry** R package [212], that brings together all the algorithms associated with each approach.

In this way, from Chapter 2 to Chapter 6 all the methodologies and results included in this PhD thesis are presented.

At last, Chapter 7 provides the most important conclusions.

Resumen

Objetivos

La Ergonomía es la disciplina que estudia las interacciones entre los seres humanos y los elementos de un sistema y presenta numerosas aplicaciones en ámbitos como el diseño de indumentaria y calzado o entornos tanto laborales como relacionado con el hogar. En cada uno de estos sectores, el conocimiento de las dimensiones antropométricas de la población objetivo actual es fundamental para que los productos que van a ser desarrollados se adapten lo mejor posible a la mayor parte de los usuarios que componen dicha población. La Antropometría se refiere al estudio de las medidas y dimensiones del cuerpo humano y está considerada como una rama muy importante de la Ergonomía por su considerable influencia en el diseño ergonómico del producto [162].

Las medidas corporales de las personas han sido habitualmente tomadas utilizando reglas, calibradores o cintas métricas. Estos procedimientos son sencillos y baratos de utilizar. Sin embargo, presentan una gran desventaja: cada medida antropométrica extraída y, en consecuencia la información relacionada con la forma de las personas, es imprecisa e inexacta. Además, requieren siempre la interacción con sujetos reales, lo que incrementa el tiempo de toma de medidas y de recogida de datos. El desarrollo de nuevas técnicas de escaneado tridimensional (3D) ha supuesto un gran paso adelante en la manera de obtener datos antropométricos. Esta tecnología permite capturar las imágenes 3D de las formas de las personas que son escaneadas y al mismo tiempo genera con una gran precisión sus medidas antropométricas, que además son reproducibles.

El gran potencial de estos nuevos sistemas de escaneado para la digitalización del cuerpo humano ha contribuido a llevar a cabo nuevos estudios antropométricos en diferentes países, como por ejemplo, EEUU, Alemania, el Reino Unido, Francia o Australia, con el fin de obtener datos antropométricos

precisos de sus actuales poblaciones. En este contexto, el Ministerio de Sanidad y Consumo del Gobierno de España promovió en 2006 el Estudio Antropométrico de la Población Femenina en España, tras el acuerdo suscrito por el propio Ministerio con las asociaciones y empresas de los sectores de confección, distribución, diseño de moda y género de punto en España. En total, se recogieron las medidas corporales de 10.415 mujeres españolas, con un rango de edad comprendido entre los 12 y 70 años.

Los dos objetivos principales de este Estudio Antropométrico, el cual fue realizado por el Instituto de Biomecánica de Valencia, eran los siguientes: por un lado, caracterizar la forma y dimensiones del cuerpo de la actual población de mujeres españolas para desarrollar un sistema de tallaje estándar que pudiera ser utilizado por todos los diseñadores de ropa. Por otro, fomentar una imagen de belleza saludable mediante la representación de maniquíes adecuados a las dimensiones reales de la misma población [3].

Para acometer ambos objetivos, la Estadística juega un papel esencial. De este modo, las metodologías estadísticas que se presentan en esta tesis doctoral han sido aplicadas sobre la base de datos de este Estudio.

Metodología

Un sistema de tallaje de ropa clasifica a la población en grupos homogéneos utilizando el conjunto de dimensiones antropométricas que se consideran más relevantes para tal fin. Todos los miembros de un mismo grupo (talla) son similares en su forma y tamaño corporal, por lo que deberían poder usar la misma prenda de ropa. Asimismo, los miembros de distintos grupos son muy distintos en sus dimensiones corporales. Todo sistema de tallaje óptimo y eficiente tiene como objetivo acomodar el máximo porcentaje posible de la población, en el menor número de tallas posible que mejor describan la variabilidad en la forma de la misma población. Además, el ajuste de las prendas de ropa para los individuos cubiertos por el sistema de tallaje debe ser el mejor posible.

Cada talla de ropa se define a partir de una persona cuyas medidas corporales están localizadas alrededor del valor central para cada una de las dimensiones consideradas en el análisis. Dicha persona central (llamada prototipo), a la que se considera como la representante de esa talla, se convierte en el patrón básico a partir del cual se diseña la línea de ropa en esa misma talla. Una referencia muy valiosa sobre sistemas de tallaje es el libro *Sizing*

in clothing: Developing effective sizing systems for ready-to-wear clothing, escrito por Susan Ashdown [7].

El análisis de conglomerados (en inglés, clustering) es la herramienta estadística que divide un conjunto de elementos en grupos (clusters) de manera que los individuos de un mismo grupo sean tan similares entre sí como sea posible y al mismo tiempo, tan diferentes a los miembros de los demás grupos como sea posible [115]. Además, los métodos clustering permiten definir cada grupo mediante un individuo representativo. Por lo tanto, nos parece natural tratar de definir un sistema de tallaje óptimo utilizando técnicas clustering. En concreto, cuatro de las metodologías que se presentan en esta memoria con el objetivo de segmentar a la población en tallas eficientes, utilizan diferentes métodos de clustering.

El primero de estos métodos, al que llamaremos *trimowa*, ha sido publicado en *Expert Systems with Applications* [102]. Se basa en utilizar una distancia especialmente definida para estudiar las diferencias entre mujeres de acuerdo a sus medidas corporales. El segundo y tercer método (denominados *biclustAnthropom* y *TDDclust*, respectivamente) serán próximamente sometidos en un mismo artículo [214]. *BiclustAnthropom* adapta al ámbito de la Antropometría un método clustering desarrollado para trabajar con datos de expresión de genes [140]. Por su parte, *TDDclust* utiliza el concepto de profundidad estadística [132] para agrupar en función de las mujeres más centrales (profundas) en cada talla. Como se ha comentado anteriormente, los sistemas de tallaje actuales se basan en utilizar un conjunto adecuado de medidas antropométricas, por lo que los procedimientos clustering se desarrollan en un espacio Euclídeo. En las tres anteriores propuestas se ha trabajado de esta manera. Por el contrario, en el cuarto y último enfoque, llamado *kmeansProcrustes*, proponemos un procedimiento clustering que permite agrupar según la forma de las mujeres, la cual se representa por una serie de marcadores anatómicos (en inglés, landmarks). Para ello, el análisis estadístico de formas (statistical shape analysis, en inglés) [47] será básico. Esta última contribución se encuentra sometida [216].

Todo sistema de tallaje está pensado para cubrir a la población conocida como "estándar", en la que se han eliminado aquellos individuos con tallas extremas (tanto grandes como pequeñas). En el lenguaje matemático, estos individuos pueden ser considerados outliers. Un outlier es una observación que es numéricamente distante del resto de los datos. En nuestro caso de estudio, una persona que tuviera medidas antropométricas extremas se consideraría un outlier estadístico. La política habitual de las empresas de ropa

consiste en diseñar prendas para las tallas "estándar" con el fin de optimizar su cuota de mercado. Sin embargo, con la expansión al extranjero, muchas marcas están ampliando su colección de tallas y ya cuentan con una sección de tallas especiales. En los últimos años, la compra por Internet ha supuesto una alternativa para los consumidores con tallas extremas que buscan ropa que siga las tendencias. La fabricación artesanal y a medida es otra opción, con la ventaja de poder confeccionar las prendas al gusto del cliente.

Las cuatro metodologías comentadas anteriormente (*trimowa*, *biclustAnthropom*, *TDDclust* y *kmeansProcrustes*) se han adaptado para acomodar únicamente a la "población estándar".

Una vez una determinada prenda ha sido diseñada, la evaluación y análisis de su ajuste se realiza utilizando uno o varios modelos llamados modelos de ajuste o, en inglés, fit models. Un modelo de ajuste representa las proporciones y medidas corporales que el/la diseñador/a de una empresa de ropa ha especificado para fabricar y producir sus prendas con el ajuste que él/ella mismo/a ha fijado. La definición de un sistema de tallaje eficiente depende en gran medida de lo representativos y precisos que sean los modelos de ajuste de la población a la que va dirigido.

En este trabajo doctoral, se propone una metodología para identificar modelos de ajuste representativos de la población basada en otro método clustering, el cual fue creado originalmente para agrupar datos de expresión de genes. Este método, al cual llamaremos *hipamAnthropom*, ha sido publicado en *Decision Support Systems* [215]. A partir de tanto un modelo de ajuste bien definido, como de un prototipo, se pueden construir maniquíes adecuados a la población.

A diferencia del diseño de indumentaria, donde los casos representativos se corresponden con individuos centrales, en el diseño de entornos de trabajo o domésticos, la variabilidad de la forma humana se describe mediante individuos extremos, que son aquellos que presentan los valores más grandes, más pequeños o combinaciones extremas, en las dimensiones implicadas en el estudio. Esto es lo que se conoce como el problema de acomodación. Una referencia muy interesante en este área es el libro titulado *Guidelines for Using Anthropometric Data in Product Design*, publicado por The Human Factors and Ergonomics Society [93]. La idea detrás de esta manera de proceder consiste en que si un producto se acomoda a los individuos extremos de la población, también se acomodará los demás. Para tal fin, se proponen otras dos aportaciones metodológicas basadas en el análisis estadístico de arquetipos. Un arquetipo en Estadística es un individuo extremo que se

obtiene como una combinación convexa de los demás sujetos de la muestra [37]. La primera de estas metodologías ha sido publicada en *Computers & Industrial Engineering* [54], mientras que la segunda se encuentra sometida [213].

Todos los algoritmos asociados a los métodos presentados en esta tesis doctoral se han reunido en un paquete de R llamado **Anthropometry** [165, 212], el cual está disponible para cualquier usuario en el CRAN de R, http://cran.r-project.org/package=Anthropometry.

La estructura de la tesis se indica a continuación:

En el capítulo 1 se introduce el estado del arte de las disciplinas de la Ergonomía y la Antropometría y se detallan los aspectos básicos del Estudio Antropométrico de la Población Femenina en España.

En el capítulo 2 se presentan los métodos *trimowa*, *biclustAnthropom* y *hipamAnthropom*.

En el capítulo 3 se detalla la propuesta *kmeansProcrustes*.

La metodología *TDDclust* se explica en el capítulo 4.

El capítulo 5 incluye las dos metodologías relacionadas con el análisis estadístico de arquetipos.

En el capítulo 6 se presenta el paquete de R **Anthropometry**.

Por tanto, los capítulos 2, 3, 4, 5 y 6 presentan toda la metodología y los resultados incluidos en esta tesis doctoral.

Por último, el capítulo 7 recoge las conclusiones más importantes de este trabajo.

Conclusiones

La presente tesis doctoral se ha planteado con el fin de ser una aportación científica rigurosa, desde el punto de vista matemático y estadístico, a las disciplinas de la Ergonomía y de la Antropometría. A lo largo de esta memoria, se han desarrollado diferentes metodologías estadísticas que puedan ser de utilidad para mejorar el diseño ergonómico del producto, en especial para el diseño óptimo y eficiente de ropa y de lugares de trabajo.

Los sistemas de tallaje de ropa utilizados hoy en día por la industria de indumentaria no se encuentran optimizados para ajustar de manera correcta a la población a la que van dirigidos. Como consecuencia, una gran parte de dicha población, especialmente entre las mujeres, no encuentra ropa que le acomode bien (que le *venga* bien), incluso tras probarse varias prendas. Esta

situación genera que las tiendas de ropa acumulen una gran cantidad de ropa sin vender, incluyendo aquellas prendas que han sido devueltas por el consumidor porque no quedó satisfecho/a con su compra. Un llamativo efecto de esta circunstancia en España es la proliferación de tiendas *outlet*. Uno de los principales problemas para desarrollar nuevos patrones y diseños de ropa es la falta de datos actualizados de la población actual. La utilización de tablas de tallaje basadas en medidas de la población anticuadas provoca una gran diferencia en el tallaje ofrecido por las distintas compañias de ropa. En este contexto, el Instituto Nacional de Consumo (INC) del Ministerio de Sanidad y Consumo del Gobierno de España firmó un acuerdo con las principales empresas de ropa españolas para llevar a cabo el Estudio Antropométrico de la Población Femenina en España, el cual fue realizado en 2006 por el Instituto de Biomecánica de Valencia. La información antropométrica obtenida se generó tanto de manera unidimensional (1D) como tridimensional (3D). La principal motivación era caracterizar la forma y dimensiones del cuerpo de la población femenina actual de España.

Cada uno de los nuevos enfoques estadísticos aquí presentados ha utilizado los datos de este Estudio Antropométrico. De esta manera, uno de los objetivos esenciales de este trabajo doctoral ha consistido en crear y desarrollar técnicas matemáticas y estadísticas que permitan explotar grandes bases de datos del cuerpo humano orientadas al diseño ergónomico del producto. Del mismo modo, esta tesis doctoral se enmarca dentro de las actividades realizadas por el proyecto de investigación relacionado con el Estudio Antropométrico de la Población Femenina en España.

En Ergonomía y Antropometría, la variabilidad en el tamaño corporal dentro de la población objetivo se caracteriza mediante la definición de un número concreto de *casos antropométricos*. Un caso puede ser, o bien un ser humano en particular, o una combinación de medidas corporales. En función del problema, hay tres tipos de casos: centrales, extremos o distribuidos. Los métodos propuestos en esta tesis persiguen identificar tanto casos centrales como extremos.

Las metodologías desarrolladas utilizando algoritmos de agrupamiento (o clustering), es decir, *trimowa*, *biclustAnthropom*, *TDDclust*, *kmeansProcrustes* y *hipamAnthropom*, permiten definir casos centrales 1D y 3D, los cuales se corresponden con modelos estadísticos o prototipos (y fit models o modelos de ajuste en el caso de *hipamAnthropom*) del cuerpo humano que representan a la población objetivo. Tanto los modelos de ajuste como los prototipos pueden ser utilizados para fabricar maniquíes de pasarela y es-

caparate adecuados a las dimensiones de la población real, que ayuden a fomentar una imagen de belleza saludable. Las cinco propuestas comentadas anteriormente han seguido la misma línea metodológica. En primer lugar, el conjunto de datos seleccionado se segmenta utilizando una primera dimensión control (el perímetro de busto en el caso de *trimowa*, *TDDclust*, *kmeansProcrustes* y *hipamAnthropom* y el perímetro de cintura en el caso de *biclustAnthropom*). A continuación, se lleva a cabo una subsiguiente partición utilizando otras variables antropométricas control secundarias. De este modo, la primera segmentación permite al usuario elegir su talla de una manera sencilla, mientras que los grupos resultantes basados en el busto (o la cintura) y otras dimensiones corporales optimizan el tallaje. Mediante la aplicación de un enfoque puramente estadístico como es el clustering, se obtienen grupos homogéneos teniendo en cuenta la variabilidad antropométrica de esas dimensiones secundarias que influyen de manera relevante en el ajuste de las prendas. Cada uno de estos métodos ha sido adaptado para acomodar únicamente a la población estándar. Para elegir las variables control primarias y secundarias se utilizó la norma EN 13402-3-2004 [59]. Este texto fue elaborado por la Unión Europea y pretende ser una guía de orientación para la industria textil. El texto, el cual no es de obligado cumplimiento, promueve la implantación de un sistema de tallaje basado en la consideración de tres variables: perímetro de busto, cintura y cadera, en función de la estatura.

Por otro lado, las metodologías basadas en el análisis estadístico de arquetipos permiten identificar casos extremos, es decir aquellos individuos que presentan medidas corporales extremas. La idea básica de este procedimiento radica en que acomodar estos casos extremos permitirá la acomodación del resto de la población (con unas medidas menos extremas). Esta estrategia es de gran valor en todos aquellos problemas de interacción hombre-máquina, como por ejemplo, el diseño de cabinas de aviones o camiones (los así llamados problemas de acomodación). Cuando se diseñan estaciones de trabajo, es común emplear solamente un número reducido de modelos humanos (que son casos extremos) como modelos de prueba virtuales. En esta tesis doctoral, hemos demostrado que el análisis de arquetipos representa una mejor alternativa para determinar casos extremos, en comparación con el enfoque habitualmente utilizado basado en el análisis de componentes principales (PCA). A diferencia del PCA, el análisis de arquetipos asegura alcanzar el nivel deseado de acomodación. Además, el usuario puede decidir el número de arquetipos que desea calcular, tanto de manera subjetiva como utilizando

un criterio matemático. En la literatura, hay un debate en curso acerca de si los arquetipos deben corresponderse con una observación real, puesto que en el análisis de arquetipos tradicional los arquetipos pueden ser individuos reales o no. Sin embargo, en algunos problema es crucial que los arquetipos sí sean observaciones concretas de la muestra. En este trabajo doctoral, se ha introducido un nuevo concepto arquetípico para abordar este problema: el arquetipoide. Se ha presentado un algoritmo eficiente para calcularlos y se ha demostrado algunas de sus ventajas con respecto a los arquetipos clásicos. El análisis de arquetipos y arquetipoides podría suponer una mejora en aquellas prácticas de la industria en las que se utilizan modelos humanos para el diseño de productos y entornos de trabajo.

Todos los algoritmos computacionales asociados a los métodos presentados en esta memoria se han recopilado en un paquete de R llamado **Anthropometry**, el cual se puede descargar libremente desde el CRAN de R, http://cran.r-project.org/package=Anthropometry.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: anthropometric data and statistical methods

1.1 Motivation

When designing a product that is going to be used by many different users, such as clothes or workstations, engineers and statisticians working in Ergonomics must have databases with relevant anthropometric measurements of the so-called target population or target audience. The main goal of researchers in this field is to achieve a successful fit, understood as *the best possible* fit between the product and its users, taking into account other criteria such as task performance, ease of use and comfort [162]. Anthropometry plays a prominent role in achieving this purpose because variations in body dimensions have a great influence on fit and other criteria [100]. To ensure the intended level of accommodation, it is necessary to consider anthropometric diversity [20, 14]. Indeed, anthropometric measurements are the basic and mandatory information for a suitable design and manufacture of industrial products [87]. The type of anthropometric data varies according to the product to be designed and the field of application [170].

An anthropometric database is a collection of body dimensions taken from a sample of people [100]. A major issue when developing new products and equipment that adapt to the current population and that fit well is the lack of up-to-date anthropometric data. Improvements in health, nutrition and living conditions and the transition to a sedentary life style have changed the body dimensions of people over recent decades. The data measurements of

anthropometric databases must therefore be updated regularly. Traditionally, the same set of anthropometric tools has been used to measure a body manually [189, 137, 185]. As expounded in [188], this set consists of seven instruments: a camera for photographing subjects, a weighing scale for obtaining weight, a tape measure for measuring circumferences and curvatures, an anthropometer for determining height and different traverse diameters of the body, a caliper for measuring diameters, a sliding compass for measuring short diameters and a head spanner for measuring head height. These procedures are simple (user-friendly), non-invasive and cheap. The measurements obtained are often called traditional measurements because the instruments used to take them have been used for hundreds of years [169]. Nevertheless, they have important methodological problems. Manual measurements are usually affected by several potential sources of error, so the set of measurements and consequently the shape information, is imprecise and inaccurate. In addition, this process is time-consuming since it involves interaction with real subjects [218, 137].

1.2 Novel scanning methods and anthropometric surveys

In recent years, advances and improvements in modern optical technologies, such as new 3D body scanner measurement systems, has opened up a great range of opportunities for collecting and updating anthropometric data. 3D full body scanning technologies have been successfully applied to the measurement and scanning of the human body in several industrial sectors. They provide accurate and reproducible anthropometric data from which 3D shape images of the people being measured can be obtained [107, 128, 218, 38]. Another important advantage is the speed of the process. Several studies have been published comparing 3D body scan measurements and manual measurements, as discussed in [88]. The rapidly emerging 3D body scanning techniques constitute a true breakthrough in realistically characterizing people and they have made it possible to conduct new large-scale size surveys in different countries (for instance, in the USA, France, the UK, Germany and Australia).

Most of the anthropometric surveys performed in the past included only military populations and surveys of civilians were quite unusual [171]. The

Civilian American and European Surface Anthropometry Resource study (CAESAR study) was the first anthropometric survey to provide 3D human models aimed at civil society [168]. Data were collected from North America, the Netherlands and Italy. SizeUK was the first national survey of the UK adult population since the 1950's and used 3D body scanners to extract measurements [61]. In the USA, a comprehensive sizing survey of the U.S. population was done in 2004. It was the first such survey in over 40 years [190, 157]. In this particular context, the Spanish Ministry of Health sponsored a 3D anthropometric study of the Spanish female population in 2006 [3].

1.3 Anthropometric survey of the Spanish female population

The Spanish National Institute of Consumer Affairs (INC according to its Spanish acronym) of 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 main 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. The two main objectives of the project 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.

The target sample was made up of 10415 women grouped into 10 age groups ranging from 12 to 70 years, see Table 1.1.

Age group	1	2	3	4	5	6	7	8	9	10
Range	12-13	14-15	16-17	18-19	20-24	25-29	30-39	40-49	50-59	60-70

Table 1.1: Definition of the 10 age groups in which the women were divided.

They were randomly selected from the official Postcode Address File (the census of the population of each town and city) in 61 different locations that represent the seven Spanish NUT areas (statistical territorial units defined by the European Community), see Fig. 1.1.

Figure 1.1: Spanish locations where the women's measurements were taken.

Women enrolled in the study were scanned using a Vitus Smart 3D body scanner from Human Solutions, a non-intrusive laser system consisting of four columns containing the optical system, which moves from head to feet in ten seconds performing a sweep of the body. Two scanning postures were registered for each subject (standard and standing) and the sitting posture was scanned for a random percentage of 25%, see Fig. 1.2.

Figure 1.2: Scanning postures: standard (left), standing (center) and sitting (right).

Associated software provided by the scanner manufacturers made a triangulation providing 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 [101]. The location is extracted by translating each image to the origin in such a way that its centroid coincides with the origin. Each trunk is also rotated to make its principal inertia axis coincide with the canonical axis of coordinates.

All the women wore a standard white garment, a swimming cap, a top and shorts, see Fig. 1.3. These items were designed and scaled in 5 sizes in order to harmonize the measurements. From the 3D mesh, 95 body measurements were extracted using a semi-automatic methodology that combined automatic measurements (Fig. 1.4) with a manual review (Fig. 1.5).

Figure 1.3: Standard garments worn by the women in the measuring process.

Figure 1.4: Automatic measuring process.

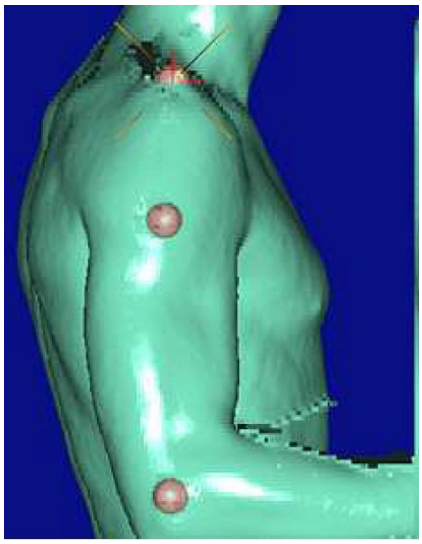


Figure 1.5: Semi-automatic detection of anthropometric markers.

In addition, the shape of all the women was represented by a set of points that were comparable between individuals, called landmarks. These landmarks were placed in three different ways:

- Automatic landmarks: automatically calculated with scanner program algorithms, based on geometrical features of the body.
- Manual landmarks: points which are not reflected on the external body geometry; they were located through palpation by expert personnel and identified by a physical marker.
- Digital landmarks: detected on the computer screen in the 3D scanned image. They are not robust on the automatic calculation but are easy to detect on the screen.

Furthermore, a socio-demographic questionnaire was included, collecting the user characterization (her age, weight, place of birth, etc.), aspects related to health and eating habits, the woman's satisfaction with her body and questions about clothes shopping habits. The important point to note here is that women were also asked about their size in the current Spanish sizing system. Because of the lack of rigor in its definition, their answers were in some cases numerical and in other cases qualitative (small, large, etc.) and

was only an approximation in all cases. This study was presented in more detail in [3].

1.4 Objectives of Anthropometry (sizing systems) and Ergonomics (accommodation problem)

In the design process, anthropometric variability among users is commonly summarized using a small number of cases that accommodate a certain percentage of the population [111]. A case represents the most relevant combination of body dimensions for a particular design problem. Depending on the product being designed and according to the distribution of the set of dimensions chosen, three types of strategies can be distinguished for searching for cases [100]:

1. *Central cases*: points located toward the middle of the distribution of dimensions selected. These cases are required for designing a multiple-size product (n sizes to fit n groups of people within a designated accommodation percentage of the population), apparel sizing system design being the most common application. Central cases represent the basic proportions in a clothing line and they are critical in defining a sizing system.

2. *Boundary cases*: determined on the boundary of the distribution of dimensions, that is to say, they are extreme cases. In designing a one-size product (one-size to fit people within a specified portion of accommodation) such as working environments or the passenger compartment of any vehicle including aircraft cockpits, the representative subjects are boundary cases. These design problems fall into a more general category: the accommodation problem. If the product is properly designed to fit the extreme individuals, then it will also fit other less extreme users. Therefore, it is worth pointing out that these cases can also be used in apparel design in combination with central cases.

3. *Distributed cases*: spread throughout the distribution. Central and boundary cases can be considered particular distributed cases. However, distributed cases need not include the former two cases.

In this PhD thesis, we aim to define central and boundary cases to tackle the apparel sizing system design problem and the workplaces design problem (focusing on the particular case of an aircraft cockpit).

1.5 Sizing systems: multiple-size products

1.5.1 Background of sizing systems

The use of anthropometric databases to enhance apparel design and fit has mainly been aimed at defining new sizing systems. A sizing system divides a population into homogeneous subgroups with similar body measurements (size groups), in such a way that all individuals in a size group can wear the same garment [7, 34]. In current sizing systems, the body dimensions used to obtain the size groups are called control dimensions or key dimensions. The primary control dimension separates the set of individuals into major size groups along the anthropometric measurement that is considered the most important dimension for designing a particular garment. Next, these major size groups are divided into subgroups according to a secondary control dimension that is considered the second most important dimension for the same garment (more than one secondary control variable can be used). Each subgroup is described by certain values of the control dimensions selected, thus a body of specific proportions, called body shape or body type, is defined. Of course, subgroups can be further split according to a tertiary control dimension, etc. Each further subdivision of the groups identifies the body shape of the size group more closely. The set of values of each control dimension to be covered in the sizing chart is called the size range along that control dimension. The percentage of the population that is accommodated by the sizing system is called the accommodation rate of the sizing system. The size range of each control dimension is divided into a set of sizes known as the size scale. The size scale depends on the increment between adjacent sizes. This increment is called the size interval, size step or size grade and can have a fixed or variable value. Once the size groups are determined, the values of certain other body dimensions needed for garment manufacturing can be added to the sizing chart. These additional measurements are called secondary dimensions (not to be confused with secondary control dimensions). Finally, the coding system used to identify the body dimensions for which the garment was designed is called size designation (or better still,

labeling). The combination of the size designation method and the sizing charts constitutes the sizing system, size roll or tariff system of the specified garment. In particular, a sizing system is specified with a table of numbers that details the values of the body dimensions used to define each size group. Table 1.2 summarizes the decision steps to be made when defining a sizing chart.

1. Primary and secondary control dimensions to classify the population.
2. Range of values covered by each control dimension (size range and size scale).
3. Division of the size scale of each variable into segments (size interval).
4. Number of size groups to produce.
5. Additional secondary dimensions that are relevant for creating the garment.
6. Labeling to identify the dimensions of each size group.

Table 1.2: Decisions to be made when defining a sizing system for a specified type of garment, see [7].

Most manufacturers from different apparel companies create and adjust their own size charts by trial and error using the information collected from sales studies, returned goods reports and small-scale customer surveys. Further changes to the dimensions of garments of specific sizes are made in a stepwise manner, often without adapting the size designation.

The sizing systems resulting from this process are the reason for the lack of fit of the clothes that companies offer, a large amount of unsold and returned garments and a less competitive business [30]. Furthermore, because each apparel company creates its own sizing system, garments from one company may fit differently to those of another with the same size-label. All makes for a highly unsatisfactory and confusing clothes shopping experience for customers. In order to remedy this situation, several standardization organizations have arisen, which propose a regulation of the sizing system. In 1968 the Swedish member body of the International Organization for Standardization (ISO) proposed that a Technical Committee (ISO/TC 133, which was established in 1969) should be set up to create a global sizing system for clothing [221, 7]. This committee reached the conclusion that developing a unique sizing system to accommodate the world's population would be ineffective because of the variability inherent to different countries or ethnic groups. However, they concentrated on deciding which the most relevant elements were for defining a sizing system. In 1991, ISO/TC 133 published a report providing the preferred control dimensions, values and intersize in-

tervals for defining a sizing system based on anthropometric data from a particular population [103, 7]. Several countries have revised their size designation systems in accordance with the standards published by ISO. The European Committee for Standardization has developed several standards (EN 13402 Size Designation of Clothes) from the ISO standards. In this report, we will use two of those EN reports to propose the definition of an efficient standard sizing system which may be used for every Spanish apparel company: the *Size designation of clothes. Part 2: Primary and secondary dimensions* [58] and the *Size designation of clothes. Part 3: Measurements and intervals* [59].

1.5.2 Background of fit models

The final evaluation of garment fit requires models to test every new design before the production phase. These models are the dress form, the human fit model and the virtual fit model. Of these three, the human fit model plays the most important role. Companies try to enhance the quality of fit by scanning their fit models and deriving dress forms from those scans [7, 193]. The fit model represents the commercial measurements established by each company to define the proportional relationships needed to achieve the company's fit [225]. Beyond merely wearing the garment for examination, a fit model is a person who provides objective feedback about fit, movement, comfort and visual appearance of a garment in place of the consumer. A fit model therefore acts as a live mannequin.

The current practice in apparel fit analysis is based on using expert panels [8]. An expert panel is an experienced working team that judges the fit of a garment. In the apparel industry, fit analysis is tested with a live fit model. Almost every apparel company develops its own sizing system by using a different fit model which covers their whole target market [224]. This means that apparel companies only attempt to fit one body type, generating base patterns and grade rules that match the proportions of their fit model [6]. However, there might be many shapes and body types within a size and this single idealized fit model may not adequately address the differences between them [7, page 133]. Furthermore, there is little information available to help choose a fit model whose body size and shape are consistent with the body characteristics of the target market [6].

During recent years, research has been done to examine the reliability of using virtual 3D scan models instead of fit models to improve garment fit

[8, 24]. These virtual 3D models come from scanned live fit models. They can be used similarly to fit models but offer many benefits in different areas of apparel design and manufacture [157, 8]. Indeed, the tailoring procedure followed by fashion designers and manufacturers needs real individuals to be scanned to generate 3D clothes from 2D patterns [142, 217]. Consequently, a representative fit model of the target population, whether a live fit model or a 3D scan model, is critical for improving garment fit and has become an integral part of the design process. Good fit models are basic for defining an accurate sizing system.

1.5.3 Literature review and our statistical proposals

Three types of approaches can be distinguished for creating a sizing system: traditional step-wise sizing, multivariate methods and optimization methods. The main difference between the traditional approach regarding multivariate and optimization methods is that the size groups that it defines form a fixed regular pattern along each control dimension, while the other approaches define size groups that are spaced randomly (without constraints) in the space defined by the key dimensions. Traditional methods use bivariate distributions to define a sizing chart and cross tabulation to select the sizes gradually, covering the highest percentage of population. The size interval is set according to common practice or fit and style considerations of the designers. This approach is too simplistic. It is not possible to cover the different body types of the population because other relevant anthropometric dimensions are not considered.

More recently, more advanced mathematical methods have been developed. From the statistical point of view, Principal component analysis (PCA) and clustering methods have been widely used. PCA has been used as a dimensionality reduction technique. The usual procedure consists of selecting the first two principal components that explain the bulk of the data variance and generating the bivariate distribution in which to define the sizing chart [85, 97, 138, 179]. Partitioning clustering methods, especially the k-means algorithm, have been used to classify the target population into different morphologies by using every anthropometric measurement available as an input [96, 34, 230, 155, 9]. Other alternatives combining data mining and decision trees have also been proposed [98].

The first proposal using an optimization method was put forward by Peter Tryfos in [204]. He developed an integer programming procedure to

optimize the number of sizes in such a way garment sales were maximized. An alternative to Tryfos'proposal was introduced in [147], where a nonlinear optimization technique was used to maximize the quality of fit instead of sales. More recently, a linear programming approach to divide the population into homogenous size groups has been proposed in [86].

In this PhD work, we propose several methodologies to divide the population into efficient sizes from a central case in each size. They are based on clustering, statistical shape analysis and the statistical concept of data depth. On the other hand, to the best of our knowledge, no statistical method has been developed for the purpose of defining representative fit models. A clustering methodology is developed with this goal in mind.

1.6 Accommodation problem in human modelling: one-size products

1.6.1 Background

As regards one-size products, the most common approach is to search for boundary cases. In the design of workplaces or also household environments, the primary goal is to fit the majority of individuals in terms of the structural size of the human body. Use of boundary representative human models (extreme cases) provides designers with an efficient way to do this. The design for extreme cases makes it possible to determine the minimum and maximum value for the target population to be accommodated. The supposition is that the accommodation of boundaries will facilitate the accommodation of interior points with less-extreme dimensions [14, 161]. 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.

A major advantage of considering boundary points is that a large range of accommodation is achieved while using a relative small number of cases. For example, in [16] it is showed that in using only 17 cases (16 boundary and 1 centroid) they were able to get the same accommodation percentage as with 400 distributed cases. For final evaluation, it is important to assess the product with mock-ups corresponding as much as possible to actual individuals or even better with "live" test subjects [175]. Physical mock-ups can only be built when enough information is available about the design. In case of difficulties, the mock-up should be modified or even a new mock-up

should be built [122]. However, building a mock-up involves considerable time, effort and money [18]. For instance, for the ergonomic design of a car many mock-ups are usually built and each mock up costs between $500,000 and $1,000,000 [21]. In addition, the number of test individuals that are needed for assessing the mock-ups is large (about 30 individuals in the cockpit design [120]). Given this problem, if we were able to identify these "hard to fit" individuals prior to assessing mock-ups, we would improve the design from the beginning and would cut down on the time and cost of the design process. Another related benefit would be the reduction in the number of real individuals needed to build the mock-ups, which represents a very significant advance in practice.

1.6.2 Literature review and our statistical proposals

The most common approaches used to define boundary cases have been percentile analysis, regression analysis and especially PCA. Percentile analysis is useful when there is only one key anthropometric dimension. The drawbacks of this simple approach have been highlighted in [228, 151, 172]. An alternative to percentile analysis is regression analysis [172, 64, 144]. It selects one or two key measurements and predicts values for other dimensions. Today, the most widely used method of obtaining extreme cases is PCA [228, 16, 83, 71, 99, 173]. The typical procedure is to build a probability ellipse that includes any desired percentage of the population ellipse in the first two or three components that explain as much variability as possible. Ref. [70] provides critical analysis of this approach. We propose to use other statistical tool in this work: archetypal analysis.

1.7 Final remark: selecting cases

All the methodologies we have developed have been applied to the data obtained from the anthropometric study of the Spanish female population. In addition, archetypal analysis has been also applied to an aircraft pilot database, which will be presented in Chapter 5.

Fig. 1.6 shows a decision tree, analogous to that one shown in [100], that helps us to decide which statistical approach is best suited to obtain valid and representative anthropometric cases. In this tree, the first decision to be made is whether determining boundary cases is useful in our particular

design problem. If it is not, we probably face a problem where determining central cases will be the most suitable approach. On the contrary, the second step is to assess whether the number of boundary points we are going to determine is enough to represent the whole population. If it is not enough, we must define distributed cases scattered throughout the distribution. But if they are enough, the third question is related to the number of relevant anthropometric dimensions. If only one dimension is important, percentiles could suffice. For two or more dimensions, our proposal based on archetypal analysis should be used.

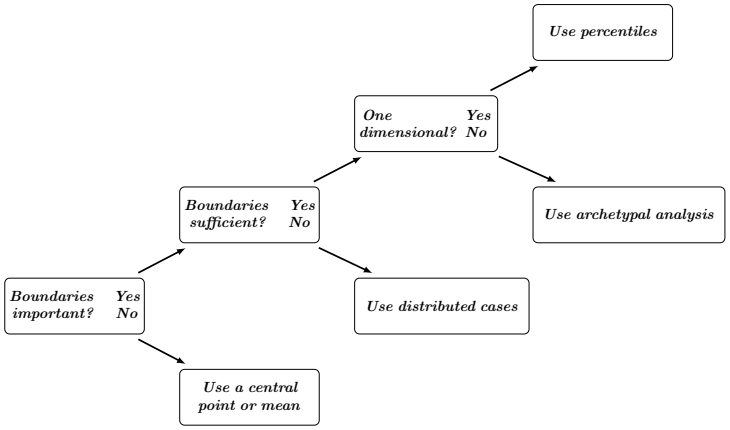


Figure 1.6: Decision tree for case selection methods based on [100].

Chapter 2

Antropometric dimensions based clustering

2.1 Introduction [1]

Clustering methods can be classified in two categories: partitioning methods and hierarchical methods.

Partitioning procedures classify the objects into k clusters, where k is usually fixed in advance (k can also be data-adaptively selected). The most well-known and commonly used partitioning techniques are k-means and k-medoids (also called Partitioning Around Medoids, PAM). Both methods are based on the assumption that a central point represents each cluster [12]. With k-means the notion of a centroid is used, which is the mean of a set of points. With k-medoids the concept of medoid is used. Note that the centroids do not have to correspond to actual data points, whereas the medoids are restricted to be one of them. The k-means method has been proposed by several scientists in different forms. The original algorithms are [134, 66, 139, 91] and, although there have been many attempts to improve the performance of the partition, see e.g. [114, 154], they are still used as standard methods. In short, k-means aims at partitioning the observations into k sets in such a way that the average squared separation of objects to their closest centroid is minimized.

Let x_1, ..., x_n be n observations of dimension p. Let k be the number of

[1] Section 2.3 is published in [102], Section 2.4 belongs to a paper in progress [214] and Section 2.5 is published in [215].

groups. The k-means method searches for a set of k points, m_1^*, ..., m_k^*, the centroids, verifying

$$\{m_1^*, ..., m_k^*\} = argmin_{m_1,...,m_k} \frac{1}{n} \sum_{i=1}^{n} \inf_{1 \leq j \leq k} \|x_i - m_j\|^2, \tag{2.1}$$

and each point x_i is assigned to its closest centroid m_j^* ($argmin$ is the argument of the minimum ($argmin_x f(x)$ is the value of x for which $f(x)$ is minimized)). The Minkowski metric is defined as

$$\|\mathbf{x} - \mathbf{y}\|_q = \left(\sum_{i=1}^{p} |x_i - y_i|^q \right)^{1/q} \tag{2.2}$$

The Euclidean distance is a particular case, with $q = 2$ [1].

Regarding PAM, its goal is to minimize the average separation of objects to their closest medoid. PAM has two phases [115]. In the first phase, called BUILD, a sequential selection of k objects is done forming the initial partition. The first object is the one for which its separation to all other objects is minimal. Afterwards, the object that decreases the objective function as much as possible is chosen. This procedure is repeated until k elements are found. In the second phase, called SWAP, one tries to improve the set of k representatives and consequently, the quality of clustering, by exchanging selected objects with unselected objects.

Hierarchical clustering methods seek to build a hierarchy of clusters. There are two types of hierarchical algorithms: agglomerative and divisive. Agglomerative methods start with as many clusters as objects and pairs of clusters are merged as one moves up the hierarchy. On the contrary, in divisive methods all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. AGglomerative NESting (AGNES) and DIvisive ANAlysis (DIANA) [115] are, respectively, examples of an agglomerative and a divisive hierarchical method.

In some contexts, especially in the analysis of gene expression data, conventional clustering methods are not suitable. Gene expression data are organized in a data frame where rows refer to genes and columns to experimental samples (conditions). Clustering applied to this kind of data is only able to identify groups of genes that show a similar pattern under all the conditions and groups of conditions that are defined for the whole of genes. However, it is well known that some sets of genes are only expressed in a

subset of conditions. In order to overcome this drawback, a novel clustering method, called biclustering, was developed. Biclustering aims at identifying subgroups of rows and subgroups of columns in a rectangular or square data matrix, by performing a simultaneous clustering of the rows and columns. Each row in a conventional row cluster is defined using all the columns that belong to that cluster. On the other hand, each column in a column cluster is defined using all the rows of that cluster. However, regarding biclustering, each row in a bicluster is defined using only a subset of columns and vice versa. In this way, clustering defines a global model but biclustering defines a local one. This interesting feature can be used to look for groups in other databases different from gene expression databases, such as anthropometric databases.

Clustering and outlier detection are very related problems. Outliers may seriously influence the results of the standard clustering procedures [75], so their possible detection and further deletion should be a primary step in any clustering application in order to make it robust. Robustness is a very desirable property for clustering methods if they want to be useful in practice. One of the strategies to remove outlier observations is the *trimming approach*. It consists in removing ("trimming") a proportion α (between 0 and 1) of the most outlying observations.

The k-means method is not a robust procedure because their clustering results can be influenced by outliers and extreme data, or bridging points between clusters. By incorporating a trimmed procedure into k-means, its robustness increases. Given k and the trimming size α, trimmed k-means searches k points, m_1^*, ..., m_k^* such that

$$\{m_1^*, ..., m_k^*\} = argmin_{\mathbf{Y}, \{\mathbf{m_1}, ..., \mathbf{m_k}\}} \frac{1}{\lceil n(1-\alpha) \rceil} \sum_{x_i \in \mathbf{Y}} \inf_{1 \leq j \leq k} \|x_i - m_j\|^2, \quad (2.3)$$

where $\mathbf{Y}$ ranges on subsets of x_1, ..., x_n containing $\lceil n(1-\alpha) \rceil$ data points ($\lceil \cdot \rceil$ denotes the integer part of a given value). Each non-trimmed point x_i is assigned to its closest centroid m_j^*. An algorithm for computing trimmed k-means is introduced in [74] and it is available in the **tclust** R package [72].

In this chapter, we introduce the methodologies called *trimowa*, *biclustAnthropom* and *hipamAnthropom*, that cluster individuals according to their anthropometric measurements. The outline of this chapter is as follows: Section 2.2 introduces the background. Sections 2.3, 2.4 and 2.5 give, respectively, the theoretical details, experimental results and summary of the three

mentioned approaches. Finally, the most important conclusions of this chapter are presented in Section 2.6.

2.2 Background

This opening section explains some methods that help the user to select the number of clusters, introduces the ordered weighted average operators and details the dissimilarity used in *trimowa* and *hipamAnthropom*.

2.2.1 Methods for selecting the number of clusters

The determination of the number of clusters, k, in the data is one of the hardest problems to solve when applying a clustering method. Partitioning methods usually require the specification of the k parameter. For hierarchical clustering, the common approach is to look at the tree of clusters and choose the level of the tree at which the clusters are still meaningful.

One method for selecting the number of groups is called the average silhouette width (asw). It can deal with any kind of data (continuous, binary or qualitative). It is used not only to estimate the correct number of clusters, but also to evaluate the quality of a classification [115]. Let us introduce its mathematical definition:

For a given clustering, $C = \{C_1, \ldots, C_k\}$, the silhouette width for the ith observation $x_i \in C(x_i), i = 1, \ldots, p$, is defined as follows:

$$
\begin{aligned}
s(x_i) &= 1 - \frac{a_i}{b_i} \quad if \;\; a_i < b_i \\
&= 0 \qquad\quad if \;\; a_i = b_i \\
&= \frac{b_i}{a_i} - 1 \quad if \;\; a_i > b_i
\end{aligned}
$$

In one single formula:

$s(x_i) = \dfrac{b_i - a_i}{max\{a_i, b_i\}}$, where

$$\begin{cases} a_i = \dfrac{1}{|C(x_i)|} \displaystyle\sum_{x_j \in C(x_i)} d(x_j, x_i) \\ \\ b_i = \displaystyle\min_{C \neq C(x_i)} \dfrac{1}{|C|} \sum_{x_j \in C} d(x_j, x_i) \end{cases}$$

From the definition, we can see that $-1 \leq s(x_i) \leq 1$. The higher the value of $s(x_i)$ is, the better the clustering of x_i in $C(x_i)$ is. The average silhouette width is obtained as $asw(C) = \dfrac{1}{p} \displaystyle\sum_{i=1}^{p} s(x_i)$. The partition with the maximum average silhouette width can be considered as the optimal partition. This maximum is called the silhouette coefficient (SC).

The authors of [115] claim that SC is a useful measure to evaluate the amount of clustering structure that has been found by the clustering method used. From their experience, they established a subjective interpretation of SC, which is summarized in Table 2.1.

SC	Proposed interpretation
$0.71 - 1.00$	A strong structure has been discovered.
$0.51 - 0.70$	A reasonable structure has been found.
$0.26 - 0.50$	The structure is weak and could be artificial. Additional clustering methods should be used to compare.
≤ 0.25	No substantial structure has been found.

Table 2.1: Subjective interpretation of the silhouette coefficient (SC).

In addition, the silhouette plot displays the silhouettes of all the elements that belong to the same cluster, ranked in decreasing order. As an example, Fig. 2.1 shows the silhouette of four clusters discovered in the Ruspini data used in [115]. The average silhouette width of this particular partition is equal to 0.74, so, according to Table 2.1, a strong clustering structure has been discovered. This plot has been generated using the *silhouette* function of the **cluster** R package [141].

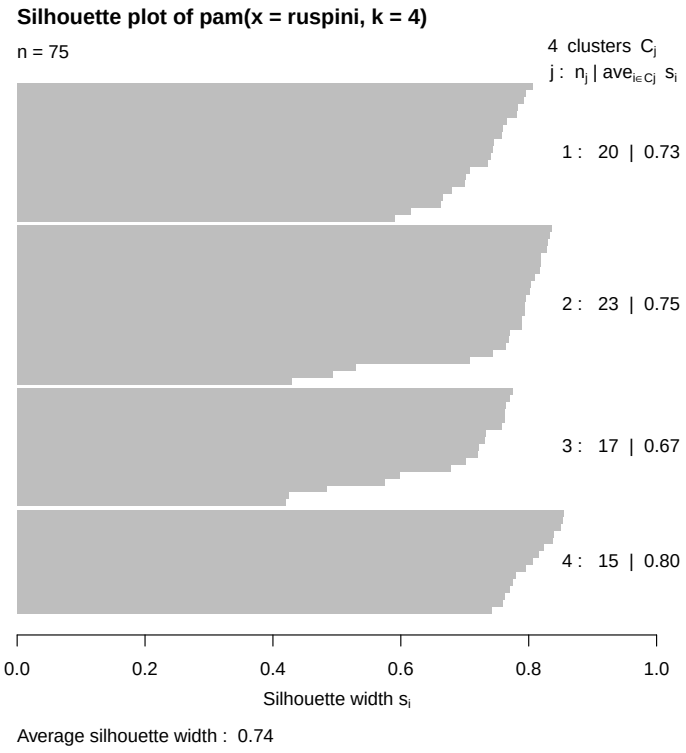


Figure 2.1: Silhouette plot for four clusters using the Ruspini data [115].

A new measure of cluster structure has been recently developed: the *index number clusters atypical* (INCA) criterion [104]. It has been defined not only to assess the number of clusters, but also to determine whether a new observation belongs to one of those identified clusters or it may be considered as an outlier. Next we give its definition.

First, we start with n observartions assigned into k clusters $C_1, \ldots, C_k$ of sizes $n_1, \ldots, n_k$. The data of the cluster C_i can be considered as a random sample of a continuous random vector Y_i. Any new point y_0 should be classified into one of the previously defined clusters C_j, $j = 1, \ldots, k$. However, it may happen that y_0 would actually be an outlier regarding those fixed clusters. In that case, y_0 would belong to a new cluster, whose center would be a convex combination of the previous centers: $\sum_{i=1}^{k} \alpha_i E(Y_i)$, where the weights α_i are assessed by minimizing the following $L(y_0)$ objective function and $E(Y_i)$ refers to the expected value of Y_i.

Let $\delta(x, y)$ be a particular distance function. Given the point y_0, the following function, called the INCA statistic, is evaluated:

$$W(y_0) = \min_{(\alpha_1,\ldots,\alpha_k):\sum_{i=1}^{k}\alpha_i=1} L(y_0), \tag{2.4}$$

where

$$L(y_0) = \sum_{i=1}^{k} \alpha_i \phi_i^2(y_0) - \sum_{1\le i<j\le k} \alpha_i \alpha_j \Delta_{ij}^2 \tag{2.5}$$

being

$$\phi_i^2(y_0) = \frac{1}{n_i} \sum_{l\in C_i} \delta^2(y_0, y_l) - V_\delta(C_i), \tag{2.6}$$

$$V_\delta(C_i) = \frac{1}{2n_i^2} \sum_{l,m\in C_i} \delta^2(y_l, y_m), \tag{2.7}$$

$$\Delta_{ij}^2 = \frac{1}{n_i n_j} \sum_{l\in C_i, m\in C_j} \delta^2(y_l, y_m) - V_\delta(C_i) - V_\delta(C_j). \tag{2.8}$$

All the theoretical details and more references are given in [104] and [5]. The procedure for determining the number of clusters is then as follows. Given a fixed cluster C_j, the value $W(y)$ is calculated for every observation y of the data set considering all clusters excepting C_j. The value $W(y)$ is denoted as $W_{C_j}(y)$. If $W_{C_j} = \max_{z\notin C_j} W_{C_j}(z)$ (the maximum of the squared orthogonal distances for all the observations not belonging to C_j), the following rule is stated:

- Observation $y \in C_j$ is well classified in C_j if $W_{C_j}(y) > W_{C_j}$.
- Observation $y \in C_j$ is not well classified in C_j if $W_{C_j}(y) \le W_{C_j}$.

The basic idea is that if y is at a greater distance from any cluster (different from C_j) even than the more distant z from any cluster, then y must be located in C_j.

Then, the INCA index, $INCA_k$, to estimate the number of clusters can be introduced. It is defined as the probability of properly classified individuals and it is estimated with the following expression:

$$INCA_k = \frac{1}{k} \sum_{j=1}^{k} \frac{N_j}{n_j} \tag{2.9}$$

where N_j is the total number of units in C_j which are well classified.

The closer $INCA_k$ to 1 is, the more units correctly classified are. But the closer $INCA_k$ to 0 is, the more units not correctly classified are. Therefore, the value of k helps to select the approppriate number of clusters. The rule proposed in [104] is that k should be chosen as the value of k preceding the first biggest slope decrease.

It is worth pointing out that if the $INCA_k$ is small for all k, then there is only one cluster made up of all the points.

The INCA criterion is implemented in the **ICGE** R package [106].

2.2.2 Ordered weighted average operators

In multicriteria decision making problems, a decision has to be made based on several alternatives. The different criteria are attributes or features, which are expressed numerically.

As an example taken from [13], different attributes such as the price, quality, fuel consumption, brand, etc, should be evaluated before buying a car. In order to choose which car to buy, the customer should assign to each attribute of each car a score indicating how important to her/him is that attribute. Then, she/he should combine in some way all the scores related to each particular car. After combining them, a general score representing each car would arise. Then, the person should compare all the general scores to finally make a decision.

In this way, the efficient combination (or better said, aggregation) of information constitutes a very important task in these types of problems. Aggregation is the procedure that transforms a set of elements into a summary measure describing the whole set. There are different types of functions (called aggregation function or operators [13]) for this purpose. The arithmetic mean, minimum or maximum are three simple examples. Another type is the ordered weighted averaging (OWA) function, introduced for the first time by Yager [227].

Let $d_1, \ldots, d_p$ the values to be aggregated. An OWA operator of dimension p is a mapping $f : \mathbb{R}^p \to \mathbb{R}$ where $f(d_1, \ldots, d_p) = w_1 b_1 + \ldots + w_p b_p$, being b_j the jth largest element in the collection $d_1, \ldots, d_p$ (i.e., these values are ordered in decreasing order) and $W = (w_1, \ldots, w_p)$ an associated weighting vector such that:

- $w_i \in [0, 1], 1 \leq i \leq p$

- $\sum_{j=1}^{p} w_j = 1$

From this definition, the maximum, minimum and arithmetic mean can be easily obtained from a particular set of weights:

- If $W = (1, 0, \ldots, 0)$, then $f(d_1, \ldots, d_p) = \max_i d_i$
- If $W = (0, 0, \ldots, 1)$, then $f(d_1, \ldots, d_p) = \min_i d_i$
- If $W = (\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p})$, then $f(d_1, \ldots, d_p) = \frac{1}{p} \sum_{j=i}^{p} d_i$

A fact to be emphasized of OWA is regarding the re-ordering step: d_i is not associated with a weight w_i, but the w_i is associated with the ith largest element b_i. In other words, the weights are not associated with a particular input, but rather with its value.

Because the OWA operators are bounded between the max and min operators, a measure called orness was defined in [227] to classify the OWA operators between these two. The orness quantity adjusts the importance to be attached to the values $d_1, \ldots, d_p$, depending on their ranks:

$$\text{orness}(W) = \frac{1}{p-1} \sum_{i=1}^{p} (p - i) w_i. \tag{2.10}$$

In particular:

- If $W = (1, 0, \ldots, 0)$, then $\text{orness}(W) = 1$
- If $W = (0, 0, \ldots, 1)$, then $\text{orness}(W) = 0$
- If $W = (\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p})$, then $\text{orness}(W) = 0.5$

Thus, when orness = 1, the highest importance is given to the largest aggregated value. On the contrary, when orness = 0, the highest importance is given to the smallest aggregated value. In addition, when orness = 0.5, all aggregated values are equally important.

A major issue in the OWA operators theory is the determination of the associated weights. In [227] two ways of doing this are discussed, but since

then, other methods have been developed to the same end, as it is discussed in [127]. In our case, the procedure that we will use to generate the set of weights $W = (w_1, \ldots, w_p)$ is the following. According to the corollary 1 explained in [127, p.2626], the orness α of a set of weights $W = (w_1, \ldots, w_p)$ is calculated as $orness(W) = \alpha = \lambda(1 - prob) + (1 - \lambda)0.5$ where $\lambda \in [0, 1]$ is a mixture parameter and $prob$ is the success probability in a binomial distribution $Bi(p - 1, prob)$.

A direct consequence of this equation is that, for a given α, we can generate a vector of weights using a mixture of binomial and discrete uniform distributions, $\lambda Bi(p - 1, prob) + (1 - \lambda)U(1, p)$.

The relationship between α, λ and $prob$ can be formulated as $2\alpha - 1 = \lambda(1 - 2prob)$. Therefore, fixed α, we have two options to determine the other two parameters:

1. Fixed α and $prob$, λ is calculated.

2. Fixed α and λ, $prob$ is calculated.

We choose the second option. We give orness a value of $\alpha = 0.7$ in order to highlight the largest aggregated values. In addition, we fix $\lambda = 0.5$ to give equal importance to both binomial and discrete uniform. Thus, $prob = 0.1$ and our binomial can be expressed as $Bi(p - 1, prob = 1.5 - 2 \cdot orness)$. It is worth pointing out that this binomial only exists for $orness \in [0.3, 0.7]$. For our application, these values are the most relevant.

Specifically, each weight is calculated as $w_i = \lambda \cdot \pi_i + (1 - \lambda) \cdot \frac{1}{p}$, where π_i is the binomial probability for each $i = 0, \ldots, p - 1$. This way of proceeding allows us to give every aggregated value at least a small influence in the overall computation of the OWA operator. All the theoretical properties are given with great detail in [127].

We decide to use this particular procedure because the weights are easily obtained and are also easy to interpret. Furthermore, our practical experiments have shown that it works well for this case.

We are going to use an OWA operator to combine dissimilarities.

2.2.3 Dissimilarity measure

Clustering methods require the selection of a function that determines how far every pair of elements are. In our everyday life we use the concept of distance to quantify the degree of closeness between two objects [41]. Here we focus on its mathematical definition.

Let X be a subset of $\mathbb{R}^p$. A distance (also called metric) is defined as a function $d : X \times X \to \mathbb{R}$ that satisfies [183]:

1. **Non-negativity:** $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$ for all $x, y \in X$
2. **Symmetry:** $d(x, y) = d(y, x)$ for all $x, y \in X$
3. **Triangle inequality:** $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$

When any of these three axioms is not met, d is called a dissimilarity.

In the definition of a sizing system, a dissimilarity function allows to mathematically represent the idea of garment fit. Let $x = (x_1, \ldots, x_p)$ be an individual of the user population represented by a feature vector of size p of his/her body measurements. In the same way, let $y = (y_1, \ldots, y_p)$ be the p measurements of the prototype of a particular size. Then, $d(x, y)$ measures the misfit between a particular individual and the prototype. In other words, $d(x, y)$ indicates how far a garment made for prototype y would be from the measurements for a given person x. In this way, the function used to quantify the misfit between an individual and the prototype is a key element to define an efficient sizing system. Next, we introduce the dissimilarity that we use in *trimowa* (Section 2.3) and *hipamAnthropom* (Section 2.5). It is defined considering the same ideas stated in [147]:

i. The larger the individual differences between a person and the prototype, the worse the fit.

ii. Fit is better predicted by proportional rather than absolute differences between individual and prototype features.

iii. There is a range of values where there is no difference between the measurements x_i and y_i $(i = 1, \ldots, p)$, probably because the fit is perfect although those values do not exactly coincide.

iv. A too small garment may not affect fit in the same way as one which is too large.

v. Discrepancies in certain features are more important than others.

Several functional forms may satisfy the above criteria. We will assume the same as in [147], which has the following expression:

$$d_i(x_i, y_i) = \begin{cases} a_i^l(ln(y_i) - b_i^l - ln(x_i)) & \text{if } ln(x_i) < ln(y_i) - b_i^l \\ 0 & \text{if } ln(y_i) - b_i^l \leq ln(x_i) \leq ln(y_i) + b_i^h \\ a_i^h(ln(x_i) - b_i^h - ln(y_i)) & \text{if } ln(x_i) > ln(y_i) + b_i^h \end{cases} \tag{2.11}$$

where a_i^l, b_i^l, a_i^h and b_i^h are constants for each dimension and have the following meaning: b_i corresponds to the range in which there is a perfect fit (condition iii.); a_i indicates the rate at which fit deteriorates outside this range, i.e., it reflects the misfit rate. In addition, measurements are log transformed to meet the condition ii. Fig. 2.2 illustrates this function.

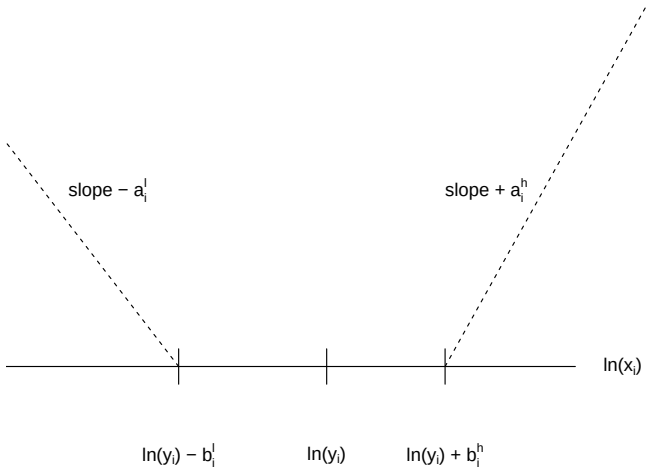


Figure 2.2: This plot, based on [147], illustrates the marginal dissimilarity between the prototypes and each individual for the i*th* dimension.

From condition iv., we can state that this functional form (and any satisfying these conditions) is not symmetric. In particular, for a given value of $|x_i - y_i|$, the discrepancy may be smaller if $x_i < y_i$ than if $x_i > y_i$. The global dissimilarity between individuals x and y was finally defined in [147] as a sum of squared discrepancies over each of the p body measurements:

$$d(x, y) = \sum_{i=1}^{p} \big(d_i(x_i, y_i)\big)^2 \tag{2.12}$$

2.3 Trimowa

This methodology was introduced aimed at developing an efficient apparel sizing system [102]. It is close to that one proposed in [147]. However, there are two main differences. First, when searching for the k prototypes, we use a more statistical approach instead of the continuous optimization problem developed in [147]. Specifically, we use a trimmed version of the PAM (or k-medoids) algorithm. We aim at looking for medoids i.e., for typical people within the sample, which means that our final prototypes will correspond to real subjects of the data set. We use a trimmed procedure because an apparel sizing system is intended to cover only what we could call standard population, leaving out those individuals who might be considered outliers with respect to a set of measurements. Second, we propose to modify the dissimilarity measure defined in [147], by taking into account to the people morphology using an OWA operator.

2.3.1 Methodology

2.3.1.1 Global dissimilarity measure

As mentioned in Section 2.2.3, the global dissimilarity described in [147] is merely defined as a sum of squared discrepancies over each of the p anthropometric measurements considered:

$$d(x, y) = \sum_{i=1}^{p} \big(d_i(x_i, y_i)\big)^2 \tag{2.13}$$

In this way, the different dissimilarities $d_i(x_i, y_i)$'s are being *aggregated* and a lot of possibilities can be opened by looking at the problem under this

point of view. At first glance, it could be more natural consider:

$$d(x, y) = \max_i d_i(x_i, y_i) \tag{2.14}$$

or:

$$d(x, y) = \min_i d_i(x_i, y_i) \tag{2.15}$$

However, if we chose the dissimilarity defined as in eq. (2.14), we would consider the maximum discrepancy between the individual and the prototype among all the measurements, which means that we would consider the worse fit. Otherwise, if we decided to work with the dissimilarity defined as in eq. (2.15), we would consider the best fit, but again only for one feature. These two situations are not neither reliable nor practical. We would prefer an intermediate scenario. To that end, we use an OWA operator. We adjust the importance of each one of the p discrepancies by assigning to each one of them a particular weight. The largest discrepancy gets the largest weight, the second largest discrepancy gets the second largest weight and so on for the p measurements. On consequence, the dissimilarity we will use in this methodology and also with *hipamAnthropom* is defined as follows:

$$d(x, y) = \sum_{i=1}^{p} w_i \big(d_i(x_i, y_i)\big)^2 \tag{2.16}$$

where $d_i(x_i, y_i)$ is the ith largest element of the collection of aggregated dissimilarities $d_1(x_1, y_1), \ldots, d_p(x_p, y_p)$ and w_i the weight associated with $d_i(x_i, y_i)$. As indicated in [31], the weights of the OWA operator can be used to achieve a better balance between the fashion style of garments and the general comfort of wearers.

In addition, the δ distance function related to the definition of the INCA criterion (see Section 2.2.1) is the distance of eq. (2.16):

$$\delta(x, y) = d(x, y) = \sum_{i=1}^{p} w_i \big(d_i(x_i, y_i)\big)^2 \tag{2.17}$$

The core of this methodology is presented in Section 2.3.1.2, while its results and a brief summary are given in Sections 2.3.2 and 2.3.3, respectively.

2.3.1.2 Clustering procedure

Trimmed k-medoids (or trimmed PAM) is analogous to k-medoids but a trimming procedure is added. It includes the advantages of the k-medoids and the trimmed k-means defined in eq. (2.3).

Let $x_1, \ldots, x_n$ be n observations of dimension p and let k be the number of groups. Let $d(x_i, x_j)$ be the dissimilarity defined in eq. (2.16) between individuals i and j. For a given number of clusters k and a trimming proportion α, trimmed k-medoids searches for k individuals, $x^*_{i_1} \ldots x^*_{i_k}$, such that

$$\{x^*_{i_1}, ..., x^*_{i_k}\} = argmin_{\mathbf{Y}, \mathbf{x_{i_1}}, ..., \mathbf{x_{i_k}}} \frac{1}{\lceil n(1-\alpha) \rceil} \sum_{x_i \in \mathbf{Y}} \inf_{1 \leq j \leq k} d(x_i, x_{i_j}), \qquad (2.18)$$

where $\mathbf{Y}$ ranges on subsets of x_1, ..., x_n containing $\lceil n(1-\alpha) \rceil$ data points, $\lceil \cdot \rceil$ refers to the integer part of a given value and $argmin$ is the argument of the minimum ($argmin_x f(x)$ is the value of x for which $f(x)$ is minimized). Each non-trimmed point x_i is assigned to its closest medoid $x^*_{i_j}$.

Then, the trimmed k-medoids algorithm involves the following steps:

1. Initialization: Random selection of k starting points that will serve as seed medoids.

2. Given the assumption that x_{i_1}, ..., x_{i_k} are the k medoids obtained in the previous iteration, then:

   (a) Associate each observation to the closest medoid:

   $$d_i = \min_{j=1,...k} d(x_i, x_{i_j}), \qquad i = 1, ..., n,$$

   and keep the set H having the $\lceil n(1-\alpha) \rceil$ points with lowest d_i's.

   (b) Split H into $H = \{H_1, ..., H_k\}$ where the observations in H_j are those closer to x_{i_j} than to any of the other medoids.

   (c) The medoid x_{i_j} for the next iteration will be the medoid of observations belonging to group H_j.

3. Repeat the step 2 a few times. After these iterations, compute the final evaluation function.

This algorithm is repeated a few times and the best solution is preserved. Regarding programming, the algorithm of García Escudero et al. [74] was adapted for computing trimmed k-medoids. The medoid of each cluster is computed with the *pam* function (with $k = 1$ for each group) of the **cluster** R package. The detailed algorithm is given in Algorithm 1.

2.3.2 Results

Before presenting our experimental results, we first introduce the database, the procedure used to apply the trimmed k-medoids and the parameters for both the dissimilarity and algorithm.

From the whole database introduced in Section 1.3, a selection of 6013 women is done, leaving out pregnant women, breastfeeding women at the time they participated in the survey, women who have undergone any type of cosmetic surgery, and the ones younger than 20 or older than 65. In addition, from all the anthropometric dimensions of the original database, only the five most relevant in the clothing development process were considered: bust circumference, chest circumference, neck to ground length, waist circumference and hip circumference. They are our control or key dimensions. They were chosen for these three reasons:

- Recommendations of experts.
- They are commonly used in the literature.
- They appear in the *European standard to sizing system. Size designation of clothes. Part 2: Primary and secondary dimensions* [58].

According to the European standard. Part 2 [58], bust circumference will be considered the primary control dimension to define each major size group and the other four measurements as secondary control dimensions to divide each major size group into subgroups. Summarizing, the database is made up of 6013 Spanish women and their body measurements for 5 dimensions. Summary statistics of these five characteristics are given in Table 2.2.

Algorithm 1 An algorithm for trimmed k-medoids

Set k, number of groups; ns, (for instance, $ns = 10$) and nr (for instance, $nr = 100$).
Select k starting points that will serve as seed medoids (e.g., draw at random k subjects from the whole data set).
for $r = 1 \to nr$ **do**
  for $s = 1 \to ns$ **do**
    Assume that x_{i_1}, ..., x_{i_k} are the k medoids obtained in the previous iteration.
    Assign each observation to its nearest medoid:

$$d_i = \min_{j=1,\ldots k} d(x_i, x_{i_j}), \qquad i = 1, \ldots, n,$$

    and keep the set H having the $\lceil n(1-\alpha) \rceil$ observations with lowest d_i's.
    Split H into $H = \{H_1, ..., H_k\}$ where the points in H_j are those closer to x_{i_j} than to any of the other medoids.
    The medoid x_{i_j} for the next iteration will be the medoid of observations belonging to group H_j.
    Compute

$$F_0 = \frac{1}{\lceil n(1-\alpha) \rceil} \sum_{j=1}^{k} \sum_{x_i \in H_j} d(x_i, x_{i_j}). \tag{2.19}$$

    if $s == 1$ **then**
      $F_1 = F_0$.
      Set M the set of medoids associated with F_0.
    else
      if $F_1 > F_0$ **then**
        $F_1 = F_0$.
        Set M the set of medoids associated with F_0.
      end if
    end if
  end for
  if $r == 1$ **then**
    $F_2 = F_1$.
    Set M the set of medoids associated with F_1.
  else
    if $F_2 > F_1$ **then**
      $F_2 = F_1$.
      Set M the set of medoids associated with F_1.
    end if
  end if
end for
return M and F_2.

Measurement (cm)	Minimum	First Quantile	Median	Mean	Third Quantile	Maximum
Neck to ground length	116.4	132.9	136.8	137	140.8	161.9
Bust circumference	73	87.4	93.3	95.02	100.7	145.7
Chest circumference	45.91	90.78	96.37	97.92	103.7	150.30
Waist circumference	58.60	75.6	83.10	84.98	92.40	167.6
Hip circumference	72.8	98.3	103.3	104.9	109.9	170.8

Table 2.2: Summary statistics for the five variables considered.

Regarding the constants that define the dissimilarity in eq. (2.11), their values were chosen keeping in mind the following facts:

1. As in [147], an individual's anthropometric dimension being larger than the prototype's one is penalized three times more than that being smaller ($b_i^l = 3b_i^h$ and $a_i^h = 3a_i^l$).

2. The dissimilarity consistent with a perfect fit (b_i^h) was chosen within each bust segment in which the population is first divided, to cover all the range of values of each dimension in such a way that all the subjects would be perfectly fitted in exactly one group. In other words, for each segment j, $b_i^h = \frac{3 \cdot Range(\{x_{j_1 i}, \ldots x_{j_n i}\})}{4k}$, where k is the number of groups into the population is classified.

3. The values of a_i^h were chosen in the same way as in [147]. They are shown in Table 2.3.

In addition, the aggregation weights to be assigned to the five dissimilarities (one per dimension) computed, are given in Table 2.4. The value of orness is 0.7. This orness is close to 1 in order to highlight high dissimilarities in any of the considered measurements.

	a_i^l	a_i^h
Chest circumference	7.5	22.5
Bust circumference	8.3	25
Neck to ground length	9.5	28.5
Waist circumference	6.7	20
Hip circumference	8.3	25

Table 2.3: Constants that define the dissimilarity function in eq. (2.11) .

w_1	w_2	w_3	w_4	w_5
0.42805	0.24580	0.12430	0.10180	0.10005

Table 2.4: Aggregation weights.

The procedure that we propose to obtain size groups is the following: First, the data set is divided into twelve major bust size groups (also called segments or classes) according to the bust sizes defined in the *European standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals* [59]. Table 2.5 shows the size range, the size scale and the size interval along the bust, waist and hip control dimensions.

Bust	74-78	78-82	82-86	86-90	90-94	94-98	98-102	102-107	107-113	113-119	119-125	125-131
Waist	58-62	62-66	66-70	70-74	74-78	78-82	82-86	86-91	91-97	97-103	103-109	109-115
Hip	82-86	86-90	90-94	94-98	98-102	102-106	106-110	110-115	115-120	120-125	125-130	130-135

Table 2.5: Size range, size scale and size interval for bust, waist and hip, used to define the size groups according to the *European standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals* [59].

Then, the trimmed k-medoids is applied to each bust segment with $k = 3$ clusters, so a total of 36 sizes is obtained. The number of random initializations was 600, with seven steps per initialization. The trimmed proportion was prefixed to $\alpha = 0.01$ per segment (therefore, the accommodation rate in each bust size will be 99%).

2.3.2.1 Experimental results

Now we are in disposition to illustrate our results. Fig. 2.3 shows the scatter plots of bust circumference against neck to ground, waist, hip and chest (from left to right and from top to bottom) jointly with the three medoids obtained for each bust segment.

A careful examination of the distribution of medoids leads to different conclusions. In particular, we analyze the medoids corresponding to bust intervals $[82, 86[$ cm (green crosses in Fig. 2.3) and $[94, 98[$ cm (brown inverted triangles). They are described in Tables 2.6 and 2.7.

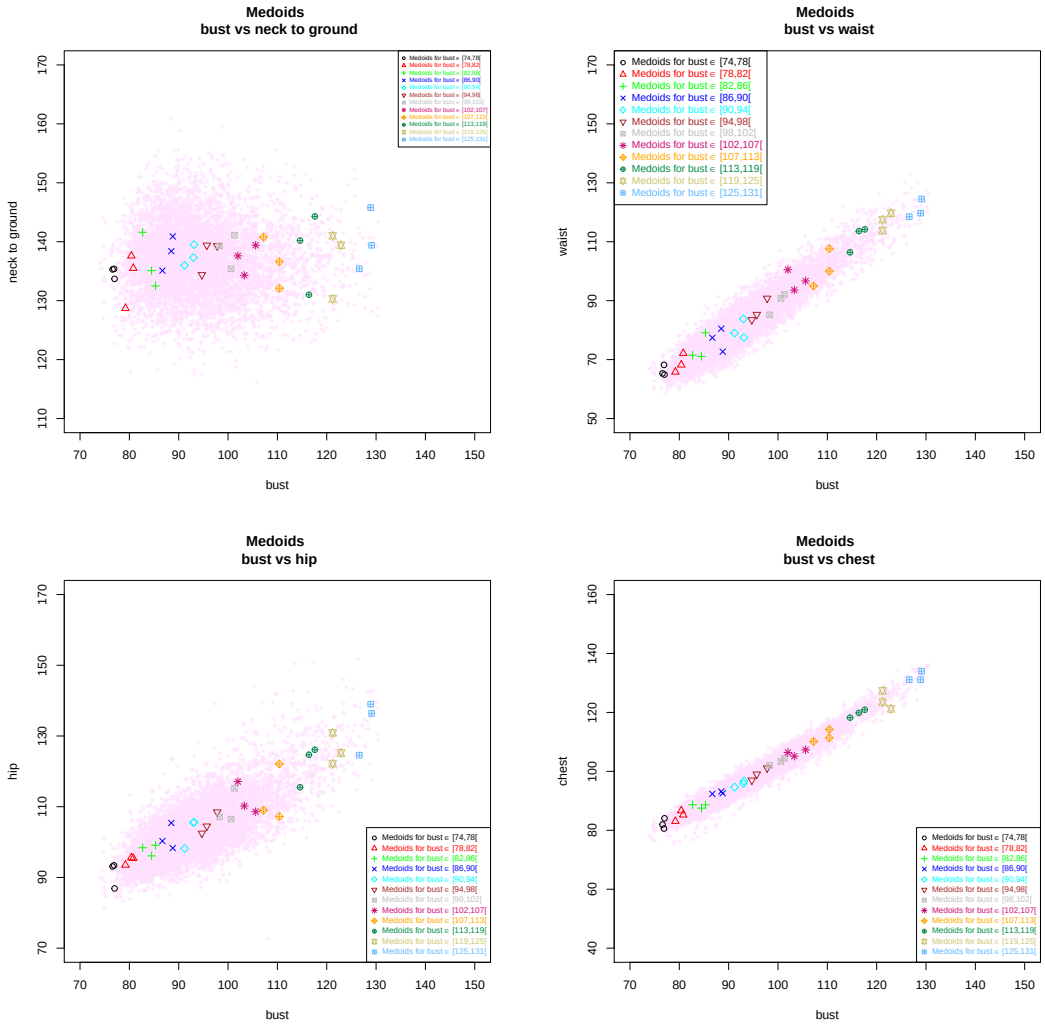


Figure 2.3: Bust vs neck to ground (top left), waist (top right), hip (bottom left) and chest (bottom right) for each one of the medoids. $[82, 86[$ medoids are represented with a green cross, while $[94, 98[$ medoids are represented with a brown facing down triangle.

Woman Code	Chest	Neck to ground	Waist	Hip	Bust	Hip - Waist	Bust - Waist
CANDE021	88.6283	132.5	79.1	99.1	85.3	20	6.2
SEVI132	88.6745	141.6	71.5	98.4	82.7	26.9	11.2
LLEID074	87.5182	135.1	71.1	96.1	84.5	20	13.4

Table 2.6: Medoids measurements for bust size $[82, 86[$ cm.

Woman Code	Chest	Neck to ground	Waist	Hip	Bust	Hip - Waist	Bust - Waist
SILLE034	96.9951	134.4	83.5	102.5	94.7	19	11.2
JAEN075	101.129	139.3	90.8	108.5	97.8	17.7	7
CANDE068	99.0432	139.4	85.3	104.5	95.7	19.2	10.4

Table 2.7: Medoids measurements for bust size $[94, 98[$ cm.

Medoids JAEN075 and CANDE068 of $[94, 98[$ cm have similar neck to ground values, so two sizes for length could be enough for this bust class. On the contrary, medoids of range $[82, 86[$ cm are more distributed for this dimension, which suggests that three sizes with different lengths would be more appropriate for this bust range. However, these same medoids just show an opposite pattern regarding their waist measurements. For bust range $[82, 86[$ cm, medoids SEVI132 and LLEID074 have a similar waist, while for range $[94, 98[$ cm medoids present a greater variation for this dimension. Therefore, medoids of $[94, 98[$ cm are more different in waist, while in range $[82, 86[$ cm the variability of length predominates. Regarding hip and chest dimensions, the most relevant feature is that for $[82, 86[$ cm, CANDE021 and SEVI132 are quite overlapped.

Figs. 2.4 and 2.5 show the frontal and lateral perspective of the body shape of the medoids obtained for the bust size $[82, 86[$ cm, while Figs. 2.6 and 2.7 show the body shape of the medoids for $[94, 98[$ cm. The same patterns just explained above can be more clearly appreciated with these images.

In addition, we would like to know the improvement we would achieve in garment fit if we considered the 36 sizes defined by the medoids obtained in our work instead of those defined by the European standard. Part 3 [59]. First, we must define the measurements of the prototypes for the five dimensions we are considering in this study. As detailed in Table 2.5, this standard establishes 12 sizes according to the combinations of bust (from 74 to 131 cm), waist (from 58 to 115 cm) and hip (from 82 to 135 cm) measurements. However, it does not provide any indication about chest and height standard measurements.

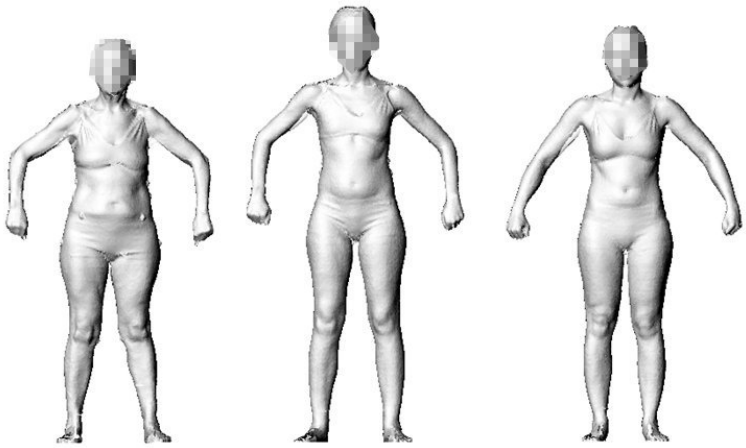


Figure 2.4: Front body shape of medoids for size [82,86[cm (left to right, CANDE021, SEVI132 and LLEID074).

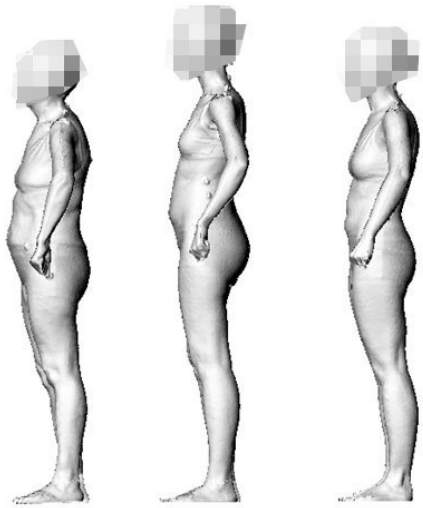


Figure 2.5: Lateral body shape of medoids for size [82,86[cm (left to right, CANDE021, SEVI132 and LLEID074).

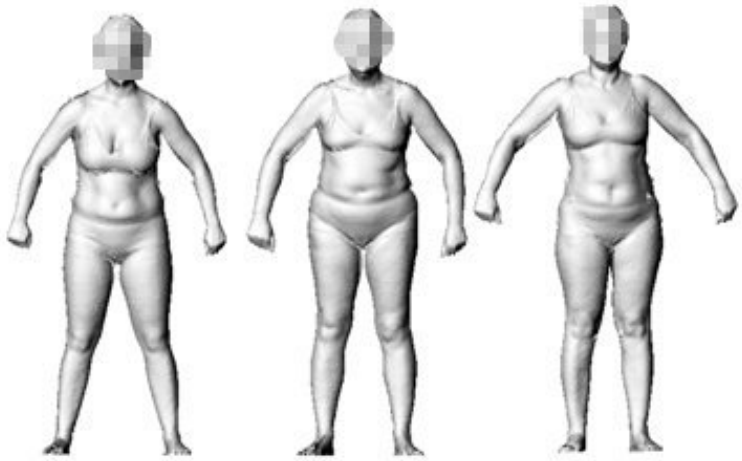


Figure 2.6: Front body shape of medoids for size [94,98[cm (left to right, SILLE034, JAEN075 and CANDE068).

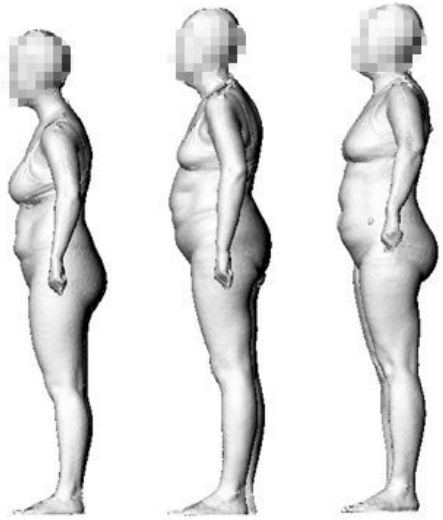


Figure 2.7: Lateral body shape of medoids for size [94,98[cm (left to right, SILLE034, JAEN075 and CANDE068).

Regarding chest, we can easily approximate their measurements through a linear regression analysis because bust and chest measurements are highly correlated in the women of our data set. This can be seen in the bottom right plot of Fig. 2.3. We consider the bust dimension as the independent variable. The obtained values are detailed in Table 2.8.

Bust	76	80	84	88	92	96	100	104	110	116	122	128
Waist	60	64	68	72	76	80	84	88	94	100	106	112
Hip	84	88	92	96	100	104	108	112	117	122	127	132
Chest	79.50	83.38	87.26	91.14	95.02	98.90	102.78	106.66	112.46	118.30	124.12	129.94

Table 2.8: Measurements to define the prototypes (central individuals) for each size of the European standard to sizing system. Part 3 [59] (see Table 2.5), including the calculated values for chest.

Then, Table 2.8 details the measurements the prototypes would have for bust, waist, hip and chest according to the European standard. Part 3 [59]. Regarding height, since neck to ground shows no correlation with the other four variables, we take a different aproach. We consider as neck to ground values for the standard sizing system the values 132, 136 and 140 cm because those are the most repeated measurements, and according to the bust vs neck to ground plot of Fig. 2.8 (top left plot), they are the measurements which best cover our data set. Once the 12 groups and 3 different neck to ground measurements per group for the standard prototypes are specified, we are now in the position to compare the adequacy of the sizing system defined from our medoids with those standard ones. To that end, Fig. 2.8 shows the same scatter plots as Fig. 2.3, but incorporating the corresponding measurements of the prototypes obtained following the European standard. Part 3 [59]. In the four plots, the prototypes would be located just in the center of the corresponding boxes.

Another valuable analysis of our methodology's performance could be to compare how far are the individuals with respect to the medoids obtained with our approach and the standard prototypes defined by the European standard. Part 3 [59]. Fig. 2.9 allows this comparison to be made. Fig. 2.9 displays the cumulative distribution functions for the dissimilarities between all the women and our medoids and for the dissimilarities between all the women and the standard prototypes. Cumulative distribution functions show the probability that a distance will be less or equal than a certain value.

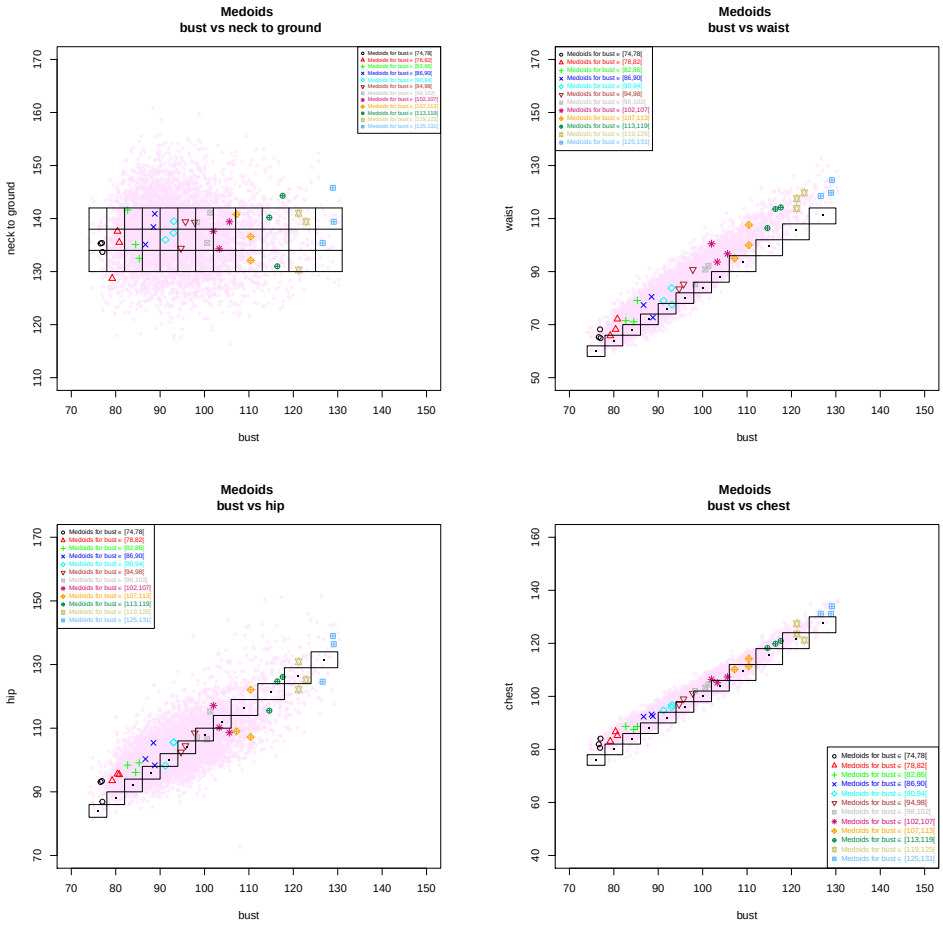


Figure 2.8: Bust vs neck to ground (top left), waist (top right), hip (bottom left) and chest (bottom right), jointly with our medoids and the prototypes defined by the European standard. Part 3 [59].

In both cases, dissimilarities have been computed by using the dissimilarity of Section 2.2.3. The first thing we see in Fig. 2.9 is that there is a percentage of population rounding 60%, which gets a perfect fit taking into account our dissimilarity criteria in both sizing systems. Nevertheless, this percentage increases until 80% with the sizing system we propose. Because the cumulative distribution function for our method increases faster than the cumulative distribution function for the standard system, we can state that women are closer to their corresponding medoids computed by our methodo-

logy with respect to the standard prototypes. This can be also seen by computing the expected range of the dissimilarities, that is to say, the values between the 10th and 90th percentiles. The range for the dissimilarities between women and our medoids is [0,0.15], while the range for the dissimilarities between women and standard prototypes is [0,0.36].

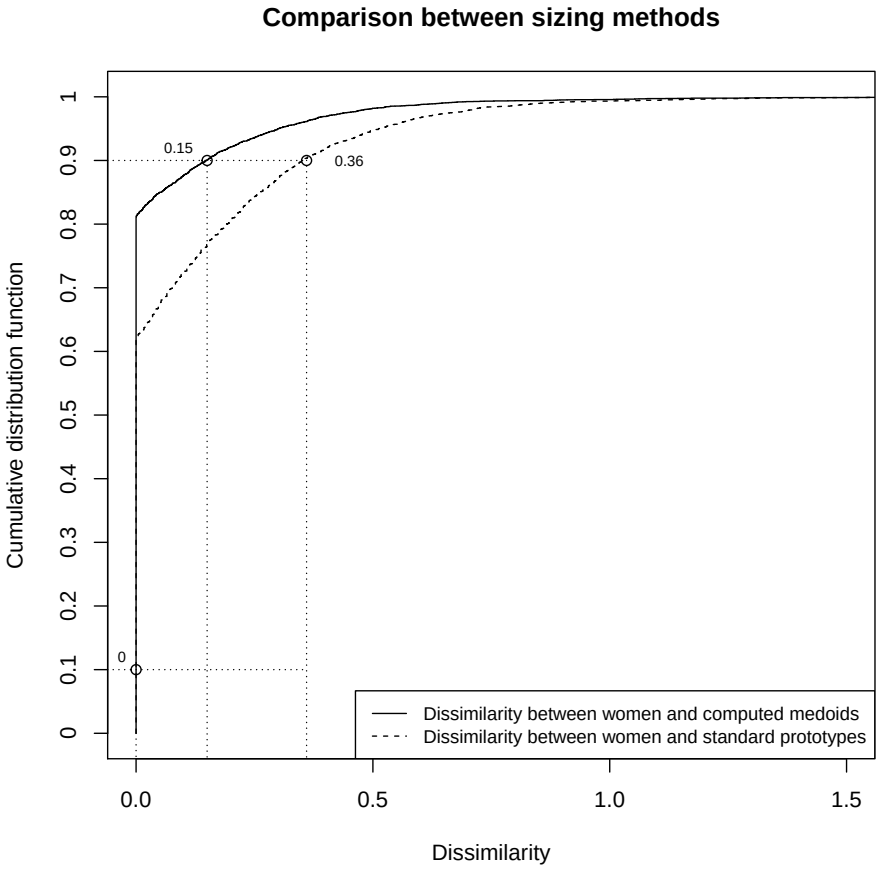


Figure 2.9: Cumulative distribution function for the dissimilarities between women and computed medoids and for the dissimilarities between women and standard prototypes.

All the considerations from the analysis of Fig. 2.8 and Fig. 2.9 suggest that our method actually defines more efficient sizes where the women are better accommodated.

2.3.3 Summary

A new statistical approach using clustering has been developed aimed at defining an optimal and efficient clothing sizing system. It was conceived as an extension and improvement of the optimization problem proposed in [147].

Traditional methodologies are based on segmentation of bivariant distributions of two independent variables. Then, a sizing chart and cross tabulation is defined to select the sizes in a stepwise format that accommodate as large a percentage of the population as possible. The main advantage of these types of methods is that they can communicate in an effective way their size to consumers. However, the variability of other key anthropometric dimensions is not taken into account. As a consequence, it is very difficult to fit all the individuals in the same group well and consumer dissatisfaction may arise.

In the clothing development process, the first step is to divide the population into different major size groups usually using one or two control dimensions that are considered to be the most important dimensions for a specific garment. Each group is represented by a body of specific proportions, called body shape or body type. Further subdivisions of the groups using some other secondary dimensions define in more detail the body shape of the size group. Our proposal is consistent with this procedure: First, we segment the data set in 12 segments using bust circumference as the principal control dimension following the recommendations of the European standard. Part 3: Measurements and intervals [59]. Next, we apply a trimmed k-medoids clustering algorithm to each segment using as secondary control dimensions the waist, hip and chest circumference and the neck to ground length. In this way, the pre-segmentation using bust dimension values, which is the primary dimension for upper garment fitting, provides a first easy input for customers to choose the size, while the more realistic prototypes optimize sizing using the anthropometric variability of other secondary dimensions, which are also very important to define the corresponding sizes.

From the theoretical point of view, the question about which number k of clusters to choose is a major issue in clustering methods and in particular, in our application. In defining a sizing system, deciding the number of groups into a population should be divided, in order to optimize benefits and user satisfaction, represents an important problem. On the one hand, it is not profitable to design many sizes because it would increase a lot the production

and distribution costs. On the other hand, to define too few sizes would cause a poor accommodation index. Accordingly, we have fixed the number of groups (sizes) to search for within each bust class to $k = 3$, because this number of sizes is quite well aligned to these considerations. Other theoretical aspects are that we have fixed a disaccommodating rate in order to discard outlier women and an OWA operator has been incorporated to the dissimilarity between women and prototypes.

This approach presents many advantages with regard to currently used systems. In the same way as explained in [147], our methodology selects simultaneously the disaccommodated individuals and prototypes, and assigns the individuals to size classes. But, as main findings, the prototypes returned by our methodology are more realistic because they correspond to specific women of the data set and the use of OWA operators has resulted in a more realistic dissimilarity measure.

2.4 BiclustAnthropom

This section presents the most interesting methodologies and results among those included in my final project for the Master's degree in Biostatistics, granted by the University of Valencia. My Master's thesis, entitled *Biclustering methods applied to anthropometric data: Exploring its possible application in clothing design* (written in Spanish), was aimed at reviewing some biclustering methods implemented in the **biclust** R package and assessing their potential usefulness in the definition of an efficient sizing system oriented toward the clothing design [211].

Biclustering is a data mining technique for clustering rows and columns of a matrix simultaneously. Boris Mirkin coined this term in his 1996 book [149] but the earliest biclustering formulation is the direct clustering introduced in [90]. Biclustering is also known in the literature as co-clustering [15, 123] or two-mode clustering [208], among others names [140].

Given a data set of n rows (observations) in m columns (variables or attributes) (i.e., an $n \times m$ matrix), classical clustering algorithms group by rows using all the m columns. However, working with such data sets, there is always the opportunity to investigate not only properties of samples, but also of their attributes [23]. A well-known clustering procedure on variables is *VARCLUS*, which divides a set of variables into hierarchical clusters [181]. Variable clustering is essentially used for separating variables into clusters

that can be considered as a single variable, thus resulting in data reduction. *VARCLUS* is implemented in the *varclus* function of the **Hmisc** R package [89]. Accordingly, clustering can be applied to either the rows or the columns of the data matrix, separately. When clustering is applied, each row in a given row cluster is defined using all the columns. Similarly, each variable in a variable cluster is characterized by all the observations that belong to it. Instead, biclustering identifies subgroups of rows and subgroups of columns, by performing simultaneous clustering of both rows and columns of the data matrix. In this way, a bicluster is a group of observations that show similar behavior under a specific subset of the attributes. Hence, the main difference between clustering and biclustering is that clustering derives a global model while biclustering, a local one [140].

Mathematically, let $A_{n\times m}$ be a matrix with a row set $X = \{x_1, \ldots, x_n\}$ and a column set $Y = \{y_1, \ldots, y_m\}$, where the element a_{ij} corresponds to a value representing the relation between row i and column j. A bicluster $A'_{k\times s} \equiv A_{IJ} = (I, J)$ is a submatrix of the matrix $A_{n\times m}$, with a row subset $I = \{i_1, \ldots, i_k\}$ ($I \subseteq X$ and $k \leq n$) and a column subset $J = \{j_1, \ldots, j_s\}$ ($J \subseteq Y$ and $s \leq m$).

An interesting approach to classify a particular biclustering algorithm is regarding the type of biclusters obtained. Four predominant classes are identified [140]:

1. Biclusters with constant values.
2. Biclusters with constant values on rows or columns.
3. Biclusters with coherent values.
4. Biclusters with coherent evolutions.

Classes 1, 2 and 3 analyze the numeric values in the data matrix and try to find out subsets of rows and subsets of columns that show similar behaviors. These behaviors may occur on the rows, on the columns, or in both dimensions at the same time.

A perfect constant bicluster is a submatrix (I, J), where all values are equal, $\forall i \in I$ and $\forall j \in J$:

$$a_{ij} = \mu \tag{2.20}$$

Table 2.9a is an example of this type of bicluster.

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

(a)

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

(b)

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

(c)

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

(d)

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

(e)

S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1

(f)

S1	S1	S1	S1
S2	S2	S2	S2
S3	S3	S3	S3
S4	S4	S4	S4

(g)

S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4

(h)

70	13	19	10
49	40	49	35
40	20	27	15
90	15	20	12

(i)

↗	↗	↘	↗
↘	↘	↗	↘
↗	↗	↘	↗
↘	↘	↗	↘

(j)

Table 2.9: Examples of different types of biclusters according to [140]. (a) Constant bicluster, (b) constant rows, (c) constant columns, (d) coherent values (additive model), (e) coherent values (multiplicative model), (f) overall coherent evolution, (g) coherent evolution on the rows, (h) coherent evolution on the columns, (i) coherent evolution on the columns, and (j) coherent sign changes on rows and columns.

A perfect bicluster with constant rows is a submatrix (I, J), where all the values within the bicluster can be calculated by means of one of the following expressions:

$$a_{ij} = \mu + \alpha_i \tag{2.21}$$

$$a_{ij} = \mu \times \alpha_i \tag{2.22}$$

where μ is the typical value within the bicluster and α_i is the adjustment for row $i \in I$. This adjustment can be obtained either in an additive, see eq. (2.21) or multiplicative way, see eq. (2.22).

In the same way, a perfect bicluster with constant columns is a submatrix (I, J), where all the values within the bicluster can be calculated by means of one of the following expressions:

$$a_{ij} = \mu + \beta_j \tag{2.23}$$

$$a_{ij} = \mu \times \beta_j \tag{2.24}$$

where μ is the typical value within the bicluster and β_j is the adjustment for column $j \in J$. Again, this adjustment is obtained either in an additive, see eq. (2.23) or multiplicative way, see eq. (2.24).

Table 2.9b and Table 2.9c show examples of perfect biclusters with constant rows and columns, respectively, for the additive case.

Following the same reasoning, biclusters with coherent values are different depending on whether an additive model or a multiplicative one is evaluated. For the additive case, a perfect bicluster with coherent values is defined as a subset of rows and a subset of columns, whose values a_{ij} can be predicted with this expression:

$$a_{ij} = \mu + \alpha_i + \beta_j \tag{2.25}$$

where μ is the typical value within the bicluster, α_i is the adjustment for row $i \in I$ and β_j is the adjustment for column $j \in J$. An illustration of this type of bicluster can be seen in Table 2.9d. The biclusters of Tables 2.9b and 2.9c are especial cases of this general additive model. This means that eq. (2.21) and eq. (2.23) are particular cases of eq. (2.25) considering $\beta_j = 0$ and $\alpha_i = 0$, respectively.

For the multiplicative case (see Table 2.9e), the values a_{ij} of a perfect bicluster with coherent values can be predicted with this expression:

$$a_{ij} = \mu^{'} \times \alpha_i^{'} \times \beta_j^{'} \tag{2.26}$$

This model is equivalent to the additive one of eq. (2.25) when $\mu = log(\mu^{'})$, $\alpha_i = log(\alpha_i^{'})$ and $\beta_j = log(\beta_j^{'})$.

On the other hand, class 4 tries to find coherent evolutions across the rows and/or columns of the data matrix , without taking into account of its exact numeric values. Hence, the matrix elements can be represented with symbols. There are three options:

- Symbols can be nominal, as in Tables 2.9f, 2.9g and 2.9h.
- Symbols can correspond to a given order, as in Table 2.9i.
- Symbols can represent coherent positive and negative changes relatively to a *normal* value (Table 2.9j).

Biclustering algorithms can also be classified according to the structure of the biclusters they return [140]. There are situations where only one bicluster is obtained, as in Fig. 2.10a.

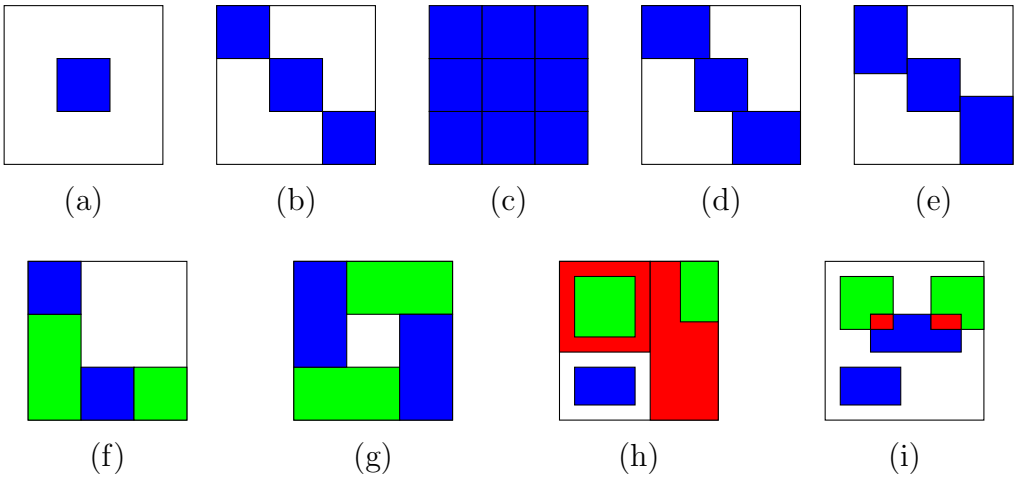


Figure 2.10: Examples of different bicluster structure according to [140]. (a) Single bicluster, (b) exclusive row and column biclusters, (c) checkerboard structure, (d) exclusive rows biclusters, (e) exclusive columns biclusters, (f) nonoverlapping biclusters with tree structure, (g) nonoverlapping nonexclusive biclusters, (h) overlapping biclusters with hierarchical structure, and (i) arbitrarily positioned overlapping biclusters.

Besides, when the existence of several biclusters in the data matrix is assumed, the following bicluster structures can be found:

1. Exclusive row and column biclusters (rectangular diagonal blocks after row and column reorder) (Fig. 2.10b).
2. Nonoverlapping biclusters with checkerboard structure (Fig. 2.10c).
3. Exclusive-rows biclusters (Fig. 2.10d).
4. Exclusive-columns biclusters (Fig. 2.10e).
5. Nonoverlapping biclusters with tree structure (Fig. 2.10f).
6. Nonoverlapping nonexclusive biclusters (Fig. 2.10g).
7. Overlapping biclusters with hierarchical structure (Fig. 2.10h).
8. Arbitrarily positioned overlapping biclusters (Fig. 2.10i).

A third way of classifying biclustering methods is regarding the specific routine used to identify each bicluster. Given the complexity of the problem, five approaches can be distinguished [140]:

1. Iterative row and column clustering combination.
2. Divide and conquer.
3. Greedy iterative search.
4. Exhaustive bicluster enumeration.
5. Distribution parameter identification.

A simple way to obtain biclusters is to apply clustering algorithms to the rows and columns of the data matrix, separately, and then to combine the results using a type of iterative procedure. A divide and conquer algorithm breaks the problem into two or more smaller sub-problems of the same (or related) type, solves the problems recursively and finally, these solutions are combined to give a solution to the original problem. The greedy iterative search typically consists in iterations made up from a locally optimal choice to try to find a globally good solution. Exhaustive bicluster enumeration approaches assume that the best biclusters can only be identified by means of an exhaustive enumeration of all possible biclusters inside the matrix. At last, distribution parameter identification methods are based on the idea that the biclusters are obtained using a given statistical model and try to identify the distribution parameters used to generate the data by minimizing a certain criterion through an iterative loop.

During last years, a number of biclustering algorithms have been developed. A very detailed survey about such methods can be found in [140]. Two other informative surveys are [198] and [23]. Biclustering has been widely used for the analysis of gene expression data [140, 164] but it has also been applied to other different fields, such as market segmentation (in tourism [45], with different data sets from the marketing area [77] and for cosmetic products [186]), in information retrieval and text mining [42, 43], electoral data [90], or to nutrition data and some foreign exchange data [126]. Several biclustering algorithms are available from different sources, including R. For instance, Barkow et al [10] created a Graphical User Interface (GUI) called BicAT (Biclustering Analysis Toolbox) to help researchers with the analysis

and exploration of genetic data. BicAT provides five biclustering and two standard clustering algorithms. This interface is very easy to use. However, the programming code of the algorithms is not free, so the users cannot change and adapt them to analyze their own data. Another pitfall is that the results generated by BicAT cannot be directly used as input for other statistical softwares. So far, the most complete R package for biclustering is **biclust** [112, 113]. It includes the algorithms related to five methods: *Bimax* [164], *Cheng & Church* [32], *plaid model* [206, 126], *Spectral* [123] and *Xmotifs* [153]. The main function of **biclust** is *biclust*, that allows to execute the corresponding biclustering algorithm specified in its method-argument [113, 112]:

- method=BCBimax() or method=BCrepBimax(): *Bimax*.
- method=BCCC(): *Cheng & Church*.
- method=BCPlaid(): *plaid model*.
- method=BCSpectral(): *Spectral*.
- method=BCXmotifs(): *Xmotifs*.

Excepting *Bimax*, that only works with binary and multiple choice data, the other four deal with continuous data.

Other biclustering R packages such as **BicARE** [80] and **fabia** [95] have been developed in the framework of Bioconductor, which is an open source software for the analysis and comprehension of genomic data. Its website is http://www.bioconductor.org/.

The main conclusion of my Master's final project was that the proposed methodology using *Cheng & Church* could be considered as a potential statistical approach to be used for the clothing design. Besides, *Bimax* was helpful to examine the eating habits of the Spanish women. This way of using *Bimax* was inspired by the strategy followed in [45].

Section 2.4.1 presents the foundation of the *Cheng & Church* biclustering algorithm (from now on, CC). Section 2.4.2 focuses on the methodology developed using CC (called *biclustAnthropom*) and on the results obtained. These results are discussed in Section 2.4.3. The *Bimax* algorithm and its application can be found in Appendix.

2.4.1 Methodology

Before introducing the CC algorithm, some notation is needed:

- Mean of the ith row in the bicluster:

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij} \tag{2.27}$$

- Mean of the jth column in the bicluster :

$$a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij} \tag{2.28}$$

- Mean of all elements in the bicluster:

$$a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij} = \frac{1}{|I|} \sum_{i \in I} a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{Ij} \tag{2.29}$$

The CC biclustering algorithm was introduced by Yizong Cheng and George M. Church in [32]. For them, a bicluster is a subset of rows and a subset of columns with a high similarity score. In order to check the similarity of a bicluster, they consider the mean squared residue, called H, that measures the coherence of the rows and columns in the bicluster. The CC algorithm aims at finding a maximum bicluster with the quantity H lower than a certain threshold $\delta \geq 0$. A δ-bicluster is *perfect* if $\delta = 0$. Biclusters of Tables 2.9b, 2.9c, 2.9d and 2.9e are examples of perfect biclusters. A perfect bicluster meets this relation:

$$a_{ij} = a_{iJ} + a_{Ij} - a_{IJ} \tag{2.30}$$

However, δ-biclusters are not usually perfect biclusters due to the presence of noise in data. Consequently, the concept of residue arises to quantify the difference between each element a_{ij} and its expected value predicted from a_{iJ}, a_{Ij} y a_{IJ}:

$$r(a_{ij}) = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}. \tag{2.31}$$

Then, CC defines a δ-bicluster as the subset of rows and the subset of columns where the following expression is verified:

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} r(a_{ij})^2 < \delta, \ for\ some\ \delta \geq 0. \tag{2.32}$$

In other words, the overall quality of a δ-bicluster is evaluated by means of the mean squared residue.

CC returns arbitrarily positioned (possibly overlapping) biclusters with coherent values (corresponding with Tables 2.9d and 2.9e and Fig. 2.10i). Besides, the CC algorithm is a greedy iterative search method. The authors proposed several greedy row/column removal/addition algorithms that are combined in an overall approach that allows to identify a given number of δ-biclusters. Specifically, the algorithm itself has the next major steps [32, 112]:

1. Let A be a matrix of real numbers, let $\delta \geq 0$ be the maximum acceptable mean squared residue score and let $\alpha > 1$ be the threshold for multiple node deletion. Compute a_{iJ}, a_{Ij}, a_{IJ} and $H(I,J)$ $\forall i \in I \ \ \forall j \in J$, with row subset I and column subset J of A.

   Note: If this $H(I,J)$ is fewer than δ, return the bicluster A_{IJ}.

2. Deleting rows $i \in I$ with $\frac{1}{|J|} \sum_{j \in J} r(a_{ij})^2 > \alpha H(I,J)$

3. Deleting columns $j \in J$ with $\frac{1}{|I|} \sum_{i \in I} r(a_{ij})^2 > \alpha H(I,J)$

4. Adding rows $i \notin I$ verifying $\frac{1}{|J|} \sum_{j \in J} r(a_{ij})^2 < \delta$

5. Adding columns $j \notin J$ verifying $\frac{1}{|I|} \sum_{i \in I} r(a_{ij})^2 < \delta$

These steps are repeated until a maximum number of biclusters is reached or no bicluster is found.

2.4.2 Results

The data set used here is made up of the same 6013 women as in *trimowa*. However, the CC algorithm is applied to a different set of anthropometric dimensions. The idea behind this approach is to use the fact that designing and manufacturing lower body garments depend 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 identifies groups of observations with a similar pattern in a subset of attributes instead of in the whole of them. Therefore, it seems more interesting to use CC with lower body dimensions. In this way, only those measurements related to the lower body part (36 numeric variables) are finally chosen. Our way of proceeding is the following: first, the data set is divided into twelve segments (classes) using waist circumference values according to the European standard. Part 3: Measurements and intervals [59]. Then, CC is applied to each waist class. In this way, for each waist size, each bicluster may be defined for a different number of anthropometric dimensions which will be the most relevant for designing the garments for the women belonging to that bicluster. All individuals in a same bicluster (group) can wear a garment with dimensions specific to that group.

In order to use the CC algorithm within the *biclust* function of **biclust**, the maximum number of biclusters to be found must be indicated. We propose to fix this number for each waist size according to the number of women it contains. We have guided us by the clothing companies' policy regarding the production and distribution time and costs, already mentioned in Section 2.3.3. Table 2.10 summarizes this suggestion.

Number of women in each size	Maximum number of biclusters to be found
< 150	2
$151 - 300$	3
$351 - 450$	4
> 451	5

Table 2.10: Proposed number of biclusters to be found in each size.

From the definition of the mean squared residue H, the results provided by CC may be very influenced by a_{iJ} y a_{IJ} in case of variables involved in the study are on very different scales. In fact, standardizing the data by subtracting the mean and by dividing by the standard deviation is widely

recommended before applying this algorithm [32, 112, 113]. The selected variables related to lower body part were indeed on different scales. As a first practical attempt, CC was applied after standardazing the data. However, no biclusters were found for any waist size. The same happened with some other types of preprocessing methods implemented in the **clusterSim** R package [49]. To overcome this problem, we propose to select in each waist class the variables that have a similar scale. We think this is the best choice so that the influence of a_{iJ} y a_{IJ} dissappear. We mean by a similar scale a difference among the variables ranges less or equal than 7 cm. We decided to use an upper limit of 7 cm, because the number of variables which had a difference of ranges less or equal than 7 in each segment, was an enough and suitable number to be able to design a garment in a given size, following the recommendations of experts. Thus, each data matrix referred to each waist size contains those women with a similar waist measurement and with those secondary control dimensions with a similar scale. Table 2.11 details each waist size.

Waist segment	$[58, 62[$	$[62, 66[$	$[66, 70[$	$[70.74[$	$[74, 78[$	$[78, 82[$
Number of variables	14	10	13	7	10	15
Number of women	15	121	414	673	809	785
Waist segment	$[82, 86[$	$[86, 91[$	$[91, 97[$	$[97, 103[$	$[103, 109[$	$[109, 115[$
Number of variables	11	9	11	9	12	12
Number of women	804	786	676	464	277	162

Table 2.11: Number of women and number of variables with a similar scale of each waist segment.

In addition to the number of biclusters to search for, two other arguments are required to call the CC algorithm with the function *biclust*. They are δ ($\delta \geq 0$), which is the maximum acceptable mean squared residue score, and α, which is a threshold for multiple node deletion ($\alpha > 1$). We maintain the same default value fixed for α within *biclust*. On the other hand, the CC algorithm might not group every woman into a bicluster (this property is called nonexhaustivity [140]), so we can consider those no grouped women as disaccommodated women. Therefore, the value of δ can be iteratively adapted to the number of disaccommodated women we want to discard in each size. The proportion of no accommodated sample was prefixed to 0.01 per segment. In this way, a number of women between 0 and the previous

fixed proportion will not be assigned to any group. The detailed methodology is given in Algorithm 2.

Algorithm 2 Algorithm to find biclusters and no accommodated women with CC

We define for each size the objects **nc**, **delta** and **disac**.
nc is the proposed number of biclusters to be found in each size.
delta is the parameter δ of the Cheng and Church method. Its initial value is 1 (by default for *method=BCCC()*).
disac is the number of women who will not form part of any group. At the beginning, it is equally to the number of women belonging to each size.
The proportion of disaccommodated sample is prefixed to 1% per segment.
while *disac* > *ceiling*(0.01 ∗ *number of women belonging to the size*) **do**
  biclust(SizeData, method = BCCC(), delta = delta, alpha = 1.5, number = nc)
  disac = number of women not grouped.
  delta = delta + 1
end while

2.4.2.1 Experimental results

Once the theoretical and practical details of the CC algorithm are explained, we can examine the experimental results obtained for each waist size. From Table 2.12 to Table 2.23 the corresponding results for each waist size are detailed.

Women and dimensions of this class: 15 14	
Disaccommodated women: 0	
δ value: 2	
	BC 1
Number of women:	15
Number of dimensions:	13

Table 2.12: CC results for waist size $[58, 62[$ cm.

Women and dimensions of this class: 121 10		
Disaccommodated women: 0		
δ value: 2		
	BC 1	BC 2
Number of women:	77	44
Number of dimensions:	9	8

Table 2.13: CC results for waist size $[62, 66[$ cm.

Women and dimensions of this class: 414 13			
Disaccommodated women: 0			
δ value: 6			
	BC 1	BC 2	BC 3
Number of women:	323	81	10
Number of dimensions:	13	7	7

Table 2.14: CC results for waist size $[66, 70[$ cm.

Women and dimensions of this class: 673 7					
Disaccommodated women: 2					
δ value: 3					
	BC 1	BC 2	BC 3	BC 4	BC 5
Number of women:	415	111	90	46	9
Number of dimensions:	7	7	7	5	6

Table 2.15: CC results for waist size $[70, 74[$ cm.

Women and dimensions of this class: 809 10					
Disaccommodated women: 4					
δ value: 3					
	BC 1	BC 2	BC 3	BC 4	BC 5
Number of women:	388	119	136	93	69
Number of dimensions:	10	10	10	8	8

Table 2.16: CC results for waist size $[74, 78[$ cm.

Women and dimensions of this class: 785 15					
Disaccommodated women: 0					
δ value: 4					
	BC 1	BC 2	BC 3	BC 4	BC 5
Number of women:	382	123	130	97	53
Number of dimensions:	15	15	15	11	11

Table 2.17: CC results for waist size $[78, 82[$ cm.

Women and dimensions of this class: 804 11					
Disaccommodated women: 0					
δ value: 5					
	BC 1	BC 2	BC 3	BC 4	BC 5
Number of women:	497	98	82	90	37
Number of dimensions:	11	11	11	8	8

Table 2.18: CC results for waist size $[82, 86[$ cm.

Women and dimensions of this class: 786 9					
Disaccommodated women: 8					
δ value: 6					
	BC 1	BC 2	BC 3	BC 4	BC 5
Number of women:	488	104	114	43	29
Number of dimensions:	9	9	9	8	6

Table 2.19: CC results for waist size $[86, 91[$ cm.

Women and dimensions of this class: 676 11					
Disaccommodated women: 6					
δ value: 5					
	BC 1	BC 2	BC 3	BC 4	BC 5
Number of women:	427	88	99	47	9
Number of dimensions:	11	11	10	8	11

Table 2.20: CC results for waist size $[91, 97[$ cm.

Women and dimensions of this class: 464 9					
Disaccommodated women: 5					
δ value: 7					
	BC 1	BC 2	BC 3	BC 4	BC 5
Number of women:	302	67	66	15	9
Number of dimensions:	9	8	8	8	6

Table 2.21: CC results for waist size $[97, 103[$ cm.

Women and dimensions of this class: 277 12			
Disaccommodated women: 0			
δ value: 5			
	BC 1	BC 2	BC 3
Number of women:	134	96	47
Number of dimensions:	12	8	8

Table 2.22: CC results for waist size $[103, 109[$ cm.

Women and dimensions of this class: 162 12			
Disaccommodated women: 2			
δ value: 7			
	BC 1	BC 2	BC 3
Number of women:	97	57	6
Number of dimensions:	11	7	8

Table 2.23: CC results for waist size $[109, 115[$ cm.

When analyzing Tables 2.12 to 2.23, some interesting results are found. Firstly, for almost all waist sizes, it can be observed that some biclusters are defined for the whole set of variables considered in each size. These particular groups could have been identified by any conventional clustering method, like k-means. However, there are also other biclusters defined for a smaller number of dimensions regarding the total number. Traditional clustering could not have defined them. This particular behavior can be seen for example in Table 2.17: the first three biclusters are defined for the whole set of 15 dimensions considered in the waist size $[78, 82[$ cm, whereas the last two are defined only for 11. Even more, for the $[109, 115[$ cm segment, no bicluster is identified for the whole set of 12 dimensions, see Table 2.23.

This is one of the main advantages of biclustering: It allows to define very homogeneous and restrictive groups. We mean by homogeneous groups those ones whose members agree on all the variables that are characteristic for that group. In addition, we mean by restrictive groups those ones defined by only those variables, from the total number of them, that are truly relevant for defining the group. Classical clustering algorithms weigh each variable equally, so they are viewed as equally important in providing a segmentation. However, this might not be a desirable option for many scenarios. It may happen that some attributes included in the database are not actually critical to the construction of groups. Biclustering can solve this problem without needing to preprocess data, by using variable selection methods before segmenting. By looking for observations that show a similar pattern, non-informative dimensions are automatically ignored because they do not demonstrate such systematic patterns. This property of biclustering is very valuable to data analysts because they can safely assume that less relevant variables do not bias the entire segmentation results. In anthropometric terms, we are finding women with very similar features for specific body dimensions. These specific variables describe a body in the detail necessary to construct a garment to fit that body.

Another relevant result is that, in some waist classes, the number of disaccommodated women is zero (see for example Tables 2.12, 2.13), so CC manages to find groups for which every woman can be adapted. Moreover, the δ parameter is a reduced value in all cases, demonstrating not only the speed with which this algorithm has been able to find biclusters, given the relative restriction concerning the percentage of disaccommodated women, but also and what is more important, that all biclusters have a high similarity coefficient. A third interesting property of these results is that, in this case, the returned biclusters by CC are not overlapped. This is very important in our application because each individual must be assigned to a single size. Since it is a deterministic algorithm, these results are reproducible.

Next, we show some graphical results. Fig. 2.11 (resp. Fig. 2.12) shows the scatter plots of waist circumference against neck to ground (resp. hip), jointly with the median woman of all the biclusters obtained for each waist class. Plotting waist against neck to ground and hip is the usual analysis in the literature for lower body part clothing sizing. Fig. 2.11 shows quite distributed biclusters for all the waist segments, while for Fig. 2.12 this behavior is only appreciated for certain segments.

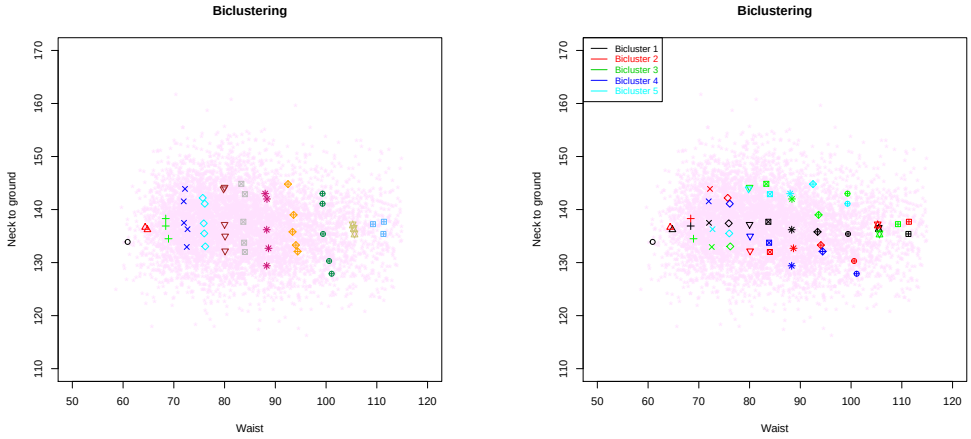


Figure 2.11: Waist circumference against neck to ground, jointly with the median woman of all the biclusters obtained for each waist class. Right plot helps to identify the biclusters of each waist class.

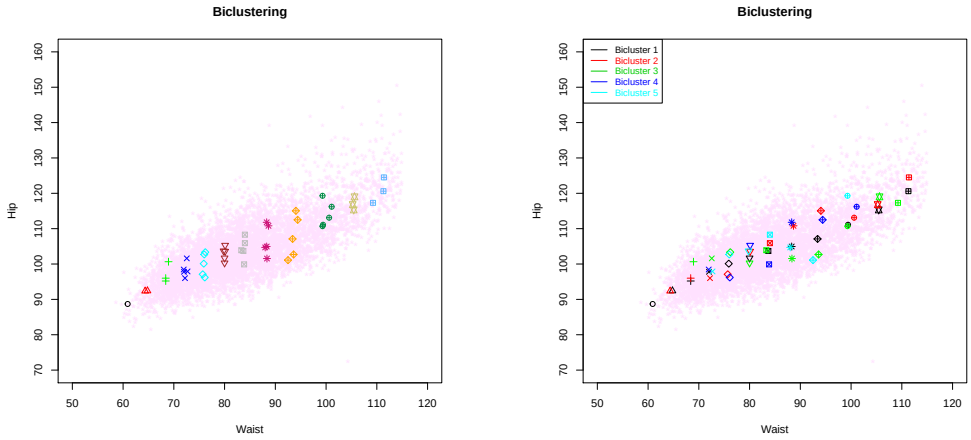


Figure 2.12: Waist circumference against hip circumference, jointly with the median woman of all the biclusters obtained for each waist class. Right plot helps to identify the biclusters of each waist class.

It is also meaningful to examine separately the differences that may exist among biclusters for each waist segment. As an illustrative example, the results for the $[74, 78[$ cm size (Table 2.16) will be analyzed because they

are representative of the results for the rest of sizes. Fig. 2.13 shows the boxplots for two variables that influence on the lower body clothing design, such as buttock girth and thigh horizontal girth. Boxplots are ordered from lower to greater median. Differences among biclusters can be seen for both variables. We also note that all the boxplots, with the exception of bicluster 4 for buttock girth and bicluster 5 for thigh girth, present an almost symmetric distribution of the data.

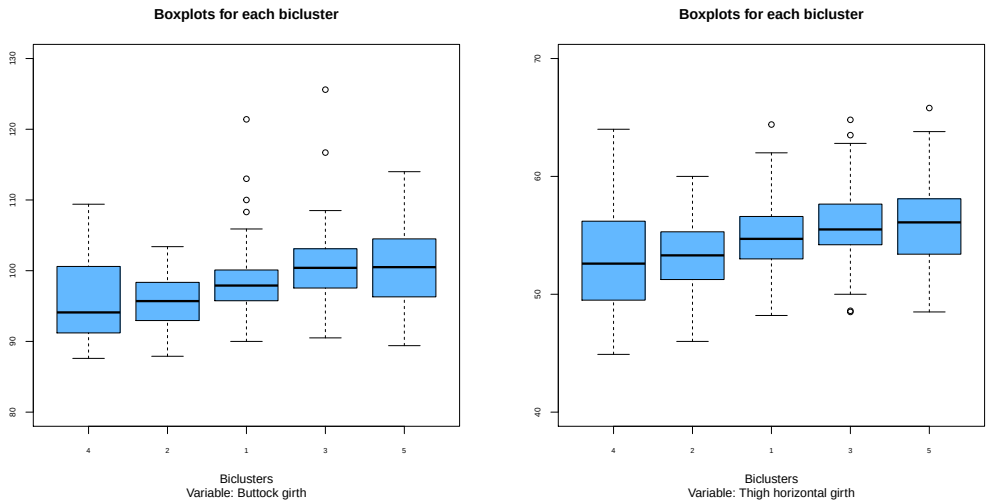


Figure 2.13: Boxplots of buttock girth and thigh horizontal girth for waist size $[74, 78[$ cm.

2.4.3 Summary

The goal of my Master's thesis was to explore if any of the biclustering algorithms included in the **biclust** R package may be useful to deal with anthropometric data in order to try to define an efficient sizing system oriented towards the apparel design. The main conclusion reached there was that the *Cheng & Church* algorithm (CC) provided the most promising results. In this section we have presented the foundation and the application of this biclustering method.

We have proposed a way of proceeding to enable CC to discard some individuals who might be considered as disacommodated regarding a set of measurements. Results returned by CC show that biclustering, unlike conventional clustering methods, obtains groups not always defined for the whole

number of anthropometric dimensions and at the same time, identifies which are the most relevant to define each group. In addition, results are reproducible and biclusters are not only very restrictive but also not overlapped.

These features are interesting in garment design. For each waist size, we find women groups with similar body measurements for a specific number of dimensions. As known, the women of a same group will wear the same garment. That garment should be designed considering only those dimensions identified as the most relevant ones, in addition to the waist circumference.

Moreover, we have included in Appendix the results provided by the *Bimax* biclustering algorithm when analyzing eating habits.

According to our analysis, we can say that biclustering, at least some of biclustering algorithms, can be considered as a potential alternative to conventional clustering procedures when analyzing both anthropometric and sociological data.

2.5 HipamAnthropom

This methodology is concerned with the generation of optimal and representative fit models for use in apparel sizing [215]. We use the hierarchical partitioning around medoids (HIPAM) clustering algorithm, originally developed to work with gene expression data [222]. We modify it to deal with anthropometric features by incorporating the dissimilarity explained in Section 2.2.3. HIPAM is a divisive hierarchical clustering algorithm using PAM. We propose two HIPAM algorithms. The first one, called $HIPAM_{MO}$, is a slightly modification of HIPAM that uses the dissimilarity of Section 2.2.3. $HIPAM_{MO}$ uses asw as a measure of cluster structure and the maximization of the asw as the rule to subdivide each already accepted cluster. The use of asw could be too restrictive. Therefore, we propose a second algorithm, $HIPAM_{IMO}$, where the differences regarding the original HIPAM are even deeper. It incorporates a different criterion: the INCA statistic criterion (see Section 2.2.1) to decide the number of child clusters and as a stopping rule. Both algorithms return a set of representative objects (the medoids). As mentioned, medoids correspond to real individuals of the database, so they could be used as representative fit models of the target market. The HIPAM algorithm also makes it possible to identify outliers. Given its hierarchical nature, HIPAM returns clusters which contain only one or two women. In the hierarchical clustering methods, all clusters with only one element (singleton

clusters) are consider as outliers. Regarding clusters with two elements, this is because three is the minimum number of elements for clustering with PAM. Extending the property that singleton clusters are considered as outliers in the hierarchical clustering algorithms, we can also consider the clusters with two elements as outliers. Remind that discarding the individuals who might be considered as outliers regarding their measurements is particularly important in the apparel sizing context, where the target population is not the whole population.

Functions of the **smida** R package [223] will be used to implement both $HIPAM_{MO}$ and $HIPAM_{IMO}$. This package is freely available from the authors website: http://www.stats.gla.ac.uk/ microarray/book/smida.html.

2.5.1 Methodology

We detail next both $HIPAM_{MO}$ and $HIPAM_{IMO}$. They start with all objects in one single cluster, the highest or top node, T. At each level of the classification tree, each node corresponds with a cluster. End nodes define the final partition.

First, we describe the algorithm $HIPAM_{MO}$. We refer to [207, 222, 223] for more details. For a given node P, the algorithm must decide whether it is advisable to split this (parent) cluster into new (child) clusters, or to stop. If $|P| \leq 2$, then it is an end (or terminal) node; otherwise, PAM is applied to P with k_1 groups, where k_1 is chosen by maximizing the asw of the new partition. After a post-processing step, where further partitioning or collapsing procedures for the k_1 clusters try to improve the asw, a partition $C = \{C_1, \ldots, C_k\}$ is finally obtained from P (k is not necessarily equal to k_1). Next, the asw of C, asw_C, is obtained, and the same steps used to generate C are applied to each C_i to generate a new partition. If we denote the asw of the new partition with $i = 1, \ldots, k$, using SS_i (if $|C_i| \leq 2$ then $SS_i = 0$), then the Mean Split Silhouette (MSS) is defined as the mean of the SS_i. If $MSS(k) < asw_C$, then these new k child clusters of the partition C are included in the classification tree. Otherwise, P is a terminal node. MSS is a rule to evaluate the average homogeneity of the partition C.

Fig. 2.14 helps to understand this procedure: *Node 2* corresponds to the previous given node P, the partition p_2 (consisted of *Node 4* and *Node 5*, red-colored to indicate that it is the candidate partition) to C and SS_i, $i = 1, 2$, would be calculated for the partitions p_3 and p_4, respectively.

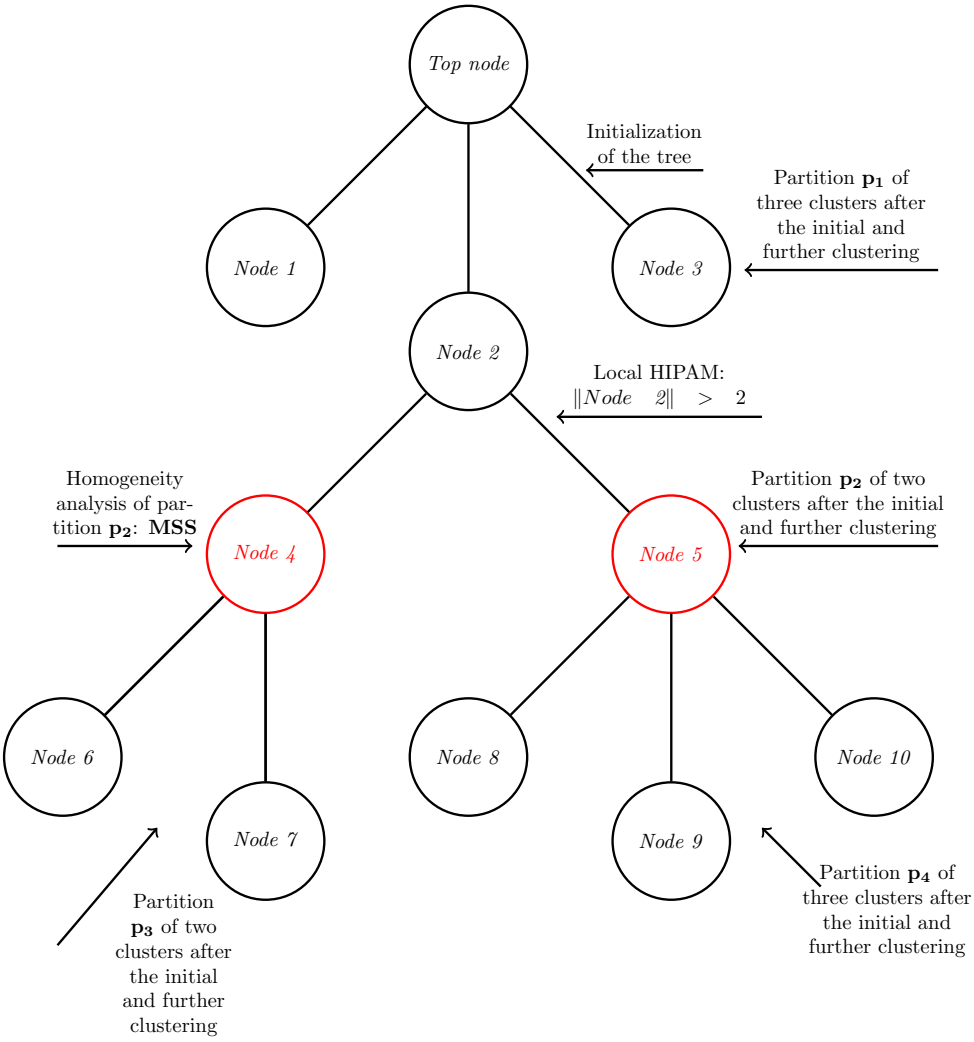


Figure 2.14: Mean Split Silhouette (MSS) procedure within the *HIPAM* algorithm.

The algorithm $HIPAM_{IMO}$ is summarized in Algorithm 3.

Algorithm 3 $HIPAM_{IMO}$

1. Initialization of the tree:
Let the top cluster with all the elements be T.
1.1. Initial clustering: Apply PAM to T with the number of clusters, k_1, provided by the INCA statistic with the following rule:
if $INCA_{k_1} < 0.2\ \forall k_1$ **then**
  $k_1 = 3$
else
  k_1 as the value preceding the first largest slope decrease.
end if
An initial partition with k_1 clusters is obtained.
1.2. Post-processing: Apply several partitioning or collapsing procedures to the k_1 clusters to try to improve the asw.
A partition with k clusters from T is obtained.
2. Local HIPAM:
while there are active clusters **do**
  Generation of the candidate clustering partition: *PHASE I FOR* $HIPAM_{IMO}$
  Evaluation of the candidate clustering partition: *PHASE II FOR* $HIPAM_{IMO}$
end while

Subroutine 1 PHASE I FOR $HIPAM_{IMO}$

For each cluster, P, of a partition:
1.
if $|P| \leq 2$ **then**
  STOP (P is a terminal node).
else
  if $INCA_{k_1} < 0.2\ \forall k_1$ **then**
    STOP (P is a terminal node).
  else
    2. Initial clustering: Apply PAM to P with the number of clusters, k_1, provided by the INCA statistic as the value preceding the first largest slope decrease. An initial partition with k_1 clusters is obtained.
    3. Post-processing: Apply several partitioning or collapsing procedures to the k_1 clusters to try to improve the asw.
    The candidate partition, $C = \{C_1, \ldots, C_k\}$, from P is obtained.
  end if
end if

Subroutine 2 PHASE II FOR $HIPAM_{IMO}$

Let the candidate clustering partition be $C = \{C_1, \ldots, C_k\}$ obtained from P.
1. Calculate the asw of C, asw_C.
2. For each C_i, generate a new partition using steps **1.1.** and **1.2.** of the initialization of the tree and calculate its SS_i.
3.
if $MSS(k) = \frac{1}{k}\sum_{i=1}^{k} SS_i < asw_C$ **then**
C is accepted.
else
C is rejected. STOP (P is a terminal node).
end if

The two algorithms differ mainly in terms of the use of the INCA criterion.

1. At each node P, if there is k such that $INCA_k > 0.2$, then we select the k prior to the first largest slope decrease.

2. On the other hand, if $INCA_k < 0.2$ for all k, then P is a terminal node.

However, this procedure does not apply either to the top node T or to the generation of the new partitions from which the MSS is calculated. In these cases, even when all $INCA_k < 0.2$, $k = 3$ is fixed as the number of groups to divide and proceed. Fig. 2.15 shows a flowchart representing the steps of $HIPAM_{IMO}$.

2.5.2 Results

We use the same database (6013 individuals and 5 body dimensions), parameters for the dissimilarity and procedure as in Section 2.3.2, where *trimowa* was explained. Two preliminary pre-segmentations have been performed. The first one uses bust circumference (see Section 2.5.2.1) and the second one is based on geographical location, i.e., takes into account the region of provenance of women, see Section 2.5.2.3.

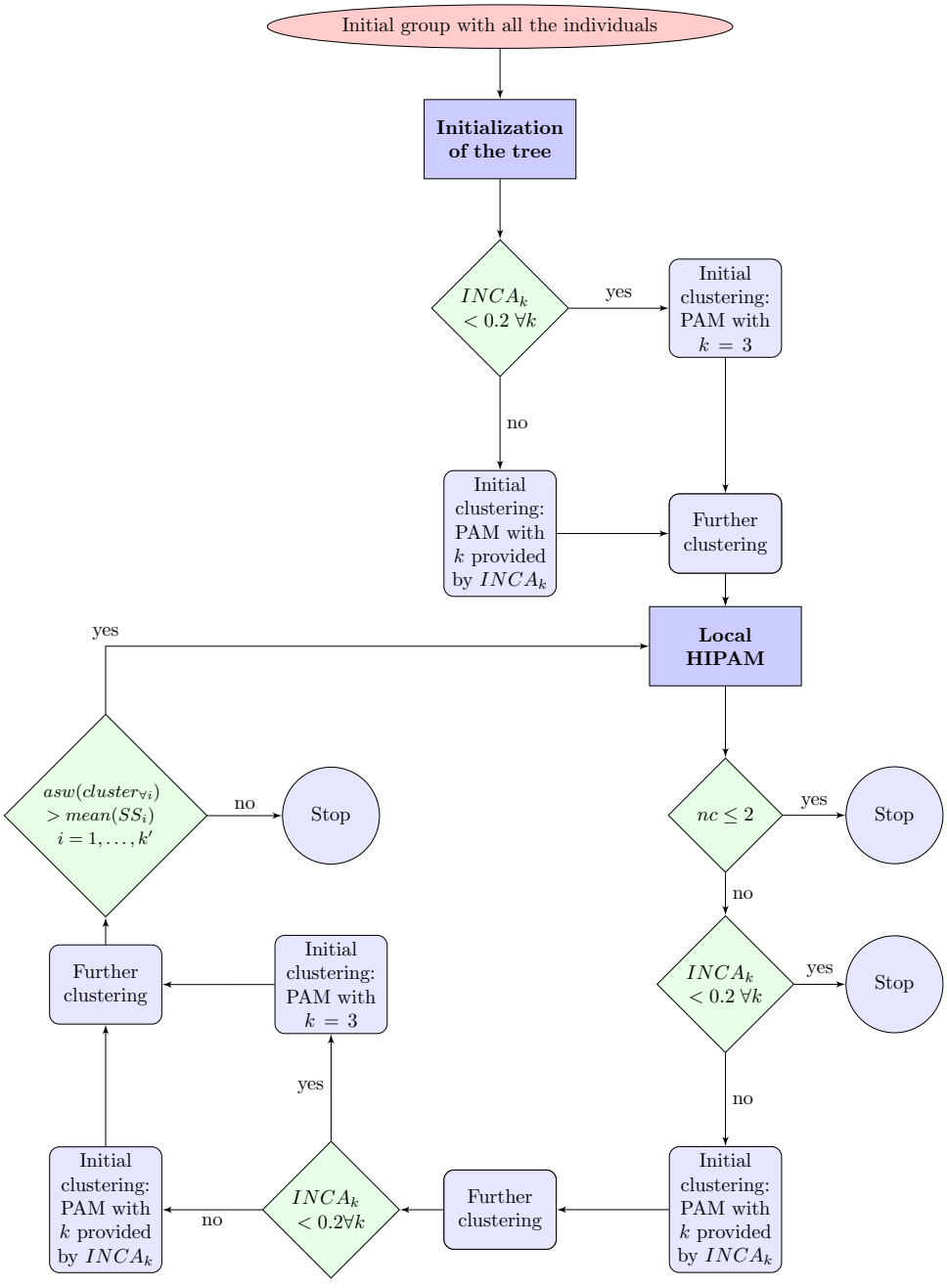


Figure 2.15: Flowchart as a guide to explain how $HIPAM_{IMO}$ works.

2.5.2.1 Bust circumference

For this particular study we have only considered the classes corresponding to the 8 central bust sizes (from 78-82 to 107-113, see Table 2.24) with more than 250 women per class. This is because there are a few women with a bust circumference of less than 78 cm or larger than 113 cm. Then, $HIPAM_{MO}$ and $HIPAM_{IMO}$ have been applied to each class. Table 2.25 details the number of clusters with more than two women returned with each bust class and Table 2.26 shows the number of women that contains each one of those clusters.

Bust	78-82	82-86	86-90	90-94	94-98	98-102	102-107	107-113
Waist	62-66	66-70	70-74	74-78	78-82	82-86	86-91	91-97
Hip	86-90	90-94	94-98	98-102	102-106	106-110	110-115	115-120

Table 2.24: Measurements used to define the eight bust sizes of European standard. Part 3 [59] involved in the analysis of the HIPAM methodology.

Bust class	[78,82[	[82,86[	[86,90[	[90,94[	[94,98[	[98,102[	[102,107[	[107,113[
Count	287	732	1028	952	818	633	547	356
Num. medoids $HIPAM_{MO}$	5	3	10	5	4	4	9	2
Num. medoids $HIPAM_{IMO}$	3	3	3	3	3	3	3	5

Table 2.25: Counts and number of clusters with more than two women obtained using the algorithms $HIPAM_{MO}$ and $HIPAM_{IMO}$.

Algorithm \ Bust class	[78,82[	[82,86[	[86,90[	[90,94[	[94,98[	[98,102[	[102,107[	[107,113[
$HIPAM_{MO}$	88 65 25 67 42	201 304 227	309 94 70 94 71 113 191 27 27 24	318 277 114 160 83	280 282 83 173	163 181 170 119	141 31 19 39 65 20 99 53 72	202 154
$HIPAM_{IMO}$	98 79 110	234 289 209	366 329 333	318 339 295	331 292 195	182 229 222	213 140 194	140 130 21 47 18

Table 2.26: Size of the clusters with more than two women obtained by $HIPAM_{MO}$ and $HIPAM_{IMO}$.

Fig. 2.16 (resp. Fig. 2.17) shows the scatter plots of bust versus hip circumference (resp. neck to ground), while Fig. 2.18 shows the scatter plots of hip versus waist circumference, together with the representation of the medoids in the clusters with more than two elements (left plot for $HIPAM_{MO}$ and right plot for $HIPAM_{IMO}$).

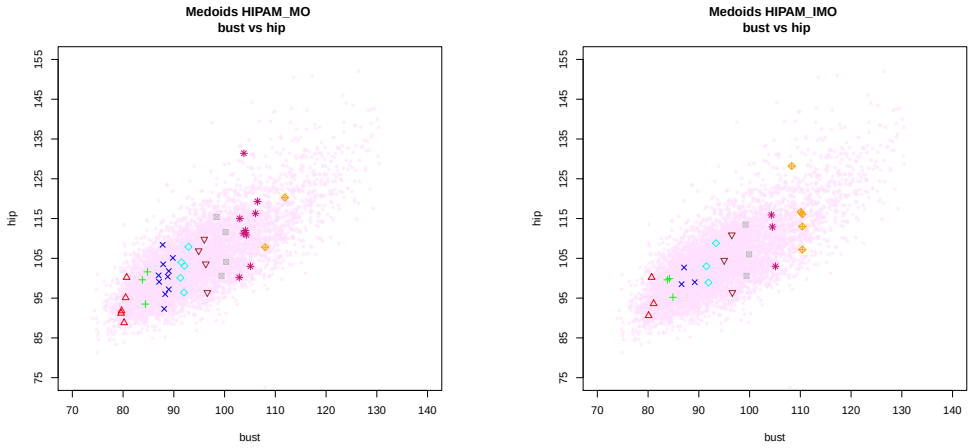


Figure 2.16: Bust vs. hip in the medoids obtained using $HIPAM_{MO}$ (left) and $HIPAM_{IMO}$ (right).

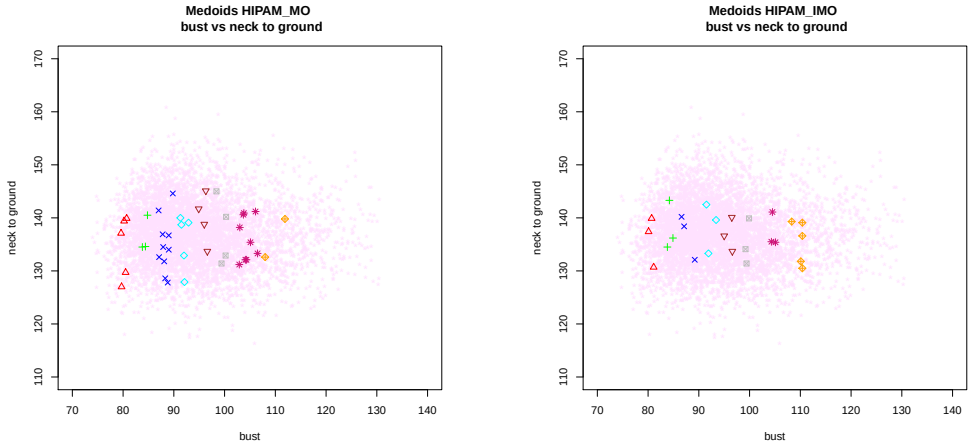


Figure 2.17: Bust vs neck to ground in the medoids obtained using $HIPAM_{MO}$ (left) and $HIPAM_{IMO}$ (right).

The main difference between the two algorithms is in the rule which selects the number of clusters to be found. If we pay attention for Fig. 2.16 where bust and hip (one of the critical dimensions for determining body morphotypes) are displayed, we see that $HIPAM_{IMO}$ (right plot) identifies fewer clusters (around 3) which are more balanced regarding the number of elements assigned to each group and are better distributed.

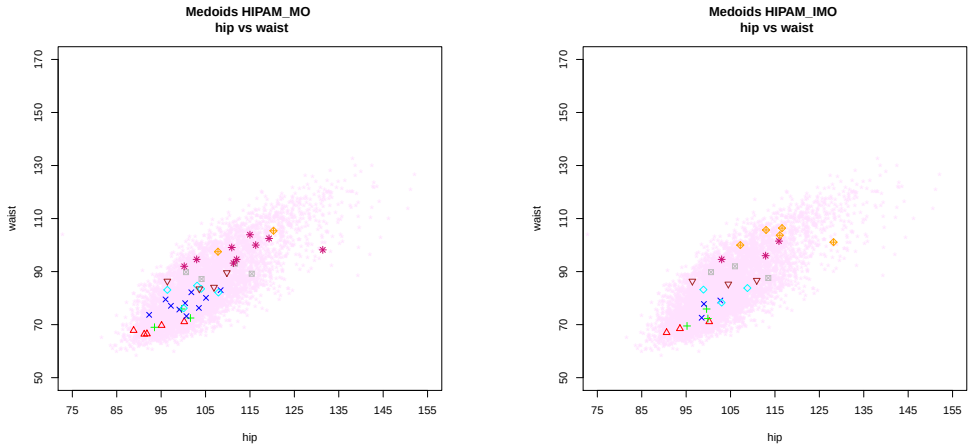


Figure 2.18: Hip vs waist in the medoids obtained using $HIPAM_{MO}$ (left) and $HIPAM_{IMO}$ (right).

However, for the same two dimensions, $HIPAM_{MO}$ (left plot) shows closer medoids for each class and some of them are really quite close or even overlapping. A similar behavior can be seen in Fig. 2.17. These results can be compared with the PCA-based sizing method, which is the most commonly used approach in the literature for generating sizes from anthropometric data [85, 96, 230]. As an example, Fig. 2.19 shows the locations of the medoids in the $[86, 90[$ cm class over the first two principal components, that represent 60% of the variance. As we expected, the first PC groups the four body circumferences considered (bust, chest, waist and hip) and the second PC refers to the neck to ground variable.

Regarding outliers, $HIPAM_{MO}$ detects 92 whereas $HIPAM_{IMO}$ detects 82. Forty of them match. Table 2.27 shows the percentage of women that are considered as outliers by both algorithms, for each bust size where outliers are founded out.

Algorithm \ Bust class	[74,78[	[86,90[	[102,107[	[113,119[	[119,125[	[125,131[
$HIPAM_{MO}$	30%	0.77%	1.5%	6.5%	23%	78%
$HIPAM_{IMO}$	17%	—	—	14%	9%	100%

Table 2.27: Percentage of outliers in each class.

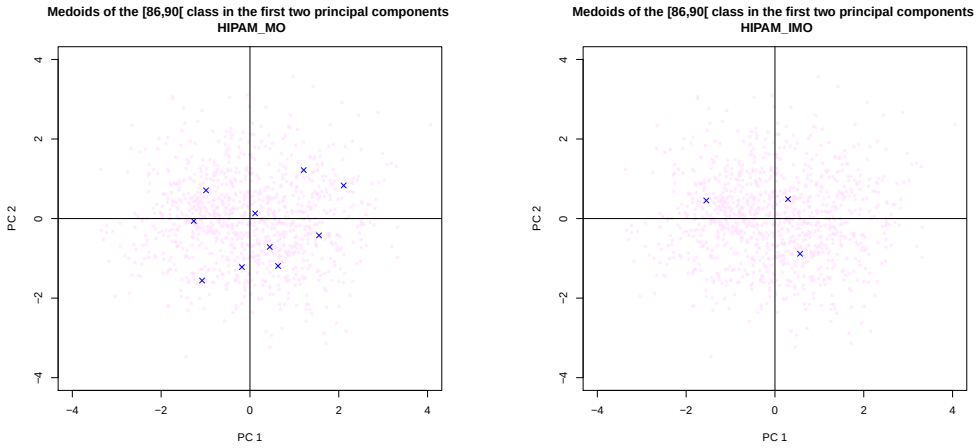


Figure 2.19: The first two principal components of the medoids in the [86,90[cm class obtained using both algorithms.

Figure 2.20 (resp. Fig. 2.21) shows the scatter plots of bust versus hip (resp. neck to ground), together with the outliers detected by $HIPAM_{MO}$ (left plot) and $HIPAM_{IMO}$ (right plot), while Fig. 2.22 represents the scatter plots of hip vs waist with the same outliers. As we can see, $HIPAM_{IMO}$ only identifies outliers in the four bust classes corresponding to small and large sizes.

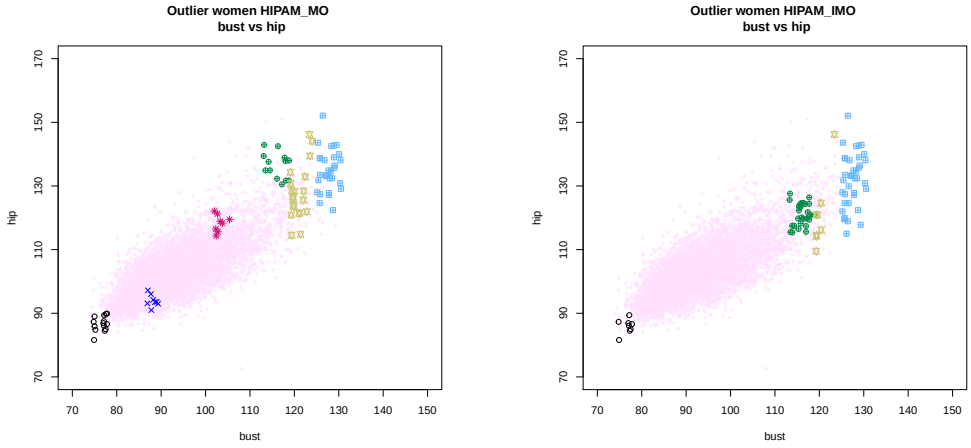


Figure 2.20: Bust vs. hip in the outliers obtained using $HIPAM_{MO}$ (left) and $HIPAM_{IMO}$ (right).

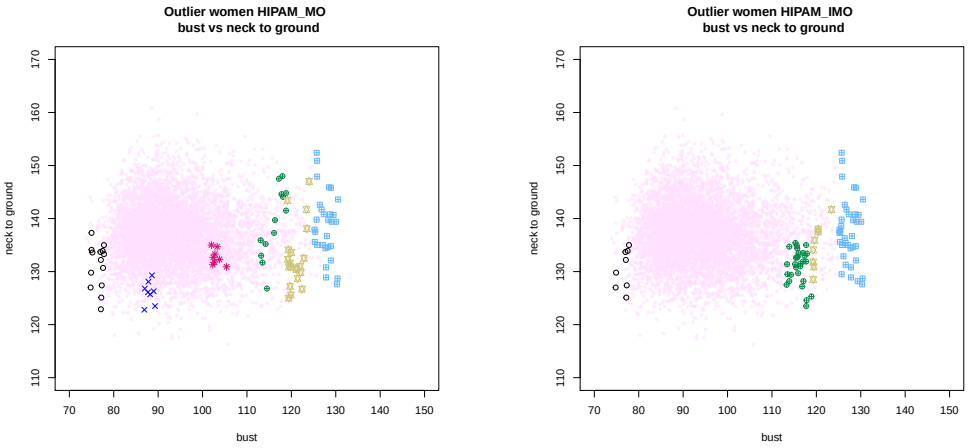


Figure 2.21: Bust vs neck to ground in the outliers obtained using $HIPAM_{MO}$ (left) and $HIPAM_{IMO}$ (right).

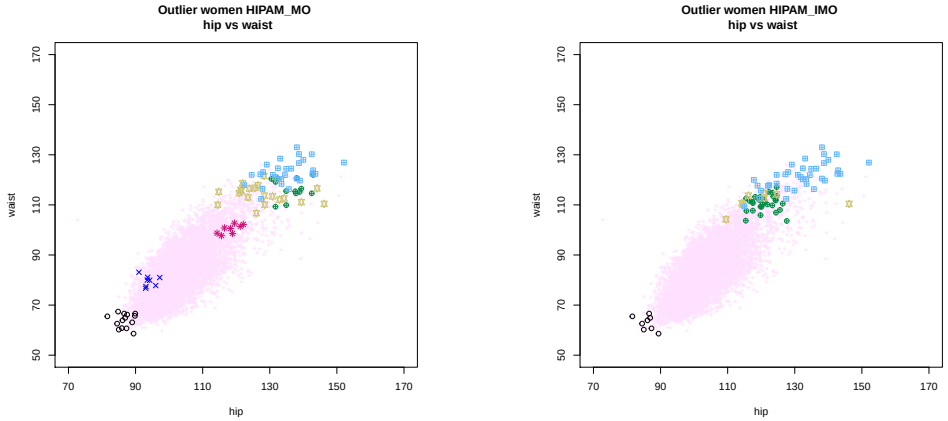


Figure 2.22: Hip vs waist in the outliers obtained using $HIPAM_{MO}$ (left) and $HIPAM_{IMO}$ (right).

Next, we want to compare the perfomance of our algorithm to identify outliers with two other approaches developed with the same goal: confidence ellipses and the **mvoutlier** R package [84].

Confidence ellipses is a quite common used method in the clothing literature to remove abnormal data (see for instance [96]). The usual procedure is

to define a confidence ellipse that covers a specified percentage of the data, assuming a bivariate normal distribution for height and weight.

Assuming bivariate normality, the portion of points falling inside the ellipse should closely agree with the fixed confidence level. Data falling outside the ellipse are considered as abnormal and are discarded. Using the function *dataEllipse* of the **car** R package [67], we have generated these types of ellipses at a 99% confidence level for each bust segment.

As an illustration, Fig. 2.23 identifies the outliers detected by the confidence ellipse for the bust segment $[78, 82[$ cm. Height is in mm. and weight in kg.

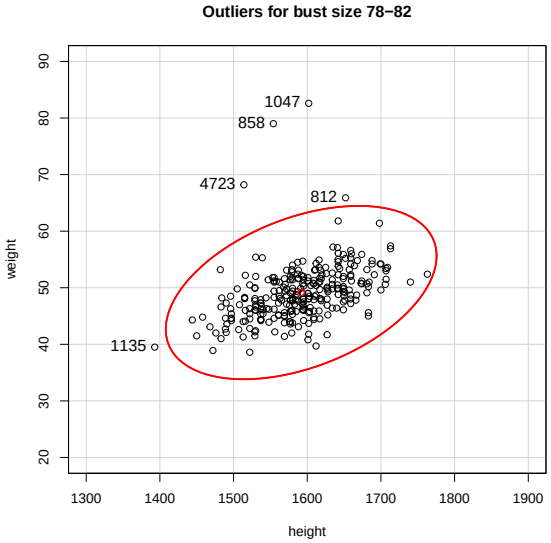


Figure 2.23: Outlier women returned by the normality ellipse for the bust segment $[78, 82[$ cm.

We would like to point out that by using this procedure, a woman is considered as outlier at the time she deviates substantially from the mean of only one of those variables. In addition, Fig. 2.24 represents the bust and hip measurements of the outliers detected by the confidence ellipses represented for every bust size (we consider again all the twelve bust classes of Table 2.5). To make this clear, we use the height and weight of the women belonging to each bust size, to represent the confidence ellipse with which identifying outliers and then, we examine some of their body measurements (such as bust

and hip). If we look at Fig. 2.24, we see oppossite results regarding both HIPAM algorithms. According to Fig. 2.24, no outliers have been identified in the most extreme sizes $[74, 78[$ cm, $[119, 125[$ cm and $[125, 131[$ cm, but there are a lot of women considered as outliers in the central sizes.

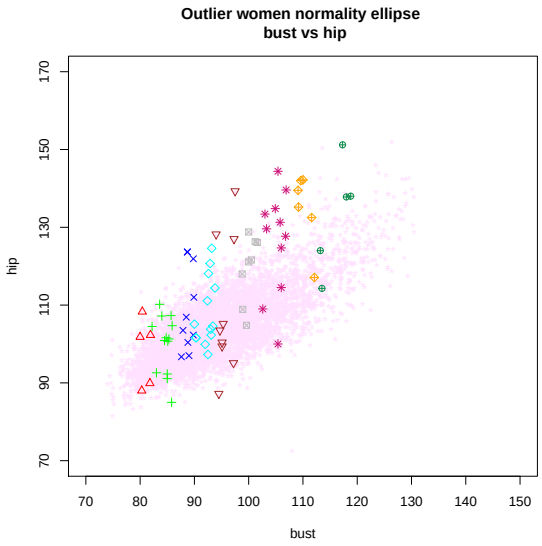


Figure 2.24: Bust vs hip of the outliers obtained using the normality ellipse within each bust segment.

However, at least for bust and hip measurements, they do not seem outliers because they are clearly located over the points cloud.

After finishing the previous comparison, a thorough survey of the R packages implemented to detect multivariate outliers was done. To the best of our knowledge, all those packages are based on robust methods. Among them, a well-known and commonly used R package is **mvoutlier** [84], which contains all the programs for the methods developed in [63]. In [63], Filzmoser et al. proposes to detect outliers by using the minimum covariance determinant (MCD) estimator to make the Mahalanobis distance robust. To properly compare **mvoutlier** with both HIPAM algorithms, we should replace the Mahalanobis distance within the corresponding functions of **mvoutlier** by the particular dissimilarity we are working with. However, given the implementation of this R package, this is not possible. This seems to mean that **mvoutlier** cannot be used for all types of data. Anyway, we have applied

mvoutlier for the twelve bust segments to analyze its results. As an illustration, Fig. 2.25 identifies the outliers detected for the bust segment $[78, 82[$ cm (marked with red circles).

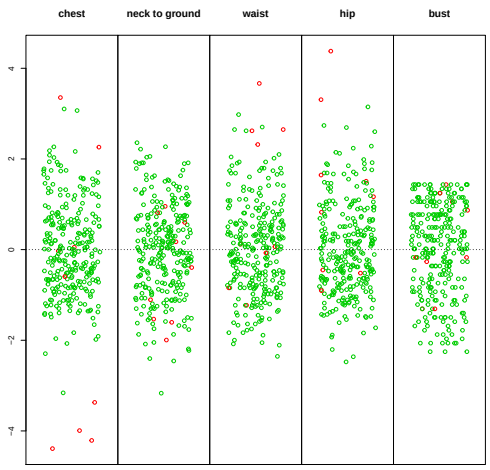


Figure 2.25: Outlier women returned by the **mvoutlier** R package for the segment $[78, 82[$ cm.

Fig. 2.26 joints together the outliers (bust vs hip) detected by **mvoutlier** for every bust size. Again, Fig. 2.26 shows that this R package is overestimating the number of outlier women.

Anthropometric analysis of the outliers

To close the study about outliers, we carry out an anthropometric analysis of the outlier women provided by the $HIPAM_{MO}$ and $HIPAM_{IMO}$ algorithms. Table 2.28 shows their answers to the question about whether they have had problems in finding their correct size.

Algorithm \ Answer	Never	Almost never	Yes, sometimes	Yes, always
$HIPAM_{MO}$	16	7	27	42
$HIPAM_{IMO}$	14	11	23	34

Table 2.28: Outlier women problems in finding their correct size.

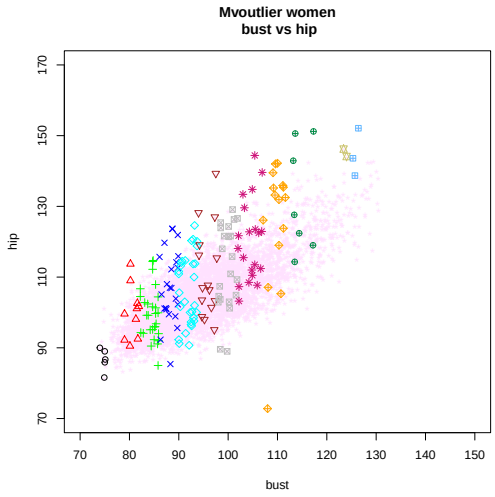


Figure 2.26: Bust vs hip in the outliers obtained using the **mvoutlier** R package.

Noteworthy is most of the outlier women answer that they sometimes or always have had problems in finding their correct size. To the knowledge of the experts, there might be several reasons to this:

- HYPOTHESIS 1: The most likely scenario is that they are overweight and therefore they fall outside the usual sizing range. This is a key health and sizing issue. Manufacturers usually offer the most common sizes because they are sent in greater quantities and consequently the unsold stock is minimized. Following this trend, extremely thin women will also have problems in finding their size, but these cases will be a minority since the increasing tendency is to be overweight.

- HYPOTHESIS 2: Another reasonable explanation could be that this kind of woman, despite not being overweight, presents a less frequent morphotype, with a narrow waist or wide hip. In this case, the women would find their clothes size, but items would not fit correctly and they would need to get the clothes altered.

- HYPOTHESIS 3: Finally, a third possibility would be that these women could fall between sizes, the difference usually being around 5 cm bust

or waist circumference depending on whether upper or lower body garments. In a real situation, when a woman does not fit one brand, it is probable that she will find her size in another store.

We want to evaluate if some of these hypotheses are met by the outliers returned by $HIPAM_{MO}$ and $HIPAM_{IMO}$:

- HYPOTHESIS 1: Regarding the 92 outliers obtained by $HIPAM_{MO}$, 70 of them are overweight ($BMI > 25$, BMI is the Body Mass Index), 12 are underweight ($BMI < 18.5$) and 10 are normal weight ($18.5 \leq BMI < 24.99$). Regarding the 82 outliers obtained by $HIPAM_{IMO}$, 74 are overweight, 7 are underweight and there is only one woman who is normal weight. Accordingly, the majority of the outliers are overweight, while underweight women are the minority. Indeed, the outlier population tends to be overweight.

- HYPOTHESIS 2: We consider those 10 outlier women with normal weight according to previous hypothesis (the only woman with normal weight returned by $HIPAM_{IMO}$ is one of the 10 obtained by $HIPAM_{MO}$). We analyze their body shape using the *drop value* [85, 60], which allows to identify different relationships between key anthropometric dimensions that determine body shape and morphotype. The drop value is defined as the difference between a woman's hip circumference and bust circumference. The population can be classified into the following categories (see for instance [85]):

  - Triangular or pear-shaped (bust much smaller than hip).
  - Inverted triangle (bust much bigger than hip).
  - Rectangular (bust equal to hip).
  - The other categories lie in between these.

  Table 2.29 details hip and bust perimeter and drop value of those ten outlier women that are normal weight. BARNA066, ELGOI023, MELGA066, CAMB037, CAMAR142 and SILLE148 can be considered pear shaped. The other women are rectangle shaped because they have very similar hip and bust measurements.

Woman Code	BARNA066	ELGOI023	MELGA066	CAMB037	ABAD071	MALAG131	CAMAR142	SILLE148	MADY165	MADY229
Hip measurement	89.4	87.5	96.0	94.2	93.6	91.0	93.1	97.2	93.5	93.0
Bust measurement	77.2	77.1	87.7	88.2	89.0	87.8	86.9	87.0	88.6	89.3
Drop value	12.2	10.4	8.3	6.0	4.6	3.2	6.2	10.2	4.9	3.7

Table 2.29: Drop values for the outlier women with a normal weight for $HIPAM_{MO}$.

- HYPOTHESIS 3: We analyze the four women with a rectangular shape (ABAD071, MALAG131, MADY165 and MADY229). Table 2.30 shows their bust and waist circumferences.

Woman Code	ABAD071	MALAG131	MADY165	MADY229
Bust measurement	89.0	87.8	88.6	89.3
Waist measurement	81.1	83.1	80.0	76.7

Table 2.30: Bust and waist measurements for the outlier women with a rectangular shape.

  By examining Table 2.30, ABAD071 and MADY229 might be considered women between two bust sizes (88 and 92) when comparing their bust measurement with the sizes detailed in Tables 2.5 and 2.8. For MALAG131 and MADY165 is not so easy to make a similar statement for their bust or waist.

2.5.2.2 Comparison with standard adopted method

Now, we would like to check the accuracy of our methodology in obtaining representative fit models. It is not straightforward to compare our approach to many others in the apparel sizing literature because their objective is the creation of a sizing system instead of looking for fit models. We think that, in our case, the most interesting and valuable study is to compare the main measurements (bust, hip and waist circumference) of the fit models obtained by $HIPAM_{MO}$ and $HIPAM_{IMO}$ with the corresponding standard measurements defined by the European standard. Part 3 [59], as we explain next. Tables 2.31 and 2.32 (resp. Tables 2.33 and 2.34) display the hip and waist measurements corresponding to the fit models obtained by $HIPAM_{MO}$ (resp. $HIPAM_{IMO}$) for the eight bust classes considered before. These tables also indicates the total number of fit models for each bust class (in parentheses and bold). These quantities appear in the column corresponding to the *expected* measurements according to the European standard. Part 3 [59].

Bust class \ Hip class	[86,90[	[90,94[	[94,98[	[98,102[	[102,106[	[106,110[	[110,115[	[115,120[	>120
[78, 82[	1 (**5**)	2	1	1	0	0	0	0	0
[82, 86[	0	1 (**3**)	0	2	0	0	0	0	0
[86, 90[	0	1	2 (**10**)	4	2	1	0	0	0
[90, 94[	0	0	1	1 (**5**)	2	1	0	0	0
[94, 98[	0	0	1	0	1 (**4**)	2	0	0	0
[98, 102[	0	0	0	1	1	0 (**4**)	1	1	0
[102, 107[	0	0	0	1	1	0	3 (**9**)	3	1
[107, 113[	0	0	0	0	0	1	0	0 (**2**)	1

Table 2.31: Hip measurements corresponding to the fit models obtained by $HIPAM_{MO}$. The total number of fit models for each bust segment is displayed in parentheses and bold in the column corresponding to the *expected* measurements.

Bust class \ Waist class	[62,66[	[66,70[	[70,74[	[74,78[	[78,82[	[82,86[	[86,91[	[91,97[	>97
[78, 82[	0 (**5**)	4	1	0	0	0	0	0	0
[82, 86[	0	1 (**3**)	1	1	0	0	0	0	0
[86, 90[	0	0	2 (**10**)	3	3	2	0	0	0
[90, 94[	0	0	0	1 (**5**)	0	4	0	0	0
[94, 98[	0	0	0	0	0 (**4**)	2	2	0	0
[98, 102[	0	0	0	0	0	0 (**4**)	3	1	0
[102, 107[	0	0	0	0	0	0	0 (**9**)	4	5
[107, 113[	0	0	0	0	0	0	0	0 (**2**)	2

Table 2.32: Waist measurements corresponding to the fit models obtained by $HIPAM_{MO}$. The total number of fit models for each bust segment is displayed in parentheses and bold in the column corresponding to the *expected* measurements.

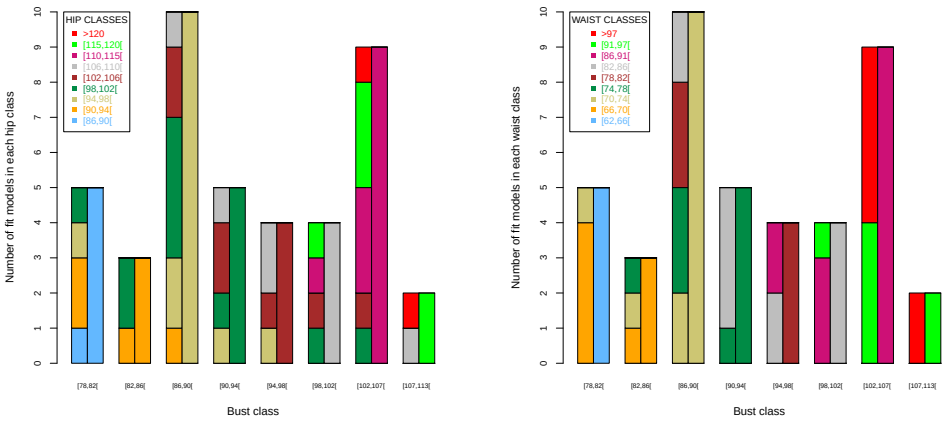
Bust class \ Hip class	[86,90[	[90,94[	[94,98[	[98,102[	[102,106[	[106,110[	[110,115[	[115,120[	>120
[78, 82[	0 (**3**)	2	0	1	0	0	0	0	0
[82, 86[	0	0 (**3**)	1	2	0	0	0	0	0
[86, 90[	0	0	0 (**3**)	2	1	0	0	0	0
[90, 94[	0	0	0	1 (**3**)	1	1	0	0	0
[94, 98[	0	0	1	0	1 (**3**)	0	1	0	0
[98, 102[	0	0	0	1	0	1 (**3**)	1	0	0
[102, 107[	0	0	0	0	1	0	1 (**3**)	1	0
[107, 113[	0	0	0	0	0	1	1	2 (**5**)	1

Table 2.33: Hip measurements corresponding to the fit models obtained by $HIPAM_{IMO}$. The total number of fit models for each bust segment is displayed in parentheses and bold, in the column corresponding to the *expected* measurements.

Bust class \ Waist class	[62,66[	[66,70[	[70,74[	[74,78[	[78,82[	[82,86[	[86,91[	[91,97[	>97
[78, 82[	0 (**3**)	2	1	0	0	0	0	0	0
[82, 86[	0	1 (**3**)	1	1	0	0	0	0	0
[86, 90[	0	0	1 (**3**)	1	1	0	0	0	0
[90, 94[	0	0	0	0 (**3**)	1	2	0	0	0
[94, 98[	0	0	0	0	0 (**3**)	1	2	0	0
[98, 102[	0	0	0	0	0	0 (**3**)	2	1	0
[102, 107[	0	0	0	0	0	0	0 (**3**)	2	1
[107, 113[	0	0	0	0	0	0	0	0 (**5**)	5

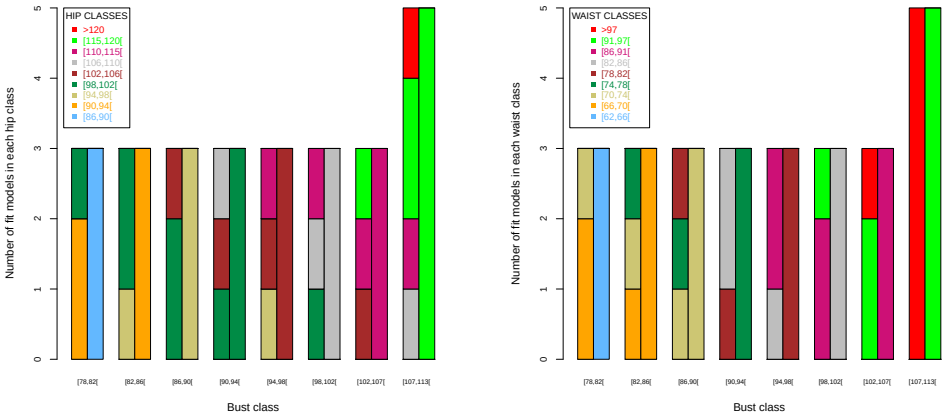
Table 2.34: Waist measurements corresponding to the fit models obtained by $HIPAM_{IMO}$. The total number of fit models for each bust segment is displayed in parentheses and bold, in the column corresponding to the *expected* measurements.

Fig. 2.27 (resp. Fig. 2.28) displays the results of $HIPAM_{MO}$ (resp. $HIPAM_{IMO}$) as bar charts for hip and waist.



(a) (b)

Figure 2.27: For each bust class, the left bar of (a) (resp. (b)) refers to the hip (resp. waist) of the fit models obtained by $HIPAM_{MO}$ while the right bar refers to the *expected* hip (resp. waist).



(a) (b)

Figure 2.28: For each bust class, the left bar of (a) (resp. (b)) refers to the hip (resp. waist) of the fit models obtained by $HIPAM_{IMO}$ while the right bar refers to the *expected* hip (resp. waist).

When analyzing these numerical and graphical displays, we reach the same conclusions for both $HIPAM_{MO}$ and $HIPAM_{IMO}$. First, we see that for all bust sizes, our fit models have a greater waist than expected. For hip this feature is not so clear, but it does also tend to be greater. Furthermore, an even more important aspect is that for some bust sizes there is no fit model with the expected hip or waist size, see e.g. bust size $[98, 102[$ cm for the results of $HIPAM_{MO}$ or bust size $[78, 82[$ cm for the results of $HIPAM_{IMO}$.

This analysis may suggest that the European standard. Part 3 [59] is not accurately representing the real measurements of the current female population (at least for Spanish women).

2.5.2.3 Obtaining medoids for different regions

As was said before, as a second approach to apply this methodology, we propose to segment the population based on geographical location. Geographic segmentation is common to apparel industries aimed at exploring national and regional influences on size and fit and defining guidelines. Clothing manufacturers are interested to know the variations in body size and shape (between and within national populations) to improve sizing and to offer

consistent labelling.

For our particular study, we analyze two Spanish regions, one located in the south and the other in the north. We only consider the bust size $[90, 94[$ cm because it contains the largest number of women for both regions, 166 women from the southern group and 64 from the northern group. See Table 2.35 for summary statistics of both groups.

Summary statistics for the southern group

Measurement (cm)	Minimum	First Quantile	Median	Mean	Third Quantile	Maximum
Neck to ground length	118.2	134.5	137.4	137.6	140.9	155.8
Bust circumference	90.00	90.90	91.70	91.85	92.80	93.90
Chest circumference	89.63	93.20	94.75	94.83	96.10	101.63
Waist circumference	68.90	78.92	81.55	81.98	84.67	95.00
Hip circumference	91.60	99.78	103.05	103.38	106.75	117.20

Summary statistics for the northern group

Measurement (cm)	Minimum	First Quantile	Median	Mean	Third Quantile	Maximum
Neck to ground length	123.6	131.8	136.0	135.8	139.4	152.4
Bust circumference	90.00	91.20	92.45	92.15	93.12	93.90
Chest circumference	85.47	94.31	95.32	95.58	97.17	101.39
Waist circumference	71.40	78.38	81.65	81.33	83.72	92.20
Hip circumference	88.60	98.12	102.20	102.42	106.12	117.60

Table 2.35: Summary statistics of the groups located in the south and north of Spain.

Table 2.36 (resp. Table 2.37) describes the identification codes and main measurements of the medoids returned by $HIPAM_{MO}$ (resp. $HIPAM_{IMO}$).

Code	Chest	Neck to ground	Waist	Hip	Bust
ANTAS052	93.8267	139.0	81.1	103.3	92.0
CHIPI018	96.0296	137.5	81.2	101.1	91.6
BILB085	97.4198	131.8	86.6	103.9	93
ELGOI111	94.3988	141.9	81.0	108.0	91

Table 2.36: Measurements of the medoids from the southern group (first and second row) and the northern group (third and fourth row), obtained with $HIPAM_{MO}$.

Fig. 2.29 (resp. Fig. 2.30) shows the 3D representation of the medoids of $HIPAM_{MO}$ (resp. $HIPAM_{IMO}$). These medoids have been identified using the five anthropometric variables considered in the analysis.

Code	Chest	Neck to ground	Waist	Hip	Bust
ABAD024	94.0821	139.1	80.1	101.7	92.4
PRTOS159	95.0924	140.3	79.0	108.2	91.5
MALAG004	93.9341	131.5	86.0	101.8	91.1
ELGOI020	94.4474	134.7	73.1	96.3	90.5
ERNAD110	97.1320	139.6	83.8	108.8	93.4
BILB132	94.3286	128.7	87.3	101.2	93.3

Table 2.37: Measurements of the medoids from the southern group (first to third row) and the northern group (fourth to sixth row), obtained with $HIPAM_{IMO}$.

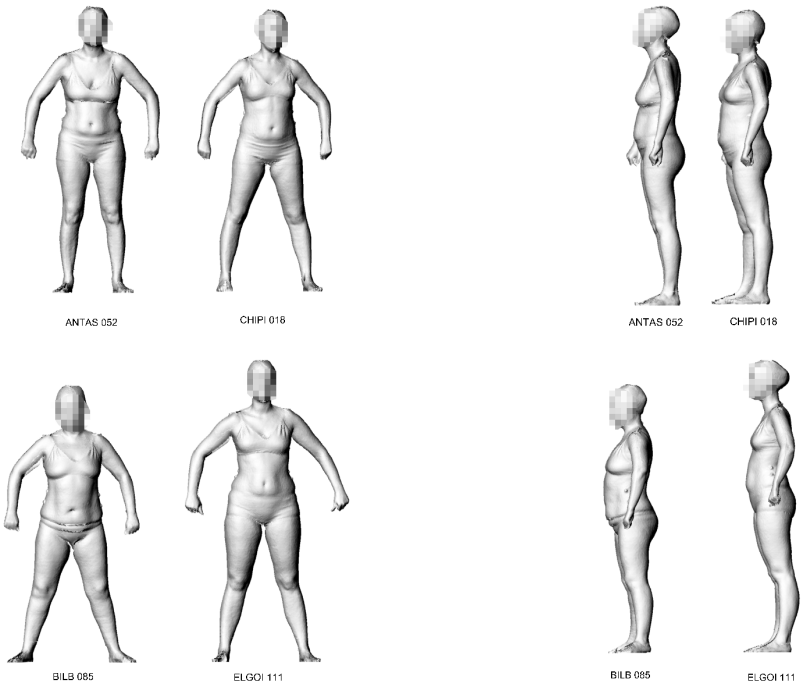


Figure 2.29: Front and lateral 3D representations of the medoids obtained with $HIPAM_{MO}$ from southern group (first row) and the northern group (second row).

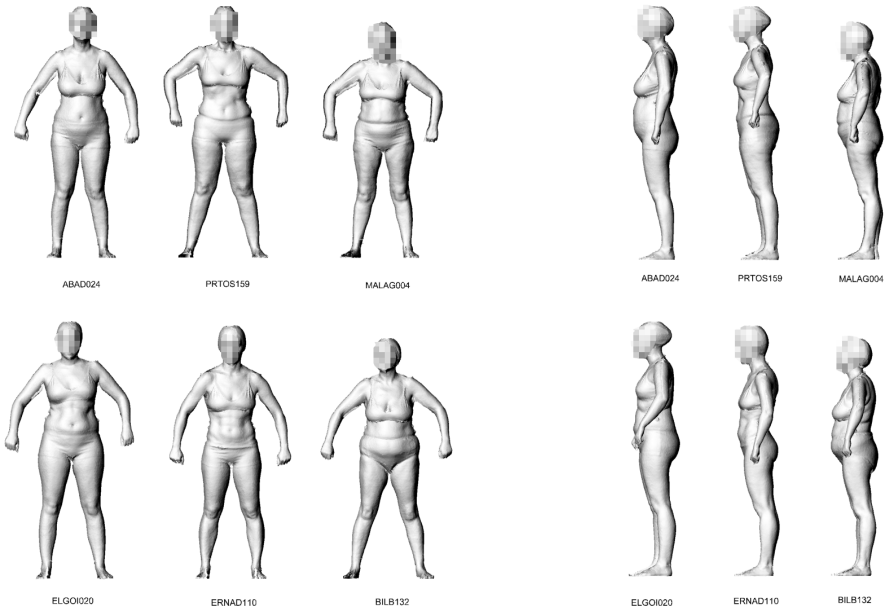


Figure 2.30: Front and lateral 3D representations of the medoids obtained with $HIPAM_{IMO}$ from the southern group (first row) and northern group (second row).

$HIPAM_{MO}$ finds two medoids while $HIPAM_{IMO}$ obtains three for the two regions. Let us examine their features for the southern group. Two of the $HIPAM_{MO}$ medoids, ANTAS052 and CHIPI018, are very similar, see Table 2.36. However, as it can be seen in Table 2.35, there is relevant anthropometric variability for this region; for instance, the range for the neck to ground variable is 37 cm. In contrast, the three $HIPAM_{IMO}$ medoids present high anthropometric differences. Neck to ground differs by 8.8 cm between PRTOS159 and MALAG004 and hip circumference differs by 6.5 cm between ABAD024 and PRTOS159, see Table 2.37.

For the northern group the two $HIPAM_{MO}$ medoids are more diverse, see Table 2.36. In this case, the neck to ground variable differs by 10.1 cm and waist differs by 5.6 cm. Even so, $HIPAM_{IMO}$ shows even more anthropometric dissimilarity between medoids, see Table 2.37. ELGOI020 and ERNAD110 has a difference in waist circumference of 10.7 cm and neck to ground between BILB132 and ERNAD110 differs by 10.9 cm. This brief analysis for these two particular Spanish regions shows that the $HIPAM_{IMO}$ algorithm performs better at finding representative models for apparel sizing

and design. The $HIPAM_{IMO}$ medoids represent a subset of the population with anthropometric diversity and therefore with different fitting requirements.

2.5.2.4 HIPAM as an alternative to current sizing system Standard

Finally, we would like to study the possibility of using a HIPAM algorithm to define a clothing sizing system, as an alternative to current sizing system standards. We apply both $HIPAM_{MO}$ and $HIPAM_{IMO}$ to the whole database we use in this methodology, without presegmenting into standard sizes. $HIPAM_{MO}$ returned ten clusters, while $HIPAM_{IMO}$ returned three, so we will analyze only $HIPAM_{MO}$ groups in order to see if they can be considered a good approximation of the sizes defined by the European standard. Part 3 [59]. Fig. 2.31 shows the classification tree generated by $HIPAM_{MO}$ and Table 2.38 details the key anthropometric measurements of the cluster medoids ordered by bust circumference in an increasing order.

Table 2.39 shows the basic descriptives for bust dimension for those 10 clusters, while Table 2.40 shows the same descriptives but for the 12 sizes defined by the European standard. Part 3 [59].

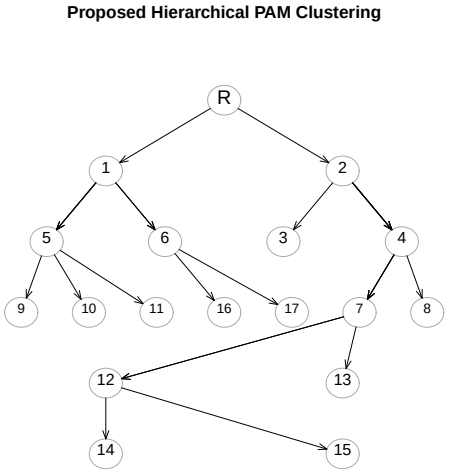


Figure 2.31: Clustering results returned by $HIPAM_{MO}$ applied to the whole database.

Label medoid	chest	neck to ground	waist	hip	bust	cluster label
869	84.55	133.20	66.70	89.30	79.50	13
150	86.53	135.10	69.10	94.60	82.30	14
232	88.42	134.90	73.20	96.00	84.60	15
302	91.36	134.80	73.10	97.40	86.30	8
5503	98.68	140.90	86.50	105.70	94.00	3
1120	103.24	130.90	90.40	104.10	99.60	9
1506	102.35	141.60	88.70	114.20	100.00	10
3988	109.26	136.30	100.30	110.20	106.30	11
431	113.83	131.00	105.10	117.40	110.50	16
501	120.86	141.20	112.30	128.30	115.50	17

Table 2.38: Cluster medoids returned by $HIPAM_{MO}$ applied to the whole database, ordered by bust circumference value.

	13	14	15	8	3	9	10	11	16	17
Min.	74.00	74.90	80.00	79.10	81.00	92.50	88.50	97.70	97.40	103.30
1st Qu.	79.20	82.90	84.70	87.80	91.90	99.00	97.93	104.10	110.00	115.60
Median	80.80	84.60	86.50	89.80	94.50	100.40	99.80	106.40	113.60	120.40
Mean	80.96	84.51	86.32	89.38	93.84	100.50	99.60	106.70	113.50	119.90
3rd Qu.	82.98	86.50	88.20	91.40	96.30	102.20	102.00	109.00	117.70	125.80
Max.	86.60	89.60	90.60	95.10	108.00	108.50	106.70	117.00	126.10	130.50
Number women	350	574	396	1106	1389	510	350	575	366	111

Table 2.39: Basic descriptives for bust dimension for each one of the ten clusters returned by $HIPAM_{MO}$ applied to the whole database.

	[74,78[	[78,82[	[82,86[	[86,90[	[90,94[	[94,98[	[98,102[	[102,107[	[107,113[	[113,119[	[119,125[	[125,131[
Min	74.00	78.00	82.00	86.00	90.00	94.00	98.00	102.00	107.00	113.00	119.00	125.20
1st Qu.	76.30	79.55	83.30	87.00	90.90	94.90	98.70	103.00	108.30	113.80	119.80	125.80
Median	77.00	80.50	84.20	88.00	91.80	95.80	99.70	104.20	109.60	115.40	121.20	127.30
Mean	76.64	80.39	84.13	87.97	91.89	95.88	99.74	104.20	109.70	115.60	121.40	127.40
3rd Qu.	77.60	81.30	85.10	89.00	92.80	96.90	100.80	105.40	111.10	117.00	122.70	128.90
Max.	77.90	81.90	85.90	89.90	93.90	97.90	101.90	106.90	112.90	118.90	124.90	130.50
Number women	47	287	732	1028	952	818	633	547	356	203	87	37

Table 2.40: Basic descriptives for bust dimension for the first twelve bust sizes defined by the European standard. Part 3 [59].

When comparing both tables, we realize that the 10 clusters obtained by $HIPAM_{MO}$ are somewhat different with respect to the sizes defined by the current standard. But we also appreciate that those 10 clusters are quite different to each other, except for clusters labelled as 9 and 10, which are very similar. For these specified two clusters, we see that the bust measurement of the medoid in cluster 10 is greater than the corresponding one for

the medoid in cluster 9 (see Table 2.38), but all the descriptives in cluster 9 are greater than those in cluster 10, (see in this case Table 2.39). As a conclusion, we have seen that $HIPAM_{MO}$ applied to the whole database, does not approximate well to the sizes defined by the current European standard. Part 3 [59]. However, we have proved that any HIPAM algorithm does serve to segment a huge database in different groups according to the most relevant anthropometric dimensions.

2.5.3 Summary

We have adapted the HIPAM algorithm to deal with anthropometric data aimed at identifying representative fit models. We have proposed two HIPAM algorithms, $HIPAM_{MO}$ and $HIPAM_{IMO}$, where the main differences regarding the original one are the dissimilarity used and the criterion used to divide clusters. In our first analysis we have followed the same procedure as in Section 2.3: we have segmented the data set into twelve standard bust sizes according to European standard. Part 3 [59] and we have applied $HIPAM_{MO}$ and $HIPAM_{IMO}$ to each one of them. According to the analysis of the medoids obtained by both algorithms, we conclude that $HIPAM_{IMO}$ provides a more cost-effective performance taking into account the fitting tolerances of clothing. In addition, we also observed that $HIPAM_{IMO}$ only identifies outliers in the four bust classes corresponding to small and large sizes (extreme sizes). This is quite well aligned to the clothing industry practice for the mass production of clothing where the objective is to optimize sizes by addressing only the most profitable. Extreme sizes are usually offered as "special sizes". In this way, we could state that $HIPAM_{IMO}$ shows better performance. However, it is also true that both algorithms provide consistent results. They identify more realistic fit models, especially when comparing the results with the European standard. Part 3 [59], and also detect outliers. These fit models could be shown to experts in a practical situation to help them in their task.

As a complementary analysis of our methodology we have compared its performance to detect multivariate outliers with two common used methods. We reached the conclusion that these two common approaches take the risk to overestimate the number of outliers. The HIPAM algorithm is based on hierarchical features to discover outliers and return true outliers. In addition, the exploratory anthropometric analysis of the outlier women obtained by both HIPAM algorithms has helped to confirm that those women who state

that they always have problems in finding their size verify the hypotheses suggested by the experts to try to explain the reasons for their problems.

We think that this methodology based on HIPAM is a useful tool to help fashion designers and apparel manufacturers to hire accurate and representive fit models. They will be used to test the size specifications of their clothes before the production phase with a consequent improvement of garment fit. A good fit model is the basis for an accurate sizing system.

2.6 Chapter conclusions

In this chapter we have presented three methodologies based exclusively on clustering procedures.

The first one, *trimowa*, has been developed aimed at defining an efficient sizing system. To achieve this, a percentage of women with extreme anthropometric measurements has been removed. It was found that the segmentation proposed by this first method is an improvement with respect to the approaches previously published in the literature. Each one of the bust size groups described by the European standard has been divided into three groups using a pure clustering procedure. This way of proceeding allows that each woman belongs to a bust class and within it, she is part of any of the obtained three groups, which are defined for specific anthropometric measurements of chest, hip, waist and neck to ground. Thus, with such a sizing system, a woman wishing to buy an upper body garment, should first selects those garments whose bust label coincide with her bust measurement. Finally, she should choose the garment whose label indicating the measurements of the other four dimensions were as approximately as possible to her measurements. This kind of scenario could facilitate women to find the garment that fit them correctly in less time, implying greater satisfaction in the buying process. Commercially, this situation would mean that clothing stores would not have so many clothes without selling, which also includes returned clothes. The *trimowa* methodology obtains representative subjects for each group which are real women of the data set. As a result, their measurements should be helpful for labelling the garments corresponding to each size group.

The second methodology, *biclustAnthropom*, has also been proposed to define an efficient sizing system, but in this case for lower body garments. The clustering algorithm on which it is based, is called biclustering and has

been mainly used with gene expression data. Its main feature is that groups are described by a non previously determined number of variables. This number does not have to coincide with the total number of variables used in the study. This fact has a particular interest in the definition of efficient sizes since each group may be different from the others according to the set of dimensions that belong to it, which furthermore are the most relevant to define this group. On the other hand, it has also investigated the possibility of using other biclustering method for grouping people according to their eating habits, following a reference of the literature on market segmentation (see Appendix). Both approaches using biclustering have had an exploratory and descriptive nature. In both cases, original and profitable results have been obtained. For future work, from the point of view of the definition of an optimal sizing system, we aim at developing a biclustering method that incorporates a specific distance to deal with anthropometric data. From the point of view of market segmentation, we think that the type of biclustering method used could be a convenient alternative for researchers related to the field of Sociology.

At last, the third methodology, *hipamAnthropom*, has been raised in order to identify representative fit models of the population. As far as we know, no methodology has been proposed in the literature for the same purpose. With this proposal, we wanted to open a new line of research related not only to the definition of an efficient sizing system, but also to the identification of people who according to their morphology, should be considered as the best fit models to develop clothes for the population that they represent. Specifically, fit models represent the body dimensions which a company designer has determined to provide the proportional relationships needed to achieve the company fit. However, these models may not be so representative of the target population as might be thought at first glance. This third approach tries to statistically identify a target set of individuals who best represent the user population. With them, fit, comfort and visual appearance of the manufactured clothes could improve. The fashion designer can visually check the fit of a design on the fit model, effectively acting as a live mannequin.

Chapter 3

Statistical shape analysis

3.1 Introduction [1]

The statistical shape analysis (also called geometric morphometrics or simply Morphometrics) concerns with the statistical study of the variation and covariation of the shape of objects. The seminal paper on this field was [116]. The word "shape" is frequently used in our day by day to refer the appearance of an object. We often describe unknown shapes by using known shapes, e.g. "Italy has the shape of a boot". Mathematically, shape is defined as the geometrical information of the object that remains when location, scale and rotational (orientation) effects are removed from that object [116]. There are three major approaches to describe an object's shape [196]:

1. Objects can be treated as subsets of $\mathbb{R}^m$ (figures). The basic features to describe shapes in $\mathbb{R}^2$ are area and perimeter. In $\mathbb{R}^3$ we could use volume, for example.

2. They can be described by using functions representing their contours.

3. They are described by using a finite number of points, called landmarks, that are given by certain geometrical or anatomical properties.

In this chapter we focus on the third approach. A landmark is a point of correspondence on each object that matches between and within populations. The configuration is the set of landmarks on an object. In the beginning,

[1] The methodology presented in this chapter is submitted for publication [216].

shape was considered an attribute of the structure of the object, which could be described by using distances, ratios or angles between landmarks, rather than landmarks themselves. Traditional methods are based on multivariate analysis of these measurements [145]. However, with these types of methods, part of the geometric relationships among the measured variables is omitted, so shape cannot be quantitatively analyzed. It was in the 20th century that the statistical shape analysis was greatly expanded with the development of new statistical and numerical methods. These modern techniques are able to deal directly with the entire geometric information contained in the configuration, describing shape variation in both qualitative and quantitative terms.

A basic technique for comparing shapes and quantifying shape differences is the Procrustes superimposition approach or Procrustes method. Nowadays, it is the most widely used method in geometric morphometrics. It consists in superimposing configurations by translating, scaling and rotating them, in such a way that the distance among them (a Procrustes-type distance) is as small as possible. Once they are superimposed, the differences in the positions of the landmarks can be easily observed. Fig. 3.1 shows an illustrative example of the performance of the Procrustes superimposition (based on [202]). A fundamental operation is the computation of the average configuration of the set of objects. It is called the Procrustes mean and it is defined as the shape whose sum of squared Procrustes distances to the other objects is minimal [174].

Clustering of objects according to their shape information is an important task related to the statistical shape analysis with direct implications in many scientific areas, such as Biology, Archaeology, Medicine and, in recent decades, Computer Vision and Pattern Recognition. An unsupervised detection of elements with similar shape is usually required to facilitate the analysis of the entire collection of observations.

In this chapter, we present the *kmeansProcrustes* clustering methodology. We have used the **shapes** R package [48], which contains the main routines for the statistical analysis of shapes. Other interesting R packages related to Morphometrics are **geomorph** [2] and **Morpho** [182]. Our work is based on [47]. Refs. [35, 174, 191] have been also very helpful.

The outline of this chapter is as follows: Section 3.2 introduces all the needed theoretical background. Section 3.3 presents the data and procedure used, the experimental results and a comprehensive summary of *kmeansProcrustes*. Finally, the conclusions of this chapter and future work are discussed

in Section 3.4.

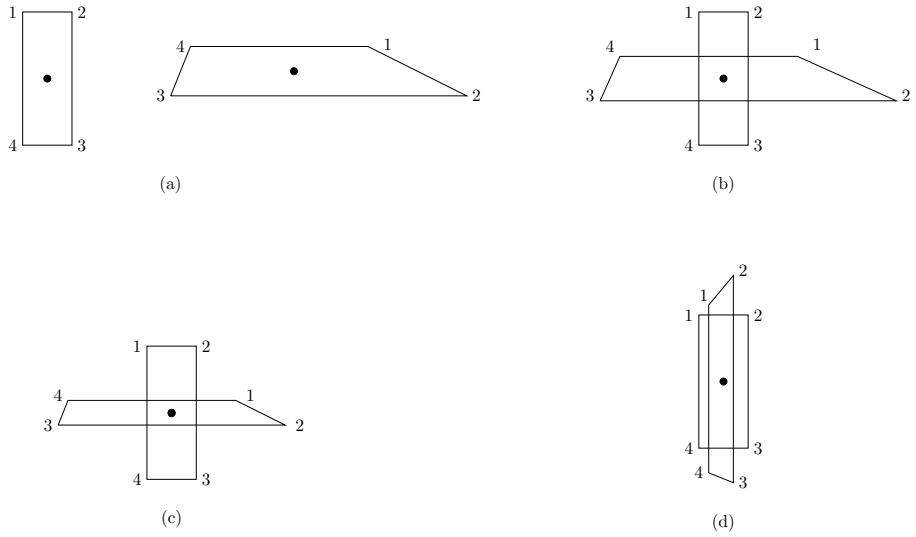


Figure 3.1: Procrustes superimposition following [202]: (a) Original position, (b) centered figures, (c) scaled figures, (d) rotated figures.

3.2 Background

This introductory section describes the basic concepts of the statistical shape analysis and explains how to adapt the k-means algorithm for use in shape space.

3.2.1 Shape spaces and Procrustes superimposition

The following general notation will be used: n refers to the number of objects, h to the number of landmarks and m to the number of dimensions (in our case, $m = 3$). Then, each object is described by an $h \times m$ configuration matrix X containing the m Cartesian coordinates of its h landmarks. X has $h \times m$ dimensions. Ref. [192] is a good consultation material to introduce the most important notions related to shape analysis.

Definition 1 *An $m \times m$ rotation matrix Γ satisfies $\Gamma^T\Gamma = \Gamma\Gamma^T = I_m$ and $|\Gamma| = \pm 1$, where I_m is the $m \times m$ identity matrix. The set of all $m \times m$ rotation matrices is known as the special orthogonal group of rotations $SO(m)$.*

Definition 2 *The Euclidean similarity transformations are the set of translations, scaling and rotations procedures that are applied to a configuration X:*

$$\{\beta X\Gamma + 1_h\gamma^T : \beta \in \mathbb{R}^+, \Gamma \in SO(m), \gamma \in \mathbb{R}^m\} \tag{3.1}$$

where $\beta \in \mathbb{R}^+$ is the scale parameter, Γ is an $(m \times m)$ rotation matrix, γ is an $(m \times 1)$ location vector and 1_h is a column vector of h ones.

Definition 3 *A size measure is a positive, real-valued function $g(X)$ satisfying:*

$$g(aX) = ag(X) \tag{3.2}$$

where a is some magnification positive factor.

The most commonly used measure of size for a configuration is the centroid size.

Definition 4 *The centroid size is defined as:*

$$S(X) = \sqrt{\sum_{i=1}^{h}\sum_{j=1}^{m}(X_{ij} - \bar{X}_j)^2} \tag{3.3}$$

where X_{ij} is the (i,j)th element of X and $\bar{X}_j$ is the arithmetic mean for the jth dimension, $\bar{X}_j = \dfrac{1}{h}\displaystyle\sum_{i=1}^{h} X_{ij}$.

Definition 5 *The centroid coordinates of the configuration X are the arithmetic mean for each dimension $j = 1, \ldots, m$:*

$$CC_X = (\bar{X}_1, \ldots, \bar{X}_m) \tag{3.4}$$

In our case of $m = 3$, CC_X is a vector with three elements.

A possibility to remove the location effect from X would consist of subtracting the centroid coordinates of the configuration. In this way, the coordinates of the centroid define the translation parameter γ. This yields the translated (or centered) configuration, X_C, in such a way that the centroid is sent to the origin: $X_C = X - 1_h CC_X$. Then, the centroid coordinates of X_C are equal to zero, $CC_{X_C} = (0, 0, 0)$ and its centroid size is the same as the original one: $S(X) = S(X_C)$.

Another approach to get the translated configuration is premultiplying X by the centering matrix C, $X_C = CX$. The matrix C has a diagonal equal to $1-1/h$ and lower and upper triangle cells equal to $-1/h$, i.e. $C = I_h - \frac{1}{h}1_h 1_h^T$, being I_h the $h \times h$ identity matrix. C satisfies $C^T = C$ because it is symmetric and $C^T C = CC = C$ because it is idempotent.

Definition 6 *The centroid size of X_C is expressed as follows:*

$$\begin{aligned} S(X_C) = \|CX\| &= \sqrt{trace((CX)^T CX)} \\ &= \sqrt{trace(X^T C^T CX)} \\ &= \sqrt{trace(X^T CX)} \end{aligned}$$

being $\|X\| = \sqrt{trace(X^T X)}$ the Euclidean norm.

The problem of this alternative is that the centered configuration X_C is a $h \times m$ matrix with a range of $h - 1$. Instead, it is mathematically more convenient to work with $X_H = HX$, where H is the Helmert sub-matrix, a $(h-1) \times h$ orthogonal matrix that satisfies $H^T H = C$ and consequently $H^T X_H = H^T HX = CX$. Next, Helmert matrices will be fully defined:

Definition 7 *The full Helmert matrix H^F is a square $h \times h$ orthogonal matrix with its first row of elements equal to $1/\sqrt{h}$ and the remaining rows are orthogonal to the first row. Specifically, the jth row of H^F is given by:*

$$(\underbrace{h_j, \ldots, h_j}_{j-1}, -(j-1)h_j, \underbrace{0, \ldots, 0}_{h-j}), \quad hj = -1/\sqrt{j(j-1)} \tag{3.5}$$

Then, the Helmert sub-matrix, H, is a Helmert matrix with its first row removed.

Any centered configuration has $hm - m$ dimensions. Removing the size effect (scale) can be done by dividing the coordinates of the configuration by the centroid size. For the centered configuration, X_H, this operation has the following expression:

$$Z = \frac{X_H}{S(X_H)} = \frac{HX}{\|HX\|} = \frac{HX}{\sqrt{trace((HX)^T HX)}} \tag{3.6}$$

Z is called the pre-shape of the configuration matrix X because all information about location and scale are removed, but rotation information remains. This terminology was first introduced in [118]. Z has unit centroid size, i.e. $S(Z) = 1$.

Definition 8 *The pre-shape of X is all the geometrical information of X invariant to location and scale:*

$$Z = \{\beta X + 1_h \gamma^T : \beta \in \mathbb{R}^+, \gamma \in \mathbb{R}^m\} \tag{3.7}$$

Definition 9 *The pre-shape space S_m^h is the set of all possible pre-shapes. The dimension of S_m^h is $m(h-1) - 1$.*

In order to remove the rotation effect, all the rotated versions of the pre-shape with each other must be identified.

Definition 10 *The shape of X is all the geometrical information of X invariant under the Euclidean similarity transformations:*

$$[X] = \{Z\Gamma : \Gamma \in SO(m)\} \tag{3.8}$$

Definition 11 *The shape space Σ_m^h (named Kendall shape space) is the set of all possible shapes, which are represented by a single point in this space. The dimension of Σ_m^h is $hm - m - 1 - \frac{m(m-1)}{2}$.*

Fig. 3.2 shows a summary of the hierarchy of the different spaces explained in this chapter, following [47].

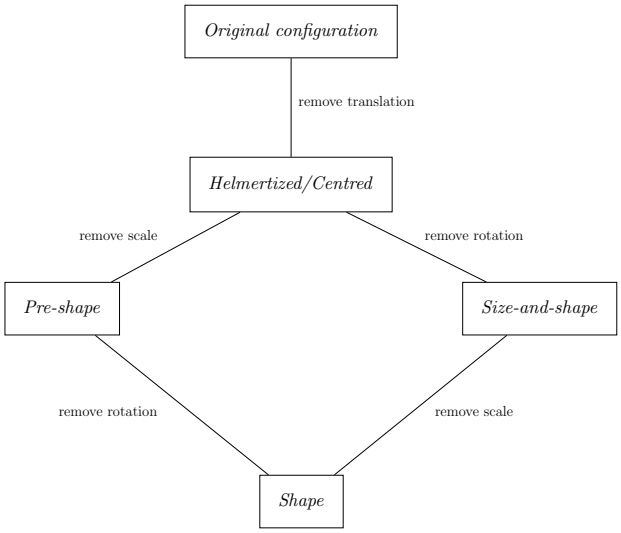


Figure 3.2: The hierarchy of shape spaces explained in this chapter (following [47]).

The pre-shape and shape spaces are not Euclidean spaces because there are less dimensions than geometric coordinates. In order to completely define a non-Euclidean shape space, a concept of distance between shapes is needed. The full Procrustes distance is one of the best known distances. It is a measure of shape difference between two configurations. Given X_1 and X_2 and their corresponding pre-shapes Z_1 and Z_2, the full Procrustes distance is a least-squares type metric where we minimize over rotations and scale to find the nearest Euclidean distance between Z_1 and Z_2.

Definition 12 *The full Procrustes distance between X_1 and X_2 is:*

$$d_F(X_1, X_2) = \inf_{R \in SO(m), \beta \in \mathbb{R}^+} \|Z_2 - \beta Z_1 R\|, \tag{3.9}$$

Unlike the shape space Σ_m^h, which it is not a familiar space when $m > 2$, the pre-shape space S_m^h is a hypersphere of unit radius in $(h-1)m$ dimensions. Therefore, we can analyze other distance functions that measure how far two points on a sphere are, such as the great circle distance. The great circle

distance is a particular distance which it is commonly used in the shape space as an alternative to d_F.

In S_m^h, all possible rotations are organized along an orbit called a fiber [35, 47]. S_m^h is partitioned into fibers by the rotation group $SO(m)$. Two pre-shapes on a given fiber differ by a rotation [82]. A fiber in S_m^h corresponds to a shape in Σ_m^h, that is to say, the shapes of the configurations are represented by fibres on S_m^h. Finding the rotation parameters to superimpose X_1 on X_2 is equivalent to finding the shortest distance between both fibers in S_m^h. Therefore, the distance between two shapes can be defined in the following way:

Definition 13 *The Procrustes distance $\rho(X_1, X_2)$ is the closest (over rotations) great circle distance between Z_1 and Z_2 on the pre-shape hypersphere S_h^m.*

Fig. 3.3 is the same figure as Fig. 4.5 of Ref. [35] which is represented to illustrate the geometric meaning of the pre-shape space and the related Procrustes-type distances introduced in this chapter. The relationship between d_F and ρ is:

$$d_F(X_1, X_2) = \sin \rho. \tag{3.10}$$

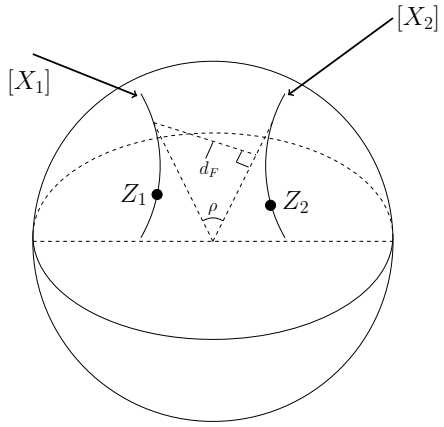


Figure 3.3: Illustration of the pre-shape space according to [35] and [47]. Z_1 and Z_2 represent the pre-shapes on their fibers $[X_1]$ and $[X_2]$. Only the two distances explained in this chapter are indicated. ρ corresponds to the smallest angle between Z_1 and Z_2, while d_F is the shortest distance between Z_1 and the radius of the fiber $[X_2]$ to which Z_2 belongs.

Consider now the case where a set of configuration matrices $X_1, \dots X_n$, $n \geq m$, are available. The concept of mean shape can be introduced. Because we are not working with Euclidean spaces, there is not a single concept of mean that corresponds to the arithmetic average of realizations. Otherwise, we need to use a Fréchet type mean [68], i.e. one that minimizes the sum of squared full Procrustes distances from any shape in the set.

Definition 14 *Given a set of configuration matrices $X_1, \dots, X_n$, the full Procrustes mean is given by $[\widehat{\mu}]$, where:*

$$[\widehat{\mu}] = \arg \inf_{\mu: S(\mu)=1} \sum_{l=1}^{n} d_F^2(X_l, \mu). \tag{3.11}$$

Definition 15 *The full Generalized Procrustes Analysis (full GPA) method involves matching the configurations, $X_1, X_2, \dots, X_n$, by estimating the similarity parameters γ, Γ and β that minimizes a quantity proportional to the sum of squared norms of pairwise differences:*

$$G(X_1, \dots, X_n) = \frac{1}{n} \sum_{l=1}^{n} \sum_{j=l+1}^{n} \|(\beta_l X_l \Gamma_l + 1_h \gamma_i^T) - (\beta_j X_j \Gamma_j + 1_h \gamma_j^T)\|^2 \tag{3.12}$$

under the constraint of the size of the average configuration, $S(\bar{X}) = 1$, being $\bar{X} = \frac{1}{n} \sum_{l=1}^{n} (\beta_l X_l \Gamma_l + 1_h \gamma_l^T)$.

Note also that $G(X_1, \dots, X_n) = \inf_{\mu: S(\mu)=1} \sum_{l=1}^{n} sin^2 \rho(X_l, \mu)$.

Definition 16 *The full Procrustes coordinates of each of the X_i is given by:*

$$X_l^P = \hat{\beta}_l X_l \hat{\Gamma}_l + 1_h \hat{\gamma}_l^T, \;\; l = 1, \dots, n, \tag{3.13}$$

where $\hat{\Gamma}_l \in SO(m)$ (rotation matrix), $\hat{\beta}_l > 0$ (scale parameter), $\hat{\gamma}_l^T$ (location parameter), $l = 1, \dots, n$, are the minimizing parameters.

Result. The point in shape space corresponding to the arithmetic mean of the Procrustes coordinates:

$$\bar{X} = \frac{1}{n} \sum_{l=1}^{n} X_l^P \tag{3.14}$$

has the same shape as the full Procrustes mean.

Because of the non-Euclidean feature of the shape space Σ_m^h, it is not recommended to use the standard multivariate statistical methods in this space. On the contrary, especial statistical techniques should be developed. However, when variation in shape is small, Σ_m^h can be projected to a tangent space, called the Procrustes tangent space (also Kendall or Kent tangent space), where usual Euclidean statistics are acceptable [121]. The point of tangency corresponds to the mean shape. This approach is summarized in the following definition:

Definition 17 *The Procrustes tangent space is the linearized version of the shape space in the vicinity of the Procrustes mean. If the data are fairly concentrated around this mean, the Euclidean distance in the tangent space is a good approach to the Procrustes distances d_F and ρ, so standard statistical techniques in this space can be performed.*

This is an approach to inference on shape space that is extremely important and useful for practical shape analysis and it is widely used in many applications. The projection onto the tangent space can be orthogonal or stereographic. A detailed explanation of this point is given in Refs. [174, 35]. Fig. 3.4 shows an illustration of the tangent space, shape space and pre-shape space for the triangle case, following [174, 35]. In the triangle case, the shape space corresponds to the surface of a sphere of radius $r = 1/2$ [117] and the pre-shape space corresponds to a hemisphere of pre-shapes aligned to the mean reference shape [174, 35]. In Fig. 3.4, O corresponds to the mean reference shape, point M_k represents the position of a shape in Σ_m^h and M_p is its corresponding position in S_m^h. M_s (M_o) is the stereographic (orthogonal) projection of M_k (M_p) onto the tangent space.

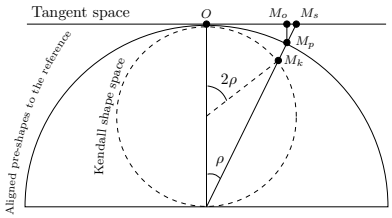


Figure 3.4: Illustration of a cross-section of the construction of the tangent space, shape space and aligned pre-shape space (hemisphere with a radius of 1) for triangles, following [174, 35]. In this case, Σ_m^h is a circle with $r = 1/2$.

In addition, a way to evaluate if shape variation is small, consists in plotting Procrustes distances in shape space against Euclidean distances in the tangent space and checking if they have a highly positive correlation.

3.2.2 The k-means algorithm in the shape space

The objective function of k-means detailed in eq. (2.1) of Chapter 2 can be easily rewritten in these other terms:

$$W(\mathcal{C}) = \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - m_j\|^2, \tag{3.15}$$

where $\mathcal{C} = (C_1, \ldots, C_k)$ is the k-partition of the set $\mathcal{O} = \{x_1, \ldots, x_n\}$ that provides the minimum value of $W(\mathcal{C})$. This type of one-parameter optimization problem is equivalent to the following two-parameter optimization problem:

$$W(\mathcal{C}, m_1, \ldots, m_k) = \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - m_j\|^2, \tag{3.16}$$

where minimization is also with respect to all vectors $M = (m_1, \ldots, m_k)$ of k points $m_1, \ldots, m_k$ from $\mathbb{R}^p$, which we call class representatives or prototypes.

The k-means method tries an approximation to this optimum k-partition by iterating two partial minimization steps [19], see Algorithm 4.

Algorithm 4 Lloyd k-means algorithm

(i) Given a centroid vector M, we minimize with respect to $\mathcal{C}$, assigning each point to the class whose cluster center (centroid) has the Euclidean minimum distance to it.

(ii) Given $\mathcal{C}$, we minimize with respect to M, considering the new centroid vector $M = (\bar{x}_1, \ldots, \bar{x}_k)$, the sample means.

By construction, this algorithm yields a sequence $M(0), \mathcal{C}(1), M(1), \mathcal{C}(2), M(2), \mathcal{C}(3), \ldots$ of centroids and partitions with decreasing values of the objective function of eq. (3.16) that converges to a (typically local) minimum value. The new centroid vector M obtained at each step (ii) of the Algorithm

4 decreases the value of the objective function because of the fact that the sample mean minimizes the Euclidean distance of any point in the cluster.

Consequently, the k-means algorithm to cluster $X_1, \ldots, X_n$ configuration matrices, each of them containing the coordinates of h landmarks of an object in $\mathbb{R}^m$, that arises by integrating to it the Procrustes distance of definition (13) and the Procrustes mean of definition (14) is shown in Algorithm 5.

Algorithm 5 Lloyd k-means algorithm adapted to the shape space

(i) Given a centroid vector $M = ([M_1], \ldots, [M_k])$ $[M_i] \in \Sigma_m^h$ $i = 1, \ldots, k$, we minimize with respect to $\mathcal{C}$, assigning each shape $([X_1], \ldots, [X_n])$ to the class whose centroid has the Procrustes minimum distance to it.

(ii) Given $\mathcal{C}$, we minimize with respect to M, taking $M = ([\widehat{\mu_1}], \ldots, [\widehat{\mu_k}])$, and $[\widehat{\mu_i}]$ $i = 1, \ldots, k$, the Procrustes mean of shapes in C_i.

Steps (i) and (ii) are repeated until convergence of the algorithm.

At this point, we would like to make a few clarifications on the Algorithm 5:

- In the minimization step (i) of Algorithm 5, it is equivalent to use the full Procrustes distance d_F of eq. (3.9) or Procrustes distance ρ of definition (13), because the function $\sin(\theta)$ is an increasing function in $0 \leq \theta \leq \pi/2$. However, we have used the Procrustes distance for programming and running this algorithm because its computational time is significantly lower with the **shapes** R package.
- We choose our starting centroid vector at random.
- Since this algorithm converges to a local optimum, trying several random starts is recommended. Then, the solutions corresponding with the minimum of the objective function (3.16) must be selected.

The original and standard k-means algorithm, known as the Lloyd algorithm in the computer science and pattern recognition fields, is the one presented in Algorithm 4. Some years later, a more efficient version was published by Hartigan and Wong in [91]. Its proposal involves looking for a k-partition with a locally optimal within-cluster sum of squares, by moving

points from one cluster to another. The obvious distinction with Lloyd is that Hartigan-Wong proceeds point by point [199]. The Hartigan-Wong algorithm is shown in Algorithm 6. Its adaptation to the shape space is given in [4].

3.3 kmeansProcrustes

The idea behind 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. Two basic concepts of the statistical shape analysis are the Procrustes distance and the Procrustes mean. Our aim in this methodology has been adapting k-means to the shape analysis field by replacing the Euclidean distance and the sample mean by the Procrustes distance and mean, respectively. After a literature search, two papers were found related to this topic. One the one hand, in [79] a type of k-means algorithm is used (in some aspects similar to its Lloyd version) to cluster fuzzy shapes. On the other hand, in [4] the use of the Hartigan-Wong k-means algorithm for clustering shapes was proposed.

Our approach is very similar to that one of [4] but we use the Lloyd version of k-means and our application is different. In Section 3.3.2.1 , we will show that Lloyd k-means performs computationally better than Hartigan-Wong k-means in the shape analysis context. In addition, in Section 3.3.2.2 the Lloyd k-means is used as an attempt to define an efficient apparel sizing system using the 3D shape information of women. Besides, the shape variability for the clustering results obtained will be analyzed in Section 3.3.2.3.

3.3.1 Methodology

The data set we use is made up of the same women than in the *trimowa*, *biclustAnthropom* and *hipamAnthropom* methodologies. The women shape is represented by a configuration matrix of landmarks. Table 3.1 describes the anthropometric meaning of each of the 66 landmarks we use in this study. In addition, Fig. 3.5 shows the position of the landmarks on the woman's body.

The procedure we use to try to define the sizing system is analogous to those used with *trimowa* and *hipamAnthropom*. However, we segment now not only by bust circumference, but also by height (following the *European*

Algorithm 6 Hartigan-Wong k-means algorithm

(i) Given a centroid vector $M = (m_1, \ldots, m_k)$, for each point x_j $(j = 1, 2, ..., n)$, find its closest and second closest cluster centroids, and denote these clusters by $C1(j)$ and $C2(j)$, respectively. Assign point j to cluster $C1(j)$.

(ii) Update the cluster centroids to be the averages of points contained within them.

(iii) Initially, all clusters belong to the live set.

(iv) This stage is called the *optimal-transfer stage*: Consider each point x_j $(j = 1, 2, ..., n)$ in turn. If cluster l $(l = 1, 2, ..., k)$ is updated in the last quick-transfer stage, then it belongs to the live set throughout that stage. Otherwise, at each step, it is not in the live set if it has not been updated in the last n optimal-transfer steps. Let point x_j be in cluster l_1. If l_1 is in the live set, do Step (iv-a). Otherwise, do Step (iv-b).

(iv-a) Compute the minimum of the quantity, $R2 = \frac{n_l \|x_j - m_l\|^2}{n_l + 1}$, over all clusters l $(l \neq l_i,\ l = 1, 2, ..., k)$. Let l_2 be the cluster with the smallest $R2$. If this value is greater than or equal to $\frac{n_{l_1} \|x_j - m_{l_1}\|^2}{n_{l_1} + 1}$, no reallocation is necessary and C_{l_2} is the new $C2(j)$. Otherwise, point j is allocated to cluster l_2 and C_{l_1} is the new $C1(j)$. Cluster centroids are updated to be the means of points assigned to them if reallocation has taken place. The two clusters that are involved in the transfer of point j at this particular step are now in the live set.

(iv-b) This step is the same as Step (iv-a), except that the minimum $R2$ is only computed over clusters in the live set.

(v) Stop if the live set is empty. Otherwise, go to Step (vi) after one pass through the data set.

(vi) This is the *quick-transfer stage*: Consider each point x_j $(j = 1, 2, ..., n)$ in turn. Let $l_1 = C1(j)$ and $l_2 = C2(j)$. It is not necessary to check point j if both clusters l_1 and l_2 have not changed in the last n steps. Compute the values $R1 = \frac{n_{l_1} \|x_j - m_{l_1}\|^2}{n_{l_1} + 1}$ and $R2 = \frac{n_{l_2} \|x_j - m_{l_2}\|^2}{n_{l_2} + 1}$. If $R1$ is less than $R2$, point j remains in cluster l_1. Otherwise, switch $C1(j)$ and $C2(j)$ and update the centroids of clusters l_1 and l_2. The two clusters are also noteworthy for their involvement in a transfer at this step.

(vii) If no transfer took place in the last n steps, go to Step (iv). Otherwise, go to Step (vi).

Landmark	Description	Landmark	Description
1. Head back	Most prominent point of the head in the sagittal plane	34. Armpit back left	Left back armpit
2. Head front	Glabella (most prominent point of the forehead)	35. Armpit back right	Right back armpit point
3. Forearm wrist left	Maximum girth of left forearm	36. Armpit left	Left front armpit point
4. Forearm girth left	Maximum girth of left forearm just under left elbow	37. Armpit right	Right front armpit point
5. Forearm wrist right	Maximum girth of right forearm	38. Breast front	Most prominent point of the sagittal plane breast area
6. Forearm girth right	Maximum girth of the right forearm just under the right elbow	39. Crotch	Crotch at the anterior side
7. Wrist girth left	Styloid apophysis of left radius	40. Crotch center	Crotch at the coronal plane level
8. Wrist girth right	Styloid apophysis of right radius	41. Crotch center left	Midpoint of the left thigh at crotch level coronal plane
9. Elbow left	Most prominent point of left elbow	42. Crotch center right	Midpoint of right thigh at crotch level coronal plane
10. Elbow right	Most prominent point of the right elbow	43. Crotch back	Crotch at the posterior side
11. Nipple left	Left nipple	44. Sideseam left	Projection over the floor of the left lateral malleolus
12. Inseam right	Most medial point of right foot at floor level	45. Sideseam right	Projection over floor of right lateral malleolus
13. Scapula	Most prominent point of back sagittal plane	46. Hip girth front	Medial front point of maximum girth of buttock area
14. Knee left	Most prominent point of left knee sagittal plane	47. Mid neck girth	Frontal point of neck mean section perpendicular to line between chin-suprasternal notch
15. Knee right	Most prominent point of right knee sagittal plane	48. Mid neck girth sternum	Mid point of line between neck and shoulders
16. Belly front	Most prominent point of sagittal plane in belly area	49. Midriff girth	Frontal point of underbust
17. Inseam left	Most medial point of the left foot at floor level	50. Neck back	Most upper y-axis extreme point of section defined by suprasternal notch
18. Nipple right	Right nipple	51. Neck front	Suprasternal notch
19. Waist girth left	Most left x-axis extreme point of section defined by minimum girth of torso	52. Neck left	Left point of line between clavicular notches just above suprasternal notch
20. Waist girth right	Most right x-axis extreme point of section defined by minimum girth of torso	53. Neck right	Right point of the line between clavicular notches just above suprasternal notch
21. Waist girth back	Most upper y-axis extreme point of section defined by minimum girth of the torso	54. Thigh crease left	Most prominent frontal point of the line between crotch and left knee
22. Waist girth front	Most lower y-axis extreme point of section defined by minimum girth of torso	55. Thigh crease right	Most prominent frontal point of line between crotch and right knee
23. High waist girth back	Most upper y-axis extreme point of section defined by narrowest points of torso in frontal view	56. Mid neck girth left	Left point of neck mean section perpendicular to line between chin and suprasternal notch
24. High waist girth front	Most lower y-axis extreme point of section defined by narrowest points of torso in frontal view	57. Mid neck girth right	Right point of neck mean section perpendicular to line between chin and suprasternal notch
25. High waist girth left	Most left x-axis extreme point of section defined by narrowest points of torso in frontal view	58. CV	7th cervical
26. High waist girth right	Most right x-axis extreme point of section defined by narrowest points of torso in frontal view	59. Floor	Point indicating the floor
27. Max belly circ. front	Extreme y-axis point of section defined by maximum waist girth	60. Vertex	Highest point of head
28. Max hip girth front	Extreme y-axis point of section defined by maximum hip girth	61. Right acromion	Physical marker of right shoulder
29. Calf girth left	Maximum girth of left calf	62. Left acromion	Physical marker of left shoulder
30. Calf girth right	Maximum girth of right calf	63. Right side of rib cage	Physical marker of right side of rib cage
31. Scapula	Most prominent point of back in sagittal plane	64. Left side of rib cage	Physical marker of left side of rib cage
32. Ankle left	Left medial malleolus	65. Right iliac crest	Physical marker on the right of iliac crest
33. Ankle right	Right medial malleolus	66. Left iliac crest	Physical marker on the left of iliac crest

Table 3.1: Anthropometric landmarks used in the analysis.

standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals [59], as usual). With this type of segmentation, all women belonging to the same group are different in shape but similar in size, so the size effect is filtered out in an easy way. We select those 10 groups composed of a reasonable number of women according to the apparel companies policy. Both bust and height measurements and number of women of each groups are described in Table 3.2. Finally, we apply the Lloyd algorithm to each one of the 10 groups with $k = 3$.

One important drawback to k-means is that k is a tuning parameter and the question of which number k of clusters to choose is one of the most difficult problems in data clustering and in particular, in our application. We initially chose $k = 3$ because this number of sizes is quite well aligned to the strategy of designing sizes. However, it may happen in some cases that $k = 2$ would be enough to accommodate the whole population. For those cases, we need some objective measure to help make a decision. Some papers review the methods to estimate k (see e.g. [194, 108]). In this analysis, we are going to consider the widely used silhouette plot [115] (see Section 2.2.1). An appropriate choice of k can be made on the basis of the validity index the silhouette computes.

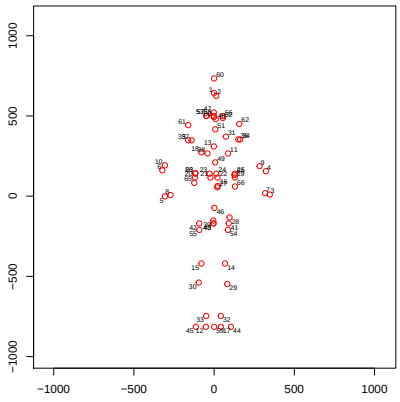


Figure 3.5: Set of 66 body landmarks used in the study. Each number identifies the corresponding landmark described in Table 3.1. This plot represents the projection in the xy plane of a woman who belongs to the group of height less than 162 cm and bust between 74 and 82 cm, see Table 3.2.

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

Table 3.2: Bust and height measurement ranges used to segment our data set, together with the number of women in each group.

3.3.2 Results

3.3.2.1 Simulation study

Hartigan-Wong k-means requires the mean to be updated many more times than Lloyd because it proceeds point by point. Although the calculus of the sample mean in the Euclidean space has a negligible computational cost, the calculus of the Procrustes mean needs much more time. Consequently, the Hartigan-Wong should have a high computational cost in the shape space, losing efficiency. This problem should be stressed when the sample size increases. In order to demonstrate empirically these hypothesis, we carried out the following simulation study with controlled data.

Configurations are described by l landmarks, each one of them playing the role of a random variable. Accordingly, we must generate random data following multivariate distributions. First of all, we build two compact geometric figures, a cube and a parallelepiped, with a number of landmarks equal to $l = 8$ and $l = 34$. Figs. 3.6 and 3.7 display the cube and the parallelepiped with $l = 8$ and $l = 34$ landmarks, respectively.

Once this is done, we simulate n_1 cubes corresponding to one cluster and n_2 parallelepipeds corresponding to another cluster. Mathematically, cluster 1 (resp. cluster 2) is defined by a multivariate normal distribution of a $3l$-dimensional mean vector represented by the previously generated cube (resp. parallelepiped), and an $l \times l$ covariance matrix $\Sigma_1 = \sigma_1 I_{3l}$ (resp. $\Sigma_2 = \sigma_2 I_{3l}$), $l = 8, 34$. Tables 3.3 and 3.4 describes the coordinates of both mean vectors for $l = 8$ (Fig. 3.6 shows the landmark labels). For $l = 34$, its corresponding mean vectors are too long to show.

Figure 3.6: Cube and parallelepiped formed by 8 landmarks. Each number indicates the label of the corresponding landmark according to Tables 3.3 and 3.4.

Figure 3.7: Cube and parallelepiped formed by 34 landmarks.

Landmark label \ Dimension	x	y	z
1	0	0	10
2	0	10	10
3	0	0	0
4	0	10	0
5	10	10	10
6	10	10	0
7	10	0	10
8	10	0	0

Table 3.3: Coordinates of the mean shape for cluster 1 in the case of $l = 8$.

Landmark label \ Dimension	x	y	z
1	0	0	10
2	10	0	10
3	0	0	0
4	10	0	0
5	10	20	10
6	10	20	0
7	0	20	10
8	0	20	0

Table 3.4: Coordinates of the mean shape for cluster 2 in the case of $l = 8$.

For each l, we first apply the Lloyd k-means and then the Hartigan-Wong k-means, to the combination of both clusters for different values of n_1, n_2, σ_1 and σ_2. We select the same values for n_1 and n_2 regarding different sample sizes: small sample (50), medium sample (500) and large sample (900). These sample sizes are chosen in this way because they approximate the number of women that belong to each group to we will apply the Lloyd k-means in Section 3.3.2.2. The values for σ_1 and σ_2 are selected in such a way that the data are more or less dispersed (0.1, 3 and 6) and they are equal in each case as well. Regarding the programming of the algorithms, we fix both the number of random initializations and maximum number of steps per initialization to 10. We also fix a relative stopping criteria equal to 0.0001. For each iteration, we save the random initial values used by Lloyd and next Hartigan-Wong is executed at the same iteration with the same values.

The clustering effectiveness of both algorithms has been evaluated by computing the allocation rate. The allocation rate is defined as the proportion of correctly allocated observations from that population in the sample. The computational time is the time we waited for the code to run. Tables 3.5 and 3.6 describe the results obtained in terms of the basic descriptives (mean, $\bar{x}$, and standard deviation, sd) of the allocation rate and computational time returned at each random initialization for both $l = 8$ and $l = 34$.

				l=8 landmarks							
				Lloyd algorithm				Hartigan-Wong algorithm			
				Alloc. rate		Comput. time		Alloc. rate		Comput. time	
n_1	n_2	σ_1	σ_2	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
25	25	0.1	0.1	1	0	4,67s.	1,88s.	1	0	9,03s.	10,06s.
25	25	3	3	1	0	5,77s.	1,77s.	1	0	8,54s.	9,29s.
25	25	6	6	0.97	0.01	11,52s.	3,59s.	0.98	0	21,09s.	10,67s.
250	250	0.1	0.1	1	0	57,69s.	33,84s.	1	0	14m.15s	14m.
250	250	3	3	1	0	56,47s.	18,87s.	1	0	18m.30s.	25m.05s.
250	250	6	6	0.948	0.0013	2m.12s.	24s.	0.948	0	29m.	13m.25s.
450	450	0.1	0.1	1	0	2m.	1m.	1	0	1h.20m.	58m.36s.
450	450	3	3	0.998	0	1m.30s.	30s.	0.998	0	29m.40s.	46m.25s.
450	450	6	6	0.95	0	3m.30s.	1m.35s.	0.95	0	1h.52m.	1h.03m.

Table 3.5: Allocation rate and computational time of both k-means algorithms applied to the cube and parallelepiped represented by 8 landmarks. s stands for *seconds*, m stands for *minutes* and h stands for *hours*.

				l=34 landmarks							
				Lloyd algorithm				Hartigan-Wong algorithm			
				Alloc. rate		Comput. time		Alloc. rate		Comput. time	
n_1	n_2	σ_1	σ_2	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
25	25	0.1	0.1	1	0	20s.	3,47s.	1	0	23,59s.	17,71s.
25	25	3	3	0.924	0.05	39,61s.	15,15s.	0.96	0	46s.	19,66s.
25	25	6	6	0.674	0.078	24,78s.	5,79s.	0.712	0.11	43,22s.	9,56s.
250	250	0.1	0.1	1	0	3m.12s.	26,16s.	1	0	22m.08s.	19m.50s.
250	250	3	3	0.9866	0.0013	5m.	38,27s.	0.9863	0.00076	57m.43s.	15m.
250	250	6	6	0.86	0.017	8m.30s.	1m.	0.87	0.009	1h.18m.	11m.
450	450	0.1	0.1	0.988	0.0008	9m.	42s.	0.988	0	2h.45m.	50m.
450	450	3	3	0.989	0.0009	7m.27s.	1m.10s.	0.989	0	2h.29m.	55m.
450	450	6	6	0.903	0.0032	15m.	2m.	0.886	0.004	3h.30m.	1h.10m.

Table 3.6: Allocation rate and computational time of both k-means algorithms applied to the cube and parallelepiped represented by 34 landmarks. s stands for *seconds*, m stands for *minutes* and h stands for *hours*.

The analysis of the results lead to similar conclusions for both algorithms: When the sample size is small ($n_1 = n_2 = 25$), we see that the Hartigan-Wong version has a larger computational time than the associated with the Lloyd version, but its performance is actually quite reasonable. In fact, it obtains an allocation rate slightly better than the Lloyd algorithm for $l = 34$. Nevertheless, when the sample size is larger ($n_1 = n_2 = 250$ and $n_1 = n_2 = 450$) for both $l = 8$ and $l = 34$, Hartigan-Wong has a poor performance. Its computational time increases seriously and its allocation

rate is only the same or even worse than the corresponding one provided by the Lloyd version in the majority of the cases (see for example, $n_1 = n_2 = 450$ with $\sigma_1 = \sigma_2 = 6$ for both $l = 8, 34$). The allocation rate of Hartigan-Wong is slighlty better for $l = 34$ when $n_1 = n_2 = 250$ with $\sigma_1 = \sigma_2 = 6$ but again, its computational time is quite larger.

According to this analysis, we would expect that for a larger number of landmarks and a medium or large sample size the Hartigan-Wong algorithm will be totally computationally unoperative. We did some checks for $l = 44$ and $l = 56$ landmarks with a medium or large sample size and we confirmed this thought: Hartigan-Wong took so much time to compute only for the first step, that it was pointless to continue. On the contrary, the Lloyd algorithm continued performing well.

After this comprehensive study we are able to claim that the Lloyd version of k-means represents a noticeable reduction in the computation involved in the context of shape analysis. We will refer the Lloyd k-means adapted to shape analysis to as *kmeansProcrustes*.

3.3.2.2 Application of *kmeansProcrustes* for clustering human body shapes

Next, we focus on using *kmeansProcrustes* to try to propose an efficient sizing system by clustering human body shapes. As an illustrative example of the global clustering results obtained, we detail the results for the group that contains women of bust between 90 and 98 cm and height between 162 and 174 cm (671 women). This is the group that shows greater differences among clusters for the key variables bust, waist and hip circumference and neck to ground length, also taking into account the number of women belonging to it. These results can be seen in Table 3.7. In addition, the 3D mean shapes for each one of the three clusters can be found in Fig. 3.8. In order to examine the differences among clusters, we represent their boxplots for the four aforementioned anthropometric dimensions. As it can be seen in Fig. 3.9, the three clusters are very different, especially cluster 2, which it is the one that presents more different values in all variables.

Cluster 1	Cluster 2	Cluster 3
299	184	188

Table 3.7: Partition for the group with bust $\in [90, 98[$ cm and height of $\in [162, 174[$ cm using the same Lloyd k-means used in the simulation study.

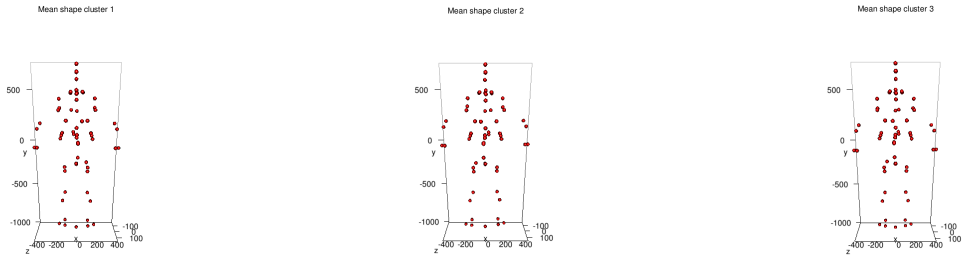


Figure 3.8: 3D mean shapes for each one of the three clusters.

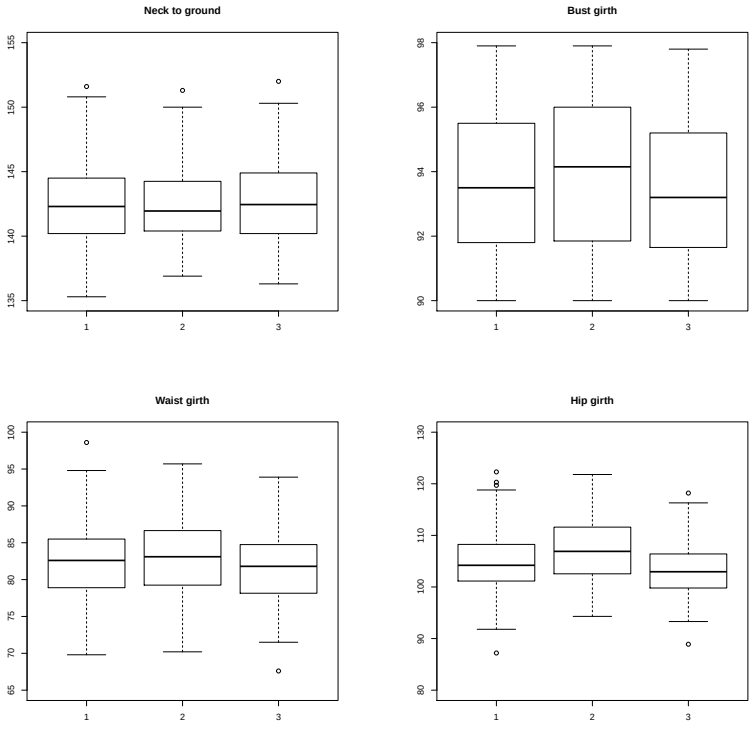


Figure 3.9: Boxplots for neck to ground, bust, waist and hip measurements for the three clusters obtained with the Lloyd k-means applied to the group with bust of $[90, 98[$ cm and height of $[162, 174[$ cm. The three clusters are very different among themselves.

Specifically, cluster 2 includes women with lower neck to ground and larger bust, waist and hip girth measurements regarding the women of clusters 1 and 3. Because these clusters are different to each other for some key anthropometric variables and include a reasonable number of women, they could be consider as efficient sizes. Fig. 3.10 displays the Procrustes rotated data for all the women that belongs to each cluster, with their Procrustes mean shape superimposed and projected in the xy plane. The point clouds corresponding to the feet, elbows and wrists present the most variation.

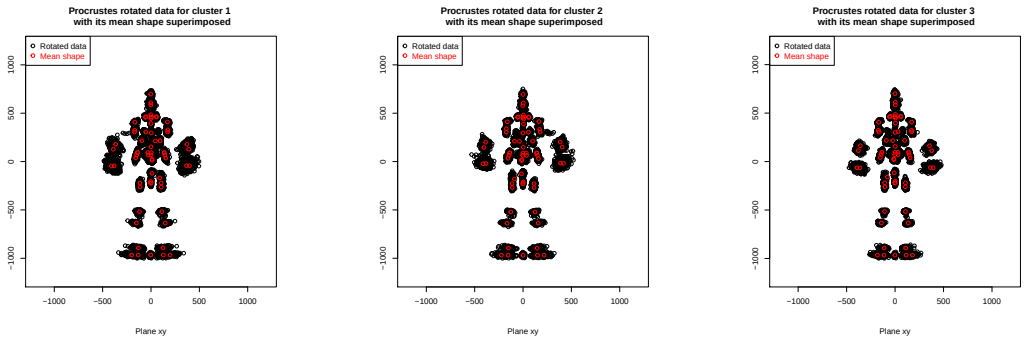


Figure 3.10: Projection on the plane xy of the rotated points and mean shape for clusters 1, 2 and 3. The point clouds corresponding to the feet, elbows and wrists presents the most variation.

The plots of Fig. 3.9 suggest the possibility of choosing only two size groups because the second cluster seems different than clusters one and three, but one and three seem quite similar. Accordingly, we investigate whether $k = 2$ would be enough to accommodate this population segment. For that purpose, the silhouette plots for $k = 2$ and $k = 3$ are compared.

Table 3.8 shows the clustering results for $k = 2$.

Cluster 1	Cluster 2
319	352

Table 3.8: Partition for the group with bust $\in [90, 98[$ cm and height of $\in [162, 174[$ cm using the same Lloyd k-means used with k=3 but now with $k = 2$.

These two clusters are also quite balanced. The silhouette plots for both $k = 2$ and $k = 3$ are displayed in Fig. 3.11. We see that both results are quite similar but slightly better for $k = 2$. Taking into account that an efficient sizing system aims at accommodating as large a percentage of the population as possible, we think that this small difference is not enough to consider $k = 2$. However, we also think that the final decision should be made by an expert.

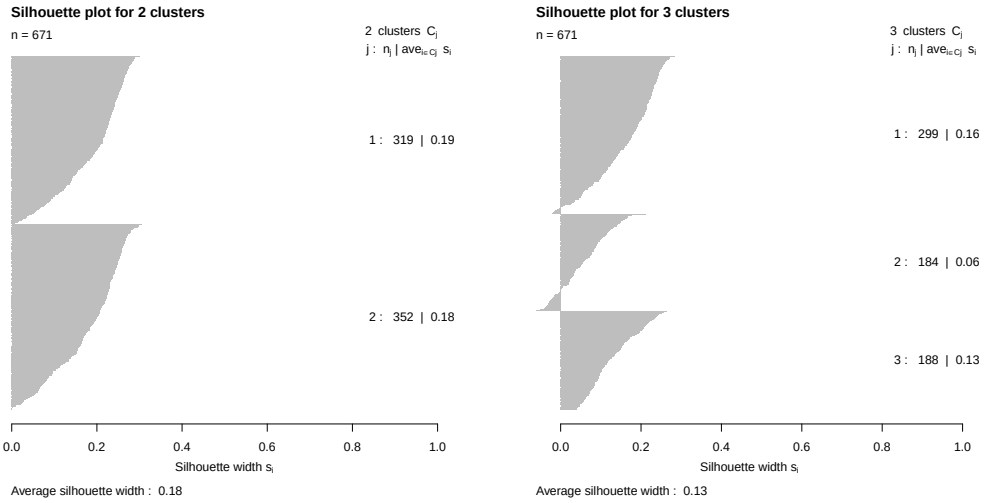


Figure 3.11: Silhouette plot associated with two (left) and three (right) clusters.

3.3.2.3 Analysis of shape variability

After calculating the mean shape of each cluster (its average configuration), the next step is to examine the variability in shape within the clusters. As indicated in [47], PCA in Procrustes tangent space coordinates is a very effective way of analyzing the main modes of variation in shape. For illustration purposes, we are going to analyze the cluster 1 and the analysis for the other clusters would be analogous. A first way to visualize the effect of each PC consists in drawing vectors from the mean shape to +3 sd's along the first three PCs. This is represented in Fig. 3.12, where the structure of shape variability in each one of the first three PCs can be directly evaluated. The black lines are related to the first PC and we see that the landmarks associated with arms and wrists present the most variation. The red lines represent

the second PC. In this case the landmarks associated with feet and knees are the most variables. The green lines correspond to the third PC. In Fig. 3.12 we only appreciate a small green line in the landmarks related to nipples, so the most relevant variation for this component occurs in the chest.

Regarding the analysis of the first and second component, it is worth pointing out that the relative position of women when they were scanned, has contributed mostly to the shape variability. This is similar to the example 5.5 and figure 64 explained in [47] for describing shape variability in hands. However, our case study is not so extreme. In the example shown in [47] it is clearly appreciated that hands differs in position. Instead, in the Spanish anthropometric survey, every effort was made to place women in the same position before scanning. A complementary analysis could be done without landmarks related to extremities.

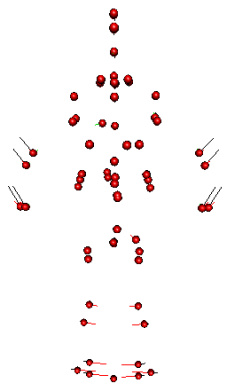


Figure 3.12: Vectors from the mean shape to +3 sd's along the first three PCs. The black lines refer to the first PC, the red lines refer to the second PC and the green line refer to the third PC.

Figs. 3.13 shows four scatterplots: ρ_i vs s_i (top left), ρ_i vs c_{i1} (top right), ρ_i vs c_{i2} (bottom left) and ρ_i vs c_{i3} (bottom right), where ρ_i are the Procrustes distances to the mean shape, s_i are the centroid sizes of the configuration and c_{i1}, c_{i2}, c_{i3} are the first three standardized PC scores, being $i = 1, \ldots, 299$

(the first cluster has 299 women). Plots ρ_i vs c_{ij}, $j = 1, 2, 3$ are called PCA plots and each point represents the shape of a specific women. The closer two women are, the more similar in shape they are. We see in all plots of Fig. 3.13 that there is a potential outlier. She is a woman with a large Procrustes distance to the mean shape of her cluster. This woman is the number 139 in the cluster 1 (299 women) and the woman 327 in her height and bust group (671 women). When we only plot these first three components, see Fig. 3.14, this woman appears also as an extreme woman for the first two principal components. On closer inspection of this extreme woman by means of Fig. 3.15, we observe that the landmarks related to the head, forehead, neck and shoulders are poorly placed. Thus, the Procrustes distance can be used to identify outliers.

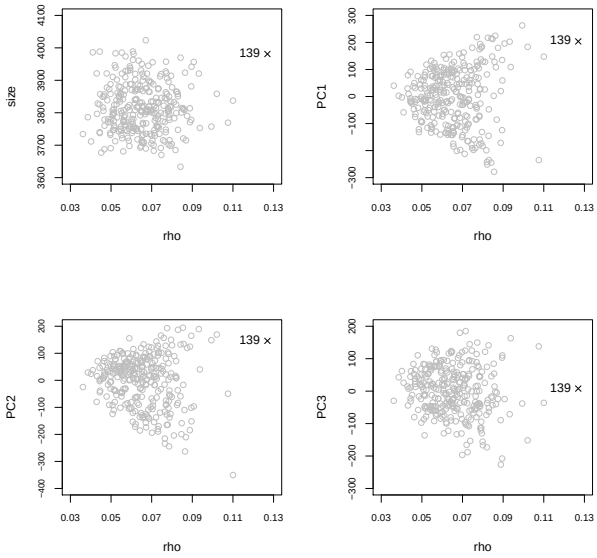


Figure 3.13: rho vs. size and rho vs. principal component scores plots. Woman 139 of cluster 1 (woman 327 in the whole height and bust class) is marked with a cross.

3.3.2.4 Trimmed *kmeansProcrustes*

At this point, we propose to use a trimmed procedure for clustering shapes. Adapted to the shape analysis context, the trimmed k-means algorithm re-

places the step (i) of the Lloyd k-means stated in Section 3.3, with:

(i) Given a centroid vector $M = ([M_1], \ldots, [M_k])$ $[M_i] \in \Sigma_m^h$ $i = 1, \ldots, k$, we calculate the Procrustes distances of each shape $([X_1], \ldots, [X_n])$ to its closest centroid. The $n\alpha$ shapes with largest distances are removed, the $n(1-\alpha)$ left are assigned to the class whose centroid has the minimum full Procrustes distance to it.

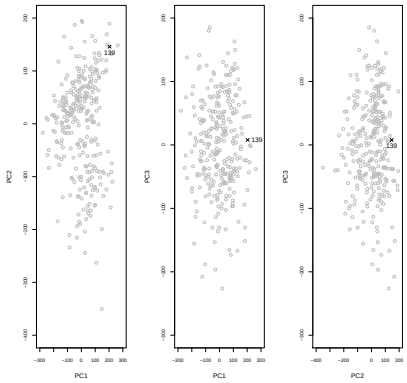


Figure 3.14: Principal component scores plots. Woman 139 of cluster 1 (woman 327 in the whole height and bust class) is marked with a cross.

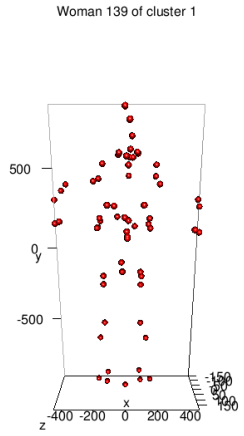


Figure 3.15: 3D shape for woman 139.

Incorporating the trimming approach to the Lloyd k-means is very easy. On the contrary, more difficulties appear when trying to do the same into the Hartigan-Wong version. We apply the trimmed Lloyd to the same above explained group (bust $\in [90, 98[$ cm and height $\in [162, 174[$ cm) with the same algorithm parameters and the proportion of the trimmed sample fixed at 1%. Hence, the number of women being deleted will be seven. We observe in Table 3.9 that a similar clustering partition in terms of individuals in each cluster has been obtained regarding Table 3.7.

Cluster 1	Cluster 2	Cluster 3
187	265	212

Table 3.9: Clustering partition for the group with bust $[90, 98[$ cm and height $[162 - 174[$ cm after applying the trimmed version of the Lloyd k-means algorithm with random initial values and 10 iterations.

By examining Table 3.10, we analyze the seven trimmed women in anthropometric terms. The most relevant aspect is that the woman 327 has been trimmed. Another interesting fact is that woman 506 has the largest waist for the considered height and bust group. Fig. 3.16 shows the boxplots for each cluster with respect to the four same anthropometric dimensions. Because we have deleted the most extreme individuals, the anthropometric differences between clusters are now more evident. In addition, the cluster means are more representative.

Woman label	Neck to ground	Bust	Waist	Hip	Chest
73	144.2	94.5	84.2	104.1	96.3695
74	141.3	93.9	83.2	97.9	96.3930
205	138.1	92.0	77.7	106.3	92.7016
327	147.0	94.2	82.2	104.2	94.3778
383	143.1	96.5	84.6	101.6	98.7048
463	146.4	96.3	85.9	106.1	99.9697
506	143.8	97.7	98.6	109.9	98.6261

Table 3.10: Anthropometric dimensions of the trimmed women. Woman 327 is highlighted.

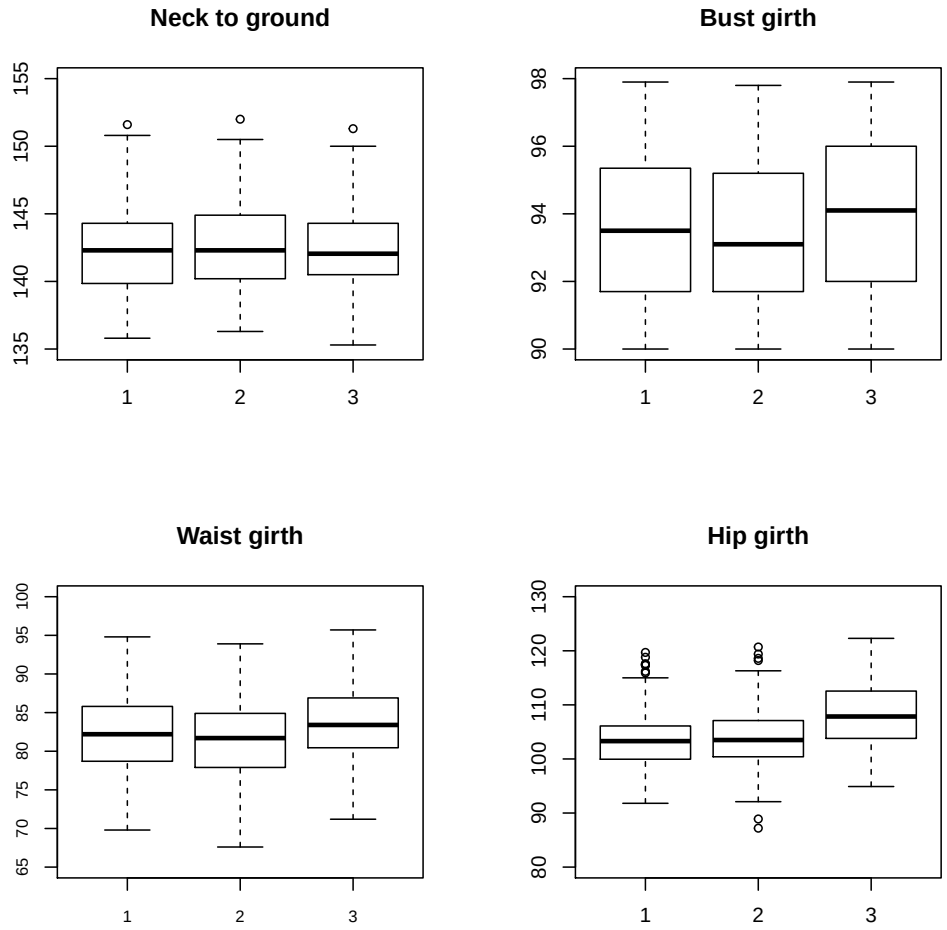


Figure 3.16: Boxplots for neck to ground, bust, waist and hip measurements for the three clusters obtained with the trimmed k-means algorithm applied to shape analysis in the group with bust $[90, 98[$ cm and height $[162, 174[$ cm.

3.3.3 Summary

We have adapted the k-means clustering algorithm to the shape analysis context. Two versions of k-means have been compared and one of them has been used to try to define an efficient clothing sizing system (with and without trimming). In addition, the shape variation of the clustering results

obtained has been analyzed.

In two previous papers it was suggested how k-means could be generalised in the field of statistical shape analysis for clustering elements according to their shapes. Each one of them used a different version of k-means algorithm: Amaral et al. [4] adapted its Hartigan-Wong version, whereas Georgescu [79] proposed a similar but not identical Lloyd k-means on a different context. Bearing in mind that the computation of the Procrustes mean requires a high computational cost, we wanted to empirically demonstrate that Lloyd should perform better than Hartigan-Wong for grouping shapes. To that end, we have carried out a numerical simulation comparing both k-means versions and we have concluded that the Lloyd version clearly has a better performance (considering both computational cost and clustering effectiveness), especially when the sample size is large. The Hartigan-Wong algorithm is only feasible for small problems. In fact, in [4] a database made up of only 49 individuals represented by only 11 landmarks was used. That's why Hartigan-Wong was suitable there. However, in case of larger samples, the Lloyd version should be used. Therefore, our proposal represents a valuable contribution to the field of statistical shape analysis. In addition, to the best of our knowledge, our work represents the first attempt at adding a trimmed approach for clustering shapes.

From the point of our application, we have observed that the clustering results provided by the Lloyd algorithm show meaningful differences among the anthropometric dimensions considered, while containing a reasonable number of women. Hence, these clusters could be consider as efficient sizes and therefore, this clustering method could be useful to define optimal sizes groups when clustering human body shapes represented by landmarks.

3.4 Chapter conclusions

In this chapter we have proposed an approach that represents a novelty in terms of integrating concepts of the statistical shape analysis in clustering procedures.

Unlike the *trimowa* and *biclustAnthropom* clustering methods presented in chapter 2, which are used to define an apparel sizing system from a multivariate perspective taking into account some anthropometric dimensions, the method developed in this chapter makes it possible to divide the women sample into efficient size groups by using the information of their body shape

represented by anatomical markers called landmarks. The proposed algorithms not only allow human beings to be grouped, but also any animal or fossil species whose shape is clearly determined by an enough amount of landmarks. Therefore, researchers from a lot of scientific fields, such as Archaeology or Oceanography, can use our algorithms as a statistically robust tool for finding common patterns and grouping individuals according to their morphology, regardless of the sample size.

Three future work possibilities are open: firstly, we would like to consider the size component and consequently to work at the size-and-shape space. Size is recognized as an important component of the comparison of structures. Secondly, we aim at applying our proposal to a larger number of landmarks representing the whole human body or a single part of the body, such as its trunk (see Fig. 3.17 for examples of both cases). A big amount of landmarks represents the shape information of the objects on a more accurate manner. Besides, we will analyze the results obtained with PAM using as input the Procrustes distances.

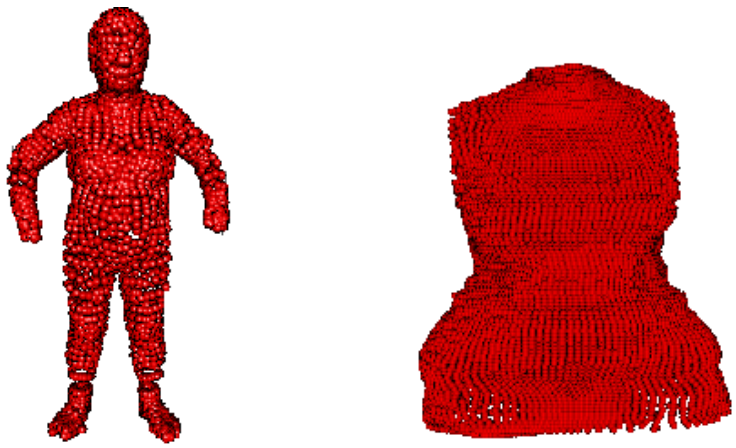


Figure 3.17: Whole human body (left) and trunk (right) represented by a larger amount of landmarks.

Chapter 4

Statistical data depth

4.1 Introduction [1]

Both the univariate median and the univariate mean summarize quantitative data by means of a single numerical quantity that reflects the central tendency of data. The univariate median is the value that separates the data into two equal halves: 50% of the numbers are below the median and the other 50% are above. It is obtained by putting the values in ascending order and choosing the middle one [220].

Recently, the ordering of multivariate data and consequently the multivariate extension of the median concept has received close attention. The statistical notion of data depth has emerged as a powerful generalization of the order-related univariate statistical methods. A data depth measures how deep a point is regarding a multivariate probability distribution or regarding a data cloud. This leads to a natural center-outward ranking of the sample points: The deepest points are those closest to the center of the data cloud and as we move away from the center, the depth of the points decreases.

Given a probability distribution F in $\mathbb{R}^d$, $d \geq 1$, a data depth is a way of measuring how deep (or central) a given point $x \in \mathbb{R}^d$ is regarding F or regarding a given data cloud $\{x_1, \ldots, x_n\}$. We give its formal definition following [234]:

Definition 1 *Denote by $\mathbb{F}$ the class of distributions on $\mathbb{R}^d$. A statistical depth function is a bounded nonnegative mapping $D(\cdot;\cdot)\colon \mathbb{R}^d \times \mathbb{F} \longrightarrow \mathbb{R}^1$ satisfying:*

[1] Section 4.3 belongs to a paper in progress [214].

1. **Affine invariance:** $D(Ax + b; F_{AX+b}) = D(x; F_X)$, *where* F_X *is the induced probability by the random vector* $X \in \mathbb{R}^d$, A *is a* $d \times d$ *nonsingular matrix and* $b \in \mathbb{R}^d$. *This means that the depth of a point* $x \in \mathbb{R}^d$ *should not depend on the underlying coordinate system.*

2. **Maximality at center:** $D(\theta; F) = sup_{x \in \mathbb{R}^d} D(x; F)$, *for any* $F \in \mathbb{F}$ *whose center is* θ. *That is, for a distribution having a uniquely defined "center", the depth function should attain maximum value at this center.*

3. **Monotonicity regarding the deepest point:** $D(x; F) \leq D(\theta + \lambda(x - \theta); F)$ *holds for any* F *having a deepest point* θ *and any* $\lambda \in [0, 1]$. *This means that as a point* $x \in \mathbb{R}^d$ *moves away from* θ *along any fixed ray through the center, the depth at* x *should decrease monotonically.*

4. **Vanishes at infinity:** $\lim_{\|x\| \to \infty} D(x; F) = 0$, *i.e., the depth of a point* $x \in \mathbb{R}^d$ *should approach zero as its norm approaches infinity.*

From this definition, different particular cases of depth functions have been proposed in the last decades [131, 178, 132, 22, 27]. The following are some of the most broadly used. In [135] (in Spanish), some detailed comments about them are given.

- **Mahalanobis depth** [143] at $x \in \mathbb{R}^d$ with respect to (w.r.t) F is defined to be:

$$M_h D(x; F) = \left[1 + (x - \mu_F)\Sigma_F^{-1}(x - \mu_F)\right]^{-1} \tag{4.1}$$

  where μ_F and Σ_F are the mean vector and the variance-covariance matrix of F, respectively. This depth function is based on the Mahalanobis distance [143].

- **Tukey depth** (also known as halfspace depth [205]) at $x \in \mathbb{R}^d$ w.r.t F is defined to be:

$$HD(x; F) = \inf_H \{P(H) : H \; is \; a \; closed \; halfspace \; in \; \mathbb{R}^d \; and \; x \in H\} \tag{4.2}$$

  It is the minimal probability (P in the above definition) which can be achieved in a closed halfspace that contains x [36] (a halfspace is

either of the two parts into which a hyperplane divides an affine space).

- **Convex hull peeling depth** [11]. Given $\{x_1, \ldots, x_n\}$ a multivariate data set, the procedure to calculate this depth function is the following (it begins with $i = 1$):

  1. The convex hull is built. The set of its vertices is called the convex layer, C_i.
  2. The points belonging to C_i are removed.
  3. i increases one unit and the process is repeated until there are no more remaining points.

  The outermost points will be those of C_1, and with further C_i ($i = 2, \ldots$), increases the degree of centrality of the points. In this way, the convex hull peeling depth at x_k regarding the data set $\{x_1, \ldots, x_n\}$ is the level of the convex layer to which x_k belongs.

- **Oja depth** [156] at $x \in \mathbb{R}^d$ w.r.t F is defined to be:

$$OD(x; F) = \Big[1 + E_F\big\{volume(S[x, X_1, \ldots, X_d])\big\}\Big]^{-1} \tag{4.3}$$

  where $S[x, X_1, \ldots, X_d]$ is the closed simplex with vertices x and d random observations $X_1, \ldots, X_d$ from F.

- **Simplicial depth** (also known as Liu depth [131]) at $x \in \mathbb{R}^d$ w.r.t F is defined to be:

$$SD(x; F) = P\big\{x \in S[X_1, \ldots, X_{d+1}]\big\} \tag{4.4}$$

  where $S[X_1, \ldots, X_{d+1}]$ is a closed simplex formed by $(d + 1)$ random observations from F. It is the probability (P in the above definition) that x belongs to a simplex formed by a sample of F of size $(d + 1)$.

- **Projection depth** [231]. Let μ and σ be univariate location and scale measures, respectively. Define the *outlyingness* of a point $x \in \mathbb{R}^d$ regarding F as:

$$O(x;F) = \sup_{\|u\|=1} \frac{|u'x - \mu(F_u)|}{\sigma(F_u)} \tag{4.5}$$

  where $u'x$ is the cross product $<u, x>$, F_u is the distribution of $u'X$ and X is a random vector from F.

  The projection depth of $x \in \mathbb{R}^d$ w.r.t F is defined to be:

$$PD(x;F) = 1/(1 + O(x;F)) \tag{4.6}$$

  The projection depth and its associated estimators depend on the choice of $(\mu(F_u),\ \sigma(F_u))$. Different choices of $(\mu(F_u),\ \sigma(F_u))$ provide different estimators relevant to robustness and efficiency. The most commonly used alternatives are $\mu(F_u) = Med(u'X)$, being $Med(u'X)$ the univariate median, and $\sigma(F_u) = MAD(u'X)$, where $MAD(u'X) = Med(|u'X - Med(u'X)|)$ the median absolute deviation is [133].

- **Zonoid depth** [51] at $x \in \mathbb{R}^d$ w.r.t $\{x_1, \ldots, x_n\}$ is defined to be:

$$ZD(x; x_1, \ldots, x_n) = sup\{\alpha : x \in D_\alpha(x_1, \ldots, x_n)\} \tag{4.7}$$

  where $D_\alpha(x_1, \ldots, x_n) = \left\{ \sum_{i=1}^{n} \lambda_i x_i : \sum_{i=1}^{n} \lambda_i = 1,\ 0 \le \lambda_i,\ \alpha\lambda_i \le \frac{1}{n}\ \ \forall i \right\}$.

- **Spatial depth** [29] at $x \in \mathbb{R}^d$ w.r.t F is defined to be:

$$SPD(x;F) = 1 - \|E_F S(x - X)\| \tag{4.8}$$

  where $S(x) = x/\|x\|$ is the spatial sign function ($S(0) = 0$) with Euclidean norm $\|.\|$.

Given a notion of data depth, one can compute the depth value associated with all the sample points $x_1, \cdots, x_n$ and order them in decreasing order,

getting a center-outward ranking of the sample points. A larger rank will always be related to a more outlying position regarding the data cloud. Thus, this ordering starts at the "middle" sample point and moves outwards in all directions. In [132], Liu et al. defined several parameters to characterize a multivariate distribution in terms of its location, scale, skewness and kurtosis, taking into account the data depth ordering. Specifically, they proposed to estimate the median or center of the underlying distribution as the deepest point (or the average of the deepest points, if there were more than one). In fact, they proved that all deepest points derived from the depths listed above are unbiased estimators of the mean of a multivariate normal distribution.

Computing data depth is a non-trivial problem. When the number of dimensions is higher than three and the sample size is large, only approximations of most depths can be given [197, 105]. Some theoretical notions of statistical depth are presented in [132] and [234]. Recent work has been focused on the projection depth [233, 133].

Nowadays, two R packages compute statistical depths. The **depth** package [78] includes algorithms related to three types of depths: the Tukey or halfspace depth, the simplicial or Liu depth and the Oja depth. The **ExPD2D** package provides an exact computation of bivariate projection depth (although it is no longer available from the CRAN repository).

Several data analysis statistical methods based on data depth have been developed for constructing confidence regions, p-values, quality indices and control charts [177]. The notion of depth in the regression setting is introduced in [176]. In addition, some clustering and classification methods have been developed in recent years, based on the concept of data depth. For instance, in Ref. [33], a comparison between modern classification methods based on support vector machines and on the regression depth method, and classical discriminant analyisis is done. In Ref. [44], a new divisive clustering algorithm based on the statistical spatial depth is proposed. Dutta & Ghosh use the projection depth for robust classification of multivariate points in [50]. In Ref. [124], a new classification method using the zonoid depth is developed. Besides, another type of classification method based on the DD-plot (depth vs. depth plot, see [132]) is introduced in [129]. Other attempts can be found in [136] and [94].

In a straightforward way, an object can be classified to the group where it is deepest, that is to say, according to its maximum depth. The author Rebecka Jörnsten followed this idea in [109] (see also [110]). Her clustering methodology is called *DDclust*. In short, she divides all the observations in

clusters, and assigns to each point z in the data space, the depth value with respect to its cluster. As depth function, she considers the L_1 data depth (see [209]). We will define the L_1 depth in Section 4.2. This depth provides robust representatives of the cluster and it is non-zero outside the convex hull of the data cluster, being therefore meaningful when comparing multiple clusters. The L_1 depth also has a closed form which makes it an efficient building block in complex algorithms [109]. Torrente et al. [203] propose to improve k-means using bootstrap and data depth, as an alternative to [109].

In this chapter, the application of the Tukey, Oja, Mahalanobis, convex hull peeling and L_1 depths to anthropometric data is evaluated. In addition, we propose a new algorithm, called *TDDclust*, which is based on *DDclust*. The *DDclust* algorithm was programmed in R and its code was available from http://www.stat.rutgers.edu/home/rebecka/DDcl/ (however, the link to this page doesn't currently exist as a result of a website redesign). The *TDDclust* algorithm is an extension of *DDclust* where a trimmed approach is incorporated to discard those individuals who might be considered outliers regarding a set of measurements, in line with *trimowa* (see Section 2.3).

The outline of this chapter is as follows: Section 4.2 introduces the foundation of *DDclust* and *TDDclust*. Section 4.3 focuses on the results provided by the above mentioned data depths and by *TDDclust*. Finally, Section 4.4 includes the conclusions of this chapter and future work.

4.2 Background

In this introductory section, the definition of the L_1 multivariate median and its associated statistical depth function are reviewed. Besides, the clustering algorithms *DDclust* and *TDDclust* are presented.

4.2.1 Clustering based on data depth

In [110, 109], R. Jörnsten presented two new methods for clustering and classification based on the concept of data depth, and in particular based on the depth function associated with the L_1 multivariate median of Vardi and Zhang [209]. Data depth based clustering methods had not appeared in the literature before [109].

4.2.2 L_1 multivariate median

First of all, we define the L_1 depth from the L_1 multivariate median, which is defined as the solution of the Weiszfeld problem. For that purpose, we follow [209].

Weiszfeld problem. Consider the problem of minimizing the weighted sum of the Euclidean distances from different points $x_1, \ldots, x_m$ in $\mathbb{R}^d$. Let $\eta_1, \ldots, \eta_m$ be the weights or multiplicities of each x_i and let $C(y)$ be the "cost function":

$$C(y) = \sum_i \eta_i d_i(y) \tag{4.9}$$

where $d_i(y) = \|y - x_i\|$ is the Euclidean distance between y and x_i.

Then, the objective is to find a point $y \in \mathbb{R}^d$ (or a set of points) that minimize the "cost function" $C(y)$, i.e., to find:

$$M = M(x_1, \ldots, x_m; \eta_1, \ldots, \eta_m) = argmin\{C(y) : y \in \mathbb{R}^d\} \tag{4.10}$$

The solution of this problem is the spatial median or L_1 multivariate median (from now on, L_1-MM).

Despite this median being an easy-to-understand concept, computing it poses a challenge, and only numerical or symbolic approximations to the solution of this problem are possible. In this way, it is proposed to calculate an approximation to the L_1-MM using an iterative procedure in which each step produces a more accurate approximation. Procedures of this type can be derived from the fact that the sum of distances to the sample points is a convex function, since the distance to each sample point is convex and the sum of convex functions remains convex. Therefore, procedures that decrease the sum of distances at each step cannot get trapped in a local optimum.

One common approach of this type, called Weiszfeld's algorithm after the work of E. Weiszfeld [219], is a form of iteratively re-weighted least squares. This algorithm defines a set of weights that are inversely proportional to the distances from the current estimate to the samples, and creates a new estimate that is the weighted average of the samples according to these weights. That is,

$$y_{i+1} = \begin{cases} \tilde{T}(y_i) = \left\{ \sum_{x_j \neq y_i} \frac{\eta_j}{\|y_i - x_j\|} \right\}^{-1} \sum_{x_j \neq y_i} \frac{\eta_j x_j}{\|y_i - x_j\|} & if \;\; y_i \notin \{x_1, \ldots, x_m\}, \\ \\ x_k & if \;\; \exists k \in \{1, \ldots, m\} : y_i = x_k. \end{cases}$$

It converges to the L_1-MM for a given initial point, if the algorithm never reaches the set $\{x_k : k = 1, \ldots, m; x_k \neq M\}$. To guarantee the convergence to the L_1-MM from any starting point in $\mathbb{R}^d$, this algorithm has been modified, defining $\forall y_i \in \mathbb{R}^d$ the new y_{i+1} as a weighted average of $\tilde{T}(y_i)$ and y_i. The definition of the weights can be found in [209].

Given a definition of a multivariate median θ and a distribution function F, Vardi et al. [209] defined the corresponding depth function (DD) as:

$$D_{\theta,F}(y) \equiv D_F(y) = 1 - inf\left\{ w \geq 0 : \theta\left(\frac{w\delta_y + F}{1+w} \right) = y \right\} \tag{4.11}$$

where δ_y is a point mass at y. That is, $1 - D_F(y)$ is the amount of probability mass w needed at y to make y the multivariate median of the mixture $(w\delta_y + F)/(1+w)$.

From this definition, Vardi et al. proved that the depth function associated with the L_1-MM is:

$$D(y) = \begin{cases} 1 - \|\bar{e}(y)\| & if \;\; y \notin \{x_1, \ldots, x_m\}, \\ \\ 1 - (\|\bar{e}(y)\| - f_k) & if \;\; y = x_k. \end{cases} \tag{4.12}$$

where $e_i(y) = (y - x_i)/\|y - x_i\|$ (unit vector from y to x_i) and $\bar{e}(y) = \sum_{x_k \neq y} e_i(y) f_i$ (average of the unit vectors from y to all observations), with $f_i = \eta_i \,/\, \sum_{j=1}^{k} \eta_j$ and $\|\bar{e}(y)\|$ is close to 1 if y is close to the edge of the data, close to 0 if y is close to the center.

In simple terms, for a point y and an observation $x_i \neq y$, take the unit vector, pointing in the direction from y to x_i. Then, compute the average of all the unit vectors from y to x_i in the points cloud. Finally, define the data depth as $D(y) = 1 - \|Average\ of\ unit\ vectors\|$ [158]. Because $e_i(y)$ are vectors of unit length for $y \neq x_i$ with $\|\bar{e}(y)\| \leq \sum_{x_k \neq y} f_k \leq 1$, it is verified:

$$0 \leq D(y) \leq 1 \tag{4.13}$$

D is near 0 if the point y is close to edge of the points cloud, and D is near 1 if the point y is close to the center of mass of the points cloud.

4.2.3 Clustering based on L_1 depth: *DDclust*

This clustering algorithm, due to R. Jörnsten [110, 109], iterates between median computations via the modified Weiszfeld algorithm [219] and a Nearest-Neighbor allocation scheme with simulation annealing. In the K-median method, the cluster representatives are multivariate medians. PAM is an approximation of the exact K-median. A Nearest-Neighbor criterion is used to generate a partition, given the K medians. To prevent convergence to a local maximum a standard simulated annealing approach is applied. The clustering criterion function used is the maximization of:

$$C(I_1^K) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I(k)} (1-\lambda) sil_i + \lambda ReD_i \tag{4.14}$$

with respect to a partition $I_1^K = \{I(1), \ldots, I(K)\}$, where:

- $D(y|k)$, the L_1 data depth of each point y regarding to the kth cluster (see eq. (4.12)), is computed as:

$$D(y|k) = 1 - max[0, \|\bar{e}(y|k)\| - f(y|k)] \tag{4.15}$$

  where $e_i(y) = (y - x_i)/\|y - x_i\|$, $\bar{e}(y|k) = \sum_{i \in I(k), y \neq x_k} \eta_i e_i(y) / \sum_{j \in I(k)} \eta_j$, $\eta(y) \sum_i \eta_i I\{y = x_i\}$ and $f(y|k) = \eta(y) / \sum_{i \in I(k)}^{k} \eta_i$.

- The within cluster data depth of observations $x_i : i \in I(k)$ is defined as $D_i^w = D(x_i|k)$ (see eq. (4.15)).

- If $I(l)$ is the nearest cluster of an observation $x_i : i \in I(k)$, the between cluster data depth of x_i is defined as $D_i^b = D(x_i|l)$ (see eq. (4.15)).

- $ReD_i = D_i^w - D_i^b$

  ReD is the difference between the depths with respect to the cluster an observation has been allocated to, and the nearest competing cluster.

- $\lambda \in [0, 1]$ is a parameter that controls the influence the data depth has over the clustering.

- sil_i is the silhouette width of the point i, see Section 2.2.1.

For practical implementation of an algorithm that maximizes $C(I_1^K)$, an iterative procedure is followed. A starting point is generated from PAM. At each iteration, observations x_i are moved to new clusters and a new partition is accepted if $C(\widetilde{I}_1^K) > C(I_1^K)$. The multivariate medians are the cluster representatives in *DDclust*. The *DDclust* algorithm is detailed in Algorithm 7.

4.2.4 Trimmed clustering based on L_1 depth: *TDDclust*

Several authors have developed multidimensional trimmed algorithms based on data depth [76, 46, 232] and trimmed clustering algorithms based on the Euclidean distance between points [74], but as far as we know, this is the first time that a trimmed clustering algorithm based on data depth is proposed in the literature. Following [232], for any $0 < \alpha < \alpha^* = sup_x(DD_F(x)) \leq 1$, the α-th trimmed depth region is:

$$DD_F^\alpha = \{x : DD_F(x) \geq \alpha\}. \tag{4.16}$$

The idea behind our algorithm is to define trimmed regions at each step of the iterative algorithm and to apply the *DDclust* algorithm to the remaining set of observations.

As we did with *trimowa* in Chapter 2, we propose to add a trimmed procedure to *DDclust* because an apparel sizing system aims at covering only the standard population.

The procedure will be analogous to the followed with *DDclust*. First, a starting point is generated from PAM. Then, at each iteration, a proportion α (between 0 and 1) is discarded. Let R be the set of $\lceil n(1-\alpha) \rceil$ non-trimmed

Algorithm 7 DDclust [109]

1. Start with an initial partition I_1^K obtained with PAM. Set $\beta = \beta_{init}$.

2. Compute:

   - The L_1-MM of the K clusters, $y_0(1), \cdots, y_0(K)$.
   - The silhouette widths, $sil_i \ \forall i = 1, \cdots, n$.
   - The relative data depths, $ReD_i \ \forall i = 1, \cdots, n$.
   - The total value of the partition, $C(I_1^K)$.

3. Compute $c_i = (1-\lambda)sil_i + \lambda ReD_i \ \forall i = 1, \cdots, n$. Identify a set of observations $S = \{i : c_i \leq T\}$, where T is a prefixed threshold.

4. For a random subset $E \subset S$, identify the nearest competing clusters. Define the partition with E relocated as $\widetilde{I}_1^K$.

5. Compute the value of the new partition $C(\widetilde{I}_1^K)$.
if $C(\widetilde{I}_1^K) > C(I_1^K)$ **then**
  set $I_1^K \leftarrow \widetilde{I}_1^K$.
else
  if $C(\widetilde{I}_1^K) \leq C(I_1^K)$ **then**
    set $I_1^K \leftarrow \widetilde{I}_1^K$ with probability $Pr(\beta, \Delta(C))$, being b a tuning parameter, and $\Delta(C) = C(\widetilde{I}_1^K) - C(I_1^K)$.
  end if
else
  Keep I_1^K.
end if
Set $S = S_{-E}$ removing the subset E form S.

6. Iterate 4-5 until set S is empty.

7. If no moves were accepted for the last M iterations and $\beta < \infty$, set $\beta = \infty$ and iterate 2-6. If no moves were accepted for the last M iterations and $\beta = \infty$. Otherwise, set $\beta = 2\beta$ and iterate 2-6.

points. Observations $x_i \in R$ are moved to new clusters and a new partition is accepted if $C(\widetilde{I}_1^K) > C(I_1^K)$. In addition, we have fixed a relative stopping criteria equal to 0.01. This extension of *DDclust* is called *TDDclust* and it is detailed in Algorithm 8.

4.3 Depth measures and *TDDclust*

4.3.1 Data and methodology

The database used here is made up of the same women as in *trimowa*, *biclustAnthropom* and *hipamAnthropom*. The data set is divided into twelve bust group sizes according to the bust sizes defined in the *European standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals* [59], see Table 2.5 of Section 2.3.

Two different applications are going to be worked on the same data set. Firstly, we apply several depth functions to one of the predetermined bust groups. We aimed at showing the utility of the depth paradigm to identify prototypes when available groups already exist. The depth measures used are Tukey (both exact and approximate), Oja, Mahalanobis, convex hull peeling and L_1. The Tukey and Oja depths are calculated with **depth**. We programmed the Mahalanobis and convex hull peeling depths. The L_1 depth is computed using the code which was accesible from http://www.stat.rutgers.edu/home/rebecka/DDcl/. Results are shown in Section 4.3.2.

Secondly, we will focus on defining an efficient apparel sizing system. Specifically, we will consider a subset of women of our data set and we will use *TDDclust* to get homogeneous groups, to define the prototypes of each one of them and to identify the disaccommodated women. This application will be developed in Section 4.3.3.

4.3.2 Statistical depth to get prototypes of prefixed sizes

For this first example, the largest bust class is selected: it is the bust class between 86 and 90 cm and contains 1028 women. From the total number of considered anthropometric variables, we choose three: bust and waist

Algorithm 8 TDDclust

1. Start with an initial partition I_1^K obtained with PAM. Set $\beta = \beta_{init}$.

2. Compute:

   - The L_1-MM of the K clusters, $y_0(1), \cdots, y_0(K)$.
   - The silhouette widths, sil_i $\forall i = 1, \cdots, n$.
   - The relative data depths, ReD_i $\forall i = 1, \cdots, n$.
   - The total value of the partition, $C(I_1^K)$.

3. Compute $c_i = (1-\lambda)sil_i + \lambda ReD_i$ $\forall i = 1, \cdots, n$. **Remove** $R = \{i : c_i \leq \alpha\}$**, being** α **the trimming size. Let** R **be the set of** $n(1-\alpha)$ **non-trimmed points.**

4. Identify a set of observations $S = \{i \in R : c_i \leq T\}$, where T is a prefixed threshold.

5. For a random subset $E \subset S$, identify the nearest competing clusters. Define the partition with E relocated as $\widetilde{I}_1^K$.

6. Compute the value of the new partition $C(\widetilde{I}_1^K)$.
if $C(\widetilde{I}_1^K) > C(I_1^K)$ **then**
  set $I_1^K \leftarrow \widetilde{I}_1^K$.
else
  if $C(\widetilde{I}_1^K) \leq C(I_1^K)$ **then**
    set $I_1^K \leftarrow \widetilde{I}_1^K$ with probability $Pr(\beta, \Delta(C))$, being b a tuning parameter, and $\Delta(C) = C(\widetilde{I}_1^K) - C(I_1^K)$.
  end if
else
  Keep I_1^K.
end if
Set $S = S_{-E}$ removing the subset E form S.

7. Iterate 5-6 until set S is empty.

8. $\forall j \in \{1, \cdots, n\, x_j \in R\}$ compute $k_j = argmax\{c_j^k\}$ being c_j^k the value of c_j as in eq. (4.14), assuming that the j-th point belongs to cluster k. Assign x_j to the k_j-th cluster.

9. If no moves were accepted for the last M iterations and $\beta < \infty$, set $\beta = \infty$ and iterate 2-8. If no moves were accepted for the last M iterations and $\beta = \infty$. Otherwise, set $\beta = 2\beta$ and iterate 2-8.

circumference and neck to ground length. In short, our training database is made up of 1028 women and 3 dimensions.

The Tukey (both exact and approximate), Oja, Mahalanobis, convex hull peeling and L_1 depths identify the same most centered woman: ABAD101. The convex hull peeling depth also identifies MALAG103 as the deepest woman, with the same depth value as ABAD101. The rest of depth measures considered MALAG103 as the second deepest woman. It is worth pointing out that Liu's depth can be only calculated on bivariate datasets, therefore it is not used in this example. The Oja depth is computationally very costly. It lasted roughly 30 hours to provide results. Tukey, Mahalanobis and the convex hull peeling depths took only a few seconds, while L_1 depth took just 5 minutes.

Table 4.1 shows the anthropometric dimensions and Fig. 4.1 shows the body shape of these two particular women. We see that they have very similar body measurements.

Dimension / Woman	Neck to ground	Waist	Bust
ABAD101	137.4	76.6	88
MALAG103	137.2	76.4	87.9

Table 4.1: Body measurements of the most centered (deepest) women, according to the considered depth measures: Tukey, Oja, Mahalanobis, the convex hull peeling and L_1 data depths, applied to the $[86, 90[$ cm bust class.

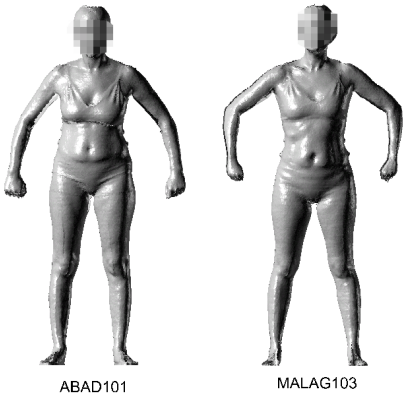


Figure 4.1: Front body shape of deepest women ABAD101 and MALAG103 (left to right).

Fig. 4.2 represents a three-dimensional scatterplot where ABAD101 and MALAG103 are marked with a star. They are clearly located at the center of the points cloud.

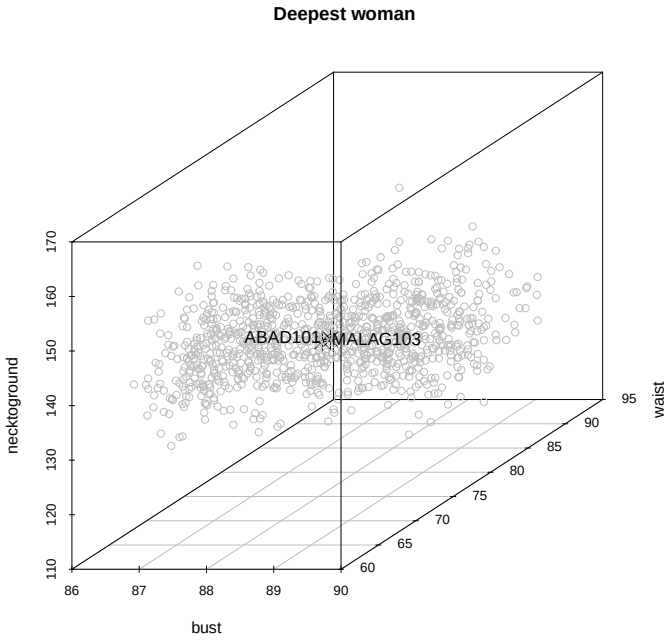


Figure 4.2: 3D scatterplot of the $[86, 90[$ cm bust class. The most centered (deepest) women, ABAD101 and MALAG103, are marked with a star.

Fig. 4.3 shows a density plot for the Tukey depth with the percentile lines incorporated. It serves to examine how the depth scores are distributed. A right-skeweded distribution can be appreciated, where the right tail is longer and the mass of the distribution is concentrated on the left of the figure (percentiles 10, 25 and 50 are quite close).

4.3.3 Statistical depth to get an efficient apparel sizing system

In order to illustrate this methodology and to obtain results in a relative short period of time, we looked now for a small sized data set, selecting to work with the second bust segment, ranging between 78 and 82 cm. It contains 287 women.

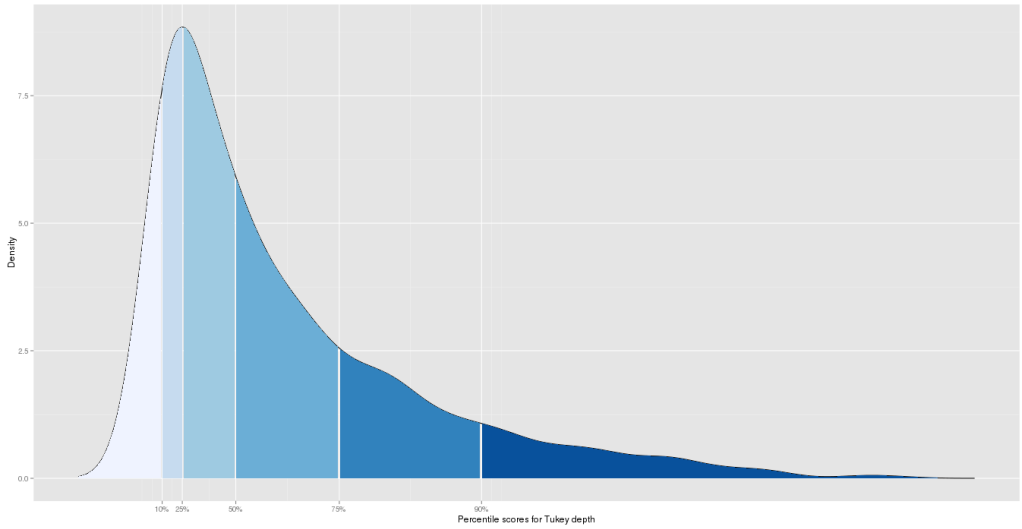


Figure 4.3: Percentile lines in a density plot for the Tukey depth applied to the $[86, 90[$ cm bust class.

In case of larger samples, this algorithm takes much longer to complete. The three same variables (neck to ground, waist and bust) as the previous example were selected. Before applying *TDDclust*, these variables were standardized to make sure all variables contribute evenly to the clustering. We did this because we appreciated that the neck to ground dimension dominated the segmentation. The number of random initializations was 5 and the number of clusters, 3. In this case, the trimmed proportion was prefixed to 0.1, therefore, the accommodation rate is 90%. The λ parameter was set to 0.5 to weight between the average silhouette width and average relative data depth.

Table 4.2 shows the clustering results. It can be observed that clusters are quite balanced.

Cluster 1	Cluster 2	Cluster 3
93	88	77

Table 4.2: Number of women in each cluster for the bust group $[78, 82[$ cm after applying the *TDDclust* algorithm.

Fig. 4.4 displays a 3D scatterplot of the returned clusters together with

the discarded individuals.

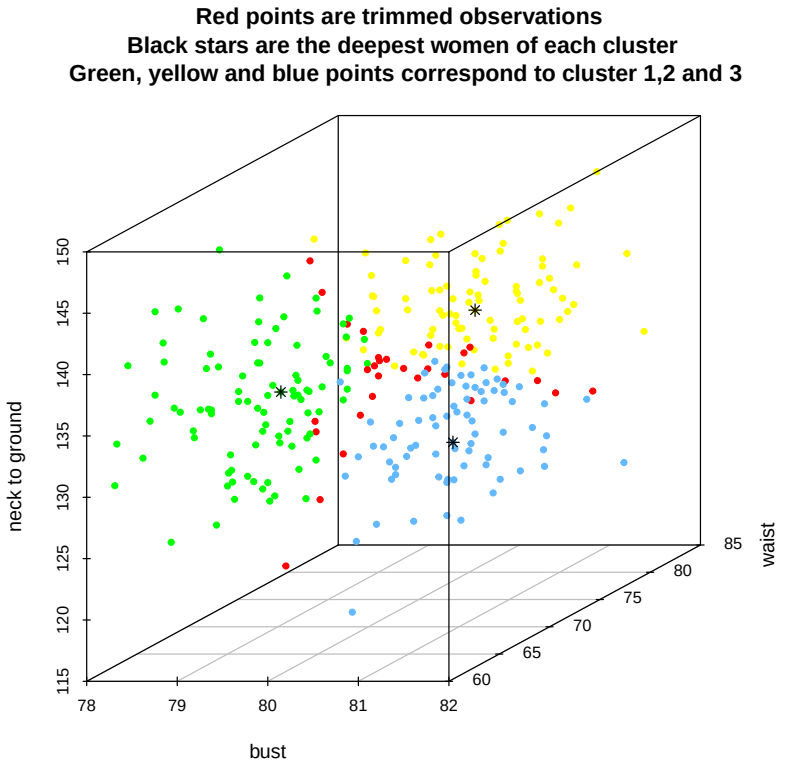


Figure 4.4: 3D scatterplot of the clusters obtained for the $[78, 82[$ cm bust class by applying the *TDDclust* algorithm.

We see very interesting results. The three clusters are clearly separated. Cluster 2 (yellow points) is related to largest waist and neck to ground. Regarding cluster 3 (blue points), it also contains women with large waist circumferences but they are not so tall. In cluster 1 (green points) predominates small bust values. Finally, and most important, the trimmed women (red points) are located along the border among clusters, that is to say, they are indeed extreme women for every cluster. Black stars are the deepest women of each cluster. Fig. 4.5 shows the clustering results for every variable combination. As expected because the data preprocessing, no variable seems to be more important than other one in constructing the segmentation solution.

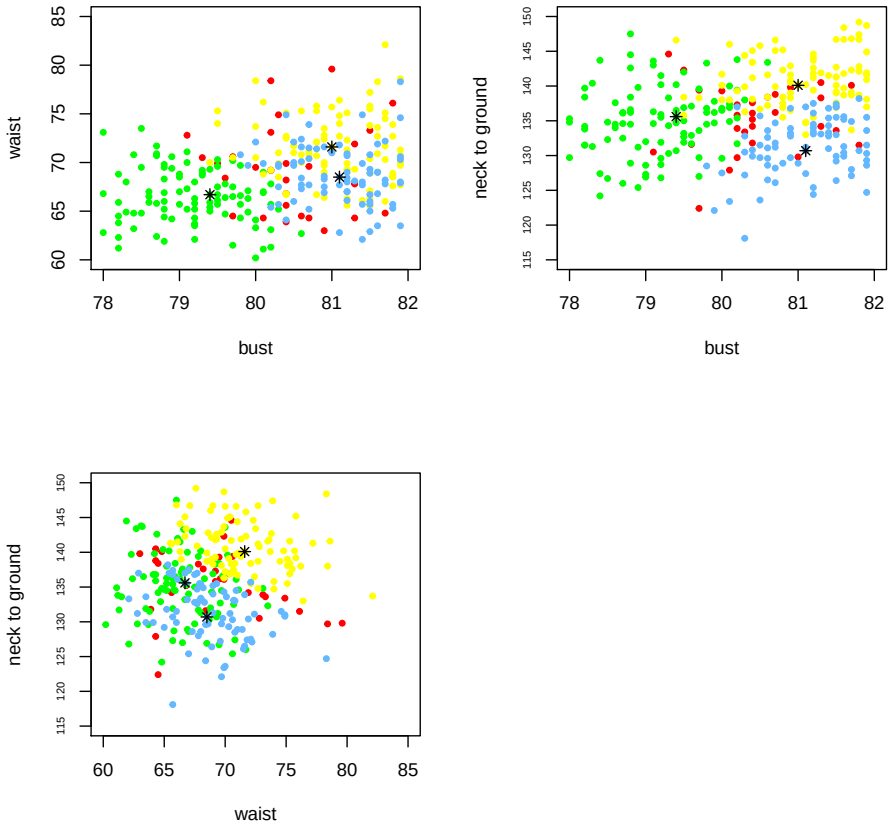


Figure 4.5: Separated plots displaying the clusters obtained for the $[78, 82[$ cm bust class by applying the *TDDclust* algorithm.

4.3.4 Summary

Linear order induces an ordering for 1-dimensional points. Median is the most centered point in the data set. On the contrary, there is no natural order for dimensions $d \geq 2$. As compensation, it is suitable to orient to a "center" (the so-called the deepest point), which corresponds to the multivariate median. This involves a center-outward ordering of observations. Statistical data depth allows the univariate median to be generalized with respect to a multivariate probability distribution or a data cloud.

The definition of optimal sizes involves the identification of one or even

several centered individuals, that is, representative prototypes of the body size groups. Data depth procedures could be very useful to that purpose since data depth measures the centrality of a point in the data. We have validated the utility of this approach from two different perspectives. The first attempt is aimed at identifying prototypes when available size groups already exist. We have used different well-known data depths and the computational cost of some of them is highlighted.

However, when the goal is searching for optimal size groups of individuals with similar body dimensions, a clustering algorithm should be used, in line with *trimowa*. Some clustering methodologies based on data depth have been recently developed. One of them, the *DDclust* algorithm, is proposed in [109], where the L_1 depth is used. The idea behind [109] is to classify a point to the class where it has the largest depth value. We have proposed an extension of *DDclust*, called *TDDclust*, where a trimmed procedure is incorporated. The bust classes defined by the current European Standard to sizing system are used as data sets. As a preliminar investigation, we have only applied *TDDclust* with small bust sizes. According to the results obtained, we have reached the conclusion that *TDDclust* provides promising results in terms of sizing segmentation and the identification of disaccommodated women. In this way, it could be used with the rest of bust sizes, although computations will take for a long time. Anyway, we think that this type of algorithm should be used with small samples, such as "special sizes" (small or larger sizes).

It is well-known that the specifical depths used in this approach, as well as other classical ones, have serious issues regarding the dimensionality of the database used in the analysis. Their routines have a high computational cost when the number of rows or the number of colums or both increases. A major challenge in this field is to develop computationally efficient algorithms to calculate depths when the dimensionality is large. A convenient depth could be the projection depth [231].

4.4 Chapter conclusions

In this chapter we have proposed the use of another statistical approach, the statistical data depth, to find central individuals oriented at apparel design. This work can be considered as a preliminar approach to investigate the use of data depth with anthropometric data. Data depth provides an alternative way to define efficient sizes because allows the observations near the middle

of the distribution (the deepest observations) to be identified, from which designing the garments of the corresponding size. Data depth represents an alternative approach to find the "center" of multivariate data sets and in addition, it is useful and robust for clustering.

As a future work, we aim at applying the *TDDclust* methodology to a combination of different bust sizes in order to check whether *TDDclust* is able to divide the population according to the bust variable. We could compare its results with the segmentation proposed by the European Standard, by computing for example, the success rate of people belonging to the corresponding group. Further investigations should be also done for combining clustering and data depth to develop efficient, computationally tractable algorithms and even combining biclustering and data depth measures, following [229].

Chapter 5

Archetypal analysis

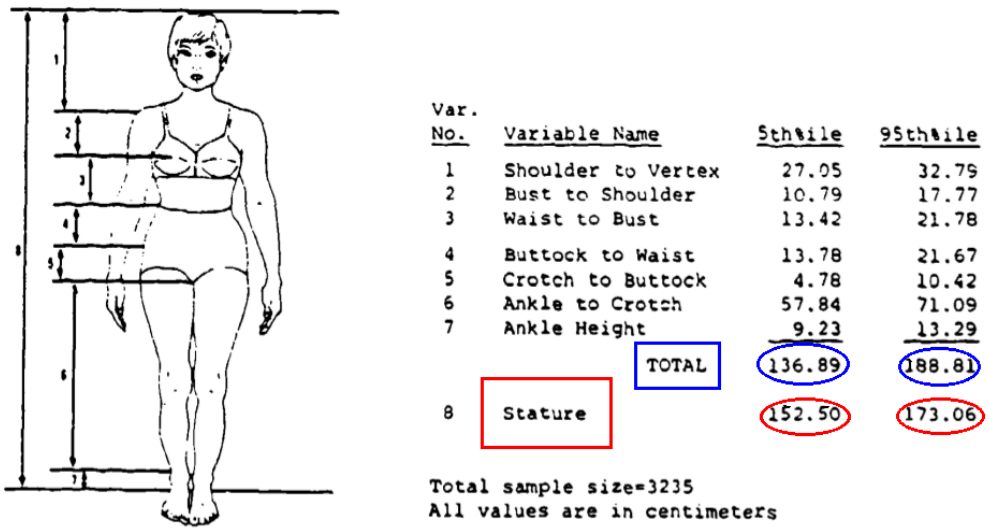
5.1 Introduction [1]

In the multivariate accommodation problem in human modelling and ergonomic design, it is fundamental to find representative individuals in a database using a few observations. These few observations are the extreme individuals (boundaries) of a sample [14]. The percentage of accommodation, that is to say, the percentage of excluded people, is a major issue to determine. The typical required percentage of accommodation is between 90% and 95%. The most commonly used statistical approaches in this field have been percentile analysis, regression analysis and principal component analysis (PCA).

Percentile analysis can be considered as the traditional procedure used to accommodate a portion of the population [228]. A percentile is a very simple statistic that says the percentage of people who are smaller than a given individual for a single measurement. For example, if we wanted to accommodate the 90% of the population, we could define the percentiles 5 and 95. However, percentiles are only relevant from a univariate point of view. This means that percentiles calculated on one dimension tell us nothing about the variability of other dimensions involved in the study. Furthermore, they are not additive [151, 228, 172]. Fig. 5.1, which has been originally taken from [172], helps to understand this fact: The stature of the individuals of a population is divided into seven measurements and the percentile 5 value for each one of them is calculated. The sum of them is equal to 136.89. However,

[1] Section 5.3 is published in [54] and Section 5.4 is submitted for publication[213].

the actual percentile 5 of the stature for the whole population is 152.5. The same thing occurs with the percentile 95 ($188.81 \neq 173.06$). This problem is particulary relevant when trying to use anthropometric data to develop human body models.



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

Total sample size=3235
All values are in centimeters

Figure 5.1: Image taken of [172] that illustrates the reason why percentiles are not additive.

An alternative to percentile analysis has been regression analysis [172, 228, 64, 144]. This method selects one or two key measurements and predicts values for other dimensions. Its main advantage regarding percentiles is that the predicted values are additive. Specifically, this means that if we used stature and weight in a regression equation to predict the seven derived variables explained in Fig. 5.1, the resulting values would add up to exactly the value of stature. On the contrary, the big drawback is that it only provides average quantities for the predicted dimensions. This is not interesting when the intention is to look for extreme patterns.

Today, the most used strategy is based on the principal component analysis (PCA) [228, 16, 83, 71, 99, 173]. PCA works as a data reduction technique by considering only the first two or three components that explain as much variability as possible. Then, several extreme points are selected in a probability ellipse (or in a circle if the scores are standardized) that includes any

desired percentage of the population (see Fig. 5.2). However, this method also presents several pitfalls, as rightly pointed out in [70]. Because it only chooses the first PCs, a portion of the data variation is eliminated, but this variation may represent cases difficult to accommodate. Consequently, when building the ellipse (or the circle), the covered level of accommodation is not really the desired percentage (for example, 95%). Therefore, an improved version of this approach would require the use of a greater number of components (even all). However, the more PCs there are, the more cases there will be. For the first two PCs, eight cases are selected and for the first three PCs, fourteen (see Figs. 5.2 and 5.3). That is, if we were interested in representing more than three components to consider more variation, the number of cases would be too many. In practice, it would not be useful.

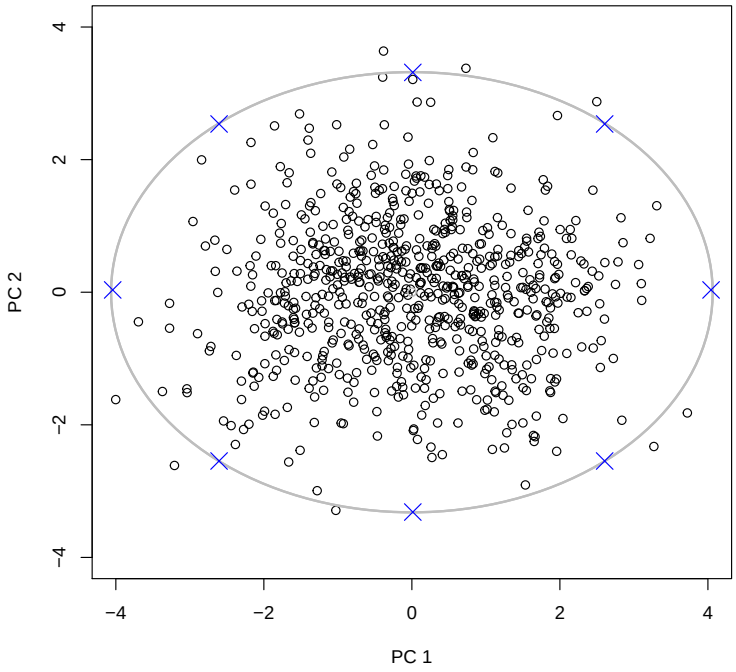


Figure 5.2: PCA procedure for the accommodation problem with two PCs. The first two PCs are considered and the eight most extreme cases are selected (marked with blue crosses).

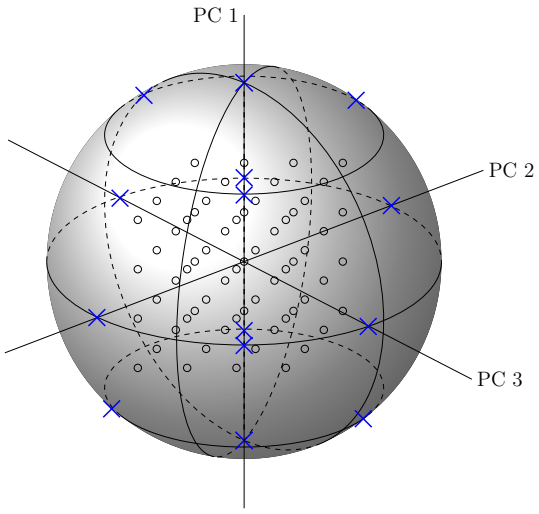


Figure 5.3: PCA procedure for the accommodation problem with three PCs. The first three PCs are considered and the fourteen most extreme cases are selected (marked with blue crosses).

Because none of these procedures correctly identifies the extreme cases in data that covers a certain portion of the population, a different statistical approach is proposed in this PhD dissertation: archetypal analysis (from now on AA). AA was proposed in [37]. It aims at finding some pure types (the archetypes) in such a way the rest of observations are a mix of them. They can be easily computed by means of the **archetypes** R package [56]. This method was demonstrated useful with head-dimension data, air-pollution data and for tracking spatio-temporal dynamics. In recent years, AA has been used in different fields such as market research and benchmarking [130, 163, 148], the evaluation of scientists [184], e-learning [200], the analysis of astronomy spectra [28, 167], face recognition in vision problems [226], sports [55], biology [40], multi-document summarization [25, 26] or different machine learning problems [152, 195]. This is the first approximation that uses AA with anthropometric data. We emphasize that since appearing **archetypes**, AA has increased its activity and the number of applications is even growing faster.

In this chapter, two methodologies based on AA to tackle the accommodation problem are presented. In a first approach, the advantages of AA regarding the currently most used PCA-based procedure will be shown. In the second approach, a new archetypal concept will be introduced: the archety-

poid. The outline of this chapter is as follows: Section 5.2 introduces the archetypal and archetypoid analysis and provides some theoretical properties of archetypoids. Sections 5.3 and 5.4 give the details, results and a comprehensive summary of the two methodologies proposed. Finally, the most important conclusions of this chapter are given in Section 5.5.

5.2 Background

This introductory section gives an overview of archetypal and archetypoid analysis. In addition, it presents and discusses some theoretical aspects of archetypoids.

5.2.1 Archetypal analysis

We begin with an $n \times m$ matrix X that represents a multivariate data set with n observations and m variables. Given a number of archetypes equal to k, the goal of AA is to find a $k \times m$ matrix Z that characterizes the archetypal patterns in the data, such that data can be approximated by convex combinations of the archetypes. In other words, AA is aimed at obtaining the two $n \times k$ coefficient matrices α and β which minimize the following residual sum of squares, resulting from the combination of (i) the equation that shows x_i as being approximated by a convex combination of z_j's (archetypes) and (ii) the equation that shows z_j's as convex combinations of the data:

$$\left.\begin{array}{ll} (i) & \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 \\ (ii) & \mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l \end{array}\right\} \Rightarrow \begin{array}{l} RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 \\ \qquad = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2 \end{array}$$

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 $j = 1, \ldots, k$

Constraint 1) implies that $\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$. This means that the predictors of $\mathbf{x}_i$ are finite mixtures of the archetypes (not confusing with a mixture of distributions). Each α_{ij} is the weight of the archetype j for the individual i, i.e., the α coefficients represent how much each archetype contributes to the approximation of each individual.

Constraint 2) implies that $\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$, i.e., the archetypes are finite mixtures of the observations.

Archetypes are located on the convex hull of the data, except when $k = 1$, where the obtained archetype is the sample mean.

5.2.2 Archetypoid analysis

Archetypes do not have to be exactly sampled individuals. However, in some circumstances it is crucial that they are. To fill this gap, a novel archetypal concept is presented: the archetypoid.

Archetypes would correspond to specific individuals when $\mathbf{z}_j = \mathbf{x}_l$ for any $l = 1, \ldots, n$, which is the same as saying that only one β_{jl} is equal to 1 in the previous constraint 2) of the AA problem, for each j. This implies that β_{jl} should only take on the value 0 or 1. This line of reasoning leads to the assumption that in the analysis of archetypoids, the original AA optimization problem becomes:

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2 \tag{5.1}$$

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 $\boxed{\beta_{jl} \in \{0, 1\}}$ and $j = 1, \ldots, k$. Hence, $\beta_{jl} = 1$ for one and only one l and $\beta_{jl} = 0$ otherwise.

5.2.3 Location of the archetypoids

Let $Conv(\mathbf{X})$ be the convex hull of the n observations in $\mathbb{R}^m$ of the set $\mathbf{X}$. The convex hull of $\mathbf{X}$ in the Euclidean space is the smallest convex set that contains $\mathbf{X}$. As the number of points in $\mathbf{X}$ is finite, $Conv(\mathbf{X})$ is a convex polytope, which is the convex hull of its vertices. A vertex of $Conv(\mathbf{X})$ is an observation $\mathbf{x}_i$ of $\mathbf{X}$ for which $\mathbf{x}_i$ does not belong to $Conv(\mathbf{X}\backslash\{\mathbf{x}_i\})$. A vertex of $Conv(\mathbf{X})$ is also called an extremal point of $\mathbf{X}$. Let $\mathbf{V}$ be the set of vertices of $Conv(\mathbf{X})$ and N be the number of vertices.

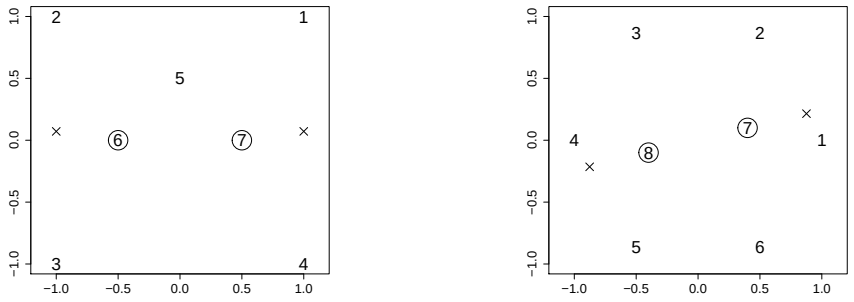
Next, we investigate where the archetypoids are located and the differences with the archetype locations for different values of k.

1. If $k = 1$, the archetypoid is the medoid (with one cluster) of $\mathbf{X}$ considering the squared Euclidean distance as dissimilarity, since the minimization of RSS coincides with the definition of the medoid [115]. As said, the mean is the archetype with $k = 1$.

2. If $k = N$ (or $> N$), the archetypoids are $\mathbf{V}$ (or $\mathbf{V}$ plus any other observation), as $RSS = 0$, since $Conv(\mathbf{V}) = Conv(\mathbf{X})$.

3. If $1 < k < N$, it is not possible to say as true that archetypoids are on the boundary of $Conv(\mathbf{X})$, as archetypes are. This is discussed in the following artificial Example 1. It depends on the distribution of the observations. However, for Normal distributions, archetypoids seem to be vertices, as it can be seen in the Example 2, where the examples of [92, Fig. 14.35] and [37, Fig. 14] are reproduced.

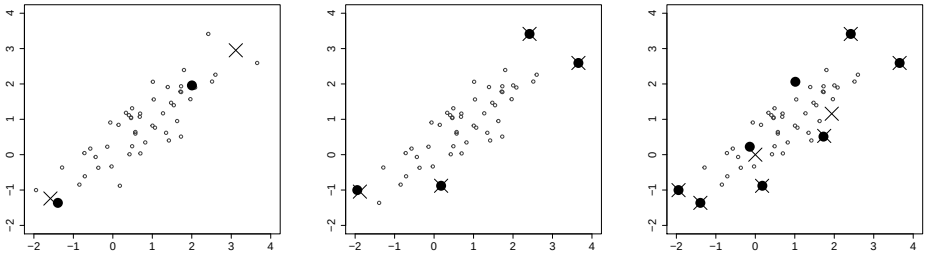
Example 1 Fig. 5.4a shows the location of 7 points in $\mathbb{R}^2$. Archetypes and archetypoids are calculated for $k = 2$ (note that with $k = 4$ the $Conv(\mathbf{X})$ is the square formed by vertices 1, 2, 3 and 4, and these are the archetypes and archetypoids). In this example, archetypoids are not vertices. In addition, they coincide with the closest points to archetypes. If we compute the RSS for archetypes and archetypoids, the elbow is at $k = 4$ (see Section 5.3.2 for details on the elbow criterion). Fig. 5.4b shows a second artificial example with $k = 2$. In this case, the nearest points to the archetypes are 1 and 4 but the archetypoids are 7 and 8 (not vertices either).



(a) (b)

Figure 5.4: Two examples where the archetypoids (circles) are not on the boundary of $Conv(\mathbf{X})$ as the archetypes (crosses) are.

Example 2 We generate a sample of size 50 from N(μ,$\mathbf{\Sigma}$), where $\mu = \binom{1}{1}$ and $\mathbf{\Sigma} = \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)$ and we calculate the archetypes and archetypoids for $k = 2$ (Fig. 5.5a), $k = 4$ (Fig. 5.5b) and $k = 8$ (Fig. 5.5c). N is equal to 7 in this example. Note that like archetypes, the archetypoids do not nest (as more archetypoids are found, the existing ones can change to better capture the shape of the data set).

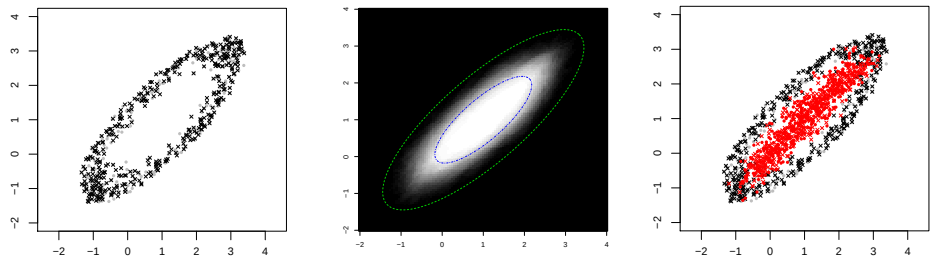


(a) (b) (c)

Figure 5.5: Archetypes (with crosses) and archetypoids (with solid circles) for simulated Bivariate Normal Data, with $k = 2$ (a), $k = 4$ (b) and $k = 8$ (c).

We have repeated the experiment 100 times and have now discarded any points outside the 95% density contour, that is, with Mahalanobis distance

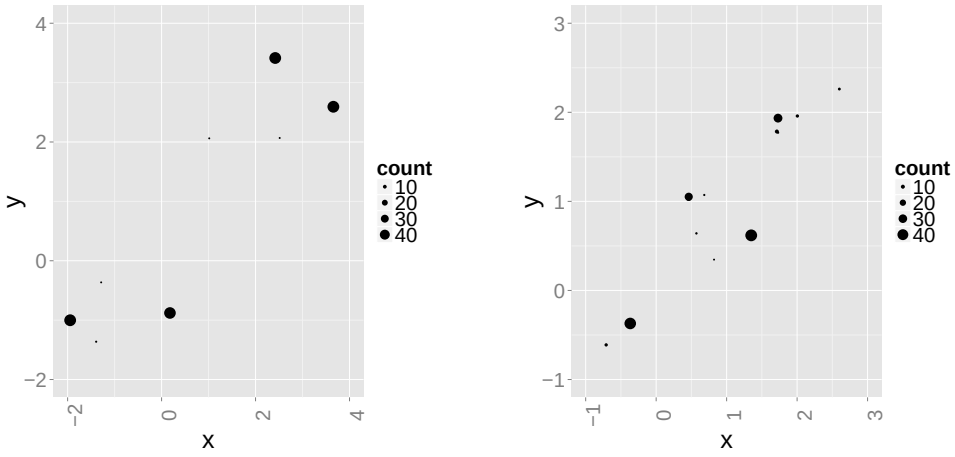
$\geq \chi^2_2(0.95)$, computing the archetypes and archetypoids with $k = 4$, as in [37] was done for archetypes. Fig. 5.6a represents the results. It can be appreciated that like archetypes, the archetypoids cluster around the ends of the major and minor axes of the 95% density contour. In addition, we have overlaid the 100 convex hulls of the 4 archetypoids and the resulting image can be seen in Fig. 5.6b. The whiter the image, the larger the overlap. In this image, the ellipse corresponding to the 95% and 50% probability regions for the bivariate Normal distribution are also added to better understand the results. Fig. 5.6c shows the centroids obtained with k-means with $k = 4$ and the four medoids obtained with PAM working with the same data used for obtaining Fig. 5.6a. In this Fig. 5.6c, we have again represented the archetypes and archetypoids to clearly show that the clustering algorithms choose the prototypes in the middle of the data cloud, not on the convex hull of the data.



(a) (b) (c)

Figure 5.6: (a) Location of the 4 archetypes (black crosses) and archetypoids (grey solid circles) for 100 simulated Bivariate Normal Data; (b) their overlapped convex hulls with contour lines for 95% (green dashed line) and 50% (blue dot-dashed line) probabilities and (c) the location of the 4 centroids (red crosses) and 4 medoids (red solid circles).

The last theoretical question we want to discuss is regarding stability, that is to say, if the solution does not change much when the data are slightly modified. For this purpose, we consider the data of Fig. 5.5b and leave out one point. Then, we compute the archetypoids and medoids with $k = 4$ and save the number of times that each point appears as archetypoid (Fig. 5.7a) or medoid (Fig. 5.7b). We can see that archetypoids are very stable. Three



(a) (b)

Figure 5.7: Frequencies of the points obtained when one point is left out for the simulated Bivariate Normal data, for 4 archetypoids (a) and 4 medoids (b).

points are archetypoids 49 times (every time except the one that is left out), and the other point is an archetypoid 48 times. The frequencies of the most repeated medoids are 44, 43, 31 and 28.

5.2.4 Comparison with other unsupervised methods

Using the same matrix notation as in [152], the goal of archetypoid analysis is to find the optimal matrices $\mathbf{C}$ and $\mathbf{S}$ that minimize some distortion measure $D(\mathbf{X}'|\mathbf{X}'\mathbf{C}\mathbf{S})$ (for example, $\|\mathbf{X}' - \mathbf{X}'\mathbf{C}\mathbf{S}\|_2$ or $\|\mathbf{X}' - \mathbf{X}'\mathbf{C}\mathbf{S}\|_F^2$), where $\mathbf{C} = \beta$ and $\mathbf{S} = \alpha'$ and $'$ denotes transpose. As an extension of [152], Table 5.1 shows the relationship between archetypoid analysis and different unsupervised methods (seen as a linear mixture type representation of data with various constraints) in terms of possible values of $\mathbf{C}$ and $\mathbf{S}$ (note that $\mathbf{X}'\mathbf{C}$ are the feature vectors, while $\mathbf{S}$ gives the weights for the predictors of $\mathbf{X}$).

Other authors have previously compared archetypal representation with other unsupervised methods. For instance, in [92, Sec. 14.6.1] a comparison among AA, k-means and Non-negative matrix factorization (NMF) was made and also AA, PCA and independent component analysis (ICA) were applied to the same database. In addition, Mørup et al. [152] analyzed several

databases with AA, PCA, NMF, ICA and k-means. Finally, in [25] AA, PCA, NMF and k-means and other multi-document summarization methods were compared.

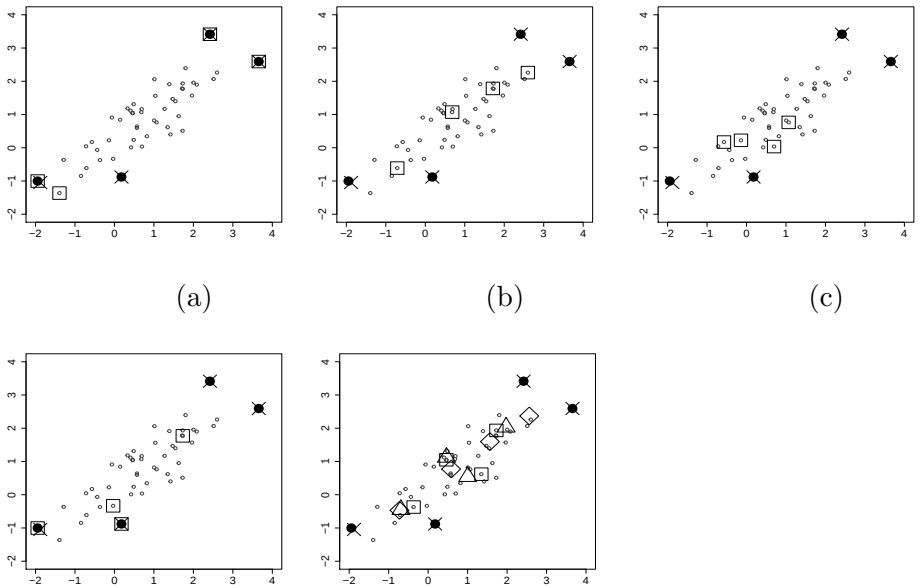
PCA	$\mathbf{C} \in \mathbb{R}$ $\mathbf{S} \in \mathbb{R}$
NMF	$\mathbf{X}'\mathbf{C} \geq 0$ $\mathbf{S} \geq 0$
CNMF	$\mathbf{C} \geq 0$ $\mathbf{S} \geq 0$
AA	$\lvert\mathbf{c}_k\rvert_1 = 1, \mathbf{C} \geq 0$ $\lvert\mathbf{s}_n\rvert_1 = 1, \mathbf{S} \geq 0$
ADA	$\lvert\mathbf{c}_k\rvert_1 = 1, \mathbf{C} \in \mathbb{B}$ $\lvert\mathbf{s}_n\rvert_1 = 1, \mathbf{S} \geq 0$
Soft k-means	$c_{k,n} = \frac{s_{k,n}}{\sum_{\tilde{n}} s_{k,\tilde{n}}}$ $\lvert\mathbf{s}_n\rvert_1 = 1, \mathbf{S} \geq 0$
k-means	$\lvert\mathbf{c}_k\rvert_1 = 1, \mathbf{C} \geq 0$ $\lvert\mathbf{s}_n\rvert_1 = 1, \mathbf{S} \in \mathbb{B}$
k-medoids	$\lvert\mathbf{c}_k\rvert_1 = 1, \mathbf{C} \in \mathbb{B}$ $\lvert\mathbf{s}_n\rvert_1 = 1, \mathbf{S} \in \mathbb{B}$

Table 5.1: Relationship between archetypoid analysis and several unsupervised methods as in [152]: Principal component analysis (PCA), Non-negative matrix factorization (NMF), Convex NMF (CNMF), Archetype analysis (AA), Archetypoid analysis (ADA), Soft k-means (i.e. fuzzy k-means or the EM-algorithm for clustering), k-means and k-medoids. $\mathbb{B}$ represents the set $\{0, 1\}$.

The foundation of AA and clustering is different. The main difference is that AA favors features that represent *corners* of the data, i.e., the extremes in the data or archetypes, while traditional clustering algorithms, like k-means or PAM, segments subjects based on centroids (averages) or medoids and focuses on the memberships in each cluster. A simple example is seeking two profiles of height. The two archetypoids will be the tallest and the smallest persons in the sample. However, with clustering techniques the profiles will be inside the data, likely the profiles will be near the men's and

women's average height.

For practical purposes, we will use the same data as in Fig. 5.5b in order to better understand the differences between some different methodologies for obtaining representative data (most of them are clustering methods), that we detail next, and AA. Fig. 5.8 shows the representatives for these above mentioned different methodologies. Specifically, we have used: a) the Sparse Modeling Representative Selection method (SMRS) developed in [52]; b) the Affinity Propagation algorithm (AP) explained in [69]; c) the HottTopixx [17] (a new approach for NMF, using the code developed in [81]); d) a Bayesian partial membership model (BPM) [73, 150] (we have represented the points with the highest membership in each group) and e) PAM, k-means and fuzzy k-means.



(a) (b) (c)

(d) (e)

Figure 5.8: Archetypes (marked with crosses) and archetypoids (with solid circles) for simulated Bivariate Normal Data, with k =4, together with the four representatives (with squares) provided by the following methods: (a) SMRS, (b) AP, (c) HottTopixx, (d) BPM and (e) classical clustering algorithms (PAM with squares, k-means with triangles and fuzzy k-means with diamonds).

As we see in the corresponding plots of Fig. 5.8, all methods excepting SMRS, return representatives around the middle of the data, rather than in the boundaries. However, regarding SMRS, each point in the data set is approximated by an affine combination of the representatives, meaning that the coefficients can be negative (in fact, for this example several coefficients are negative, the maximum value of the coefficients is 0.669, only 6 coefficients are above 0.5 and the majority of non-zero values are between 0.2 and 0.3). On the contrary, in archetypoid analysis the points are approximated by a mixture of archetypoids and the coefficients therefore add up to one and are positive. This SMRS behavior makes difficult the intuitive interpretation of its results. Furthermore, with SMRS it is not possible to select exactly how many representatives have to be obtained. In this example, SMRS only was able to return two representatives because the other two representatives were below a certain threshold. In fact, without considering the threshold, only four representatives were extracted in total. We could not have obtained five or more representatives for these particular data with the SMRS algorithm.

5.3 First methodology: AA vs PCA

The objective of this methodology is to compare the performance of both AA and the common used PCA-approach when obtaining a set of representative extreme cases to achieve body size accommodation for a specific portion of the population. We have used the **archetypes** R package. Section 5.3.1 describes the data set and the methodology used. Section 5.3.2 shows the results obtained. Finally, Section 5.3.3 gives a comprehensive summary of this approach.

5.3.1 Methodology

The anthropometric database we use here comes from the 1967 United States Air Force (USAF) Survey. It can be freely downloaded from the website http://www.dtic.mil/dtic/. The 1967 USAF Survey was undertaken from January to March 1967, planned and supervised by the Anthropology Branch of the Aerospace Medical Research Laboratory, located in Ohio. The target sample was made up of 2420 Air Force personnel between 21 and 50 years of age. The measurements were made in 17 different Air Force bases across the United States of America. A total of 202 variables (including body dimen-

sions and background variables) were reported from all the participants in the survey. From those 202 variables, we choose the same six anthropometric dimensions selected in [228]. They are called *cockpit dimensions* because they are the most important dimensions to design and build an aircraft cockpit. A description of each one of them, according to [119], can be found in Table 5.2. Table 5.3 shows their summary statistics. Fig. 5.9 displays the skeleton of an aircraft pilot with the six selected measurements detailed.

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

Table 5.2: Description of the six variables considered.

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

Table 5.3: Summary statistics for the six variables considered.

We are going to calculate the archetypes with a 95% accommodated. We propose to follow these steps: First, depending on the problem, it must be decided whether or not the data should be standardized. We does standardize our variables (as in [228]) since they measure different dimensions. Second, the more extreme 5% data (because the percentage of accommodation is fixed to 95%) must be removed. If we suppose normality, we can use the Mahalanobis distance and Chi-square distribution. If not, a non-parametric approach such as a depth procedure might be used.

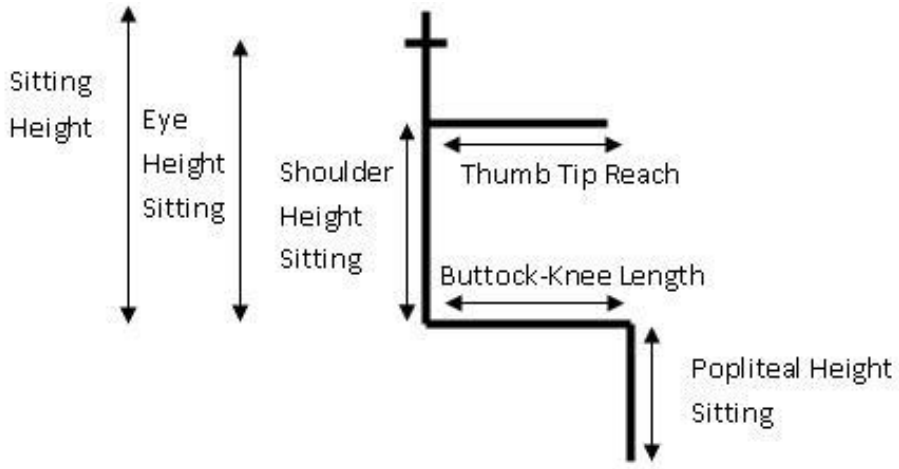


Figure 5.9: Generic skeleton for an aircraft pilot with explanations of the cockpit dimensions.

Because the PCA approach assumes normality to draw the normality ellipse or circle (it could be done because we are dealing with anthropometric measurements) we will also assume normality. Specifically, assuming that the data comes from a multivariate (m-variate) normal, we can use the fact that the Mahalanobis distance from an observation to the mean, $D^2 = (x - \hat{\mu})'\hat{\Sigma}^{-1}(x - \hat{\mu})$, where $\hat{\mu}$ is the estimated mean and $\hat{\Sigma}$ is the estimated covariance matrix, is distributed according to the Chi-square distribution with m degrees of freedom. Consequently, the observations that are more far away from the 95th percentil of the Chi-square distribution are discarded. We would like to point out that very similar results were obtained when removing the disaccommodated individuals using both the Mahalanobis distance and depth procedure for this database, although depth presented a relevant disadvantage: the desired percentage is not under control of the analyst as with the Mahalanobis procedure. For example, there was almost a 7% of less deep data in the USAF Survey ($169/2420 = 0.0698$), each one of them with the same depth. Third and last, once the more extreme 5% data are removed, AA is applied to calculate the archetypes.

5.3.2 Results

We computed 10 archetypes, $k = 1, \ldots, 10$. In fact, we compute three times the archetypes for each k and we keep the archetypes with the smallest RSS

for each k. In Fig. 5.10 we represent with a set of six bars (one per variable, from dark gray to light gray) the percentile value of each archetype for each variable, from $k = 2$ (a) to $k = 10$ (j). Then, a simple analysis of the archetypes can be done. For instance, we see in Fig. 5.10a that the first archetype is low in all variables, whereas the second archetype is high in all of them.

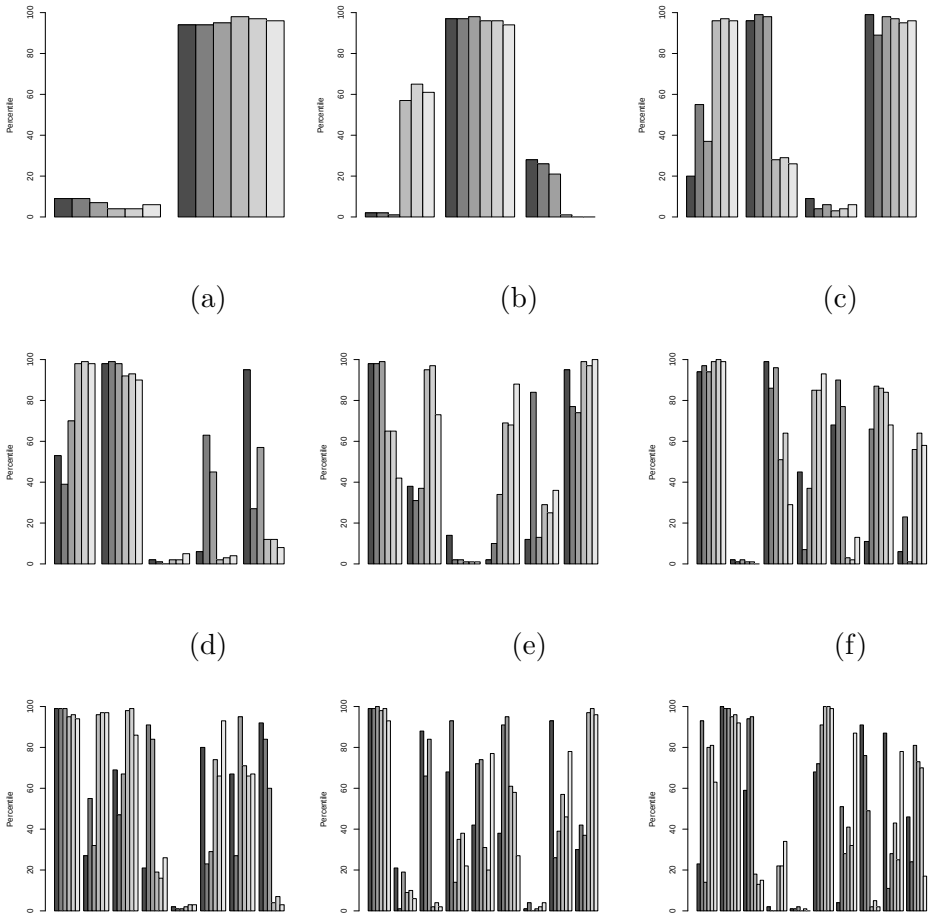


Figure 5.10: Percentiles of archetypes from $k = 2$ to $k = 10$.

As said before, the archetypes are not nested. In particular, this means

that if we first calculate three archetypes and then we calculate four, there is no reason so that these four include the first three obtained. The new four archetypes may have changed to better capture the shape of the data set. In this way, we must determine how many archetypes we want to consider. There might be two ways of doing it. On the one hand, the user may choose the number of archetypes he/she considers the best for his/her work. This would be a subjective decision. An objective alternative is based on using the elbow criterion, which is a very common method in Statistics. It simply consists in representing the RSS associated with each value of k. The correct value of k would be the one where a flattening of the curve occurs. As an illustration of this approach, Fig. 5.11 shows the RSS from $k = 2$ to $k = 15$. We see that an elbow occurs at $k = 3$, $k = 7$ and $k = 10$. In accordance with the law of parsimony (or Occam's razor) we consider that three and seven archetypes are the best number of archetypes. We think that a large number of representative cases may be counterproductive for the designer. Nevertheless, the final decision should be taken by the expert.

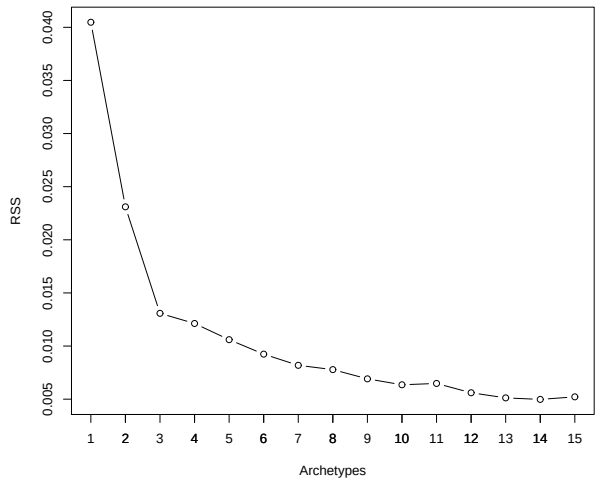


Figure 5.11: Screeplot of the residual sum of squares.

We first focus on the case of three archetypes by examining Fig. 5.10b: The first archetype presents small percentiles for the first three variables (corresponding to limb dimensions), whereas has average measures for the

last three variables (corresponding to torso dimensions). The second one represents individuals which are huge in all dimensions. Regarding the third archetype, it represents individuals which are small, although for the first three variables not very small, around the 25th percentile.

To describe the case of seven archetypes we see Fig. 5.10f: the first and second archetypes just show opposite patterns: the first one has high percentiles for all variables and the second one, low percentiles. In the same way, the third and four archetypes also are opposite: Whereas the third archetype has high percentiles in the first three variables (limb dimensions) and middle percentiles for the last three variables (torso dimensions), the fourth archetype presents middle percentiles for the first three variables and high percentiles for the last three. The fifth and seventh archetypes show contrary trends as well. In the fifth archetype, the percentiles are middle-high for the first three variables and low for the last three. This is the opposite of the seventh archetype. The sixth archetype represents a person which is huge in all measurements, but with short arms because has high percentiles for all variables excepting the first one, the only one related to arms.

In order to make a reliable comparison between the archetypes returned with our methodology with those cases obtained with PCA as in [228, 173], we have applied PCA to the whole database with the six variables standardized. Table 5.4 describes the coefficients for the six principal components, the percentage of variance explained for each component, and the cumulative percentage.

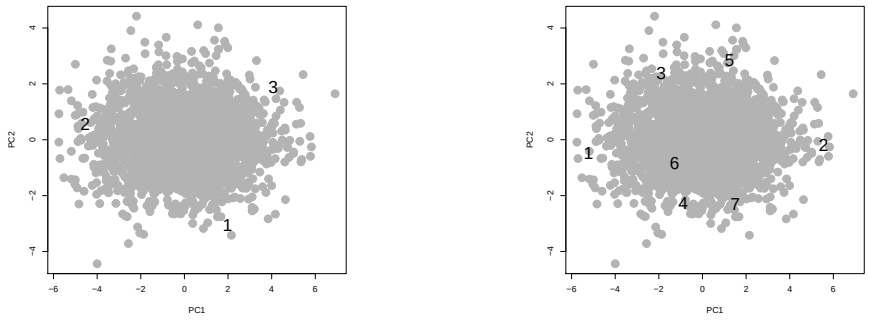
	PC1	PC2	PC3	PC4	PC5	PC6
Thumb Tip Reach	-0.364	0.453	0.697	0.418	0.04	-0.001
Buttock-Knee Length	-0.36	0.464	-0.716	0.374	0.036	-0.043
Popliteal Height Sitting	-0.39	0.408	0.025	-0.809	-0.144	0.077
Sitting Height	-0.46	-0.353	0.02	-0.082	0.305	-0.751
Eye Height Sitting	-0.449	-0.367	-0.025	0.004	0.494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

Table 5.4: PCA coefficients and percentage of explained variance.

The first two components capture the 82.6% of variability (until 89.15% with the first three). If we selected only the first two components as usual,

some variability (maybe important) would be discarded. Regarding the interpretation of the components, we see that the first component represents the overall size of the individuals. The second component contrasts (the sign is different) the limb dimensions (the first three) and the torso dimensions (the last three). The third and fourth components show a contrast inside the limbs dimensions (thumb tip reach versus buttock-knee length for the third, and thumb tip reach and buttock-knee length versus popliteal height sitting for the fourth). On the contrary, the fifth and sixth components show an opposite behavior inside the torso dimensions (sitting height and eye height sitting versus shoulder height sitting for the fifth, and sitting height versus eye height sitting for the sixth).

Fig. 5.12 displays the scores for the first two PCs of all individuals with gray color, with the scores for the three archetypes (a), and seven archetypes (b) in black squares. We appreciate in Fig. 5.12a that $k = 3$ archetypes are similar to those cases that can be obtained with the first two PCs. The second archetype (marked with a 2) is an extreme of PC1, and the first and third archetype (marked with a 1 and 3, respectively) correspond to a combination of extremes of PC1 and PC2 (octants). In the case of the archetypes obtained with $k = 7$ archetypes (Fig. 5.12b), all but the sixth archetype (marked with a 6, the one with scores -1.28 -0.86 for PC1 and PC2 respectively) correspond to extreme combinations of PC1 and PC2 (they form a circle). This means that this sixth archetype cannot be extracted as an extreme combination of the first two PCs. What is more, it cannot be obtained with any extreme combination of the PCs.



(a) (b)

Figure 5.12: PC scores for three (a) and seven (b) archetypes.

Now, we want to compare the $k = 7$ archetypes with the 8 cases extracted with PCA for 95% accommodation shown in [173]. To that end, in Table 5.5 appears the percentile values for those 8 cases, while Table 5.6 shows the percentiles for those 7 archetypes.

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

Table 5.5: Percentile values for two principal component representative cases.

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

Table 5.6: Percentile values for seven archetypes.

There are two clear correspondences: case W with archetype 1 and case Y with archetype 2. The case A is in the middle between archetype 1 and 3 but there is no case with PCA that corresponds to archetype 3. The case B is in the middle between archetype 2 and 5. The case X is the nearest to archetype 5 but there is not an exact equivalency. The case C can be considered in the middle between archetype 2 and 7. In the same way, the case Z is in the middle between archetypes 4 and 7. Finally, regarding case D, it could be seen as a mixture of archetypes 1, 4 and 6. However, there is not case for archetypes 4 and 6. After this examination, it can be stated that, except in two cases, there is no clear coincidence between the cases for PCA and archetypes.

In closing, Table 5.7 (resp. Table 5.8) collects the corresponding values for each original variable (without standardizing) for the PCA eight cases (resp. for the seven archetypes). Fig. 5.13 shows the skeletons of the seven archetypes.

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

Table 5.7: Variable values for two principal component representative cases.

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

Table 5.8: Variable values for seven archetypes.

5.3.3 Summary

This study has investigated the performance of the Archetypal Analysis to obtain representative boundary cases that entail a certain percentage of the population, compared with the most used PCA approach. In short, the procedure we have proposed is the following: first, depending on the problem, to standardize the data or not. Then, to use Mahalanobis distance and Chi-square distribution to obtain the sample in which obtaining the archetypes as the third and last step.

PCA is mainly a dimensionality reduction technique. On the contrary, Archetypal Analysis aims at obtaining extreme individuals, so it seems to be the suitable statistical tool to tackle the accommodation problem. Its application presents several advantages regarding PCA: the level of accommo-

dation is exactly reached. Some archetypes could not be obtained with PCA even if we consider all the components.

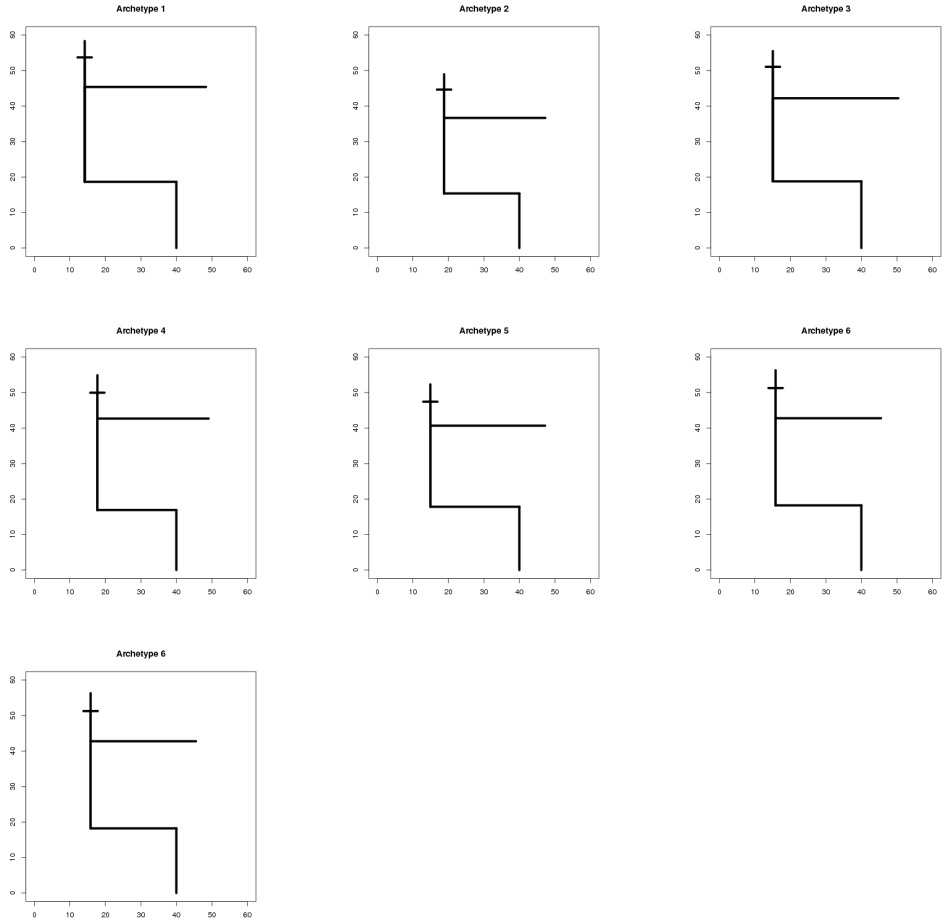


Figure 5.13: Skeleton plots visualizing the seven archetypes.

In addition, we have seen that there is not exact matches between the cases of PCA and archetypes. From a practical point of view, we have proposed to use the elbow criterion and the Occam's razor to determine the number of archetypes. A large number of archetypes, that is to say, of representative cases, may increase the time and cost of the design process unnecessarily.

5.4 Second methodology: Archetypoids

This second approach proposes an extension of AA to handle the accommodation problem. Archetypes can be any point in the convex hull. This means that they are not necessarily real observed individuals of the data set. However, in some cases it is fundamental that they are. In this context, a new archetypal concept is proposed: the archetypoid. It is a real observation, not a mixture of observations. Section 5.4.1 presents an algorithm to compute archetypoids and how to obtain them when features are unavailable. Section 5.4.2 explains the results obtained by applying the archetypoid algorithm to three problems in three different fields: sports (specifically, basketball), aircraft cockpits design and apparel design. At last, Section 5.4.3 discusses this new approximation.

5.4.1 Methodology

In order to solve the optimization problem related to the archetypoid analysis of eq. (5.1), various alternatives were analyzed. Because the archetypoid problem includes a variable that is required to be an integer, the optimization methods that are related to mixed integer programming [65] were reviewed. One widely used strategy due to its efficiency is the branch and bound method. It consists in dividing an initial unconstrained problem into a few smaller ones until a good solution is obtained. This method can be used in Matlab by means of a function called *BNB20()*, which can be downloaded from www.mathworks.com/matlabcentral/fileexchange/95-bnb/content/BNB20.m. Unfortunately, this method did not provide a good solution when trying to minimize eq. (5.1). We understand as good solution that one that verifies the problem conditions and reduces the computations required. As in [65] is pointed out, there are some situations where the branch and bound method is not appropriate. As a second possibility, we checked another type of optimization method: the genetic algorithm. It can also be used in Matlab by means of the function *ga()*. This algorithm creates a sequence of new populations from a random initial one and stops as soon as any of the stopping criteria is met. However, this method did not return a good solution either. In fact, the results provided by the genetic algorithm did not verify the equations of the archetypoid analysis problem. It is noteworthy that both the branch and bound method and genetic algorithm have a high computational cost for large sample sizes. A more naive option is to

calculate archetypoids with an exhaustive search, that is to say, to obtain the set of archetypoids that produces the minimum value of the objective function, after trying all the possible combinations. This will be called the combinatorial solution. However, this alternative increases its computational time seriously as the sample size of the database increases. In view of these procedures do not work for computing archetypoids, we decided to develop an algorithm based on PAM.

5.4.1.1 Archetypoid algorithm

Our algorithm has two phases: a BUILD phase and a SWAP phase, as PAM (see Section 2.1). In the BUILD step, an initial set of archetypoids is determined. Unlike PAM, we do not obtain this collection in a stepwise format. Instead, we suggest choosing the set made up of the nearest individuals to the archetypes returned by **archetypes** (the best archetypes are selected after running the algorithm several times, specifically, twenty times). We propose to define this set in two different ways. A first alternative is to compute the Euclidean distance between the archetypes and the individuals and choosing the nearest ones, as mentioned in [54]. Hereafter, this set will be referred to as *nearest*. The second option identifies the individuals with the maximum value of α for each archetype, i.e., the individuals with the largest relative share for the respective archetype. In this case, this set will be referred to as *which*. This is used in [55] and [184]. Accordingly, our initial set of archetypoids is either *nearest* or *which*.

The aim of the SWAP phase of our algorithm is the same as the one of the SWAP phase of PAM, but changing the objective function. Our SWAP step attempts to improve the quality of the set of archetypoids by exchanging selected individuals for unselected individuals and by checking whether these replacements reduce the objective function of eq. (5.1), namely RSS. In the inner loop of this second step, for each given set of archetypoids, S, the α coefficients are updated in order to calculate the effect of the swap. The corresponding RSS is then calculated. If this RSS is lower than the previous RSS, S turns into the new initial vector of archetypoids. The SWAP phase is repeated until there is no change in any archetypoid. As indicated above, the α coefficients have to be recalculated, so it is necessary to solve n convex least squares problems as in the algorithm implemented in **archetypes**. To deal with the n convex least squares problems, a penalized version of

the non-negativity least squares algorithm by Lawson and Hanson [125] is used, in such a way that the convexity constraints (being nonnegative and adding up to one) are fulfilled (see [56, point 2.1 on page 3]). Additionally, we would like to point out that our algorithm does not update the β coefficients in the inner loop by solving other k convex least squares problems as **archetypes** does. In our algorithm, the β coefficients are "updated" in the sense that for the individuals considered as archetypoids, their β is equal to 1, being 0 for the other unselected individuals. Because all potential swaps are considered, the results of the algorithm do not depend on the order of the objects in the database. Besides, unlike the archetype algorithm that alternates between finding the best α for given archetypes $\mathbf{Z}$ and finding the best archetypes $\mathbf{Z}$ for given α, our algorithm only focuses on finding the best α for given archetypoids $\mathbf{Z}$. This is because the archetypoids correspond to sampled objects. A brief outline of the archetypoids algorithm is given in Algorithm 9. We have considered the 2-norm for the distance $\|\mathbf{X} - \alpha\beta'\mathbf{X}\|_2$. As final points, some comments about large sample sizes, local minimum and standardization of data are listed below:

1. For very large databases, an algorithm using samples of the data like Clustering LARge Applications (CLARA) algorithm does [115], would be more suitable.

2. Our algorithm, as PAM, aims at finding good solutions in a short period of time, although they are not necessarily the best ones. The global minimum solution could always be obtained with the combinatorial solution, using as much time as necessary. Nevertheless, it would be computationally very inefficient.

3. As mentioned in Section 5.3, standardization of data depends on ones sense about the data. Variables are standardized for a variety of reasons: they measure different dimensions, their scales are not comparable or if their ranges are very different. In our practical examples, we standardize the data for the basketball and aircraft pilots databases. On the contrary, we will work with the data as they stand in the apparel design example. We have modified the *stepArchetypes* function of **archetypes** because it standardizes the data by default. Therefore, our way of proceeding in each problem is the following:

    i) Depending on the problem, to standardize or not the data.

ii) Archetypes must be calculated using a new R function called *stepArchetypesMod*, that results from modifying and adapting the original *stepArchetypes* function.

iii) Archetypoids are calculated with the archetypoid algorithm, beginning with *which* and *nearest* sets, for several values of k. As in Section 5.3, we select the k where the elbow on the RSS representation is found.

5.4.1.2 Archetypoids when features are unavailable

There are problems related to specific fields such as psychology or economy, especially those where multidimensional scaling applies, where only dissimilarities are available. In such situations, we cannot approximate the data directly as mixtures of archetypoids or archetypes. However, if the dissimilarities are Euclidean distances, they can be represented exactly in at most n - 1 dimensions [146, Theorem 14.4.1] using classical multidimensional scaling (cMDS). Multidimensional scaling takes an input matrix giving dissimilarities between pairs of objects and outputs a set of points such that the distances between the points are approximately equal to the dissimilarities, since the dimension of the space in which the data have to be represented, is usually less than n - 1. These new features can be used to find the archetypoids. In addition, we could also obtain archetypes in this new space, but we cannot establish a correspondence with the original subjects or to create artificial subjects, for which only the dissimilarities were available.

If the dissimilarity is a distance, but not an Euclidean distance, cMDS can be used as an approximation (and it is optimal for a kind of discrepancy measure [146, Theorem 14.4.2]), or we can use the h-plot [53], a recent alternative method that it also works when the dissimilarity is not a distance.

Next, we detail the phases for obtaining the archetypoids when features are unavailable. Let $\mathbf{D}$ be the $n \times n$ matrix that contains the dissimilarities between the observations i and j, d_{ij}.

1. Compute cMDS with a dimension of the space in which the data have to be represented, equal to m. For that purpose, we have used the function *cmdscale* of R. The number m is an integer less than or equal to $n-1$ and it is chosen as the first integer for which the goodness of fit

Algorithm 9 Archetypoid algorithm

Let be an $n \times m$ matrix $\mathbf{X}$.

1. PHASE BUILD: Initial vector of k archetypoids.

Select *nearest* or *which* vector:

$vect_{ini} = (x_1, \ldots, x_k)$

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

2. PHASE SWAP: Try to improve the set of archetypoids.

Set $vect_{archet} = vect_{ini}$ **and** $rss_{archet} = rss_{ini}$.

for $j = 1 \rightarrow k$ **do**

  $setposs =$ dif. between $(x_1, \ldots, x_n)$ and $(x_1, \ldots, x_k)$

  for $t \in setposs$ **do**

    $vect_{swap} = (t, x_1, \ldots, x_{k-1})$ **(without** x_j**)**

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

    if $rss_{swap} < rss_{archet}$ **then**

      $vect_{archet} = vect_{swap}$

      $rss_{archet} = rss_{swap}$

    end if

  end for

end for

return $vect_{archet}$ **and** rss_{archet}

(GOF) measure α_2 proposed in [146, eq. 14.4.8], is greater than 90%. The greater m, the more variables and the more computation time [56].

2. Compute the archetypoids of the $n \times m$ matrix $\mathbf{X}$, the matrix returned by cMDS. This matrix has the coordinates of the points computed to represent the dissimilarities.

The archetypoids returned by cMDS correspond to a specific set of points, with a direct correspondence with the original observations. We also obtain the α coefficients, indicating the contribution of each archetypoid to each original observation. However, the predictors $\sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$ of each original observation cannot be represented with the original information (the dissimilarities), in the same way that archetypes cannot be represented in that space.

5.4.2 Results

We present three applications: a sportive (basketball) example, the cockpit design problem and an apparel design problem.

5.4.2.1 Sportive example

This example is motivated by Ref. [55] where archetypes for two mass sports such as basketball and soccer are calculated. Because in [55] real players are analyzed, this reference is of particular interest for us since we are going to be able to demonstrate that archetypoids need not to be the same as *nearest* or *which* individuals. Among the different examples introduced in [55], we focus on the NBA database that collects the total minutes played and field goals made of 441 players from the season 2009/2010. Table 5.9 and Figure 5.14 show in blue color the archetypal players obtained in [55]. They are Kevin Durant, Dwayne Jones and Jason Kidd. This is both the set *nearest* and *which*.

The first thing we do is to compute the best possible set of archetypal players, the combinatorial solution. This set is made up of Kevin Durant, Jason Kidd and Travis Diener (put in a frame in both the Table 5.9 and Figure 5.14) and was obtained after 9 days of computation, using a forward sequential search procedure run on a single computer. When applying our archetypoid algorithm to the same database we have indeed obtained these

three players as the final archetypoids. The computational time in this case were a few minutes, both beginning with *nearest* and *which*.

Archetypal players of [55] (blue color) and obtained with our proposal (frame box)

	Name	Total minutes played	Field goals made
124	Kevin Durant	3241	794
236	Dwayne Jones	7	0
243	Jason Kidd	2883	284
113	Travis Diener	50	2

Table 5.9: Archetypal players obtained in [55] (blue color) and by our proposal (frame box).

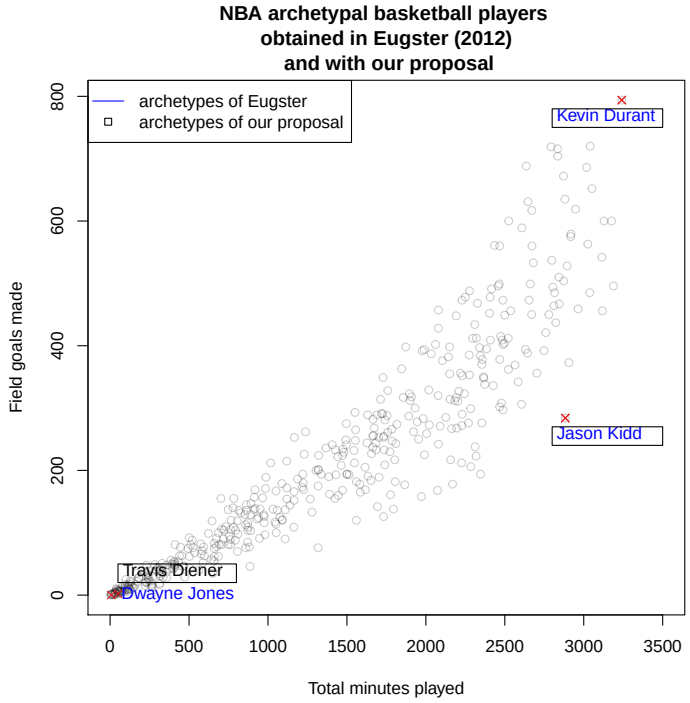


Figure 5.14: Total minutes played and field goals made of a set of NBA players from the season 2009/2010 with the archetypal players obtained in [55] (blue color) and by our proposal (frame box), also marked with a red cross.

Next, a brief description of the main features of each of these players

is introduced. In sport, a detailed analysis of the players performance may help coaches to create individualized performance profiles. Kevin Durant is a very good scorer because he scored a lot of shots the time he stayed on the court. According to these data, if he played an entire NBA game (48 minutes, without overtime periods), he would score almost 12 shots, which is a very good performance. Durant has won three NBA scoring titles to this day. Dwayne Jones was not able to score any point because he played very few minutes. The same thing occurs with Travis Diener who only scored two points since he hardly played 50 minutes. These kind of players are called "benchwarmers". In addition, Jason Kidd might be considered an "ineffective scorer" because he played a great amount minutes and he did not scored many baskets. However, it is well-known that Jason Kidd is a point guard whose main role is assisting instead of scoring. In fact, he is ranked second on the NBA's all-time assist list.

5.4.2.2 Cockpit design problem

We search for archetypes and archetypoids in the same data set from the USAF Survey explained in Section 5.3.1 (discarding the more extreme 5% data and choosing the six cockpit dimensions). For this example, it was not possible to obtain the combinatorial solution in a reasonable time because the large sample size of the database. According to what the screeplots of Fig. 5.15 suggest, we choose 3 archetypes and archetypoids.

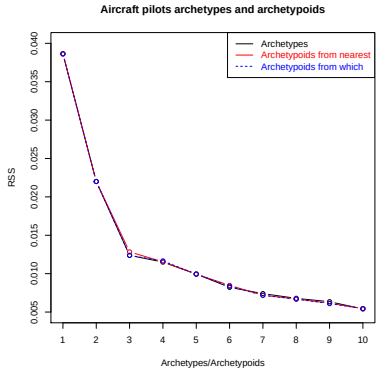


Figure 5.15: Screeplot of the RSS of the archetypes and archetypoids (from *nearest* and *which*) for the aircraft pilots of the data from the USAF survey. The elbow is at 3 in all the cases.

Table 5.10 shows the RSS associated with this number of archetypes, with the *nearest* and *which* individuals to these archetypes and with the same number of archetypoids.

	RSS
3 archetypes	0.012380776
nearest (511,314,1691)	0.01824692
which (1421,314,1691)	0.01947072
3 archetypoids from nearest (2177,2240,1691)	0.012830864
3 archetypoids from which (1632,1822,52)	0.012385042

Table 5.10: RSS associated with each set of archetypes, nearest individuals or archetypoids for the aircraft pilots of the data from the USAF survey.

We appreciate that the smallest RSS is for the archetypes. This could be expected because its set of possible solutions is the largest. However, the RSS related to the *nearest* and *which* archetypoids (the archetypoids obtained beginning from *nearest* and *which*, respectively) are pretty close to the archetype-RSS and what is more, they decrease in each case the RSS associated with the initial set of the nearest individuals to the archetypes. Although not outstanding, this reduction is remarkable.

Similarly to Fig. 5.10, we can analyze the percentiles of the three archetypoids beginning with *nearest* and *which* for each one of the six cockpit dimensions, see Fig. 5.16.

The first *nearest* archetypoid has high percentiles for the first three variables (corresponding to limb dimensions), small percentiles for the fourth and fifth and an average value for the sixth (shoulder height sitting). The third *which* archetypoid presents a similar behavior, except that the shoulder height sitting is smaller. The second *nearest* archetypoid and the first *which* archetypoid are small in all measurements (although the *which* archetypoid is a little larger). Finally, the third *nearest* archetypoid is high in the six variables. In the same way, the second *which* archetypoid shows high values, but they are not so high. We note that the percentiles of the three *which* archetypoids are not so extreme in comparison with the *nearest* ones. Fig. 5.17 allows to compare the three archetypes from which the *nearest* and *which* archetypoids are computed. We see that the first archetype is small for all variables (similarly to the second *nearest* and first *which* archetypoids). The

second archetype is similar to the first *nearest* and third *which* archetypoids, whereas the third archetype is similar to the third *nearest* and second *which* archetypoids.

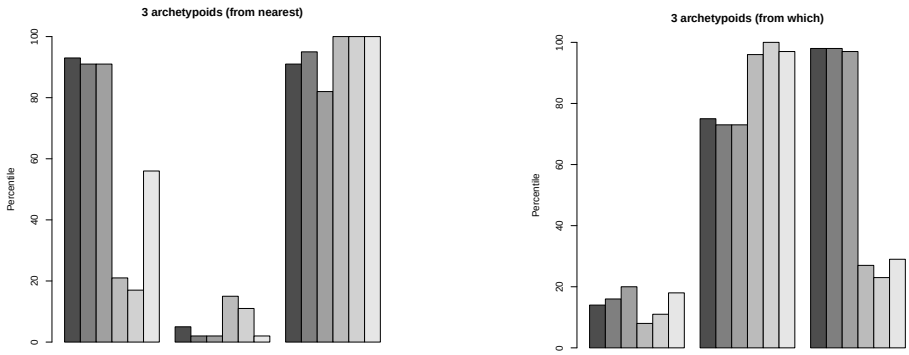


Figure 5.16: Percentiles of three archetypoids, beginning with *nearest* (left) and with *which* (right) for the aircraft pilots of the data from the USAF survey.

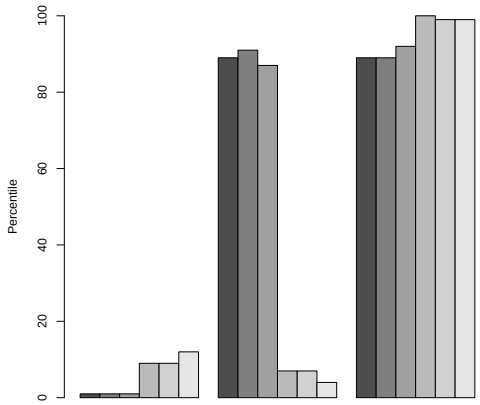


Figure 5.17: Percentiles of three archetypes for the aircraft pilots of the data from the USAF survey, from which the *nearest* and *which* archetypoids shown in Fig. 5.16, are calculated.

5.4.2.3 Apparel design problem

We use the Spanish anthropometric data (see Section 1.3). From the whole database, we choose a subsample of the 470 non-pregnant and non-lactating women, between 25 and 45 years, with a bust circumference between 86 and 90 cm. This age range represents an important potential group for the apparel market and, at the same time, includes a high variability of body shapes. As a result, women with the same size (86-90 cm bust for upper garments) may have very different body shapes [39], causing fitting problems when a garment is designed to fit a body prototype perfectly. Thus, different classifications of body types have been proposed for apparel sizing and design [187], [166], [62], [96].

Within this context, it is proposed that archetypoids should be used to identify subjects who represent the fittings problems of the target population. Central cases are used for developing initial design ideas, whereas boundary cases are useful in determining the extent of adjustment or scaling required to acommodate the full range of variability in the target population (both the boundaries and all the individuals between the boundaries). In the design process, it is important to know how close the design is to accommodating the boundary cases. If the boundaries are not acommodated, this information can be used to determine changes necessary to achieve the desired accommodation percentage [100]. Note that we are not seeking to find sub-sizes, but to accommodate women within a specific size. Central cases and clustering algorithms should be used to define sizes.

As explained in Section 1.3, a 3D binary image of the trunk of each woman is available. We can compute the dissimilarity between trunk forms and build a distance matrix $\mathbf{D}$ between women. Let A and B be two binary images associated with the trunk of two women and defined in a lattice Λ. There are several metrics for measuring the differences between A and B. We use the simplest one, which is the misclassification error: $d(A, B) = \frac{nu(A \Delta B)}{nu(\Lambda)}$, where Δ is the set symmetric difference, and nu counts the number of pixels in that set, that is to say, the volume of the set.

We have chosen $m = 4$, with 92.39% explained, according to the agreement measure for the proportion of the distance matrix $\mathbf{D}$ explained. In this way, we apply the archetypoids algorithm to a matrix with 470 rows and 4 columns. The screeplots of Fig. 5.18 suggest choosing 3 or 5 archetypes and archetypoids.

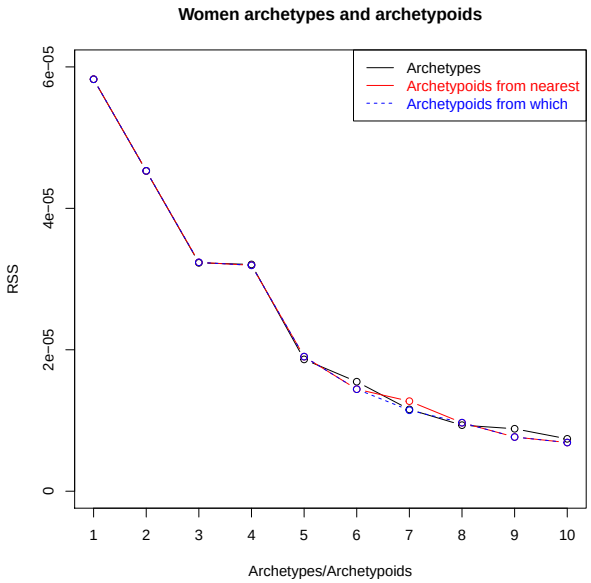


Figure 5.18: Screeplots of the residual sum of squares of the archetypes and archetypoids for the Spanish women.

In this case, the *nearest* and *which* archetypoids agree in both cases of $k = 3$ and $k = 5$. In the interests of brevity and as an illustrative example, we are going to examine the results of three archetypes and archetypoids. However, as mentioned, in a real situation the final decision about how many archetypoids to consider would correspond to the analyst. Again, the first thing we do is to calculate the combinatorial solution. We can do it in this case because the sample size is not too large (it is similar to that of the NBA database). The best possible set is formed by individuals 85, 212 and 447. It was obtained after 23 days by using a forward sequential search procedure run on a single computer. By applying our algorithm, we have also obtained these same three women as the final archetypoids (both *nearest* and *which* archetypoids), taking only a few minutes.

Table 5.11 shows the RSS associated with this number of archetypes, with the nearest individuals (*nearest* and *which*) to these archetypes and with the same number of archetypoids.

Again, the smallest RSS corresponds to the archetypes. It also occurs in this case that the RSS related to the archetypoids is smaller than the RSS of the respective nearest individuals and at the same time, is close to the

archetype-RSS.

	RSS
3 archetypes	3.229195e-05
nearest (212,445,288)	3.648499e-05
which (170,310,120)	3.825974e-05
3 archetypoids from nearest (212,447,85)	3.236335e-05
3 archetypoids from which (212,447,85)	3.236335e-05

Table 5.11: RSS associated with each set of archetypes, nearest individuals or archetypoids for the Spanish women.

In addition, Table 5.12 describes the archetypoid women according to certain easily recognized variables: weight, height, waist circumference and hip circumference. Fig. 5.19 shows these women. RAPITA026 is very different to the other two in all the measurements. ALCA163 y STAC055 are similar in terms of their waist and hip measurements but STAC055 is very tall and thin while ALCA163 is near the limit of being overweight.

Woman code	Weight	Height	Waist circumf.	Hip circumf.
RAPIT026	51.2	1.61	70.4	91.7
ALCA163	55.0	1.49	78.9	104.2
STAC055	63.9	1.72	77.2	104.3

Table 5.12: Women archetypoids.

5.4.3 Summary

This methodology introduces the concept of an archetypoid, develops an algorithm for locating them in the data and presents three applications. There are problems where it is fundamental to find extreme representative data. The archetypal analysis is a very useful tool to that end but presents a very important pitfall: the archetypes do not correspond necessarily to observed individuals. In order to overcome this fact, the usual procedure is to select those individuals who are the closest to the computed archetypes (*which* or *nearest* archetypes). However, it may even happen that those nearest subjects are not plausible individuals (as in fact ocurrs in [184], where the nearest

"economists are a mixture of different types"). Furthermore, there are some cases where is critical that the archetypes are real subjects. To tackle this problem, a new archetypal concept is introduced: the archetypoid, that is a real (observed) archetype.

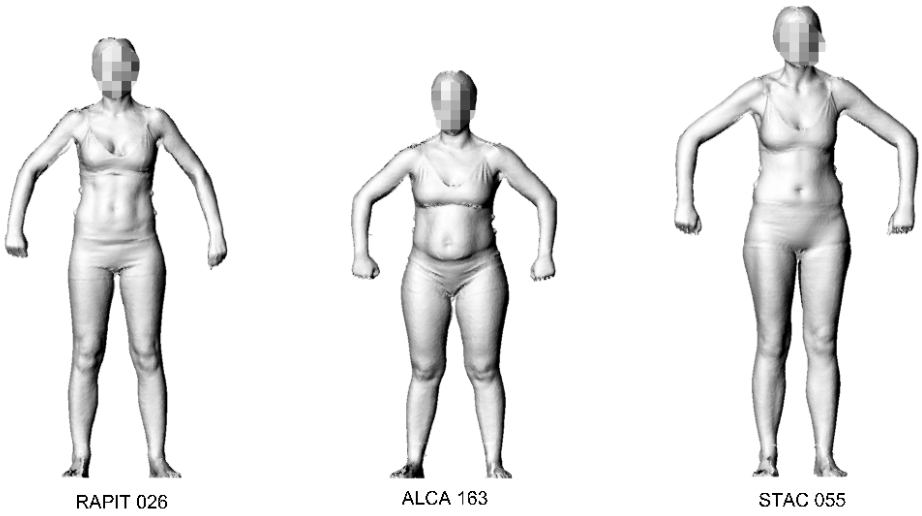


Figure 5.19: Three archetypical women: RAPIT026, ALCA163 and STAC055.

In order to develop the archetypoids algorithm we follow the idea in PAM for finding the medoids (build and swap phases), but changing the build phase for beginning near the archetypoids, and changing in the swap phase the objective function for our optimization problem, in order to know if we should make or not the swap. This algorithm is quick and efficient in terms of computational complexity. The archetypoids do not necessarily coincide with the *which* or *nearest* individuals to archetypes. In fact, in the sportive example, although similar, the archetypoids were not exactly the same as those described in [55]. In the cockpit and apparel design problems, this is more evident, since the RSS of archetypoids is decreased to the same level of the archetype-RSS, and the set of archetypoids does not coincide with those obtained by the *which* or *nearest* options.

According to our results, it is not possible to state categorically if it is

more convenient to choose the *nearest* or *which* options in the BUILD phase of the archetypoid algorithm. We have seen that for the apparel design problem both *nearest* and *which* options provided the better solution. On the contrary, the *which* alternative offered the local minimum for the cockpit design problem. In this way, both options must be checked, obviously unless *nearest* and *which* coincide, as it occurs with the NBA database.

5.5 Chapter conclusions

In this chapter, in order to tackle the accomodation problem, we have used for the first time the archetypal analysis. We have shown in a first approximation that AA performs better than the common-used method PCA. AA determines extreme patterns from a population, but does not necessarily provide specific boundary individuals associated with those patterns. Identifying real people with the anthropometric features of boundary cases is usually problematic because boundaries represent extremes in the population. Therefore, there are not many people who present the proportions. In order to overcome this limitation of AA, we have proposed a new archetypal concept, the archetypoid, which correspond to sampled individuals.

A lot of ergonomic studies require to build several mock-ups to test and validate the match between product dimensions and related physical measurements in addition to other aspects such as comfort and flexibility. It is very valuable that boundary cases be identified ir order to improve initially the mock-ups. Our archetypoid analysis and algorithm should be a very useful approach in these types of situations.

In fact, the calculus of archetypoids can be successfully applied in all the fields such as computer vision, text mining, collaborative filtering, etc, where the archetypal analysis has been used. Furthermore, the archetypoid analysis can be used beyond multivariate vectors or dissimilarity matrices. For example, it is suitable with functional data, interval data, images [201], etc.

We have some different possibilities as future work: From a practical point of view, a study about the computational complexity of the archetypoids can be done following the ideas developed in [56, Sect. 4]. Besides, we aim at implementing an archetypoid algorithm which can be used for very large databases. From a theoretical perspective, it would be interesting to perform a numerical simulation with data from different probability distri-

butions in order to analyze the location of the archetypes and archetypoids and to study their accuracies using randomization techniques. Another direct extension is to try to define weighted and robust archetypoids, similarly to the ideas explained in [57], or to consider missing values modifying the objective function as in [152] is made with archetypal analysis.

Because the main subject of this PhD thesis is the apparel design, we aim at looking further at the use of archetypoids in the design process, by considering more body sizes and age populations. In this first approximation, we have determined three archetypoids. However, it may be more interesting or appropriate to consider more representative individuals in order to achieve a better fit of garments. We expect that the archetypoid analysis could serve as a satisfactory approach to the clothing design problem and to the lack of fitting of garments.

Chapter 6

Anthropometry R package

The accelerating power of modern 3D scanning technologies has contributed to the generation of broad anthropometric databases which constitute high valued data to improve the product design and fitting adapted to the user population. Accordingly, Ergonomics and Anthropometry are two fields rapidly becoming more quantitative, so modern software tools and computer applications are demanded for a more efficient use of anthropometric data.

First of all, we undertook a search of the Internet to look for software sources related to Ergonomics and Anthropometry. We found a data repository in http://www.openerg.com/psz/, where a visual software called *PeopleSize 2008* is presented. It is a paid software that allows up-to-date data of the Western population to be downloaded. Most other available sites provide anthropometric calculators for computing simple indicators such as percentiles or z-scores of the children population. These tools help public health researchers and pediatricians to analyze their data. For instance, the World Health Organization (WHO) has developed the *WHO Anthro* software, which is available from http://www.who.int/childgrowth/software/en/. Another example is *NutriStat*, a nutritional anthropometric tool created by the Centers for Disease Control and Prevention, from USA, to examine children and adolescents (http://nutristat.codeplex.com/). *PeopleSize 2008* also computes percentiles of its data.

In recent years, different software systems have been created for the management of 3D anthropometric databases. Three examples are Nefertiti [159], Alexandria and Cleopatra [160]. Furthermore, over the past 50 years, human modelling software has became more and more available, which has helped to improve the design, development and quality of workspaces. Modern human

modelling techniques include Jack and Ramsis, which are two of the most widely used tools by a broad range of industries [18].

Regarding statistical software, we investigated if any available R package from CRAN was developed to deal with anthropometric data. Searching by the keyword *anthrop* in the list of available packages, sorted by name, (visit http://cran.r-project.org/web/packages/available_packages_by_name.html), two packages were highlighted: **AnthropMMD** [180] and **phcfM** [210]. On the one hand, **AnthropMMD** is a package addressed to anthropologists. On the other hand, **phcfM** serves for modelling anthropogenic deforestation. Anthropogenic refers to the facts caused or produced by humans. Thus, these packages are not related to the field of Anthropometry. In addition, by repeating the same process using the keyword *ergon* (referring to Ergonomics), no R package was founded.

As far as we know, there is currently no reference in the literature of sizing systems that provides the programming of the proposed algorithms. To the best of our knowledge, with the exception of human modelling tools like Jack, there are neither general software nor statistical packages available from the Internet to tackle the common problems related to Anthropometry or Ergonomics, such as the definition of an efficient sizing system or the accommodation problem.

In this context, we introduce here the **Anthropometry** R package that brings together all the algorithms and functions associated with the statistical methodologies presented in this PhD dissertation [212]. The most current version of **Anthropometry** is always available from the Comprehensive R Archive Network at http://cran.r-project.org/package=Anthropometry. It is also available from http://www.uv.es/vivigui/software. **Anthropometry** includes a vignette to assist new users in learning the purpose and use of this package. As illustrative data of the whole Spanish anthropometric survey, **Anthropometry** provides a data set called *dataDemo*, containing 600 women and their measurements for bust, chest, waist, hip and neck to ground length. Besides, another data file called *landmarks* collects the landmarks representing the shape of the 600 women. The data set of the USAF survey is also included. A short manual for helping with installation and first use of this package will be also freely downloable from the same author's website http://www.uv.es/vivigui/software.

In the following sections, the main functions to execute each one of the proposed methodologies are described (related to clustering, the statistical shape analysis, data depth and the archetypal analysis).

6.1 Antropometric dimensions based clustering

The function to compute the dissimilarity presented in eq. (2.16) in Section 2.3 is written in C and it is first exported from the NAMESPACE file. Then, the `GetDistMatrix` function calls it from R.

Both `trimowa` and `hipamAnthropom` functions, that implement the so-called methodologies, incorporate the calculus of the dissimilarity matrix within them. Besides, the `CCbiclustAnthropo` function is the R function related to the *biclustAnthropom* approach.

6.1.1 `trimowa` function

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

This function calls the `trimmedoid` function that is the programming of the trimmed k-medoids algorithm. Its arguments are the following:

- x: Data frame. In our approach, this is each one of the subframes originated after segmenting the whole anthropometric Spanish survey in twelve bust segments, following the *European standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals* [59]. Each row corresponds to an observation and each column corresponds to an anthropometric variable. All variables are numeric.
- w: The aggregation weights of the OWA operators. 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 dissimilarity function. Given the five variables considered in our study, this vector is c(23,28,20,25,25). This vector would be other according to the variables considered.

6.1.2 `CCbiclustAnthropo` function

```
CCbiclustAnthropo(data,waist,waistCirc,lowerVars,nsizes,nBic,
diffRanges,percDisac,dir)
```

Its arguments are the following:

- data: Data in which searching for biclusters. Each row corresponds to an observation and each column corresponds to an anthropometric variable. All variables are numeric.
- waist: Vector containing the waist values of the individuals.
- waistCirc: *data* is segmented in twelve waist classes following the *European standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals* [59]. This vector contains the waist values to define each one of the waist segments.
- lowerVars: Lower body dimensions involved in the analysis.
- nsizes: Number of waist sizes.
- nBic: Maximum number of biclusters to be obtained in each waist size.
- diffRanges: List in which each element is a vector whose extremes indicate the acceptable boundaries for selecting the variables that have a similar scale.
- percDisac: Proportion of no accommodated sample.
- dir: Working directory for saving the results.

6.1.3 `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 the following:

- x: Data frame. In our approach, this is each one of the subframes originated after segmenting the whole anthropometric Spanish survey in twelve bust segments, following the *European standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals* [59]. Each row corresponds to an observation and each column corresponds to an anthropometric 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 (equal to 0) is maintained. See [222, p. 154] for details.

- maxsplit: Maximum number of clusters that any cluster can be divided 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 value is also maintained, therefore this value is ignored. See [222, p. 154] for details.

- orness: Quantity to measure the degree to which the aggregation is like a min or max operation. This value is used to compute the aggregation weigths by means of `WeightsMixtureUB`.

- type: Type of HIPAM algorithm to be used. The possible options are 'MO' (for using $HIPAM_{MO}$) and 'IMO' (for using $HIPAM_{IMO}$).

- ahVect: Constants that define the dissimilarity function. Given the five variables considered in our study, this vector is c(23,28,20,25,25). This vector would be other according to the variables considered.

- ...: More arguments to be passed to the internal functions of the HIPAM algorithms.

6.2 Statistical shape analysis

The function to use the Lloyd version of k-means adapted to shape analysis (what we called *kmeansProcrustes*) is `LloydShapes`. In addition, the function to use the Hartigan-Wong version of k-means adapted to shape analysis is

`HartiganShapes`. The function to execute the trimmed *kmeansProcrustes* is `trimmedLloydShapes`.

6.2.1 `LloydShapes` function

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

Its arguments are the following:

- 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).
- Nclusters: 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 the simulation study.
- print: Logical value. If TRUE, certain messages associated with the running process are displayed.

6.2.2 `HartiganShapes` function

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

Its arguments are the following:

- 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).
- Nclusters: 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 the 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 just as an empty vector.
- print: Logical value. If TRUE, some messages associated with the running process are displayed.

6.2.3 `trimmedLloydShapes` function

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

Its arguments are the following:

- 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 observations.
- alpha: Proportion of trimmed sample.
- Nclusters: 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: A logical value. If TRUE, certain messages associated with the running process are displayed.

6.3 Statistical data depth

The `TDDclust` function corresponds to the so-called methodology.

6.3.1 `TDDclust` function

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

Its arguments are the following:

- X: Data frame. Each row corresponds to an observation and each column corresponds to an anthropometric 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 [109].
- 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 value is 0.9.
- lplot: Tracking convergence, default value is 0.
- Trimm: Proportion of no accommodated sample.
- data1: The same data frame as X, used to incorporate the trimmed observations to the rest of them for the next iteration.

6.4 Archetypal analysis

After pre-processing the data by means of the `accommodation` function (eventual standarization and removal of extreme individuals), the `archetypesUSAF` function allows some of the results presented in [54] to be reproduced, whereas the `stepArchetypoids` function calls the `archetypoids` function to run the archetypoid algorithm repeatedly.

6.4.1 `archetypesUSAF` function

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

Its arguments are the following:

- data: USAF 1967 database (see the `dataUSAF` file). Each row is an observation and each column corresponds to an anthropometric variable.
- numArchet: Number of archetypes.
- verbose: Logical value. If TRUE, it shows the progress during execution (this is the same argument of the `stepArchetypes` function of the **archetypes** R package [56]).
- nrep: For each archetype run `archetypes` nrep times (this is the same argument of the `stepArchetypes` function of **archetypes**).

6.4.2 `archetypoids` function

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

Its arguments are the following:

- i: Number of archetypoids.
- data: Data set in which looking for archetypoids. Each row corresponds to an observation and each column corresponds to an anthropometric variable. All variables must be numeric.
- huge: This is a penalization added to solve the convex least squares problems regarding the minimization problem to estimate archetypoids, see [56]. 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 been already 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 next *nearest* argument to know 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 the **archetypes** R package [56] to apply the archetype algorithm to raw data. The `stepArchetypes` function standardizes the data by default and this option is not always desired. This 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 [215] (submitted for publication), 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 was used in [54]. The *which* vector is calculated by identifying consecutively the individual with the maximum value of alpha for each archetype, until getting the number of archetypes defined. It is used in [55].

- 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.

6.4.3 `stepArchetypoids` function

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

Its arguments are the following:

- i: Number of archetypoids.
- 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 the **archetypes** R package [56] (in [215] (submitted for publication), archetypes are computed after running the archetypes algorithm twenty times). The *nearest* vector is calculated by computing the Euclidean distance between the archetypes and the individuals and choosing the nearest. It was used in [54]. The *which* vector is calculated by identifying consecutively the individual with the maximum value of alpha for each archetype, until getting the number of archetypes defined. It is used in [55].
- data: Data set where looking for archetypoids. Each row corresponds to an observation and each column corresponds to a variable. All variables must be 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 is needed to compute the nearest individuals to archetypes.

Chapter 7

Conclusions

This PhD dissertation has been raised in order to be a rigorous scientific contribution, from the mathematical and statistical point of view, to the disciplines of Ergonomics and Anthropometry. Throughout this report, we have developed different statistical methodologies that may be useful to improve the ergonomic design of products, focusing on the efficient design of clothes and working places.

Current sizing systems used by the apparel industry are not accurately optimized to properly fit the target population. As a consequence, a large part of the population, especially women, does not find clothing that fit well, even after trying several garments. This results in a poor fit, unsold garments and a less competitive business. Furthermore, many people return bought clothes because they are not satisfied with them. Because of this, there are many obsolete stocks. A striking effect of this circumstance in Spain is the proliferation of the so-called *outlet* stores. One of the main problems to develop new patterns and designs is the lack of updated anthropometric data of the current population. Outdated size charts contribute to sizing variance between companies. In this context, The Spanish National Institute of Consumer Affairs (INC, in Spanish) of the Spanish Ministry of Health and Consumer Affairs commissioned in 2006 a 3D anthropometric survey of the Spanish female population, in the frame of the agreement signed with the main Spanish clothing companies. The study was performed by the Biomechanics Insitute of Valencia and the anthropometric information recorded was both 1D and 3D. The main motivation was to characterize the shape and body dimensions of the current Spanish women.

Each of the new statistical approaches presented in this PhD work has

been applied to the anthropometric database obtained from this Spanish survey. In this way, one of the main goals of this doctoral work consisted in developing mathematical and statistical techniques and tools for the exploitation of human body databases with a focus on the ergonomic design and functional evaluation of products. This dissertation is part of the activities carried out by the research project related to the Spanish anthropometric survey and can be considered as an example of industry-academia interaction, that help to highlight top-level research and encourage excellence in science.

In Ergonomics and Anthropometry, the body size variability within the target population is characterized through the definition of a limited number of cases. An anthropometric case may be a particular human being or a combination of measurements. Depending on the design problem, there are three types of cases: central, boundary and distributed. Our proposed methods have been developed aimed at identifying central and boundary cases.

The approaches based on clustering algorithms, *trimowa*, *biclustAnthropom*, *hipamAnthropom*, *kmeansProcrustes* and *TDDclust*, allow to define 1D and 3D central cases, which actually are representative statistical models or prototypes (and fit models in the case of *hipamAnthropom*) of the human body of the target population. These prototypes and fit models can be used to make more realistic store mannequins. The five aforementioned methodologies have followed the same scheme. Firstly, the selected data matrix was segmented using a primary control dimension (bust circumference in the case of *trimowa*, *hipamAnthropom*, *kmeansProcrustes* and *TDDclust*, and waist circumference in the case of *biclustAnthropom*). Then, a further segmentation using other secondary control anthropometric variables is carried out. 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 sizing. By using a more appropriate statistical strategy, such as clustering, homogeneous subgroups are generated taking into account the anthropometric variability of the secondary dimensions that have a relevant influence on garment fit. Every method has been adapted to only accommodate the "standard" population. In order to choose the primary and control secondary body dimensions, the *European standard to sizing system. Size designation of clothes. Part 3: Measurements and intervals* [59] (EN 13402-3-2004) has been used. This Standard is drawn up by the European Union and it is a set of guidelines for the textile industry. The text, whose compliance is desirable but not obligatory, promotes the im-

plementation of a clothing sizing sytem, adapted to the users, based on the consideration of three body parameters: bust, waist and hip circumference, depending on height.

The comfort and wellbeing feeling and a trendy design are the key elements for consumers to proceed to purchase. A garment can only be comfortable to the wearer if the fit is good. In the case of protective clothing and sportswear, good fit is mandatory to ensure the safety and performance of the user. Clothing fit should be improved with a better garment labelling. Apparel companies should offer consumers a truthful and not confusing information on the garment sizes that they wish to offer for sale, so that people could recognise their size with facility. An understandable labelling system could incorporate pictograms indicating the body measurements associated with the garment, see Fig. 7.1 for an example based on [7]. This labelling proposal could speed up the purchase process.

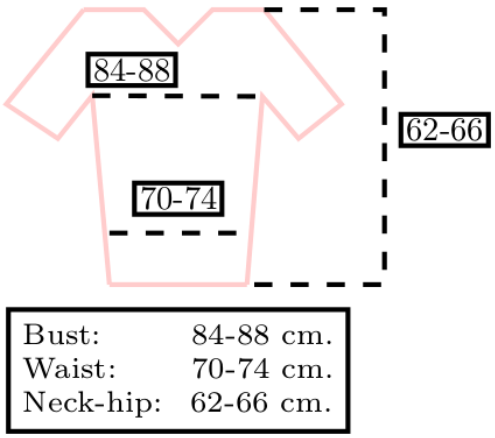


Figure 7.1: A clothing labelling proposal for upper garments, based on [7].

On the other hand, the approaches based on the statistical archetypal analysis allow to identify boundary cases, that is to say, the individuals who present extreme body measurements. The basic idea is that accommodating boundary cases will accommodate the people who fall within the boundaries (less extreme population). This strategy is valuable in all those problems of human-computer interaction, for example, the design of plane cockpits or truck cabins. When designing workstations or evaluating manual work,

it is common to use only a few human models (extreme cases) as virtual test individuals. In this PhD thesis, we have been able to demonstrate that archetypal analysis is a better statistical alternative to determine extreme cases regarding the common used PCA-approach. Unlike PCA, archetypal analysis ensures intended accommodation levels. In addition, the user can decide the number of archetypes to consider or leave the selection by a criterion. In the literature, there is an ongoing discussion about whether the archetypes should be represented by a real observation instead of the standard output, since using the archetypal method, they may or may not be. In some problems, it is crucial that the archetypes are real subjects, observations of the sample, and not fictitious. In this PhD work, a new archetypal concept has been introduced to tackle this problem: the archetypoid. It has been presented an efficient computational algorithm to calculate them and it has been demonstrated some of their advantages regarding classical archetypes. Archetypal and archetypoid analysis could improve industry practice when using human model tools for the design of products and work environments.

All computational algorithms associated with the methods presented in this PhD report have been gathered together into an R package called **Anthropometry**, which is freely available on the Comprehensive R Archive Network at http://cran.r-project.org/package=Anthropometry.

Appendix

Bimax algorithm: Theory, results and discussion

Theoretical aspects of *Bimax*

The *Bimax* algorithm was proposed in [164]. It assumes that data can be represented by a binary matrix $A_{n\times m}$, where the element a_{ij} is 1 if row i responds in the variable j, being 0 if not. In this way, a bicluster is a submatrix of A whose elements are all equal to 1. Although each a_{ij} equal to 1 represents a bicluster by itself, they are not considered since the goal of this method is to find the so-called *inclusion-maximal* biclusters, i.e., those ones that are not completely contained in other bicluster.

As a brief summary, *Bimax* works as follows [164, 45]: $A_{n\times m}$ is first divided into two column sets, C_U and C_V, by using a certain row as a template, with at least a prespecified minimum of 1s. C_U includes the columns where this row is 1 and C_V the others. Next, the rows of A are rearranged in the following way.

First come the rows, R_U, that respond only to C_U (i.e. R_U are those rows that contain only 0s in column set C_V). Then, come the rows, R_W, that respond at the same time to C_U and C_V. Finally, the rows, R_V, that respond only to C_V are considered (i.e. R_V are those rows that contain only 0s in C_U). The combination of R_U, R_W and R_V, with C_U and C_V originates the submatrices U, V and W: U is the matrix [rows = R_U + R_W, columns = C_U], V is the matrix [rows = R_W + R_V, columns = ALL] and W is the matrix [rows = R_U, columns = C_V] that contains only 0s and therefore can be deleted.

The algorithm is recursively applied to U and V until a matrix with only 1's (a bicluster) is identified. Fig. A1 helps to understand this procedure. Blue squares represent the 1s.

Application and results

As mentioned, *Bimax* requires an array of binary data. The **biclust** R package includes a function to binarize the data using a threshold.

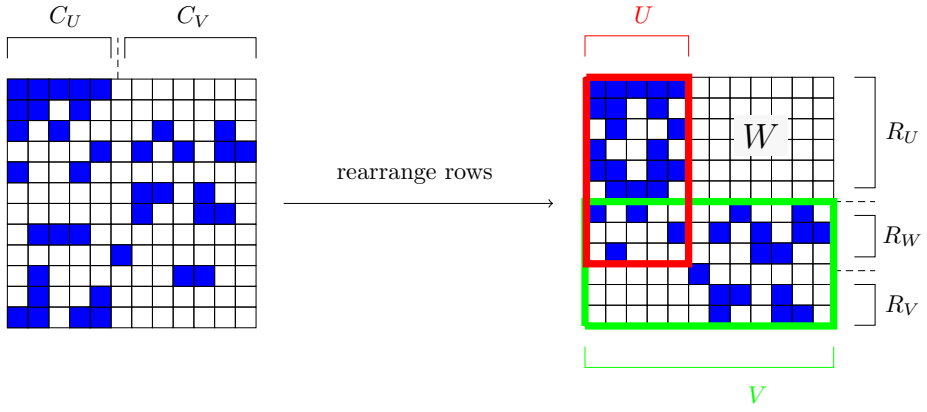


Figure A1: Illustration of the Bimax algorithm based on [164].

Values over threshold will be set to 1, the rest to 0. However, this alternative is not interesting in anthropometric terms because the original dimensions would turn into ones and zeros, losing its true value. In [45], a very interesting alternative is provided. *Bimax* is applied to binary and multiple choice data to identify market segments among tourists with similar hobbies. In order to apply a similar approach to the whole Spanish anthropometric database, the qualitative variables that are binary and multiple choice must be identified. After checking all the variables, the only one that met both conditions was the list of missing foods for women who claim not to follow a varied diet. We will use *Bimax* with this variable. In this way, we will not use *Bimax* aimed at apparel sizing and desing, but at exploring its utility with a particular sociological variable related to eating habits. For practical purposes, these results may be relevant in dietetics or nutrition. For example, by interpreting the groups returned by this method, nutrition experts can plan a type of diet for different customers, depending on the food they do not eat.

First, we must turn the selected variable into a binary variable. To do this, we define as columns each listing food and frame with a 1 those ones where each woman says that they are not included in her diet. Otherwise, frame them with a 0 (see Table A1 for a particular example). Table A1 displays a sample of the first 7 women between 20 and 24 years. For instance, ABAD116 says that meat, pasta, rice and potatoes are often missing in her diet.

Woman code	Legumes	Meat	Eggs	Fish	Pasta	Rice	Potatoes	...
ABAD025	1	0	0	0	0	0	0	...
ABAD098	1	0	0	0	0	0	0	...
ABAD114	1	0	0	0	0	0	0	...
ABAD116	0	1	0	0	1	1	1	...
ABAD117	0	0	1	0	0	0	0	...
ALCA027	0	1	0	0	0	0	0	...
ALCA061	1	1	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...

Table A1: Sample of the first 7 women between 20 and 24 years that make up of the database on which the algorithm Bimax applies.

Fig. A2 represents the percentage of women who do not feed of varied way, considering the whole database (left) and segmenting by age groups (right).

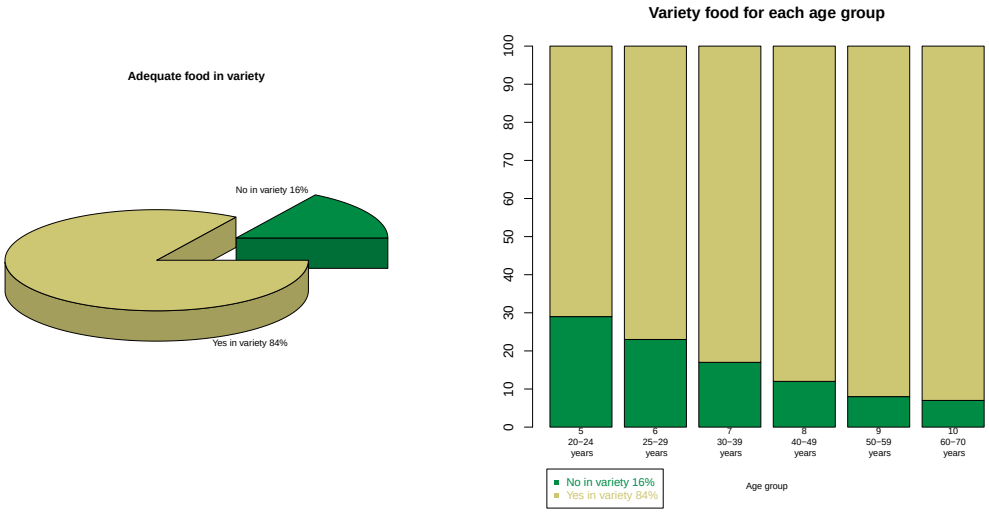


Figure A2: Descriptive plots showing responses of women participating in the study related to their eating habits, with the entire database (left) and segmented by age groups (right).

Fig. A3 shows missing foods in the diet of women who claim not to follow a varied diet, considering again the whole database (left) and segmenting by age groups (right). Each bar of the plots of Fig. A3 corresponds with each

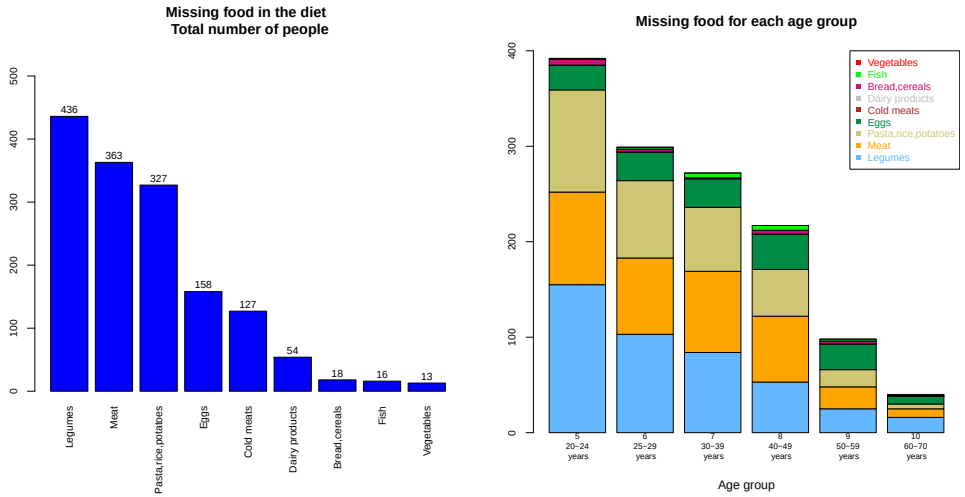


Figure A3: Descriptive plots showing missing foods in the diet of women who claim not to follow a varied diet, both with the entire database (left), such as segmented by age groups (right).

food and includes the total number of women who claim that that food is missing in their diet, which may be the only one or one among others.

With this particular variable, it is interesting to work segmenting the individuals according to the age group to which they belong. Right plots of Figs. A2 and A3 show that the greatest number of women who claim not to follow a varied diet are those aged between 20 and 24 years (in total they are 267 women). Therefore, as an example of the performance of this algorithm, we are going to apply *Bimax* only to this age group.

In order to use *Bimax* within the function *biclust*, we set in its method-argument the option *BCrepBimax()*. It is a modification of the original *Bimax* algorithm to avoid overlapping biclusters. We fixed that the minimum number of women who must belong to every bicluster is 5. Table A2 shows the results.

Number of identified biclusters: 7

	BC 1	BC 2	BC 3	BC 4	BC 5	BC 6	BC 7
Number of rows (women):	6	22	11	17	8	42	12
Number of columns:	7	5	5	4	4	3	3

Table A2: Results of the *Bimax* algorithm for the age group [20,24[years.

As shown in Table A2, there are 7 clusters with more than 5 women. This constraint can be obviously changed, but in this case it was set in this way, since the returned groups contain a 44% (118 women) of the women in this age group (Bimax is a nonexhaustive algorithm). This is the highest percentage of women who are grouped under any restriction. Furthermore, it is observed that these obtained groups mostly contain a small number of women for a number of missing foods in the diet. This is because any biclustering method calculates groups in a very restrictive way, since all the elements of a same group should show the same behavior in all the variables that belong to that group. Results are reproducible. Fig. A4 is a bicluster membership graph and shows the foods that are present in each one of the computed biclusters.

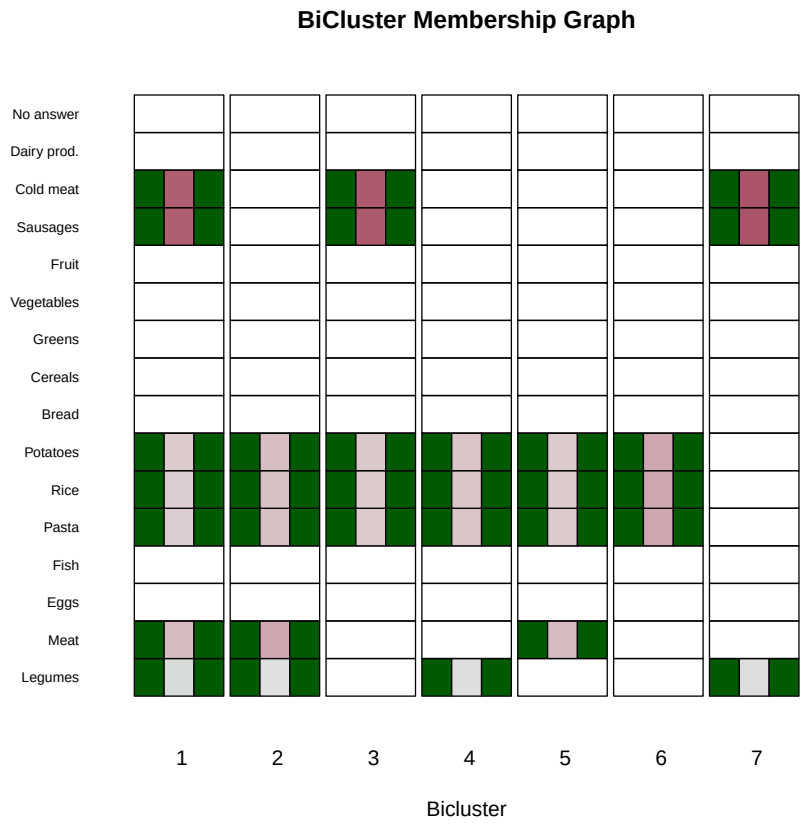


Figure A4: Foods in each bicluster.

In a bicluster membership plot, variables belonging to a bicluster appear with a different color (more or less purple). For each bicluster, the darker the color purple for a particular food is, the stronger the common pattern for women who belong to that bicluster regarding that food is (green color of this plot is non-informative). For example, the most common women in the biclusters 1, 3 and 7 are those that do not eat cold meats or sausage. However, in the bicluster 3, there are no women who do not eat meat or legumes and in the bicluster 7 there are no women who do not eat meat.

Discussion

The *Bimax* algorithm represents a very interesting approach to define biclusters using variables related to sociological aspects of a particular population. We have developed a similar analysis than in [45]. The only multiple choice variable of our anthropometric database refers to the foods that are lacking in the diet of the women who claim not to follow a varied diet. This type of variable does not have a self-interest in the fields of Ergonomics and Anthropometry, although it can be very useful for other sociological studies that relate the spending habits and health of the population. A possible alternative to implement this algorithm on qualitative variables of single answer of the anthropometric database, would join the categories of variables that have some kind of relationship among them in a same database. For example, it could be interesting to cross the women's answers to the questions *"Are you satisfied with your body?"*, *"Do you get angry in the changing room when you can't find clothes that fit you well?"* and *"Do you have problems to find your size?"*.

Bibliography

[1] Abonyi, J., and Feil, B. *Cluster Analysis for Data Mining and System Identification*. Birkhäuser Verlag AG, 2007.

[2] Adams, D. C., and Otárola Castillo, E. `geomorph`: an R package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution 4* (2013), 393–399.

[3] Alemany, S., González, J. C., Nácher, B., Soriano, C., Arnáiz, C., and Heras, H. 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, 2010), pp. 1–10.

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

[5] Arenas, C., and Cuadras, M. Recent statistical methods based on distances. *Contributions to Science 2*, 2 (2002), 183–191.

[6] Ashdown, S. & Loker, S. Improved apparel sizing: Fit and anthropometric 3D scan data. Tech. rep., National Textile Center Annual Report, November 2005.

[7] Ashdown, S. P. *Sizing in clothing: Developing effective sizing systems for ready-to-wear clothing*. Woodhead Publishing in Textiles, 2007.

[8] Ashdown, S. P., Loker, S., Schoenfelder, K., and Lyman-Clarke, L. Using 3D Scans for Fit Analysis. *Journal of Textile and Apparel, Technology and Management 4*, 1 (2004), 1–12.

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

[10] Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and Zitzler, E. Bicat: a biclustering analysis toolbox. *Bioinformatics 22* (2006), 1282–1283.

[11] Barnett, V. The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General) 139* (1976), 319–354.

[12] Batra, A. Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms. In *5th IEEE International Conference on Advanced Computing & Communication Technologies* (2011), pp. 274–279.

[13] Beliakov, G., Pradera, A., and Calvo, T. *Aggregation Functions: A Guide for Practitioners*. Studies in Fuzziness and Soft Computing, Volume 221, Springer, 2007.

[14] Bertilsson, E., Högberg, D., and Hanson, L. Using experimental design to define boundary manikins. *Work: A Journal of Prevention, Assessment and Rehabilitation 41*, Supplement 1 (2012), 4598–4605.

[15] Bisson, G., and Hussain, F. χ-sim: A New Similarity Measure for the Co-clustering Task. In *Seventh International Conference on Machine Learning and Applications* (2008), pp. 211–217.

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

[17] Bittorf, V., Recht, B., Re, C., and Tropp, J. A. Factoring nonnegative matrices with linear programs. In *In Neural Information Processing Systems (NIPS)* (2012), pp. 1–17.

[18] Blanchonette, P. Jack Human Modelling Tool: A Review. Tech. Rep. DSTO-TR-2364, Defence Science and Technology Organisation (Australia). Air Operations Division, 2010.

[19] Bock, H.-H. Origins and extensions of the k-means algorithm in cluster analysis. *Electronic Journal for History of Probability and Statistics 4*, 2 (2008), 1–18.

[20] Brolin, E. Consideration of anthropometric diversity - Methods for virtual product and production development. Thesis for the degree of licentiate of engineering. Department of Product and Production Development. Chalmers University of Technology. Gothernburg, Sweden, 2012.

[21] Brown, A. S. Role models. *Mechanical Engineering 121*, 7 (1999), 44–49.

[22] Burr, M. A., Rafalin, E., and Souvaine, D. L. Simplicial depth: an improve definition analysis, and efficiency for the finite sample case. *Data depth: robust multivariate analysis 72* (2006), 195–209.

[23] Busygin, S., Prokopyev, O., and Pardalos, P. M. Biclustering in data mining. *Computers and Operations Research 35* (2008).

[24] Bye, E., and McKinney, E. Fit analysis using live and 3D scan models. *International Journal of Clothing Science and Technology 22*, 2 (2010), 88–100.

[25] Canhasi, E., and Kononenko, I. Multi-document summarization via Archetypal Analysis of the content-graph joint model. *Knowledge and Information Systems* (2013), 1–22.

[26] Canhasi, E., and Kononenko, I. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications 41*, 2 (2014), 535 – 543.

[27] Cascos, I. *Data Depth: Multivariate Statistics and Geometry. In New Perspectives in Stochastic Geometry.* Oxford Scholarship Online. Kendall, W.S. and Molchanov, I., 2010, ch. Part III, 12, pp. 398–423.

[28] Chan, B. H. P., Mitchell, D. A., and Cram, L. E. Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society 338* (2003).

[29] Chaudhuri, P. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association 91* (1996), 862–872.

[30] Chen, C.-M. Fit evaluation within the made-to-measure process. *International Journal of Clothing Science and Technology 19*, 2 (2007), 131–144.

[31] Chen, Y., Zeng, X., Happiette, M., Bruniaux, P., Ngb, R., and Yu, W. Optimisation of garment design using fuzzy logic and sensory evaluation techniques. *Engineering Applications of Artificial Intelligence 22* (2009), 272–282.

[32] Cheng, Y., and Church, G. M. Biclustering of expression data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology 8* (2000), 93–103.

[33] Christmann, A. Classification Based on the Support Vector Machine and on Regression Depth. *Statistics in Industry and Technology: Statistical Data Analysis* (2002), 341–352.

[34] Chung, M.-J., Lin, H.-F., and Wang, M.-J. J. The development of sizing systems for Taiwanese elementary- and high-school students. *International Journal of Industrial Ergonomics 37* (2007), 707–716.

[35] Claude, J. *Morphometrics with R.* Use R! Springer, 2008.

[36] Cuesta-Albertos, J. A., and Nieto-Reyes, A. The random Tukey depth. *Computational Statistics and Data Analysis 52* (2008), 4979–4988.

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

[38] D'Apuzzo, N. *Recent Advances in 3D full body scanning with applications to fashion and apparel.* Gruen, A., Kahmen, H. (eds.). Optical 3-D Measurement Techniques IX, Vienna, Austria, 2009.

[39] De Raeve, A., De Smedt, M., and Bossaer, H. Mass customization, business model for the future of fashion industry. In *3rd Global Fashion International Conference* (2012), pp. 1–17.

[40] D'Esposito, M. R., Palumbo, F., and Ragozini, G. Interval Archetypes: A New Tool for Interval Data Analysis. *Statistical Analysis and Data Mining 5*, 4 (2012), 322–335.

[41] Deza, M.-M., and Deza, E. *Dictionary of Distances*. Elsevier, 2006.

[42] Dhillon, I. S. Co-clustering documents and words using Bipartite Spectral Graph Partitioning. In *Proceedings of the Seventh ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD'01)* (2001), pp. 269–274.

[43] Dhillon, I. S., Mallela, S., and Modha, D. S. Information-Theoretic Co-clustering. In *Proceedings of the Ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD '03)* (2003), pp. 89–98.

[44] Ding, Y., Dang, X., Peng, H., and Wilkins, D. Robust clustering in high dimensional data using statistical depths. *BMC Bioinformatics 8*, Suppl 7:S8 (2007), 1–16.

[45] Dolnicar, S., Kaiser, S., Lazarevski, K., and Leisch, F. Biclustering: Overcoming Data Dimensionality Problems in Market Segmentation. *Journal of Travel Research* (2011).

[46] Donoho, D. L., and Gasko, M. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics 20*, 4 (1992), 1803–1827.

[47] Dryden, I. E., and Mardia, K. V. *Statistical Shape Analysis*. John Wiley & Sons, 1998.

[48] Dryden, I. L. `shapes` *package*. R Foundation for Statistical Computing, Vienna, Austria, 2012. Contributed package.

[49] Dudek, M. W. A. `clusterSim`*: Searching for optimal clustering procedure for a data set*, 2011. R package version 0.40-6.

[50] Dutta, D., and Ghosh, A. K. On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics 64*, 3 (2012), 657–676.

[51] Dyckerhoff, R., Koshevoy, G., and Mosler, K. Zonoid data depth: Theory and computation. In *A. Prat, ed., COMPSTAT 1996. Proceedings in Computational Statistics* (Heidelberg, 1996), Physica-Verlag, pp. 235–240.

[52] Elhamifar, E., Sapiro, G., and Vidal, R. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012), pp. 1–8.

[53] Epifanio, I. h-plots for displaying nonmetric dissimilarity matrices. *Statistical Analysis and Data Mining 6*, 2 (2013), 136–143.

[54] Epifanio, I., Vinué, G., and Alemany, S. Archetypal analysis: Contributions for estimating boundary cases in multivariate accommodation problem. *Computers & Industrial Engineering 64* (2013), 757–765.

[55] Eugster, M. J. A. Performance Profiles based on Archetypal Athletes. *International Journal of Performance Analysis in Sport 12*, 1 (2012), 166–187.

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

[57] Eugster, M. J. A., and Leisch, F. Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis 55*, 3 (2011), 1215–1225.

[58] European Committee for Standardization. Size designation of clothes. Part 2: Primary and secondary dimensions, 2002.

[59] European Committee for Standardization. Size designation of clothes. Part 3: Measurements and intervals, 2005.

[60] Fan, J., Yu, W., and Hunter, L. *Clothing appearance and fit: Science and technology*. Woodhead Publishing in Textiles, 2004.

[61] Fashion Technology, Size & Fit Solutions. *UK National Sizing Survey Information Document*, 2001-2. `http://www.sizemic.eu/products-and-services/46-size-survey-data`.

[62] Faust, M. E., and Carrier, S. 3D body scanning's contribution to the use of apparel as an identity construction tool. In *ICDHM '09 Proceedings of the 2nd International Conference on Digital Human Modeling: Held as Part of HCI International* (2009), pp. 19–28.

[63] Filzmoser, P., Garrett, R. G., and Reimann, C. Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences 31* (2005), 579–587.

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

[65] Fletcher, R. *Practical Methods of Optimization*, second ed. John Wiley & Sons, 2000.

[66] Forgy, E. W. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics 21* (1965), 768–769.

[67] Fox, J., and Weisberg, S. *An R Companion to Applied Regression*, second ed. Sage, Thousand Oaks CA, 2011.

[68] Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincare. Probabilités et Statistiques 10*, 4 (1948), 215–310.

[69] Frey, B. J., and Dueck, D. Clustering by Passing Messages Between Data Points. *Science 315* (2007), 972–976.

[70] Friess, M. Multivariate Accommodation Models using Traditional and 3D Anthropometry. Tech. rep., SAE, 2005.

[71] Friess, M., and Bradtmiller, B. 3D Head Models for Protective Helmet Development. Tech. rep., SAE, 2003.

[72] Fritz, H., García Escudero, L. A., and Mayo-Iscar, A. An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software 47*, 12 (2012), 1–26.

[73] Galyardt, A. *Interpreting mixed membership models: Implications of Erosheva's representation theorem.* In Edoardo M. Airoldi, David Blei, Elena Erosheva, and Stephen E. Fienberg, editors. Handbook of Mixed Membership Models., Chapman and Hall, in press, 2014.

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

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

[76] Gasko, M., and Donoho, D. L. Multivariate generalization of the median and trimmed mean, I. Tech. Rep. 133, Department of Statistics. University of California, Berkeley, December 1987.

[77] Gaul, W., and Schader, M. *A New Algorithm for Two-Mode Clustering.* Data Analysis and Information Systems, 1996, ch. Section 1, pp. 15–23.

[78] Genest, M., Masse, J.-C., and Plante, J.-F. `depth`*: Depth functions tools for multivariate analysis*, 2012. R package version 2.0-0.

[79] Georgescu, V. Clustering of Fuzzy Shapes by Integrating Procrustean Metrics and Full Mean Shape Estimation into K-Means Algorithm. In *IFSA-EUSFLAT Conference* (20th-24th July 2009), pp. 1679–1684.

[80] Gestraud, P. *Biclustering Analysis and Results Exploration*, 2008. R package version 1.16.0.

[81] Gillis, N. Robustness Analysis of Hottopixx, a Linear Programming Model for Factoring Nonnegative Matrices. *SIAM Journal on Matrix Analysis and Applications 34*, 3 (2013), 1189–1212.

[82] Goodall, C. Procrustes Methods in the Statistical Analysis of Shape. *Journal of the Royal Statistical Society, Series B (Methodological) 53*, 2 (1991), 285–339.

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

[84] Gschwandtner, M., and Filzmoser, P. `mvoutlier`*: Multivariate outlier detection based on robust methods*, 2011. R package version 1.8.

[85] Gupta, D., and Gangadhar, B. R. A statistical model for developing body size charts for garments. *International Journal of Clothing Science and Technology 16*, 5 (2004), 458–469.

[86] Gupta, D., Garg, N., Arora, K., and Priyadarshini, N. Developing body measurement charts for garments manufacture based on a linear programming approach. *Journal of Textile and Apparel Technology and Management 5*, 1 (2006), 1–13.

[87] Han, H., and Nam, Y. Automatic body landmark identification for various body figures. *International Journal of Industrial Ergonomics 41*, 6 (2011), 592–606.

[88] Han, H., Nam, Y., and Choi, K. Comparative analysis of 3D body scan measurements and manual measurements of size Korea adult females. *International Journal of Industrial Ergonomics 40* (2010), 530–540.

[89] Harrell, F. E. J. `Hmisc`*: Harrell Miscellaneous*, 2013. R package version 3.13-0.

[90] Hartigan, J. A. Direct clustering of a data matrix. *Journal of the American Statistical Association 67*, 337 (1972), 123–129.

[91] Hartigan, J. A., and Wong, M. A. A k-Means Clustering Algorithm. *Applied Statistics* (1979), 100–108.

[92] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning. Data mining, inference and prediction.* 2nd ed., Springer-Verlag, 2009.

[93] HFES 300 Committee. *Guidelines for Using Anthropometric Data in Product Design*. Human Factors and Ergonomics Society, 2004.

[94] Hoberg, R. *Cluster Analysis Based on Data Depth*. Data Analysis, Classification, and Related Methods. Springer, 2000, ch. Part I, pp. 17–22.

[95] Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., and Clevert, D.-A. FABIA: Factor Analysis for Bicluster Acquisition. *Bioinformatics 26*, 12 (2010), 1520–1527.

[96] Hsu, C.-H. Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry. *Expert Systems with Applications 36* (2009), 4185–4191.

[97] Hsu, C.-H. Developing accurate industrial standards to facilitate production in apparel manufacturing based on anthropometric data. *Human Factors and Ergonomics in Manufacturing 19*, 3 (2009), 199–211.

[98] Hsu, C.-H., and Wang, M.-J. J. Using decision-tree based data mining to establish a sizing system for the manufacture of garments. *The International Journal of Advanced Manufacturing Technology 26* (2005), 669–674.

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

[100] Human Factors and Ergonomics Society and HFES 300 Committee. *Guidelines For Using Anthropometric Data In Product Design*. Human Factors and Ergonomics Society, 2004.

[101] Ibañez, M. V., Simó, A., Domingo, J., Durá, E., Ayala, G., Alemany, S., Vinué, G., and Solves, C. 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* (Vilamoura, Algarve, Portugal, 2012), vol. 1, pp. 370–374.

[102] Ibáñez, M. V., Vinué, G., Alemany, S., Simó, A. Epifanio, I., Domingo, J., and Ayala, G. Apparel sizing using trimmed PAM and OWA operators. *Expert Systems with Applications 39*, 12 (2012), 10512–10520.

[103] International Organization for Standardization. Standard Sizing Systems for Clothes. International Organization for Standardization, Geneva, 1991.

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

[105] Irigoien, I., Mestres, F., and Arenas, C. The Depth Problem: Identifying the Most Representative Units in a Data Group. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 10*, 1 (2013), 161–172.

[106] Irigoien, I., Sierra, B., and Arenas, C. `ICGE`: an R package for detecting relevant clusters and atypical units in gene expression. *Bioinformatics 13*, 30 (2012), 1–11.

[107] Istook, C. L., and Hwang, S.-J. 3D body scanning systems with application to the apparel industry. *Journal of Fashion Marketing and Management 5* (2001), 120–132.

[108] Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31* (2010), 651–666.

[109] Jörnsten, R. Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis 90* (2004), 67–89.

[110] Jörnsten, R., Vardi, Y., and Zhang, C. H. A robust clustering method and visualization tool based on data depth. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods (Neuchâtel, 2002), Statistics for Industry and Technology* (2002), pp. 353–366.

[111] Jung, K., Kwon, O., and You, H. Evaluation of the multivariate accommodation performance of the grid method. *Applied Ergonomics 42* (2010), 156–161.

[112] Kaiser, S., and Leisch, F. A Toolbox for Bicluster Analysis in R. Tech. rep., Department of Statistics (University of Munich), 2008.

[113] Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., Theron, R., Quintales, L., and Leisch, F. `biclust`*: BiCluster Algorithms*, 2011. R package version 1.0.1.

[114] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. An Efficient k-means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 7 (2002), 881–892.

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

[116] Kendall, D. G. The diffusion of shape. *Advances in Applied Probability 9* (1977), 428–430.

[117] Kendall, D. G. *The shape of Poisson-Delaunay triangles*. In: Studies in Probabilities and related topics, Demetrescu M. C. and Iosifescu M., editors, Nagard, Montreal, 1983.

[118] Kendall, D. G. Shape manifolds, Procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society 16* (1984), 81–121.

[119] Kennedy, K. W. *Anthropometric accommodation in aircraft cockpits*, June 2001.

[120] Kennedy, K. W., and Zehner, G. F. Assessment of anthropometric accommodation in aircraft cockpits. *SAFE Journal 25*, 1 (1995), 51–57.

[121] Kent, J. T. The Complex Bingham Distribution and Shape Analysis. *Journal of the Royal Statistical Society. Series B (Methodological) 56*, 2 (1994), 285–299.

[122] Kiel, S., and van der Meulen, P. Benefits and advantages of ergonomic studies in digital 3D. Tech. rep., Human Solutions GmbH, Kaiserlautern, 2007.

[123] Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Research 13*, 4 (2003), 703–716.

[124] Lange, T., Mosler, K., and Mozharovskyi, P. Fast nonparametric classification based on data depth. *Statistical Papers 5* (2012), 1–22.

[125] Lawson, C. L., and Hanson, R. J. *Solving Least Squares Problems*. Prentice Hall, 1974.

[126] Lazzeroni, L., and Owen, A. Plaid models for gene expression data. *Statistica Sinica 12* (2002), 61–86.

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

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

[129] Li, J., Cuesta-Albertos, J. A., and Liuc, R. Y. DD-Classifier: Nonparametric Classification Procedure Based on DD-plot. *Journal of the American Statistical Association 107*, 498 (2012), 737–753.

[130] Li, S., Wang, P., Louviere, J., and Carson, R. Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals. In *ANZMAC 2003 Conference Proceedings* (December 2003), pp. 1674–1679.

[131] Liu, R. Y. On a notion of data depth based on random simplices. *The Annals of Statistics 18* (1990), 405–414.

[132] Liu, R. Y., Parelius, J. M., and Singh, K. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics 27*, 3 (1999), 783–858.

[133] Liu, X., and Zuo, Y. *Computing projection depth and its associated estimators*. Statistics and Computing. Springer., 2012.

[134] Lloyd, S. P. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory 28*, 2 (1982), 129–137.

[135] López, A. *Similaridad y Contraste Mediante Profundidad Estadística*. PhD thesis, Departamento de Estadística, Universidad Carlos III de Madrid (Spain), 2010.

[136] López, A., and Romo, J. Simplicial similarity and its application to hierarchical clustering. Tech. rep., Universidad Carlos III de Madrid. Departamento de Estadística. Working papers. Statistics and Econometrics, 2010.

[137] Lu, J.-M., and Wang, M.-J. J. Automated anthropometric data collection using 3D whole body scanners. *Expert Systems with Applications 35* (2008), 407–414.

[138] Luximon, A., Zhang, Y., Luximon, Y., and Xiao, M. Sizing and grading for wearable products. *Computer-Aided Design 44* (2012), 77–84.

[139] MacQueen, J. Some methoods for classification and analysis of mulivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Statistical Laboratory of the University of California, Berkeley, 1965/66), vol. 1, pp. 281–297.

[140] Madeira, S. C., and Oliveira, A. L. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE Transactions on Computational Biology and Bioinformatics 1* (2004), 24–45.

[141] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. `cluster`*: Cluster Analysis Basics and Extensions*, 2012. R package version 1.14.3 — For new features, see the "Changelog" file (in the package source).

[142] Magnenat-Thalmann, N., Ed. *Modeling and simulating bodies and garments*. Springer, 2010.

[143] Mahalanobis, P. C. On the generalized distance in statistics. In *Proceedings National Institute of Science, India* (1936), vol. 12, pp. 49–55.

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

[145] Marcus, L. F. Traditional morphometrics. In *Proceedings of the Michigan Morphometrics Workshop* (1990), pp. 77–122.

[146] Mardia, K. V., Kent, J. T., and Bibby, J. M. *Multivariate Analysis*. Academic Press, 1979.

[147] McCulloch, C. E., Paal, B., and Ashdown, S. P. An optimization approach to apparel sizing. *Journal of the Operational Research Society 49* (1998), 492–499.

[148] Midgley, D., and Venaik, S. Marketing strategy in MNC subsidiaries: pure versus hybrid archetypes. In *P. McDougall-Covin and T. Kiyak, Proceedings of the 55th Annual Meeting of the Academy of International Business* (2013), pp. 215–216.

[149] Mirkin, B. *Mathematical Classification and Clustering*. Kluwer Academic Publishers, 1996.

[150] Mohamed, S., Heller, K. A., and Ghahramani, Z. *A simple and general exponential family framework for partial membership and factor analysis*. In Edoardo M. Airoldi, David Blei, Elena Erosheva, and Stephen E. Fienberg, editors. Handbook of Mixed Membership Models., Chapman and Hall, in press, 2014.

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

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

[153] Murali, T. M., and Kasif, S. Extracting conserved gene expression motifs from gene expression. *Pacific Symposium on Biocomputing 8* (2003), 77–88.

[154] Nazeer, K. A. A., and Sebastian, M. P. Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. In *Proceedings of the World Congress on Engineering* (July 2009), pp. 1–5.

[155] Ng, R. and Ashdown, S. P. and Chan, A. Intelligent size table generation, 2007.

[156] Oja, H. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters 1* (1983), 327–332.

[157] Outling, C. D. S. *Process, fit and appearance analysis of three-dimensional to two-dimensional automatic pattern unwrapping technology*. PhD thesis, Graduate Faculty of North Carolina State University, 2007.

[158] Pan, J. Z., Jörnsten, R., and Hart, R. P. Screening anti-inflammatory compounds in injured spinal cord with microarrays: a comparison of bioinformatics analysis approaches. *Physiological Genomics 17* (2004), 201–214.

[159] Paquet, E., and Rioux, M. Nefertiti: A tool for 3-D shape databases management. *SAE transactions 108*, 1 (1999), 387–393.

[160] Paquet, E., Robinette, K. M., and Rioux, M. Management of three-dimensional and anthropometric databases: Alexandria and Cleopatra. *Journal of Electronic Imaging 9*, 4 (2000), 421–431.

[161] Parkinson, M. B., Reed, M. P., Kokkolaras, M., and Papalambros, P. Y. Optimizing Truck Cab Layout for Driver Accommodation. *Journal of Mechanical Design 129*, 11 (2006), 1110–1117.

[162] Pheasant, S. *Bodyspace: Anthropometry, Ergonomics and the Design of Work*. Taylor & Francis, Ltd, 2003.

[163] Porzio, G. C., Ragozini, G., and Vistocco, D. On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry 24* (2008), 419–437.

[164] Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bhlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics 22*, 9 (2006), 1122–1129.

[165] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[166] Rasband, J. A., and Liechty, E. L. *Fabulous Fit: Speed Fitting and Alteration*. 2nd ed., Fairchild Publication, New York, USA, 2006.

[167] Richards, J. W., Lee, A. B., Schafer, C. M., and Freeman, P. E. Prototype selection for parameters in complex models. *The Annals of Applied Statistics 6*, 1 (2012), 383–408.

[168] Robinette, K., Blackwell, S., Daanen, H., Boehmer, M., Fleming, S., Brill, T., Hoeferlin, D., , and Burnsides, D. Civilian American and European surface anthropometry resource (CAESAR), final report, volume I: Summary. Tech. rep., Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, 2255 H Street, Wright-Patterson AFB OH 45433-7022 and Society of Automotive Engineers International, 400 Commonwealth Drive, Warrendale PA, 15096, 2002.

[169] Robinette, K. M. *An investigation of 3-D anthropometric shape descriptors for database mining*. PhD thesis, Division of Epidemiology and Biostatistics of the Department of Environmental Health of the College of Medicine, 2003.

[170] Robinette, K. M. Maximizing Anthropometric Accommodation and Protection. Tech. rep., Biosciences and Protection Division Biomechanics Branch. Air Force Research Laboratory, 2007.

[171] Robinette, K. M., Daanen, H., and Paquet, E. The Caesar project: A 3-D surface anthropometry survey. In *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on* (Ottawa, Canada, 1999), pp. 380–386.

[172] Robinette, K. M., and McConville, J. T. Alternative to Percentile Models. Tech. rep., SAE, 1981.

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

[174] Rohlf, J. F. Shape Statistics: Procrustes Superimpositions and Tangent Spaces. *Journal of Classification 16* (1999), 197–223.

[175] Rothwell, P. L., and Hickey, D. T. Three-dimensional computer models of man. In *Proceedings of the Human Factors Society 30th Annual Meeting* (1986), pp. 216–220.

[176] Rousseeuw, P. J., and Hubert, M. Regression Depth. *Journal of the American Statistical Association 94*, 446 (1999), 388–433.

[177] Rousseeuw, P. J., and Ruts, I. Algorithm as 307: Bivariate location depth. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 45*, 4 (1996), 516–526.

[178] Rousseeuw, P. J., and Struyf, A. Computing location depth and regression depth in higher dimensions. *Statistical Computation 8* (1998), 193–203.

[179] Salusso-Deonier, C. J., DeLong, M. R., Martin, F. B., and Krohn, K. R. A multivariate method of classifying body form variation for women's apparel. *Clothing and Textiles Research Journal 4*, 1 (1985-1986), 38–45.

[180] Santos, F. `AnthropMMD`*: A GUI for Mean Measures of Divergence*, 2013. R package version 0.9.5.

[181] Sarle, W. S. *The VARCLUS Procedure. SAS/STAT User's Guide*, 4th ed., 1990.

[182] Schlager, S. `Morpho`*: Calculations and visualisations related to Geometric Morphometrics*, 2013. R package version 0.25-1.

[183] Seber, G. A. F. *A Matrix Handbook for Statisticians*. Wiley Series in Probability and Statistics, John Wiley & Sons, 2008.

[184] Seiler, C., and Wohlrabe, K. Archetypal scientists. *Journal of Informetrics 7* (2013), 345–356.

[185] Shu, C., Wuhrer, S., and Xi, P. Geometric and Statistical Methods for Processing 3D Anthropometric Data. In *International Symposium on Digital Human Modeling* (2011), pp. 1–5.

[186] Shuyong, L., Yan, C., Ming, Y., and Rui, D. Bicluster Algorithm and Used in Market Analysis. In *Second International Workshop on Knowledge Discovery and Data Mining* (2009), pp. 504–507.

[187] Simmons, K., Istook, C. L., and Devarajan, P. Female figure identification technique (FFIT) for apparel. part I: Describing female body shapes. *Journal of Textile and Apparel, Technology and Management 4* (2004), 1–16.

[188] Simmons, K. P. *Body measurement techniques: A comparison of three-dimensional body scanning and physical anthropometric methods.* PhD thesis, North Carolina State University, Textile Technology Management, 2002.

[189] Simmons, K. P., and Istook, C. L. Body measurement techniques: Comparing 3D body-scanning and anthropometric methods for apparel applications. *Journal of Fashion Marketing and Management 7*, 3 (2003), 306–332.

[190] U.S anthropometric survey. Size USA. `http://www.tc2.com/sizeusa.html`, 2004.

[191] Slice, D. E. Modern Morphometrics. In *Modern Morphometrics in Physical Anthropology*, Developments in Primatology: Progress and Prospects. Springer US, 2005, pp. 1–45.

[192] Slice, D. E., Bookstein, F. L., Marcus, L. F., and Rohlf, F. J. *A Glossary for Geometric Morphometrics*, 2002.

[193] Song, H. K., and Ashdown, S. P. An Exploratory Study of the Validity of Visual Fit Assessment From Three-Dimensional Scans. *Clothing and Textiles Research Journal 28*, 4 (2010), 263–278.

[194] Steinley, D. k-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology 59* (2006), 1–34.

[195] Stone, E. Exploring archetypal dynamics of pattern formation in cellular flames. *Physica D 161* (2002), 163–186.

[196] Stoyan, L. A., and Stoyan, H. *Fractals, Random Shapes and Point Fields.* John Wiley and Sons, 1995.

[197] Struyf, A., and Rousseeuw, P. J. High-dimensional computation for the deepest location. *Computational Statistics and Data Analysis 34* (2000), 415–436.

[198] Tanay, A., Sharan, R., and Shamir, R. Biclustering Algorithms: A Survey. Handbook of bioinformatics, 2004.

[199] Telgarsky, M., and Vattani, A. Hartigan's Method: k-means Clustering without Voronoi. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* (Sardinia, Italy, 2010), pp. 820–827.

[200] Theodosiou, T., Kazanidis, I., Valsamidis, S., and Kontogiannis, S. Courseware usage archetyping. In *Proceedings of the 17th Panhellenic Conference on Informatics* (New York, NY, USA, 2013), PCI '13, ACM, pp. 243–249.

[201] Thurau, C., and Bauckhage, C. Archetypal Images in Large Photo Collections. In *Proceedings of the 3rd IEEE International Conference on Semantic Computing* (2009), pp. 129–136.

[202] Toro Ibacache, M. V., Manriquez Soto, G., and Suazo Galdames, I. Geometric Morphometry and the Biologic Shapes Study: From the Descriptive Morphology to the Quantitative Morphology. *International Journal of Morphology 28*, 4 (2010), 977–990. In Spanish.

[203] Torrente, A., and Romo, J. Refining k-means by data depth and bootstrap. Departamento de Estadística, Universidad Carlos III, Madrid, Spain, 2008.

[204] Tryfos, P. An integer programming approach to the apparel sizing problem. *The Journal of the Operational Research Society 37*, 10 (1986), 1001–1006.

[205] Tukey, J. W. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (Montreal, 1975), C. M. Congress, Ed., pp. 523–531.

[206] Turner, H., Bailey, T., and Krzanowski, W. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis 48*, 2 (2005), 235–254.

[207] van der Laan, M. J., and Pollard, K. S. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference 117* (2003), 275–303.

[208] Van Mechelen, I., Bock, H. H., and De Boeck, P. Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research 13*, 5 (2004), 363–394.

[209] Vardi, Y., and Zhang, C.-H. The multivariate l_1-median and associated data depth. *Proceedings of the National Academy of Sciences 97* (2000), 1423–1426.

[210] Vieilledent, G. `phcfM`*: Modelling anthropogenic deforestation*, 2013. R package version 1.2.

[211] Vinué, G. 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), 2012. In Spanish.

[212] Vinué, G. `Anthropometry`: An R package for Analysis of Anthropometric Data. In progress, 2013.

[213] Vinué, G., Epifanio, I., and Alemany, S. Archetypoids: A new approach to define representative archetypal data. Submitted, 2013.

[214] Vinué, G., and Ibañez, M. V. Data depth and Biclustering applied to anthropometric data: Exploring their utility in apparel design. In progress, 2013.

[215] Vinué, G., León, T., Alemany, S., and Ayala, G. Looking for representative fit models for apparel sizing. *Decision Support Systems 57* (2014), 22–33.

[216] Vinué, G., Simó, A., and Alemany, S. The k-means algorithm for 3D shapes with an application to apparel design. Submitted, 2013.

[217] Volino, P., and Magnenat-Thalmann, N. Accurate Garment Prototyping and Simulation. *Computer-Aided Design & Applications 2*, 5 (2005), 645–654.

[218] Wang, M.-J. J., Wu, W.-Y., Lin, K.-C., Yang, S.-N., and Lu, J.-M. Automated anthropometric data collection from three-dimensional digital human models. *The International Journal of Advanced Manufacturing Technology 32* (2007), 109–115.

[219] Weiszfeld, E., and Plastria, F. On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research 167* (2009), 7–41.

[220] Wilcox, R. R. *Basic Statistics: Understanding Conventional Methods and Modern Insights*. Oxford University Press, 2009.

[221] Winks, J. M. *Clothing Sizes: International Standardization*. Textile Institute, 1997.

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

[223] Wit, E., and McClure, J. *Statistics for Microarrays: Inference, Design and Analysis*, 2006. R package version 0.1.

[224] Workman, J. Body measurement specifications for fit models as a factor in clothing size variation. *Clothing and Textiles Research Journal 10*, 1 (1991), 31–36.

[225] Workman, J. E., and Lentz, E. S. Measurement specifications for manufacturers' prototype bodies. *Clothing and Textiles Research Journal 18*, 4 (2000), 251–259.

[226] Xiong, Y., Liu, W., Zhao, D., and Tang, X. Face Recognition via Archetype Hull Ranking. In *IEEE International Conference on Computer Vision (ICCV)* (Sidney, Australia, 2013), pp. 585–592.

[227] Yager, R. R. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics 18* (1988), 183–190.

[228] Zehner, G. F., Meindl, R. S., and Hudson, J. A. A Multivariate Anthropometric Method For Crew Station Design: Abridged. Tech. rep., Kent State University, April 1993.

[229] Zhang, Z., Cui, X., Jeske, D. R., and Borneman, J. Biclustering scatter plots using data depth measures. *Statistical Analysis and Data Mining 6*, 2 (2013), 102–115.

[230] Zheng, R., Yu, W., and Fan, J. Development of a new chinese bra sizing system based on breast anthropometric measurements. *International Journal of Industrial Ergonomics 37* (2007), 697–705.

[231] Zuo, Y. Projection-based depth functions and associated medians. *The Annals of Statistics 31*, 5 (2003), 1460–1490.

[232] Zuo, Y. Multidimensional trimming based on projection depth. *The Annals of Statistics 34*, 5 (2006), 2211–2251.

[233] Zuo, Y., and Lai, S. Exact computation of bivariate projection depth and the Stahel-Donoho estimator. *Computational Statistics and Data Analysis 55* (2011), 1173–1179.

[234] Zuo, Y., and Serfling, R. General notions of statistical depth function. *The Annals of Statistics 28*, 2 (2000), 461–482.