Some Multivariate techniques
Principal components analysis (PCA)
Factor analysis (FA)
Structural equation models (SEM)
Applications: Personality

Boulder
March 2006

Dorret I. Boomsma
Danielle Dick
Marleen de Moor
Mike Neale
Conor Dolan

Presentation in dorret\2006

Multivariate statistical methods; for example

-Multiple regression

-Fixed effects (M)ANOVA

-Random effects (M)ANOVA

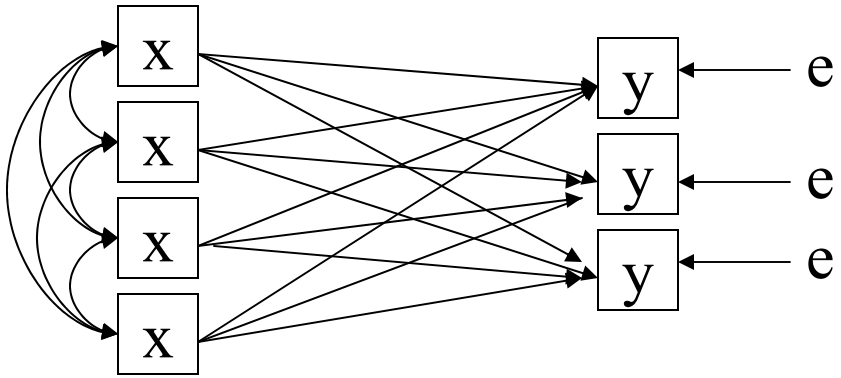
-Factor analysis / PCA

-Time series (ARMA)

-Path / LISREL models

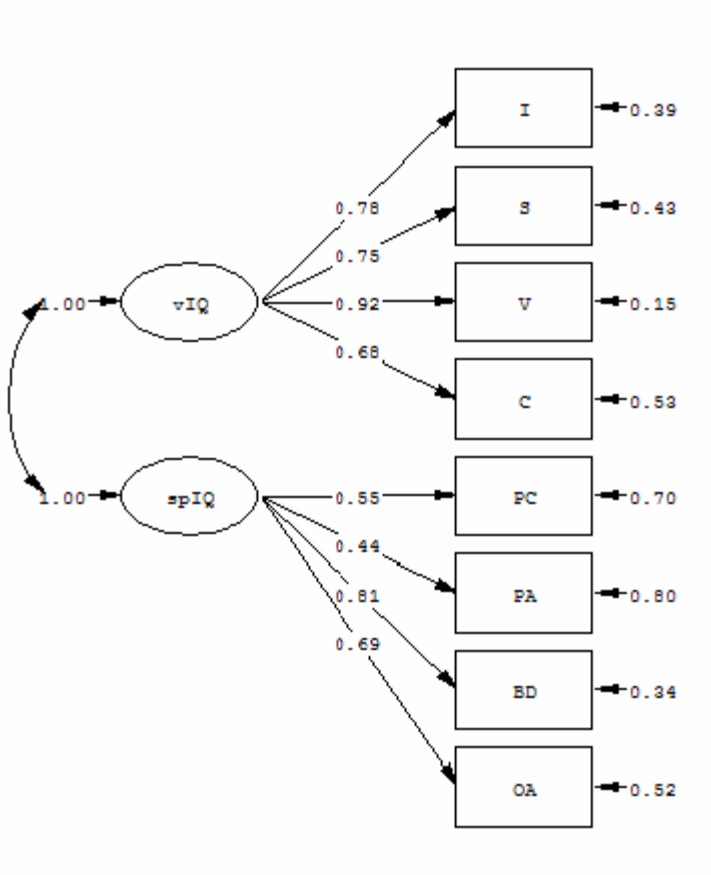
Multiple regression

x predictors (independent), e residuals, y dependent;
both x and y are observed

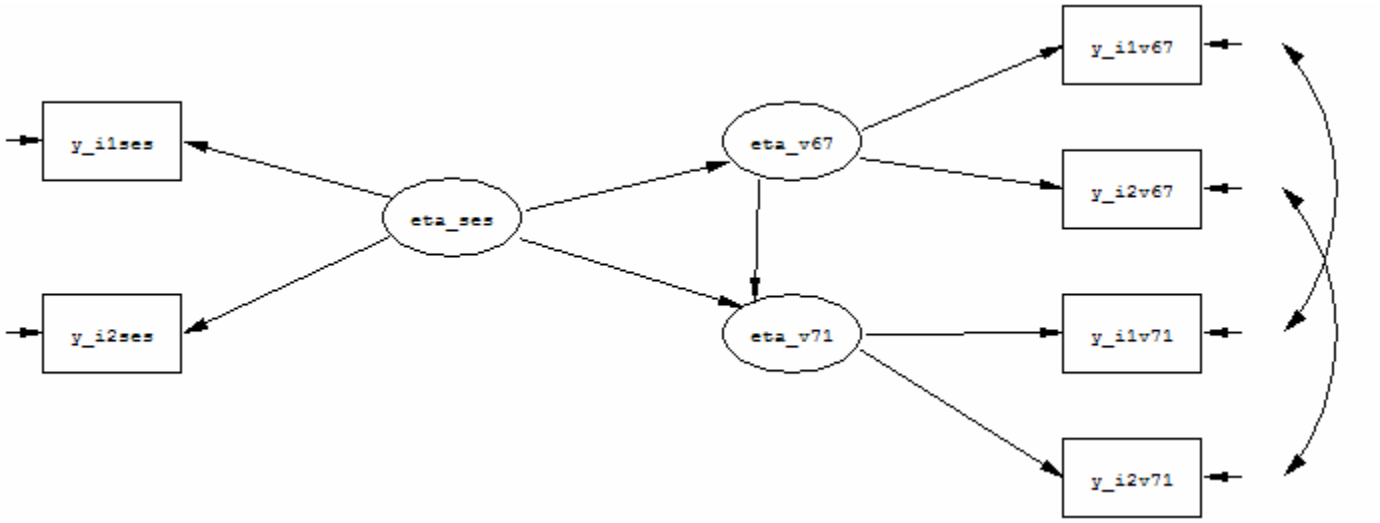


Factor analysis:

measured and unmeasured (latent) variables. Measured variables can be “indicators” of unobserved traits.



Path model / SEM model



Latent traits can influence other latent traits

Measurement and causal models in non-experimental research

- Principal component analysis (PCA)
- Exploratory factor analysis (EFA)
- Confirmatory factor analysis (CFA)
- Structural equation models (SEM)
- Path analysis

These techniques are used to analyze multivariate data that have been collected in *non-experimental* designs and often involve *latent constructs* that are not directly observed.

These latent constructs underlie the observed variables and account for inter-correlations between variables.

Models in non-experimental research

All models specify a covariance matrix Σ and means vector μ:

$$\Sigma = \Lambda\Psi\Lambda^t + \Theta$$

total covariance matrix [Σ] =
factor variance [$\Lambda\Psi\Lambda^t$] + residual variance [Θ]

means vector μ can be modeled as a function of other (measured) traits e.g. sex, age, cohort, SES

Outline

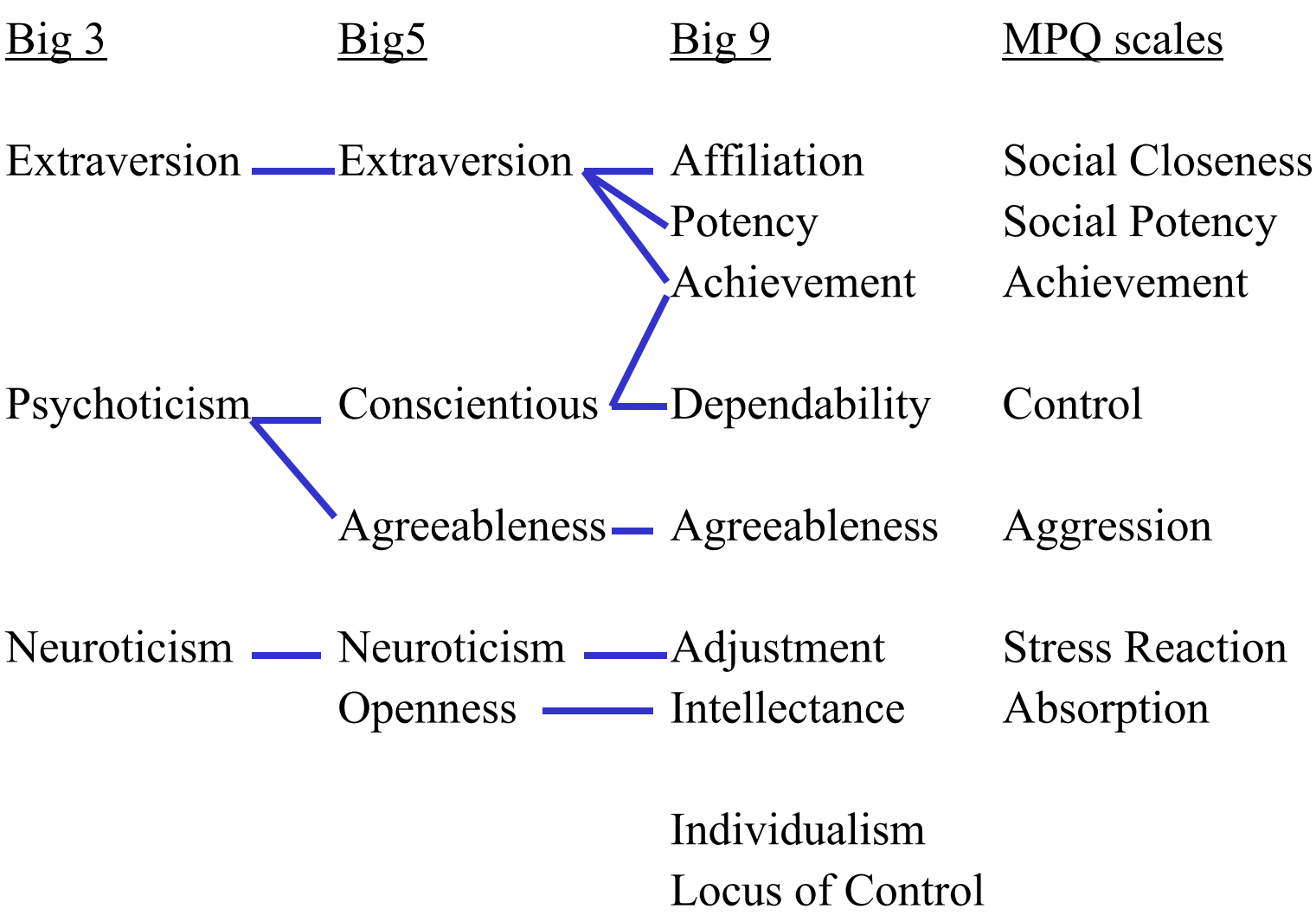
- Cholesky decomposition
- PCA (eigenvalues)
- Factor models (1,..4 factors)
- Application to personality data
- Scripts for Mx, [Mplus, Lisrel]

Application: personality

- Personality (Gray 1999): a person's general style of interacting with the world, especially with other people – whether one is withdrawn or outgoing, excitable or placid, conscientious or careless, kind or stern.

- Is there one underlying factor?
- Two, three, more?

Personality: Big 3, Big 5, Big 9?



Data:

Neuroticism, Somatic anxiety, Trait Anxiety, Beck Depression, Anxious/Depressed, Disinhibition, Boredom susceptibility, Thrill seeking, Experience seeking, Extraversion, Type-A behavior, Trait Anger, Test attitude (13 variables)

Software scripts

- **Mx** MxPersonality (also includes data)
- (Mplus) Mplus
- (Lisrel) Lisrel

- **Copy from dorret\2006**

Cholesky decomposition for 13 personality traits

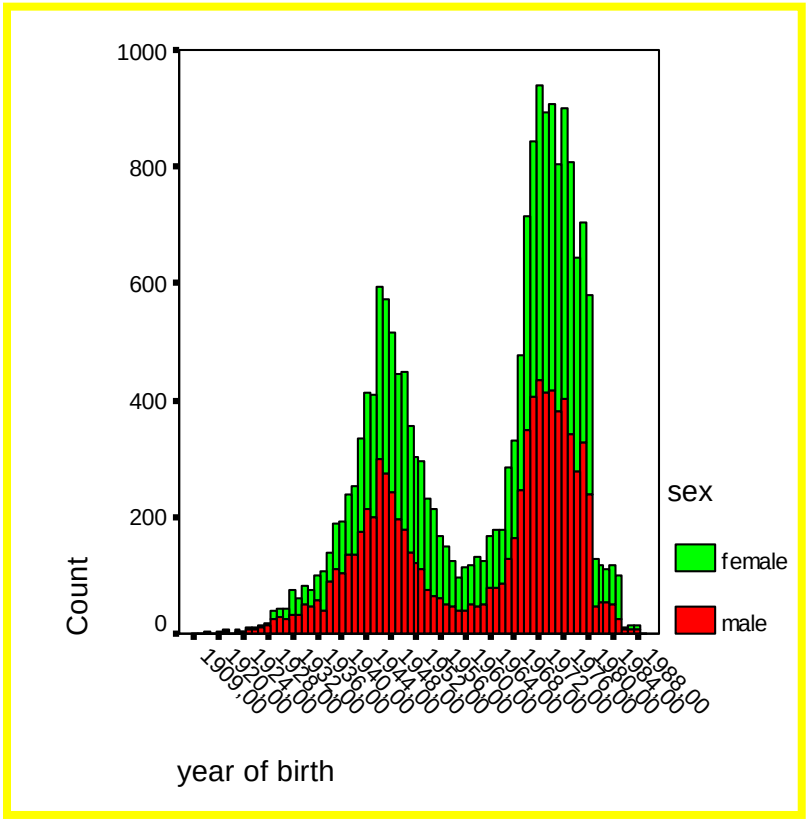
Cholesky decomposition: S = Q Q'
where Q = lower diagonal (triangular)

For example, if S is 3 x 3, then Q looks like:

$$\begin{pmatrix} f_{11} & 0 & 0 \\ f_{21} & f_{22} & 0 \\ f_{31} & f_{32} & f_{33} \end{pmatrix}$$

I.e. # factors = # variables, this approach gives a transformation of S; completely determinate.

Subjects: Birth cohorts (1909 – 1989)



Four data sets were created:

1 Old male	(N = 1305)
2 Young male	(N = 1071)
3 Old female	(N = 1426)
4 Young female	(N = 1070)

What is the structure of personality?
Is it the same in all datasets?

Total sample: 46% male, 54% female

Application: Analysis of Personality in twins, spouses, sibs, parents from Adult Netherlands Twin Register: longitudinal participation

	1x	**2x**	**3x**	**4x**	**5x**	**6x**	**Total**
Twin	2835	2189	1471	1145	867	446	8953
Sib	1069	844	611	323			2847
Father	955	664	725	402			2739
Mother	1071	696	797	468	1		3033
Spouse of twin	1598	352					1950
Total	**7528**	**4745**	**3604**	**5942**	**868**	**446**	**19529**

Data from multiple occasions were averaged for each subject;
Around 1000 Ss were quasi-randomly selected for each sex-age group

Because it is March 8, we use data set 3 (personShort sexcoh3.dat)

dorret\2006\Mxpersonality (docu.doc)

- **Datafiles** for Mx (and other programs; free format)
- personShort_sexcoh1.dat old males N=1035 (average yr birth 1943)
- personShort_sexcoh2.dat young males N=1071 (1971)
- personShort_sexcoh3.dat old females N=1426 (1945)
- personShort_sexcoh4.dat young females N=1070 (1973)

- **Variables (53 traits): (averaged over time survey 1 – 6)**

trappreg trappext sex1to6 gbdjr twzyg halfsib id_2twns drieli: *demographics*
neu ext nso tat tas es bs dis sbl jas angs boos bdi : *personality*
ysw ytrg ysom ydep ysoc ydnk yatt ydel yagg yoth yint yext ytot yocd: *YASR*
cfq mem dist blu nam fob blfob scfob agfob hap sat self imp cont chck urg obs com*: other*

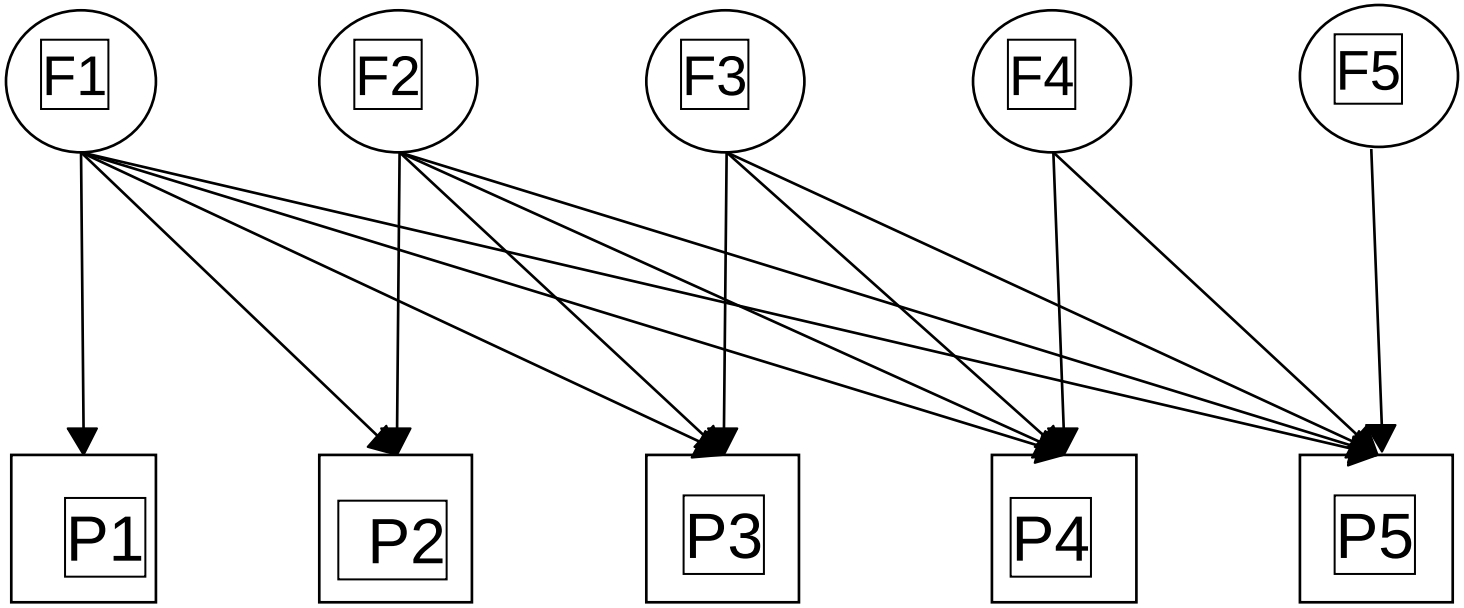
- **Mx Jobs**
- Cholesky 13vars.mx : cholesky decomposition (saturated model)
- Eigen 13vars.mx: eigenvalue decomposition of computed correlation matrix (also saturated model)
- Fa 1 factors.mx: 1 factor model
- Fa 2 factors.mx : 2 factor model
- Fa 3 factors.mx: 3 factor model (constraint on loading)
- Fa 4 factors.mx: 1 general factor, plus 3 trait factors
- Fa 3 factors constraint dorret.mx
- Fa 3 factors constraint dorret.mx: alternative constraint to identify the model

```
title cholesky for sex/age groups
data ng=1 Ni=53              !8 demographics, 13 scales, 14 yasr, 18 extra
missing=-1.00                !personality missing = -1.00
rectangular file =personShort_sexcoh3.dat
labels
trappreg trappext sex1to6 gbdjr twzyg halfsib id_2twns drieli neu ext nso etc.

Select NEU NSO ANX BDI YDEP  TAS ES BS DIS  EXT JAS ANGER TAT /
begin matrices;
A lower 13 13 free           !common factors
M full 1 13 free             !means
end matrices;

covariance A*A'/
means M  /
start 1.5 all etc.
option nd=2
end
```

NEU NSO ANX BDI YDEP TAS ES BS DIS EXT JAS ANGER TAT /

MATRIX A: This is a LOWER TRIANGULAR matrix of order 13 by 13

23.74												
3.55	4.42											
6.89	0.96	5.34										
1.70	0.72	0.80	2.36									
2.79	0.32	0.68	-0.08	2.87								
-0.30	0.03	-0.01	0.16	0.18	7.11							
0.28	0.13	0.17	-0.04	0.24	3.32	6.03						
1.29	-0.08	0.30	-0.15	-0.09	0.96	1.52	6.01					
0.83	-0.07	0.35	-0.30	0.15	1.97	0.91	1.16	5.23				
-4.06	-0.11	-1.41	-0.20	-0.90	2.04	1.07	3.14	0.94	14.06			
1.85	-0.02	0.70	-0.28	0.01	0.47	0.00	0.43	-0.08	1.11	3.98		
1.86	-0.09	0.80	-0.49	-0.18	0.13	0.04	0.21	0.18	0.51	0.97	3.36	
-1.82	0.16	-0.34	0.02	-1.26	-0.16	-0.46	-0.80	-0.53	-1.21	-1.20	-1.64	7.71



To interpret the solution, standardize the factor loadings both with respect to the latent and the observed variables.
In most models, the latent variables have unit variance;
standardize the loadings by the variance of the observed variables
(e.g. $\lambda 21$ is divided by the SD of P2)

Group 2 in Cholesky script

```
Calculate Standardized Solution
Calculation
Matrices = Group 1
I Iden 13 13
End Matrices;

Begin Algebra;
 S=(\sqrt(I.R))~;          ! diagonal matrix of standard deviations
 P=S*A;                    ! standardized estimates for  factors loadings
End Algebra;
End
```

(R=(A*A'). i.e. R has variances on the diagonal)

Standardized solution: standardized loadings

NEU NSO ANX BDI YDEP TAS ES BS DIS EXT JAS ANGER TAT /

- 1.00
- **0.63** 0.78
- **0.79** 0.11 0.61
- **0.55** 0.23 0.26 0.76
- **0.69** 0.08 0.17 -0.02 0.70
- -0.04 0.00 0.00 0.02 0.03 **0.99**
- 0.04 0.02 0.02 -0.01 0.04 **0.48** 0.87
- 0.20 -0.01 0.05 -0.02 -0.01 **0.15** 0.24 0.94
- 0.14 -0.01 0.06 -0.05 0.02 **0.34** 0.15 0.20 0.89
- -0.27 -0.01 -0.09 -0.01 -0.06 0.13 0.07 0.21 0.06 **0.92**
- 0.40 0.00 0.15 -0.06 0.00 0.10 0.00 0.09 -0.02 **0.24** **0.86**
- 0.45 -0.02 0.19 -0.12 -0.04 0.03 0.01 0.05 0.04 0.12 **0.24** 0.82
- -0.22 0.02 -0.04 0.00 -0.15 -0.02 -0.05 -0.09 -0.06 -0.14 -0.14 -0.19 0.91

NEU NSO ANX BDI YDEP TAS ES BS DIS EXT JAS ANGER TAT /

- Your model has104 estimated parameters :
- 13 means
- 13*14/2 = 91 factor loadings
-
- -2 times log-likelihood of data >>>108482.118

Eigenvalues, eigenvectors & principal component analyses (PCA)

1) data reduction technique
2) form of factor analysis
3) very useful transformation

Principal components analysis (PCA)

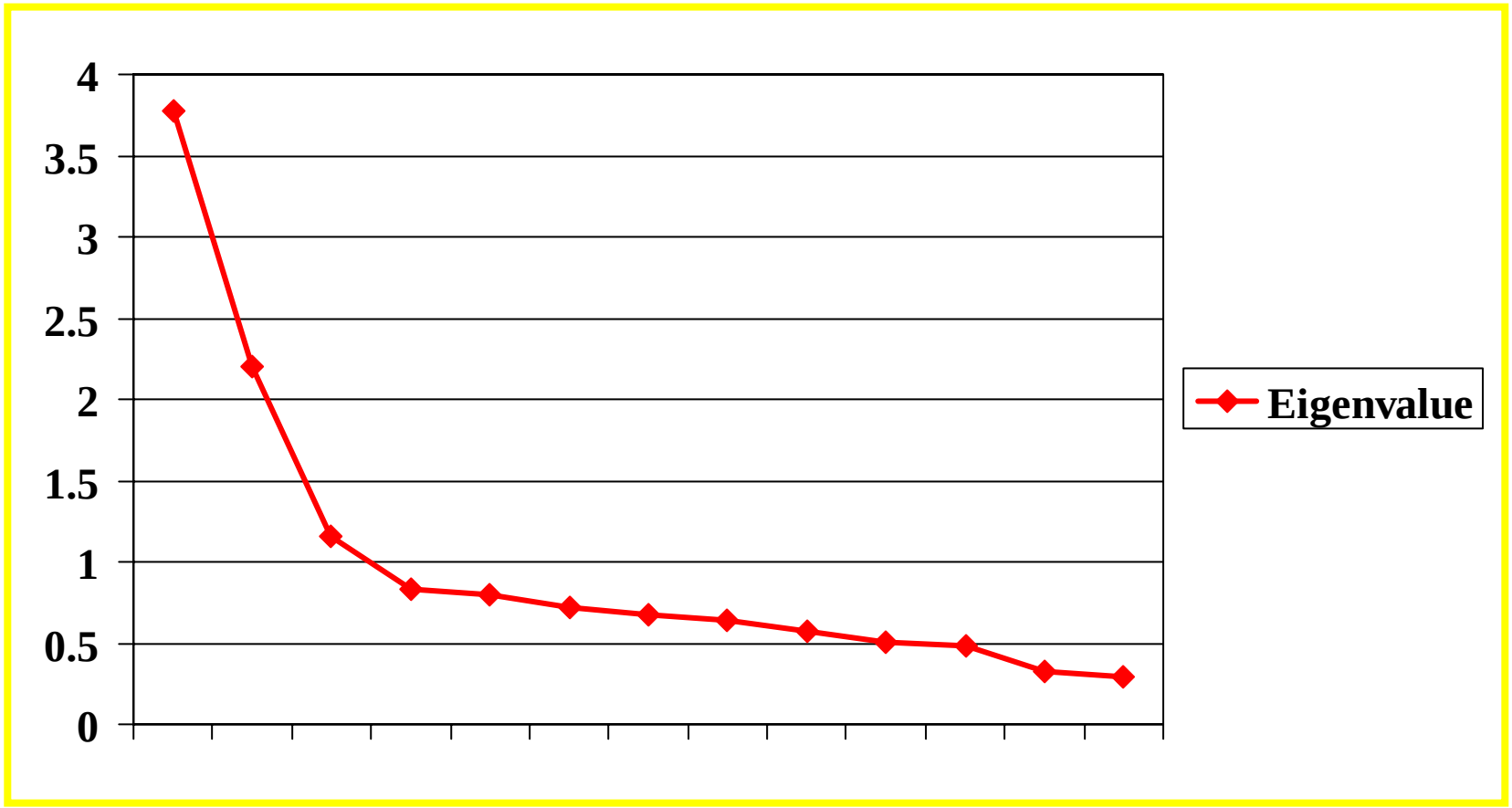
PCA is used to reduce large set of variables into a smaller number of uncorrelated components.

Orthogonal transformation of a set of variables (x) into a set of uncorrelated variables (y) called *principal components* that are linear functions of the x-variates.

The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

Principal component analysis of 13 personality / psychopathology inventories: 3 eigenvalues > 1

(Dutch adolescent and young adult twins, data 1991-1993; SPSS)



Principal components analysis (PCA)

PCA gives a transformation of the correlation matrix R and is a completely determinate model.

R (q x q) = P D P', where
P = q x q orthogonal matrix of eigenvectors
D = diagonal matrix (containing eigenvalues)

$y = P' x$ and the variance of y_j is p_j

The first principal component $y_1 = p_{11}x_1 + p_{12}x_2 + \ldots + p_{1q}x_q$
The second principal component $y_2 = p_{21}x_1 + p_{22}x_2 + \ldots + p_{2q}x_q$
etc.

$[p_{11}, p_{12}, \ldots, p_{1q}]$ is the first eigenvector
d_{11} is the first eigenvalue (variance associated with y_1)

Principal components analysis (PCA)

The principal components are linear combinations of the x-variables which maximize the variance of the linear combination and which have zero covariance with the other principal components.

There are exactly q such linear combinations (if R is positive definite).

Typically, the first few of them explain most of the variance in the original data. So instead of working with X1, X2, ..., Xq, you would perform PCA and then use only Y1 and Y2, in a subsequent analysis.

PCA, Identifying constraints: transformation unique

Characteristics:

1) var(d_{ij}) is maximal

2) d_{ij} is uncorrelated with d_{kj}

are ensured by imposing the constraint:

PP' = P'P = I (where ' stands for transpose)

Principal components analysis (PCA)

The objective of PCA usually is not to account for covariances among variables, but to summarize the information in the data into a smaller number of (orthogonal) variables.
No distinction is made between common and unique variances. One advantage is that factor scores can be computed directly and need not to be estimated.

- H. Hotelling (1933): Analysis of a complex of statistical variables into principal component. Journal Educational Psychology, 417-441, 498-520

PCA

Primarily data reduction technique, but often used as form of exploratory factor analysis:

- Scale dependent (use only correlation matrices)!
- Not a "testable" model, no statistical inference
- Number of components based on rules of thumb (e.g. # of eigenvalues > 1)

```
title eigen values
data ng=1 Ni=53
missing=-1.00
rectangular file =personShort_sexcoh3.dat
labels
trappreg trappext sex1to6 gbdjr twzyg halfsib id_2twns drieli neu ext nso  tat  tas  etc.
Select NEU NSO ANX BDI YDEP  TAS ES BS DIS  EXT JAS ANGER TAT /

begin matrices;
R stand 13 13 free   !correlation matrix
S diag   13 13 free   !standard deviations
M full 1 13 free      !means
end matrices;
begin algebra;
E = \eval(R);         !eigenvalues of R
V = \evec(R);         !eigenvectors of R
end algebra;

covariance S*R*S'/
means M  /

start 0.5 all etc.
end
```

Correlations NEU NSO ANX BDI YDEP TAS ES BS DIS EXT JAS ANGER TAT /

MATRIX R: This is a STANDARDISED matrix of order 13 by 13

•	1.000										
•	0.625	1.000									
•	0.785	0.576	1.000								
•	0.548	0.523	0.612	1.000							
•	0.685	0.490	0.648	0.421	1.000						
•	-0.041	-0.023	-0.033	-0.005	-0.011	1.000					
•	0.041	0.040	0.049	0.028	0.059	0.480	1.000				
•	0.202	0.116	0.186	0.102	0.136	0.140	0.288	1.000			
•	0.142	0.080	0.146	0.052	0.125	0.329	0.305	0.306	1.00		
•	-0.266	-0.172	-0.266	-0.181	-0.239	0.143	0.110	0.172	0.108		
•	0.400	0.247	0.406	0.211	0.301	0.083	0.070	0.191			
•	0.451	0.265	0.470	0.201	0.312	0.009	0.045	0.159	ETC		
•	-0.216	-0.120	-0.192	-0.123	-0.258	-0.013	-0.071	-0.148			

Eigenvalues

- MATRIX E: This is a computed FULL matrix of order 13 by 1, [=\EVAL(R)]

1	0.200
2	0.263
3	0.451
4	0.457
5	0.518
6	0.549
7	0.677
8	0.747
9	0.824
10	0.856
11	1.300
12	2.052
13	4.106

What is the fit of this model?

It is the same as for Cholesky

Both are saturated models

Principal components analysis (PCA): $S = P\ D\ P' = P^*\ P^{*\prime}$

where S = observed covariance matrix

P'P = I (eigenvectors)

D = diagonal matrix (containing eigenvalues)

$P^* = P\ (D^{1/2})$

Cholesky decomposition: S = Q Q'

where Q = lower diagonal (triangular)

For example, if S is 3 x 3, then Q looks like:

$$\begin{pmatrix} f_{11} & 0 & 0 \\ f_{21} & f_{22} & 0 \\ f_{31} & f_{32} & f_{33} \end{pmatrix}$$

If # factors = # variables, Q may be rotated to P*. Both approaches give a transformation of S. Both are completely determinate.

PCA is based on the eigenvalue decomposition.

$$S=P*D*P'$$

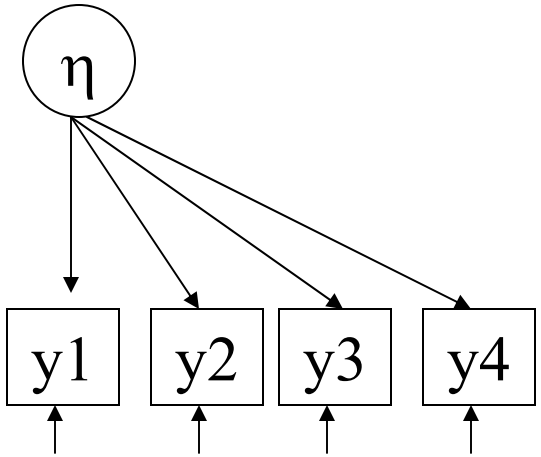
If the first component approximates S:

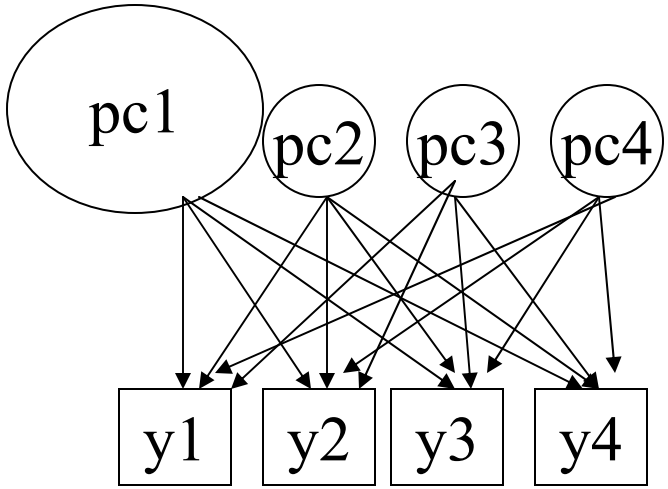
$$S\approx P_1*D_1*P_1'$$

$$S\approx \Pi_1*\Pi_1', \Pi_1 = P_1*D_1^{1/2}$$

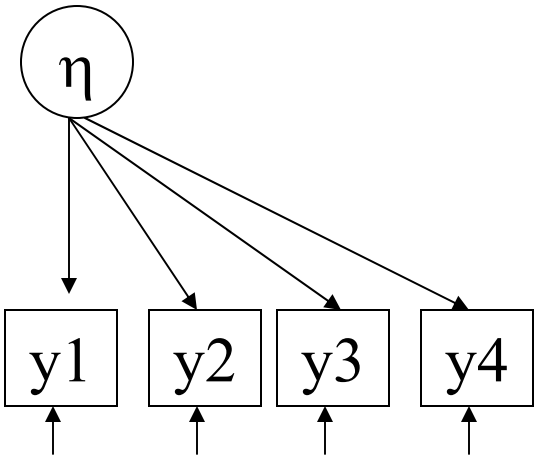
It resembles the common factor model

$$S\approx \Sigma=\Lambda*\Lambda' +\Theta, \Lambda\approx\Pi_1$$

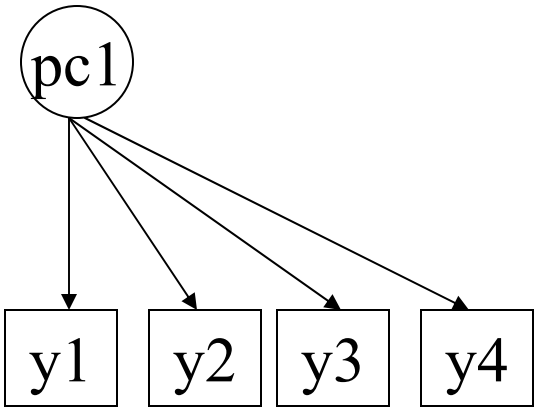




If pc1 is large, in the sense that it accounts for much variance



=>



Then it resembles the common factor model (without unique variances)

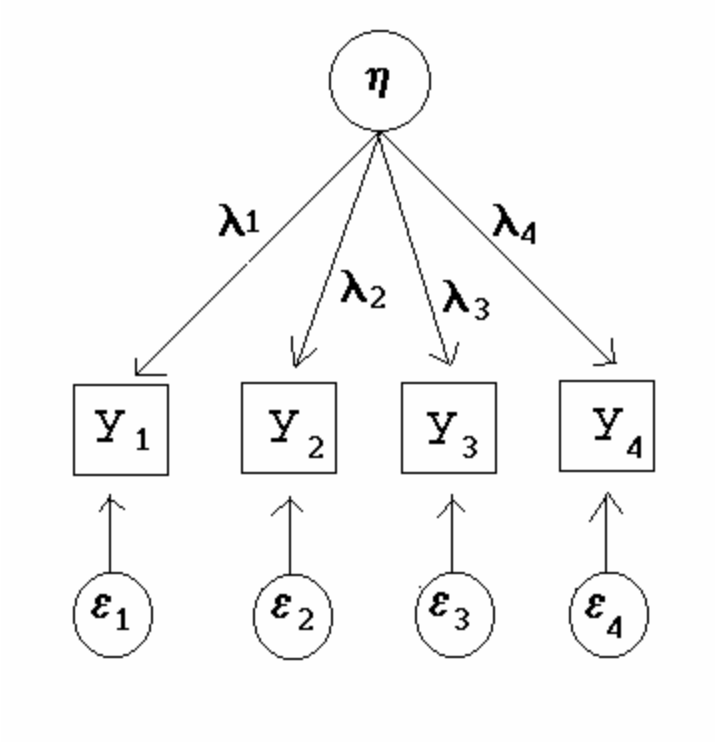
Factor analysis

Aims at accounting for covariances among observed variables / traits in terms of a smaller number of latent variates or common factors.

Factor Model: $x = \Lambda f + e$,

where x = observed variables
f = (unobserved) factor score(s)
e = unique factor / error
Λ = matrix of factor loadings

Factor analysis: Regression of observed variables (x or y) on latent variables (f or η)



One factor model with specifics

Factor analysis

Factor Model: $x = \Lambda f + e$,

With covariance matrix: $\Sigma = \Lambda \Psi \Lambda' + \Theta$

where Σ = covariance matrix
Λ = matrix of factor loadings
Ψ = correlation matrix of factor scores
Θ = (diagonal) matrix of unique variances

To estimate factor loadings we do not need to know the individual factor scores, as the expectation for Σ only consists of Λ, Ψ, and Θ.

•C. Spearman (1904): General intelligence, objectively determined and measured. American Journal of Psychology, 201-293
•L.L. Thurstone (1947): Multiple Factor Analysis, University of Chicago Press

One factor model for personality?

- Take the cholesky script and modify it into a 1 factor model (include unique variances for each of the 13 variables)
- Alternatively, use the FA 1 factors.mx script
- NB think about starting values (look at the output of eigen 13 vars.mx for trait variances)

Confirmatory factor analysis

An initial model (i.e. a matrix of factor loadings) for a confirmatory factor analysis may be specified when for example:

– its elements have been obtained from a previous analysis in another sample.

– its elements are described by a clinical model or a theoretical process (such as a simplex model for repeated measures).

Mx script for 1 factor model

```
title factor
data ng=1 Ni=53
missing=-1.00
rectangular file =personShort_sexcoh3.dat
labels
trappreg trappext sex1to6 gbdjr twzyg halfsib id_2twns drieli neu ext ETC
Select NEU NSO ANX BDI YDEP  TAS ES BS DIS  EXT JAS ANGER TAT  /
begin matrices;
A full 13 1 free              !common factors
B iden  1 1                   !variance common factors
M full 13 1  free             !means
E diag 13 13 free             !unique factors (SD)
end matrices;
specify A
1 2  3 4  5  6  7 8 9 10 11 12 13
covariance A*B*A' + E*E'/
means M /
Starting values
end
```

Mx output for 1 factor model

loadings	1
• neu	21.3153
• nso	3.7950
• anx	7.7286
• bdi	1.9810
• ydep	3.0278
• tas	-0.1530
• es	0.4620
• bs	1.4337
• dis	0.9883
• ext	-3.9329
• jas	2.1012
• anger	2.1103
• tat	-2.1191

Unique loadings are found on the Diagonal of E.

Means are found in M matrix

Your model has 39 estimated parameters
-2 times log-likelihood of data 109907.192

13 means
13 loadings on the common factor
13 unique factor loadings

Factor analysis

Factor Model: $x = \Lambda f + e$,
Covariance matrix: $\Sigma = \Lambda \Psi \Lambda' + \Theta$

Because the latent factors do not have a "natural" scale, the user needs to scale them. For example:

If $\Psi = I$: $\Sigma = \Lambda\Lambda' + \Theta$

- factors are standardized to have unit variance
- factors are independent

Another way to scale the latent factors would be to constrain one of the factor loadings.

In confirmatory factor analysis:

- a model is constructed in advance
- that specifies the number of (latent) factors
- that specifies the pattern of loadings on the factors
- that specifies the pattern of unique variances specific to each observation
- measurement errors may be correlated
- factor loadings can be constrained to be zero (or any other value)
- covariances among latent factors can be estimated or constrained
- multiple group analysis is possible

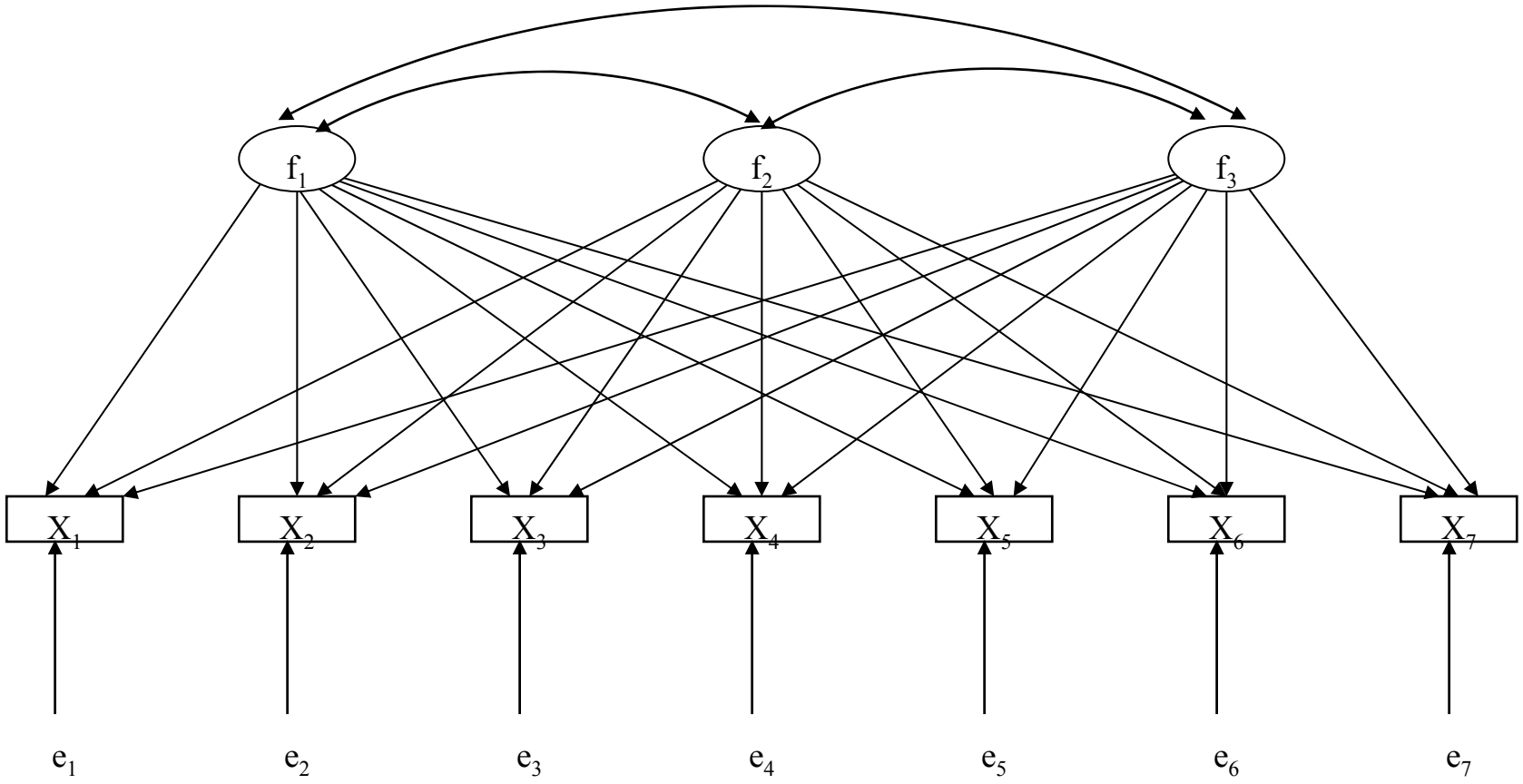
We can TEST if these constraints are consistent with the data.

Distinctions between exploratory (SPSS/SAS) and confirmatory factor analysis (LISREL/Mx)

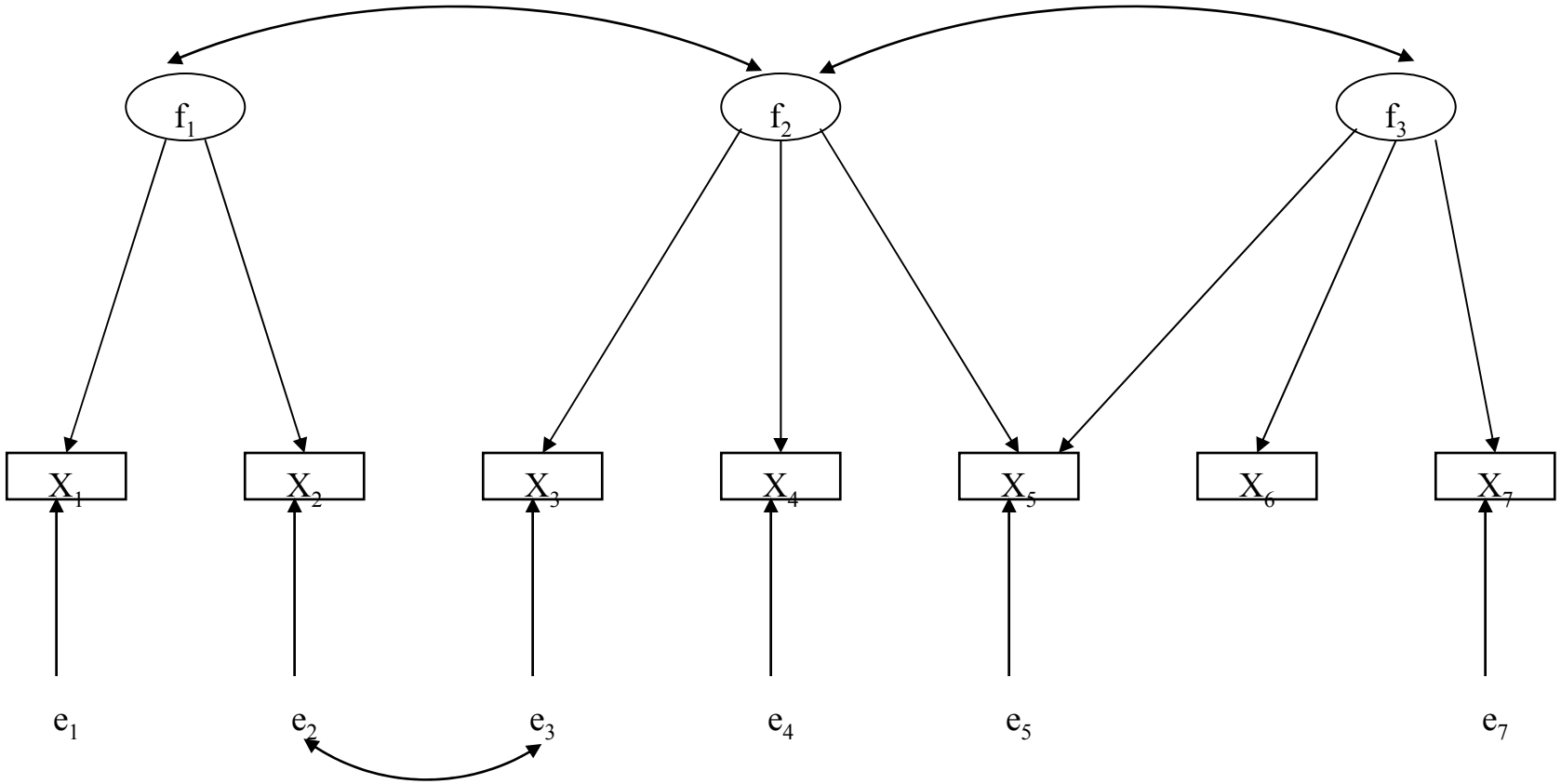
In exploratory factor analysis:

- no model that specifies the number of latent factors
- no hypotheses about factor loadings (usually all variables load on all factors, factor loadings cannot be constrained)
- no hypotheses about interfactor correlations (either no correlations or all factors are correlated)
- unique factors must be uncorrelated
- all observed variables must have specific variances
- no multiple group analysis possible
- under-identification of parameters

Exploratory Factor Model



Confirmatory Factor Model



Confirmatory factor analysis

A maximum likelihood method for estimating the parameters in the model has been developed by Jöreskog and Lawley (1968) and Jöreskog (1969).

ML provides a test of the significance of the parameter estimates and of goodness-of-fit of the model.

Several computer programs (Mx, LISREL, EQS) are available.

• K.G. Jöreskog, D.N. Lawley (1968): New Methods in maximum likelihood factor analysis. British Journal of Mathematical and Statistical Psychology, 85-96

•K.G. Jöreskog (1969): A general approach to confirmatory maximum likelihood factor analysis Psychometrika, 183-202

• D.N. Lawley, A.E. Maxwell (1971): Factor Analysis as a Statistical Method. Butterworths, London

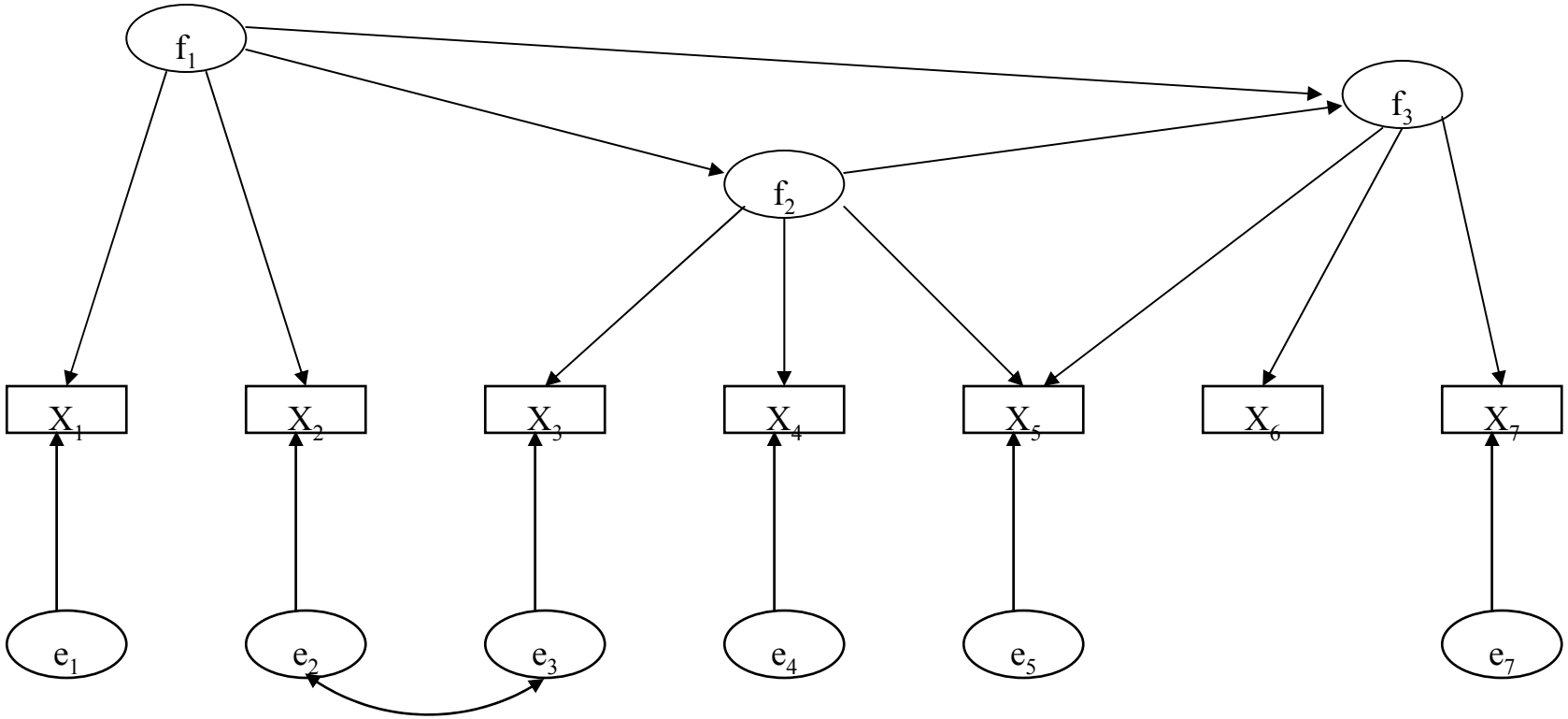
• S.A. Mulaik (1972): The Foundations of Factor analysis, McGraw-Hill Book Company, New York

• J Scott Long (1983): Confirmatory Factor Analysis, Sage

Structural equation models

Sometimes $x = \Lambda f + e$ is referred to as the measurement model, and the part of the model that specifies relations among latent factors as the covariance structure model, or the structural equation model.

Structural Model



Path Analysis & Structural Models

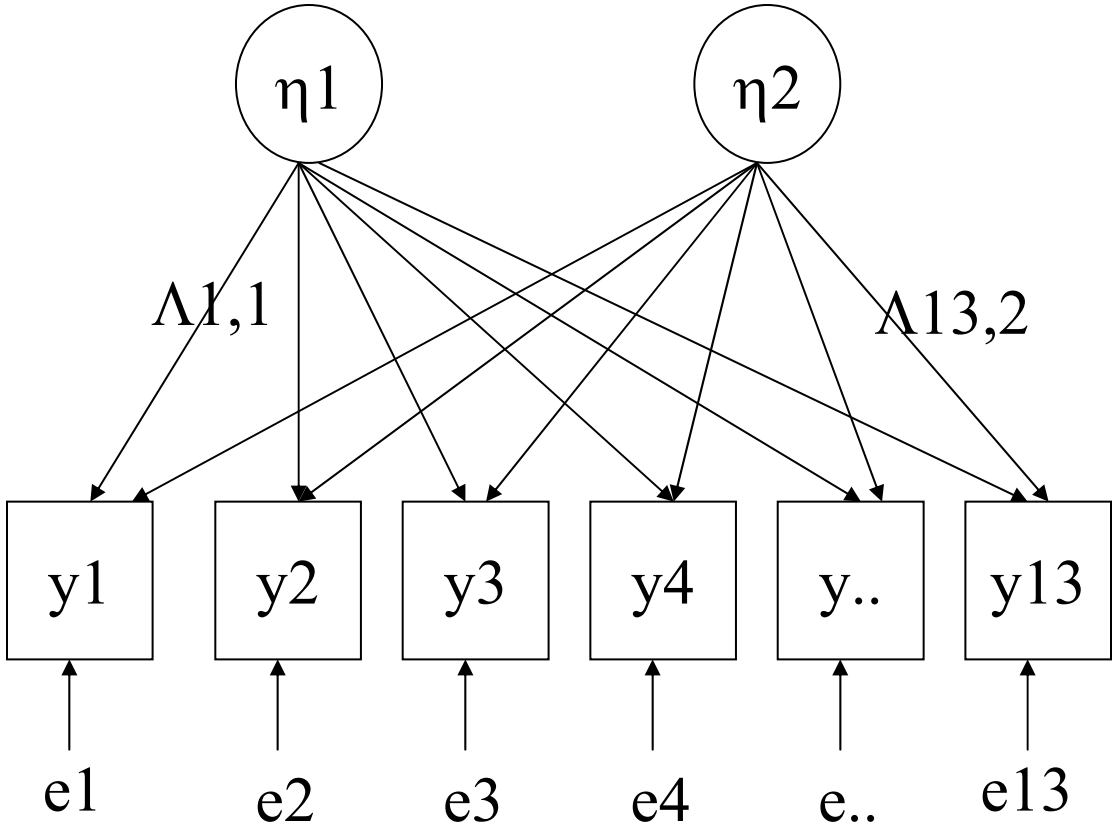
Path analysis diagrams allow us

- to represent linear structural models, such as regression, factor analysis or genetic models.
- to derive predictions for the variances and covariances of our variables under that model.

Path analysis is not a method for discovering causes, but a method applied to a causal model that has been formulated in advance. It can be used to study the direct and indirect effects of *exogenous v*ariables ("causes") on e*ndogenous* variables ("effects").

- C.C. Li (1975): Path Analysis: A primer, Boxwood Press
- E.J. Pedhazur (1982): Multiple Regression Analysis Explanation and Prediction, Hold, Rinehart and Wilston

Two common factor model



Two common factor model

y_{ij}, i=1...P tests, j=1...N cases

$Y_{ij} = \lambda_{i1j}\, \eta_{1j} + \lambda_{i2j}\, \eta_{2j} + e_{ij}$

Λ matrix of factor loadings:

$$\begin{matrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \dots & \dots \\ \lambda_{P1} & \lambda_{P2} \end{matrix}$$

Identification

The factor model in which all variables load on all (2 or more) common factors is not identified. It is not possible in the present example to estimate all 13x2 loadings.

But how can some programs (e.g. SPSS) produce a factor loading matrix with 13x2 loadings?

Identifying constraints

Spss automatically imposes the identifying constraint similar to:

$L^t\Theta^{-1}L$ is diagonal,

Where L is the matrix of factor loadings and Θ is the diagonal covariance matrix of the residuals (eij).

Other identifying constraints

3 factors			2 factors	
λ_{11}	0	0	λ_{11}	0
λ_{21}	λ_{22}	0	λ_{21}	λ_{22}
λ_{31}	λ_{32}	λ_{33}	λ_{31}	λ_{32}
…	…	…	…	…
λ_{P1}	λ_{P2}	λ_{P3}	λ_{P1}	λ_{P2}

Where you fix the zero is not important!

Confirmatory FA

Specify expected factor structure directly and fit the model.

Specification should include enough fixed parameter in Λ (i.e., zero's) to ensure identification.

Another way to guarantee identification is the constraint that $\Lambda\ \Theta^{-1}\Lambda'$ is diagonal (this works for orthogonal factors).

2, 3, 4 factor analysis

- Modify an existing script (e.g. from 1 into 2 and common factors)
- ensure that the model is identified by putting at least 1 zero loading in the second set of loading and at least 2 zero's in the third set of loadings
- Alternatively, do not use zero loadings but use the constraint that $\Lambda \Theta^{-1} \Lambda'$ is diagonal

- Try a CFA with 4 factors: 1 general, 1 Neuroticism, 1 Sensation seeking and 1 Extraversion factor

3 factor script

```
BEGIN MATRICES;
 A FULL 13 3 FREE              !COMMON FACTORS
 P IDEN  3 3                   !VARIANCE COMMON FACTORS
 M FULL 13 1  FREE             !MEANS
 E DIAG 13 13 FREE             !UNIQUE FACTORS
 END MATRICES;

 SPECIFY A
 1  0    0
 2  14   98
 3  15   28
 4  16   29
 5  17   30
 6  18   0
 7  19   31
 8  20   32
 9  21   33
 10 22   34
 11 23   35
 12 24   36
 13 25   37

 COVARIANCE A*P*A' + E*E'/
 MEANS M /
```

3 factor output:

NEU NSO ANX BDI YDEP TAS ES BS DIS EXT JAS ANGER TAT /

- MATRIX A
- 1 **21.3461** 0.0000 0.0000
- 2 **3.8280** 0.0582 -0.6371
- 3 **7.7261** 0.0621 0.0936
- 4 **1.9909** 0.0620 -0.5306
- 5 **3.0229** 0.1402 -0.1249
- 6 -0.2932 **4.6450** 0.0000
- 7 0.3381 **4.9062** -0.1884
- 8 1.3199 **2.3474** 1.1847
- 9 0.8890 **2.8024** 0.6020
- 10 -4.3455 3.3760 **5.8775**
- 11 2.0539 0.4507 **2.2805**
- 12 2.0803 0.1255 **1.8850**
- 13 -2.0109 -0.6641 **-3.0246**

Analyses

- 1 factor -2ll = 109,097 parameters = 39
- 2 factor -2ll = 109,082 51
- 3 factor -2ll = 108,728 62
- 4 factor -2ll = 108,782 52

- saturated -2ll = 108,482 104

$\chi 2$ = -ll(model) - -2ll(saturated;
e.g. -2ll(model3) - -2ll(sat) = 108,728-108,482 = 246; df = 104-62 = 42

Genetic Structural Equation Models

Confirmatory factor model: $x = \Lambda f + e$, where

x = observed variables

f = (unobserved) factor scores

e = unique factor / error

Λ = matrix of factor loadings

"Univariate" genetic factor model

$$P_j = hG_j + e E_j + c C_j \text{ , } j = 1, ..., n \text{ (subjects)}$$

where P = measured phenotype

f = G: unmeasured genotypic value

C: unmeasured environment common to family members

E: unique environment

Λ = h, c, e (factor loadings/path coefficients)

Genetic Structural Equation Models

$x = \Lambda f + e$

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

Genetic factor model

$P_{ji} = hG_{ji} + c\,C_{ji} + e\,Eji$, j=1,..., n (pairs) and i=1,2 (Ss within pairs)

The correlation between latent G and C factors is given in Ψ (4x4)

Λ contains the loadings on G and C: $\begin{bmatrix} h & 0 & c & 0 \\ 0 & h & 0 & c \end{bmatrix}$

And Θ is a 2x2 diagonal matrix of E factors.

Covariance matrix: (MZ pairs)

$$\begin{bmatrix} h*h + c*c & h*h + c*c \\ h*h + c*c & h*h + c*c \end{bmatrix} + \begin{bmatrix} e*e & 0 \\ 0 & e*e \end{bmatrix}$$

Structural equation models, summary

The covariance matrix of a set of observed variables is a function of a set of parameters: $\Sigma = \Sigma(\Theta)$

where Σ is the population covariance matrix,
Θ is a vector of model parameters and
Σ is the covariance matrix as a function of Θ

Example: $x = \lambda f + e$,

The observed and model covariances matrices are:

$$\begin{bmatrix} \text{Var}(x) & \\ \text{Cov}(x,f) & \text{Var}(f) \end{bmatrix} \qquad \begin{bmatrix} \lambda^2 \text{Var}(f)+\text{Var}(e) & \\ \lambda \text{Var}(f) & \text{Var}(f) \end{bmatrix}$$

KA Bollen (1990): Structural Equation with Latent Variables, John Wiley & Sons

Five steps characterize structural equation models:

1. Model Specification
2. Identification
3. Estimation of Parameters
4. Testing of Goodness of fit
5. Respecification

K.A. Bollen & J. Scott Long: Testing Structural Equation Models, 1993, Sage Publications

1: Model specification

Most models consist of systems of *linear equations*. That is, the relation between variables (latent and observed) can be represented in or transformed to linear structural equations. However, the covariance structure equations can be non-linear functions of the parameters.

2: Identification: do the unknown parameters in Θ have a unique solution?

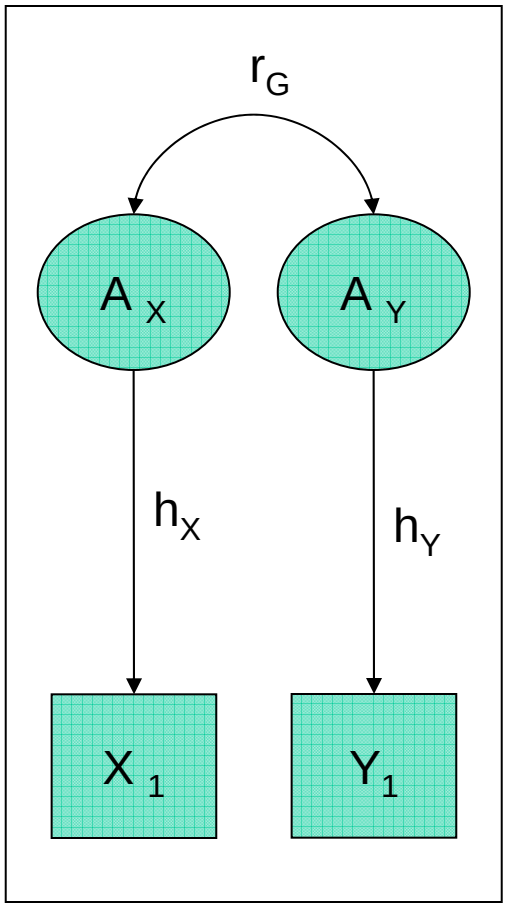
Consider 2 vectors Θ_1 and Θ_2, each of which contains values for unknown parameters in Θ.

If $\Sigma(\Theta_1) = \Sigma(\Theta_2)$ then the model is identified if $\Theta_1 = \Theta_2$

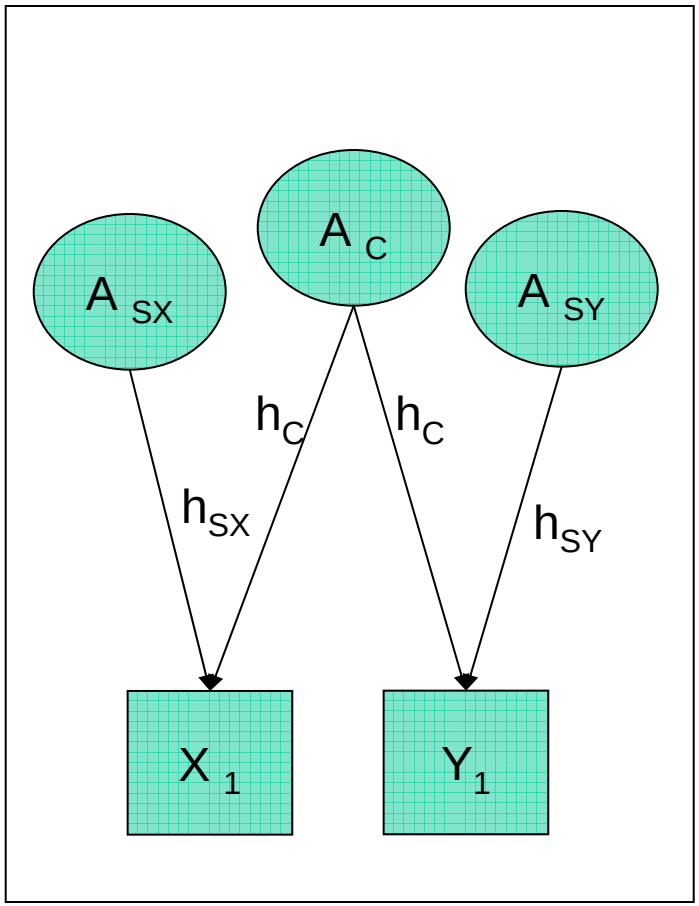
One necessary condition for identification is that the number of observed statistics is larger than or equal to the number of unknown parameters.
(use different starting values; request CI)

Identification in "twin" models depends on the multigroup design

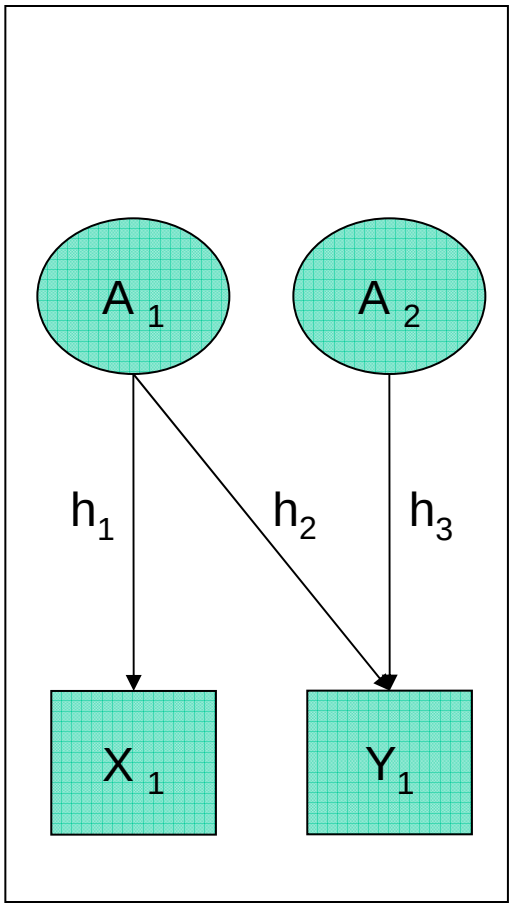
Identification: Bivariate Phenotypes: 1 correlation and 2 variances



Correlation

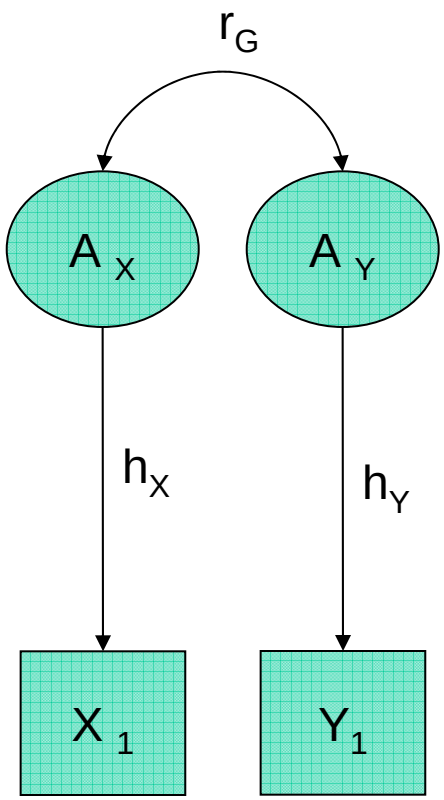


Common factor



Cholesky decomposition

Correlated factors

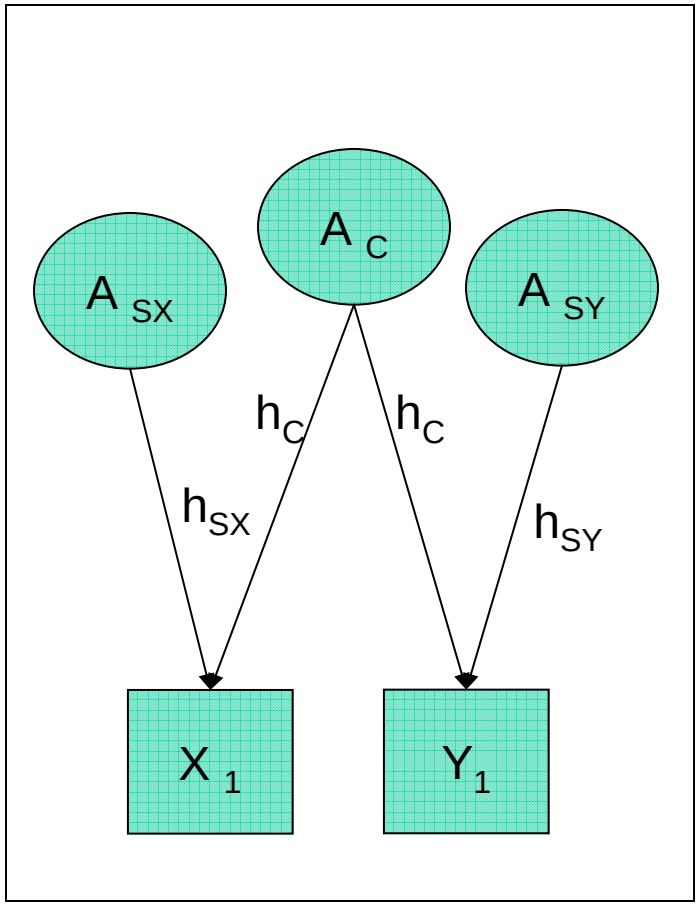


- Two factor loading (h_x and h_y) and one correlation r_G

- Expectation:

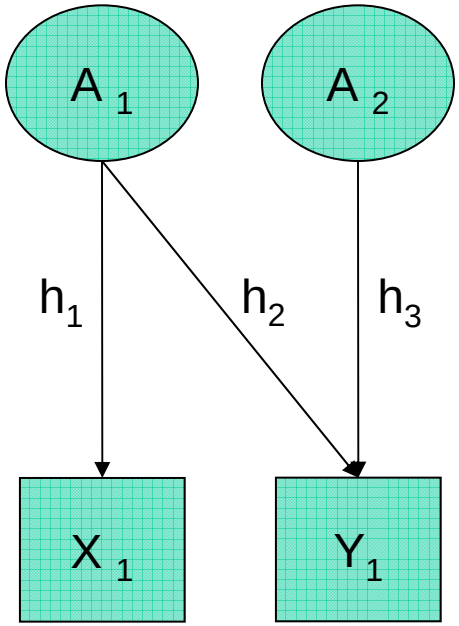
  $r_{XY} = h_X r_G h_Y$

Common factor



Four factor loadings:
A constraint on the factor loadings is needed to make this model identified.
For example: loadings on the common factor are the same.

Cholesky decomposition



- Three factor loadings
- If $h_3 = 0$: no influences specific to Y
- If $h_2 = 0$: no covariance

3: Estimation of parameters & standard errors

Values for the unknown parameters in Θ can be obtained by a fitting function that minimizes the differences between the model covariance matrix $\Sigma(\Theta)$ and the observed covariance matrix S.

The most general function is called Weighted Least Squares (WLS): $F = (s - \sigma)^t W^{-1} (s - \sigma)$
where s and σ contain the non-duplicate elements of the input matrix S and the model matrix Σ.
W is a positive definite symmetric weight matrix.
The choice of W determines the fitting function.

Rationale: the discrepancies between the observed and the model statistics are squared and weighted by a weight matrix.

Maximum likelihood estimation (MLE)

Choose estimates for parameters that have the highest likelihood given the data.

A good (genetic) model should make our empirical results likely, if a theoretical model makes our data have a low likelihood of occurrence then doubt is cast on the model.

Under a chosen model, the best estimates for parameters are found (in general) by an iterative procedure that maximizes the likelihood (minimizes a fitting function).

4: Goodness-of-fit & 5: Respecification

The most widely used measure to assess goodness-of-fit is the chi-squared statistic: $\chi^2 = F\ (N-1)$, where F is the minimum of the fitting function and N is the number of observations on which S is based.

The overall $\chi 2$ tests the agreement between the observed and the predicted variances and covariances.

The degrees of freedom (df) for this test equal the number of independent statistics minus the number of free parameters. A low χ^2 with a high probability indicates that the data are consistent with the model.

Many other indices of fit have been proposed, eg Akaike's information criterion (AIC): χ^2-2df or indices based on differences between S and Σ.

Differences in goodness-of-fit between different structural equation models may be assessed by likelihood-ratio tests by subtracting the chi-square of a properly nested model from the chi-square of a more general model.

Compare models by chi square (χ^2) tests:

A disadvantage is that χ^2 is influenced by the unique variances of the items (Browne et al., 2002).

If a trait is measured reliably, the inter-correlations of items are high, and unique variances are small, the χ^2 test may suggest a poor fit even when the residuals between the expected and observed data are trivial.

The Standardized Root Mean-square Residual (SRMR; is a fit index that is based on the residual covariation matrix and is not sensitive to the size of the correlations (Bentler, 1995).

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software

Browne, M. W., MacCallum, R. C., Kim, C., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods, 7*, 403-421.

Finally: factor scores

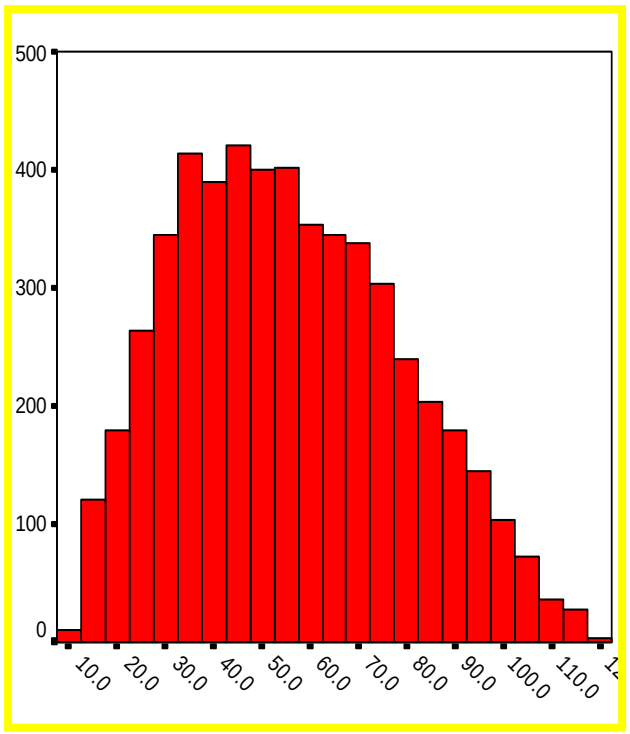
Estimates of factor loadings and unique variances can be used to construct individual factor scores: f = A'P, where A is a matrix with weights that is constant across subjects, depending on the factor loadings and the unique variances.

- R.P. McDonald, E.J. Burr (1967): A comparison of four methods of constructing factor scores. Psychometrika, 381-401
- W.E. Saris, M. dePijper, J. Mulder (1978): Optimal procedures for estimation of factor scores. Sociological Methods & Research, 85-106

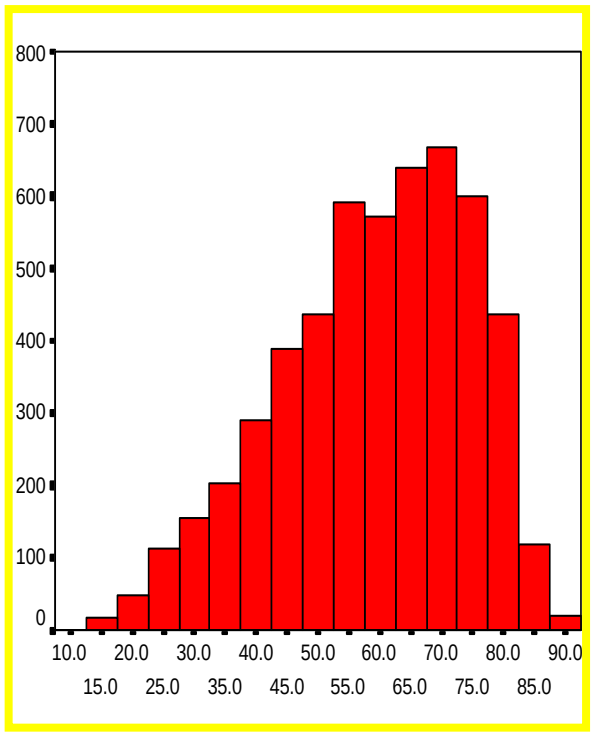
Issues

- Distribution of the data
- Averaging of data over time (alternatives)
- Dependency among cases (solution: correction)
- Final model depends on which phenotypes are analyzed (e.g. few indicators for extraversion)
- Do the instruments measure the same trait in e.g. males and females (measurement invariance)?

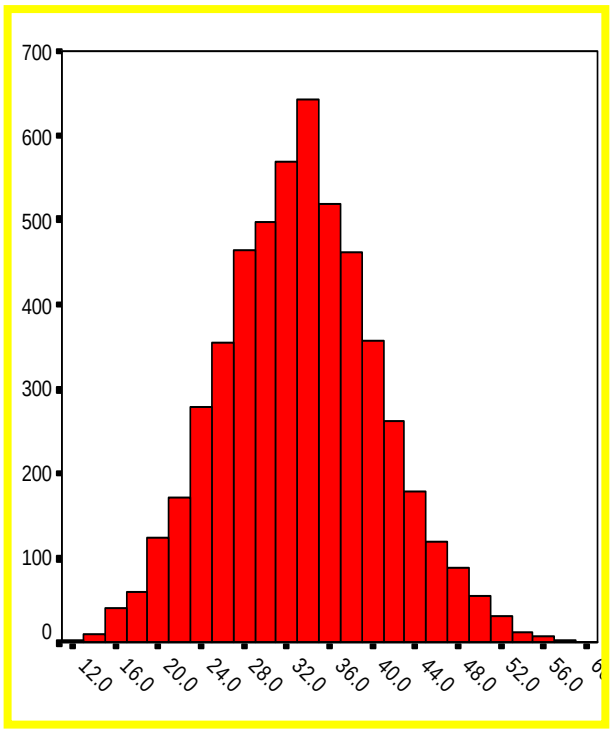
Distribution personality data
(Dutch adolescent and young adult twins, data 1991-1993)



Neuroticism (N=5293 Ss)

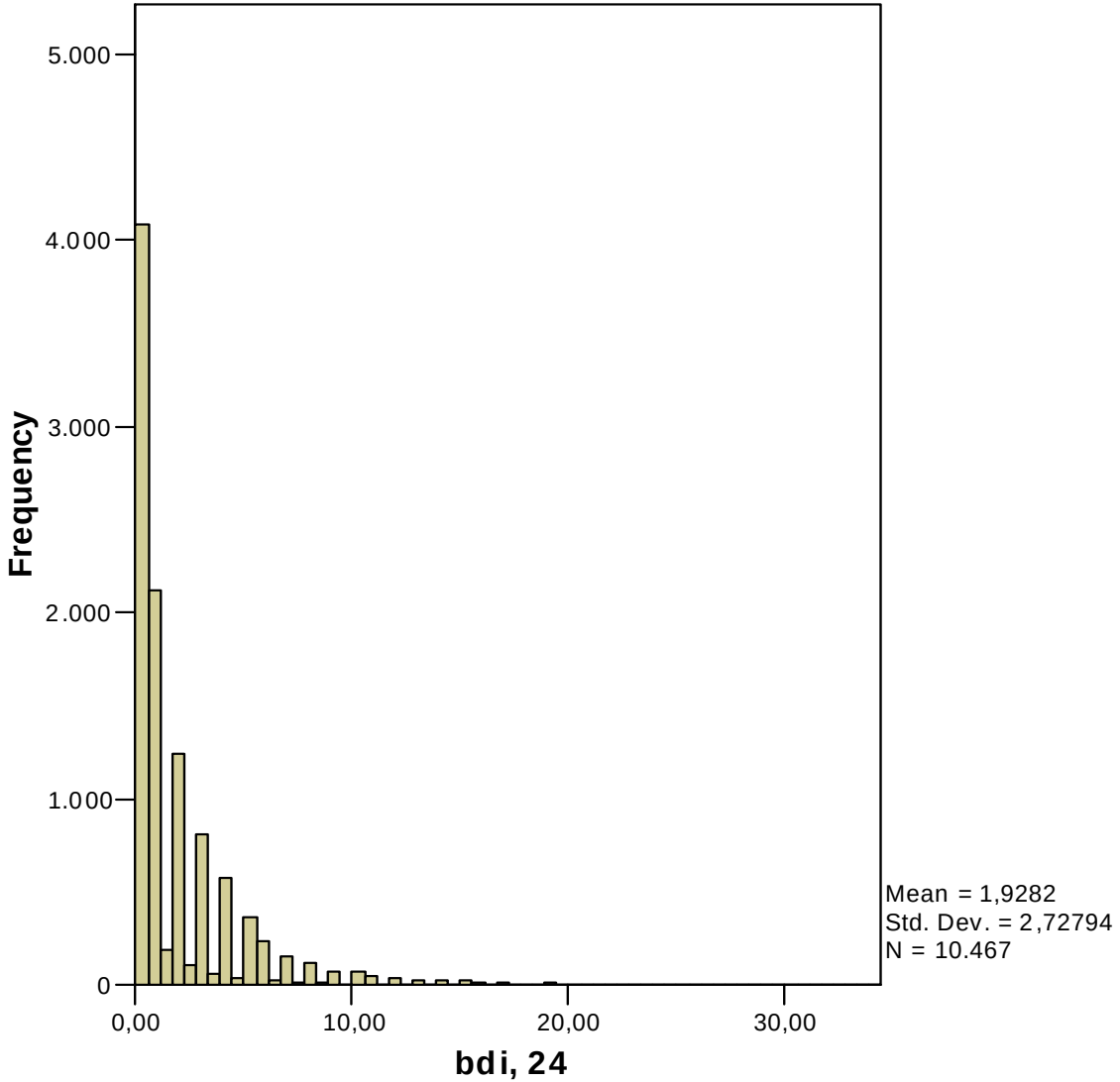


Extraversion (N=5299 Ss)

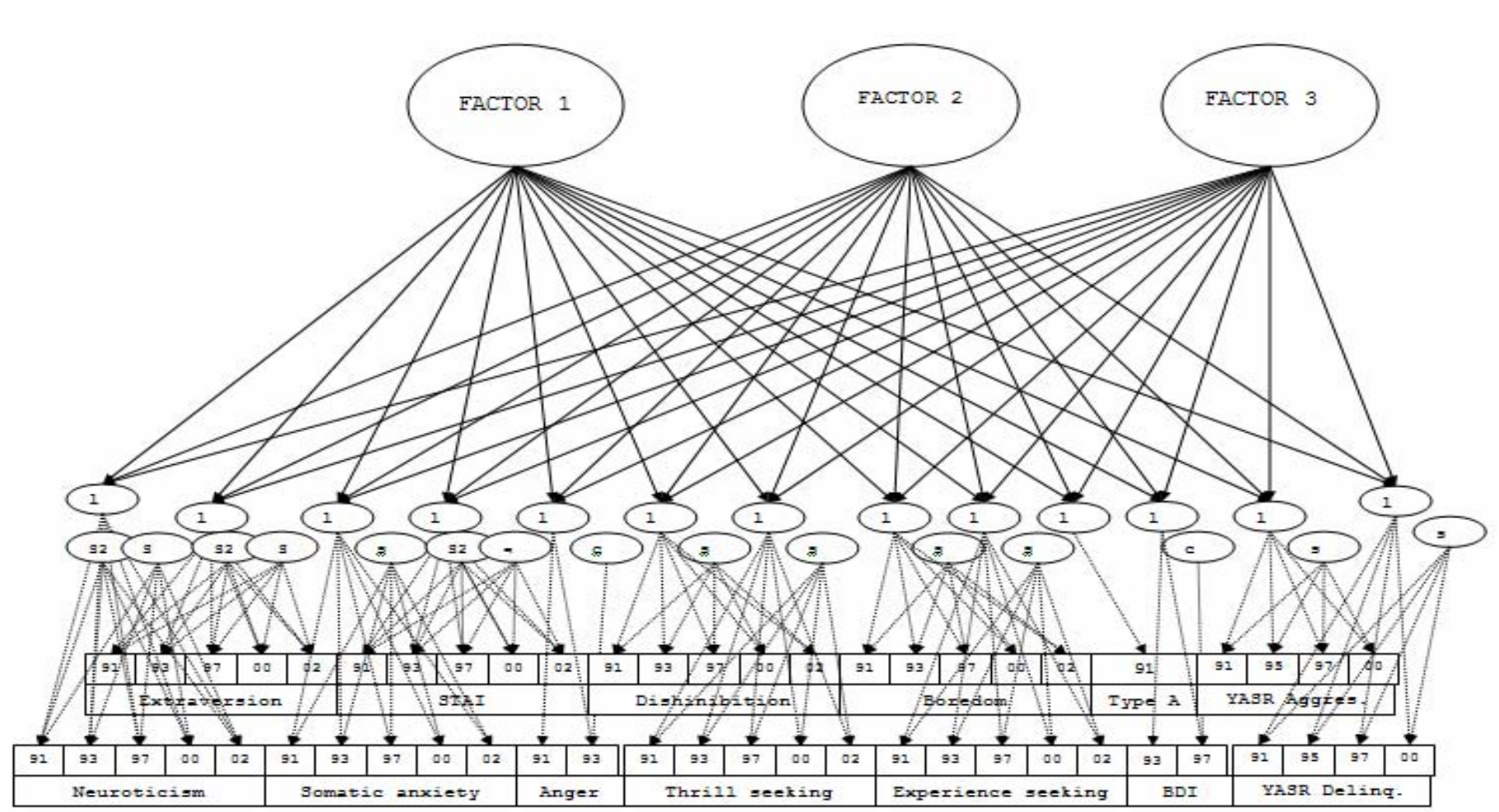


Disinhibition (N=52813 Ss)

Beck Depression Inventory



Alternative to averaging over time



Rebollo, Dolan, Boomsma

The end

- Scripts to run these analyses in other programs: Mplus and Lisrel