# STAT 5020

# Chapter 1&2: Introduction of Structural Equation Modelling

Department of Statistics
2021/2022 Term 2

# 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

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

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

# 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.

# 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
```

# 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
```

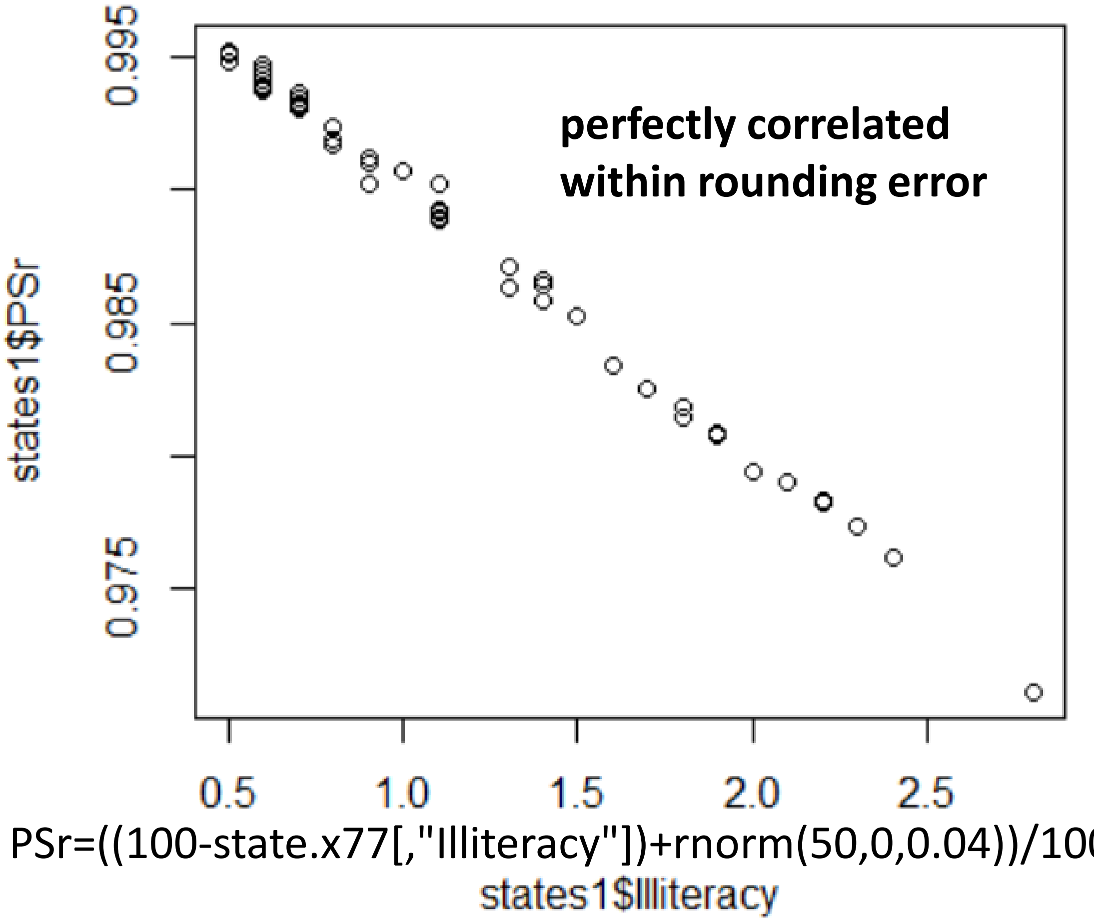
# Regression

## *Multicollinearity*

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

```
> 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

# Regression

## *Multicollinearity*

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

```
> 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***

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

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

$$X_2 = k X_1$$

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

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

# Regression

## *Multicollinearity*

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

## *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**

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

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

## Regression

***Multicollinearity***

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

***Variance inflation factor (VIF)***

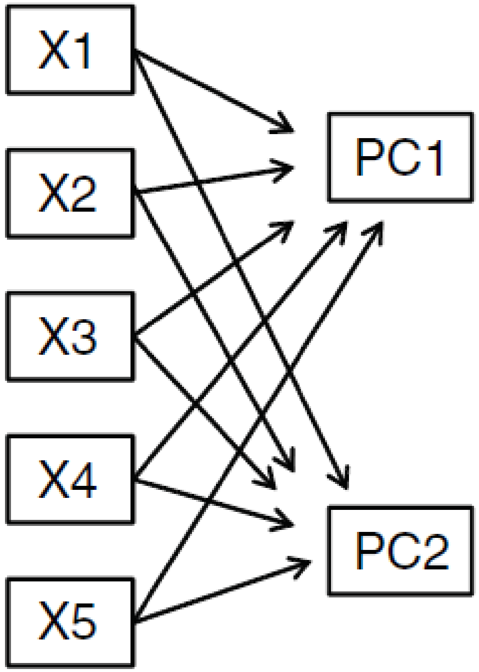
```
> vif(fit)
Population Illiteracy     Income      Frost
  1.245282   2.165848   1.345822   2.082547
> vif(fit1)
Population Illiteracy     Income      Frost        PSr
  1.247204 242.705554   1.348271   2.130582 246.684222
```

**Principal Components Analysis (PCA)**

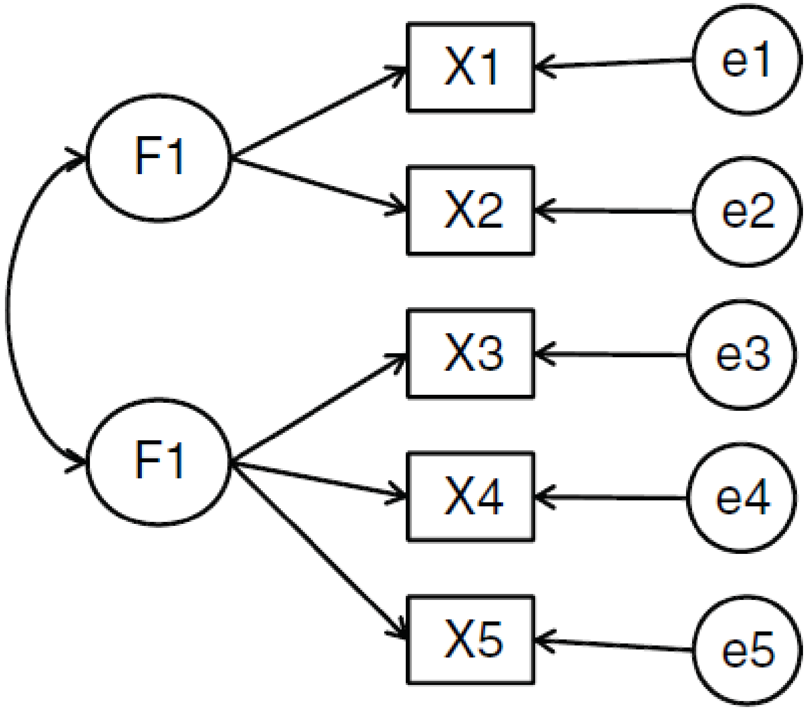
**Exploratory Factor Analysis (EFA)**

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

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




(a) Principal Components Model




(b) Factor Analysis Model

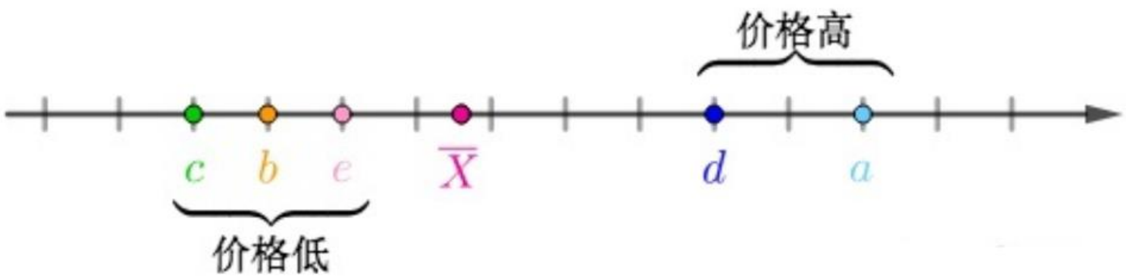
# Section 1: Principal Component Analysis and Exploratory Factor Analysis

# Principal Components Analysis (PCA)

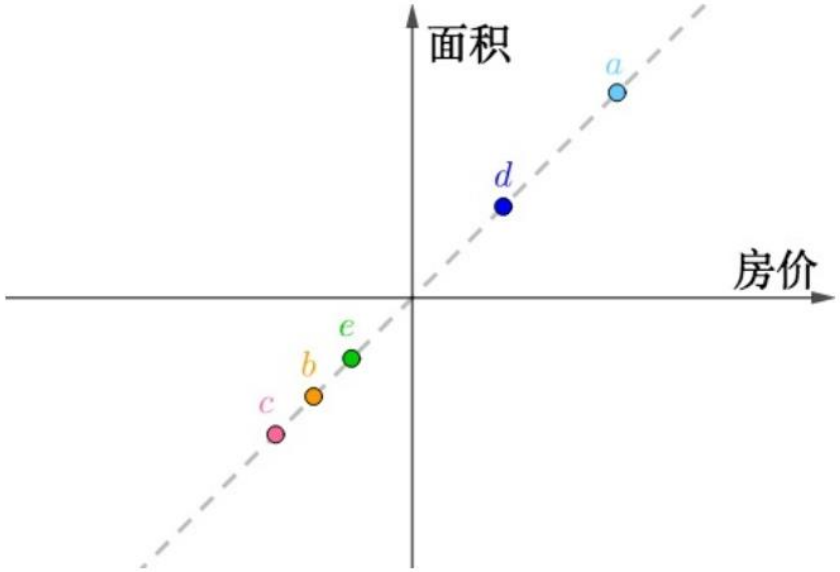
## Data-reduction

| | 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 |

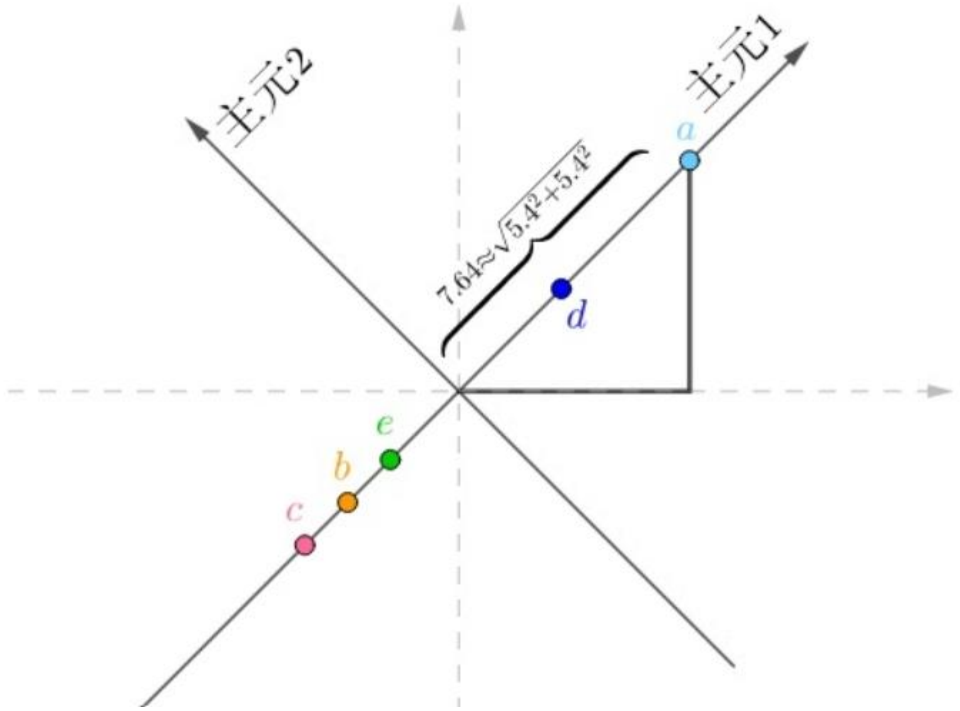


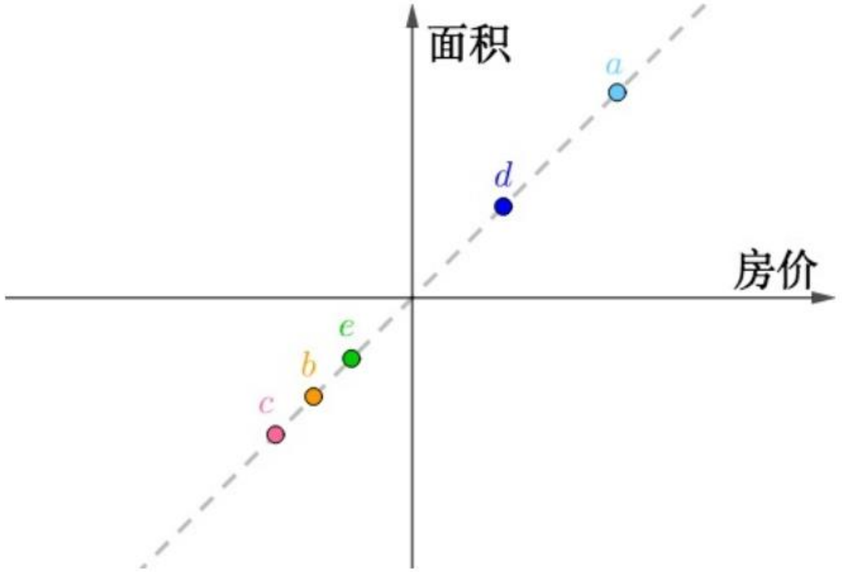
# Principal Components Analysis (PCA)

## 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 |

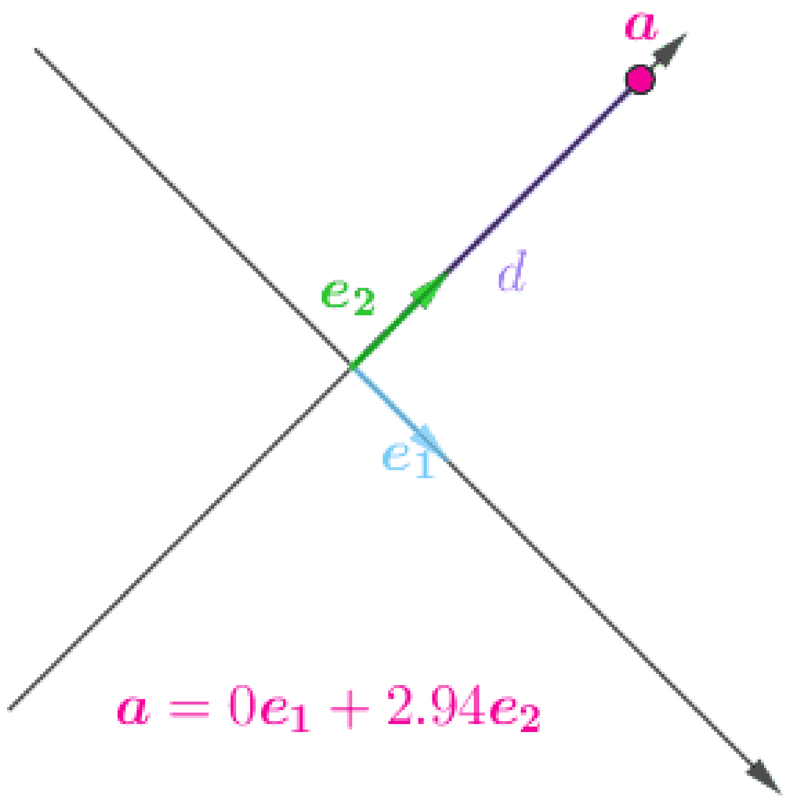
| | 房价(百万元) | 面积(百平米) |
|---|---|---|
| $a$ | 10 | 10 |
| $b$ | 2 | 2 |
| $c$ | 1 | 1 |
| $d$ | 7 | 7 |
| $e$ | 3 | 3 |

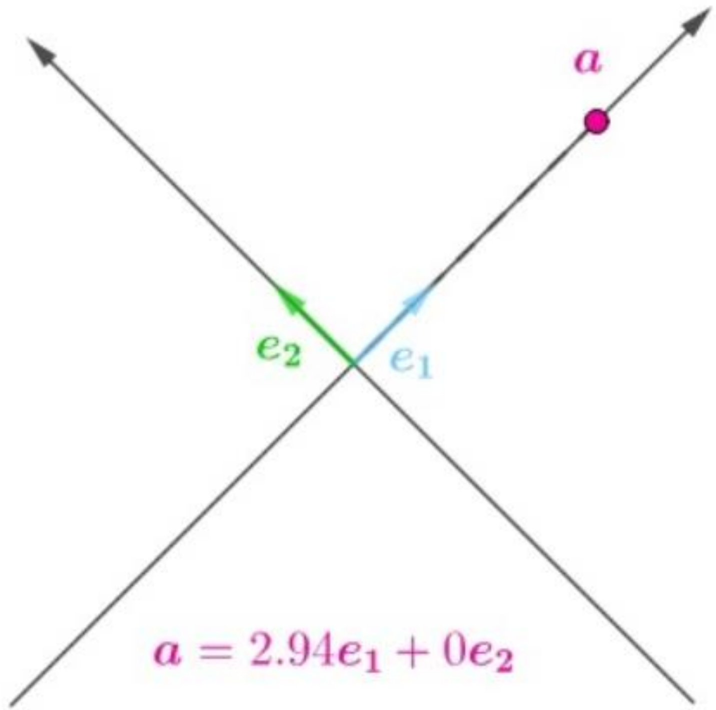


# Principal Components Analysis (PCA)

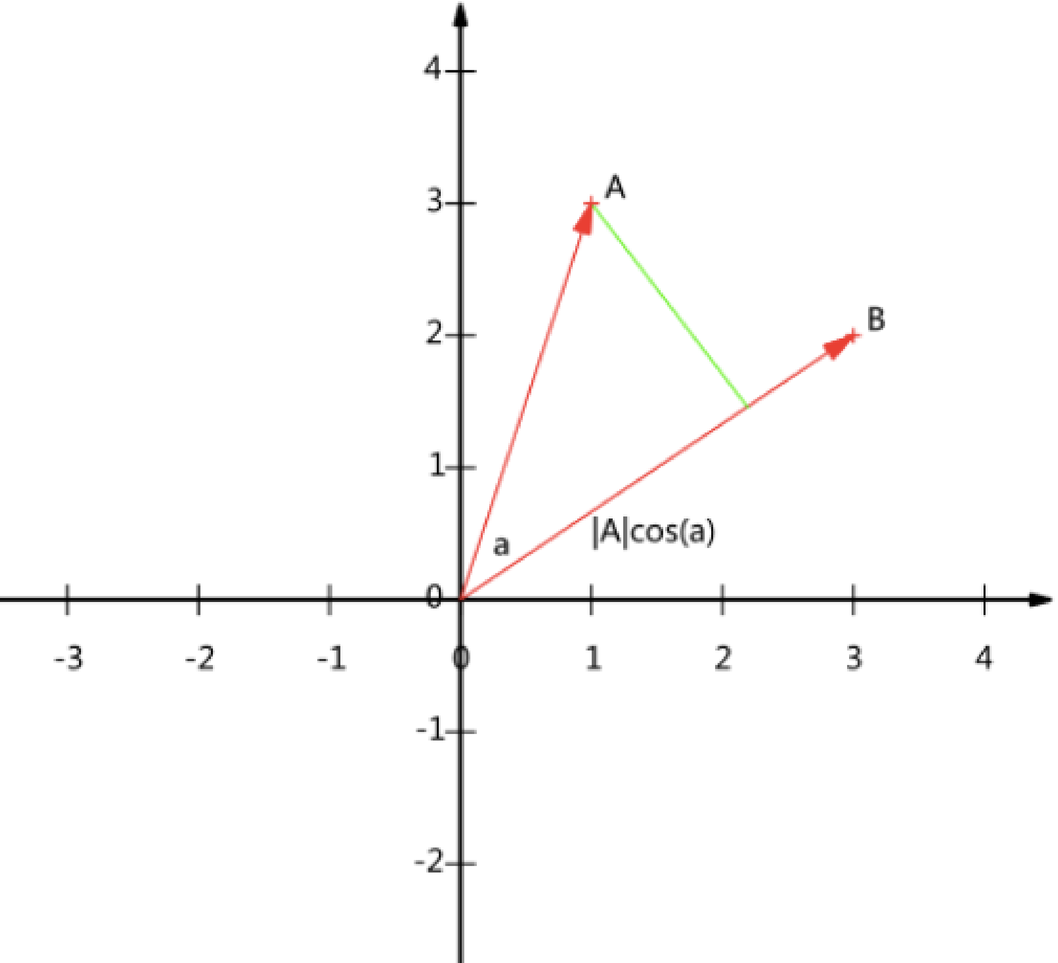
Data-reduction



# Principal Components Analysis (PCA)

## Change of Basis in Matrix Form




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

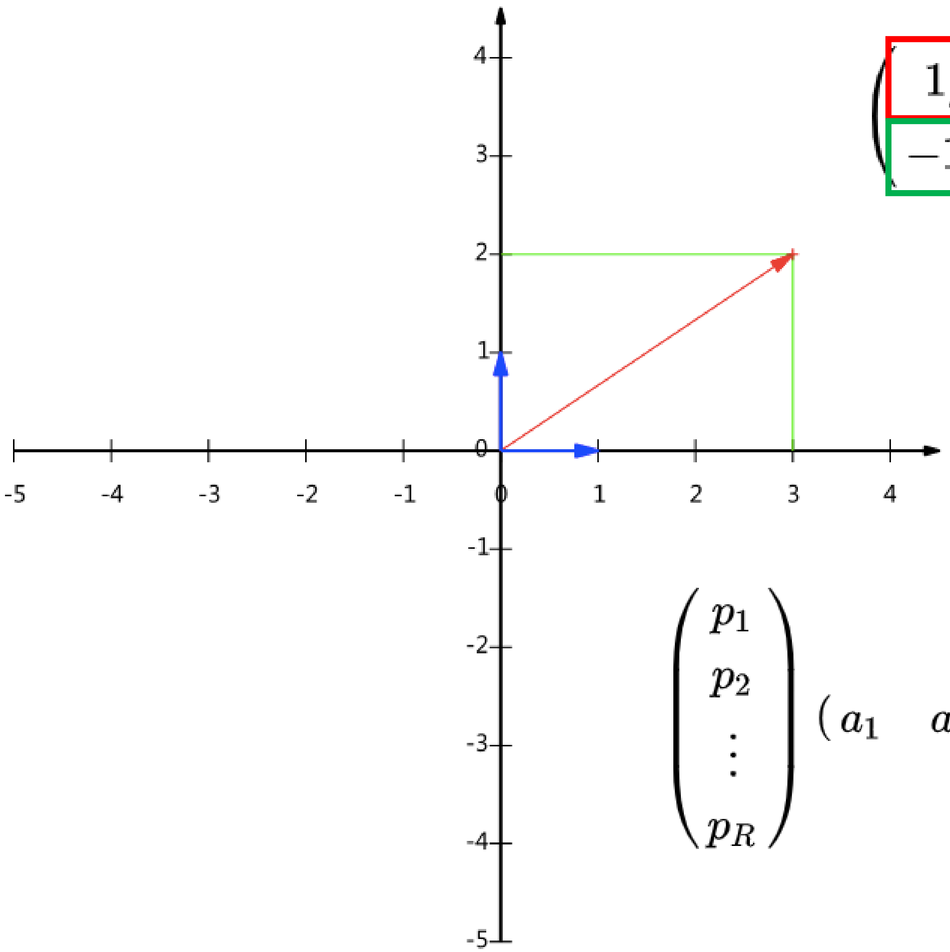
$$A \cdot B = x_1 y_1 + x_2 y_2$$

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

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

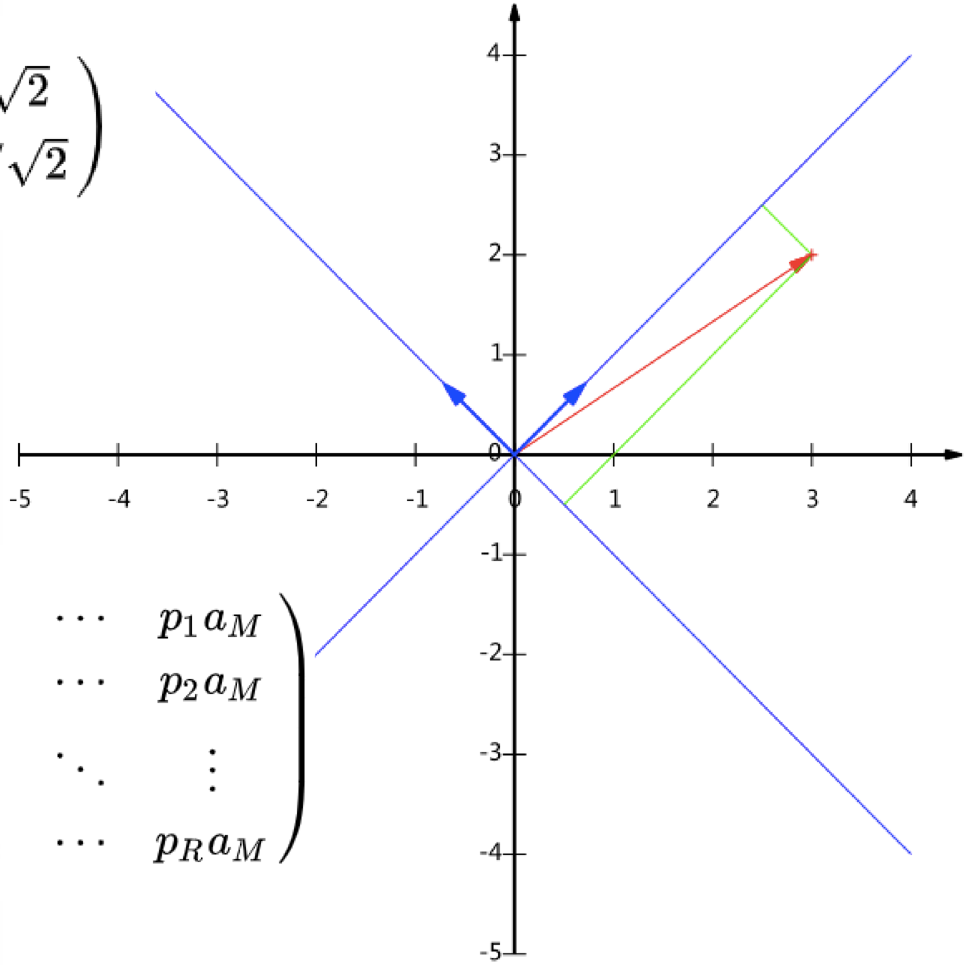
# Principal Components Analysis (PCA)

## Change of Basis in Matrix Form




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

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



# Principal Components Analysis (PCA)

Covariance Matrix

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

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

Change of Basis

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

$$Y = PX$$

**Eigenvector**

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

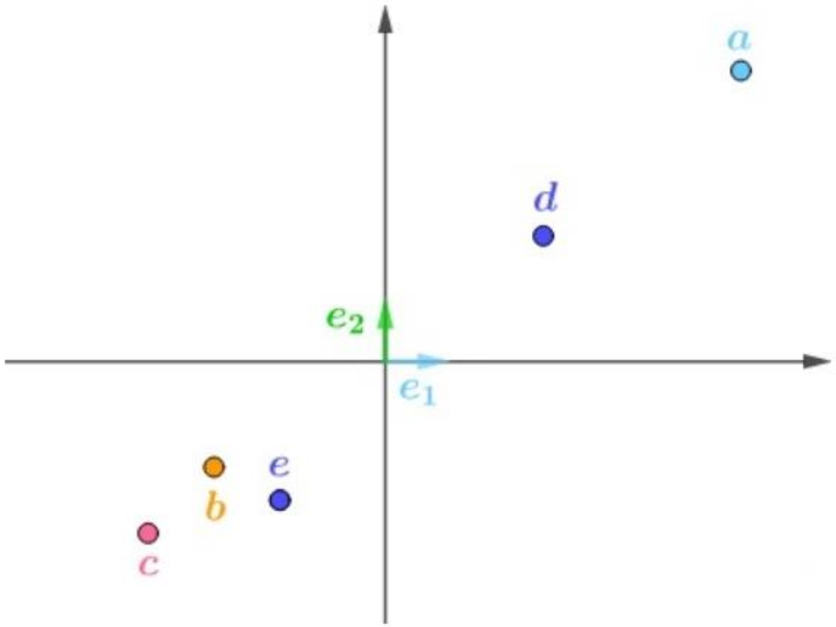
**Diagonal**

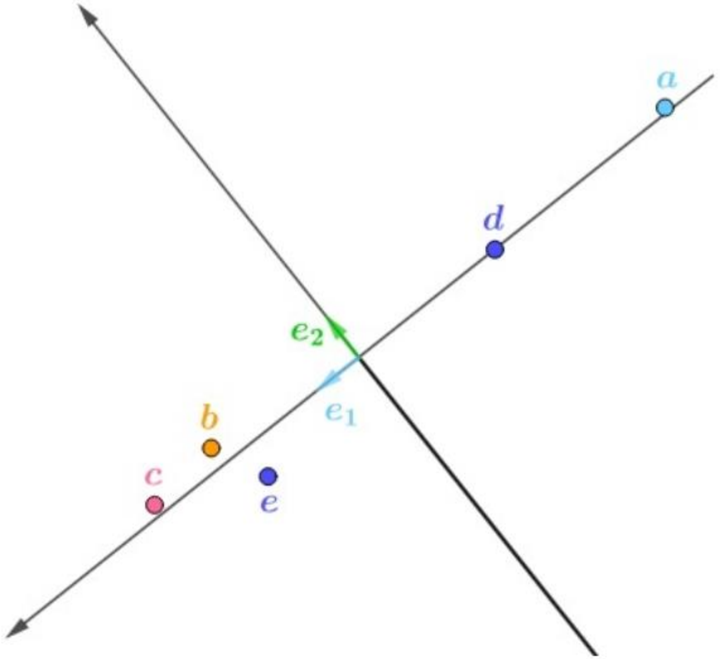
A square matrix is diagonal if and only if it is triangular and normal.

# Principal Components Analysis (PCA)

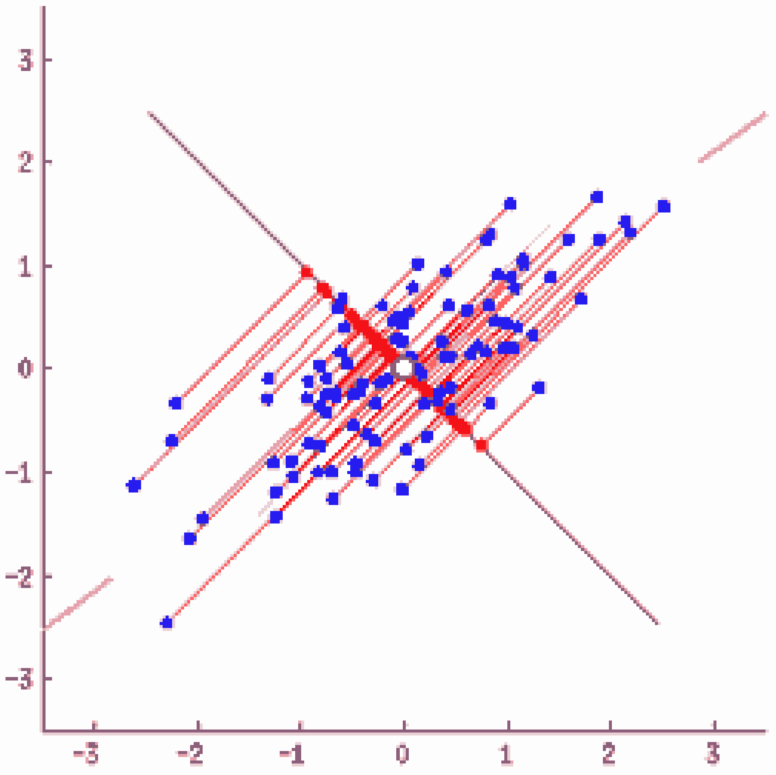
Data-reduction

| | 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

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

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)

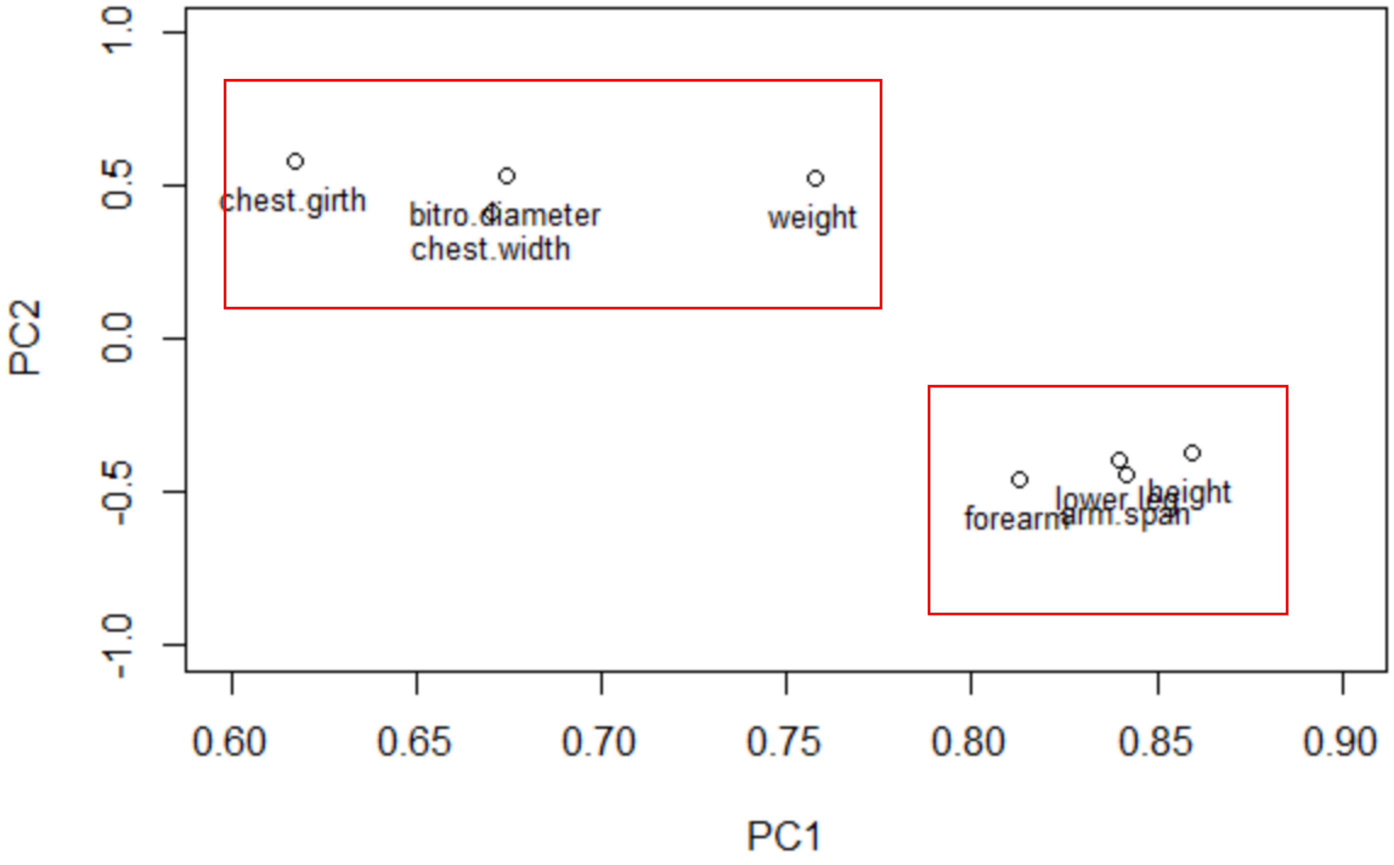
**Table 14.3 Correlations among body measurements for 305 girls (`Harman23.cor`)**

| | 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.

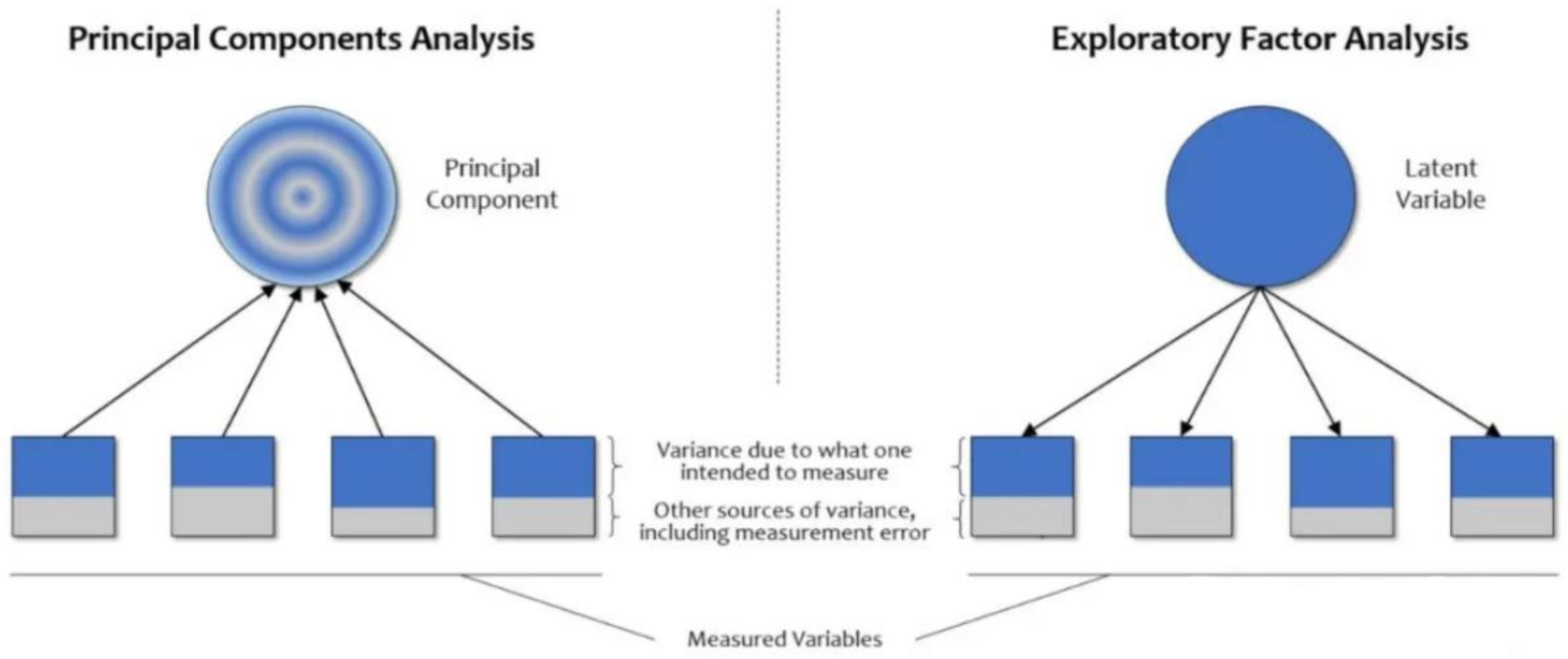
```
> pc1
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 3, rotate = "none"
Standardized loadings (pattern matrix) based upon correlation matri
                PC1   PC2   PC3   h2    u2 com
height         0.86 -0.37 -0.07 0.88 0.118 1.4
arm.span       0.84 -0.44  0.08 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)



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

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

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

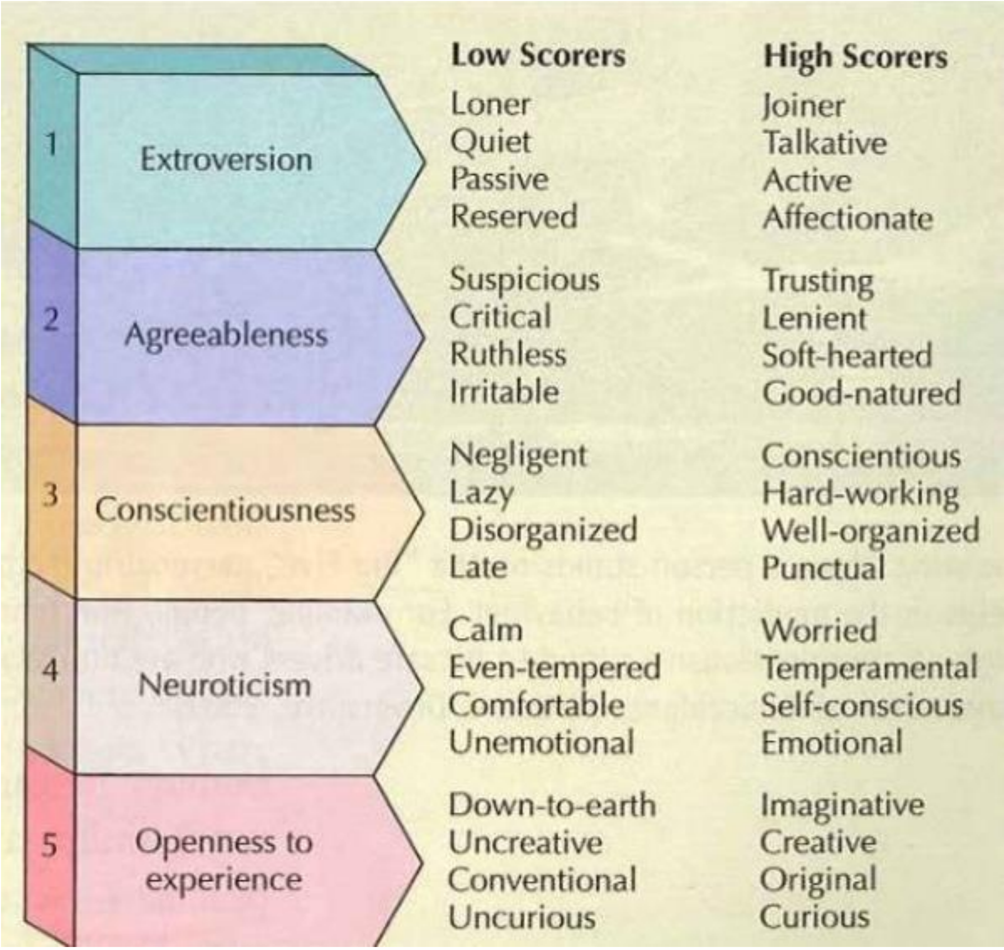
The goal of EFA is to explain the correlations among a set of observed variables by uncovering a smaller set of more fundamental **unobserved variables underlying the data**.

- ✓ 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)

## The Basic Factor Analysis Model

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

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

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

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

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

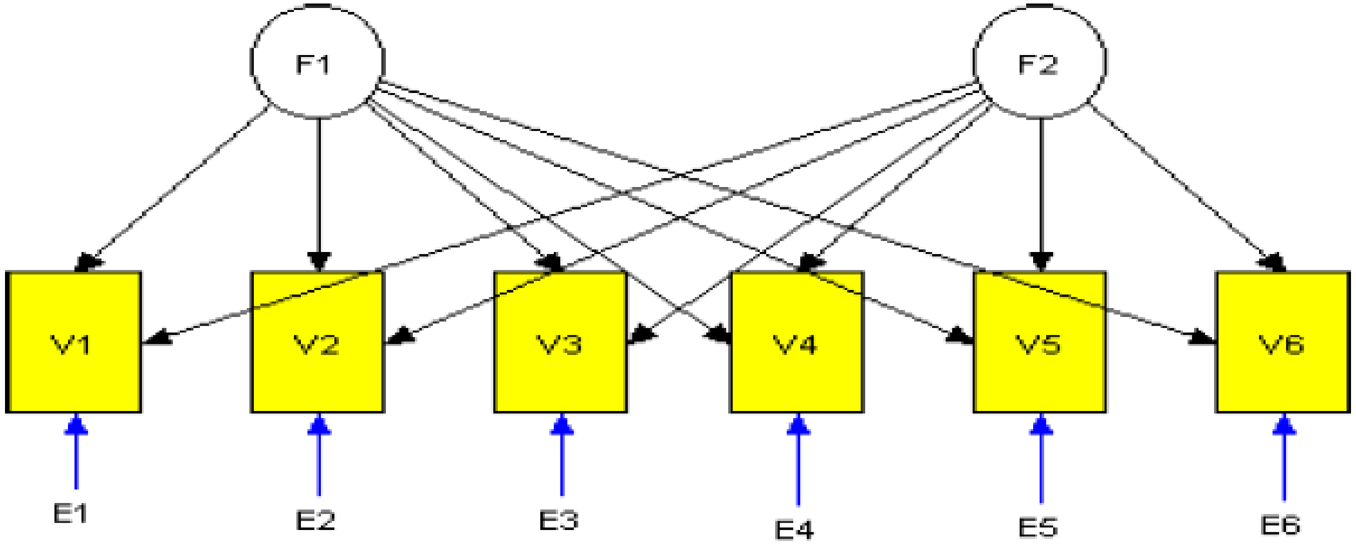
Matrix Form

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

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

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

<u>**$p \times k$ factor loading (pattern) matrix**</u>



# Exploratory Factor Analysis (EFA)

## The Basic Factor Analysis Model

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

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

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

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

<u>**$p \times k$ factor loading (pattern) matrix**</u>

## Assumption

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

Means of <u>**errors**</u> are <u>**zero**</u> and errors are <u>**uncorrelated**</u> of each other

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

Means of <u>**Factors**</u> are <u>**zero**</u> and factors are <u>**Independent**</u> of each other <u>**(orthogonal)**</u>

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

<u>**Common factors and errors are uncorrelated**</u>

# Exploratory Factor Analysis (EFA)

## The Basic Factor Analysis Model

$$\boxed{X} = \mu + AF + \varepsilon,$$

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

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

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

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

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

$a_{ij}$ 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 ( $f_j$)

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

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

$$\boxed{X} = \mu + AF + \varepsilon,$$

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

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

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

<u>**$p \times k$ factor loading (pattern) matrix**</u>

**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

$$\boxed{X} = \mu + AF + \varepsilon,$$

**Uniqueness (specific variance) $\sigma_i^2$**

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

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

$i = 1, \cdots, p$

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

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

$\sigma_i^2$ 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

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

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

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

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

Diagonal

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

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

# Exploratory Factor Analysis (EFA)

## Estimation

The Basic Factor Analysis Model

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

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

Diagonal

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

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

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

Principal component Method

# Exploratory Factor Analysis (EFA)

## Estimation

The Basic Factor Analysis Model

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

Diagonal

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

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

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.

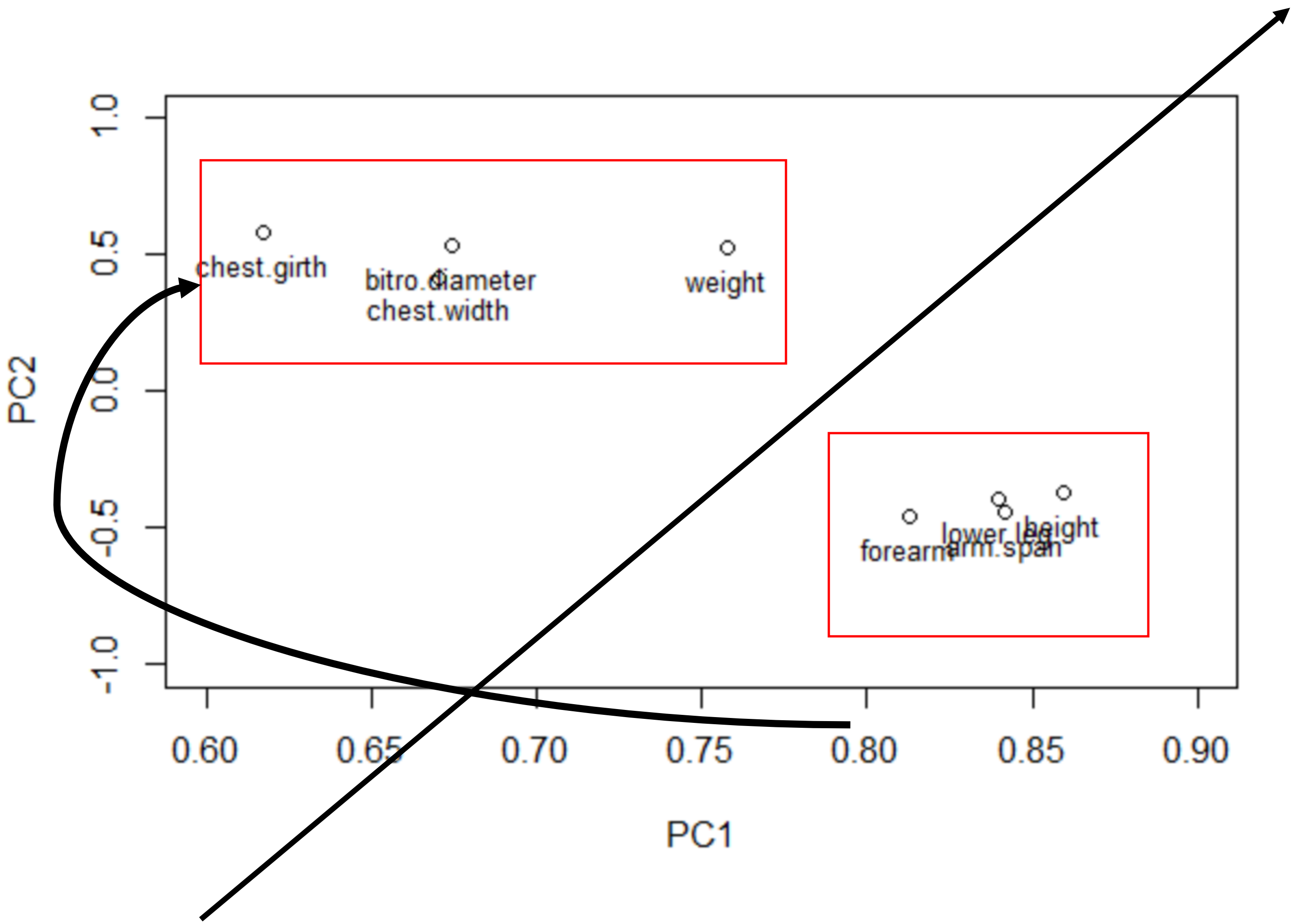
```
                PC1   h2   u2 com
height         0.86 0.74 0.26   1
arm.span       0.84 0.71 0.29   1
forearm        0.81 0.66 0.34   1
lower.leg      0.84 0.70 0.30   1
weight         0.76 0.57 0.43   1
bitro.diameter 0.67 0.45 0.55   1
chest.girth    0.62 0.38 0.62   1
chest.width    0.67 0.45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

The Basic Factor Analysis Model

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

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

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

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

$$X = A\Gamma\boxed{Z} + \varepsilon,$$ **Factor**

$A\Gamma$ ------ Loading matrix

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

# 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

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

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

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

Let $Z = \Gamma^T F$ $\quad \cancel{\Gamma^T\Gamma = I}$ **Oblique rotations**

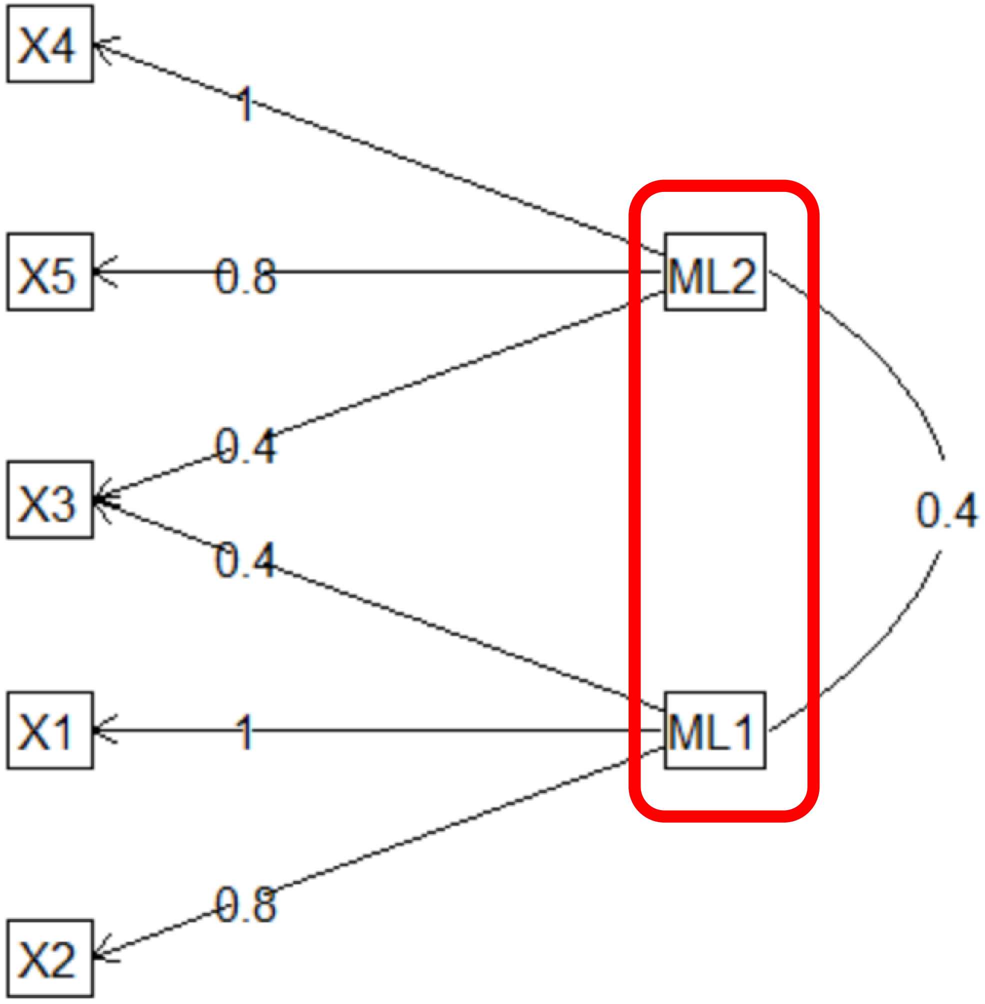
$X = A\Gamma\boxed{Z} + \varepsilon,$ **Factor** $\quad A\Gamma$ ------ Loading matrix

$Var(Z) = \Phi$

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

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

# 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**

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)

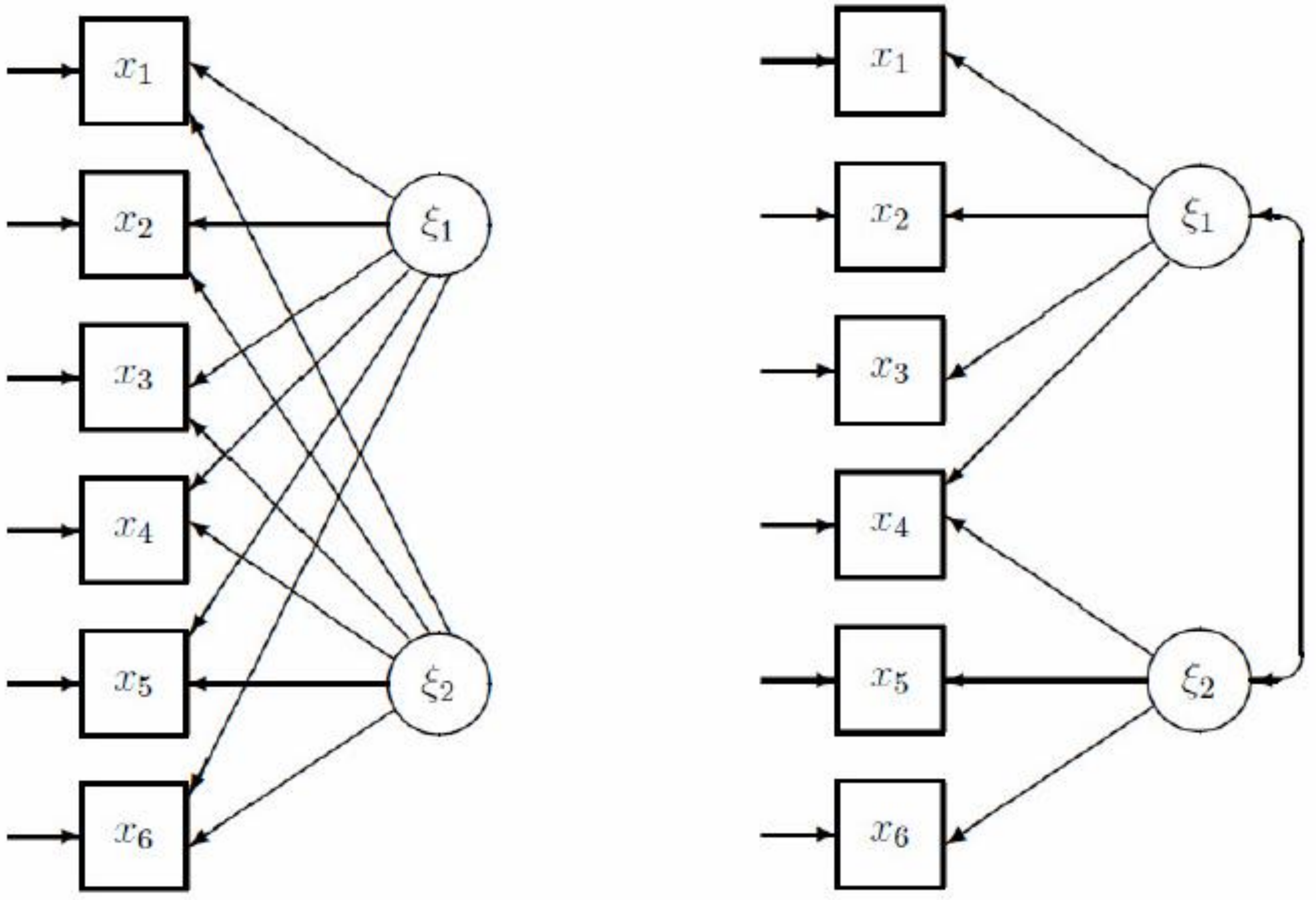
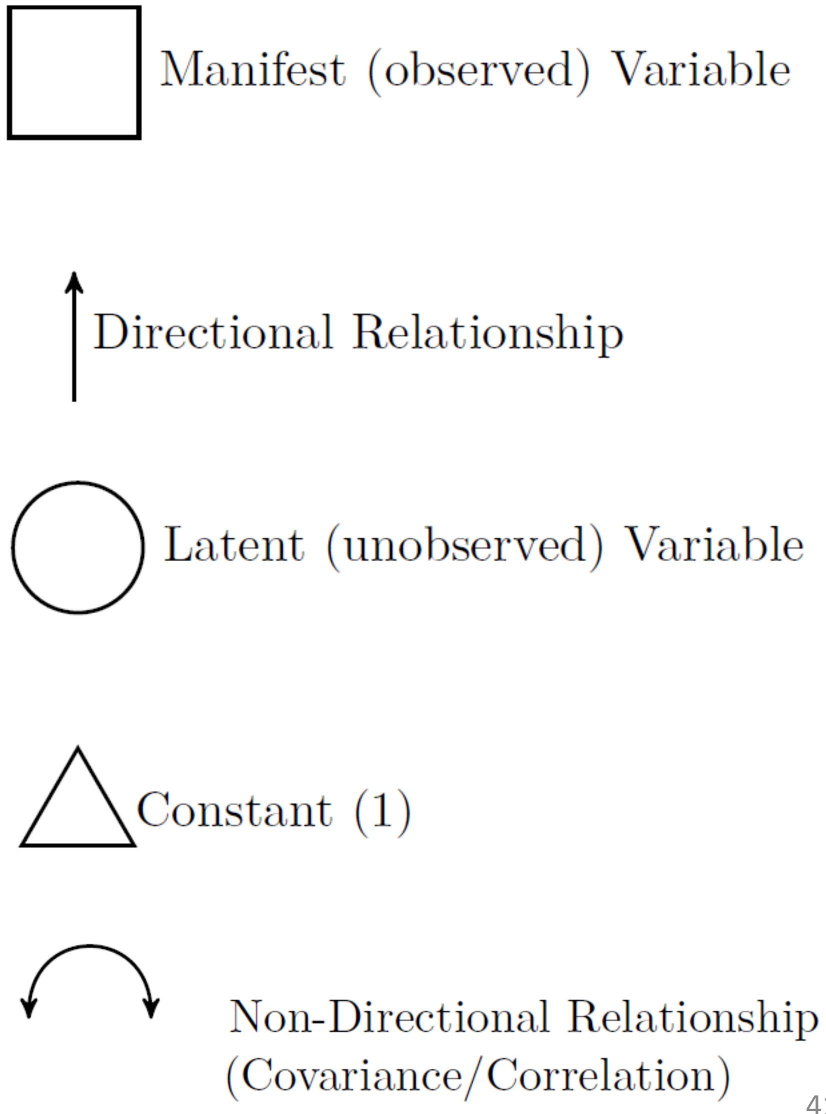



Figure 1: Exploratory Factor Analysis Figure 2: Confirmatory Factor Analysis



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

# Section 2: Confirmatory Factor Analysis and Structural Equation Models

# Confirmatory Factor Analysis (CFA)

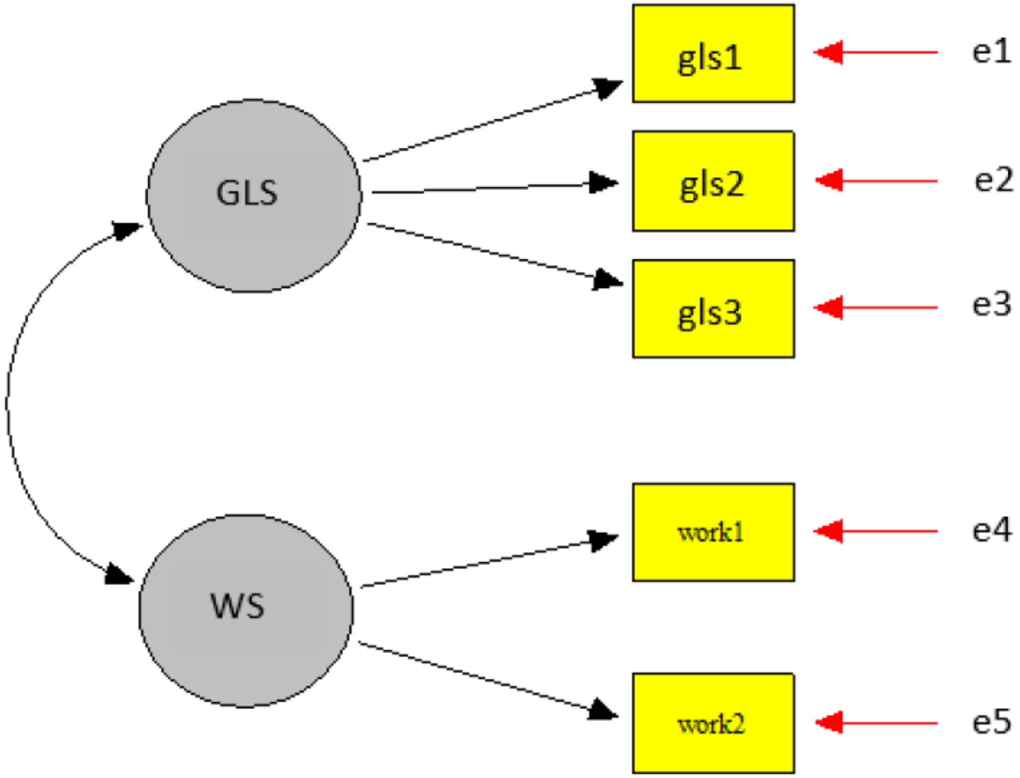
## An Example: Subjective Well Being (SWB) Model

- 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**

$$F_1\text{=GLS},\ F_2\text{=WS}$$

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

# Confirmatory Factor Analysis (CFA)

## 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
$\mu$ is $p \times 1$ vector of intercepts (means of $v$)
$\Lambda$ 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 $\Lambda, \Psi, \Theta$**

## 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)

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

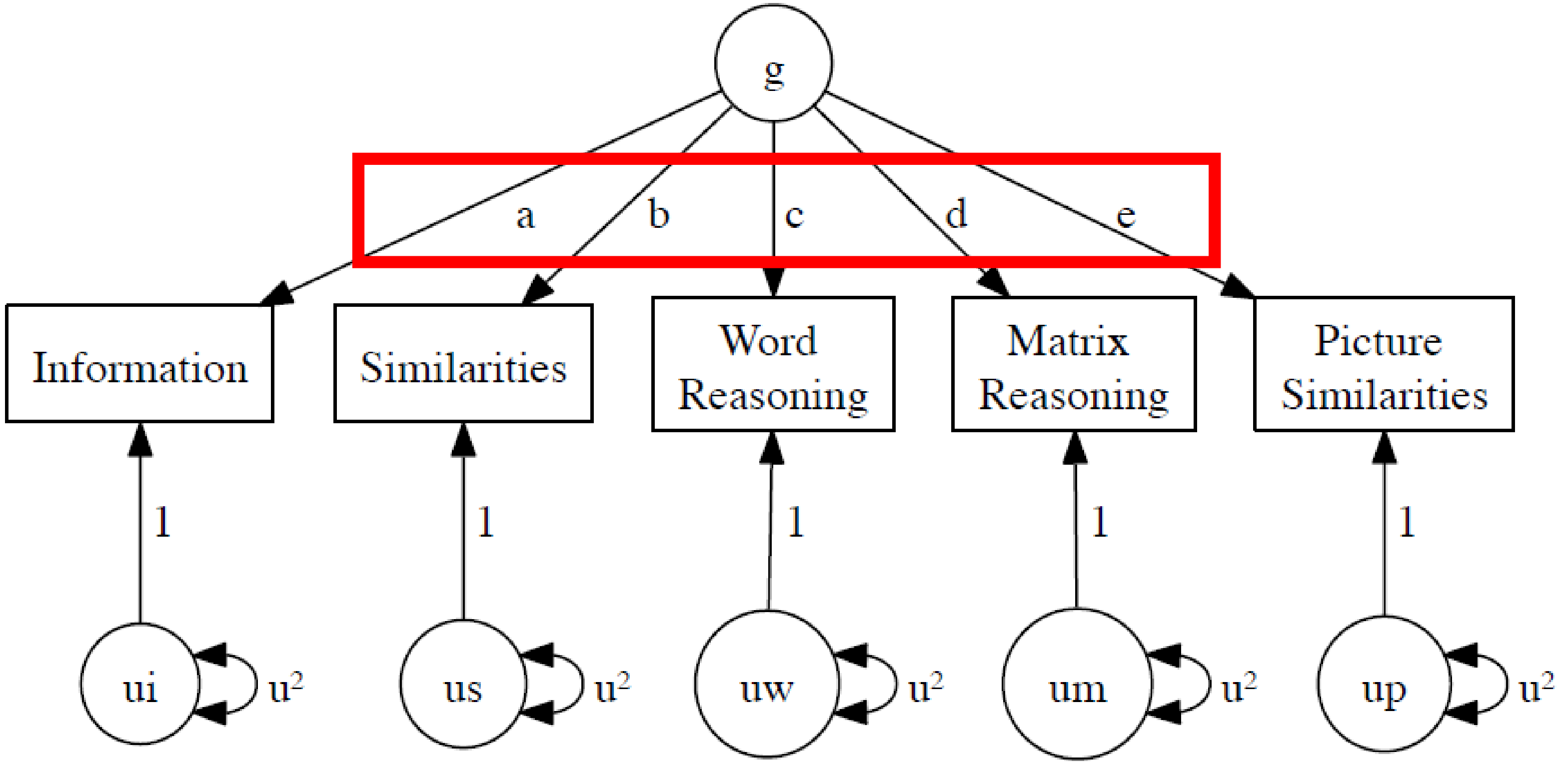
Means of **Factors** are **zero,** $\Psi$ is a **general covariance matrix**

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

**Common factors and errors are uncorrelated**

# Confirmatory Factor Analysis (CFA)

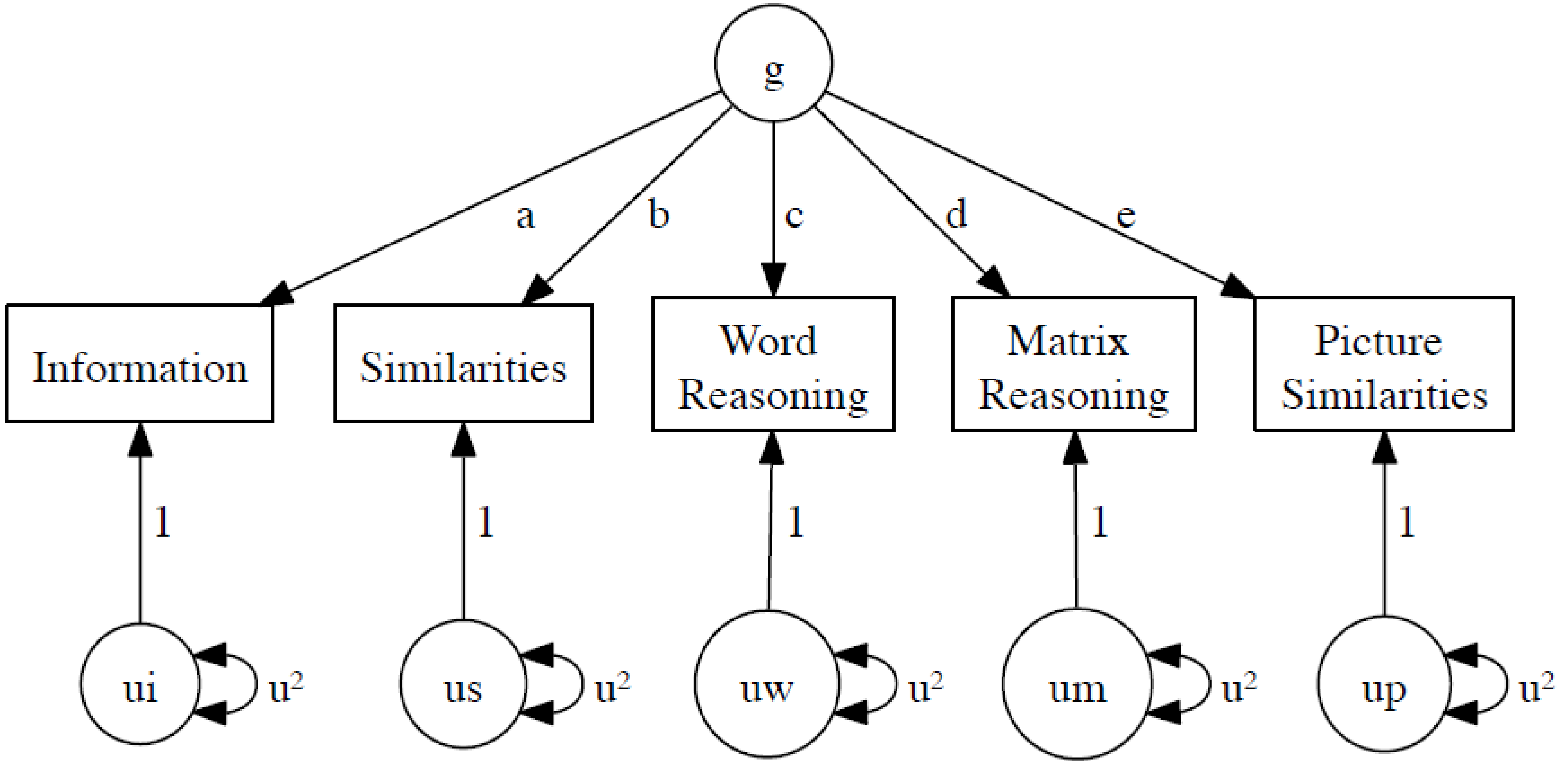
**Wechsler Intelligence Scale for Children-Fourth Edition subscales**




- 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 $Var(g) = 1$

$$\mathit{Communality}$$

$$h_1^2 = a^2$$

$$h_2^2 = b^2$$

$$h_3^2 = c^2$$

$$\ldots$$

$$\mathit{Uniqueness}$$

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

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

$$\ldots$$

# Confirmatory Factor Analysis (CFA)

## Wechsler Intelligence Scale for Children-Fourth Edition subscales

Correlations for the WISC-IV data

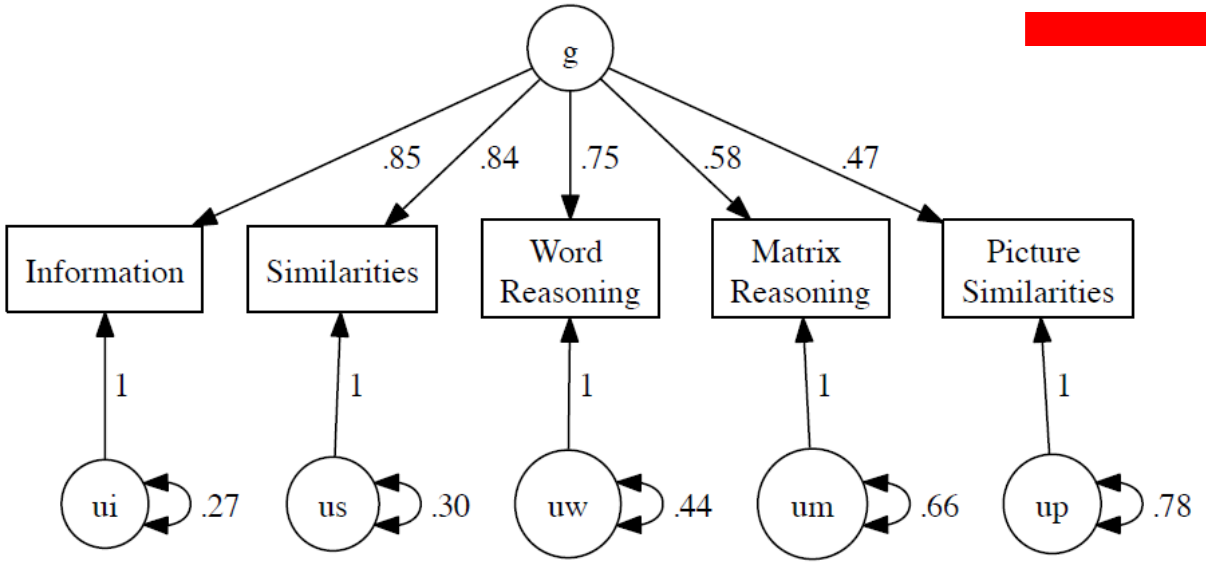
| | Info | Sim | Word Reas | Matrix Reas | Picture Sim |
|---|---|---|---|---|---|
| inss | 1.00 | 0.72 | 0.64 | 0.51 | 0.37 |
| siss | 0.72 | 1.00 | 0.63 | 0.48 | 0.38 |
| wrss | 0.64 | 0.63 | 1.00 | 0.37 | 0.38 |
| mrss | 0.51 | 0.48 | 0.37 | 1.00 | 0.38 |
| psss | 0.37 | 0.38 | 0.38 | 0.38 | 1.00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

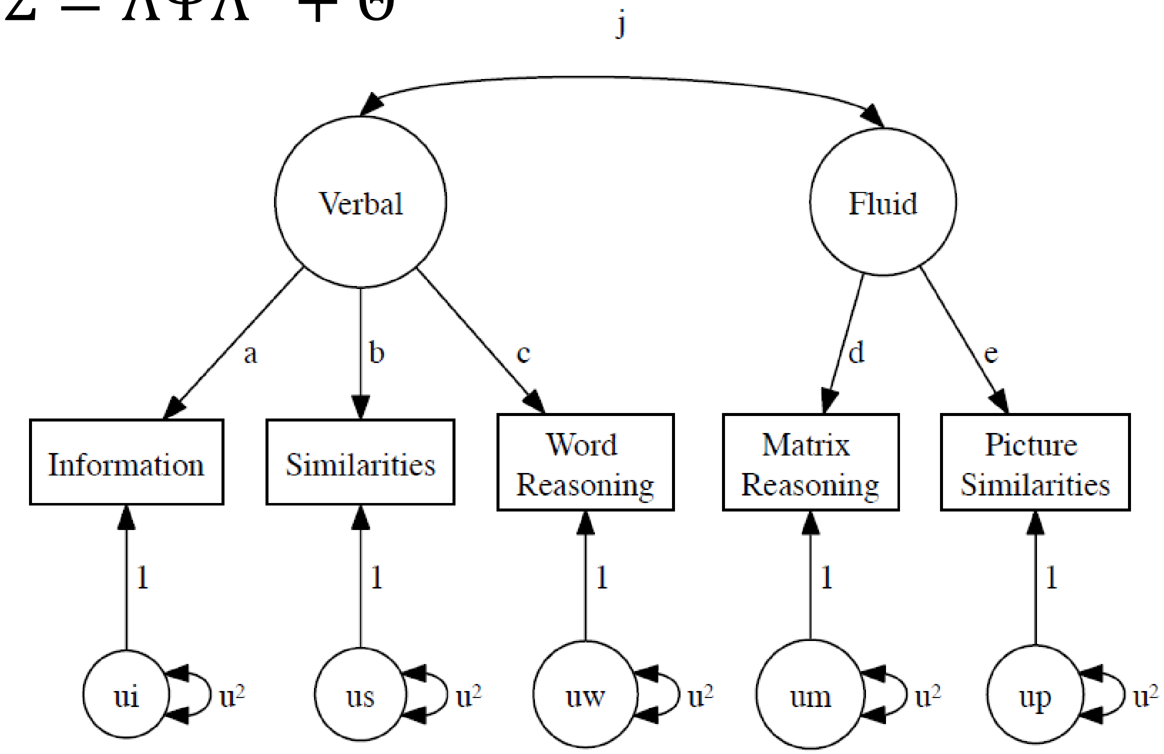
```
> fa(R, nfactors=1,rotate="none",n.obs=550,fm="mle")
Factor Analysis using method =  ml
Call: fa(r = R, nfactors = 1, n.obs = 550, rotate = "none", f
Standardized loadings (pattern matrix) based upon correlation
        ML1   h2   u2 com
Info   0.86 0.74 0.26   1
Sim    0.84 0.70 0.30   1
Word  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

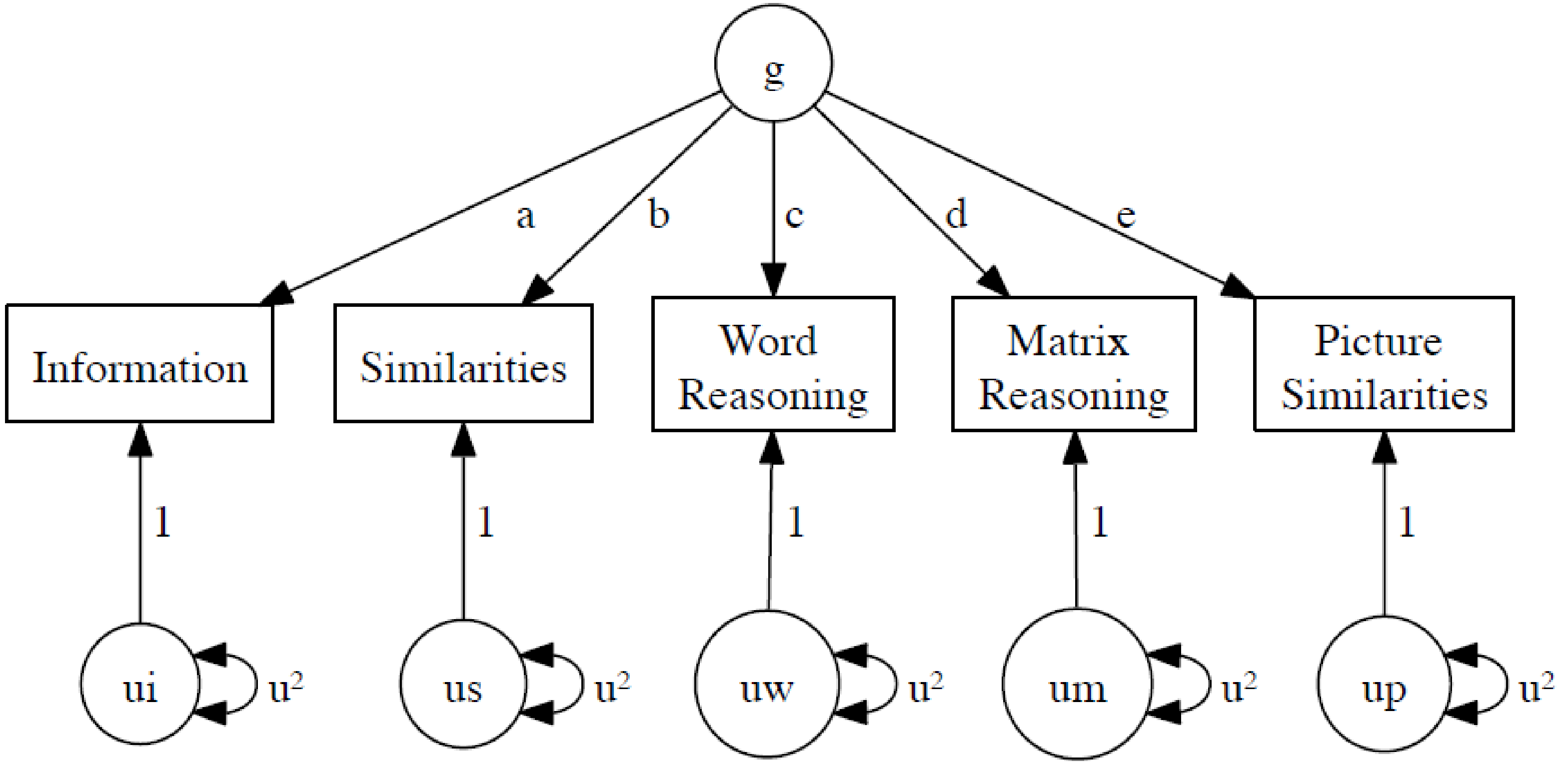
$$\Sigma = \Lambda\Lambda^T + \Theta \qquad \Sigma = \Lambda\Psi\Lambda^T + \Theta$$



# Confirmatory Factor Analysis (CFA)

**Wechsler Intelligence Scale for Children-Fourth Edition subscales**




**Identification**

Assume $Var(g) = 2$

$$Communality$$
$$h_1^2 = a'^2 = 2a^2$$
$$h_2^2 = b'^2 = 2b^2$$
$$h_3^2 = c'^2 = 2c^2$$
$$\ldots$$

$$Uniqueness$$
$$u_1^2 = 1 - a'^2$$
$$u_2^2 = 1 - b'^2$$
$$\ldots$$

# 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
$\mu$ is $p \times 1$ vector of intercepts (means of $v$)
$\Lambda$ 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
$$\Lambda^* = \Lambda D$$
$$\Psi^* = D^{-1}\Psi D^{-1'}$$
$$\Theta^* = \Theta$$

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

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

- The parameters cannot be **uniquely determined** even $\Sigma$ 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.

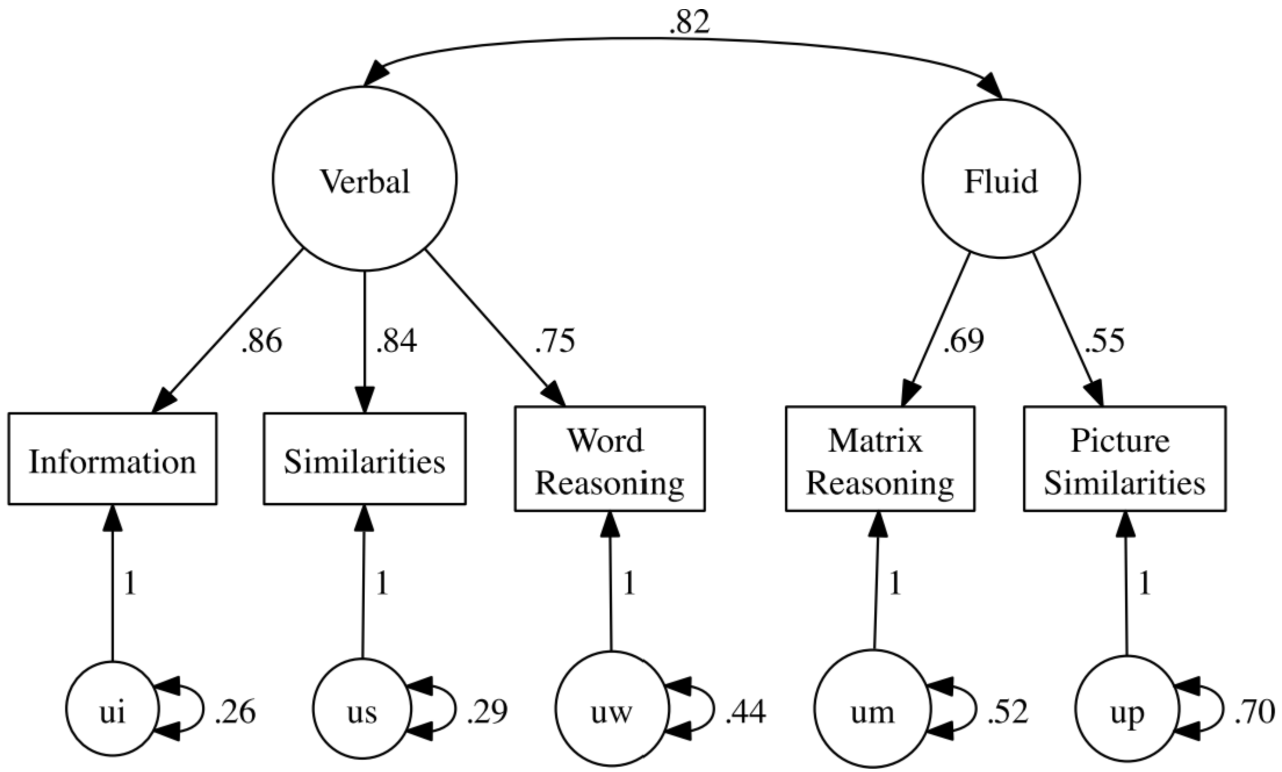
1. **Standardized latent variable.** This method constrains the **latent variable's variance to 1.0**. This, in effect, makes the latent variable a standardized variable. Moreover, if there is more than one Latent Variables, then the covariance among the Latent Variables becomes a **correlation**.
2. **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

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

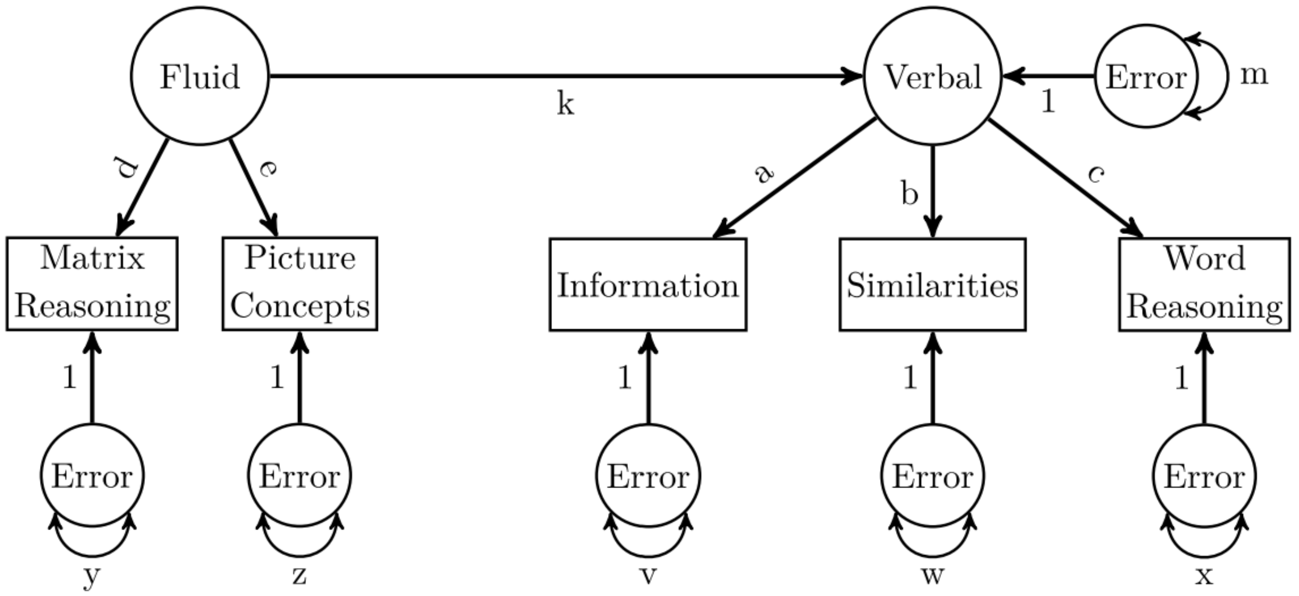
$$\underset{(5\times 2)}{\boldsymbol{\Lambda}} = \begin{bmatrix} .86 & 0 \\ .84 & 0 \\ .75 & 0 \\ 0 & .69 \\ 0 & .55 \end{bmatrix}, \ \& \ \underset{(2\times 2)}{\boldsymbol{\Psi}} = \begin{bmatrix} 1 & .82 \\ .82 & 1 \end{bmatrix}$$

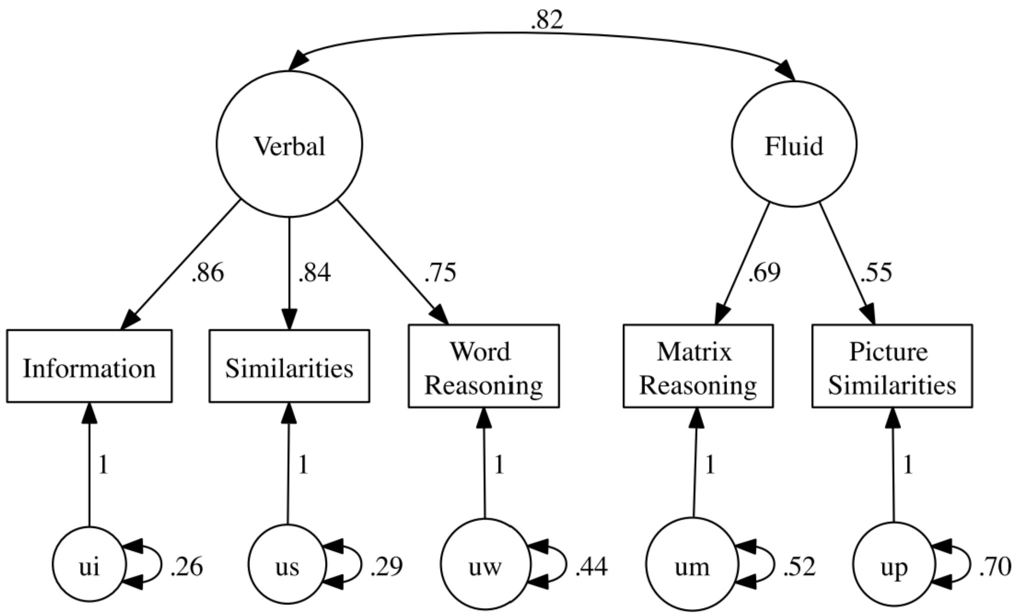


# Confirmatory Factor Analysis (CFA)

**Wechsler Intelligence Scale for Children-Fourth Edition subscales**

Regression $Verbal = k * Fluid + error$ **Structural Equation Model**



# 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.

$$
\boxed{\begin{aligned}
\mathrm{PCr} &= \alpha_1 \mathrm{SBP} + \alpha_2 \mathrm{DBP} + \alpha_3 \mathrm{BMI} + \alpha_4 \mathrm{WHR} + \epsilon_1 \\
\mathrm{ACR} &= \beta_1 \mathrm{SBP} + \beta_2 \mathrm{DBP} + \beta_3 \mathrm{BMI} + \beta_4 \mathrm{WHR} + \epsilon_2
\end{aligned}}
$$

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)

$$\mathrm{PCr} = \mu_1 + \lambda_{11}\mathrm{KD} + \epsilon_1, \quad \mathrm{DBP} = \mu_4 + \lambda_{42}\mathrm{BP} + \epsilon_4$$

$$\mathrm{ACR} = \mu_2 + \lambda_{21}\mathrm{KD} + \epsilon_2, \quad \mathrm{BMI} = \mu_5 + \lambda_{53}\mathrm{OB} + \epsilon_5$$

$$\mathrm{SBP} = \mu_3 + \lambda_{32}\mathrm{BP} + \epsilon_3, \quad \mathrm{WHR} = \mu_6 + \lambda_{63}\mathrm{OB} + \epsilon_6$$

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

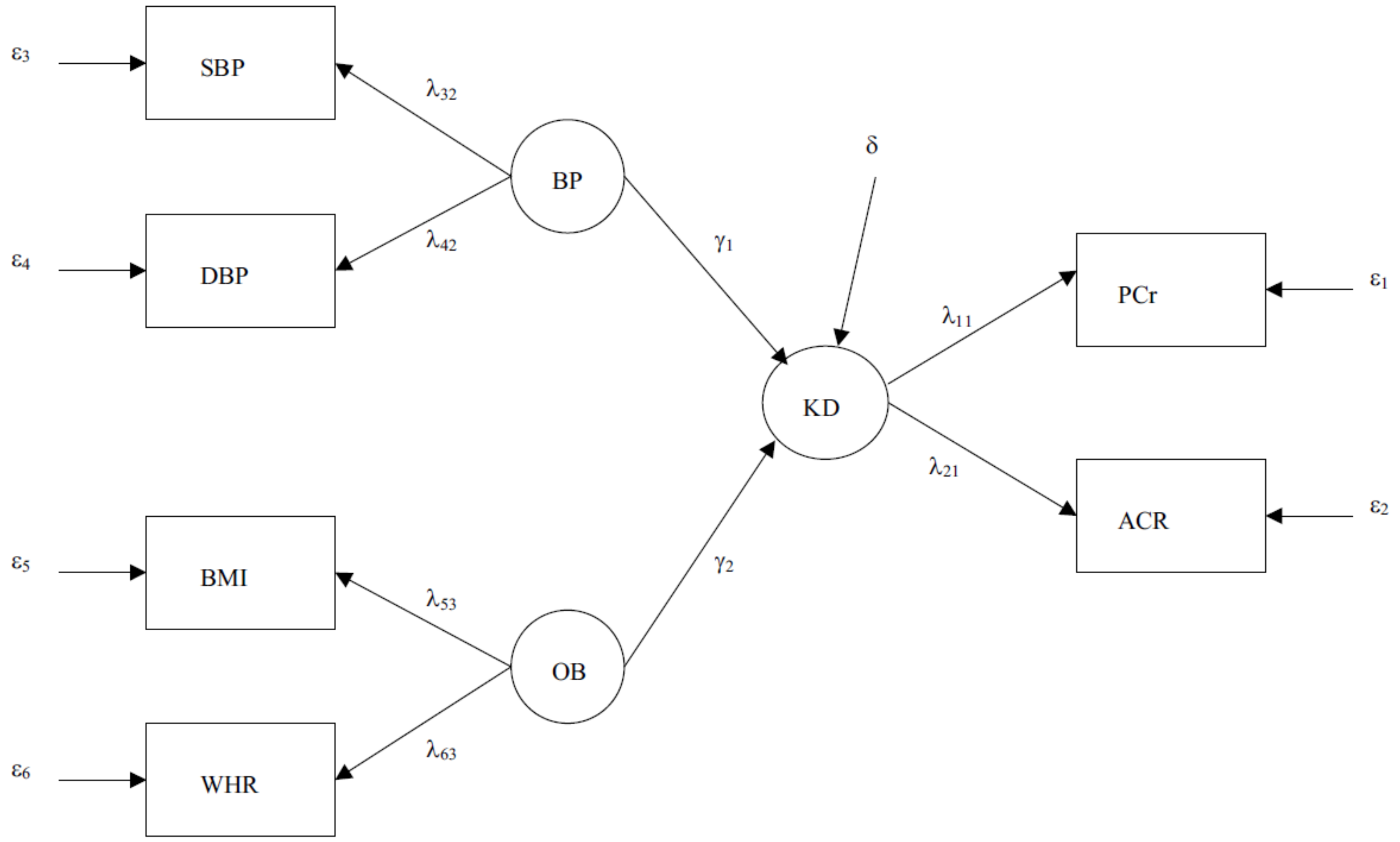
$$\begin{bmatrix} \mathrm{PCr} \\ \mathrm{ACR} \\ \mathrm{SBP} \\ \mathrm{DBP} \\ \mathrm{BMI} \\ \mathrm{WHR} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & \lambda_{32} & 0 \\ 0 &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

**Structural Equation**

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

# Structural Equation Model (SEM)

## Type 2 diabetic patients data



# Structural Equation Model (SEM)

## Matrix Form

Outcome latent variables

$$\omega = (\boxed{\eta^T}\ \boxed{\xi^T})^T$$

$\eta$ and $\xi$ are $q_1 \times 1$ and $q_2(= q - q_1) \times 1$ random vectors

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

Explanatory latent variables

latent variables

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

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

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

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

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

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

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

$\epsilon$ and $\delta$ 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**

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

**Method 2**

allow $\lambda_{11}$, $\lambda_{52}$, and/or $\lambda_{83}$ in $\boldsymbol{\Lambda}$ to be unknown parameters and fix the diagonal elements of $\boldsymbol{\Phi}^+$ as 1

hence $\boldsymbol{\Phi}^+$ is a correlation matrix

**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)

## 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** $\eta = \boxed{\mathbf{Bd}} + \Gamma\xi + \delta$

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

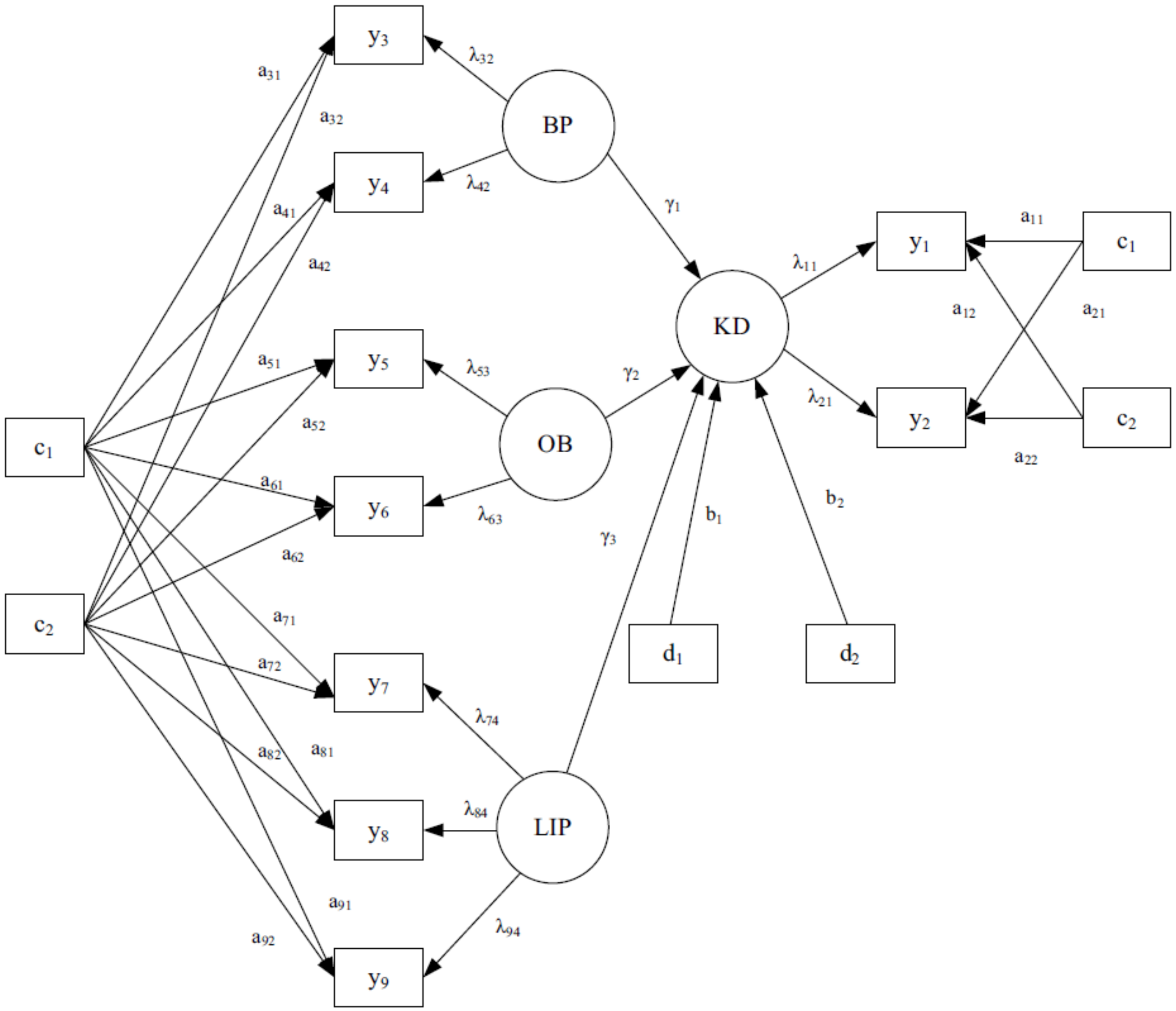
$$
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \\ a_{71} & a_{72} \\ a_{81} & a_{82} \\ a_{91} & a_{92} \end{bmatrix} \begin{bmatrix}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

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

where $a_{jk}$, $\lambda_{jk}$, $b_1$, $b_2$, $\gamma_1$, $\gamma_2$, and $\gamma_3$ are unknown regression coefficients

# Structural Equation Model (SEM)

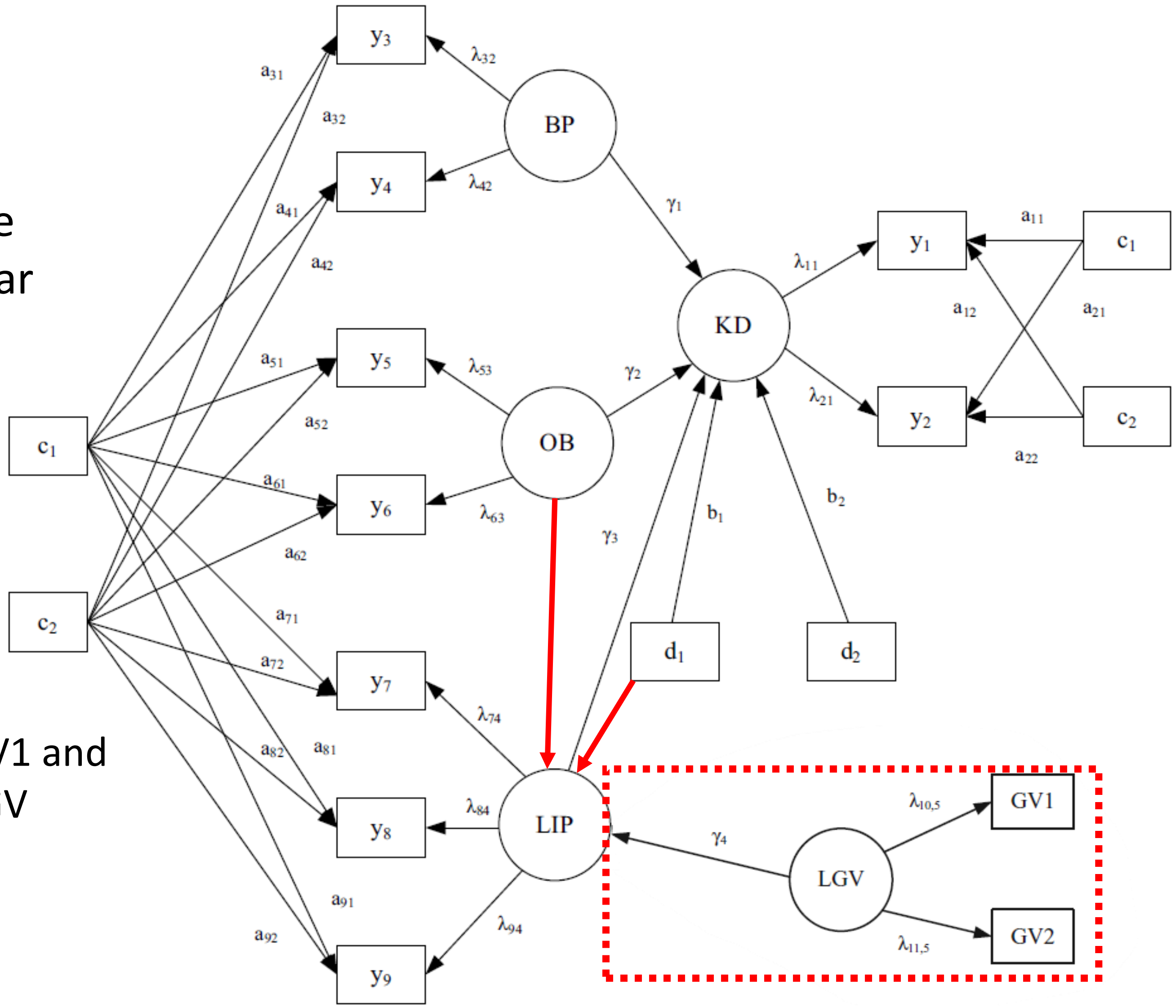
## Type 2 diabetic patients data



# 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 $\Pi$. Particularly useful in business and social-psychological research

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

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

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

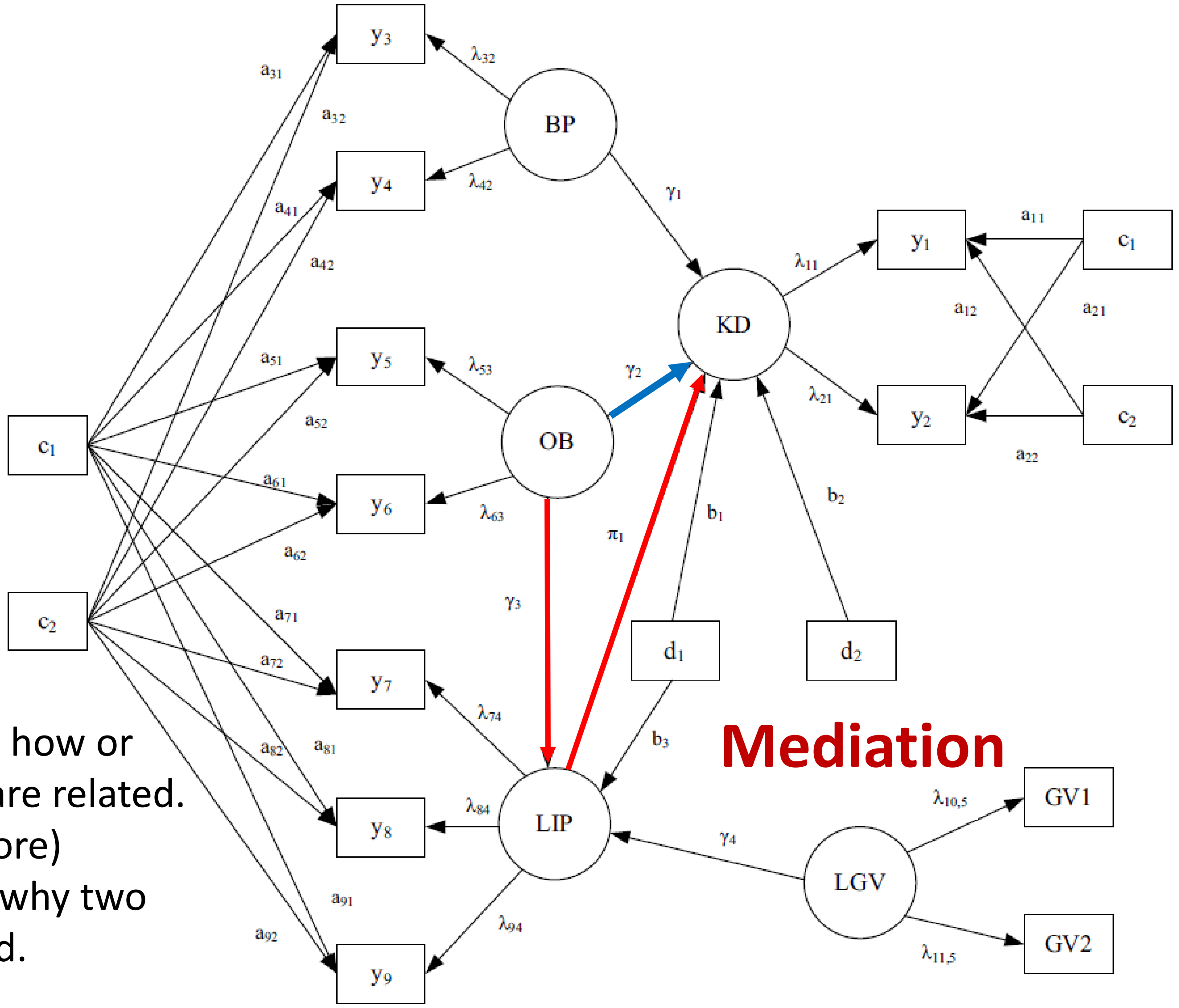
$\mathbf{I} - \mathbf{\Pi}$ is nonsingular

diagonal elements of $\mathbf{\Pi}$ are zero

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

# Structural Equation Model (SEM)

## Extension

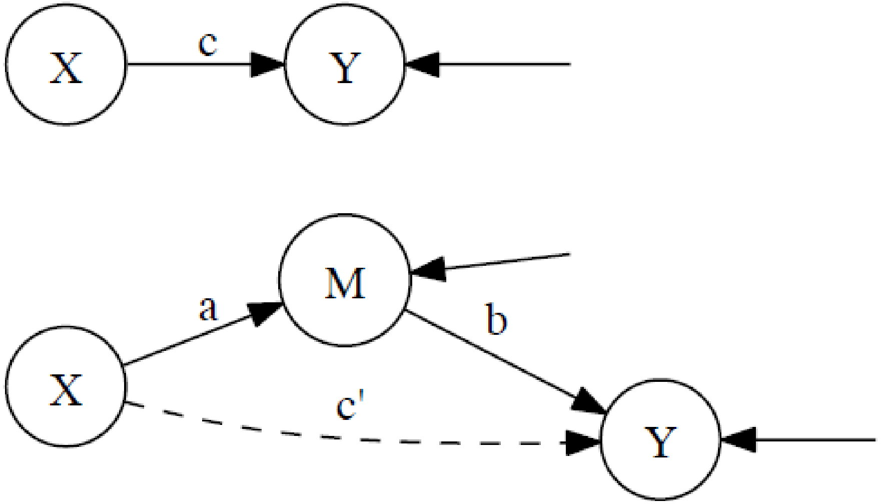



- Mediation models investigate how or why two (or more) variables are related.
- Mediation is when one (or more) variables explains the reason why two (or more variables) are related.

# Structural Equation Model (SEM)

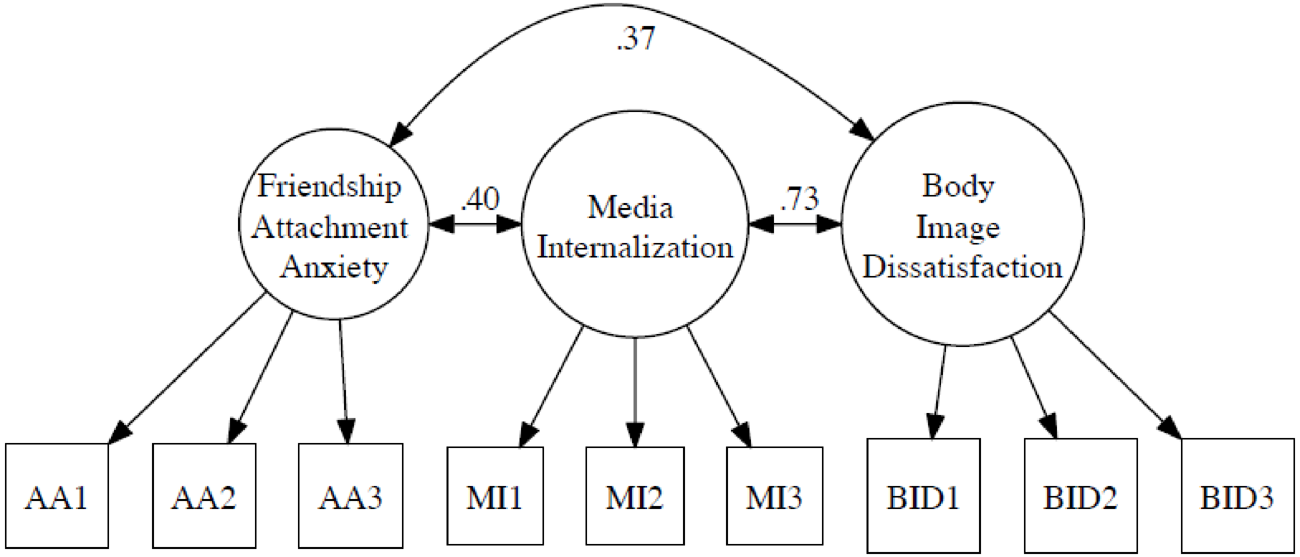
## Mediation

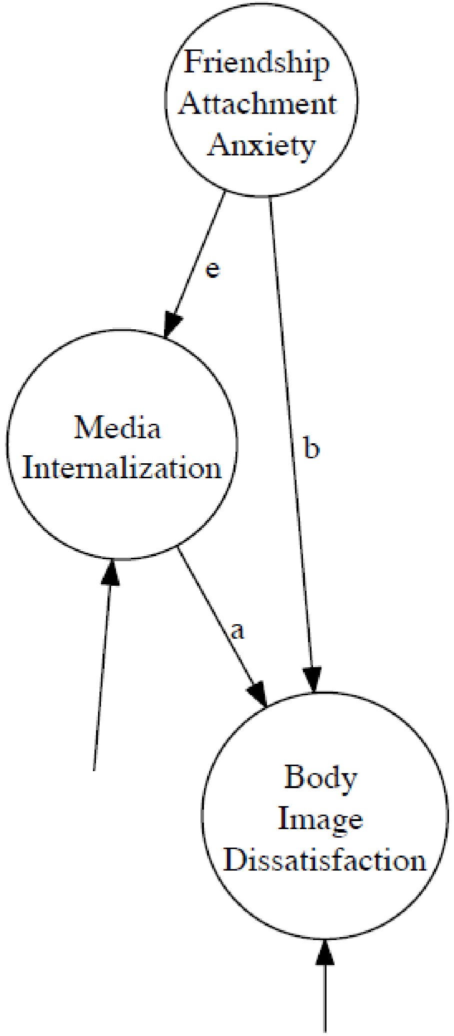
1. First, there is a relationship (via $c$) between variables $X$ and $Y$
2. Then, $M$ is put into the model and is related to both $X$ (via $a$) and $Y$ (via $b$).
3. After $M$ was put into the model, then the relationship between $X$ and $Y$ dwindles (i.e., $c' < c$).



# Structural Equation Model (SEM)

## Mediation

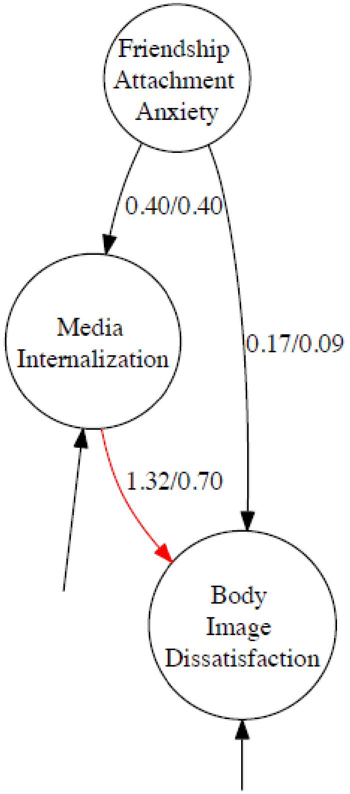



Media internalization (awareness and attitudes toward prevailing sociocultural standards of attractiveness) was hypothesized to mediate the positive association between attachment anxiety in friendships and body image dissatisfaction

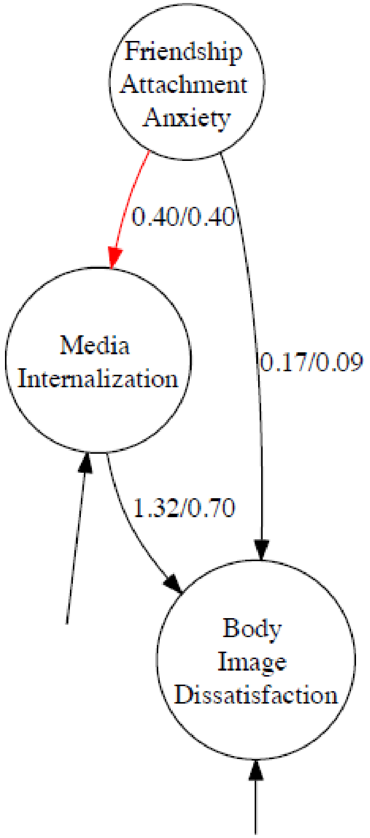
# Structural Equation Model (SEM)

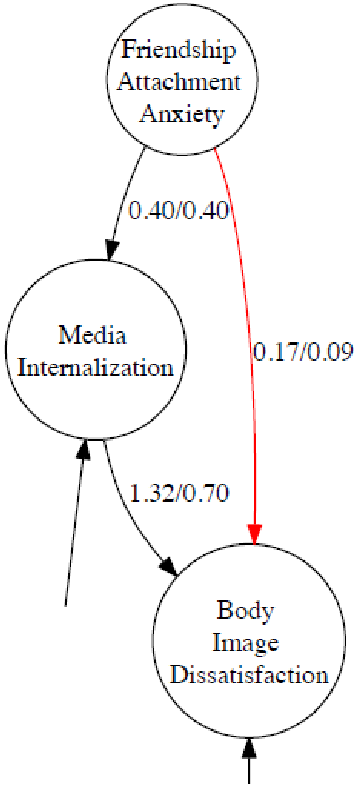
## Mediation




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

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



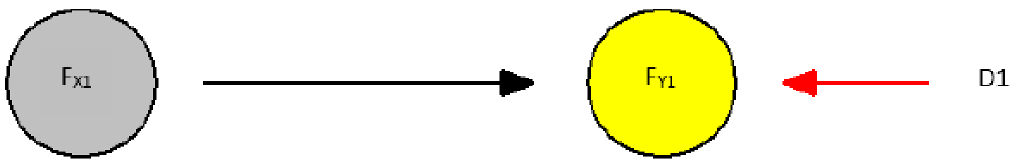




The attachment anxiety-body image dissatisfaction relationship (path b), dwindles to almost 0 (b = .09) in the presence of these variables

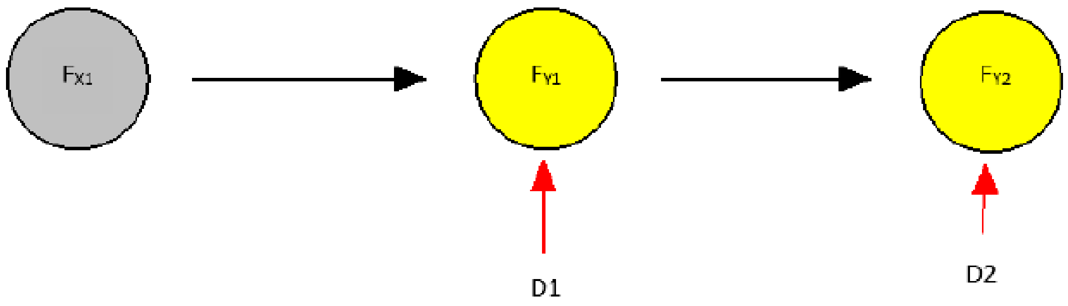
# Structural Equation Model (SEM)

## Mediation

• ***Direct effect***: Influence of one variable on another that is unmediated by any other variables in a path model



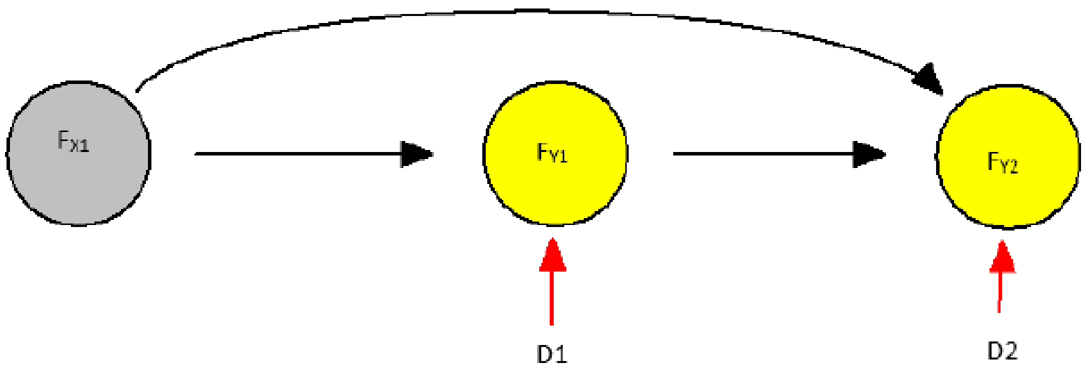

• ***Indirect effect:*** Influence of one variable on another is mediated by at least one intervening variable (mediator)



# Structural Equation Model (SEM)

## Mediation

- ***Total effect:*** The total of the direct effect and all indirect effects




$F_{X1}$ to $F_{Y2}$: direct effect $= \gamma_{21}$
indirect effect $= \gamma_{11}\beta_{21}$
total effect $= \gamma_{21} + \gamma_{11}\beta_{21}$

End of Chapter 1&2