STAT 5020

Chapter 1&2: Introduction of Structural Equation Modelling

Department of Statistics

2022/2023 Term 2

Motivation:

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \qquad \varepsilon \sim N(0,\ \sigma^2)$$

n Observations

$$(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i),\ i = 1, 2, \cdots, n$$

ordinary least squares

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},\ X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{2n} & \cdots & x_{np} \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

$$Y = X\beta + \varepsilon, \qquad E(\varepsilon) = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2 I_n$$

Motivation:

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \qquad \varepsilon \sim N(0,\ \sigma^2)$$

Statistical assumptions

- ✓ **Normality**—For fixed values of the independent variables, the dependent variable is normally distributed.
- ✓ **Independence**—The y_i values are independent of each other.
- ✓ **Linearity**—The dependent variable is linearly related to the independent variables.
- ✓ **Homoscedasticity**—The variance of the dependent variable doesn't vary with the levels of the independent variables.

Motivation:

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \qquad \varepsilon \sim N(0,\ \sigma^2)$$

```
> state.x77
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073
Hawaii	868	4963	1.9	73.60	6.2	61.9	0	6425

```
> cor(states)
```

	Murder	Population	Illiteracy	Income	Frost
Murder	1.0000000	0.3436428	0.7029752	-0.2300776	-0.5388834
Population	0.3436428	1.0000000	0.1076224	0.2082276	-0.3321525
Illiteracy	0.7029752	0.1076224	1.0000000	-0.4370752	-0.6719470
Income	-0.2300776	0.2082276	-0.4370752	1.0000000	0.2262822
Frost	-0.5388834	-0.3321525	-0.6719470	0.2262822	1.0000000

Motivation:

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

```
> fit <- lm(Murder ~ Population + Illiteracy + Income + Frost,data=states)
> summary(fit)

Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
    data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7960 -1.6495 -0.0811  1.4815  7.6210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e+00  3.866e+00   0.319   0.7510
Population  2.237e-04  9.052e-05   2.471   0.0173 *
Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
Income      6.442e-05  6.837e-04   0.094   0.9253
Frost       5.813e-04  1.005e-02   0.058   0.9541
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.535 on 45 degrees of freedom
Multiple R-squared:  0.567,     Adjusted R-squared:  0.5285
F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

```
> PSr
     Alabama       Alaska      Arizona     Arkansas
   0.9790317    0.9852373    0.9818092    0.9808451
  California     Colorado  Connecticut     Delaware
   0.9889506    0.9934678    0.9892711    0.9912333
     Florida      Georgia       Hawaii        Idaho
   0.9871121    0.9793954    0.9807257    0.9938642
     Illinois      Indiana         Iowa       Kansas
   0.9902405    0.9931182    0.9951797    0.9938780
     Kentucky    Louisiana        Maine     Maryland
   0.9833668    0.9711480    0.9930349    0.9912289
Massachusetts     Michigan    Minnesota  Mississippi
   0.9902111    0.9910377    0.9947019    0.9761897
     Missouri      Montana     Nebraska       Nevada
   0.9923995    0.9942787    0.9944699    0.9948049
New Hampshire   New Jersey   New Mexico     New York
   0.9936008    0.9891653    0.9782266    0.9857946
North Carolina North Dakota         Ohio     Oklahoma
   0.9814884    0.9918751    0.9916563    0.9888762
       Oregon   Pennsylvania Rhode Island South Carolina
   0.9938496    0.9907103    0.9863271    0.9773651
   South Dakota    Tennessee        Texas         Utah
   0.9950658    0.9824860    0.9783674    0.9938433
      Vermont     Virginia   Washington West Virginia
   0.9940978    0.9865749    0.9938347    0.9864402
    Wisconsin      Wyoming
   0.9932858    0.9937658
```

Motivation:

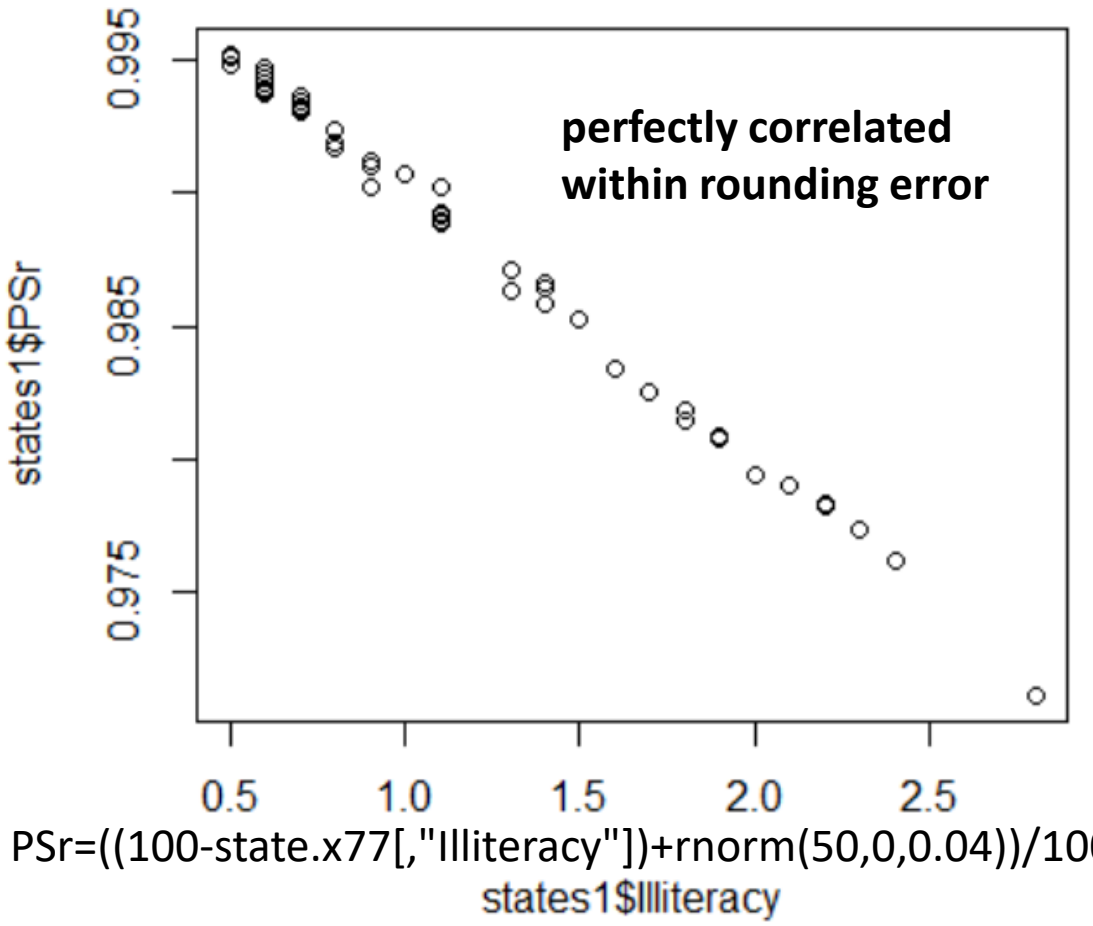
Multicollinearity

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

```
> fit1 <- lm(Murder ~ Population + Illiteracy + Income + Frost + PSr,data=states1)
> summary(fit1)

Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost +
    PSr, data = states1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8701 -1.5750 -0.3795  1.2215  7.0095

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.466e+03  9.000e+02   1.629   0.1104
Population   2.294e-04  8.897e-05   2.578   0.0134 *
Illiteracy  -1.059e+01  9.091e+00  -1.165   0.2504
Income       1.111e-04  6.721e-04   0.165   0.8695
Frost        3.022e-03  9.988e-03   0.303   0.7636
PSr         -1.465e+03  9.002e+02  -1.628   0.1107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.49 on 44 degrees of freedom
Multiple R-squared:  0.5915,    Adjusted R-squared:  0.5451
F-statistic: 12.74 on 5 and 44 DF,  p-value: 1.118e-07
```

nonsignificant



PSr=((100-state.x77[,"Illiteracy"])+rnorm(50,0,0.04))/100

Motivation:

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

```
> fit1 <- lm(Murder ~ Population + Illiteracy + Income + Frost + PSr,data=states1)
> summary(fit1)

Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost +
    PSr, data = states1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8701 -1.5750 -0.3795  1.2215  7.0095

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.466e+03  9.000e+02   1.629   0.1104
Population   2.294e-04  8.897e-05   2.578   0.0134 *
Illiteracy  -1.059e+01  9.091e+00  -1.165   0.2504
Income       1.111e-04  6.721e-04   0.165   0.8695
Frost        3.022e-03  9.988e-03   0.303   0.7636
PSr         -1.465e+03  9.002e+02  -1.628   0.1107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.49 on 44 degrees of freedom
Multiple R-squared:  0.5915,    Adjusted R-squared:  0.5451
F-statistic: 12.74 on 5 and 44 DF,  p-value: 1.118e-07
```

nonsignificant

Multicollinearity

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$Y = \beta_0 + \beta'_1 X_1 + \beta'_2 X_2 + \varepsilon$$

$$X_2 = kX_1$$

$$Y = \beta_0 + \beta'_1 X_1 + \beta'_2 k X_1 + \varepsilon$$

$$\beta_1 = \beta'_1 + \beta'_2 k$$

7

Motivation:

Multicollinearity

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

Variance inflation factor (VIF)

Step one

First we run an ordinary least square regression that has X_i as a function of all the other explanatory variables in the first equation.

If $i = 1$, for example, equation would be

$$X_1 = \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_p X_p + \varepsilon$$

Step two

$$VIF_1 = \frac{1}{1 - R_1^2} = \frac{SS_{tot}}{SS_{res}} = \frac{\sum\left(x_{1j} - \overline{x_1}\right)^2}{\sum e_j^2}$$

A rule of thumb is that if $VIF_i > 10$ then multicollinearity is high (a cutoff of 5 is also commonly used).

8

Motivation:

Multicollinearity

Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

Variance inflation factor (VIF)

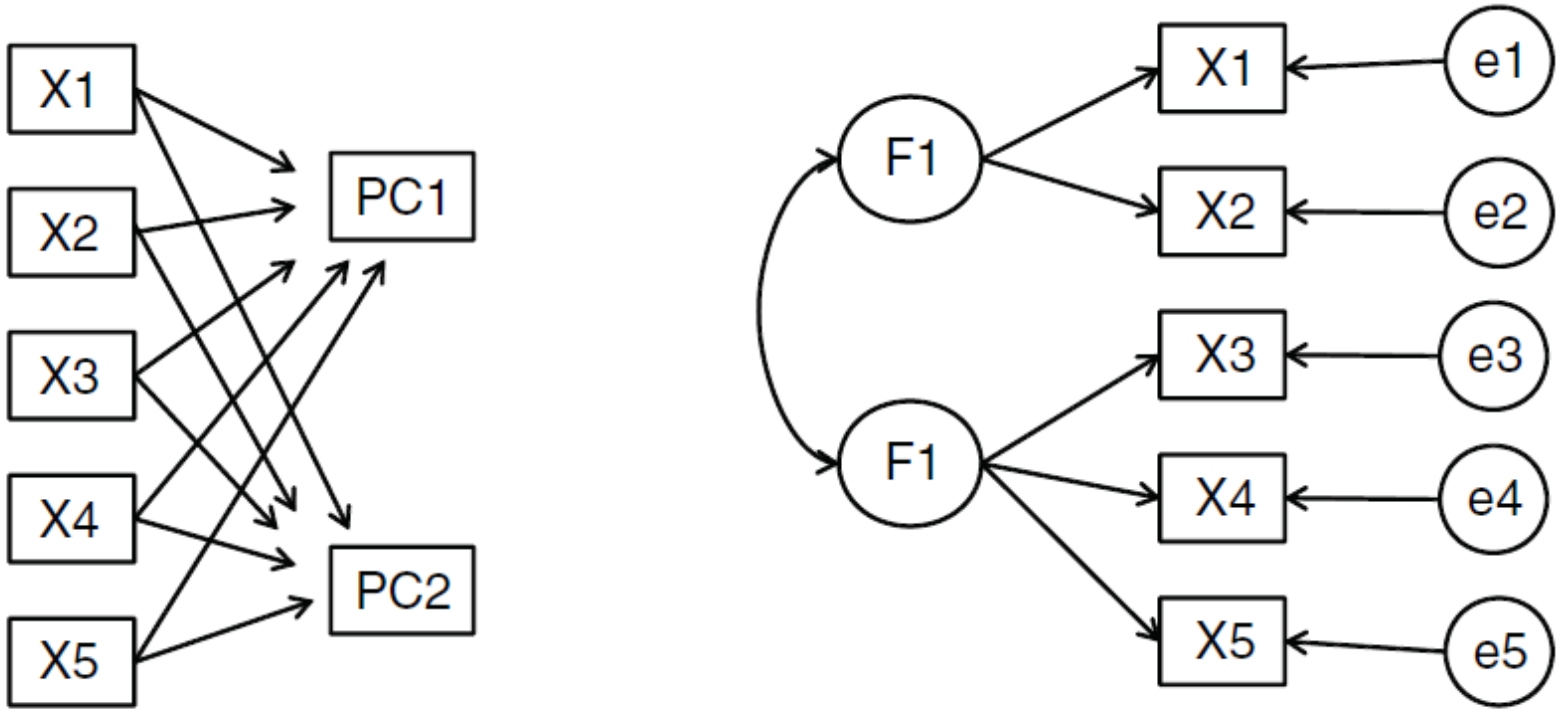
```
> vif(fit)
Population Illiteracy     Income      Frost
  1.245282   2.165848   1.345822   2.082547
> vif(fitl)
Population Illiteracy     Income      Frost        PSr
  1.247204 242.705554   1.348271   2.130582 246.684222
```

Principal Components Analysis (PCA)

Exploratory Factor Analysis (EFA)

- ✓ Delete
- ✓ Data-reduction technique
- ✓ Latent structure

Principal Components Analysis (PCA) VS Exploratory Factor Analysis (EFA)



(a) Principal Components Model

(b) Factor Analysis Model

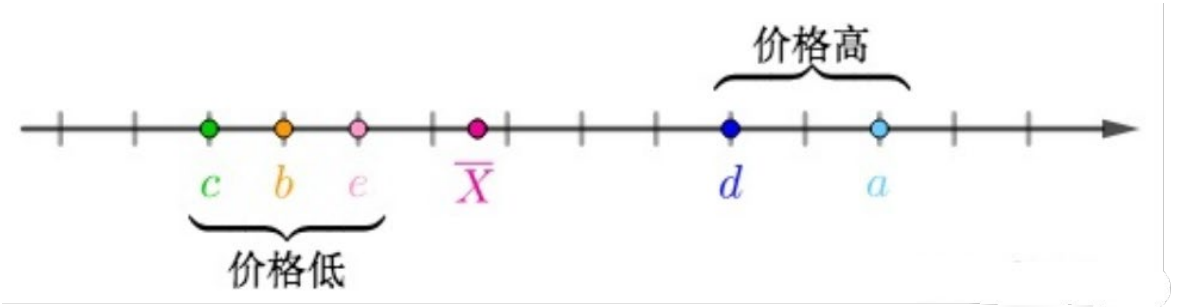
Section 1: Principal Component Analysis and Exploratory Factor Analysis

Principal Components Analysis (PCA)

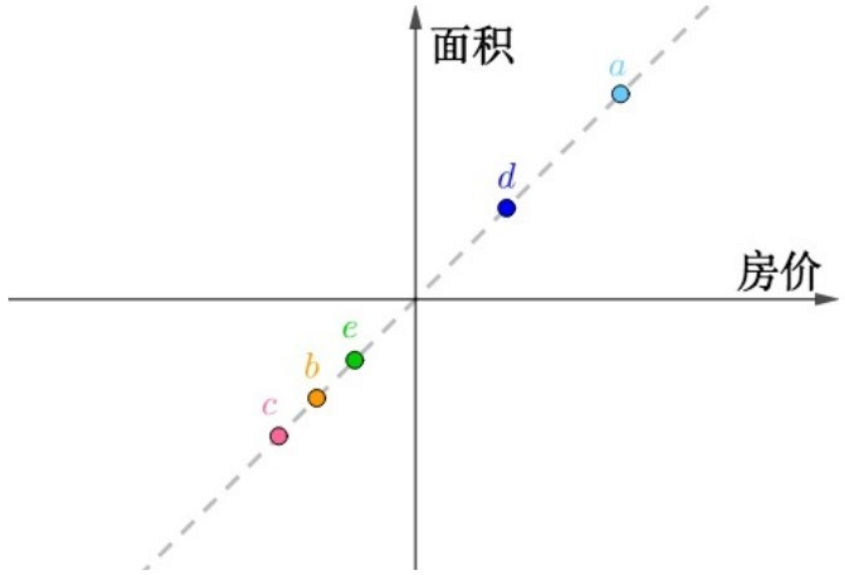
Data-reduction

	房价(百万元)
a	10
b	2
c	1
d	7
e	3

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} = \frac{10 + 2 + 1 + 7 + 3}{5} = 4.6$$



	房价(百万元)	面积(百平米)
a	10	10
b	2	2
c	1	1
d	7	7
e	3	3

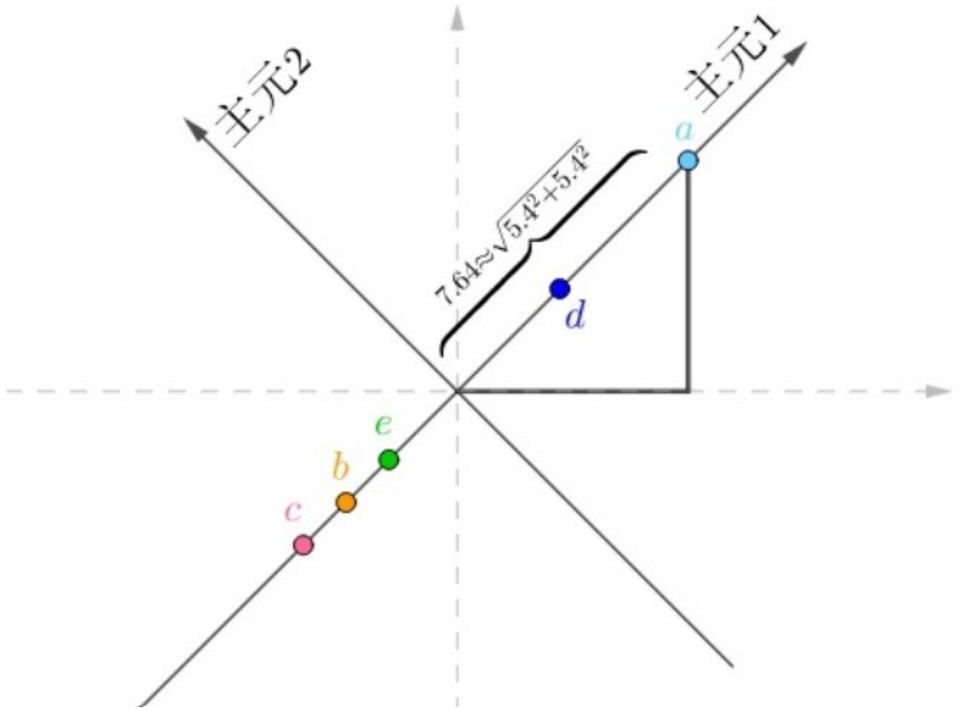


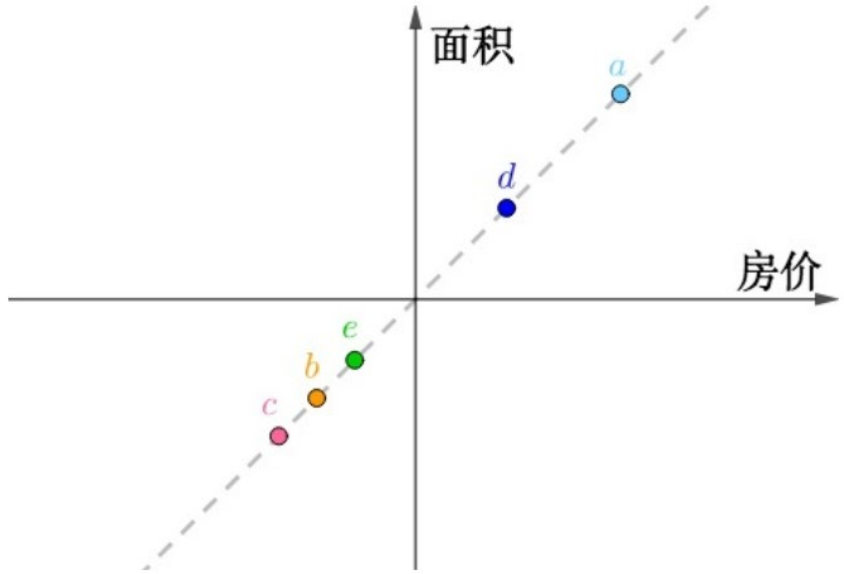
Principal Components Analysis (PCA)

Data-reduction

	主元1	主元2
a	7.64	0
b	−3.68	0
c	−5.09	0
d	3.39	0
e	−2.26	0

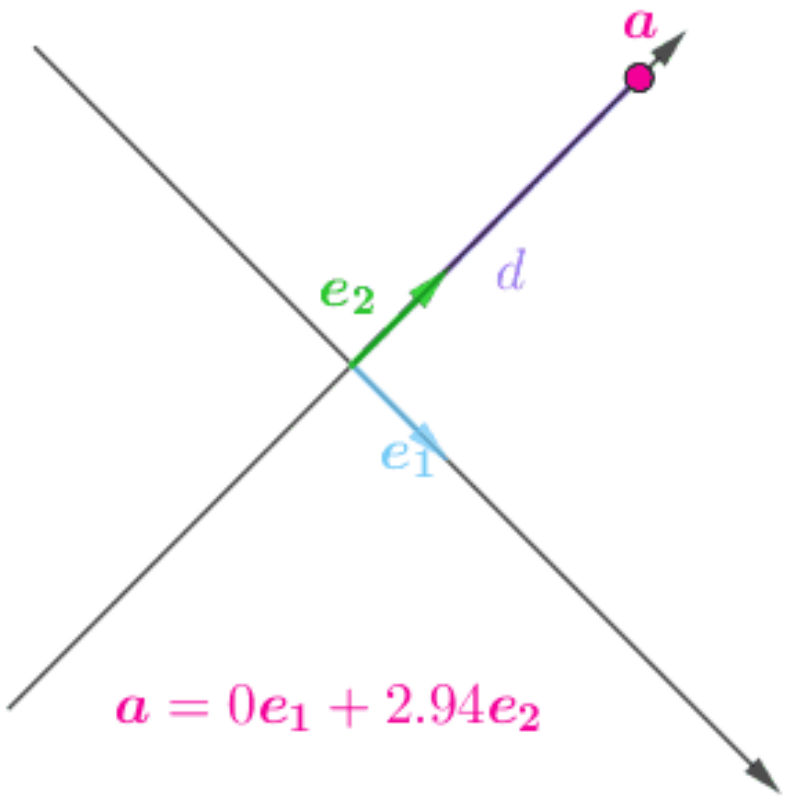
	房价(百万元)	面积(百平米)
a	10	10
b	2	2
c	1	1
d	7	7
e	3	3

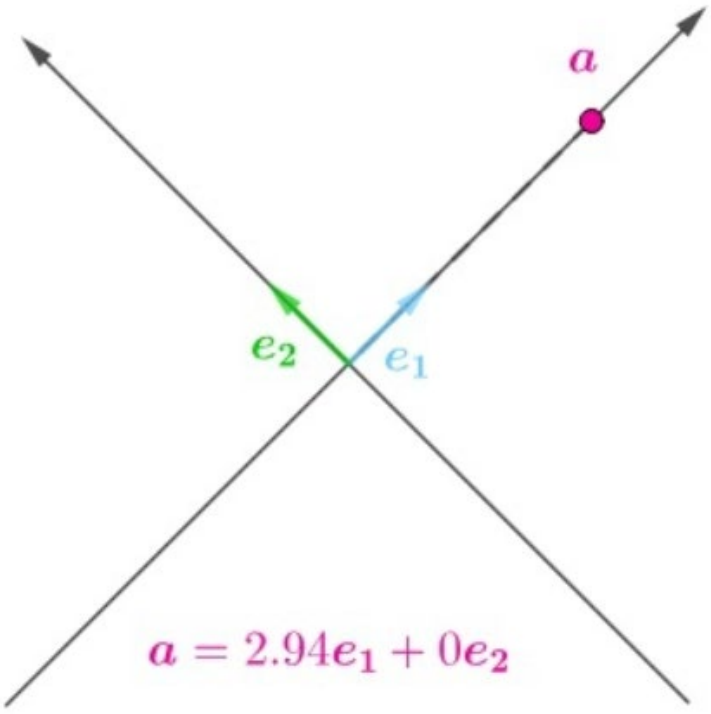




Principal Components Analysis (PCA)

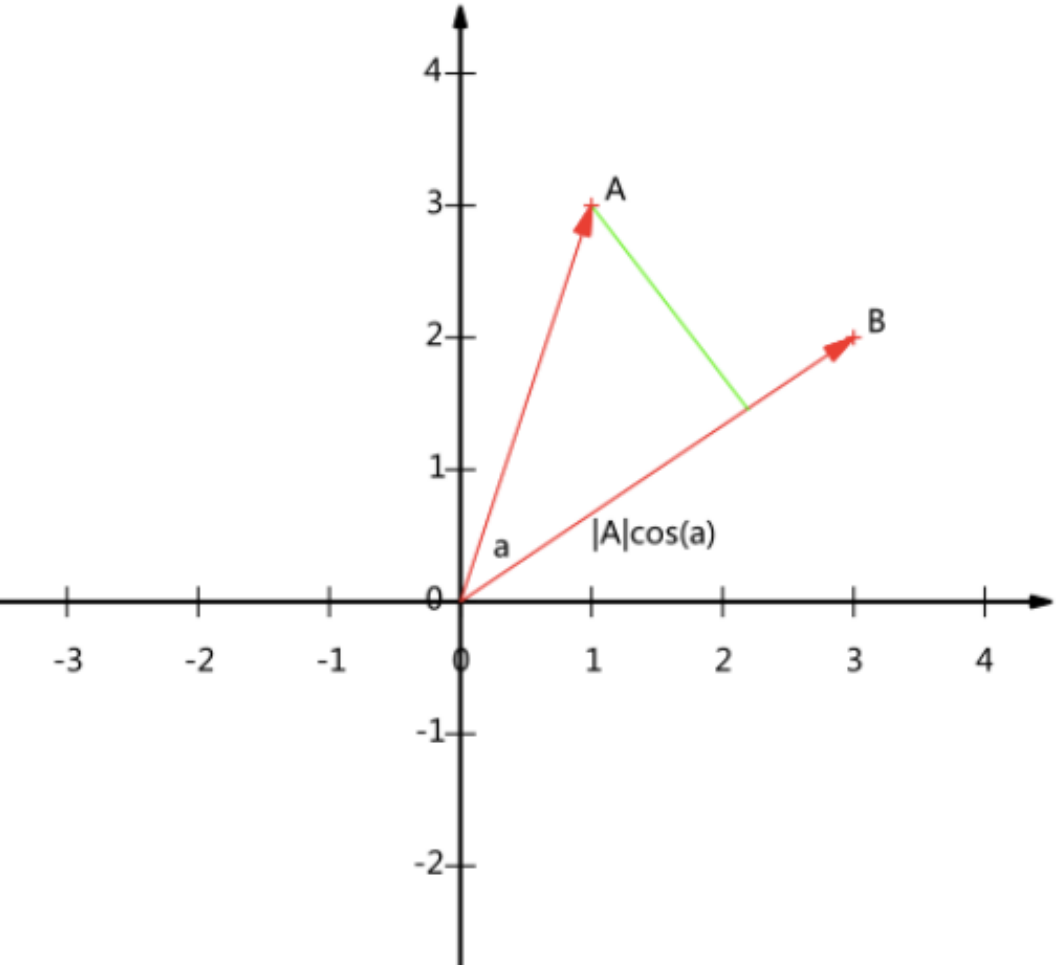
Data-reduction





Principal Components Analysis (PCA)

Change of Basis in Matrix Form



$$A = (x_1, y_1), B = (x_2, y_2)$$

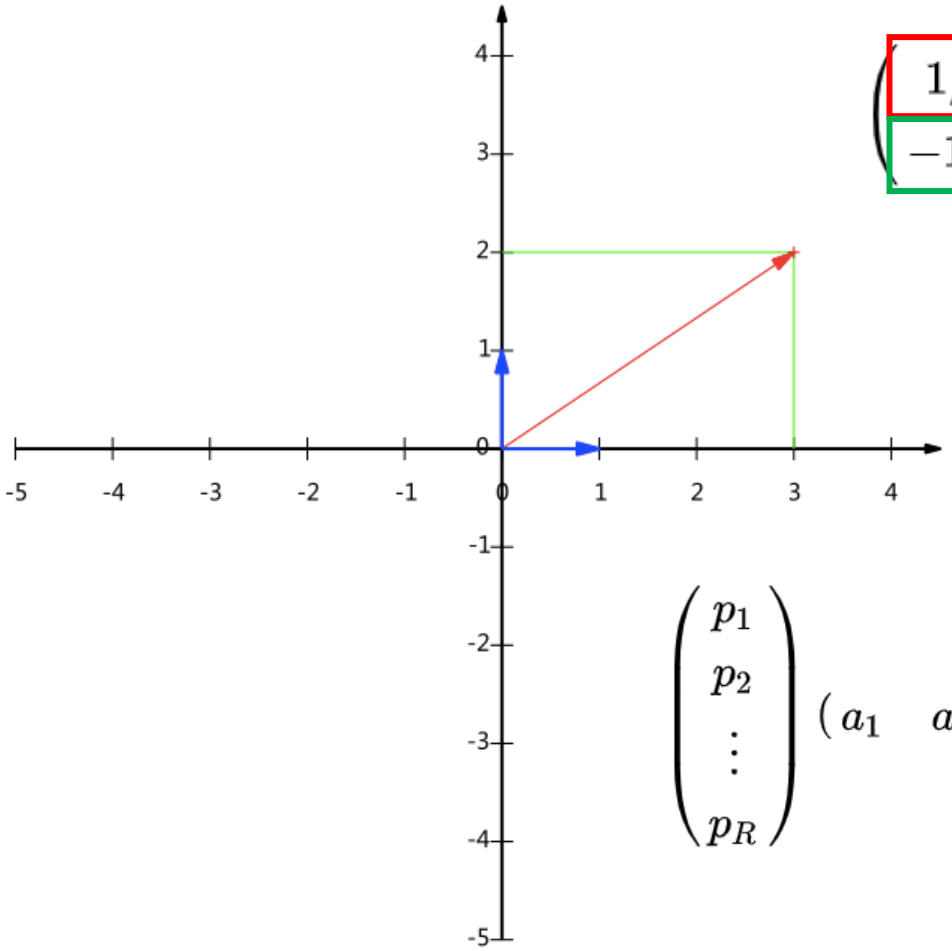
$$A \cdot B = x_1 y_1 + x_2 y_2$$

$$A \cdot B = |A||B|\cos(\alpha)$$

$$\text{If } |B| = 1, A \cdot B = |A|\cos(\alpha)$$

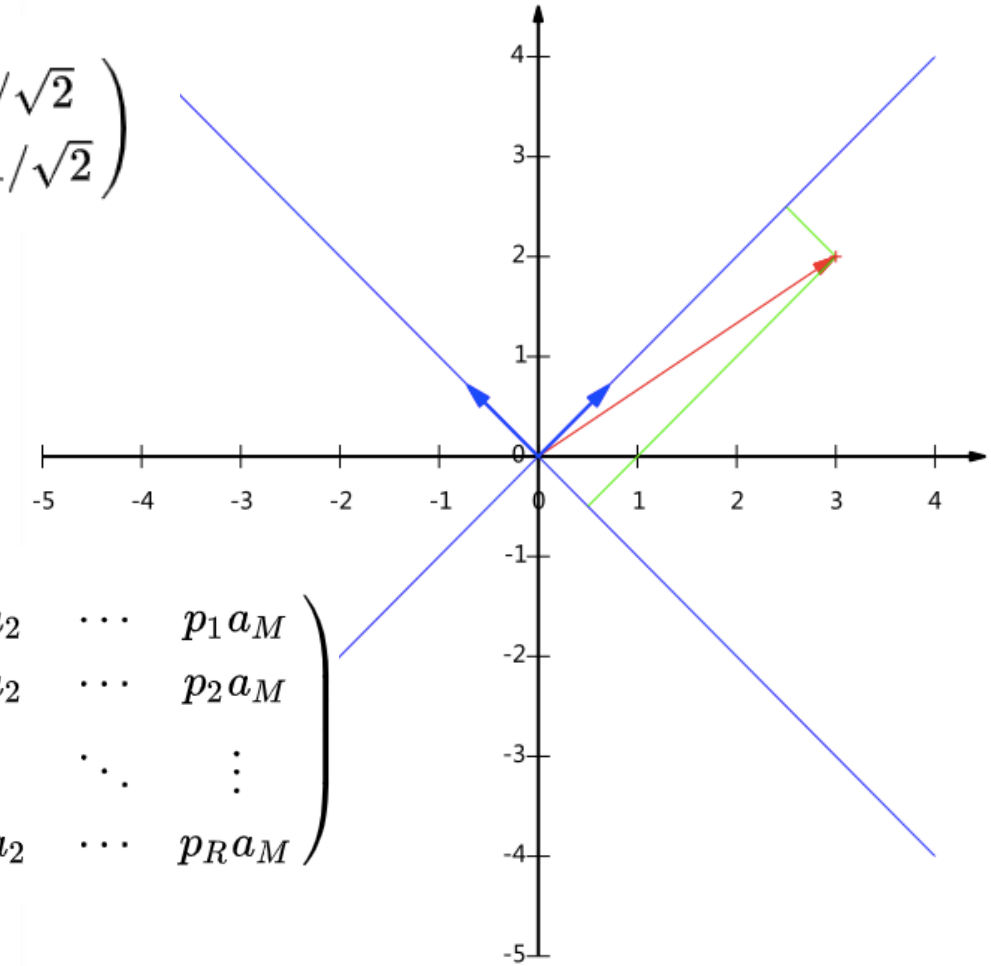
Principal Components Analysis (PCA)

Change of Basis in Matrix Form $Y = PX$



$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \cdots & a_M \end{pmatrix} = \begin{pmatrix} p_1 a_1 & p_1 a_2 & \cdots & p_1 a_M \\ p_2 a_1 & p_2 a_2 & \cdots & p_2 a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_R a_1 & p_R a_2 & \cdots & p_R a_M \end{pmatrix}$$



Principal Components Analysis (PCA)

Covariance Matrix

$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix}$$

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m} a_i^2 \qquad Cov(a,b) = \frac{1}{m}\sum_{i=1}^{m} a_i b_i$$

$$\frac{1}{m}XX^{\mathsf{T}} = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{pmatrix}$$

Change of Basis

$$Y = PX$$

Eigenvector

$$\begin{aligned} \frac{1}{m}YY^{\mathsf{T}} &= \frac{1}{m}(PX)(PX)^{\mathsf{T}} \\ &= \frac{1}{m}PXX^{\mathsf{T}}P^{\mathsf{T}} \\ &= P(\frac{1}{m}XX^{\mathsf{T}})P^{\mathsf{T}} \end{aligned}$$

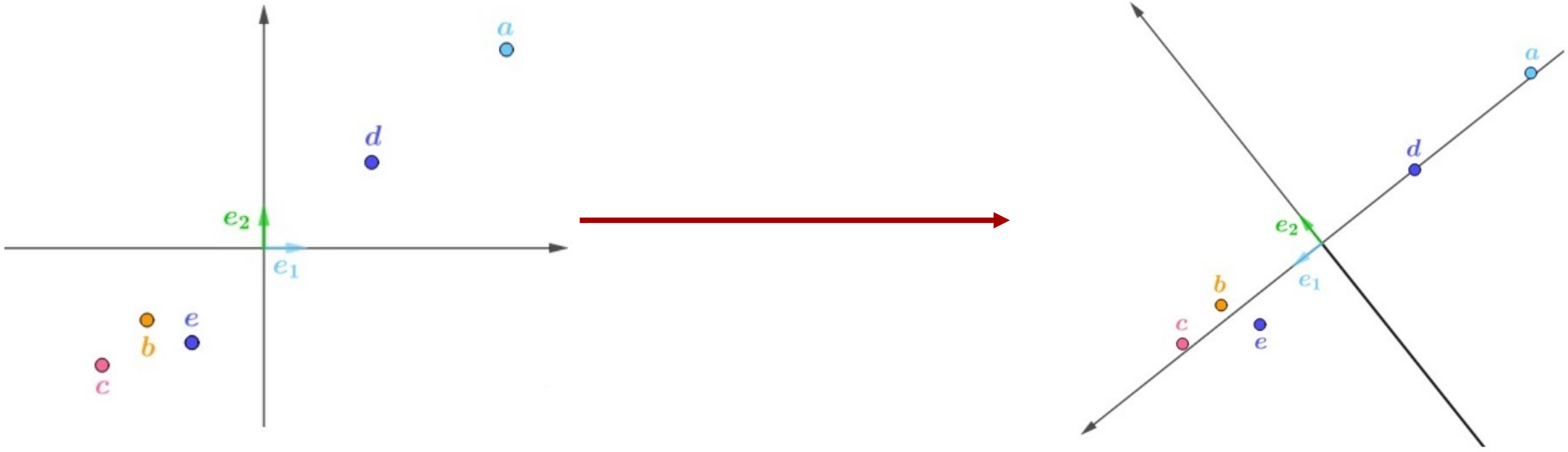
Diagonal

A square matrix is diagonal if and only if it is triangular and normal.

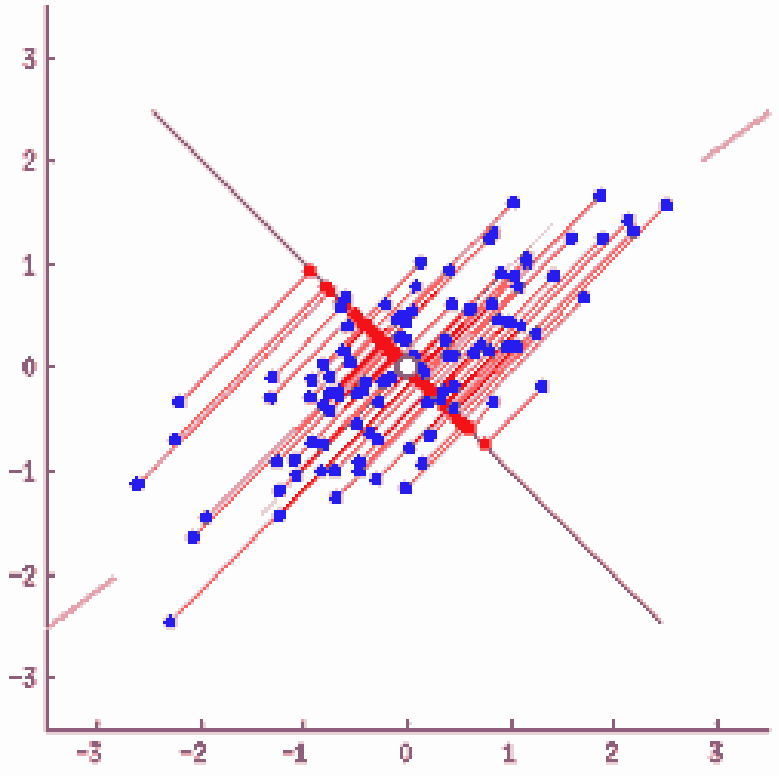
Principal Components Analysis (PCA)

Data-reduction

	房价(百万元)	面积(百平米)
a	5.4	4.4
b	−2.6	−1.6
c	−3.6	−2.6
d	2.4	1.9
e	−1.6	−2.1



Principal Components Analysis (PCA)



The first principal component weighted combination of the k observed variables that accounts for the most variance in the original set of variables

$$PC_1 = a_1X_1 + a_2X_2 + \ldots + a_kX_k$$

The second principal component is the linear combination that accounts for the most variance in the original variables, under the constraint that it's **orthogonal** (uncorrelated) to the first principal component

Theoretically, you can extract as many principal components as there are variables

Principal Components Analysis (PCA)

Variable	Description	Variable	Description
CONT	Number of contacts of lawyer with judge	PREP	Preparation for trial
INTG	Judicial integrity	FAMI	Familiarity with law
DMNR	Demeanor	ORAL	Sound oral rulings
DILG	Diligence	WRIT	Sound written rulings
CFMG	Case flow managing	PHYS	Physical ability
DECI	Prompt decisions	RTEN	Worthy of retention

```
> USJudgeRatings
```

	CONT	INTG	DMNR	DILG	CFMG	DECI	PREP	FAMI	ORAL	WRIT	PHYS	RTEN
AARONSON,L.H.	5.7	7.9	7.7	7.3	7.1	7.4	7.1	7.1	7.1	7.0	8.3	7.8
ALEXANDER,J.M.	6.8	8.9	8.8	8.5	7.8	8.1	8.0	8.0	7.8	7.9	8.5	8.7
ARMENTANO,A.J.	7.2	8.1	7.8	7.8	7.5	7.6	7.5	7.5	7.3	7.4	7.9	7.8
BERDON,R.I.	6.8	8.8	8.5	8.8	8.3	8.5	8.7	8.7	8.4	8.5	8.8	8.7
BRACKEN,J.J.	7.3	6.4	4.3	6.5	6.0	6.2	5.7	5.7	5.1	5.3	5.5	4.8
BURNS,E.B.	6.2	8.8	8.7	8.5	7.9	8.0	8.1	8.0	8.0	8.0	8.6	8.6
CALLAHAN,R.J.	10.6	9.0	8.9	8.7	8.5	8.5	8.5	8.5	8.6	8.4	9.1	9.0
COHEN,S.S.	7.0	5.9	4.9	5.1	5.4	5.9	4.8	5.1	4.7	4.9	6.8	5.0

Principal Components Analysis (PCA)

To avoid the Unit affect, we scale the data first

Or just use the Correlation Matrix.

```
> cov(scale(USJudgeRatings[,-1]))
          INTG      DMNR      DILG      CFMG      DECI      PREP
INTG 1.0000000 0.9646153 0.8715111 0.8140858 0.8028464 0.8777965
DMNR 0.9646153 1.0000000 0.8368510 0.8133582 0.8041168 0.8558175
DILG 0.8715111 0.8368510 1.0000000 0.9587988 0.9561661 0.9785684
CFMG 0.8140858 0.8133582 0.9587988 1.0000000 0.9811359 0.9579140
DECI 0.8028464 0.8041168 0.9561661 0.9811359 1.0000000 0.9570883
PREP 0.8777965 0.8558175 0.9785684 0.9579140 0.9570883 1.0000000
FAMI 0.8688580 0.8412415 0.9573634 0.9354684 0.9428045 0.9898634
ORAL 0.9113992 0.9067729 0.9544758 0.9505657 0.9482564 0.9831004
WRIT 0.9088347 0.8930611 0.9592503 0.9422470 0.9461009 0.9867992
PHYS 0.7419360 0.7886804 0.8129211 0.8794874 0.8717628 0.8486735
RTEN 0.9372632 0.9437002 0.9299652 0.9270827 0.9249924 0.9502926
          FAMI      ORAL      WRIT      PHYS      RTEN
INTG 0.8688580 0.9113992 0.9088347 0.7419360 0.9372632
DMNR 0.8412415 0.9067729 0.8930611 0.7886804 0.9437002
DILG 0.9573634 0.9544758 0.9592503 0.8129211 0.9299652
CFMG 0.9354684 0.9505657 0.9422470 0.8794874 0.9270827
DECI 0.9428045 0.9482564 0.9461009 0.8717628 0.9249924
PREP 0.9898634 0.9831004 0.9867992 0.8486735 0.9502926
FAMI 1.0000000 0.9813391 0.9906956 0.8437444 0.9416450
ORAL 0.9813391 1.0000000 0.9934294 0.8911639 0.9821323
WRIT 0.9906956 0.9934294 1.0000000 0.8559400 0.9675564
PHYS 0.8437444 0.8911639 0.8559400 1.0000000 0.9065478
RTEN 0.9416450 0.9821323 0.9675564 0.9065478 1.0000000
```

```
> cor(USJudgeRatings[,-1])
          INTG      DMNR      DILG      CFMG      DECI      PREP
INTG 1.0000000 0.9646153 0.8715111 0.8140858 0.8028464 0.8777965
DMNR 0.9646153 1.0000000 0.8368510 0.8133582 0.8041168 0.8558175
DILG 0.8715111 0.8368510 1.0000000 0.9587988 0.9561661 0.9785684
CFMG 0.8140858 0.8133582 0.9587988 1.0000000 0.9811359 0.9579140
DECI 0.8028464 0.8041168 0.9561661 0.9811359 1.0000000 0.9570883
PREP 0.8777965 0.8558175 0.9785684 0.9579140 0.9570883 1.0000000
FAMI 0.8688580 0.8412415 0.9573634 0.9354684 0.9428045 0.9898634
ORAL 0.9113992 0.9067729 0.9544758 0.9505657 0.9482564 0.9831004
WRIT 0.9088347 0.8930611 0.9592503 0.9422470 0.9461009 0.9867992
PHYS 0.7419360 0.7886804 0.8129211 0.8794874 0.8717628 0.8486735
RTEN 0.9372632 0.9437002 0.9299652 0.9270827 0.9249924 0.9502926
          FAMI      ORAL      WRIT      PHYS      RTEN
INTG 0.8688580 0.9113992 0.9088347 0.7419360 0.9372632
DMNR 0.8412415 0.9067729 0.8930611 0.7886804 0.9437002
DILG 0.9573634 0.9544758 0.9592503 0.8129211 0.9299652
CFMG 0.9354684 0.9505657 0.9422470 0.8794874 0.9270827
DECI 0.9428045 0.9482564 0.9461009 0.8717628 0.9249924
PREP 0.9898634 0.9831004 0.9867992 0.8486735 0.9502926
FAMI 1.0000000 0.9813391 0.9906956 0.8437444 0.9416450
ORAL 0.9813391 1.0000000 0.9934294 0.8911639 0.9821323
WRIT 0.9906956 0.9934294 1.0000000 0.8559400 0.9675564
PHYS 0.8437444 0.8911639 0.8559400 1.0000000 0.9065478
RTEN 0.9416450 0.9821323 0.9675564 0.9065478 1.0000000
```

Principal Components Analysis (PCA)

Eigenvalue

```
> eigen(cor(USJudgeRatings[,-1]))
eigen() decomposition
$values
 [1] 10.133417426  0.424628969  0.255279532  0.091303054  0.037300167  0.019761364
 [7]  0.018474499  0.008291420  0.006091924  0.003360869  0.002090776

$vectors
             [,1]          [,2]         [,3]        [,4]        [,5]         [,6]
 [1,] -0.2885122  0.5744682517  0.117763148  0.08380834  0.37493974 -0.50952871
 [2,] -0.2868395  0.5763568072 -0.176986952  0.23977262 -0.39860809  0.51407811
 [3,] -0.3043623 -0.1385605824  0.334740068  0.26555601  0.59149417  0.29806148
 [4,] -0.3026194 -0.3100115588  0.019545609  0.47773553 -0.08202695  0.10089374
 [5,] -0.3019234 -0.3364674872  0.054443551  0.38036525 -0.39888902 -0.44826185
 [6,] -0.3094144 -0.1252540296  0.229233996 -0.20132809  0.08469611  0.33583565
 [7,] -0.3066761 -0.1228593988  0.227525865 -0.52405105 -0.09943784 -0.03818923
 [8,] -0.3127088  0.0052082558 -0.005507203 -0.22936834 -0.14642044  0.01945629
 [9,] -0.3110520 -0.0002999784  0.148245297 -0.31656247 -0.23702291 -0.07288963
[10,] -0.2807447 -0.2347983520 -0.820161360 -0.15475146  0.29791670  0.03755338
[11,] -0.3097836  0.1527808928 -0.201053522  0.01114254  0.03729716 -0.23409315
              [,7]         [,8]         [,9]        [,10]          [,11]
 [1,] -0.229705308 -0.284903977  0.145484887 -0.10273495 -0.0006869163
 [2,]  0.167067325 -0.169286228 -0.005467441  0.10539158 -0.0764809505
 [3,]  0.367529033  0.004789352 -0.354685540  0.02389188 -0.0735829555
 [4,] -0.722336184  0.035844452 -0.026425045  0.20704699 -0.0131126895
 [5,]  0.452351620 -0.199576677  0.150276288 -0.13826398 -0.0422633237
 [6,] -0.006823921  0.068955312  0.717150217 -0.25188457  0.3049299442
 [7,] -0.002372688 -0.222092249  0.060538415  0.54400573 -0.4518559528
 [8,] -0.163555968  0.274475348 -0.252450059 -0.66684780 -0.4660731103
 [9,] -0.060729628 -0.099198510 -0.492809116 -0.01152927  0.6804727629
[10,]  0.042123360 -0.272363503 -0.001096901 -0.03061736  0.0487848868
[11,]  0.159967574  0.797241835  0.071824676  0.33262222  0.0835119540
```

Principal Components Analysis (PCA)

Princomp

```
> princomp(USJudgeRatings[,-1],cor=T)
Call:
princomp(x = USJudgeRatings[, -1], cor = T)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
3.18330291 0.65163561 0.50525195 0.30216395 0.19313251 0.14057512 0.13592093 0.09105723
    Comp.9    Comp.10    Comp.11
0.07805078 0.05797301 0.04572500

 11  variables and  43 observations.
> sqrt(eigen(cor(USJudgeRatings[,-1]))$value)
 [1] 3.18330291 0.65163561 0.50525195 0.30216395 0.19313251 0.14057512 0.13592093
 [8] 0.09105723 0.07805078 0.05797301 0.04572500
```

Principal Components Analysis (PCA)

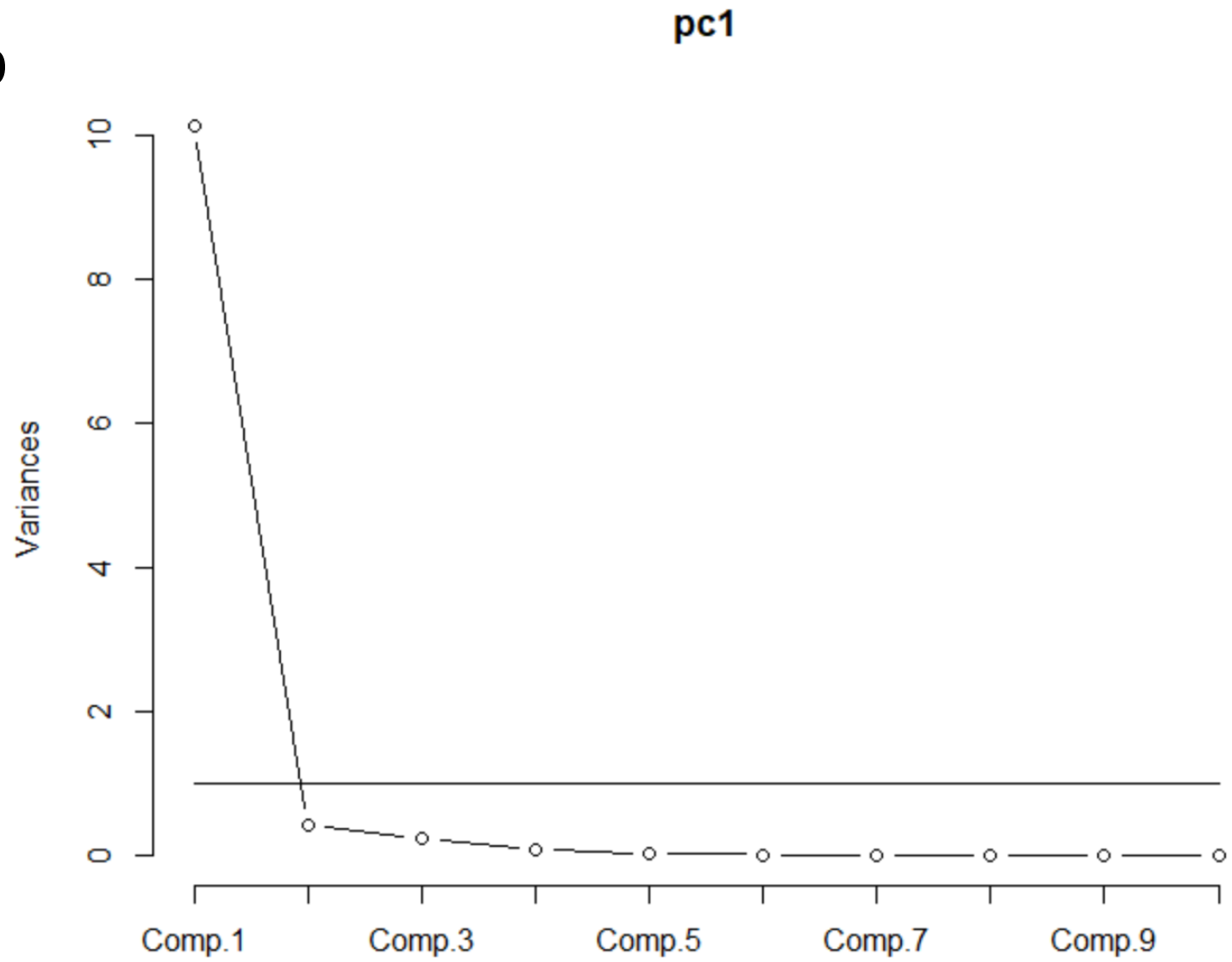
Princomp

```
> summary(princomp(USJudgeRatings[,-1],cor=T),loading=TRUE,score=TRUE)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
Standard deviation     3.1833029 0.65163561 0.50525195 0.302163952 0.193132512
Proportion of Variance 0.9212198 0.03860263 0.02320723 0.008300278 0.003390924
Cumulative Proportion  0.9212198 0.95982240 0.98302963 0.991329907 0.994720832
                           Comp.6     Comp.7       Comp.8       Comp.9      Comp.10
Standard deviation     0.140575118 0.1359209 0.0910572348 0.0780507762 0.0579730055
Proportion of Variance 0.001796488 0.0016795 0.0007537655 0.0005538112 0.0003055336
Cumulative Proportion  0.996517319 0.9981968 0.9989505847 0.9995043959 0.9998099295
                          Comp.11
Standard deviation     0.0457250007
Proportion of Variance 0.0001900705
Cumulative Proportion  1.0000000000

Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
INTG   0.289  0.574  0.118         0.375  0.510  0.230  0.285  0.145  0.103
DMNR   0.287  0.576 -0.177  0.240 -0.399 -0.514 -0.167  0.169        -0.105
DILG   0.304 -0.139  0.335  0.266  0.591 -0.298 -0.368        -0.355
CFMG   0.303 -0.310         0.478         -0.101  0.722                -0.207
DECI   0.302 -0.336         0.380 -0.399  0.448 -0.452  0.200  0.150  0.138
PREP   0.309 -0.125  0.229 -0.201         -0.336                0.717  0.252  -0.305
FAMI   0.307 -0.123  0.228 -0.524                               0.222        -0.544   0.452
ORAL   0.313                -0.229 -0.146         0.164 -0.274 -0.252  0.667   0.466
WRIT   0.311         0.148 -0.317 -0.237                      -0.493         -0.680
PHYS   0.281 -0.235 -0.820 -0.155  0.298                       0.272
RTEN   0.310  0.153 -0.201                0.234 -0.160 -0.797        -0.333
```

$$PC_1 = 0.289 * INTG + 0.287 * DMNR \cdots$$

Principal Components Analysis (PCA)

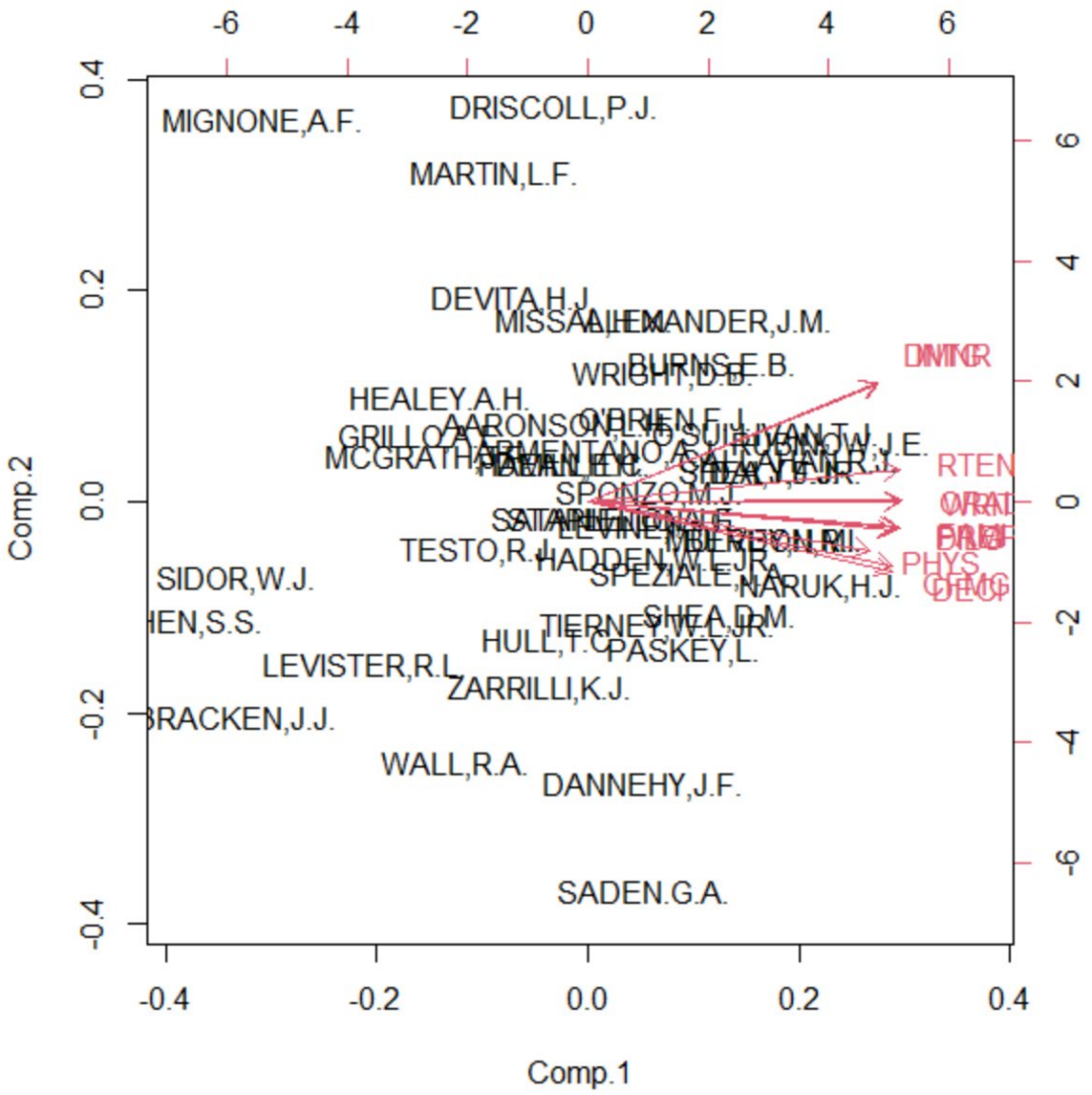
Princomp




```
screeplot(princomp(USJudgeRatings[,-1],cor=T),type="lines")
```

Principal Components Analysis (PCA)

Princomp




```
biplot(screeplot(princomp(USJudgeRatings[,-1],cor=T))
```

Principal Components Analysis (PCA)

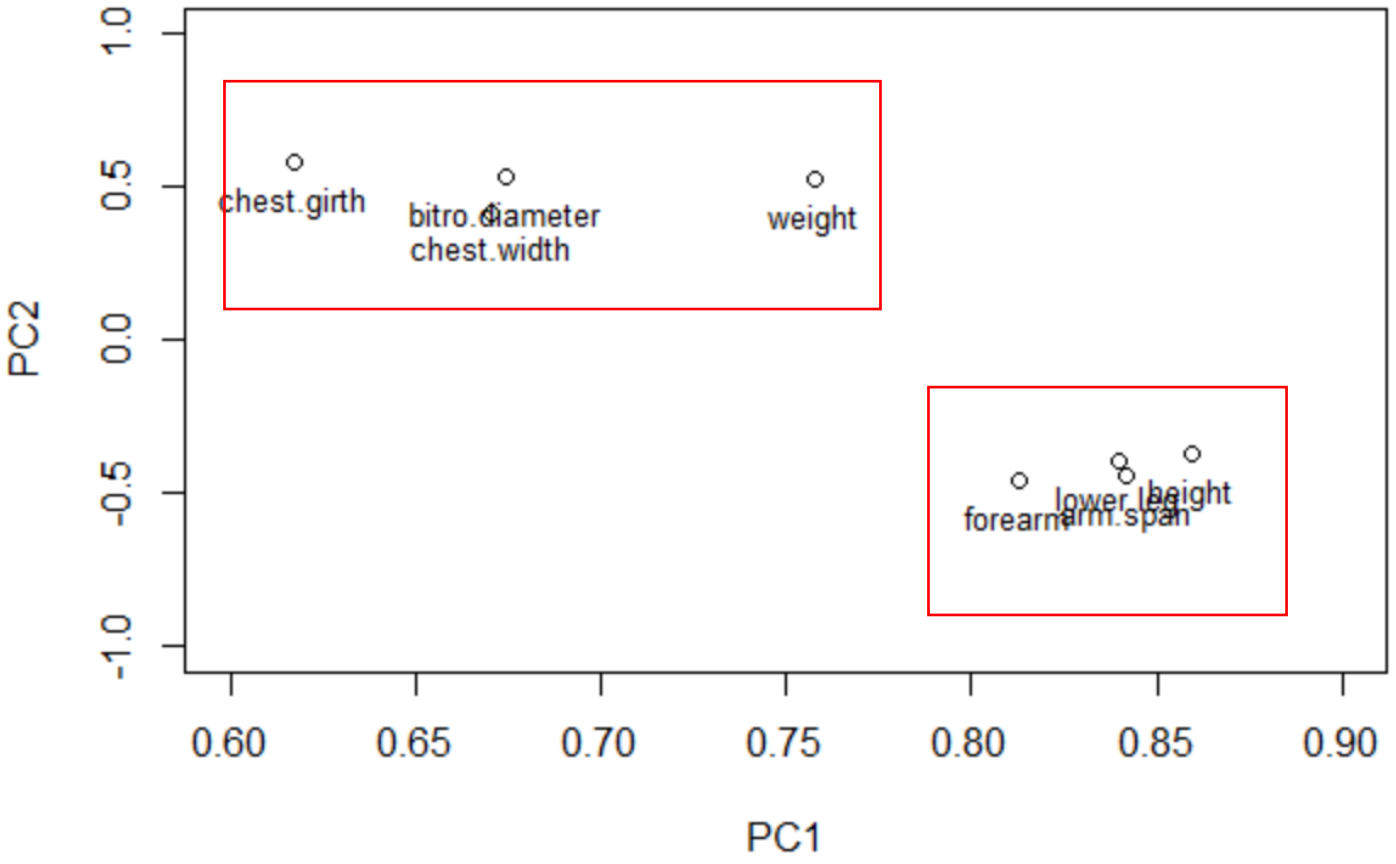
Table 14.3 Correlations among body measurements for 305 girls (`Harman23.cor`)

	Height	Arm span	Forearm	Lower leg	Weight	Bitro diameter	Chest girth	Chest width
Height	1.00	0.85	0.80	0.86	0.47	0.40	0.30	0.38
Arm span	0.85	1.00	0.88	0.83	0.38	0.33	0.28	0.41
Forearm	0.80	0.88	1.00	0.80	0.38	0.32	0.24	0.34
Lower leg	0.86	0.83	0.8	1.00	0.44	0.33	0.33	0.36
Weight	0.47	0.38	0.38	0.44	1.00	0.76	0.73	0.63
Bitro diameter	0.40	0.33	0.32	0.33	0.76	1.00	0.58	0.58
Chest girth	0.30	0.28	0.24	0.33	0.73	0.58	1.00	0.54
Chest width	0.38	0.41	0.34	0.36	0.63	0.58	0.54	1.00

Source: H. H. Harman, *Modern Factor Analysis, Third Edition Revised*, University of Chicago Press, 1976, Table 2.3.

```
> pc1
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 3, rotate = "none"
Standardized loadings (pattern matrix) based upon correlation matri
                PC1   PC2   PC3   h2    u2 com
height         0.86 -0.37 -0.07 0.88 0.118 1.4
arm.span       0.84 -0.44  0.08 0.91 0.091 1.5
forearm        0.81 -0.46  0.01 0.87 0.128 1.6
lower.leg      0.84 -0.40 -0.10 0.87 0.129 1.5
weight         0.76  0.52 -0.15 0.87 0.128 1.9
bitro.diameter 0.67  0.53 -0.05 0.74 0.258 1.9
chest.girth    0.62  0.58 -0.29 0.80 0.197 2.4
chest.width    0.67  0.42  0.59 0.97 0.025 2.7

                       PC1  PC2  PC3
SS loadings           4.67 1.77 0.48
Proportion Var        0.58 0.22 0.06
Cumulative Var        0.58 0.81 0.87
Proportion Explained  0.67 0.26 0.07
Cumulative Proportion 0.67 0.93 1.00

Mean item complexity =  1.9
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.05

Fit based upon off diagonal values = 0.99
```

Principal Components Analysis (PCA)



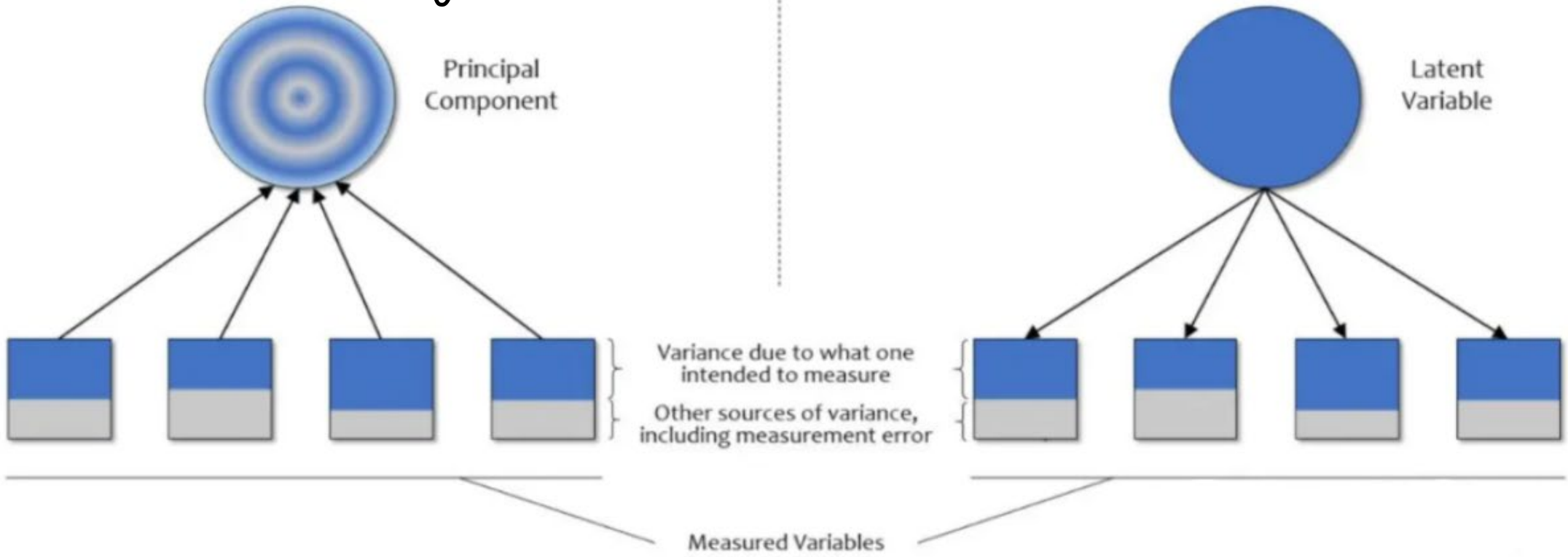

```
> pc1
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 3, rotate = "none"
Standardized loadings (pattern matrix) based upon correlation matri
                PC1   PC2   PC3   h2    u2 com
height         0.86 -0.37 -0.07 0.88 0.118 1.4
arm.span       0.84 -0.44  0.08 0.91 0.091 1.5
forearm        0.81 -0.46  0.01 0.87 0.128 1.6
lower.leg      0.84 -0.40 -0.10 0.87 0.129 1.5
weight         0.76  0.52 -0.15 0.87 0.128 1.9
bitro.diameter 0.67  0.53 -0.05 0.74 0.258 1.9
chest.girth    0.62  0.58 -0.29 0.80 0.197 2.4
chest.width    0.67  0.42  0.59 0.97 0.025 2.7

                       PC1  PC2  PC3
SS loadings           4.67 1.77 0.48
Proportion Var        0.58 0.22 0.06
Cumulative Var        0.58 0.81 0.87
Proportion Explained  0.67 0.26 0.07
Cumulative Proportion 0.67 0.93 1.00

Mean item complexity =  1.9
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.05

Fit based upon off diagonal values = 0.99
```

Principal Components Analysis (PCA) VS Exploratory Factor Analysis (EFA)

$Y = PX$ we hope YY^T diagonal, i.e.

eigenvalue decomposition $Av = \lambda v \Rightarrow AV = V\Lambda$ constraint $|v| = 1$

then $V^TV = I_p$ V is a orthogonal matrix. $YY^T = PXX^TP^T = [\ddots]$



Principal Components Analysis (PCA) VS Exploratory Factor Analysis (EFA)

Transformation of variable

$$PC_1 = a_1X_1 + a_2X_2 + \ldots + a_kX_k$$

generative latent probabilistic model.

$$X_i = a_1F_1 + a_2F_2 + \ldots + a_pF_p + U_i$$

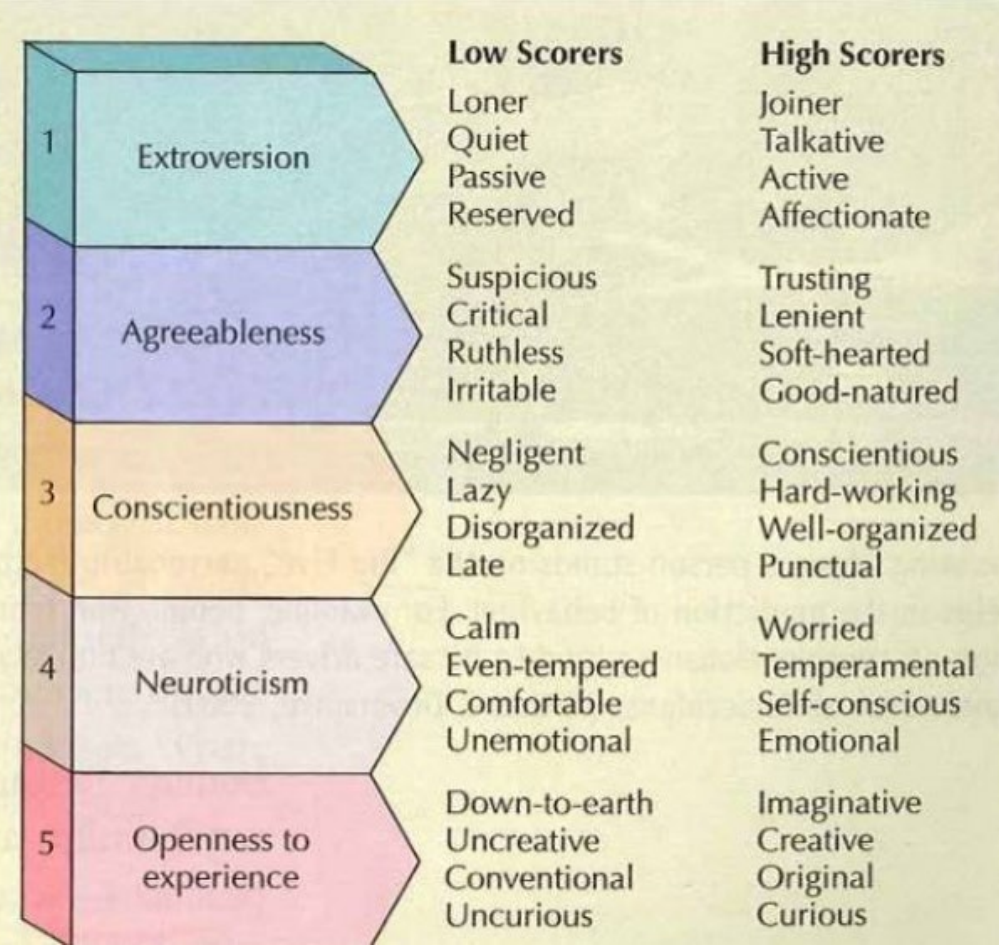
The goal of EFA is to explain the correlations among a set of observed variables by uncovering a smaller set of more fundamental **unobserved variables underlying the data**.

- ✓ data simplification/dimension reduction
- ✓ theory development/construct validation

Factors/ Common Factors/ Latent variable

Each factor is assumed to explain the variance shared (relationships correlation, covariance) among two or more observed variables

Principal Components Analysis (PCA) VS Exploratory Factor Analysis (EFA)



Relevant theory is the Five-Factor Model of Personality

Dimension/Scale	Subtests (WAIS-IV)
Verbal Comprehension	Similarities[a]
	Vocabulary[a]
	Information[a]
	Comprehension[b]
Perceptual Reasoning	Block Design[a]
	Matrix Reasoning[a]
	Visual Puzzles[a]
	Picture Completion[b]
	Figure Weights[b]
Working Memory	Digit Span[a]
	Arithmetic[a]
	Letter-Number Sequencing[b]
Processing Speed	Symbol Search[a]
	Coding[a]
	Cancellation[b]

a Core subtest.
b Supplemental subtest.

Study of Intelligence

Exploratory Factor Analysis (EFA)

Table 14.3 Correlations among body measurements for 305 girls (`Harman23.cor`)

	Height	Arm span	Forearm	Lower leg	Weight	Bitro diameter	Chest girth	Chest width
Height	1.00	0.85	0.80	0.86	0.47	0.40	0.30	0.38
Arm span	0.85	1.00	0.88	0.83	0.38	0.33	0.28	0.41
Forearm	0.80	0.88	1.00	0.80	0.38	0.32	0.24	0.34
Lower leg	0.86	0.83	0.8	1.00	0.44	0.33	0.33	0.36
Weight	0.47	0.38	0.38	0.44	1.00	0.76	0.73	0.63
Bitro diameter	0.40	0.33	0.32	0.33	0.76	1.00	0.58	0.58
Chest girth	0.30	0.28	0.24	0.33	0.73	0.58	1.00	0.54
Chest width	0.38	0.41	0.34	0.36	0.63	0.58	0.54	1.00

Source: H. H. Harman, *Modern Factor Analysis, Third Edition Revised*, University of Chicago Press, 1976, Table 2.3.

- How many factors?
- How to interpret the factors?
- Are the factors interrelated?

Exploratory Factor Analysis (EFA)

The Basic Factor Analysis Model

$$X = (X_1, X_2, \cdots, X_p)^T$$ **observed/measured/indicator variables**

intercepts

$$E(X) = \mu = (\mu_1, \mu_2, \cdots, \mu_p)^T, \quad \mathrm{Var}(X) = \Sigma = (\sigma_{ij})_{p\times p}.$$

$$\begin{cases} X_1 - \mu_1 = a_{11} f_1 + a_{12} f_2 + \cdots + a_{1m} f_m + \varepsilon_1 \\ X_2 - \mu_2 = a_{21} f_1 + a_{22} f_2 + \cdots + a_{2m} f_m + \varepsilon_2 \\ \quad\vdots \\ X_p - \mu_p = a_{p1} f_1 + a_{p2} f_2 + \cdots + a_{pm} f_m + \varepsilon_p \end{cases}$$

Measurement errors
unique factors

unobserved/latent/common factors

factor loading (regression coefficient) of variable *i* on factor *j*

Each measured variable can be expressed as a linear combination of common factors plus error

Exploratory Factor Analysis (EFA)

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

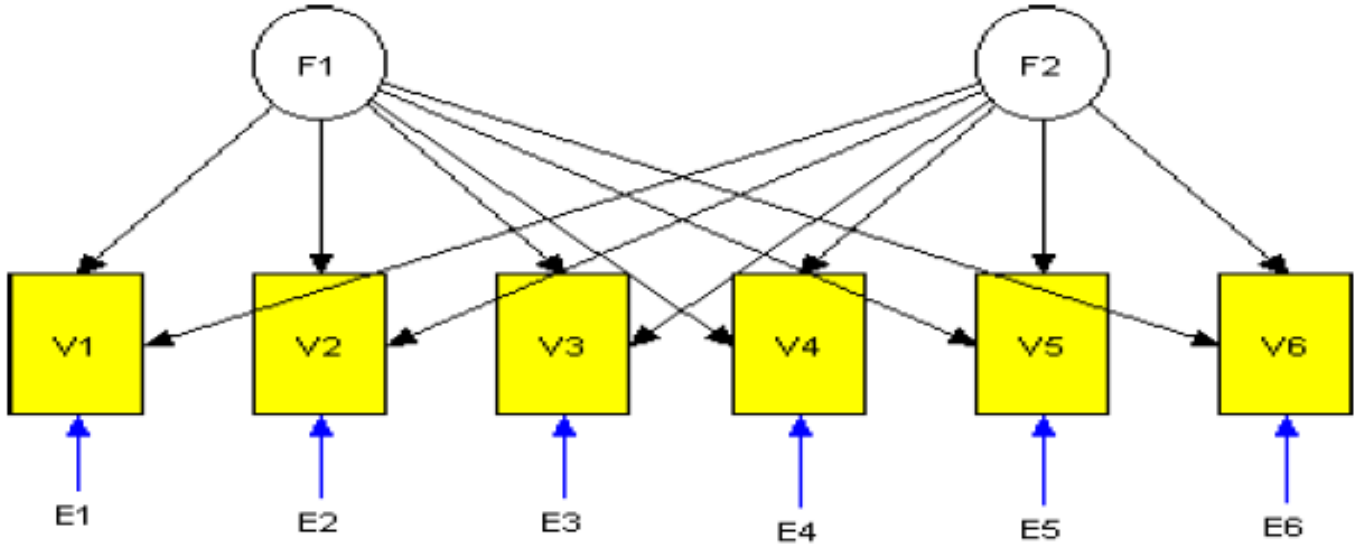
Matrix Form

$$F = (f_1, f_2, \cdots, f_m)^T$$

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)^T$$

$$A = (a_{ij})_{p \times m}$$

$p \times k$ factor loading (pattern) matrix



Exploratory Factor Analysis (EFA)

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

$$F = (f_1, f_2, \cdots, f_m)^T$$

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)^T$$

$$A = (a_{ij})_{p \times m}$$

$p \times k$ factor loading (pattern) matrix

Assumption

$$E(\varepsilon) = 0, \quad \mathrm{Var}(\varepsilon) = D = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2),$$

Means of **errors** are **zero** and errors are **uncorrelated** of each other

$$E(F) = 0, \quad \mathrm{Var}(F) = I_m,$$

Means of **Factors** are **zero** and factors are **Independent** of each other **(orthogonal)**

$$\mathrm{Cov}(F, \varepsilon) = 0.$$

Common factors and errors are uncorrelated

Exploratory Factor Analysis (EFA)

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

$$\Sigma = Var(X) = AA^T + D$$

$$Cov(X, F) = A$$

$$Cov(X_i, f_i) = a_{ij}$$

a_{ij} indicates the effect of f_j on X_i, with the influence of other factors partial out (regression coefficient)

If variables are standardized, which is usually the case in EFA, a_{ij} can be interpreted as the estimated correlation between the variable (X_i) and the factor (f_j)

$$A = (a_{ij})_{p\times m} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pm} \end{pmatrix}$$

$p \times k$ factor loading (pattern) matrix

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2 \qquad i = 1, \cdots, p$$

Amount (proportion) of variance of variable X_i that is accounted by the common factors

Communality (common variance): h_i^2

Exploratory Factor Analysis (EFA)

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

$$\Sigma = Var(X) = AA^T + D$$

$$Cov(X, F) = A$$

$$Cov(X_i, f_i) = a_{ij}$$

$$A = (a_{ij})_{p\times m} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pm} \end{pmatrix}$$

$p \times k$ factor loading (pattern) matrix

Amount of variance that is accounted for by factor *j*

$$\sum_{i=1}^{p} a_{ij}^2 \qquad j = 1, \cdots, m$$

Percentage of variance accounted for by factor *j* $= \dfrac{\sum_{i=1}^{p} a_{ij}^2}{total\ variance}$

(Standardize variables, p)

Exploratory Factor Analysis (EFA)

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

$$\Sigma = Var(X) = AA^T + D$$

$$\mathrm{Var}(\varepsilon) = D = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2),$$

$$Var(X_i) = h_i^2 + \sigma_i^2$$

$$i = 1, \cdots, p$$

Uniqueness (specific variance) σ_i^2

$$i = 1, \cdots, p$$

σ_i^2 measures the amount (proportion) of unexplained variance of variable X_i (variance not accounted for by the common factors)

Exploratory Factor Analysis (EFA)

Estimation

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

Diagonal

$$E = \begin{pmatrix} e_1 & e_2 & \cdots & e_n \end{pmatrix}$$

$$\Sigma = Var(X) = AA^T + D$$

$$\Sigma = E \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} E^T$$

like PCA, Σ is decomposed and use the first m PC as the latent factors and the remaining small PC is put into error term

$$E^T C E = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

$$= e_1 \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} e_1^T + e_2 \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \lambda_2 & \vdots \\ 0 & \cdots & 0 \end{pmatrix} e_2^T + \cdots + e_n \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} e_n^T$$

factors

error.

Exploratory Factor Analysis (EFA)

Estimation

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

Diagonal

$$E = (e_1 \quad e_2 \quad \cdots \quad e_n)$$

$$\Sigma = Var(X) = AA^T + D$$

$$A = (a_{ij})_{p\times m} = (\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, \cdots, \sqrt{\lambda_m}e_m)$$

$$D = diag(s_{11} - h_1^2, s_{22} - h_2^2, \cdots, s_{pp} - h_p^2)$$

Principal component Method

Recall the definition

λ_i is not λ_i, the diagonal of $A^TA \in \mathbb{R}^{m\times m}$

h_i^2 is the diagonal of $AA^T \in \mathbb{R}^{p\times p}$

Exploratory Factor Analysis (EFA)

Estimation

The Basic Factor Analysis Model

Principal component Method

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

$$= \begin{pmatrix} 2 & 0 \\ 0 & 2/5 \end{pmatrix}$$

$$\begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 2/5 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

$$\approx \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} (1/\sqrt{2} \quad 1/\sqrt{2}) + D = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} (1/\sqrt{2} \quad 1/\sqrt{2}) + D$$

AA^T

Exploratory Factor Analysis (EFA)

Estimation

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

Diagonal

$$E = (e_1 \quad e_2 \quad \cdots \quad e_n)$$

$$\Sigma = Var(X) = AA^T + D$$

How many factors (m) ? $m < p$

Rule #1 : Examine the percentage of variance explained by each factor. Ignore any additional factor if it can only explain a small percentage

Rule #2 : Examine the communalities of the variables. Make sure they are high enough. The presence of low communalities suggests more factors should be extracted.

Rule #3 : The extracted factors should be interpretable **(most important)**

Exploratory Factor Analysis (EFA)

Table 14.3 Correlations among body measurements for 305 girls (`Harman23.cor`)

	Height	Arm span	Forearm	Lower leg	Weight	Bitro diameter	Chest girth	Chest width
Height	1.00	0.85	0.80	0.86	0.47	0.40	0.30	0.38
Arm span	0.85	1.00	0.88	0.83	0.38	0.33	0.28	0.41
Forearm	0.80	0.88	1.00	0.80	0.38	0.32	0.24	0.34
Lower leg	0.86	0.83	0.8	1.00	0.44	0.33	0.33	0.36
Weight	0.47	0.38	0.38	0.44	1.00	0.76	0.73	0.63
Bitro diameter	0.40	0.33	0.32	0.33	0.76	1.00	0.58	0.58
Chest girth	0.30	0.28	0.24	0.33	0.73	0.58	1.00	0.54
Chest width	0.38	0.41	0.34	0.36	0.63	0.58	0.54	1.00

Source: H. H. Harman, *Modern Factor Analysis, Third Edition Revised*, University of Chicago Press, 1976, Table 2.3.

```
> pc1
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 3, rotate = "none
Standardized loadings (pattern matrix) based upon correlation matr
                 PC1   PC2   PC3   h2    u2 com
height          0.86 -0.37 -0.07 0.88 0.118 1.4
arm.span        0.84 -0.44  0.08 0.91 0.091 1.5
forearm         0.81 -0.46  0.01 0.87 0.128 1.6
lower.leg       0.84 -0.40 -0.10 0.87 0.129 1.5
weight          0.76  0.52 -0.15 0.87 0.128 1.9
bitro.diameter 0.67  0.53 -0.05 0.74 0.258 1.9
chest.girth     0.62  0.58 -0.29 0.80 0.197 2.4
chest.width     0.67  0.42  0.59 0.97 0.025 2.7

                        PC1  PC2  PC3
SS loadings            4.67 1.77 0.48
Proportion Var         0.58 0.22 0.06
Cumulative Var         0.58 0.81 0.87
Proportion Explained   0.67 0.26 0.07
Cumulative Proportion 0.67 0.93 1.00

Mean item complexity =  1.9
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.05

Fit based upon off diagonal values = 0.99
```

Exploratory Factor Analysis (EFA)

Table 14.3 Correlations among body measurements for 305 girls (`Harman23.cor`)

	Height	Arm span	Forearm	Lower leg	Weight	Bitro diameter	Chest girth	Chest width
Height	1.00	0.85	0.80	0.86	0.47	0.40	0.30	0.38
Arm span	0.85	1.00	0.88	0.83	0.38	0.33	0.28	0.41
Forearm	0.80	0.88	1.00	0.80	0.38	0.32	0.24	0.34
Lower leg	0.86	0.83	0.8	1.00	0.44	0.33	0.33	0.36
Weight	0.47	0.38	0.38	0.44	1.00	0.76	0.73	0.63
Bitro diameter	0.40	0.33	0.32	0.33	0.76	1.00	0.58	0.58
Chest girth	0.30	0.28	0.24	0.33	0.73	0.58	1.00	0.54
Chest width	0.38	0.41	0.34	0.36	0.63	0.58	0.54	1.00

Source: H. H. Harman, *Modern Factor Analysis, Third Edition Revised*, University of Chicago Press, 1976, Table 2.3.

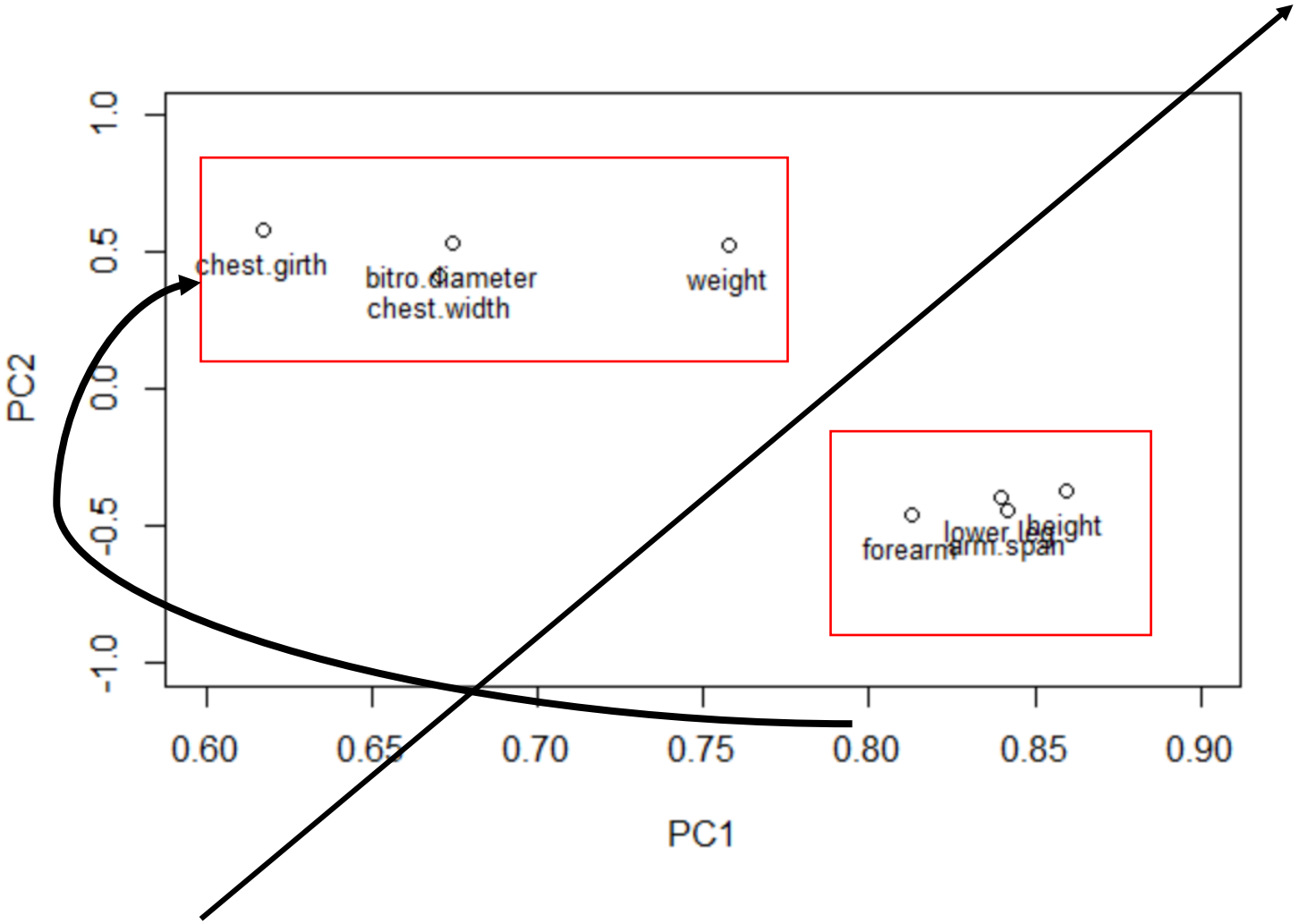
```
                PC1   h2   u2 com
height         0.86 0.74 0.26   1
arm.span       0.84 0.71 0.29   1
forearm        0.81 0.66 0.34   1
lower.leg      0.84 0.70 0.30   1
weight         0.76 0.57 0.43   1
bitro.diameter 0.67 0.45 0.55   1
chest.girth    0.62 0.38 0.62   1
chest.width    0.67 0.45 0.55   1
```

```
                PC1   PC2   h2    u2 com
height         0.86 -0.37 0.88 0.123 1.4
arm.span       0.84 -0.44 0.90 0.097 1.5
forearm        0.81 -0.46 0.87 0.128 1.6
lower.leg      0.84 -0.40 0.86 0.139 1.4
weight         0.76  0.52 0.85 0.150 1.8
bitro.diameter 0.67  0.53 0.74 0.261 1.9
chest.girth    0.62  0.58 0.72 0.283 2.0
chest.width    0.67  0.42 0.62 0.375 1.7
```

```
                PC1   PC2   PC3   h2    u2 com
height         0.86 -0.37 -0.07 0.88 0.118 1.4
arm.span       0.84 -0.44  0.08 0.91 0.091 1.5
forearm        0.81 -0.46  0.01 0.87 0.128 1.6
lower.leg      0.84 -0.40 -0.10 0.87 0.129 1.5
weight         0.76  0.52 -0.15 0.87 0.128 1.9
bitro.diameter 0.67  0.53 -0.05 0.74 0.258 1.9
chest.girth    0.62  0.58 -0.29 0.80 0.197 2.4
chest.width    0.67  0.42  0.59 0.97 0.025 2.7
```

Exploratory Factor Analysis (EFA)

Factor Rotation



- Simple structure is achieved when (Thurstone, 1947)
  - each variable is only related to "a few" factors, preferably one
  - each factor is only related to "a few" variables

To transform the initial pattern matrix into simple structure for easier interpretation

Exploratory Factor Analysis (EFA)

Factor Rotation

Why the rotation is needed?

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon, \qquad E(F) = 0, \quad \mathrm{Var}(F) = I_m,$$

$$\Sigma = Var(X) = AA^T + D$$

Means of **Factors** are **zero** and factors are **Independent** of each other **(orthogonal)**

$F \sim Dist(0, I_m)$

Let $Z = \Gamma^T F \qquad \Gamma^T\Gamma = I$ **orthogonal rotation**

$Z = \Gamma^T F$. $\Gamma \in \mathbb{R}^{m \times m}$ $X = A\Gamma Z + \varepsilon,$ **Factor** $A\Gamma$ ------ Loading matrix

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathrm{Var}(\Gamma^T F) = \Gamma^T \mathrm{Var}(F)\Gamma = I_m, \\
\mathrm{Cov}(Z, \varepsilon) &= \mathrm{Cov}(\Gamma^T F, \varepsilon) = \Gamma^T \mathrm{Cov}(F, \varepsilon) = 0, \\
\mathrm{Var}(X) &= \mathrm{Var}(A\Gamma Z) + \mathrm{Var}(\varepsilon) = A\Gamma \mathrm{Var}(Z)\Gamma^T A^T + D \\
&= AA^T + D.
\end{aligned}$$

$\mathbb{E}(ZZ^T) = \mathbb{E}(\Gamma^T F F^T \Gamma)$

$= \Gamma^T(\mathbb{E}FF^T)\Gamma$

$= \Gamma^T I_m \Gamma$

$= \Gamma^T\Gamma = I_m$ ← the solution of FA is not unique

Exploratory Factor Analysis (EFA)

Factor Rotation

```
> pc1
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 3, rota
Standardized loadings (pattern matrix) based upon correl
                PC1   PC2   PC3   h2    u2 com
height         0.86 -0.37 -0.07 0.88 0.118 1.4
arm.span       0.84 -0.44  0.08 0.91 0.091 1.5
forearm        0.81 -0.46  0.01 0.87 0.128 1.6
lower.leg      0.84 -0.40 -0.10 0.87 0.129 1.5
weight         0.76  0.52 -0.15 0.87 0.128 1.9
bitro.diameter 0.67  0.53 -0.05 0.74 0.258 1.9
chest.girth    0.62  0.58 -0.29 0.80 0.197 2.4
chest.width    0.67  0.42  0.59 0.97 0.025 2.7

                       PC1  PC2  PC3
SS loadings           4.67 1.77 0.48
Proportion Var        0.58 0.22 0.06
Cumulative Var        0.58 0.81 0.87
Proportion Explained  0.67 0.26 0.07
Cumulative Proportion 0.67 0.93 1.00

Mean item complexity =  1.9
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.05

Fit based upon off diagonal values = 0.99
```

```
> principal(r=Harman23.cor$cov,nfactors=3,rotate="varimax")
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 3, rotate = "varimax"
Standardized loadings (pattern matrix) based upon correlation matrix
                RC1  RC2  RC3   h2    u2 com
height         0.90 0.25 0.09 0.88 0.118 1.2
arm.span       0.92 0.13 0.20 0.91 0.091 1.1
forearm        0.92 0.13 0.13 0.87 0.128 1.1
lower.leg      0.90 0.23 0.05 0.87 0.129 1.1
weight         0.26 0.87 0.23 0.87 0.128 1.3
bitro.diameter 0.18 0.79 0.30 0.74 0.258 1.4
chest.girth    0.12 0.88 0.07 0.80 0.197 1.1
chest.width    0.21 0.45 0.85 0.97 0.025 1.7

                       RC1  RC2  RC3
SS loadings           3.48 2.50 0.95
Proportion Var        0.43 0.31 0.12
Cumulative Var        0.43 0.75 0.87
Proportion Explained  0.50 0.36 0.14
Cumulative Proportion 0.50 0.86 1.00

Mean item complexity =  1.2
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.05

Fit based upon off diagonal values = 0.99
```

Exploratory Factor Analysis (EFA)

Factor Rotation

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon, \qquad E(F) = 0, \quad \mathrm{Var}(F) = I_m,$$

$$\Sigma = Var(X) = AA^T + D$$

Means of **Factors** are **zero** and factors are **Independent** of each other **(orthogonal)**

$\Gamma^T\Gamma = I \Rightarrow$ Orthogonal rotations

~~$\Gamma^T\Gamma = I$~~ **Oblique rotations**

Let $Z = \Gamma^T F$

$\mathrm{Var}\, Z = \mathbb{E}\, ZZ^T = \Gamma^T(\mathbb{E}FF^T)\Gamma = \Gamma^T\Gamma = \phi$

$X = A\Gamma Z + \varepsilon,$ **Factor** $A\Gamma$ ------ Loading matrix

$\Gamma^T\Gamma = \phi \neq I_m$

$Var(Z) = \Phi$

but $\Gamma\Gamma^T\Gamma\Gamma^T = I_m$

$$\mathrm{Cov}(Z, \varepsilon) = \mathrm{Cov}(\Gamma^T F, \varepsilon) = \Gamma^T \mathrm{Cov}(F, \varepsilon) = 0,$$

$$\mathrm{Var}(X) = \mathrm{Var}(A\Gamma Z) + \mathrm{Var}(\varepsilon) = A\Gamma \mathrm{Var}(Z)\Gamma^T A^T + D$$

$$= AA^T + D.$$

$\Gamma\Phi\Gamma^T = I$

Exploratory Factor Analysis (EFA)

Factor Rotation

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon, \qquad E(F) = 0, \quad \mathrm{Var}(F) = I_m,$$

$$\Sigma = Var(X) = AA^T + D$$

Means of **Factors** are **zero** and factors are **Independent** of each other **(orthogonal)**

Let $Z = \Gamma^T F$ ~~$\Gamma^T\Gamma = I$~~ **Oblique rotations**

$$X = A\Gamma Z + \varepsilon,$$ **Factor** $A\Gamma$ ------ Loading matrix

$$Var(Z) = \Phi$$

$$\begin{aligned}
\mathrm{Cov}(Z, \varepsilon) &= \mathrm{Cov}(\Gamma^T F, \varepsilon) = \Gamma^T \mathrm{Cov}(F, \varepsilon) = 0, \\
\mathrm{Var}(X) &= \mathrm{Var}(A\Gamma Z) + \mathrm{Var}(\varepsilon) = A\Gamma \mathrm{Var}(Z)\Gamma^T A^T + D \\
&= AA^T + D.
\end{aligned}$$

$$\Gamma\Phi\Gamma^T = I$$

Exploratory Factor Analysis (EFA)

The Basic Factor Analysis Model

$$X = \mu + AF + \varepsilon,$$

$$\Sigma = Var(X) = AA^T + D$$

reduced correlation matrix

$$\Sigma - D = E' \begin{pmatrix} \lambda_1' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n' \end{pmatrix} E'^T$$

$$\hat{\lambda}_1^* \geq \hat{\lambda}_2^* \geq \cdots \geq \hat{\lambda}_m^* > 0,$$

$$A = (a_{ij})_{p \times m} = (\sqrt{\lambda_1'}e_1', \sqrt{\lambda_2'}e_2', \cdots, \sqrt{\lambda_m'}e_m')$$

$$D = diag(s_{11} - h_1^2, s_{22} - h_2^2, \cdots, s_{pp} - h_p^2)$$

Estimation

- Principal component Method
- Principal axis (PAF) Method

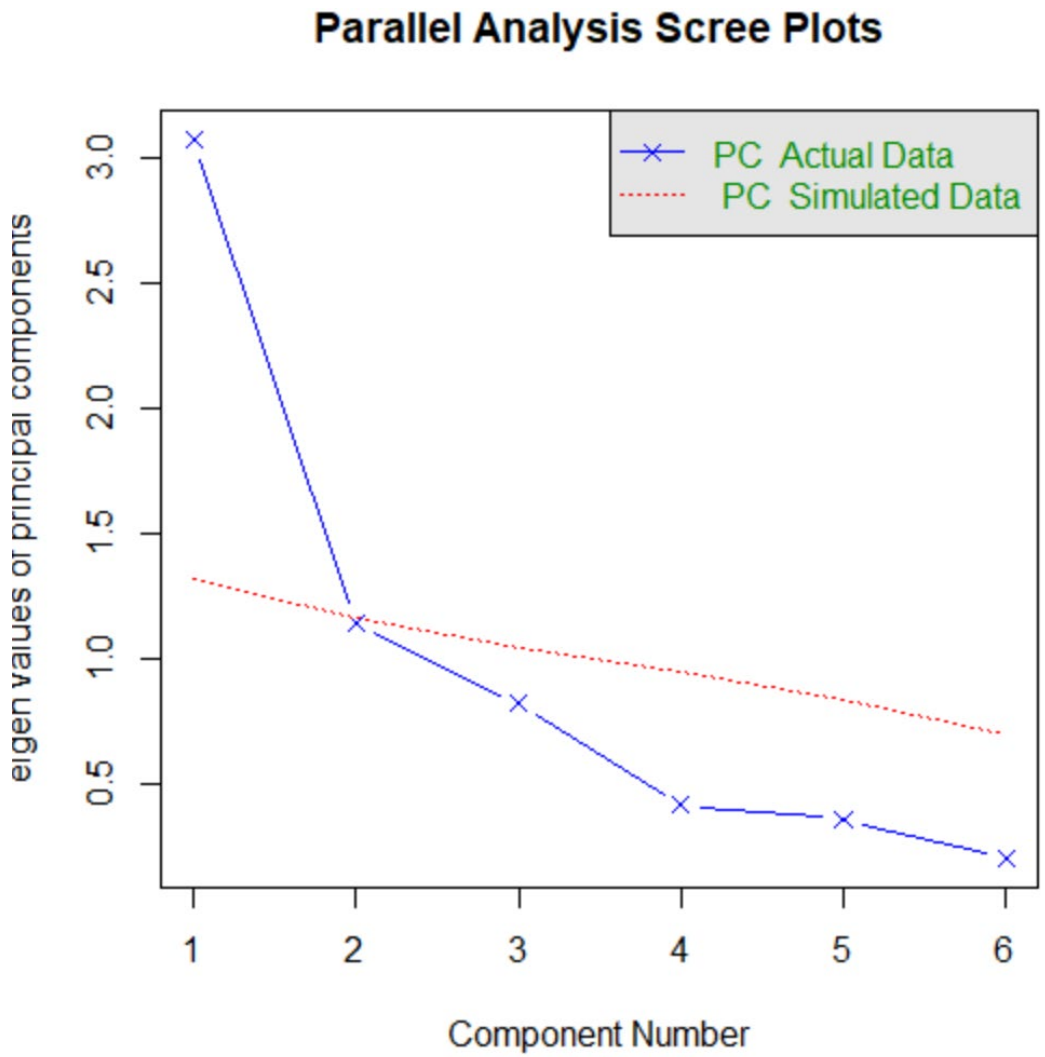
?

Iteration

Exploratory Factor Analysis (EFA)

```
> ability.cov
$cov
        general picture  blocks   maze reading   vocab
general  24.641   5.991  33.520  6.023  20.755  29.701
picture   5.991   6.700  18.137  1.782   4.936   7.204
blocks   33.520  18.137 149.831 19.424  31.430  50.753
maze      6.023   1.782  19.424 12.711   4.757   9.075
reading  20.755   4.936  31.430  4.757  52.604  66.762
vocab    29.701   7.204  50.753  9.075  66.762 135.292

$center
[1] 0 0 0 0 0 0

$n.obs
[1] 112
```

```
> cov2cor(ability.cov$cov)
          general   picture    blocks      maze   reading     vocab
general 1.0000000 0.4662649 0.5516632 0.3403250 0.5764799 0.5144058
picture 0.4662649 1.0000000 0.5724364 0.1930992 0.2629229 0.2392766
blocks  0.5516632 0.5724364 1.0000000 0.4450901 0.3540252 0.3564715
maze    0.3403250 0.1930992 0.4450901 1.0000000 0.1839645 0.2188370
reading 0.5764799 0.2629229 0.3540252 0.1839645 1.0000000 0.7913779
vocab   0.5144058 0.2392766 0.3564715 0.2188370 0.7913779 1.0000000
```



Exploratory Factor Analysis (EFA)

```
factor.pa<-function(S, m){
  p<-nrow(S)
  diag_S<-diag(S)
  sum_rank<-sum(diag_S)
  rowname<-paste("X", 1:p, sep="")
  colname<-paste("Factor", 1:m, sep="")
  A<-matrix(0, nrow=p, ncol=m, dimnames=list(rowname, colname))
  eig<-eigen(S)
  for (i in 1:m)
    A[,i]<-sqrt(eig$values[i])*eig$vectors[,i]
  h<-diag(A%*%t(A))

  rowname<-c("SS loadings", "Proportion Var", "Cumulative Var")
  B<-matrix(0, nrow=3, ncol=m, dimnames=list(rowname, colname))
  for (i in 1:m){
    B[1,i]<-sum(A[,i]^2)
    B[2,i]<-B[1,i]/sum_rank
    B[3,i]<-sum(B[1,1:i])/sum_rank
  }
  method<-c("Principal Component Method")
  list(method=method, loadings=A,
       var=cbind(common=h, spcific=diag_S-h), B=B)
}
```

Using squared multiple correlations (SMC) as the initial estimates of the communalities and proceed as PC

```
d = 1-diag(1/solve(S))
kmax=200; k<-1; h <- diag_S-d
repeat{
  diag(S)<- h; h1<-h; eig<-eigen(S)
  for (i in 1:m)
    A[,i]<-sqrt(eig$values[i])*eig$vectors[,i]
  h<-diag(A %*% t(A))
  if ((sqrt(sum((h-h1)^2))<1e-4)|k==kmax) break
  k<-k+1
}
```

- PC is the easiest, but it extracts the total variances instead of the common variances. So, it tends to overestimate the factor loadings, esp. when correlations are small.
- PAF is a modified approach of PC, and it overcomes some of the drawbacks of PC

Exploratory Factor Analysis (EFA)

```
> pa2=factor.pa(R,2)
> pa2
$method
[1] "Principal Axis Method"

$loadings
     Factor1     Factor2
X1 0.7497858 -0.07027341
X2 0.5239144 -0.31943134
X3 0.7500878 -0.52173392
X4 0.3923568 -0.22016964
X5 0.8179396  0.51821423
X6 0.7266817  0.37692981

$var
      common    spcific
X1 0.5671172 0.43288284
X2 0.3765226 0.62347736
X3 0.8348379 0.16516206
X4 0.2024185 0.79758150
X5 0.9375712 0.06242876
X6 0.6701423 0.32985765

$B
                 Factor1   Factor2
SS loadings    2.7503321 0.8382778
Proportion Var 0.4583887 0.1397130
Cumulative Var 0.4583887 0.5981016

$iterative
[1] 68
```

```
> pa1=fa(R,nfactors=2,fm="pa",rotate = "none",n.obs=112)
> pa1
Factor Analysis using method =  pa
Call: fa(r = R, nfactors = 2, n.obs = 112, rotate = "none", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
          PA1   PA2   h2    u2 com
general 0.75  0.07 0.57 0.432 1.0
picture 0.52  0.32 0.38 0.623 1.7
blocks  0.75  0.52 0.83 0.166 1.8
maze    0.39  0.22 0.20 0.798 1.6
reading 0.81 -0.51 0.91 0.089 1.7
vocab   0.73 -0.39 0.69 0.313 1.5

                       PA1  PA2
SS loadings           2.75 0.83
Proportion Var        0.46 0.14
Cumulative Var        0.46 0.60
Proportion Explained  0.77 0.23
Cumulative Proportion 0.77 1.00
```

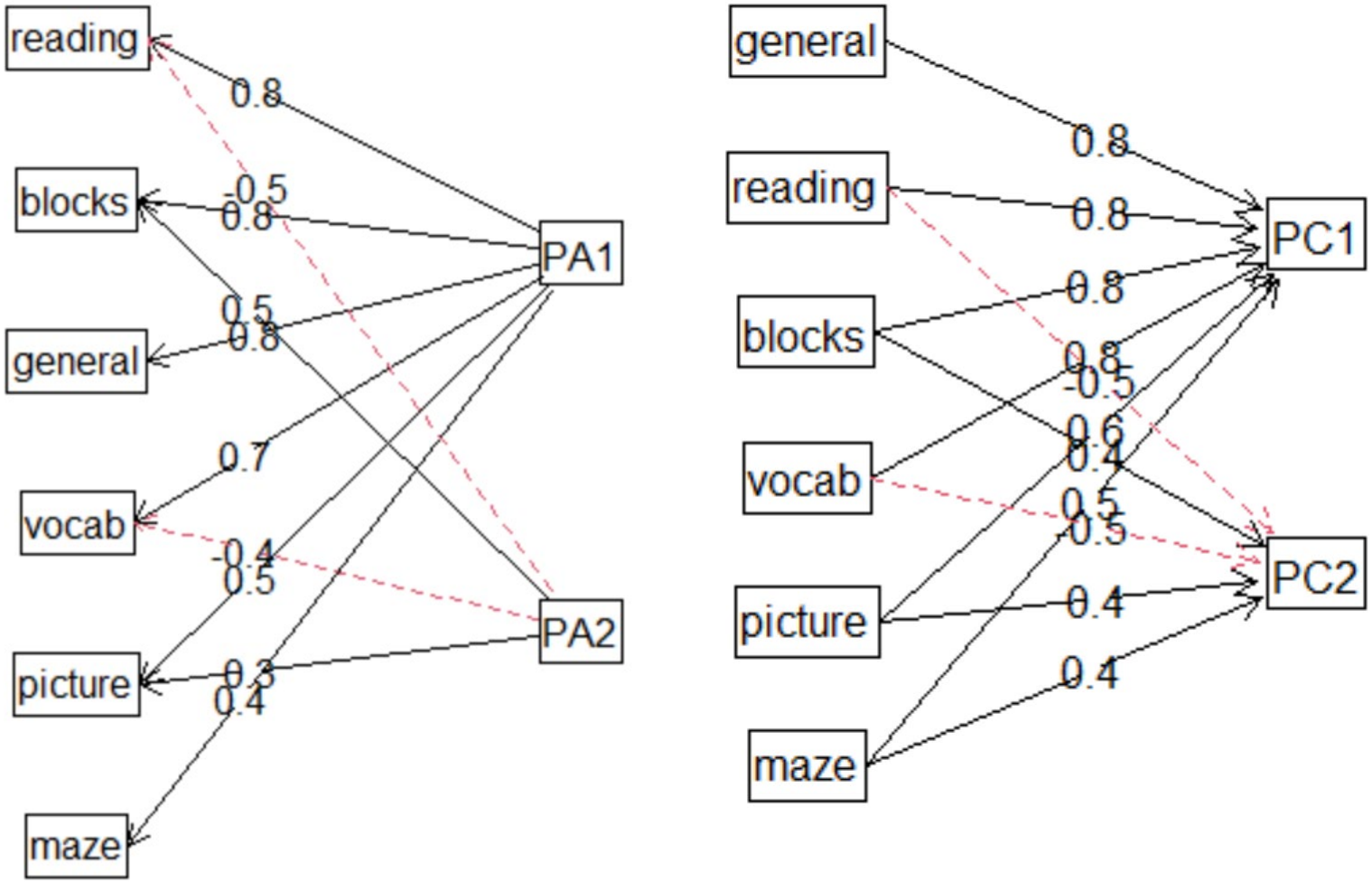
```
> Epa1=R-pa1$lo
> sum(Epa1^2)
[1] 0.0311694
> Epa2=R-pa2$lo
> sum(Epa2^2)
[1] 0.03108547
```

Exploratory Factor Analysis (EFA)

```
> fa(R,nfactors=2,fm="pa",rotate = "varimax",n.obs=112)
Factor Analysis using method =  pa
Call: fa(r = R, nfactors = 2, n.obs = 112, rotate = "varimax",
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
          PA1  PA2   h2    u2 com
general 0.49 0.57 0.57 0.432 2.0
picture 0.16 0.59 0.38 0.623 1.1
blocks  0.18 0.89 0.83 0.166 1.1
maze    0.13 0.43 0.20 0.798 1.2
reading 0.93 0.20 0.91 0.089 1.1
vocab   0.80 0.23 0.69 0.313 1.2

                       PA1  PA2
SS loadings           1.83 1.75
Proportion Var        0.30 0.29
Cumulative Var        0.30 0.60
Proportion Explained  0.51 0.49
Cumulative Proportion 0.51 1.00
```

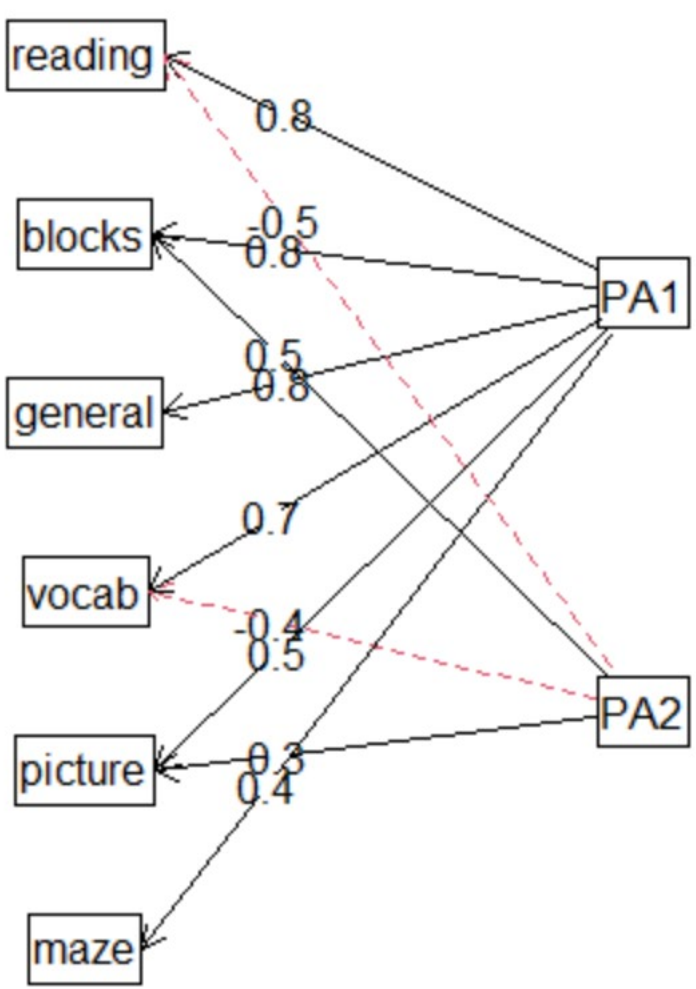
```
> fa(R,nfactors=2,fm="pa",rotate = "promax",n.obs=112)
Factor Analysis using method =  pa
Call: fa(r = R, nfactors = 2, n.obs = 112, rotate = "promax",
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
           PA1   PA2   h2    u2 com
general  0.37  0.48 0.57 0.432 1.9
picture -0.03  0.63 0.38 0.623 1.0
blocks  -0.10  0.97 0.83 0.166 1.0
maze     0.00  0.45 0.20 0.798 1.0
reading  1.00 -0.09 0.91 0.089 1.0
vocab    0.84 -0.01 0.69 0.313 1.0

                       PA1  PA2
SS loadings           1.83 1.75
Proportion Var        0.30 0.29
Cumulative Var        0.30 0.60
Proportion Explained  0.51 0.49
Cumulative Proportion 0.51 1.00

 With factor correlations of
     PA1  PA2
PA1 1.00 0.55
PA2 0.55 1.00
```

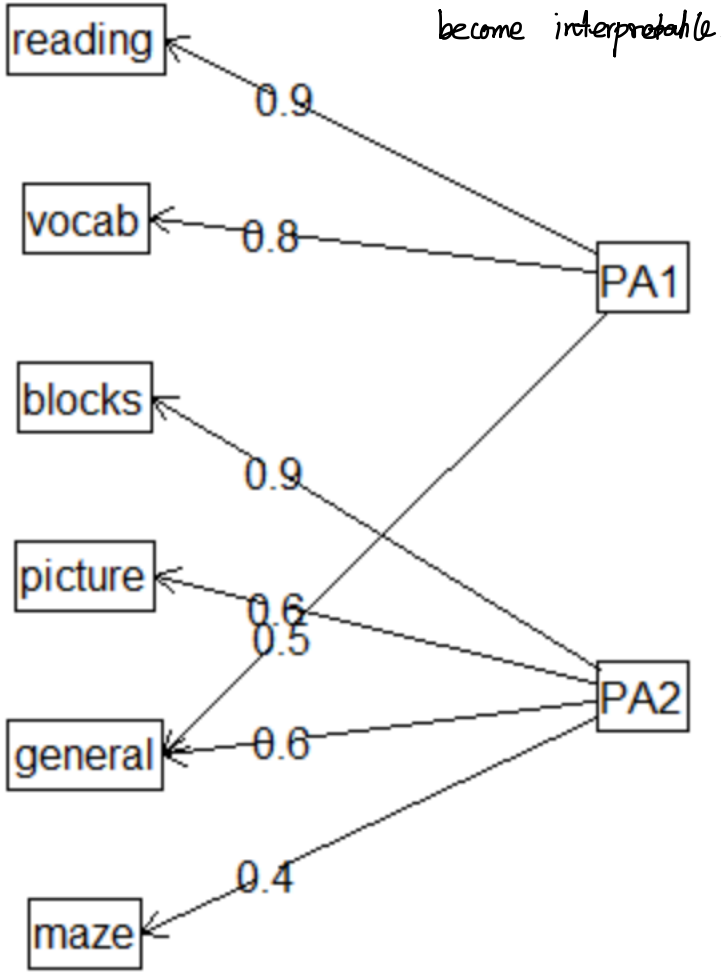
Exploratory Factor Analysis (EFA)

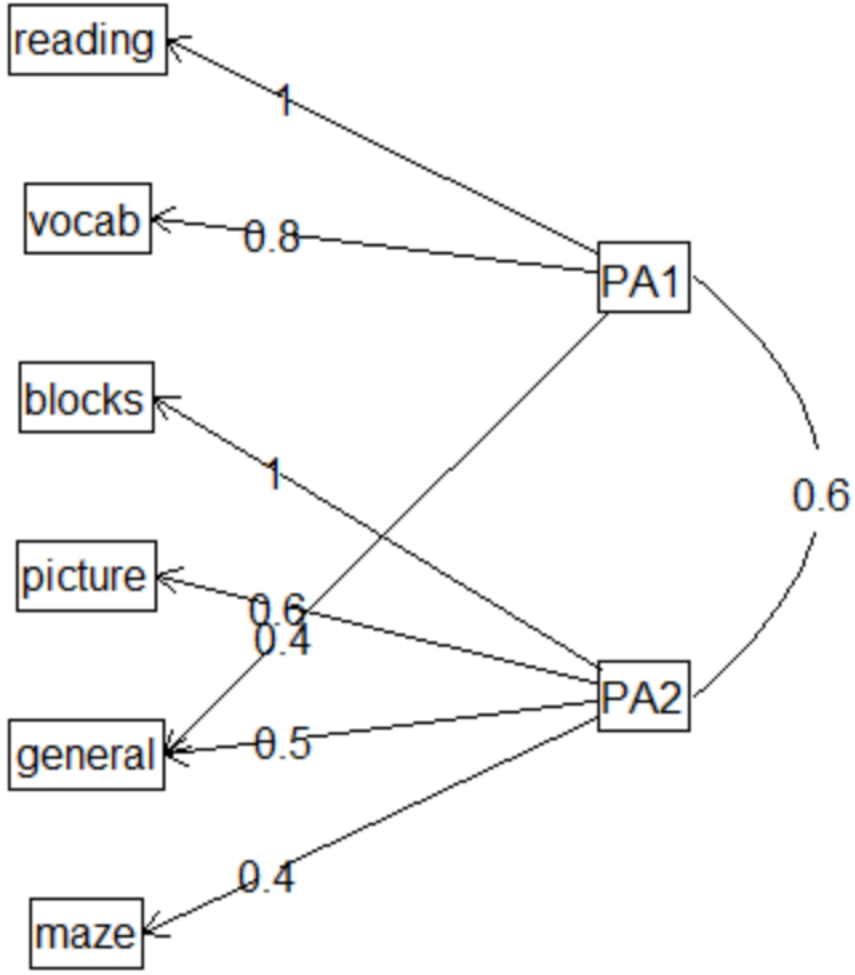


Exploratory Factor Analysis (EFA)

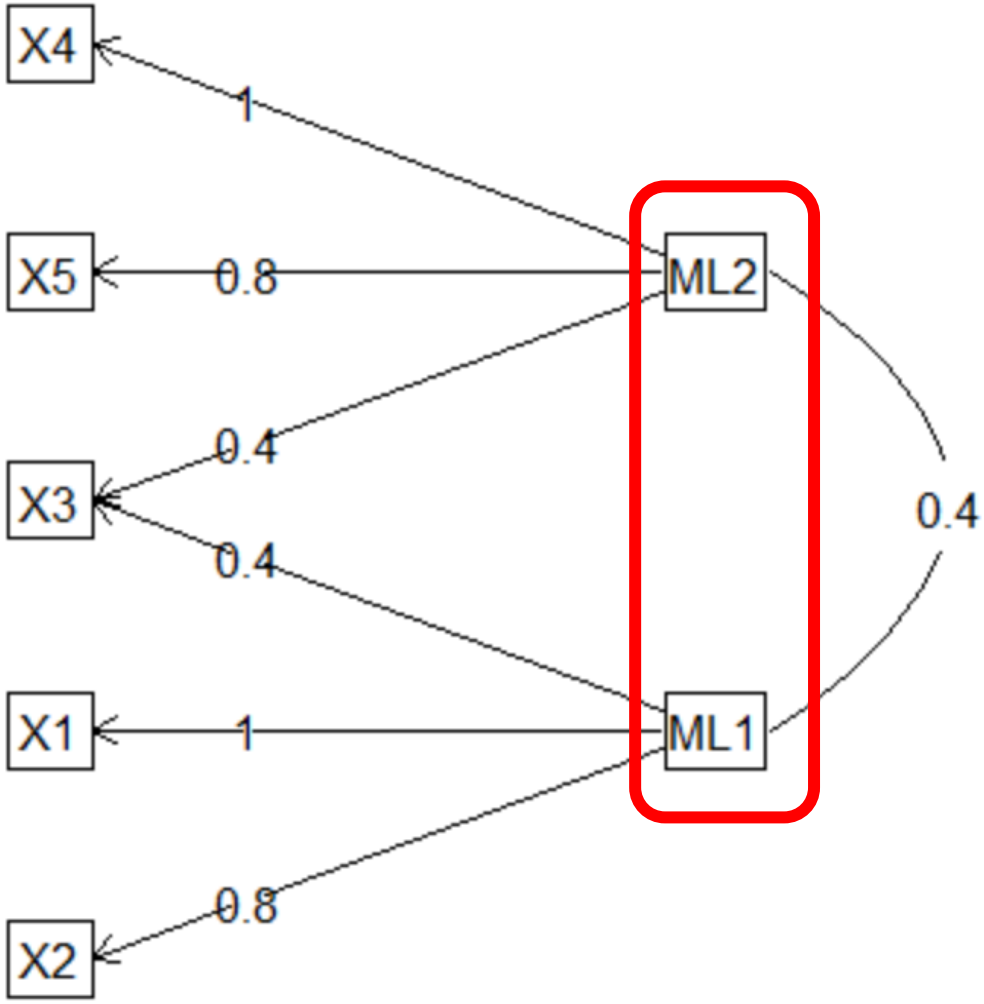
the orthogonal latent factors may be hard to interpret, which is the motivation to perform rotation. Although the factors are correlated with each other, but the meaning of factors become interpretable.







Confirmatory Factor Analysis (CFA) vs Exploratory Factor Analysis (EFA)



In contrast to exploratory factor analysis, a confirmatory factor analysis **begins by defining the latent variables one would like to measure**

number of latent factors is unknown before fitting.

This is based on **substantive theory and/or previous knowledge**. One then constructs observable variables to measure these latent variables. Thus, in a confirmatory factor analysis, the number of factors is known and equal to the number of latent variables.

EFA as a preliminary step before CFA

Confirmatory Factor Analysis (CFA) vs Exploratory Factor Analysis (EFA)

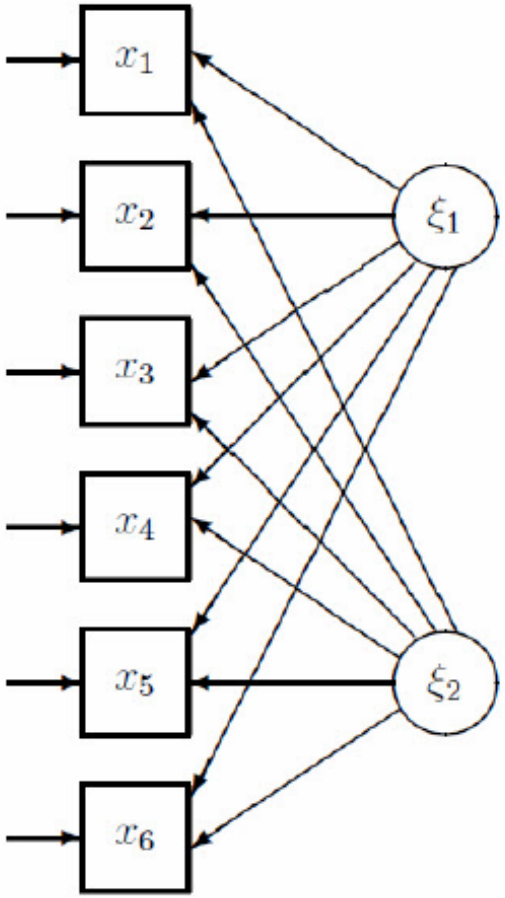


Figure 1: Exploratory Factor Analysis

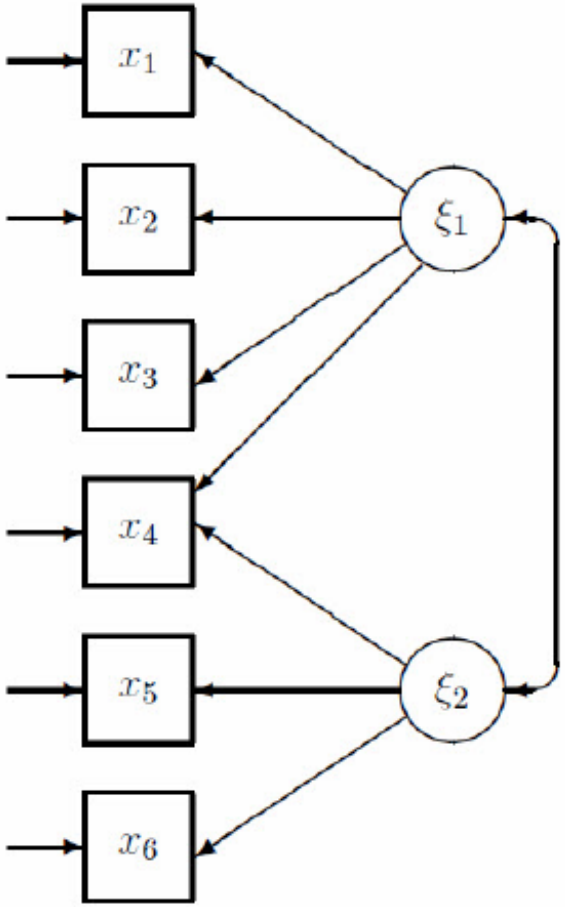
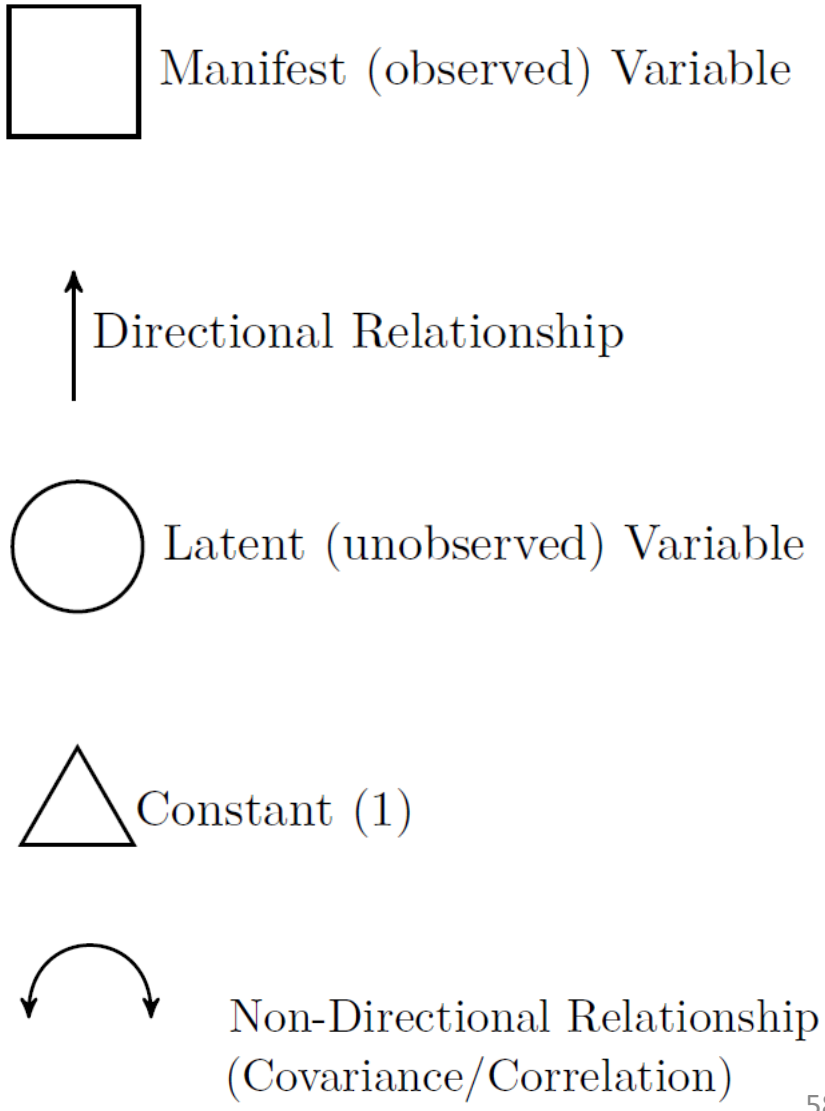


Figure 2: Confirmatory Factor Analysis



Confirmatory Factor Analysis (CFA) vs Exploratory Factor Analysis (EFA)

EFA	CFA
theory development	theory testing
no. of factors not fixed	fixed no. of factors
orthogonal factors	usually correlated
rotation	not necessary
variables load on all factors	load on specific factors

Use CFA for

What is the null hypothesis?

- testing single model (strictly confirmatory)
- comparing alternative models

Section 2: Confirmatory Factor Analysis and Structural Equation Models

Confirmatory Factor Analysis (CFA)

An Example: Subjective Well Being (SWB) Model

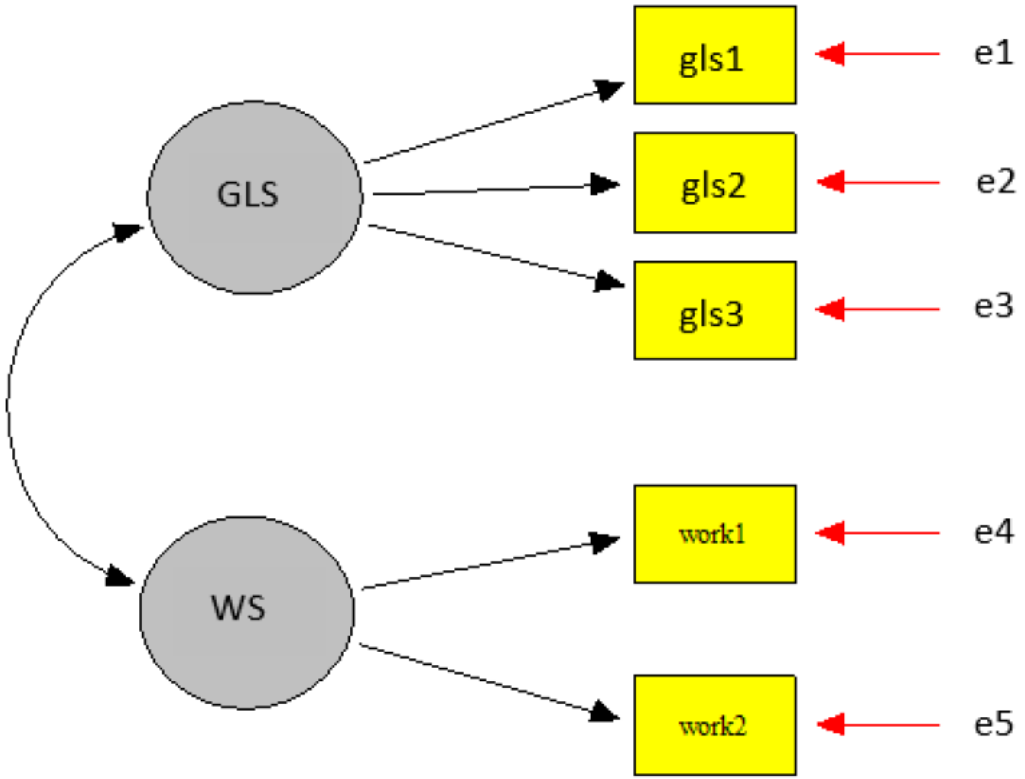
主观幸福感

- To examine the hypothesis that subjective well being is a multidimensional construct composed of **general life satisfaction (GLS)** and **work-related satisfaction (WS)**
- Data: 5 variables were measured in a sample of size 500

	V1	V2	V3	V4	V5
V1 (gls1)	198				
V2 (gls2)	82	86			
V3 (gls3)	54	28	24		
V4 (work1)	52	30	18	151	
V5 (work2)	16	10	7	44	28

Confirmatory Factor Analysis (CFA)

An Example: Subjective Well Being (SWB) Model



Path diagrams

F_1=GLS, F_2=WS

$$
\begin{aligned}
\text{gls1} &= V_1 = \mu_1 + \lambda_{11} F_1 + e_1 \\
\text{gls2} &= V_2 = \mu_2 + \lambda_{21} F_1 + e_2 \\
\text{gls3} &= V_3 = \mu_3 + \lambda_{31} F_1 + e_3 \\
\text{work1} &= V_4 = \mu_4 + \lambda_{42} F_2 + e_4 \\
\text{work2} &= V_5 = \mu_5 + \lambda_{52} F_2 + e_5
\end{aligned}
$$

Confirmatory Factor Analysis (CFA)

for SWB model $\Lambda = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \end{bmatrix}$

Matrix Form

$$E(F) = 0, \quad \mathrm{Var}(F) = I_m,$$

$$v = \mu + \Lambda f + e$$

v	is $p \times 1$ vector of observed variables
μ	is $p \times 1$ vector of intercepts (means of v)
Λ	is $p \times k$ factor loading matrix
f	is $k \times 1$ vector of latent factors
e	is $p \times 1$ vector of measurement errors

$$\Sigma = \mathrm{E}[(v\text{-}\mu)(v\text{-}\mu)'] = \Lambda\Psi\Lambda' + \Theta$$

Covariance matrix of observed variable

Estimate the unknown parameters Λ, Ψ, Θ

Assumption

$\mathrm{E}(e) = 0 \qquad Var(e) = \Theta\ (= diag(\sigma_1^2, \cdots, \sigma_p^2))$

Means of **errors** are **zero** and errors are (usually **uncorrelated** of each other)

$\mathrm{E}(f) = 0 \qquad Var(f) = \Psi$

Means of **Factors** are **zero,** Ψ is a **general covariance matrix**

$\geq$ not orthogonal.

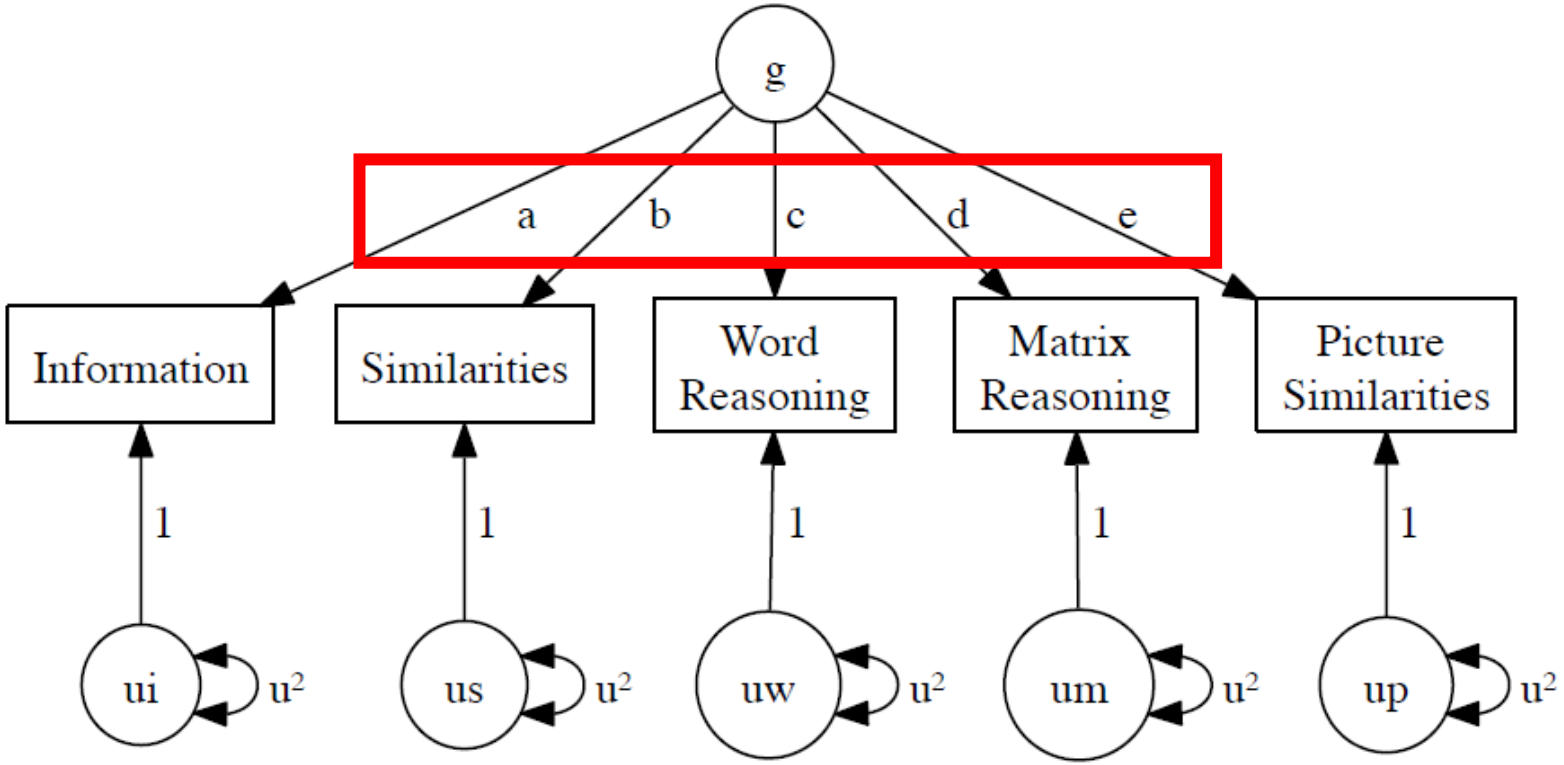
the only difference with the assumption of EFA.

$\mathrm{E}(fe') = 0$

Common factors and errors are uncorrelated

Confirmatory Factor Analysis (CFA)

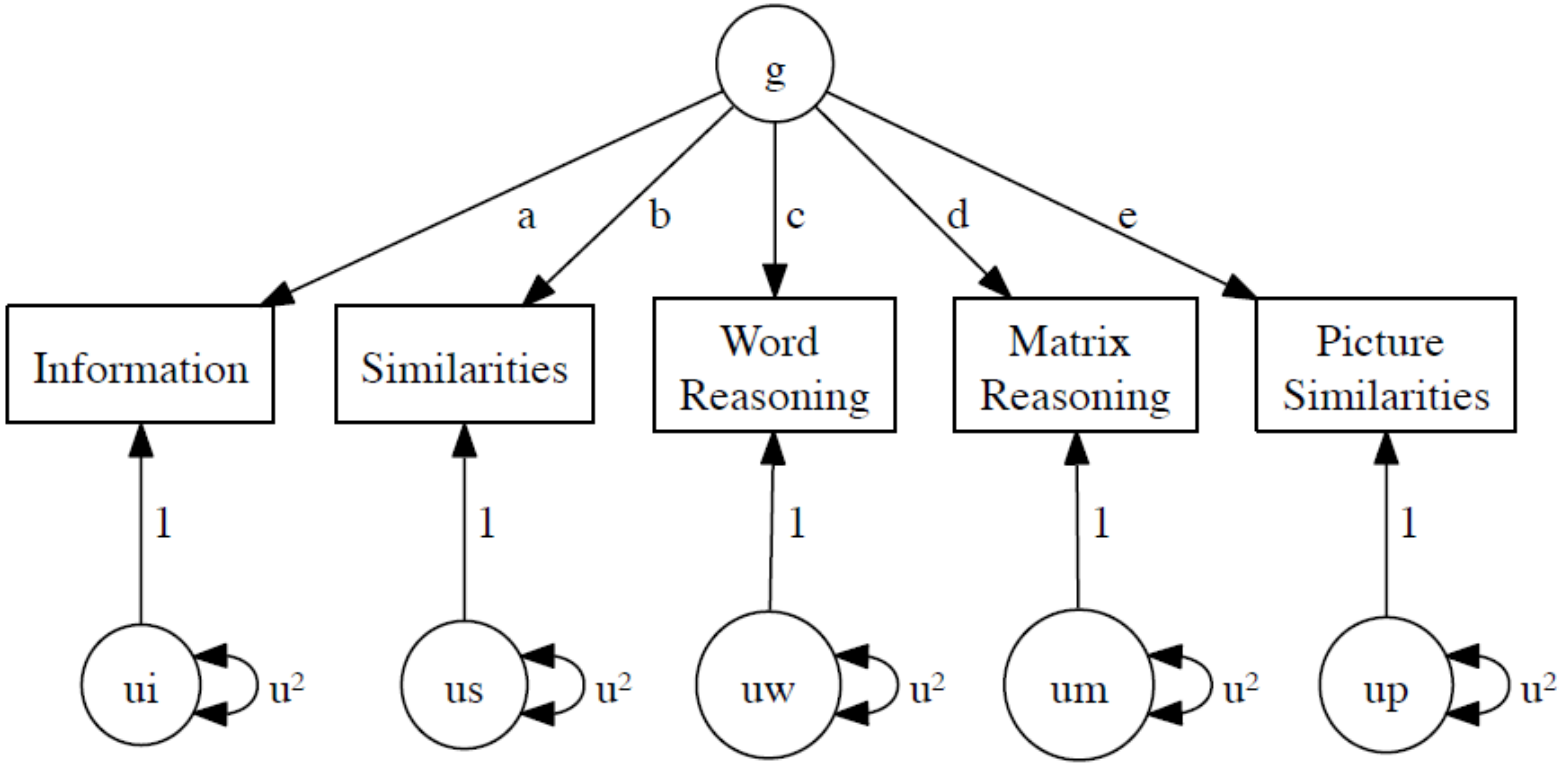
Wechsler Intelligence Scale for Children-Fourth Edition subscales



- The amount that common factors influence observed variable is measured by factor loadings
- a, b, c, d and e are all factor loadings.

Confirmatory Factor Analysis (CFA)

Wechsler Intelligence Scale for Children-Fourth Edition subscales



Assume $Var(g) = 1$

$$Communality$$

$$h_1^2 = a^2$$

$$h_2^2 = b^2$$

$$h_3^2 = c^2$$

$$\dots$$

$$Uniqueness$$

$$u_1^2 = 1 - a^2$$

$$u_2^2 = 1 - b^2$$

$$\dots$$

Confirmatory Factor Analysis (CFA)

Wechsler Intelligence Scale for Children-Fourth Edition subscales

Correlations for the WISC-IV data

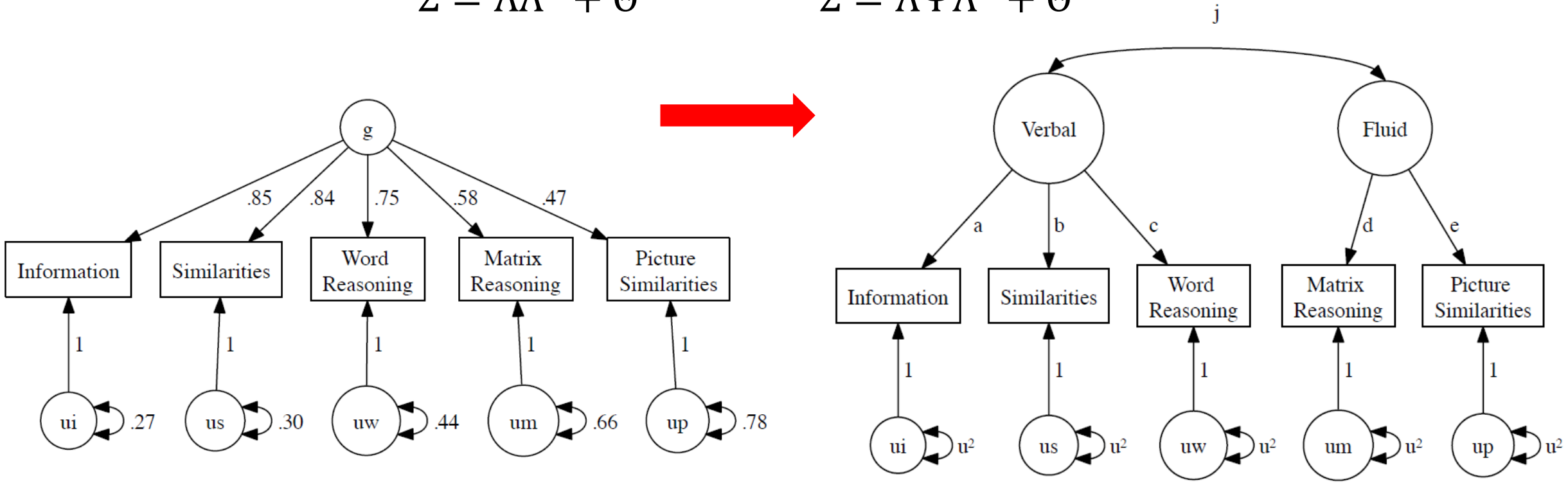
	Info	Sim	Word Reas	Matrix Reas	Picture Sim
inss	1.00	0.72	0.64	0.51	0.37
siss	0.72	1.00	0.63	0.48	0.38
wrss	0.64	0.63	1.00	0.37	0.38
mrss	0.51	0.48	0.37	1.00	0.38
psss	0.37	0.38	0.38	0.38	1.00

```
> fa(R, nfactors=1,rotate="none",n.obs=550,fm="mle")
Factor Analysis using method =  ml
Call: fa(r = R, nfactors = 1, n.obs = 550, rotate = "none", f
Standardized loadings (pattern matrix) based upon correlation
        ML1   h2   u2 com
Info   0.86 0.74 0.26   1
Sim    0.84 0.70 0.30   1
Word   0.74 0.55 0.45   1
Matrix 0.58 0.33 0.67   1
Pict   0.47 0.22 0.78   1

                ML1
SS loadings    2.55
Proportion Var 0.51
```

Confirmatory Factor Analysis (CFA)

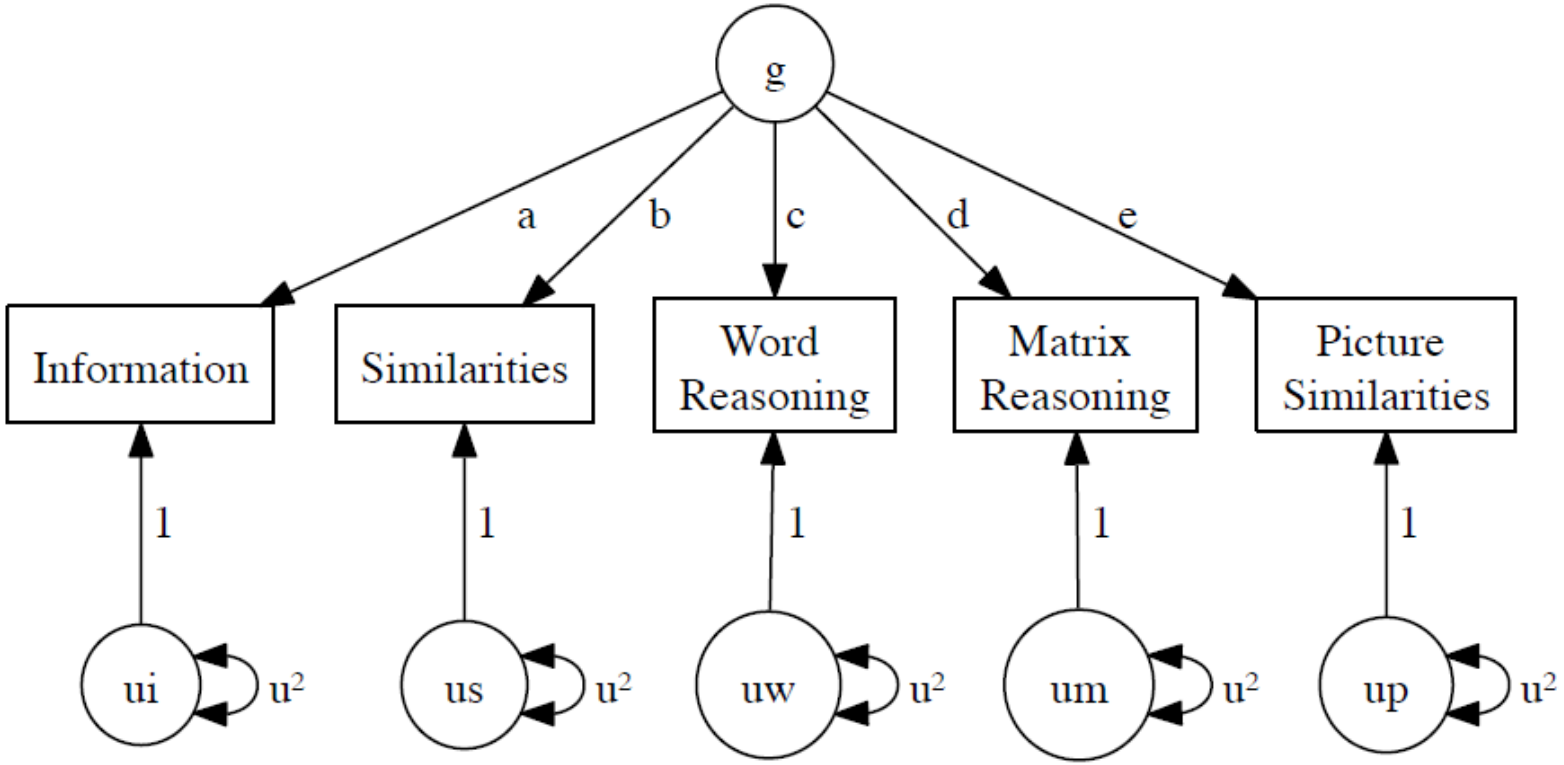
Wechsler Intelligence Scale for Children-Fourth Edition subscales

$$\Sigma = \Lambda\Lambda^T + \Theta \qquad \Sigma = \Lambda\Psi\Lambda^T + \Theta$$



Confirmatory Factor Analysis (CFA)

Wechsler Intelligence Scale for Children-Fourth Edition subscales



Identification

Assume $Var(g) = 2$

$$Communality$$

$$h_1^2 = a'^2 = 2a^2$$

$$h_2^2 = b'^2 = 2b^2$$

$$h_3^2 = c'^2 = 2c^2$$

$$\dots$$

$$Uniqueness$$

$$u_1^2 = 1 - a'^2$$

$$u_2^2 = 1 - b'^2$$

$$\dots$$

Confirmatory Factor Analysis (CFA)

Matrix Form

$$E(F) = 0, \quad \mathrm{Var}(F) = I_m,$$

$$v = \mu + \Lambda f + e$$

v	is $p \times 1$ vector of observed variables
μ	is $p \times 1$ vector of intercepts (means of v)
Λ	is $p \times k$ factor loading matrix
f	is $k \times 1$ vector of latent factors
e	is $p \times 1$ vector of measurement errors

Identification

Let $\quad \Lambda^* = \Lambda\, \mathrm{D}$
$\Psi^* = \mathrm{D}^{-1}\Psi\mathrm{D}^{-1'}$
$\Theta^* = \Theta$

(D is an arbitrary $k \times k$ square matrix such that $DD^{-1} = I$)

Then $\quad \Lambda^*\Psi^*\Lambda^{*'}+\Theta^* = \Lambda\Psi\Lambda' + \Theta = \Sigma$

- The parameters cannot be **uniquely determined** even Σ is known. This is the **identification problem** or factor indeterminacy problem in CFA
- That means every model **parameter has to be uniquely solved** in terms of the population variances and covariance of the observed variables

Confirmatory Factor Analysis (CFA)

Latent Variable's Scale

Identification

Because Latent Variables are not directly observed, there are no inherent units by which to measure them. Consequently, the model is not identified unless some parameter estimates are constrained to set the latent variable's scale. There are two common ways to set this scale.

1. **Standardized latent variable.** This method constrains the **latent variable's variance to 1.0**. This, in effect, makes the latent variable a standardized variable. Moreover, if there is more than one Latent Variables, then the covariance among the Latent Variables becomes a **correlation**.
2. **Marker variable.** This method requires a single factor loading for each the latent variable be **constrained to an arbitrary value (usually 1.0).** The indicator variable whose loading is constrained is called the marker variable. This method uses the marker variable to define the LV's variance.

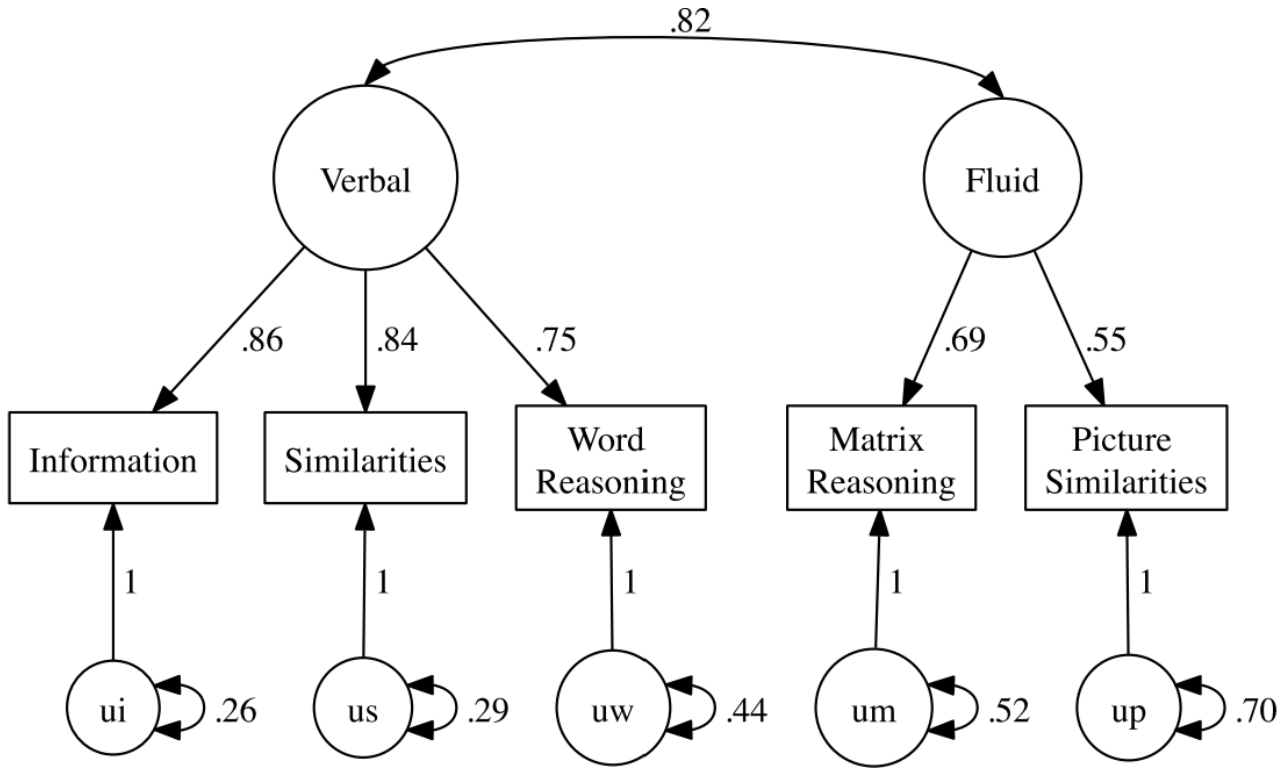
Confirmatory Factor Analysis (CFA)

Wechsler Intelligence Scale for Children-Fourth Edition subscales

$$v = \mu + \Lambda f + e \qquad Var(f) = \Psi$$

$$\underset{(5\times 2)}{\Lambda} = \begin{bmatrix} .86 & 0 \\ .84 & 0 \\ .75 & 0 \\ 0 & .69 \\ 0 & .55 \end{bmatrix}, \; \& \; \underset{(2\times 2)}{\Psi} = \begin{bmatrix} 1 & .82 \\ .82 & 1 \end{bmatrix}$$

non-overlapping loadings.



Confirmatory Factor Analysis (CFA)

lavaan

R package

LAtent VAriable ANalysis

To compute a model in lavaan requires two steps:

1. **specify the path model**
2. analyze the model.

Specification commands

Syntax	Command	Example
~	Regress onto	Regress B onto A: `B ~ A` → ?
~~	(Co)variance	Variance of A: `A ~~ A` Covariance of A and B: `A ~~ B`

~1	Constant/mean/intercept	Regress B onto A, and and include the intercept in the model: `B ~ 1 + A` or
		`B ~ A`
		`B ~ 1`

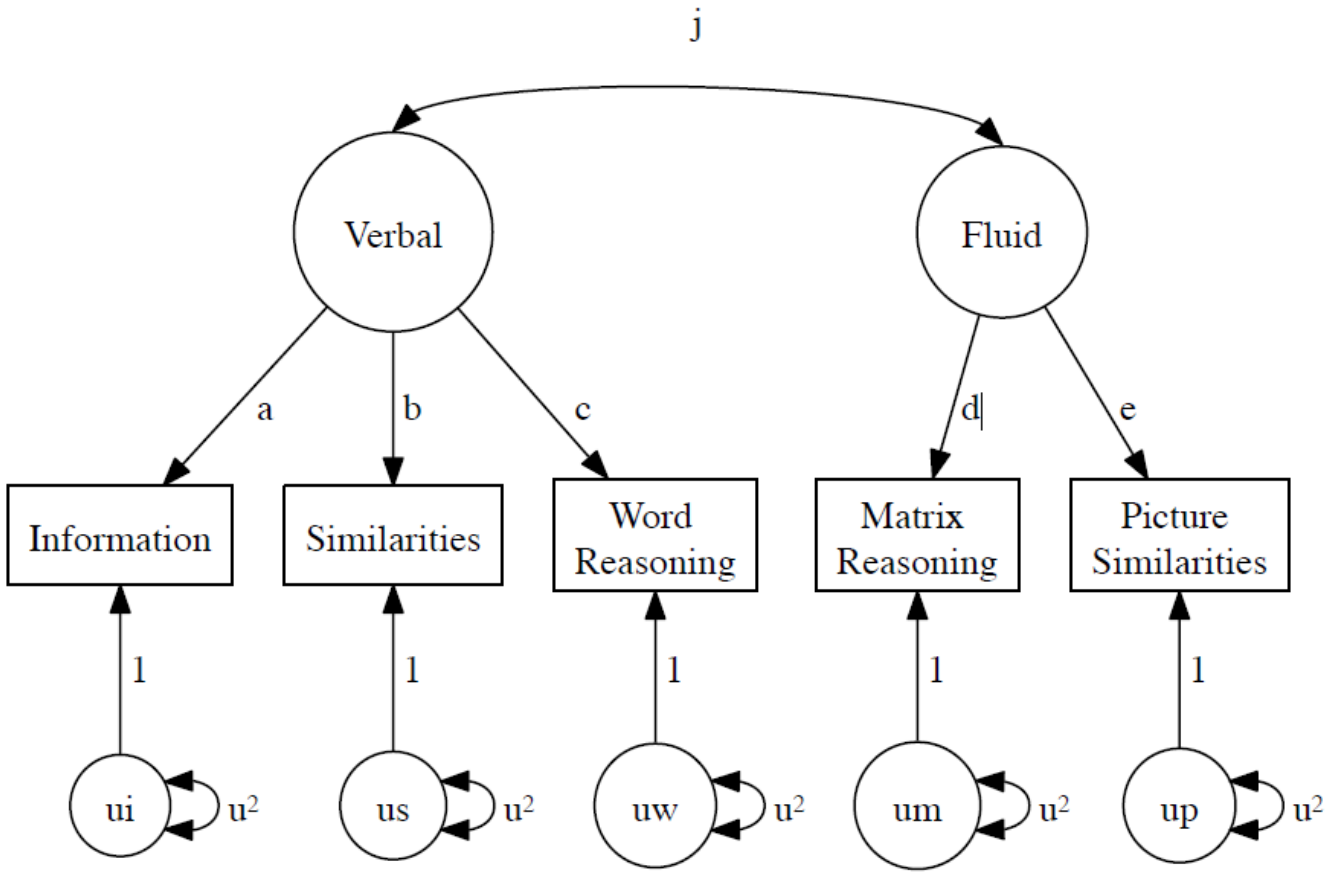
=~	Define reflective latent variable	Define Factor 1 by A-D: `F1 =~A+B+C+D`
<~	Define formative latent variable	Define Factor 1 by A-D: `F1 <~ 1*A+B+C+D`
:=	Define non-model parameter	Define parameter u2 to be twice the square of u: `u2 := 2*(u^2)`
*	Label parameters (the label has to be pre-multiplied)	Label the regression of Z onto X as b: `Z ~b*X`
|	Define the number of thresholds (for categorical endogenous variables)	Variable u has three thresholds: `u | t1 + t2 + t3`

In reflective models, the indicators are affected by the latent variable, whereas in formative models the indicators define the latent variable. Reflective indicators are interchangeable since the concept is reflected in the different indicators, sharing a common theme.

What's the different?

a variable relies on other variable.

Confirmatory Factor Analysis (CFA)




```
model3 <- '
# Measurement model
Verbal=~ a*Info + b*Sim + c*Word
Fluid =~ d*Matrix + e*Pict
# error Variance and Covariance (psi)
Verbal ~~ Fluid
'
```

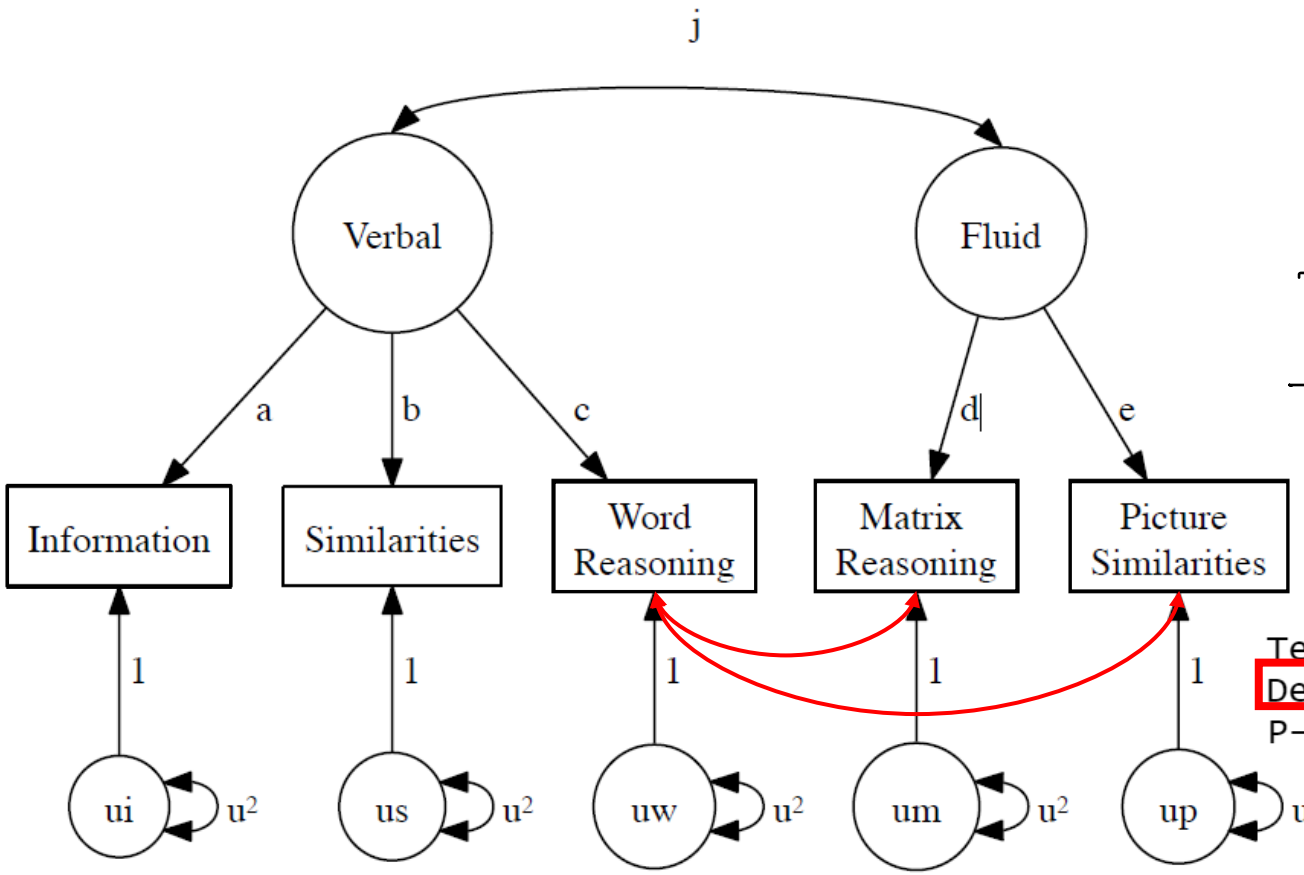
```
model31 <- '
# Measurement model
Verbal=~ NA*Info + b*Sim + c*Word
Fluid =~ NA*Matrix + e*Pict
# error Variance and Covariance (psi)
Verbal ~~ 1*Verbal
Fluid ~~ 1*Fluid
Verbal ~~ Fluid
'
```

Confirmatory Factor Analysis (CFA)

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Verbal =~
    Info              0.859    0.036   23.613    0.000
    Sim        (b)    0.840    0.037   22.867    0.000
    Word       (c)    0.742    0.038   19.303    0.000
  Fluid =~
    Matrix            0.688    0.049   13.920    0.000
    Pict       (e)    0.551    0.047   11.699    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  Verbal ~~
    Fluid             0.823    0.043   19.280    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    Verbal            1.000
    Fluid             1.000
   .Info              0.260    0.028    9.295    0.000
   .Sim               0.292    0.028   10.282    0.000
   .Word              0.447    0.033   13.555    0.000
   .Matrix            0.524    0.055    9.557    0.000
   .Pict              0.695    0.051   13.673    0.000
```

```
Model Test User Model:

  Test statistic                                12.687
  Degrees of freedom                                4
  P-value (Chi-square)                          0.013

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Verbal =~
    Info       (a)    1.000
    Sim        (b)    0.978    0.045   21.625    0.000
    Word       (c)    0.864    0.046   18.958    0.000
  Fluid =~
    Matrix     (d)    1.000
    Pict       (e)    0.801    0.082    9.747    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  Verbal ~~
    Fluid             0.487    0.046   10.604    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .Info              0.260    0.028    9.295    0.000
   .Sim               0.292    0.028   10.282    0.000
   .Word              0.447    0.033   13.555    0.000
   .Matrix            0.524    0.055    9.557    0.000
   .Pict              0.695    0.051   13.673    0.000
    Verbal            0.739    0.063   11.807    0.000
    Fluid             0.474    0.068    6.960    0.000
```

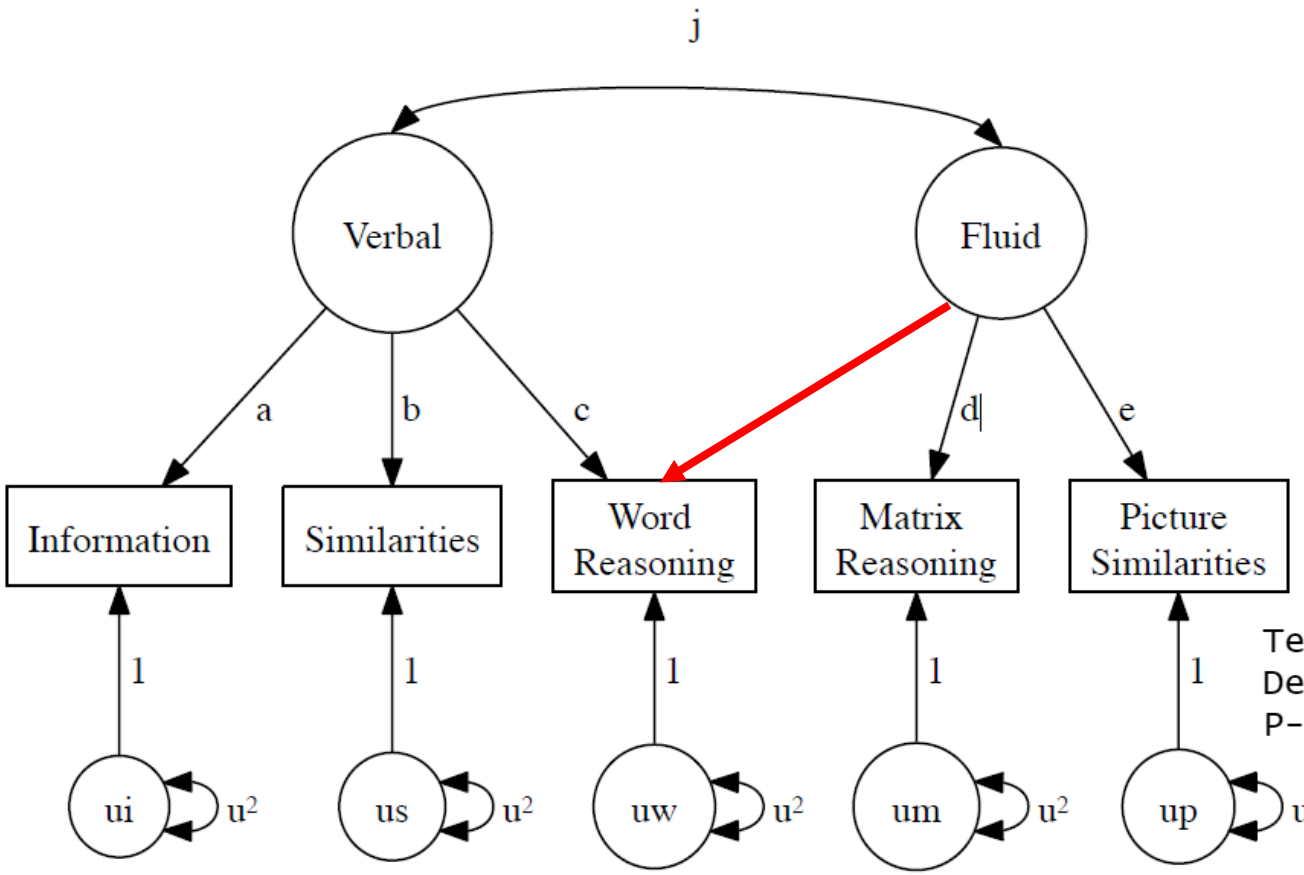
Confirmatory Factor Analysis (CFA)




```
model32 <- '
# Measurement model
Verbal=~ NA*Info + b*Sim + c*Word
Fluid =~ NA*Matrix + e*Pict
# error Variance and Covariance (psi)
Verbal ~~ 1*Verbal
Fluid ~~ 1*Fluid
Verbal ~~ Fluid
# error covariances
Word ~~ Matrix
Word ~~ Pict
'
```

```
Test statistic                                1.117
Degrees of freedom                            2
P-value (Chi-square)                          0.572
```

Confirmatory Factor Analysis (CFA)




```
model34 <- '
# Measurement model
Verbal=~ NA*Info + b*Sim + c*Word
Fluid =~ NA*Matrix + e*Pict + f*Word
# error Variance and Covariance (psi)
Verbal ~~ 1*Verbal
Fluid ~~ 1*Fluid
Verbal ~~ Fluid
'
```

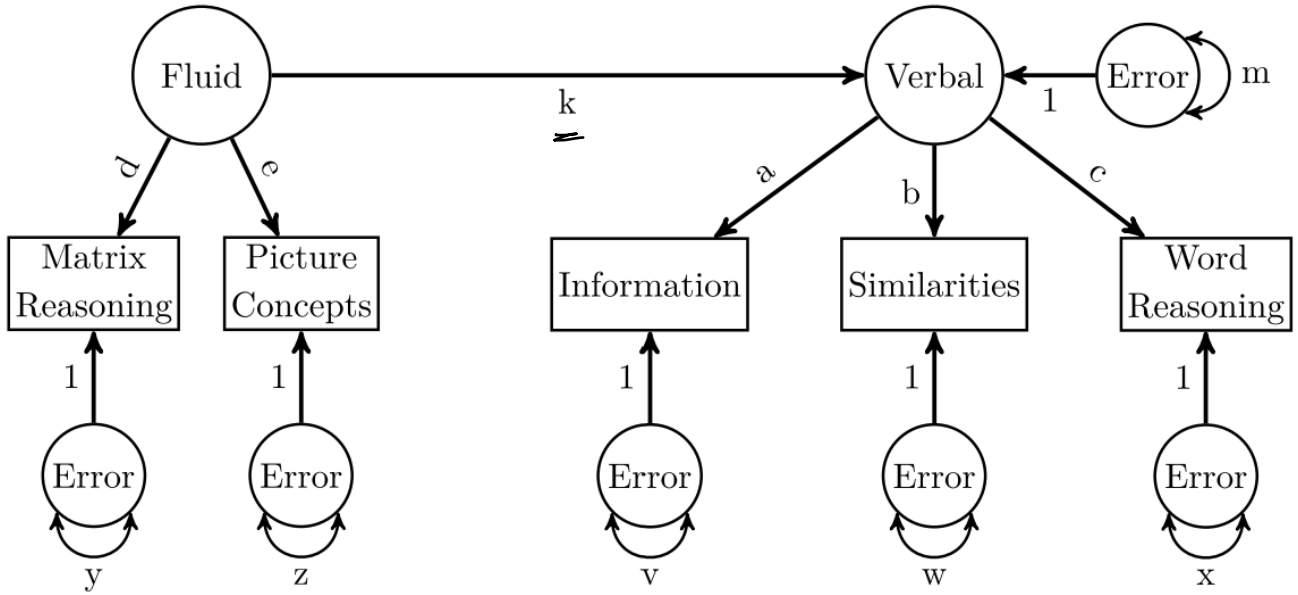
Test statistic	11.612
Degrees of freedom	3
P-value (Chi-square)	0.009

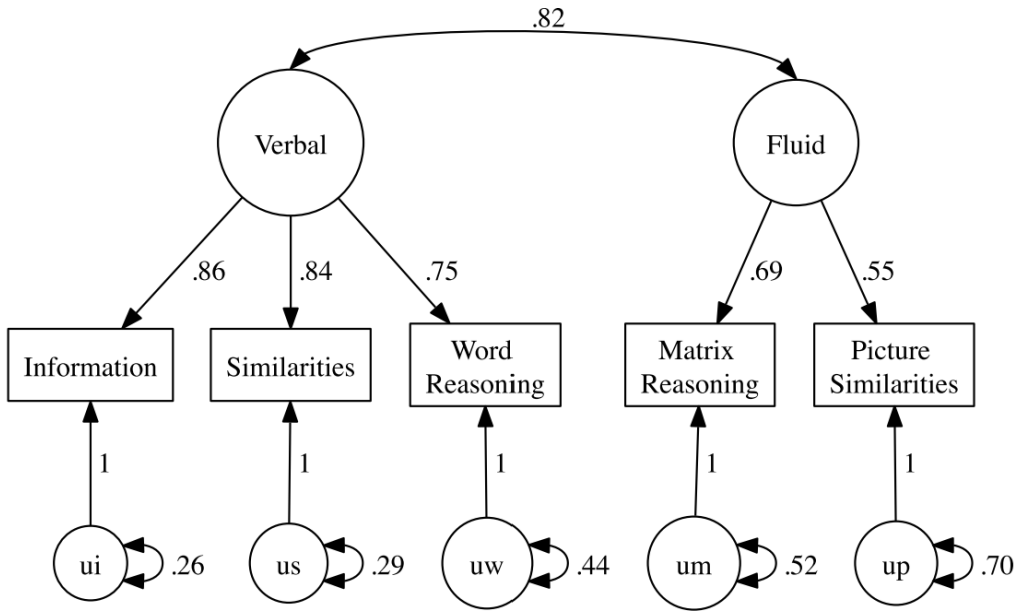
Confirmatory Factor Analysis (CFA)

Wechsler Intelligence Scale for Children-Fourth Edition subscales

Regression $Verbal = k * Fluid + error$

Structural Equation Model





Confirmatory Factor Analysis (CFA)

The correlations from McIver, Carmines, and Zeller's (1980) study of attitudes toward police. They are based on telephone interviews with a total of some 11,000 respondents in 60 neighborhoods in three U.S. metropolitan areas..

General dimension of attitude toward police

Likelihood of burglary, vandalism, robbery in the neighborhood

	1	2	3	4	5	6	7	8	9
1. Police service	1.00	.50	.41	.33	.28	.30	−.24	−.23	−.20
2. Responsiveness		1.00	.35	.29	.26	.27	−.19	−.19	−.18
3. Response time			1.00	.30	.27	.29	−.17	−.16	−.14
4. Honesty				1.00	.52	.48	−.13	−.11	−.15
5. Courtesy					1.00	.44	−.11	−.09	−.10
6. Equal treatment						1.00	−.15	−.13	−.13
7. Burglary							1.00	.58	.47
8. Vandalism								1.00	.42
9. Robbery									1.00

Confirmatory Factor Analysis (CFA)

The correlations from McIver, Carmines, and Zeller's (1980) study of attitudes toward police. They are based on telephone interviews with a total of some 11,000 respondents in 60 neighborhoods in three U.S. metropolitan areas..

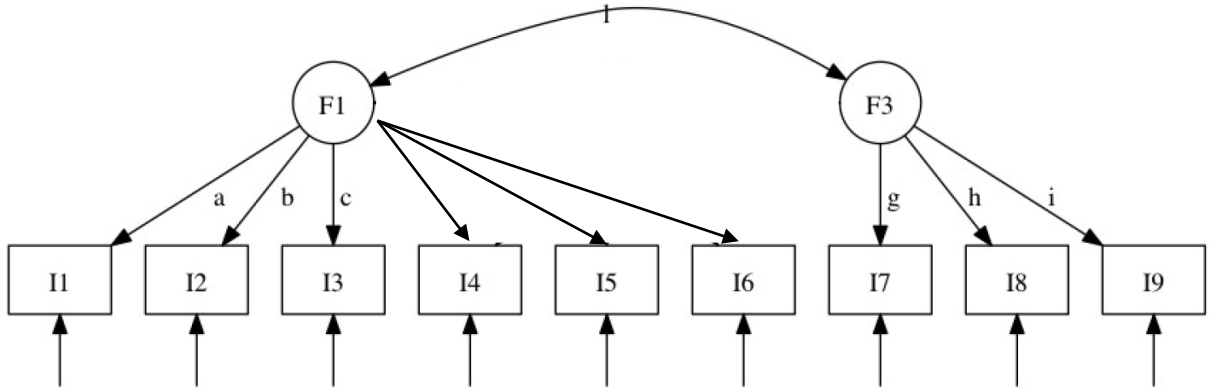
Reflecting attitudes toward the quality of police services

Likelihood of burglary, vandalism, robbery in the neighborhood

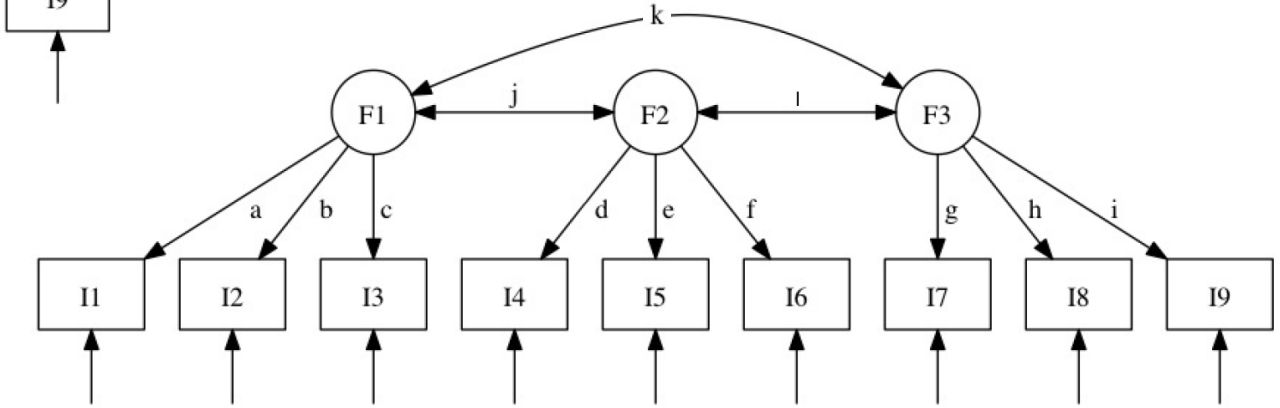
	1	2	3	4	5	6	7	8	9
1. Police service	1.00	.50	.41	.33	.28	.30	−.24	−.23	−.20
2. Responsiveness		1.00	.35	.29	.26	.27	−.19	−.19	−.18
3. Response time			1.00	.30	.27	.29	−.17	−.16	−.14
4. Honesty				1.00	.52	.48	−.13	−.11	−.15
5. Courtesy					1.00	.44	−.11	−.09	−.10
6. Equal treatment						1.00	−.15	−.13	−.13
7. Burglary							1.00	.58	.47
8. Vandalism								1.00	.42
9. Robbery									1.00

Personal qualities of the police

Confirmatory Factor Analysis (CFA)



VS



Confirmatory Factor Analysis (CFA)

$\Lambda\Psi\Lambda^T +$

```
> residuals(p1.1.fit)
$type
[1] "raw"

$cov
   PS     RE     RT     HO     CO     ET     BU     VA     RO
PS  0.000
RE  0.173  0.000
RT  0.094  0.062  0.000
HO -0.069 -0.074 -0.052  0.000
CO -0.086 -0.074 -0.052  0.112  0.000
ET -0.068 -0.066 -0.034  0.070  0.064  0.000
BU -0.072 -0.037 -0.022  0.057  0.061  0.022  0.000
VA -0.079 -0.052 -0.027  0.059  0.065  0.025  0.003  0.000
RO -0.076 -0.067 -0.031 -0.012  0.026 -0.003 -0.002 -0.006  0.000
```

some cors are large

3-factor model is better.

```
> residuals(p1.fit)
$type
[1] "raw"

$cov
   PS     RE     RT     HO     CO     ET     BU     VA     RO
PS  0.000
RE  0.016  0.000
RT -0.009 -0.019  0.000
HO -0.015 -0.013  0.038  0.000
CO -0.033 -0.015  0.032  0.009  0.000
ET  0.001  0.007  0.063 -0.008 -0.003  0.000
BU  0.000  0.021  0.013  0.013  0.019 -0.026  0.000
VA -0.011  0.002  0.006  0.020  0.028 -0.017  0.003  0.000
RO -0.022 -0.023 -0.005 -0.044 -0.004 -0.038  0.000 -0.007  0.000
```

Confirmatory Factor Analysis (CFA)

Item	Factor pattern F_1	F_2	F_3	h^2
1. Police service	.74	.00	.00	.55
2. Responsiveness	.65	.00	.00	.43
3. Response time	.56	.00	.00	.32
4. Honesty	.00	.75	.00	.56
5. Courtesy	.00	.68	.00	.46
6. Equal treatment	.00	.65	.00	.42
7. Burglary	.00	.00	.80	.63
8. Vandalism	.00	.00	.72	.52
9. Robbery	.00	.00	.59	.35

Factor correlations

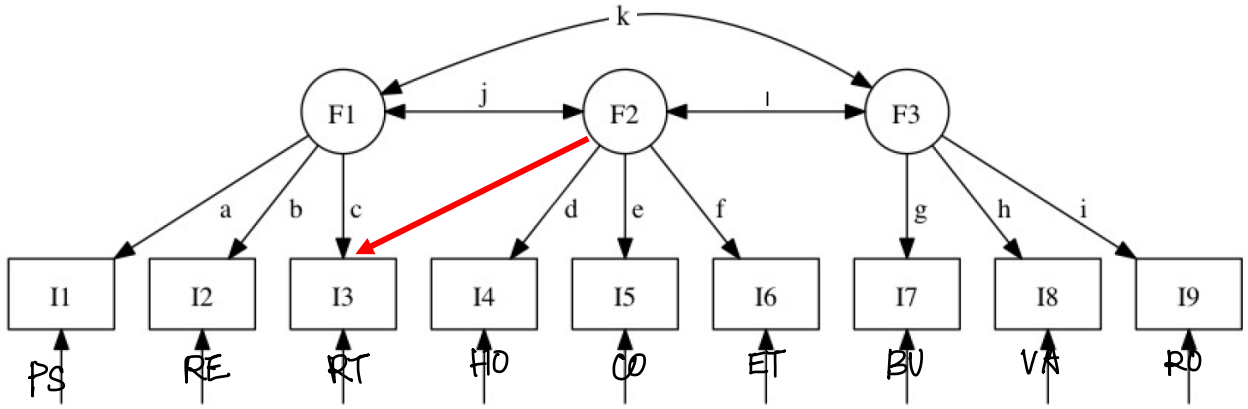
	F_1	F_2	F_3
F_1	1.00	.62	−.41
F_2		1.00	−.24
F_3			1.00

```
> modificationindices(p1.fit,sort=T)
   lhs op rhs      mi    epc sepc.lv sepc.all sepc.nox
33  F2 =~  RT 100.693  0.160   0.160    0.160    0.160
43  PS ~~  RE  69.015  0.102   0.102    0.201    0.201
31  F2 =~  PS  43.204 -0.122  -0.122   -0.122   -0.122
36  F2 =~  RO  31.732 -0.057  -0.057   -0.057   -0.057
68  HO ~~  RO  30.842 -0.036  -0.036   -0.068   -0.068
60  RT ~~  ET  29.191  0.038   0.038    0.061    0.061
```

```
> residuals(p1.fit)
$type
[1] "raw"

$cov
    PS     RE     RT     HO     CO     ET     BU     VA     RO
PS  0.000
RE  0.016  0.000
RT -0.009 -0.019  0.000
HO -0.015 -0.013  0.038  0.000
CO -0.033 -0.015  0.032  0.009  0.000
ET  0.001  0.007  0.063 -0.008 -0.003  0.000
BU  0.000  0.021  0.013  0.013  0.019 -0.026  0.000
VA -0.011  0.002  0.006  0.020  0.028 -0.017  0.003  0.000
RO -0.022 -0.023 -0.005 -0.044 -0.004 -0.038  0.000 -0.007  0.000
```

Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis




```
modelp2 <- '
# Measurement model
F1 =~ a*PS + b*RE + c*RT
F2 =~ d*HO + e*CO + f*ET + RT
F3 =~ g*BU + h*VA + i*RO
# error Variance and Covariance (psi)
F1 ~~ j*F2
F1 ~~ k*F3
F2 ~~ l*F3
'
p2.fit=sem(modelp2,sample.cov= police.cor,sample.nobs =11000,std.lv=T)
```

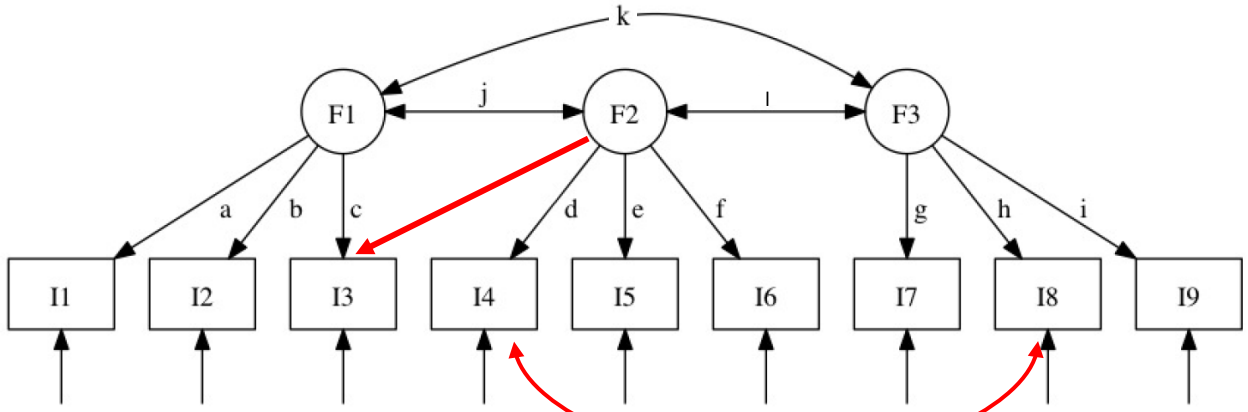
Why no coeff

```
Latent Variables:
                   Estimate
  F1 =~
    PS        (a)     0.760
    RE        (b)     0.657
    RT        (c)     0.451
  F2 =~
    HO        (d)     0.749
    CO        (e)     0.681
    ET        (f)     0.651
    RT                0.148
  F3 =~
    BU        (g)     0.796
    VA        (h)     0.725
    RO        (i)     0.590

Covariances:
                   Estimate
  F1 ~~
    F2        (j)     0.582
    F3        (k)    -0.404
  F2 ~~
    F3        (l)    -0.239
```

```
> anova(p1.fit,p2.fit)
Chi-Squared Difference Test

       Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
p2.fit 23 256978 257139 127.32
p1.fit 24 257075 257228 226.23     98.915       1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis



Honesty ~~ Robbery

```
> modificationindices(p2.fit,sort=T)
   lhs op rhs     mi    epc sepc.lv sepc.all sepc.nox
68  HO ~~  RO 31.322 -0.036  -0.036   -0.068   -0.068
36  F2 =~  RO 31.013 -0.056  -0.056   -0.057   -0.057
64  HO ~~  CO 30.481  0.067   0.067    0.138    0.138
42  F3 =~  ET 28.044 -0.053  -0.053   -0.053   -0.053
28  F1 =~  ET 24.836  0.074   0.074    0.074    0.074
76  BU ~~  VA 19.161  0.101   0.101    0.241    0.241
```

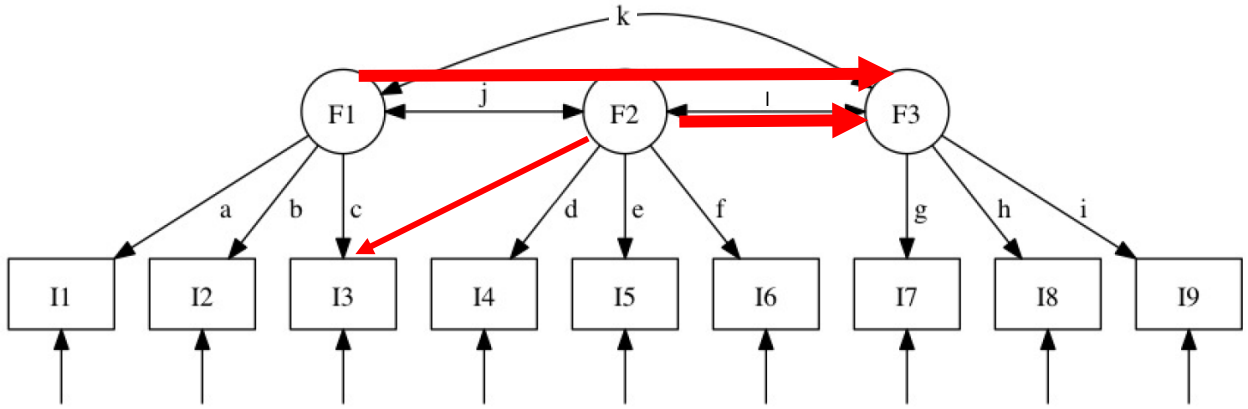
```
modelp2.1 <- '
# Measurement model
F1 =~ a*PS + b*RE + c*RT
F2 =~ d*HO + e*CO + f*ET + RT
F3 =~ g*BU + h*VA + i*RO
# error Variance and Covariance (psi)
F1 ~~ j*F2
F1 ~~ k*F3
F2 ~~ l*F3
#covariance
HO ~~ RO
'
```

```
Covariances:
                   Estimate
  F1 ~~
    F2         (j)    0.582
    F3         (k)   -0.405
  F2 ~~
    F3         (l)   -0.234
 .HO ~~
   .RO              -0.036
```

Confirmatory Factor Analysis (CFA)

Structural Equation Model

$CFA + Regression$




```
modelp3 <- '
# Measurement model
F1 =~ a*PS + b*RE + c*RT
F2 =~ d*HO + e*CO + f*ET + RT
F3 =~ g*BU + h*VA + i*RO
# error Variance and Covariance (psi)
F1 ~~ j*F2
#Structural model
F3 ~ k*F1+l*F2
# error Variance
F3 ~~ m*F3
'
```

causual structure.
but in CFA, there is only correlation relationship.

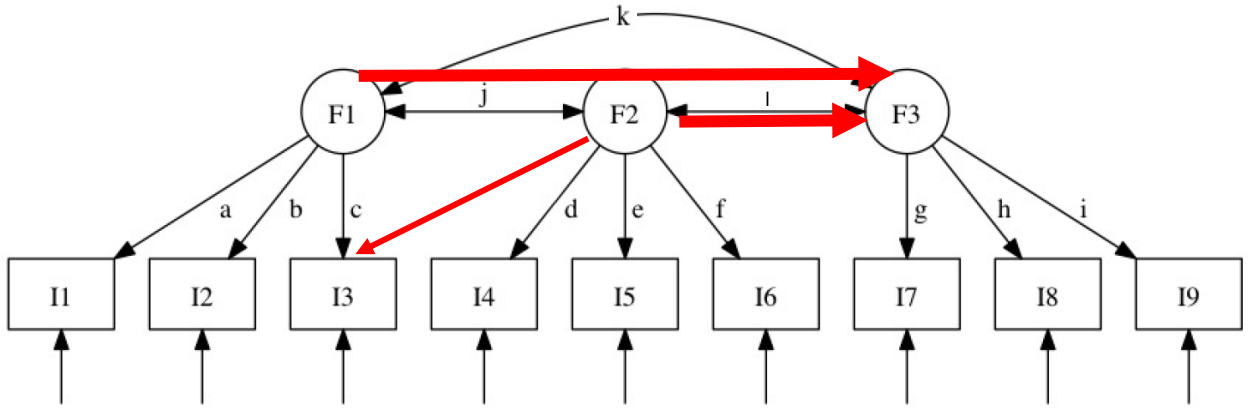
Reflecting attitudes toward the quality of police services

Likelihood of burglary, vandalism, robbery in the neighborhood

Personal qualities of the police

Confirmatory Factor Analysis (CFA)

Structural Equation Model



Reflecting attitudes toward the quality of police services

Likelihood of burglary, vandalism, robbery in the neighborhood

Personal qualities of the police

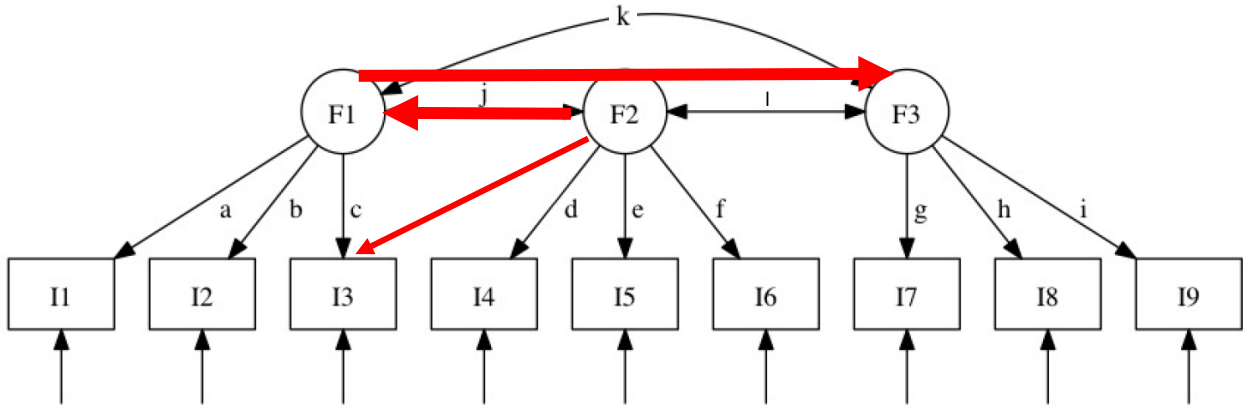
$CFA + Regression$

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 =~
    PS        (a)     1.000
    RE        (b)     0.864    0.018   49.141    0.000
    RT        (c)     0.593    0.021   28.017    0.000
  F2 =~
    HO        (d)     1.000
    CO        (e)     0.908    0.017   55.034    0.000
    ET        (f)     0.869    0.016   53.810    0.000
    RT                0.197    0.019   10.325    0.000
  F3 =~
    BU        (g)     1.000
    VA        (h)     0.910    0.017   54.714    0.000
    RO        (i)     0.741    0.015   50.410    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  F3 ~
    F1        (k)    -0.420    0.020  -21.243    0.000
    F2        (l)    -0.006    0.018   -0.304    0.761

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 ~~
    F2        (j)     0.332    0.009   36.520    0.000
```

Confirmatory Factor Analysis (CFA)

Structural Equation Model

$CFA + Regression$



Reflecting attitudes toward the quality of police services

Likelihood of burglary, vandalism, robbery in the neighborhood

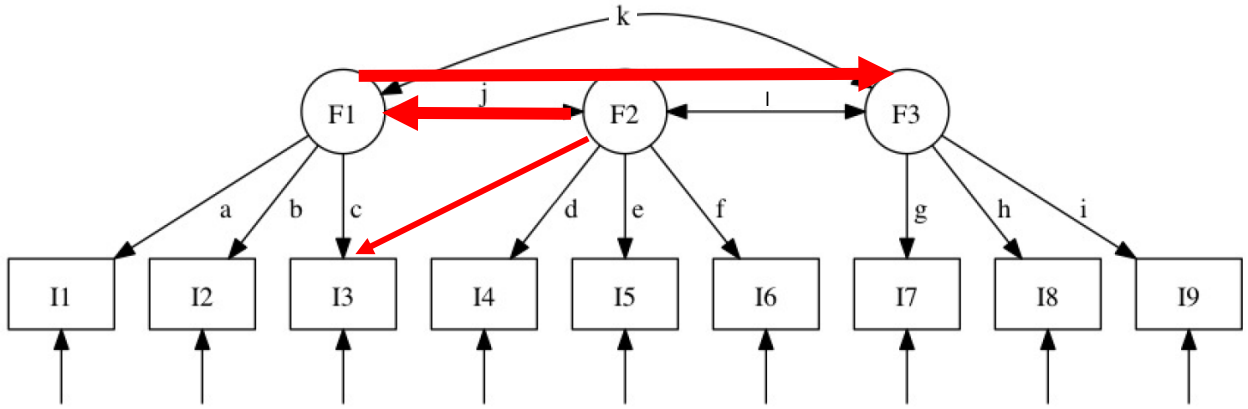
Personal qualities of the police

```
modelp3.1 <- '
# Measurement model
F1 =~ a*PS + b*RE + c*RT
F2 =~ d*HO + e*CO + f*ET + RT
F3 =~ g*BU + h*VA + i*RO
#Structural model
F1 ~ j*F2
F3 ~ k*F1+l*F2
#define a new parameter
ind := j*k
# error Variance
F3 ~~ m*F3
F1 ~~ n*F1
'
```

Confirmatory Factor Analysis (CFA)

Structural Equation Model

$CFA + Regression$




```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 ~
    F2        (j)     0.591    0.015   40.155    0.000
  F3 ~
    F1        (k)    -0.420    0.020  -21.243    0.000
    F2        (l)    -0.006    0.018   -0.304    0.761

Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)
    ind              -0.248    0.013  -19.197    0.000
```

Reflecting attitudes toward the quality of police services

Likelihood of burglary, vandalism, robbery in the neighborhood

Personal qualities of the police

Confirmatory Factor Analysis (CFA)

Structural Equation Model

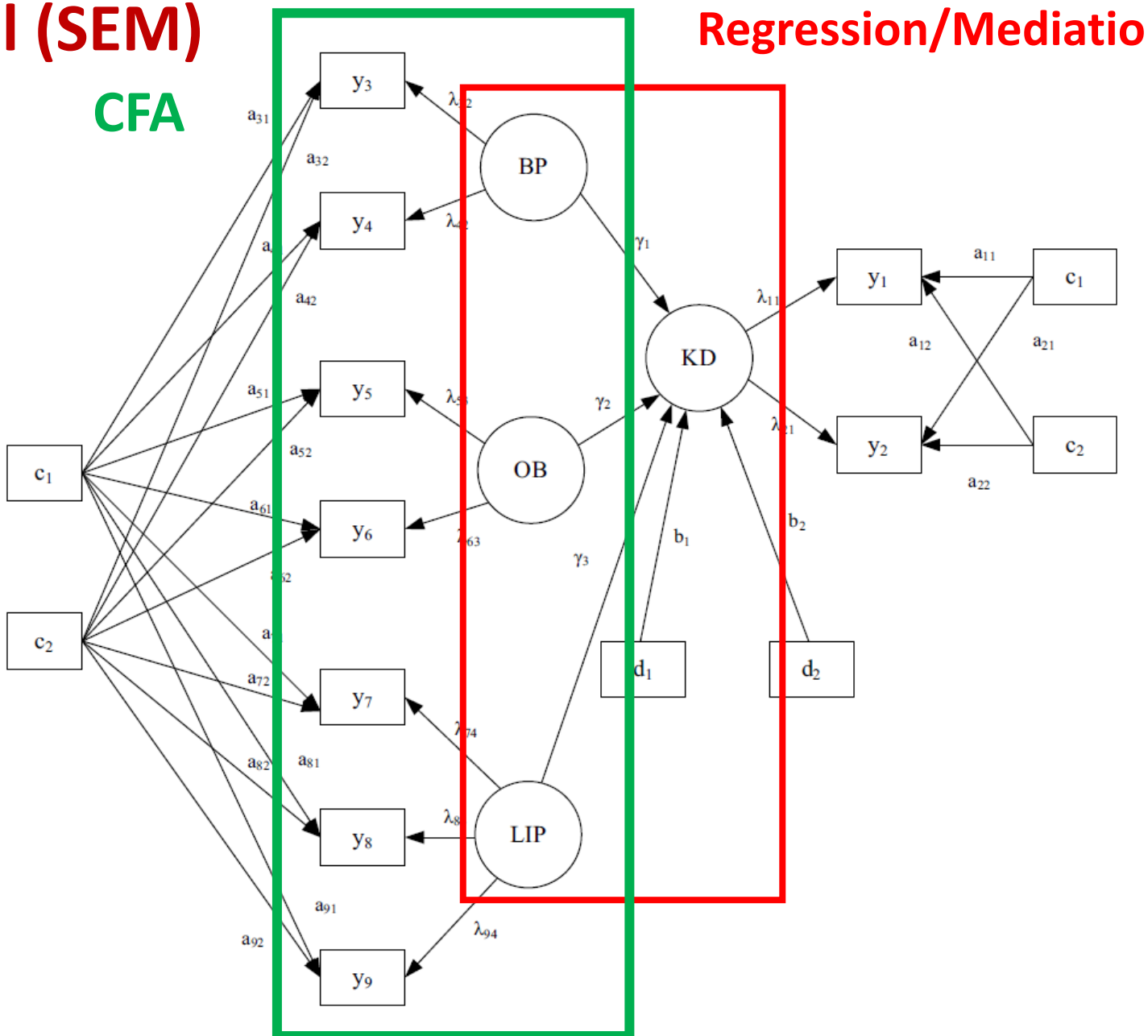
```
modelp3.1 <- '
# Measurement model
F1 =~ a*PS + b*RE + c*RT
F2 =~ d*HO + e*CO + f*ET + RT
F3 =~ g*BU + h*VA + i*RO
#Structural model
F1 ~ j*F2
F3 ~ k*F1+l*F2
#define a new parameter
ind := j*k
# error Variance
F3 ~~ m*F3
F1 ~~ n*F1
'
```

```
modelp3.2 <- '
# Measurement model
F1 =~ a*PS + b*RE + c*RT
F2 =~ d*HO + e*CO + f*ET +
F3 =~ g*BU + h*VA + i*RO
#Structural model
F1 ~ j*F2
F3 ~ k*F1
# error Variance
F3 ~~ m*F3
F1 ~~ n*F1
'
```

```
> anova(p3.1.fit,p3.2.fit)
Chi-Squared Difference Test

          Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
p3.1.fit 23 256978 257139 127.32
p3.2.fit 24 256976 257129 127.41   0.091983       1     0.7617
```

```
Latent Variables:
                   Estimate  Std.Err  z
  F1 =~
    PS        (a)    1.000
    RE        (b)    0.865    0.018
    RT        (c)    0.594    0.021
  F2 =~
    HO        (d)    1.000
    CO        (e)    0.908    0.017
    ET        (f)    0.869    0.016
    RT               0.197    0.019
  F3 =~
    BU        (g)    1.000
    VA        (h)    0.910    0.017
    RO        (i)    0.741    0.015

Regressions:
                   Estimate  Std.Err  z
  F1 ~
    F2        (j)    0.591    0.015
  F3 ~
    F1        (k)   -0.425    0.014  -
```

Structural Equation Model (SEM)



Structural Equation Model (SEM)

Type 2 diabetic patients data

The data set was collected from an applied genomics program conducted by the Institute of Diabetes, the Chinese University of Hong Kong. It aims to examine the clinical and molecular epidemiology of type 2 diabetes in Hong Kong Chinese, with particular emphasis on diabetic nephropathy. A consecutive cohort of 1188 type 2 diabetic patients was enrolled into the Hong Kong Diabetes Registry

- ■ Plasma creatine (PCr)
- ■ Urinary albumin creatinine ratio (ACR)
- ☐ Systolic blood pressure (SBP)
- ☐ Diastolic blood pressure (DBP)
- ☐ Body mass index (BMI)
- ☐ Waist hip ratio (WHR)

Kidney disease is reflected by both PCr and ACR.

Regression model by treating PCr and ACR as outcome (dependent) variables and regressing them on the observed explanatory (independent) variables

Structural Equation Model (SEM)

Type 2 diabetic patients data

The data set was collected from an applied genomics program conducted by the Institute of Diabetes, the Chinese University of Hong Kong. It aims to examine the clinical and molecular epidemiology of type 2 diabetes in Hong Kong Chinese, with particular emphasis on diabetic nephropathy. A consecutive cohort of 1188 type 2 diabetic patients was enrolled into the Hong Kong Diabetes Registry

Kidney disease is reflected by both PCr and ACR.

$$\text{PCr} = \alpha_1 \text{SBP} + \alpha_2 \text{DBP} + \alpha_3 \text{BMI} + \alpha_4 \text{WHR} + \epsilon_1$$
$$\text{ACR} = \beta_1 \text{SBP} + \beta_2 \text{DBP} + \beta_3 \text{BMI} + \beta_4 \text{WHR} + \epsilon_2$$

However, the effects of observed explanatory variables on kidney disease cannot be **directly assessed** from results obtained from regression analysis

Regression model by treating PCr and ACR as outcome (dependent) variables and regressing them on the observed explanatory (independent) variables

Structural Equation Model (SEM)

Type 2 diabetic patients data

The data set was collected from an applied genomics program conducted by the Institute of Diabetes, the Chinese University of Hong Kong. It aims to examine the clinical and molecular epidemiology of type 2 diabetes in Hong Kong Chinese, with particular emphasis on diabetic nephropathy. A consecutive cohort of 1188 type 2 diabetic patients was enrolled into the Hong Kong Diabetes Registry

- ■ Plasma creatine (PCr)
- ■ Urinary albumin creatinine ratio (ACR)
- □ Systolic blood pressure (SBP)
- □ Diastolic blood pressure (DBP)
- □ Body mass index (BMI)
- □ Waist hip ratio (WHR)

Kidney disease is reflected by both PCr and ACR.

A better approach is to appropriately group observed variables to form latent variables

$$\mathrm{KD} = \gamma_1 \mathrm{BP} + \gamma_2 \mathrm{OB} + \delta.$$

Simple regression equation with latent variables

Structural Equation Model (SEM)

Type 2 diabetic patients data

The data set was collected from an applied genomics program conducted by the Institute of Diabetes, the Chinese University of Hong Kong. It aims to examine the clinical and molecular epidemiology of type 2 diabetes in Hong Kong Chinese, with particular emphasis on diabetic nephropathy. A consecutive cohort of 1188 type 2 diabetic patients was enrolled into the Hong Kong Diabetes Registry

- ■ Plasma creatine (PCr)
- ■ Urinary albumin creatinine ratio (ACR)
- ☐ Systolic blood pressure (SBP)
- ☐ Diastolic blood pressure (DBP)
- ☐ Body mass index (BMI)
- ☐ Waist hip ratio (WHR)

Advantages of incorporating latent variables

1. It can reduce the number of variables in the key regression equation.
2. As highly correlated observed variables are grouped into latent variables, the problem induced by multicollinearity is alleviated.
3. It gives better assessments on the interrelationships of latent constructs.

Structural Equation Model (SEM)

Type 2 diabetic patients data

The data set was collected from an applied genomics program conducted by the Institute of Diabetes, the Chinese University of Hong Kong. It aims to examine the clinical and molecular epidemiology of type 2 diabetes in Hong Kong Chinese, with particular emphasis on diabetic nephropathy. A consecutive cohort of 1188 type 2 diabetic patients was enrolled into the Hong Kong Diabetes Registry

- ■ Plasma creatine (PCr)
- ■ Urinary albumin creatinine ratio (ACR)
- □ Systolic blood pressure (SBP)
- □ Diastolic blood pressure (DBP)
- □ Body mass index (BMI)
- □ Waist hip ratio (WHR)

Structural Equation

$$\mathrm{KD} = \gamma_1 \mathrm{BP} + \gamma_2 \mathrm{OB} + \delta.$$

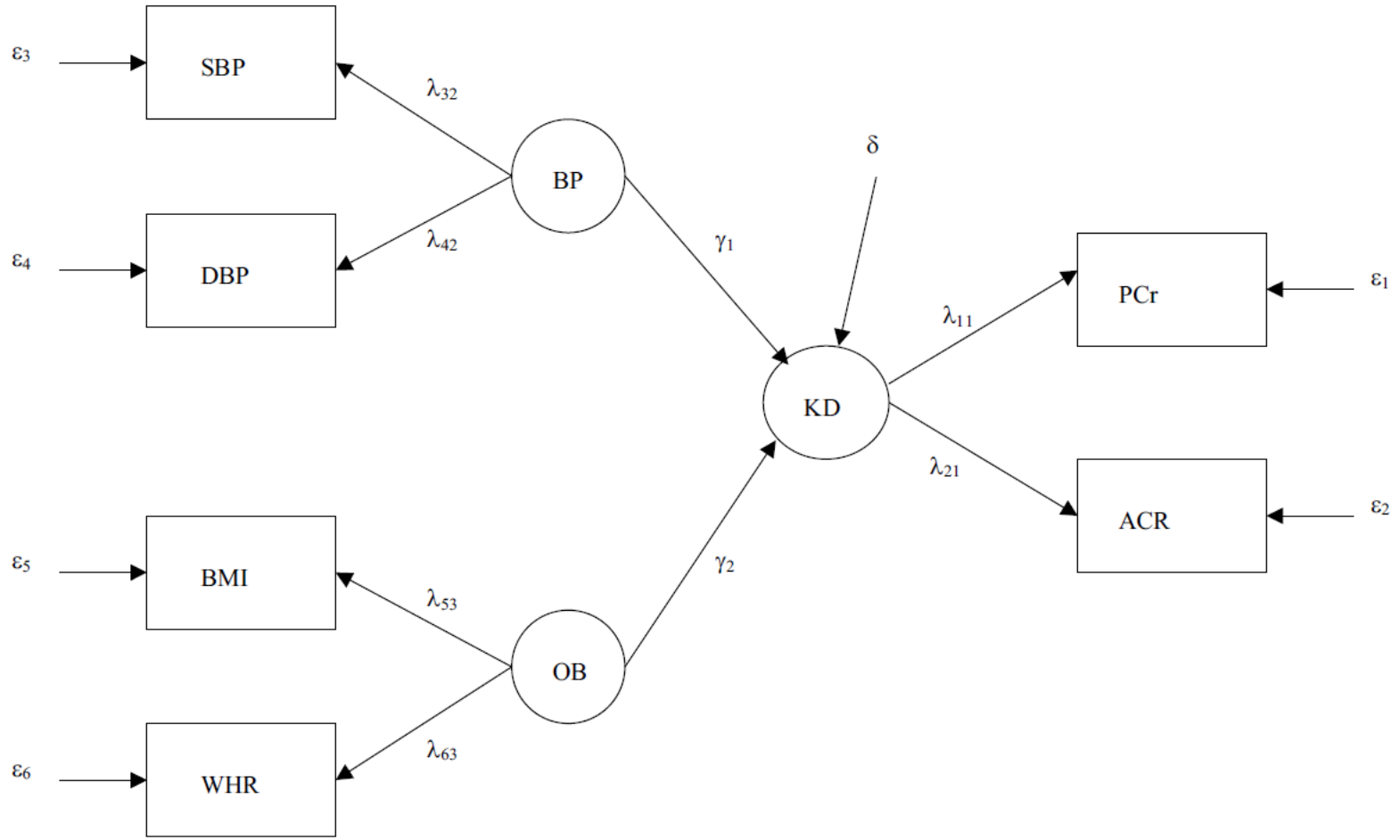
$$\mathrm{PCr} = \mu_1 + \lambda_{11}\mathrm{KD} + \epsilon_1,\ \mathrm{DBP} = \mu_4 + \lambda_{42}\mathrm{BP} + \epsilon_4$$
$$\mathrm{ACR} = \mu_2 + \lambda_{21}\mathrm{KD} + \epsilon_2,\ \ \mathrm{BMI} = \mu_5 + \lambda_{53}\mathrm{OB} + \epsilon_5$$
$$\mathrm{SBP} = \mu_3 + \lambda_{32}\mathrm{BP} + \epsilon_3,\ \mathrm{WHR} = \mu_6 + \lambda_{63}\mathrm{OB} + \epsilon_6$$

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega} + \boldsymbol{\epsilon}$$ **Measurement Equation**

$$\begin{bmatrix} \mathrm{PCr} \\ \mathrm{ACR} \\ \mathrm{SBP} \\ \mathrm{DBP} \\ \mathrm{BMI} \\ \mathrm{WHR} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & \lambda_{32} & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & \lambda_{53} \\ 0 & 0 & \lambda_{63} \end{bmatrix} \begin{bmatrix} \mathrm{KD} \\ \mathrm{BP} \\ \mathrm{OB} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

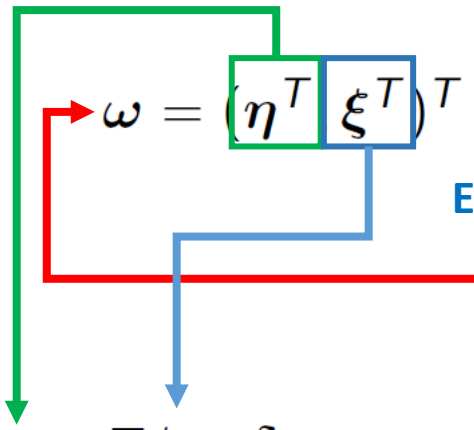
Structural Equation Model (SEM)

Type 2 diabetic patients data



Structural Equation Model (SEM)

Matrix Form



η and ξ are $q_1 \times 1$ and $q_2 (= q - q_1) \times 1$ random vectors

$\xi = (\xi_1, \cdots, \xi_{q_2})^T \qquad \eta = (\eta_1, \cdots, \eta_{q_1})^T$

Structural Equation $\eta = \Gamma\xi + \delta$

$$\eta_j = \gamma_{j1}\xi_1 + \cdots + \gamma_{jq_2}\xi_{q_2} + \delta_j$$
$$j = 1, \cdots, q_1$$

Γ is a $q_1 \times q_2$ unknown matrix of regression coefficients

Measurement Equation $\mathbf{y} = \mu + \Lambda\omega + \epsilon$

$\mathbf{y}$ is a $p \times 1$ random vector of observed variables

μ is a $p \times 1$ vector of intercepts

Λ is a $p \times q$ unknown matrix of factor loadings

ϵ and δ are $p \times 1$ and $q_1 \times 1$ random vectors of measurement (residual) errors, respectively

Structural Equation Model (SEM)

Matrix Form

The standard linear SEMs have some assumptions: For $i = 1, \cdots, n$,

- A1: The random vectors of residual errors $\boldsymbol{\epsilon}_i$ are i.i.d. $N[\mathbf{0}, \boldsymbol{\Psi}_\epsilon]$, where $\boldsymbol{\Psi}_\epsilon$ is a diagonal covariance matrix.
- A2: The random vectors of explanatory latent variables $\boldsymbol{\xi}_i$ are i.i.d. $N[\mathbf{0}, \boldsymbol{\Phi}]$, where $\boldsymbol{\Phi}$ is a general covariance matrix.
- A3: The random vectors of residual errors $\boldsymbol{\delta}_i$ are i.i.d. $N[\mathbf{0}, \boldsymbol{\Psi}_\delta]$, where $\boldsymbol{\Psi}_\delta$ is a diagonal covariance matrix.
- A4: $\boldsymbol{\delta}_i$ is independent of $\boldsymbol{\xi}_i$, and $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\omega}_i$ and $\boldsymbol{\delta}_i$.

Structural Equation $\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\delta}$

$$\eta_j = \gamma_{j1}\xi_1 + \cdots + \gamma_{jq_2}\xi_{q_2} + \delta_j$$
$$j = 1, \cdots, q_1$$

$\boldsymbol{\Gamma}$ is a $q_1 \times q_2$ unknown matrix of regression coefficients

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega} + \boldsymbol{\epsilon}$$ **Measurement Equation**

$\mathbf{y}$ is a $p \times 1$ random vector of observed variables

$\boldsymbol{\mu}$ is a $p \times 1$ vector of intercepts

$\boldsymbol{\Lambda}$ is a $p \times q$ unknown matrix of factor loadings

$\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are $p \times 1$ and $q_1 \times 1$ random vectors of measurement (residual) errors, respectively

Structural Equation Model (SEM)

Matrix Form

Identification

Method 1

$$\mathbf{\Lambda}^T = \begin{bmatrix} 1 & \lambda_{21} & \lambda_{31} & \lambda_{41} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{62} & \lambda_{72} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{93} & \lambda_{10,3} \end{bmatrix}$$

Method 2

allow λ_{11}, λ_{52}, and/or λ_{83} in $\mathbf{\Lambda}$ to be unknown parameters and fix the diagonal elements of $\mathbf{\Phi}^+$ as 1
hence $\mathbf{\Phi}^+$ is a correlation matrix

Structural Equation $\boldsymbol{\eta} = \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\delta}$

$$\eta_j = \gamma_{j1}\xi_1 + \cdots + \gamma_{jq_2}\xi_{q_2} + \delta_j$$
$$j = 1, \cdots, q_1$$

$\mathbf{\Gamma}$ is a $q_1 \times q_2$ unknown matrix of regression coefficients

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{\Lambda}\boldsymbol{\omega} + \boldsymbol{\epsilon}$$ **Measurement Equation**

$\mathbf{y}$ is a $p \times 1$ random vector of observed variables

$\boldsymbol{\mu}$ is a $p \times 1$ vector of intercepts

$\mathbf{\Lambda}$ is a $p \times q$ unknown matrix of factor loadings

$\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are $p \times 1$ and $q_1 \times 1$ random vectors of measurement (residual) errors, respectively

Structural Equation Model (SEM)

Extension

To develop better models, it is often desirable to incorporate explanatory observed variables on the right-hand sides of the measurement and structural equations. In the field of SEM, these explanatory observed variables are regarded as **fixed covariates**.

Fixed covariates give **more ingredients** to account for the outcome latent variables, in addition to the explanatory latent variables.

The residual errors in both equations can be reduced by incorporating fixed covariates

Provides **additional information** about the latent exposure and thus reduces estimation uncertainty for the latent variables

Structural Equation $\eta = \mathbf{Bd} + \mathbf{\Gamma}\xi + \delta$

$\mathbf{y} = \mathbf{Ac} + \mathbf{\Lambda}\omega + \epsilon$ **Measurement Equation**

$\mathbf{B}$ is a $q_1 \times r_2$ matrix of unknown coefficients

$\mathbf{d}$ is an $r_2 \times 1$ vector of fixed covariates

$\mathbf{A}$ is a $p \times r_1$ matrix of unknown coefficients

$\mathbf{c}$ is an $r_1 \times 1$ vector of fixed covariates

Note that $\mathbf{c}$ and $\mathbf{d}$ may have common elements

Structural Equation Model (SEM)

Type 2 diabetic patients data

Suppose that the main objective is on studying the complex diabetic kidney disease, with emphasis on assessing effects of blood pressure, obesity, lipid control as well as some covariates on that disease

- ■ Plasma creatine (PCr)
- ■ Urinary albumin creatinine ratio (ACR)
- ☐ Systolic blood pressure (SBP)
- ☐ Diastolic blood pressure (DBP)
- ☐ Body mass index (BMI)
- ☐ Waist hip ratio (WHR)
- ☐ Non-high-density lipoprotein cholesterol (non-HDL-C)
- ☐ Low-density lipoprotein cholesterol (LDL-C)
- ☐ Plasma triglyceride (TG)

Incorporate **‘smoking (c1)’ and ‘alcohol (c2)’** in the measurement equation, and ‘**age (d1)’ and ‘gender (d2)’** in the structural equation

Structural Equation Model (SEM)

Type 2 diabetic patients data

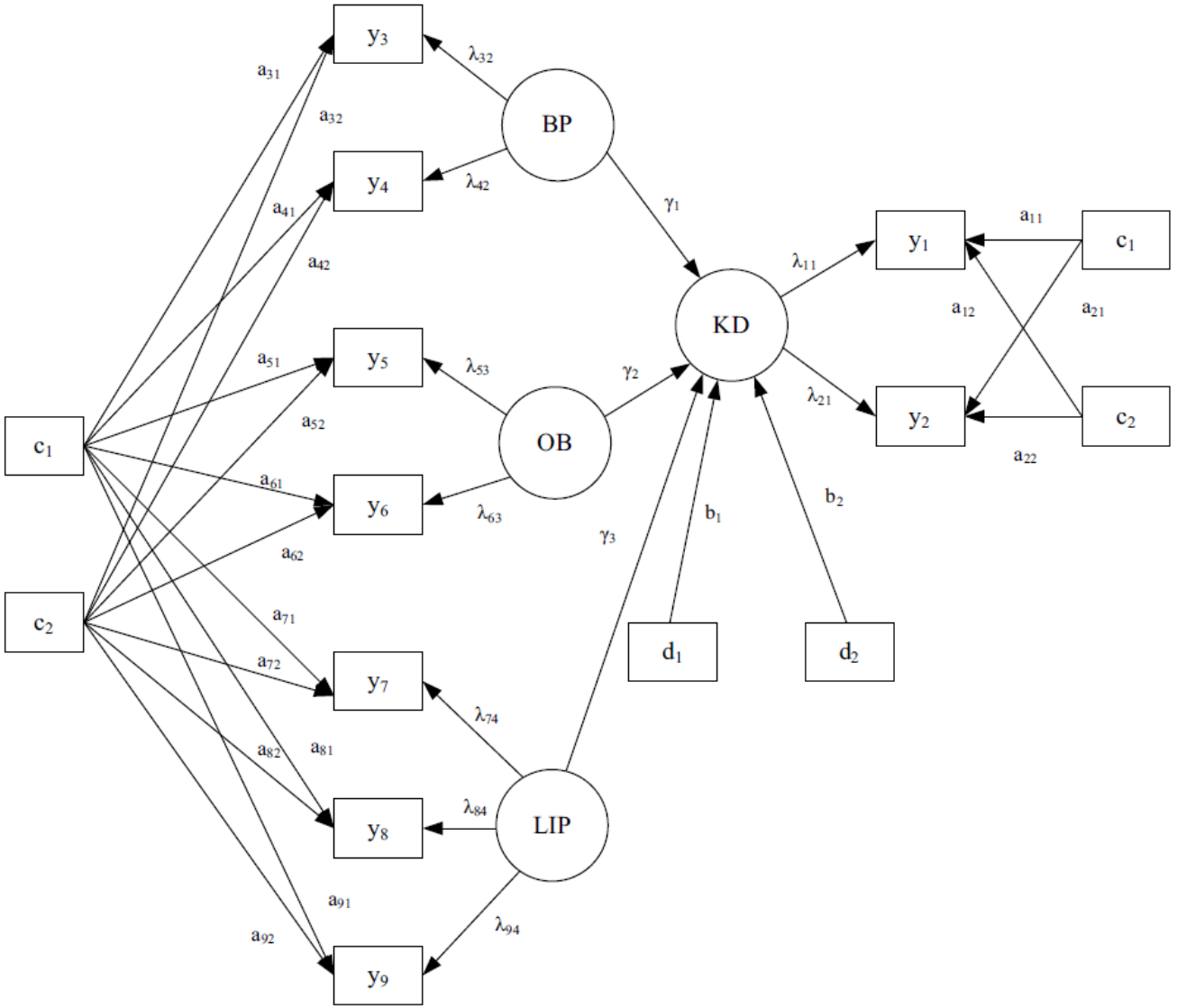
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \\ a_{71} & a_{72} \\ a_{81} & a_{82} \\ a_{91} & a_{92} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 \\ 0 & \lambda_{32} & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 \\ 0 & 0 & \lambda_{53} & 0 \\ 0 & 0 & \lambda_{63} & 0 \\ 0 & 0 & 0 & \lambda_{74} \\ 0 & 0 & 0 & \lambda_{84} \\ 0 & 0 & 0 & \lambda_{94} \end{bmatrix} \begin{bmatrix} \mathrm{KD} \\ \mathrm{BP} \\ \mathrm{OB} \\ \mathrm{LIP} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix}$$

$$\mathrm{KD} = b_1 \mathrm{age} + b_2 \mathrm{gender} + \gamma_1 \mathrm{BP} + \gamma_2 \mathrm{OB} + \gamma_3 \mathrm{LIP} + \delta,$$

where a_{jk}, λ_{jk}, b_1, b_2, γ_1, γ_2, and γ_3 are unknown regression coefficients

Structural Equation Model (SEM)

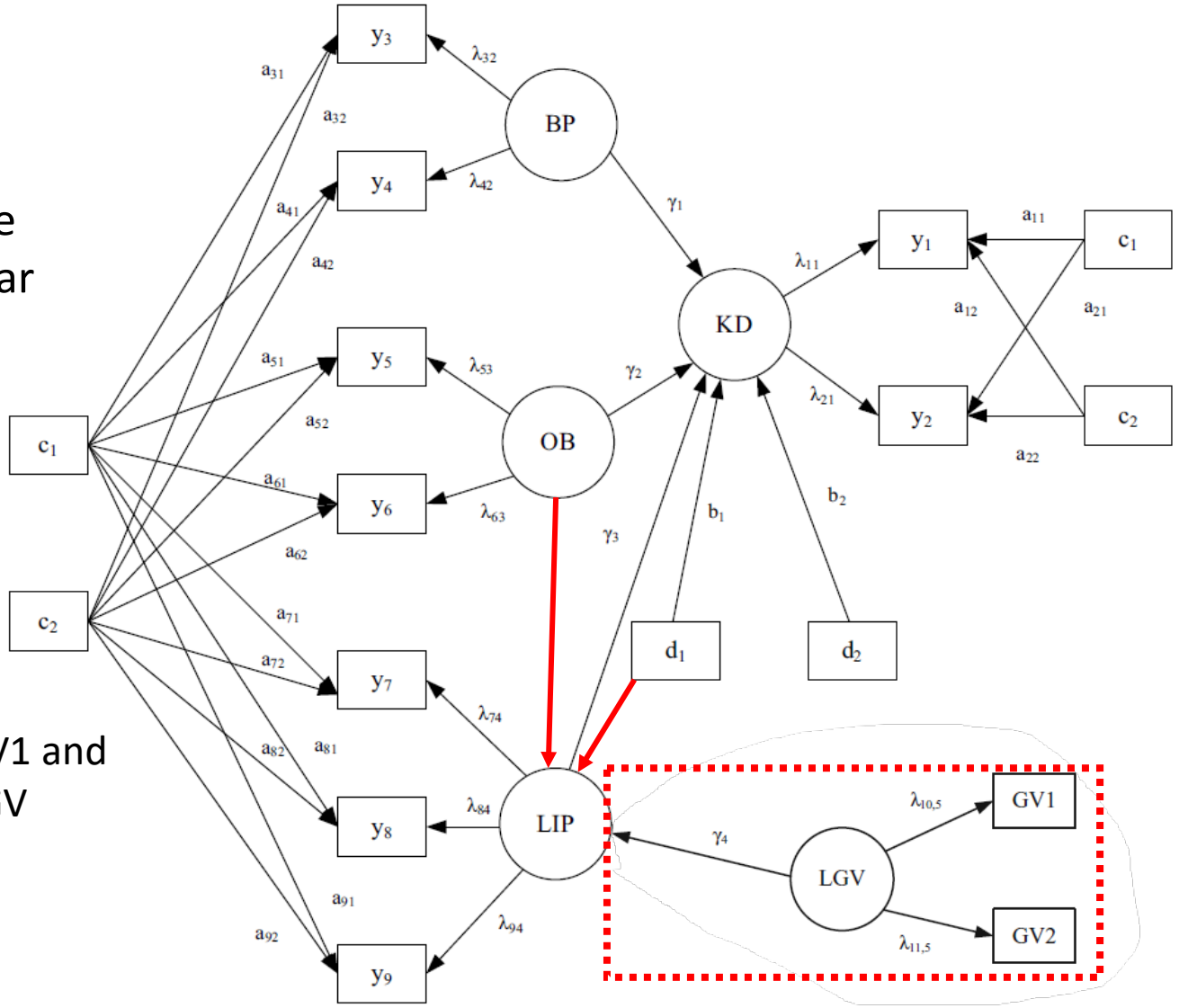
Type 2 diabetic patients data



Structural Equation Model (SEM)

Extension

Although the emphasis is on assessing the effects of explanatory latent variables on the key outcome latent variables, some particular **explanatory latent variables may be significantly related to other explanatory latent variables and/or fixed covariates**.



- ✓ Two additional observed genetic variables GV1 and GV2 which correspond to a latent variable LGV
- ✓ A path from LGV to LIP
- ✓ A path from OB to LIP
- ✓ A path from age (d1) to LIP.

Structural Equation Model (SEM)

Extension

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ \mathrm{GV}_1 \\ \mathrm{GV}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \\ a_{71} & a_{72} \\ a_{81} & a_{82} \\ a_{91} & a_{92} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & \lambda_{32} & 0 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{53} & 0 & 0 \\ 0 & 0 & \lambda_{63} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{74} & 0 \\ 0 & 0 & 0 & \lambda_{84} & 0 \\ 0 & 0 & 0 & \lambda_{94} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{10,5} \\ 0 & 0 & 0 & 0 & \lambda_{11,5} \end{bmatrix} \begin{bmatrix} \mathrm{KD} \\ \mathrm{BP} \\ \mathrm{OB} \\ \mathrm{LIP} \\ \mathrm{LGV} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \end{bmatrix}$$

$$\begin{pmatrix} \mathrm{KD} \\ \mathrm{LIP} \end{pmatrix} = \begin{pmatrix} b_1 & b_2 \\ b_3 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} + \begin{pmatrix} 0 & \pi_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathrm{KD} \\ \mathrm{LIP} \end{pmatrix} + \begin{pmatrix} \gamma_1 & \gamma_2 & 0 \\ 0 & \gamma_3 & \gamma_4 \end{pmatrix} \begin{pmatrix} \mathrm{BP} \\ \mathrm{OB} \\ \mathrm{LGV} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$$

Structural Equation Model (SEM)

Extension

This structural equation allows some outcome latent variables depend on the other outcome latent variables through an appropriately defined Π. Particularly useful in business and social-psychological research

Structural Equation $\eta = \mathbf{Bd} + \mathbf{\Pi}\eta + \mathbf{\Gamma}\xi + \delta$ $\quad\quad$ $\mathbf{y} = \mathbf{Ac} + \mathbf{\Lambda}\omega + \epsilon$ Measurement Equation

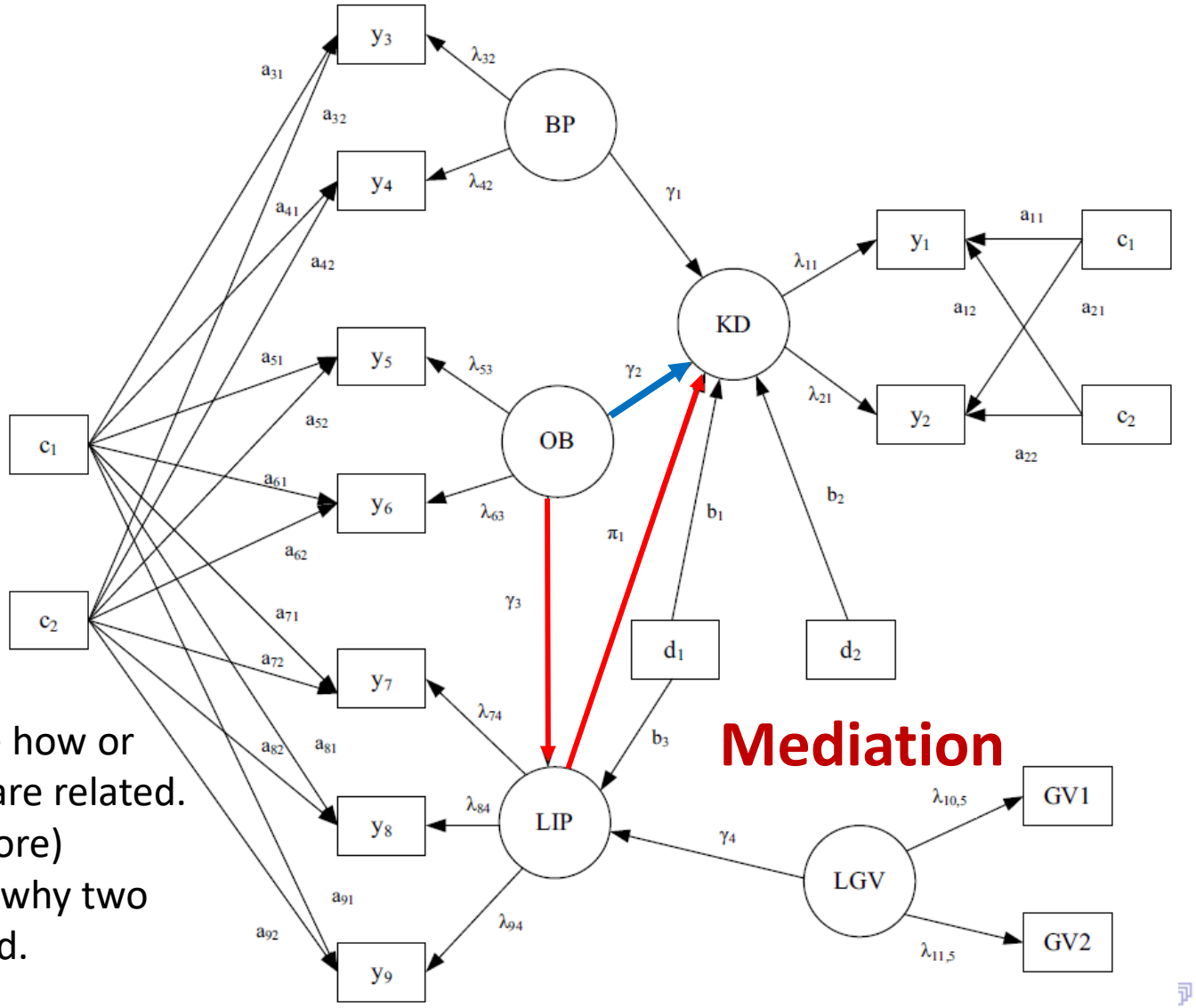
$\mathbf{\Pi}$ is a $q_1 \times q_1$ matrix of unknown coefficients

$\mathbf{I} - \mathbf{\Pi}$ is nonsingular $\quad\quad$ diagonal elements of $\mathbf{\Pi}$ are zero

$$\begin{pmatrix} \text{KD} \\ \text{LIP} \end{pmatrix} = \begin{pmatrix} b_1 & b_2 \\ b_3 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} + \begin{pmatrix} 0 & \pi_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \text{KD} \\ \text{LIP} \end{pmatrix} + \begin{pmatrix} \gamma_1 & \gamma_2 & 0 \\ 0 & \gamma_3 & \gamma_4 \end{pmatrix} \begin{pmatrix} \text{BP} \\ \text{OB} \\ \text{LGV} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$$

Structural Equation Model (SEM)

Extension

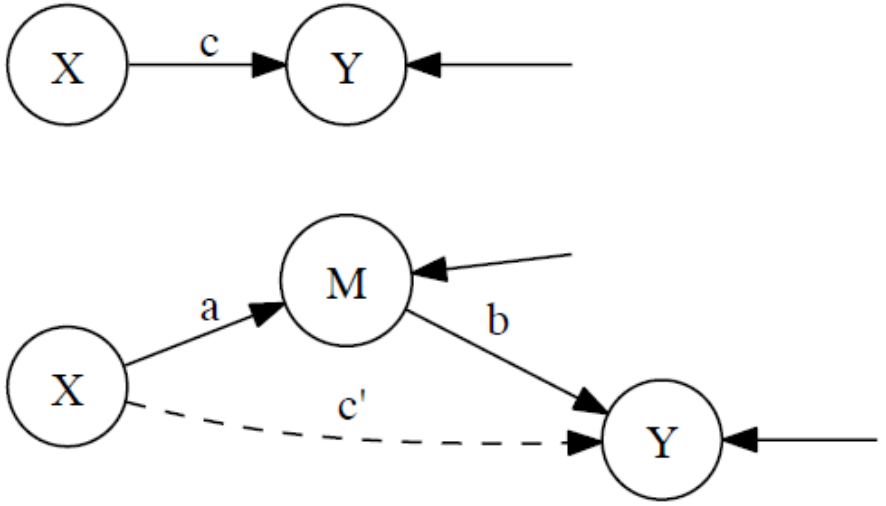


- Mediation models investigate how or why two (or more) variables are related.
- Mediation is when one (or more) variables explains the reason why two (or more variables) are related.

Structural Equation Model (SEM)

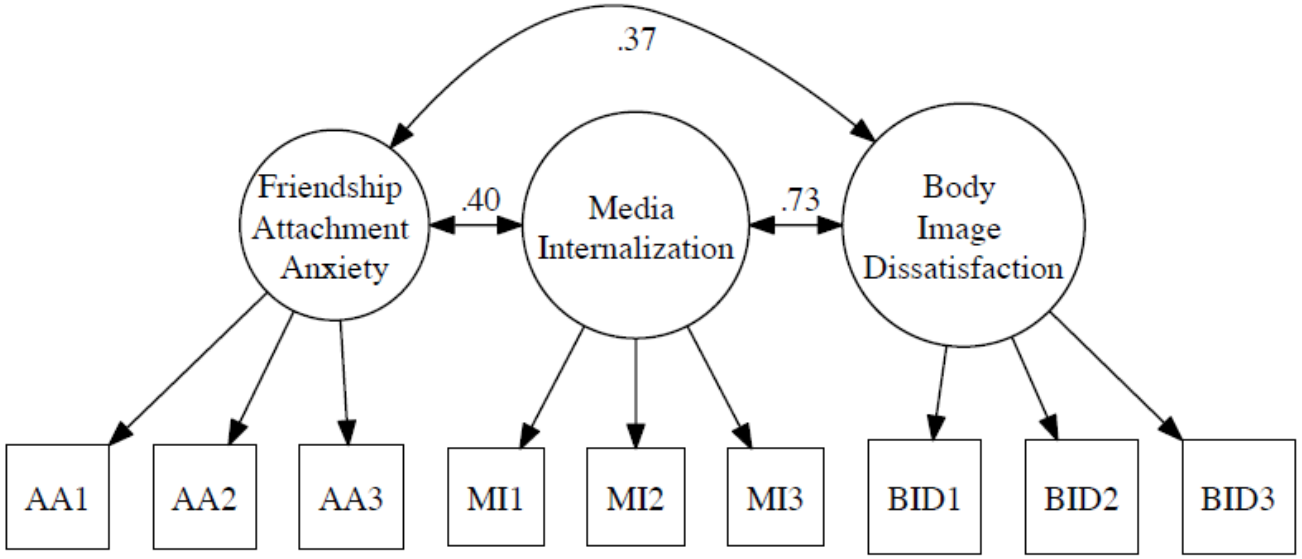
Mediation

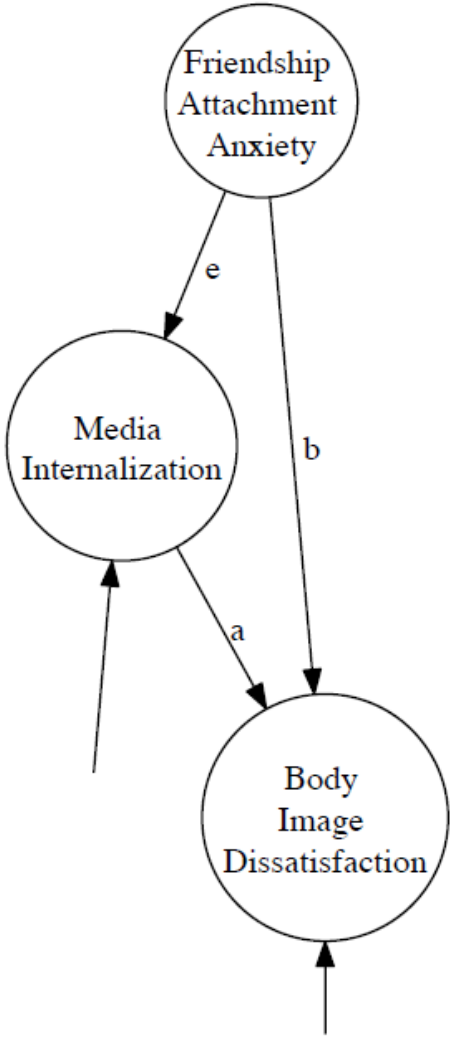
1. First, there is a relationship (via c) between variables X and Y
2. Then, M is put into the model and is related to both X (via a) and Y (via b).
3. After M was put into the model, then the relationship between X and Y dwindles (i.e., $c' < c$).



Structural Equation Model (SEM)

Mediation

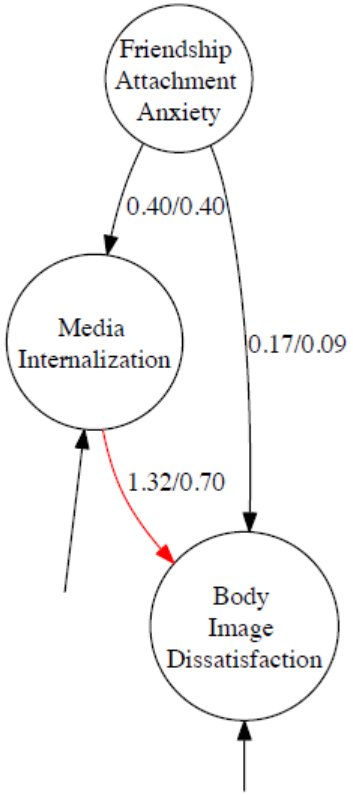




Media internalization (awareness and attitudes toward prevailing sociocultural standards of attractiveness) was hypothesized to mediate the positive association between attachment anxiety in friendships and body image dissatisfaction

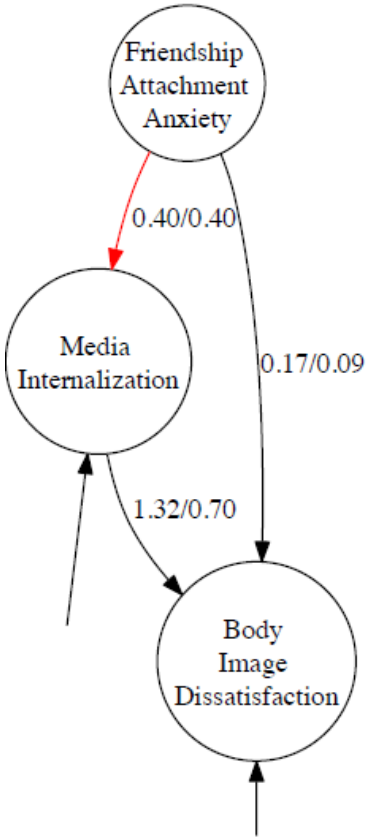
Structural Equation Model (SEM)

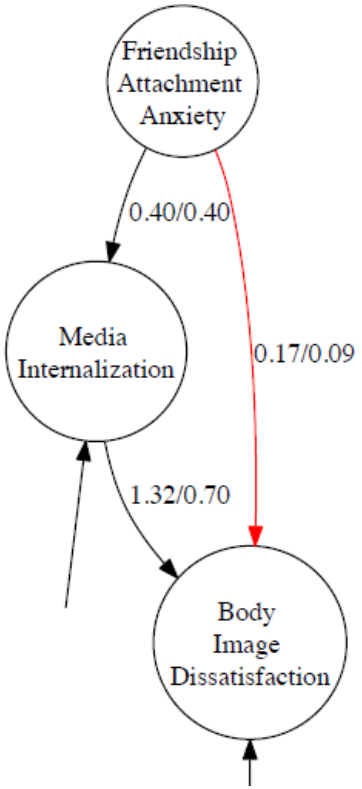
Mediation



Media internalization is strongly related to body image dissatisfaction (path a: 0.70)

Media internalization is moderately related to attachment anxiety (path e: 0.40)





The attachment anxiety-body image dissatisfaction relationship (path b), dwindles to almost 0 (b = .09) in the presence of these variables

harder one Basic & Advanced ...
easier one Structural Equation ...

End of Chapter 1&2