An Introduction to Principal Component Analysis

Marco R. Steenbergen

Chair of Political Methodology
University of Zurich

October 14, 2018

1 Motivation

Imagine that we collected data on economic left-right and GAL-TAN party positions in Germany in 2017.[1] We obtain the data shown in Table 1. (For now we shall ignore the EU positioning of parties.) A geometric representation of the mean-centered party positions can be found in Figure 1.

The question before us is whether we can reduce the 2-dimensional issue space to a smaller, i.e., 1-dimensional, space. This task can be accomplished using principal component analysis or PCA for short. In this case, PCA reveals that a single component accounts for 78 percent of the total variance in the two ideology items. We could thus conclude that a single dimension underlies economic left-right and GAL-TAN. The party scores on this dimension are shown in Figure 2. Given the party locations, this may be interpreted as a general left-right dimension.

2 How We Got There

We extracted two principal components, as many as we had variables. The first component accounts for most of the variance in the economic left-right and GAL-TAN items. The second component is orthogonal to the first component and accounts for the remainder of the variance. Once we got the components, the next step was to decide whether to retain both or just a single component. We decided for the latter.

In general, then, PCA involves two related steps.

[1] GAL-TAN stands for green-alternative-libertarian versus traditional-authoritarian-nationalist.

Table 1: 2017 German Party Data

Party	Ec. L-R	GAL-TAN	EU
AfD	7.53	9.47	1.81
CDU	6.06	5.80	6.06
CSU	6.13	7.47	4.56
FDP	8.25	3.80	4.56
Grünen	3.25	1.40	6.73
Linke	1.06	4.13	4.19
SPD	3.44	3.67	6.56

Notes: Flash Chapel Hill Expert Survey.

Figure 1: German Party Positions in a 2-Issue Space

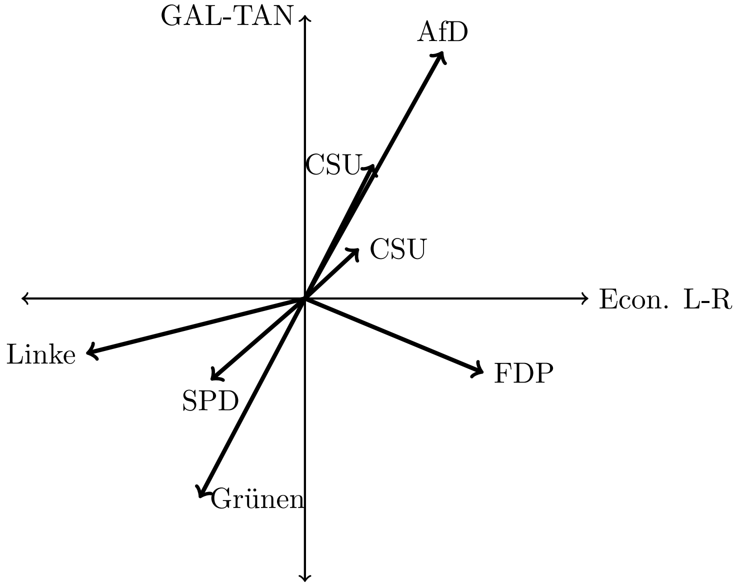
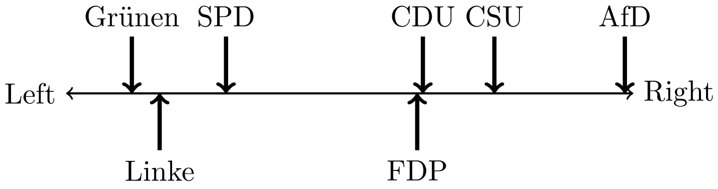


Figure 2: Party Placements on a Single Component



1. Extract the principal components. This replaces the original K variables with K uncorrelated new variables, the principal components.
2. Decide if the $\mathbb{R}^K$-space can be simplified to a lower-dimensional space. This is the data reduction aspect of PCA.

3 Extracting Principal Components

How we extract principal components depends in part on how we construct the estimation problem. The end result, however, is always the same.

3.1 Two Conceptualizations

Let $\mathbf{D}$ be a matrix of deviation scores such as

$$\mathbf{D} = \begin{bmatrix} 2.43 & 4.36 \\ 0.95 & 0.88 \\ 1.21 & 2.37 \\ 3.15 & -1.31 \\ -1.86 & -3.52 \\ -3.86 & -0.97 \\ -1.66 & -1.44 \end{bmatrix}$$

These are the deviation scores for economic left-right and GAL-TAN for the seven German parties, in the order in which those parties are listed in Table 1. Each row in $\mathbf{D}$ may be denoted as $\mathbf{d}_i^\top$ and includes the deviation scores for a particular unit.

We extract principal components such that unit i's score on the jth component is given by

$$s_{ij} = \mathbf{d}_i^\top \mathbf{c}_j \tag{1}$$

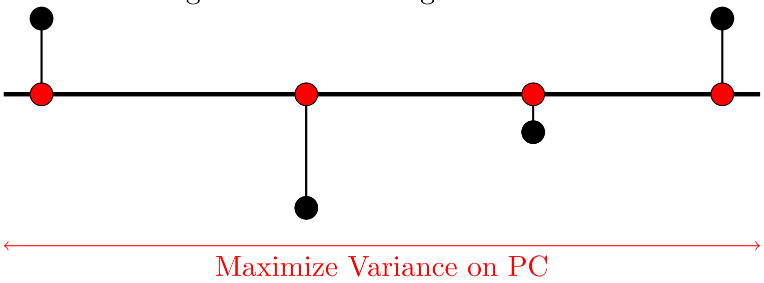
Collecting all of the units, we may also write this as $\mathbf{s}_j = \mathbf{D}\mathbf{c}_j$. Note that the expectation over the component scores is 0 because the data are given as deviation scores. This means that the variance in component score j may be written as $\mathbb{V}[\mathbf{s}_j] = \mathbb{E}\left[\mathbf{s}_j^\top \mathbf{s}_j\right] = \mathbf{c}_j^\top \mathbf{D}^\top \mathbf{D}\mathbf{c}_j$. Here, $\mathbf{D}^\top\mathbf{D}$ is a matrix of sums-of-squares and cross-products (SSCP).[2]

One perspective of PCA is now that we try to maximize the variance. Hence, we choose the first principal component in such a way that

$$\mathbf{c}_1 = \underset{\|\mathbf{c}_j\|_2=1}{\arg\max}\, \mathbf{c}_j^\top \mathbf{D}^\top \mathbf{D}\mathbf{c}_j \tag{2}$$

[2] Note that $\mathbf{D}^\top\mathbf{D} = (n-1)\mathbf{S}$, where $\mathbf{S}$ is the covariance matrix.

Figure 3: Maximizing the Variance



Here, we constrain the principal component to have a unit length. Being a unit length vector, $\mathbf{c}_1$ also maximizes Rayleigh's quotient:

$$\mathbf{c}_1 \quad = \quad \arg\max_{\mathbf{c}_j} \left(\frac{\mathbf{c}_j^\top \mathbf{D}^\top \mathbf{D} \mathbf{c}_j}{\mathbf{c}_j^\top \mathbf{c}_j} \right) \tag{3}$$

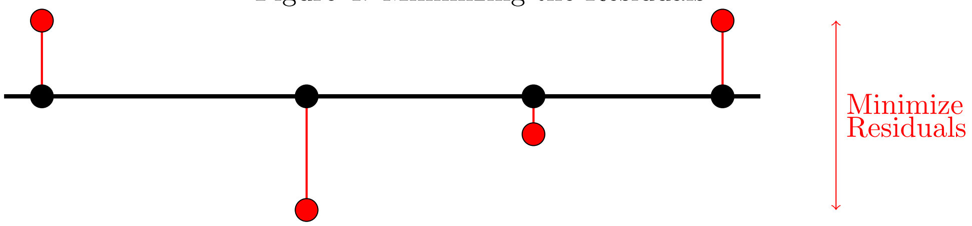
A well-known result is that the quotient's maximum possible value is the largest eigenvalue of $\mathbf{D}^\top \mathbf{D}$, provide that the matrix is positive semidefinite. Proper SSCP matrices satisfy this property. Thus, the link to spectral decomposition has been made.

Figure 3 illustrates the idea. There is greater variance along the horizontal dimension and hence, the first principal component corresponds to it.

To extract the second principal component, we look at the residual variance. The second principal component maximizes the residual variance. It is subject to the requirement that it is orthogonal to the first principal component: $\mathbf{c}_2^\top \mathbf{c}_1 = 0$.

There is a second conception of PCA, one that is based on minimization of the residuals. This is depicted in Figure 4. Here, we consider the discrepancy between the observed scores of a unit and those that would be predicted from the principal components. We select as our first component a dimension such that the residuals with respect to it are smaller than those with respect to an alternative component. As Figure 4 shows, drawing the first component as a horizontal dimension produces relatively small residuals. had we drawn a vertical dimension, then the residuals would have tended to be larger. Just imagine drawing the vertical line, projecting the black dots onto this line, and computing the horizontal distances from the data points to the principal component. You will find that those distances will be larger.

Figure 4: Minimizing the Residuals



The two conceptualizations are equivalent, in the sense that they generate exactly the same solution. We shall say a bit more about the second conceptualization once we have seen how to make predictions from PCA.

3.2 Spectral Decomposition

Given the SSCP matrix $\mathbf{M} = \mathbf{D}^\top \mathbf{D}$, we can apply spectral decomposition to derive the eigenvalues and the eigenvectors. Specifically,

$$\mathbf{M} \quad = \quad \mathbf{CLC}^\top = \sum_{j=1}^{K} \lambda_j \mathbf{c}_j \mathbf{c}_j^\top \tag{4}$$

In lieu of using the SSCP matrix, it is also possible to use the covariance matrix or even the correlation matrix.

For the German party data,

$$\mathbf{M} \quad = \quad \begin{bmatrix} 39.30 & \\ 22.86 & 42.55 \end{bmatrix}$$

Spectral decomposition of this matrix yields

$$\mathbf{L} \quad = \quad \begin{bmatrix} 68.84 & 0.00 \\ 0.00 & 18.01 \end{bmatrix}$$

and

$$\mathbf{C} \quad = \quad \begin{bmatrix} 0.68 & -0.73 \\ 0.73 & 0.68 \end{bmatrix}$$

It is easily verified that the columns, representing the principal components, are orthogonal. Further, the L_2-norm for each column is unity.

You see that the diagonal elements of $\mathbf{L}$ are arranged in descending order. The principal components are listed in the same order. Hence, the first column in $\mathbf{C}$ corresponds to the first principal component, while the second column corresponds to the second component. The entries in each column of $\mathbf{C}$ are weights or loadings. These are ordered the same way as the variables in $\mathbf{D}$. In our case, the first column in $\mathbf{D}$ is economic left-right. Hence, the first loading reflects the relationship between this variable and a particular component. The second column in $\mathbf{D}$ is GAL-TAN, so that the second loading pertains to the relationship between this variable and a component.

3.3 Singular Value Decomposition

In lieu of performing a spectral decomposition on $\mathbf{M}$, one can also perform a singular value decomposition (SVD) of $\mathbf{D}$. Specifically,

$$\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \tag{5}$$

Here, $\boldsymbol{\Sigma}$ is a $K \times K$ diagonal matrix of singular values, which equal the square roots of the eigenvalues. Further, $\mathbf{U}$ is an $n \times K$ orthonormal matrix and $\mathbf{V}$ is an $K \times K$ orthonormal matrix. It turns out that $\mathbf{V} = \mathbf{C}$.

SVD is equivalent to spectral decomposition. This is easily seen when we evaluate $\mathbf{M}$:

$$\begin{aligned}
\mathbf{M} &= \mathbf{D}^\top\mathbf{D} \\
&= \left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{C}^\top\right)^\top \left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{C}^\top\right) \\
&= \mathbf{C}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{C}^\top \\
&= \mathbf{C}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{C}^\top \\
&= \mathbf{C}\mathbf{L}\mathbf{C}^\top
\end{aligned} \tag{6}$$

The fourth line follows from the fact that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. The last line captures the fact that $\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ amounts to squaring the diagonal elements of $\boldsymbol{\Sigma}$. That, in turn, produces a diagonal matrix of eigenvalues.

The matrix of component scores is given by $\mathbf{W} = \mathbf{X}\mathbf{C}$. The polar decomposition of $\mathbf{W}$ follows from the SVD:

$$\begin{aligned}
\mathbf{W} &= \mathbf{X}\mathbf{C} \\
&= \mathbf{U}\boldsymbol{\Sigma}\mathbf{C}^\top\mathbf{C} \\
&= \mathbf{U}\boldsymbol{\Sigma}
\end{aligned} \tag{7}$$

In our example, we obtain the following results:

$$
\begin{aligned}
\mathbf{\Sigma} &= \begin{bmatrix} 8.05 & 0.00 \\ 0.00 & 4.37 \end{bmatrix} \\
\mathbf{U} &= \begin{bmatrix} -0.60 & 0.27 \\ -0.14 & -0.05 \\ -0.30 & 0.20 \\ -0.15 & -0.73 \\ 0.49 & -0.27 \\ 0.43 & 0.52 \\ 0.27 & 0.05 \end{bmatrix} \\
\mathbf{V} &= \begin{bmatrix} -0.68 & -0.73 \\ -0.73 & 0.68 \end{bmatrix}
\end{aligned}
$$

Further,

$$
\mathbf{W} = \begin{bmatrix} -0.60 & 0.27 \\ -0.14 & -0.05 \\ -0.30 & 0.20 \\ -0.15 & -0.73 \\ 0.49 & -0.27 \\ 0.43 & 0.52 \\ 0.27 & 0.05 \end{bmatrix} \begin{bmatrix} 8.05 & 0.00 \\ 0.00 & 4.37 \end{bmatrix} = \begin{bmatrix} -4.85 & 1.20 \\ -1.16 & -0.23 \\ -2.43 & 0.86 \\ -1.19 & -3.19 \\ 3.97 & -1.17 \\ 3.47 & 2.29 \\ 2.18 & 0.24 \end{bmatrix}
$$

Notice that the diagonal elements of $\mathbf{\Sigma}$ are equal to the square roots of the eigenvalues we derived before. Also notice that the first column of $\mathbf{V}$ is equal to -1 times the first column of $\mathbf{C}$ that we derived earlier. This difference is inconsequential from a mathematical point of view. From a substantive point of view, it amounts to flipping the poles of the principal component.

R and many other statistical programs perform PCA using SVD. This is generally a good approach but it is worth noting that it breaks down for tensor matrices, which are often used in neural networks, or when data are missing.

4 Data Reduction

Spectral decomposition and SVD always extract as many principal components as there are variables. With K variables, then, we obtain K components—new variables that are designed to be uncorrelated. Data reduction comes about when one eliminates certain components. In this case, $\mathbb{R}^K$ is reduced to $\mathbb{R}^P$ where $P < K$. This entails setting one or more columns on $\mathbf{C}$ to zero.

A key question in PCA is how many components should be retained. That is, how should one choose P. Several criteria have been proposed, whereby the solution of (Gavish and Donoho, 2014) is particularly interesting.

4.1 Kaiser's Rule

Kaiser (1960) proposed that we retain only those principal components that have eigenvalues greater than 1. By this criterion, we should not engage in any data reduction: for the German party data, both eigenvalues exceed 1. The rule, however, is not very reliable. As already pointed out by Horn (1965), it is best to compare the eigenvalues with those observed from a random matrix. In a sample, such a matrix could easily generate eigenvalues greater than 1, even if there is no correlation to account for. Thus, I would recommend against using this rule.

4.2 Scree Plot

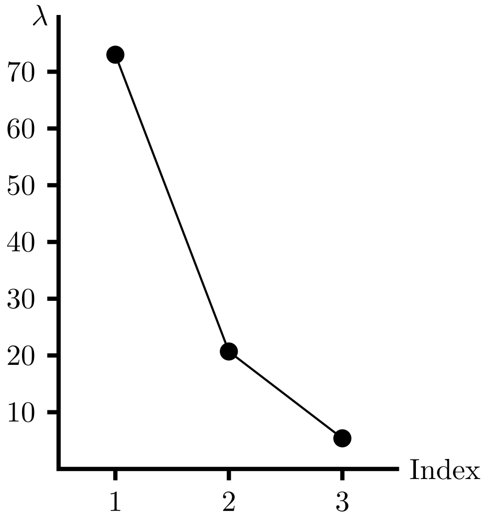
A different solution is found in the so-called scree plot (Cattell, 1966). This plots the component indices on the horizontal axis and the eigenvalues/singular values on the vertical axis. A combination of a particular index and eigenvalue appears as a point in the plot. Those points are connected via lines. Often a clear inflection point emerges in this plot. The index number at which this "elbow" appears indicates the last principal component that should be extracted.

Figure 5 shows the idea for the German party data. In this case, we have performed a spectral decomposition of economic left-right, GAL-TAN, and the EU position (see Table 1). The inflection point occurs at the second component. Thus, we are inclined to retain two principal components.

Cattell (1966, 249) describes the intuition of the scree plot as follows:

> As everyone knows such a plot falls first in a steep curve but then straightens out in a line which runs with only trivial and irregular deviations from straightness to the nth factor [...] This straight end portion we began calling the *scree*—from the straight line of rubble and boulders which forms at the pitch of sliding stability at the foot of a mountain. The initial implication was that this scree represents a "rubbish" of small error factors.

Figure 5: Scree Plot of German Party Data



Thus everything past the inflection point appears to be trivial. Those components pick up on idiosyncratic aspects of individual variables and, in any case, do not account for much of the variance.

Variance Accounted For

PCA accounts for the total variation in a set of variables. Let

$$\mathbf{S} = \frac{1}{n-1}\mathbf{M} \tag{8}$$

be the covariance matrix, then the total variance is given by $\mathrm{Tr}\,\mathbf{S}$, which is the sum of the diagonal elements of $\mathbf{S}$. It can be demonstrated that

$$\mathrm{Tr}\,\mathbf{S} = \mathrm{Tr}\,\mathbf{L} = \sum_{j=1}^{K} \lambda_j \tag{9}$$

Thus, the eigenvalues sum to the total variance.

Let us now define normalized eigenvalues:

$$\lambda_j^* = \frac{\lambda_j}{\mathrm{Tr}\,\mathbf{S}} \tag{10}$$

These give the portion of the total variance across a set of variables that is accounted for by a principal component. Following Cattell (1966), one seeks to retain those components that account for a substantive portion of the variance.

Returning to the German party data, the analysis of economic left-right and GAL-TAN yielded eigenvalues of 68.84 and 18.01. The sum of those eigenvalues is actually $(n-1)$ times the total variance because they are based on $\mathbf{M}$ and not on $\mathbf{S}$. However, if we replace the denominator in equation (10) by $\operatorname{Tr}\mathbf{M}$, the same multiplier of $n-1$ appears in both the numerator and denominator and the result may still be interpreted as the portion of explained variance. In our case,

$$\begin{aligned}\lambda_1^* &= \frac{68.84}{68.84+18.01} = 0.78 \\ \lambda_2^* &= \frac{18.01}{68.84+18.01} = 0.22\end{aligned}$$

Thus, the first principal component accounts for 78 percent of the total variance. We may find this to be sufficient and decide to retain only this component, as we did for the construction of Figure 2.

Compared to the other methods, thinking about how much explained variance suffices is obviously far more subjective. Still, this may be one of the more meaningful criteria. It forces you to think about the data. It also forces you to compare your research with the general state of the field. How many dimensions do researchers typically extract? How much variance do they explain? Does it make sense to extract additional dimensions or does this mostly add noise? These are important questions and more than some of the other criteria, thinking about the variance you account for requires you to engage with those questions.

4.3 Hard Threshold Criteria

The question of whether there is a hard threshold on the number of components has eluded researchers for decades. Recently, Gavish and Donoho (2014) proposed an answer to that question, which is equally plausible as it is elegant. Several versions of the hard thresholding rule are being proposed in their paper. Here, I shall focus on thresholds that derive entirely from empirical data.

Consider a square matrix. Let $\tilde{\sigma}$ denote the median singular value from the SVD of the matrix. Then the threshold for the number of factors is roughly

$$\tau \approx 2.858\tilde{\sigma} \tag{11}$$

When performing SVD, the matrix will generally not be square but rectangular and of order $m \times n$. In this case, the threshold is

$$\tau \approx f(\beta)\tilde{\omega} \tag{12}$$

Here $\beta = n/m$[3] and

$$f(\beta) \approx 0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43 \tag{13}$$

The approach works well even when n is small. It also takes into account the noisiness of the data. However, it depends on a few assumptions. For example, there can be no missing data and constraints on the SVD are not allowed. Another potential disadvantage is that the approach is very new and does not seem to have made its way into the social sciences yet. Using it, then, may require a lot of explanation about its workings, eating up precious paper space.

Interpretation

From a mathematic perspective, PCA ends with decisions about data reduction. For social scientists, however, interpretation of the reduced space $\mathbb{R}^P$ is essential. If the reduced space cannot be interpreted, then it may be of little practical use for our understanding of a phenomenon.

For the interpretation of principal components, the basic procedure involves two steps:

1. Retain only those columns of **C** that correspond to the principal components one has chosen to retain.

2. Use the loadings to assess the meaning of those components.

Often this suffices for obtaining an interpretable solution. In some cases, one may wish to use a rotation method such as varimax. This is the topic of discussion in the next section.

Let us return to the analysis of German parties and their positions on economic left-right and GAL-TAN. Spectral decomposition of the SSCP matrix showed that a single principal component accounts for 78 percent of

[3] Gavish and Donoho (2014) define $\beta = m/n$ but seem to assume $m < n$. I have adjusted the formula assuming $m > n$, since we typically have more units than variables. This change allows perfect reconstruction of the results in Table IV of their paper.

the total variance in party positions. Imagine we decide to retain only this component. The corresponding eigenvector is the first column in $\mathbf{C}$:

$$\mathbf{c}_1 \quad = \quad \begin{bmatrix} 0.68 \\ 0.73 \end{bmatrix}$$

The value 0.68 captures the relationship between economic positions and the principal component. This is a positive loading, suggesting that parties favoring tax reduction over government spending tend to score higher on the component. The value of 0.73 captures the relationship between GAL-TAN and the first principal component. Being positive, we know that parties with a strict immigration stance tend to score higher on the component. For both variables, then more rightist positions tend to produce higher scores on the principal component. Moreover, the loadings are roughly identical, suggesting that both variables equally drive the principal component. All of this allows us to interpret the first dimension as a general ideological dimension, as we have done in Figure 2.

When there are many variables, there are also many loadings. In this case, it may pay off to restrict some of the loadings to zero so that fewer variables flow into the interpretation. A cutoff of absolute loadings of .30 is often applied. With this cutoff, we consider only those variables for the interpretation where the loadings exceed the cutoff.

5 Rotation

Rotation is not commonly applied in PCA but can help to improve interpretability (Joliffe, 1986). It is the process of rotating the axes such that a clearer picture merges of how variables relate to principal components. A wide range of rotation methods have been proposed. In the context of PCA, we probably want to opt for a method that preserves orthogonality. Varimax rotation is one such method (Kaiser, 1958).

The idea of Varimax rotation is that the loadings are not unique and can be changed without altering the implied SSCP or covariance matrix. Consider a matrix $\mathbf{Q}$ such that $\mathbf{QC} = \mathbf{P}$ is another orthonormal matrix. Then it can be demonstrated that $\mathbf{PLP}^\top = \mathbf{M}$.[4] Thus, we can change the loadings and still recover the SSCP matrix.

[4] First, we note that $\mathbf{PP}^\top = \mathbf{I}$. It then also follows that $\mathbf{QQ}^\top = \mathbf{I}$. Next, we evaluate $\mathbf{PLP}^\top = \mathbf{QCLC}^\top\mathbf{Q}^\top$ and ask whether this yields $\mathbf{M}$. The diagonalization theorem states that $\mathbf{L} = \mathbf{P}^\top\mathbf{MP} = \mathbf{C}^\top\mathbf{Q}^\top\mathbf{MQC}$. Hence, $\mathbf{QCLC}^\top\mathbf{Q}^\top = \mathbf{QCC}^\top\mathbf{Q}^\top\mathbf{MQCC}^\top\mathbf{Q}^\top = \mathbf{QIQ}^\top\mathbf{MQIQ}^\top = \mathbf{QQ}^\top\mathbf{MQQ}^\top = \mathbf{IMI} = \mathbf{M}$. Hence, $\mathbf{CLC}^\top = \mathbf{M} = \mathbf{PLP}^\top$.

To illustrate the algorithm, consider all three variables from Table 1. Based on Figure 5, we decide to retain two principal components. The loadings are given by

$$\mathbf{C} = \begin{bmatrix} -0.59 & 0.80 \\ -0.71 & -0.48 \\ 0.38 & 0.36 \end{bmatrix}$$

The interpretation would be that we have an economic dimension and an openness dimension, which is captured by openness to immigration and European integration. The question is whether can we bring this picture into even greater relief when we perform a rotation.

The rotation process requires that we start by row-normalizing the loadings. Define the communality as the sum of the squared loadings over all retained components:

$$h_i = \sum_{j=1}^{M} c_{ij}^2 \tag{14}$$

In our example,

$$\mathbf{h} = \begin{bmatrix} 1.00 \\ 0.73 \\ 0.28 \end{bmatrix}$$

Now transform the original loadings by dividing by the square roots of the communalities:

$$r_{ij} = \frac{c_{ij}}{\sqrt{h_i}} \tag{15}$$

It is easily demonstrated that the squared loadings now add to 1 in each row. In our example,

$$\mathbf{R} = \begin{bmatrix} -0.60 & 0.80 \\ -0.76 & -0.52 \\ 0.53 & 0.50 \end{bmatrix}$$

The process then requires that we generate a number of quantities, which are shown in Table 2. We now proceed in the following manner.

1. Define

$$X = 2 \cdot (K \cdot D - A \cdot B) \tag{16}$$

$$Y = K \cdot C - \left(A^2 - B^2\right) \tag{17}$$

In our case, $X = 3.21$ and $Y = -4.77$

Table 2: Quantities for Varimax Rotation

	$\mathbf{r}_1$	$\mathbf{r}_2$	$\mathbf{u} =$ $\mathbf{r}_1^\top \mathbf{r}_1 - \mathbf{r}_2^\top \mathbf{r}_2$	$\mathbf{v} =$ $2 \cdot \mathbf{r}_1 \circ \mathbf{r}_2$	$\mathbf{u}^\top \mathbf{u} - \mathbf{v}^\top \mathbf{v}$	$\mathbf{u} \circ \mathbf{v}$
	-0.60	0.80	-0.29	-0.96	-0.83	0.27
	-0.76	-0.52	0.32	0.79	-0.53	0.25
	0.53	0.50	0.03	0.53	-0.27	0.02
Sum			$A = 0.06$	$B = 0.36$	$C = -1.63$	$D = 0.54$

Notes: The symbol ∘ indicates the Hadamard product.

2. We compute the angle of rotation as

$$\theta = \frac{1}{4} \arctan\left(\frac{X}{Y}\right) \tag{18}$$

In our case $X/Y = -0.67$ and $\theta = -0.15$..

3. We now define the rotation matrix as

$$\mathbf{Q} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{19}$$

In our case, we obtain

$$\mathbf{Q} = \begin{bmatrix} 0.99 & 0.15 \\ -0.15 & 0.99 \end{bmatrix}$$

4. The transformed loadings are given by

$$\mathbf{P} = \mathbf{RQ}$$

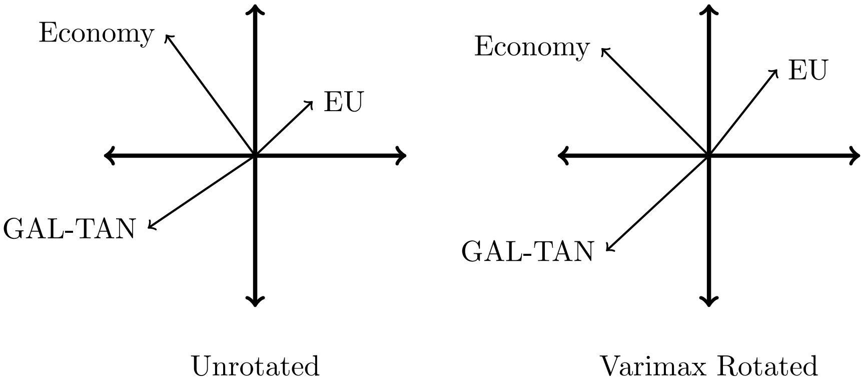
In our case,

$$\begin{bmatrix} -0.60 & 0.80 \\ -0.76 & -0.52 \\ 0.53 & 0.50 \end{bmatrix} \begin{bmatrix} 0.99 & 0.15 \\ -0.15 & 0.99 \end{bmatrix} = \begin{bmatrix} -0.71 & 0.71 \\ -0.68 & -0.63 \\ 0.45 & 0.57 \end{bmatrix}$$

The original and varimax rotated loadings are shown in Figure 6. In this case, the rotation has not cleared up the interpretation by much. This is also not surprising, given the small rotation angle.

We have now seen varimax rotations in the simple case of two remaining principal components. But what happens if there are more components?

Figure 6: Unrotated and Rotated Solutions



In that case, we apply the aforementioned procedure in a pairwise manner. After row-normalizing with respect to all of the remaining principle components, we start by rotating components 1 and 2, followed by 1 and 3, etc. For example, with four principal components, then the order would be 1-2, 1-3, 1-4, 2-3, 2-4, and 3-4.

6 Predictions and Residuals

6.1 Predictions

Predicted values for $\mathbf{D}$ come about by taking the outer-products $\mathbf{w}_j\mathbf{c}_j^\top$ and summing them over the principal components that have been retained:

$$\hat{\mathbf{D}} = \sum_{j=1}^{P} \mathbf{w}_j\mathbf{c}_j^\top \tag{20}$$

For example, imagine that we retain only the first principal component in an analysis of economic left-right and GAL-TAN. From before, we know that $\mathbf{w}_1^\top = (-4.85, -1.16, -2.43, -1.19, 3.97, 3.47, 2.18)$. We also know that $\mathbf{c}_1^\top = (-0.68, -0, 73)$. Hence,

$$\hat{\mathbf{D}} = \begin{bmatrix} -4.85 \\ -1.16 \\ -2.43 \\ -1.19 \\ 3.97 \\ 3.47 \\ 2.18 \end{bmatrix} \begin{bmatrix} -0.68 & -0.73 \end{bmatrix} = \begin{bmatrix} 3.30 & 3.55 \\ 0.79 & 0.85 \\ 1.66 & 1.78 \\ 0.81 & 0.87 \\ -2.71 & -2.91 \\ -2.36 & -2.54 \\ -1.49 & -1.60 \end{bmatrix}$$

These are the predicted deviation scores if only the first component is considered.

If we consider both principal components, then we add a second outer-product to the result we just obtained, namely $\mathbf{w}_2\mathbf{c}_2^\top$. Thus,

$$\hat{\mathbf{D}} = \begin{bmatrix} 3.30 & 3.55 \\ 0.79 & 0.85 \\ 1.66 & 1.78 \\ 0.81 & 0.87 \\ -2.71 & -2.91 \\ -2.36 & -2.54 \\ -1.49 & -1.60 \end{bmatrix} + \begin{bmatrix} -0.88 & 0.82 \\ 0.17 & -0.15 \\ -0.63 & 0.59 \\ 2.34 & -2.18 \\ 0.86 & -0.80 \\ -1.68 & 1.56 \\ -0.17 & 0.16 \end{bmatrix} = \begin{bmatrix} 2.43 & 4.36 \\ 0.96 & 0.69 \\ 1.03 & 2.36 \\ 3.15 & -1.31 \\ -1.85 & -3.71 \\ -4.04 & -0.98 \\ -1.66 & -1.44 \end{bmatrix} = \mathbf{D}$$

This result makes sense: if we consider all of the principal components, then we should be able to reconstruct all of the data since no data reduction has taken place.

6.2 Residuals

The quantity $\mathbf{E} = \mathbf{D} - \hat{\mathbf{D}}$ captures the residuals from a principal component analysis. When we retain a single principal component for the German data, for example, we obtain

$$\mathbf{E} = \begin{bmatrix} 2.43 & 4.36 \\ 0.96 & 0.69 \\ 1.03 & 2.36 \\ 3.15 & -1.31 \\ -1.85 & -3.71 \\ -4.04 & -0.98 \\ -1.66 & -1.44 \end{bmatrix} - \begin{bmatrix} 3.30 & 3.55 \\ 0.79 & 0.85 \\ 1.66 & 1.78 \\ 0.81 & 0.87 \\ -2.71 & -2.91 \\ -2.36 & -2.54 \\ -1.49 & -1.60 \end{bmatrix} = \begin{bmatrix} -0.88 & 0.82 \\ 0.17 & -0.15 \\ -0.63 & 0.59 \\ 2.34 & -2.18 \\ 0.86 & -0.80 \\ -1.68 & 1.56 \\ -0.17 & 0.16 \end{bmatrix}$$

Mathematically, the residuals are given by

$$\mathbf{E} = \mathbf{D} - \mathbf{W}\mathbf{C}^\top$$

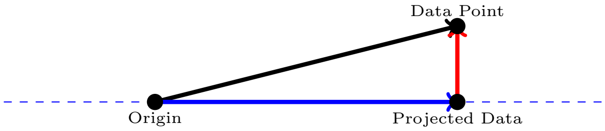
Since $\mathbf{W} = \mathbf{D}\mathbf{C}$, we can also write this as

$$\mathbf{E} = \mathbf{D} - \mathbf{D}\mathbf{C}\mathbf{C}^\top \tag{21}$$

If all principal components are retained, then $\mathbf{E} = \mathbf{0}$, a matrix of 0s.[5]

[5] In that case, $\mathbf{C}$ is a square matrix and $\mathbf{C}\mathbf{C}^\top = \mathbf{I}$. It then follows that $\mathbf{E} = \mathbf{D} - \mathbf{D}\mathbf{I} = \mathbf{0}$.

Figure 7: How Maximizing the Variance Implies Minimizing the Residuals



6.3 PCA as a Minimization Problem

At the beginning of these notes, I stated that PCA can be alternatively conceptualized as a problem of maximizing the variance or of minimizing the residuals. Let us now look at the latter proposition and see why it is equivalent to the first one.

When conceived of as a minimization problem, PCA consists of minimizing the squared Frobenius norm of the residuals:

$$\|\mathbf{D} - \mathbf{D}\mathbf{C}\mathbf{C}^\top\|_F^2 \tag{22}$$

subject to $\mathbf{C}\mathbf{C}^\top = \mathbf{I}$.

The relationship to the variance becomes clearer when we visualize the decomposition of the variance, as is done in Figure 7. The Euclidean norm of the black line reflects the total variation and is fixed. Per Pythagoras' theorem, this norm can be decomposed into the norms of the blue and the red lines. The blue line represents the prediction (or projected data), whereas the red line represents the residual from that prediction. Since the norm of the black line is fixed by the data, decreasing the norm of the red line automatically means increasing the norm of the blue line. That norm reflects the variation in the projections, so that minimizing the residuals amounts to maximizing the variation that is accounted for.

References

Cattell, Raymond B. 1966. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1(2):245–276.

Gavish, Matan and David L. Donoho. 2014. "The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$." *IEEE Transactions of Information Theory* 60(8):5040–5053.

Horn, John L. 1965. "A Rationale and Test for the Number of Factors in Factor Analysos." *Psychometrika* 30(2):179–185.

Joliffe, Ian T. 1986. *Principal Component Analysis.* New York: Springer.

Kaiser, Henry F. 1958. “The Varimax Criterion for Analytic Rotation in Factor Analysis.” *Psychometrika* 23(3):187–200.

Kaiser, Henry F. 1960. “The Apllication of Electronic Computers to Factor Analysis.” *Educational and Psychological Measurement* 20(1):141–151.