Principal Components Analysis

Variable redundancy and reduction

- Variable redundancy: some variables are correlated with one another, possibly because they measure the same "construct"
  - Poverty, education, income and unemployment
    - Should be possible to combine these variables into a smaller number that will account for most of the variance in the observed data
- Variable reduction: reducing a large set of variables into a much smaller set
  - Naturally leads to loss of some information, but we try to minimize this!

Principle Component Analysis

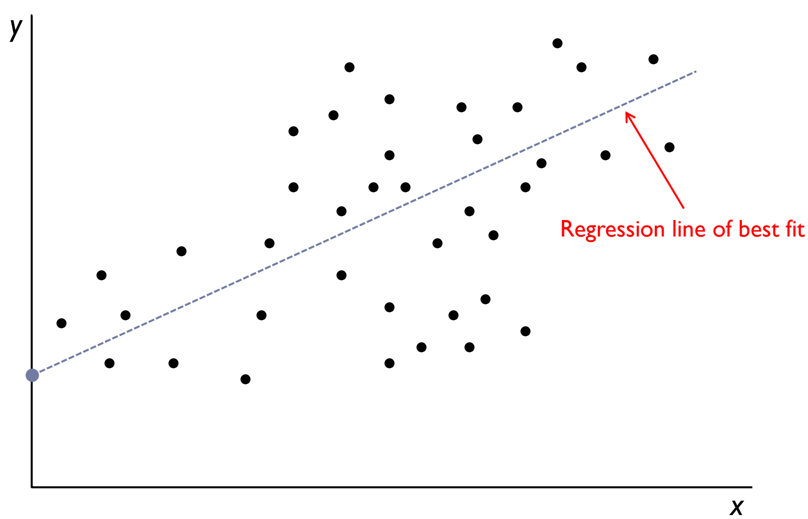
- A statistical technique used to examine the interrelations among a set of variables in order to identify the underlying structure of those variables
- Combine (reduce) a set of observed variables into a smaller set of "artificial" variables called principal components
  - The resulting PCs can be used in subsequent analyses
    - Regression

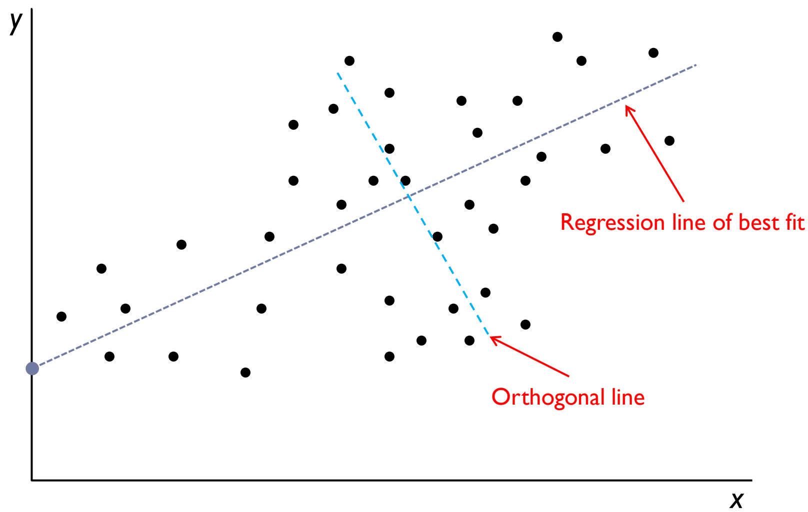
The assumptions of PCA

- Linearity
  - Assumes the data set to be linear combinations of the variables
- The importance of mean and covariance
  - There is no guarantee that the directions of maximum variance will contain good features for discrimination
- That large variances have important dynamics
  - Assumes that components with larger variance correspond to interesting dynamics and lower ones correspond to noise

PCA

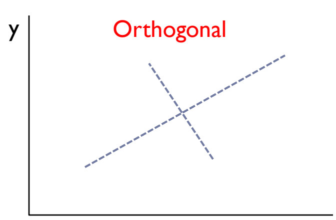
- Where regression determines a line of best fit to a data set, PCA determines several orthogonal lines of best fit
- Orthogonal: meaning "at right angles"
  - Actually the lines are perpendicular to each other in *n*-dimensional space
    - *n*-dimensional space is the variable sample space
    - There are as many dimensions as there are variables, so in a data set with 4 variables the sample space is 4-dimensional

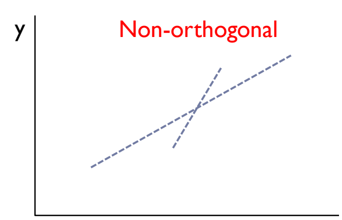


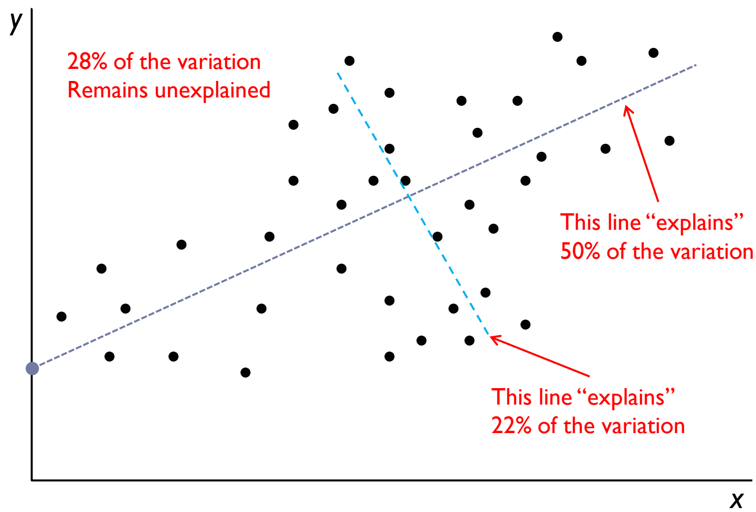


Components

- A linear combination of weighted variables:
  - The greatest variance of the data set is captured by the first axis (called the first principal component)
  - The second greatest variance on the second axis (the second principal component)
    - Note that components are uncorrelated since in the sample space they are orthogonal to each other







Components

- The general form for the formula to compute scores on a components created using PCA is:

$$c_1 = \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1p}x_p$$

- Where:
  - c_l = the subject's score on principal component 1 (the first component extracted)
  - β_{lp} = the regression coefficient (or weight) for observed variable p, as used in creating principal component 1
  - x_p = the subject's score on observed variable p

- You will have as many c's (components) as variables in the dataset

Variable "loading"

- An observed variable "loads" on a factor if it is highly correlated with the factor (has a large eigenvalue)

- How much weight is given to a variable when constructing a principle component

$$c_1 = .44x_1 + .40x_2 + .47x_3 + .32x_4 + .02x_5 + .01x_6 + .03x_7$$

  - x_1 has a loading of .44 (large) while x_2 has a loading of .02 (small)
  - So, x_1 determines more of the variance explained by PC1

Eigenequations and eigenvalues

- The regression weights (loadings) are determined using a type of equation called an eigenequation
  - These weights are optimal because no other set of weights could produce a set of components that are more successful in explaining the variation in the observed variables
    - Sort of like maximum likelihood estimation (MLE)
    - Sometimes called eigenvector

- The eigenvalue is a numeric estimation of how much of the variation each component explains

Steps in conducting a PCA

- Initial extraction of the components
- Determining the number of components to retain
  - Eigenvalue-one criterion
  - Scree test
  - Proportion of variance accounted for
  - Interpretability criteria
- Rotation to a final solution
- Interpreting the rotated solution
- Creating factor scores

PCA in R

- There are numerous ways of conducting PCA in R
  - `prcomp()` and `princomp()` are the most common
- We will focus on the `principal()` function in the `psych` package because it has the best options

```
> install.packages("psych")
> library(psych)
```

Example: Swiss fertility

- Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland

- 47 observations on 6 variables
  - Fertility - 'common standardized fertility measure'
  - Agriculture - % of males involved in agriculture as occupation
  - Examination - % draftees receiving highest mark on army examination
  - Education - % education beyond primary school for draftees
  - Catholic - % 'Catholic' (as opposed to 'protestant')
  - Infant.Mortality - % live births that live less than 1 year

Example

- First, let's create a new dataset with only the variables we want to use in our PCA

```
> swiss2<-swiss[c(2:6)]

> names(swiss2)
[1] "Agriculture"   "Examination"   "Education"    "Catholic"
[5] "Infant.Mortality"
```

Initial extraction of the components

```
> swpca <- principal(swiss2, nfactors=5, rotate="none")

Principal Components Analysis
Call: principal(r = swiss2, nfactors = 5, rotate = "none")
                 item   PC1   PC2   PC3   PC4   PC5 h2 u2
Agriculture         1 -0.85              0.45        1  0
Examination         2  0.93                          1  0
Education           3  0.80        0.49              1  0
Catholic            4 -0.63  0.38  0.66              1  0
Infant.Mortality    5        0.90 -0.38              1  0

                PC1  PC2  PC3  PC4  PC5
SS loadings    2.63 1.07 0.82 0.31 0.17
Proportion Var 0.53 0.21 0.16 0.06 0.03
Cumulative Var 0.53 0.74 0.90 0.97 1.00
```

Variable loadings

Eigenvalues (amount of variance accounted for by each PC)

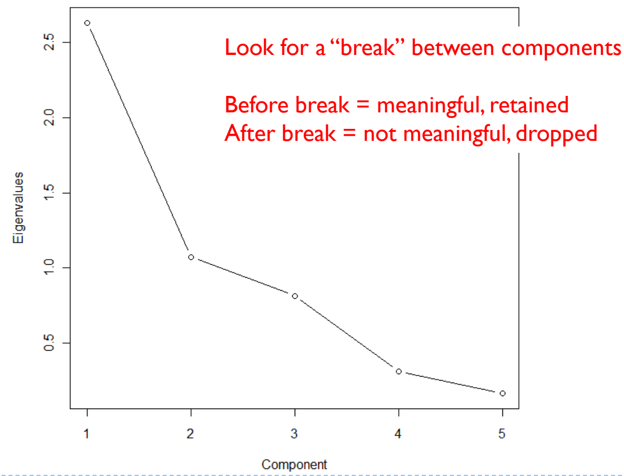
Determine number of components to retain Eigenvalue-one criteria

```
                PC1  PC2  PC3  PC4  PC5
SS loadings    2.63 1.07 0.82 0.31 0.17
Proportion Var 0.53 0.21 0.16 0.06 0.03
Cumulative Var 0.53 0.74 0.90 0.97 1.00
```

- We're lucky here, PC3 is 0.82 which is enough below 1 that we don't feel the need to include it
  - More challenging decision if PC3=0.95

Determine number of components to retain The scree test

```
> plot(swpca$values, type="b", ylab="Eigenvalues",
    xlab="Component", lab=c(5,5,5))
```



Determine number of components to retain Proportion of variance

	PC1	PC2	PC3	PC4	PC5
SS loadings	2.63	1.07	0.82	0.31	0.17
Proportion Var	0.53	0.21	0.16	0.06	0.03
Cumulative Var	0.53	0.74	0.90	0.97	1.00

- Retain components that account for at least x% of the total variance
  - 5% or 10%, etc.
- Retain components that *combined* account for x% of the cumulative variance
  - Usually at least 70%

Determine number of components to retain Interpretability

	item	PC1	PC2	PC3	PC4	PC5	h2	u2
Agriculture	1	-0.85			0.45		1	0
Examination	2	0.93					1	0
Education	3	0.80		0.49			1	0
Catholic	4	-0.63	0.38	0.66			1	0
Infant.Mortality	5		0.90	-0.38			1	0

- Do variables that load on a component share a conceptual meaning?
- Do variables that load on different components seem to measure a different construct?
- How many PC's would you choose?

Rotation to a Final Solution

- After initially deciding which PCs to retain, create a rotated factor pattern
- We do this for ease of interpretation

```
> swpca.r <- principal(swiss2, nfactors = 2, rotate = "varimax", scores = T)

Principal Components Analysis
Call: principal(r = swiss2, nfactors = 2, rotate = "varimax", scores = T)
                 item   RC1   RC2   h2   u2
Agriculture         1 -0.89       0.79 0.21
Examination         2  0.90       0.86 0.14
Education           3  0.82       0.68 0.32
Catholic            4 -0.51  0.52 0.54 0.46
Infant.Mortality    5        0.91 0.84 0.16

                RC1  RC2
SS loadings    2.54 1.16
Proportion Var 0.51 0.23
Cumulative Var 0.51 0.74

Test of the hypothesis that 2 factors are sufficient.
The number of observations was 47 with Chi Square = 33.2 with prob < 8.3e-09
```

Interpreting the rotated solution

- Determining just what is measures by each of the retained components

	item	RC1	RC2	h2	u2
Agriculture	1	-0.89		0.79	0.21
Examination	2	0.90		0.86	0.14
Education	3	0.82		0.68	0.32
Catholic	4	-0.51	0.52	0.54	0.46
Infant.Mortality	5		0.91	0.84	0.16

h^2 is called the communality estimate

Measures the % of variance in an observed variable accounted for by the retained components

- The first component seems to measure socioeconomic status
- The second component seems to measure beliefs and experiences
  - May choose to remove catholic from interpretation because it loads highly on two different components

Creating factor scores

- Linear composite of the weighted observed variables
  - Determine weights
  - Multiply variable for each observation by these weights
  - Sum the products

```
> swpca.r <- principal(swiss2, nfactors=2, rotate="varimax",
  scores=T)
> sw.scores<-swpca.r$scores
> sw.scores
                    RC1         RC2
Courtelary     0.74892706  0.61472668
Delemont      -0.46078328  1.21119279
Franches-Mnt  -0.68659489  0.73075268
Moutier       -0.05433337  0.14329745
Neuveville     0.43894928 -0.07097574
Porrentruy    -0.03838465  2.53479768
```

Summarizing the results

	PC1	PC2	h2
Agriculture	-0.89		0.79
Examination	0.90		0.86
Education	0.82		0.68
Catholic	-0.51	0.52	0.54
Infant.Mortality		0.91	0.84

- Only the first 2 components displayed eigenvalues greater than 1, chose to retain these. Together, these two components accounted for 74% of the total variance.
- Variables and corresponding factor loading are presented in the table.
- Four items were found to load on PC1, which was labeled the "socioeconomic" component. Two items loaded on PC2, which was labeled the "beliefs and experiences" component.

Using the factor scores

```
> sw.scores<-data.frame(swpca.r$scores)
> sw.lm<-lm(swiss$Fertility~sw.scores$RC1 + sw.scores$RC2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     70.143      1.242  56.472  < 2e-16 ***
sw.scores$RC1   -7.255      1.256  -5.779 7.13e-07 ***
sw.scores$RC2    5.835      1.256   4.648 3.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.515 on 44 degrees of freedom
Multiple R-squared: 0.5555,     Adjusted R-squared: 0.5353
F-statistic:  27.5 on 2 and 44 DF,  p-value: 1.789e-08
```

Some terminology

- Latent construct or unobserved variable
  - A variable that cannot be measured directly
  - Capture the variable (infer it) indirectly using other variables that are observed
  - Factors are the underlying latent variables that are responsible for the covariation between observed variables
- Unique variance
  - Variance of each variable unique to that variable and not explained or associated with other variables

What's the difference between PCA and Factor Analysis?

- Fundamentally the same, both analyze correlation matrices
- Difference is mainly in how the variance is analysed:
  - PCA: all variance of observed variables is analysed
    - Shared, unique and error
  - FA: only shared variance is analysed
- And the interpretation:
  - PCA: components are empirically determined aggregates of the variables without presumed theory
    - Labels are used but they are just a short hand for the component
  - FA: factors are the underlying (*latent*) variables that CAUSE the covariation between observed variables
    - Labels for factors are attempts to name these causal latent variables

FA vs. PCA conceptually

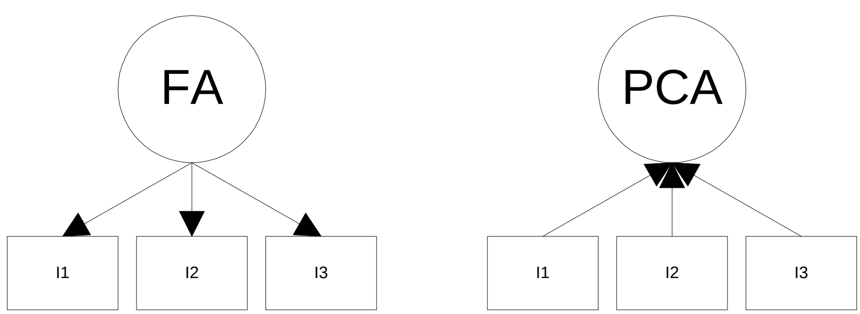
Factor Analysis

- Produces factors
- Factors cause variables

PCA

- Produces components
- Components are aggregates of the variables

Conceptual FA and PCA



FA vs. PCA conceptually

Factor Analysis

- Analyzes only the variance shared among the variables
  - common variance without error or unique variance
- "What are the underlying processes that could produce these correlations?"

PCA

- Analyzes all of the variance
- Just summarize empirical associations, very data driven

Example: Swiss data

- I believe that fertility in Switzerland is related to the type of job a person has and their religious beliefs surrounding family size
  - BUT, I don't have data specifically on these things
- Instead I have variables I measured as "proxies" for these concepts:
  - Agricultural employment, level of education, aptitude for military service, percent catholic and infant mortality
  - I think employment, education and military will group together to measure "Job Potential"
  - Catholic and IMR will group together to measure "Beliefs"

Factor Analysis in R

```
> sw.fa<-factanal(swiss2, factors=2, rotation="varimax")
> print(sw.fa, cutoff = .2, sort = TRUE)
Uniquenesses:
     Agriculture      Examination         Education         Catholic  Infant.Mortality
           0.408            0.190             0.202            0.005             0.969

Loadings:
                 Factor1 Factor2
Agriculture      -0.713   0.290
Examination       0.778  -0.453
Education         0.894
Catholic                  0.984
Infant.Mortality

               Factor1 Factor2
SS loadings      1.940   1.287
Proportion Var   0.388   0.257
Cumulative Var   0.388   0.645

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 2.98 on 1 degree of freedom.
The p-value is 0.0843
```

The p-value for the χ^2 test (0.08) indicates that the hypothesis of perfect fit cannot be rejected

What did we learn?

```
Uniqueness:
 Agriculture    Examination     Education      Catholic  Infant.Mortality
       0.408          0.190         0.202         0.005             0.969

                Factor1  Factor2
Agriculture      -0.713    0.290
Examination       0.778   -0.453
Education         0.894
Catholic                   0.984
Infant.Mortality
```

- There is too much unexplained (by other factors) variation in the Infant.Mortality measures to group it with other latent construct
- Agriculture, examination and education all appear to capture some underlying construct, perhaps on related to education and fertility (we'll call it Job Potential)
- Catholic appears to also capture some underlying latent structure, perhaps about beliefs regarding family size(so we'll call it Beliefs)

Which to use PCA vs. FA?

Factor Analysis

- Purpose is to identify the latent variables which are contributing to the common variance in a set of measured variables

PCA

- Purpose is to reduce the information in many variables into a set of weighted linear combinations of those variables