Lec 06: Dimension Reduction using Principal Components

MATH 456 - Spring 2016

Assigned Reading & additional references

Assigned Reading

- Afifi Chapter 14 (not quite in linear order. See section details below)
- A tutorial on Principal Component Analysis http://arxiv.org/pdf/1404.1100.pdf
- PCA A How-To Manual for R by Emily Mankin (builds off the above paper) http://psych.colorado.edu/wiki/lib/exe/fetch.php?media=labs:learnr:emily_-_principal_components_analysis_in_r:pca_how_to.pdf

Additional references

- Quick-R http://www.statmethods.net/advstats/factor.html
- Computing and visualizing PCA in R via R Bloggers http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/
- Little book of R for Multivariate Analysis http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html#principal-component-analysis
- https://www.youtube.com/watch?v=Heh7Nv4qimU

Introduction (Afifi 14.1, 14.2)

- Primarily used as an exploratory technique
- Restructure interrelated variables
- Simplify description
- Reduce dimensionality
- Avoid multicollinearity problems in regression

It's**"Principal"** Components (adjective), not **"Principle"** Components (noun)

From Grammerist:

> As a noun, principal refers to (1) one who holds a presiding position or rank, and (2) capital or property before interest, and it's also an adjective meaning (3) first or most important in rank
>
> Principle is only a noun. In its primary sense, it refers to a basic truth, law, assumption, or rule.

This third definition (3) is the context in which we will be using this term.

Purpose

Principal Components Analysis (PCA) differs from variable selection in two ways:

1. No dependent variable exists
2. Variables are not eliminated but rather summary variables, i.e., principal components, are computed from all of the original variables.

We are trying to understand a phenomenon by collecting a series of component measurements, but the underlying mechanics is complex and not easily understood by simply looking at each component individually. The data could be redundant and high levels of multicolinearity may be present.

Basic Idea

- Transform correlated variables X_1 and X_2 to uncorrelated variables C_1 and C_2

[top]

Hypothetical Data Example (Afifi 14.3, 14.4)

Consider a hypothetical data set that consists of 100 random pairs of observations X_1 and X_2 where $X_1 \sim \mathcal{N}(100, 100)$, $X_2 \sim \mathcal{N}(50, 50)$, and $\rho_{12} = \frac{1}{\sqrt{2}}$.

In matrix notation this is written as: $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_x\sigma_y \\ \rho_{12}\sigma_x\sigma_y & \sigma_2^2 \end{pmatrix}$$

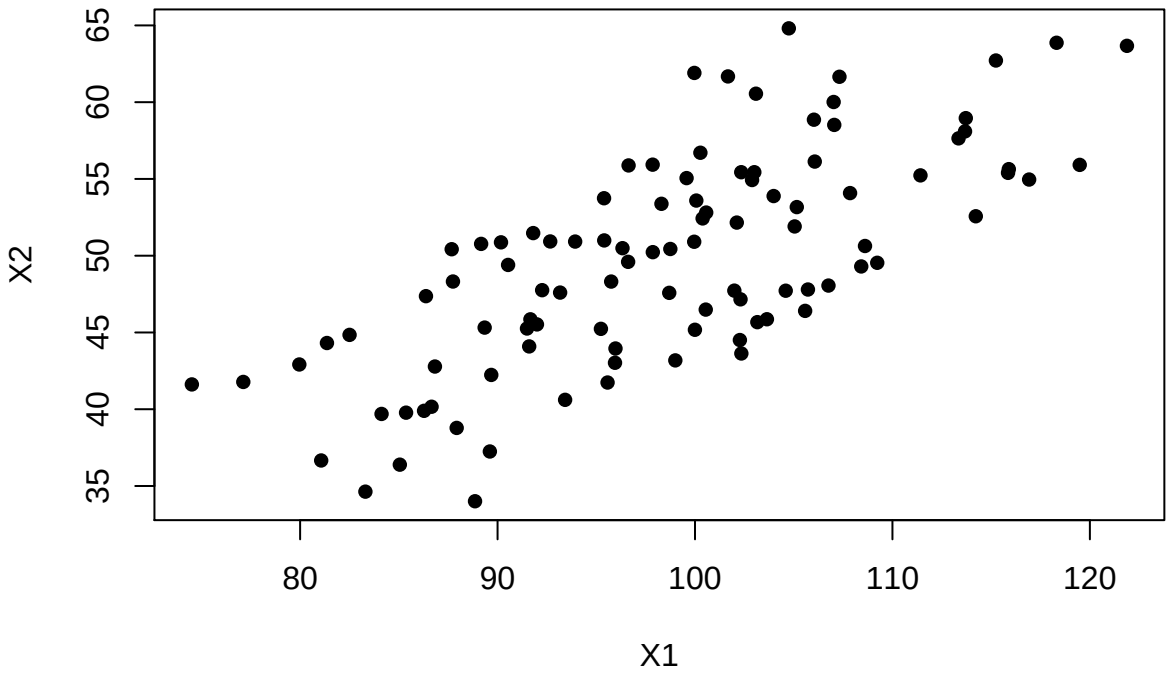
.

Generate hypothetical data Using the `mvrnorm()` function in the `MASS` package to simulate this data,

```
set.seed(456)
library(MASS)
m <- c(100, 50)
s <- matrix(c(100, sqrt(.5*100*50), sqrt(.5*100*50), 50), nrow=2)
data <- data.frame(mvrnorm(n=100, mu=m, Sigma=s))
colnames(data) <- c("X1", "X2")
```

Let's plot this sample data,

```
plot(X2 ~ X1, data=data, pch=16)
```



and calculate the summary statistics (compare to Table 14.1).

```
apply(data, 2, mean)
```

```
##       X1       X2
## 98.55837 49.70860
```

```
apply(data, 2, sd)
```

```
##        X1        X2
## 10.037004  6.971032
```

```
apply(data, 2, var)
```

```
##        X1        X2
## 100.74146  48.59528
```

```
cor(data[,1], data[,2])
```

```
## [1] 0.7187811
```

Generating PC's from this data. Goal: Create two new variables C_1 and C_2 as linear combinations of x_1 and x_2 where

- the x's have been centered by subtracting their mean ($x_1 = X_1 - \bar{X}_1$)
- $Var(C_1)$ is as large as possible
- C_1 and C_2 are uncorrelated

$$C_1 = a_{11}x_1 + a_{12}x_2$$
$$C_2 = a_{21}x_1 + a_{22}x_2$$

or more simply $\mathbf{C} = \mathbf{aX}$.

Calculating the principal components in R is as easy as a call to the function `prcomp`. _Be sure to read the "PCA How To" for the difference between `prcomp` and `princomp`.

```
pr <- princomp(data)
summary(pr)
```

```
## Importance of components:
##                             Comp.1    Comp.2
## Standard deviation     11.4019265 4.2236767
## Proportion of Variance  0.8793355 0.1206645
## Cumulative Proportion   0.8793355 1.0000000
```

- The summary output above shows the first PC (`Comp.1`) has the highest variance.
- The values for the matrix **A** is contained in `pr$loadings`.

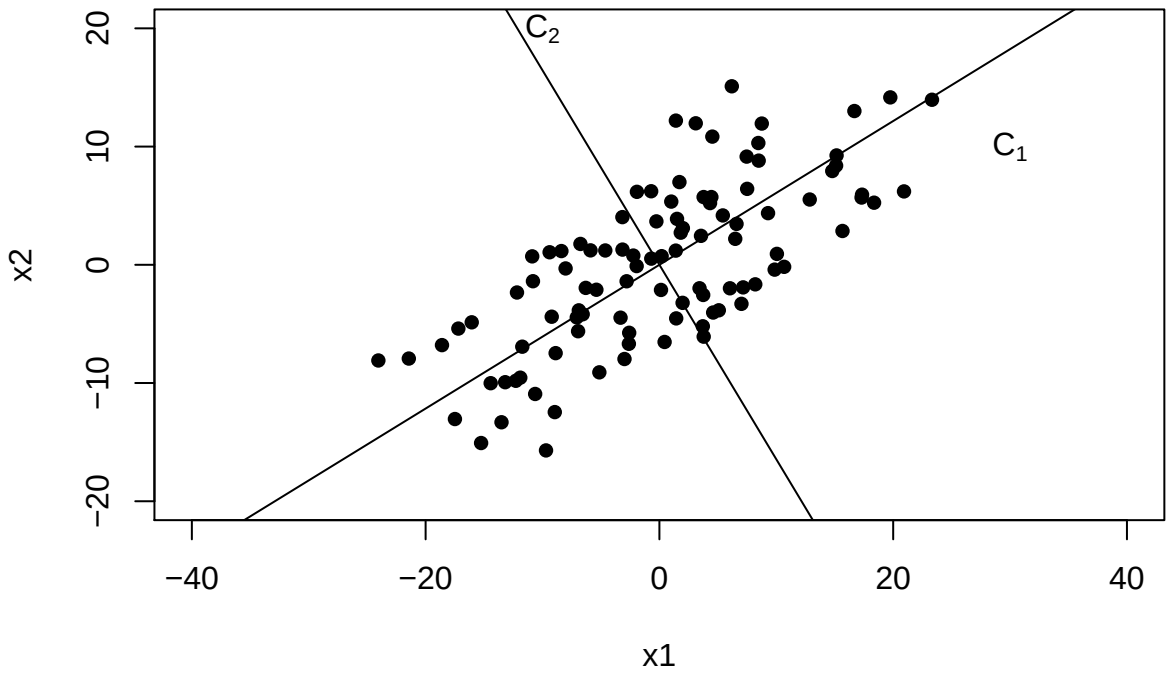
```
pr$loadings
```

```
##
## Loadings:
##    Comp.1 Comp.2
## X1 -0.854  0.519
## X2 -0.519 -0.854
##
##                Comp.1 Comp.2
## SS loadings       1.0    1.0
## Proportion Var    0.5    0.5
## Cumulative Var    0.5    1.0
```

To visualize these new axes, we plot the centered data.

```
a <- pr$loadings
x1 <- with(data, X1 - mean(X1))
x2 <- with(data, X2 - mean(X2))

plot(c(-40, 40), c(-20, 20), type="n",xlab="x1", ylab="x2")
points(x=x1, y=x2, pch=16)
abline(0, a[2,1]/a[1,1]); text(30, 10, expression(C[1]))
abline(0, a[2,2]/a[1,2]); text(-10, 20, expression(C[2]))
```

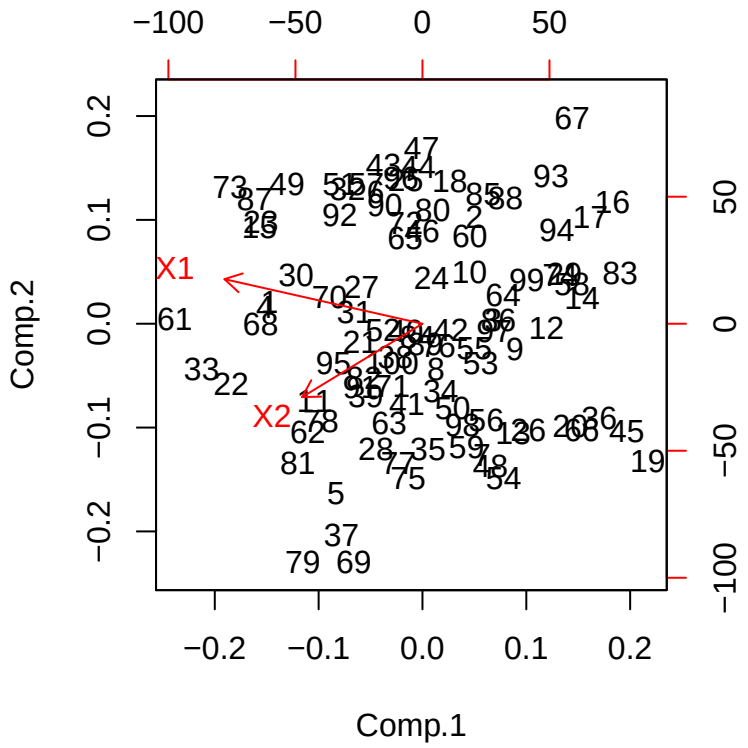


This helps to illustrate that PCA also has a geometric interpretation. From Wikipedia:

> PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

Plot the original data on the new axes we see that PC1 and PC2 are uncorrelated.

```
biplot(pr)
```



[top]

PC Solution on Multivariate Data

We want

- From P original variables $X_1, \ldots, X_P$ get P principal components $C_1, \ldots, C_P$
- Where each C_j is a linear combination of the X_i's: $C_j = a_{j1}X_1 + a_{j2}X_2 + \ldots + a_{jP}X_P$
- The coefficients are chosen such that $Var(C_1) \geq Var(C_2) \geq \ldots \geq Var(C_P)$
- Any two PC's are uncorrelated: $Cov(C_i, C_j) = 0, \quad \forall i \neq j$

We have

$$\begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_P \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1P} \\ a_{21} & a_{22} & \ldots & a_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ a_{P1} & a_{P2} & \ldots & a_{PP} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{bmatrix}$$

.

- Hotelling (1933) showed that the a_{ij}'s are solutions to $(\mathbf{\Sigma} - \lambda \mathbf{I})\mathbf{a}$.
- This means λ is an eigenvalue and $\mathbf{a}$ an eigenvector
- Problem: There are infinite number of possible $\mathbf{a}$'s

- Solution: Choose a_{ij}'s such that the sum of the squares of the coefficients for any one eigenvector is = 1.
  - P unique eigenvalues and P corresponding eigenvectors.

So,

- Principal components are the eigenvectors
- and their variances are the eigenvalues of the covariance matrix $\mathbf{\Sigma}$ of the X's.
- Variances of the C_j's add up to the sum of the variances of the original variables (total variance).

Solving for PC's using the correlation matrix (Afifi 14.5 *Using standardized variables*)

- Standardizing: Take X and divide each element by σ_x.
  - $Z = X/\sigma_X$
- Side note: Standardizing and centering == normalizing
  - $Z = (X - \bar{X})/\sigma_X$
- Equivalent to analyzing the correlation matrix instead of covariance matrix.

Standardizing your data prior to analysis aids the interpretation of the PC's in a few ways

1. The total variance is the number of variables P
2. The proportion explained by each PC is the corresponding eigenvalue / P
3. The correlation between PC C_i and standardized variable x_j can be written as $r_{ij} = a_{ij}SD(C_i)$

This last point means that for any given C_i we can quantify the relative degree of dependence of the PC on each of the standardized variables. This is a.k.a. the **factor loading** (we will return to this key term later).

Hypothetical Data Example continued

Our simulated data has the following correlation matrix

```
cor(data)
```

```
##           X1        X2
## X1 1.0000000 0.7187811
## X2 0.7187811 1.0000000
```

To calculate the principal components using the correlation matrix, you just need to specify that you want `cor=TRUE`.

```
pr_corr <- princomp(data, cor=TRUE)
pr_corr
```

```
## Call:
## princomp(x = data, cor = TRUE)
##
## Standard deviations:
##    Comp.1    Comp.2
## 1.3110229 0.5303008
##
##  2  variables and  100 observations.
```

The first principal component explains `pr_corr$sdev[1]^2/sum(pr_corr$sdev^2)` $= 85.94\%$ of the total variance. Another way to view the proportion, and cumulative proportion, of the total variance each PC explains is to use the `summary` function.

```
summary(pr_corr)
```

```
## Importance of components:
##                           Comp.1    Comp.2
## Standard deviation     1.3110229 0.5303008
## Proportion of Variance 0.8593906 0.1406094
## Cumulative Proportion  0.8593906 1.0000000
```

- If we use the covariance matrix and change the scale of a variable (i.e. in to cm) that will change the results of the PC's
- Many researchers prefer to use the correlation matrix
  - It compensates for the units of measurements for the different variables.
  - Interpretations are made in terms of the standardized variables.

[top]

PC as a Dimension reduction tool (Afifi 14.5 *Number of components retained*)

- Keep the first m PC's as representatives of the original P variables.
- Keep enough PC's to explain a large percentage of total variance.
- How many to keep?
  - Existing theory
  - Explain a given % of variance
  - The "Elbow rule" on a Scree plot.

Example: Analysis of depression data set (Afifi 14.5)

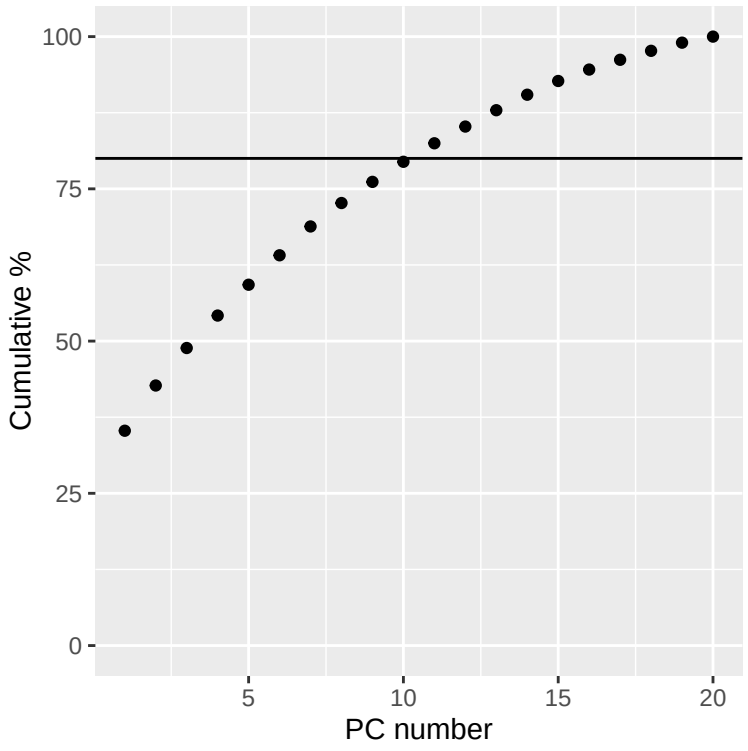
I just show here how to conduct a PCA on the depression data set in R. Carefully read the *Analysis of depression data set* section in Afifi 14.5 for information on what the questions are and how they are used in practice.

```
depress <- read.delim("C:/GitHub/MATH456/data/depress_030816.txt")
pc_dep  <- princomp(depress[,9:28], cor=TRUE)
```

The following plots are helpful in determining how many PC's to retain. Ideally you want a small number of PC's that explain a large percentage of the total variance.

In the cumulative percentage plot below, I drew a horizontal line at 80%. So the first 9PC's explain around 75% of the total variance, and the first 10 can explain around 80%.

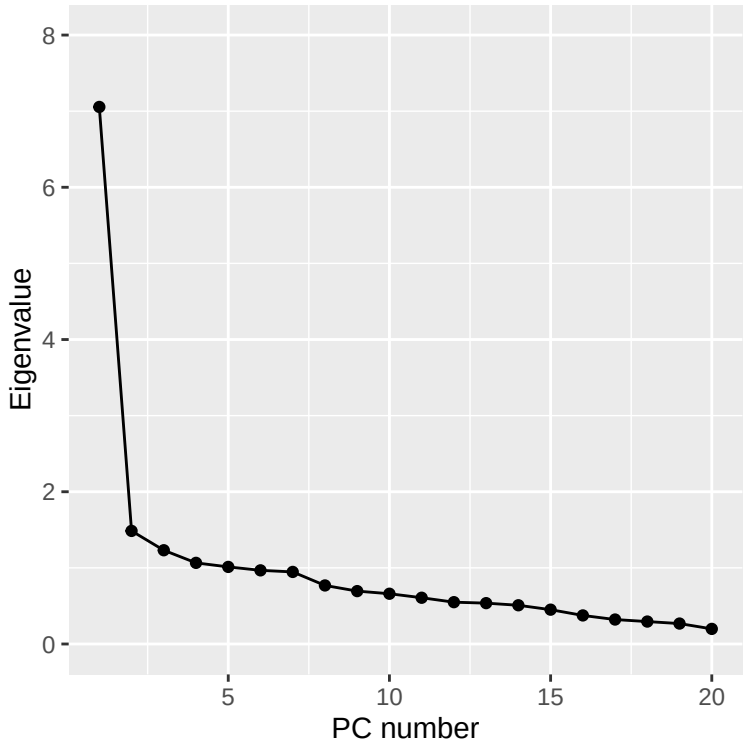
```
qplot(x=1:20, y=cumsum((pc_dep$sdev)^2/20)*100, geom="point") +
  xlab("PC number") + ylab("Cumulative %") + ylim(c(0,100)) +
  geom_hline(aes(yintercept=80))
```



Elbow Rule

- Create a **Scree plot** by plotting the eigenvalue against the PC number.
- Use a cutoff point where the lines joining consecutive points are steep to the left of the cutoff point and flat right of the cutoff point.
- Point where the two slopes meet is the elbow.
- In this plot $m = 2$.

```
qplot(x=1:20, y=(pc_dep$sdev)^2, geom=c("point", "line")) +
  xlab("PC number") + ylab("Eigenvalue") + ylim(c(0,8))
```



Variable groupings

PC's are a very good way to see how component variables group, or cluster together. Or if a set of questions measure the "same" thing.

I do a little data management here to extract the loadings for the first five PC's,

```
pc15 <- data.frame(round(pc_dep$loadings[,1:5],4))
```

then use Table 14.2 to identify which variables are in which theoretical group,

```
item.group <- c(rep("Negative", 7), rep("Positive", 4), rep("Somatic", 7), rep("Interpersonal", 2))
```

combine the item group, an variable number index (1 to 20), and the reshaped loading data. Here I am using `melt()` to reshape the loading data from wide to long.

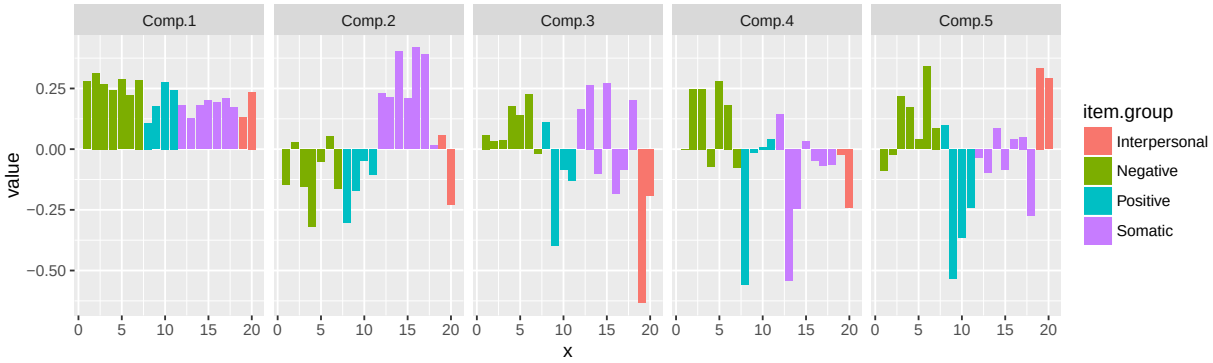
```
library(reshape2)
melted <- cbind(item.group, x=1:20, melt(pc15))
head(melted)
```

```
##   item.group x variable  value
## 1   Negative 1   Comp.1 0.2774
## 2   Negative 2   Comp.1 0.3132
## 3   Negative 3   Comp.1 0.2678
## 4   Negative 4   Comp.1 0.2436
## 5   Negative 5   Comp.1 0.2868
## 6   Negative 6   Comp.1 0.2206
```

I can now create a visualization where

- the x-axis is the question number (C1-C20)
- the height of the bar represents the loading value
- Each principal component is shown on it's own panel
- the questions are colored by the theoretical variable groupings.

```
ggplot(data=melted) +
  geom_bar(aes(x=x, y=value, fill=item.group), stat="identity") +
  facet_wrap(~variable, ncol=5)
```



This shows that the first PC is almost an average of all questions, the Somatic questions load heavily on PC2, so PC2 can be thought of as a measure of Somatic, retarded activity. Interpersonal questions load strongly negative on PC3, and positive on PC5.

Reminder

- Principal components are *linear combinations* of the component variables.
- We've seen methods to create combined scales as averages or sums.
- These averages are *unweighted* averages.
  - Each component variable contributes equally
- PC's are *weighted averages* of the component variables.

Use in Multiple Regression

- Discard last few principal components, and perform regression on remaining. Leads to more stable regression estimates.
- Alternative to variable selection
  - Ex: several measures of behavior.
  - Use PC_1 or PC_1 and PC_2 as summary measures of all.

Multicollinearity

The size of variance of last few principal components can be useful as indicator of multicollinearity among original variables

- $PC_{10} = .5X - .2Y - .25Z, \lambda_{10} = .01 \approx 0$
- So $X \approx .4Y + .5Z$
- Decision: discard X

Things to watch out for

- Eigenvalues are estimated variances of the PC's and so are subject to large sample variations.

  Arbitrary cutoff points should not be taken too seriously

- Principal components derived from standardized variables differ from those derived from original variables
- Interpretation is easier if data arise from or are transformed to have symmetric distribution
- Important that measurements are accurate, especially for detection of collinearity

PCA vs Factor Analysis

These are two similar techniques, sometimes used for similar purposes. However the underlying theory behind each is very different. We will not cover Factor Analysis in this class but it is very important for you to understand the differences, and when you would use one over the other.

Here are some references that can start to explain the differences. Consider them **additional readings**

- **Univ Wisconsin Madison School of Psychology** http://psych.wisc.edu/henriques/pca.html
- **Minitab** http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/multivariate/principal-components-and-factor-analysis/differences-between-pca-and-factor-analysis/
- **Stack Overflow** http://stats.stackexchange.com/questions/1576/what-are-the-differences-between-factor-analysis-and-p

[top]

On Your Own

On Your Own

1. Confirm the total variance in the hypothetical example is preserved by calculating the sum of the variances for each X_1 and X_2, and C_1 and C_2.
2. Confirm the total variance of the principal components when using the correlation matrix is equal to the number of principal components.
3. According to the PCA How-To manual by Emily Mankin, what are the two Principles of PCA?
4. PCA vs FA
   a. In your own words, what is the differnce between PCA and FA?
   b. Explain what is meant by a "latent factor".
5. For the depression data set, perform a PCA on the last seven variables `DRINK - CHRONILL`. Interpret the results.
6. Using the family lung function dat,a perform a PCA on mother's height, weight, age, FEV1, and FVC. Use the covariance matrix, then repeat using the correlation matrix. Compare the results and comment.