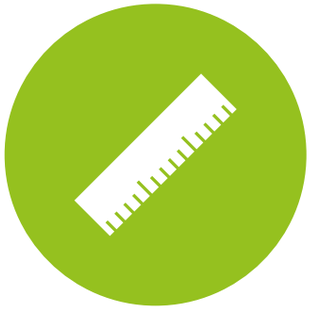
EFA

Exploratory Factor Analysis



EFA

Today's goal:

Teach Exploratory Factor Analysis (EFA)

Outline

- EFA theory
- EFA in R

EFA theory

Exploratory Factor Analysis

Why EFA?

In CFA, we specify the factor structure

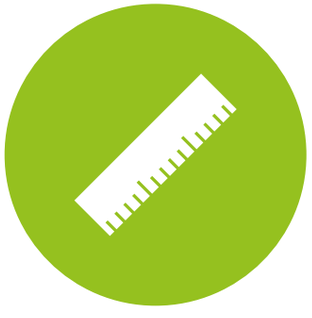
CFA will tell you how well this structure fits to the data

CFA will give you suggestions on how to improve fit

In EFA, the factor structure is “free”

EFA will “extract” factors and then “rotate” them to fit

Effectively, it infers the structure from the data



Why EFA?

Use EFA when you have no idea about the factor structure

- E.g. semi-related behaviors (see example at the end)
- E.g. A (large) factor that didn't fit and might consist of multiple dimensions instead

Many HCI researchers use EFA instead of CFA

- Why? Because it is available in SPSS...
- Using EFA instead of CFA is a crutch
- Moreover, the default EFA settings of SPSS are almost always wrong!

EFA

Steps in EFA:

Factor Extraction

Factor Rotation

Determining the number of factors

Extraction

R	A	B	C	D	E	F
A	1.00	0.48	0.44	0.52	0.28	0.24
B	0.48	1.00	0.33	0.39	0.21	0.18
C	0.44	0.33	1.00	0.47	0.35	0.30
D	0.52	0.39	0.47	1.00	0.49	0.42
E	0.28	0.21	0.35	0.49	1.00	0.42
F	0.24	0.18	0.30	0.42	0.42	1.00

Communalities

Rr	A	B	C	D	E	F
A	0.64	0.48	0.44	0.52	0.28	0.24
B	0.48	0.36	0.33	0.39	0.21	0.18
C	0.44	0.33	0.37	0.47	0.35	0.30
D	0.52	0.39	0.47	0.61	0.49	0.42
E	0.28	0.21	0.35	0.49	0.49	0.42
F	0.24	0.18	0.30	0.42	0.42	0.36

Total shared variance = sum(diagonal) = 2.83

Communalities

In Principal Component Analysis, the diagonal remains 1

No uniqueness!

Components include error!

PCA factor is more a “summary” of the underlying items than a psychological trait

Extract factor I

Try to match Rr and explain a lot of variance

Factor loadings: sqrt(diagonal)

Explained variance: sum(diagonal) = 2.36

(this is the eigenvalue of the matrix)

impR1	A	B	C	D	E	F
A	0.50	0.37	0.43	0.55	0.42	0.36
B	0.37	0.28	0.32	0.41	0.31	0.27
C	0.43	0.32	0.37	0.47	0.36	0.31
D	0.55	0.41	0.47	0.61	0.46	0.40
E	0.42	0.31	0.36	0.46	0.36	0.30
F	0.36	0.27	0.31	0.40	0.30	0.26

	I
A	0.704
B	0.528
C	0.607
D	0.778
E	0.596
F	0.510

Extract factor I

Trick: You only have to find the diagonal!

Once you have the loadings, you can calculate the other values: R_{AB} = $loading_A$ * $loading_B$

How to find this diagonal? Several methods possible...

impR1	A	B	C	D	E	F
A	0.50	0.37	0.43	0.55	0.42	0.36
B	0.37	0.28	0.32	0.41	0.31	0.27
C	0.43	0.32	0.37	0.47	0.36	0.31
D	0.55	0.41	0.47	0.61	0.46	0.40
E	0.42	0.31	0.36	0.46	0.36	0.30
F	0.36	0.27	0.31	0.40	0.30	0.26

Subtract from Rr

Rr	A	B	C	D	E	F
A	0.64	0.48	0.44	0.52	0.28	0.24
B	0.48	0.36	0.33	0.39	0.21	0.18
C	0.44	0.33	0.37	0.47	0.35	0.30
D	0.52	0.39	0.47	0.61	0.49	0.42
E	0.28	0.21	0.35	0.49	0.49	0.42
F	0.24	0.18	0.30	0.42	0.42	0.36

−

impR1	A	B	C	D	E	F
A	0.50	0.37	0.43	0.55	0.42	0.36
B	0.37	0.28	0.32	0.41	0.31	0.27
C	0.43	0.32	0.37	0.47	0.36	0.31
D	0.55	0.41	0.47	0.61	0.46	0.40
E	0.42	0.31	0.36	0.46	0.36	0.30
F	0.36	0.27	0.31	0.40	0.30	0.26

=

resR1	A	B	C	D	E	F
A	0.14	0.11	0.01	-0.03	-0.14	-0.12
B	0.11	0.08	0.01	-0.02	-0.11	-0.09
C	0.01	0.01	0.00	0.00	-0.01	-0.01
D	-0.03	-0.02	0.00	0.00	0.03	0.02
E	-0.14	-0.10	-0.01	0.03	0.13	0.12
F	-0.12	-0.09	-0.01	0.02	0.12	0.10

Subtract from Rr

resR1	A	B	C	D	E	F
A	0.14	0.11	0.01	-0.03	-0.14	-0.12
B	0.11	0.08	0.01	-0.02	-0.11	-0.09
C	0.01	0.01	0.00	0.00	-0.01	-0.01
D	-0.03	-0.02	0.00	0.00	0.03	0.02
E	-0.14	-0.10	-0.01	0.03	0.13	0.12
F	-0.12	-0.09	-0.01	0.02	0.12	0.10

Extract factor II

Try to match resR1 and explain a lot of variance

Explained variance: sum(diagonal) = 0.465

	II
A	-0.379
B	-0.284
C	-0.032
D	0.073
E	0.368
F	0.315

impR2	A	B	C	D	E	F
A	0.14	0.11	0.01	-0.03	-0.14	-0.12
B	0.11	0.08	0.01	-0.02	-0.10	-0.09
C	0.01	0.01	0.00	0.00	-0.01	-0.01
D	-0.03	-0.02	0.00	0.01	0.03	0.02
E	-0.14	-0.10	-0.01	0.03	0.14	0.12
F	-0.12	-0.09	-0.01	0.02	0.12	0.10

Subtract from resR1

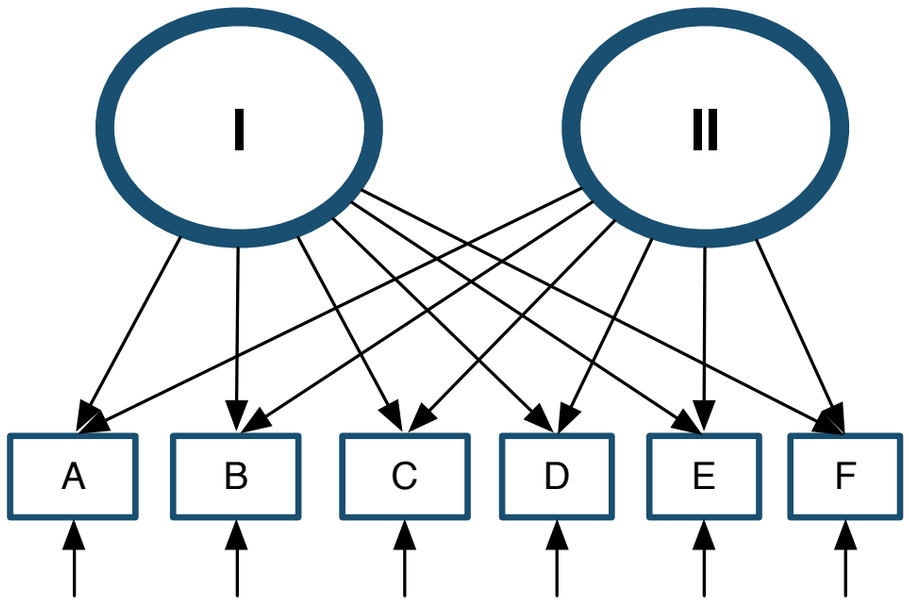
resR2	A	B	C	D	E	F
A	0.00	0.00	0.00	0.00	0.00	0.00
B	0.00	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00

Rotation

Current solution:

Complicated! Can we simplify this?

P0	I	II
A	0.704	-0.379
B	0.528	-0.284
C	0.607	-0.032
D	0.778	0.073
E	0.596	0.368
F	0.510	0.315



Rotation

Guess what? There is actually more than one solution!

- This model is underidentified!

Normally this sucks, but in this case, we are going to use it to our advantage

- How? We are going to find a solution that is more parsimonious

Rotation

Make the solution more parsimonious by spreading the explained variance over the factors in a “smart” way

- So that each item loads only on one factor, as much as possible
- Solution does not improve, just becomes easier to interpret!

Two methods:

- Orthogonal (no correlations between factors allowed)
- Oblique (correlations allowed)

Orthogonal

Multiply P0 with a transformation matrix T

TT' = I, so the explained variance remains the same

How? Different methods exist

P0	I	II
A	0.704	-0.379
B	0.528	-0.284
C	0.607	-0.032
D	0.778	0.073
E	0.596	0.368
F	0.510	0.315

→

T	1	2
I	0.736	0.677
II	-0.677	0.736

Varimax

→

P0	1	2
A	0.78	0.20
B	0.58	0.15
C	0.47	0.39
D	0.52	0.58
E	0.19	0.67
F	0.16	0.58

Oblique

Multiply P0 with a transformation matrix T, and inter-factor correlation matrix F

$$TFT' = I$$

P0	I	II
A	0.704	-0.379
B	0.528	-0.284
C	0.607	-0.032
D	0.778	0.073
E	0.596	0.368
F	0.510	0.315

→

T	1	2
I	0.575	0.555
II	-1.071	1.081

Oblimin

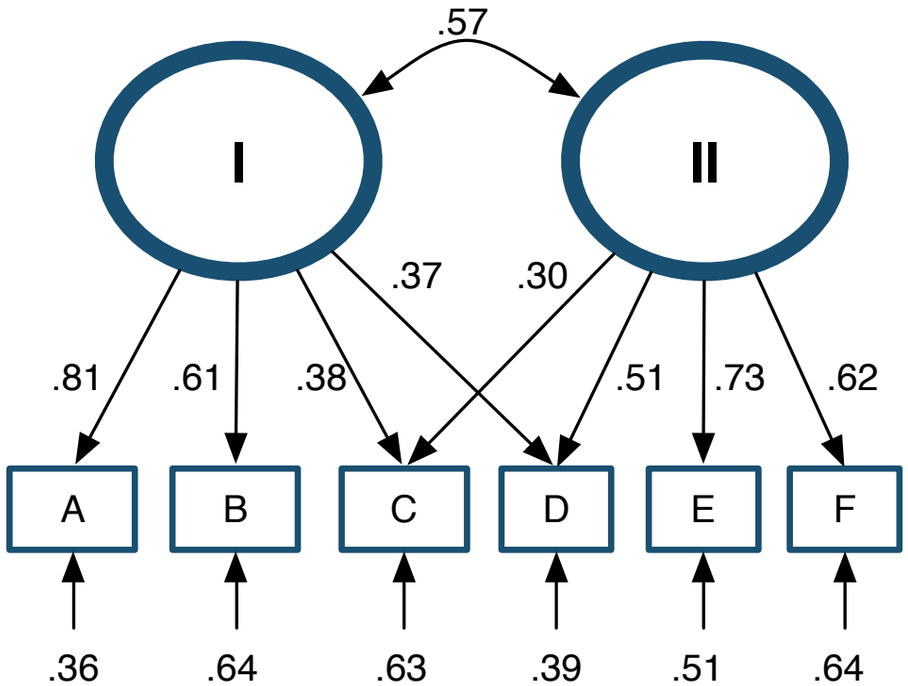
→

P0	1	2
A	0.81	0.02
B	0.61	-0.01
C	0.38	0.30
D	0.37	0.51
E	-0.05	0.73
F	-0.04	0.62

+

F	1	2
I	1.00	0.57
II	0.57	1.00

Final result



P0	1	2
A	0.81	0.02
B	0.61	-0.01
C	0.38	0.30
D	0.37	0.51
E	-0.05	0.73
F	-0.04	0.62

+

F	1	2
I	1.00	0.57
II	0.57	1.00

Rotation

Orthogonal or oblique?

Ask yourself: can these factors be correlated?

If you have a path model in mind, are there paths between these factors?

Oblique is almost always better

Unless you want to rule out any correlation

This can be useful for "data science" purposes, but the resulting factors are very hard to interpret

Reflection

SPSS default settings: PCA with varimax rotation

Not an EFA! Not an oblique rotation!

Mostly used as a crutch

"Our data is multi-collinear (high VIFs), so we apply PCA to reduce the dimensionality"

For psychometrics, use CFA

(or EFA if you really have to)

Number of factors?

Method 1 (quick):

Obtain the eigenvalue of each factor

The sum of the communalities

Do this for as many factors as possible (in most cases, same as number of items)

Build a "scree plot" of eigenvalues

Find the inflection point

Where the eigenvalue levels off

Number of factors?

Method 2 (thorough):

Run with increasing number of factors

Compare each model against the previous model

Test whether model shows significant misfit against the covariance matrix (saturated model)

(a good model can be, but may not be, non-significant)

EFA in R

An example EFA

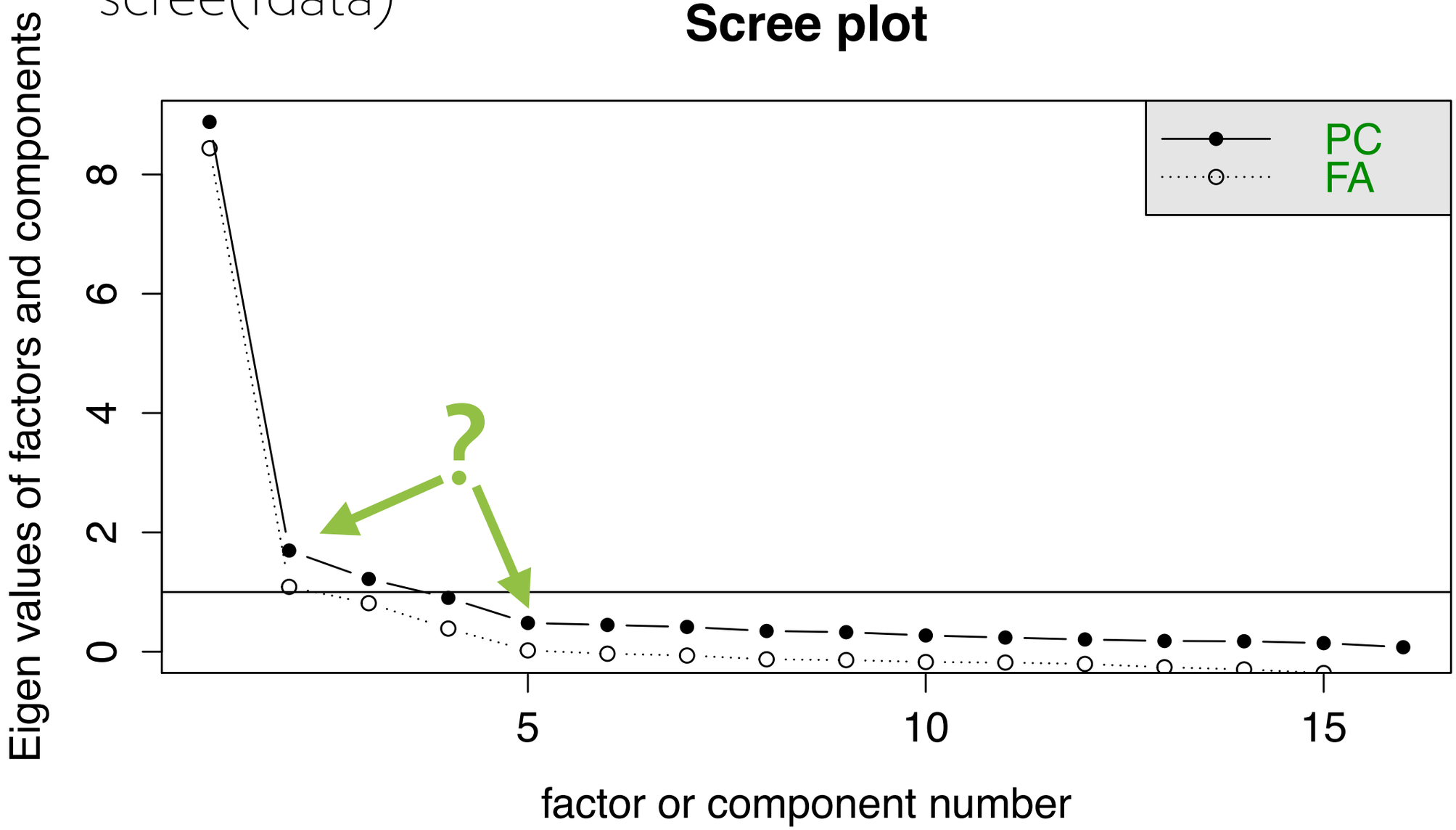
Dataset

ID	Items
1	Wall
2	Status updates
3	Shared links
4	Notes
5	Photos
6	Hometown
7	Location (city)
8	Location (state/province)
9	Residence (street address)
10	Employer
11	Phone number
12	Email address
13	Religious views
14	Interests (favorite movies, etc.)
15	Facebook groups
16	Friend list

Number of factors

Get the number of factors: use “scree” in the psych package:

```
scree(fdata)
```



Find # of factors

Method 1: factanal in stats package:

f2 <- factanal(fdata, factors=2, rotation="oblimin")

Oblimin rotation (in the GPArotation package) is used because factors may be correlated!

inspect f2

2-factor model doesn't seem to fit well

Some uniquenesses are high

Model shows significant misfit

Find # of factors

f3 <- factanal(fdata, factors=3, rotation="oblimin")

Uniquenesses are going down

Model shows significant misfit, but getting less

1-pchisq(f2$STATISTIC-f3$STATISTIC, f2$dof-f3$dof)

Model is significantly better than f2

Find # of factors

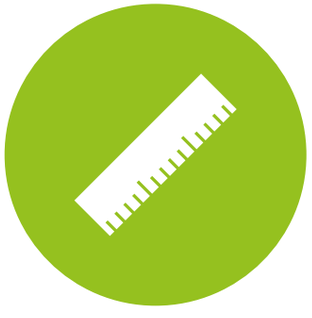
f4 <- factanal(fdata, factors=4, rotation="oblimin")

Only one uniqueness > 0.50 left (employer)

Model at p = .011 (good job for the chi-square test)

1-pchisq(f3$STATISTIC-f4$STATISTIC, f3$dof-f4$dof)

Model is significantly better than f3



Find # of factors

factanal(fdata, factors=5, rotation="oblimin")

5th factor has only a single item

1-pchisq(f4$STATISTIC-f5$STATISTIC, f4$dof-f5$dof)

p = .011

Model is significantly better than f4, but getting closer

Improve solution

	Factor1	Factor2	Factor3	Factor4
cwall	0.810			
cstatus	0.942			
clinks	0.776		0.146	
cnotes	0.790			0.125
cphoto	0.569	0.209		0.140
ctown	0.145	0.698	0.116	
cloccity		0.976		
clocstate		0.960		
clocadress		0.111	-0.105	0.746
cemployer	-0.156	0.311	0.297	0.403
cphone				0.934
cemail			0.211	0.648
creligious			0.810	
cinterest			0.858	
cgroups	0.138		0.755	
cfriends	0.306	0.112	0.462	

Improve solution

Remove “employer” (item 10):

Why? Maybe because not everyone is employed!

factanal(fdata[-c(10)],factors=4,rotation=“oblimin”)

Remove “friends” (item 16):

Why? Maybe because not everyone has the same number of friends!

factanal(fdata[-c(10,16)],factors=4,rotation=“oblimin”)

Inspect solution

Model seems to be a good fit!

All uniqueness < 0.50

No large cross-loadings

p-value = 0.0144

Inspect solution

	Factor1	Factor2	Factor3	Factor4
cwall	0.831			
cstatus	0.963			
clinks	0.793		0.136	
cnotes	0.795			0.120
cphoto	0.569	0.207		0.138
ctown	0.148	0.705	0.105	
cloccity		0.979		
clocstate		0.961		
clocadress		0.130	-0.101	0.724
cphone				0.957
cemail			0.209	0.632
creligious			0.812	
cinterest			0.882	
cgroups	0.155		0.716	

Inspect solution

Let's try some alternatives using the "fa" function in psych:

Geomin rotation instead of Oblimin, WLS estimator instead of ML:

```
f4 <- fa(fdata[,-c(10,16)],nfactors=4,rotate="geominQ",
fm="wls")
```

Inspect solution

communalities and uniquenesses

	WLS1	WLS3	WLS4	WLS2	h2	u2	com
cwall	0.79	0.01	0.00	-0.01	0.62	0.377	1.0
cstatus	0.95	0.00	0.00	-0.01	0.90	0.103	1.0
clinks	0.78	-0.03	0.16	-0.03	0.73	0.266	1.1
cnotes	0.78	0.02	0.02	0.12	0.77	0.225	1.1
cphoto	0.56	0.23	-0.01	0.13	0.65	0.353	1.5
ctown	0.15	0.70	0.12	0.01	0.77	0.229	1.1
cloccity	-0.03	0.98	-0.02	0.04	0.94	0.064	1.0
clocstate	0.00	0.96	0.02	-0.04	0.90	0.096	1.0
clocadress	0.06	0.13	-0.07	0.73	0.66	0.336	1.1
cphone	-0.05	-0.03	0.03	0.96	0.87	0.132	1.0
cemail	0.04	0.00	0.23	0.63	0.57	0.429	1.3
creligious	-0.06	-0.03	0.80	0.05	0.58	0.425	1.0
cinterest	0.03	0.03	0.89	-0.01	0.85	0.153	1.0
cgroups	0.14	0.09	0.72	0.00	0.78	0.223	1.1

complexity
(amount of cross-loading)

Inspect solution

Model fit

Chi-square = 65.16, df = 41, p-value = 0.0095

TLI = 0.987

RMSEA = 0.042, 90% CI: [0.020, 0.058]

Model comparison

```
f5 <- fa(fdata[,-c(10,16)],nfactors=5,rotate="geominQ",
fm="wls")
```

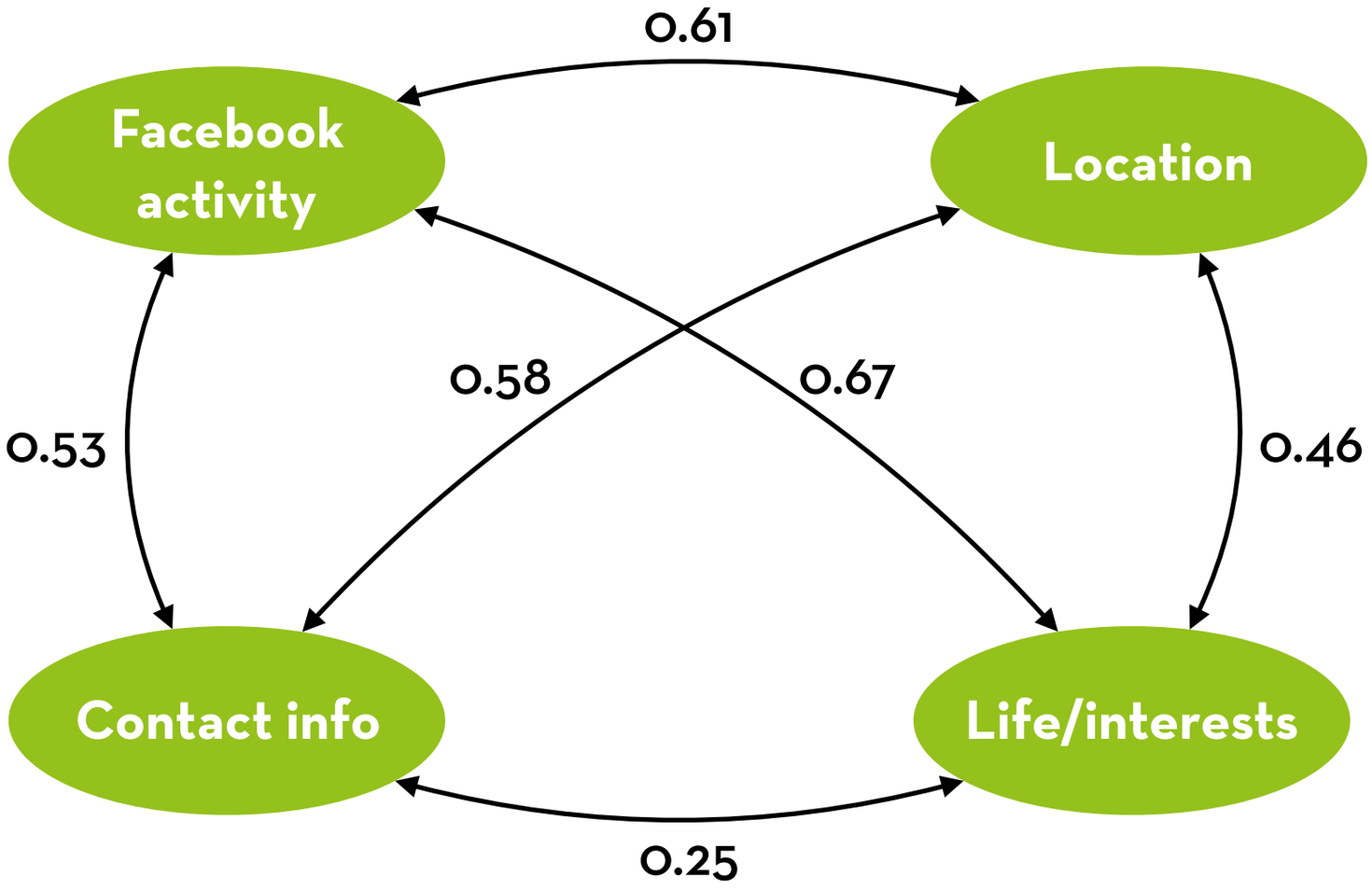
```
1-pchisq(f4$STATISTIC-f5$STATISTIC,f4$dof-f5$dof)
```

f5 is not significantly better than f4!

Name factors

Type of data	ID	Items
Facebook activity	1	Wall
	2	Status updates
	3	Shared links
	4	Notes
	5	Photos
Location	6	Hometown
	7	Location (city)
	8	Location (state/province)
Contact info	9	Residence (street address)
	11	Phone number
	12	Email address
Life/interests	13	Religious views
	14	Interests (favorite movies, etc.)
	15	Facebook groups

Factor correlation



“It is the mark of a truly intelligent person
to be moved by statistics.”

T H A N K S !

George Bernard Shaw