

Principal Components An Introduction

- Exploratory factoring
- Meaning & application of “principal components”
- Basic steps in a PC analysis
- PC extraction process
- # PCs determination
- PC rotation & interpretation
- PC Scores
- Tour of PC MAtrices

Exploratory vs. Confirmatory Factoring

Exploratory Factoring – when we do not have RH: about . . .

- the number of factors
- what variables load on which factors
- we will “explore” the factor structure of the variables, consider multiple alternative solutions, and arrive at a *post hoc* solution

Weak Confirmatory Factoring – when we have RH: about the # factors and factor memberships

- we will “test” the proposed weak *a priori* factor structure

Strong Confirmatory Factoring – when we have RH: about relative strength of contribution to factors by variables

- we will “test” the proposed strong *a priori* factor structure



Meaning of “Principal Components”

“Component” analyses are those that are based on the “full” correlation matrix

- 1.00s in the diagonal
 - yep, there's other kinds, more later

“Principal” analyses are those for which each successive factor...

- accounts for maximum available variance
- is orthogonal (uncorrelated, independent) with all prior factors
- full solution (as many factors as variables) accounts for all the variance

Component Scores

- ☞ A principal component is a composite variable formed as a linear combination of measure variables
- ☞ A component SCORE is a person's score on that composite variable -- when their variable values are applied to the formulas shown below
 - ☞ usually computed from Z-scores of measured variables
 - ☞ the resulting PC scores are also Z-scores ($M=0, S=1$)
$$PC_1 = \beta_{11}Z_1 + \beta_{21}Z_2 + \dots + \beta_{k1}Z_k$$
$$PC_2 = \beta_{12}Z_1 + \beta_{22}Z_2 + \dots + \beta_{k2}Z_k \quad (\text{etc.})$$
- ☞ Component scores have the same properties as the components they represent (e.g., orthogonal or oblique)

Proper & Improper Component Scores

- ☞ A proper component score is a linear combination of all the variables in the analysis
 - ☞ the appropriate β s applied to variable Z-scores
 - ☞ An improper component score is a linear combination of the variables which "define" that component
 - ☞ usually an additive combination of the Z-scores of the variables with structure weights beyond the chosen cut-off value
- (Note: improper doesn't mean "wrong" -- it means "not derived from optimal OLS weightings")

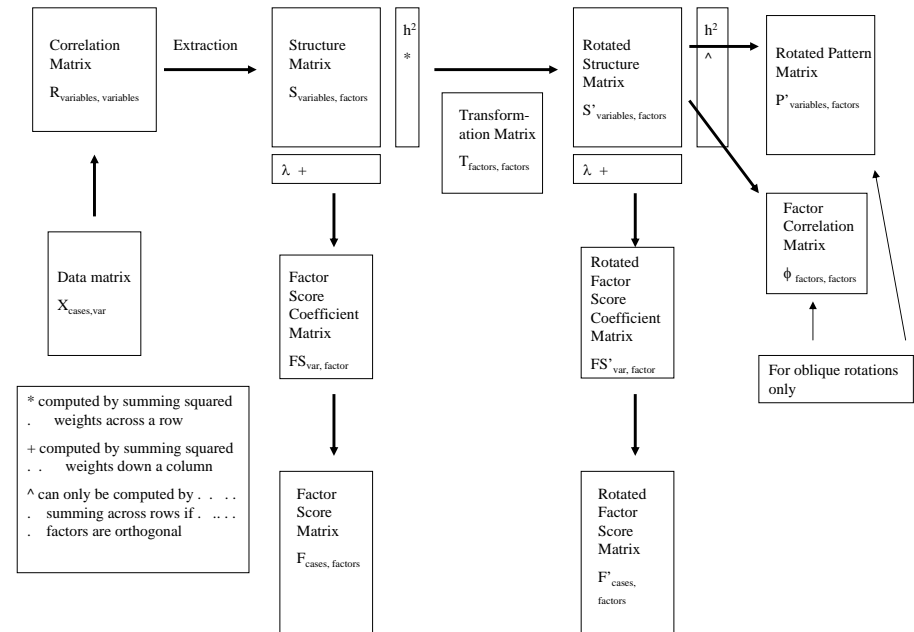
Proper Component Scores

- ☞ Proper component scores are the "instantiation" of the components *as they were mathematically derived from R* (a linear combination of all the variables)
- ☞ Proper component scores have the same properties as components
 - ☞ they are correlated with each other the same as are the PCs
 - ☞ PC scores from orthogonal components are orthogonal
 - ☞ PC scores from oblique components have $r = \phi$
 - ☞ they can be used to produce the structure matrix (corr of component scores and variables scores), communalities, variance accounted for, etc.

Improper Component Scores

- ☞ Improper component scores are the “instantiation” of the components *as they were interpreted by the researcher* (a linear combination of the variables which define that component)
- ☞ Improper component scores usually don’t have exactly the same properties as components
 - ☞ they are usually correlated with each other whether based on orthogonal or oblique solutions
 - ☞ they can not be used to produce the structure matrix (corr of component scores and variables scores), communalities, variance accounted for, etc.

Tour of PC matrices



Applications of PC analysis

Components analysis is a kind of “data reduction”

- start with an inter-related set of “measured variables”
- identify a smaller set of “composite variables” that can be constructed from the “measured variables” and that carry as much of their information as possible

A “Full components solution” ...

- has as many PCs as variables
- accounts for 100% of the variables’ variance
- each variable has a final communality of 1.00 – all of its variance is accounted for by the full set of PCs

A “Truncated components solution” ...

- has fewer PCs than variables
- accounts for <100% of the variables’ variance
- each variable has a communality < 1.00 -- not all of its variance is accounted for by the PCs

The basic steps of a PC analysis

- Compute the correlation matrix
- Extract a full components solution
- Determine the number of components to “keep”
 - total variance accounted for
 - variable communalities
 - interpretability
 - replicability
- “Rotate” the components and “interpret” (name) them
 - Structure weights $> |.3|-.4|$ define which variables “load”
- Compute “component scores”
- “Apply” components solution
 - theoretically -- understand meaning of the data reduction
 - statistically -- use the component scores in other analyses



PC Factor Extraction

- Extraction is the process of forming PCs as linear combinations of the measured variables

$$PC_1 = b_{11}X_1 + b_{21}X_2 + \dots + b_{k1}X_k$$

$$PC_2 = b_{12}X_1 + b_{22}X_2 + \dots + b_{k2}X_k$$

$$PC_f = b_{1f}X_1 + b_{2f}X_2 + \dots + b_{kf}X_k$$

- Here’s the thing to remember...
 - We usually perform factor analyses to “find out how many groups of related variables there are” ... however ...
 - The mathematical goal of extraction is to “reproduce the variables’ variance, efficiently”

PC Factor Extraction, cont.

- Consider R on the right
- Obviously there are 2 kinds of information among these 4 variables
 - X_1 & X_2 X_3 & X_4
- Looks like the PCs should be formed as,

	X_1	X_2	X_3	X_4
X_1	1.0			
X_2	.7	1.0		
X_3	.3	.3	1.0	
X_4	.3	.3	.5	1.0

$$PC_1 = b_{11}X_1 + b_{21}X_2 + 0X_3 + 0X_4$$

$$PC_2 = 0X_1 + 0X_2 + b_{32}X_3 + b_{42}X_4$$

But remember, PC extraction isn’t trying to “group variables” it is trying to “reproduce variance”

- notice that there are “cross correlations” between the “groups” of variables !!

PC Factor Extraction, cont.

- So, because of the cross correlations, in order to maximize the variance reproduced, PC1 will be formed more like ...

$$PC_1 = .5X_1 + .5X_2 + .4X_3 + .4X_4$$

- Notice that all the variables contribute to defining PC₁
- Notice the slightly higher loadings for X₁ & X₂
- Because PC₁ didn't focus on the X₁ & X₂ variable group or X₃ & X₄ variable group, there will still be variance to account for in both, and PC₂ will be formed, probably something like ...

$$PC_2 = .3X_1 + .3X_2 - .4X_3 - .4X_4$$

- Notice that all the variables contribute to defining PC₂
- Notice the slightly higher loadings for X₃ & X₄

PC Factor Extraction, cont.

- While this set of PCs will account for lots of the variables' variance -- it doesn't provide a very satisfactory interpretation
 - PC₁ has all 4 variables loading on it
 - PC₂ has all 4 variables loading on it and 2 of them have negative weights, even though all the variables are positively correlated with each other
- The goal here was point out what extraction does (maximize variance accounted for) and what it doesn't do (find groups of variables)

Determining the Number of PCs

Determining the number of PCs is arguably the most important decision in the analysis ...

- rotation, interpretation and use of the PCs are all influenced by the how many PCs are "kept" for those processes
- there are many different procedures available – none are guaranteed to work !!
- probably the best approach to determining the # of PCs...
 - remember that this is an exploratory factoring -- that means you don't have decent RH: about the number of factors
 - So ... Explore ...
 - consider different "reasonable" # PCs and "try them out"
 - rotate, interpret &/or tryout resulting factor scores from each and then decide

To get started we'll use the SPSS "standard" of $\lambda > 1.00$

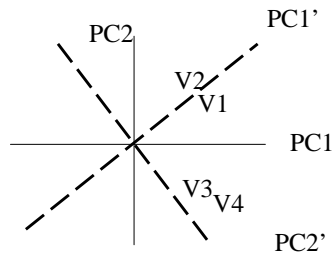


Rotation – finding “groups” in the variables

Factor Rotations

- changing the “viewing angle” or “head tilt” of the factor space
- makes the groupings visible in the graph apparent in the structure matrix

Unrotated Structure		
	PC ₁	PC ₂
V ₁	.7	.5
V ₂	.6	.6
V ₃	.6	-.5
V ₄	.7	-.6



Rotated Structure		
	PC ₁	PC ₂
V ₁	.7	-.1
V ₂	.7	.1
V ₃	.1	.5
V ₄	.2	.6


Interpretation – Naming “groups” in the variables

Usually interpret factors using the rotated solutions using the rotated

- Factors are named for the variables correlated with them
- Usual “cutoffs” are +/- .3 - .4
- So ... a variable that shares at least 9-16% of its variance with a factor is used to name that factor
- Variables may “load” on none, 1 or 2+ factors

Rotated Structure		
	PC ₁	PC ₂
V ₁	.7	-.1
V ₂	.7	.1
V ₃	.1	.5
V ₄	.2	.6

This rotated structure is easy – PC₁ is V₁ & V₂ PC₂ is V₃ & V₄

It is seldom this easy !?!?! 

“Kinds” of Factors

- General Factor
 - all or “almost all” variables load
 - there is a dominant underlying theme among the set of variables which can be represented with a single composite variable
- Group Factor
 - some subset of the variables load
 - there is an identifiable sub-theme in the variables that must be represented with a specific subset of the variables
 - “smaller” vs. “larger” group factors (# vars & % variance)
- Unique Factor
 - single variable loads

“Kinds” of Variables

- Univocal variable -- loads on a single factor
- Multivocal variable -- loads on 2+ factors
- Nonvocal variable -- doesn't load on any factor

You should notice a pattern here...

- a higher “cutoff” (e.g., .40) tends to produce ...
 - fewer variables loading on a given factor
 - less likely to have a general factor
 - fewer multivocal variables
 - more nonvocal variables
- a lower “cutoff” (e.g., .30) tends to produce ...
 - more variables loading on a given factor
 - more likely to have a general factor
 - more multivocal variables
 - fewer nonvocal variables



Component Scores

☞ A principal component is a composite variable formed as a linear combination of measure variables

☞ A component SCORE is a person's score on that composite variable -- when their variable values are applied to the formulas shown below

☞ usually computed from Z-scores of measured variables

☞ the resulting PC scores are also Z-scores (M=0, S=1)

$$PC_1 = \beta_{11}Z_1 + \beta_{21}Z_2 + \dots + \beta_{k1}Z_k$$

$$PC_2 = \beta_{12}Z_1 + \beta_{22}Z_2 + \dots + \beta_{k2}Z_k \text{ (etc.)}$$

☞ Component scores have the same properties as the components they represent (e.g., orthogonal or oblique)

Proper & Improper Component Scores

☞ A proper component score is a linear combination of all the variables in the analysis

☞ the appropriate β s applied to variable Z-scores

☞ An improper component score is a linear combination of the variables which “define” that component

☞ usually an additive combination of the Z-scores of the variables with structure weights beyond the chosen cut-off value

(Note: improper doesn't mean “wrong” -- it means “not derived from optimal OLS weightings”)

Proper Component Scores

- ☞ Proper component scores are the “instantiation” of the components *as they were mathematically derived from R* (a linear combination of all the variables)
- ☞ Proper component scores have the same properties as components
 - ☞ they are correlated with each other the same as are the PCs
 - ☞ PC scores from orthogonal components are orthogonal
 - ☞ PC scores from oblique components have $r = \phi$
 - ☞ they can be used to produce the structure matrix (corr of component scores and variables scores), communalities, variance accounted for, etc.

Improper Component Scores

- ☞ Improper component scores are the “instantiation” of the components *as they were interpreted by the researcher* (a linear combination of the variables which define that component)
- ☞ Improper component scores usually don't have exactly the same properties as components
 - ☞ they are usually correlated with each other whether based on orthogonal or oblique solutions
 - ☞ they can not be used to produce the structure matrix (corr of component scores and variables scores), communalities, variance accounted for, etc.

Tour of PC matrices

