# PLSC 597: Modern Measurement

## Principal Components and Factor Analysis

February 8, 2018

# The Plan

- Principal Components
- Biplots!
- Exploratory Factor Analysis
- Diagnostics, etc.

```
> X <- data.frame(X1=c(0,1,2),X2=c(6,5,3),X3=c(7,9,10))
> X
  X1 X2 X3
1  0  6  7
2  1  5  9
3  2  3 10

> CX <- sweep(M,2,colMeans(M),"-")  # "centered" M
> CX
  X1      X2      X3
1 -1  1.3333 -1.6667
2  0  0.3333  0.3333
3  1 -1.6667  1.3333
```

```
> Sigma <- cov(CX)
> Sigma
     X1     X2     X3
X1  1.0 -1.500  1.500
X2 -1.5  2.333 -2.167
X3  1.5 -2.167  2.333

> R <- cor(CX)
> R
       X1      X2      X3
X1  1.000 -0.9820  0.9820
X2 -0.982  1.0000 -0.9286
X3  0.982 -0.9286  1.0000
```

# Eigenvalues and Eigenvectors

For the variance-covariance matrix $\Sigma$ of (centered) $\mathbf{X}$, we can diagonalize:

$$\Sigma = \mathbf{VLV}'$$

where

- $\mathbf{V}$ is the matrix of *eigenvectors* ("principal axes"), and
- $\mathbf{L}$ is the (diagonal) matrix of *eigenvalues*.

Things:

- The sum of the eigenvalues equals the trace of $\Sigma$
- The product of the eigenvalues is $|\Sigma|$

## Eigenvalues and Eigenvectors

```
> E <- eigen(Sigma)
> E
$values
[1] 5.5000000000 0.1666666666667407 0.00000000000001776

$vectors
        [,1]   [,2]    [,3]
[1,]  0.4264 0.0000  0.9045
[2,] -0.6396 0.7071  0.3015
[3,]  0.6396 0.7071 -0.3015

> L <- E$values
> V <- E$vectors
>
> sum(E$values)
[1] 5.667
> tr(Sigma)
[1] 5.667
```

# Singular Value Decomposition

The *singular value decomposition* (SVD) of $\mathbf{X}$ is:

$$\mathbf{X} = \mathbf{USV}'$$

where $\mathbf{S}$ is the diagonal matrix of *singular values*, $\mathbf{U}$ is a unitary (orthogonal) matrix, and $\mathbf{V}$ is again the matrix of eigenvectors.

Note:

- Elements of $\mathbf{S}$ $s_i$ are related to the eigenvalues $v_i$ according to $v_i = s_i^2/(N-1)$.
- The *principal components* are equal to $\mathbf{US}$ ($\equiv \mathbf{XV}$).

# SVD

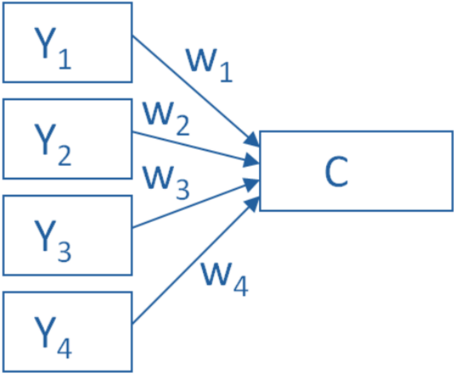
```
> SVD <- svd(CX)
> SVD
$d
[1] 3.3166247903553993659 0.5773502691896256200 0.0000000000000004209

$u
                      [,1]    [,2]   [,3]
[1,] -0.7071067811865470176  0.4082 0.5774
[2,]  0.0000000000000001665 -0.8165 0.5774
[3,]  0.7071067811865475727  0.4082 0.5774

$v
        [,1]       [,2]    [,3]
[1,]  0.4264  3.332e-17 -0.9045
[2,] -0.6396 -7.071e-01 -0.3015
[3,]  0.6396 -7.071e-01  0.3015

> S <- SVD$d
> U <- SVD$u
> otherV <- SVD$v
>
> # Eigenvalues:
>
> (S^2)/(2)
[1] 5.500e+00 1.667e-01 8.858e-32
```

## Principal Components (PCA)

PCA is:

- an orthogonal transformation, that
- converts a set of variables $\underset{N \times K}{\mathbf{X}}$ into a set of $K$ linearly-uncorrelated values, where
- the first principal component has the largest possible variance, and
- the second has the second-highest (subject to orthogonality),
- etc.

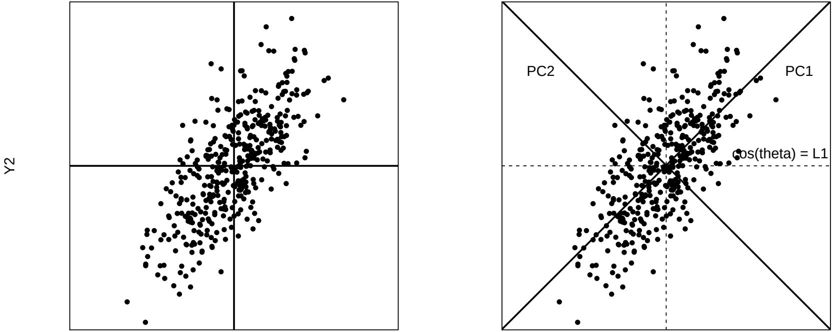
# PCA, Conceptually




$$C = w_1 Y_1 + w_2 Y_2 + w_3 Y_3 + w_4 Y_4$$

(Source)

# PCA Intuition




"(Principal components) can be considered as a rotation of original variable coordinate system to new (orthogonal) axes... such that the new axes coincide with the directions of maximum variation in the original observations." (Campbell and Atchley 1981)

```
> princomp(CX) # via eigenvalues
Call:
princomp(x = CX)

Standard deviations:
      Comp.1       Comp.2       Comp.3
1.9148542155 0.3333333333 0.0000000365

 3  variables and  3 observations.

> prcomp(CX) # via SVD
Standard deviations:
[1] 2.345e+00 4.082e-01 6.833e-18

Rotation:
       PC1        PC2     PC3
X1  0.4264  1.071e-17  0.9045
X2 -0.6396 -7.071e-01  0.3015
X3  0.6396 -7.071e-01 -0.3015

> otherV # from -svd-
        [,1]       [,2]    [,3]
[1,]  0.4264  3.332e-17 -0.9045
[2,] -0.6396 -7.071e-01 -0.3015
[3,]  0.6396 -7.071e-01  0.3015
```

# PCA Steps

- *Extract* the principal components
- *Interpret* the components...
- Consider *rotation*
- Choosing the *number of components* (dimensions)
- Generating *scores*

# PCA: A Simulation Example

```
> N <- 20
> set.seed(7222009)
> Name <- randomNames(N, which.names="first")
> Z <- rnorm(N)
> Z1 <- Z + 0.2*rnorm(N)
> Z2 <- Z + 0.5*rnorm(N)
> Z3 <- Z + 1*rnorm(N)
> Z4 <- Z + 1.5*rnorm(N)
> Z5 <- Z + 2*rnorm(N)
> Z6 <- Z + 3*rnorm(N)
>
> X <- rnorm(N)
> X1 <- X + rnorm(N)
> X2 <- X + rnorm(N)
> X3 <- X + rt(N,5)
> X4 <- X + rt(N,5)
>
> df <- data.frame(Z1,Z2,Z3,Z4,Z5,Z6,X1,X2,X3,X4)
> rownames(df)<-Name
```

# PCA Simulation (continued)

```
> head(df)
               Z1       Z2      Z3      Z4      Z5     Z6      X1      X2       X3      X4
Guillermo -1.2792 -1.19146 -2.6476 -0.4796 -2.5831 -2.499  1.8278  4.7884  0.00842  2.7039
Rachel     0.6212  0.76287  0.8448  0.4931  0.8397 -5.000 -2.8699 -2.3021 -1.93999  2.7174
Deidra     1.0345  0.90827  1.2779 -0.5136  0.1952 -1.010 -0.5812 -2.1720  0.34358  1.1017
Quaton    -0.2520 -0.09033 -0.4566  2.4858 -0.1768 -1.144  2.6244  1.2543  3.40352  0.7601
Alicia    -0.3546 -0.88944  1.4155  1.7511 -3.3356  4.756 -0.7656  0.9322  0.25697  1.0309
Angelique  0.6942 -0.19473  1.6100 -0.8439  2.8762  2.454 -3.5113 -0.7967 -0.47199 -2.3341
```

```
> cor(df)
         Z1       Z2      Z3      Z4      Z5       Z6       X1      X2       X3       X4
Z1  1.00000  0.84292  0.7952 0.29558  0.5756  0.23931 -0.46737 -0.3180 -0.02362  0.07671
Z2  0.84292  1.00000  0.5092 0.38871  0.3820  0.03809 -0.28794 -0.3016 -0.04371  0.10192
Z3  0.79518  0.50921  1.0000 0.16477  0.5474  0.49522 -0.66286 -0.4519 -0.23495 -0.19329
Z4  0.29558  0.38871  0.1648 1.00000  0.1804  0.23241  0.04063  0.2556  0.20279  0.26906
Z5  0.57556  0.38201  0.5474 0.18041  1.0000  0.14471 -0.46899 -0.2477 -0.16443 -0.21052
Z6  0.23931  0.03809  0.4952 0.23241  0.1447  1.00000 -0.09680  0.2533  0.00289 -0.27962
X1 -0.46737 -0.28794 -0.6629 0.04063 -0.4690 -0.09680  1.00000  0.5941  0.57578  0.25633
X2 -0.31804 -0.30161 -0.4519 0.25556 -0.2477  0.25327  0.59415  1.0000  0.48818  0.35794
X3 -0.02362 -0.04371 -0.2350 0.20279 -0.1644  0.00289  0.57578  0.4882  1.00000  0.50038
X4  0.07671  0.10192 -0.1933 0.26906 -0.2105 -0.27962  0.25633  0.3579  0.50038  1.00000
```

# PCA Simulation (continued)

```
> PCE <- princomp(df)
> PCE
Call:
princomp(x = df)

Standard deviations:
 Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9 Comp.10
 3.4376  2.9921  2.1887  1.5840  1.3081  1.1704  0.8357  0.7074  0.4936  0.1870

 10  variables and  20 observations.

> PCS <- prcomp(df,retx=FALSE)
> PCS
Standard deviations:
 [1] 3.5269 3.0698 2.2455 1.6252 1.3420 1.2008 0.8574 0.7258 0.5064 0.1919

Rotation:
        PC1      PC2       PC3      PC4       PC5       PC6       PC7     PC8      PC9      PC10
Z1  0.14329 -0.08291  0.221066 -0.16453  0.116250 -0.076287 -0.295228  0.3737  0.01414 -0.806425
Z2  0.07769 -0.08154  0.195924 -0.16972  0.006413 -0.265866 -0.400862  0.4689  0.45834  0.508674
Z3  0.35128 -0.14139  0.145255 -0.30868  0.229943  0.004829 -0.094717  0.1058 -0.76632  0.282703
Z4  0.09075  0.13402  0.371057 -0.16379 -0.563079 -0.613069  0.289656 -0.1405 -0.09645 -0.058551
Z5  0.37041 -0.32978  0.541407  0.61985 -0.006412  0.151054 -0.066350 -0.2057  0.07598  0.036762
Z6  0.73986  0.55782 -0.190833 -0.08766  0.062539  0.066331 -0.050980 -0.1848  0.22778 -0.020532
X1 -0.26633  0.37528  0.004739  0.31738  0.171189 -0.382348 -0.611576 -0.2882 -0.23434 -0.002257
X2 -0.09196  0.49655  0.148655  0.29742 -0.387299  0.371810  0.001255  0.5323 -0.24248  0.063619
X3 -0.15372  0.34087  0.401467  0.04308  0.652815 -0.138273  0.475659  0.1274  0.09122  0.032585
X4 -0.23278  0.15817  0.491046 -0.48895 -0.084182  0.465670 -0.228032 -0.3879  0.11616  0.026319
```

# Friendly PCA using principal

```
> PCSim1 <- principal(df, nfactors=1,rotate="none")
> PCSim1
Principal Components Analysis
Call: principal(r = df, nfactors = 1, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1    h2   u2 com
Z1  0.83 0.694 0.31   1
Z2  0.67 0.454 0.55   1
Z3  0.89 0.798 0.20   1
Z4  0.17 0.030 0.97   1
Z5  0.69 0.482 0.52   1
Z6  0.29 0.083 0.92   1
X1 -0.79 0.618 0.38   1
X2 -0.61 0.372 0.63   1
X3 -0.44 0.194 0.81   1
X4 -0.31 0.093 0.91   1

                PC1
SS loadings    3.82
Proportion Var 0.38

Mean item complexity =  1
Test of the hypothesis that 1 component is sufficient.

The root mean square of the residuals (RMSR) is  0.2
 with the empirical chi square  72.9  with prob <  0.00018

Fit based upon off diagonal values = 0.71
```

# Let's break that down...

- It's a PCA, where we're extracting the first principal component (`nfactors = 1`)
- No rotation (`rotate = "none"`)
- `PC1` are the "loadings" of each variable on the first principal component (think of these as the $w_k$ in the conceptual figure)
- `h2` are *communalities*; the sums of the squared factors loadings (so, here, $\texttt{PC1}^2$)
- `u2` is *uniqueness*; simply $1-$ `h2`
- `SS Loadings` is the value(s) of the principal component(s)
- `Proportion Var` is the proportion of the total variance in **X** that that principal component accounts for

# PCA Scores

```
> PCSim1$scores

               PC1
Guillermo -2.11824
Rachel     0.78544
Deidra     0.63716
Quaton    -0.92891
Alicia    -0.33762
Angelique  1.10640
Johnaton  -0.49378
Javan      0.23606
Khulood    1.06142
Cody      -0.07695
Cameron   -0.02655
Heidi      1.41409
Maahir     1.57304
Rogelio   -0.15324
Erica     -0.11558
Barren    -1.57311
Kiana      1.04688
Elyse     -1.19026
Chadrick  -0.41816
Tahani    -0.42809
```

# PCA with `nfactors = 2`

```
> PCSim2 <- principal(df, nfactors=2,rotate="none")
> PCSim2
Principal Components Analysis
Call: principal(r = df, nfactors = 2, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1  PC2   h2   u2 com
Z1  0.83 0.44 0.89 0.11 1.5
Z2  0.67 0.45 0.66 0.34 1.8
Z3  0.89 0.13 0.81 0.19 1.0
Z4  0.17 0.71 0.53 0.47 1.1
Z5  0.69 0.11 0.49 0.51 1.1
Z6  0.29 0.24 0.14 0.86 1.9
X1 -0.79 0.32 0.72 0.28 1.3
X2 -0.61 0.52 0.64 0.36 1.9
X3 -0.44 0.68 0.65 0.35 1.7
X4 -0.31 0.63 0.49 0.51 1.4

                       PC1  PC2
SS loadings           3.82 2.21
Proportion Var        0.38 0.22
Cumulative Var        0.38 0.60
Proportion Explained  0.63 0.37
Cumulative Proportion 0.63 1.00

Mean item complexity =  1.5
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.12
 with the empirical chi square  25.8  with prob <  0.47

Fit based upon off diagonal values = 0.9
```

# Let's break that down again...

- It's a PCA, where now we're extracting the first two principal components (`nfactors = 2`)
- No rotation (`rotate = "none"`)
- `PC1`, `PC2`, `h2`, and `u2` are the same as above
- `com` is the *complexity* $c_k$ of each measure; $c_k = \frac{(\sum PC_k^2)^2}{\sum PC_k^4}$
- `SS Loadings` are again the value(s) of the principal component(s)
- There are now both total and cumulative variance explained statistics
- The model fit statistic now suggests that the model

```
> PCSim2$scores
               PC1     PC2
Guillermo -2.11824  0.3897
Rachel     0.78544 -0.0387
Deidra     0.63716  0.3501
Quaton    -0.92891  1.5769
Alicia    -0.33762  0.5989
Angelique  1.10640 -0.6102
Johnaton  -0.49378 -0.7382
Javan      0.23606 -0.1504
Khulood    1.06142 -1.1449
Cody      -0.07695 -1.1076
Cameron   -0.02655  1.5239
Heidi      1.41409  1.0112
Maahir     1.57304  1.0410
Rogelio   -0.15324  0.4666
Erica     -0.11558 -0.7050
Barren    -1.57311 -1.3553
Kiana      1.04688 -0.8571
Elyse     -1.19026  0.8022
Chadrick  -0.41816  0.7941
Tahani    -0.42809 -1.8473
```

# PCA with `nfactors = 3`

```
> PCSim3 <- principal(df, nfactors=3,rotate="none")
> PCSim3
Principal Components Analysis
Call: principal(r = df, nfactors = 3, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1  PC2   PC3   h2   u2 com
Z1  0.83 0.44 -0.15 0.91 0.09 1.6
Z2  0.67 0.45 -0.33 0.77 0.23 2.3
Z3  0.89 0.13  0.19 0.85 0.15 1.1
Z4  0.17 0.71  0.11 0.54 0.46 1.2
Z5  0.69 0.11  0.05 0.50 0.50 1.1
Z6  0.29 0.24  0.88 0.91 0.09 1.4
X1 -0.79 0.32  0.07 0.73 0.27 1.3
X2 -0.61 0.52  0.40 0.80 0.20 2.7
X3 -0.44 0.68 -0.05 0.65 0.35 1.7
X4 -0.31 0.63 -0.48 0.72 0.28 2.4

                       PC1  PC2  PC3
SS loadings           3.82 2.21 1.35
Proportion Var        0.38 0.22 0.14
Cumulative Var        0.38 0.60 0.74
Proportion Explained  0.52 0.30 0.18
Cumulative Proportion 0.52 0.82 1.00

Mean item complexity =  1.7
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.08
 with the empirical chi square  11.12  with prob <  0.89

Fit based upon off diagonal values = 0.96
```

# Biplots!

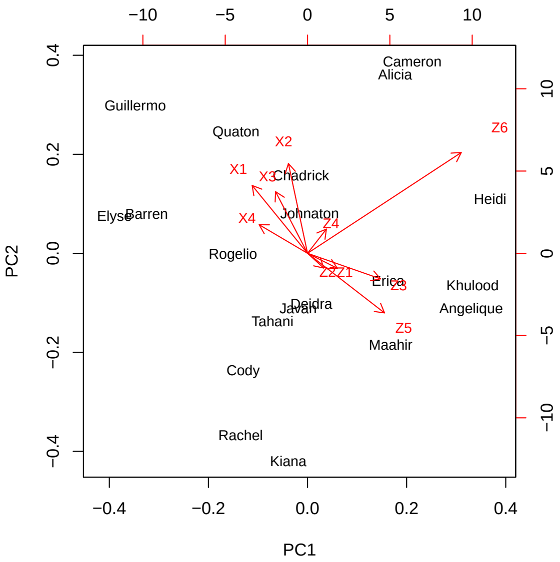
A *biplot* is a graphical representation of a two-axis PCA.

- It plots both loadings (of variables) and scores (of observations)
- It represents the former as vectors from the origin, and the latter as points in the (transformed) space
- Interpretation:
    - Angles between item vectors represent degrees of correlation/covariance
    - Distances between points reflect dissimilarities between those observations
- Details are in Gower and Hand (1996) and Jacoby (1998, Chapter 7)

# Biplot Basics (Simulation Data)

```
> foo<-prcomp(df)

> foo$rotation[,1:2]
         PC1      PC2
Z1  0.14329 -0.08291
Z2  0.07769 -0.08154
Z3  0.35128 -0.14139
Z4  0.09075  0.13402
Z5  0.37041 -0.32978
Z6  0.73986  0.55782
X1 -0.26633  0.37528
X2 -0.09196  0.49655
X3 -0.15372  0.34087
X4 -0.23278  0.15817
```

# A Biplot...



−10
−5
0
5
10
0.4
0.2
0.0
−0.2
−0.4
PC2
Cameron
Alicia
Guillermo
Quaton
X2
Z6
X1
X3
Chadrick
Heidi
Elyse
Barren
X4
Johnaton
Z4
Rogelio
Z2
Z1
Erica
Z3
Khulood
Deidra
Javan
Angelique
Tahani
Z5
Maahir
Cody
Rachel
Kiana
−0.4
−0.2
0.0
0.2
0.4
PC1

# (Exploratory) Factor Analysis

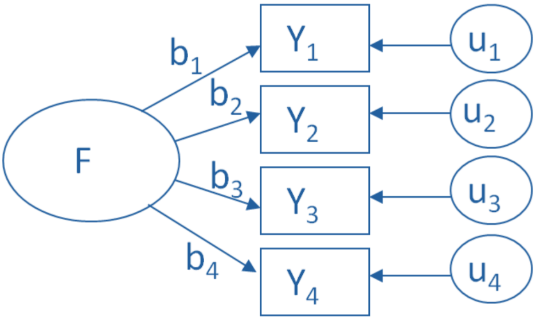
*Factor analysis* (FA) is a model for the measurement of a latent variable using manifest / observable indicators.

- Observable indicators are manifestations of one or more latent / unobservable *factors*

- Extant indicators are differentially caused by the latent factor(s), and are observed with error

- The goal of FA is to derive measures of the latent factor from the observed data, by estimating factor *loadings* (associations between latent factors and observable variables)

# Factor Analysis, Conceptually




$$
\begin{aligned}
Y_1 &= b_1 F + u_1 \\
Y_2 &= b_2 F + u_2 \\
Y_3 &= b_3 F + u_3 \\
Y_4 &= b_4 F + u_4
\end{aligned}
$$

(Source)

Formally:

$$\mathbf{Y} = \Lambda \mathbf{F} + \mathbf{U}$$

This implies that the observed covariance matrix $\Sigma$ can be written:

$$\Sigma = \Lambda\Lambda' + \Psi$$

where

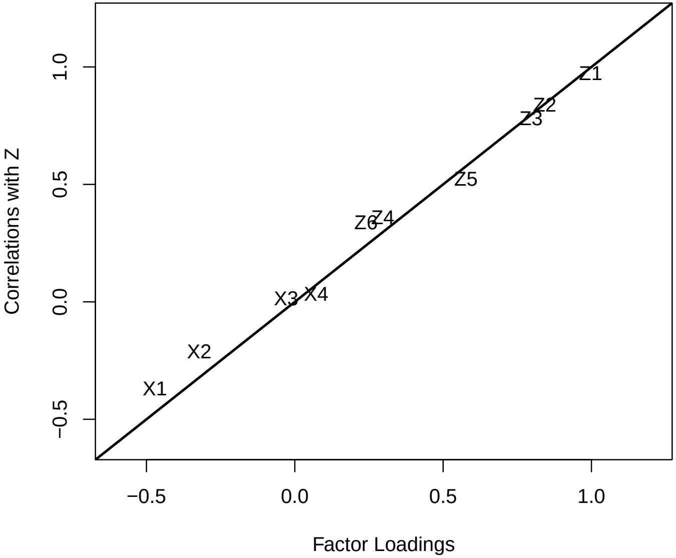
$$\Psi = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_K^2 \end{bmatrix}$$

- Choose the *number of factors* (dimensions)
- Consider *rotation*
- *Estimate* the factor loadings $\hat{\Lambda}$
- *Interpret* the factors...
- Generate *factor scores*

## Factor Analysis Simulation

```
> FASim1 <- factanal(df,factors=1,scores="regression",
+                     rotation="none")
> print(FASim1,cutoff=0)

Call:
factanal(x = df, factors = 1, scores = "regression", rotation = "none")

Uniquenesses:
   Z1    Z2    Z3    Z4    Z5    Z6    X1    X2    X3    X4 
0.005 0.290 0.365 0.912 0.667 0.942 0.778 0.896 0.999 0.995 

Loadings:
   Factor1
Z1  0.998 
Z2  0.843 
Z3  0.797 
Z4  0.297 
Z5  0.577 
Z6  0.240 
X1 -0.471 
X2 -0.322 
X3 -0.029 
X4  0.072 

               Factor1
SS loadings      3.150
Proportion Var   0.315

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 57.3 on 35 degrees of freedom.
The p-value is 0.0101 
```

# Factor Loadings vs. Correlations with $Z$



# Factor Analysis Simulation: Two Factors

```
> FASim2 <- factanal(df,factors=2,scores="regression",
+                    rotation="none")
> print(FASim2,cutoff=0)

Call:
factanal(x = df, factors = 2, scores = "regression", rotation = "none")

Uniquenesses:
   Z1    Z2    Z3    Z4    Z5    Z6    X1    X2    X3    X4
0.005 0.276 0.226 0.810 0.608 0.938 0.271 0.525 0.444 0.667

Loadings:
   Factor1 Factor2
Z1  0.997   0.013
Z2  0.841   0.129
Z3  0.801  -0.364
Z4  0.295   0.320
Z5  0.579  -0.238
Z6  0.242  -0.060
X1 -0.478   0.707
X2 -0.327   0.607
X3 -0.034   0.745
X4  0.068   0.573

               Factor1 Factor2
SS loadings      3.166   2.064
Proportion Var   0.317   0.206
Cumulative Var   0.317   0.523

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 31.42 on 26 degrees of freedom.
The p-value is 0.213
```

# Factor Analysis Simulation: Three Factors

```
> FASim3 <- factanal(df,factors=3,scores="regression",
+                    rotation="none")
> print(FASim3,cutoff=0)

Call:
factanal(x = df, factors = 3, scores = "regression", rotation = "none")

Uniquenesses:
   Z1    Z2    Z3    Z4    Z5    Z6    X1    X2    X3    X4
0.005 0.246 0.111 0.758 0.621 0.005 0.312 0.276 0.518 0.582

Loadings:
   Factor1 Factor2 Factor3
Z1  0.809   0.021   0.583
Z2  0.582   0.136   0.629
Z3  0.836  -0.379   0.216
Z4  0.334   0.359   0.036
Z5  0.475  -0.207   0.333
Z6  0.761   0.002  -0.645
X1 -0.382   0.674  -0.297
X2 -0.071   0.704  -0.473
X3 -0.024   0.693  -0.032
X4 -0.125   0.567   0.285

               Factor1 Factor2 Factor3
SS loadings      2.775   2.086   1.706
Proportion Var   0.277   0.209   0.171
Cumulative Var   0.277   0.486   0.657

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 15.13 on 18 degrees of freedom.
The p-value is 0.653
```

# Real Data: ANES 2016 Feeling Thermometers

```
> describe(Therms,range=FALSE)
                              vars    n  mean    sd  skew kurtosis   se
Asian-Americans                  1 2387 70.17 20.20 -0.38     0.02 0.41
Hispanics                        2 2387 69.35 20.91 -0.41     0.01 0.43
Blacks                           3 2387 69.00 21.19 -0.35    -0.24 0.43
Illegal Immigrants               4 2387 42.54 27.31  0.13    -0.71 0.56
Whites                           5 2387 71.63 19.40 -0.46     0.08 0.40
Dem. Pres. Candidate             6 2387 44.12 34.91  0.12    -1.42 0.71
GOP Pres. Candidate              7 2387 40.53 35.65  0.23    -1.43 0.73
Libertarian Pres. Candidate      8 2387 43.61 19.92 -0.58     0.25 0.41
Green Pres. Candidate            9 2387 43.20 20.87 -0.54     0.22 0.43
Dem. VP                         10 2387 48.24 25.91 -0.22    -0.44 0.53
GOP VP                          11 2387 49.59 33.42 -0.10    -1.21 0.68
John Roberts                    12 2387 53.75 18.39 -0.41     1.44 0.38
Pope Francis                    13 2387 69.55 25.17 -0.73     0.14 0.52
Christian Fundamentalists       14 2387 48.59 28.48 -0.07    -0.72 0.58
Feminists                       15 2387 56.94 26.65 -0.24    -0.47 0.55
Liberals                        16 2387 52.27 27.35 -0.24    -0.67 0.56
Labor Unions                    17 2387 56.70 24.74 -0.27    -0.29 0.51
Poor People                     18 2387 72.20 19.63 -0.36    -0.06 0.40
Big Business                    19 2387 49.34 22.52 -0.15    -0.18 0.46
Conservatives                   20 2387 55.22 25.91 -0.24    -0.45 0.53
SCOTUS                          21 2387 59.34 19.38 -0.32     0.54 0.40
Gays & Lesbians                 22 2387 62.83 26.86 -0.46    -0.20 0.55
Congress                        23 2387 41.17 22.32  0.02    -0.34 0.46
Rich People                     24 2387 53.53 20.69 -0.13     0.52 0.42
Muslims                         25 2387 55.80 25.64 -0.29    -0.23 0.52
Christians                      26 2387 74.40 23.80 -0.87     0.35 0.49
Jews                            27 2387 72.20 21.19 -0.45    -0.14 0.43
Tea Party                       28 2387 42.97 27.08 -0.06    -0.70 0.55
Police                          29 2387 75.57 22.50 -1.15     1.13 0.46
Transgender People              30 2387 57.29 26.88 -0.28    -0.31 0.55
Scientists                      31 2387 77.74 19.23 -0.77     0.39 0.39
BLM                             32 2387 48.26 32.66 -0.06    -1.15 0.67
```

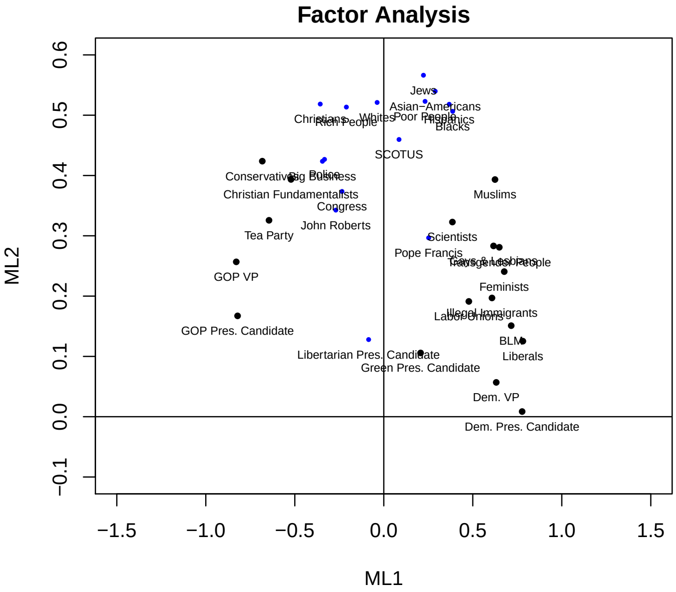
# Factor Analysis: One Factor

```
> FTFA1 <- fa(Therms,nfactors=1,fm="ml",rotate="none")
> print(FTFA1)
Factor Analysis using method =  ml
Call: fa(r = Therms, nfactors = 1, rotate = "none", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
                              ML1      h2   u2 com
Asian-Americans              0.29 0.08306 0.92   1
Hispanics                    0.37 0.13456 0.87   1
Blacks                       0.39 0.15227 0.85   1
Illegal Immigrants           0.61 0.37552 0.62   1
Whites                      -0.03 0.00066 1.00   1
Dem. Pres. Candidate         0.79 0.62770 0.37   1
GOP Pres. Candidate         -0.81 0.65791 0.34   1
Libertarian Pres. Candidate -0.07 0.00476 1.00   1
Green Pres. Candidate        0.22 0.05026 0.95   1
Dem. VP                      0.65 0.42135 0.58   1
GOP VP                      -0.80 0.64779 0.35   1
John Roberts                -0.24 0.05942 0.94   1
Pope Francis                 0.27 0.07253 0.93   1
Christian Fundamentalists   -0.49 0.23650 0.76   1
Feminists                    0.69 0.47926 0.52   1
Liberals                     0.80 0.63513 0.36   1
Labor Unions                 0.49 0.24414 0.76   1
Poor People                  0.25 0.06198 0.94   1
Big Business                -0.31 0.09877 0.90   1
Conservatives               -0.65 0.42099 0.58   1
SCOTUS                       0.11 0.01287 0.99   1
Gays & Lesbians              0.62 0.38096 0.62   1
Congress                    -0.20 0.04024 0.96   1
Rich People                 -0.18 0.03379 0.97   1
Muslims                      0.63 0.39894 0.60   1
Christians                  -0.32 0.10381 0.90   1
Jews                         0.23 0.05481 0.95   1
Tea Party                   -0.62 0.38321 0.62   1
Police                      -0.31 0.09796 0.90   1
Transgender People           0.65 0.42375 0.58   1
```

```
(...continued)

                ML1
SS loadings    8.09
Proportion Var 0.25

Mean item complexity =  1
Test of the hypothesis that 1 factor is sufficient.

The degrees of freedom for the null model are 496 and the objective function was 16.99 with
    Chi Square of 40352
The degrees of freedom for the model are 464 and the objective function was 8.87

The root mean square of the residuals (RMSR) is  0.15
The df corrected root mean square of the residuals is  0.16

The harmonic number of observations is  2387 with the empirical chi square  53448  with prob <  0
The total number of observations was  2387  with MLE Chi Square =  21052  with prob <  0

Tucker Lewis Index of factoring reliability =  0.448
RMSEA index =  0.137  and the 90 % confidence intervals are  0.135 0.138
BIC =  17443
Fit based upon off diagonal values = 0.74
Measures of factor score adequacy
                                                ML1
Correlation of scores with factors             0.97
Multiple R square of scores with factors       0.94
Minimum correlation of possible factor scores  0.88
```

```
> FTFA2 <- fa(Therms,nfactors=2,fm="ml", rotate="none")
> print(FTFA2)
Factor Analysis using method =  ml
Call: fa(r = Therms, nfactors = 2, rotate = "none", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
                              ML1  ML2    h2   u2 com
Asian-Americans              0.29 0.54 0.375 0.63 1.5
Hispanics                    0.37 0.52 0.404 0.60 1.8
Blacks                       0.39 0.51 0.406 0.59 1.9
Illegal Immigrants           0.61 0.20 0.408 0.59 1.2
Whites                      -0.04 0.52 0.273 0.73 1.0
Dem. Pres. Candidate         0.78 0.01 0.604 0.40 1.0
GOP Pres. Candidate         -0.82 0.17 0.703 0.30 1.1
Libertarian Pres. Candidate -0.09 0.13 0.024 0.98 1.7
Green Pres. Candidate        0.21 0.11 0.054 0.95 1.5
Dem. VP                      0.63 0.06 0.402 0.60 1.0
GOP VP                      -0.83 0.26 0.753 0.25 1.2
.
.
.
Police                      -0.33 0.43 0.293 0.71 1.9
Transgender People           0.65 0.28 0.500 0.50 1.4
Scientists                   0.39 0.32 0.253 0.75 1.9
BLM                          0.72 0.15 0.535 0.46 1.1

                       ML1  ML2
SS loadings           8.16 4.29
Proportion Var        0.26 0.13
Cumulative Var        0.26 0.39
Proportion Explained  0.66 0.34
Cumulative Proportion 0.66 1.00

Mean item complexity =  1.5
.
.
```

# Factor Analysis: Two Factors



PCA / FA are *data reduction* techniques...

- Rotation is exactly that: Rotation of the axes in the transformed space to make the results more interpretable.
- Two broad types:
  - *Orthogonal* rotation (maintains orthogonality of the axes)
  - *Oblique* rotation (allows components / factors to be correlated)
- **The goal of rotation is to improve the interpretability of the PCA/FA results. ("simple structure")**

# Rotation Methods

Orthogonal rotations:

- **Varimax** (minimizes the number of variables that have high loadings on each factor.)
- **Quartimax** (minimizes the number of factors needed to explain each variable)
- **Equamax** (a combination of varimax and quartimax)
- Others...

Oblique rotations (less easily interpretable):

- **Direct Oblimin** (the de facto standard for oblique rotation)
- **Promax** (simpler / faster than oblimin)
- Others...

# Rotation: Considerations

"Simple structure": "A condition in which variables load at near 1 (in absolute value) or at near 0 on an eigenvector (factor). Variables that load near 1 are clearly important in the interpretation of the factor, and variables that load near 0 are clearly unimportant." (Bryant and Yarnold 1995)

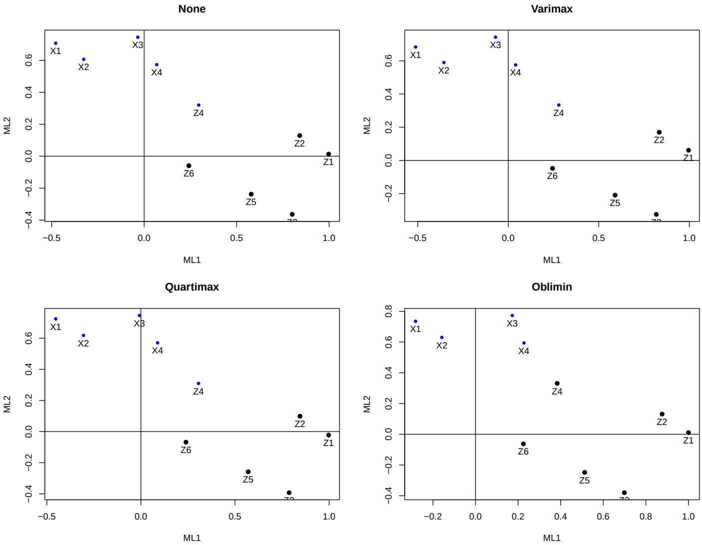
Factor Loading $\ell$ Guidelines:

- $0.10 < \ell < -0.10$ are unimportant
- $|\ell| > 0.30$ are important with $N \geq 100$
- Variables with $\ell > 0.30$ on more than one factor are *complex*

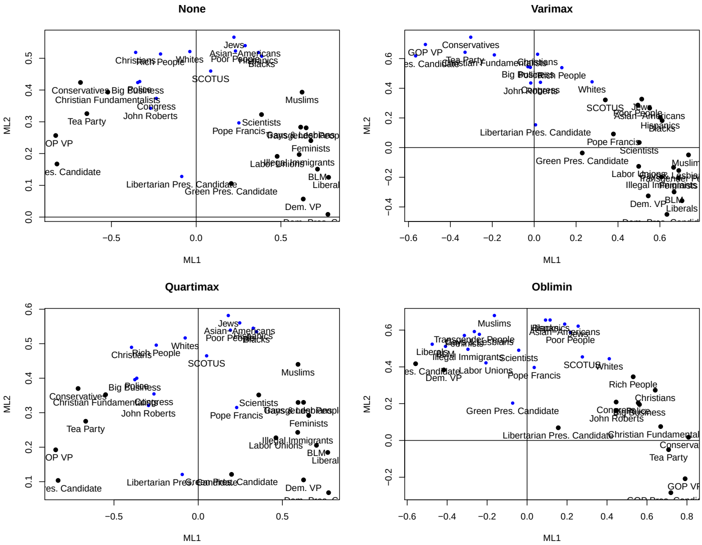
Thurstone's Criteria:

- Each variable should produce at least one zero loading on some factor.
- Each factor should have at least as many zero loadings as there are factors.
- Each pair of factors should have variables with significant loadings on one and zero loadings on the other.
- Each pair of factors should have a large proportion of zero loadings on both factors (if there are say four or more factors total).
- Each pair of factors should have only a few complex variables.

# Rotation: Simulated Data



# Rotation: Feeling Thermometers



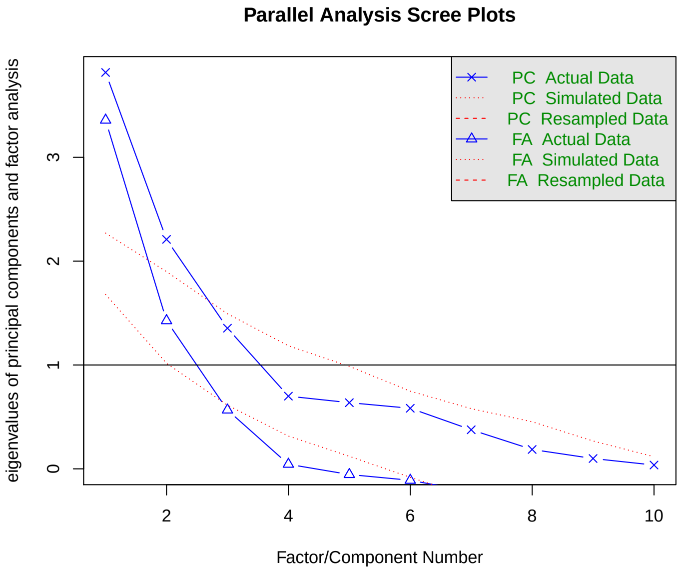
PCA/FA are *data reduction* techniques...

- The sum of the eigenvalues equals $K$; so...
- A factor / component with en eigenvalue less than 1.0 isn't even "explaining itself"
- "Kaiser criterion"

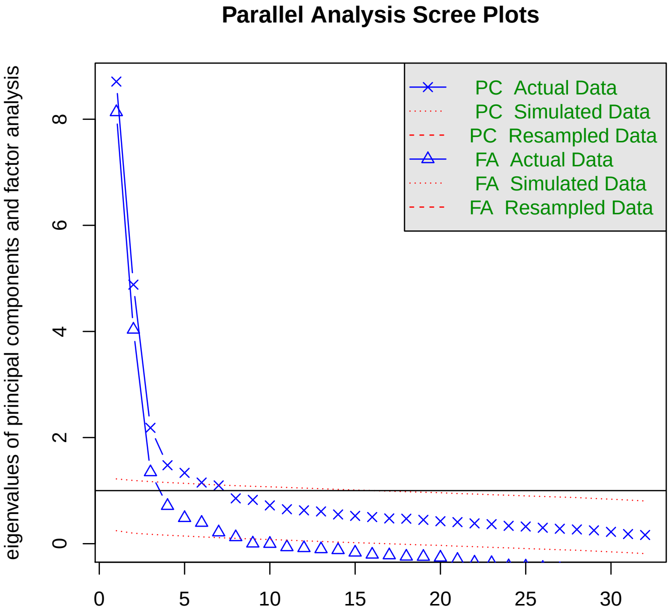
Other approaches:

- *Theory*...
- "Scree plot" (look for the "elbow")
- Target variance explained
- Others...

# “Parallel” Scree Plot (Simulated Data)



# “Parallel” Scree Plot (Feeling Thermometer Data)



# Topic: Non-Continuous Items

- PCA / FA are *linear* models; FA in particular makes OLS-like assumptions.
- These assumptions are often difficult to justify when items are non-continuous / nominal / ordinal
- One solution: PCA/FA on *polychoric* / *tetrachoric* matrices of item associations
    - *Polychoric* correlations are based on binary/ordinal realizations of underlying bivariate normal latent variables
    - *Tetrachoric* correlation for binary items:

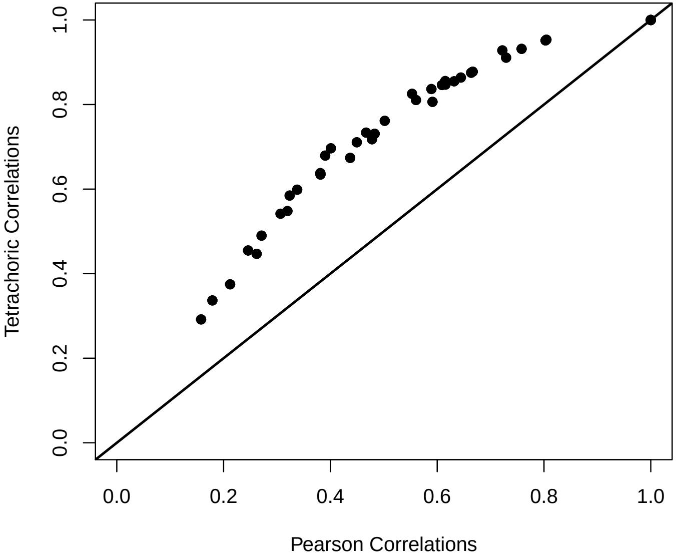
$$r_{tet} \approx cos(180/(1 + \sqrt{(BC/AD)})).$$

- PCA / FA can be applied to polychoric/tetrachoric correlation matrices.

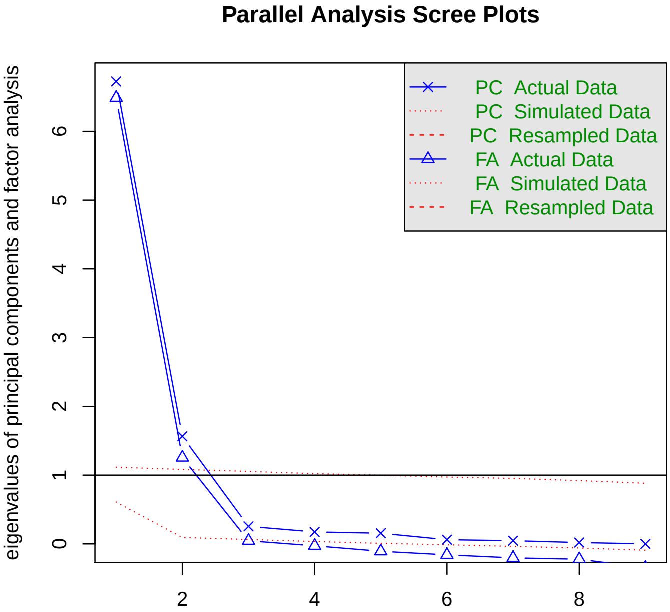
# Example: SCOTUS Voting Data

```
> url <- getURL("https://raw.githubusercontent.com/PrisonRodeo/MM-git/master/Data/SCOTUS-IRT.csv")
> SCOTUS <- read.csv(text = url)
> rm(url)
>
> SCOTUS <- na.omit(SCOTUS)
> head(SCOTUS)
  id Rehnquist Stevens OConnor Scalia Kennedy Souter Thomas Ginsburg Breyer
1  1         0       1       0      0       0      0      0        0      0
2  2         0       0       0      0       0      0      0        0      0
3  3         0       0       0      0       0      0      0        0      0
4  4         1       1       1      1       1      1      1        1      1
5  5         0       1       0      0       1      1      0        1      1
7  7         1       1       1      0       1      1      0        1      1
```

# Pearson and Tetrachoric Correlations: SCOTUS Data



# "Parallel" Scree Plot (SCOTUS Data)



# Tetrachoric FA Example: SCOTUS Data

```
> SCOTUSFA <- fa(SCOTUS[,2:10],nfactors=2,rotate="varimax",fm="ml",cor="tet")

> SCOTUSFA
Factor Analysis using method =  ml
Call: fa(r = SCOTUS[, 2:10], nfactors = 2, rotate = "varimax", fm = "ml",
    cor = "tet")
Standardized loadings (pattern matrix) based upon correlation matrix
           ML2  ML1   h2    u2 com
Rehnquist 0.43 0.84 0.88 0.117 1.5
Stevens   0.88 0.16 0.79 0.206 1.1
OConnor   0.64 0.64 0.82 0.182 2.0
Scalia    0.22 0.94 0.93 0.066 1.1
Kennedy   0.48 0.79 0.86 0.145 1.7
Souter    0.86 0.44 0.94 0.060 1.5
Thomas    0.17 0.97 0.97 0.032 1.1
Ginsburg  0.91 0.35 0.95 0.054 1.3
Breyer    0.94 0.25 0.94 0.056 1.1

                       ML2  ML1
SS loadings           4.12 3.96
Proportion Var        0.46 0.44
Cumulative Var        0.46 0.90
Proportion Explained  0.51 0.49
Cumulative Proportion 0.51 1.00

Mean item complexity =  1.4
.
.
.
```

# Useful References


- Gorsuch, Richard L. 1983. *Factor Analysis*, 2nd Ed. NJ: Lawrence Erlbaum.

- Cudek, Robert and Robert C. MacCallum, Eds. 2007. *Factor Analysis at 100*. NJ: Lawrence Erlbaum.

- Mulaik, Stanley A. 2010. *Foundations of Factor Analysis*, 2nd Ed. Boca Raton, FL: CRC Press.

- Fabrigar, Leandre R., and Duane T. Wegener. 2014. *Exploratory Factor Analysis*. New York: Oxford University Press.

# Useful R Packages and Routines

PCA and Biplots

- `stats::prcomp` (principal components via SVD)
- `biplot` (biplots)
- `psych::principal` (User-friendly PCA routine)
- Others...

Factor Analysis

- `nFactors` (Routines for assessing dimensionality / number of factors)
- `FactoMineR` (Hugely expanded FA package...)
- `GPARotation` (Many, many rotation options)