FEMALE BREAST SHAPE CLASSIFICATION BASED ON ANALYSIS OF CIVILIAN AMERICAN AND EUROPEAN SURFACE ANTHROPOMETRY RESOURCE (CAESAR) DATA

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

by

Jie Pei

August 2016

ABSTRACT

This study explored the variation in female breast shape across the younger (age: 18-45), non-obese Caucasian population, and proposed a classification method for breast shape. 41 relative measurements, i.e. ratios and angles, were constructed from 66 raw measurements, extracted from 478 CAESAR (Civilian American and European Surface Anthropometry Resource) scans through Matlab programs. Data were examined through Shiny app for outliers and skewed distribution. Multiple data mining techniques, multivariate statistical methods, including Principal component analysis, Cluster analysis, Random forest, Discriminant analysis, MANOVA, were applied to the data. Moreover, an algorithm was proposed to visualize clustering outcomes. In the end, breast shapes were categorized into three and five groups by two different cluster number selection criteria proposed by the study: 1) based on misclassification rate; 2) based on Goodness-of-Fit. Several body measurements were identified to be critical in defining breast shape. Findings of the study can help the design of bras with improved fit and comfort.

BIOGRAPHICAL SKETCH

Jie Pei was born and raised at a city in Northern China. Her strict yet loving parents gave all that they had to support her dreams, taught her the merits of hard work, and always encourage her to tackle challenges and get out of comfort zone. She attended Donghua University in Shanghai (Originally China Textile University), majored in Fashion Design and Engineering, and obtained the degree of Bachelor of Engineering in 2014. Right after graduation, she came to the U.S. and attended Cornell University, in the department of Fiber Science and Apparel Design, to pursue her Master's degree. Her research had a concentration on intimate apparel, anthropometry, garment sizing and fit. Jie was accepted into the Ph.D. program by the same department in 2016.

This thesis is dedicated to my parents and God,

for their endless love and support

ACKNOWLEDGMENTS

I would like to express my very great appreciation to my advisor Prof. Huiju Park for all his efforts over the past two years in helping me to grow as a researcher and training my writing skills. His valuable guidance, patience, helpfulness, and willingness to give his time generously have been very much appreciated.

I would also like to offer my special thank to my minor committee member, Prof. Susan Ashdown for her support and inspiring advice throughout my research process. Her encouraging words and timely instructions were indeed very helpful and have been greatly appreciated.

I also gained insight from the feedback of Prof. Denise Green, Prof. Jintu Fan, Prof. Tasha Lewis, Prof. Paul Velleman, and fellow graduate students David Clark, Mengyun Shi, Qinwen Xu, Menglin Jia, Manwen Li, Marion Schelling, Diego Alzate Sanchez, Sarah Portway, Helen Trejo, Keith Fraley, Larissa Shepherd, Lina Sanchez Botero, Fangfang Weng, Ying Ji, Yingying Wu and Amanda Denham. I truly appreciate their support. I would also like to thank administrative staff Karen Steffy, Judy Wiiki, Michele Draiss and Timothy Snyder for their helpfulness through out the years.

I wish to acknowledge my parents, Jun Pei and Qianru Liu, my grandmother Yujin Wang, and late grandfather Huancheng Pei. They have been the best family, always loving and supportive.

My grateful thanks also extend to my brothers and sisters in First Ithaca Chinese Christian Church, and Cornell Chinese Christian Fellowship for their love, tolerance and prayers. I would like to offer my special thank to Xiaoshan Li for guiding me to Jesus. I would also like to thank her and Hui Zhu, Xiaohong Wang, Jianhua Chen, Jyying Kan, Lishih and Wenming Luh, Wenchao Wang, Wei Liu, Hui Wang, Mingqi Ge, Jing Zhang, Jingzhen Guo, Chou Li, Huiqin Chen, Haiping Li, Sarah and Joseph Cheng, Pastor Paul Epp, Pastor Andrew Lin, Pastor Bin Tang, for their guidance in my spiritual growth. My

special thanks also goes to my best friends, spiritual companions in Christ: Ye Song, Chunli Yuan, and Wei Xu. They have always been there in good and bad times, to encourage me and to light up my day. I would also like to thank Moyao Xue, Xin Huang, Yiju Wang, Neil Lin, Yue Ma, Xinzeng Feng, Yi-Chun Yeh, Yuguang Gao, Zechen Zhang, En-Ting Hsu, Chieh-Ren Hsia, Jing Jiang, Suki Zhang, Hua Sui, Yuan Luo, Siyu Chen, Chenruo Zhang, Lu Han, Wenxian Liu, Xiyue Zhang, Zongjie Wang, Yuling Huang, Yi Zhang, Qiannan Wen, Yating Ru, Xuan Yi, Junhao Li, Binghan He, Shuo Pan, Mingxuan Cai, Xiaoyang He, Yashan Li, Jiawei Yeh, Wenting Liu, Yifen Liu, Yifei Su, Jie Ding, for their friendship and love.

Last but not least, I would like to thank God my Lord for all His blessings in my life and for His eternal love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

APPENDICES

CHAPTER 1

INTRODUCTION

1.1. Intimate Apparel

Intimate apparel is a general term for garments that are worn underneath outer clothing and next to skin (Yu, 2011; Law, Wong & Yip, 2012). Examples of intimate apparel include panties, briefs, bras, boxer shorts, comfort wear, thongs, bikini, and body shaper (Yu & So, 2001; Hume & Mills, 2013). Intimate apparel can provide hygiene, tactile and thermal comfort, as well as support of certain area of the body. (Farnworth & Dolhan, 1985). In particular, brassiere functions as the cover and support of female breasts to provide concealment, and to protect the breasts from sagging or other unsightly deformation (Hardaker & Fozzard, 1997; Farrell-Beck, Poresky, Paff & Moon, 1998). Sports bra, as a special type of bra, provides protection and comfort by restraining the movement of breasts during exercise and athletic activity (Lorentzen & Lawson, 1987; Mason, Page & Fallon, 1999). Other functional bras are designed for various purposes: to facilitate nursing, to function as prosthesis, or simply to enhance wearers' perceived shape or size of their breasts (Kembering, 1979; Fanelli, 2001).

1.2. Fit and Comfort of Intimate Apparel

Fit issue is a common and long-existing issue for intimate apparel, especially for bras. According to previous studies literature, up to 70% of women are wearing bras with wrong sizes (Greenbaum, Heslop, Morris & Dunn, 2003; Wood, Cameron & Fitzgerald, 2008; McGhee & Steele, 2010). White and Scurr (2012) investigated fit of underwire bras sold in the U.K. adopting a bra fitting criteria suggested by McGhee and Steele (2010), and reported the followings as common fit issues: a) bra band on the back not level to the front; b) bra band too tight and digging in; c) cup size too small or baggy; d) underwire resting too low or too high (directly on breast tissue); e) shoulder straps digging in; and f) front of bra pulling away from the chest. A survey conducted by

Nethero (2007) on 1500 U.S. females showed that more than 65% of them had encountered discomfort during bra wearing; more than half of the respondents had back fat bulging issues; 50% of them reported straps sliding away while 28% reported straps digging into the shoulders; over 50% of them complained about insufficient support of their bras, and 25% of all participants felt that breast lifting was insufficient; 35% of the 1500 women had experienced underwire hurt; approximately 27% suffered from bra riding up.

Nonetheless, fit and comfort of intimate apparel is of great importance, not only with regard to consumers' satisfaction and buying decisions (Law, Wong & Yip, 2012), but also to the wearers' health. Studies found that the excessive pressure resulting from ill-fitting bras can cause breast pain and increase the risk of breast cancer (Singer & Grismaijer, 1995; Chen, LaBat & Bye, 2010). Singer and Grismaijer (1995) concluded that compared with women who did not wear bras, women who wear bras unceasingly every day were 125 times more likely to develop breast cancer, based on their investigation on 4730 women. Lee et al. (2005) also claimed that wearing bras during sleeping or long-time wearing (more than 12 hours every day) would significantly increase the risk of developing breast cancer. Moreover, excessive bra strap compression can cause deep bra furrows and upper limb neural symptoms (Kaye, 1972; Ryan, 2000; Greenbaum, Heslop, Morris & Dunn, 2003). Additionally, ill-fitting bra can lead to neck and back pain or poor posture (Lorentzen & Lawson, 1987; Mason, Page & Fallon, 1999). Some women with large breasts may even seek reduction mammoplasty because of the severity of these symptoms (Hadi, 2000; Ryan, 2000; Greenbaum, Heslop, Morris & Dunn, 2003).

Poor fit of brassieres result from various factors. Small errors caused by rounding of measurements during the calculation of band and cup sizes could accumulate to a large error of up to three cup sizes based on the traditional bra sizing method (Wright, 2002; McGhee & Steele, 2006, 2010;). Absence of professional fitting guide or service at

shopping sites, and consumers' lack of fit evaluation knowledge can also be the reasons for poor fit and comfort (McGhee & Steele, 2010; White & Scurr, 2012). Page and Steele (1999) claimed that if size was wrongly chosen, bra would not be able to provide sufficient support for the wearer regardless of how well the design was. For instance, due to the bulging of fat on the back during bra wearing, some full figure women preferred to buy larger band size which would later result in breast discomfort caused by insufficient support, and the sense of insecurity caused by straps falling down from shoulders (Yu, 2011). Meanwhile, despite sharing the same size on garment labels, measurements adopted by different manufactures during product development can be very different, which leads to high level of inconsistency in sizing and fit of apparel products in the market (Hardaker & Fozzard, 1997). Additionally, fitting of intimate apparel relies largely on life models, which makes it more difficult to achieve or maintain size uniformity (Hardaker & Fozzard, 1997).

More importantly, the complexity in breast shape, variations in size and shape among women, all contribute to the difficulty in the design of intimate apparel with good-fit (Hart & Dewsnap, 2001). Affected by pregnancy, nursing and menopause throughout life stages, women's upper body shape including breast shape goes through changes, which adds to the complexity of the issue (Goldsberry, Shim & Reich, 1996). Age and ethnic category can also be responsible for the diversity in female breast shape. Researchers have found that ageing could cause breasts to sag and the distance between bust points to increase, meanwhile the relative density of breasts to decrease (Brown, Ringrose, Hyland, Cole & Brotherston, 1999; Soares, Reid, & James, 2002; Haars, van Noord, van Gils, Grobbee & Peeters, 2005; Ashdown & Na, 2008). Shin (2009) conducted an investigation on 90 Asian females and 90 Caucasian females, and found significant difference in breast configuration between the two groups, for instance, Asian women had wider sternum and more centered bust points (smaller bust points distance) in general compared with the Caucasian counterparts.

The study of body shape variation is a typical way of improving garment fit and comfort, and of developing sizing systems. Feather, Ford and Herr (1996) claimed that body shapes could directly affect the satisfaction level of garment fit. LaBat (1987) classified a group of college female students into short, well proportioned, and long, by the length of their upper body, and discovered a significant difference in fit satisfaction among the three groups. Likewise, a better understanding of breast shape can help with the improvement of the designs of bras. Chen, LaBat and Bye (2011) claimed that different bust prominence resulted in contrasting bra fit perception, and that the study of breast measurements could contribute to the design of a better fitting bra and a reliable bra classification system. Zheng, Yu and Fan (2009) suggested that an improved bra design could fit the complex contours of the breasts, and provide support and appropriate strain by proper use of cups, shoulder straps and bottom band. Oh and Chun (2014) emphasized on the importance of measuring breast size precisely in order to achieve good fit in bra design. Lee and Hong (2007) studied the geometrical shape of under-bust curve and came up with an optimal design of underwire that could provide better support of breast mass. Zheng, Yu and Fan (2007) developed an enhanced bra sizing systems with higher accommodation rate for Chinese females via the anthropometric analysis of 456 nude breasts.

Although research on female breast shape has been done quite intensively for the Asian population, not many literatures investigating 3D shapes of female breasts for the general Caucasian population (with sufficient number of subjects included) have been found. In addition, the outcomes of the previous breast shape studies have rarely been put into practice. The traditional bra sizing system and fitting method, which adopts the body measurements of bust circumference and underbust circumference, is still widely used by many intimate apparel companies (White & Scurr, 2012). However, despite a wide range of sizes provided by the traditional sizing system (from 28AA to 56FF), the sizing system still cannot provide satisfactory fit for a large proportion of consumers because of its

inadequacy in approximating breast volume, ambiguity in measurement definition, and insufficiency in differentiating breast shape (for not taken factors such as the relative position of the breasts on chest wall into account) (Pechter, 1998; Nethero, 2007; Zheng, Yu & Fan, 2007). Furthermore, a few limitations in previous research on breast shape have been identified through in-depth review of existing literatures (details are in Chapter 2). Improvements need to be made in the methodologies.

1.3. Purpose of Study

Based on research gap identified through literature review, this study aims to understand the variation in female breast shape across the Caucasian population, and to propose a classification method for breast shape. The followings are some research questions to be answered.

- Question 1: What are the most critical body measurements that best define breast shape?
- Question 2: What is the best way to classify breast shapes? How many groups should they be classified into?
- Question 3: How to present and validate the final classification outcomes?

The structure of the thesis is as follows:

- Chapter 1 introduces the purpose of the study and a few research questions.
- Chapter 2 includes the literature review of body shape and breast shape studies, and the summary of limitations of previous research.
- Chapter 3 presents the methodology including the national-scale dataset and data mining of 66 breast-related anthropometric measurements.
- Chapter 4 contains the descriptions of the programs developed in Matlab, including 3D body scan preparation, extraction of measurements, and algorithms to exhibit classification outcomes.

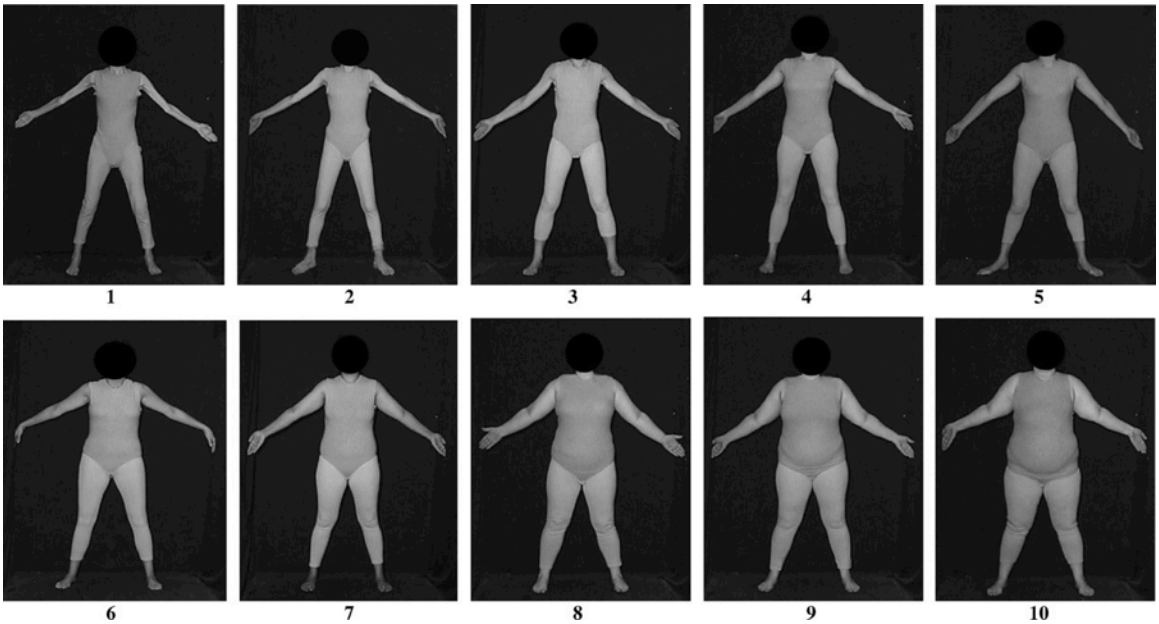
- Chapter 5 explains the importance of data examination, and presents the details of a Shiny App, created for the generation of interactive plots and for data examination and preparation. Lastly, the outcomes of outlier detection and data transformation are included.
- Chapter 6 discusses statistical analysis and results.
- Chapter 7 presents conclusions, practical implications and suggestions for future studies.

CHAPTER 2

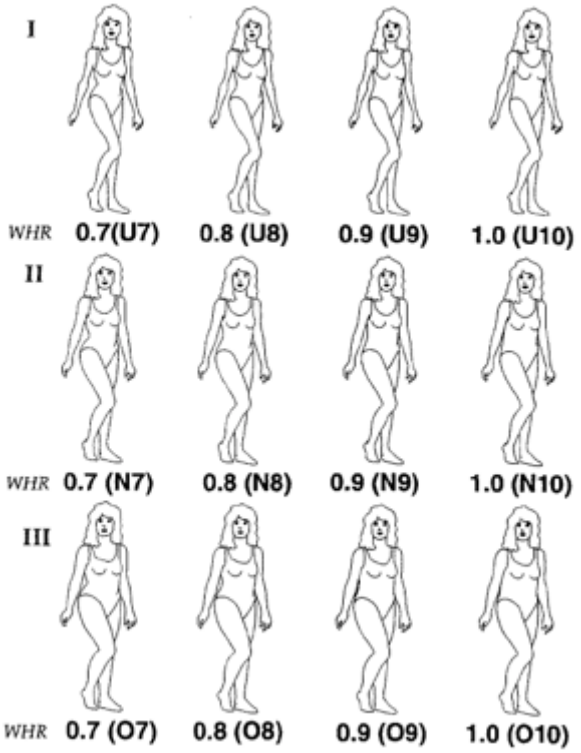
LITERATURE REVIEW

2.1. Studies of Anthropometry and Body Shapes

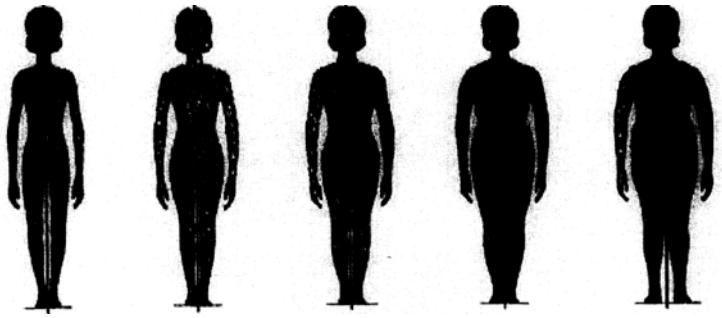
Before 3D scanning, researchers had to use photographs or silhouette figure drawings (Figure 2-1) to demonstrate or to categorize human body types (Sheldon, Stevens & Tucker, 1940; Douty, 1968; Jones, 1972; Stunkard, Sorenson & Schulsinger, 1982; Singh, 1993; Henss, 1995). The development of 3D body scanning technology, accompanied with powerful algorithms (for extracting measurements automatically and for discovering measurements that describe body proportions and postures as well as basic dimensions), and fast processing speed in computers, allows researchers to conduct in-depth work to analyze the complicated human body shape, and to quantify the variations in body shapes and sizes across population. Nowadays, a wide variety of anthropometric studies and body shape studies, based on analysis of 3D body scans, have been conducted.



a). Photographic figure rating scale (Swami, Salem, Furnham & Tovée, 2008)



b). Body shape illustrations used by Singh (1993)



c). Douty body build scale (Jones, 1972)

Figure 2-1. Silhouettes, photographs, and figure drawings used by researchers for the study of human body shape

2.1.1. Body Shape Variation and Assessment Scale

Simmons, Istook and Devarajan (2004) analyzed the body scans of 222 female participants, and classified their body shape into the category of Bottom Hourglass,

Hourglass, Spoon, Rectangle, Oval and Triangle. They developed a program (FITT, Female Figure Identification Technique) to automatically sort new case into their defined shape categories. In the study, participants of the Bottom Hourglass shape type had small bust and hip circumferential difference, and their bust-to-waist and hip-to-waist ratios were significant and similar (Figure 2-2a). Those of Bottom Hourglass shape type also had significant bust-to-waist and hip-to-waist ratios, but their bust circumferences were slightly smaller than hip circumferences (Figure 2-2b). Those of Spoon shape type had large hip-to-waist ratio, lower bust-to-waist ratio and larger bust and hip circumferential difference, compared with the Hourglass shape (Figure 2-2c). The Rectangle shape had low bust-to-waist and hip-to-waist ratios and similar bust and hip circumferences (Figure 2-2d). Participants with the Oval shape type were those who did not belong to the previously defined four shape types, and had larger bust girth than the average of waist, abdomen, and stomach girths (Figure 2-2e). The Triangle shape had larger hip girth than bust girth, and small hip-to-waist ratio (Figure 2-2f). The author also added the category of Top Hourglass, Inverted Triangle and Diamond into their program. They suggested that their system could facilitate the development of new sizing systems that could provide better garment fit to consumers.

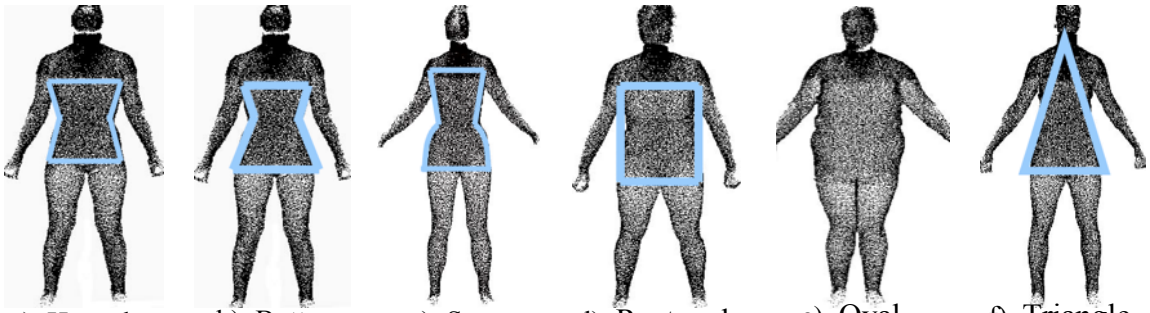


Figure 2-2. Body shape types defined by Simmons, Istook & Devarajan (2004)

Connell et al. (2006) had five experts analyze the body scans of 42 female participants aged 20 to 55, and came up with a nine-section body shape assessment scale (BSAS©). Three of the nine sections were for whole body analysis, which included analysis of body build, body shape and posture. In terms of body build, participants were classified into four types, namely Slender, Average, Full and Heavy; for overall body shape, four categories were adopted: the types of Hourglass, Pear, Rectangle and Inverted Triangle; and there were three categories of postures: Aligned, Forward alignment and Compensating alignment. The other six sections were evaluations of parts of body: back shapes were classified into four types (Flat, High, Middle and Low); shoulder shapes, defined by shoulder slope, were classified into three (Square, Average, and Sloped); bust shape had three types (Flat, Average and Prominent), same with buttocks shape; and hip shape had four types (Straight, High, Mid and Low). Moreover, 100 additional scans were rated by the same experts to find links between the categorical scales with anthropometric measurements extracted from the scans. A program (BMS©) was built using the proposed scale and the ratings from the experts, to classify body scans.

Azouz, Rioux, Shu and Lepage (2006) proposed an approach to convert 3D body scans into 3D models with main variations of shape retained. They worked on 300 male scans randomly selected from CAESAR (Civilian American and European Surface Anthropometry Resource) database. Firstly, they compared several hole-filling methods and proposed a method by fitting gaps on horizontal slices with second-order Bezier curves (Figure 2-3). They then converted every polygonal mesh into a volumetric representation using a set of voxels and their corresponding distances from the nearest point on the scan surface. The distances were expressed in a vector form for every scan. Then the authors applied Principal component analysis (PCA) on the distance vectors of the 300 scans. They found the first few eigenvectors accounted for the majority of scan variations, and the first five eigenvectors were related with the variations in body weight,

leaning posture, muscularity, arm-torso spacing and head position respectively (Figure 2-4).

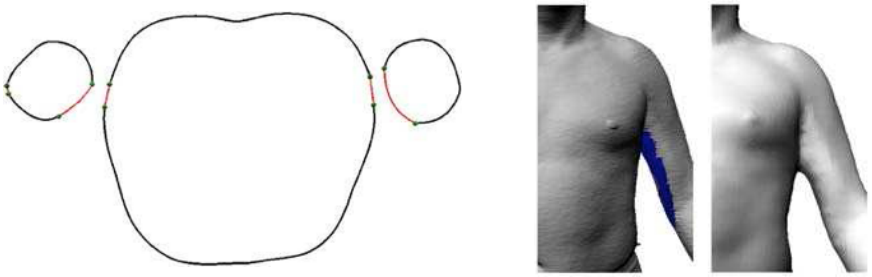
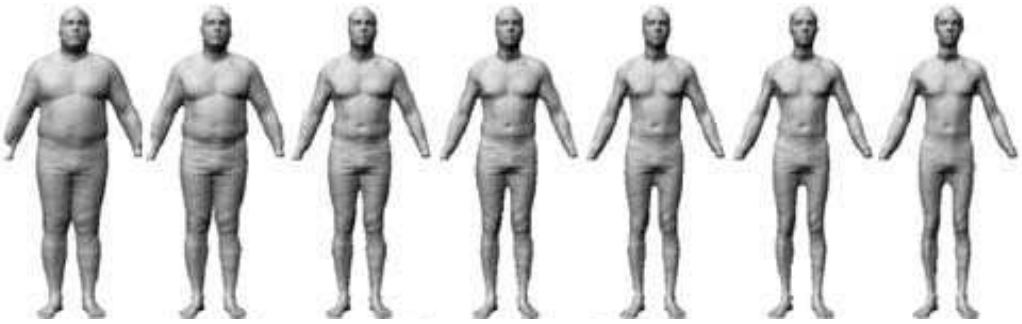
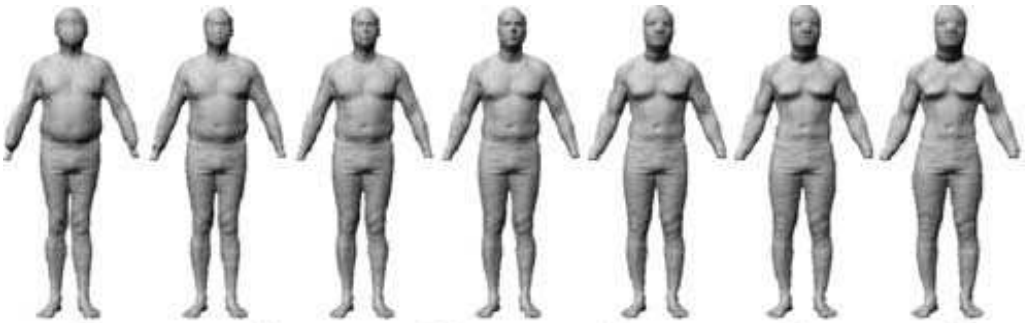


Figure 2-3. Hole-filling method adopted by Azouz, Rioux, Shu & Lepage (2006)

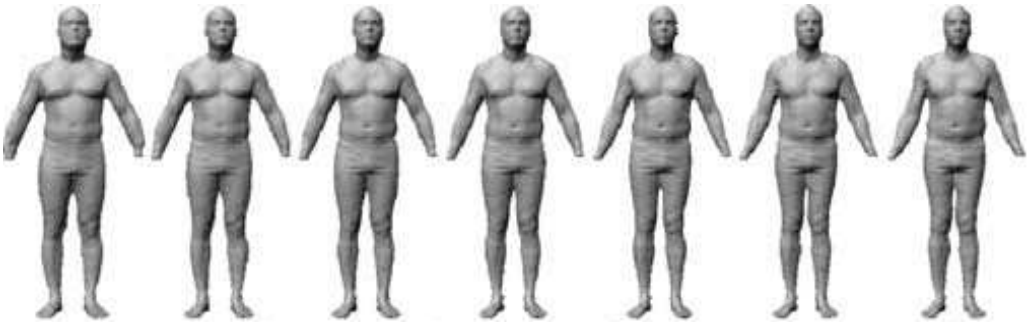


a). Variation in body weight

b). Variation in leaning posture



c). Variation in muscularity



d). Variation in arm-torso spacing

e). Variation in head position

Figure 2-4. Five main variations in body scan discovered by Azouz, Rioux, Shu & Lepage (2006)

2.1.2. Key Landmarks and Key Body Measurements

Landmarks, usually anatomical landmarks, are feature points that describe body shape and size. They are also the reference points, identified before 3D scanning or traditional anthropometrics, to facilitate extraction of body measurements. Hence,

rigorous landmarking is important and responsible for the accurate extraction of measurements, and for the precision and validity of anthropometric studies.

Lu and Wang (2008) developed a system to detect landmarks and characteristic lines on 3D body scans and to extract body dimension data automatically (Figure 2-5). They recruited 189 participants for scanning. Each of the participants was scanned for five times. The landmark recognition results and measurement extraction results obtained from the five scans were compared and the results showed high consistency. In addition, landmarks were placed onto participant's body manually with their locations recorded. 12 anthropometric data were collected manually using traditional method. The manual results were compared with the automatic results, 5 out of the 12 body dimensions were found to be significantly different by paired t-test, namely shoulder breadth, chest girth, waist girth, sleeve length, and cervicale to waist length (anterior). Although the differences were within the acceptable threshold required by ISO (International Organization for Standardization), the authors considered that automatic extraction to be more reliable than manual extraction.

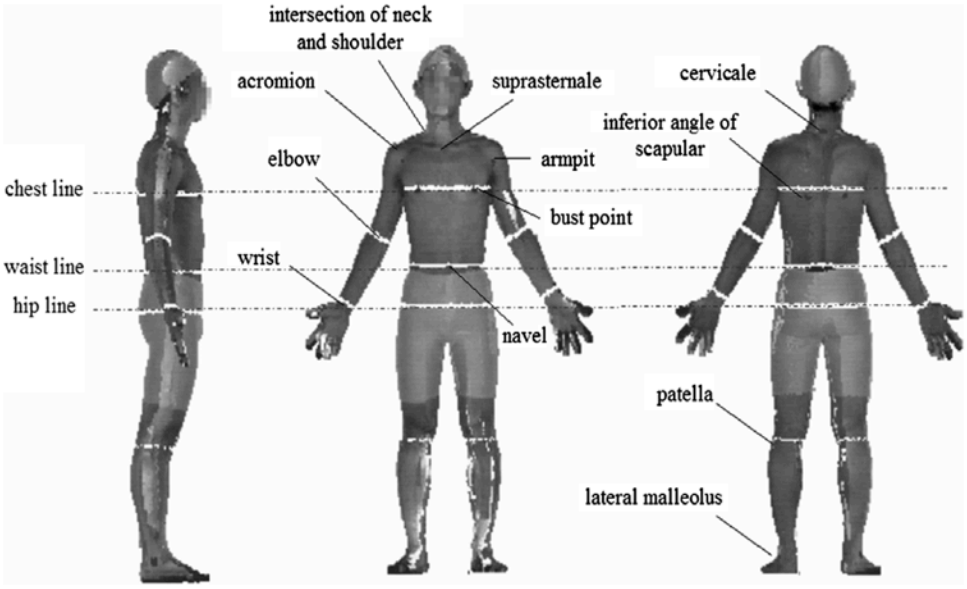


Figure 2-5. 12 landmarks identified by the measurement extraction system developed by Lu & Wang (2008)

Han and Nam (2011) pointed out that variations in body shape could cause a significant difference in landmark locations. Accordingly, they proposed six landmark detection algorithms (one algorithm for one landmark) that worked for different body shapes, including the overweight population. The six corresponding landmarks were crotch point, armpit point, front neck point, side neck point, back neck point and shoulder point. For each landmark, the authors divided the population into three groups based on body shapes. They treated BMI as the grouping factor for crotch point, armpit point and side neck point; front neck angle as the grouping factor of front neck point; back neck angle as that of back neck point; and shoulder forward angle as that of shoulder point. Through statistical analysis on empirical data, the authors narrowed down the searching range for each landmark, thus improved the efficiency in landmark identification. They examined the accuracy of their algorithms for each body shape and concluded that their algorithms were more precise than method that used lateral pit points, especially for the overweight population.

Cho et al. (2006) developed an interactive 3D virtual model of female torso. The shape and posture of the model could be adjusted by changing a few parameters, including circumferences, lengths, angles, depths, etc., in a control panel. The authors regarded ten points on the side profile of a scan as key landmarks that determined body shape and body posture (Figure 2-6). A few control parameters were based on the relative locations of those landmark points. In their paper, the authors presented three types of back shapes, classified by Angle A and Angle B, namely flat back shape, average back shape and stoop back shape (Figure 2-7a); and three types of hip shapes, classified by Angle C, namely flat hip shape, average hip shape and protruding hip shape (Figure 2-7b). Moreover, they compared the side profile of actual body with that of the virtual model and found ideal match. Therefore, they concluded that their virtual model was good representation of real body. Lastly, the authors suggested that the model had many

potential uses (e.g. virtual draping) for the apparel industry, and could be a helpful tool in pattern making.

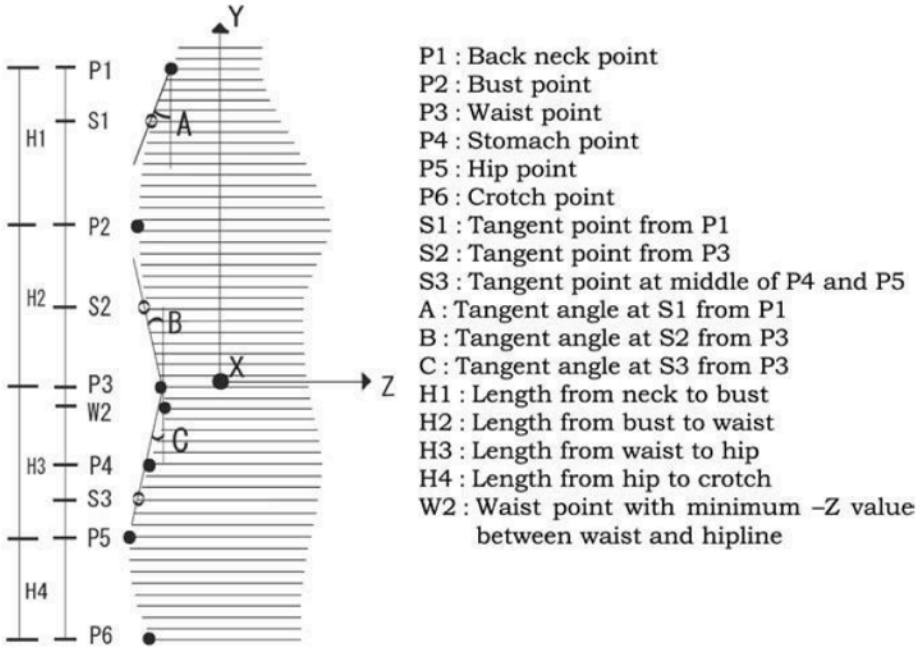
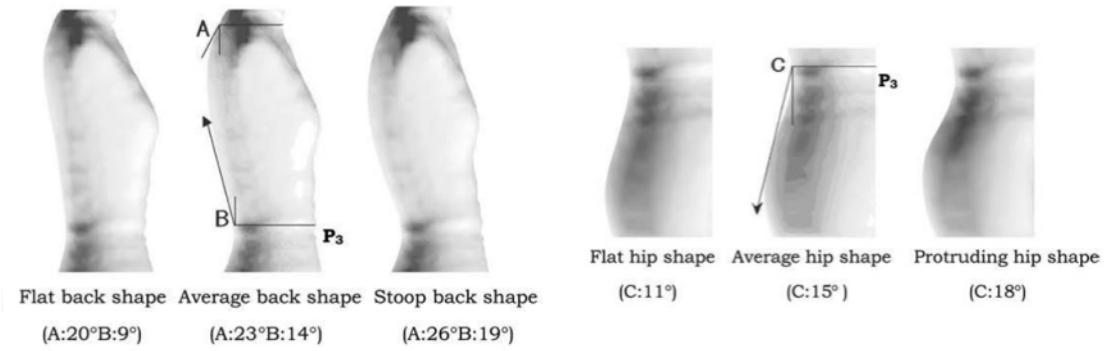


Figure 2-6. Control points and measurements for the adjustable 3D model (Cho et al., 2006)

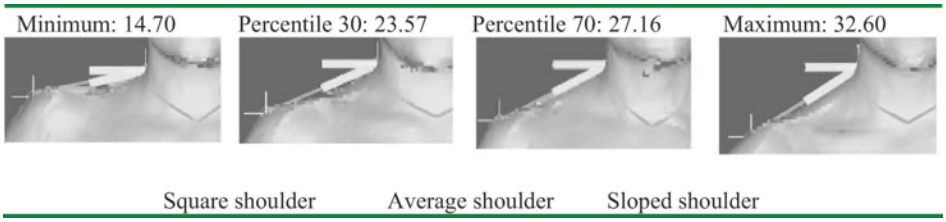


a). Three types of back shapes b). Three types of hip shapes

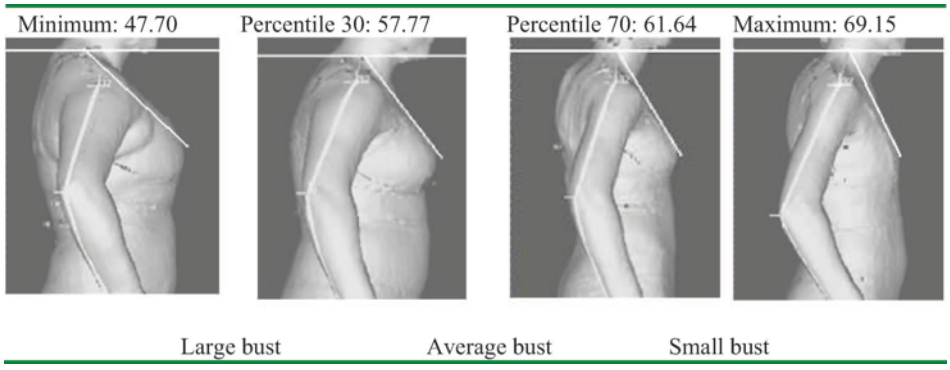
Figure 2-7. Three types of back shapes and hip shapes (Cho et al., 2006)

2.1.3. Body Shape Classification

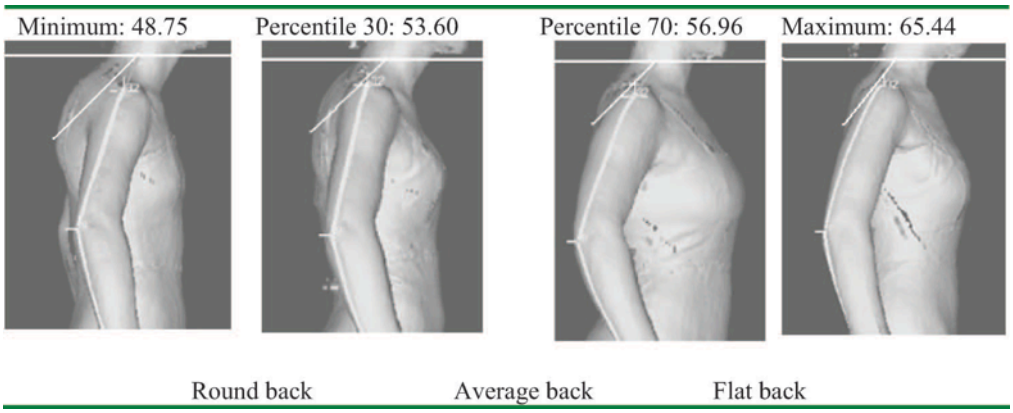
Chen, LaBat and Bye (2010) recruited 103 Caucasian female participants for 3D scanning. Participants were scanned wearing their daily bra underneath a body-scan suit. Body measurements, mostly angles that associated with the design of bras, were extracted using ScanWorX™ and Polyworks™. The authors utilized the descriptive statistics of percentiles to classify upper body shape into three types. The 30th and 70th percentiles were chosen as the threshold values between shape types. Based on the measurement of shoulder slope, participants were classified into the types of square shoulder, average shoulder and sloped shoulder (Figure 2-8a). Similarly, based on bust prominence, participants were classified into the types of large bust, average bust, and small bust (Figure 2-8b). The measurement of back curvature corresponded to the types of round back, average back, and flat back (Figure 2-8c); and that of acromion placement corresponded to the types of backward acromion, average acromion and forward acromion (Figure 2-8d).



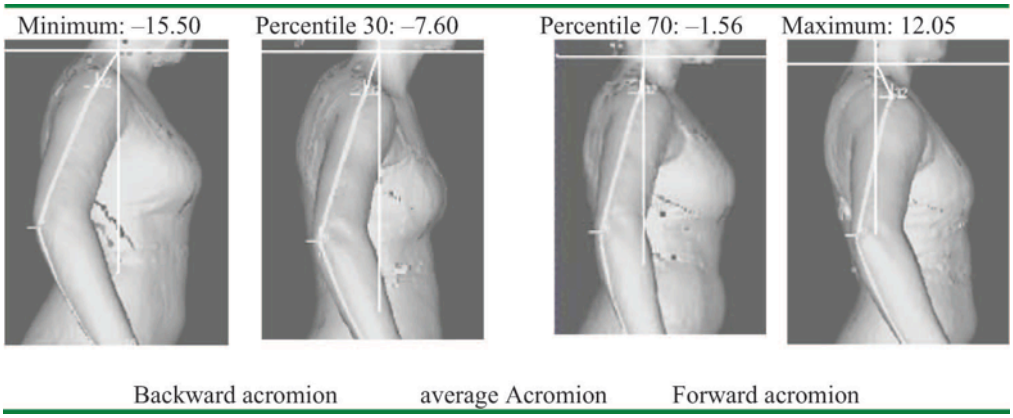
a). Based on the measurement of shoulder slope



b). Based on the measurement of bust prominence



c). Based on the measurement of back curvature



d). Based on the measurement of acromion placement

Figure 2-8. Upper body classification based on four body measurements (Chen, LaBat & Bye, 2010)

Song and Ashdown (2011) developed a classification method for female lower body shapes, by applying PCA and Cluster analysis to the body measurement data of 2488 participants, selected from SizeUSA female database. Initially the authors chose 18 raw measurements (e.g. front and back arc at waist level, buttocks angle, etc.) and 15 drops (e.g. hip to waist girth drop, abdomen to waist depth drops, etc.) as variables of interest. Then they conducted bivariate correlation analysis between one of the variables and participant body weight. Variables that had weak linear and quadratic relationship with body weight were regarded as shape-related variables that were not affected by body size,

and a total of 15 variables (14 drops and buttocks angle) were kept for PCA. The first three Principal components (PC's), and two z-scores calculated from 2 variables (these 2 variables had large loading for either the 4th or the 5th PC), were used for K-means cluster analysis. The authors ended up classifying buttock shape into three types, namely curvy shape, hip tilt shape, and straight shape (Figure 2-9). Lastly, they proposed Discriminant functions based on cluster membership for the prediction of future cases.

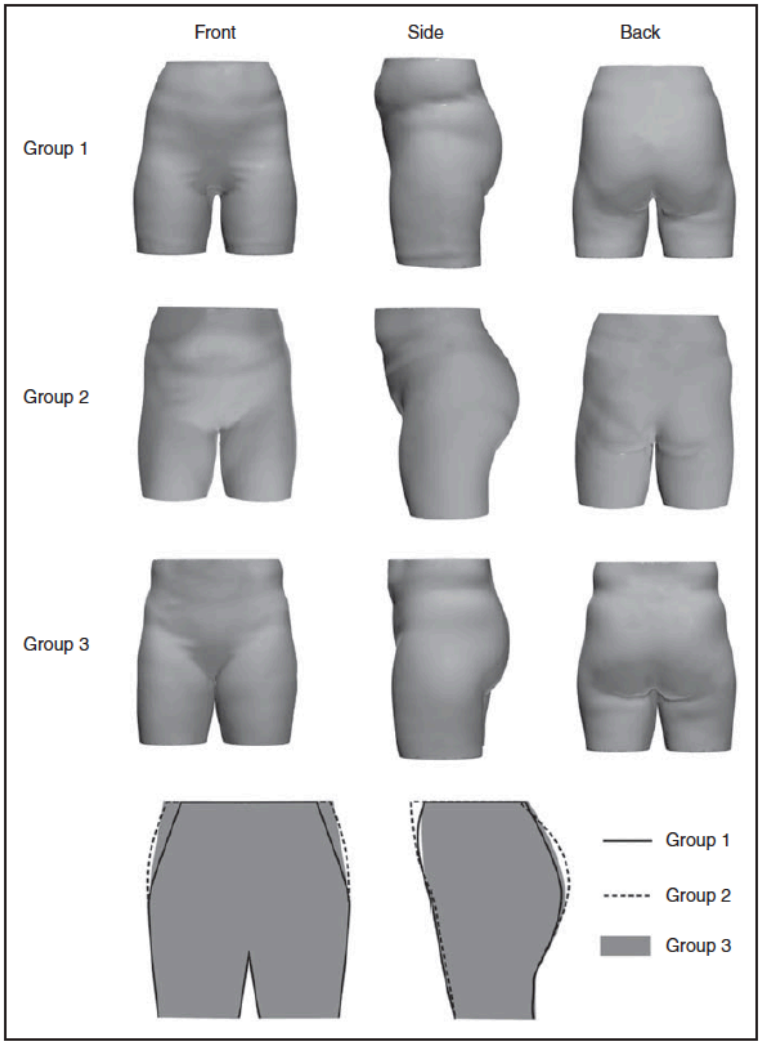


Figure 2-9 Three types of buttocks shape classified by Song & Ashdown (2011)

2.2. Female Breast Shape Studies

Female breast has a complicated structure (Figure 2-10). The inner structure relates closely to the outer appearance. For instance, the stretching of the Cooper's ligaments associates with the sagging of breasts (Page & Steel, 1999). Moreover, the amount and distribution of the adipose tissue relates to the size and shape of the breast. It was recorded that generally breast has a diameter of 10 to 12 centimeters, and a central thickness of 5 to 7 centimeters (Lawrence & Lawrence, 2010). However, because of the complexity in structure, female breasts can vary greatly in size and shape. In fact, there are multiple ways to define breast types, and breast shape classification has been done both in academia and in commerce. According to Bratabase.com (Figure 2-11), breast can be round or pointy, full on top or full on bottom, have wide root or narrow root, have high nipples or low nipples, etc. According to Herroom.com (Figure 2-12), breast shapes can be divided into the following types: archetype shape, Omega shape, thin shape, conical shape, uneven shape and reduced projection shape.

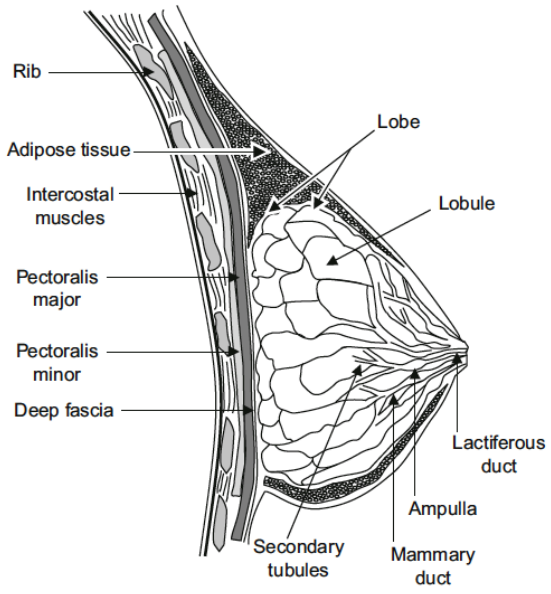


Figure 2-10. Structure of female breast from sagittal section (Page & Steel, 1999)

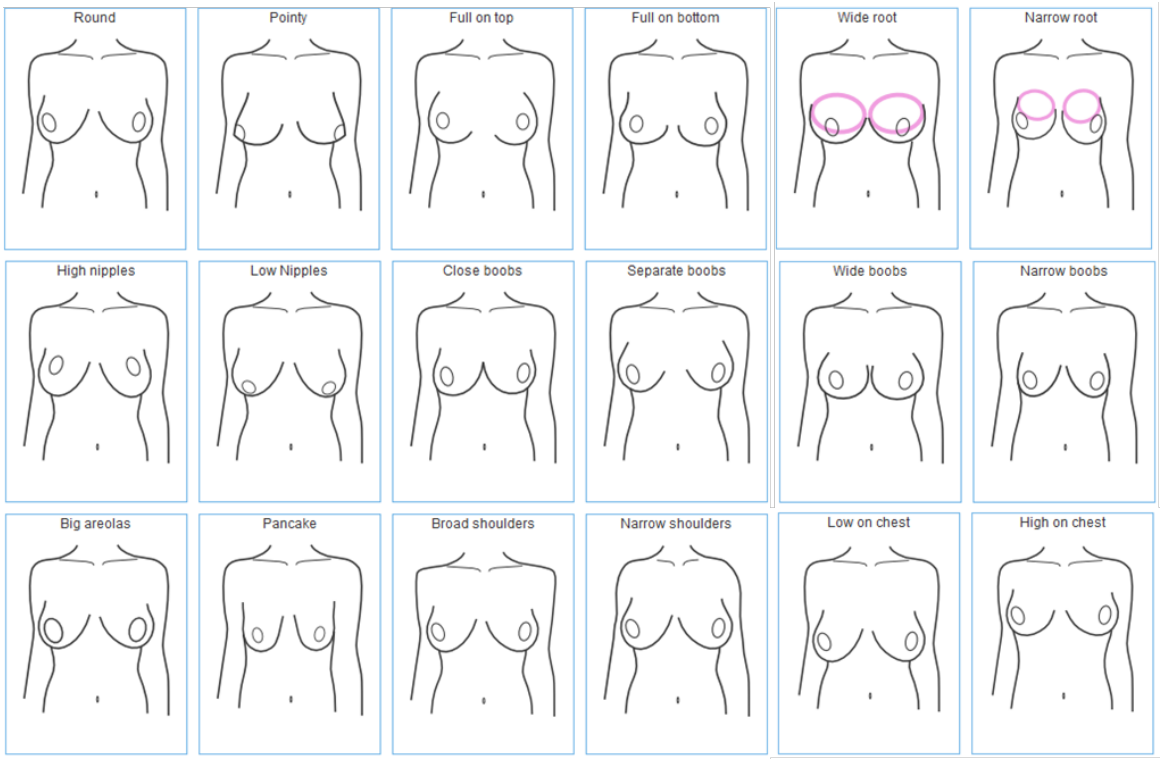


Figure 2-11. Female breast shape types (Bratabase, n.d.)

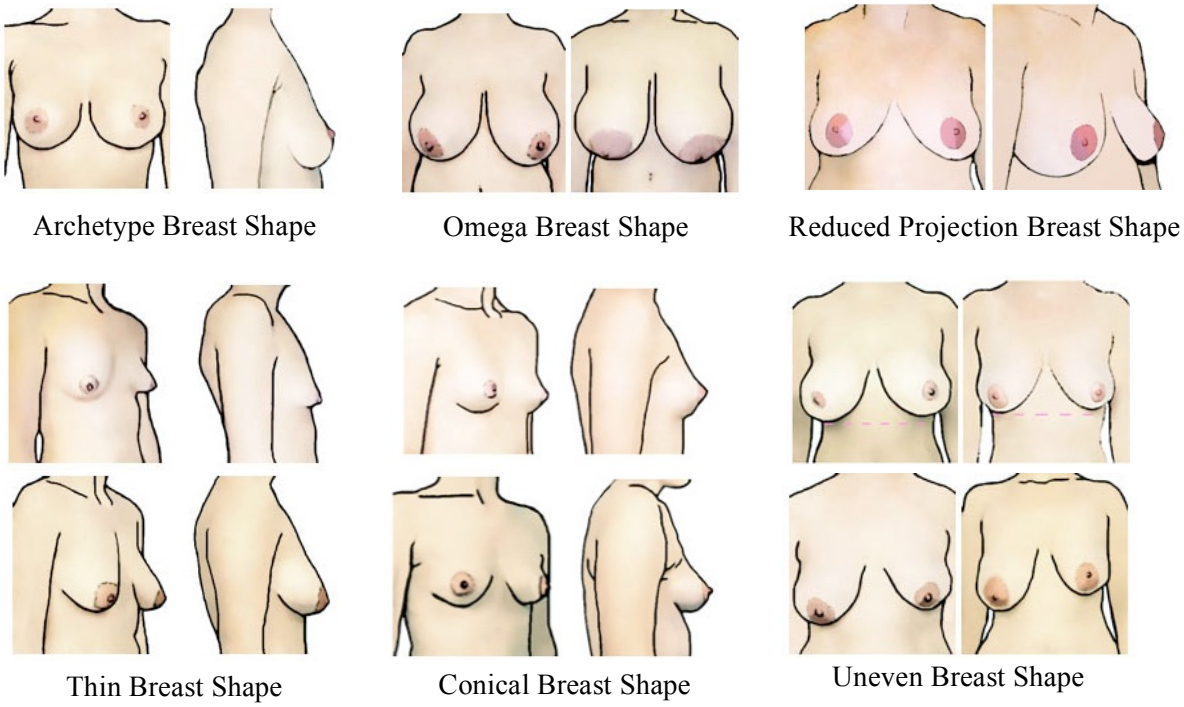


Figure 2-12. Female breast shape types (Herroom, n.d.)

2.2.1. Breast Shape Assessment Scale

Hsia and Thomson (2003) studied the difference in breast shape preference, in terms of the upper-pole contour, among plastic surgeons, breast augmentation surgery patients, and ordinary people. They created a 12-scale lateral breast profiles (Figure 2-13) for 66 participants to rate, based on their preferences and their own judgments on the attractiveness and naturalness of breasts. In Figure 2-13, breast shapes with concave upper-pole were marked with negative sign, those with linear upper-pole were labeled with zero, and those with convex upper-pole were marked with positive sign. Their research found significant distinction in ideal breast shape between plastic surgeons and patients. Moreover, their results showed that breasts with linear (0) or convex (+) upper-pole profiles were generally considered by female patients to be more attractive and natural.

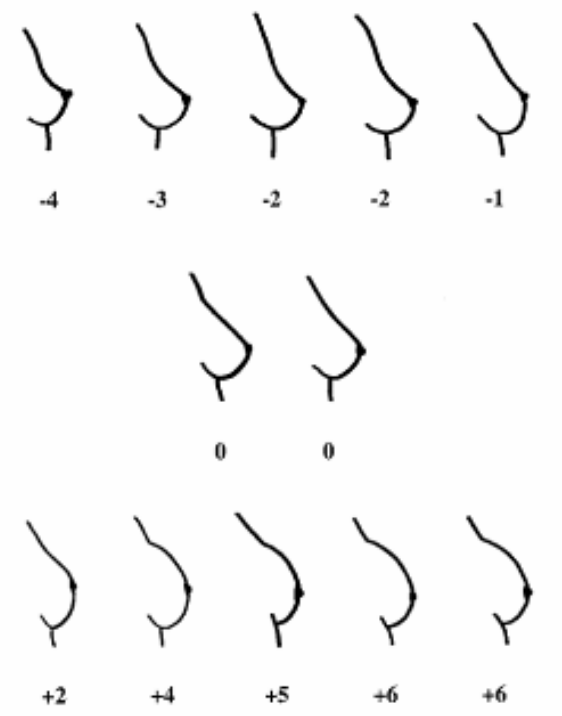


Figure 2-13. The 12-scale lateral breast profiles used by Hsia and Thomson (2003)

Lau (2014) pointed out in her research that four attributes were of importance in defining female breast shape, namely breast size, breast separation, breast projection and breast height. She proposed a 9-point multi-dimensional breast rating scales on the basis of the four attributes via the use of graphical illustrations (Figure 2-14). She also defined a new index, called "Bust Satisfaction Index" (calculated by dividing the ideal self image rating score by the current self-perceived image score), to quantify the gap between each female participant's ideal breast shape with her actual breast shape. A total of 36 Chinese females were recruited to validate the proposed 9-point rating scaled, and the rating scale was proved to have good repeatability. The researcher then investigated on the relationship between the bra fit preference of wearers and their satisfactory level of their breast shape images. 100 Chinese females participated in the breast rating, accompanied with a series of wear trials of seamless knitted bras. The results showed that the shaping effect provided by bras, i.e. the changes in breast shape, was the most essential factor related to bra fit perception.

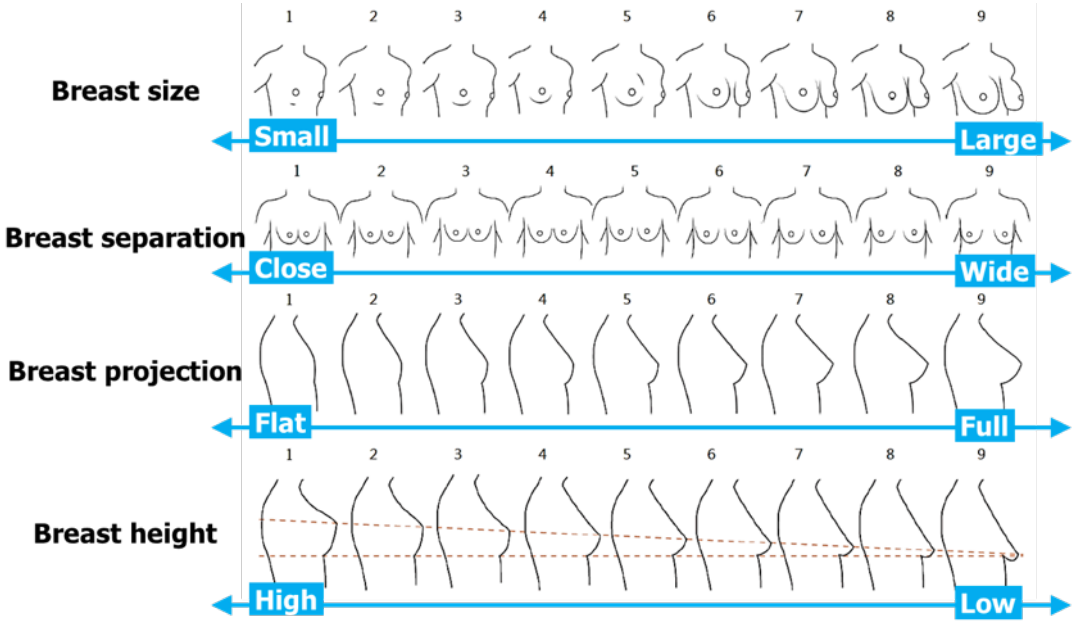


Figure 2-14. The 9-point breast rating scales proposed by Lau (2014)

2.2.2. Key landmarks and Key Parameters for Breasts

Brown, Ringrose, Hyland, Cole, and Brotherston (1999) proposed a few key landmarks at bust area (Figure 2-15) and a few morphometric measurements that they considered to be essential in defining breast shape. They measured 31 female patients who sought for reduction or augmentation mammoplasty, and 60 ordinary female volunteers, using set square, tape measure and ruler. Certain distinction between the volunteer group and patient group had been found. In addition, they tested the reproducibility of their proposed measurements by having the same observer measure the same 10 participants on two different occasions. Same set of measurements was recorded and the differences in measurement values between the two separate occasions were within 5 millimeters. Moreover, the authors studied the asymmetry of breasts, and the impact of age, weight and height on breast shape. No significantly different means between the measurements of right breast and those of left breast had been found, although the authors believed in the existence of asymmetry between right and left breasts (they referred to this kind of asymmetry as fluctuating asymmetry). Additionally, they found that areolar diameter could be negatively related to age and positively related to weight; weight also had an impact on vertical positions of the landmarks; and height did not associate with any of the breast shape measurements.

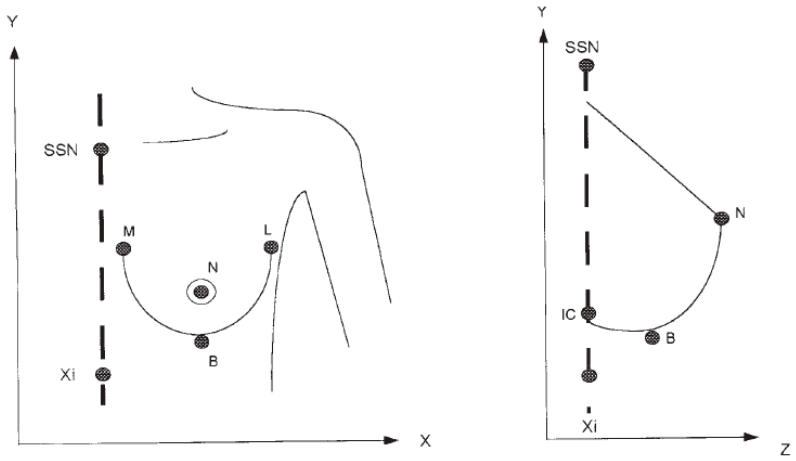
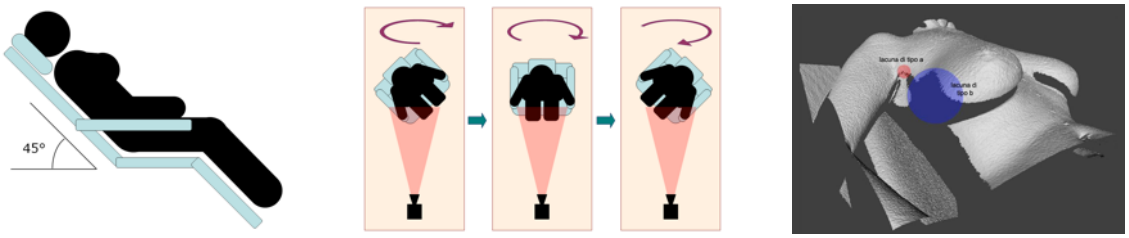


Figure 2-15. Key landmarks at bust area defined by Brown et.al (1999)

Farinella et al. (2006) studied geometric shape of female breasts by analyzing real clinical cases from breast reconstructive surgery. Patients involved were scanned sitting in a chair leaning backwards with one fixed camera in the front (Figure 2-16a). The chair was also rotated 45 degrees, both to the right and to the left, for the camera to capture more regions on the bust (Figure 2-16b). Seven reference points (Figure 2-17a), namely Sternal notch point (P_j), Xiphoid point (P_x), Nipple (P_c), Pectoralis insertion in the arm (P_{aa}), Acromial extremity of clavicle (P_s), Mid-axillary point (P_{pa}), and Lowest breast point with respect to the vertical body axis (P_d), were selected by a professional surgeon, and were considered anatomically important. The authors concluded from their study that partitioning breast into four subunits was helpful, and they developed an algorithm to divide breast surface by breast meridian curve and breast equator curve (Figure 2-17b), constructed from the above-mentioned reference points.

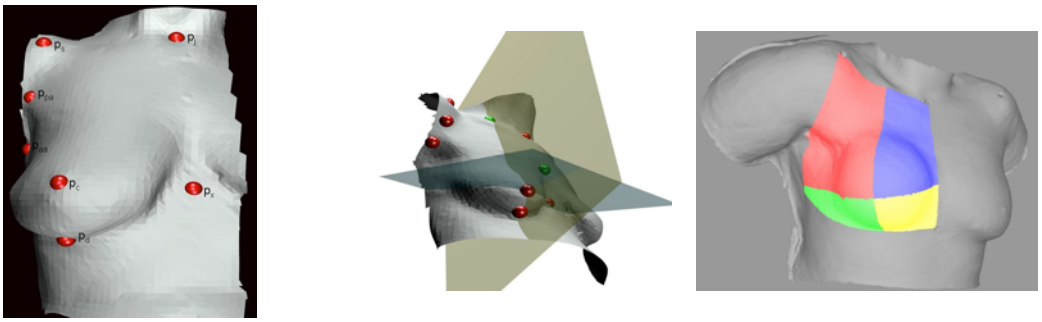


a). Patient's posture

b). Rotation of the chair with respect to the camera

c). Original scan

Figure 2-16. Scanning posture adopted by Farinella et al. (2006)



a). Landmarks b). Partition of breast

Figure 2-17. Reference points and four sub-surfaces of breast
(Farinella et al., 2006)

Catanuto et al. (2008) proposed a way to display the curvature of the thoracic surface using color map (Figure 2-18). They applied their technology to seven female patients who sought for breast reconstructive surgery, and meanwhile obtained the patients' breast measurements, such as surface area, a few point-to-point distances and angles, during patients' pre- and post-operative visits. The changes in breast measurements of three clinical cases were presented in the paper (the first two cases were implant-based reconstruction after mastectomy, and the third one was augmentation mammoplasty). Moreover, the authors believed that the degree of curvature could be a more scientific way to describe breast shape, compared with the descriptive terminology, for instance, the fullness of breast.

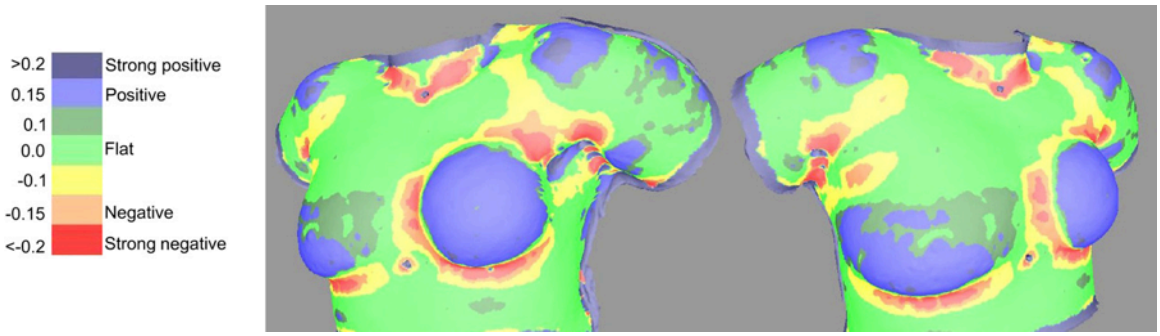
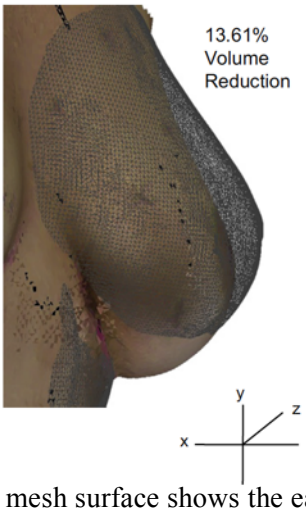
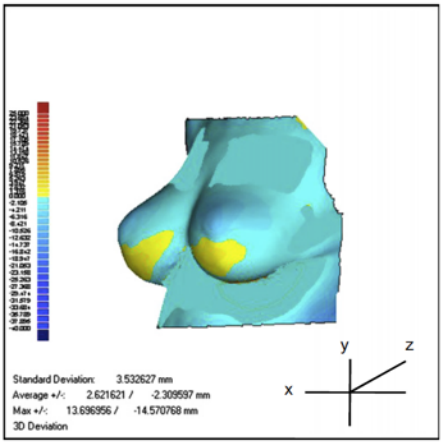


Figure 2-18. Color map of curvature (Catanuto et al., 2008)

Small et al. (2010) evaluated breasts shapes of 15 reduction mammoplasty patients, during their early (60 to 120 days after operation) and late (400 to 500 days after operation) post-operative period. Participants were scanned facing a 3D laser scanner at the angles of +90°, +45°, 0°, -45°, and -90°. Breast models were built for each participant in Geomagic Studio. The models obtained from early and late post-operative period were plotted together for comparison (Figure 2-19a). A few parameters, such as volumetric tissue distribution and maximum projection, were assessed, to study the bottoming-out effect. In addition, the authors adopted vector sum analysis to quantify the changes in breast shape and generated a color map where darker blue color corresponded to decrease in volume, and yellow or red corresponded to increase in volume (Figure 2-19b). In the end, the authors pointed out the need of a database of 3D scans of nude breasts in pre-surgical planning and in studying breast tissue re-distribution over time.



a). The mesh surface shows the early post-operative breast image; the solid surface shows the late post-operative breast image



b). Color map generated from vector sum analysis

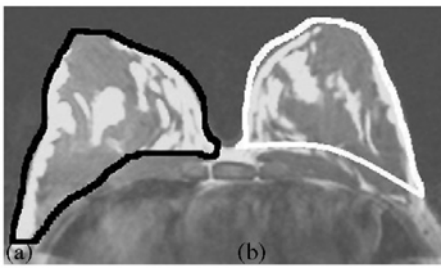
Figure 2-19. Comparison of breast shapes in early and late post-operative period (Small et al., 2010)

Qiao, Zhou and Ling (1997) measured the breasts of 125 Chinese young females, and proposed a formula to calculate breast volume (Eq. 2-1). The authors applied the formula to 178 breast cosmetics surgery patients, and claimed that the estimation of breast volume was helpful for their clinical work. Moreover, they obtained the average breast volume for Chinese female aged 18-26, and investigated the relationship of breast volume with parameters such as body weight, body height and bust circumferential difference (bust girth subtracted by chest girth at axilla level). They found that breast volume is positively correlated to body weight, and to bust, waist and hip circumferences.

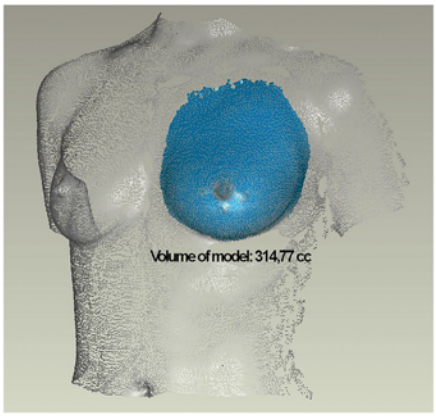
$$V = 1/3 \times \pi \times MP^2 \times (MR + LR + IR - MP) \quad (2\text{-}1)$$

where MP is short for Mammy projection (maximum horizontal protrusion observed from side profile); MR is short for Medial breast radius (nipple to medial terminal crest); LR is Lateral breast radius (nipple to lateral terminal crest); and IR is Inferior breast radius (nipple to inframammary fold).

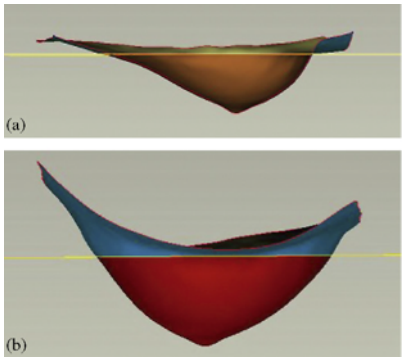
Kovacs et al. (2007) selected four most commonly used breast volume calculation methods and applied them to the same 6 female volunteers. The four methods were based on magnetic resonance imaging (MRI), 3D surface scanning, thermoplastic casting, and mathematics equation (Qiao, Zhou & Ling's formula, see Eq. 2-1) respectively (Figure 2-20). The breast volumes obtained by the four methods did not agree well. One of the reasons could be caused by different body postures (participants were in a prone position for the MRI scanning, in a standing position for the 3D scanning, and an upright seated position for the casting). Nonetheless, the authors concluded that 3D scanning was a promising approach to measuring breast volume, with acceptable measurement accuracy, good participant acceptance, low operative difficulty and relatively low cost.



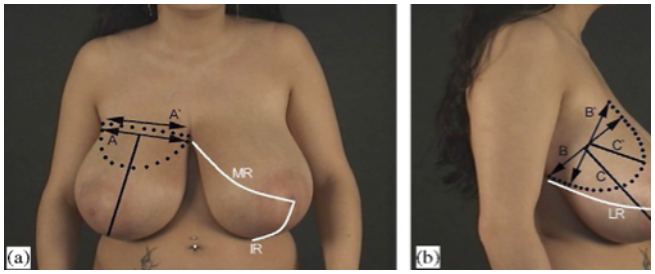
a). MRI scanning



b). 3D scanning



c). Thermoplastic casting



d). Manual measurement

Figure 2-20. Four common breast volume estimation methods (Kovacs et al., 2007)

2.2.3. Anthropometric Studies of Breasts for Bra Design and Sizing

Lee, Hong, and Kim (2004) also utilized the 3D scanning technology on female nude breasts, but to seek for helpful information for the design of form-fitting brassiere. The authors scanned 37 Korean females, who normally wore bras of size 80A, using 3D phase shifting moirés topography, with their upper body scanned in nude. They proposed an approach to find the natural boundary between breasts and chest wall by upward pushing and inward pushing of the breast (Figure. 2-21). Flat landmarks and 3D landmarks had been placed onto each participant's body before scanning (Figure. 2-22). Based on the anthropometric measurements that they extracted, the authors calculated the

radii of the breast base curvatures for each subject. They suggested that the global average radii of curvatures, calculated from four curve segments, were useful design parameters for underwire. Participants were classified into two groups using Cluster analysis by the radii of curvatures. The authors implied that specific shapes of underwire designed separately for either group, were necessary. In addition, the volume of breast was calculated, and regarded by the authors, a helpful design parameter for bra cup.

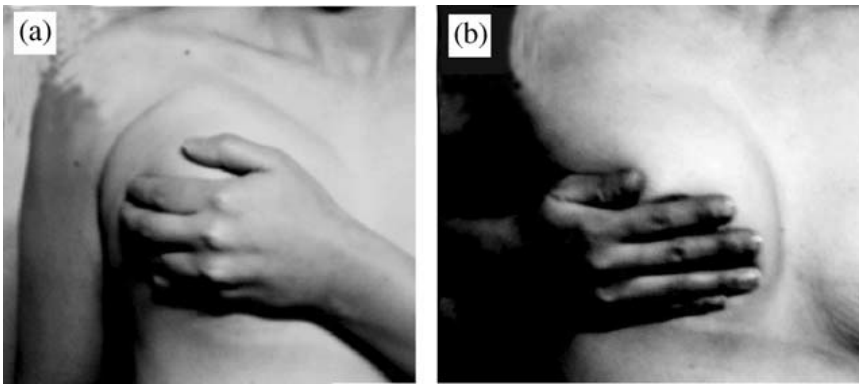


Figure 2-21. The folding line method to detect breast outline
(Lee, Hong & Kim, 2004)

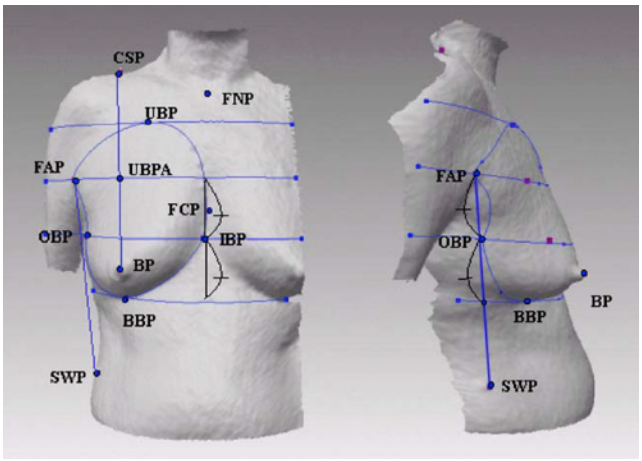


Figure 2-22. Reference points adopted by Lee, Hong & Kim (2004)

Oh and Chun (2014) analyzed the 3D body scans of 32 Korean females, who wore bras with band size 75 (in centimeters), selected from Size Korea study. The scans were sliced horizontally at bust level and underbust level to create two cross-section planes (i.e. transverse planes). Five measurements, including 1 measurement of depth, 1 measurement of width, and 3 measurements of arc-length, were extracted from the scans (Figure. 2-23). The authors proposed a new sizing system specifically for bra cup by referring to breast arc-length. They suggested an interval of 1.5 centimeters for each cup size (e.g. arc-length between 14 to 15.5 cm belonged to AA cup; arc-length between 15.5 to 17 cm belonged to A cup, etc.). They compared their proposed method with the traditional sizing system using bust-underbust-difference. However, the comparison did not clearly show whether the traditional or the proposed sizing strategy was better.

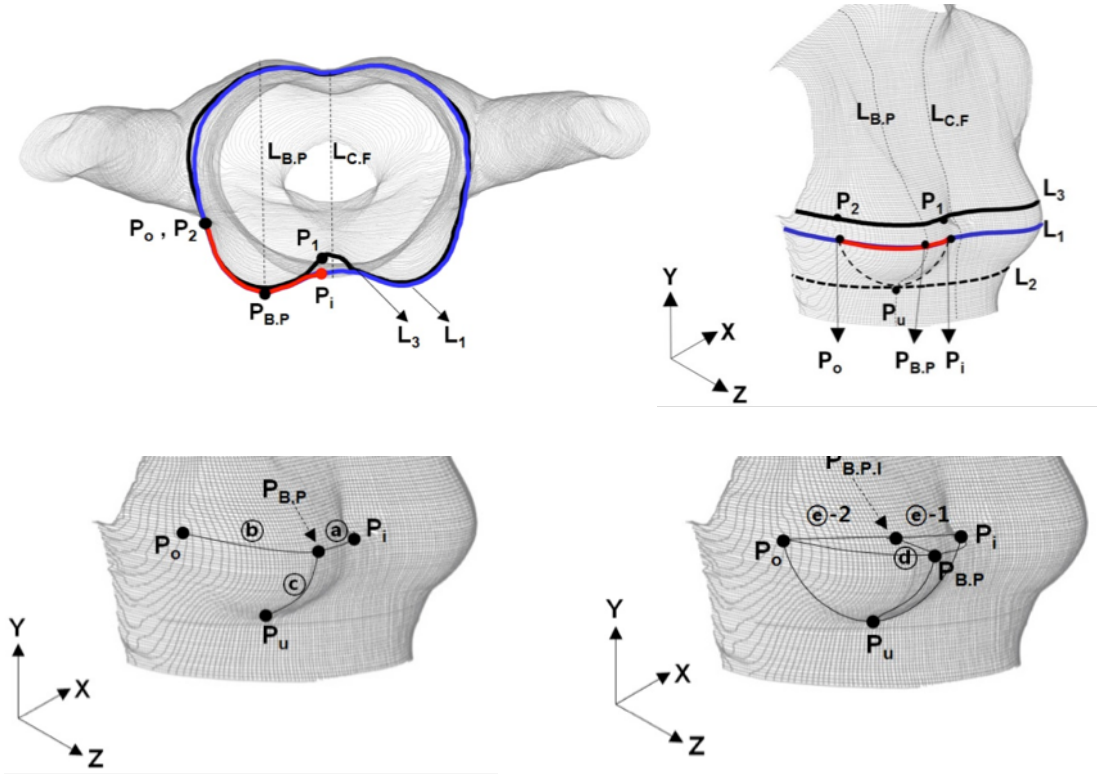


Figure 2-23. Anthropometric measurements used by Oh & Chun (2014)

Zheng, Yu and Fan (2007) studied the 3D breast shape of 456 Chinese female subjects. 98 breast-shape-related measurements, including bust girth, body width and circumference, etc., were extracted from each of the 3D body scans (Figure. 2-24). Three different software programs were used to acquire the 98 measurements, namely 3D-rugle, Rapidform, and Shapeline-3D. Five additional measurements were manually obtained and included in their analysis. Factor analysis was applied to the 103 measurements with varimax rotation. The authors claimed that the first factor, which explained 23.5% of the total variance, were relevant to the overall body build of each participant (Figure. 2-25a), and the second factor, which explained 19.8% of the total variance, were relevant to the volume of breast (Figure. 2-25b). Based on the amount of total variance that factors could explain, they decided that the first two factors were sufficient in classifying breast shapes. Underbust girth and the depth-width-ratio of the breast were chosen and regarded as the main parameters for the design of a new bra sizing system for Chinese women. The new sizing system was considered as an improvement of the existing sizing system with regard to accommodation rates.

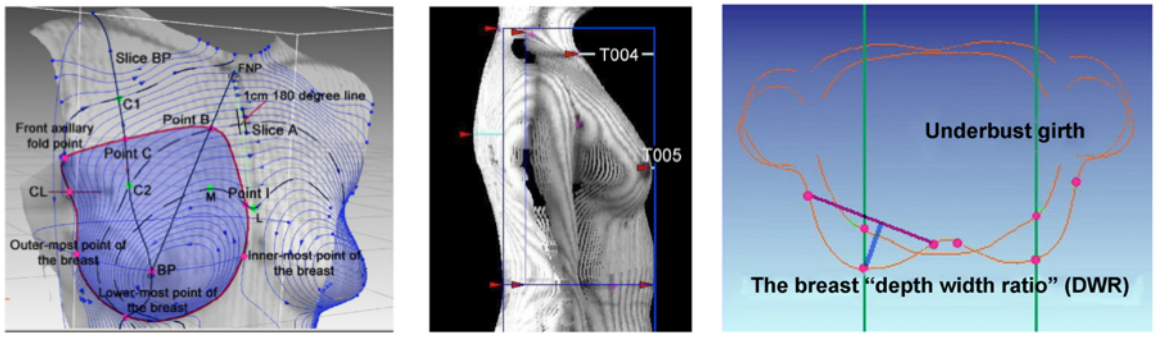
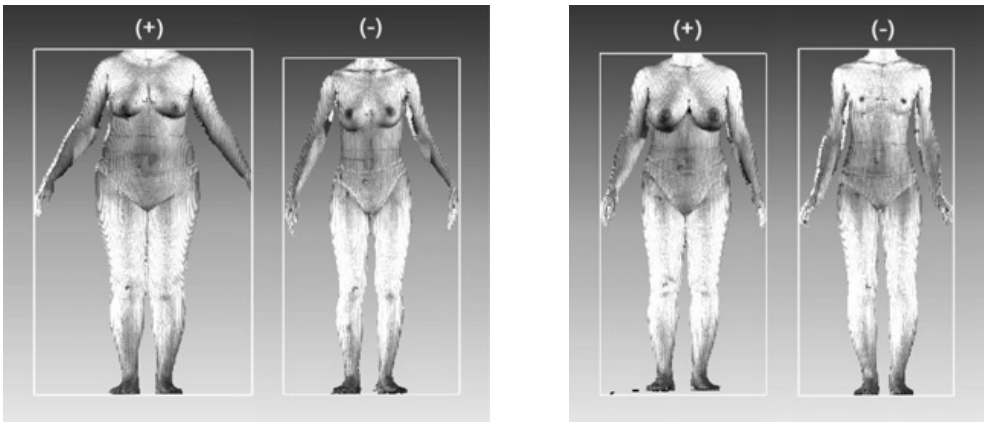


Figure 2-24. A few measurements used by Zheng, Yu & Fan



a). Factor 1. Overall body build b). Factor 2. Volume of breast

Figure 2-25. Extreme figures of the first two factors (Zheng, Yu & Fan, 2007)

2.3. Limitations in Previous Research

As shown in previous sections, body shape studies and breast shape studies have been intensively done in the past. However, some limitations still exist.

Firstly, many of today's body shape studies, especially with respect to apparel design and clothing development, still rely on manual placement of 3D landmarks and manual extraction of body measurements. Admittedly, manual measuring process played an important role for the anthropometric research in the past, and automatic extraction does have some drawbacks. Looking at the geometry of the scan is not the same as having a living body in front of the measurer: interacting with the participant (e.g., palpation to find bony landmarks) makes it easier for the measurer to find desired body landmarks. However, for the analysis of a large scale of body-scans (e.g. scans from a national database), and with a large number of measurements to be extracted, the manual process can be time-consuming. Moreover, sometimes apparel design researchers had to use multiple software programs on the same set of scans in order to acquire all their desired measurements. However, due to the divergence in calibrations between programs and human operation errors, the manually extracted anthropometric data can be error prone.

Secondly, although breast shape study has been done intensively for the Asian population, limited literatures have been found for the Caucasian population, especially for the study of nude breasts. In those limited studies, data were mostly collected from female patients seeking for breast plastic surgery. However, on the one hand, the sample sizes were too small to reach to generalized conclusions. On the other hand, the breast shapes of female patients (even though data were collected before surgery) cannot be considered as representative samples for the general Caucasian population. In addition, the typical scanning posture of breast reconstruction research is different from that of apparel design research. Patients were mostly scanned in a chair leaning backwards, some of them were scanned in a prone position. These scanning postures could result in significantly different results in breast shapes (McGhee & Steele 2006; Pandarum, Yu & Hunter, 2011), compared with those that were obtained from standing posture (the posture adopted by most apparel design researchers).

Thirdly, in some studies, body measurements were analyzed separately, despite that some of them are highly correlated (for instance, bust circumference is positively correlated with hip circumference). Even though some interaction terms may have been added into a statistical model to monitor the effect of correlation, the output from the statistical analysis was simply some p-values, indicating whether each of the interaction terms made a difference. In other words, a p-value can suggest whether correlation exists between the measurements, but it cannot demonstrate in what way the measurements are influencing each other.

Lastly, some of the previous shape studies could be more convincing if validation and justification was made on their statistical analysis results. There is a possibility that researchers sometimes might have over-interpreted their results. For example, many researchers (not limited to the apparel design field) were inclined to interpret PCA result by checking the summary table, which consists of the percentage of variance each PC accounts for, and by checking the loadings (coefficients) of eigenvectors to determine

which variables are important. However, judgments on the following two issues can be difficult to make: a) How many percentages of variance to be included can be considered sufficient? b) How to decide the thresholds between “large” and “small” loadings? Another example with regard to Cluster analysis is that decisions upon the choice of clustering algorithm and the number of clusters can be even more difficult to make. In general, there is no consensus even among statisticians on how to deal with these problems. Nonetheless, these issues are very influential on the analysis results and the final conclusions, and thus deserve more attention and consideration.

CHAPTER 3

METHODOLOGY

3.1. Introduction of CAESAR Data

Carried out by the U.S. Air Force, the CAESAR (Civilian American and European Surface Anthropometry Resource) project collected 3D body scan data and demographic information of participants from three countries, namely the USA, The Netherlands, and Italy. One site in Canada (Ottawa, Ontario) was also included when the survey was being conducted. Hence, the data from the USA population and the Ontario data combined is referred to as the North American population. Demographic information includes gender, age (in years), race, education level, occupation, fitness level, marital status, etc. In addition to the scan data, some traditional apparel measurements, such as chest circumference, neck base circumference and total crotch length, were measured by traditional tools manually. Participants were scanned wearing snug-fitting shorts and a soft sports bra. It is possible the sports bra may influence the shape of the nude breasts to some extent. However, the sports bra does not have as much constraint to the breasts as a daily bra does. Figure 3-1 shows a scan image of one participant from the CAESAR project.

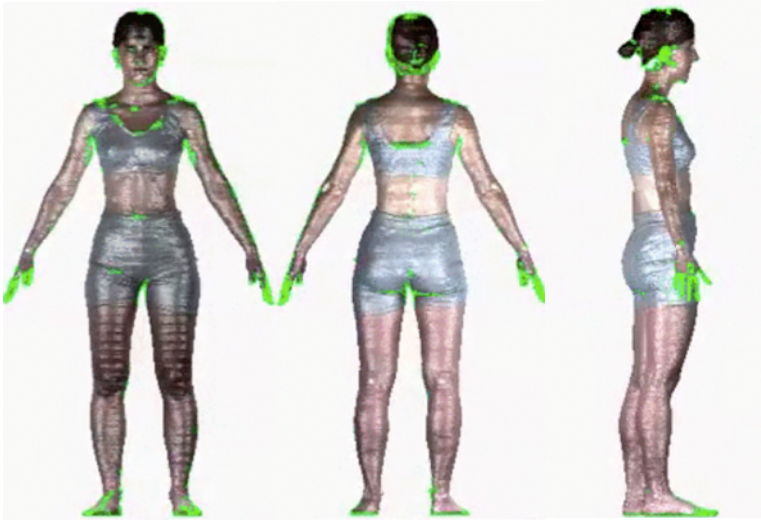


Figure 3-1. 3D body scan image

A total of 4431 participants (2375 from North America; 1255 from The Netherlands and 801 from Italy) joined the survey. Table 3-1 shows the number of participants in each strata. This study focused on the North American female population.

Table. 3-1

Number of Participants in Each Strata

The Netherlands									
	Females					Males			
Age	18-29	30-44	45-65	Total	Age	18-29	30-44	45-65	Total
Dutch	167	200	177	544	Dutch	156	152	172	480
Other	41	48	58	147	Other	29	23	32	84
Total	208	248	235	691	Total	185	175	204	564
Italy									
	Females					Males			
Age	18-29	30-44	45-65	Total	Age	18-29	30-44	45-65	Total
Italian	252	67	57	376	Italian	235	103	50	388
Other	5	4	1	10	Other	14	7	1	22
Total	257	74	58	386	Total	249	110	51	410
North America									
	Females					Males			
Age	18-29	30-44	45-65	Total	Age	18-29	30-44	45-65	Total
White	188	373	394	957	White	191	353	320	867
Black	61	48	56	477	Black	39	52	25	116
Other	58	56	37	151	Other	51	56	30	137
Total	307	477	469	1255	Total	281	461	375	1120

All participants were scanned in two postures: standing and sitting, but only scans of standing pose were used in this study. Moreover, 72 landmarks were manually placed onto their body before scanning. The 3D coordinates of the landmarks were recorded and can be directly accessed. This study concerned only about upper body, thus only landmarks on torso were reserved. There are 21 landmarks around torso (10 in the front, 3 in the back, 8 in the side, see Figure 3-2).

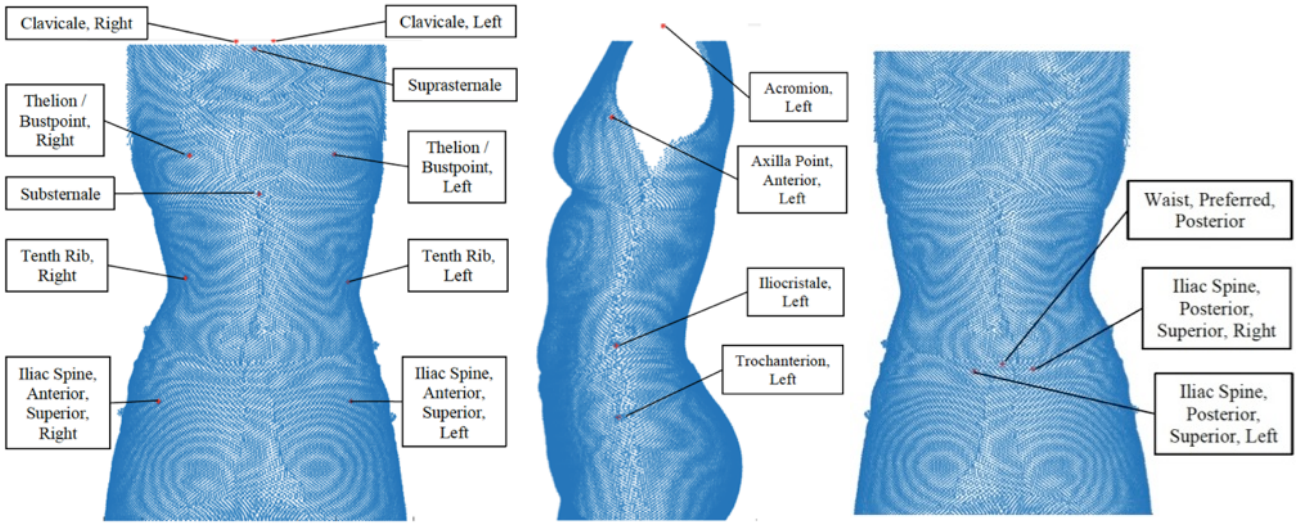


Figure 3-2. Landmarks placed around the torso

3.2. Target Population

Factors such as age, race, and ethnicity are likely to have impact on breast shape. The elderly tends to have more sagged breasts compares with younger population. Different race and ethnicity can also result in significantly difference in breast size and shape. Factors such as fitness level, percentage of body fat, should also be taken into consideration. However, certain amount of variation across population is necessary, not only for the data to be representative, but also to ensure the classification process to be smoothly conducted (If every participant falls into the same breast shape type, there is no point in doing the classification). On the one hand, accurate classification of breast shape is desired, and thus the elimination of the influence of extreme cases is necessary. On the other hand, valid classification is of more importance. This study attempted to find distinctive breast shape categories with no significant breast shape type left out, thus a considerably large sample is required.

Accordingly, the target population of this study is chosen to be the North American female Caucasians, whose age are between 18 to 45, and whose BMI (Body Mass Index)

is below 30. The World Health Organization (WHO) regards BMI over 30 as obese. This study included overweight participants but not the obese population. BMI was not included in the original CAESAR data, but can be calculated via the following formula (Eq. 3-1). Figure 3-3 shows the distribution of BMI of the target population. It approximates a normal distribution. It also shows that sufficient participants with healthy body weight (BMI between 18.5 to 25) were included. 75.7% of the target population falls into the healthy BMI level.

$$\text{BMI} = \frac{\text{Body weight [kg.]}}{(\text{Body height [m]})^2} \tag{3-1}$$

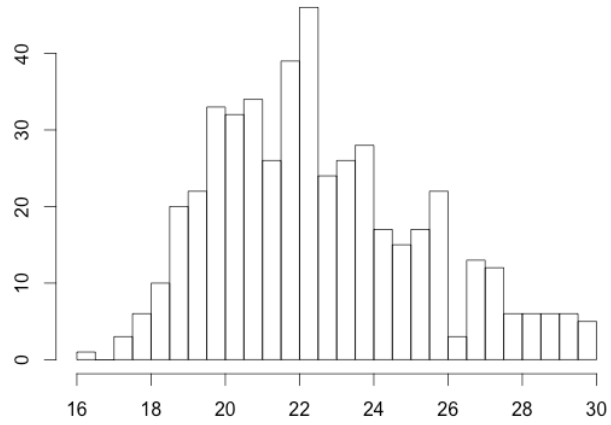


Figure 3-3. Histogram of Body Mass Index of the target population

Figure 3-4 shows the histogram of age of the target population. It is uniformly distributed, which means similar proportion of subjects at each age.

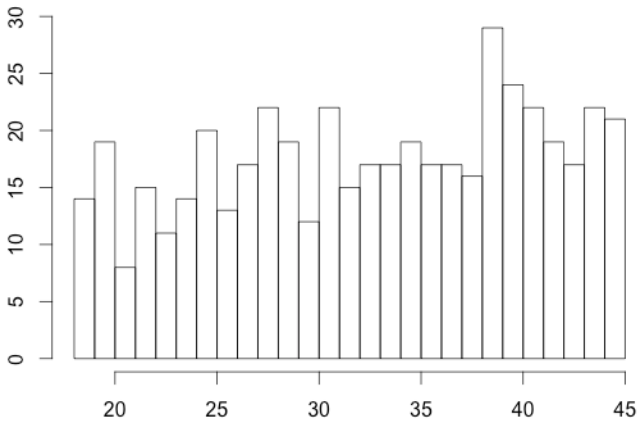
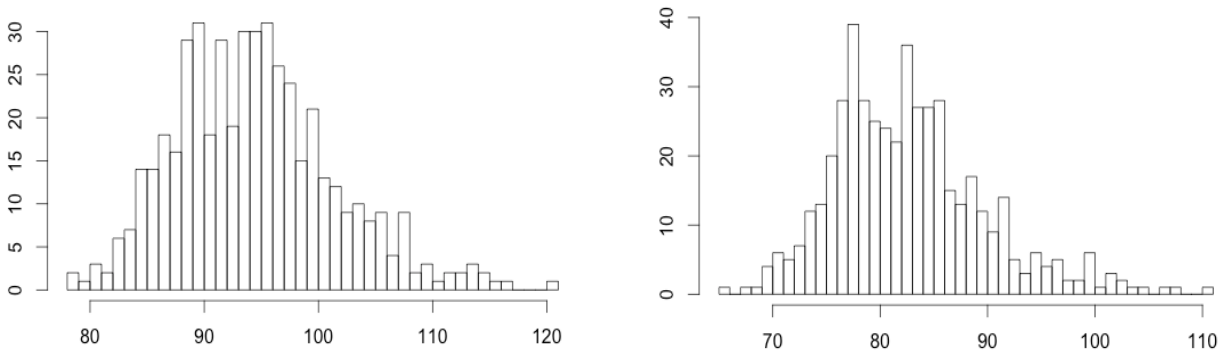


Figure 3-4. Histogram of age (in years) of the target population

Both of the distributions of bust circumference and underbust circumference approximate normal distributions (Figure 3-5).



a) Bust circumference (cm)

b) Underbust circumference (cm)

Figure 3-5. Histograms of two bust measurements

3.3. Body Measurements

All measurements of interest included in the analysis were from bust area. However, the original CAESAR measurements, collected by researchers of CAESAR project, include limited measurements at bust area. The traditional anthropometric bust measurements, bust girth and underbust girth, cannot fully describe the complicated 3-dimentional breast shape. Hence, a majority of measurements in this study were directly extracted from body scans via some self-developed programs. Several transverse planes

sliced at different levels (z-coordinates), i.e. at bust level (averaged z-coordinates of the right and left Thelion/Bust-points), at underbust level (z-coordinate of the Substernale Point) and at armscye level (averaged z-coordinates of the right and left Anterior Axilla Points), were plotted together (Figure 3-6) for each subject. Several sagittal planes sliced at different x-coordinates, i.e. at front central line (x=0), and at left bust point (x-coordinate of the left Bust-point), together with the overall side profile were plotted together (Figure 3-7).

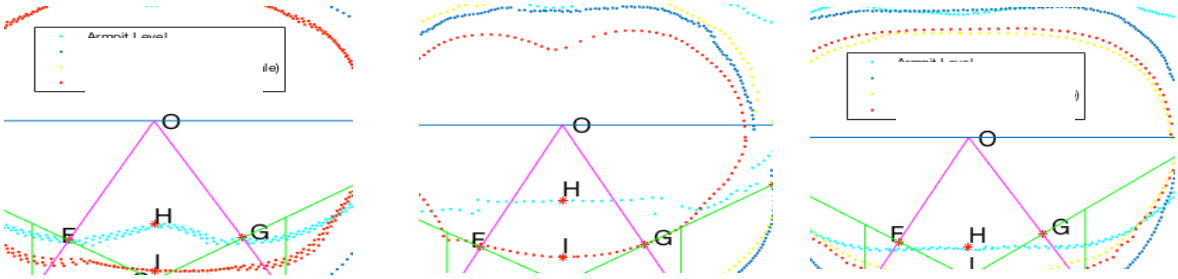


Figure 3-6. Transverse planes sliced at different z-coordinates

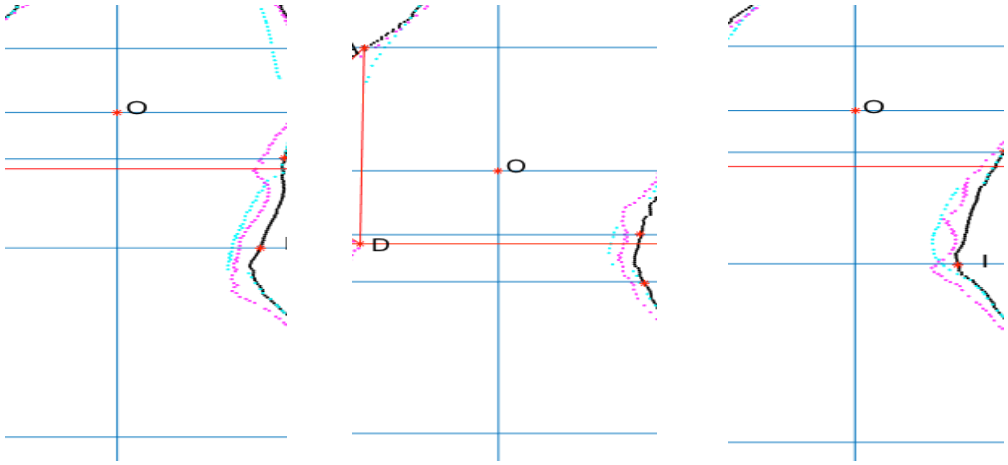


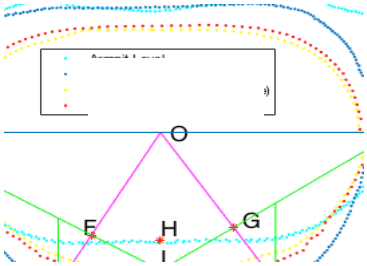
Figure 3-7. Sagittal planes sliced at different x-coordinates

As shown in Figure 3-6 and 3-7, a few auxiliary lines and points were added onto the merged planes. Table 3-2 shows the detailed description of the placement of the auxiliary points. Note that each scan had been shifted before the extraction of planes so

that the pivot axis passes through the origin at (0, 0, 0) (see Figure 4-5 in Section 4.2). The x-coordinate of the pivot axis was calculated by averaging the x-coordinates of all points on torso. Likewise, the y-coordinate of the pivot axis was obtained by averaging all y-coordinates. The z-coordinate of the pivot axis ranges from the neck level to the crotch level.

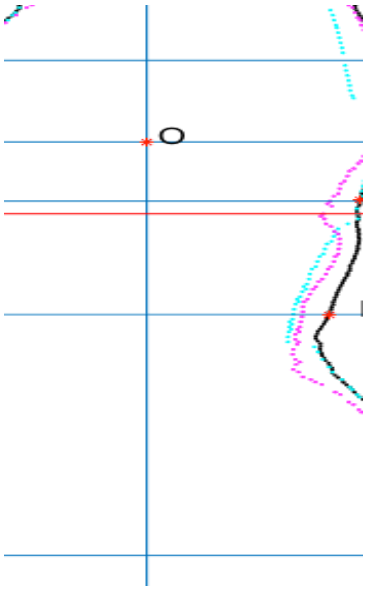
Table 3-2.

Placement of auxiliary points

Auxiliary points on transverse planes		
Illustration	Name	Description
	Point O	Point O is where the pivot axis locates on the x-y plane (at x=0 and y=0)
	Point C	The intersection between Line y=0 (the vertical line passing through Point O) & the outline of the bust plane
	Point I	The intersection between Line y=0 & the outline of the underbust plane
	Point H	The intersection between Line y=0 & the outline of the armscye plane
	Point A	The left bust point
	Point B	The right bust point
	Point D	Point D is the intersection between the line, which is perpendicular to Line AO meanwhile passes through Point C, & the outline of the bust plane;
	Point E	Point E is the intersection between the line, which is perpendicular to Line BO meanwhile passes through Point C, & the outline of the bust plane
	Point F	Point F is the intersection between Line CD & Line AO;

	Point G	Point G is the intersection between Line CE & Line BO

Auxiliary points on sagittal planes

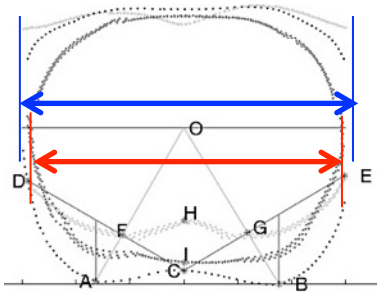
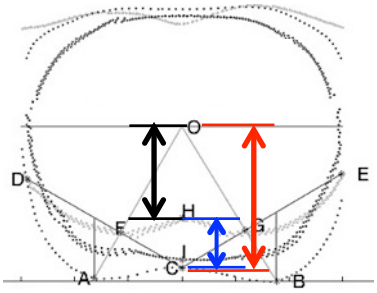
Illustration	Name	Description
	Point O	The intersection between the pivot axis & the horizontal line at bust level (The pivot axis divides the body into the anterior half and posterior half)
	Point B	The anterior intersection between the horizontal line at bust level & the outline of the overall side profile
	Point F	The posterior intersection between the horizontal line at bust level & the outline of the overall side profile
	Point L	The anterior intersection between the horizontal line at bust level & the outline of the sagittal plane sliced at front central line
	Point A	The anterior intersection between the horizontal line at armscye level & the outline of the overall side profile
	Point E	The posterior intersection between the horizontal line at armscye level & the outline of the overall side profile
	Point C	The anterior intersection between the horizontal line at underbust level & the outline of the overall side profile
	Point G	The posterior intersection between the horizontal line at underbust level & the outline of the overall side profile
	Point D	Point D is regarded as the (left) breast root point. It is the turning point on the outline of the sagittal plane sliced at left bust point that connects the breast and the chest wall. Note: The breast root point is a self-defined point. It is different from the underbust point, which is one of the landmarks (referred to as the Substernale

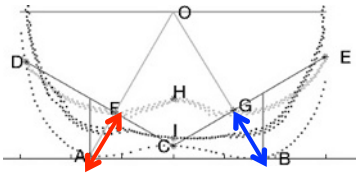
	Point by CAESAR researchers).
Point H	The anterior intersection between the horizontal line at waist level & the outline of the overall side profile
Point I	The posterior intersection between the horizontal line at waist level & the outline of the overall side profile
Point J	The most protruded point at buttock

With the help of the auxiliary points, 30 raw measurements were extracted from transverse planes: 2 of them are widths; 5 of them are depths; 2 are angles; 7 are distances; and 14 are areas (Table 3-3).

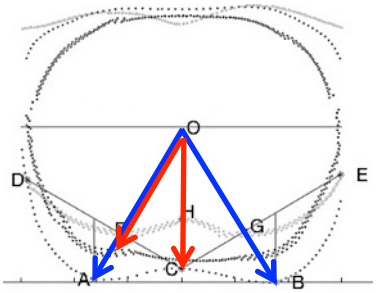
Table 3-3.

Raw Measurements Extracted from Transverse Planes

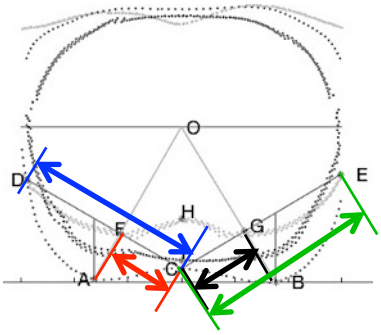
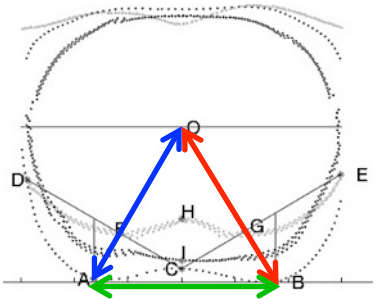
Measure of width		
Illustration	Name	Description
	width_bust	Width of the transverse plane sliced at bust point level (bust plane)
	width_ub	Width of the transverse plane sliced at underbust level (underbust plane)
Measure of depth		
Illustration	Name	Description
	depthHO	Anterior depth of the armscye plane at central front (Length of Line OH)
	depthCO	Anterior depth of the bust plane at central front (Length of Line CO)
	depthIH	Anterior depth difference between underbust plane and armscye plane

		(Length of Line IH)
	depthAF	Depth of right bust point (Length of Line AF)
	depthBG	Depth of left bust point (Length of Line BG)

Measure of angles

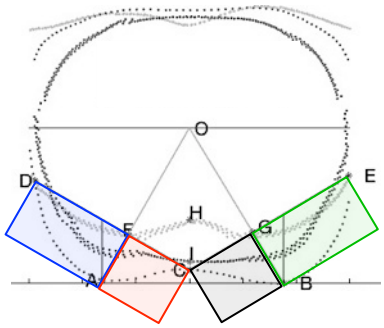
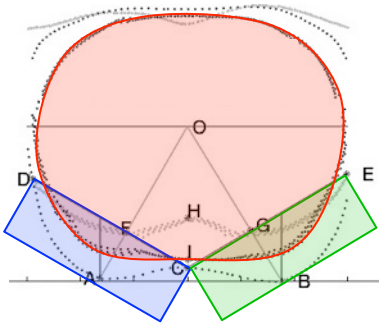
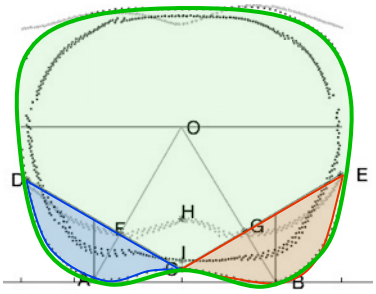
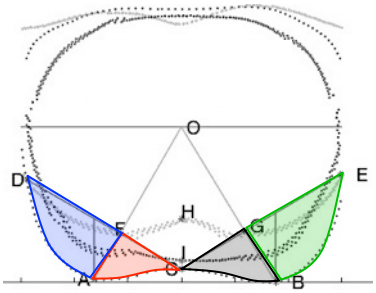
Illustration	Name	Description
	angleBP	Pointing of bust points (Angle AOB)
	angleBP_rt	Pointing of right bust point (Angle AOC)

Measure of distance

Illustration	Name	Description
	distCD	Distance between Point C & Point D
	distCE	Distance between Point C & Point E
	distCF	Distance between Point C & Point F
	distCG	Distance between Point C & Point G
	distAO	Distance between Point A & Point O
	distBO	Distance between Point B & Point O
	BP2BP	Distance between bust points (Distance between Point A & Point B)

Measure of area

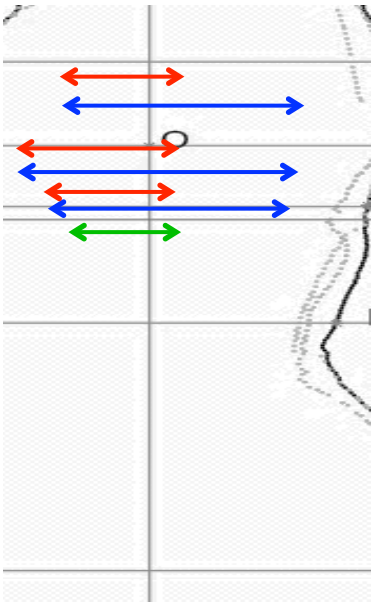
Illustration	Name	Description

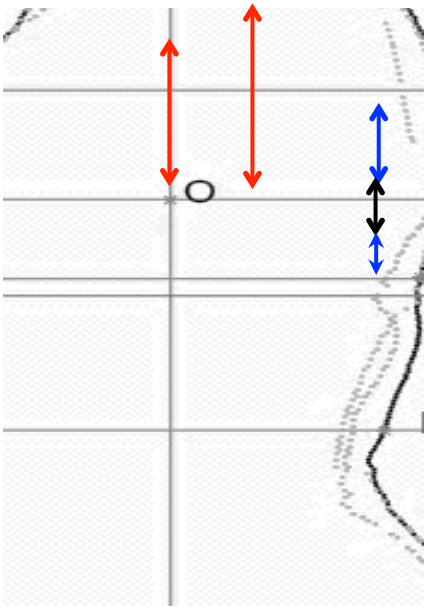
	rec_rt_inner	Area of the rectangle at right inner bust (with width CF & height AF)
	rec_lt_inner	Area of the rectangle at left inner bust (with width CG & height BG)
	rec_rt_outer	Area of the rectangle at right outer bust (with width DF & height AF)
	rec_lt_outer	Area of the rectangle at left outer bust (with width EG & height BG)
	rec_right	Area of the rectangle at right bust (with width CD & height AF)
	rec_left	Area of the rectangle at left bust (with width CE & height BG)
	area_ub	Area of the underbust plane
	area_bust	Area of the bust plane
	area_curveCAD	Area of the arc enclosed by Curve CAD & Line CD
	area_curveCBE	Area of the arc enclosed by Curve CBE &d Line CE
	area_curveACF	Area of the fan enclosed by Curve AC & Line CF & Line AF
	area_curveBCG	Area of the fan enclosed by Curve BC & Line CG & Line BG
	area_curveADF	Area of the fan enclosed by Curve AD & Line DF & Line AF
	area_curveBEG	Area of the fan enclosed by Curve BE & Line GE & Line BG

36 raw measurements were extracted from sagittal planes: 7 of them are thicknesses; 7 of them are height differences; 5 are angles; 6 are distances; and 11 are areas (Table 3-4).

Table 3-4.

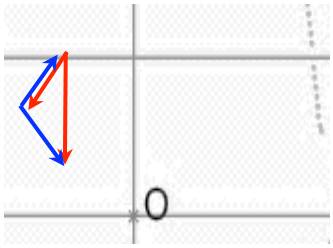
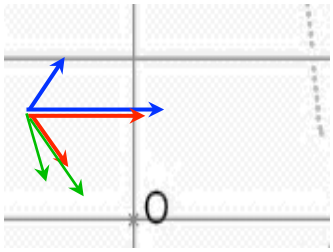
Raw Measurements Extracted from Sagittal Planes

Measure of thickness		
Illustration	Name	Description
	thickAE	Body thickness at armscye level (Length of Line AE)
	thickBF	Body thickness at bust level (Length of Line BF)
	thickCG	Body thickness at underbust level (Length of Line CG)
	thickAO	Anterior body thickness at armscye level
	thickBO	Anterior body thickness at bust level
	thickCO	Anterior body thickness at underbust level
	thickDO	Anterior body thickness at breast root level

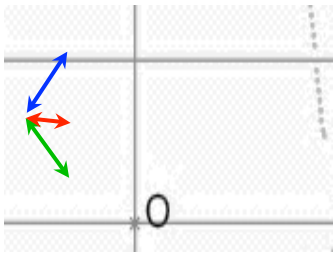
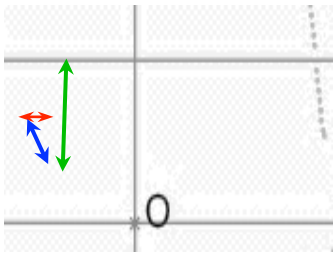
Measure of height difference		
Illustration	Name	Description
	heightEF	Height difference between armscye level & bust level
	heightFG	Height difference between bust level & underbust level
	heightGD	Height difference between underbust level & breast root level
	height_acro2BP_rt	Height difference between acromion point & bust point measured from right side
	height_acro2BP_lt	Height difference between acromion point & bust point measured from left side
	height_shd2armp_rt	Height difference between acromion point & bust point measured from right side

	height_shd2armp_lt	Height difference between acromion point & bust point measured from left side

Measure of angles

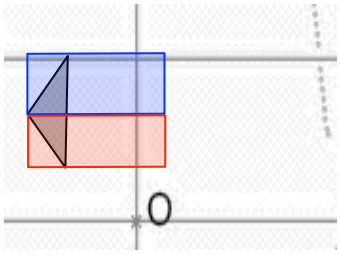
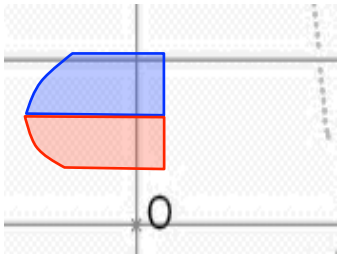
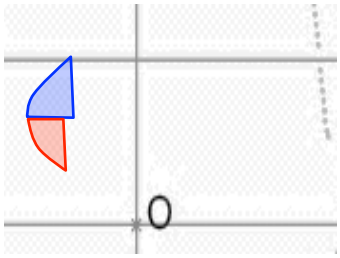
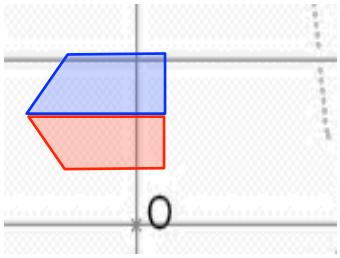
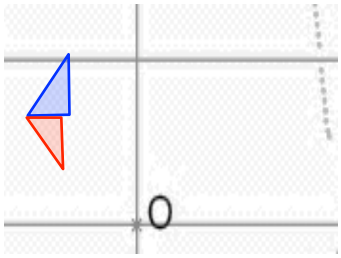
Illustration	Name	Description
	angleBAD	Angle BAD from the self-constructed Triangle ABD
	angleABD	Angle ABD from the self-constructed Triangle ABD
	angleABF	Angle ABF (upper bust)
	angleCBF	Angle CBF (lower bust)
	angleCBD	Angle CBD

Measure of distance

Illustration	Name	Description
	distAB	Distance between Point A & Point B
	distBD	Distance between Point B & Point D
	distB_AD	Distance between Point B & Line AD
	distAD	Distance between Point A & Point D
	distBC	Distance between Point B & Point C
	distBL	Distance beween Point B & Point L

Measure of area

Illustration	Name	Description

	areatriABD	Area of the self-constructed Triangle ABD
	area_rec_upper	Area of the upper rectangle (with width BO & height EF)
	area_rec_lower	Area of the lower rectangle (with width BO & height FG)
	areaBupper	Area enclosed by Curve AB, y-axis, Line AE & Line BF
	areaBlower	Area enclosed by Curve BD, y-axis, Line BF & the horizontal line passing through D
	area_curvAB_linA	Area of the fan enclosed by Curve AB, the vertical line passing through Point A, & Line BF
	area_curvBD_linD	Area of the fan enclosed by Curve BD, the vertical line passing through Point D, & Line BF
	area_trap_upper	Area of the trapezoid enclosed by Line AB, Line BF, y-axis & Line AE
	area_trap_lower	Area of the trapezoid enclosed by Line DB, Line BF, y-axis & the horizontal line passing through Point D
	area_tri_upper	Area of the triangle enclosed by Line AB, Line BF & the vertical line passing through Point A
	area_tri_lower	Area of the triangle enclosed by Line BD, Line BF & the vertical line passing through Point D

One of the main body shape theories suggested that shape is independent from size (Mossimann, 1988). This theory was supported by many studies where researchers found that people who wear the same size of garments are not necessarily to have the same type of body shape; but people who wear different sizes may have similar body shape (Apeagyei, 2008; Faust, Carrier & Baptist, 2006; Simmons, Istook & Devarajan, 2004). This study also adopted this theory, assuming female breast shape does not depend on breast size, and concentrated on the shape factors.

To eliminate the effect of size, instead of using the raw measurements, 34 ratios were constructed. For instance, thickR_underb_bust is the ratio between the body thickness at underbust level (thickCG) and thickness at bust level (thickBF) from sagittal plane. The 34 ratios were regarded and referred to as 34 variables for the analysis. In addition, angle is another way of measuring shape that is not influenced by the absolute size. Thus the 7 angles extracted from the scans were directly employed as variables. Angles are in radians rather than in degrees, although it makes little difference for the analysis (The conversion between the two metrics is just a multiple of a scalar, either $\pi/180$ or $180/\pi$). The list of all 41 variables and how they were calculated can be found in Appendix A. Note that Variable 1 (circmR_deltaB_bust) is the only ratio calculated from the traditional apparel measurements that are directly available in CAESAR data. It is the ratio between deltaB and bust circumference, where deltaB is the difference between bust circumference and underbust circumference.

3.4. Data Mining

Data of this study include 41 variables and 478 observations (478 subjects). It is essentially a large data matrix with 41 columns and 478 rows. Multivariate analysis, and many data mining technics are suitable for this kind of data in terms of understanding the relationship among the variables (it is inappropriate to assume independence among variables as most of the univariate analysis requires), and discovering patterns in the data

if there is any. The goal is to classify the subjects into a few groups which have significantly different multivariate means across the 41 variables, and to propose some classification rules to assign future cases into the groups using fewer variables.

The first multivariate statistical method adopted is Principle component analysis (PCA) to remove the correlations in variables so that the original 41 variables can be converted into 41 uncorrelated Principal components (PCs). 41 variables suggest there are 41 axis directions. Each of the observations lies in this 41-dimensional space. PCA can also provide information to facilitate judgments regarding dimension reduction (some of the dimensions could have so little variation that can be ignored).

Cluster analysis is another common multivariate analysis method. It is used to identify heterogeneous groups (or clusters) and place homogeneous observations into the same group. There are numerous clustering methods and it's difficult to claim one clustering technique to be superior to another. Different clustering methods applied to the same data can lead to very different cluster outcomes, especially when inappropriate cluster number is chosen. Therefore, instead of adopting only one clustering method, three most commonly used methods, namely K-means, K-medoids, and Hierarchical Clustering, were included in the analysis. The outcomes of the three methods were compared and the comparison result contributes to the decision making in selecting cluster number.

Determining the number of clusters is an important yet challenging task. Lee, Hong and Kim (2011) classified their breast shape data into two clusters. However, they did not provide their reasoning in choosing cluster number. Zheng, Yu and Fan (2007) utilized K-means cluster analysis on their data and decided to set the cluster number to be eight, the same number of cup size choices in traditional Chinese bra sizing system (from AA cup to G cup). Song and Ashdown (2011) selected the number of clusters based on the assumption that similar number of people should be included in each cluster. Unlike previous research, this study proposed two different criteria to address this issue, based

on two different perspectives aiming at different purposes. The first criterion is based on misclassification rate, and focuses on checking how well a future case can be classified into the correct group. Discriminant analysis was conducted and discriminant functions were created from the clustering results. Decision about the ideal cluster number is made according to the misclassification rates obtained from Linear discriminant analysis (LDA), and the OOB error rates obtained from Random forest (RF) analysis. The second criterion is based on goodness-of-fit of model and concerns about how well the model can capture and explain the majority of variations in the data. Three goodness-of-fit measures were used as reference, namely BIC, AIC and WSS statistics.

After the number of clusters was finalized, MANOVA (Multivariate analysis of variance) was applied to examine whether the multivariate means of different clusters are significantly distinctive. Four major MANOVA test statistics (i.e. Wilks' Lambda, Roy's Maximum Root, Hotelling-Lawley Trace, and Pillai's Trace) were used in calculating p-values.

Lastly, this study proposed an approach to reduce data dimension and the number of variables. Principal component loadings provide information on how influential each variable is on each PC direction. In addition, Random forest analysis returns the importance of each variable in placing observations into the correct groups (by measuring the decrease in node impurity). Therefore, with the reference of PC loadings and Random forest importance measures, a few key variables were selected from the 41 to start with. Then multiple trials were undergone. Each time one more variable would be excluded and K-means clustering would be applied to the new PCs calculated from the remaining key variables. The new clustering result would be compared with the original. A similar result would lead to further variable deduction. A significantly distinctive result would lead to the reservation of the variable and the attempt of deleting another variable. (The similarity or dissimilarity was judged by side profiles of the breasts, with the help of a Matlab program, developed to visualize clustering outcomes, see Section 4.4). The

deletion sequence also referred to the PC loadings and RF importance measures. Variables with lower importance were the first ones to be considered for exclusion. The iteration of this process stopped when nothing more could be deleted.

CHAPTER 4

PROGRAM DEVELOPMENT IN MATLAB

4.1. Introduction of Matlab

Matlab (MathWorks®) is a programming language and a computing environment widely used in the fields of engineering and science. Scientists and engineers use Matlab to develop programs, to process text, and image, and to graphically display their ideas and research outcomes. Matlab has extensive libraries of mathematical, statistical, simulation, and other tools. It also allows interfacing with program written in other languages, such as Java, Python and C++. In addition, because it has been so widely used in engineering and sciences, abundant resources, instructions and tutorials can be found online. Online communication community has been maturely developed. Some programs written and shared by others can be easily found and accessed through those communication platforms. In conclusion, Matlab is a useful and helpful tool for researchers.

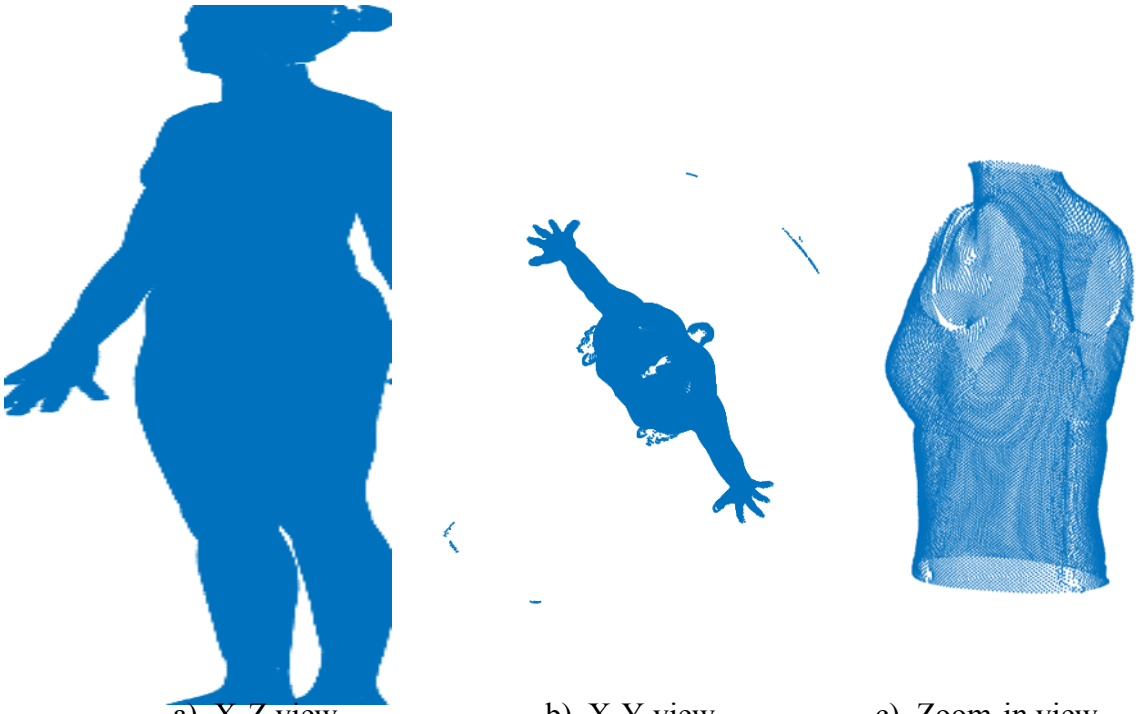
This study involves the 3D body scans of 478 subjects, and 66 unconventional raw measurements for each scan. Manual extraction of this large amount of measurements would not only be time-consuming, but would also be error prone. An alternative way is by developing programs to achieve automatic extraction. The larger the sample size is, the more time it can save, compared with manual extraction. Therefore, for the purpose of feasibility, high efficiency and accuracy, several programs were developed in Matlab to deal with the scans and to obtain body measurements.

4.2. Preparation of the Body Scans

Body scans originally included in the CAESAR data are in the format of PLY (Polygon File Format). PLY format supports storage of 3-dimensional data from 3D scanner as a list of flat polygons (triangles for most cases). Properties such as color, transparency, texture coordinates are stored. However, some of the information (e.g.

color) is not the primary concern for this study. What is truly important is the 3-dimensional coordinates of the vertexes of each of those polygons. Inclusion of redundant information will bring extra workload to computer, and will prolong operating time. Moreover, the PLY files cannot be directly opened in the Matlab environment and requires importing and format conversion. A Matlab function called read_ply was found from the online discussion community. It successfully read data from each PLY file and returned vertex coordinates and connectivity information of the triangular meshes. Nothing but the coordinates were exported and saved as MAT files. Each file was essentially an N by 3 numeric matrix with N being the number of points (vertexes). The format conversion was a very time-consuming process. It took approximately 14 seconds to convert one body scan file. Figure 4-1 shows an example of the direct output from the format conversion function.

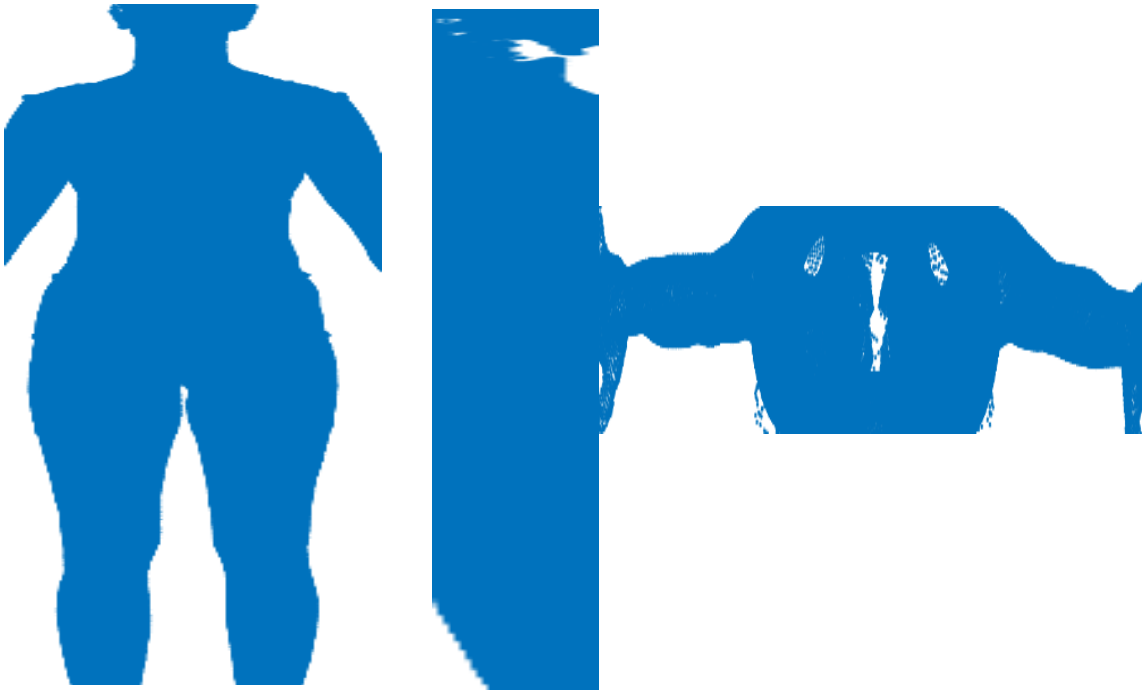
According to CAESAR report, also as shown in Figure 4-1a and b, participants were scanned facing at an angle (approximately 45 degrees) from the cameras, which means that the reference point for all the coordinates (the (0, 0, 0) point for the axes) is not optimal for further analysis. Therefore, rotation of the scans is necessary. In order to find the precise angle that each participant was facing, a program was written to align three pairs of landmarks (each pair consists of one landmark on right side of body, and the corresponding landmark on left side of body). The first pair was the left bust point and the right bust point. Angle between the X-axis and the line passing through the two bust points was recorded as the first reference angle. The second and the third pair were landmarks at tenth rib (left and right) and landmarks at radial styloid (left and right) respectively (see Figure 3-2). Their corresponding angles were recorded as the second and the third reference angle respectively. The averaged value calculated from the three reference angles was treated as the rotation angle. A short function was written to rotate each of the scans according to their rotation angles.



a). X-Z view b). X-Y view c). Zoom-in view

Figure 4-1. Original body scan plotted in Matlab

Furthermore, the raw scans often include some error points that do not belong to participant's body. For instance, it can be observed from Figure 4-1a and b that part of the edge of the round platform that participants were standing on has been included in the body scan file. Another short function was written to deal with this issue. Figure 4-2 shows how the same scan in Figure 4-1 looks like after the rotation and cleaning (exclusion of error points).

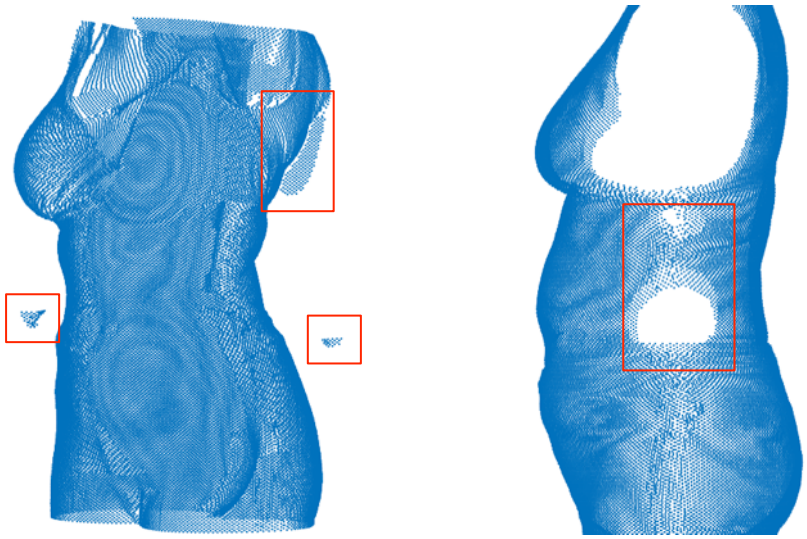


a). X-Y view b). Y-Z view c). X-Z view

Figure 4-2. Body scan after rotation and cleaning

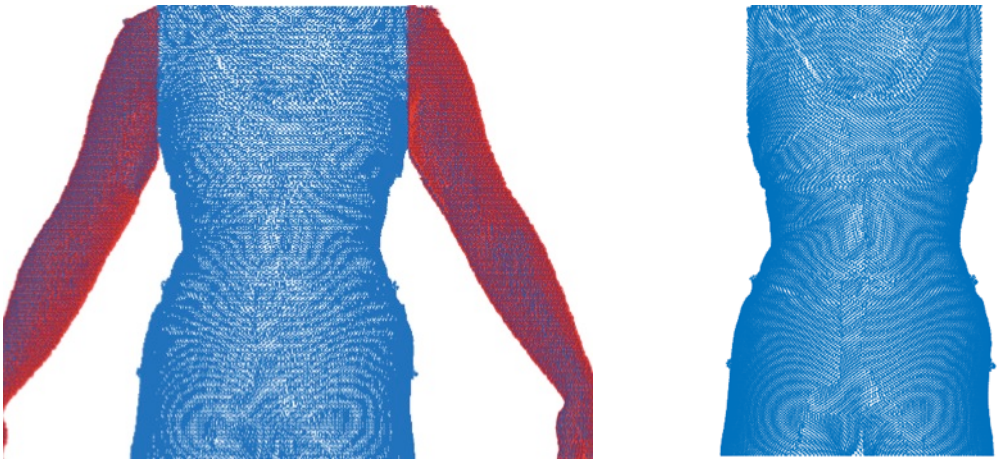
Measurements and landmarks on limbs and head were not the concern of this study. In order to enable and improve the extraction of circumferential measurements on torso, only the torso section was kept while the rest was removed for each scan. The truncated version of body scan can also significantly lower the number of vertex points and, more importantly, their coordinates, thus it can relieve computation and storage pressure on computer for large data matrices. Initially, Codes were developed to remove limbs and head. However, automatic removal only worked for legs and head, but not the arms. Some part of the arms could not be removed completely (Figure 4-3a). In other cases, some part of the torso could be wrongly cut off (Figure 4-3b). Therefore, arms were manually selected and removed for all scans. There is an interactive function available in Matlab figure window that allows selection (brushing) of points. Brushed points are

colored red and can either be deleted or saved as a new variable. Accordingly, points on arms were carefully selected (Figure 4-4a) and deleted (Figure 4-4b).



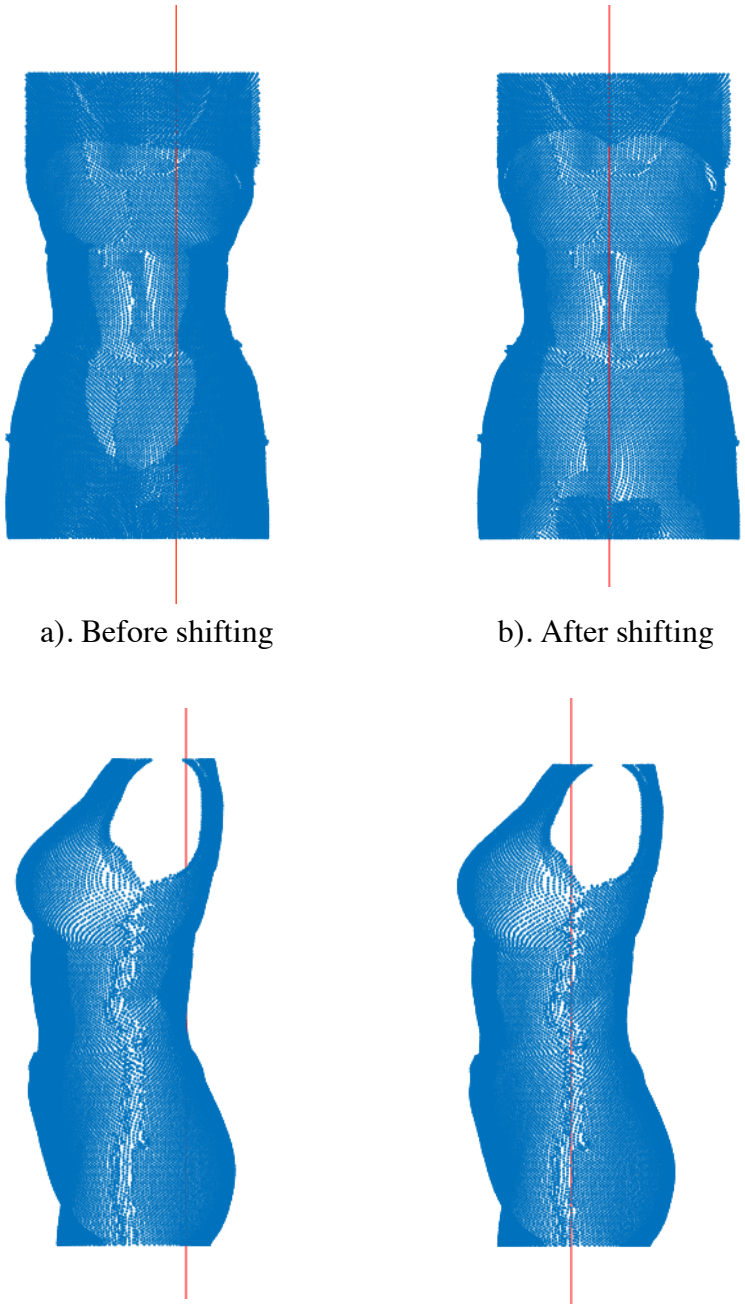
a). Failure Case 1 b). Failure Case 2

Figure 4-3. Cases when automatic removal of arms fails



a). Manual selection of vertex points b). Torso after arms manually removed

Figure 4-4. Manual selection and removal of arms



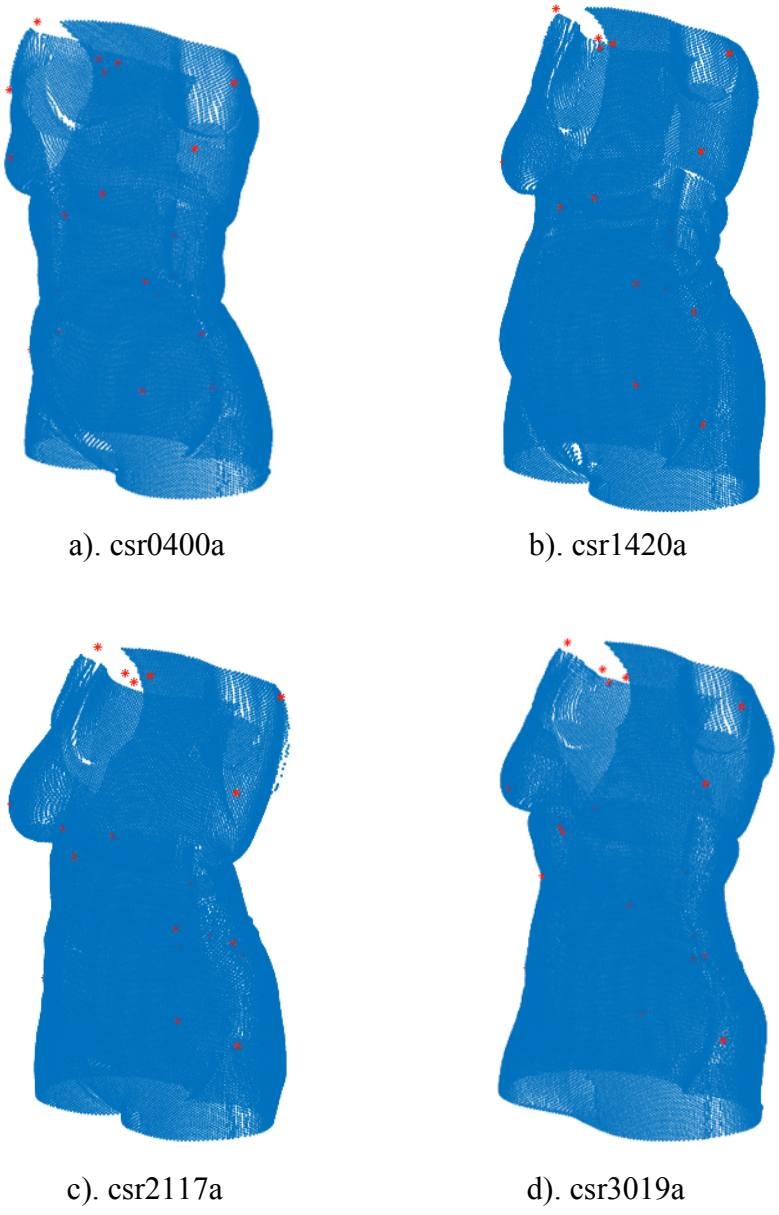
a). Before shifting b). After shifting

c). Before shifting d). After shifting

Figure 4-5. Shifting of torso along X-axis and Y-axis

To simplify calculation in further programing procedure, torso of each scan was shifted (along X-axis and Y-axis) to ensure its pivot axis passes through the origin point at (0,0,0), as demonstrated in Figure 4-5 (the red lines are the vertical line passing

through the origin). The x-coordinate of the pivot axis was defined as the averaged value of x-coordinate values from all vertex points on torso. The y-coordinate of the pivot axis was the averaged value of all y-coordinate values on torso. Figure 4-6 shows four examples of the final outputs of torso, with the manually identified landmarks from the study highlighted in red. To this point, scans were ready for measurement extraction.



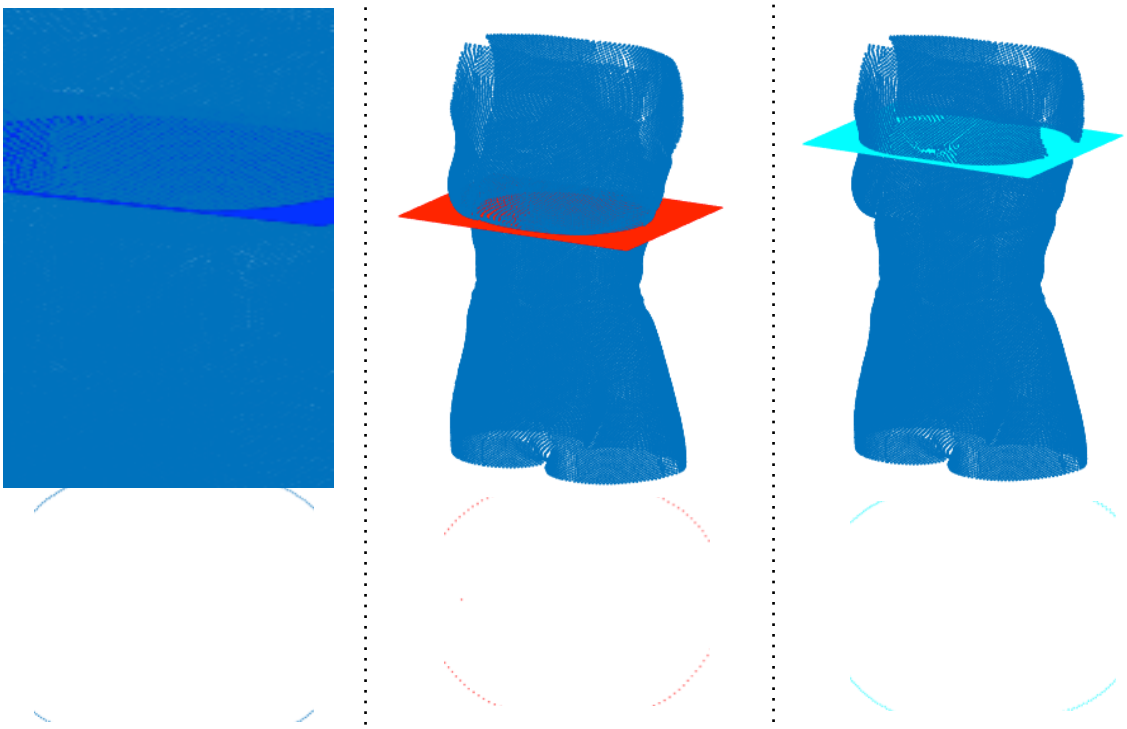
a). csr0400a b). csr1420a

c). csr2117a d). csr3019a

Figure 4-6. Torsos with the manually identified landmarks (four examples)

4.3. Extraction of Planes

Before body measurements were extracted, planes were firstly extracted from each of the torsos. Transverse planes were sliced horizontally at three different levels (i.e. z-coordinates) as shown in Figure 4-7. Three transverse planes were obtained for each scan, namely bust plane (sliced at bust level, i.e. at the averaged z-coordinates of the Thelion/Bust-points, see Figure 4-7a), underbust plane (sliced at underbust level, i.e. at the z-coordinate of the Substernale Point, see Figure 4-7b), and armscye plane (sliced at armscye level, i.e. at the averaged z-coordinates of the Anterior Axilla Points, see Figure 4-7c). Plotting planes together allows for both visual and quantitative investigation on the relationship among planes at different levels. Therefore, the three planes were plotted together (Figure 4-8). A few auxiliary points and lines were added onto the plot to facilitate the extraction of measurements.



a). Transverse plane sliced at bust level

b). Transverse plane sliced at underbust level

c). Transverse plane sliced at armscye level

Figure 4-7. Transverse planes sliced at different levels (z-coordinates)

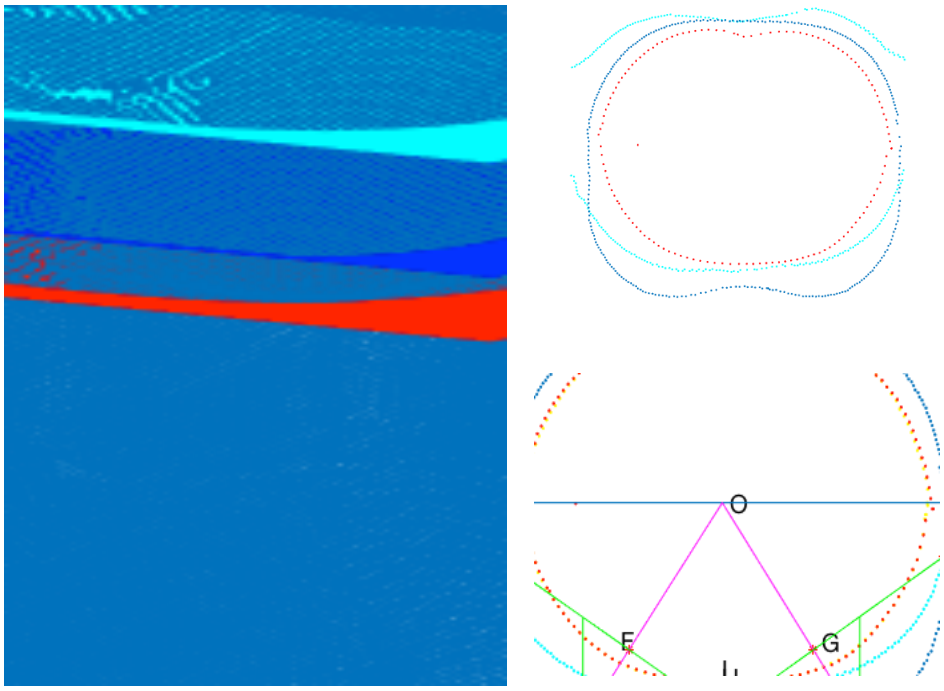
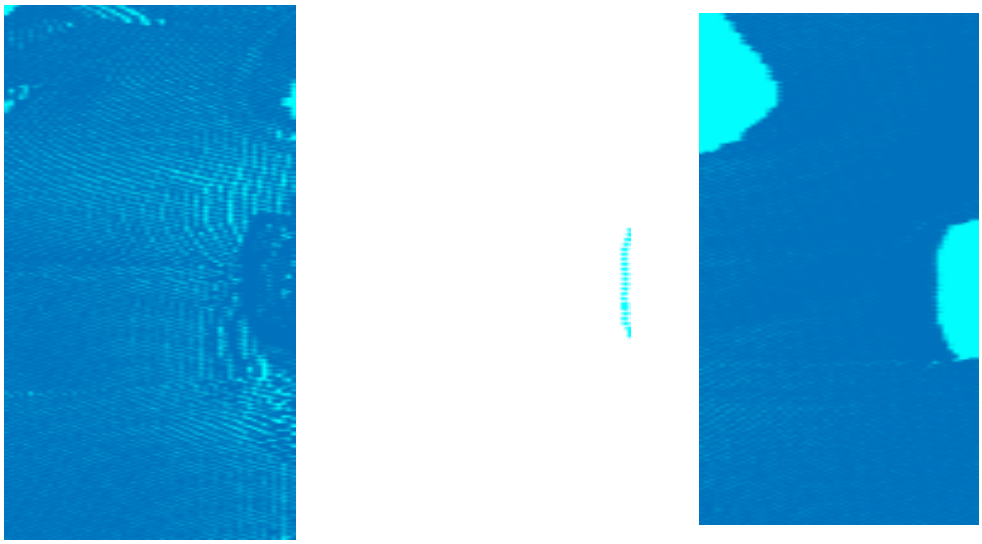


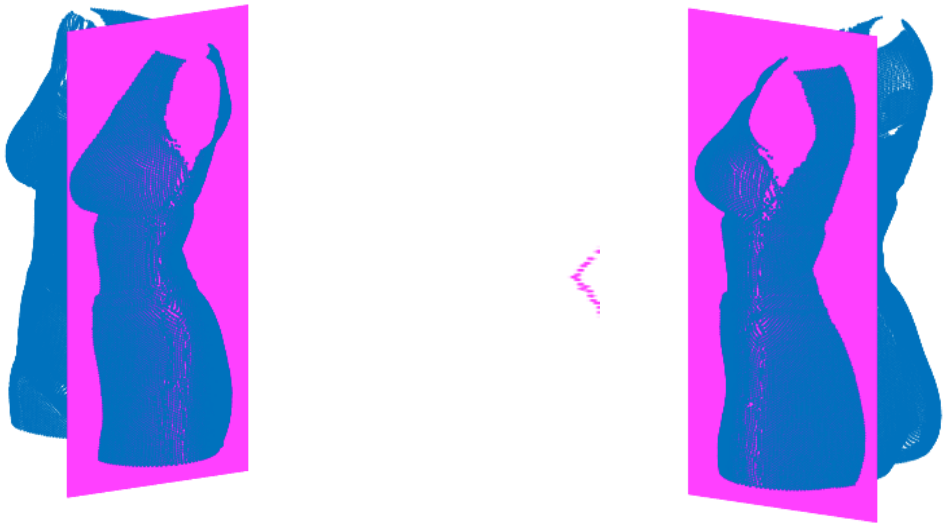
Figure 4-8. Transverse planes plotted together (with auxiliary lines and points added)

Similarly, sagittal planes were sliced vertically at two different x-coordinates, i.e. at central line (x=0), and at left bust point, as shown in Figure 4-9b and c. The overall side profile of each torso was also obtained (Figure 4-9a).

a). Overall side profile



b). Sagittal plane sliced at central line (x=0)



c). Sagittal plane sliced at left bust point

Figure 4-9. Sagittal planes sliced at different x-coordinates

As shown in Figure 4-9b, there are some gaps on the outline of the sagittal plane at both the anterior and posterior body, around bust area. The gaps were caused by the sports bra that participant was wearing. In other words, the real front central line of the (nude) body does not have the bulge in front as this bulge is actually the bra being suspended away from the body. Similarly, part of the outline at posterior body is actually

the sports bra, suspended away from the body. Because of the existence of the bra, the actual sternum point at the bust level cannot be obtained.

The three sagittal planes were then plotted together (Figure 4-10) with a few auxiliary points and lines added.

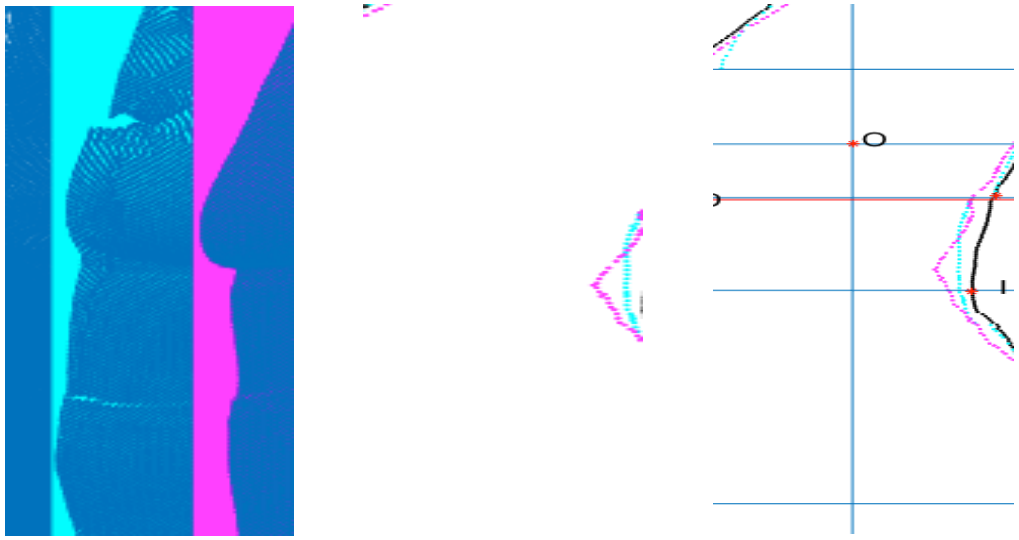


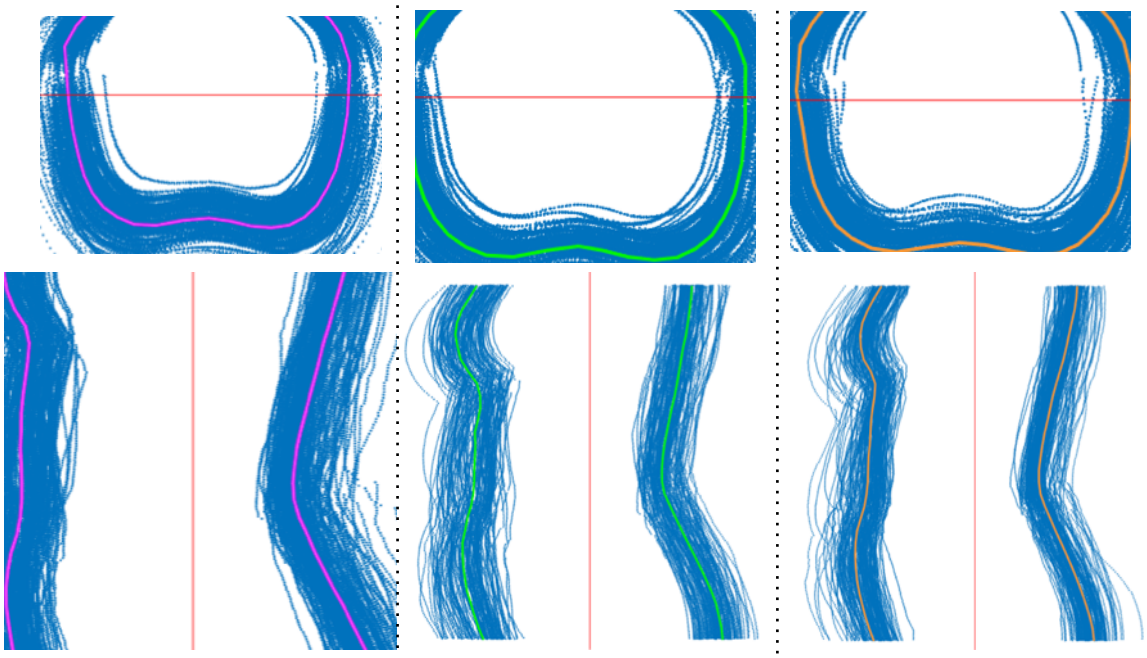
Figure 4-10. Sagittal planes plotted together (with auxiliary lines and points added)

4.4. Algorithm to Acquire Median Curves

To visually present the breasts classification results, an algorithm was developed to acquire the median curves of each group.

Figure 4-11 demonstrates one grouping (clustering) result, where subjects were classified into three groups. Subjects that belong to the same group were plotted together. The figure shows both the plotting of the overall side profiles and that of the bust planes. Considering that this study focused only on shape instead of size, each individual side profile was scaled so that the height difference between armscye level and hip level was equal to 1. The side profiles were then shifted so that their armscye lines were at the same height level. The algorithm for side profile searches for the median point among all body curves, which belong to the same group, at a constantly decreasing height level (from armscye level to hip level). The median point searching is done separately for the anterior

body and the posterior body. After the searching is finished, the anterior median points are connected together in sequence. So are the posterior median points.



a). Group 1 b). Group 2 c). Group 3

Figure 4-11. Obtaining median curves for each group

Meanwhile, each individual bust plane was scaled to have the thickness of posterior body equal to 1. The bust planes were shifted so that their centroid points were at the origin. Then the Cartesian coordinates (i.e. X-Y numerical coordinates) of every point on the bust planes were converted to polar coordinates using Eq. 3-2. In a polar coordinate system, the location of a point is determined by an angle (α) from a reference direction (in this case, the positive x-axis direction), and a distance (r) from a reference point (in this case, the origin).

$$r = \sqrt{x^2 + y^2}$$

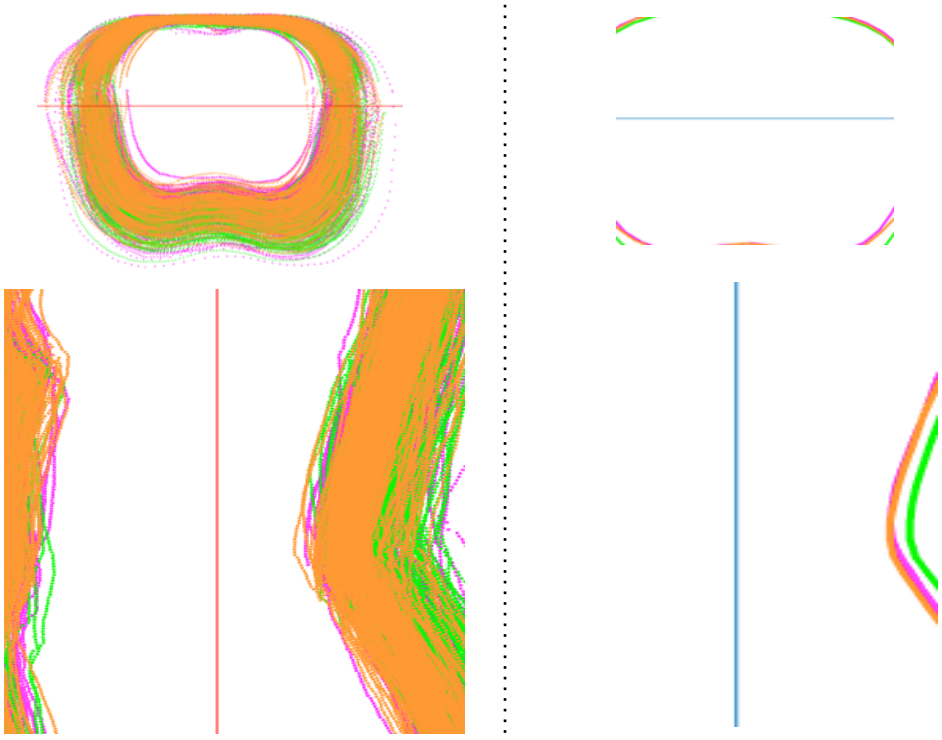
$$\alpha = \begin{cases} arctan(\frac{x}{y}) & \text{if } x > 0 \\ arctan(\frac{x}{y}) + \pi & \text{if } x < 0 \text{ and } y \geq 0 \\ arctan(\frac{x}{y}) - \pi & \text{if } x < 0 \text{ and } y < 0 \\ \frac{\pi}{2} & \text{if } x = 0 \text{ and } y > 0 \\ -\frac{\pi}{2} & \text{if } x = 0 \text{ and } y < 0 \\ \text{undefined} & \text{if } x = 0 \text{ and } y = 0 \end{cases} \quad (3\text{-}2)$$

where x is the first coordinate of a point in Cartesian coordinate system, and y is the second coordinate.

The algorithm for bust plane searches for the median distance r among all planes, which belong to the same group, at a constantly increasing angle α (from 0 to $2 \times \pi$). After the searching is done, the coordinates of the median points are changed back to the Cartesian coordinates (Eq. 3-3) for plotting.

$$x = r \cos \alpha$$
$$y = r \sin \alpha \quad (3\text{-}3)$$

From Figure 4-11 alone, it is difficult to declare whether the three groups are distinctive enough. For that reason, subjects in all three groups were plotted together with colors (Figure 4-12a). The three median curves were also plotted together (Figure 4-12b), and distinctions in shape can be observed at and only at bust area.



a). Subjects belong to all three groups plotted together

b). Three median curves plotted together

Figure 4-12. Three groups plotted together

CHAPTER 5

DATA EXAMINATION AND PREPARATION

5.1. Importance of Data Examination

Multivariate analysis allows for investigation on inter-correlation among numerous variables, both predictor (independent) and response (dependent) variables. Compared with a univariate method such as a paired t-test and ANOVA, it raises the analysis of data into a high dimensional level. Because of these facts, outliers in data can be much more influential for a multivariate method, resulting in biased, if not completely different, analysis outcomes. Outliers can also be much more difficult to find. An outlier in a one-dimensional histogram or two-dimensional scatter plot does not necessarily make it an outlier in high-dimensional space. For some analysis (e.g. Principal component analysis), outliers are also able to make themselves less outstanding by distorting the transformation of the entire data matrix into the directions that they are pointing (points in data can be and are often expressed in the form of vectors). Moreover, most of statistical metrics for multivariate outlier detection have unsatisfactory behavior in handling the masking (unable to detect outlier) effect or the swamping (mislabeling points as outliers) effect. Some methods have calculation-intensive algorithms and require too much computational power for the use of the methods on large datasets to be feasible. The difficulty in finding outliers is probably one of the reasons why outlier detection was rarely done by previous studies.

Nor was the examination of data distribution sufficiently done in the past. Most statistical methods have some underlying assumption, such as linearity, normality, homoscedasticity (constant variance), etc. Violation of the assumptions can lead to biased models, over complicated models, or failure in the fitting of models. Assumption violations can be fixed by data transformation (also called data re-expression). Data transformation can help to obtain homogeneity in variance, to achieve linearity, to improve normality of a variate or of model residuals, and to improve simplicity of

structure for models. It can improve the results of analysis even if it does not require the aforementioned assumptions. In fact, some statisticians argued that data re-expression should be considered first before any other analysis is done. Some common re-expression methods are: logarithm, inverse, square root, 2^{nd} power, 3^{rd} power, etc. In conclusion, careful data examination and diagnostics is necessary. Data preparation, including outlier deletion and transformation, can be very helpful.

5.2. Shiny App for Interactive Plots

There are a total of 41 variables included in this study. Hence, data examination requires plotting of 41 histograms, and plotting of 1681 (41×41) pairwise scatter plots. In order to make this data probing process feasible and to improve efficiency, a Shiny app was built specifically for this dataset to display interactive plots. Shiny (RStudio™) provides an app development framework for R language, and allows users to turn analyses into interactive web applications. The programing codes of the Shiny app developed for this study can be found in Appendix B.

Figure 5-1 shows the initial user interface when the codes were run. The main panel displays the histogram of the first variable (circmR_deltaB_bust) on the left and the corresponding Q-Q plot (Quantile-Quantile plot) on the right. The sidebar panel on the left is where users can operate to make interactive changes. The slider on top of the panel allows users to choose any of the 41 variables for plotting. Underneath the slider are four buttons. The first one navigates to the original dataset which includes all North American female subjects aged 18-45 (variables are the original ratios and angles, and have not been transformed). When the first button has been selected, the histogram is colored grey (Figure 5-2a). The second button navigates to the re-expressed version of the dataset with the same population as the first button (variables that were skewed have been transformed; all variables are in their standardized form). When it has been selected, the histogram is colored blue (Figure 5-2b). The third and fourth buttons link to datasets for

the target population (North American female Caucasians, aged 18-45, with obese population excluded). The third button associates with the untransformed version of the data of the target population while the fourth button associates with the transformed and standardized data. Their corresponding colors are yellow and red respectively (Figure 5-2c and d). At the very initial phase when none of the button is selected, the histogram has no color (Figure 5-1).

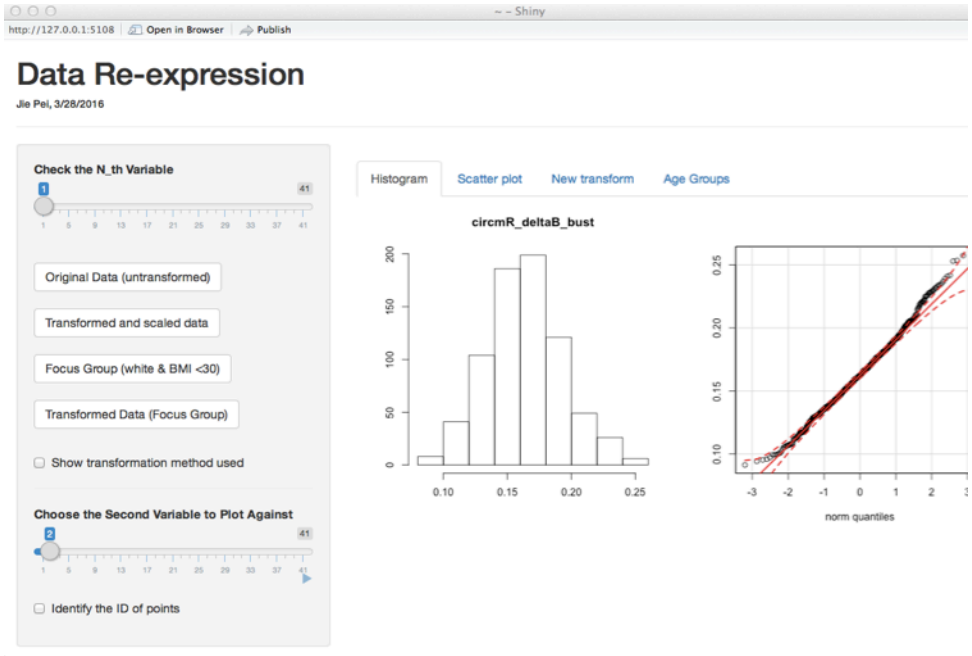
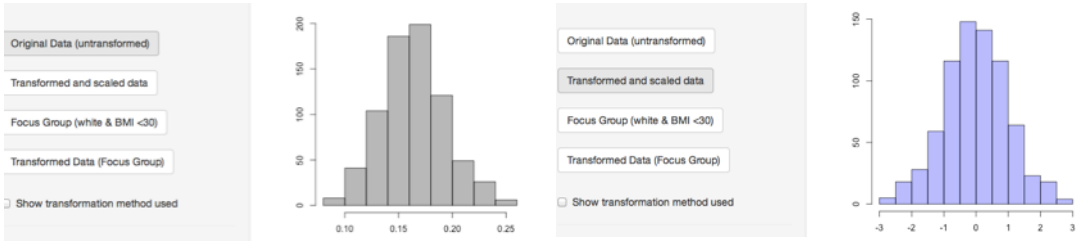
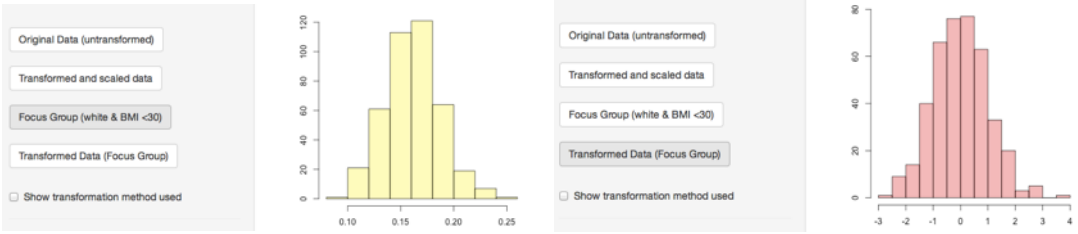


Figure 5-1. Initial user interface of the Shiny app



a). The first button

b). The second button



c). The third button d). The fourth button

Figure 5-2. The four buttons in the sidebar panel and their associated colors

By checking or unchecking the "Show transformation method used" option, users can decide whether the transformation method applied to the data should be displayed (Figure 5-3). It will always demonstrate "no transformation" when the first or third button is selected (because they are linked to the untransformed data).

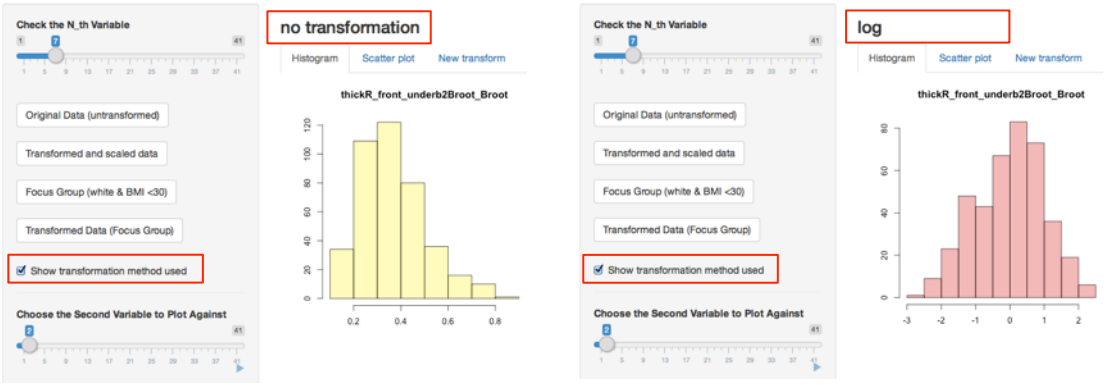


Figure 5-3. Display of the transformation method used (log stands for logarithm)

In addition to histogram and Q-Q plot, users can choose a second variable (via the other slider at the bottom of the sidebar panel) and plot it against the first variable (selected via the top slider) in a scatter plot (Figure 5-4). The use of the two sliders significantly improved the efficiency for the data examination. Furthermore, by checking the "Identify the ID of points" option (inside the sidebar panel region) and by selecting one or more points (inside the scatter plot region), users can identify subject IDs that associate with

the points (Figure 5-5). The IDs, as well as the x and y coordinates, can be found under the scatter plot. In the opposite way, users can type in the ID of one subject and the associated point will be highlighted in red inside the scatter plot (Figure 5-6). Unless the text (subject ID number) is removed from the textbox, its associated points will be highlighted in other scatter plots as well. This function facilitates the process of outlier detection (if a subject is truly an extreme case in high-dimensional space, it is very likely that the subject will appear to be an outlier in many two-dimensional scatter plots).

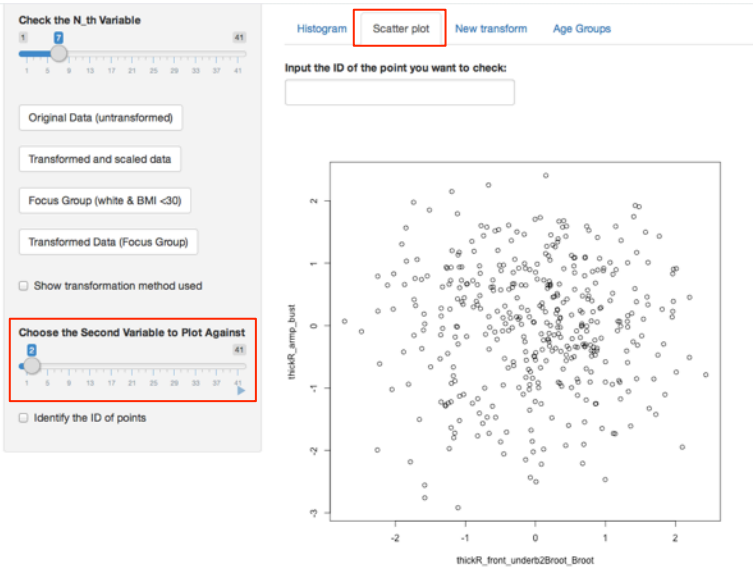


Figure 5-4. Scatter plot demonstration example (Variable 7 against Variable 2)

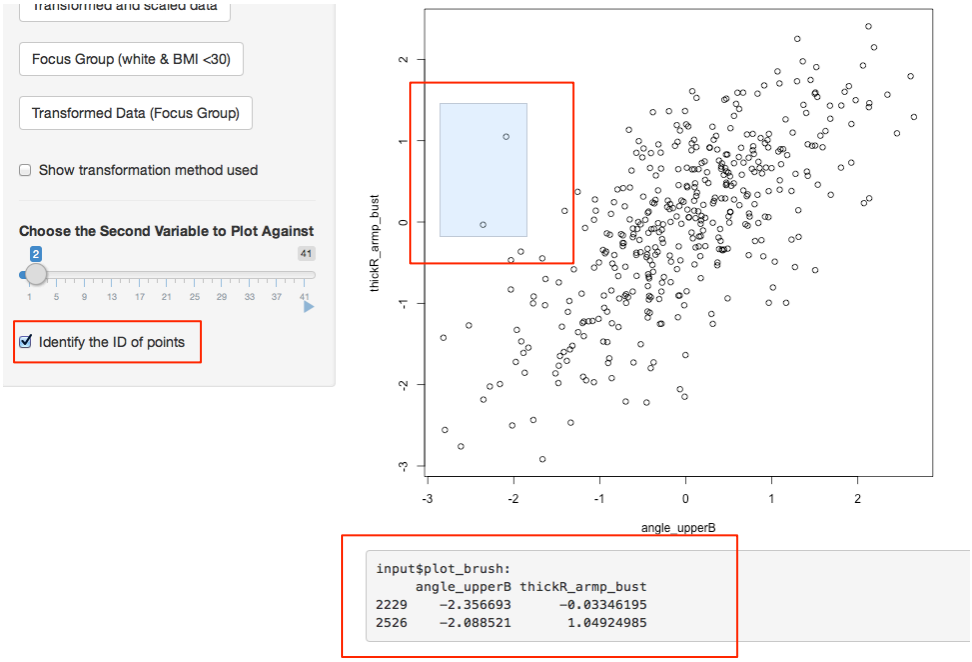


Figure 5-5. Identification of points in a scatter plot

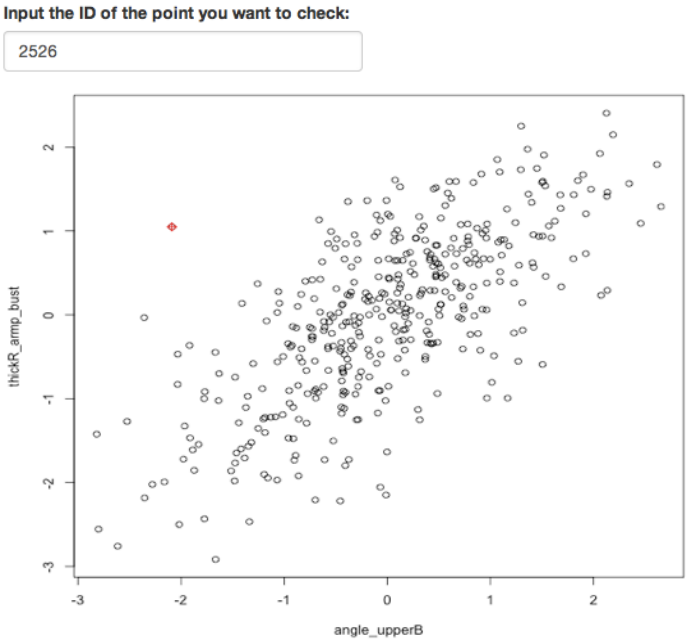


Figure 5-6. Highlight one point according to subject ID

Initially, transformation methods were decided via Box-Cox method. The Box-Cox function in R suggests the best λ to use for the power transformation ($\lambda = 2$ suggests that

the second power should be applied; $\lambda = 3$ suggests that the third power should be applied, $\lambda = 0.5$ suggests the square root, etc.; note that $\lambda = -1$ suggests the inverse and $\lambda = 0$ suggests the logarithm). The function generates a log-likelihood plot according to each value of λ. Theoretically λ can be any real number, but the integer λ that maximize the likelihood is preferred (sometimes $\lambda =$ multiples of 0.5 is also acceptable). As an example shown in Figure 5-7, λ around zero has the maximum log-likelihood, and $\lambda = 0$ is within the 95% confidence interval region. Therefore, the logarithm transformation is reasonable. The issue about Box-Cox function is that it is not very robust, and can sometimes be influenced by extreme outliers. In addition, it sometimes gives out very large λ value (e.g. $\lambda = 10$). However, unless there is a very strong reason, it is not preferable to use a λ value this large. Data re-expression is not data manipulation only to make results look better. In most cases, some rationales can be observed behind the scene. Here are some examples: sometimes ratios are taken in arbitrary order and reciprocal may be just as meaningful; the measure of growth (e.g. population, bacteria, income, salary, etc.) is usually based on current size (or value) and increase exponentially, thus the logarithm transformation makes good sense; in the case of this study, the square root of an area ratio may be more helpful than the area ratio itself. On the contrary, the rationale for the 10th power is too weak to be actually adopted.

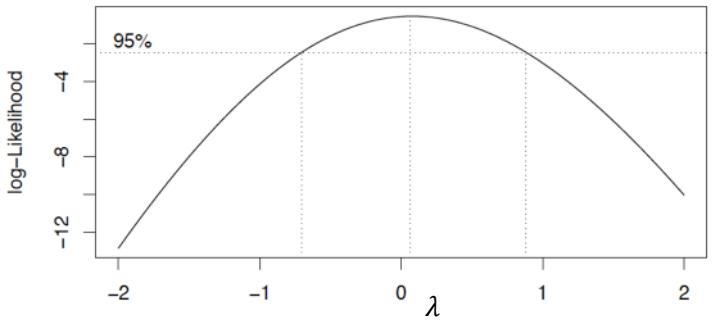


Figure 5-7. Box-Cox power transformation log-likelihood plot

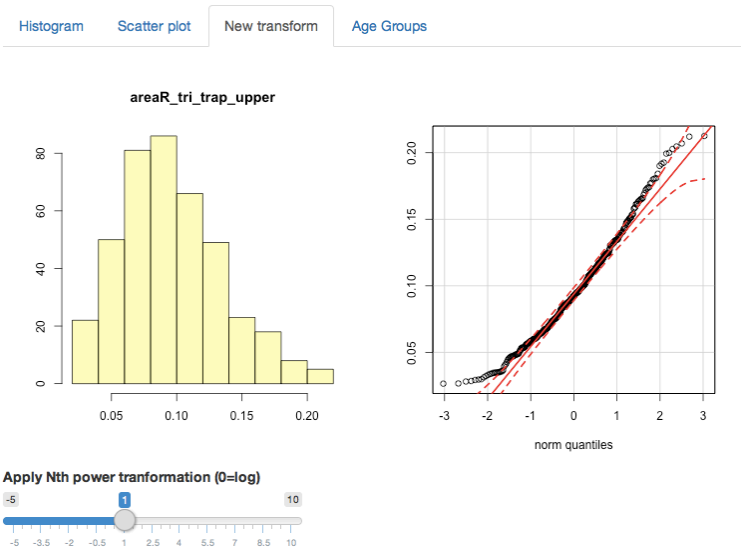
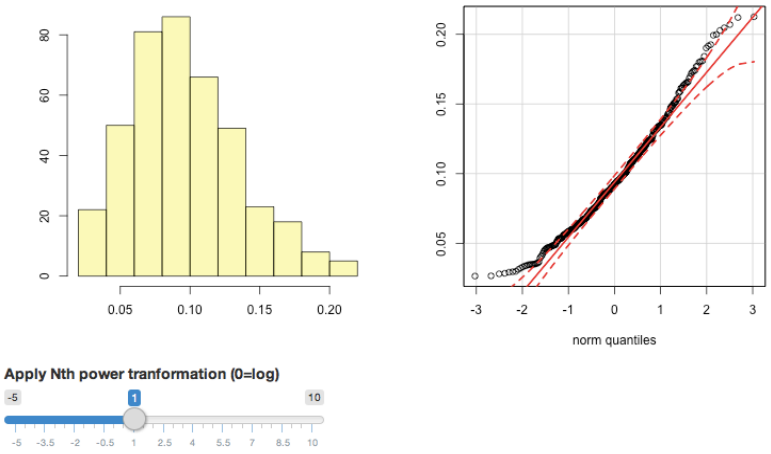
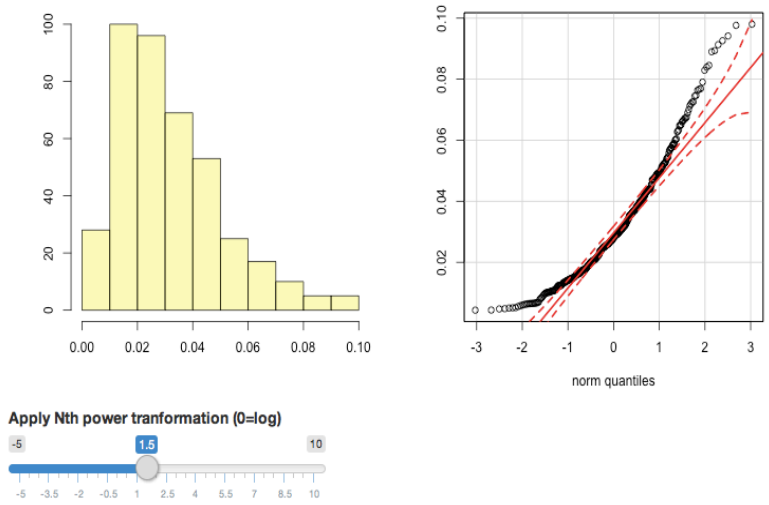


Figure 5-8. Interface under the "New transform" tablet

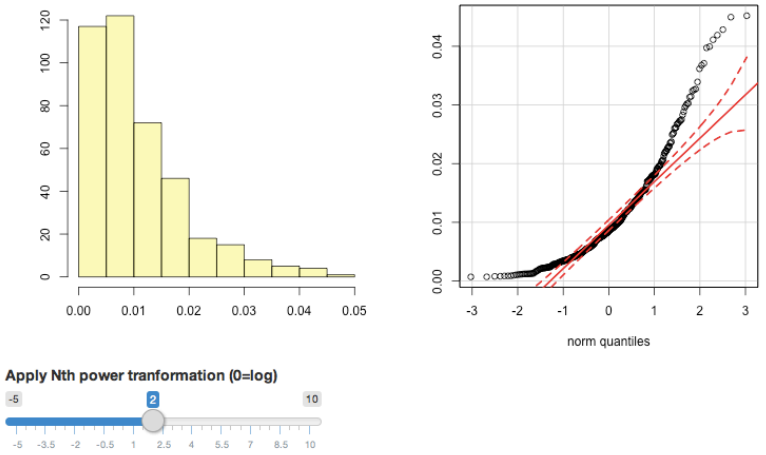
To fix the problems of Box-Cox method, the Shiny app provides an option for users to manually select λ (from multiples of 0.5) and watch changes in distribution via histogram and Q-Q plot. As shown in Figure 5-8, the interface looks like the initial interface presented in Figure 5-1, but with an additional slider below the histogram for users to choose λ value. Initially the value of the slider is set to be 1, suggesting no power transformation being applied (anything to its first power equals itself). Figure 5-9 is the demonstration of the function.



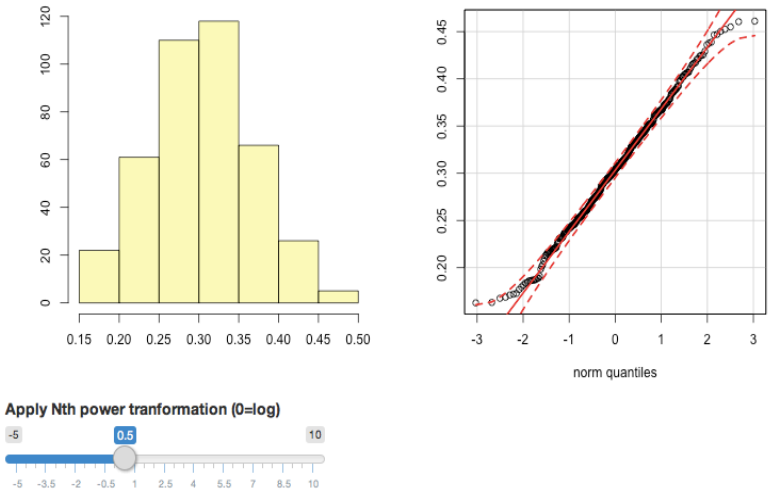
a). 1st power (no transformation)



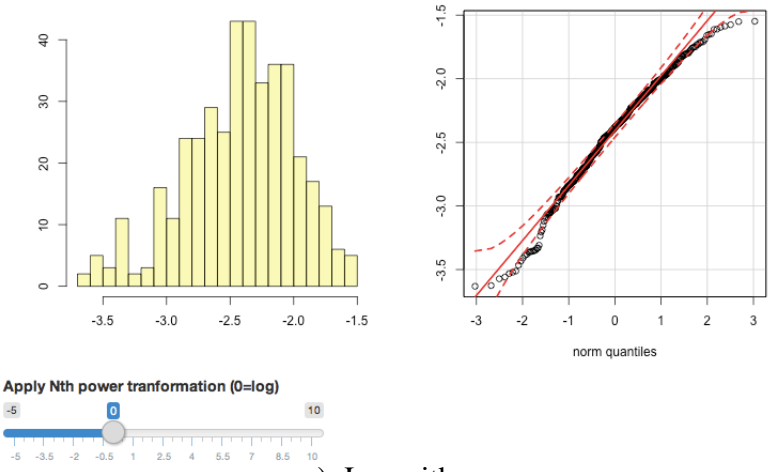
b). 1.5th power



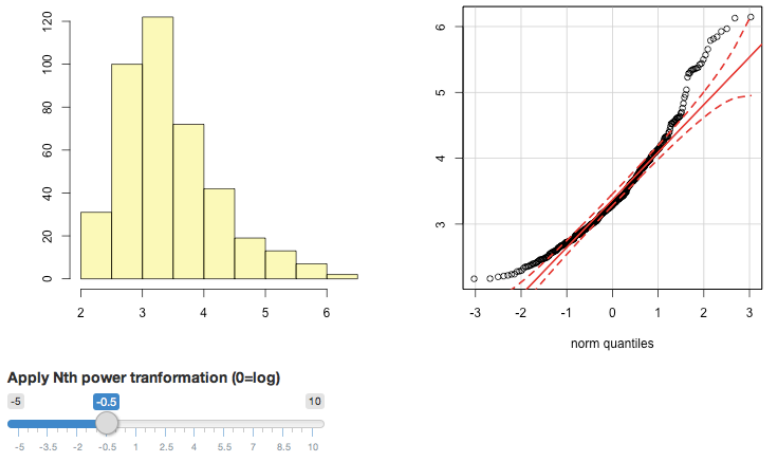
c). 2nd power (square)



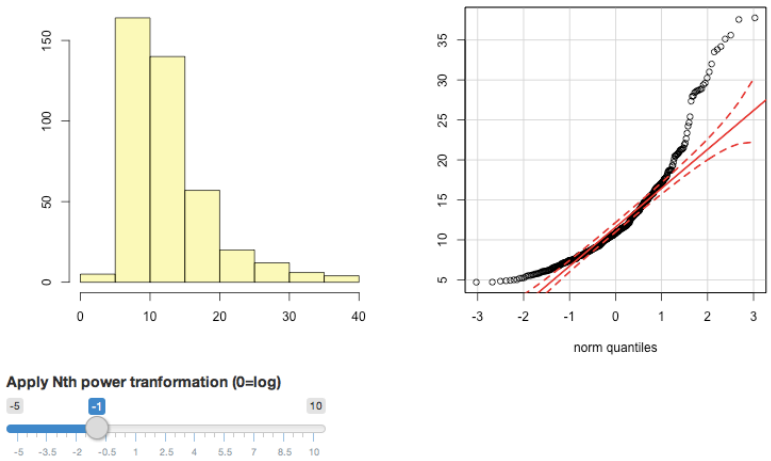
d). 0.5th power (square root)



e). Logarithm



f). -0.5th power



g). -1th power (inverse)

Figure 5-9. Multiple power transformations applied to Variable 24 (areaR_tri_trap_upper)

As shown in Figure 5-9, the following transformations were applied to one of the 41 variables (in this case Variable 24, areaR_tri_trap_upper): 1st power (no transformation), 1.5th power, 2nd power (square), 0.5th power (square root), logarithm ($\lambda = 0$), power of -0.5, and power of -1 (inverse). It can be observed that after the 0.5th power (square root) transformation has been applied, distribution of the variable (Figure 5-9d) approximates a normal distribution the most.

Lastly, although this study has a focused age group (i.e. 18-45) aiming to remove the effect on breast shape caused by age, it is still necessary to investigate on whether different age levels within this range (18-45) have remarkably large distinctions in breast shape measurements. The last function written for the Shiny app allows for plotting of points in color. Three age levels were created, namely Age18-25, Age 25-35, and Age 35-45. Points in a scatter plot were colored red for those whose age is between18 to 25. Points were colored green or blue for subjects who belong to Age 25-35 or Age 35-45 respectively. When the "Age Groups" tablet is initially activated, no point has color (Figure 5-10).

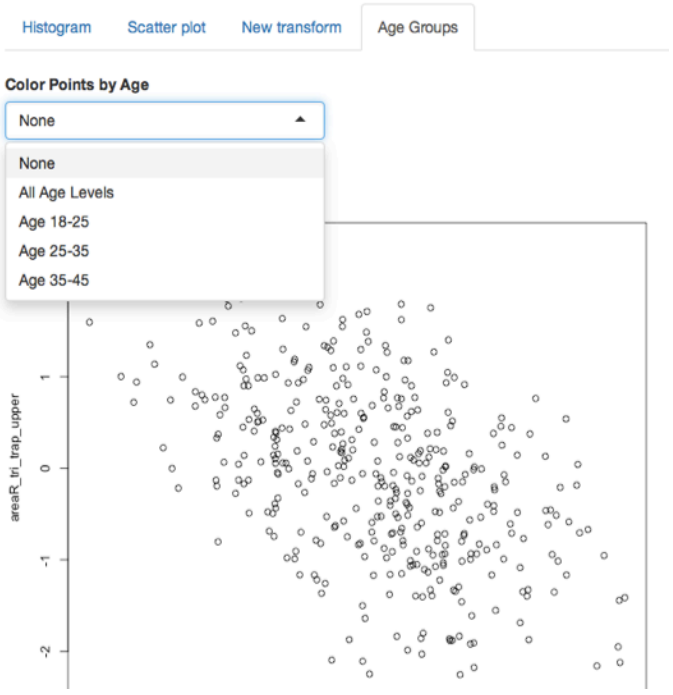
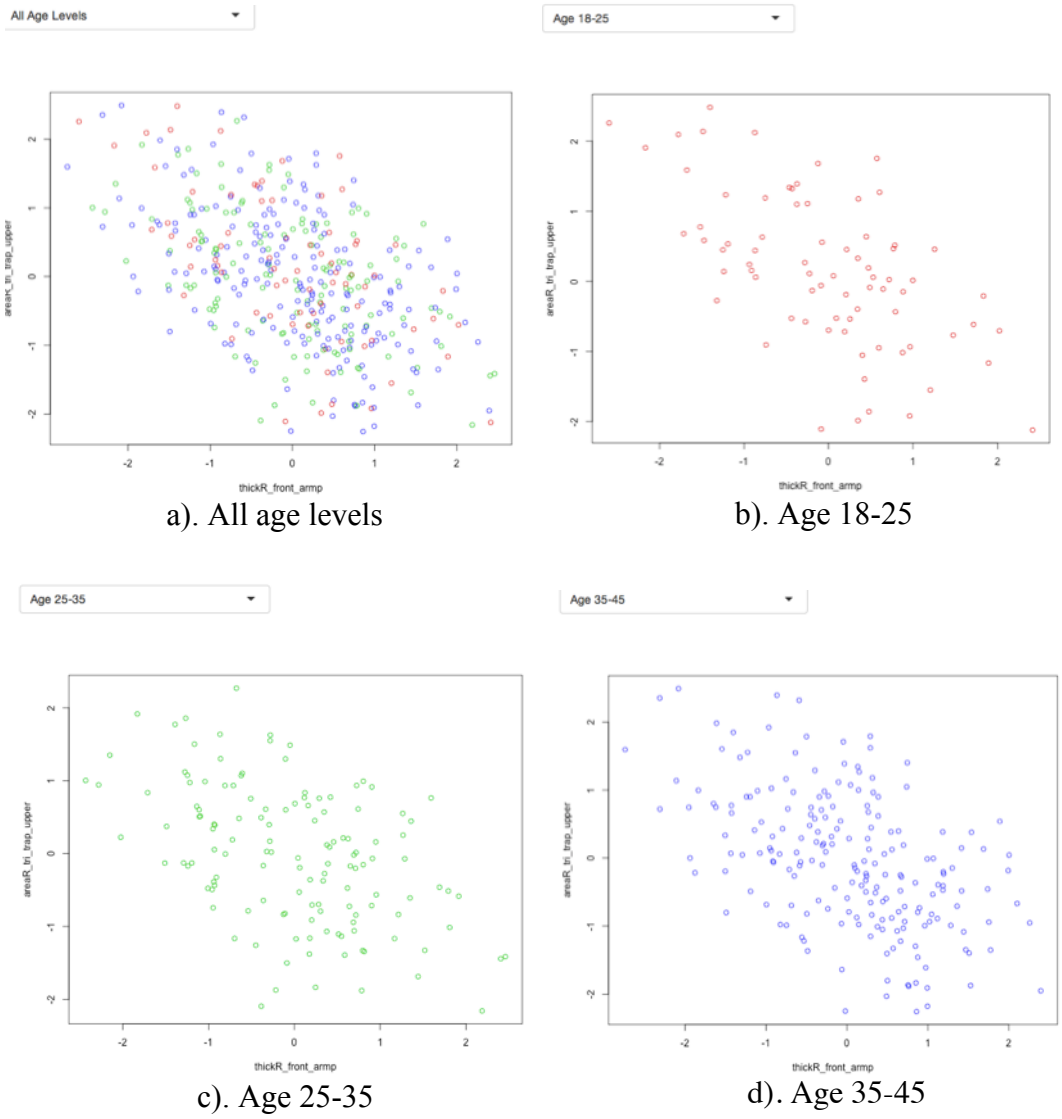


Figure 5-10. Interface under the "Age Groups" tablet

Users may select a specific age level from a dropdown menu on top of the scatter plot (Figure 5-10). The dropdown menu also includes the option of "None" which will remove all colors, and the option of "All Age Levels" which will plot all three age levels together with colors (Figure 5-11). It can be observed that the three colors mix well with each other rather than gather in separate clusters (Figure 5-11a). In addition, points in different age levels follow similar trends (Figure 5-11 b, c and d). Therefore, it is safe to assume that there is no remarkably large distinction among the three age levels.



a). All age levels

b). Age 18-25

c). Age 25-35

d). Age 35-45

Figure 5-11. Selection of age level via dropdown menu

5.3. Outlier Detection and Data Re-expression Outcomes

All of the 1681 scatter plots had been carefully examined for several times for outliers. If one point lay extremely far from the others, its ID number would be recorded. If the same ID stood out for more than five times in different plots, the corresponding subject would be regarded as an extreme case and thus excluded from the data. The scans of the extreme cases had been examined before removal. Most of those scans have bad scanning posture, with the subject: a) leaning too much to the front; b) leaning too much to the back; or c) having twisted torso. A handful of those scans appear to be very different from the majority of the scans by: a) having unusually flat breasts; or b) having unusually high or low bust points. In the end, 70 subjects (14.6% of the 478 subjects) were removed from the data, and sample size reduced to 408.

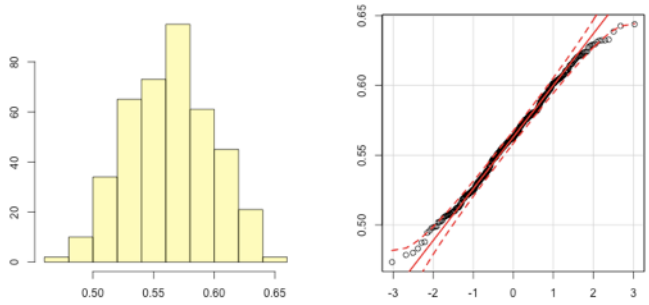
Table 5-1 listed all the transformation methods used on the data. Some variables do not need to be transformed (they are already approximately normally distributed), thus are not listed in the table. Figure 5-12 shows two examples of the untransformed variables (Variable 5, thickR_front_bust and Variable 29, areaR_underb_bust).

Table 5-1.

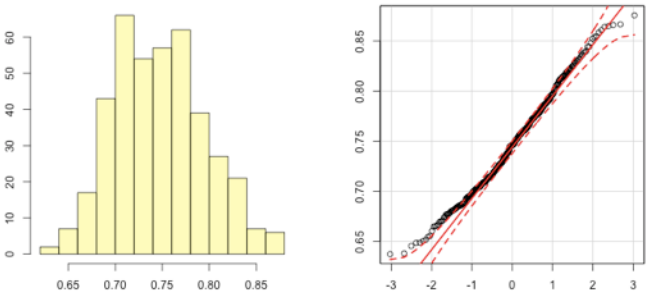
Transformation Methods Used for Variables That Need to Be Re-Expressed

Variable serial number	**Name of the variable**	**Transformation (re-expression) method used**
2	thickR_armp_bust	3rd power
7	thickR_front_underb2Broot_Broot	Logarithm
8	heightR_upperB_fullB	2nd power (square)
9	heightR_underb2Broot_lowerB	0.5th power (square root)
11	angle_lowerB	Inverse
12	angle_lower_diff	0.5th power (square root)
18	distR_upperB_lowerB	0.5th power (square root)
19	distR_upperB_lowerBroot	0.5th power (square root)

20	areaR_curv_rec_upper	log(log(1/y)): inverse was firstly applied, then logarithm was applied twice
22	areaR_curvUp_curvLow	0.5th power (square root)
24	areaR_tri_trap_upper	0.5th power (square root)
25	areaR_tri_trap_lower	0.5th power (square root)
26	areaR_fanUp_fanLow	0.5th power (square root)
34	depth_width_ratio	2nd power (square)
36	areaR_fan_rec_inner	3rd power
37	areaR_innerfan_rt_lt	Logarithm
39	areaR_outerfan_rt_lt	0.5th power (square root)
40	areaR_fullarc_rec	-0.5th power
41	areaR_fullarc_rt_lt	0.5th power (square root)



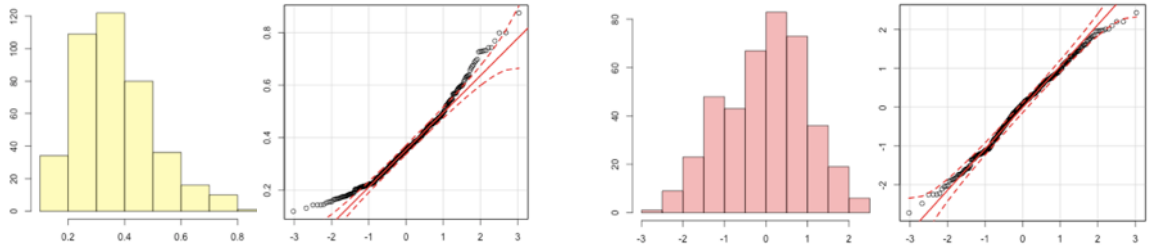
a). Variable 5 - thickR_front_bust



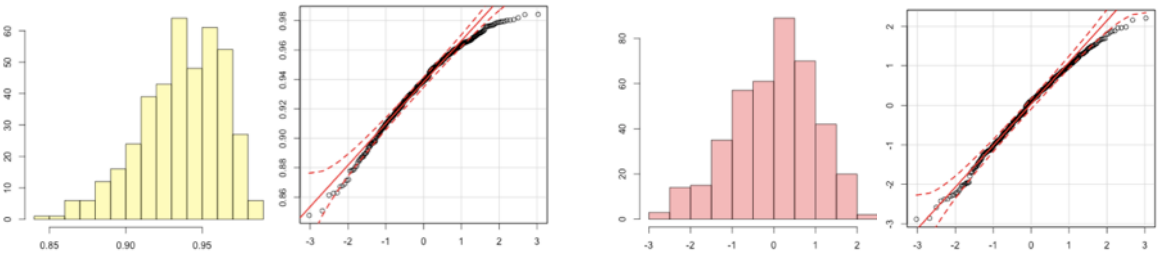
b). Variable 29 - areaR_underb_bust

Figure 5-12. Variables that do not need to be transformed (two examples)

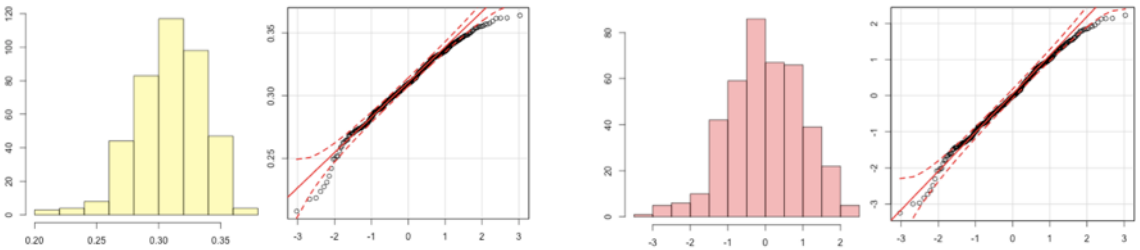
Figure 5-13 shows the comparison between transformed distribution and untransformed distribution. Three examples were presented in the figure, namely Variable 7 (Logarithm transformation), Variable 20 (log(log(1/y))), and Variable 34 (2nd power transformation). It can be observed, from the histogram and Q-Q plot, that although perfect normality was not achieved, normality was improved to some extent after the transformation. Figure 5-14 shows the comparison of scatter plots between transformed and untransformed data. Improvement on linearity can be seen in Figure 5-14a. Improvement on homoscedasticity can be seen in Figure 5-14b.



a). Variable 7 - thickR_front_underb2Broot_Broot

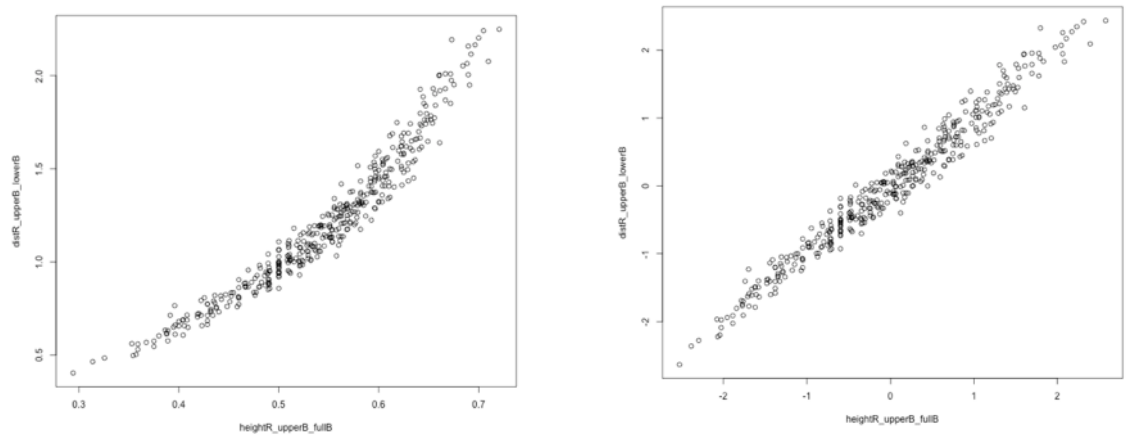


b). Variable 20 - areaR_curv_rec_upper

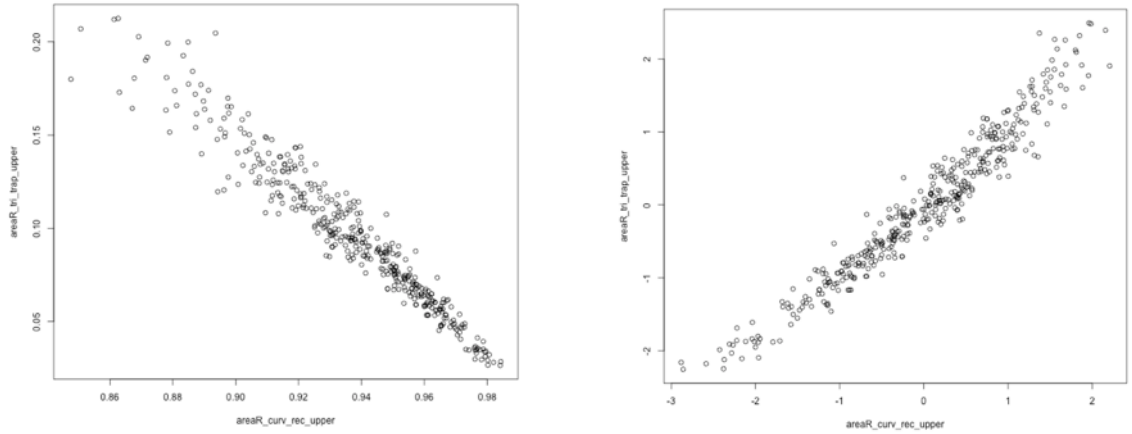


c). Variable 34 - depth_width_ratio

Figure 5-13. Impact of transformation on normality (three examples)



a). Variable 8 (heightR_upperB_fullB) against
Variable 18 (distR_upperB_lowerB)



b). Variable 20 (areaR_curv_rec_upper) against
Variable 24 (areaR_tri_trap_upper)

Figure 5-14. Impact of transformation on linearity and homoscedasticity

CHAPTER 6

STATISTICAL ANALYSIS

6.1. Introduction of R and RStudio

R is a statistical computing language widely used by statisticians and researchers for statistical analysis and data mining. R is strong at data visualization and is capable of plotting both static and dynamic graphics. It supports matrix arithmetic and is good at manipulation of large data. A wide variety of advanced techniques, such as machine learning (Cluster analysis, General tree structures, Neural networks, etc.), signal processing (Convolutions, Filters, etc.), optimization (General purpose optimization, One dimensional optimization, etc.), statistical modeling (Generalized linear models, Non-linear models, Principal component analysis, Factor analysis, etc.) can be achieved. In addition, RStudio® is an integrated development environment (IDE) for R. It provides a user-friendly interface, thus widely used by R users.

All data mining techniques, statistical modeling and testing methods adopted in this study was done in RStudio, coded in R language.

6.2. Principal Component Analysis (PCA)

6.2.1. Correlation and Covariance

The 41 variables in this study are likely to be correlated with each other. A wide variety of multivariate methods interest in not only the variances of variables, but also the covariance's between variables. Their algorithms often involve the calculation of covariance matrix or correlation matrix. Covariance of two variables (vectors), **x** and **y** can be calculated via Eq. 6-1. Correlation can be considered as a standardized version of the covariance, and the conversion between the two can be carried out according to Eq. 6-2.

$$cov(\mathbf{x}, \mathbf{y}) = E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y} - E[\mathbf{y}])] \tag{6-1}$$

where E[**x**] and E[**y**] are the expected values (often treated as the mean values) of **x** and **y** respectively.

$$cor(\mathbf{x}, \mathbf{y}) = \frac{cov(\mathbf{x}, \mathbf{y})}{sd(\mathbf{x})\, sd(\mathbf{y})} \tag{6-2}$$

where sd(**x**) and sd(**y**) are the standard deviations of **x**, and **y** respectively. cov(**x**, **y**) is the covariance of **x** and **y**.

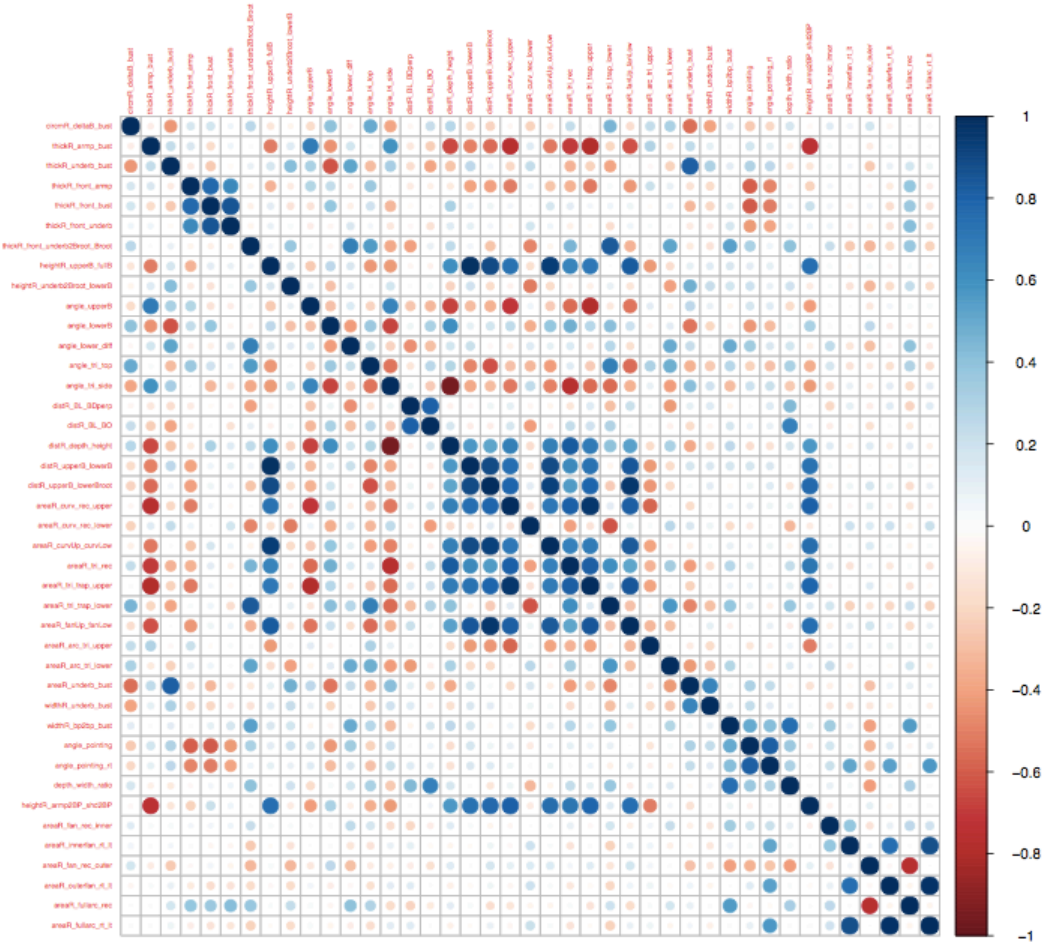


Figure 6-1. Graphical display of variable correlation matrix

The correlation matrix of variables (41 × 41) was examined (Figure 6-1). Viewing a data matrix of this size filled with numbers is neither intuitive nor efficient. However, R

provides a function for graphical display of the correlation matrix (Figure 6-1). According to the heat map on right side of the figure, dark blue color suggests a strong positive correlation, and dark red color suggests a strong negative correlation. The lighter the color is, the weaker the correlation is. The size of the circles relates to the covariance values. The larger the circle is, the greater the covariance is. It can be seen that the main diagonal of the matrix contains deep dark blue circles. This is because the correlation of a variable against itself is always equal to 1.

6.2.2. Results of Principal Component Analysis

Principal component analysis (PCA) seeks for a new set of mutually orthogonal variables, called Principal components (PC's), transformed from the original interrelated variables, where each PC is a linear combination of the original variables with varying coefficients, or loadings (Jolliffe, 2002). The 41 original variables are essentially 41 vectors pointing at non-orthogonal directions (The non-orthogonality is caused by correlations). According to Figure 6-1, there is no collinearity among variables (no correlation equals to 1 or -1 can be observed except for the diagonal). Collinearity in variables means that the corresponding vectors are pointing at exactly the same direction (correlation equals to 1), or at exactly the opposite directions (correlation equals to -1). As there is no collinearity, the original vectors are pointing at 41 different directions. On the other hand, the transformed versions of the original variables, i.e. the PC's, are pointing at 41 mutually orthogonal directions. While the correlation structure is retained via the PC loading matrix, the PC's themselves are rid of the correlations (and covariance's), allowing the analysis to concentrate on variances. The main idea of PCA is to reduce the dimensionality of data while preserving as much variation in the data as possible. PC's with extremely small variation can be considered for exclusion. When PC's are being generated, they are sorted in a decreasing order by the variation that each of them retains. In other words, the first PC direction has the largest variation while the

last PC direction has the smallest. Usually the first few PC's alone preserve the majority of the variation for all variables.

Table 6-1.

Principal Component Analysis Summary Table (Partial)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	3.1511	2.5573	2.2338	1.9150	1.7784	1.6626
Proportion of Variance	0.2422	0.1595	0.1217	0.0894	0.0771	0.0674
Cumulative Proportion	0.2422	0.4017	**0.5234**	0.6129	**0.6900**	0.7574
	PC7	**PC8**	**PC9**	**PC10**		
Standard deviation	1.2715	1.1352	1.0566	1.0250		
Proportion of Variance	0.0394	0.0314	0.0272	0.0256		
Cumulative Proportion	0.7968	**0.8283**	0.8555	**0.8811**		

Table 6-1 is the partial PCA summary table, which contains some summary statistics of the first 10 PC's. The standard deviations are the square roots of the variances. Proportion of variance is calculated via Eq. 6-3 and sums up to 1 for all 41 PC's. Cumulative proportion is the sum of current and all preceding proportions of variances, and the last entry of cumulative proportion (the 41st PC) is equal to 1. The complete PCA summary table can be found in Appendix C. According to Table 6-1, the first 3 PC's are able to explain over 50% of total variance; the first 5 PC's can explain approximately 70% of total variance; the first 8 PC's can explain over 80% of variance and the first 10 can explain around 90%.

$$\text{Proportion of Variance} = \frac{\text{Variance included in a specific PC}}{\text{Sums of variances for all PC's}} \qquad (6\text{-}3)$$

Figure 6-2 is the scree plot presenting the variances of the PC's. There seems to be a natural break point at the 8th PC where 80% of total variance is explained. The red horizontal line in the plot has a y-coordinate of 1, which equals to the average variance of the individual variables (only for the case when PCA works with correlations, i.e. standardized data, instead of working with covariance's). The first 10 PC's are above the line of average variance.

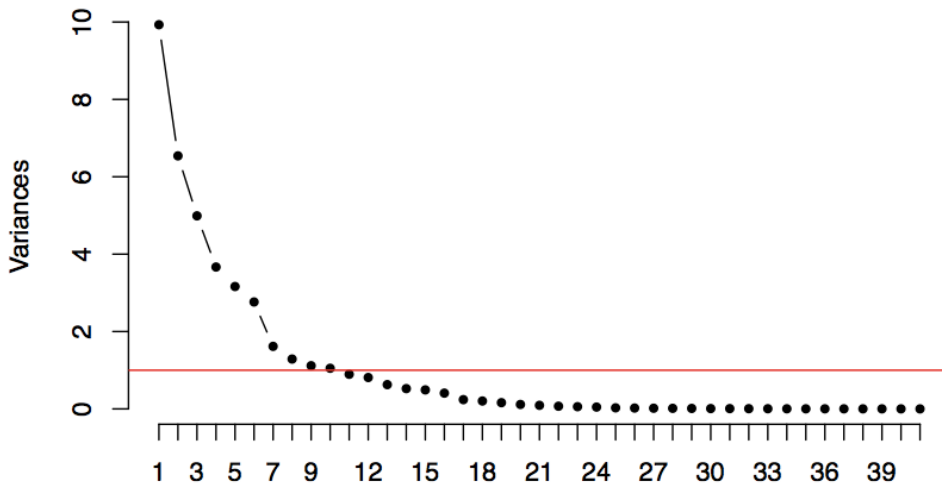


Figure 6-2. Scree plot of PC variances

There is no established rule in choosing the number of PC's for dimension reduction. There are, however, some suggested instructions. Generally, at least 80% of total variance is required for the PC's to retain, suggesting the first 8 PC's are required to be preserved for this case. The scree plot shows similar conclusion: the natural break point seems to appear at the 8th PC. Nonetheless, some statisticians suggest that all PC's with variances higher than the average variance are necessary to be included. Based on this standard, the first 10 PC's need to be preserved. Conclusively, 10 PC's (90% of total variance) were kept instead of 8 to be conservative. Further dimensionality reduction will be discussed in later section.

The complete table consists of PC loadings can be found in Appendix D.

6.2.3. PCA Applied to Unstandardized Data

As shown in Eq. 6-1 and Eq. 6-2, correlation is the scaled (standardized) version of covariance. The finalized data for this study has been transformed and standardized. Therefore, the PCA results discussed in the previous section utilized the correlations between variables. In fact, whether to standardize the data or not does make a significant difference in PCA. PC's depend highly on the scale of the data. Multiplying a variable by a scalar value can alter the PC's completely.

Table 6-2.

PCA Summary Table for Unstandardized Data (Partial)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	0.4367	0.3720	0.3015	0.1886	0.1317	0.1167
Proportion of Variance	0.3564	0.2586	0.1699	0.0665	0.0324	0.0254
Cumulative Proportion	0.3564	0.6150	0.7849	0.8513	0.8837	0.9092
	PC7	**PC8**	**PC9**	**PC10**		
Standard deviation	0.0920	0.0886	0.0837	0.0920		
Proportion of Variance	0.0158	0.0147	0.0131	0.0158		
Cumulative Proportion	0.9250	0.9397	0.9528	0.9250		

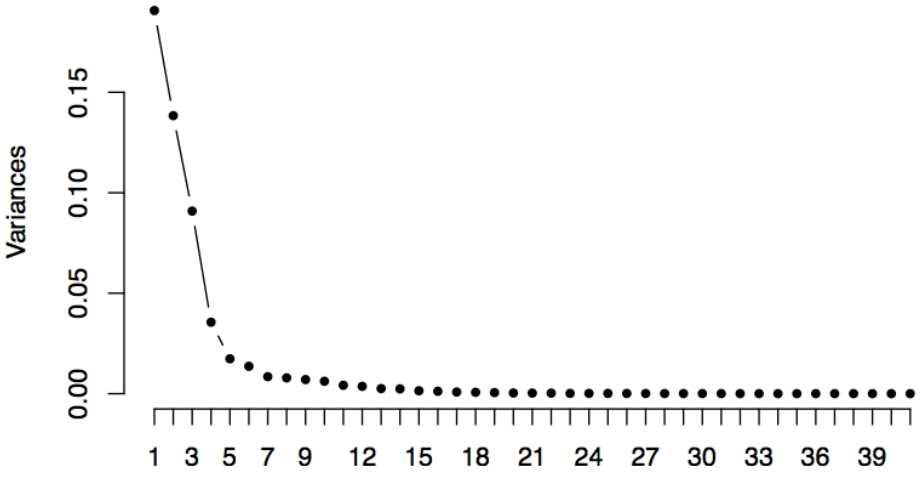


Figure 6-3. Scree plot for PCA applied on unstandardized data

In order to fully understand the effect of standardization on this specific dataset, PCA was also applied to the unstandardized data. Table 6-2 shows the summary table of the variance-related statistics for the first 10 PC's. Figure 6-3 is the corresponding scree plot. Clearly, the table is very different from Table 6-1. For the new case, only 3 PC's are needed to explain nearly 80% of total variance while at least 7 PC's are required for the previous analysis. The scree plots are very different as well. The variances drop more quickly for the first few PC's and the natural break point is more obvious compared with Figure 6-2. Although the analysis on the unstandardized data has a plausibly more promising table and nicer-looking scree plot, it does not work well in terms of classification (Figure 6-4a). Figure 6-4a and b adopted the same classification method (K-means applied to first 10 PCs), and the only difference was whether standardization was performed before calculating PC's. It can be seen that the three breast shapes are not very distinguishable from each other in Figure 6-4a, whereas Figure 6-4b shows outstanding distinctions between breast shapes.

a). PCA on unstandardized data (working with covariance)

b). PCA on standardized data (working with correlation)

Figure 6-4. Classification results calculated from unstandardized and standardized data (K-means applied to first 10 PC's)

Researchers in the past often tried to use PCA summary table and scree plot to imply the effectiveness of their analysis, but this example shows that providing the table and scree plot alone may not be sufficient enough in showing how good the result is.

6.2.4. PCA Applied to Untransformed Data

In order to understand the impact of data re-expression for this specific dataset, PCA was applied to the untransformed data (essentially standardized ratios and angles). Table 6-3 is the corresponding PCA summary table and Figure 6-5 is the scree plot. Surprisingly, the table is very similar to Table 6-1 and the scree plot is almost identical to Figure 6-2.

Table 6-3.

PCA Summary Table for Untransformed Data (Partial)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	3.1609	2.5938	2.2226	1.8905	1.7720	1.6781
Proportion of Variance	0.2437	0.1641	0.1205	0.0872	0.0766	0.0687
Cumulative Proportion	0.2437	0.4078	0.5283	0.6154	0.6920	0.7607
	PC7	**PC8**	**PC9**	**PC10**		
Standard deviation	1.2359	1.1266	1.0397	0.9861		
Proportion of Variance	0.0373	0.0310	0.0264	0.0237		
Cumulative Proportion	0.7980	0.8289	0.8553	0.8790		

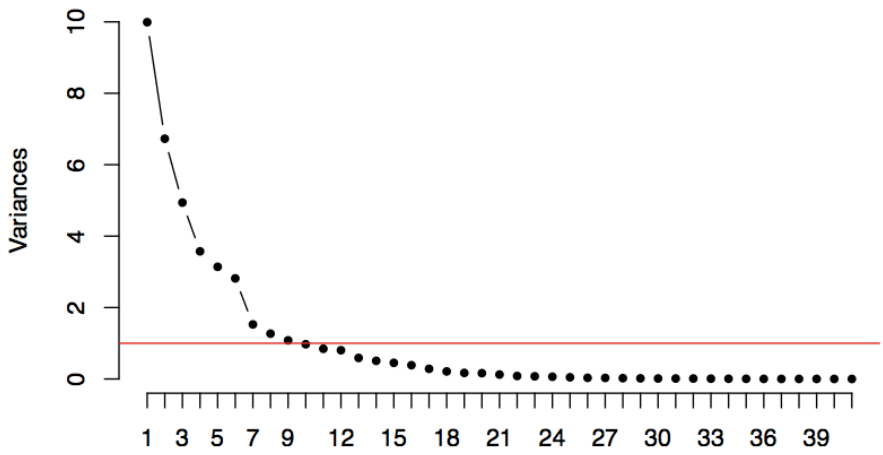


Figure 6-5. Scree plot for PCA applied on untransformed data

Figure 6-6a and b adopted the same classification method (K-means applied to first 10 PCs), and the only difference was whether data re-expression (transformation) was performed before calculating PC's. Clearly, the two outcomes look very similar. However, the similarity in results does not mean that the re-expression process is unnecessary. Whether the transformation can make a difference and how much difference it can make is determined by the configuration of the data, which can be known of only after the transformation has been performed. In most cases, data transformation will make a difference. Variables in this data, however, are not heavily skewed, and this is probably the reason why data transformation does not have a strong influence here. Furthermore, not all body measurement data is as good as this one, especially for those that include raw measurements (absolute measurements such as circumferences, widths and heights). All body measurements involved in this study are relative measurements (ratios and angles), and thus have less chance of containing extreme values (An extremely large absolute measurement is very likely to be divided by another large absolute measurement, resulting in a ratio much less extreme). In conclusion, despite no significant difference has been observed between transformed and untransformed data of this study, it is still important to re-express data before any sophisticated statistical analysis is done.

a). PCA on untransformed data b). PCA on transformed data

Figure 6-6. Classification results calculated from untransformed and transformed data (K-means applied to first 10 PC's)

6.3. Cluster Analysis

Cluster analysis is a general task, rather than a specific algorithm, aiming at grouping objects in a way such that objects in the same group (cluster) are similar, whereas distinction can be observed between groups (Rencher, 2003a). From statistical standpoint, clusters that have small "within" variation and large "between" variation are desired. Clustering is a common statistic technique, and has been actively involved in data mining, pattern recognition, machine learning, and many other fields. As there is no precise definition on cluster, there are numerous clustering algorithms. Despite the extensive use of clustering, there is no optimal clustering method. An optimal method would need to access every possible way of partitioning objects into a certain number of clusters. However, there are so many options that, even for a small sample size and small cluster number, it is not feasible to perform all calculations and assessments (even with modern computational capability). Moreover, due to the diversity in clustering algorithms, different methods can sometimes end up with very different cluster results when applied to the same data.

In fact, Cluster analysis has been used in classifying body shapes in the past. However, few of previous researchers have mentioned the extensive choices of clustering methods. Nor did they provide any reasoning on why they preferred certain clustering method than the others. In addition, few of them was able to bring up a strong validation on the clustering result based on the method of their choice, in showing how distinctive the clusters were. To fix those issues and improve previous methodology, and to find the most promising clustering method that works for this specific data, three most commonly used clustering algorithms were chosen for the analysis, namely Hierarchical clustering, K-means clustering, and K-medoids clustering. The "acquiring median curves" algorithm developed for this study (see Section 4.4) makes it possible for the visual presentation of the clustering results, which enables the comparison between the three clustering methods.

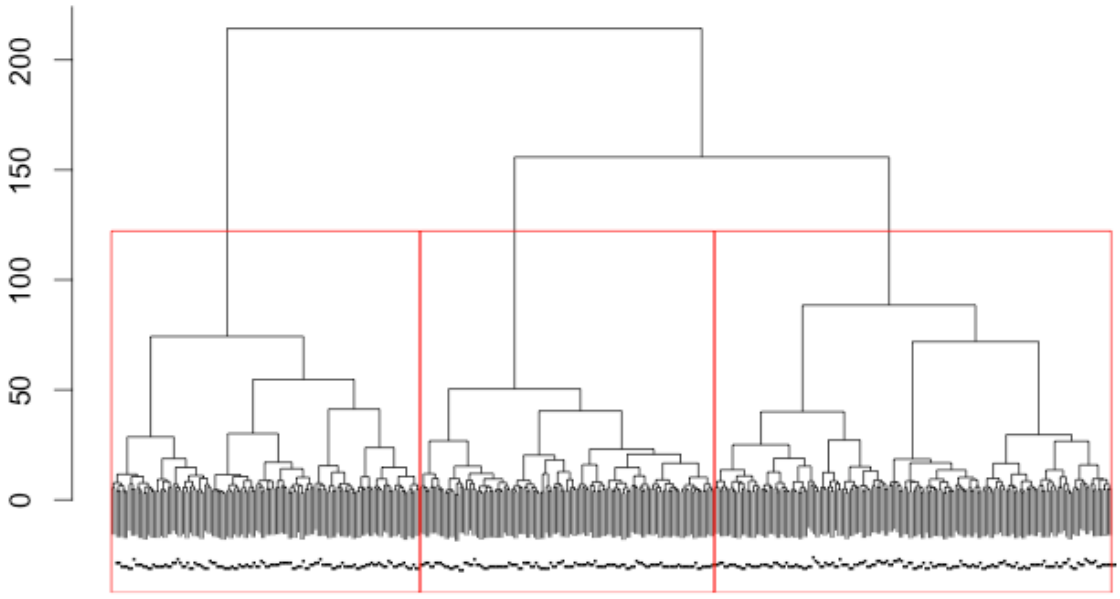
Furthermore, a measure of similarity or dissimilarity needs to be pre-determined as a guideline to differentiate clusters, and different ways of choosing the measure is one of the reasons that cause the divergence in clustering algorithms. Two commonly used measures of similarity are Euclidean distance and Manhattan distance (City block distance). Euclidean distance is a direct measurement on any vector, and Manhattan distance expresses the distance between points as measured on an x and y axis (and on other Cartesian axes as well, if there are any, depending on data dimension). Nevertheless, Euclidean distance is more appropriate for this study, and also for not further complicating the analysis, all three clustering methods chosen for the study utilized Euclidean distance as the measure of similarity. Euclidean distance is a common distance function and can be calculated from the generalized Minkowski distance formula (Eq. 6-4).

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{\mathrm{p}} \left|\mathbf{x_i} - \mathbf{y_i}\right|^{\mathrm{r}}\right]^{1/\mathrm{r}} \tag{6-4}$$

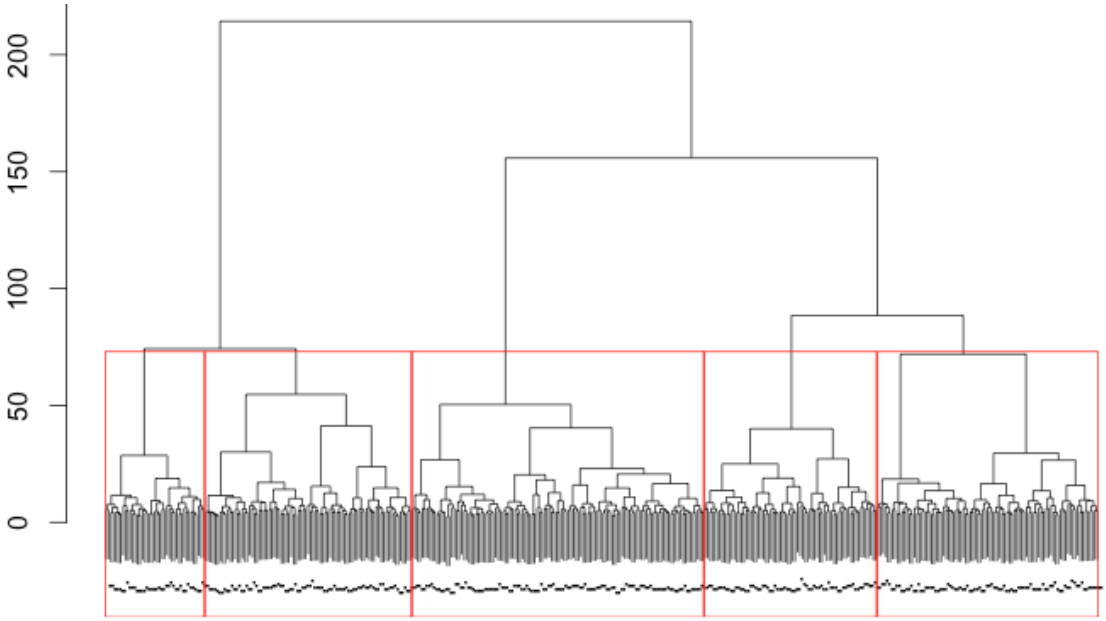
where $\mathbf{x}= (x_1, x_2, \ldots, x_p)^T$ and $\mathbf{y}= (y_1, y_2, \ldots, y_p)^T$ are two vectors; when r= 2, it is the Euclidean distance; when r= 1 and p= 2, it is the Manhattan distance.

6.3.1. Hierarchical Clustering

Hierarchical clustering builds a hierarchy of clusters sequentially. Generally, there are two major strategies for hierarchical clustering: a) divisive clustering builds a hierarchy in a "top down" fashion (all objects start in one cluster and get partitioned progressively along the hierarchy); b) agglomerative clustering, on the contrary, adopts a "bottom up" fashion in building clusters (every individual object is regarded as one cluster to start with, and the method seeks to combine pairs of clusters as it moves up the hierarchy) (Rencher, 2003a). This study adopted the agglomerative hierarchical clustering method and, as mentioned earlier, refers to the Euclidean distances when merging cluster pairs (two clusters with the smallest Euclidean distance get merged each time).



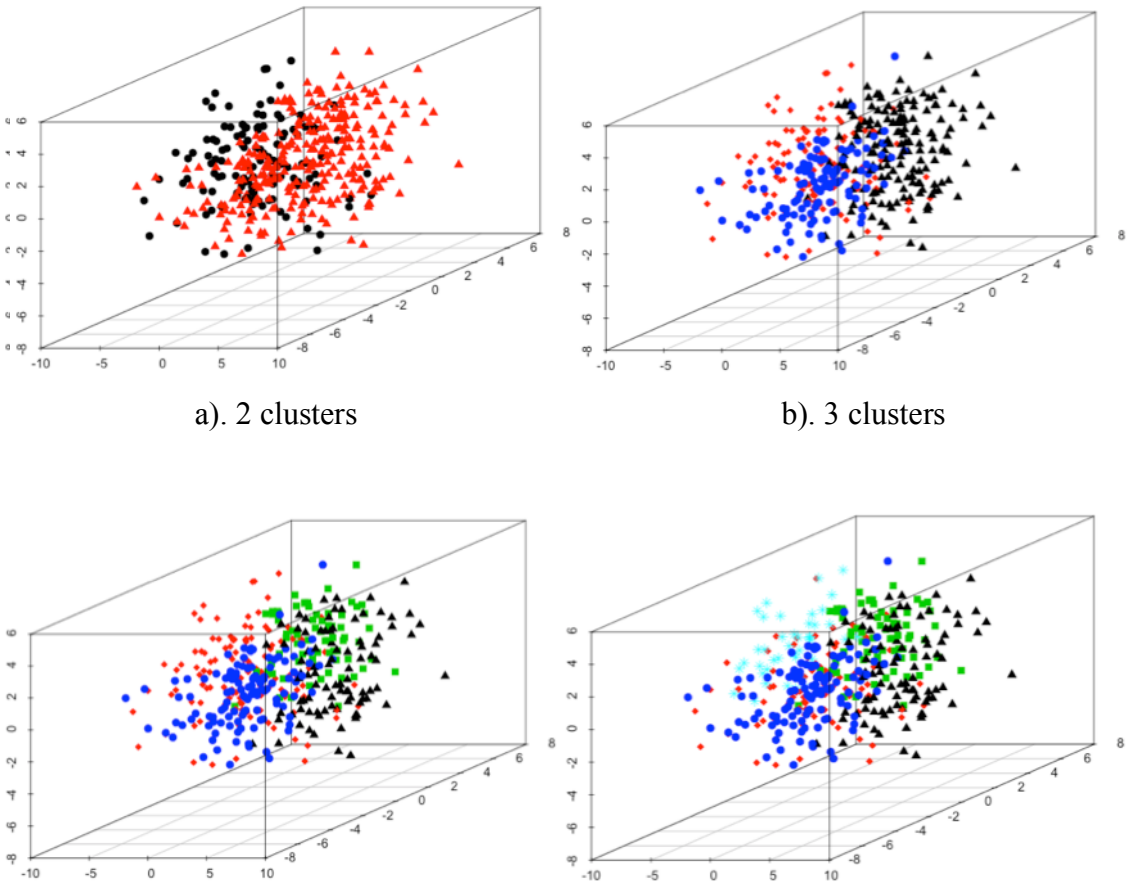
a). 3 clusters being selected



b). 5 clusters being selected

Figure 6-7. Hierarchical clustering dendrogram

A dendrogram is a good way to display the progress and result of the algorithm. Figure 6-7 shows the dendrogram of the hierarchical clustering applied to the data of this study (note that 6-7a and b have the same dendrogram, only with different number of clusters being selected). The X-axis contains all of the 408 observations (therefore, initially there were 408 clusters). The Y-axis gives the distance at which cluster pairs were merged. The number of clusters was not automatically given or suggested by the statistical package, thus requires judgments from the data analyst. Selection of cluster number will be discussed in later section. For now, Figure 6-7 presents two examples where cluster numbers were chosen to be 3 and 5 respectively.



a). 2 clusters

b). 3 clusters

c). 4 clusters

d). 5 clusters

Figure 6-8. Clusters plotted in 3-D space (Hierarchical clustering)

The cluster assignments based on hierarchical clustering were recorded for a series of cluster numbers (from 2 to 9). Figure 6-8 shows the outcomes for the cases of 2, 3, 4 and 5 clusters. Observation points were plotted in 3-dimensional space, and colored by clusters. The X-axis of the 3D space was chosen to be the direction of the first PC. The Y-axis is the direction of the second PC, and the Z-axis is that of the third PC.

Figure 6-9 shows the results for all 8 cases of cluster number, displayed in a more intuitive way where breast shape for each cluster has been visualized.

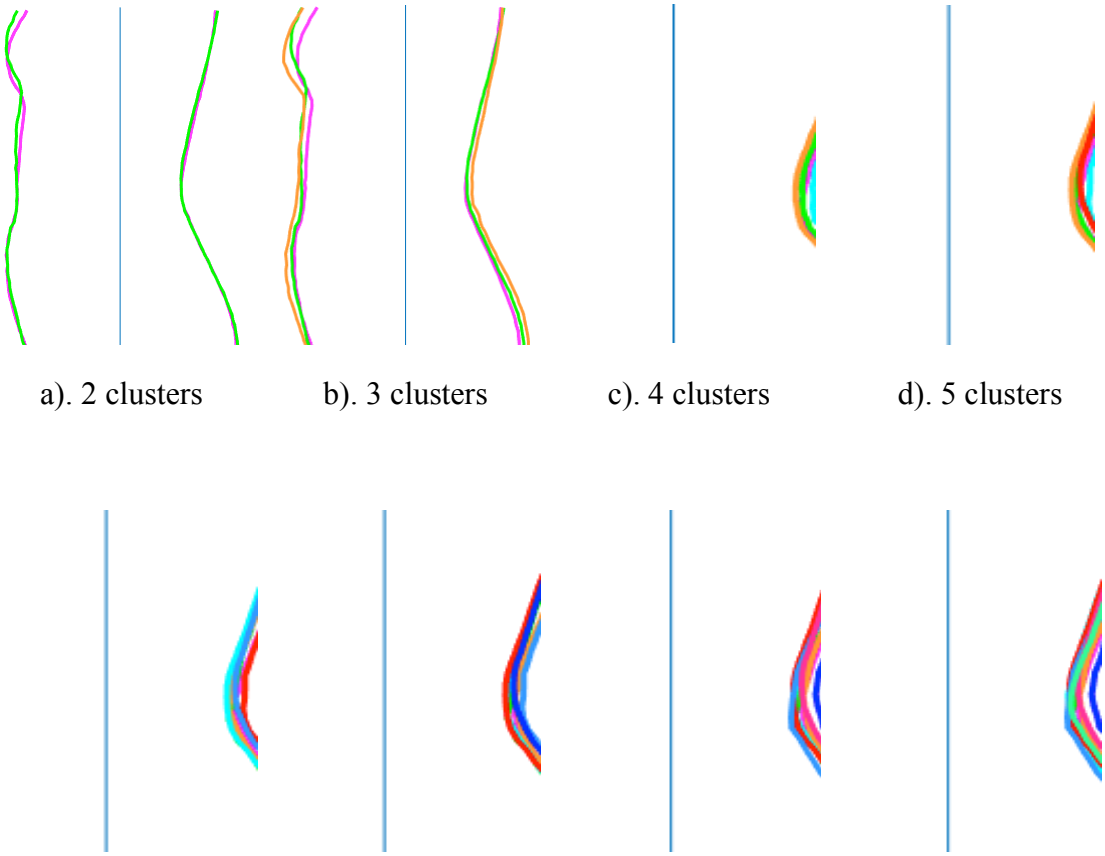


Figure 6-9. Breast shape clustering results (Hierarchical clustering)

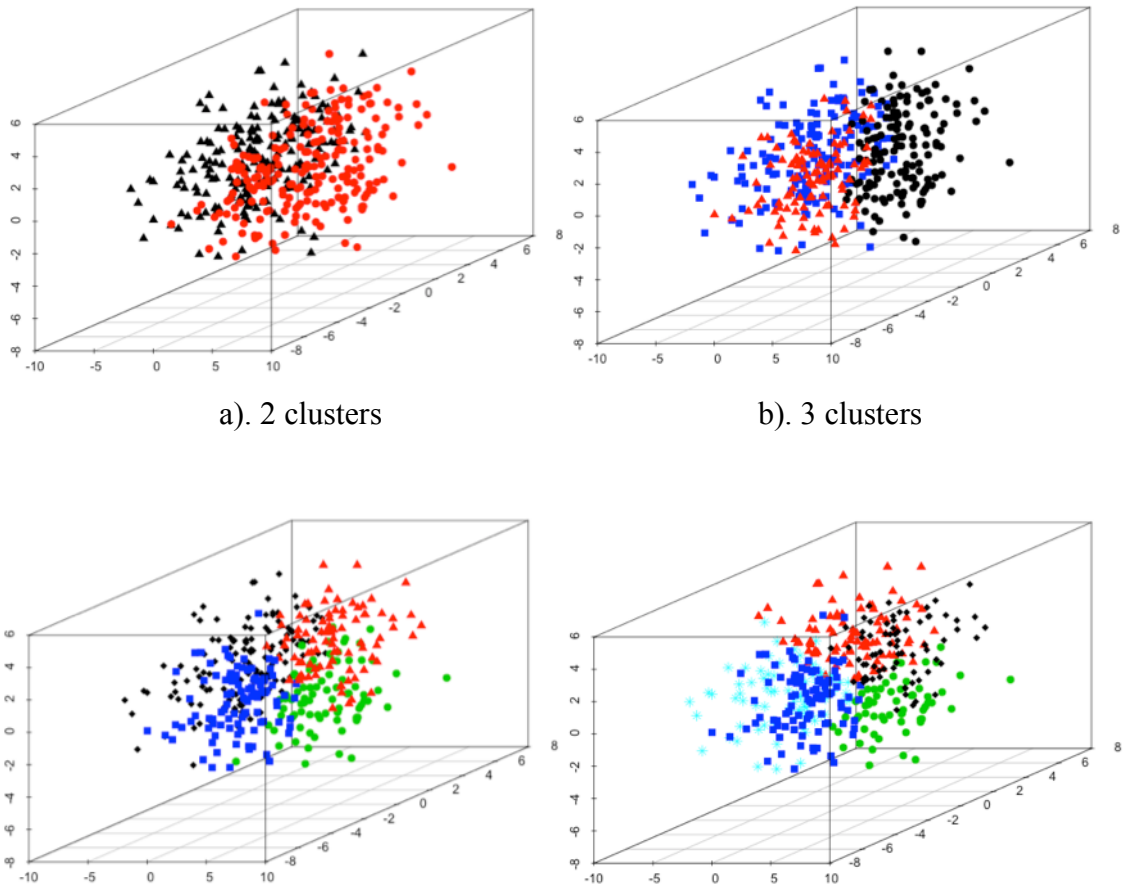
The hierarchical clustering algorithm has a drawback, which cannot be fixed easily. Once the clusters have been merged or partitioned, the algorithm does not re-consider other possible assignment of observations to clusters, even though a different assignment can lead to a better clustering result. In spite of this, the algorithm gave an acceptable clustering outcome for this study. Distinctive shapes at bust area can be observed (Figure 6-9).

6.3.2. K-means Clustering

K-means clustering is based on a quite different algorithm. It separates objects into a few clusters (the number of clusters is often referred to as k) so that each object is sorted to the cluster with the nearest centroid (the distance from the object to the mean vector of the cluster is the shortest) (Hartigan & Wong, 1979). The algorithm first selects k observations to serve as seeds (one seed for each of the k clusters). The seeds can be randomly selected, but they are regarded as the initial centroids of clusters. Then each observation is assigned to the cluster whose centroid it is closest to. Right after the assignment, the centroids are re-computed and can shift away from the initial seeds. After the new centroids have been calculated, observations get re-assigned into the k clusters. The process of re-computing centroids and re-assigning observations will keep iterating until convergence is reached, i.e. until the cluster assignments no longer change. Again, the distance between an observation and a centroid is measured by Euclidian distance.

One issue of the K-means algorithm is that different starting seeds can result in very different cluster results. This issue can be fixed by running K-means multiple times with various sets of seeds. For this study, 25 sets of random seeds were used and the K-means function returned the best (in terms of the minimum overall observation-to-centroid variance) clustering result among the 25. Moreover, during the process of exploratory study, it was found that K-means applied to the first 10 PC's functions equally well as K-means applied to the whole data set. Therefore, everything presented in this section is the outcome of K-means applied to the first 10 PC's.

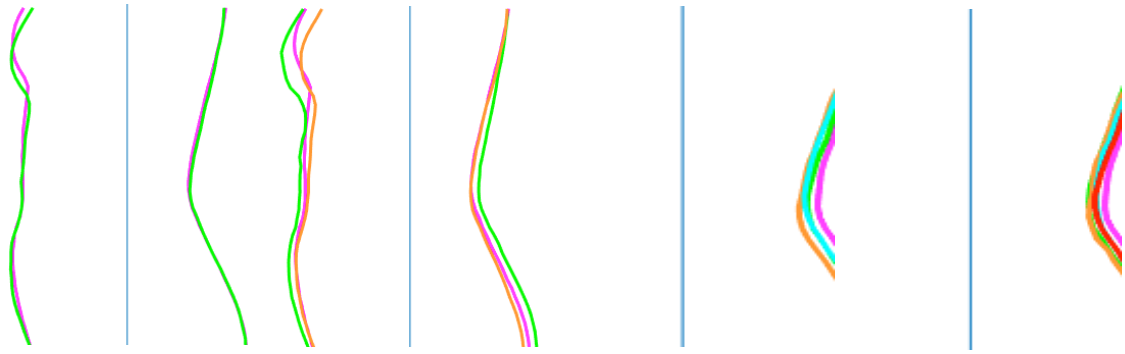
Figure 6-10 shows the clustering results obtained from K-means analysis. Observations were colored by the K-means cluster assignments, and plotted in the 3-dimensional space constructed by the first three PC directions. Figure 6-11 presents the breast shape for each cluster. Distinctive shapes at bust area can be observed.



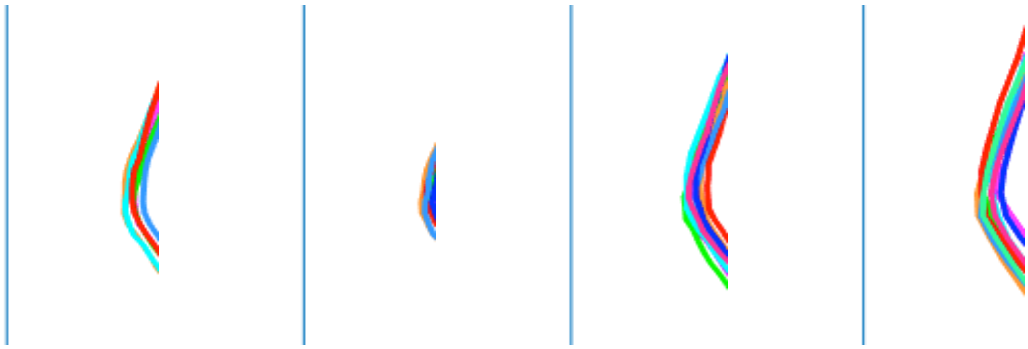
a). 2 clusters b). 3 clusters

c). 4 clusters d). 5 clusters

Figure 6-10. Clusters plotted in 3-D space (K-means clustering)



a). 2 clusters b). 3 clusters c). 4 clusters d). 5 clusters



e). 6 clusters f). 7 clusters g). 8 clusters h). 9 clusters

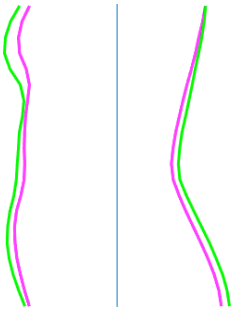
Figure 6-11. Breast shape clustering results (K-means clustering)

6.3.3. K-medoids Clustering

K-medoids is another iterative partitioning algorithm, very similar to K-means. The only difference is that for each step of the iteration, the medoids of clusters (the median case within each cluster) are obtained, rather than the centroids (the mean of each cluster). The medoids are constrained to be the actual observations (Kaufman & Rousseeuw, 2009). This constraint is good in the sense that it can provide medoid cases that actually exist. However, for the fact that the medians are now high dimensional vectors and each contains 41 scalars, there is no guarantee that each of the individual scalars of a medoid case equals (or is similar to) the median value of the corresponding variable (calculated within the cluster). The majority of them can be close, but some of them might be far off. For instance, the third variable is called thickR_underb_bust. The median value of this measurement within a cluster can be very different from the third scalar of the medoid vector of this cluster.

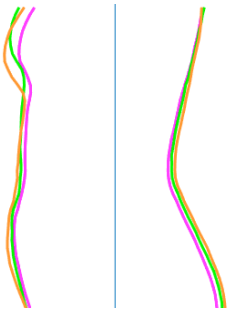
Nevertheless, the medoids provide another way of testing the "acquiring median curves" algorithm that was developed to visualize breast shapes for this study (see Section 4.4). As shown in Figure 6-12, side profiles obtained from the self-developed algorithm (Figure 6-12a, c, e, and g) and side profiles extracted from the medoid cases (Figure 6-12b, d, f, and h) were plotted side-by-side for comparison. It can be seen that

the breast shapes obtained from the algorithm are good representations for the real cases (Note that only the bust area is comparable because all measurements involved in the classification were extracted from bust area). Moreover, the idea of using ratios and angles to concentrate on the study of shape, as well as the scaling process of the algorithm, seems to be effective.



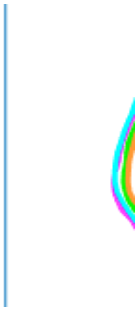
a). Outcomes from algorithm (2-cluster case)

b). Real cases (medoids) (2-cluster case)



c). Outcomes from algorithm (3-cluster case)

d). Real cases (medoids) (3-cluster case)



e). Outcomes from algorithm (4-cluster case)

f). Real cases (medoids) (4-cluster case)

g). Outcomes from algorithm (5-cluster case)

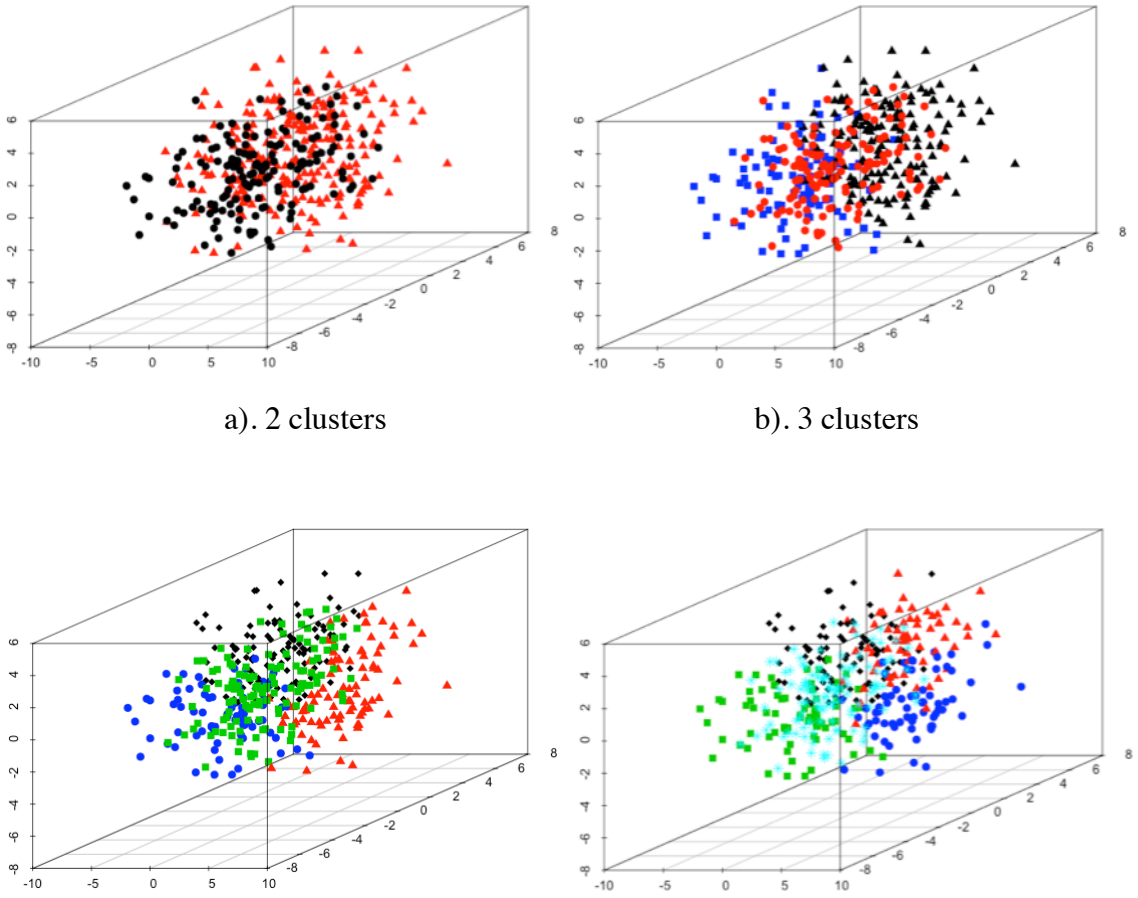
h). Real cases (medoids) (5-cluster case)

Figure 6-12. Comparison between side profiles obtained from algorithm and real cases

In fact, there is more than one way to realize K-medoids clustering. This study adopted the Partitioning Around Medoids (PAM) algorithm, which is one of the most commonly used algorithms for K-medoids. Figure 6-13 shows the clustering outcomes obtained from K-medoids analysis. Observations were colored by the K-medoids cluster assignments, and plotted in the 3-dimensional space constructed by the first three PC directions. Figure 6-14 presents the breast shape for each cluster.

K-medoids is known for its robustness to outliers and skewed distributions, in contrast to the K-means algorithm. However, surprisingly it does not function as well as the K-means does on the data of this study, according to the comparison between Figure 6-14 and Figure 6-11. Distinctions among clusters are less significant for K-medoids than for K-means. This is probably due to the constraint, as mentioned earlier, that forces the algorithm to select real cases as medoids, and the fact that the real cases are not sufficient enough as replacements of the median vectors, with each scalar calculated from every individual variable. In addition, the data of this study has been carefully examined, with outliers removed and skewed distribution transformed, and with all variables standardized. Hence, it is very likely that the mean vectors are very similar to the median

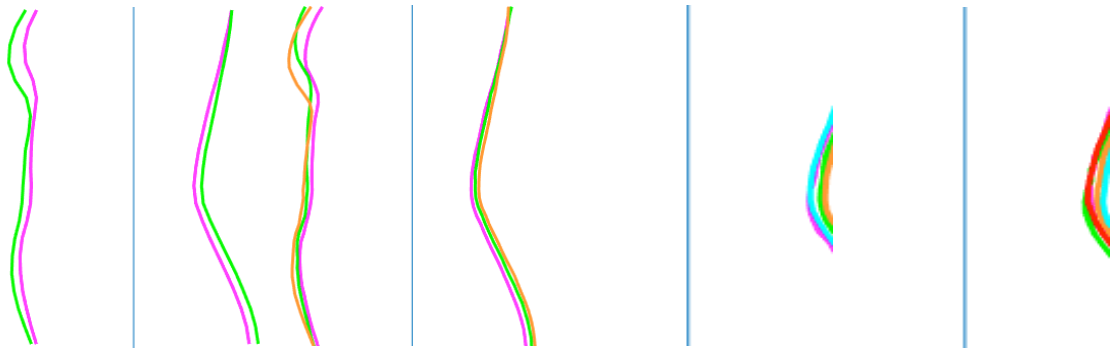
vectors, thus the disadvantages of the K-means algorithm do not apply for this specific data.



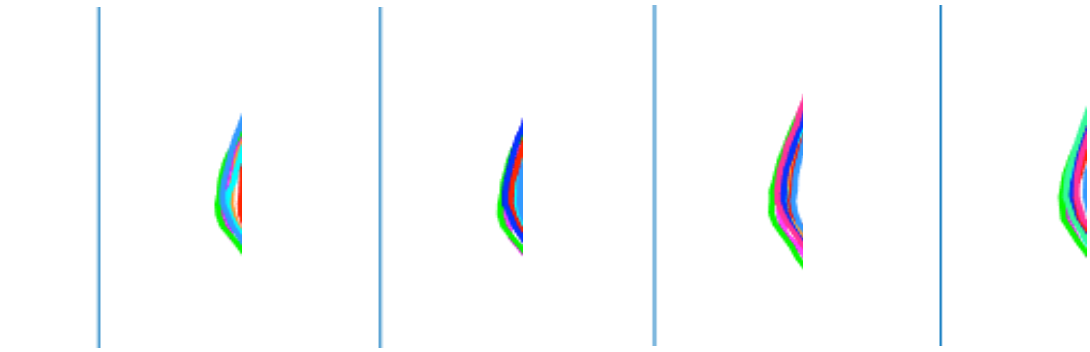
a). 2 clusters b). 3 clusters

c). 4 clusters d). 5 clusters

Figure 6-13. Clusters plotted in 3-D space (K-medoids clustering)



a). 2 clusters b). 3 clusters c). 4 clusters d). 5 clusters

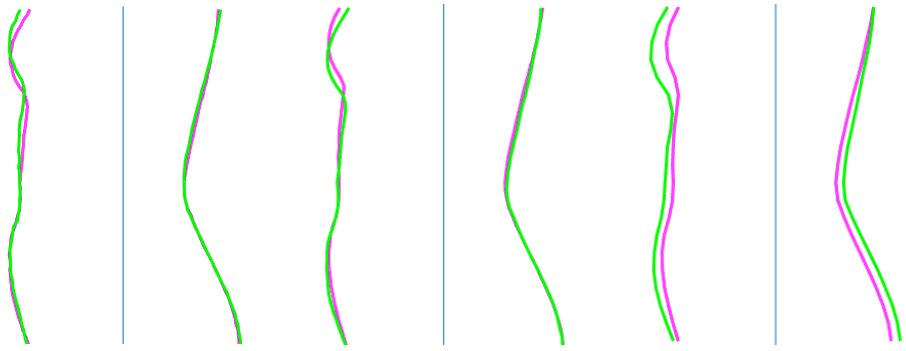


e). 6 clusters f). 7 clusters g). 8 clusters h). 9 clusters

Figure 6-14. Breast shape clustering results (K-medoids clustering)

6.3.4. Comparison of the Three Clustering Methods

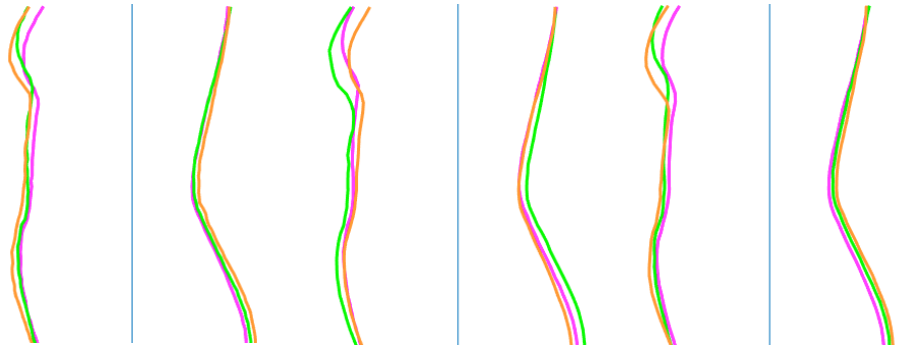
All three methods have advantages and drawbacks, as discussed earlier. The figures displaying the clustering results of breast shapes have been re-arranged in this section so that the three methods, Hierarchical clustering, K-means clustering, and K-medoids clustering, can be easily compared.



a). Hierarchical clustering b). K-means clustering c). K-medoids clustering

Figure 6-15. Comparison between the three clustering methods (2-cluster case)

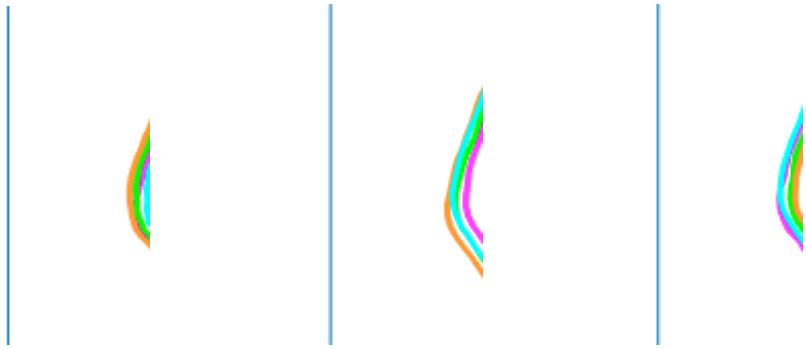
When there are 2 clusters, Hierarchical clustering and K-means clustering present similar outcomes (Figure 6-15a and b), whereas K-medoids has a different result (Figure 6-15c).



a). Hierarchical clustering b). K-means clustering c). K-medoids clustering

Figure 6-16. Comparison between the three clustering methods (3-cluster case)

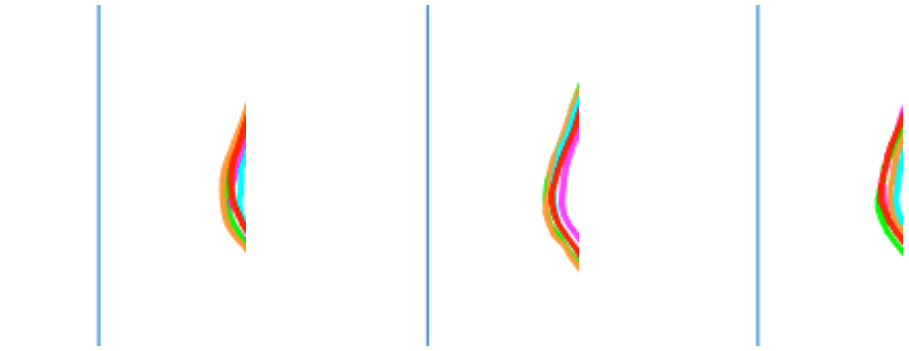
For the 3-cluster case, all three methods present similar results (Figure 6-16), although breast shapes of Hierarchical clustering (Figure 6-16a) are slightly less distinguishable.



a). Hierarchical clustering b). K-means clustering c). K-medoids clustering

Figure 6-17. Comparison between the three clustering methods (4-cluster case)

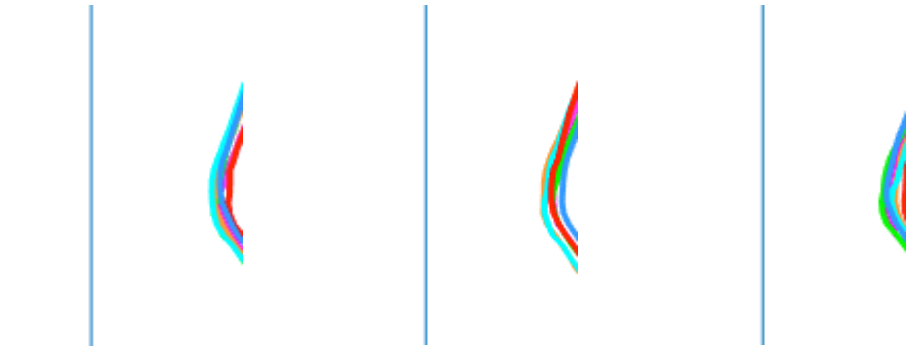
For the 4-cluster case, Hierarchical and K-means (Figure 6-17a and b) have similar outcomes, but again, breast shapes of Hierarchical clustering (Figure 6-17a) are less distinguishable. As for K-medoids in this case, two clusters (colored with green and light blue respectively) appear to be too similar (Figure 6-17c).



a). Hierarchical clustering b). K-means clustering c). K-medoids clustering

Figure 6-18. Comparison between the three clustering methods (5-cluster case)

For the case of 5 clusters (Figure 6-18), the three methods disagree with each other, but in general breast shapes of K-means clustering are the most distinguishable, while two clusters in Figure 6-18a (colored with red and light blue) look too similar, and three clusters (colored with red, pink and orange) in Figure 6-18c seem to be undistinguishable.



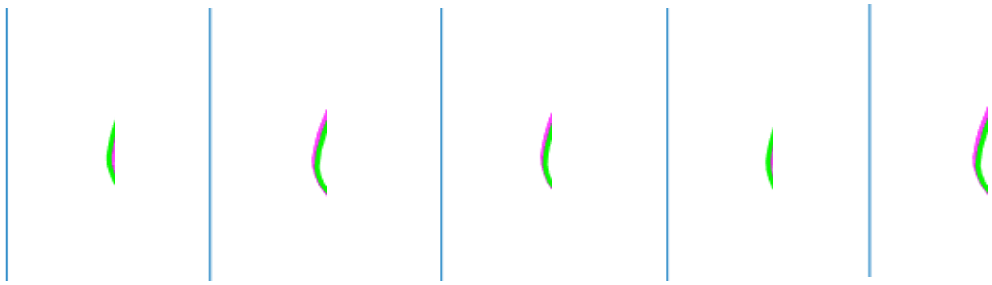
a). Hierarchical clustering b). K-means clustering c). K-medoids clustering

Figure 6-19. Comparison between the three clustering methods (4-cluster case)

For the case of 6 clusters (Figure 6-19) or more, the differences among clusters become so trivial that even for K-means, the shapes are difficult to distinguish.

In conclusion, among the three methods, K-means clustering is the best in giving the most distinctive clusters.

Furthermore, K-means works well in the reduction of dimensionality, and presents good stability and repeatability. Figure 6-20 displays the clustering results when k, the number of clusters, was set to be 2, and K-means was applied to all 41 PC's (i.e. the whole data), the first 10 PC's, the first 8 PC's, the first 5 PC's and the first 3 PC's respectively. Clearly, all plots in Figure 6-20 are very similar, if not identical. For the reason that 3 PC's are sufficient in reaching the similar clustering results as those that were obtained from the whole data, the dimensionality of the data can be reduced to 3. This also explains why observation points can be plotted in 3-D space, colored by cluster membership (as shown in Figure 6-8, 6-10, and 6-13).



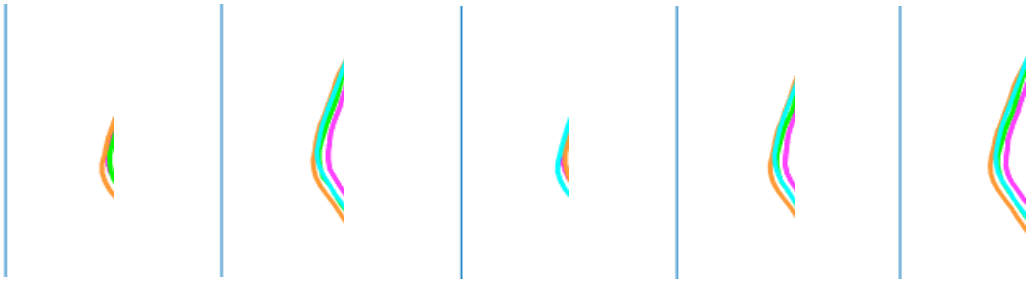
a). 41 PC's (Whole data) b). 10 PC's c). 8 PC's d). 5 PC's e). 3 PC's

Figure 6-20. K-means applied to different number of PC's (k=2)

Same conclusion can be drawn for Figure 6-21, 6-22 and 6-23 as well, where k was set to be 3, 4 and 5 respectively.

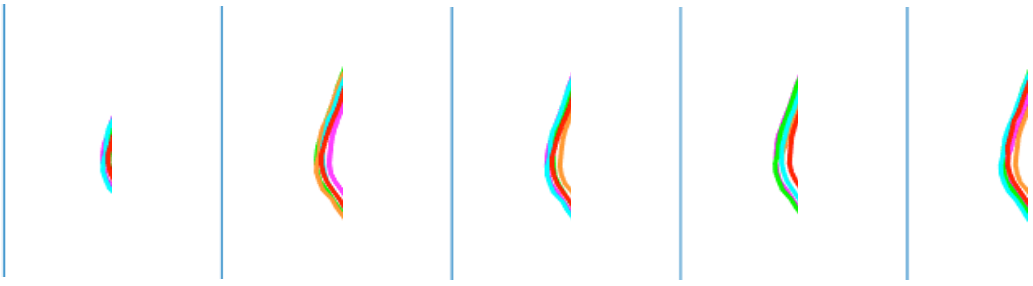
a). 41 PC's (Whole data) b). 10 PC's c). 8 PC's d). 5 PC's e). 3 PC's

Figure 6-21. K-means applied to different number of PC's (k=3)



a). 41 PC's (Whole data) b). 10 PC's c). 8 PC's d). 5 PC's e). 3 PC's

Figure 6-22. K-means applied to different number of PC's (k=4)



a). 41 PC's (Whole data) b). 10 PC's c). 8 PC's d). 5 PC's e). 3 PC's

Figure 6-23. K-means applied to different number of PC's (k=5)

As for the cases where the numbers of clusters are 6 and higher, different curves at bust area are much harder to distinguish visually. Distinctions in breast shapes among clusters are too trivial to be actually useful in practical term.

6.4. Selection of Cluster Number

Determination of the number of clusters is a common task and statistically challenging problem. For K-means and K-medoids clustering algorithms, the selection of k is an unavoidable start-up step for the algorithms to be able to perform. A wrong choice of cluster number can lead to poor clustering results that do not reflect the real homogeneity and heterogeneity in the data. Although Hierarchical clustering does not require a pre-determined cluster number, the same question will come up after the hierarchy of clusters has been built. Practically speaking, researchers are generally more interesting in classifying objects into a definite, usually small, number of groups, rather than knowing every possible way of grouping. In the case of agglomerative hierarchical clustering, every individual observation is regarded as one cluster at the beginning of the algorithm. However, if a dataset has been classified into hundreds of groups, the classification result is of no use. An ideal cluster number needs to be small enough for the result to be practically useful, while large enough for most of the variation in the data to be captured and explained. Despite the importance of selecting a right cluster number, few of the researchers in the past provided a strong justification in their choice of cluster number when classifying body shapes. This study, on the contrary, proposed two different ways for the selection of cluster number, based on two different criteria that serve different purposes.

Before the discussion of the selection criteria, the choices in cluster number were narrowed down from eight to three by referring to the side profiles. Although similar side profiles cannot guarantee that the corresponding 3D shapes of breasts look similar in very view, significantly distinctive side profiles ensure the corresponding shapes to be

different enough. Now that this study aimed at capturing the major differences in breast shape, side profile is the most efficient view in presenting the clustering outcomes. Initially, cluster assignments for k=2 to 9 have all been recorded. However, according to Section 6.3.4, for the case of 6 clusters or more, distinctions in breast shapes are too trivial (Figure 6-19) and some shapes, obtained from different clusters, look almost identical (their side profiles overlap when plotted together), which implies that classifying breast shapes into this large number of groups is unnecessary. On the other hand, for the 2-cluster case, due to the insufficiency in cluster choices, the impact of some variables got counterbalanced within cluster, rather than got identified and summarized by the cluster. It can be seen from Figure 6-15 that the corresponding two breast shapes are not significantly different enough. Therefore, the choice of k=2 for the number of clusters is also not appropriate. In conclusion, only the 3-cluster case, 4-cluster case and 5-cluster case have been kept for further discussion.

6.4.1. Criterion 1: Based on Misclassification Rate

Discriminant analysis and classification are often used together. Discriminant analysis creates discriminant functions that separate groups of observations from each other, based on existing group assignments (Rencher, 2003b). Classification usually refers to the process of allocating new cases into the previously defined groups (Rencher, 2003c). (Note that the term "classification" in statistics refers to a different method from clustering: in contrast with clustering, group identities, as well as the number of groups, are known in classification).

Although Discriminant analysis seeks for the optimum way to separate groups, it is often imperfect. New cases can be misclassified to wrong groups. The apparent error rate (AER) is the proportion of misclassified cases in a sample when observations in the sample have been regarded as new cases and got re-classified based on the discriminant

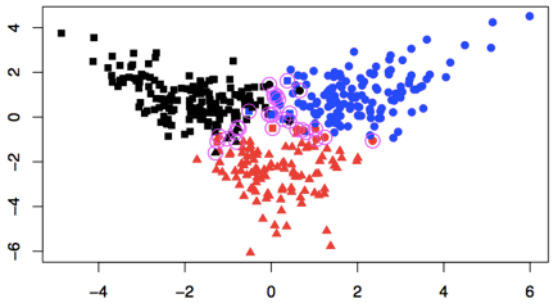
rules built on the same data (see Eq. 6-5). Clearly, a low AER, essentially a low misclassification rate, is preferable.

$$\text{AER}= \frac{\text{The number of misclassified cases}}{\text{The total number of observations}} \tag{6-5}$$

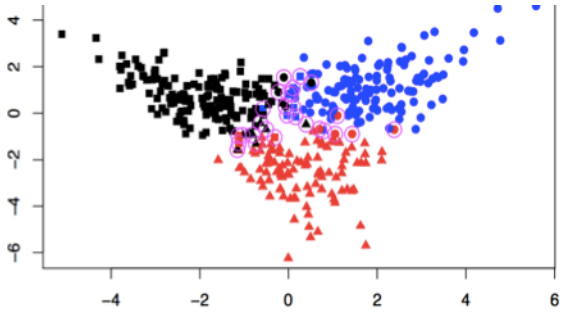
However, using the same cases in building classification rules makes it very likely for AER to underestimate the real misclassification rate. This bias can be avoided by cross validation, where data are divided into a training dataset, upon which discriminant rules are built, and a testing dataset, upon which the correctness of group assignment is tested. Complete information, especially for group membership, is required for each of the training data. For this study, the group membership obtained from clustering was used to create the discriminant functions.

Linear discriminant analysis (LDA), accompanied by linear classification rules, is a commonly used method in Discriminant analysis. Figure 624 shows the results of LDA on the data of this study, for the cases of 3, 4 and 5 clusters. In terms of the estimation of misclassification rate, N-1 cases were treated as the training data, where N=408 is the total number of observations, and the remaining case was regarded as a new case for testing and got classified into one of the groups. Then another N-1 cases got trained and the remaining case got re-assigned until every possible N-1 combination had been attended to. Each plot in Figure 6-24 displays the plane (or projection) that maximizes the dissimilarity between the previously found clusters, extracted from the high dimensional space. The color of points associates with the clustering result, i.e. cluster membership obtained from K-means cluster analysis, applied to either the whole data (Figure 6-24a, c and e) or the first 10 PC's (Figure 6-24b, d and f). The shape of points (e.g. circle, triangle, square, etc.) associates with the new classification result obtained from LDA. When the shape of a point did not match with the color that the point was supposed to

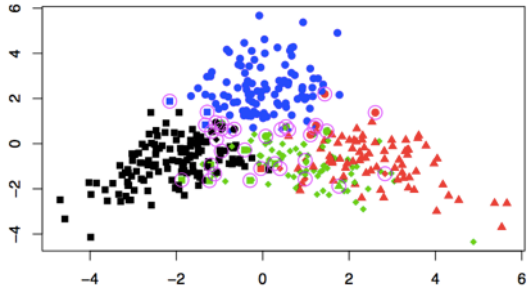
have, it revealed a wrongly classified case. All the misclassification cases were circled in pink in the figure. Clearly, the 3-cluster case has the lowest misclassification rate while 5-cluster case has the highest. In addition, whether K-means clustering was performed on the whole data or on the first 10 PC's leads to similar misclassification rates.



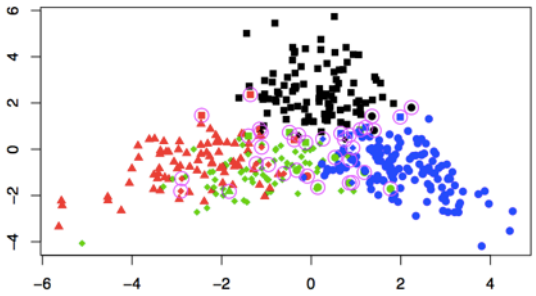
a). K-means on whole data (k=3)
Misclassification rate: 7.35%



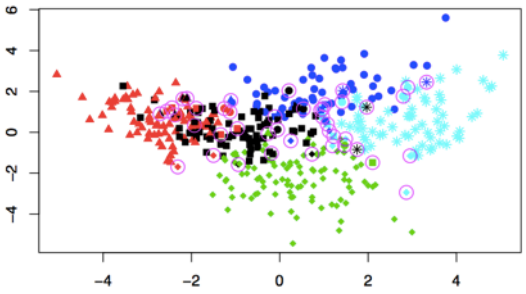
b). K-means on first 10 PC's (k=3)
Misclassification rate: 6.86%



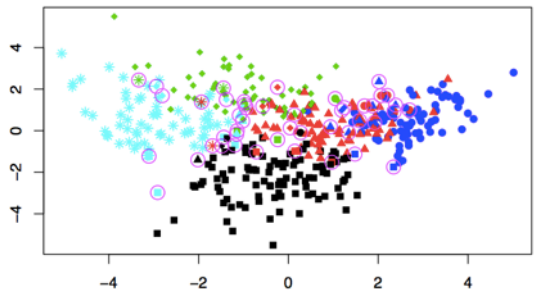
c). K-means on whole data (k=4)
Misclassification rate: 8.33%



d). K-means on first 10 PC's (k=4)
Misclassification rate: 8.82%



e). K-means on whole data (k=5)
Misclassification rate: 10.78%



f). K-means on first 10 PC's (k=5)
Misclassification rate: 11.27%

Figure 6-24. Misclassification rate of Linear discriminant analysis

Decision tree is another popular tool in data mining and machine learning. It belongs to the general method of recursive partitioning and produces a tree-like output in which each parent node has two succeeding child nodes, splitting the observations in the parent node based on certain criteria. Moreover, decision tree is the generalized term for classification tree and regression tree. Regression tree manages data with a numeric outcome (i.e. the dependent variable is continuous), whereas classification tree works with categorical responses (e.g. levels of groups). Classification tree is able to predict a group to which an observation belongs, and thus it is another way of building classification rules, other than the linear classification rules. However, classification tree often suffers from the issue of overfitting (constructing overly complicated trees that cannot generalize). Moreover, a single tree can be unstable, especially with large number of predictors. Minor changes to the data can result in a very different tree. Nonetheless, Random forest (RF) is a good way of fixing the issue. RF is essentially building a multitude of tree models. Each time it draws a bootstrap (random sampling of observations with replacement) sample, usually referred to as in-bag sample (IB sample), and uses the bootstrap sample as the training data to build each tree. Observations that are not included in the bootstrap sample are often called out-of-bag sample (OOB sample), and are treated as the testing data. The misclassification rate can be estimated from the OOB samples and their corresponding trees. Table 6-4 shows the OOB estimates of misclassification rate from 1000 trees.

Table 6-4.

OOB Estimate of Misclassification Rate from 1000 Trees

Clustering method	**3-Cluster Case**	**4-Cluster Case**	**5-Cluster Case**
Hierarchical	13.73%	18.14%	17.89%
K-medoids	9.31%	14.22%	18.38%

K-means (whole data)	7.84%	8.09%	10.78%
K-means (10 PC's)	6.86%	10.05%	9.80%
K-means (8 PC's)	7.11%	9.31%	11.52%
K-means (5 PC's)	7.35%	9.31%	12.25%
K-means (3 PC's)	7.11%	8.09%	9.31%

It can be seen from Table 6-4 that K-means clustering has the lowest misclassification rates among the three clustering methods. The 3-cluster case has the lowest misclassification rates, compared with 4-cluster case and 5-cluster case. Likewise, whether K-means was applied to the whole data or to the first few PC's leads to similar misclassification rates.

The first criterion in choosing cluster number is based on the misclassification rate, which focuses on how well a new case can be classified into the correct group. LDA and Random forest are methods based on different algorithms. Both of them have a different cross-validation approach to estimate the misclassification rate. However, both of them ended up with the same conclusion that the 3-cluster case has the lowest misclassification rate. Therefore, the ideal number of clusters based on Criterion 1 is 3.

6.4.2. Criterion 2: Based on Goodness-of-Fit of Model

Typically, how well a statistical model fits to a set of data is evaluated by the Goodness-of-Fit. A well-fitted model can capture and explain most of the variations among the observations. However, an overfitted model can run into the problem of generalization: the model that fits one specific sample of data perfectly will fail to predict the behavior of a larger population or of another set of observations. In terms of clustering, a saturated model (perfectly fitted model) is the case when each observation has its own cluster. Many statistical methods have been proposed to seek for models with reasonably high Goodness-of-Fit while at the same time avoiding overfitting. Three of

the commonly used references are the statistics of BIC (Bayesian Information Criterion), AIC (Aikaike's Information Criterion) and WSS (Within-groups Sum of Squares, also known as residual variance). However, none of the statistics is sufficient enough to be regarded as a comprehensive guideline for model selection. Hence, instead of referring to only one model selection criterion, all of the three statistics were used as the measures of Goodness-of-Fit in this study.

As mentioned earlier in Section 6.3.4, K-means clustering is the best among the three clustering methods in giving the most distinctive breast shapes. Moreover, K-means applied to the first 10 PC's behaves equally well as to the whole data set. Therefore, models involved in this section all refer to K-means clustering models, with different numbers of clusters, performed on the first 10 PC's.

One of the R functions (the "find.cluster" function in the package named "adegenet") implements the procedure of running successive K-means with a series of cluster numbers (k), and returning the suggested optimal k, determined by the Goodness-of-Fit of the model. For this study, k was set to be no more than 20 and no less than 1. Furthermore, for the reason that K-means clustering selects random seeds (starting points) as the initial centroids of clusters (see Section 6.3.2), the cluster assignments can slightly differ each time K-means has been conducted, accompanied by a slightly different model. Occasionally, although seldomly, an unsatisfactory clustering outcome can be caused by bad choice of initial seeds rather than the bad choice of k. In order to fix this problem, the "find.cluster" function was run repeatedly for 200 times. All of the three aforementioned statistics were calculated every time the function had been run. Figure 6-25 shows the votes for optimal k, proposed by AIC, BIC and WSS respectively. Clearly, k=5 received the highest votes by all three statistics. Therefore, the ideal number of clusters based on Criterion 2 is 5.

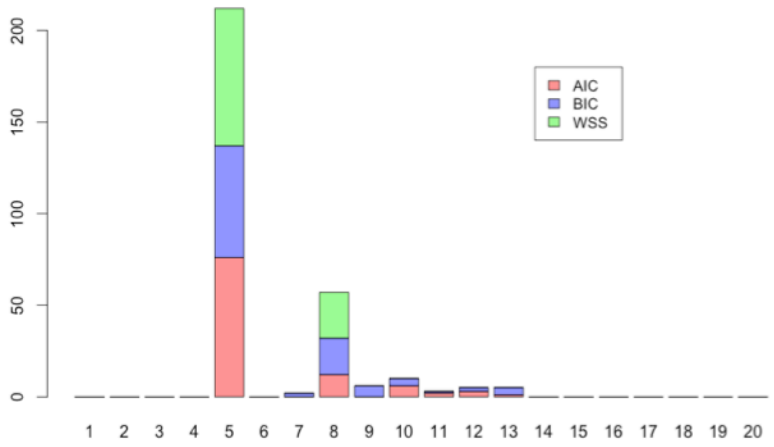


Figure 6-25. Votes by three different statistics for optimal cluster number

6.5. Multivariate Analysis of Variance (MANOVA)

Similar to the usage of ANOVA (Analysis of variance) for univariate data, MANOVA also tests the dissimilarity among the means of different groups, only that the group means are no longer some scalar values, but instead, a few 1 by p vectors, where p is the number of response variables recorded for each observation. The vectors of means are often called the multivariate means (Rencher, 2003d). For this study, MANOVA was used to examine whether the multivariate means of different clusters are significantly different for both the 3-cluster case and 5-cluster case. There are four major MANOVA test statistics, namely the Wilks' Lambda, the Roy's Maximum Root, the Hotelling-Lawley Trace, and the Pillai's Trace (see Eq. 6-6 to Eq. 6-9) (Rencher, 2003d). Each of the test statistics corresponds to a different approximation of the F-value and thus a different p-value. All four of the test statistics have been involved in this study. The null hypothesis of MANOVA is that all multivariate means are equal. A small p-value implies the rejection of the null hypothesis.

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \prod_{j=1}^{p}\left(\frac{1}{1+\lambda_j}\right) \tag{6-6}$$

where Λ refers to the Wilks' Lambda; **W** is the Error matrix (representing the within-group variation) and **B** is the Treatment matrix (or Hypothesis matrix, representing the between-group variation); $|\mathbf{W}|$ refers to the determinant of **W**, and $|\mathbf{B}+\mathbf{W}|$ refers to the determinant of (**B**+**W**); λ_j is the j-th eigenvalue of matrix $\mathbf{BW}^{-1}$.

$$\theta = \frac{\lambda_1}{1+\lambda_1} \tag{6-7}$$

where θ refers to the Roy's Maximum Root; λ_1 is the first and largest eigenvalue of matrix $\mathbf{BW}^{-1}$ (**B** is the Treatment matrix and **W** is the Error matrix).

$$L = (n-g)tr(\mathbf{BW}^{-1}) = \sum_{j=1}^{p} \lambda_j \tag{6-8}$$

where L refers to the Hotelling-Lawley Trace; n is the total number of observations, g is the total number of groups, and (n-g) refers to the degree of freedom; $tr(\mathbf{BW}^{-1})$ is the trace of the matrix $\mathbf{BW}^{-1}$ (**B** is the Treatment matrix and **W** is the Error matrix); and λ_j is the j-th eigenvalue of matrix $\mathbf{BW}^{-1}$.

$$P = tr\left[\mathbf{B}\,(\mathbf{B}+\mathbf{W})^{-1}\right] = \sum_{j=1}^{p}\left(\frac{1}{1+\lambda_j}\right) \tag{6-9}$$

where P refers to the Pillai's Trace; $tr\left[\mathbf{B}\,(\mathbf{B}+\mathbf{W})^{-1}\right]$ refers to the trace of the matrix $\mathbf{B}\,(\mathbf{B}+\mathbf{W})^{-1}$ (**B** is the Treatment matrix and **W** is the Error matrix); and λ_j is the j-th eigenvalue of matrix $\mathbf{BW}^{-1}$.

Table 6-5 is the summary table of MANOVA for the 3-cluster case. Clearly, the p-values calculated from all four of the test statistics are very small (much smaller than the 0.05 significance level). Therefore, it is safe to conclude that the null hypothesis has been rejected, and that the three clusters have statistically significantly different multivariate

means. Now that the three multivariate means are different, it is worthwhile inspecting which variables (i.e. breast measurements) are responsible for the difference. Hence, univariate ANOVA significance test has been conducted for every individual dependent variable. The full list of the test results can be found in Appendix E. All variables, except for Variable 16 (distR_BL_BO), Variable 36 (areaR_fan_rec_inner), and Variable 38 (areaR_fan_rec_outer), appear to be significantly different at a 0.05 level.

Table 6-5.

MANOVA Summary Table for 3-Cluster Case

	Df	**Pillai's Trace**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
3-Cluster Case	2	1.4102	21.344	82	732	< 0.0001
Residuals	405					

	Df	**Wilks' Lambda**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
3-Cluster Case	2	0.0867	21.337	82	730	< 0.0001
Residuals	405					

	Df	**Hotelling-Lawley Trace**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
3-Cluster Case	2	4.8048	21.329	82	728	< 0.0001
Residuals	405					

	Df	**Roy's Maximum Root**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
3-Cluster Case	2	2.5994	23.204	41	366	< 0.0001
Residuals	405					

Table 6-6 is the summary table of MANOVA for the 5-cluster case. Similarly, all of the four p-values are much smaller than 0.05. Thus, it can be concluded that the five clusters have significantly different multivariate means. Appendix F contains the full list of univariate ANOVA tests for individual variables. All variables appear to be significantly different among the five clusters at a 0.05 level.

Table 6-6.

MANOVA Summary Table for 5-Cluster Case

	Df	**Pillai's Trace**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
5-Cluster Case	4	2.5231	15.251	164	1464	< 0.0001
Residuals	403					

	Df	**Wilks' Lambda**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
5-Cluster Case	4	0.0160	16.112	164	1450	< 0.0001
Residuals	403					

	Df	**Hotelling-Lawley Trace**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
5-Cluster Case	4	7.7039	16.981	164	1446	< 0.0001
Residuals	403					

	Df	**Roy's Maximum Root**	**Approximate F-value**	**Numerator Df**	**Denominator Df**	**P-value**
5-Cluster Case	4	3.2411	28.933	41	366	< 0.0001
Residuals	403					

6.6. Reduction of Dimensionality

6.6.1. Importance of Variables and PC Loadings

As mentioned in Section 6.4.1, Random forest is the process of building lots of tree models (classification trees for the case of this study). During the process, each of the trees chooses the variables that are considered by the tree itself to be most important in terms of splitting the nodes in accordance with the pre-determined group membership (cluster membership for this case). Within each tree, the root node contains all observations and gets split in two parts for a few times so that each of the child nodes can be as homogeneous as possible. Meanwhile, the homogeneity of a node is often measured by Node Impurity, and a pure node is the case when every observations included by the node belongs to the same group (cluster). Moreover, the decrease in Node Impurity (defined via Eq. 6-10) is often regarded as the measure of Goodness-of-Split.

$$\Delta S = i(N) - P_1 \times i(N_1) - P_2 \times i(N_2) \tag{6-10}$$

where $i(N)$ refers to the impurity of the parent node; $i(N_1)$ is the impurity of one of the two child nodes, and $i(N_2)$ is the impurity of the other child node; P_1 and P_2 refer to the proportion of cases included by the first child node, and by the other child node respectively.

Furthermore, the Random forest package can generate plots for the importance of variables based on their impact on the Goodness-of-Split, measured by the decrease in Node Impurity (see Figure 6-26 and 6-27). Larger value in the plot represents higher importance. The importance measure is often used for variable selection to obtain a simpler model. Hence, it is a helpful reference for the reduction of dimensionality.

On the other hand, each PC is a linear combination of the 41 variables, and within each linear combination, variables with larger loadings (or coefficients) have greater impact on determining the direction of the corresponding PC. Moreover, according to Table 6-1, the

first three PC's account for 52.3% of the total variance. In addition, it can be seen from Figure 6-21 and 6-23 that 3 PC's are sufficient enough in reaching the similar clustering outcomes as those that were obtained from the whole data, for both cases when k=3 and k=5. Therefore, the loadings of variables for the first 3 PC's were also regarded as a reference for the dimension-reduction process. Accordingly, in the importance plots, variables that have large loadings for the first PC were colored in red; variables that have large loadings for the second or third PC were colored in green or in blue respectively.

Figure 6-26 presents the importance plot for the 3-cluster case (note that the cluster membership designated for Random forest was the outcome of K-means clustering performed on the first 10 PC's). Clearly, a few variables, colored in red or green, have outstanding importance, but not those that were colored in blue. Variables mainly responsible for the direction of the third PC seem not to be involved much in splitting the nodes (i.e. separating the groups).

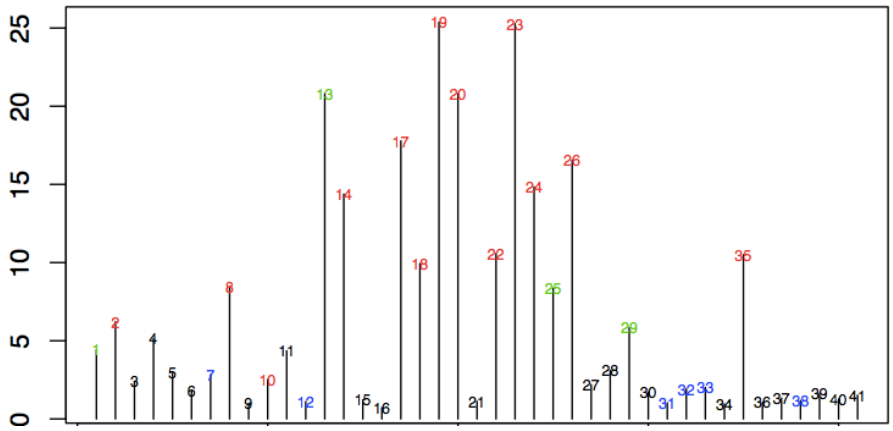


Figure 6-26. Measure of importance for variables based on decrease in Node Impurity (K-means applied to 10 PC's, k=3)

Similarly, Figure 6-27 shows the importance plot for the 5-cluster case. Not surprisingly, as the number of clusters increases, the number of variables (and also the

number of PC's) that are required to accurately place observations into the correct clusters increases. Several variables colored in blue stand out in the plot, suggesting that the third PC has been involved, in addition to the first two PC's.

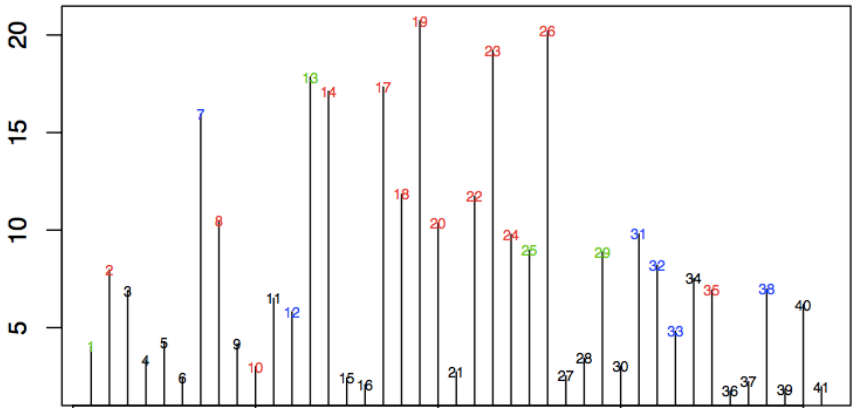


Figure 6-27. Measure of importance for variables based on decrease in Node Impurity (K-means applied to 10 PC's, k=5)

6.6.2. The 3-Cluster Case

With the reference of the PC loadings and the importance measures, a few key variables were selected from the 41 to start with. According to Figure 6-26, the following 8 variables appear to be more important: Variable #23, #19, #13, #20, #17, #26, #14 and #24. They were regarded as and will be referred to as the key variables. Then PCA was performed on the 8 key variables alone, with all 408 observations included. Table 6-7 lists the summary statistics for all the 8 PC's obtained from the new PCA. It can be seen that the first two PC's account for over 90% of total variance. Hence, K-means clustering was performed on the first 2 PC's of the 8 key variables, and Figure 6-28b shows the clustering result demonstrated through breast shape.

Table 6-7.

PCA Summary Table (8 Key Variables for the 3-Cluster Case)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	2.2850	1.4756	0.5912	0.3999	0.2305	0.1660
Proportion of Variance	0.6526	0.2722	0.0437	0.0200	0.0066	0.0034
Cumulative Proportion	0.6526	0.9248	0.9685	0.9885	0.9952	0.9986
	PC7	**PC8**				
Standard deviation	0.0867	0.0607				
Proportion of Variance	0.0009	0.0005				
Cumulative Proportion	0.9995	1.0000				

Figure 6-28a displays the clustering outcome when K-means was applied to the first 10 PC's, obtained from the original PCA with all 41 variables included. It can be observed that Figure 6-28b and 6-28a present similar outcomes. In other words, the combination of the 8 key variables can function almost equally well as the 41 variables as a whole can. Hence, the number of variables involved in the clustering can be reduced from 41 to 8 without losing too much information.

a). K-means applied to the first 10 PC's from 41 variables

b). K-means applied to the first 2 PC's from 8 key variables

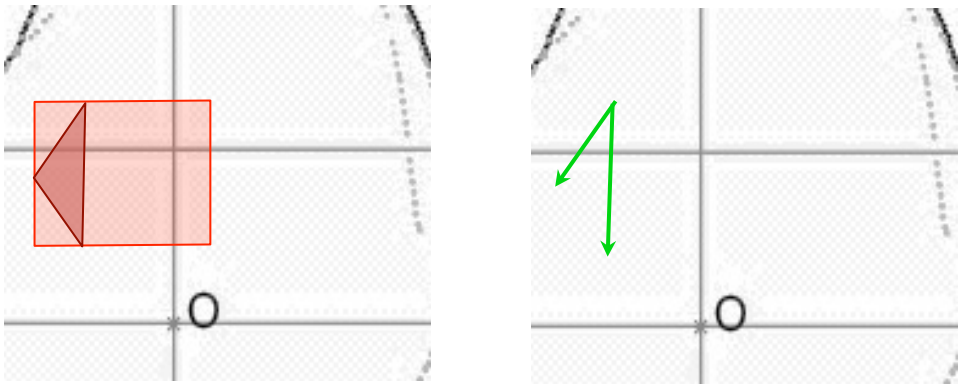
c). K-means applied to the first 2 PC's from 2 variables (#23, #13)

d). K-means applied to the first 2 PC's from 2 variables (#23, #19)

Figure 6-28. Reduction of the number of variables (3-cluster case)

Then multiple trials were undergone. Each time one more variable (from the 8 key variables) was excluded and K-means clustering was applied to the new PC's, calculated from the remaining key variables. The new clustering result was compared with the original (Figure 6-28a). Similar results led to further variable deduction whereas significantly distinctive results led to the reservation of that variable and the attempt of deleting another variable (from the 8 key variables). The sequence of the deletion referred to Figure 6-26: variables with lower importance were the first ones to be considered for exclusion. In the end, the number of variables was successfully reduced to 2 (see Figure 6-28c), and the corresponding variables are Variable #23 and Variable #13.

Moreover, it is worthwhile mentioning that despite Variable #19 has higher importance than Variable #13 (according to Figure 6-26), it does not give a satisfying clustering result (Figure 6-28d). This is probably because Variable #13 is responsible for the direction of the second PC (it was colored in green in Figure 6-26), while #23 and #19 are both influential variables for the direction of the first PC (both of them were colored in red in Figure 6-26). It is a proof showing that retaining the second dimension is essential.



a). Variable #23 b). Variable #13

Figure 6-29. Demonstration of the two finalized variables (3-cluster case)

As demonstrated in Figure 6-29, Variable #23 (areaR_tri_rec) is the area ratio between Triangle ABD and the rectangle at anterior body; Variable #13 (angle_tri_top) is the Angle BAD. Both of them relate to the Triangle ABD. These two variables alone are sufficient enough in partitioning observations into 3 clusters.

6.6.3. The 5-Cluster Case

With the reference of the PC loadings and the importance measures (Figure 6-27), 17 key variables (much more than the number of key variables initially chosen for the 3-cluster case) were selected to start with, namely Variable #19, #26, #23, #13, #14, #7, #17, #18, #22, #20, #8, #25, #24, #31, #29, #32, and #2. Then PCA was performed on the 17 key variables, with all 408 observations included. Table 6-8 lists the summary statistics for the 17 PC's obtained from the new PCA. It can be seen that the first three PC's account for over 80% of total variance. Hence, K-means clustering was performed on the first 3 PC's of the 17 key variables, and Figure 6-30b shows the clustering result demonstrated through breast shape.

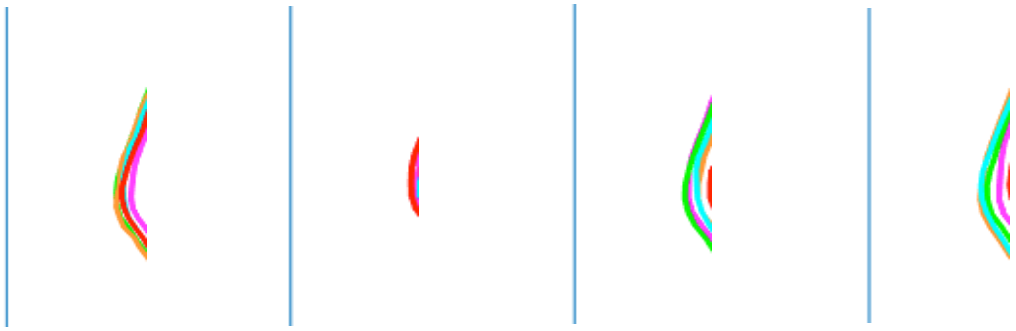
Table 6-8.

PCA Summary Table (17 Key Variables for the 5-Cluster Case)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	2.8543	2.0022	1.4309	0.9191	0.8058	0.7606
Proportion of Variance	0.4792	0.2358	0.1204	0.0497	0.0382	0.0340
Cumulative Proportion	0.4792	0.7150	0.8355	0.8852	0.9234	0.9574
	PC7	**PC8**	**PC9**	**PC10**	**PC11**	**PC12**
Standard deviation	0.5428	0.4367	0.2849	0.2489	0.2045	0.1837
Proportion of Variance	0.0173	0.0112	0.0048	0.0036	0.0025	0.0020
Cumulative Proportion	0.9747	0.9860	0.9907	0.9944	0.9968	0.9988
	PC13	**PC14**	**PC15**	**PC16**	**PC17**	

Standard deviation	0.0866	0.0766	0.0631	0.0442	0.0309
Proportion of Variance	0.0004	0.0003	0.0002	0.0001	0.0001
Cumulative Proportion	0.9993	0.9996	0.9998	0.9999	1.0000

Figure 6-30a displays the clustering outcome when K-means was applied to the first 10 PC's, obtained from the original PCA with all 41 variables included (k=5). It can be observed that Figure 6-30b and 6-30a present similar outcomes. Hence, the number of variables involved in the clustering can be reduced from 41 to 17 without losing too much information.



a). Kmeans applied to the first 10 PCs from

b). Kmeans applied to the first 3 PCs from 17 key variables

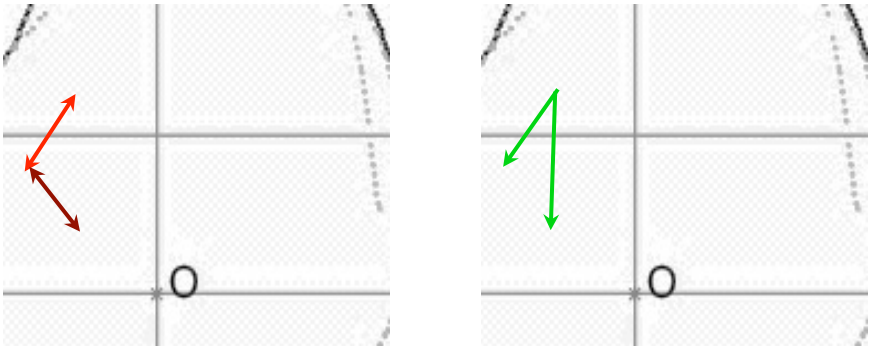
c). Kmeans applied to the first 3 PCs from 4 variables (#19, #13, #14, #32)

d). Kmeans applied to the first 3 PCs from 3 variables (#19, #13, #32)

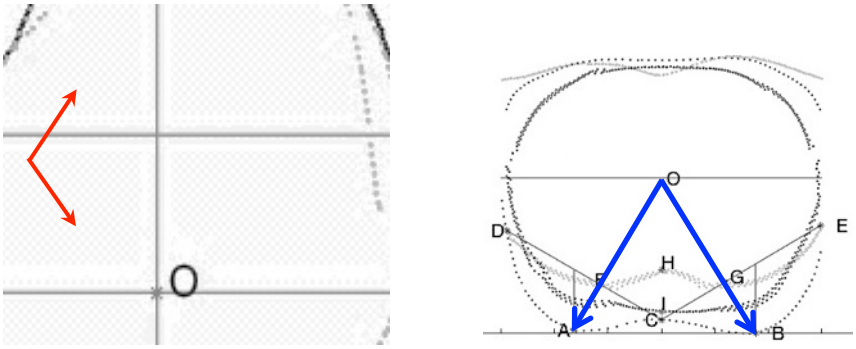
Figure 6-30. Reduction of the number of variables (5-cluster case)

After multiple trials with one of the 17 key variables excluded each time, the number of variables was successfully reduced to 4 (see Figure 6-30c), and the corresponding variables are Variable #19, #13, #14 and #13. Any further exclusion of the key variables led to a different clustering result (Figure 6-30d) even when all three dimensions had been retained (according to the colors in Figure 6-27, Variable #19, #13

and #32 are influential variables for the directions of the first PC, the second PC and the third PC respectively).



a). Variable #19 b). Variable #13



c). Variable #14 d). Variable #32

Figure 6-31. Demonstration of the four finalized variables (5-cluster case)

As demonstrated in Figure 6-31a, b and c, Variable #19 (distR_upperB_lowerBroot) is the length ratio between Line Segment AB and Line Segment BD; Variable #13 (angle_tri_top) is the Angle BAD; and Variable #14 (angle_tri_side) is the Angle ABD. All three of them relate to the Triangle ABD. In addition, Variable # 32 (angle_pointing) is the Angle AOB demonstrated in Figure 6-31d, representing the pointing of bust points inspected from the transverse plane. The four variables alone are sufficient enough in partitioning observations into 5 clusters.

CHAPTER 7

CONCLUSIONS

7.1. Answers to the Research Questions

The purpose of this study is to understand the variation in female breast shape across the Caucasian population, and to propose a classification method for breast shape. Three research questions were posed and the answers are as following:

Question 1: What are the most critical body measurements that best define breast shape?

Among all variables (41 ratios and angles), two measurements, obtained from the merged sagittal planes, were found sufficient in categorizing breast shapes into three groups: 1) Variable 23- areaR_tri_rec, the area ratio between the self-constructed bust triangle and the rectangle of upper anterior body, and 2) Variable 13- angle_tri_top, the top angle of the bust triangle, showing the slope of the upper breasts.

Likewise, four measurements were found sufficient in categorizing breast shapes into five groups: 1) Variable 19- distR_upperB_lowerBroot, the length ratio between two sides of the bust triangle; 2) Variable 13- angle_tri_top, the top angle of the bust triangle; 3) Variable 14- angle_tri_side, the side angle of the bust triangle; and 4) Variable 32- angle_pointing, representing the pointing of bust points inspected from the merged transverse planes. Note that the first three key measurements were extracted from the sagittal planes, and associates with the self-constructed bust triangle.

Question 2: What is the best way to classify breast shapes? How many groups should they be classified into?

This study selected three most commonly used clustering methods, namely Hierarchical clustering, K-means clustering, and K-medoids clustering, and applied them to the breast shape data. It was found that K-means is the best in giving the most distinctive breast shapes and presenting good stability and repeatability.

As for the choice of cluster number, this study proposed two selection criteria based on different considerations.

- Criterion 1. Based on the misclassification rate.
- Criterion 2. Based on the Goodness-of-Fit of model

The first criterion focuses on how well a new case can be classified into the correct group. The second criterion focuses on how well the model can capture and explain the majority of variations in the data, meanwhile, avoid overfitting. In terms of the breast shape classification, three and five are the ideal group numbers according to Criterion 1 and Criterion 2 respectively. This study also found that six or more clusters could end up with many of the clusters having only minimal differences; and less than three clusters could end up with clusters that have poor distinctions, due to the counterbalance of data variation caused by the insufficiency in cluster choices. In general, a wrong choice of cluster number can lead to poor clustering results that do not reflect the real homogeneity and heterogeneity in the data.

Question 3: How to present and validate the final classification outcomes?

When categorizing breast shapes into three or fewer groups, the dimensionality of the data can be reduced to two, thus a 2-dimensional scatter plots with points colored by cluster is sufficient in presenting clustering results. Similarly, when categorizing into five or fewer groups, the data dimension can be reduced to three, thus a rotating 3-dimensional scatter plots is sufficient. Generally, more groups require the plotting of higher dimensional scatter plots, but plotting points in 4-dimensional or higher is not intuitive, if not impossible.

In order to provide better visualization of the clustering outcomes, an algorithm was developed in Matlab for this study to acquire the median curves of body side profile and bust plane for each breast shape group. It was proven to be valid and helpful through many of the analysis results, especially for the K-medoids clustering analysis: when the

result of the algorithm was compared with the real median case of each group (i.e. the medoid case), the two displayed satisfying similarity.

7.2. Discussions and Conclusions

A few limitations in previous research were found and mentioned in Section 2.3. The followings are what this study did to deal with those limitations so that the methodology of the study can be more robust and comprehensive, and the results can be more convincing.

While many of previous body shape studies relied on manual extraction of body measurements from 3D body scans and the use of multiple software, this study developed a Matlab program to achieve automatic extraction of all measurements desired. Using one program could avoid the inconsistency and error caused by different calibrations or settings in various software programs. Automatic extraction could save time when dealing with large scale of scans, and avoid human error due to unintentional mistakes in operation.

Meanwhile, this is one of the first female breast shape studies that involve large scale of body scans of the Caucasian population. A total of 478 CAESAR scans were involved, and all of the subjects were scanned in a standard standing position to avoid changes in breast shapes caused by posture. Additionally, unlike some studies in the past that analyzed body measurements separately, this study utilized various multivariate statistical methods, data mining and machine learning techniques, to retain and study the correlations between body measurements.

Moreover, despite that the importance of data examination was underestimated in the past (outlier detection, data re-expression and assumption diagnostics was rarely done), this study dealt with the initial data with extra care. The data were thoroughly examined for outliers and skewed distributions. A Shiny app, which was able to display interactive histograms, pairwise scatter plots, Q-Q plots, etc., was developed to improve

the efficiency of this process. In the end, 70 subjects were removed, and a few variables were transformed for better normality and linearity. Although no significant difference in the Principal component analysis (PCA) results were observed between the transformed and the untransformed data, it was still a meaningful step. The similarity in PCA results is probably because all body measurements involved in this study are relative measurements (ratios and angles), and thus have less chance of containing extreme values or being super skewed. However, for studies that include absolute measurements (e.g. circumferences, widths, heights, etc.), data preparation may alter the analysis results entirely.

Furthermore, there were some limitations in previous studies regarding the application of statistical methods.

With regard to PCA, some researchers in the past tried to use PCA summary table and scree plot to justify their analysis. However, through the comparison between standardized and unstandardized data, this study found that although the unstandardized data had a plausibly more promising summary statistics table for PC's and nicer-looking scree plot, it did not work well in terms of classifying breast shapes. Hence, providing the table and scree plot alone may not be sufficient enough in showing how good the result is.

With regard to Cluster analysis, few of previous researchers mentioned the extensive choices of clustering algorithms. In spite of this, this study compared three clustering methods with the help of the self-developed Matlab program, which can visualize how distinctive the breast shapes are for different clusters. In addition, this study utilized Multivariate analysis of variance (MANOVA) to examine the multivariate means of different clusters, and all four types of test statistics agreed that the means were significantly distinctive. Univariate Analysis of variance (ANOVA) tests on individual variables also showed that the majority of body measurements were significantly different among clusters at a 0.05 level (with three exceptions out of 41 for the 3-cluster

case, namely Variable 16- distR_BL_BO, Variable 36- areaR_fan_rec_inner, and Variable 38- areaR_fan_rec_outer).

7.3. Implications to the Apparel Industry

A well-fitting bra is able to conform to the breasts and present a neat appearance on the surface. Therefore, a better understanding of breast shape can help with the improvement of the design of bras and other types of intimate apparel, especially in terms of fit and comfort.

Traditional bra sizing system uses the difference between bust and underbust circumferences as the only reference for cup size. However, this measurement alone cannot fully describe the 3D shape of breasts. For this reason, this study extracted a total of 66 raw measurements, including width, depth, thickness, angles, distances, etc., and constructed 41 shape-related variables, with the traditional measurement (Variable 1- circmR_deltaB_bust) included, aiming to find a measurement, or combination of measurements, that captures the variations in breast shape more efficiently. It turned out that the efficient combinations of measurements do not include the traditional measurement (Variable 1).

Furthermore, not only did this study found the most representative breast shapes based on all 41 variables, which can be directly referred to when building dress forms or designing bras, this study also proposed an approach to reduce the number of variables so that the key body measurements can be identified. The same methodology can be adopted in the shape study of other parts of the body that apparel designers are interested in, for instances, buttock shapes.

Moreover, the key measurements identified for breasts are relatively easy to measure. The two key measurements required by 3-cluster case and three of the four key measurements required by 5-cluster case all relate to the same bust triangle, thus as long as the locations (coordinates) of the vertex points of the triangle are obtained, the key

measurements can be calculated easily. The only key measurement that does not relate to the triangle is the pointing of bust points, which can also be easily calculated as long as the coordinates of the bust points are collected. When working with body scans, the algorithm is even simpler than the extraction of circumferential measurements. In addition, the discriminant functions built to differentiate the clusters, and their corresponding classification rules can be used to allocate a new human subject into one of the breast shape groups based on the aforementioned key measurements.

The findings and the proposed methods of this study make it possible to develop real bra products that work for both manufacturers and consumers, and to build an improved sizing system for bras or other type of intimate apparel.

7.4. Limitations and Suggestions for Future Research

This study had several limitations. Firstly, each participant of the CAESAR project was scanned wearing a soft sports bra, rather than had her breasts scanned in nude. It is possible that a soft sports bra may have altered the shape of the breasts and the influence of the sports bra remains unknown. The outcomes of this study could be more convincing if the same methodology was applied to real nude breast scans. However, it is important to note that nude scans can also be unsatisfactory as the shape of a nude breast can be very different from the desired shape provided by the support of the bra, and the amount of difference between the nude and desired shape can vary greatly among individuals. Another way of fixing this issue is by conducting a thorough research on the impacts of different types of bras, including soft sports bra, on the shape of nude breasts. This kind of research was also rarely done in the past, yet can provide additional meaningful information.

Secondly, side profile view of the breasts is the only reference when the clustering methods were being compared, and when the choices of cluster number were being narrowed down. It is possible that similar side profiles turn out to have different 3D

shapes. In the future, judgments and decisions can be made by additionally referring to other views of the breasts, transverse planes sliced at other levels, and sagittal planes sliced at other landmark locations.

Thirdly, although this study had taken into consideration of the impact of age, ethnicity and BMI on breast shape, there were some other factors that remain unexamined, for instance, the impact of plastic surgery, i.e. breast augmentation or reduction, the impact of breast-feeding history. It was mainly because this kind of information was not included in the CAESAR database. For future studies, more information about participants is suggested for keeping a record of.

Moreover, this study concentrated on a particular population: the North American, younger, non-obese Caucasians. In the future, the same methodology can be applied to other populations to explore to what extend the outcomes of this study can be generalized.

Furthermore, in terms of the improvement for methodology, the analysis and validation can be more persuasive if the aggregate loss of fit (McCulloch, Paal & Ashdown, 1998) is calculated and evaluated so that the difference between group means and individual subjects can be assessed.

In addition, this study focused on breast shape regardless of breast size, but the interaction between breast size and breast shape can be very informative and worthwhile studying in the future.

Lastly, with respect to further practical applications that might be useful to the apparel industry, there are a few more things that could be done. The k-medoids clustering algorithm is able to return the real cases it selects as medoids. Meanwhile, the group assignment obtained from k-means clustering is available. Hence, by applying the k-medoids algorithm to subjects that belong to the same k-means group with k set to be 1 (to make sure those subjects do not get classified again), the most appropriate fit models can be identified from the scan database. At the same time, the 3D geometric shapes of

the breasts of the fit models can be used to design molded cups, and together with the information on breast sizes, a new bra sizing system can be built with improved fit and comfort, but fewer cup size choices.

APPENDIX A. Variables constructed from raw measurements

	Name of variable	Description
1	circmR_deltaB_bust	The difference [between bust & underbust circumferences] divided by bust circumference
2	thickR_armp_bust	thickAE divided by thickBF
3	thickR_underb_bust	thickCG divided by thickBF
4	thickR_front_armp	thickAO divided by thickAE
5	thickR_front_bust	thickBO divided by thickBF
6	thickR_front_underb	thickCO divided by thickCG
7	thickR_front_underb2Broot_Broot	The difference [between thickCO & thickDO] divided by thickDO
8	heightR_upperB_fullB	heightEF divided by the sum [of heightEF & heightFG]
9	heightR_underb2Broot_lowerB	heightGD divided by heightFG
10	angle_upperB	angleABF
11	angle_lowerB	angleCBF
12	angle_lower_diff	angleCBD
13	angle_tri_top	angleBAD
14	angle_tri_side	angleABD
15	distR_BL_BDperp	distBL divided by the difference [between thickBO & thickDO]
16	distR_BL_BO	distBL divided by thickBO
17	distR_depth_height	distB_AD divided by distAD
18	distR_upperB_lowerB	distAB divided by distBC
19	distR_upperB_lowerBroot	distAB divided by distBD

20	areaR_curv_rec_upper	areaBupper divided by area_rec_upper
21	areaR_curv_rec_lower	areaBlower divided by area_rec_lower
22	areaR_curvUp_curvLow	areaBupper divided by areaBlower
23	areaR_tri_rec	areatriABD divided by the sum [of area_rec_upper & area_rec_lower]
24	areaR_tri_trap_upper	area_tri_upper divided by area_trap_upper
25	areaR_tri_trap_lower	area_tri_lower divided by area_trap_lower
26	areaR_fanUp_fanLow	area_curvAB_linA divided by area_curvBD_linD
27	areaR_arc_tri_upper	The difference [between area_curvAB_linA & area_tri_upper] divided by area_tri_upper
28	areaR_arc_tri_lower	The difference [between area_curvBD_linD & area_tri_lower] divided by area_tri_lower
29	areaR_underb_bust	area_ub divided by area_bust
30	widthR_underb_bust	width_ub divided by width_bust
31	widthR_bp2bp_bust	BP2BP divided by width_bust
32	angle_pointing	angleBP
33	angle_pointing_rt	angleBP_rt
34	depth_width_ratio	The mean of ratio_rt & ratio_lt, where ratio_rt=depthAF divided by distCD ratio_lt=depthBG divided by distCE
35	heightR_armp2BP_shd2BP	The mean of ratio_rt & ratio_lt, where ratio_rt=the difference [between height_acro2BP_rt & height_shd2armp_rt] divided by height_acro2BP_rt ratio_lt=the difference [between height_acro2BP_lt & height_shd2armp_lt] divided by height_acro2BP_lt

36	areaR_fan_rec_inner	The mean of ratio_rt & ratio_lt, where ratio_rt=area_curveACF divided by rec_rt_inner ratio_lt=area_curveBCG divided by rec_lt_inner
37	areaR_innerfan_rt_lt	area_curveACF divided by area_curveBCG
38	areaR_fan_rec_outer	The mean of ratio_rt & ratio_lt, where ratio_rt=area_curveADF divided by rec_rt_outer ratio_lt=area_curveBEG divided by rec_lt_outer
39	areaR_outerfan_rt_lt	area_curveADF divided by area_curveBEG
40	areaR_fullarc_rec	The mean of ratio_rt & ratio_lt, where ratio_rt=area_curveCAD divided by rec_right ratio_lt=area_curveCBE divided by rec_left
41	areaR_fullarc_rt_lt	area_curveCAD divided by area_curveCBE

APPENDIX B. Programming codes of the Shiny app

```
library(shiny)
library(car) # this is for Q-Q plots
load("~/breast_shape_data.RData") # load the data

ui <- fluidPage(
        tags$h1("Data Re-expression"),
        tags$h6("Jie Pei, 3/28/2016"),
        tags$hr(),
        sidebarLayout(
                sidebarPanel(
                        sliderInput(inputId = "num", label = "Check the N_th Variable",
                                    min=1,max=41, 1),
                        tags$br(),
                        actionButton(inputId = "loadOrig",label="Original Data
                                     (untransformed)"),
                        tags$br(),
                        tags$br(),
                        actionButton(inputId = "loadTrans", label="Transformed and
                                     scaled data"),
                        tags$br(),
                        tags$br(),
                        actionButton(inputId = "loadOrig.white",label="Target Population
                                     (white & BMI <30)"),
                        tags$br(),
                        tags$br(),
                        actionButton(inputId = "reduce",label="Transformed Data (Target
                                     Population)"),
                        tags$br(),
                        tags$br(),
                        checkboxInput("transmethod", "Show transformation method
                                      used", FALSE),
                        tags$hr(),
                        sliderInput(inputId = "compare", label = "Choose the Second
                                    Variable to Plot Against",min=1,max=41, 2, step = 1,
                                    animate= animationOptions(interval=1000, loop=F)),
                        checkboxInput("showid", "Identify the ID of points", FALSE),
                                      width = 4
```

```
                    ),
        mainPanel(
                tags$h3(textOutput(outputId="trans.med")),
                tabsetPanel(type = "tabs",
                        tabPanel("Histogram", plotOutput(outputId = "hist")),
                        tabPanel("Scatter plot", tags$br(), textInput(inputId="checkpoint",
                                label="Input the ID of the point you want to check:",
                                value = ""),
                               plotOutput(outputId = "scatter", width = "85%", height =
                                        "600px", brush=
                                        brushOpts("plot_brush",resetOnNew=T)),
                                verbatimTextOutput("info")),
                        tabPanel("New transform", tags$br(),
                                plotOutput(outputId = "updatehist"),
                                sliderInput(inputId = "power", label = "Apply Nth power
                                        tranformation (0=log)",min=-5,max=10,
                                        1,step=0.5)),
                        tabPanel("Age Groups", tags$br(),
                                 selectInput("ageLevel", "Color Points by Age",
                                        choices = c("None","All Age Levels","Age
                                         18-25", "Age 25-35", "Age 35-45")),
        plotOutput(outputId = "agescatter", width = "85%", height = "550px"))
                        )
                )
        )
)

server <- function(input, output) {
  # Create reactive values
        rv= reactiveValues(ratio=ratios,Id=as.character(id),
                           color=rgb(0, 0,0, 0), ageL=agelevel,
                           method=trans.method.na, binnum=25) # initialization

         observeEvent(input$loadOrig,{rv$ratio= ratios; rv$binnum=20;
                        rv$method=trans.method.na; rv$color=rgb(0, 0,0, 1/3);
                        rv$Id=as.character(id); rv$ageL=agelevel})
         observeEvent(input$loadOrig.white,{rv$ratio= focus.org; rv$binnum=20;
                        rv$method=trans.method.na; rv$color=rgb(1, 1,0, 1/3);
                        rv$Id=as.character(focus.id.ex); rv$ageL=agelevel.focus})
```

```
observeEvent(input$loadTrans,{rv$ratio= ratios4.0; rv$binnum=25;
                    rv$method=trans.method; rv$color=rgb(0,0,1, 1/3);
                    rv$Id=as.character(id); rv$ageL=agelevel})
observeEvent(input$reduce,{rv$ratio= focus.tran; rv$binnum=20;
                    rv$method=trans.method.foc; rv$color=rgb(1,0,0, 1/3);
                   rv$Id=as.character(focus.id.ex); rv$ageL=agelevel.focus})

output$hist= renderPlot({
        layout(matrix(c(1, 2), nrow = 1, ncol = 2, byrow = TRUE))
        hist(rv$ratio[,input$num], #breaks=rv$binnum,
             main = colnames(ratios)[input$num],xlab=" ",ylab=" ", col=rv$color)
        qqPlot(rv$ratio[,input$num],ylab=" ")
})

output$scatter= renderPlot({
        plot(rv$ratio[,input$num],rv$ratio[,input$compare],
                xlab=colnames(ratios)[input$num],
                ylab=colnames(ratios)[input$compare])
points(rv$ratio[which(rv$Id==input$checkpoint),input$num],
        rv$ratio[which(rv$Id==input$checkpoint),input$compare], col = "red",
        pch=9)
})

output$trans.med=renderText({
        "  "
        if (input$transmethod) {rv$method[input$num]}
})

output$info <- renderPrint({
        if (input$showid) {
                cat("input$plot_brush:\n")
                dat=as.data.frame(cbind(rv$ratio[,c(input$num,input$compare)]))
                rownames(dat)=rv$Id
                brushedPoints(df=dat, brush= input$plot_brush,
                                xvar=colnames(ratios)[input$num],
                                yvar=colnames(ratios)[input$compare])
                }
})

output$updatehist= renderPlot({
```

```
        varmin=min(rv$ratio[,input$num])
        layout(matrix(c(1, 2), nrow = 1, ncol = 2, byrow = TRUE))
        histdat=log(rv$ratio[,input$num]+abs(min(0,varmin)))
        if (input$power) {
                hist((rv$ratio[,input$num]+abs(min(0,varmin)))^input$power,
                        main = colnames(ratios)[input$num],xlab=" ", ylab=" ",
                        col=rv$color)
                qqPlot((rv$ratio[,input$num]+abs(min(0,varmin)))^input$power,
                          ylab=" ")
        }
        else { hist(log(rv$ratio[,input$num]+abs(min(0,varmin))),
                     breaks=rv$binnum, main = colnames(ratios)[input$num],
                     xlab=" ",ylab=" ", col=rv$color)
                qqPlot(log(rv$ratio[,input$num]+abs(min(0,varmin))),ylab=" ")
        }
})

ageInput <- reactive({
        switch(input$ageLevel,
                "None" = rv$ratio,
                "All Age Levels" = rv$ratio,
                "Age 18-25" = rv$ratio[rv$ageL==2,],
                "Age 25-35" = rv$ratio[rv$ageL==3,],
                "Age 35-45" = rv$ratio[rv$ageL==4,])
})

agecolor<- reactive({
        switch(input$ageLevel,
                "None" = 1,
                "All Age Levels" = rv$ageL,
                "Age 18-25" = 2,
                "Age 25-35" = 3,
                "Age 35-45" = 4)
})

output$agescatter= renderPlot({
        plot(ageInput()[,input$num],ageInput()[,input$compare],
              xlab=colnames(ratios)[input$num],
              ylab=colnames(ratios)[input$compare],col=agecolor())
})
```

```
}

shinyApp(ui = ui, server= server)
```

APPENDIX C. Principal Component Analysis summary table

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	3.15110	2.55730	2.23380	1.91499	1.77843	1.66256
Proportion of Variance	0.24220	0.15950	0.12170	0.08944	0.07714	0.06742
Cumulative Proportion	0.24220	0.40170	0.52340	0.61285	0.68999	0.75741
	PC7	**PC8**	**PC9**	**PC10**	**PC11**	**PC12**
Standard deviation	1.27154	1.13516	1.05656	1.02497	0.94605	0.90069
Proportion of Variance	0.03943	0.03143	0.02723	0.02562	0.02183	0.01979
Cumulative Proportion	0.79684	0.82827	0.85550	0.88112	0.90295	0.92274
	PC13	**PC14**	**PC15**	**PC16**	**PC17**	**PC18**
Standard deviation	0.79190	0.72389	0.70077	0.63834	0.48895	0.45146
Proportion of Variance	0.01530	0.01278	0.01198	0.00994	0.00583	0.00497
Cumulative Proportion	0.93800	0.95081	0.96279	0.97273	0.97856	0.98353
	PC19	**PC20**	**PC21**	**PC22**	**PC23**	**PC24**
Standard deviation	0.40140	0.33929	0.30374	0.26360	0.23872	0.21748
Proportion of Variance	0.00393	0.00281	0.00225	0.00170	0.00139	0.00115
Cumulative Proportion	0.98746	0.99027	0.99252	0.99420	0.99560	0.99676
	PC25	**PC26**	**PC27**	**PC28**	**PC29**	**PC30**
Standard deviation	0.16670	0.14858	0.13269	0.12016	0.11264	0.09372
Proportion of Variance	0.00068	0.00054	0.00043	0.00035	0.00031	0.00021
Cumulative Proportion	0.99743	0.99797	0.99840	0.99875	0.99906	0.99928
	PC31	**PC32**	**PC33**	**PC34**	**PC35**	**PC36**
Standard deviation	0.08495	0.07420	0.06585	0.05528	0.05096	0.04594
Proportion of Variance	0.00018	0.00013	0.00011	0.00007	0.00006	0.00005
Cumulative Proportion	0.99945	0.99959	0.99969	0.99977	0.99983	0.99988
	PC37	**PC38**	**PC39**	**PC40**	**PC41**	
Standard deviation	0.04029	0.03614	0.03091	0.02401	0.01708	
Proportion of Variance	0.00004	0.00003	0.00002	0.00001	0.00001	
Cumulative Proportion	0.99992	0.99996	0.99998	0.99999	1.00000	

APPENDIX D. Principal Component loadings

	PC1	PC2	PC3	PC4	PC5
circmR_deltaB_bust	0.0279	-0.2441	0.0586	-0.1244	0.0427
thickR_armp_bust	-0.2529	0.0336	-0.0611	0.0476	0.0136
thickR_underb_bust	-0.0747	0.2059	-0.2130	0.2040	0.0679
thickR_front_armp	-0.1130	-0.1797	0.1527	0.2001	0.1728
thickR_front_bust	0.0587	-0.1637	0.1676	0.2293	0.1793
thickR_front_underb	0.0104	-0.0809	0.0859	0.2853	0.2004
thickR_front_underb2Broot_Broot	0.0237	-0.2236	-0.3202	0.0590	-0.0368
heightR_upperB_fullB	0.2633	0.1210	-0.0798	0.1322	0.0287
heightR_underb2Broot_lowerB	-0.0377	0.0329	-0.2151	0.1150	-0.2228
angle_upperB	-0.2073	0.0691	-0.0610	0.1027	0.0555
angle_lowerB	0.1515	-0.1918	0.1787	-0.0889	-0.0123
angle_lower_diff	-0.0049	-0.0553	-0.3097	0.1956	0.1808
angle_tri_top	-0.0520	-0.3460	-0.0525	-0.0534	-0.0064
angle_tri_side	-0.2248	0.2216	0.0239	0.0030	-0.0694
distR_BL_BDperp	0.0185	0.0712	0.0813	-0.2509	-0.1992
distR_BL_BO	0.0573	-0.1105	-0.0163	-0.2912	-0.2455
distR_depth_height	0.2651	-0.1424	-0.0078	0.0160	0.0899
distR_upperB_lowerB	0.2572	0.1478	-0.1071	0.1401	0.0263
distR_upperB_lowerBroot	0.2627	0.1915	0.0333	0.0606	0.0675
areaR_curv_rec_upper	0.2946	0.0605	-0.0096	-0.0255	-0.0559
areaR_curv_rec_lower	-0.0613	0.1449	0.1144	0.0716	0.2905
areaR_curvUp_curvLow	0.2701	0.0967	-0.0059	0.0892	0.1008
areaR_tri_rec	0.2818	-0.1192	-0.0941	-0.0338	-0.0315
areaR_tri_trap_upper	0.2969	0.0463	0.0014	-0.0545	-0.0504
areaR_tri_trap_lower	0.0741	-0.3076	-0.1663	-0.0614	-0.0764
areaR_fanUp_fanLow	0.2648	0.1791	0.0659	0.0161	0.0202
areaR_arc_tri_upper	-0.1387	-0.0939	0.0449	-0.0853	0.0519
areaR_arc_tri_lower	0.0234	-0.2235	-0.0907	-0.0040	0.2052
areaR_underb_bust	-0.0859	0.2437	-0.1724	0.1442	-0.0842
widthR_underb_bust	-0.0173	0.1690	-0.0948	0.0896	-0.1500
widthR_bp2bp_bust	0.0459	-0.1122	-0.3238	-0.0120	0.1110
angle_pointing	-0.0052	0.1251	-0.3403	-0.1504	-0.0689
angle_pointing_rt	-0.0024	0.1630	-0.2595	-0.2872	0.1204
depth_width_ratio	0.0358	-0.1280	-0.2353	-0.1701	-0.0791
heightR_armp2BP_shd2BP	0.2703	0.0537	0.0184	0.0594	0.0262
areaR_fan_rec_inner	0.0039	-0.0666	-0.1377	-0.0462	0.2486

areaR_innerfan_rt_lt	-0.0118	0.0946	0.0037	-0.2809	0.3673
areaR_fan_rec_outer	0.0459	0.0200	0.2713	-0.0413	0.0019
areaR_outerfan_rt_lt	-0.0194	0.1177	-0.0251	-0.2972	0.3282
areaR_fullarc_rec	-0.0169	-0.1263	-0.1978	0.1949	0.1769
areaR_fullarc_rt_lt	-0.0170	0.1204	-0.0119	-0.3111	0.3596
	PC6	**PC7**	**PC8**	**PC9**	**PC10**
circmR_deltaB_bust	-0.0393	0.1350	-0.0931	0.2250	-0.2032
thickR_armp_bust	-0.0392	0.0732	0.1455	0.2874	-0.0325
thickR_underb_bust	0.0188	0.0228	-0.2947	0.0506	0.0533
thickR_front_armp	0.2486	0.0977	0.1100	-0.0116	0.1536
thickR_front_bust	0.3030	0.0408	0.1386	0.0249	-0.0412
thickR_front_underb	0.3165	-0.0767	0.0690	0.0949	-0.2786
thickR_front_underb2Broot_Broot	-0.1322	0.0955	-0.0396	-0.0453	-0.0423
heightR_upperB_fullB	0.0292	0.1938	0.0316	0.1620	0.0526
heightR_underb2Broot_lowerB	0.1629	0.3649	-0.2603	-0.0919	-0.2196
angle_upperB	-0.0874	0.2721	0.4071	0.1563	0.2037
angle_lowerB	0.0447	0.2292	0.0643	0.0866	0.3979
angle_lower_diff	-0.1434	-0.1664	-0.1174	0.1673	-0.0106
angle_tri_top	0.0027	-0.0449	-0.2751	-0.0174	0.1607
angle_tri_side	-0.0019	0.0612	0.2336	-0.0999	-0.1895
distR_BL_BDperp	0.3778	-0.2002	0.0139	0.1404	0.0208
distR_BL_BO	0.2854	-0.1071	0.0761	0.0891	0.0691
distR_depth_height	-0.0092	-0.0577	-0.1866	0.1251	0.1435
distR_upperB_lowerB	0.0332	0.1301	-0.0188	0.1346	-0.0352
distR_upperB_lowerBroot	-0.0010	0.0053	0.1007	0.1158	-0.0045
areaR_curv_rec_upper	-0.0195	-0.0911	0.0028	-0.1785	-0.1264
areaR_curv_rec_lower	-0.0806	-0.4852	-0.0174	0.0491	0.0215
areaR_curvUp_curvLow	-0.0260	0.1105	0.1353	0.2197	0.1507
areaR_tri_rec	-0.1010	0.0467	0.0408	-0.0484	-0.0190
areaR_tri_trap_upper	-0.0086	-0.1057	-0.0786	-0.1067	-0.1834
areaR_tri_trap_lower	-0.1434	0.1540	0.0964	-0.0936	0.0821
areaR_fanUp_fanLow	0.0634	-0.0286	-0.0564	0.0970	-0.0901
areaR_arc_tri_upper	0.0333	0.0215	-0.3053	0.4091	-0.1780
areaR_arc_tri_lower	-0.2703	-0.3161	0.1105	-0.0740	0.1190
areaR_underb_bust	0.1336	-0.0184	-0.2775	0.0006	0.2710
widthR_underb_bust	0.1086	-0.1188	-0.1040	0.0555	0.5045
widthR_bp2bp_bust	0.1777	-0.1041	0.1498	0.2034	-0.0042
angle_pointing	-0.1051	-0.0875	0.2237	0.1355	-0.0733
angle_pointing_rt	0.0015	0.0251	0.1513	0.0517	-0.0341
depth_width_ratio	0.3123	-0.1721	0.1264	0.1764	-0.0196

heightR_armp2BP_shd2BP	0.0111	0.0397	0.1509	-0.1762	0.0128
areaR_fan_rec_inner	0.0004	0.1122	-0.0319	0.0735	-0.1269
areaR_innerfan_rt_lt	0.1081	0.1834	-0.0965	-0.0741	-0.0293
areaR_fan_rec_outer	-0.2507	0.0731	-0.1125	0.3670	-0.0055
areaR_outerfan_rt_lt	0.0606	0.1179	-0.1016	-0.1308	0.1308
areaR_fullarc_rec	0.2669	-0.0186	0.0214	-0.2851	0.0149
areaR_fullarc_rt_lt	0.0795	0.1494	-0.1080	-0.1176	0.0847
	PC11	**PC12**	**PC13**	**PC14**	**PC15**
circmR_deltaB_bust	0.0731	-0.1167	0.0436	-0.7686	0.1966
thickR_armp_bust	0.0098	-0.0524	0.0772	0.1069	0.5167
thickR_underb_bust	0.1693	0.0056	-0.2991	-0.0452	-0.1912
thickR_front_armp	0.0563	-0.0744	-0.0866	0.0518	-0.2527
thickR_front_bust	0.0408	-0.1753	0.0425	0.1624	-0.0603
thickR_front_underb	0.0654	-0.2414	0.1363	0.1563	0.1654
thickR_front_underb2Broot_Broot	0.1826	-0.1210	-0.0525	0.0489	-0.0146
heightR_upperB_fullB	0.0551	0.0563	-0.0686	-0.0503	0.0643
heightR_underb2Broot_lowerB	0.0110	-0.1859	0.0359	-0.1055	-0.0157
angle_upperB	0.1217	0.0631	-0.1418	-0.1112	-0.0753
angle_lowerB	-0.1361	0.1625	0.0236	-0.0426	0.0384
angle_lower_diff	0.2250	-0.0422	-0.1765	0.0812	0.1293
angle_tri_top	-0.0317	-0.0346	-0.0145	0.1136	0.1006
angle_tri_side	0.0276	-0.0590	0.0284	-0.1029	-0.1460
distR_BL_BDperp	0.0856	0.0018	-0.3163	-0.0261	0.0843
distR_BL_BO	0.1393	-0.0390	-0.3401	0.0180	0.0012
distR_depth_height	-0.0182	0.0760	0.0114	0.1203	0.1349
distR_upperB_lowerB	0.0675	0.0247	-0.0544	-0.0307	0.0542
distR_upperB_lowerBroot	0.0177	0.0884	-0.0207	-0.0264	0.0289
areaR_curv_rec_upper	-0.0360	-0.1381	0.0303	0.0398	0.0926
areaR_curv_rec_lower	-0.0098	0.0012	-0.0767	-0.2399	-0.0144
areaR_curvUp_curvLow	0.0539	0.1455	-0.0773	-0.0102	0.0722
areaR_tri_rec	0.0427	-0.0185	0.0424	0.0686	-0.0402
areaR_tri_trap_upper	-0.0316	-0.0331	0.0927	0.0360	-0.0212
areaR_tri_trap_lower	0.0741	-0.0642	-0.0104	0.0690	-0.0834
areaR_fanUp_fanLow	-0.0266	0.1189	0.0389	-0.0202	0.0180
areaR_arc_tri_upper	0.0360	0.4685	0.2735	0.0720	-0.3854
areaR_arc_tri_lower	0.0863	-0.0936	-0.0797	-0.1515	-0.1164
areaR_underb_bust	0.0009	-0.0860	-0.1048	-0.0603	-0.0949
widthR_underb_bust	-0.2528	-0.2957	0.4085	-0.1626	0.0513
widthR_bp2bp_bust	-0.1200	-0.0881	0.2567	-0.0402	-0.1908
angle_pointing	-0.1157	-0.0006	0.1122	0.1334	-0.0176

angle_pointing_rt	-0.0381	0.0134	0.1353	0.0699	-0.0420
depth_width_ratio	0.0985	-0.0882	0.0413	-0.0596	-0.1533
heightR_armp2BP_shd2BP	0.0924	-0.0856	0.0174	-0.1975	-0.3610
areaR_fan_rec_inner	-0.7260	-0.0731	-0.4180	0.0405	-0.0498
areaR_innerfan_rt_lt	-0.1141	-0.1477	-0.0483	-0.0699	-0.0384
areaR_fan_rec_outer	0.0209	-0.4678	0.0418	0.1348	-0.2373
areaR_outerfan_rt_lt	0.3006	-0.0453	0.1101	0.1090	0.1014
areaR_fullarc_rec	-0.0853	0.3508	0.1296	-0.1748	0.1035
areaR_fullarc_rt_lt	0.1759	-0.0830	0.0607	0.0473	0.0576
	PC16	**PC17**	**PC18**	**PC19**	**PC20**
circmR_deltaB_bust	-0.0793	0.2688	-0.2415	-0.0107	-0.0067
thickR_armp_bust	0.0839	-0.2002	0.1361	0.0174	-0.5035
thickR_underb_bust	-0.2274	0.0562	-0.1031	0.1084	-0.0694
thickR_front_armp	-0.1195	0.2631	-0.0241	0.0498	0.1409
thickR_front_bust	0.0191	0.2520	-0.0080	0.0409	-0.0837
thickR_front_underb	0.1436	0.1377	0.0518	0.0220	-0.0150
thickR_front_underb2Broot_Broot	0.1868	-0.0159	-0.0022	0.0790	0.0320
heightR_upperB_fullB	0.0234	-0.0166	0.1705	-0.0464	0.0855
heightR_underb2Broot_lowerB	0.0091	0.0056	0.5485	-0.1895	0.0917
angle_upperB	0.1225	-0.0483	0.0029	0.0054	0.1095
angle_lowerB	-0.1590	0.0241	0.2380	-0.1157	-0.0448
angle_lower_diff	0.0915	-0.0384	-0.2181	0.1711	0.0926
angle_tri_top	-0.1658	0.0212	0.0175	0.0135	-0.1293
angle_tri_side	0.1382	-0.0158	0.0057	-0.0237	0.0549
distR_BL_BDperp	0.0725	0.0504	0.0497	0.0347	-0.0740
distR_BL_BO	0.1923	0.0609	0.0390	0.0553	-0.0419
distR_depth_height	-0.1132	0.0348	0.0794	0.0522	-0.0619
distR_upperB_lowerB	0.0462	-0.0197	0.1160	-0.0308	0.0590
distR_upperB_lowerBroot	0.0509	-0.0052	-0.0276	0.0093	0.0787
areaR_curv_rec_upper	-0.0230	0.0250	-0.0097	0.0154	0.0049
areaR_curv_rec_lower	-0.0210	0.0037	0.3864	-0.1407	0.0196
areaR_curvUp_curvLow	0.0292	-0.0198	-0.0047	0.0223	0.0749
areaR_tri_rec	0.1450	0.0550	0.0464	0.0389	-0.0030
areaR_tri_trap_upper	0.0906	0.0360	-0.0218	-0.0010	0.0155
areaR_tri_trap_lower	0.1578	0.0309	-0.0160	0.0555	-0.0153
areaR_fanUp_fanLow	0.0310	-0.0278	-0.1029	-0.0163	0.0549
areaR_arc_tri_upper	0.3655	0.0817	0.0824	0.0656	-0.0959
areaR_arc_tri_lower	0.1988	0.0693	0.3956	-0.0646	-0.0011
areaR_underb_bust	-0.0374	0.1359	-0.0349	-0.1510	-0.2516
widthR_underb_bust	0.3889	0.0783	-0.0909	0.1984	0.1300

widthR_bp2bp_bust	-0.3190	-0.2104	-0.1002	-0.2064	0.0121
angle_pointing	-0.1404	0.3982	-0.0094	-0.0830	-0.0077
angle_pointing_rt	-0.2549	0.3626	0.1808	0.2549	-0.1396
depth_width_ratio	-0.0317	-0.4077	-0.0071	-0.0943	0.1663
heightR_armp2BP_shd2BP	0.1050	-0.0996	-0.0871	0.0269	-0.6904
areaR_fan_rec_inner	0.2606	0.0223	-0.1257	-0.2290	-0.0460
areaR_innerfan_rt_lt	-0.0327	-0.2415	0.1566	0.6064	0.0568
areaR_fan_rec_outer	-0.1503	-0.2308	-0.0055	-0.0954	-0.0289
areaR_outerfan_rt_lt	0.1603	0.0763	-0.1463	-0.4732	0.0099
areaR_fullarc_rec	-0.0349	-0.2169	-0.0293	-0.0046	-0.0657
areaR_fullarc_rt_lt	0.0971	-0.0278	-0.0390	-0.1189	0.0335
	PC21	**PC22**	**PC23**	**PC24**	**PC25**
circmR_deltaB_bust	0.0763	0.0282	-0.0430	-0.0437	-0.0101
thickR_armp_bust	0.1950	0.2120	-0.0035	0.0700	0.0607
thickR_underb_bust	0.1371	0.4855	0.0689	0.4058	-0.0035
thickR_front_armp	0.0461	-0.1078	-0.1635	0.0649	0.0960
thickR_front_bust	0.2002	0.2847	0.1386	0.0052	-0.0641
thickR_front_underb	-0.0922	-0.1717	0.0789	-0.0832	-0.0028
thickR_front_underb2Broot_Broot	-0.1248	0.0318	0.1050	-0.0024	0.1180
heightR_upperB_fullB	-0.0281	-0.0625	-0.0393	0.0096	0.0030
heightR_underb2Broot_lowerB	-0.1110	-0.0947	0.0037	0.1329	-0.0281
angle_upperB	-0.0679	-0.0298	-0.0614	-0.0547	0.0395
angle_lowerB	0.0534	0.0549	0.2336	0.0298	-0.0230
angle_lower_diff	-0.1623	-0.2488	0.0224	-0.0384	-0.1382
angle_tri_top	-0.0650	-0.1664	0.1976	0.0550	0.1267
angle_tri_side	0.0505	0.1198	-0.2075	-0.0401	0.0750
distR_BL_BDperp	-0.2865	0.0238	-0.0709	0.0770	0.3352
distR_BL_BO	-0.1629	0.1680	-0.0372	-0.1182	-0.1895
distR_depth_height	0.0812	-0.0882	-0.5707	0.1347	-0.0467
distR_upperB_lowerB	-0.0395	-0.0582	-0.0076	-0.0098	-0.0076
distR_upperB_lowerBroot	0.0306	0.0553	0.0578	-0.0572	-0.0101
areaR_curv_rec_upper	0.1463	0.1226	-0.1834	-0.0398	0.1109
areaR_curv_rec_lower	-0.0290	-0.0445	-0.0057	0.0573	0.1544
areaR_curvUp_curvLow	0.0138	-0.0361	0.0000	-0.0397	0.0446
areaR_tri_rec	0.1314	0.1647	-0.2954	-0.0834	0.1820
areaR_tri_trap_upper	0.0775	0.2121	0.0780	-0.0782	-0.0421
areaR_tri_trap_lower	0.0284	0.1296	0.1253	-0.1136	0.1990
areaR_fanUp_fanLow	-0.0705	0.0489	0.5061	-0.0370	0.0739
areaR_arc_tri_upper	0.0383	0.0282	-0.0539	-0.0251	0.0568
areaR_arc_tri_lower	0.0605	0.1695	0.1153	-0.0744	-0.0855

areaR_underb_bust	0.2692	-0.1070	-0.0128	-0.6805	0.0777
widthR_underb_bust	-0.0942	0.0726	-0.0153	0.2164	-0.0529
widthR_bp2bp_bust	-0.2557	0.1304	-0.0484	-0.0364	0.2815
angle_pointing	0.1234	-0.1105	0.0747	0.1061	0.3427
angle_pointing_rt	-0.2014	0.0261	-0.0040	-0.1010	-0.5038
depth_width_ratio	0.5152	-0.1578	0.0542	0.0907	-0.2856
heightR_armp2BP_shd2BP	-0.0983	-0.3099	0.0172	0.2373	-0.0237
areaR_fan_rec_inner	0.0054	-0.0051	-0.0313	0.0650	-0.1150
areaR_innerfan_rt_lt	0.1213	-0.0660	0.0556	-0.1171	0.2663
areaR_fan_rec_outer	-0.2673	0.1970	-0.0811	-0.1948	-0.0939
areaR_outerfan_rt_lt	-0.0430	0.0256	-0.0243	0.0985	-0.0399
areaR_fullarc_rec	-0.2777	0.2845	-0.1194	-0.2103	-0.0766
areaR_fullarc_rt_lt	0.0155	0.0103	-0.0093	0.0342	0.0644
	PC26	**PC27**	**PC28**	**PC29**	**PC30**
circmR_deltaB_bust	-0.0046	0.0187	0.0150	0.0072	0.0042
thickR_armp_bust	-0.1845	-0.0771	-0.0079	0.0291	0.0050
thickR_underb_bust	0.0810	0.0164	-0.0476	0.0210	-0.0180
thickR_front_armp	-0.4183	-0.0630	0.0309	0.1805	0.0315
thickR_front_bust	0.1136	0.0685	0.1010	-0.1786	-0.0139
thickR_front_underb	0.2309	0.0249	-0.1303	0.0889	-0.0111
thickR_front_underb2Broot_Broot	0.0357	0.1167	-0.4369	0.4541	-0.0315
heightR_upperB_fullB	-0.1217	-0.0266	0.0669	-0.0744	-0.0208
heightR_underb2Broot_lowerB	0.0081	-0.0495	0.1033	-0.0194	0.0411
angle_upperB	0.1252	0.1793	0.0028	-0.1658	-0.1342
angle_lowerB	0.2866	0.0504	-0.1687	0.3012	0.0442
angle_lower_diff	0.0866	-0.0144	0.1206	0.0172	0.0773
angle_tri_top	-0.3292	0.1624	0.0592	-0.2001	-0.3705
angle_tri_side	-0.0306	0.0518	-0.1469	0.1212	-0.0659
distR_BL_BDperp	-0.0454	0.2752	-0.0580	-0.1290	0.4595
distR_BL_BO	0.0778	-0.3451	0.1556	0.1675	-0.4764
distR_depth_height	0.2543	-0.0003	-0.0286	0.0613	0.1204
distR_upperB_lowerB	-0.1319	0.0130	0.1555	-0.2035	-0.1091
distR_upperB_lowerBroot	-0.1539	-0.0843	0.0150	-0.0224	-0.0542
areaR_curv_rec_upper	-0.4291	-0.1232	-0.1123	0.2629	0.0255
areaR_curv_rec_lower	-0.0236	0.2408	-0.1757	0.0470	-0.3681
areaR_curvUp_curvLow	-0.1662	-0.1118	-0.0692	0.1335	0.0276
areaR_tri_rec	0.1187	0.3117	-0.0959	-0.2442	-0.2330
areaR_tri_trap_upper	0.1512	0.0469	0.0341	-0.0828	-0.0694
areaR_tri_trap_lower	-0.1383	0.1267	-0.0117	-0.1997	0.1203
areaR_fanUp_fanLow	-0.0599	0.1024	-0.0723	0.0659	0.0520

areaR_arc_tri_upper	-0.0983	-0.0545	-0.0077	0.0463	0.0017
areaR_arc_tri_lower	-0.0348	-0.2602	0.1833	-0.0601	0.3620
areaR_underb_bust	0.0546	-0.0104	-0.0834	0.0039	0.0578
widthR_underb_bust	-0.0460	0.0241	0.0217	-0.0029	-0.0223
widthR_bp2bp_bust	0.0769	-0.4000	-0.2227	-0.2132	-0.0339
angle_pointing	0.1403	0.1030	0.4797	0.2999	-0.0524
angle_pointing_rt	-0.1689	0.1707	-0.2540	-0.0991	0.0489
depth_width_ratio	-0.0825	0.2597	0.0072	0.0572	0.0531
heightR_armp2BP_shd2BP	0.0303	-0.0057	0.0622	0.0422	-0.0247
areaR_fan_rec_inner	-0.0434	0.0505	-0.0634	-0.0036	0.0170
areaR_innerfan_rt_lt	0.0983	-0.0776	0.1118	0.0371	-0.0245
areaR_fan_rec_outer	-0.0588	0.2447	0.2378	0.1951	0.0313
areaR_outerfan_rt_lt	-0.0503	0.0398	0.0116	0.0049	0.0077
areaR_fullarc_rec	-0.0651	0.2783	0.3396	0.2350	0.0183
areaR_fullarc_rt_lt	0.0439	-0.0608	0.0232	0.0042	-0.0159
	PC31	**PC32**	**PC33**	**PC34**	**PC35**
circmR_deltaB_bust	0.0005	0.0086	0.0040	0.0047	0.0032
thickR_armp_bust	-0.1056	-0.0644	0.0019	0.0206	-0.0231
thickR_underb_bust	0.1825	0.1556	-0.1190	0.0406	0.0359
thickR_front_armp	-0.1309	-0.0545	0.0074	0.0613	-0.0643
thickR_front_bust	-0.2350	-0.2684	0.1787	-0.1457	0.0094
thickR_front_underb	0.3541	0.3414	-0.2030	0.1022	0.0405
thickR_front_underb2Broot_Broot	-0.0735	0.0678	0.3907	0.0532	-0.0719
heightR_upperB_fullB	0.0986	0.0468	-0.0049	-0.3870	-0.0718
heightR_underb2Broot_lowerB	-0.1419	-0.1956	-0.1838	0.1439	0.1445
angle_upperB	0.0966	-0.0677	-0.2570	0.1959	-0.3037
angle_lowerB	0.1684	-0.2820	0.0255	-0.0452	-0.0860
angle_lower_diff	-0.0709	-0.5406	-0.1221	-0.1057	-0.0040
angle_tri_top	0.2881	-0.0044	-0.0827	0.1562	-0.1641
angle_tri_side	0.0431	-0.0641	-0.0748	-0.0515	-0.2237
distR_BL_BDperp	0.1039	-0.1067	0.0945	0.0192	0.0039
distR_BL_BO	-0.1105	0.0782	-0.0980	-0.0211	-0.0062
distR_depth_height	-0.3156	0.2012	-0.1452	0.1664	-0.2468
distR_upperB_lowerB	0.0245	0.2187	0.5094	-0.1400	-0.3521
distR_upperB_lowerBroot	-0.0657	-0.1216	0.2231	0.7563	0.2148
areaR_curv_rec_upper	0.1635	-0.2373	-0.2333	-0.0688	-0.2213
areaR_curv_rec_lower	-0.2624	-0.0411	0.0058	-0.1255	0.1403
areaR_curvUp_curvLow	0.1658	0.1076	-0.1621	-0.1805	0.5062
areaR_tri_rec	0.1715	-0.1271	-0.1003	0.0603	0.1485
areaR_tri_trap_upper	0.1176	-0.2104	0.0232	-0.0299	-0.0966

areaR_tri_trap_lower	-0.3655	0.2472	-0.1885	-0.1045	0.2038
areaR_fanUp_fanLow	-0.3437	0.1082	-0.3900	0.0746	-0.3475
areaR_arc_tri_upper	0.0672	-0.0774	-0.0129	-0.0289	-0.0294
areaR_arc_tri_lower	0.2064	0.0995	-0.0332	0.1163	-0.1604
areaR_underb_bust	-0.0120	0.0182	0.0035	0.0207	-0.0143
widthR_underb_bust	0.0000	0.0057	-0.0037	-0.0131	0.0094
widthR_bp2bp_bust	-0.0223	-0.0796	0.0019	-0.0172	-0.0332
angle_pointing	0.0136	0.0406	0.0239	0.0117	0.0171
angle_pointing_rt	-0.0082	0.0143	-0.0285	-0.0072	0.0098
depth_width_ratio	0.0239	0.0454	-0.0069	0.0161	0.0158
heightR_armp2BP_shd2BP	-0.0017	-0.0347	-0.0221	0.0076	0.0199
areaR_fan_rec_inner	0.0071	0.0024	-0.0033	0.0040	0.0030
areaR_innerfan_rt_lt	-0.0047	-0.0115	0.0151	-0.0052	0.0052
areaR_fan_rec_outer	0.0183	0.0312	-0.0027	0.0107	0.0289
areaR_outerfan_rt_lt	-0.0054	-0.0161	-0.0199	-0.0083	0.0377
areaR_fullarc_rec	0.0196	0.0433	-0.0056	0.0154	0.0352
areaR_fullarc_rt_lt	0.0068	0.0113	0.0198	0.0085	-0.0477
	PC36	**PC37**	**PC38**	**PC39**	**PC40**
circmR_deltaB_bust	0.0028	0.0012	0.0001	0.0014	0.0026
thickR_armp_bust	-0.0536	-0.1094	-0.0069	0.0050	-0.1545
thickR_underb_bust	-0.1246	0.0622	-0.0199	0.0131	-0.0040
thickR_front_armp	-0.1343	-0.2186	-0.0177	0.0084	-0.3163
thickR_front_bust	0.3109	0.0470	0.0374	-0.0232	0.2506
thickR_front_underb	-0.2202	0.1275	-0.0235	0.0131	-0.0179
thickR_front_underb2Broot_Broot	0.2876	-0.0167	0.0014	-0.0052	-0.0740
heightR_upperB_fullB	0.1383	0.3051	0.4635	0.1158	-0.4442
heightR_underb2Broot_lowerB	0.0595	-0.0737	-0.0725	-0.0353	0.0778
angle_upperB	0.2417	0.2091	-0.3249	-0.0942	-0.0157
angle_lowerB	-0.3627	0.0446	0.0307	0.0417	0.0833
angle_lower_diff	-0.2460	-0.0353	0.0794	0.0165	0.0864
angle_tri_top	0.1849	0.0380	0.2324	0.0563	0.2331
angle_tri_side	-0.1299	-0.0822	0.5435	0.1919	0.4252
distR_BL_BDperp	0.0421	0.0234	0.0055	-0.0151	0.0267
distR_BL_BO	-0.0496	-0.0275	-0.0044	0.0202	-0.0200
distR_depth_height	0.1409	0.0715	0.0878	0.0536	0.1561
distR_upperB_lowerB	-0.3086	-0.1477	-0.2527	-0.0190	0.2680
distR_upperB_lowerBroot	-0.1033	0.2190	0.2026	0.0592	-0.0028
areaR_curv_rec_upper	-0.0490	0.2566	-0.3074	-0.1070	0.1139
areaR_curv_rec_lower	-0.0551	0.1349	-0.0852	-0.0373	0.0020
areaR_curvUp_curvLow	0.2961	-0.2090	-0.0490	-0.0491	0.3821

areaR_tri_rec	-0.1487	-0.4934	0.0184	-0.0106	-0.2310
areaR_tri_trap_upper	0.1470	0.1897	-0.0544	0.0269	-0.1571
areaR_tri_trap_lower	-0.3588	0.3582	0.0182	0.0027	0.1230
areaR_fanUp_fanLow	0.0587	-0.3400	0.0573	0.0113	-0.0638
areaR_arc_tri_upper	-0.0133	0.0990	-0.0716	-0.0246	0.0228
areaR_arc_tri_lower	0.0846	-0.1609	0.0786	0.0399	0.0046
areaR_underb_bust	0.0299	0.0041	0.0016	0.0001	-0.0074
widthR_underb_bust	-0.0088	-0.0032	0.0021	-0.0009	0.0043
widthR_bp2bp_bust	-0.0047	-0.0181	-0.0072	0.0270	-0.0034
angle_pointing	0.0235	0.0186	0.0154	-0.0115	0.0067
angle_pointing_rt	-0.0176	-0.0087	-0.0058	-0.0084	-0.0041
depth_width_ratio	-0.0048	0.0092	0.0035	-0.0175	-0.0019
heightR_armp2BP_shd2BP	0.0000	0.0064	0.0021	-0.0088	0.0032
areaR_fan_rec_inner	-0.0056	-0.0019	0.0019	-0.0073	-0.0006
areaR_innerfan_rt_lt	0.0343	0.0017	-0.0754	0.2702	-0.0014
areaR_fan_rec_outer	0.0011	0.0040	0.0105	-0.0174	0.0035
areaR_outerfan_rt_lt	0.0193	0.0028	-0.1578	0.5256	-0.0030
areaR_fullarc_rec	0.0063	0.0072	0.0135	-0.0236	0.0061
areaR_fullarc_rt_lt	-0.0412	0.0012	0.2209	-0.7458	0.0056

	PC41
circmR_deltaB_bust	0.0001
thickR_armp_bust	-0.1510
thickR_underb_bust	0.0063
thickR_front_armp	-0.3220
thickR_front_bust	0.2381
thickR_front_underb	0.0008
thickR_front_underb2Broot_Broot	0.0226
heightR_upperB_fullB	0.1704
heightR_underb2Broot_lowerB	-0.0361
angle_upperB	-0.0250
angle_lowerB	-0.0202
angle_lower_diff	-0.0272
angle_tri_top	-0.0603
angle_tri_side	-0.1364
distR_BL_BDperp	-0.0219
distR_BL_BO	0.0310
distR_depth_height	-0.0831
distR_upperB_lowerB	-0.0728
distR_upperB_lowerBroot	0.0787
areaR_curv_rec_upper	0.2688

areaR_curv_rec_lower	-0.0286
areaR_curvUp_curvLow	-0.2230
areaR_tri_rec	0.1841
areaR_tri_trap_upper	-0.7493
areaR_tri_trap_lower	-0.1102
areaR_fanUp_fanLow	0.0551
areaR_arc_tri_upper	0.0613
areaR_arc_tri_lower	0.0289
areaR_underb_bust	-0.0056
widthR_underb_bust	0.0021
widthR_bp2bp_bust	-0.0069
angle_pointing	0.0198
angle_pointing_rt	-0.0119
depth_width_ratio	-0.0042
heightR_armp2BP_shd2BP	-0.0032
areaR_fan_rec_inner	-0.0032
areaR_innerfan_rt_lt	0.0213
areaR_fan_rec_outer	0.0090
areaR_outerfan_rt_lt	0.0303
areaR_fullarc_rec	0.0142
areaR_fullarc_rt_lt	-0.0427

APPENDIX E. Univariate ANOVA for individual variables (3-Cluster Case)

#1. circmR_deltaB_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	77.85	38.924	47.894	$< 2.2\times10^{-16}$
Residuals	405	329.15	0.813		
#2. thickR_armp_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	157.49	78.747	127.82	$< 2.2\times10^{-16}$
Residuals	405	249.51	0.616		
#3. thickR_underb_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	56.29	28.143	32.5	8.127×10^{-14}
Residuals	405	350.71	0.866		
#4. thickR_front_armp					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	87.6	43.801	55.539	$< 2.2\times10^{-16}$
Residuals	405	319.4	0.789		
#5. thickR_front_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	43.49	21.7465	24.229	1.151×10^{-10}
Residuals	405	363.51	0.8975		
#6. thickR_front_underb					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	11.56	5.7794	5.9191	0.002926

Residuals	405	395.44	0.9764		
#7. thickR_front_underb2Broot_Broot					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	69.71	34.856	41.854	$< 2.2\times10^{-16}$
Residuals	405	337.29	0.833		
#8. heightR_upperB_fullB					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	217.05	108.523	231.38	$< 2.2\times10^{-16}$
Residuals	405	189.95	0.469		
#9. heightR_underb2Broot_lowerB					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	6.47	3.235	3.2711	0.03897
Residuals	405	400.53	0.989		
#10. angle_upperB					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	99.444	49.722	65.475	$< 2.2\times10^{-16}$
Residuals	405	307.556	0.759		
#11. angle_lowerB					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	99.969	49.985	65.934	$< 2.2\times10^{-16}$
Residuals	405	307.031	0.758		
#12. angle_lower_diff					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	7.05	3.5247	3.5692	0.02907
Residuals	405	399.95	0.9875		
#13. angle_tri_top					
	Df	Sum of	Mean	F-value	P-value

		squares	squares		
3-Cluster Case	2	177.02	88.512	155.87	$< 2.2\times10^{-16}$
Residuals	405	229.98	0.568		

#14. angle_tri_side

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	196.83	98.415	189.65	$< 2.2\times10^{-16}$
Residuals	405	210.17	0.519		

#15. distR_BL_BDperp

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	19.63	9.8157	10.262	4.491×10^{-5}
Residuals	405	387.37	0.9565		

#16. distR_BL_BO

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	5.23	2.6149	2.636	**0.07288**
Residuals	405	401.77	0.9920		

#17. distR_depth_height

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	220.13	110.067	238.55	$< 2.2\times10^{-16}$
Residuals	405	186.87	0.461		

#18. distR_upperB_lowerB

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	215.35	107.674	227.54	$< 2.2\times10^{-16}$
Residuals	405	191.65	0.473		

#19. distR_upperB_lowerBroot

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	242.87	121.435	299.65	$< 2.2\times10^{-16}$

Residuals	405	164.13	0.405		
#20. areaR_curv_rec_upper					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	237.28	118.639	283.1	$< 2.2\times10^{-16}$
Residuals	405	169.72	0.419		
#21. areaR_curv_rec_lower					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	24.71	12.3530	13.087	3.109×10^{-6}
Residuals	405	382.29	0.9439		
#22. areaR_curvUp_curvLow					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	222.15	111.073	243.35	$< 2.2\times10^{-16}$
Residuals	405	184.85	0.456		
#23. areaR_tri_rec					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	235.5	117.748	278.06	$< 2.2\times10^{-16}$
Residuals	405	171.5	0.423		
#24. areaR_tri_trap_upper					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	224.4	112.198	248.85	$< 2.2\times10^{-16}$
Residuals	405	182.6	0.451		
#25. areaR_tri_trap_lower					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	132.15	66.075	97.364	$< 2.2\times10^{-16}$
Residuals	405	274.85	0.679		
#26. areaR_fanUp_fanLow					
	Df	Sum of	Mean	F-value	P-value

		squares	squares		
3-Cluster Case	2	232.15	116.074	268.86	$< 2.2\times10^{-16}$
Residuals	405	174.85	0.432		

#27. areaR_arc_tri_upper

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	91.988	45.994	59.133	$< 2.2\times10^{-16}$
Residuals	405	315.012	0.778		

#28. areaR_arc_tri_lower

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	73.64	36.819	44.732	$< 2.2\times10^{-16}$
Residuals	405	333.36	0.823		

#29. areaR_underb_bust

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	96.392	48.196	62.843	$< 2.2\times10^{-16}$
Residuals	405	310.608	0.767		

#30. widthR_underb_bust

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	45.7	22.848	25.611	3.362×10^{-11}
Residuals	405	361.3	0.8921		

#31. widthR_bp2bp_bust

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	13.65	6.825	7.027	9.997×10^{-4}
Residuals	405	393.35	0.9712		

#32. angle_pointing

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	26.59	13.293	14.152	1.145×10^{-6}

Residuals	405	380.41	0.9393		
#33. angle_pointing_rt					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	40.94	20.472	22.649	4.739×10^{-10}
Residuals	405	366.06	0.9038		
#34. depth_width_ratio					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	7.76	3.881	3.937	0.02026
Residuals	405	399.24	0.9858		
#35. heightR_armp2BP_shd2BP					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	212.22	106.110	220.63	$< 2.2\times10^{-16}$
Residuals	405	194.78	0.481		
#36. areaR_fan_rec_inner					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	2.1	1.0482	1.0485	**0.3514**
Residuals	405	404.9	0.9998		
#37. areaR_innerfan_rt_lt					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	16.95	8.4744	8.7991	1.817×10^{-4}
Residuals	405	390.05	0.9631		
#38. areaR_fan_rec_outer					
	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	5.4	2.7003	2.7232	**0.06687**
Residuals	405	401.6	0.9916		
#39 areaR_outerfan_rt_lt					
	Df	Sum of	Mean	F-value	P-value

		squares	squares		
3-Cluster Case	2	20.42	10.2111	10.698	2.969×10^{-5}
Residuals	405	386.58	0.9545		

#40. areaR_fullarc_rec

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	18.49	9.2446	9.637	8.152×10^{-5}
Residuals	405	388.51	0.9593		

#41. areaR_fullarc_rt_lt

	Df	Sum of squares	Mean squares	F-value	P-value
3-Cluster Case	2	22.41	11.2071	11.802	1.043×10^{-5}
Residuals	405	384.59	0.9496		

APPENDIX F. Univariate ANOVA for individual variables (5-Cluster Case)

#1. circmR_deltaB_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	98.991	24.7478	32.38	$< 2.2\times10^{-16}$
Residuals	403	308.009	0.7643		
#2. thickR_armp_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	201.5	50.376	98.793	$< 2.2\times10^{-16}$
Residuals	403	205.5	0.510		
#3. thickR_underb_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	117.26	29.314	40.773	$< 2.2\times10^{-16}$
Residuals	403	289.74	0.719		
#4. thickR_front_armp					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	79.29	19.8235	24.378	$< 2.2\times10^{-16}$
Residuals	403	327.71	0.8132		
#5. thickR_front_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	74.9	18.7247	22.722	$< 2.2\times10^{-16}$
Residuals	403	332.1	0.8241		
#6. thickR_front_underb					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	22.18	5.5449	5.8069	1.495×10^{-4}

Residuals	403	384.82	0.9549		
#7. thickR_front_underb2Broot_Broot					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	191.06	47.764	89.138	$< 2.2\times10^{-16}$
Residuals	403	215.94	0.536		
#8. heightR_upperB_fullB					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	242.6	60.651	148.68	$< 2.2\times10^{-16}$
Residuals	403	164.4	0.408		
#9. heightR_underb2Broot_lowerB					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	77.39	19.3466	23.654	$< 2.2\times10^{-16}$
Residuals	403	329.61	0.8179		
#10. angle_upperB					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	127.51	31.879	45.967	$< 2.2\times10^{-16}$
Residuals	403	279.49	0.694		
#11. angle_lowerB					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	149.77	37.443	58.662	$< 2.2\times10^{-16}$
Residuals	403	257.23	0.638		
#12. angle_lower_diff					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	105.35	26.3381	35.188	$< 2.2\times10^{-16}$
Residuals	403	301.65	0.7485		
#13. angle_tri_top					
	Df	Sum of	Mean	F-value	P-value

		squares	squares		
5-Cluster Case	4	202.88	50.721	100.14	$< 2.2\times10^{-16}$
Residuals	403	204.12	0.506		
#14. angle_tri_side					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	225.61	56.402	125.31	$< 2.2\times10^{-16}$
Residuals	403	181.39	0.450		
#15. distR_BL_BDperp					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	36.66	9.1645	9.9727	1.053×10^{-7}
Residuals	403	370.34	0.9190		
#16. distR_BL_BO					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	42.02	10.5038	11.598	6.362×10^{-9}
Residuals	403	364.98	0.9057		
#17. distR_depth_height					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	243.03	60.758	149.33	$< 2.2\times10^{-16}$
Residuals	403	163.97	0.407		
#18. distR_upperB_lowerB					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	244.08	61.019	150.93	$< 2.2\times10^{-16}$
Residuals	403	162.92	0.404		
#19. distR_upperB_lowerBroot					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	265.99	66.498	190.05	$< 2.2\times10^{-16}$

Residuals	403	141.01	0.350		
#20. areaR_curv_rec_upper					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	245.1	61.276	152.53	$< 2.2\times10^{-16}$
Residuals	403	161.9	0.402		
#21. areaR_curv_rec_lower					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	54.09	13.5220	15.441	9.23×10^{-12}
Residuals	403	352.91	0.8757		
#22. areaR_curvUp_curvLow					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	237.45	59.363	141.1	$< 2.2\times10^{-16}$
Residuals	403	169.55	0.421		
#23. areaR_tri_rec					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	243.34	60.835	149.8	$< 2.2\times10^{-16}$
Residuals	403	163.66	0.406		
#24. areaR_tri_trap_upper					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	243.97	60.993	150.77	$< 2.2\times10^{-16}$
Residuals	403	163.03	0.405		
#25. areaR_tri_trap_lower					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	178.7	44.675	78.862	$< 2.2\times10^{-16}$
Residuals	403	228.3	0.567		
#26. areaR_fanUp_fanLow					
	Df	Sum of	Mean	F-value	P-value

		squares	squares		
5-Cluster Case	4	263.46	65.865	184.92	$< 2.2\times10^{-16}$
Residuals	403	143.54	0.356		
#27. areaR_arc_tri_upper					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	88.51	22.1271	27.998	$< 2.2\times10^{-16}$
Residuals	403	318.49	0.7903		
#28. areaR_arc_tri_lower					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	88.89	22.2227	28.153	$< 2.2\times10^{-16}$
Residuals	403	318.11	0.7894		
#29. areaR_underb_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	151.83	37.957	59.946	$< 2.2\times10^{-16}$
Residuals	403	255.17	0.633		
#30. widthR_underb_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	62.02	15.504	18.111	1.082×10^{-13}
Residuals	403	344.98	0.856		
#31. widthR_bp2bp_bust					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	121.3	30.3246	42.775	$< 2.2\times10^{-16}$
Residuals	403	285.7	0.7089		
#32. angle_pointing					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	108.62	27.1540	36.675	$< 2.2\times10^{-16}$

Residuals	403	298.38	0.7404		
#33. angle_pointing_rt					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	74.41	18.6019	22.54	$< 2.2\times10^{-16}$
Residuals	403	332.59	0.8253		
#34. depth_width_ratio					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	91.339	22.8346	29.153	$< 2.2\times10^{-16}$
Residuals	403	315.661	0.7833		
#35. heightR_armp2BP_shd2BP					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	216.48	54.121	114.48	$< 2.2\times10^{-16}$
Residuals	403	190.52	0.473		
#36. areaR_fan_rec_inner					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	16.8	4.1998	4.3376	0.001908
Residuals	403	390.2	0.9682		
#37. areaR_innerfan_rt_lt					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	25.52	6.3796	6.7394	2.94×10^{-5}
Residuals	403	381.48	0.9466		
#38. areaR_fan_rec_outer					
	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	114.33	28.5812	39.355	$< 2.2\times10^{-16}$
Residuals	403	292.68	0.7262		
#39 areaR_outerfan_rt_lt					
	Df	Sum of	Mean	F-value	P-value

		squares	squares		
5-Cluster Case	4	23.79	5.9471	6.2542	6.856×10^{-5}
Residuals	403	383.21	0.9509		

#40. areaR_fullarc_rec

	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	93.146	23.2866	29.901	$< 2.2\times10^{-16}$
Residuals	403	313.854	0.7788		

#41. areaR_fullarc_rt_lt

	Df	Sum of squares	Mean squares	F-value	P-value
5-Cluster Case	4	28.96	7.2400	7.718	5.326×10^{-6}
Residuals	403	378.04	0.9381		

REFERENCES

Azouz, Z. B., Rioux, M., Shu, C., & Lepage, R. (2006). Characterizing human shape variation using 3D anthropometric data. *The Visual Computer, 22*(5), 302-314.

Ashdown, S. P., & Na, H. (2008). Comparison of 3-D body scan data to quantify upper-body postural variation in older and younger women. *Clothing and Textiles Research Journal, 26*(4), 292-307.

Brown, T. L. H., Ringrose, C., Hyland, R. E., Cole, A. A., & Brotherston, T. M. (1999). A method of assessing female breast morphometry and its clinical application. *British Journal of Plastic Surgery*, *52*(5), 355-359.

Bratabase. (n.d.). Determining your breast shape. Retrieved December 2013 from http://www.bratabase.com

Connell, L. J., Ulrich, P. V., Brannon, E. L., Alexander, M., & Presley, A. B. (2006). Body shape assessment scale: Instrument development foranalyzing female figures. *Clothing and Textiles Research Journal*, *24*(2), 80-95.

Catanuto, G., Spano, A., Pennati, A., Riggio, E., Farinella, G. M., Impoco, G., ... & Nava, M. B. (2008). Experimental methodology for digital breast shape analysis and objective surgical outcome evaluation. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *61*(3), 314-318.

Chen, C. M., LaBat, K., & Bye, E. (2010). Physical characteristics related to bra fit. *Ergonomics*, *53*(4), 514-524.

Chen, C. M., LaBat, K., & Bye, E. (2011). Bust prominence related to bra fit problems. *International Journal of Consumer Studies, 35*(6), 695-701.

Douty, H. I. (1968). Silhouette photography for the study of visual somatometry and body image. In *National Textile and Clothing Meeting, Minneapolis, Minnesota*.

Farnworth, B., & Dolhan, P. A. (1985). Heat and water transport through cotton and polypropylene underwear. *Textile Research Journal*, *55*(10), 627-630.

Feather, B. L., Ford, S., & Herr, D. G. (1996). Female collegiate basketball players' perceptions about their bodies, garment fit and uniform design preferences. *Clothing and Textiles Research Journal, 14*(1), 22-29.

Farrell-Beck, J., Poresky, L., Paff, J., & Moon, C. (1998). Brassieres and women's health from 1863 to 1940. *Clothing and Textiles Research Journal, 16*(3), 105-115.

Fanelli, M. (2001). *U.S. Patent No. 6,234,867*. Washington, DC: U.S. Patent and Trademark Office.

Farinella, G. M., Impoco, G., Gallo, G., Spoto, S., & Catanuto, G. (2006, February). Unambiguous analysis of woman breast shape for plastic surgery outcome evaluation. Paper presented at *Eurographics Italian Chapter Conference* (pp. 255-261).

Goldsberry, R. E., Shim, S., & Reich, N. (1996). Women 55 years and older: Part II overall satisfaction and dissatisfaction with the ready-to-wear. *Clothing and Textiles Research Journal*, *14*(2), 121-132.

Greenbaum, A.R., Heslop, T., Morris, J., & Dunn, K.W., (2003). An investigation of the suitability of bra fit in women referred for reduction mammaplasty. *British Journal of Plastic Surgery, 56*(3), 230–236.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108.

Henss, R. (1995). Waist-to-hip ratio and attractiveness. Replication and extension. *Personality and Individual Differences, 19(4)*, 479-488.

Hardaker, C. H. M., & Fozzard, G. J. W. (1997). The bra design process-a study of professional practice. *International Journal of Clothing Science and Technology*, *9*(4), 311-325.

Hadi, M. S. (2000). Sports brassiere: is it a solution for mastalgia? *The breast journal*, *6*(6), 407-409.

Hart, C., & Dewsnap, B. (2001). An exploratory study of the consumer decision process for intimate apparel. *Journal of Fashion Marketing and Management: An International Journal*, *5*(2), 108-119.

Hsia, H. C., & Thomson, J. G. (2003). Differences in breast shape preferences between plastic surgeons and patients seeking breast augmentation. *Plastic and Reconstructive Surgery*, *112*(1), 312-320.

Haars, G., van Noord, P. A., van Gils, C. H., Grobbee, D. E., & Peeters, P. H. (2005). Measurements of breast density: no ratio for a ratio. *Cancer Epidemiology Biomarkers & Prevention*, *14*(11), 2634-2640.

Hasler, N., Stoll, C., Rosenhahn, B., Thormählen, T., & Seidel, H. P. (2009). Estimating body shape of dressed humans. *Computers and Graphics*, *33*(3), 211-216.

Han, H., & Nam, Y. (2011). Automatic body landmark identification for various body figures. *International Journal of Industrial Ergonomics, 41*(6), 592-606.

Hume, M., & Mills, M. (2013). Uncovering Victoria's Secret: Exploring women's luxury perceptions of intimate apparel and purchasing behaviour. *Journal of Fashion Marketing and Management: An International Journal*, *17*(4), 460-485.

Herroom. (n.d.). Classify your breasts. Retrieved May 2016 from http://www.herroom.com

Jones, V. (1972). The relationship of body-image, anxiety, and achievement of female high school students. (Unpublished master's thesis). Auburn University, Auburn, AL, United States.

Jolliffe, I. (2002). *Principal component analysis*. New York, NY, United States: John Wiley & Sons.

Kaye, B. L. (1972). Neurologic changes with excessively large breasts. *South Medical Journal*, *65*(2), 177-180

Kemberling, S. R. (1979). Supporting breast-feeding. *Pediatrics*, *63*(1), 60-63.

Kovacs, L., Eder, M., Hollweck, R., Zimmermann, A., Settles, M., Schneider, A., ... & Biemer, E. (2007). Comparison between breast volume measurement using 3D surface imaging and classical techniques. *The Breast*, *16*(2), 137-145.

Kaufman, L., & Rousseeuw, P. J. (2009). Partitioning around medoids (Program PAM). In *Finding groups in data: an introduction to cluster analysis*. New York, NY, United States: John Wiley & Sons.

Lorentzen, D. & Lawson, L. (1987). Selected sports bras: A biomechanical analysis of breast motion. *Physician and Sports Medicine 15*(5), 128-139.

LaBat, K. (1987). Consumer satisfaction/dissatisfaction with the fit of ready-to-wear clothing. (Unpublished doctoral dissertation). University of Minnesota, Saint Paul, MN, United States.

Lee, H. Y., Hong, K., & Kim, E. A. (2004). Measurement protocol of women's nude breasts using a 3D scanning technique. *Applied Ergonomics*, *35*(4), 353-359.

Lee, M. M., Chang, I. Y. H., Horng, C. F., Chang, J. S., Cheng, S. H., & Huang, A. (2005). Breast cancer and dietary factors in Taiwanese women. *Cancer Causes & Control*, *16*(8), 929-937.

Lee, H. Y., & Hong, K. (2007). Optimal brassiere wire based on the 3D anthropometric measurements of under breast curve. *Applied Ergonomics, 38*(3), 377-384.

Lu, J. M., & Wang, M. J. J. (2008). Automated anthropometric data collection using 3D whole body scanners. *Expert Systems with Applications, 35*(1), 407-414.

Lawrence, R. A., & Lawrence, R. M. (2010). *Breastfeeding: a guide for the medical professional*. Philadelphia, PA, United States: Elsevier Health Sciences.

Law, D., Wong, C., & Yip, J. (2012). How does visual merchandising affect consumer affective response? An intimate apparel experience. *European Journal of Marketing*, *46*(1/2), 112-133.

Lau, W. F. (2014). *Development of seamless knitted bra for optimum fit*. (Doctoral dissertation, Hong Kong Polytechnic University, Hong Kong, China). Retrieved from http://hdl.handle.net/10397/6859

Mosimann, J. E. (1988). Size and shape analysis. In *Encyclopedia of Statistical Sciences*. New York, NY, United States: John Wiley & Sons.

McCulloch, C. E., Paal, B., & Ashdown, S. P. (1998). An optimization approach to apparel sizing. *Journal of the Operational Research Society, 49*(5), 492-499.

Mason, B. R., Page, K. A., & Fallon, K. (1999). An analysis of movement and discomfort of the female breast during exercise and the effects of breast support in three cases. *Journal of Science and Medicine in Sport*, *2*(2), 134-144.

McGhee, D.E., & Steele, J.R. (2006). How do respiratory state and measurement method affect bra size calculations? *British Journal of Sports Medicine, 40*(12), 970–974.

McGhee, D.E., & Steele, J.R. (2010). Optimising breast support in female patients through correct bra fit: a cross-sectional study. *Journal of Science and Medicine in Sport, 13* (6), 568–572.

Nethero, S. (2007). *U.S. Patent Application 11/829,568*.

Oh, S., & Chun, J. (2014). New breast measurement technique and bra sizing system based on 3D body scan data. *Journal of the Ergonomics Society of Korea*, *33*(4), 299–311.

Pechter, E.A. (1998). A new method for determining bra size and predicting postaugmentation breast size. *Plastic and Reconstructive Surgery, 102* (4), 1259–1265.

Page, K. A., & Steele, J. R. (1999). Breast motion and sports brassiere design. *Sports Medicine*, *27*(4), 205–211.

Pandarum, R., Yu, W., & Hunter, L. (2011). 3-D breast anthropometry of plus-sized women in South Africa. *Ergonomics*, *54*(9), 866-875.

Qiao, Q., Zhou, G., & Ling, Y. C. (1997). Breast volume measurement in young Chinese women and clinical applications. *Aesthetic plastic surgery*, *21*(5), 362-368.

Ryan, E. L. (2000). Pectoral girdle myalgia in women: a 5-year study in a clinical setting. *The Clinical Journal of Pain*, *16*(4), 298-303.

Rencher, A. C. (2003a). Cluster analysis. In *Methods of multivariate analysis* (pp. 451-503). New York, NY, United States: John Wiley & Sons.

Rencher, A. C. (2003b). Discriminant analysis: description of group separation. In *Methods of multivariate analysis* (pp. 270-298). New York, NY, United States: John Wiley & Sons.

Rencher, A. C. (2003c). Classification analysis: allocation of observations to groups. In *Methods of multivariate analysis* (pp. 299-321). New York, NY, United States: John Wiley & Sons.

Rencher, A. C. (2003d). Multivariate analysis of variance. In *Methods of multivariate analysis* (pp. 156-184). New York, NY, United States: John Wiley & Sons.

Sheldon, W. H., Stevens, S. S., & Tucker, W. B. (1940). *The varieties of human physique*. Oxford, England: Harper.

Stunkard, A. J., Sørensen, T., & Schulsinger, F. (1982). Use of the Danish Adoption Register for the study of obesity and thinness. *Research publications-Association for Research in Nervous and Mental Disease*, *60*(1), 115-120.

Singh, D. (1993). Adaptive significance of female physical attractiveness: role of waist-to-hip ratio. *Journal of Personality and Social Psychology*, *65*(2), 293.

Singer, S. R., & Grismaijer, S. (1995). Dressed to kill. The link between breast cancer & bras. *Journal of Applied Nutrition*, *47*(3), 90-92.

Soares, D., Reid, M., & James, M. (2002). Age as a predictive factor of mammographic breast density in Jamaican women. *Clinical Radiology*, *57*(6), 472-476.

Simmons, K., Istook, C. L., & Devarajan, P. (2004) Female figure identification technique (FFIT) for apparel part ii: development of shape sorting software. *Journal of Textile and Apparel, Technology and Manangement, 4*(1), 1-15

Sook Cho, Y., Komatsu, T., Takatera, M., Inui, S., Shimizu, Y., & Park, H. (2006). Posture and depth adjustable 3D body model for individual pattern making. *International Journal of Clothing Science and Technology*, *18*(2), 96-107.

Swami, V., Salem, N., Furnham, A., & Tovée, M. J. (2008). Initial examination of the validity and reliability of the female Photographic Figure Rating Scale for body image assessment. *Personality and Individual Differences, 44*(8), 1752-1761.

Shin, K. (2009). The origins and evolution of the bra. (Doctoral dissertation, University of Northumbria, Newcastle, United Kingdom).

Retrieved from http://nrl.northumbria.ac.uk/id/eprint/3040.

Small, K. H., Tepper, O. M., Unger, J. G., Kumar, N., Feldman, D. L., Choi, M., & Karp, N. S. (2010). Re-defining pseudoptosis from a 3D perspective after short scar-medial pedicle reduction mammaplasty. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *63*(2), 346-353.

Song, H.K., & Ashdown, S.P. (2011). Categorization of lower body shapes for adult females based on multiple view analysis. *Textile Research Journal, 81*(9), 914–931.

Wright, M.C.M., (2002). Graphical analysis of bra size calculation procedures. *International Journal of clothing Science and Technology, 14* (1), 41–45.

Wood, K., Cameron, M., & Fitzgerald, K. (2008). Breast size, bra fit and thoracic pain in young women: a correlational study. *Chiropractic and Osteopathy, 16* (1), 1–7.

White, J., & Scurr, J. (2012). Evaluation of professional bra fitting criteria for bra selection and fitting in the UK. *Ergonomics*, *55*(6), 704-711.

Yu, W., & So, Y. K. (2001). Special techniques for making push-up bra. *ATA Journal*, *12*(3), 69-70.

Yu, W. (2011). Achieving comfort in intimate apparel. In *Improving comfort in clothing* (pp. 427-448). Sawston, Cambridge, United Kindom: Woodhead Publishing.

Zheng, R., Yu, W., & Fan, J. (2007). Development of a new Chinese bra sizing system based on breast anthropometric measurements. *International Journal of Industrial Ergonomics*, *37*(8), 697-705.

Zheng, R., Yu, W., & Fan, J. (2009). Pressure evaluation of 3D seamless knitted bras and conventional wired bras. *Fibers and Polymers*, *10*(1), 124-131.