Principal Components Analysis

Measurement, Scaling, and Dimensional Analysis
2019 ICPSR Summer Program
Prof. Adam M. Enders

Assessing and Reducing Dimensionality

- The summated rating, cumulative scaling, and unfolding models all assume unidimensionality
- Obviously, there will be cases where our data contain more than one important source of variance, or dimension
- The SVD/eigendecomposition can help us distill our data down to it's fundamental characteristics
- We also saw that it can tell us how each dimension – re-expressed component of the original data matrix – contributes to variance in the total dataset
- Our goal now is to formalize how we go about selecting a smaller number of dimensions for investigating the data and conducting subsequent statistical analyses
- This is what principal components analysis helps us do

PCA: Motivation I

- Say you're an economist with a great deal of data on various indicators of economic performance of health
- Such indicators might include indices of prices, wage rates, cost of living, and so on
- When assessing changes in prices over time, the economist will wish to allow for the fact that prices of some commodities are more variable than others, or that the prices of some of the commodities are considered more important than others
  - In each case the index will need to be weighted accordingly
- In this case, the first principal component can often satisfy the researcher's requirements

PCA: Motivation II

- Perhaps you're swimming data about movie and television show preferences, organizations people are associated with, interactions with friends, political preferences, etc.
- This presents the researcher with two problems:
  - First, that's simply too much data to analyze in a comprehensible using traditional statistical models
  - Second, we probably assume that personal preferences about a wide range of objects can be distilled down to a smaller set of personality characteristics or similar latent variables
- Once again, PCA can help us cut through a ton of similar variables an reveal some simpler dimensional structure
- (Then you can go work for Cambridge Analytica and make a bunch of money!)

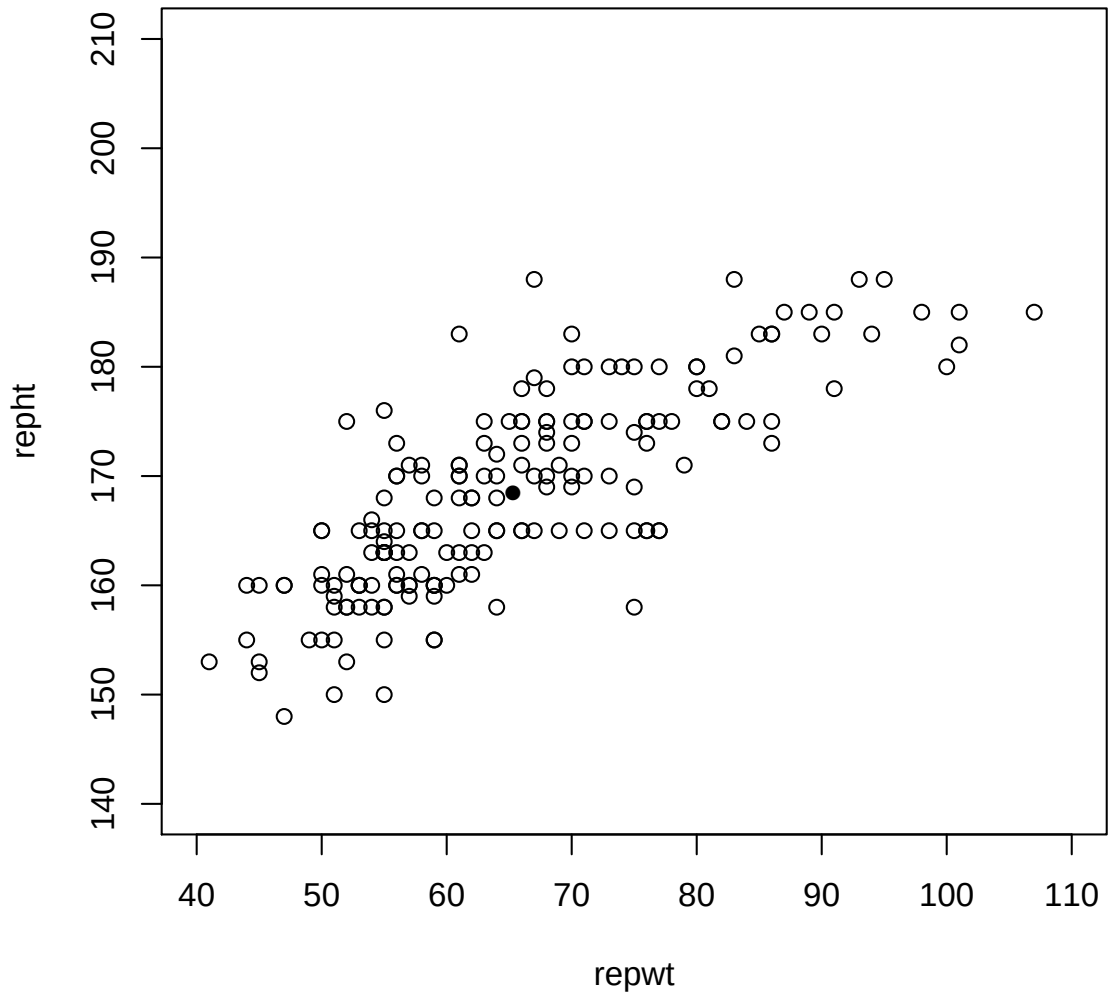
Principal Components Analysis (PCA)

- Tries to find a set of orthogonal, variance-maximizing, linear combinations of our multivariate data
- In other words:
  - The new variables capturing the information are uncorrelated
  - The first variable has the most variance and the second variable has the second most variance
- Not inherently a data reduction technique – will always get a number of principal components that is equal to the number of variables (dimensions) in the dataset

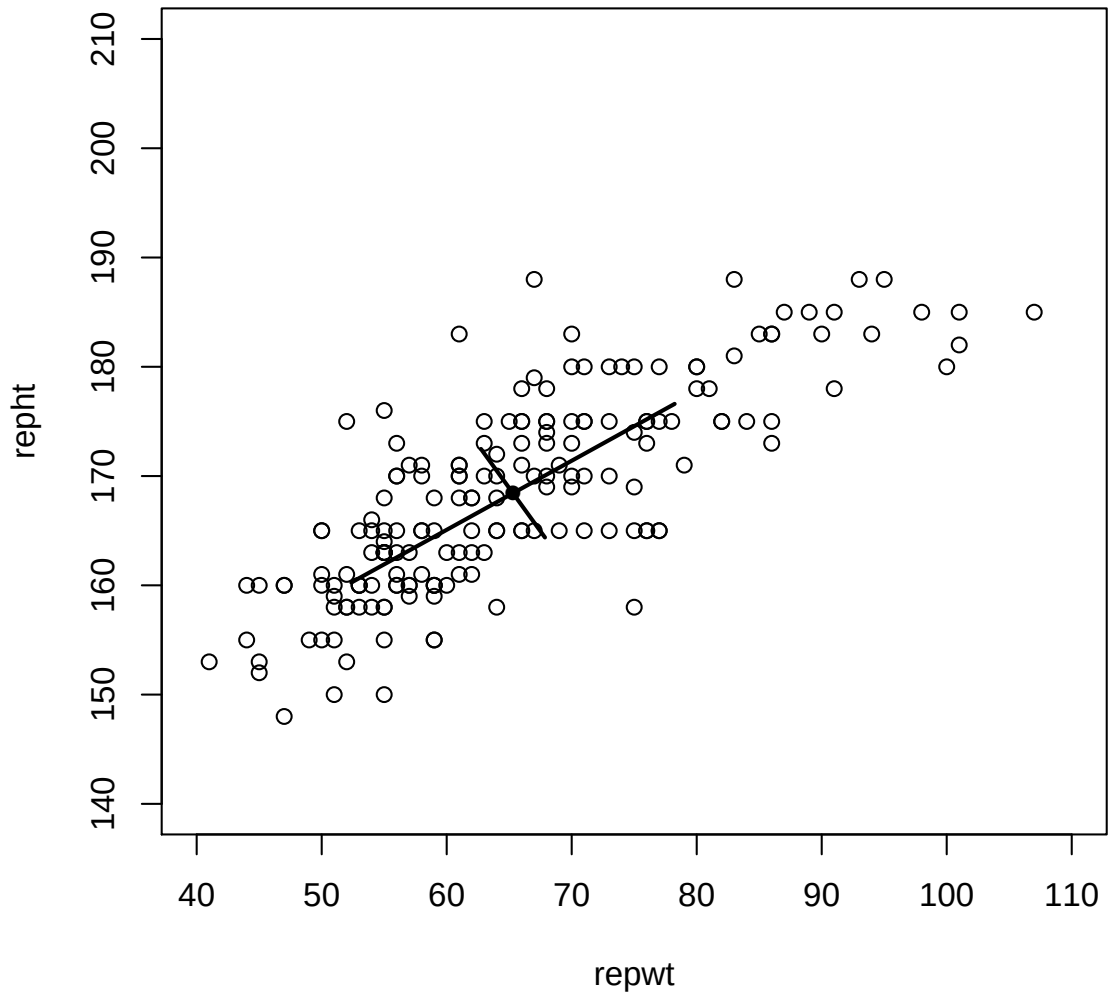
Principal Components Analysis (PCA), cont'd

- Not inherently a data reduction technique – will always get a number of principal components that is equal to the number of variables (dimensions) in the dataset
- This is also not a technique that attempts to find latent variables underlying a set of data
- What are the natural axes (sources of variation) that run through a set of variables?
  - Start by fitting the best line to the data – note that the"residuals" are constructed by drawing a *perpendicular* line from the observations to the line
  - Then draw the next best-fitting line that is orthogonal to the first, and so on

The Intuition



The Intuition



Principal Components Analysis (PCA), cont'd

- Technically an "exploratory" technique since no assumptions about the structure of the data are imposed on the "model"
- The general hope of principal components analysis is that the first few components will account for a substantial proportion of the variation in the original variables
- If this is the case, we can, consequently, use the PCs to provide a convenient lower-dimensional summary of these variables that might prove useful for a variety of reasons
- Not an estimation technique – nothing is estimated

Principal Components Analysis (PCA), cont'd

- PCA tries to find the "weights," a_k, such that the original variables, x_k, can be linearly combined to form k principal components

$$C_k = a_{k1}x_1 + a_{k2}x_2 ... + a_{kk}x_k \text{ , where } \mathbf{X}_{n\times k} \implies \mathbf{C}_{n\times k}$$

  - Note that we have no error term at the end of the above equation – again, this is not an estimation technique

- The components must have the following properties:
  1. $Var(C_1) \geq Var(C_2) \geq ... \geq Var(C_k)$
  2. $cor(C_j C_m) = 0$, where $j \neq m$
  3. $\sum_{i=1}^{k} a_i^2 = 1$
  4. As a result of the above, these also maximize: $\sum_m r^2_{c_k x_m}$ subject to the constraints above

Principal Components Analysis (PCA), cont'd

- Stated verbally:
  - The first PC captures/explains the most variance, the second explains the second most, and so on
  - The PCs must be uncorrelated with each other
  - The sum of squared loadings must equal 1
  - The above conditions maximize the sum of squared correlations between the variables and the components
- Want to find a matrix $\mathbf{C}_{n \times k}$ that equals $\mathbf{X}_{\mathbf{n} \times \mathbf{k}} \mathbf{A}_{\mathbf{k} \times \mathbf{k}}$

$$
\begin{aligned}
\boldsymbol{X} &= \boldsymbol{UDV'} \\
\boldsymbol{X} &= \boldsymbol{UDA'} \\
\boldsymbol{XA} &= \boldsymbol{UDA'A}, \text{ and } \boldsymbol{A'A} = \boldsymbol{I} \\
\boldsymbol{XA} &= \boldsymbol{UD}
\end{aligned}
$$

- SVD allows us to find matrix $\mathbf{A}$ - indeed, its just the $\mathbf{D}$ matrix of singular values (a matrix of "stretching/shrinking" values)

Arithmetical Operations: Linear Combinations

- A linear combination involves both addition and scalar multiplication
- Regression equations, and as we will see shortly, principal components analysis, are linear combinations

$$\vec{z} = b_x \vec{x} + b_y \vec{y}$$

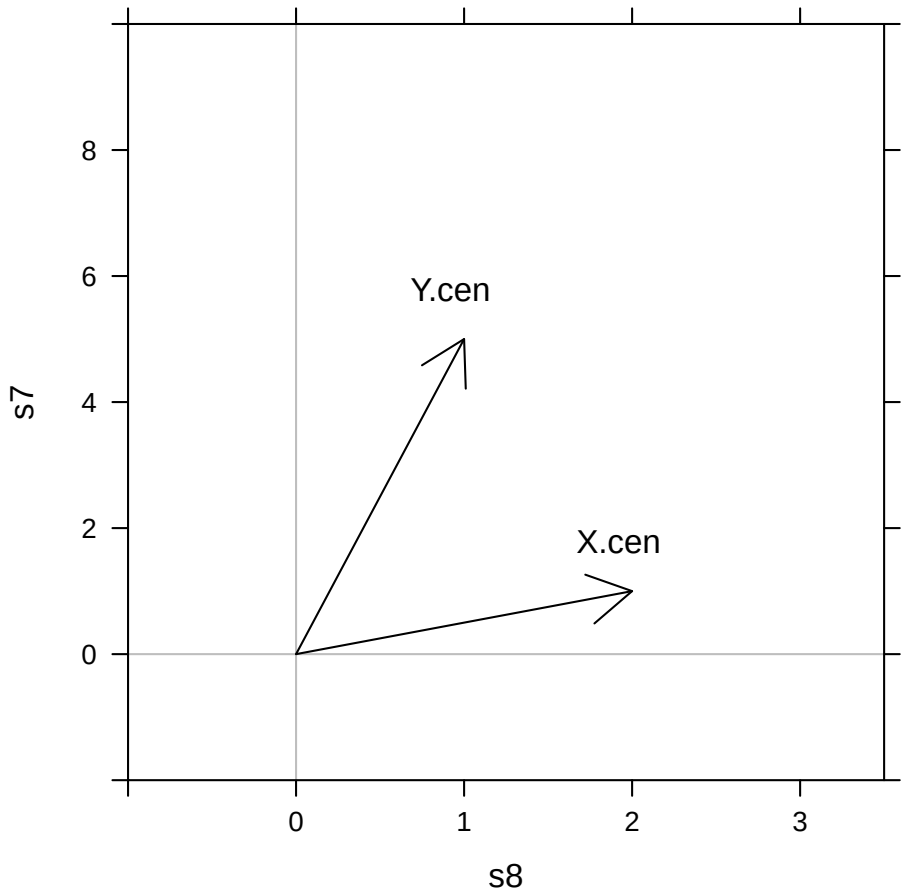
- Geometrically, the above equation entails moving along $\vec{x}$ for a distance of b_x times its length, then turning in the direction of $\vec{y}$ for for b_y times its length

Arithmetical Operations: Linear Combinations

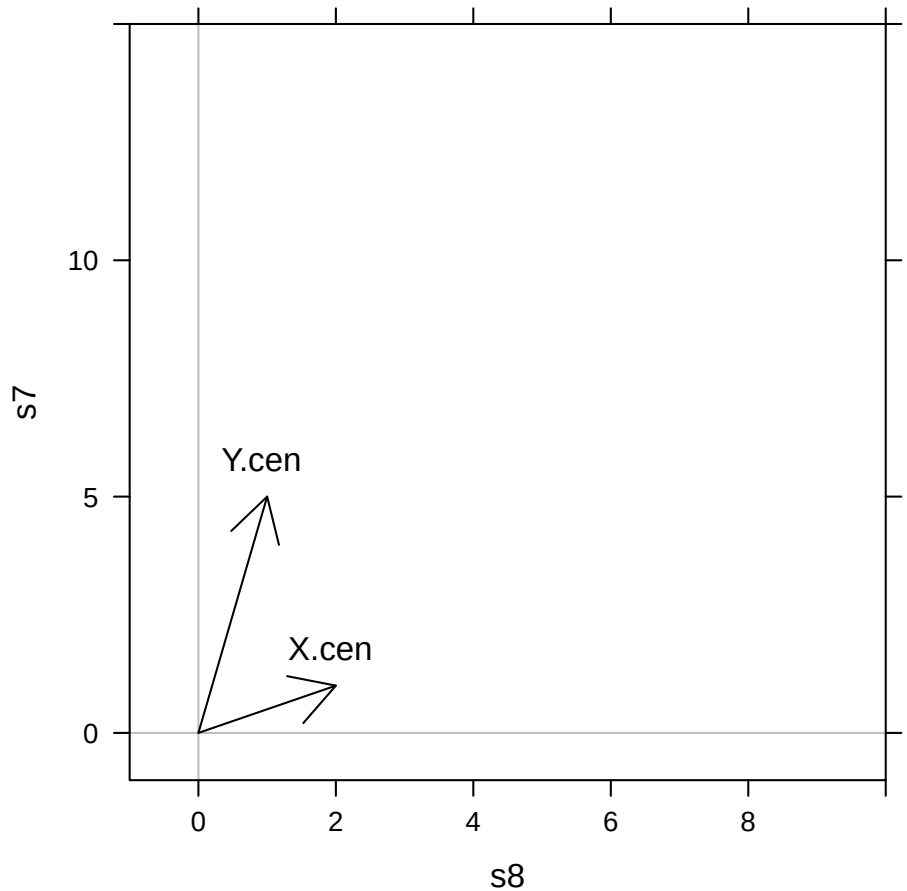
Continuing with our hypothetical vectors $\vec{x}$ and $\vec{y}$ from before...

$$\begin{aligned}
\vec{z} &= b_x\vec{x} + b_y\vec{y} \\
&= 3\vec{x} + 2\vec{y} \\
&= 3\begin{bmatrix}2\\1\end{bmatrix} + 2\begin{bmatrix}1\\5\end{bmatrix} \\
&= \begin{bmatrix}6\\3\end{bmatrix} + \begin{bmatrix}2\\10\end{bmatrix} \\
&= \begin{bmatrix}8\\13\end{bmatrix}
\end{aligned}$$

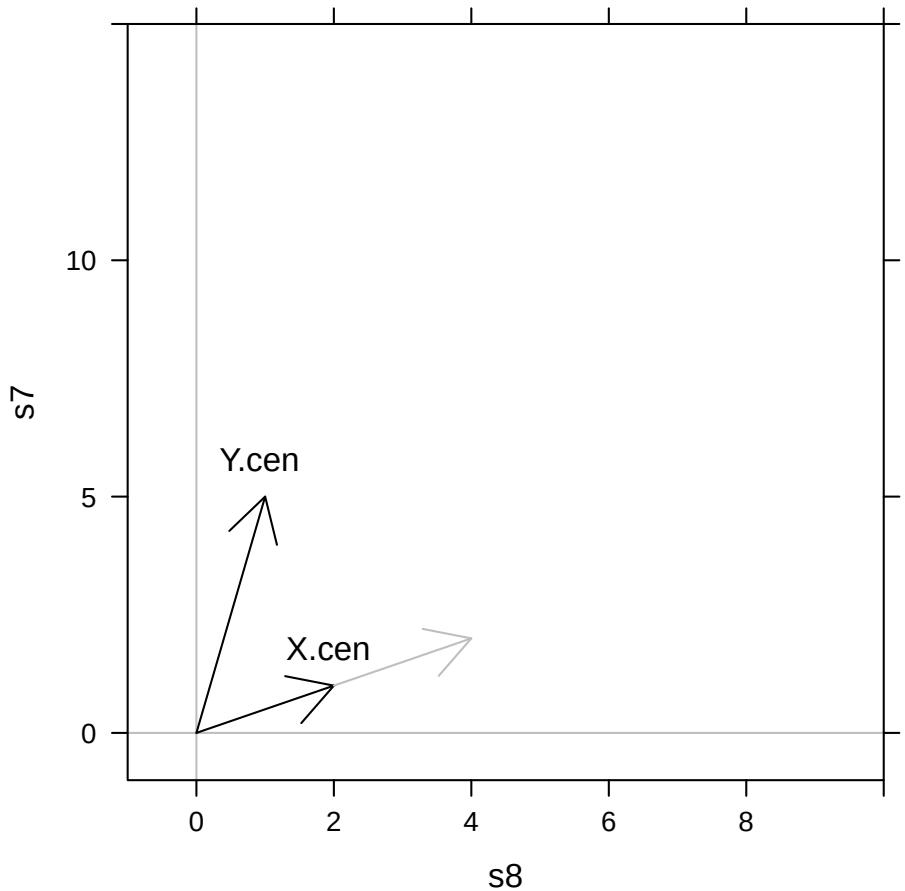
Arithmetical Operations: Linear Combinations



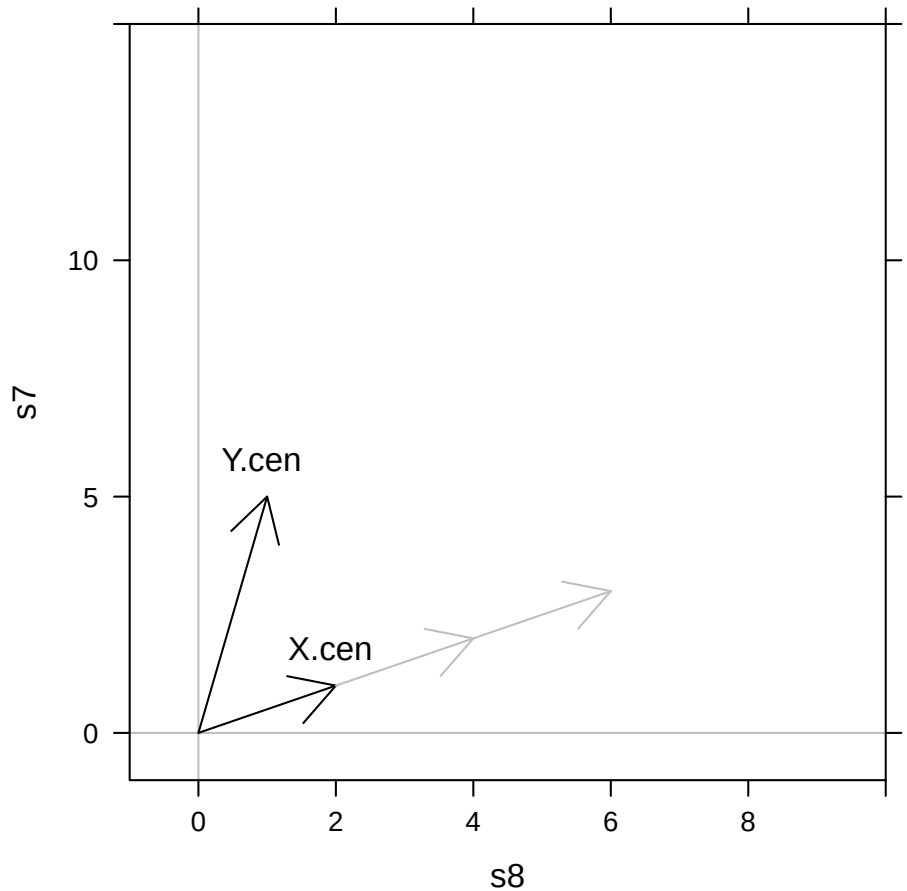
Arithmetical Operations: Linear Combinations



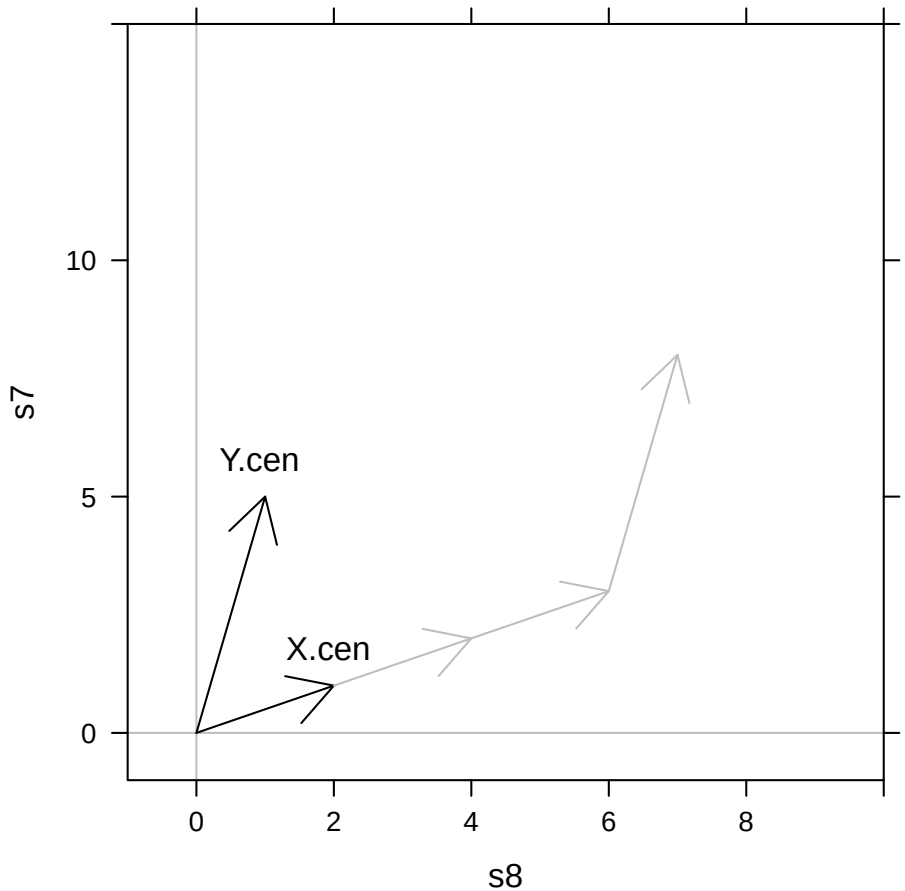
Arithmetical Operations: Linear Combinations



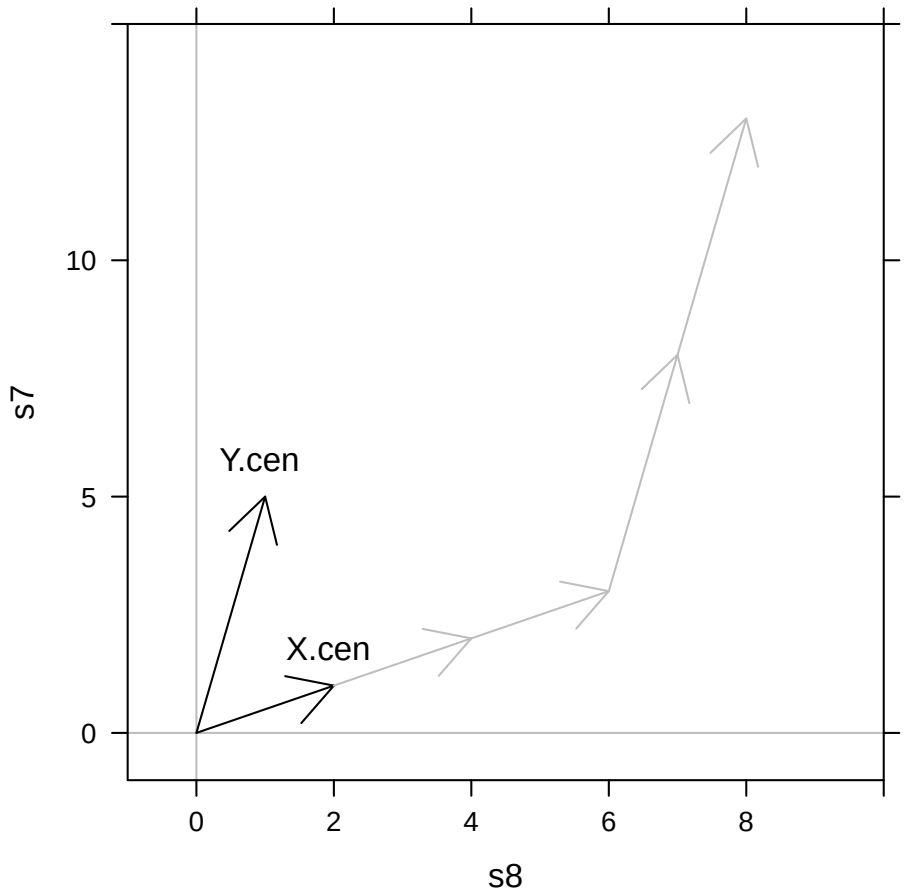
Arithmetical Operations: Linear Combinations



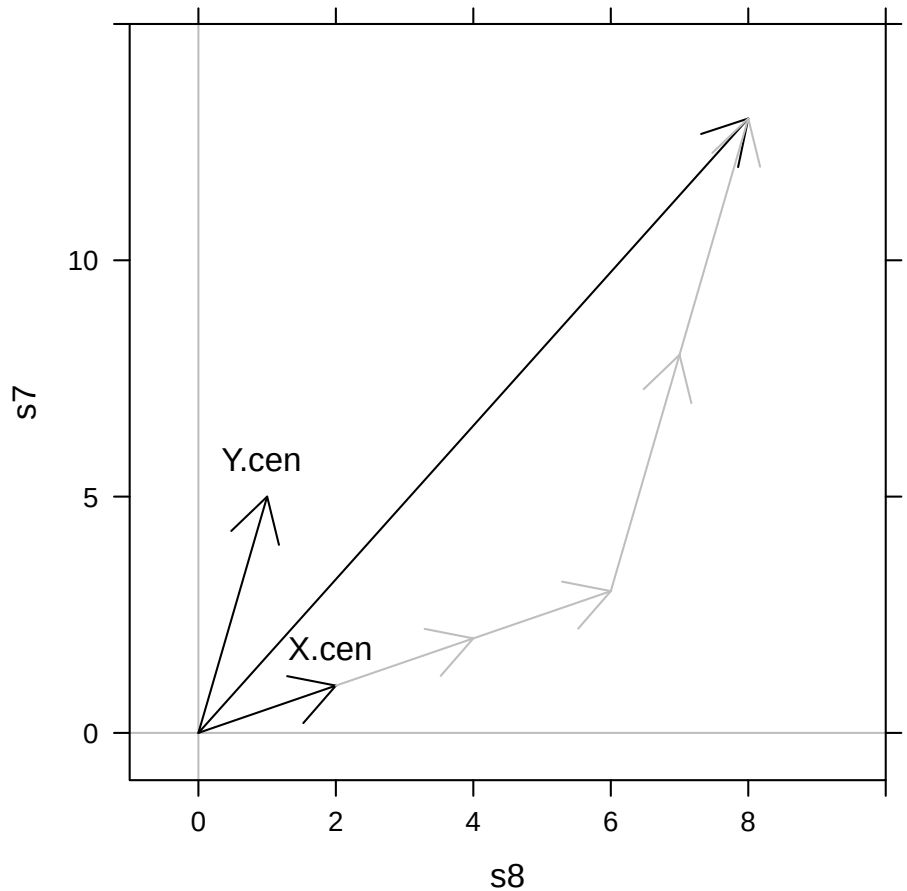
Arithmetical Operations: Linear Combinations



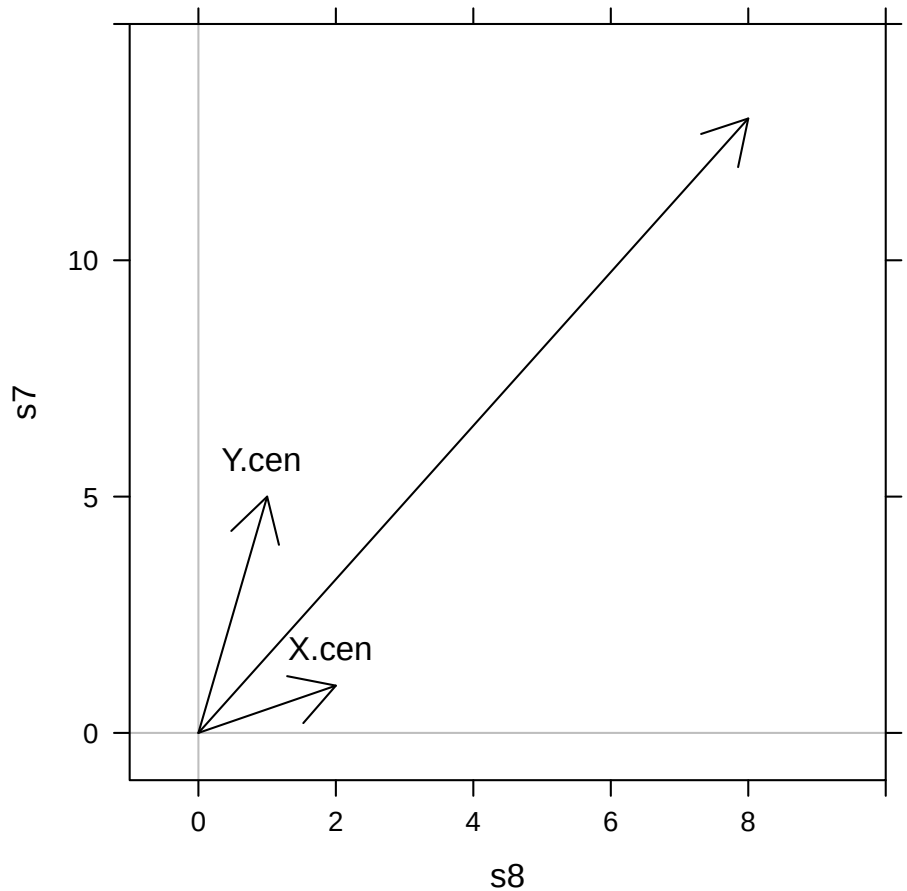
Arithmetical Operations: Linear Combinations



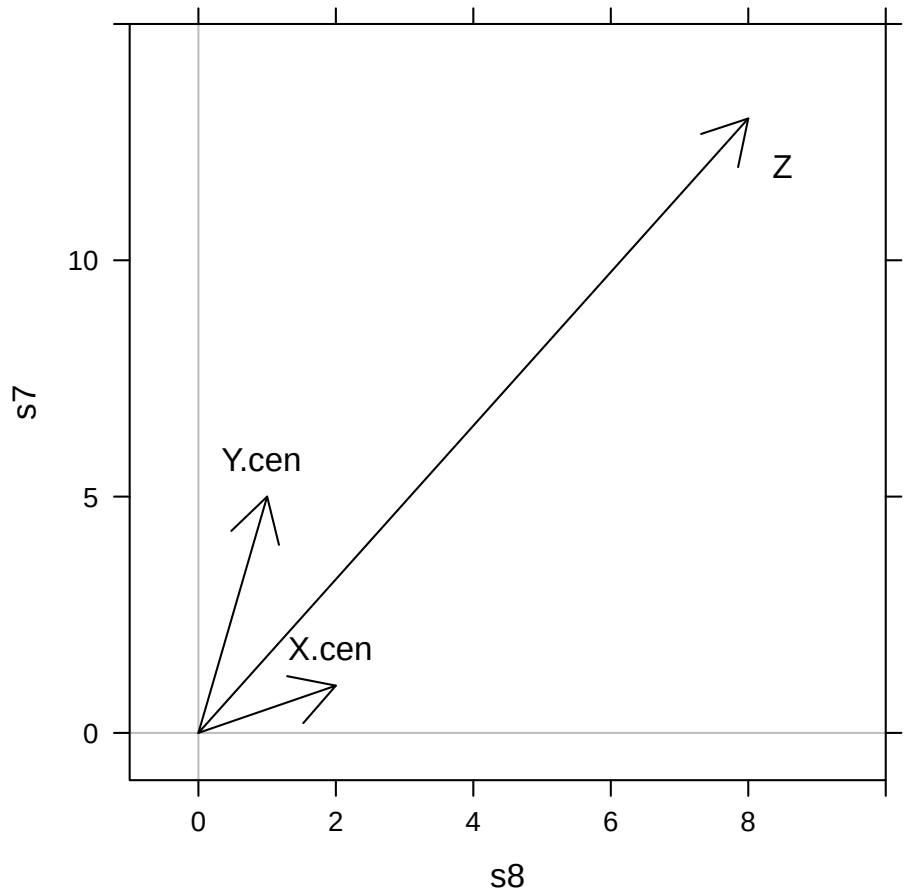
Arithmetical Operations: Linear Combinations



Arithmetical Operations: Linear Combinations



Arithmetical Operations: Linear Combinations



Principal Components Analysis (PCA), cont'd

- PCA (and the SVD that underlies it) are *scale dependent*
- This means we get different solutions depending on the underlying scale of our variables
  - So, variables with greater variance will be highly related to the first component
  - To correct for this, we standardize the data so that each variable has a variance of 1

$$
\begin{aligned}
\boldsymbol{X} &= \boldsymbol{UDV'} \\
\boldsymbol{X'X} &= (\boldsymbol{UDV'})(\boldsymbol{UDV'}) \\
&= \boldsymbol{VDU'UDV'}, \text{ where } \boldsymbol{U'U} = \boldsymbol{I} \\
&= \boldsymbol{VD^2V'}
\end{aligned}
$$

Principal Components Analysis (PCA), cont'd

- $\mathbf{D}$ is a matrix of singular values (a matrix of "stretching/shrinking" values), and $\mathbf{D}^2$ gives the variance explained by each component
- Rows = $\mathbf{UD} \longrightarrow$ this operation produces a matrix of individual PCA scores
- Columns = $\mathbf{DV}' \longrightarrow$ this operation produces the coordinates for the variable vectors in a biplot
- PCA "loadings" = $\mathbf{V}$ matrix of right singular vectors $\longrightarrow$ weight of each variable (column) one the principal components
  - These values are proportional to the correlations between the variables and the princpal components
  - This can be found by multiplying the weights/loadings by the square root of the variance of the associated principal component

Choosing the Number of Components

- The goal of PCA is (or we argue, ought to be) dimension reduction: fewer variables that capture much of the original variance across a set of observed variables
- We are not theorizing about a particular number of dimensions (i.e., there is no underlying model of the data)
- Ultimately, what we find out is that the first m components explain X% of the variance
  - The question is: "is that enough?"
  - Sort of like asking: "Can you get where you need to go on half a tank of gas?"
  - Answer: "Depends on where you are trying to go"

Choosing the Number of Components, cont'd

- That said, there are two major ways of "assessing" dimensionality:
  1. Examination of the proportion of variance explained by each component
  2. Examination of a "scree plot"
- A scree plot is a type of scatterplot where the vertical axis is the singular values or eigenvalues (or proportion of variance explained), and the horizontal axis is the number of components/dimensions/singular values/eigenvalues
- Want to look for an "elbow" – or visually-decipherable bend – in the plot
  - Generally speaking, want to retain the number of components that come before the elbow

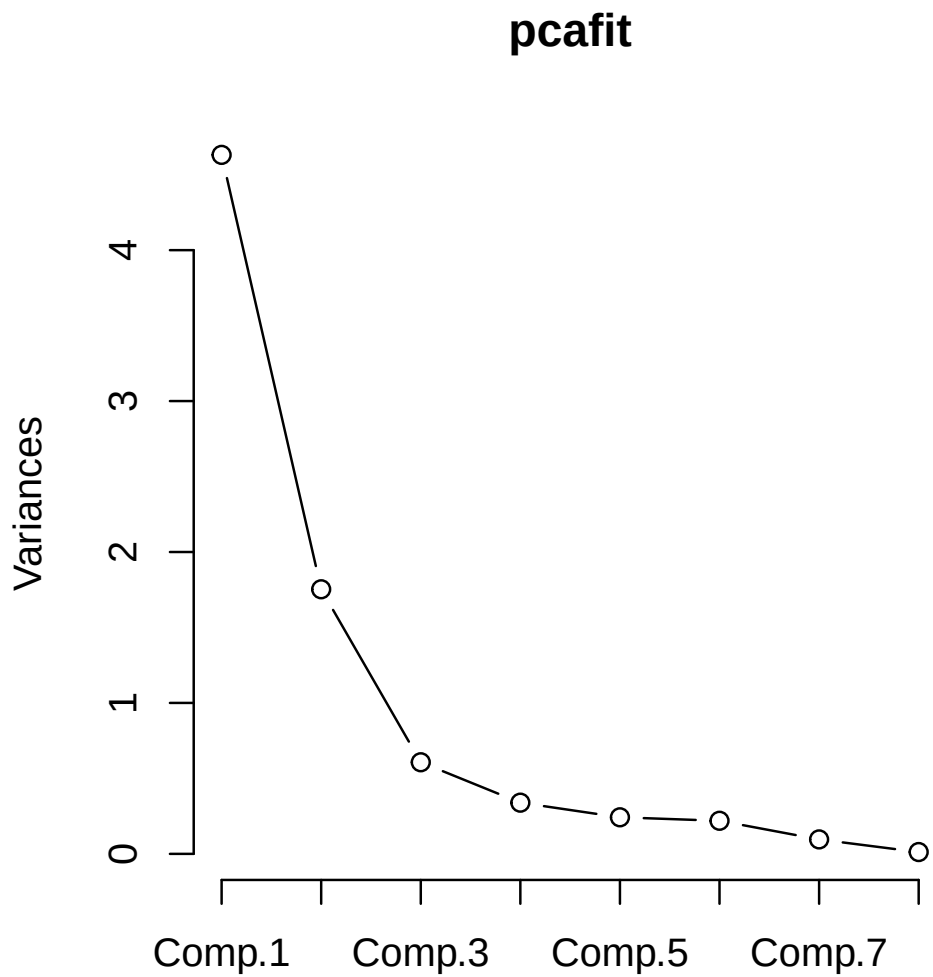
Example: Banks Data

```
> summary(pcafit)

Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     2.1520418 1.3240582 0.77911697 0.58240769 0.49216902 0.46834531
Proportion of Variance 0.5862385 0.2219152 0.07683839 0.04293655 0.03066207 0.02776548
Cumulative Proportion  0.5862385 0.8081537 0.88499207 0.92792862 0.95859069 0.98635617
                           Comp.7      Comp.8
Standard deviation     0.30948211 0.109576817
Proportion of Variance 0.01212395 0.001519883
Cumulative Proportion  0.99848012 1.000000000
```

Example: Banks Data

```
> pcafit$loadings

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
importspc  -0.374  0.187  0.627  0.126 -0.160 -0.323         -0.533
exportspc  -0.368  0.430  0.110         -0.146 -0.231  0.135  0.753
enprodkgpc -0.196  0.558 -0.643  0.108                 0.278 -0.383
enconskgpc -0.413         0.218                 0.862  0.174
wfpcgind   -0.366 -0.279 -0.127  0.622  0.595 -0.104 -0.137
newspaper  -0.348 -0.383 -0.105 -0.557  0.235 -0.294  0.519
literate   -0.283 -0.489 -0.261  0.269 -0.735
gdppcmp    -0.425        -0.195 -0.437                -0.764
```

Scree Plot: Banks Data



Using Principle Components

- PCs can be used to summarize variability across a number of variables that likely capture the same construct
- Retain the number of components necessary to "best" capture variance across the variables
  - Use scree plot, eigenvalues greater than 1 (certainly don't retain components with eigenvalues less than 1)
- Use component scores for row objects (e.g., people, states) in some statistical model, such as a regression analysis

Steps to Conducting PCA

1. Standardize data
   - Standardize columns of original dataset to mean of 0 and standard deviation of 1, or compute correlation matrix
2. Perform matrix decomposition
   - If column standardized original $(n \times m)$ data: SVD
   - If correlation matrix: eigendecomposition
3. Examine proportion of variance explained, and perhaps a scree plot, to help assess dimensionality
4. Examine and interpret component loadings
5. Retain X component loadings for use in subsequent analyses

Example: Political Sophistication

- Sophistication is best thought of not only as knowledge, but interest, participation, and engagement, as well
- Data on from 2012 ANES:
  1. `activity`: index of campaign activities one participated in
  2. `knowledge`: index of number of knowledge questions answered correctly
  3. `interest`: how interested one was in the 2012 U.S. presidential campaign
  4. `pidstrength`: strength of attachment to a political party
  5. `pidstrength`: strength of attachment to an ideological viewpoint/label
  6. `issueextreme`: average extremity of attitudes about several public policy issues

Assumptions and Caveats

- One "assumption" is that you want to explain variance (i.e., that it makes sense to get the correlation matrix among observations)
- A linear combination of observed variables is "best" or at least suitable for your purposes
  - The extent to which variables are linearly related will be the extent to which underlying structure will be identified
- No assumptions of normality (multivariate or otherwise) because we're not interested in sampling distributions here
- The principal components coefficients $\mathbf{V}$ are the values that produce sequentially variance-maximized, orthogonal linear combinations of the observed variables, period. We don't care about a population, and it may not even make sense to talk about population parameters with PCA
- Principal components don't get estimated; they get calculated. Principal components are a well-defined mathematical transformation of observed variables. As such,

Rotation?

- Some people rotate PCA solutions to make loadings more interpretable
  - By taking a new "view" of the data, can increase some loadings, decrease others
  - Can potentially get to a point where patterns in loadings or correlations are "cleaner"
- A laudable goal, but sort of problematic
- Technically PCA solutions shouldn't be rotated because then they aren't PCA solutions any more
  - PCs are orthonormal, sequentially variance-maximizing
  - By rotating, we abandon these properties – for instance, rotation redistributes component variances along the new, rotated...dimensions (can't really call them components)?
  - So...is it still PCA?
- Does it matter?
  - Depends!

Categorical (Nonlinear) PCA

- What if we can't safely assume that our variables behave in a metric fashion (i.e., interval, ratio level of measurement)?
  - Of course, as we already saw, this is empirically testable
- Might want to take steps to provide for non-linear (i.e., monotonic) measurement functions
- Two obvious options:
  1. Submit correlations appropriate for ordinal data (e.g., polychoric correlations)
     - Polychoric correlations estimate the correlation between two theoretically normally distributed continuous latent variables from two observed ordinal variables
  2. Optimal scaling transformations of original data
     - Alter variable measurements to best fit the model (i.e., maximize eigenvalues/variance explained) while preserving some measurement function

Remember ALSOS?

- **A**lternating **L**east **S**quares **O**ptimal **S**caling
  1. The variables are assigned initial optimal scale values, and the measurement characteristics are set
  2. Least-squares estimates are obtained for the parameters of the statistical model
  3. If model fit has not improved over the previous iteration, terminate the procedure; otherwise proceed with the following steps
  4. The predicted values from the statistical model are used to generate new optimal scale values for the variables
  5. Return to Step 2 and re-estimate the model using the updated optimally-scaled variable values
- Combine the `opscale` function with iterated OLS regression estimation
- Result: regression model estimates based on variables measured to maximize fit of the model to the data

Categorical (Nonlinear) PCA via Optimal Scaling

- Remember that optimal scores are estimated in the context of a particular model
- With MORALS, we're trying to calculate optimal scores based on the squared multiple correlation between the independent variables and dependent variable
- In PCA, we're trying to explain variance – maximize singular values/eigenvalues
- PRINCIPALS routine works as follows:
  1. First, determine measurement level and decide how many dimensions (components) to maximize
     - Note: have to determine number of components beforehand or there's nothing to maximize
  2. Compute SVD on original data, reconstruct data using p dimensions from the analysis
  3. Use reconstructed data as predictions by which to compute optimal scores
  4. Use new optimal scores to repeat procedure from Step 2
  5. Terminate when difference in variance explained is trivial

PCA vs. ICA

- Independent Component Analysis is a similar technique with two major differences with PCA
  1. In ICA, all components are equally important; in PCA, they are not
  2. In ICA, the components are not orthogonal; in PCA, they are
- It follows, then, that ICA is really not a data reduction technique in the same way as PCA
- In fact, a PCA or SVD are usually performed on data before an ICA is done
  - This is done precisely to reduce dimensionality and "whiten" the data (i.e., remove the correlations between variables)
  - This helps the ICA algorithm find the "signals" in the dataset
- Essentially, ICA is a form of rotation after a PCA, just like the varimax or quartimax (orthogonal) rotations used more in the social sciences