ABSTRACT

Introducing principal components to students is difficult. First, the matrix algebra and mathematical maximization lemmas are daunting, especially for students in the social and behavioral sciences. Second, the standard motivation involving variance maximization subject to unit length constraint does not directly connect to the “variance explained” interpretation. Third, the unit length and uncorrelatedness constraints of the standard motivation do not allow re-scaling or oblique rotations, which are common in practice. Instead, we propose to motivate the subject in terms of maximizing (weighted) average proportions of variance explained in the original variables; this approach is easier to understand because it links directly to the familiar “R-squared” statistic. It also removes the need for unit length and uncorrelatedness constraints, provides a direct interpretation of “variance explained,” and provides a direct answer to the question of whether to use covariance-based or correlation-based principal components. Further, the presentation can be made without matrix algebra or optimization proofs. Modern tools from data science, including heat maps and text mining, provide further help in the interpretation and application of principal components; examples are given. Together, these techniques suggest major revision of currently used methods for teaching and learning principal components in the behavioral sciences.

KEYWORDS

Factor analysis, Rotation, Optimality, Heat Map, Variance explained.
Introduction

Principal Components Analysis (PCA) is a commonly used multivariate technique for data reduction, assessing dimensionality, and identifying latent structure (examples, respectively, include Pierella et al., 2015; Debelak and Tran, 2013; and Yeh et al., 2010). PCA has a long history in psychometrics, mostly as a starting point for factor analysis, going back at least as far as Spearman (1904). PCA is used less these days as a tool for estimating factor analysis models (maximum-likelihood and related methods are preferred in the modern era); nevertheless, PCA remains a standard tool for exploratory factor analysis. Bentler and DeLeeuw (2011) show strong correspondences between PCA and factor analysis (FA), while Harlow et al. (2013) show that PCA/FA methods are prevalent in current research. Additionally, PCA remains a viable competitor for FA estimation, as it is immune to problems of non-convergence, Heywood cases, and non-identifiability that frequently occur with maximum-likelihood and other estimation procedures.

With the current “data science” revolution, PCA is more important than ever, with applications to pattern recognition and image analysis (Patel et al., 2016; El-Dahshan et al., 2014; Gao and Wilson, 2013), predictive modeling (Shaheen and Khan, 2016), gene expression (Liu and Calhoun, 2014; Yeung and Ruzzo, 2001), molecular classification (Gosav et al., 2011), social network analysis (de-Marcos et al., 2016), and bibliometrics (Wolfram and Zhao, 2014; Ebrahimy and Osareh, 2014). These new data science methods have crossed over into recent multivariate behavioral research applications as well. Dai et al. (2015) use PCA in a behavior prediction model; Li (2016) uses PCA to reduce the extreme dimensionality of electro-encephalogram (EEG) data; and Yan et al. (2014) use PCA to reduce dimensionality of the perceptions of social robots.

Researchers typically understand how to use PCA very well: Data scientists simply plug the derived principal component scores into the desired analysis (typically clustering or predictive modelling), and
psychological researchers typically examine (rotated) loading patterns to identify underlying human traits. What researchers, teachers, and students typically do not understand very well is what the scores and loadings mean, on their own merits. The standard methods for introducing PCA are not only complicated mathematically, but also seemingly unrelated to their actual use, because they do not answer questions such as “What does ‘variance accounted for’ mean”? and “Why is the unit length constraint necessary, when we scale the PCs arbitrarily after we get them”? and “When is PCA based on the correlation matrix better than PCA based on the covariance matrix”? and “What good can be stated about rotated PC solutions”? Thus the subject difficult both to teach and to learn.

In multivariate statistics classes, PCA is often taught just before FA. If PCA itself is poorly understood by students, then it is all the more difficult for them to understand FA and related methods that are presented subsequently. On the other hand, PCA is typically taught after regression analysis, so students are familiar with the concept of “proportion of variance explained,” it being the usual R-squared statistic. So rather than explain principal components in terms of variance maximization subject to unit length and uncorrelatedness constraints, we instead propose to motivate PCA entirely in terms of (weighted) average R-square maximization, with the variance maximization results being presented as a side note, if at all. This all can be done with no matrix algebra or maximization theorems, as we will show.

Correlation-based methods for introducing and interpreting PCA are well known to theorists, but the various results are scattered throughout a literature that is in some cases obscure. In this tutorial we bring these results together in a single paper, and argue they are the most natural and easiest to teach. We also introduce a correlation-based variable pairs heat map to help students, educators, and researchers identify what the PC scores specifically measure. In this presentation, we assume familiarity with basic statistics up through regression, especially correlation and coefficient of determination (R-
squared), and with correlation and covariance matrices, but deeper knowledge of matrix algebra is not needed.

The Conventional Approach to Introducing Principal Components and the “Variance Accounted For”

Analysis based on the covariance matrix

Let the \( p \) measurements be called \( Y \), with individual variables \( Y_i, i = 1, \ldots, p \), and let \( S \) denote the \((p \times p)\) covariance matrix of data consisting of \( n \) observations on the \( p \) measurements \( Y_i \). For the purposes of the presentation here, all statistics are sample-based, but the discussion is completely analogous for the “population” case.

The standard method for introducing principal components, found in Duda et al. (2001), Everitt and Hothorn, (2011), Härdle and Simar, (2007), Jackson (2003), Johnson and Wichern (2007), Jolliffe (2002), is to solve the constrained maximization problem:

Choose a linear function \( L_1 = a_{11}Y_1 + \cdots + a_{p1}Y_p \) such that the variance of \( L_1 \) is a maximum, subject to the unit length constraint

\[
a_{11}^2 + a_{21}^2 + \cdots + a_{p1}^2 = 1. \tag{1}
\]

A solution to (1) is then given as \((a_{11}, \ldots, a_{p1}) = (e_{11}, \ldots, e_{p1})\), the eigenvector corresponding to the largest eigenvalue \( \lambda_1 \) of the covariance matrix \( S \).

The remaining PCs are then defined as in (1), but recursively for \( i = 2, \ldots, p \), subject to additional uncorrelatedness constraints:
Choose a linear combination $L_i = a_{i1}Y_1 + \cdots + a_{ip} Y_p$ such that the variance of $L_i$ is a maximum, subject to the constraints that $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1$, and that the correlation between $L_i$ and all previously found $L_j$ ($j = 1, \ldots, i-1$) is 0.

A solution to (2) is then shown to be $(a_{i1}, \ldots, a_{ip}) = (e_{i1}, \ldots, e_{ip})$, the eigenvector corresponding to the $i$th largest eigenvalue $\lambda_i$ of $S$. The PC scores are then given as $PC_i = L_{ii}$ and the following mathematical identities are used to interpret them:

Identity 1. The variance of $PC_1 + \cdots + PC_i$ is equal to $\lambda_1 + \cdots + \lambda_i$.

Identity 2. The variance of $PC_1 + \cdots + PC_p$ is equal to $\lambda_1 + \cdots + \lambda_p$.

Identity 3. The sum of the $\lambda$'s is also equal to the sum of the variances of the individual $Y$ variables: $\lambda_1 + \cdots + \lambda_p = \text{Var}(Y_1) + \cdots + \text{Var}(Y_p)$.

The standard presentation is then to cobble together Identities 1.-3., and claim that first $i$ PCs “account for” $100(\lambda_1 + \cdots + \lambda_i)/(\lambda_1 + \cdots + \lambda_p)$% of the variance of the $Y$ variables.

Analysis based on the correlation matrix

First define the standardized variables as $Z$. Specifically, the standardized $j$th observation on the $i$th variable is given by

$$Z_{ji} = (Y_{ji} - m_i)/s_i,$$

and note that the covariance matrix of these $Z$ variables is identical to the correlation matrix $R$ of the $Y$ variables. The motivation then proceeds exactly as above, but with the $Y$ variables replaced by the $Z$
variables. In this case it is often stated that $100(\lambda_1 + \cdots + \lambda_i)/(\lambda_1 + \cdots + \lambda_p)\%$ of the “standardized variance” of the $Y$ variables is “accounted for” by the first $i$ PCs.

Problems with the conventional approach

Problems with this approach to presenting PCA include the following.

i. PCs are commonly presented as components that “best represent” the original variables. But the connection between maximizing variance as given in (1) and “best representation” is unclear.

ii. In practice, the coefficients $a_{ij}$ are typically adjusted to ease interpretation. They may be proportionally rescaled so that the average is 1.0, or 10, or any other number, or they may be rescaled to make the coefficient $a_{ij}$ of a variable $Y_i$ that is of interest equal to 1.0. Such coefficients lose the unit length constraint, so the optimality property specified by (1) and (2) is clearly lost. Yet still, they are considered “good.” But good in what way? The optimality property given by (1) and (2) is the wrong one for connecting PCs to this usage.

iii. How does one go from Identities 1.-3. above to get the “variance accounted for” interpretation? These mathematical identities do not make this connection. In addition, “variance accounted for” differs dramatically in the covariance-based and correlation-based cases; standard presentations state that covariance-based PCs explain a certain percentage of variance in the original variables. This interpretation is misleading at best and wrong at worst, as we will show.

iv. Related to point ii., what optimality properties can be stated for rotated coefficients that are commonly used in PC-based factor analysis? The given mathematics suggests that perhaps they are not optimal at all, even if the unit length constraint is enforced on the rotations.
v. When is correlation-based PCA better than covariance-based PCA? The optimization criteria of (1) and (2) cannot answer this question.

Similar criticisms of the standard approach to introducing PCs as variance maximizers are given by ten Berge and Kiers (1996).

A Better Approach for Introducing Principal Components

The notion of “variance explained” in regression should be well established in students’ minds by the time they have reached PCA. Under the usual regression model \( Y = L_x + \epsilon \), where \( L_x \) is the least squares fit \( L_x = b_0 + b_1X_1 + \cdots + b_kX_k \), the proportion of variance in \( Y \) that is explained by the \( X \) variables is the R-squared statistic:

\[
R^2(Y|X_1, \ldots, X_k) = \frac{\text{Var}(L_x)}{\text{Var}(Y)},
\]

which is also known as the reliability of \( L_x \) as a measure of \( Y \).

Hotelling (1933) derived the PCs using variance maximization with unit length and uncorrelatedness constraints given in (1) and (2) above, but then noted that the resulting components have the property that they maximize the sum (or, equivalently, the average) of \( R^2 \) statistics, when the PCs are based on the correlation matrix as described above. Specifically, considering all possible linear functions \( L_1 = a_{11}Y_1 + \cdots + a_{1p}Y_p \), the average \( R^2 \),

\[
\{R^2(Y_1|L_1) + R^2(Y_2|L_1) + \cdots + R^2(Y_p|L_1)\}/p,
\]

\[1 \text{ The } R^2 \text{-squared statistic is also equal to the squared correlation: } R^2(Y_1|X_1, \ldots, X_k) = \{r(Y, L_1)\}^2. \]
is maximized when \( L_1 = PC_1 \), obtained using the correlation matrix \( R \). As noted by ten Berge and Kiers (1996), this optimality criterion “reflects the possibility of other approaches to PCA”; namely, to introduce the topic by starting with a different optimality criterion than variance maximization. This optimality property was also characterized by Meredith and Millsap (1985) as “astonishing in its simplicity,” as it “provides a clear explanation of [variance] accounted for”, and “justifies correlation-based PC” (correlation-based PCs maximize the average \( R^2 \) while covariance-based PCs do not). Further, the phrase “variance accounted for” now has a concrete interpretation: On average, the first correlation-based PC explains the proportion \( \{R^2(Y_1|PC_1) + R^2(Y_2|PC_1) + \cdots + R^2(Y_p|PC_1)\}/p \) of the variance of the original variables (and also the same proportion of the variance of the standardized variables \( Z_i \)). Further, this proportion is exactly \( \lambda_1/p \), the same “proportion explained” statistic given in any computer-based PCA analysis. The unit length constraint is no longer needed because the \( R^2 \) statistic is scale-free: Any constant multiple of predictor variables in a regression will give a model with the same \( R^2 \) statistic.

Thus, to solve problems (i) – (v) noted above, we suggest that educators adopt the \( R^2 \) maximization approach, introducing PCA as follows:

Choose a linear combination \( L_1 = a_{11}Y_1 + \cdots + a_{p1}Y_p \) such that

\[
\{R^2(Y_1|L_1) + R^2(Y_2|L_1) + \cdots + R^2(Y_p|L_1)\}/p \text{ is a maximum.} \tag{3}
\]

In simple words, find the linear combination of the original variables such that the average \( R^2 \) statistic is maximized, when the linear combination is used to predict each of the original variables. This criterion immediately solves problems (i), (ii), and (iii) noted above: (i) the linear combination is a “best representation” of the original variables because it provides the highest average \( R^2 \) statistic for predicting the original variables, among all possible linear combinations; (ii) scaling (unit length or
otherwise) is irrelevant since $R^2$ statistics are scale-free, and (iii) the criterion gives a direct meaning to “variance accounted for” as the “average proportion of explained variance.”

The solution to (3), noted by Hotelling (1933), is any $L_1$ that is proportional to $PC_1$ scores obtained using the correlation matrix $R$. In this way of understanding PCA, the eigenvectors $e_i$ and the scores $PC_1$ reported by the software have no special meaning other than to provide a baseline measurement that can be re-scaled arbitrarily to suit the analyst’s needs. Hence, we suggest calling the computer-reported PC scores “baseline” scores to underscore their non-uniqueness.

The second part of the introduction of PCA extends the first linear combination to more linear combinations, and the uncorrelatedness condition is not needed. The problem is stated as follows.

Choose linear combinations $L_1, \ldots, L_q$ such that

$$\frac{R^2(Y_1 | L_1, \ldots, L_q) + R^2(Y_2 | L_1, \ldots, L_q) + \cdots + R^2(Y_p | L_1, \ldots, L_q)}{p} \text{ is a maximum.}$$

Again, in simpler words, find $q$ linear combinations of the original variables such that the average $R^2$ statistic is maximized, when these $q$ linear combinations are used to predict each of the original variables. This representation solves problem (iv) as noted above, since any $q$ non-redundant (non-singular) linear transformations of the variables $L_1, \ldots, L_q$ will give the same $R^2$ statistic when predicting $Y_i$.

The solution to (4) was given by Meredith and Millsap (1985) as any set $L_1, \ldots, L_q$ that are non-redundant (non-singular) linear transformations of the first $q$ correlation-based baseline PC scores $PC_1, PC_2, \ldots, PC_q$. For example, if $q = 2$ then an optimal solution is $L_1 = 2PC_1 + PC_2, L_2 = PC_1 - PC_2$.

However, $L_1 = 2PC_1 + PC_2, L_2 = 4PC_1 + 2PC_2$ is not optimal because the linear transformations are redundant.
Again, the eigenvectors and PC scores reported by the software have no special meaning other than to provide baseline measurements that can be re-scaled and/or linearly transformed to suit the analyst’s needs. For example, the solution to (4) states the optimality property for the rotated loadings in a PCA: The first \( q \) rotated loadings, when applied to the \( Z \) scores, also give the maximum average \( R^2 \) statistic among all possible sets of \( q \) linear combinations of the original variables. And again, the “variance accounted for” expression now has a clear interpretation: These linear combinations explain, on average, the proportion \( (1/p) \Sigma_i R^2(Y_i | PC_1, \ldots, PC_q) \) of the variances in the original (and standardized) variables. Further, this proportion is exactly \( (\lambda_1 + \cdots + \lambda_q)/p \), the proportion that is printed in the PCA software output.

It is only the correlation-based PCs that have these properties; covariance-based PCs do not. Specifically, the correlation-based PCs account for an optimal average proportion of the variances of the \textit{unstandardized} variables (and of the standardized variables as well). The next section shows that covariance-based PCs \textit{do not} optimally explain the average proportion of variances of either the unstandardized or of the standardized variables.

**Presenting Covariance-Based PCs**

The results in this section are not well known, as they are somewhat obscure in the literature. They are given directly by Okamoto (1969), and implicitly by Meredith and Millsap (1985). The problem is stated as follows:

Choose a linear combination \( L_1 = a_{11}Y_1 + \cdots + a_{p1}Y_p \) such that

\[
w_1R^2(Y_1 | L_1) + w_2R^2(Y_2 | L_1) + \cdots + w_pR^2(Y_p | L_1) \text{ is a maximum,}
\]

where \( w_i = \frac{\text{Var}(Y_i)}{\text{Var}(Y_1) + \cdots + \text{Var}(Y_p)} \).

(5)
In simple words, find a linear combination of the original variables such that the weighted average \( R^2 \) statistic is maximized, when the linear combination is used to predict each of the original variables, and where the weights are proportional to the variances of the original variables. The solution to (5) is any \( L_1 \) that is proportional to the baseline \( PC_1 \) score obtained using the covariance matrix \( S \).

This formulation solves problem (v) as noted above: It is now clear that correlation-based PCs are better than covariance-based PCs if one wishes to find linear functions to maximize the average \( R^2 \); covariance-based PCs are better if you wish to maximize the variance-weighted average \( R^2 \).

For any solution \( L_1 \) to (5), the weighted average of \( R^2 \)s, \( w_1R^2(Y_1|L_1) + w_2R^2(Y_2|L_1) + \cdots + w_pR^2(Y_p|L_1) \), is precisely equal to \( \lambda_1/\Sigma\lambda_j \), where the \( \lambda_j \) are eigenvalues of the covariance matrix \( S \), as given in any PCA software output. However, the “variance explained” interpretation commonly used in covariance-based PCA is misleading, because it is a weighted average that is explained, where the weights are proportional to the variances of the variables. Textbooks often state that covariance-based PCs “account for” a proportion of the actual variables’ variances, and that correlation-based PCs “account for” a proportion of the standardized variables’ variances. Actually, correlation-based PCs explain an optimal average proportion of actual variances of the original variables, while covariance-based PCs only can be said to optimally explain a weighted average of the proportions of actual variances of the original variables.

In light of the criterion (5), correlation-based PCA appears superior, simply because the optimality criterion is more natural. Why would anyone care about this particular weighted average of \( R^2 \) statistics? It is hard to think of a case where criterion (5) would be the one to optimize, rather than (3). This result also challenges the often-given suggestion that PCA should be performed on the covariance matrix when the variables are measured in the same scale. Unless the variances are identical (which is impossible with continuous data), covariance-based components only maximize the weighted average of
$R^2$ statistics. But again, there is no obvious reason that such components should be considered “optimal” when the criterion (5) that is maximized is itself dubious. The only justification that could possibly be given is one where larger variance for one of the variables $Y_i$ equates to greater a priori importance of $Y_i$. Otherwise, optimization criterion (3) is more natural (at least as a default) in that it weights all variables equally.

The optimization criterion (5) also makes it clear why a variable with large variance dominates the first PC when covariance-based components are used. With a relatively large variance for a particular variable, say $Y_1$ for concreteness, the weighting criterion means that the researcher essentially wants a linear combination that maximizes the $R^2$ statistic for predicting $Y_1$ alone. Thus the first baseline PC must be very highly correlated with $Y_1$ in this case.

The criteria for the remaining covariance-based PCs are as follows:

Choose linear combinations $L_1, \ldots, L_q$ such that

$$w_1 R^2(Y_1| L_1, \ldots, L_q) + w_2 R^2(Y_2| L_1, \ldots, L_q) + \cdots + w_p R^2(Y_p| L_1, \ldots, L_q)$$

is a maximum. (6)

Again, in simpler words, find $q$ linear combinations of the original variables such that the weighted average $R^2$ statistic is maximized, when these $q$ linear combinations are used to predict each of the original variables.

The solution to (6) is any set $L_1, \ldots, L_q$ that are non-redundant (nonsingular) linear transformations of the first $q$ covariance-based baseline PC scores $PC_1, PC_2, \ldots, PC_q$. As above, for example, if $q = 2$ then an optimal solution is $L_1 = 2PC_1 + PC_2, L_2 = PC_1 - PC_2$. This provides not only the optimality property of the first $q$ PCs, but also the optimality of any nonsingular rotation thereof, because any nonsingular linear transformation of the predictor variables gives the same $R^2$ statistic. Table 1 summarizes these results.
Table 1. Summary of properties of correlation-based and covariance-based PCs.

<table>
<thead>
<tr>
<th></th>
<th>Correlation-Based PCs</th>
<th>Covariance-Based PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criterion</strong></td>
<td>Choose $L_1, \ldots, L_q$ to maximize $(1/p) \sum_i R^2(Y_i \mid L_1, \ldots, L_q)$</td>
<td>Choose $L_1, \ldots, L_q$ to maximize $\sum_i w_i R^2(Y_i \mid L_1, \ldots, L_q)$, where $w_i = Var(Y_i)/{Var(Y_1) + \cdots + Var(Y_p)}$.</td>
</tr>
<tr>
<td><strong>Solution</strong></td>
<td>Any $L_1, \ldots, L_q$ that are non-redundant linear transformations of the correlation-based baseline PCs $PC_1, \ldots, PC_q$.</td>
<td>Any $L_1, \ldots, L_q$ that are non-redundant linear transformations of the covariance-based baseline PCs $PC_1, \ldots, PC_q$.</td>
</tr>
<tr>
<td><strong>(Weighted) Average $R^2$</strong></td>
<td>$(\lambda_1 + \cdots + \lambda_q)/p$, where $\lambda_1, \ldots, \lambda_q$ are the largest eigenvalues of the correlation matrix $R$.</td>
<td>$(\lambda_1 + \cdots + \lambda_q)/(\lambda_1 + \cdots + \lambda_p)$, where $\lambda_1, \ldots, \lambda_q$ are the largest eigenvalues of the covariance matrix $S$.</td>
</tr>
</tbody>
</table>

Teaching Principal Components Using the Better Approach

We now give a lesson-plan for introducing PCA as suggested above. Instructions using the R software are freely available from the authors and in the on-line supplements, but any statistical software, or even spreadsheet software, can be used.

Step 1. Introduce a data set. We will use the data set epi.bfi² from the psych package in R. This is a small data set of 5 scales from the Eysenck Personality Inventory (EPI), 5 scales from a Big 5 inventory, a

² Source: http://personality-project.org/pmc.html
Beck Depression Inventory, and State and Trait Anxiety measures. The data set contains observations on the following $p = 13$ variables for each of $n = 231$ people:

- epiE: EPI Extraversion
- epiS: EPI Sociability (a subset of Extraversion items)
- epiImp: EPI Impulsivity (a subset of Extraversion items)
- epilie: EPI Lie scale
- epiNeur: EPI Neuroticism
- bfagree: Big 5 inventory measure of Agreeableness
- bfcon: Big 5 Conscientiousness
- bfext: Big 5 Extraversion
- bfneur: Big 5 Neuroticism
- bfopen: Big 5 Openness
- bdi: Beck Depression scale
- traitanx: Trait Anxiety
- stateanx: State Anxiety

To simplify the initial presentation, we suggest using just five variables, the two extraversion measures, the Beck depression scale, and the last two anxiety measures: $Y_1 = \text{epiE}$, $Y_2 = \text{bfext}$, $Y_3 = \text{bdi}$, $Y_4 = \text{traitanx}$, and $Y_5 = \text{stateanx}$. A research question that can be posed at this point to motivate the choice of these variables is, “How many distinct personality traits are measured by these five variables”? (Maybe two, if anxiety and depression measure the same thing; if not, three or more.)

Step 2. Introduce the concept of linear functions $LF = a_1 Y_1 + \cdots + a_5 Y_5$ of the variables and explain the meaning of the coefficients $a_i$. Examples: (i) $L_1 = \text{the overall average}$, with $a_1 = a_2 = \cdots = a_5 = 0.2$; (ii) $L_2 = \text{epiE}$, for which $a_1 = 1$, $a_2 = \cdots = a_5 = 0$; (iii) $L_3 = \text{extraversion summate}$, for which $a_1 = 1$, $a_2 = 1$, $a_3 = a_4 = a_5 = 0$; (iv) $L_4 = \text{the depression/anxiety summate}$, for which $a_1 = 0$, $a_2 = 0$, $a_3 = 1$, $a_4 = 1$, $a_5 = 1$. 
Step 3. Using the statistical software, calculate and compare average $R^2$ values. With the four linear functions given in step 2, the average $R^2$ values are 0.248, 0.278, 0.319, and 0.462 respectively. Thus, on average, 24.8% of the variance in the original variables $Y_1$, $Y_2$, $Y_3$, $Y_4$, and $Y_5$ is explained when predicting each of these variables using the average of all five. Ask the students to suggest other coefficients to get a higher average $R^2$. For example, the variable $lf5 = 2\times Y_1 + 1\times Y_2 - 2\times Y_3 - 1\times Y_4 - 1\times Y_5$ gives an average $R^2$ of 0.466, better than all the previous linear combinations.

Step 4. Introduce the first baseline PC as a linear function that gives the highest average $R^2$. Do not discuss the coefficients yet. Instead, use the scores provided by the statistical software, and show how these scores look in the data set as compared to the linear functions already calculated – they are just another column of numbers that can be correlated with the columns containing the original variables. Calculate directly the average $R^2$, 0.497, obtained when predicting each of the original variables in terms of the first PC. Stress the main point that no other settings of the coefficients such as “hand-picked” in $lf1$, $lf2$, $lf3$, $lf4$, $lf5$ can give a higher average $R^2$. Note also that this average $R^2$ is identical to the “variance explained” statistic reported in the PCA software output.

Step 5. Now find the coefficients of the baseline $R^2$ maximizer in terms of both raw and standardized variables. The eigenvector reported in the software output gives weights that are applied to the Z-scores to obtain the $R^2$ maximizer. Equivalently, the eigenvector’s coefficients divided by the variables’ standard deviations give the coefficients $a_i$ that are applied to the raw variables $Y_i$. These coefficients are shown in the first column of Table 2 below. Thus, the average $R^2$ when predicting each $Y_i$ using $PC_1 = -0.310Z_1 -0.353Z_2 + 0.50Z_3 + 0.547Z_4 + 0.478Z_5$ is 0.497; this is identical to the average $R^2$ when predicting each $Y_i$ using $-0.075Y_1 -0.013Y_2 +0.087Y_3 + 0.057Y_4 + 0.042Y_5$.

Now, discuss scaling. The $R^2$ statistics are the same if you multiply the predictor variables by any non-zero constant (hence unit length scaling is not important). For example, the standardized or
unstandardized coefficients can be re-scaled to a maximum of 1.0, giving linear predictors $-0.566Z_1 - 0.645Z_2 + 0.919Z_3 + 1.000Z_4 + 0.875Z_5$, and $-0.861Y_1 -0.153Y_2 + 1.000Y_3 + 0.660Y_4 + 0.479Y_5$; these also give an average $R^2$ of 0.497 when predicting each individual variable $Y_i$.

Step 6. Ask the students what they think would happen to the average $R^2$ values if additional variables were used in the prediction equations. Then calculate some average $R^2$ values using two variables: Use the linear functions already defined such as (lf1, lf2), for which the average $R^2$ is 0.516, and (lf1, lf4), for which the average $R^2$ is 0.727. Ask students for suggestions of other pairs of linear combinations in order to improve the average $R^2$.

Step 7. Introduce the first two baseline PCs as examples of linear functions that give the highest average $R^2$ when using two variables. Again, use the scores provided by the software; do not discuss the coefficients yet. Consider the first two PCs as simply two additional columns in the data matrix which can be used to predict the original variables. Calculate directly the average $R^2$ using these columns of data, getting 0.761, and note that this is the highest possible value for any two columns of linear combinations that one might pick. Then note that 0.761 is reported automatically as “variance explained” number in the PCA software output for two components.

Step 8. Find coefficients and discuss scaling issues as above in Step 5. Then note that the $R^2$ statistic obtained by regressing any variable $Y$ on any variables $(X_1, X_2)$ is the same $R^2$ as when regressing $Y$ on any (non-redundant) linear functions of the $X's$; e.g., regression of $Y$ on $(L_1, L_2)$, where $L_1 = X_1 + X_2$ and $L_2 = X_1 - X_2$ will give the same $R^2$ as regression of $Y$ on $(X_1, X_2)$. At this point, rotations can be introduced briefly, with a “heads up” that more is to come when discussing factor analysis. Table 2 shows results using the promax rotation function of the R software. As can be seen, the rotated loadings distinguish the first two extraversion variables clearly from the last three depression and anxiety variables; these loadings provide a partial answer to the question of underlying dimensionality: It
appears from the rotated loadings that these five variables essentially measure two human traits. (At this point, a “heads up” must be given to the students that underlying dimensionality will be dealt with more formally when factor analysis is presented.)

Table 2. Original and promax-rotated loadings for the five chosen variables. Numbers in parentheses are coefficients divided by the given variable’s standard deviation. Numbers in parentheses are multiplied into the raw variables; those not in parentheses are multiplied into the standardized variables.

<table>
<thead>
<tr>
<th></th>
<th>Unrotated</th>
<th>Obliquely Rotated (Promax)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>epiE</td>
<td>-0.310 (-.075)</td>
<td>0.623 (.151)</td>
</tr>
<tr>
<td>bfext</td>
<td>-0.353 (-.013)</td>
<td>0.600 (.023)</td>
</tr>
<tr>
<td>bdi</td>
<td>0.502 (.087)</td>
<td>0.343 (.059)</td>
</tr>
<tr>
<td>traitanx</td>
<td>0.547 (.057)</td>
<td>0.124 (.013)</td>
</tr>
<tr>
<td>stateanx</td>
<td>0.478 (.042)</td>
<td>0.343 (.030)</td>
</tr>
</tbody>
</table>

The coefficients shown in Table 2 define linear functions of the standardized variables; coefficients in parentheses have been divided by the variables’ standard deviations and are applied directly to the original variables. Show by direct calculation that the two rotated linear functions \{(.010Y_1 -.000Y_2 +.105Y_3 +.056Y_4 +.051Y_5), (.168Y_1 +.026Y_2 +.011Y_3 -.016Y_4 +.006Y_5)\} explain precisely the same proportion of average variance in the original variables, 0.761, as do the two PC scores. Also, note that the rotated linear functions are correlated ($r = -0.244$), so the uncorrelatedness constraint is not
needed. Again, the PCs and loadings reported in the PCA software output can be considered as “baseline suggestions” rather than precise solutions with great import.

Also, be sure to give the students a “heads up” that when factor analysis is discussed, the rotated coefficients will be used differently: They will define effects of latent traits on the observed variables, rather than linear combinations of observed variables.

Step 9. Repeat the analyses above to show all the same properties of covariance-based PCs, with the exception that the quantity to be maximized is the weighted average of $R^2$ statistics, with weights proportional to the variances. In this example, the variables’ standard deviations are quite different, with $s_1 = 4.14$, $s_2 = 26.45$, $s_3 = 5.78$, $s_4 = 9.52$, and $s_5 = 11.48$, giving weights $w_1 = 4.14^2/(4.14^2 + \cdots + 11.48^2) = 0.018$, $w_2 = 0.719$, $w_3 = 0.034$, $w_4 = 0.093$, and $w_5 = 0.136$. Thus, the first covariance-based PC is any linear combination $L_1$ that maximizes $0.018\times R^2(Y_1|L_1) + 0.719\times R^2(Y_2|L_1) + 0.034\times R^2(Y_3|L_1) + 0.093\times R^2(Y_4|L_1) + 0.136\times R^2(Y_5|L_1)$. The software gives the baseline measurement $PC_1 = -0.085Y_1 -0.977Y_2 +0.044Y_3 +0.165Y_4 +0.095Y_5$ as a maximizing linear combination; comment on the dominance of this linear combination by $Y_2$ and relate this to the fact that the criterion to be optimized depends so heavily on $R^2(Y_2|L_1)$.

Also, distinguish the correlation-based criterion from the covariance-based criterion. Figure 1 shows the covariance/correlation issue clearly: Covariance-based PCs are suboptimal when maximizing the ordinary average of $R^2$ statistics, while correlation-based PCs are suboptimal for maximizing the variance-weighted average $R^2$ statistics.
Figure 1. Optimality and sub-optimality of principal components. Left panel: Average $R^2$ for correlation-based PCs (circles) and covariance-based PCs (triangles) showing covariance-based PCs are suboptimal for maximizing the average of $R^2$ statistics. Right panel: Weighted average $R^2$ for correlation-based PCs (circles) and covariance-based PCs (triangles) showing correlation-based PCs are suboptimal for maximizing the weighted average of $R^2$ statistics.

Interpreting Components Using Correlations: The Variable Pairs Heat Map

Given a PC, linear combination, or generally any score, behavioral researchers typically wish to identify what trait(s) it measures. Rounding or truncating component loading coefficients eases interpretation,
and is valid when the truncated/rounded component score is highly correlated with the original score (Cadima and Jolliffe, 1995; Jackson, 2003 ch.7). Similarly, it is common to calculate correlations between the individual $Y_i$ variables comprising a component score to identify which variables are most related to the component (Trendafilov and Adachi, 2015; Raykov and Marcoulides, 2012 ch.7; Rencher, 1992).

With PC scores, these correlations are given by $r(Y_i, PC_j) = \sqrt{\frac{\lambda_j e_{ij}}{s_i}}$ in the unstandardized case, and $r(Y_i, PC_j) = \sqrt{\lambda_j e_{ij}}$ in the standardized case. Hence, while the loadings $e_{ij}$ clearly tell the story as to which variables contribute more in the standardized case, the variables' variances can cloud the picture in the unstandardized case. Further, for PCs with smaller eigenvalues, all the correlations are very small, yet it is clear that the variables must be related to the PC, since the PC is a deterministic function of the variables. Thus, additional information over and above the simple correlations is needed to interpret the scores in terms of variables. We suggest expanding the usual $r(Y_i, PC_j)$ analysis to include additional information on pairs of variables $Y_i - Y_j$ and $Y_i + Y_j$ (standardized or not, at the user’s pleasure, but standardized by default). Borrowing a tool from data science, we suggest displaying all the information in a convenient variable pairs heat map.

A motivating example

A toy example clarifies the need for the additional information provided by the variable pairs. Suppose there are $p = 4$ variables having the following covariance matrix:

$$S = \begin{bmatrix} 1.00 & 0.99 & 0.98 & 0.98 \\ 0.99 & 1.00 & 0.98 & 0.98 \\ 0.98 & 0.98 & 1.00 & 0.99 \\ 0.98 & 0.98 & 0.99 & 1.00 \end{bmatrix}$$
First and second baseline PCs are, respectively, \( PC_1 = 0.5Y_1 + 0.5Y_2 + 0.5Y_3 + 0.5Y_4 \) and \( PC_2 = -0.5Y_1 - 0.5Y_2 + 0.5Y_3 + 0.5Y_4 \); these two components together explain and average proportion 99.5% of the variances of the variables. Each \( Y \) variable is strongly related to \( PC_1 \) with \( r(Y_i, PC_1) = 0.994 \), \( i = 1, \ldots, 4 \). However, each \( Y \) variable is only weakly related to \( PC_2 \), with \( r(Y_1, PC_2) = r(Y_2, PC_2) = -0.087 \) and \( r(Y_3, PC_2) = r(Y_4, PC_2) = +0.087 \). This gives an erroneous impression that none of the \( Y \) variables matters to \( PC_2 \). However, \( PC_2 \) is strongly related to differences of the \( Y \) variables: For example, \( r(Y_1 - Y_3, PC_2) = 0.866 \).

**Variable pairs heat map for the principal components of the epi.bfi data set**

For a real example we again use the data set epi.bfi from the psych package in R as introduced above, but we now include all the variables. The R code to reproduce this example, as well as the one presented in the next section, is freely available from the authors and in the on-line supplements. Table 3 shows the first three eigenvectors of the correlation matrix; the first three baseline PC scores altogether explain an average proportion 66.7% of the variances of the 13 original variables. A research question to be addressed is, “What do these scores measure”?
Table 3. Loadings $e_{ij}$ the first three PCs for the \textbf{epi.bfi} data set; these coefficients are applied to the standardized variables. Correlation between variable $Y_i$ and $PC_j$ is equal to $\sqrt{\lambda_j}e_{ij}$.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$PC_1$, $\lambda_1$ = 4.00</th>
<th>$PC_2$, $\lambda_2$ = 2.67</th>
<th>$PC_3$, $\lambda_3$ = 2.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>epiE</td>
<td>-0.287</td>
<td>0.430</td>
<td>0.246</td>
</tr>
<tr>
<td>epiS</td>
<td>-0.284</td>
<td>0.359</td>
<td>0.102</td>
</tr>
<tr>
<td>epiImp</td>
<td>-0.196</td>
<td>0.366</td>
<td>0.315</td>
</tr>
<tr>
<td>epilie</td>
<td>-0.100</td>
<td>-0.284</td>
<td>-0.202</td>
</tr>
<tr>
<td>epiNeur</td>
<td>0.359</td>
<td>0.253</td>
<td>-0.136</td>
</tr>
<tr>
<td>b❢agree</td>
<td>-0.225</td>
<td>0.107</td>
<td>-0.427</td>
</tr>
<tr>
<td>b✜con</td>
<td>-0.145</td>
<td>-0.073</td>
<td>-0.518</td>
</tr>
<tr>
<td>b✜ext</td>
<td>-0.298</td>
<td>0.329</td>
<td>-0.218</td>
</tr>
<tr>
<td>b✜neur</td>
<td>0.281</td>
<td>0.310</td>
<td>-0.270</td>
</tr>
<tr>
<td>b✜open</td>
<td>-0.112</td>
<td>0.230</td>
<td>-0.441</td>
</tr>
<tr>
<td>bdi</td>
<td>0.348</td>
<td>0.228</td>
<td>-0.044</td>
</tr>
<tr>
<td>traitanx</td>
<td>0.423</td>
<td>0.184</td>
<td>0.025</td>
</tr>
<tr>
<td>stateanx</td>
<td>0.328</td>
<td>0.217</td>
<td>-0.048</td>
</tr>
</tbody>
</table>

From Table 3 we can see that the first baseline PC does not represent an “overall average” measure as is common for the first PC in typical PCA, because there are positive and negative loadings. Instead, the first PC involves contrasts. It is easy to see the importance of individual variables to each PC simply by looking at the loadings, which are proportional to the correlations. However, it is possible that a variable
has a low correlation with the PC, but is more highly correlated when combined with another variable. A way to visualize this information is using a heat map in which the darker shades indicate higher absolute correlations.

Figure 2 shows the heat map obtained from the first baseline PC, $PC_1$, for this data set. The principal diagonal contains the absolute correlations of the individual variables with $PC_1$. The upper triangle consists of the absolute correlations of the difference between two variables $Y_i - Y_j$ (standardized by default) and $PC_1$, while the lower triangle contains the absolute correlations between $Y_i + Y_j$ (again standardized by default) and $PC_1$.

![PC1.corrHeatMap](image)

Figure 2. Variable pairs heat map for interpreting the first baseline PC in the epi.bfi data example.

The diagonal shows absolute correlations between the individual variables and the PC; the upper triangle shows absolute correlations between the differences $Z_i - Z_j$ and the PC; the lower triangle shows absolute correlations between the sums $Z_i + Z_j$ and the PC.
In Figure 2 we can see that even though $PC_1$ is not highly correlated with the first variable (epiE, which measures extraversion), it is very highly correlated with the epiE – traitanx difference ($|r| = 0.90$). Thus the first baseline PC measures something that is closely related to an extraversion/anxiety contrast scale, where people who are extraverts with low anxiety are at one end of the scale, while introverts with high anxiety are at the opposite end.

Text mining example: Abstracts in *Multivariate Behavioral Research*

A second real example comes from the Data Science field, an application of PCA in text mining. We choose to analyze the abstracts of all the articles published in the journal *Multivariate Behavioral Research* in the years 2014 and 2015. The sample consists of $n = 71$ texts. The data are first preprocessed to clean the text before generating the document-term matrix of a subset from the most frequent words, which is the data used as the input to PCA. The preprocessing includes removing all special characters, punctuation, numbers, strips whitespaces, and stop words. In addition, we minimize the noise in the analysis by considering the root or main part of each word, which is known as the stem, hence in the original documents words like “analysis” and “analyses” are considered to be the same.

The most frequent words obtained after the preprocessing process, and measured simply by counting how many times they appear in total, are shown in the following graphs (see Figures 3 and 4). These most frequent words lead us to an initial document-term matrix. Then we include a sparsity constraint of at most 0.7 to avoid including words in the analysis that seem to be important in terms of the total counting, but they are indeed not highly frequent among all the documents. In other words, imposing a maximum sparsity of 0.7 to the document-term matrix means that we will only keep those words that appear at least one time in at least 30% of the documents.
Often, the PCs from text mining analyses are used to predict an external outcome variable. In this example, it would be interesting to know if information in the abstracts can predict importance of the article, as measured, say, by article citation count. But it is not sufficient to simply predict citation count. One would rather know what are the antecedents; i.e., one would like to know what the PCs actually measure so that one can know what specifically affects the outcome variable. The variable pairs heat map is designed to help in this regard.

Figure 3. Most frequent words in the analyzed abstracts of articles published in *Multivariate Behavioral Research* in the years 2014 and 2015, including how many times they appear in total.
Figure 4. Word cloud containing the 100 most frequent words in the abstracts of articles published in *Multivariate Behavioral Research* in the years 2014 and 2015. Higher size means higher frequency.
The document-term matrix obtained after applying the sparsity constraint consists of 22 words, from which we choose to eliminate 5 words since we consider they do not provide any additional information due to the purpose of this content analysis. The deleted words are: article, can, propose, provide, and use. As a result, the variables (more important words obtained from the document-term matrix) considered in this example are given in Table 4, along with coefficients of the first four baseline correlation-based PCs scores; the average proportion of variances of the original 17 variables (in this case, word frequency measures) that is explained by these four PCs is 41.5%.

Table 4. Coefficient weights e_ij of the first four baseline PCs for the document-term data set of abstracts of articles published in *Multivariate Behavioral Research* in the years 2014 and 2015; these weights are applied to the standardized variables. Correlation between variable \(Y_i\) and \(PC_j\) is equal to \(\sqrt{\lambda_j e_{ij}}\).

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1, (\lambda_1 = 1.98)</th>
<th>PC2, (\lambda_2 = 1.89)</th>
<th>PC3, (\lambda_3 = 1.63)</th>
<th>PC4, (\lambda_4 = 1.56)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. analysis</td>
<td>0.104</td>
<td>-0.321</td>
<td>0.450</td>
<td>-0.193</td>
</tr>
<tr>
<td>2. approach</td>
<td>-0.312</td>
<td>-0.208</td>
<td>-0.135</td>
<td>-0.423</td>
</tr>
<tr>
<td>3. data</td>
<td>-0.236</td>
<td>0.121</td>
<td>-0.012</td>
<td>0.059</td>
</tr>
<tr>
<td>4. differ</td>
<td>0.131</td>
<td>-0.208</td>
<td>0.253</td>
<td>-0.342</td>
</tr>
<tr>
<td>5. effect</td>
<td>0.400</td>
<td>-0.377</td>
<td>-0.027</td>
<td>-0.016</td>
</tr>
<tr>
<td>6. empirical</td>
<td>0.139</td>
<td>0.257</td>
<td>0.106</td>
<td>-0.050</td>
</tr>
<tr>
<td>7. estimate</td>
<td>0.129</td>
<td>-0.254</td>
<td>-0.499</td>
<td>-0.201</td>
</tr>
<tr>
<td>8. factor</td>
<td>-0.071</td>
<td>0.112</td>
<td>0.255</td>
<td>-0.220</td>
</tr>
<tr>
<td>9. fit</td>
<td>-0.080</td>
<td>0.225</td>
<td>-0.316</td>
<td>-0.018</td>
</tr>
<tr>
<td>10. illustrate</td>
<td>0.345</td>
<td>-0.223</td>
<td>-0.163</td>
<td>0.239</td>
</tr>
<tr>
<td>11. method</td>
<td>0.387</td>
<td>-0.006</td>
<td>-0.154</td>
<td>0.016</td>
</tr>
<tr>
<td>12. model</td>
<td>-0.259</td>
<td>-0.122</td>
<td>-0.388</td>
<td>-0.338</td>
</tr>
<tr>
<td>13. research</td>
<td>-0.096</td>
<td>0.244</td>
<td>0.077</td>
<td>-0.290</td>
</tr>
<tr>
<td>14. result</td>
<td>0.175</td>
<td>0.280</td>
<td>0.134</td>
<td>-0.052</td>
</tr>
<tr>
<td>15. simulated</td>
<td>0.318</td>
<td>0.415</td>
<td>-0.250</td>
<td>-0.075</td>
</tr>
<tr>
<td>16. studies</td>
<td>0.315</td>
<td>0.291</td>
<td>0.039</td>
<td>-0.416</td>
</tr>
<tr>
<td>17. variable</td>
<td>-0.192</td>
<td>0.017</td>
<td>0.025</td>
<td>0.366</td>
</tr>
</tbody>
</table>
Figure 5 shows the variable pairs heat map for $PC_1$. The highest absolute correlation is 0.75, given by $Y_5 + Y_{11}$, so $PC_1$ is related to a measurement scale where documents having both “method” and “effect” in the abstract are at one end of the scale, and where documents having neither “method” nor “effect” in the abstract are at the other end of the scale. Note that the correlation 0.75 is higher than each of the individual correlations for $Y_5$ and $Y_{11}$ ($1.98^{1/2} \times 0.400 = 0.56$ and $1.98^{1/2} \times 0.387 = 0.54$, respectively), illustrating a benefit of the richer representation in terms of pairs of variables.

Figure 5. Variable pairs heat map for the first baseline PC in the text-mining analysis of abstracts of articles published in *Multivariate Behavioral Research* in the years 2014 and 2015.

In the second heat map (Figure 6) the highest absolute correlation is 0.74, with $Y_5 - Y_{15} = 0.74$, so $PC_2$ is similar to a measure of whether the abstract has either the term “effect” or the term “simulated,” but
not both. Possibly, this is a measure of whether the paper is more applied (has “effect” but not “simulated” in the abstract), or more theoretical (has “simulated” but not “effect” in the abstract).

Figure 6. Variable pairs heat map of the second baseline PC in the text-mining analysis of abstracts of articles published in *Multivariate Behavioral Research* in the years 2014 and 2015.

The heat map for the third baseline PC is not given here, but shows that the \( Y_1 - Y_7 \) is most related to \( PC_3 (|r| = 0.83) \). Since \( Y_1 = \) “analysis” and \( Y_7 = \) “estimate,” this linear combination may also be related to a “theoretical/applied” contrast, where more theoretical papers (those with “analysis” but not “estimate” in the abstract) are at one end of the scale, and where more applied papers (those with “estimate” but not “analysis”) are at the other end of the scale.

The heat map for \( PC_4 \) shown in Figure 7 shows two clear associations: One with \( Y_{16} + Y_2 \) (\( |r| = 0.76 \)), and the other with \( Y_2 - Y_{17} \) (\( |r| = 0.71 \)); here \( Y_2 = \) “approach,” \( Y_{16} = \) “studies,” and \( Y_{17} = \) “variable.” Not
surprisingly, both these pairwise functions involve \( Y_2 \) which can be seen from the diagonal elements to have a large correlation with \( PC_4 \) (\(|r| = 0.53\)). But the pairwise functions have much larger correlations, so the usual method of using the single variable correlations misses important information.

Figure 7. Variable pairs heat Map of the fourth baseline PC in the text-mining analysis of abstracts of articles published in *Multivariate Behavioral Research* in the years 2014 and 2015.

**Conclusion**

In this tutorial we present two correlation-based methods for teaching, learning and understanding principal components. First, we suggest introducing PCs as (weighted) \( R^2 \) maximizers. In this way, the eigenvectors simply define a “base” PC representation; the actual variables can be any non-redundant transformations thereof. In addition, the eigenvalues just provide a handy computing tool for the \( R^2 \)
statistics, which can just as easily be computed using regression analysis. Thus the method avoids
complicated matrix mathematics and provides simpler and more direct interpretations.

Second, we develop a correlation-based variable pairs heat map to improve the interpretation of scores
(PCs or other) in terms of variables. These heat maps display all pairwise correlations of sums and
differences with the score, as well as the correlations of the original variables with the score. The benefit
of the method is to identify the meanings of the principal components: If PC scores are used
subsequently in predictive models, the meanings of the antecedents are more easily discerned when the
PC scores are diagnosed using the heat maps.

R code for performing these tasks is freely available from the authors and in the on-line supplements.

References

470.


Dai, P., Ho, S. S., & Rudzicz, F. (2015). Sequential behavior prediction based on hybrid similarity and

Debelak, R., & Tran, U. S. (2013). Principal component analysis of smoothed tetrachoric correlation

de-Marcos, L., Garcia-López, E., Garcia-Cabot, A., Medina-Merodio, J. A., Domínguez, A., Martínez-


