Principal Components Analysis

Today . . .

- Basic goals of principal components analysis
- Similarity to other procedures
- Simple example

Principal components analysis is a method for re-expressing a set of variables. It is used for:

- converting a large set of variables to a smaller number of linear combinations that contain most of the original information
- exploring the dimensionality in a set of variables
- eliminating multicollinearity in a set of predictors
- identification of outliers

Principal components analysis creates linear combinations of the original variables. Each successive linear combination:

(a) Accounts for as much of the variance in the original variables as possible

(b) Is independent of the previously created linear combinations

Principal components proceeds by seeking a linear combination (**z**) that is a weighted combination of the original variables (in standard score form), **Xu**, such that the variance of **z** has the largest value possible:

$$\hat{\sigma}_z^2 = \frac{1}{N-1} u' X' X u$$

The weight vector, **u,** must be constrained in order to have a unique solution:

$$u' u = 1$$

$$R = \frac{1}{N-1} X' X$$

$$\hat{\sigma}_z^2 = \frac{1}{N-1} u' X' X u$$

The problem can be recast as maximizing **u`Ru**.

This is sensible because the application of a vector of weights to a variance-covariance matrix will produce the variance of the linear combination that those weights would create if applied to the original data.

The weights that satisfy the variance-maximizing goal are called an *eigenvector* (**u**). The variance of the resulting linear combination is called an *eigenvalue* (λ).

If the original correlation matrix is of full rank (no perfect dependencies), then there will be as many eigenvectors and eigenvalues as there are original variables.

The collection of eigenvectors is contained in the matrix, **U**.

The variance-covariance matrix of the principal components is a diagonal matrix, **D**, with the eigenvalues on the main diagonal:

$$D = U'RU$$

The diagonal elements are the variances of the linear combinations. The off-diagonals are zero because the procedure creates linear combinations that are independent.

Principal components analysis re-expresses the original variables. It simply rearranges the variance, shifting it so that most is contained in the first principal component, the next most in the second principal component, and so on.

This means that trace(**D**) = trace(**R**). The matrices **D** and **R** contain the same information, but in different arrangements.

The proportion of variance that each principal component accounts for in the original data can be expressed as:

$$\text{proportion of variance}(\lambda_i) = \frac{\lambda_i}{trace(D)} = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}$$

The determinant of **R** is usually complex to find. It can be found easily with **D**:

$$|R| = \prod_{j=1}^{k} d_{jj} = \prod_{j=1}^{k} \lambda_j$$

That the determinant of **R** is also the determinant of **D** is a reminder that principal components analysis simply rearranges the variance.

If $\mathbf{Z_s}$ is a matrix of standardized principal components scores, **D** is the diagonal matrix that contains variances for the principal components (the eigenvalues) and **U** is a matrix that contains the weights for creating the linear combinations (the eigenvectors), then:

$$X = Z_s D^{\frac{1}{2}} U'$$

The formula can be rearranged to provide the principal components scores:

$$Z_s = XUD^{-\frac{1}{2}}$$

These might be used in other statistical analyses because of their desirable properties.

The meaning of the new linear combinations is sometimes easier to grasp by examining the correlations of the original variables with the new linear combinations—the principal component scores.

$$R_{X,Z} = \frac{1}{N-1}X'Z_s$$

$$R_{X,Z} = \frac{1}{N-1}X'XUD^{-\frac{1}{2}}$$

$$R_{X,Z} = \left(\frac{1}{N-1}X'X\right)UD^{-\frac{1}{2}}$$

$$R_{X,Z} = (R)UD^{-\frac{1}{2}} = (UDU')UD^{-\frac{1}{2}}$$

$$R_{X,Z} = UD(U'U)D^{-\frac{1}{2}} = UDD^{-\frac{1}{2}} = UD^{\frac{1}{2}}$$

These correlations are called ***principal component loadings*** and are found by:

$$F = UD^{\frac{1}{2}}$$

The principal component loadings are just a rescaling of the eigenvectors. The matrix, **F**, is called the **structure matrix**.

The proportion of variance for a variable in **X** that is accounted for by *C* principal components is called the communality:

$$h_j^2 = \sum_{c=1}^{C} f_{j,c}$$

Each element, f_{jc}, is a principal component loading.

The communality provides an index of how well the principal components can reproduce each of the original variables.

$$h_j^2 = \sum_{c=1}^{C} f_{j,c}$$

Why does it make sense to simply sum the squares of these correlations to get the proportion of variance accounted for in X_j by the *C* components?

The variances of linear combinations can be obtained by applying the weights for the linear combinations to the covariance matrix of the original variables. In standard score form:

$$D = U'RU$$

This also means that the correlations among the original variables can be recovered from the covariance matrix of linear combinations:

$$R = UDU'$$

If all of the principal components are derived, the reconstructed correlation matrix will be exactly the original correlation matrix.

But, if only *c* components are derived, the reconstructed matrix will be an estimate of the original matrix, **R**, that is implied by the components:

$$\hat{R}_C = U_C D_C U_C'$$

The closeness of the reconstructed matrix to the original can be used to gauge how well the C components capture the variance in the original variables:

$$R_{residual} = R - \hat{(R)}$$

The correlation matrix, reconstructed correlation matrix, and the residual matrix play a role in determining how many components should be derived.

$$\chi^2_{\left[\frac{p^2-p}{2}\right]} = -\left[(N-1) - \frac{2p+5}{6}\right] ln|R|$$

This is Bartlett's test of sphericity, a test that the matrix R is an identity matrix. A modification of this test can be applied to residual matrices to test if additional components should be extracted.

Two other methods are commonly used to determine how many components are necessary to capture the information in the original variables:

- The scree test
- Kaiser's $\lambda > 1.0$ rule

The scree test is based on the idea that a meaningful component should have an eigenvalue that is noticeably different from what would emerge from random data.

The "$\lambda > 1.0$" rule is based on the idea that a component should have more variance than any random item.

Ofir and Simonson (2001) collected data (N = 195) on an individual difference measure called "need for cognition." They wanted to know if this 18-item scale was best described by one dimension. If so, a single composite score would be an appropriate summary for research purposes.

[ ] 3. Thinking is not my idea of fun.

[ ] 7. I only think as hard as I have to.

[ ] 11. I really enjoy a task that involves coming up with new solutions to problems.

[ ] 13. I prefer my life to be filled with puzzles that I must solve.

Each items is rated using the following scale:

1 = very characteristic of me

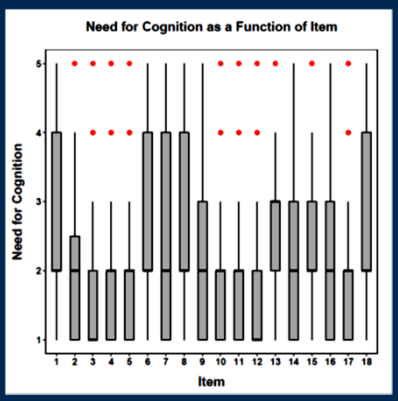
2 = somewhat characteristic of me

3 = neutral

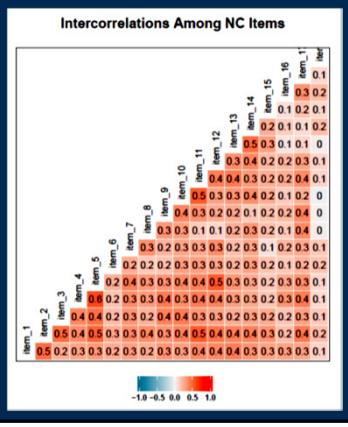
4 = somewhat uncharacteristic of me

5 = very uncharacteristic of me

To keep interpretation of the principal components analysis simpler, items are scored to be in a consistent direction.



The crude measurement scale can produce distributions that vary considerably across items. This can complicate analyses.



A heat map for the correlation matrix can assist examining a large matrix. Patterns in the data, if present, are easier to detect. This heat map suggests fairly homogeneous correlations indicative of a single principal component.

Missing data can complicate PCA analyses. Listwise deletion insures that the correlation matrix is consistent, but the available sample size will be at a minimum. Pairwise deletion conserves cases, but can produce inconsistent correlations. Imputation preserves sample size but can encounter estimation difficulties. All of these procedures have assumptions that must be carefully considered.

Two tests are available for determining if a principal components analysis should be conducted. In addition to Bartlett's test of sphericity, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (MSA) can be used to determine if common factor variance underlies the correlations in a matrix.

The MSA ranges between 0 and 1, with the following guidance for interpretation: .90 and above (undeniable evidence for factorability), .80 to .89 (very strong evidence), .70 to .79 (modest evidence), .60 to .69 (weak evidence), .50 to .59 (very weak evidence), and below .50 (unacceptable for factoring).

The correlation matrix is not an identity matrix and can be analyzed using principal components.

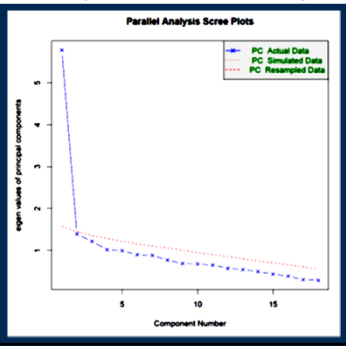
```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R)
## Overall MSA =  0.87
## MSA for each item =
##  item_1  item_2  item_3  item_4  item_5  item_6  item_7  item_8
##    0.91    0.89    0.88    0.89    0.86    0.91    0.90    0.75
##  item_9 item_10 item_11 item_12 item_13 item_14 item_15 item_16
##    0.89    0.87    0.86    0.87    0.87    0.83    0.89    0.84
## item_17 item_18
##    0.83    0.69


cortest.bartlett(R = R, n = length(NC[, 1]))

## $chisq
## [1] 1024
##
## $p.value
## [1] 8.209e-129
##
## $df
## [1] 153
```

One dominant dimension appears to underlie the data.

```
scree <- fa.parallel(NC[, c(1:18)], fa = "pc")
```



```
PCA_1 <- principal(R, nfactors = 1, rotate = "none", n.obs = 195,
    residuals = TRUE)
```

```
Standardized loadings (pattern matrix) based upon correlation matrix
          PC1   h2   u2 com
item_1   0.60 0.364 0.64   1
item_2   0.74 0.553 0.45   1
item_3   0.61 0.370 0.63   1
item_4   0.66 0.436 0.56   1
item_5   0.70 0.483 0.52   1
item_6   0.44 0.194 0.81   1
item_7   0.55 0.297 0.70   1
item_8   0.49 0.245 0.76   1
item_9   0.56 0.317 0.68   1
item_10  0.62 0.387 0.61   1
item_11  0.68 0.459 0.54   1
item_12  0.62 0.388 0.61   1
item_13  0.56 0.312 0.69   1
item_14  0.55 0.305 0.69   1
item_15  0.41 0.164 0.84   1
item_16  0.39 0.156 0.84   1
item_17  0.55 0.304 0.70   1
item_18  0.21 0.044 0.96   1

                PC1
SS loadings    5.78
Proportion Var 0.32
```

The items vary in the strength of their associations with the principal component (related to their weights in the linear combination).

The principal component accounts for 32% of the variability in the original data.

The scree test suggests a single component is required, but several additional components have eigenvalues greater than 1.00. Are those meaningful? Are they just error? The principal components analysis will capitalize on chance relations in the data and extract some components with eigenvalues greater than 1.00.

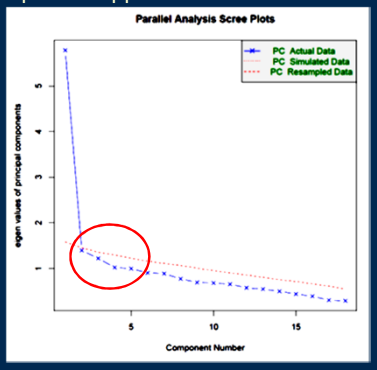
One way to address this question is to determine what pattern of eigenvalues would emerge if the data were simply random.

One approach, first suggested by Horn, is based on generating a new matrix of random, normally distributed variables. This matrix is the same size as the original data matrix.

Another approach, patterned after bootstrapping, generates a new matrix of data by resampling with replacement from the original data matrix. Each data point is independently (not case) sampled.

These random data are then analyzed with principal components analysis and serve as a baseline for the PCA of the actual data.

One component appears to be sufficient.



The ability of the principal components to account for each variable's variance is indicated by the communality.

```
Standardized loadings (p
          PC1   h2   u2
item_1   0.60 0.364 0.64
item_2   0.74 0.553 0.45
item_3   0.61 0.370 0.63
item_4   0.66 0.436 0.56
item_5   0.70 0.483 0.52
item_6   0.44 0.194 0.81
item_7   0.55 0.297 0.70
item_8   0.49 0.245 0.76
item_9   0.56 0.317 0.68
item_10  0.62 0.387 0.61
item_11  0.68 0.459 0.54
item_12  0.62 0.388 0.61
item_13  0.56 0.312 0.69
item_14  0.55 0.305 0.69
item_15  0.41 0.164 0.84
item_16  0.39 0.156 0.84
item_17  0.55 0.304 0.70
item_18  0.21 0.044 0.96
```

If the extracted components are sufficient to account for the variance in the original variables, then the correlations among the residuals should be trivial, indistinguishable from an identity matrix, and not worth additional extraction.

```
# Create a correlation matrix of the residuals by replacing the
# main diagonal with ones.
R1 <- diag(PCA_1$residual)
R2 <- diag(R1)
R3 <- PCA_1$residual - R2
R4 <- diag(18) + R3
```

```
KMO(R4)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R4)
## Overall MSA =  0.35
## MSA for each item =
##  item_1  item_2  item_3  item_4  item_5  item_6  item_7  item_8
##    0.30    0.28    0.27    0.38    0.35    0.28    0.30    0.36
##  item_9 item_10 item_11 item_12 item_13 item_14 item_15 item_16
##    0.34    0.33    0.31    0.31    0.41    0.43    0.34    0.41
## item_17 item_18
##    0.42    0.46

cortest.bartlett(R = R4, n = 195)

## $chisq
## [1] 190.1
##
## $p.value
## [1] 0.02247
##
## $df
## [1] 153
```

Mixed evidence, but probably not worth additional extraction.

Next time . . .

Finding outliers and verifying multivariate normality with PCA.