Establishment of a PCA model for skin health evaluation

YiFan He, RuiZhen Wang, Hong Meng, Li Li, Zhemin Wu and YinMao Dong

Department of Cosmetic Science, School of Sciences, Beijing Technology and Business University, Beijing, PR China; Bard Medical Device (Beijing) Co, Ltd., Beijing, PR China

ABSTRACT
The aim of this study is to establish an objective skin evaluation system through testing skin data and scoring conducted by general dermatology experts. First, 18 data points including left cheek face water evaporation, and left cheek oil content are collected. Second, comprehensive skin information is obtained from volunteers by the mathematical model of principal component analysis (PCA), following scoring the status of the volunteers’ skin by dermatology experts. Then, 99 volunteers are randomly divided into two groups, namely model information and test information. Specifically, skin model information is obtained from 80 volunteers, and test information is obtained from 19 volunteers. PCA is used to extract 18 independent variables of skin test data by means of R Studio software, followed by principal component regression by expert scoring and the extracted principal components. Finally, based on the expert score and model score, skin grading classification can be conducted by MATLAB software, in order to intuitively and effectively evaluate skin status.

Introduction
A variety of analytical methods are used to describe qualitative characteristics of the skin creating a large amount of heterogeneous data. Statistically analyzing these data through classical methods gives important information on a number of different variables [1]. Traditional statistical analysis does not provide global knowledge regarding the relationships among the different variables, and does allow the grouping of samples with homogeneous characteristics [2]. Therefore, it is useful to synthesize the trend of a particular phenomenon through a few elements. A common and legitimate complaint directed at the multivariable control literature is that while the theory appears to be strong, it is not accompanied by strong numerical tools [3]. To meet this need, we can make use of multivariate statistical methods, including principal component analysis (PCA), which makes it possible to identify the most important directions of variability in a multivariate data matrix [4]. PCA is often the method of choice for reducing a large set of correlated variables to a smaller number of uncorrelated components. PCA is a statistical method that has been used in representative features to transform the input space into a new lower dimensional space [5]. It has been widely used to identify and summarize multiple inter-relationships among original variables [6–9]. Inter-correlated variables are combined into a smaller number of new variables called principal components [10]. The first principal component accounts for much of the variability in the data, and each succeeding component accounts for the remaining variability [11]. The uncorrelated variables are linear combinations of the original variables and the last of these variables can be removed with minimum loss of real data in order to identify new meaningful underlying variables [12]. The PCA technique has previously been investigated by researchers [13].

The main purpose of exploratory data analysis is to learn about interrelationships between objects and variables [14]. In large, complex data-sets, visualization is needed to achieve this aim. A well-known method for visualization of the data patterns is PCA [1–9]. In this study we use volunteers to obtain accurate, real-time skin information.

As explained by Nis et al. [15], the results for both variables and objects are presented. The abscissa corresponds to the first principal component and the ordinate corresponds to the second principal component. Samples to the right in the score plot have high values for variables to the right in the loading plot. The same holds for samples to the left, at the top, or at the bottom. Objects close together have similar characteristics; variables close together are positively correlated, while variables lying opposite to each other in the
loading plot tend to be negatively correlated. The further a variable is away from the axis origin, the better it is represented on the considered plane.

In order to systematically analyze a project through skin state project research, we need to set up multiple statistical indicators. However, in most cases, there is a specific correlation between different indexes. The analysis is complicated because there are multiple indicators and there is a specific correlation between indicators. PCA seeks to find a few unrelated principal components, so that they can, to the extent possible, retain the original variables [16]. PCA seeks to use reduction; in order to minimize the loss of original data information, the original indexes are combined into several composite indicators [17]. Each composite indicator is linear combinations of the original indicators, which retain the main information of the original indexes and are not related to each other. This simplifies the complex problem, making it easy to grasp the main contradictions within the analysis [18].

Methodology

Procedure

There are six procedures in this study [19]: (1) Identify the analysis variables and collect data; (2) standardize the processing of the original data, and eliminate the influence of the different dimensions; (3) obtain the covariance matrix of the normalized data, specifically the correlation matrix of the original data; (4) find the characteristic root, characteristic vector and variance contribution rate of the principal components; (5) extract the principal components; (6) seek the principal component value and calculate the comprehensive score.

Data

Data description

The data used in this research were obtained from female volunteers (N = 99) aged 21–50 years old at the Chinese Cosmetics Research Center of Beijing Technology and Business University.

The test had the following requirements. First, the test was done at a temperature of 22 ± 2 °C and the humidity was held at 50% ± 1%. Second, the participants had no serious systemic diseases, immune deficiencies or autoimmune diseases. Third, in accordance with the 2007 inclusion and exclusion criteria of ‘cosmetic contact dermatitis diagnostic criteria and principles of treatment,’ all participants were required to fill out an informed consent form and questionnaire. Additionally, under the guidance of professional staff, participants cleaned the test site with fresh water, and remained in the test environment for 20 min. On the test day, participants were in a healthy state and did not have colds, headaches, fevers or other symptoms. The test indicators and equipment are shown in Table 1.

Software

All programs used here are written in Rstudio and Matlab2009b computing environment.

Results and discussion

PCA model

All skin test data were randomly extracted from 80 of the 99 volunteers. First, according to Table 2, we determined

| No. | Indicator                          | Test equipment and specifications                  |
|-----|------------------------------------|----------------------------------------------------|
| 1   | Facial skin blood perfusion volume | Laser Doppler PeriScan PIM3                         |
| 2   | Facial temperature                 | Infrared thermal imaging Vario CAM HD              |
| 3   | Melanin content of skin            | Black and red pigment probe Mexameter MX18         |
| 4   | Red pigment content of skin        | Black and red pigment probe Mexameter MX18         |
| 5   | Average roughness Rz               | Skin texture meter Derma Top VISIO 3D              |
| 6   | Depth of smoothness Rp             | Skin texture meter Derma Top VISIO 3D              |
| 7   | Arithmetic average roughness Ra    | Skin texture meter Derma Top VISIO 3D              |
| 8   | Skin moisture content              | Water content Corneometer CM 825                  |
| 9   | TEWL contents                      | Water loss Tewamater TM300                         |
| 10  | Oil content                        | Oil content Subumeter SM810                        |
| 11  | Skin colour                        | Skin color MPA 9                                  |
| 12  | Skin glossiness                    | Skin glossiness GL190                             |
| 13  | Left cheek elasticity R2           | Skin elasticityMPAS80                             |
| 14  | Left cheek elasticity R5           | Skin elasticityMPAS80                             |
| 15  | Left cheek elasticity R7           | Skin elasticityMPAS80                             |
| 16  | Left cheek pH value                | PH 905                                             |
| 17  | Average pO2                         | PeriFlux 5000                                     |
| 18  | Average pCO2                       | PeriFlux 5000                                     |

Table 2. Introduction of test item.

| Item                          | Introductions                                           |
|-------------------------------|---------------------------------------------------------|
| Melanin (MI)                  | The smaller the value, the better (take the first 5% quantile) |
| Haematochrome (EII)           | The smaller the value, the better (take the first 5% quantile) |
| Water content                 | The larger the value, the better (take the first 95% quantile) |
| Water loss                    | The smaller the value, the better (take the first 95% quantile) |
| ITA (chromaticity)            | The larger the value, the lighter the colour (take the first 95% quantile) |
| Oil                           | The better approach is in the middle (take the first 50% quantile) |
| Lustrousness                  | The larger the value, the better the glossiness (take the first 95% quantile) |
| RA/RP/RZ                      | The closer to 0, the better the skin texture (take 0)   |
| Blood perfusion volume        | The larger the value, the better the skin (take the first 95% quantile) |
| Skin temperature              | The larger the value, the better the skin (take the first 95% quantile) |
optimal values of the data from the 99 volunteers, in order to give descriptive statistics. The optimal index value and the corresponding variables $X_1-X_{18}$ are shown in Table 3.

PCA is used in data reduction, in which a set of observed variables, or factors, are analyzed in order to identify a small number of factors that underlie the pattern of correlations and explain most of the variance observed in the larger number of manifest variables. The PCA procedure is characterized by a high degree of flexibility. Finally, by using the PCA method, we can set up a relationship between the expert score and 18 indexes as follows:

$$y = 10.2612 - 0.9412X_1 - 1.336X_2 + 0.0478X_3$$
$$-0.0175X_4 - 0.0128X_5 + 0.0891X_6 + 0.6693X_7$$
$$+5.7207X_8 + 5.0790X_9 + 7.4402X_{10} + 2.2292X_{11}$$
$$-183.8901X_{12} - 44.1138X_{13} - 23.8804X_{14}$$
$$+0.0271X_{15} - 0.0948X_{16} + 0.0143X_{17} + 0.4662X_{18}$$

The goodness of fit is the fitting degree of the regression line based on the observed value. The statistics of the goodness of fit measure is $R$, which is a determining factor with a value range of $[0, 1]$. When the $R$ value is close to 1, the fitting degree of the regression line on the observations is better; on the other hand, when the value of $R$ is close to 0, the fitting degree of the regression line observations is worse [20–24].

Further, when the $R$ value for the results reaches more than 90%, we believe the method is feasible. The 18 independent variables are substituted into the main components of the regression equation, and we can obtain the corresponding model score. The 19 sets of test data and 80 sets of training data are substituted into the main components of the regression equation, resulting in a model score of 99 people matched with expert scores, for trade-off.

One can calculate the comprehensive score based on the above model of substituting the test data set (all skin test information for the 99 volunteers). Then, through comparing the model score of 80 volunteers with the scores of 99 volunteers using EXCEL, we find that the matching degree of the two groups is more than 90%, so the accuracy of the model is ideal. This means that PCA works well.

### Classification of skin

Calculating the comprehensive score of the data of the 99 volunteers, we sort the score from high to low and classify the score. The standard of classification is: Skin score greater than 64 is ‘excellent,’ greater than 36.5 but less than or equal to 64 is ‘good,’ greater than 27.5 but less than or equal to 36.5 is ‘medium,’ less than or equal to 27.5 is ‘poor.’ Using the state of the skin, we define ‘excellent’ as ‘good skin,’ the ‘good’ and ‘medium’ as ‘sub-healthy skin,’ and the ‘poor’ as ‘problem’ skin. We first compile the Matlab program and then replace the value based on the new data.

### General analysis

The starting point was from a recent contribution by Vasta et al. [25] describing three-dimensional and planar models of hyper-elastic fiber reinforced materials characterized by statistical distribution of the fiber orientation. The mechanical behaviour of the models is assessed through uniaxial, biaxial and shear tests [25]. The goal is to numerically quantify the stress induced on a scar of a human columnella by a constant load, through a fine-tuned finite elasticity continuum model by Gizzi et al. [26]. The proposed mathematical model can be applied both to theoretically designed and numerically verified new non-conventional scar geometries [26].

Then, we introduced clustered block-wise PCA for representation of visual data [27]. Clustered block-wise PCA not only takes advantage of the fact that typical visual data contains localized variations, but also exploits the correlations within the data [28]. We have shown that clustered block-wise PCA achieves higher efficiency than linear regression in terms of both storage and computational cost [29]. As a result, we have successfully scaled PCA for use in large problems without losing any of the inherent advantages of PCA [30].

In order to accurately describe the expert score using the machine score, first the 99 data-sets are divided into two categories: 80 data-sets for modeling and 19 data-sets are used as a test. Second, for the 80 data-sets, R studio was used to extract the main components of the 18 independent variables, and then the expert score and

### Table 3. Optimal index value and meaning of volunteer data.

| Index Variable                      | Optimal value |
|-------------------------------------|---------------|
| Average moisture content of left cheek | X_1           |
| Water loss amount of left cheek     | X_2           |
| Left cheek grease                   | X_3           |
| MI average of left cheek            | X_4           |
| El average of left cheek            | X_5           |
| ITA value of left cheek             | X_6           |
| Average cheek luster                | X_7           |
| Left cheek elasticity R2            | X_8           |
| Left cheek elasticity R5            | X_9           |
| Left cheek elasticity R7            | X_10          |
| pH value of left cheek              | X_11          |
| Average cheek Rz (mm)               | X_12          |
| Average cheek Rz (mm)               | X_13          |
| Average cheek Rp (mm)               | X_14          |
| Average pO2(unit)                   | X_15          |
| Average pCO2(unit)                  | X_16          |
| Average Laser Doppler               | X_17          |
| Average left cheek infrared         | X_18          |

### General analysis

The starting point was from a recent contribution by Vasta et al. [25] describing three-dimensional and planar models of hyper-elastic fiber reinforced materials characterized by statistical distribution of the fiber orientation. The mechanical behaviour of the models is assessed through uniaxial, biaxial and shear tests [25]. The goal is to numerically quantify the stress induced on a scar of a human columnella by a constant load, through a fine-tuned finite elasticity continuum model by Gizzi et al. [26]. The proposed mathematical model can be applied both to theoretically designed and numerically verified new non-conventional scar geometries [26].

Then, we introduced clustered block-wise PCA for representation of visual data [27]. Clustered block-wise PCA not only takes advantage of the fact that typical visual data contains localized variations, but also exploits the correlations within the data [28]. We have shown that clustered block-wise PCA achieves higher efficiency than linear regression in terms of both storage and computational cost [29]. As a result, we have successfully scaled PCA for use in large problems without losing any of the inherent advantages of PCA [30].

In order to accurately describe the expert score using the machine score, first the 99 data-sets are divided into two categories: 80 data-sets for modeling and 19 data-sets are used as a test. Second, for the 80 data-sets, R studio was used to extract the main components of the 18 independent variables, and then the expert score and
the extracted principal components are carried out through principal component regression. From running the program, the fitting degree equation reached above 0.9, and then using the model, we obtained the model score for the 80 data-sets. By comparing the expert scores and the model scores, we can see that the expert and model scores have a very high matching degree ($R$ is greater than 0.9). Therefore, the established model is able to explain the relationship between the expert score and the 18 independent variables.

Then, we sort the model scores from high to low, classify the scores, and define the skin as ‘good skin,’ ‘subhealthy skin,’ and ‘problem skin.’ This allows the volunteers to very clearly understand their skin state.

Therefore, this model of skin data analysis can objectively, accurately and comprehensively express the volunteers’ skin condition. Through this process, the volunteers can clearly understand their comprehensive skin condition, which can help them select skin care products in the future. Even if it does not provide analytical information derived from other statistical methods, PCA is a very effective procedure for obtaining a synthetic judgment of skin quality.

In future work, we will investigate automatic methods for estimating the optimal spatial and temporal block size, allowing for variance across data volumes. This unsolved problem hinges upon the availability of novel practical algorithms.

**Conclusions**

There are several points in this paper, which can be explored more deeply in further study. First, we address the problem of generalizing a composite indicator which is linear combinations of the original indicators. Second, this work is strongly biased towards operating directly on signals whenever possible. This means, of course, a bias against working with secondary objects such as model parameters. There are at least two reasons for this bias. First, limitations of physical hardware can usually be stated directly in terms of signals. Second, tools for coping with multiple signals (principal component analysis ingular value decomposition) are available. Overall, the present work shows that this model is more objective and time-saving than other methods. This model can help consumers to choose appropriate cosmetic products and also help cosmetic companies develop more targeted products for different consumers.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This study was supported by the Open Research Fund Program of Beijing Key Lab of Plant Resource Research and Development, Beijing Technology and Business University.

**References**

[1] Davis JC. Statistics and data analysis in geology. New York (NY): JWiley; 1973.
[2] Ziegel ER, Siegel AF, Morgan CJ. Statistics and data analysis. New York (NY): Wiley; 2006.
[3] Meredith WM, Frederiksen CH, Mclaughlin DH, et al. Statistics and data analysis. Annu Rev Psychol. 1974;39(25):453–505.
[4] Morgan, CJ. Statistics and data analysis. New York (NY): Wiley; 1996.
[5] Rice JA. Les informations continues dans cette page sont à usage strict de et ne doivent êtreutiliséesoucopiées par un tiers [Mathematical statistics and data analysis. OR Technometrics. 2012;72(462):633–649.
[6] Chatfield C, Collins AJ. Introduction to multivariate analysis. New York (NY): Chapman and Hall; 1980.
[7] Dunteman GH. Introduction to multivariate analysis. 1984;4(78):221–225.
[8] van de Geer JP. Introduction to multivariate analysis for the social sciences. Contemp Sociol. 1973;2(1).
[9] Kobayashi H. An introduction to multivariate analysis. Nara Womens Uni Sociol Stud. 1999;6:143–157.
[10] Atkinson AC. Plots, transformations and regression. New York (NY): Oxford University Press; 1985.
[11] Ferrigno G, Camenval P. Principal component analysis of chest wall movement in selected pathologie. Med Biol Eng Comp. 1998;36:445–451.
[12] Salaffi F, Manganeli P, Carotti M, et al. The differing patterns of subclinical pulmonary involvement in connective tissue diseases as shown by application of factor analysis. Clin Rheumat. 2000;19:35–41.
[13] Marek S, Pniewski, Emilia K, et al. Pattern recognition methods in evaluation of the structure of the laboratory data biominerals, antioxidant enzymes, selected biochemical parameters, and pulmonary function of welders. Biol Trace Elem Res. 2003;93:39–46.
[14] Amaz M, Robert XG. PCA-based feature selection scheme for machine defect classification. IEEE Trans Inst Meas. 2004;53:1517–1525.
[15] Niis T, Baardseth P, Helgesen H, et al. Multivariate techniques in the analysis of meat quality. Meat Sci. 1996;43(51):135–149.
[16] Islam MQ, Tiku ML. Multiple linear regression model under nonnormality. Commun Stat Theor Methods. 2005;33 (10):2443–2467.
[17] Yang J, Zhang D, Frangi AF. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Intell. 2004;26 (1):131–137.
[18] D’Apremont A, El Ghaoui L, Jordan M. A direct formulation for sparse PCA using semidefinite programming. Siam Rev. 2007;49(3):434–448.
[19] Haque MM, Rahman A, Hagare D. Principal component regression analysis in water demand forecasting: an
application to the blue mountains, NSW, Australia. Veterinary Pathol. 2013;45(6):842–848.

[20] Standardized regression coefficient. In: Michalos AC, editor. Encyclopedia of quality of life and well-being research. Dordrecht (Netherlands): Springer; 2014. p. 271.

[21] Bring J. How to standardize regression coefficients. Am Statistician. 1994;48(3):209–213.

[22] Greenland S, Maclure M, Schlesselman JJ, et al. Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology. 1991;2(5):387–392.

[23] Nimon KF, Oswald FL. Understanding the results of multiple linear regression beyond standardized regression coefficients. Organ Res Methods. 2013;16(4):650–674.

[24] Deegan J. On the occurrence of standardized regression coefficients greater than one. Edu Psychol Meas. 1978;38(4):873–888.

[25] Vasta M, Gizzi A, Pandolfi A On three- and two-dimensional fiber distributed models of biological tissues. PEM. 2014;37(4):170–179.