Introduction to Principal Components Analysis

Kristin L. Sainani, PhD

Principal components analysis (PCA) is a powerful statistical tool that can help researchers analyze datasets with many highly related predictors. PCA is a data reduction technique—that is, it reduces a larger set of predictor variables to a smaller set with minimal loss of information. PCA may be applied before running regression analyses or for exploratory purposes to help researchers understand relationships among their variables or discover patterns in their data. Although it is widely used in certain domains, such as dietary studies and questionnaire development, it is underused in other domains. In this article I first provide examples that illustrate the utility of PCA and then delve into greater detail about how it works.

PCA: AN INTRODUCTORY EXAMPLE

To illustrate the value of PCA, I will apply the technique to some data pertaining to 91 female adolescent runners. These data come from a larger real dataset [1] but have been modified for use in classroom examples, and thus the results presented here are only for teaching purposes.

The researchers measured a large number of potential predictors of bone density and stress fractures, including variables relating to body size and composition, training, performance, menstrual function, diet, and eating behaviors. PCA can be applied here to explore relationships among these many predictors and/or to reduce the total number of predictors before performing regression analyses.

To keep this example manageable, I will focus on just 11 variables that encompass 3 different domains: body size and composition (android fat ratio, fat mass, lean mass, body mass index, and height); training and performance (mile time, mileage, and interval pace); and menstrual function (periods in the past year, average periods since menarche, and age at menarche).

An examination of the correlation coefficients between these 11 variables shows that many are highly related, even across different groups of variables. For example, better menstrual function (normal menarche age and more regular periods) correlated with a higher weight and amount of fat (with correlations as high as 0.30), whereas better running performance correlated with a lower weight and amount of fat (with correlations as high as 0.57).

The application of PCA to these data before performing regression analyses has several benefits. The researchers have collected multiple measurements that reflect the same construct—for example, running performance or menstrual function. If they enter all of these variables into a regression model, they increase the risks of overfitting and type I errors (chance findings). However, if they ignore or discard variables, they lose valuable information. PCA analysis instead compresses the 11 original variables into a smaller subset of composite variables (called principal components) that capture most of the information in the original data. These new variables may then be used for further analyses. They have the added benefit of being uncorrelated with one another, which makes them easier to interpret and analyze. Besides compressing the data, PCA analysis also reveals clusters of variables that are highly related, which can give investigators a deeper understanding of their data.

Before performing PCA analysis on the example data, I first imputed missing values, because observations that have any missing data will otherwise be omitted. I also converted the variables into standard deviation units (Z scores), because all the variables need to have the same units before PCA is applied.

K.L.S. Stanford University, Department of Health Research and Policy, Division of Epidemiology, HRP Redwood Bldg, Stanford, CA 94305. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

The initial PCA analysis creates 11 composite variables (or principal components) from the 11 original variables. These composite variables are constructed in such a way that the first few contain most of the information in the original data. In this example, the first 4 components explained 76% of the variability in the original 11 variables. I decided to retain only these first 4 components. Thus I was able to reduce my predictor set from 11 to 4 while losing only a quarter of the total information. The choice of how many variables to retain is beyond the scope of this article; for more information, I direct the reader to a review by Abdi and Williams [2].

Table 1 shows the resulting 4 variables. Each new variable is a weighted sum of the original 11 variables. The weights—which can range from −1 to +1—are shown in the columns of the table. The weights indicate the relative contribution of each variable to each component, which allows us to surmise the meaning of each component. For example, android fat ratio (a marker of visceral fat), fat mass, and body mass index have weights of 0.40, 0.28, and 0.41, respectively, for principal component 1, whereas the other variables have weights close to 0 for this component. These data suggest that principal component 1 is a measure of body fatness. (An important technical detail is that I have applied a "varimax rotation" to these components. Application of this rotation is common because it forces variables to align strongly with only a single component, thus making the results more interpretable. I refer the reader to the review by Abdi and Williams [2] for further discussion.)

Principal component 2 has high weights for faster mile time (0.35), mileage (0.43), and faster interval pace (0.35) and thus represents a measure of running competitiveness. Principal component 3 has high weights for lean mass (0.43) and height (0.59) but not for the other body measures, and thus it reflects body size separate from body fatness. Component 4 has high weights for present (0.54) and past (0.55) menstrual regularity. Surprisingly, menarche age contributes strongly to component 2 (weight = 0.27) but very little to component 4 (weight = 0.06). These data provide an unexpected insight: in this population of adolescent women runners, later age at menarche was more related to running competitiveness than to subsequent menstrual function.

The 4 components are uncorrelated with one another (ie, all correlation coefficients equal 0), meaning we have disentangled the different constructs. For example, body size and body fatness, which normally overlap considerably, have now been separated into 2 independent variables. We have also separated the construct of body fatness from both menstrual function and running competitiveness.

Each woman in the dataset gets a score for each of the four new variables (body fatness, running performance, body size, and menstrual regularity) based on her values for the original 11 variables. Table 2 provides an example calculation. The scores are in standard deviation units. For example, a woman who has a value of 1.6 for principal component 1 (body fatness) is 1.6 standard deviations above average for this variable.

The new variables are then used in further analyses. For example, I tested their association with stress fractures in a logistic regression model (Table 3). I found that running competitiveness significantly increased the risk of fracture, whereas menstrual regularity reduced the risk. Bigger body size (which is strongly related to higher bone density) also protected against fracture but did not reach statistical significance. In contrast, body fatness was not significantly related to fracture.

Building a multivariate regression model with the original variables is complicated; you must decide which variables to include and which to prune. In contrast, building the regression model with principal components is simple. Because the components are uncorrelated, you get the same

Table 1. *Results of the principal components analysis showing the weights for each variable**

	Principal component 1 = body fatness	Principal component 2 = running competitiveness	Principal component 3 = body size	Principal component 4 = menstrual regularity
Android fat ratio	0.40†	0.09	−0.19	0.03
Fat mass	0.28†	−0.02	0.12	−0.02
Body mass index	0.41†	0.10	−0.07	−0.06
Lean mass	0.09	0.13	0.43†	0.01
Height	−0.19	−0.05	0.59†	0.07
Mile time (higher is faster)	−0.05	0.35†	−0.02	0.05
Mileage	0.10	0.43†	0.01	0.06
Interval pace (higher is faster)	0.05	0.35†	−0.05	0.01
Menarche age	0.08	0.27†	0.09	0.06
Periods in the past year	0.00	0.09	0.00	0.54†
Periods since menarche	−0.07	0.04	0.08	0.55†

*The meaning of each principal component (eg, body fatness) is determined after the principal components analysis. To surmise the meaning of the principal components, researchers typically examine the correlations between the original variables and the principal components (called the "factor loadings") rather than the weights, but the patterns revealed are similar.

†One of the variables that contributes most strongly to each component.

Table 2. *An example of how a principal component score is calculated for a particular woman**

Original variable	Value for a particular woman (in SD units)	Weight for principal component 1 ("body fatness")	Weight × value
Android fat ratio	1.5	0.40	0.40 × 1.5
Fat mass	1.3	0.28	0.28 × 1.3
Body mass index	1.0	0.41	0.41 × 1
Lean mass	0.3	0.09	0.09 × 0.3
Height	−0.4	−0.19	−0.19 × −0.4
Mile time (higher is faster)	−0.6	−0.05	−0.05 × −0.6
Mileage	0.2	0.10	0.10 × 0.2
Interval pace (higher is faster)	0.5	0.05	0.05 × 0.5
Menarche age	1.0	0.08	0.08 × 1
Periods in the past year	1.0	0.00	0.00 × 1
Periods since menarche	0.8	−0.07	−0.07 × 0.8
Principal component 1 ("body fatness") score	—	—	1.6

SD = standard deviation.

*The last column is summed to get the woman's total score on principal component 1. The principal component scores will have a mean of 0 and an SD of 1; thus they are in SD units. A score of 1.6 indicates that this woman is 1.6 SDs above average in body fatness.

results regardless of which components are included in the model (for example, the odds ratio for running competitiveness will remain 2.48 in a model that contains only this variable). Furthermore, because I tested 4 variables rather than 11, I substantially reduced the risk of chance findings. The results are also easier to interpret because the separate constructs have been clearly isolated. Finally, the results may be more robust because the composite variables may capture a particular construct (such as body size) better than any of the original variables.

EXAMPLE 2: USING PCA TO IDENTIFY DIETARY PATTERNS

One of the most common uses of PCA in the medical literature is for dietary studies. Researchers often collect dietary data by using food-frequency questionnaires, which ask subjects how often they eat different food items. The resulting dataset may contain tens or hundreds of variables (at least one for each food item). PCA can be used to identify a small set of common "dietary patterns"—that is, clusters of food types that tend to track together. For example, people who frequently eat hamburgers may also consume a lot of soda and fried foods.

Table 3. *Odds ratios for stress fracture for each of the four new variables (body fatness, running performance, body size, and menstrual regularity), calculated with logistic regression*

Predictor variable	Odds ratio for stress fracture	*P* value
Body fatness (component 1)	0.66	.38
Running competitiveness (component 2)	2.48	.02
Body size (component 3)	0.39	.07
Menstrual function (component 4)	0.35	.01

An example is illustrated by Link et al [3], who collected dietary information on 118,465 women from the California Teachers Study. These investigators used PCA to identify 5 predominant dietary patterns: a plant-based dietary pattern that was high in fruits and vegetables; a high-protein, high-fat pattern that was high in butter, meat, and fried foods; a high-carbohydrate pattern that was high in convenience foods, breads, and pasta; an "ethnic pattern" that was high in legumes, soy, rice, and dark-green leafy vegetables; and a "salad and wine" pattern that was high in fish, wine, lettuce, low-fat salad dressing, and coffee. Each woman received a score for each dietary pattern. Women with high scores on the "salad and wine" pattern had a significantly higher risk of developing breast cancer than did women with low scores for this pattern.

THE MECHANICS OF PCA

But how does PCA actually come up with the weights for the composite variables? The mechanics of PCA are best illustrated with a simple example. Returning to the example dataset on female adolescent runners, consider just two variables: fat mass and weight. They are highly related, with a correlation coefficient of 0.86. Given that the variables are so similar, we could simply choose to focus on one and ignore the other. However, this approach results in an unnecessary loss of information. With PCA, we can instead create a single composite variable that extracts the maximal possible information from both variables. This new variable is a weighted sum of a woman's weight and fat mass (W_{fm} = weight for fat mass; W_{wt} = weight for weight):

$$\text{Component 1} = W_{fm} \times \text{fat mass} + W_{wt} \times \text{weight}$$

PCA finds the weights W_{fm} and W_{wt} such that component 1 has the maximal possible variance. To understand this intuitively, a picture is helpful. Figure 1 (left panel) shows a scatter plot of fat mass and weight. Think of the cloud of data points as an oval and draw a line along the direction of the greatest spread (variability). This is the first principal component (represented as a solid line in the graph) because it spans the highest amount of variance in the data and thus encompasses the greatest amount of information. A woman's value for component 1 is just her distance from 0 along this line. It is easier to figure out this distance by rotating the scatter plot so that principal component 1 lines up with the X-axis (Figure 1, right panel).

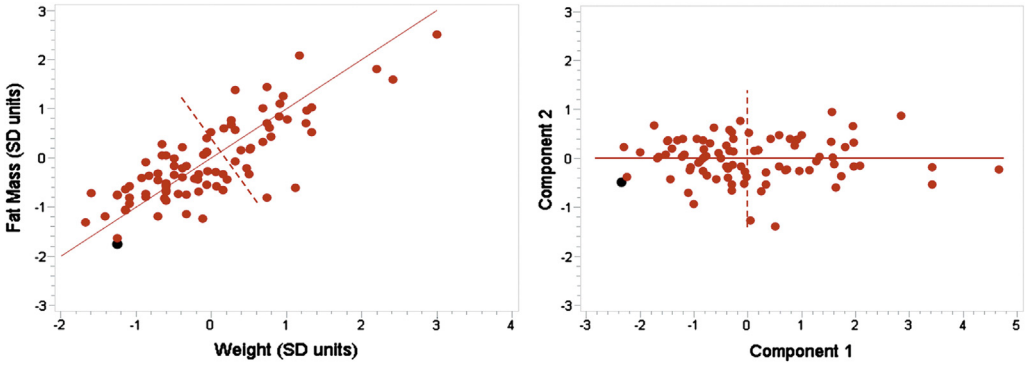

Figure 1. The left panel shows the scatter plot between fat mass and weight (both given in standard deviation (SD) units), with component 1 (solid line) and component 2 (dashed line) superimposed. In the right panel, the scatter plot has been rotated so that component 1 aligns with the X-axis. The data point with fat mass = −1.74 and weight = −1.26 (left panel) has been highlighted in black; this observation has coordinates −.34, −2.12 in the rotated plot (right panel). Thus this runner has a value of −2.12 for component 1 (X-axis value) and −.34 for component 2 (Y-axis value).

The equation $W_{fm} \times$ fat mass $+ W_{wt} \times$ weight gives a woman's X coordinate in the rotated plot.

In this example, the PCA yields weights of 0.707 for fat mass and 0.707 for weight. Thus a woman with a fat mass of −1.74 standard deviations and a weight of −1.26 standard deviations (pictured as a black dot in both panels of Figure 1) has an X coordinate in the rotated plot of X = −2.12; this is also her score on principal component 1.

Because we have 2 variables in this simple example, we can extract 2 principal components. The second component must be perpendicular to the first, and thus in 2 dimensions, there is only one possible line (Figure 1, dotted line). In higher dimensions, however, the second component will be the line that spans the greatest remaining variability in the data in a direction perpendicular to component 1. The third component will be the line of maximal variance in a direction perpendicular to the other 2 components, and so on. Because the lines are perpendicular, the extracted components will not be correlated.

Because weight and fat mass are both in standard deviation units, they each have a variance of 1, for a total variance of 2. Principal component 1 has a variance of 1.86. Thus this single variable accounts for 93% (1.86/2) of the total variance—or total information—from the original 2 variables. The second component only accounts for 7% of the variance; visually, you can see that little scatter exists in the Y direction in the rotated plot. We may choose to discard component 2 given that it does not capture much additional information.

CONCLUSIONS

PCA produces a parsimonious set of predictor variables for regression analysis and provides insights into patterns and relationships in the data. It has been used in the physical medicine and rehabilitation literature for questionnaire design and more recently for human kinematics and biomechanics studies [4]; however, it likely has many more useful applications. Researchers dealing with datasets that involve large numbers of highly correlated predictors should consider whether PCA may be helpful in their analyses.

REFERENCES

1. Tenforde AS, Sayres LC, McCurdy ML, Sainani KL, Fredericson M. Identifying sex-specific risk factors for stress fractures in adolescent runners. Med Sci Sports Exerc 2013;45:1843-1851.
2. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisc Rev Comput Stat 2010;2:433-459.
3. Link LB, Canchola AJ, Bernstein L, et al. Dietary patterns and breast cancer risk in the California Teachers Cohort Study. Am J Clin Nutr 2013;98:1524-1532.
4. Laffaye G, Bardy BG, Durey A. Principal component structure and sport-specific differences in the running one-leg vertical jump. Int J Sports Med 2007;28:420-425.