Original article

Female breast shape categorization based on analysis of CAESAR 3D body scan data

Jie Pei, Huiju Park and Susan P. Ashdown

Abstract

In this study we explore the variation in female breast shape across the younger (age: 18–45), non-obese (BMI < 30) North American Caucasian population, a population that has not previously been well-represented in studies of breast shape. A method of classifying breast shape was developed based on multiple data-mining techniques. Forty-one relative measurements (i.e., ratios and angles) were constructed from 66 raw measurements (circumferences, depths, widths, etc.), extracted from 478 CAESAR (Civilian American and European Surface Anthropometry Resource) scans, using self-developed Matlab® programs. Seventy subjects were regarded as outliers and were removed. The remaining data were transformed and standardized to ensure robust analysis. To judge results, an algorithm was developed to visualize clustering outcomes in the form of side profiles of breasts. The results of three clustering methods, namely hierarchical, K-means, and K-medoids clustering, were compared. Finally, 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 the goodness-of-fit of the model. Several of the relative body measurements were identified to be critical in defining breast shape. The findings and the proposed methods of this study can contribute to the development of improved shape and sizing systems of bra products that work for both manufacturers and consumers. The new methodology developed in this study can also be applied to other types of intimate apparel products where an understanding of body shape plays a key role in body support, comfort, and fit.

Keywords

3D body scan, breast shape, categorization, caucasian, multivariate analysis

Intimate apparel is a general term for garments that are worn underneath outer clothing and next to the skin.[1] Intimate apparel can provide hygiene, tactile and thermal comfort, and other functionalities.[2] In particular, the brassiere functions to cover and support female breasts to provide concealment, and to prevent the breasts from sagging or other undesirable configurations.[3]

Providing good fit is a common and long-existing issue for intimate apparel, especially for bras. According to the literature, up to 70% of women are wearing incorrectly sized bras.[4–6] A survey conducted by Nethero on 1500 US females showed that more than 65% of them encountered discomfort while wearing bras[7]; more than half of the respondents had issues with fit across the back; 50% reported straps sliding off the shoulders while 28% reported straps digging into their shoulders; over 50% complained about insufficient support from their bras and 25% of all participants felt that breast lifting was insufficient;

35% of the 1500 women had experienced underwire pain; and approximately 27% suffered from the bra riding up on the body as it was worn.[7]

The complexity in breast shape and variations in size and shape among women contribute to the difficulty in the design of intimate apparel with good fit.[8] Affected by pregnancy, nursing, and menopause, women's upper body shape, including breast shape, goes through changes throughout all stages of life, which adds to the complexity of the issue.[9] Variations due to age

Department of Fiber Science and Apparel Design, Cornell University, Ithaca, NY, USA

Corresponding author:
Jie Pei, 263 Human Ecology Building, Department of Fiber Science and Apparel Design, Cornell University, Ithaca, NY 14853, USA.
Email: jp2285@cornell.edu

and ethnicity can also contribute to the diversity in female breast shapes. Researchers have found that physiological changes experienced with age result in breasts sitting lower on the body, an increasing distance between bust points, and a decrease in the relative density of breasts.[10,11] Shin carried out an investigation of 90 Asian females and 90 Caucasian females, and found significant differences in breast configuration between the two groups.[12]

The study of body shape variation can contribute information for improving garment fit and comfort, and for developing effective sizing systems. Feather and colleagues claimed that different body shapes could directly affect the satisfaction level provided by garment fit.[13] LaBat classified a group of female college 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.[14] Likewise, a better understanding of breast shape can help with improvements in bra design. Chen and colleagues claimed that different bust prominence resulted in variations in 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.[15] Zheng and colleagues suggested that an improved bra design could fit the complex contours of the breast, and provide support and appropriate strain by proper use of cup design, shoulder strap, and bottom band configurations.[16] Oh and Chun emphasized the importance of measuring breast size precisely in order to achieve good fit in bra design.[17] Lee and Hong studied the geometrical shape of the underbust curve and came up with an optimal design for the underwire that could provide better support for breasts.[18] Zheng and colleagues developed an enhanced bra sizing system with higher accommodation rates for Chinese females based on the anthropometric analysis of 456 nude breasts.[19]

Although research on female breast shape has been carried out quite intensively for the Asian population, few robust studies with a sufficient number of participants have been found that investigate the 3D shapes of female breasts for the Caucasian population. In the limited number of studies that involve Caucasian females, data were mostly collected from patients seeking breast plastic surgery.[20–22] However, the breast shapes of females who desire cosmetic changes (even though data were collected before surgery) cannot be considered as a representative sample for the general Caucasian population. The traditional bra sizing system and size selection method (which adopts the body measurements of bust circumference and underbust circumference) is still widely used by intimate apparel companies.[23] 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 example, not taking factors such as the relative position of the breasts on the chest wall into account).[7,19,24].

Based on the research gap identified through the literature review, this study is designed to understand the variation in female breast shape across the Caucasian population, and we propose a categorization method for breast shape using various multivariate statistical methods and data-mining techniques. In fact, previous researchers have adopted multivariate methods [principal component analysis (PCA), cluster analysis, discriminate analysis, etc.] in the study of both breast and body shape categorization.[18,19,25,26] However, no previous studies of breast shape categorization use shape-defining ratios as distinct from linear measurements that determine size variation. The various categorization results could be further improved if the following aspects were taken into consideration: (1) the performance of data examination and assumption diagnostics (to ensure the robustness of statistical analysis);(2) the inclusion of a sufficient number of body measurements (the true influential measurements often remain unknown until the analysis is finished); (3) the consistent use of software tools to handle scans and extract body measurements (to avoid human error); (4) the establishment of a validation or justification method of the analysis outcome (most effective if it is a non-statistical method); (5) the selection of the proper algorithms or methods (for instance, numerous clustering algorithms exist for cluster analysis, but none is obviously optimal); (6) the acquisition of the key body measurements which dominate the categorization decisions (to simplify the categorization process and make it easy for the industry and consumers to understand and adopt). These are the aspects that our study contributes to the current body shape study and categorization methodologies.

Methodology

Target population

The Civilian American and European Surface Anthropometry Resource (CAESAR) project collected 3D body scans primarily from three countries: the USA, the Netherlands, and Italy. Combined body scans collected from various locations in the USA and from Ontario, Canada, are referred to as the North American scans. Participants in this study were scanned wearing well-fitting shorts and a soft sports bra. Although the sports bra may alter the shape of

the nude breasts to some extent, it does not reshape the breasts as much as a traditional bra. Seventy-two landmarks were manually placed on participants' bodies by the CAESAR anthropometrists prior to scanning. Landmarks placed by trained anthropometrists can be more reliable than those automatically derived from a 3D body scan by computer analyses based on geometric features of the body, as the manual placement of landmarks can incorporate palpation for joint interfaces and other bone protuberances that cannot be identified on the surface of the body. The coordinates of these landmarks can be directly accessed from the database. The landmarks involved in this study are as follows: the thelion/bust point, right and left; the substernale point; the acromion, right and left; and the axilla point, anterior, right, and left (see Appendix A for details).

The full set of scans of North American female Caucasians was sorted and scans of those participants aged between 18 and 45 with a BMI (body mass index) below 30 (a total of 478 scans) were retained for this study.

Body measurements

The traditional anthropometric measures initially derived from the scans – bust girth and underbust girth – cannot fully describe the complicated 3D shape of breasts. Hence, the majority of body measurements in this study were directly extracted from the CAESAR scans via self-developed Matlab® programs. The initial scans were processed (e.g., shifting, rotating, removing noisy points, etc.) before measurement extraction. The averaged x-coordinates (X_0) and the averaged y-coordinates (Y_0) of all points on the upper torso were calculated and the vertical line at $x = X_0$ and $y = Y_0$ was identified as the central axis for each scan. Each scan was shifted so that $X_0 = 0$ and $Y_0 = 0$. In other words, the central axis locates at $x = 0$ and $y = 0$ after shifting.

A total of 66 raw measurements were automatically extracted, including 2 widths, 5 depths, 7 point-to-point distances, 14 areas, and 2 angles from transverse planes (Figure 1(a)), and 7 thicknesses, 7 heights (or height differences), 6 distances, 11 areas, and 5 angles from sagittal planes (Figure 1(b)). The selection of measurements is mostly inspired by previous studies. We attempted to include all breast-shape-related measurements to the best of our knowledge. A detailed explanation of all 66 raw measurements can be found in Appendix B.

One of the main body shape theories suggests that shape is independent from size.[25] This study also adopted this theory, assuming breast shape does not depend on breast size, and therefore concentrated on the shape factors in our analysis. Song suggests that in order to eliminate the effect of size, instead of using raw measurements, ratios and body angles should be used in calculations designed to identify body shape categories.[25] Therefore, 34 ratios were constructed from the raw measurements (e.g., the ratio between body thicknesses at two different height levels), including 1 circumferential ratio, 6 thickness ratios, 3 height ratios, 5 distance ratios, 16 area ratios, 2 width ratios, and 1 depth-to-width ratio. We included all the possible ratios that in our estimation could contribute to shape variation. In addition, body angles also measure shape in a way that is not influenced by the absolute size. Therefore, 41 variables (34 ratios and 7 angles) were included in further analysis. A detailed explanation of all 41 variables can be found in Appendix C.

Examination and preparation of data

The importance of data examination seems to have been underestimated in many studies: outlier detection, data transformation, and assumption diagnostics are not generally reported in previous studies. Nevertheless, it is essential to examine the data for possible removal of outliers and for beneficial transformations to strengthen statistical outcomes. Outliers (extreme cases) in the data can influence the statistical analysis results negatively, especially for multivariate analysis. Moreover, most statistical models have some underlying assumption, such as linearity, normality, etc. Violation of assumptions can lead to biased models, over-complicated models, or failure in model fitting. Assumption violations can be improved, if not completely fixed, by data transformation, which can help to improve the homogeneity in variance, the linearity, and the normality of a variate or of model residuals.

The data for this study were thoroughly examined for outliers and skewed distributions before further analysis. For identification of outliers, the scans of each of the extreme cases were visually inspected before removal to guard against removal of scans that represent desirable variation in the population. Most of those scans exhibited a non-standard scanning posture that would preclude reliable measurement, in which the participant was: (1) leaning too far to the front; (2) leaning too far to the back; or (3) twisting the torso. A handful of the study participants appear to be very different from the population as a whole by: (1) having unusually flat breasts; or (2) having unusually high or low bust points (e.g., severe breast ptosis). In the end, 70 subjects (14.6% of the 478 subjects) were removed from the data, reducing the sample size to 408.

In addition, variables were examined for appropriateness for statistical analysis. A total of 19 of the 41 variables were transformed via one of the following transformation methods to correct for skewed
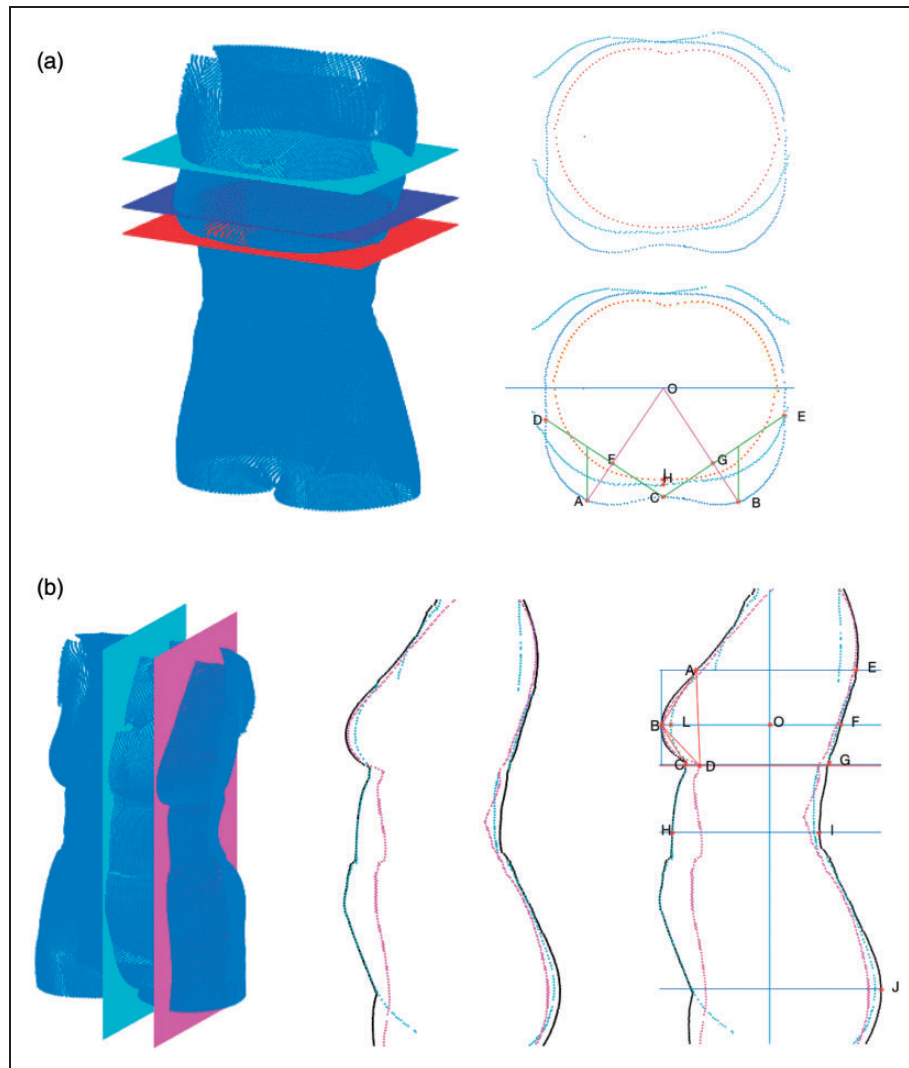
Figure 1. Extraction of body measurements in Matlab®. (a) Transverse planes sliced at different height level, (b) Sagittal planes sliced at different locations. Note: the black outline is a projection of the overall side profile.

distribution: logarithm, inverse, square root, second power, third, –0.5th power, log[log(1/y)] (first the application of inverse; then the application of logarithm twice).

Algorithm to visualize categorization outcomes

An algorithm was developed to visually present the breast shape categorization results based on two dimensional slices of the scans. Figure 2 demonstrates one result of categorization, when subjects were categorized into three groups, shown in three colors.

To remove the effect of size from the body slices and facilitate visual analysis, it is necessary to scale each individual scan to a common size. Therefore, the height difference between the axilla level and the hip level was scaled to be 1, as shown in Figure 2(a).

The scans were then aligned at both the axilla level and the hip level.

After scaling and categorization, the representative shape of each group needs to be identified. For the side profile, the algorithm developed for this purpose searches for the midpoint among all side profiles that belong to the same group, at a fixed height level. A total of 30 fixed height levels (downwards from the axilla level to the hip level) are implemented. The search is done separately for the anterior body and the posterior body. For the bust plane (transverse plane sliced at the bust-point level, determined by the averaged height of the right and left bust points), the algorithm searches for the midpoint in terms of the radial distance from the origin [i.e., point (0,0)], among all planes that belong to the same group, at a fixed angle. A total of 40 fixed angles (counterclockwise from 0 to 2 × π) were
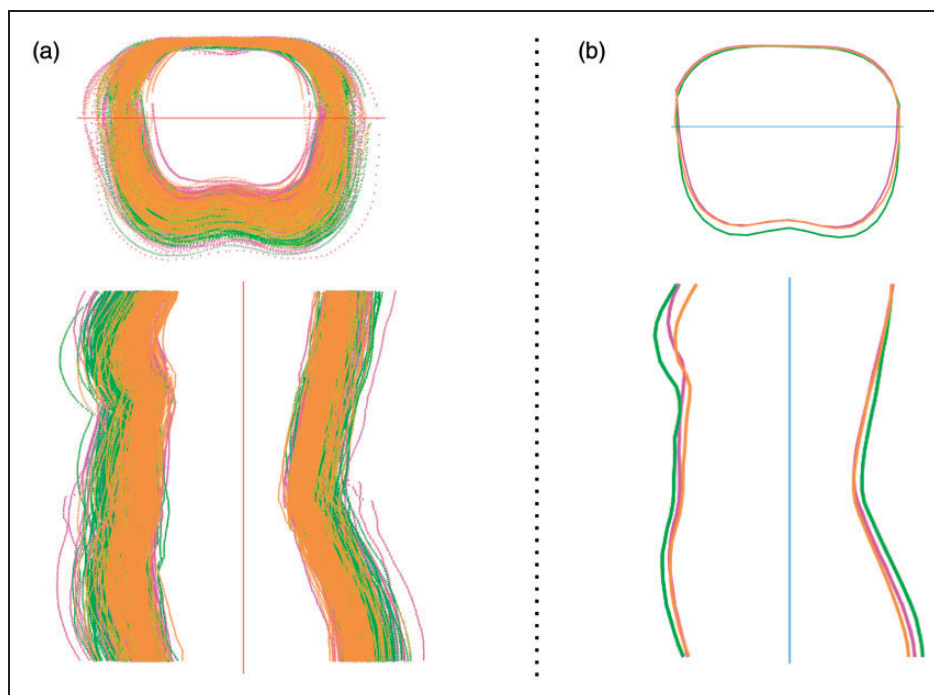
Figure 2. Demonstration of the algorithm. (a) Breast shape categorized into three groups (408 subjects included). (b) Representative shapes of the three groups.

implemented. After identification, the midpoints were connected together in sequence.

As shown in Figure 2, although distinct differences in breast shapes can be observed in the side profiles, not much distinction can be observed in the transverse plane. This is not surprising since the effect of size was removed by scaling. It is the circumferential changes across altitudes, rather than circumferences themselves, that dominate the shape differences (this is also why the difference between bust girth and under-bust girth is used to determine cup size). Therefore, the side profiles serve as a better visualization tool than the transverse planes. Hence, the following analysis used only the side profiles as the major justification mean.

Statistical analysis

Multivariate analysis allows for investigation of inter-correlation among numerous variables, which makes it suitable for the study of human body measurements.

The first multivariate statistical method adopted is PCA, which searches for a new set of mutually ortho-gonal variables, called principal components (PCs), transformed from the original interrelated variables, where each PC is a linear combination of the original variables with varying coefficients, or loadings.[27] While the correlation structure is retained via the PC loading matrix, the PCs themselves are rid of the correlations (and covariances), allowing the analysis to concentrate on variances. The main idea of PCA is to reduce the dimensionality of data while preserving as much varia-tion in the data as possible.

Cluster analysis aims at grouping objects in such a way that objects in the same group (cluster) are similar, whereas distinction can be observed between groups.[28] There are numerous clustering algorithms and different algorithms can lead to very different outcomes. Therefore, three of the most commonly used methods – K-means, K-medoids, and hierarchical clustering – were included in the analysis of this study for compar-ison. The number of clusters selected for analysis also affects the outcome in a major way. Therefore, in this study we propose two different criteria for the selec-tion of cluster number. The first criterion is based on misclassification rate, obtained from linear discriminant analysis (LDA), and from random forest (RF) analysis. The second criterion is based on the goodness-of-fit of the model, where three goodness-of-fit measures were used as a reference, namely the statistics of BIC (Bayesian information criterion), AIC (Aikaike's infor-mation criterion), and WSS (within-groups sum of squares, also known as residual variance).

MANOVA (multivariate analysis of variance) was applied to examine whether the multivariate means of different clusters are significantly distinctive. Four major MANOVA test statistics (Wilks' lambda, Roy's maximum root, Hotelling–Lawley trace, and Pillai's trace) were used in calculating p values.

Lastly, to reduce data dimensions, this study proposed an approach based on a visual judgment of the side profiles of the breasts. With the reference of PC loadings and RF importance measures, a few key variables were selected from the 41 to start with. Then multiple trials were performed. Each time one 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 then be visually compared with the original. If the new result was similar to the original clustering result, further variable exclusion would be performed. A significantly distinctive result would lead to keeping the variable and attempting to delete another variable. (The visual similarity or dissimilarity was judged by comparing side profiles of breasts. The profiles were generated from the algorithm described above for visualizing grouping outcomes.) The deletion sequence also referred to the PC loadings and RF importance measures.

Results and discussions

Principal component analysis

The 41 original variables are essentially 41 vectors pointing in non-orthogonal directions (the non-orthogonality is caused by correlations among variables). After the PCA procedure, a total of 41 PCs were obtained, pointing in 41 mutually orthogonal directions.

The original variables had not been standardized after the removal of outliers and transformation, whereas the finalized variables have undergone standardization (where each observation was subtracted from the mean of the variable it belongs to, and then was divided by the standard deviation of that variable). Table 1 shows the partial PCA summary tables for both the unstandardized and the standardized data. It contains some summary statistics of the first 10

PCs. Proportion of variance is calculated via equation 1 and sums up to 1 for all 41 PCs. Clearly, for the unstandardized data, the first five PCs can explain around 90% (88.37%) of the total variance. For the standardized data, it requires at least 10 PCs to explain 90% (88.11%) of the total variance.

$$\text{Proportion of Variance} = \frac{\text{Variance included in a specific PC}}{\text{Sums of variances for all PC's}} \quad (1)$$

According to Table 1, compared with the standardized data, unstandardized data have a plausibly more promising summary table for the PCs: they requires fewer PCs to have a high proportion of variance explained. It also has a less ambiguous scree plot (Figure 3(a) and (b)), in which the break point is more distinct. However, it does not work well in terms of classifying breast shapes (Figure 3(c) and (d)): The side profiles of breasts from different clusters are hardly distinguishable (Figure 3(c)). This example shows that the PCA table and scree plot alone may not be sufficient to show the level of success of the analysis results. Therefore, in this study, evaluations and judgments were based on the side profiles generated from the visualization algorithm described earlier.

Cluster analysis

To find the best clustering method for these specific data, three of the most commonly used clustering algorithms were chosen, namely hierarchical, K-means, and K-medoids clustering. All 41 variables were entered into the analysis. The algorithm designed for visual analysis of results (judged from side profiles of the breast shape) makes it possible for comparisons among the three clustering methods. A series of cluster numbers from k = 2 to k = 9 have been applied.

Table 1. PCA summary table (partial)

PCA applied to the **unstandardized** data										
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Standard deviation	0.4367	0.3720	0.3015	0.1886	0.1317	0.1167	0.0920	0.0886	0.0837	0.0920
Proportion of variance	0.3564	0.2586	0.1699	0.0665	0.0324	0.0254	0.0158	0.0147	0.0131	0.0158
Cumulative proportion	0.3564	0.6150	0.7849	0.8513	<u>**0.8837**</u>	0.9092	0.9250	0.9397	0.9528	0.9250
PCA applied to the finalized (transformed and **standardized**) data										
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Standard deviation	3.1511	2.5573	2.2338	1.9150	1.7784	1.6626	1.2715	1.1352	1.0566	1.0250
Proportion of variance	0.2422	0.1595	0.1217	0.0894	0.0771	0.0674	0.0394	0.0314	0.0272	0.0256
Cumulative proportion	0.2422	0.4017	0.5234	0.6129	0.6900	0.7574	0.7968	0.8283	0.8555	<u>**0.8811**</u>

Note. The boldfaced and underlined values indicate the cumulative proportion of variance is approaching 90%.
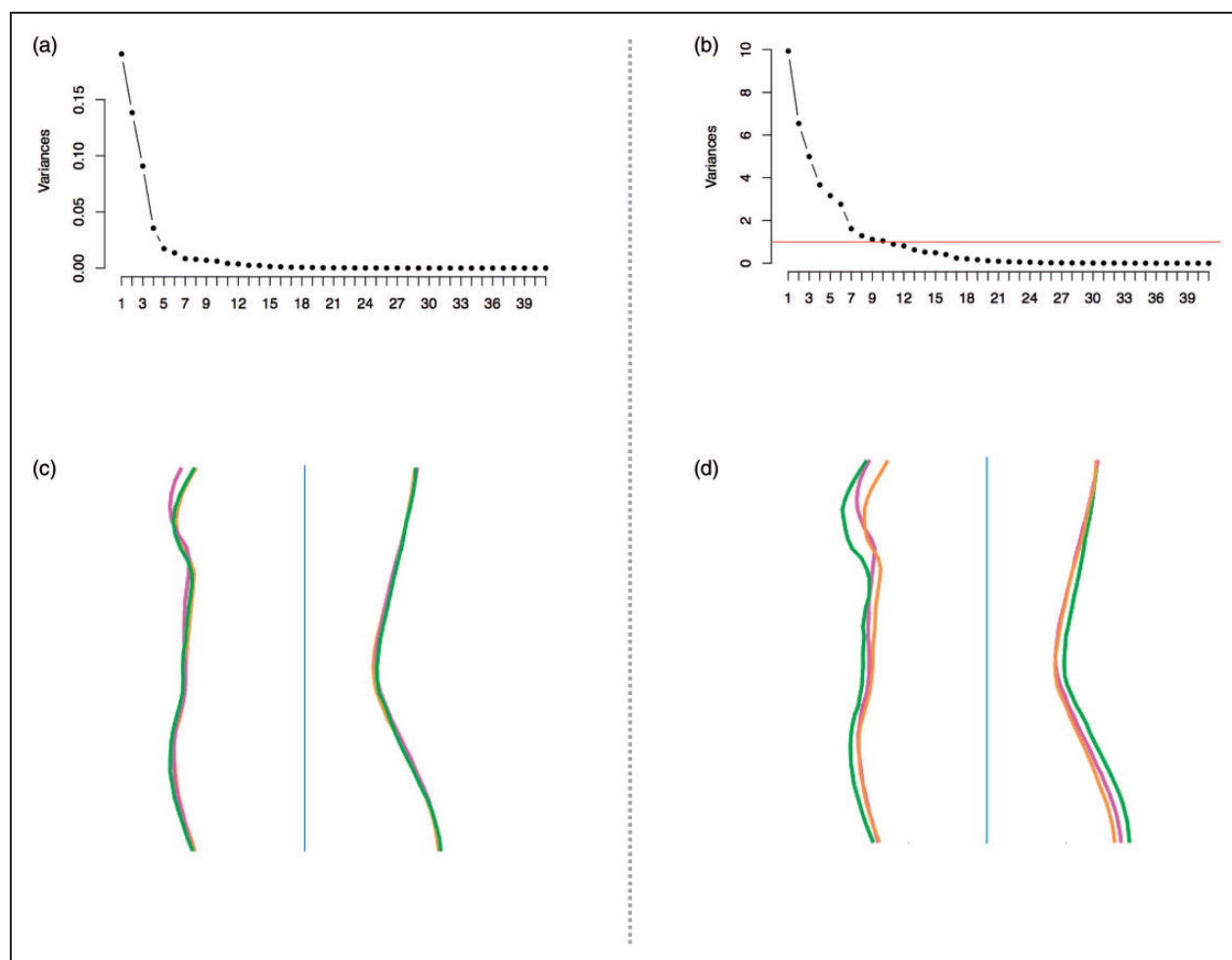
Figure 3. Categorization results calculated from unstandardized and standardized data (same procedure: K-means clustering applied to the first 10 PCs). (a) Scree plot (unstandardized data); (b) scree plot (standardized data); (c) side profiles (unstandardized data); (d) side profiles (standardized data).

Figure 4 shows the categorization results for the three-cluster case (k = 3) and the five-cluster case (k = 5), as examples only. In general, it was found that K-means is the best algorithm in giving the most distinctive breast shapes and presenting good stability and repeatability.

A wrong choice of cluster number can lead to poor clustering results that do not reflect the real homogeneity and heterogeneity in the data. The analysis of different numbers of clusters found that distinctions in breast shapes are too trivial, and several of the shapes that are obtained from different clusters look almost identical (their side profiles overlap when plotted together) for the cases of six clusters or more. This implies that classifying breast shapes into this large number of groups may add unnecessary complexity to a sizing system. On the other hand, for the two-cluster case, the two side silhouettes are not visually significantly different. This is probably because the impact of some variables was counterbalanced within clusters,

rather than being identified and summarized by the cluster. Therefore, the choices of cluster numbers were narrowed down to either three, four, or five before further analysis.

Selection of cluster number – Criterion 1: based on misclassification rate. Discriminant analysis creates discriminant functions that separate groups of observations from each other, based on existing group assignments.[28] For this study, the group membership obtained from the clustering was used to create discriminant functions. Each observation was then re-classified into new groups based on the discriminant functions. When the new group assignments do not match with the original clustering assignments, the corresponding observations are considered to be misclassified (e.g., a subject who actually belongs to Cluster 1 gets wrongly classified into Cluster 2). The misclassification rate is the proportion of misclassified cases among all observations.
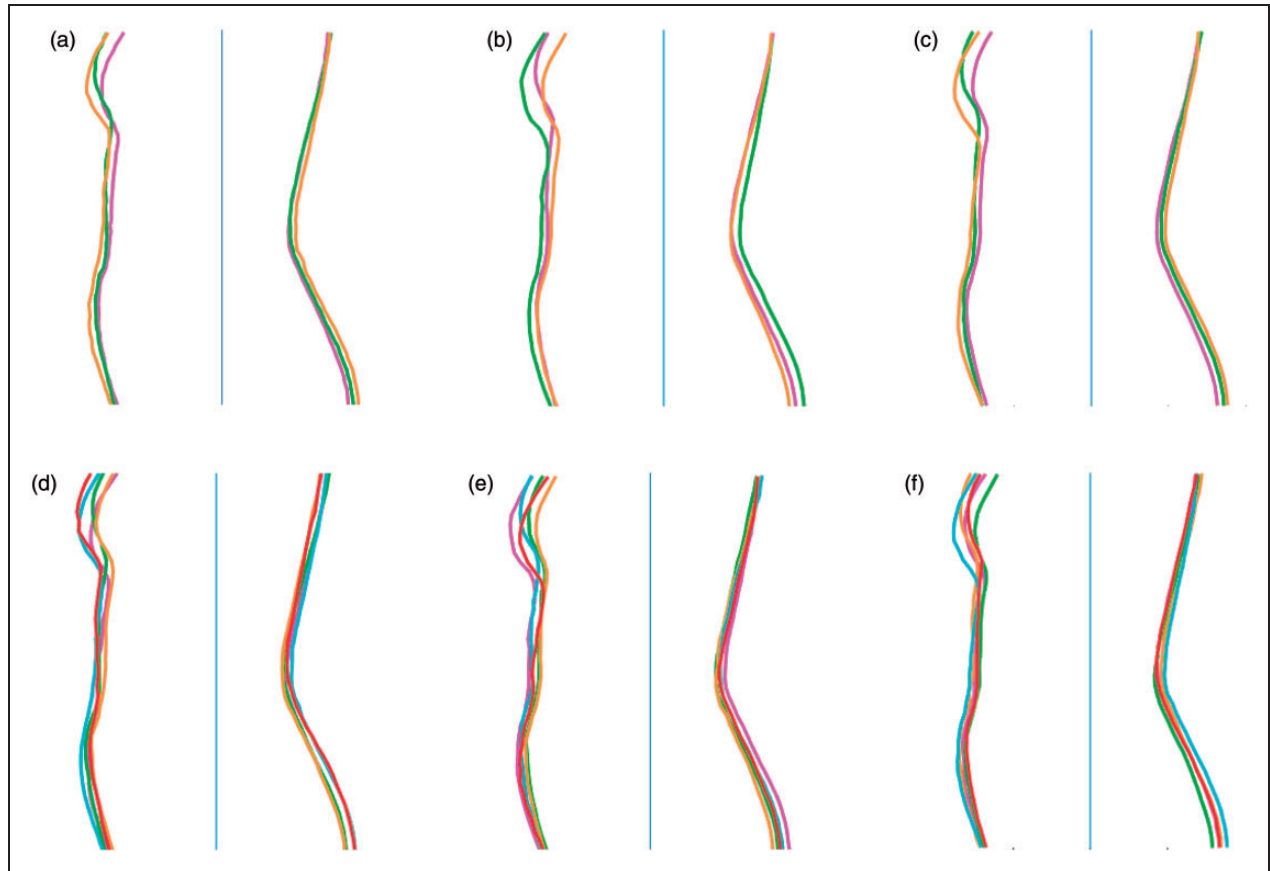
Figure 4. Comparisons among the three clustering methods. (a) Three clusters – hierarchical clustering; (b) three clusters – K-means clustering; (c) three clusters – K-medoids clustering; (d) five clusters – hierarchical clustering; (e) five clusters – K-means clustering; (f) five clusters – K-medoids clustering.
Note: Colors are not consistent among these examples due to coding choices.

Admittedly, using the same cases in building discriminant functions makes it very likely to underestimate the real misclassification rate. However, 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 assignments are tested. All these processes, including the creation of discriminating functions, the new classification, and the calculation of misclassification, etc., were done automatically by the statistical software. LDA was adopted for this study, with cross-validation applied. The misclassification rates for the three-cluster case, four-cluster case, and five-cluster case are 7.35%, 8.33%, and 10.78% respectively.

In place of using linear classification rules, classification tree learning is another way to build classification rules. However, a single tree can be unstable, especially with large numbers of predictors. Therefore, RF is a good way of finding a stable result. RF essentially builds a multitude of tree models. A bootstrap sample (a random sampling of observations with replacement, usually referred to as in-bag sample or IB sample) is repeatedly drawn from the database. The bootstrap sample functions 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 2 shows the OOB estimates of misclassification rate from 1000 trees. It can be seen that K-means clustering has the lowest misclassification rates among the three clustering methods, and the three-cluster case has the lowest misclassification rates compared with the four-cluster case and the five-cluster case.

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

Selection of cluster number – Criterion 2: based on the goodness-of-fit of the 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 in the data. However, an overfitted model

Table 2. OOB estimate of misclassification rate from 1000 trees

Clustering method	Three-cluster case	Four-cluster case	Five-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 PCs)	6.86%	10.05%	9.80%
K-means (8 PCs)	7.11%	9.31%	11.52%
K-means (5 PCs)	7.35%	9.31%	12.25%
K-means (3 PCs)	7.11%	8.09%	9.31%
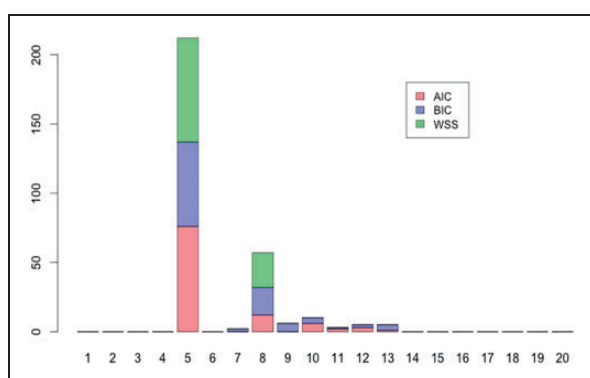
Figure 5. Votes by three different statistics for optimal cluster number.

can run into the problem of generalization. In terms of clustering, a saturated model (perfectly fitted model) is the case when each observation becomes a cluster. Many statistical methods have been proposed to search for models with reasonably high goodness-of-fit while avoiding overfitting. The BIC, AIC, and WSS statistics were used separately as references. An evaluation program was run repeatedly (200 times) for the three statistics to vote for the optimal cluster number. Figure 5 shows the results of the voting. Clearly, k = 5 received the highest number of votes from all three statistics. Therefore, the ideal number of clusters based on Criterion 2 is five.

Multivariate analysis of variance

MANOVA was adopted to examine whether the multivariate means of different clusters are significantly different. There are four major MANOVA test statistics, namely Wilks' lambda, Roy's maximum root, Hotelling–Lawley Trace, and Pillai's trace.[28] Each of the test statistics corresponds to a different approximation of the F value and accordingly a different p value. The null hypothesis of the testing is that all multivariate means are equal. A small p value implies the rejection of the null hypothesis. As shown in Table 3, for both the three-cluster case and the five-cluster case, p values calculated from all four 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 different clusters have statistically significantly different means.

Reduction of dimensionality and the number of variables

The RF package can generate plots to show the importance of variables based on their impact on the

Table 3. MANOVA summary table for the three-cluster case and five-cluster case

	Df	Pillai's trace	Approximate F value	p value
Three-cluster case	2	1.4102	21.344	<0.0001
Five-cluster case	4	2.5231	15.251	<0.0001
	Df	Wilks' lambda	Approximate F value	p value
Three-cluster case	2	0.0867	21.337	<0.0001
Five-cluster case	4	0.0160	16.112	<0.0001
	Df	Hotelling–Lawley trace	Approximate F value	p value
Three-cluster case	2	4.8048	21.329	<0.0001
Five-cluster case	4	7.7039	16.981	<0.0001
	Df	Roy's maximum root	Approximate F value	p value
Three-cluster case	2	2.5994	23.204	<0.0001
Five-cluster case	4	7.7039	16.981	<0.0001

goodness-of-split, measured by the decrease in node impurity, which describes the level of homogeneity of a node (a pure node is the case when every observation included by a node belongs to the same group). Larger values in the plot represent higher importance. The importance measure is often used for variable selection to obtain a simpler model. Furthermore, 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. Hence, both the importance measure and PC loadings are regarded as helpful references for the reduction of dimensionality. Accordingly, in the importance plots (Figure 6), variables that have large loadings for the first PC were colored red; variables that have large loadings for the second or third PC were colored green or blue respectively.

The three-cluster case. According to Figure 6(a), the following eight variables appear to be more important:

#23, #19, #13, #20, #17, #26, #14, and #24. They were regarded as the initial key variables. These initial key variables were selected through trial and error (an arbitrary cutoff value for the importance measure was chosen at first, but this was adjusted throughout the process). More variables with relatively high importance need to be included if these initial key variables do not end up with a categorization result similar to that of the 41 variables (judged *via* side profiles). Figure 7(a) displays the clustering outcome when K-means was applied to the first 10 PCs, obtained from the original PCA with all 41 variables included. It can be observed that Figures 7(b) and 7(a) present similar outcomes. In other words, the combination of the eight key variables can function almost equally well as that of 41 variables. Hence, the number of variables involved in the clustering can be reduced from 41 to 8 without losing too much information. Once these eight variables were identified, multiple trials were performed. Each time one more variable (from the eight key variables) was excluded and the new clustering
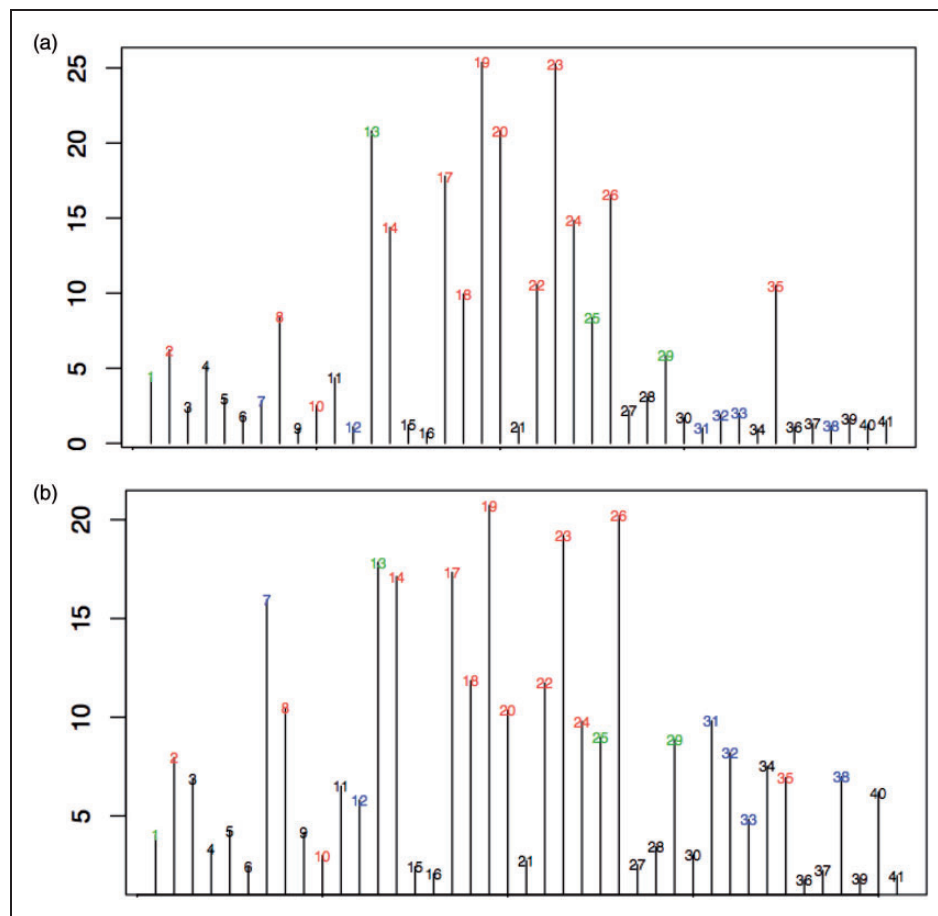

Figure 6. Measure of importance (random forest). (a) The three-cluster case; (b) the five-cluster case.
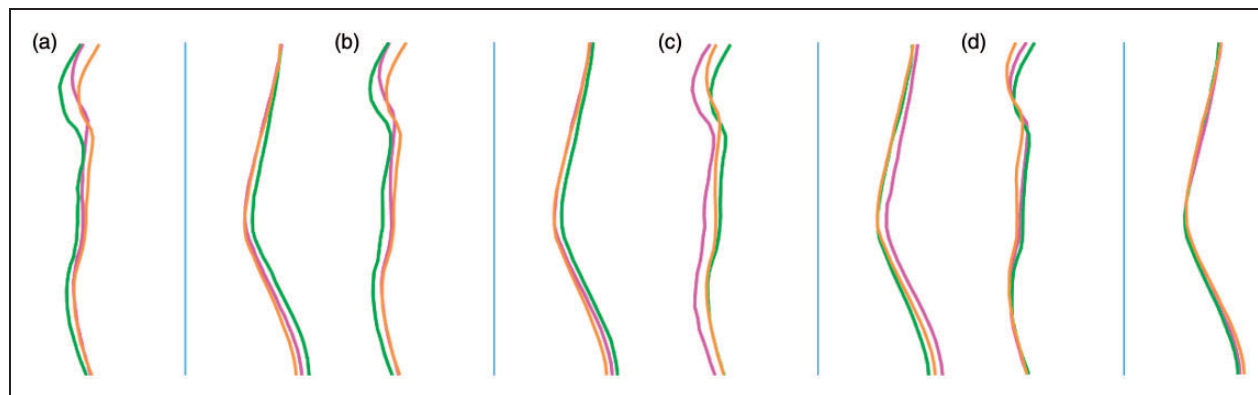
Figure 7. Reduction of the number of variables (three-cluster case). (a) K-means applied to the first 10 PCs from 41 variables; (b) K-means applied to the first 2 PCs from 8 key variables; (c) K-means applied to the first 2 PCs from 2 variables (#23, #13); (d) K-means applied to the first 2 PCs from 2 variables (#23, #19).
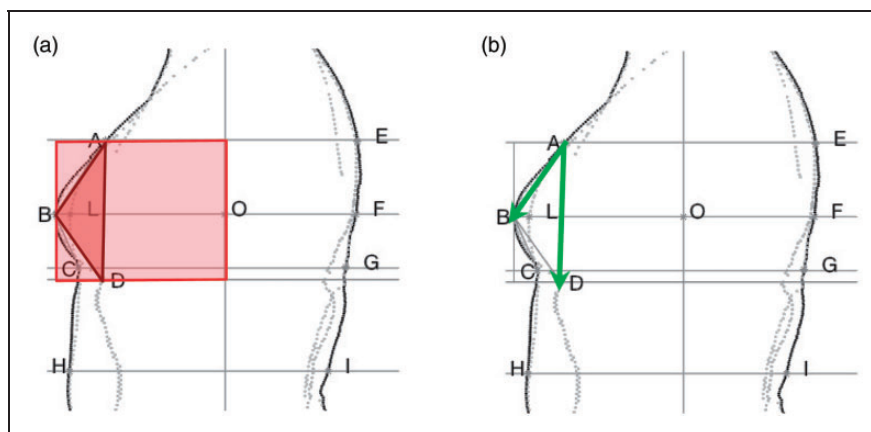

Figure 8. Demonstration of the two finalized key variables (three-cluster case). (a). Variable #23; (b) variable #13.

result was compared with the original. In the end, the number of variables was successfully reduced to two (see Figure 7(c)): #23 and #13.

Although variable #19 has higher importance than variable #13 (according to Figure 6(a)), it does not contribute greatly to a good clustering result when visually judged from a side silhouette (Figure 7(d)). This is probably because variable #13 is responsible for the direction of the second PC, while #23 and #19 are both influential variables for the direction of the first PC. This indicates that it is important to retain the second dimension. This also suggests that keeping only the first 2 PCs, in comparison with 10 PCs (as concluded from the original PCA), is sufficient to achieve similar outcomes (Figure 7(a) shows the result when 10 PCs were involved, whereas Figure 7(b–d) shows the results when only 2 PCs were involved).

As demonstrated in Figure 8, variable #23 is the area ratio between triangle ABD and the rectangle at the anterior body; variable #13 is the angle BAD.

Both of them relate to triangle ABD. These two variables alone are sufficient to partition observations into three clusters.

The five-cluster case. Referencing Figure 6(b), 17 variables were selected to start the process of variable reduction (#19, #26, #23, #13, #14, #7, #17, #18, #22, #20, #8, #25, #24, #31, #29, #32, and #2). After multiple trials, the number of variables was successfully reduced to four (see Figure 9(c)): #19, #13, #14, and #32. Any further exclusion of the key variables led to a different clustering result (Figure 9(d)) even when all three dimensions were retained. Nonetheless, keeping the first 3 PCs is sufficient (Figure 9(a) shows the result when 10 PCs were involved, whereas Figure 9(b–d) shows the results when only 3 PCs were involved).

As demonstrated in Figure 10(a–c), variable #19 is the length ratio between line segment AB and line segment BD; variable #13 is angle BAD; and variable #14 is angle ABD. All three of these relate to the bust
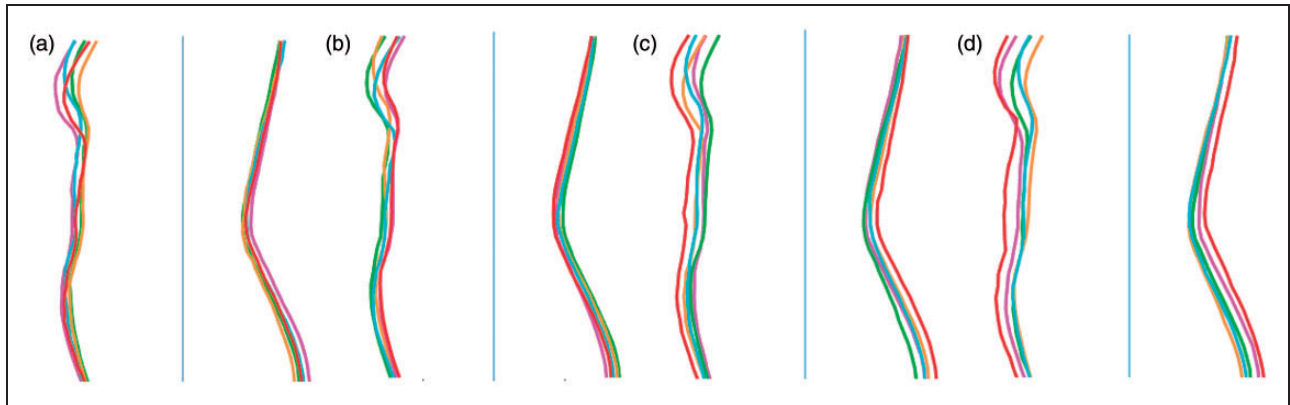
Figure 9. Reduction of the number of variables (five-cluster case). (a) K-means applied to the first 10 PCs from 41 variables; (b) K-means applied to the first 3 PCs from 17 key variables; (c) K-means applied to the first 3 PCs from 4 variables; (#19, #13, #14, #32); (d) K-means applied to the first 3 PCs from 3 variables (#19, #13, #32).
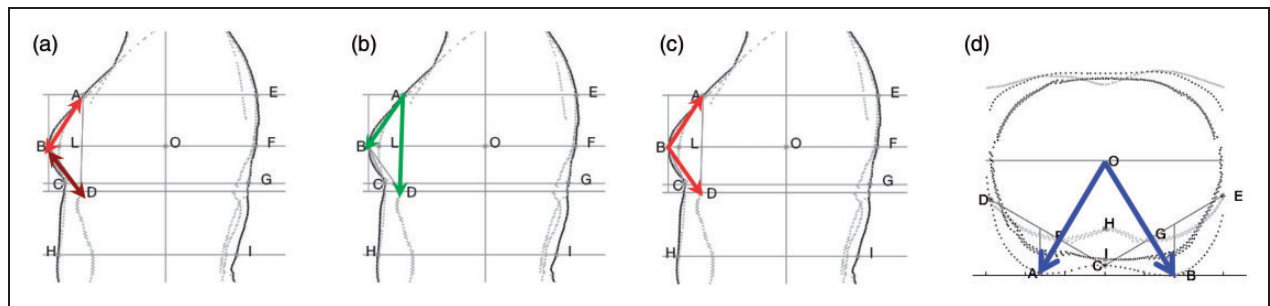

Figure 10. Demonstration of the four finalized key variables (five-cluster case). (a) Variable #19; (b) variable #13; (c) variable #14; (d) variable #32.

triangle (triangle ABD). In addition, variable #32 (Figure 10(d)) is angle AOB, representing the pointing of bust points inspected from the transverse plane. The four variables alone are sufficient to partition observations into five clusters.

Conclusions

In this study we developed a Matlab® program to achieve automatic extraction of all desired measurements. Using one program avoids the possible inconsistency and error caused by different calibrations or settings in various software programs. Automatic extraction saves time when analyzing many scans, and avoids human error due to unintentional mistakes in operation. In addition, instead of analyzing body measurements separately, this study utilized various multivariate statistical methods, data-mining, and machine learning techniques to retain and study the correlations among multiple body measurements. Moreover, the data were thoroughly examined for outliers and skewed distributions to improve the accuracy and validity of the statistical analysis.

The original PCA applied to the standardized data shows that at least 10 PCs are required to explain 90% of total variance. However, two or four critical variables were found sufficient to categorize breast shapes into three or five groups respectively, when judged by visual analysis of the side silhouette, thus reducing the number of PCs (or the dimensionality) to two or three, respectively. Most of the variables are associated with the bust triangle observed from the sagittal planes. Moreover, in this study we propose a novel way to do breast shape categorization after comparing hierarchical, K-means, and K-medoids clustering using a self-developed program that visualizes grouping outcomes. Considering that a good choice of cluster number is essential in reflecting the homogeneity and heterogeneity in the data, two cluster number selection criteria were proposed based on different considerations: (1) based on the misclassification rate (focusing on how well a new case can be classified into the correct group); and (2) based on the goodness-of-fit of the model (focusing on how well the model can capture and explain the majority of variations in the data).

In terms of application, this study found the most representative breast shapes based on all 41 variables. The measurements of these representative shapes can be directly referred to when building dress forms. Moreover, with technologies such as 3D-averaging, the averaged breast shape of each cluster can be 3D-printed and directly used in product development. Further work on size variation in combination with our analysis of shape is needed. Then the interactions and correlations between the breast size and the breast shape can be more thoroughly understood, providing data for the development of a brand-new sizing system that has the potential to improve accommodation rate or reduce aggregate-fit-loss. In this study we also propose an approach to reduce the number of variables so that the key body measurements can be identified. The key measurements make it possible to quickly allocate each consumer into her correct group (or her correct size in a shape-based sizing system) without the time-consuming and calculation-heavy clustering processes using all 41 variables. The findings and the proposed methods therefore provide information that is useful to develop bra products that work for both manufacturers and consumers. Lastly, the same methodology can be adopted in the shape study of other parts of the body of interest to apparel designers (for example, buttock shapes), and thus can be applied to the improvement of other types of intimate apparel products.

This study had several limitations. First, each participant of the CAESAR project was scanned wearing a soft sports bra. It is possible that the bra may have altered the shape of the breasts, although the influence remains unknown. The outcomes of this study could be more convincing if the same methodology was applied to 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 vastly different from the desired shape provided by the support of the bra, and can introduce undesirable variation to the study, based on different physiological measures. Second, the side profile view of the breasts is the only reference for comparison of the clustering methods. It is possible that similar side profiles may have different 3D shapes. In the future, judgments and decisions can be made by additionally referring to other views of the breasts, including transverse planes sliced at various levels, and sagittal planes from other body locations. Lastly, this study concentrated on a particular population: North American, younger, non-obese Caucasians. In the future, the same methodology can be applied to other populations (e.g., the plus-size population, the elder population, etc.) to explore to what extent the outcomes of this study can be generalized, and to test our methodology on populations with greater variation.

References

1. Law D, Wong C and Yip J. How does visual merchandising affect consumer affective response? An intimate apparel experience. *European J Mark* 2012; 46(1/2): 112–133.
2. Farnworth B and Dolhan PA. Heat and water transport through cotton and polypropylene underwear. *Text Res J* 1985; 55(10): 627–630.
3. Farrell-Beck J, Poresky L, Paff J, et al. Brassieres and women's health from 1863 to 1940. *Cloth Textiles Res J* 1998; 16(3): 105–115.
4. Greenbaum AR, Heslop T, Morris J, et al. An investigation of the suitability of bra fit in women referred for reduction mammaplasty. *Br J Plast Surg* 2003; 56(3): 230–236.
5. Wood K, Cameron M and Fitzgerald K. Breast size, bra fit and thoracic pain in young women: a correlational study. *Chiropr Osteopat* 2008; 16(1): 1.
6. McGhee DE and Steele JR. Optimising breast support in female patients through correct bra fit: a cross-sectional study. *J Sci Med Sport* 2010; 13(6): 568–572.
7. Nethero S. Method for optimization of bra fit and style for all body sizes and types and for wearing with a variety of clothing styles. Patent 11/829,568, USA, 2007.
8. Hart C and Dewsnap B. An exploratory study of the consumer decision process for intimate apparel. *J Fashion Mark Manage* 2001; 5(2): 108–119.
9. Tiggemann M. Body image across the adult life span: stability and change. *Body Image* 2004; 1(1): 29–41.
10. Haars G, van Noord PA, van Gils CH, et al. Measurements of breast density: no ratio for a ratio. *Cancer Epidemiol Biomarkers Prev* 2005; 14(11): 2634–2640.
11. Ashdown SP and Na H. Comparison of 3-D body scan data to quantify upper-body postural variation in older and younger women. *Cloth Textiles Res J* 2008; 26(4): 292–307.
12. Shin WK. *The origins and evolution of the bra*. UK: Doctoral dissertation, Northumbria University, 2009.
13. Feather BL, Ford S and Herr DG. Female collegiate basketball players' perceptions about their bodies, garment fit and uniform design preferences. *Cloth Textiles Res J* 1996; 14(1): 22–29.
14. LaBat KL. *Consumer satisfaction/dissatisfaction with the fit of ready-to-wear clothing*. USA: Doctoral dissertation, University of Minnesota, 1987.

15. Chen CM, LaBat K and Bye E. Bust prominence related to bra fit problems. *Int J Consum Stud* 2011; 35(6): 695–701.

16. Zheng R, Yu W and Fan J. Pressure evaluation of 3D seamless knitted bras and conventional wired bras. *Fiber Polym* 2009; 10(1): 124–131.

17. Oh S and Chun J. New breast measurement technique and bra sizing system based on 3D body scan data. *J Ergon Soc Korea* 2014; 33(4): 299–311.

18. Lee HY and Hong K. Optimal brassiere wire based on the 3D anthropometric measurements of under breast curve. *Appl Ergon* 2007; 38(3): 377–384.

19. Zheng R, Yu W and Fan J. Development of a new Chinese bra sizing system based on breast anthropometric measurements. *Int J Indust Ergon* 2007; 37(8): 697–705.

20. Hsia HC and Thomson JG. Differences in breast shape preferences between plastic surgeons and patients seeking breast augmentation. *Plast Reconstr Surg* 2003; 112(1): 312–320.

21. Catanuto G, Spano A, Pennati A, et al. Experimental methodology for digital breast shape analysis and object-ive surgical outcome evaluation. *J Plast Reconstr Aesthet Surg* 2008; 61(3): 314–318.

22. Small KH, Tepper OM, Unger JG, et al. Re-defining pseudoptosis from a 3D perspective after short scar-medial pedicle reduction mammaplasty. *J Plast Reconstr Aesthet Surg* 2010; 63(2): 346–353.

23. White J and Scurr J. Evaluation of professional bra fitting criteria for bra selection and fitting in the UK. *Ergonomics* 2012; 55(6): 704–711.

24. Pechter EA. A new method for determining bra size and predicting postaugmentation breast size. *Plast Reconstr Surg* 1998; 102(4): 1259–1265.

25. Song HK and Ashdown SP. Categorization of lower body shapes for adult females based on multiple view analysis. *Text Res J* 2011; 81(9): 914–931.

26. Devarajan P and Istook CL. Validation of female figure identification technique (FFIT) for apparel software. *J Text Apparel Tech Manage* 2004; 4(1): 1–23.

27. Jolliffe I. *Principal component analysis*. Hoboken, NJ: Wiley, 2002.

28. Rencher AC. *Methods of multivariate analysis*. Hoboken, NJ: Wiley, 2003.
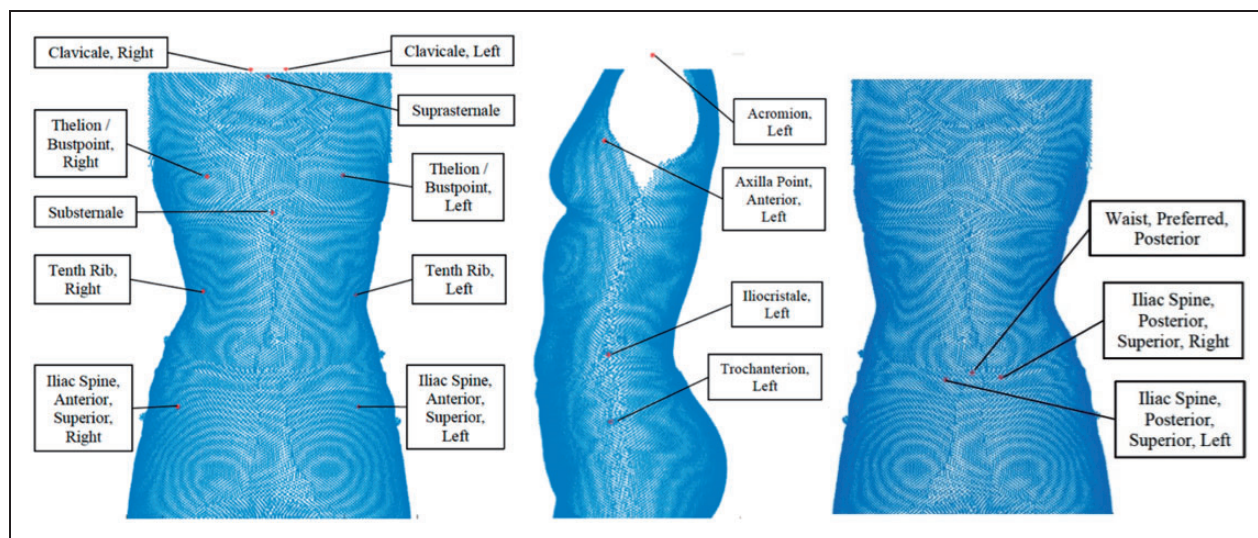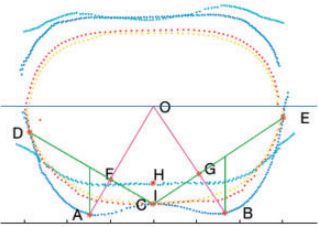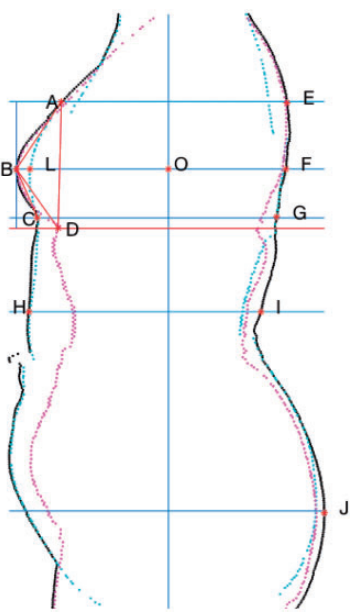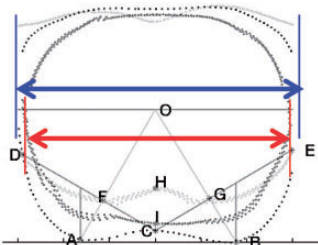
Appendix A: landmarks and auxiliary points
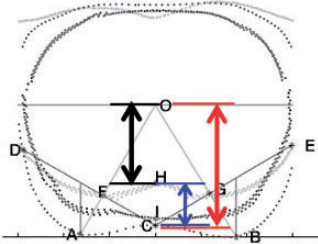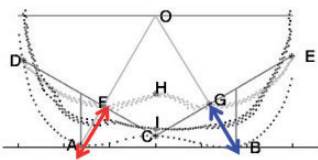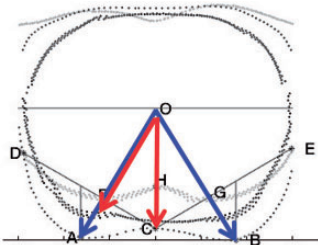
Figure A-1. Landmarks from CAESAR.

Table A-1. Placement of auxiliary points

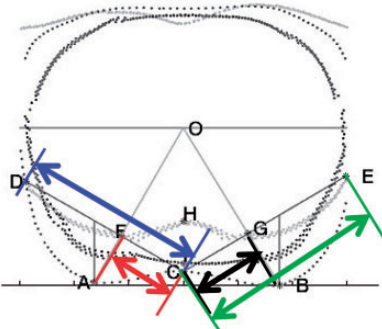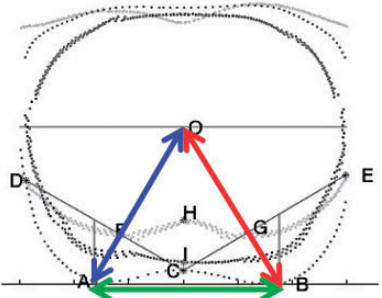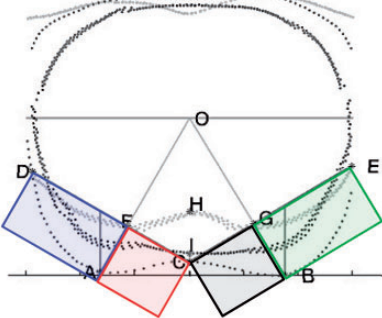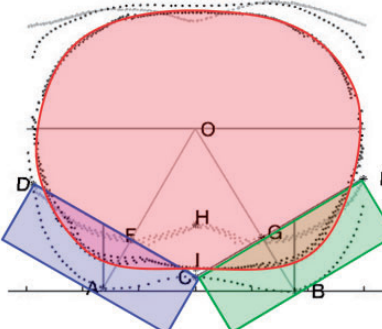| Illustration	Name	Description
Auxiliary points on transverse planes		
	Point O	Point O is where the pivot axis locates on the x–y plane (at $x = 0$ and $y = 0$)
	Point C	The intersection between line $y = 0$ (the vertical line passing through point O) and the outline of the bust plane
	Point I	The intersection between line $y = 0$ and the outline of the underbust plane
	Point H	The intersection between line $y = 0$ and the outline of the axilla plane
	Point A	The left bust point
	Point B	The right bust point
	Point D	The intersection between the line which is perpendicular to line AO, and meanwhile passes through point C, and the outline of the bust plane
	Point E	The intersection between the line that is perpendicular to line BO, which meanwhile passes through point C, and the outline of the bust plane
	Point F	The intersection between line CD and Line AO
	Point G	The intersection between line CE and line BO
Auxiliary points on sagittal planes		
	Point O	The intersection between the pivot axis and the horizontal line at bust level(The pivot axis divides the body into the anterior half and the posterior half)
	Point B	The anterior intersection between the horizontal line at bust level and the outline of the overall side profile
	Point F	The posterior intersection between the horizontal line at bust level and the outline of the overall side profile
	Point L	The anterior intersection between the horizontal line at bust level and the outline of the sagittal plane sliced at the front central line
	Point A	The anterior intersection between the horizontal line at axilla level and the outline of the overall side profile
	Point E	The posterior intersection between the horizontal line at axilla level and the outline of the overall side profile
	Point C	The anterior intersection between the horizontal line at underbust level and the outline of the overall side profile
	Point G	The posterior intersection between the horizontal line at underbust level and 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 the 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 and the outline of the overall side profile
	Point I	The posterior intersection between the horizontal line at waist level and the outline of the overall side profile
	Point J	The most protruded point at buttock

Appendix B: raw measurements

Table A-2. Raw measurements extracted from transverse planes

Illustration	Name	Description
Measure of width		
	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		
	depthHO	The anterior depth of the axilla plane at central front (length of line OH)
	depthCO	The anterior depth of the bust plane at central front (length of line CO)
	depthIH	The anterior depth difference between underbust plane and axilla 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		
	angleBP	Pointing of bust points (angle AOB)
	angleBP_rt	Pointing of right bust point (angle AOC)

(continued)

Table A-2. Continued

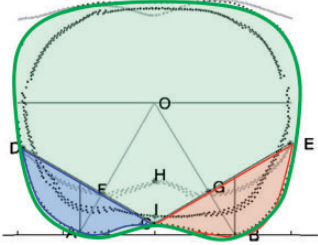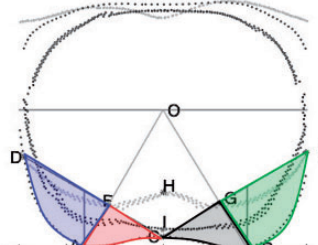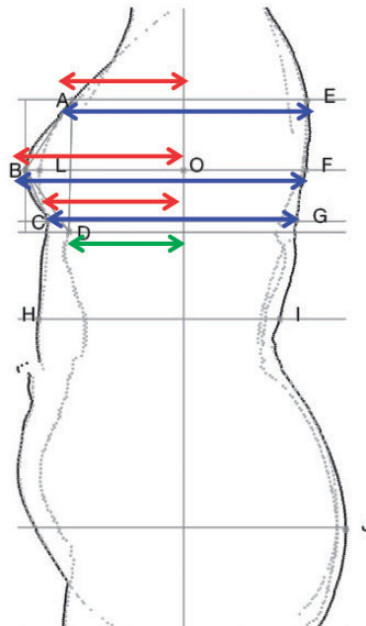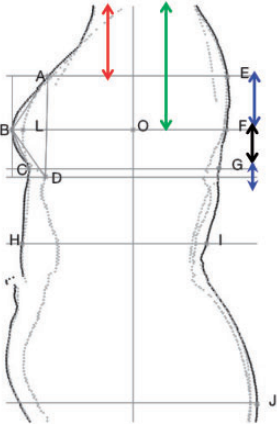| Illustration	Name	Description
Measure of distance		
	distCD	Distance between point C and point D
	distCE	Distance between point C and point E
	distCF	Distance between point C and point F
	distCG	Distance between point C and point G
	distAO	Distance between point A and point O
	distBO	Distance between point B and point O
	BP2BP	Distance between bust points (Distance between point A and point B)
Measure of area		
	rec_rt_inner	Area of the rectangle at right inner bust (with width CF and height AF)
	rec_lt_inner	Area of the rectangle at left inner bust (with width CG and height BG)
	rec_rt_outer	Area of the rectangle at right outer bust (with width DF and height AF)
	rec_lt_outer	Area of the rectangle at left outer bust (with width EG and height BG)
	rec_right	Area of the rectangle at right bust (with width CD and height AF)
	rec_left	Area of the rectangle at left bust (with width CE and 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 and line CD

(continued)

Table A-2. Continued

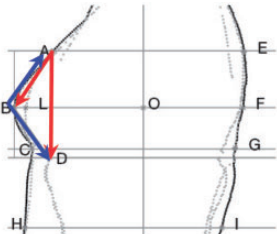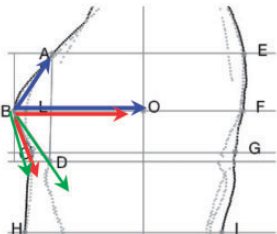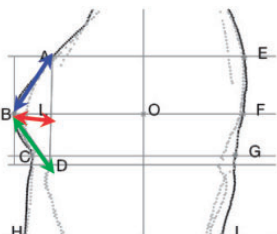| Illustration	Name	Description
	area_curveCBE	Area of the arc enclosed by curve CBE and line CE
	area_curveACF	Area of the fan enclosed by curve AC, line CF, and line AF
	area_curveBCG	Area of the fan enclosed by curve BC, line CG, and line BG
	area_curveADF	Area of the fan enclosed by curve AD, line DF, and line AF
	area_curveBEG	Area of the fan enclosed by curve BE, line GE, and line BG

Table A-3. Raw measurements extracted from sagittal planes

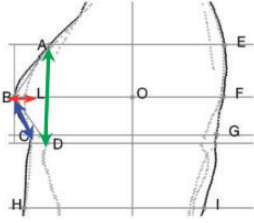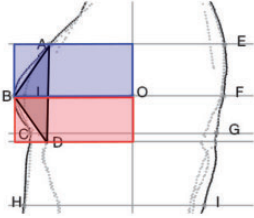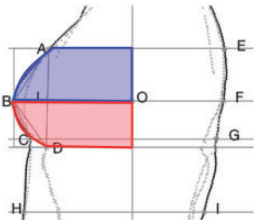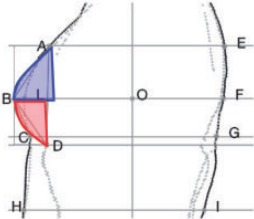| Illustration	Name	Description
Measure of thickness		
	thickAE	Body thickness at axilla level (length of line AE)
	thickBF	Body thickness at bust level (length of line BF)
	thickCG	Body thickness at underbust level (length of line CG)
	thickAO	The anterior body thickness at axilla level
	thickBO	The anterior body thickness at bust level
	thickCO	The anterior body thickness at underbust level
	thickDO	The anterior body thickness at breast root level
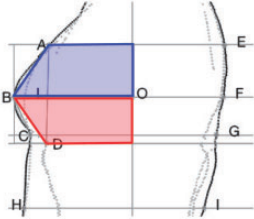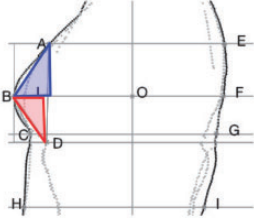
Table A-3. Continued

Illustration	Name	Description
Measure of height difference		
	heightEF	Height difference between axilla level and bust level
	heightFG	Height difference between bust level and underbust level
	heightGD	Height difference between underbust level and breast root level
	height_acro2BP_rt	Height difference between the right acromion point and the right bust point
	height_acro2BP_lt	Height difference between the left acromion point and the left bust point
	height_shd2armp_rt	Height difference between the right acromion point and the right axilla point
	height_shd2armp_lt	Height difference between the left acromion point and the left axilla point
Measure of angles		
	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		
	distAB	Distance between point A and point B
	distBD	Distance between point B and point D
	distB_AD	Distance between point B and line AD
	distAD	Distance between point A and point D

Table A-3. Continued

Illustration	Name	Description
	distBC	Distance between point B and point C
	distBL	Distance between point B and 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 and height EF)
	area_rec_lower	Area of the lower rectangle (with width BO and height FG)
	areaBupper	Area enclosed by curve AB, y-axis, line AE, and line BF
	areaBlower	Area enclosed by curve BD, y-axis, line BF, and the horizontal line passing through D
	area_curvAB_linA	Area of the fan enclosed by curve AB, the vertical line passing through point A, and line BF
	area_curvBD_linD	Area of the fan enclosed by curve BD, the vertical line passing through point D, and line BF
	area_trap_upper	Area of the trapezoid enclosed by line AB, line BF, y-axis, and line AE
	area_trap_lower	Area of the trapezoid enclosed by Line DB, line BF, y-axis, and the horizontal line passing through Point D
	area_tri_upper	Area of the triangle enclosed by line AB, line BF, and the vertical line passing through point A
	area_tri_lower	Area of the triangle enclosed by line BD, line BF, and the vertical line passing through Point D

Appendix C: variables constructed from raw measurements

	Name of variable	Description
1	circmR_deltaB_bust	The difference of bust and 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 of thickCO and thickDO, divided by thickDO
8	heightR_upperB_fullB	heightEF divided by the sum of heightEF and 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 of thickBO and 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 and 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 of area_curvAB_linA and area_tri_upper, divided by area_tri_upper
28	areaR_arc_tri_lower	The difference of area_curvBD_linD and 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 and ratio_lt, where: ratio_rt = depthAF divided by distCD ratio_lt = depthBG divided by distCE
35	heightR_armp2BP_shd2BP	The mean of ratio_rt and ratio_lt, where: ratio_rt = the difference of height_acro2BP_rt and height_shd2armp_rt, divided by height_acro2BP_rt ratio_lt = the difference of height_acro2BP_lt and height_shd2armp_lt, divided by height_acro2BP_lt
36	areaR_fan_rec_inner	The mean of ratio_rt and 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

(continued)

Continued

	Name of variable	Description
38	areaR_fan_rec_outer	The mean of ratio_rt and 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 and 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