# Principal Component Analysis and Generalizations

Marieke Timmerman
Heymans Institute for Psychology
University of Groningen, The Netherlands
m.e.timmerman@rug.nl

## Component Analysis

- Aim:
  - Summarize large number of variables into much smaller number of components, while retaining as much as possible of the variation present in the variables
- Use in Psychology:
  - data reduction: reduce set of items to smaller set of more reliable measures
  - identifying different dimensions (e.g., of personality traits, impulsive behavior)



## Principal Component Analysis (PCA): Frequency of use

Review of articles in *Personality and Social Psychology Bulletin* in 1996, 1998, 2000 (Russell, 2002):

156 Factor Analyses (FA)
19 Confirmatory FA
137 Exploratory FA

| *Type of Exploratory Factor Analysis* | *%* |
|---|---|
| Principal Component Analysis | 62 |
| Principal Axes Factoring | 21 |
| Unspecified | 26 |
| Other | 1 |



## Principal Component Analysis

**X** $(I \times J)$: observed scores of $I$ subjects on $J$ variables

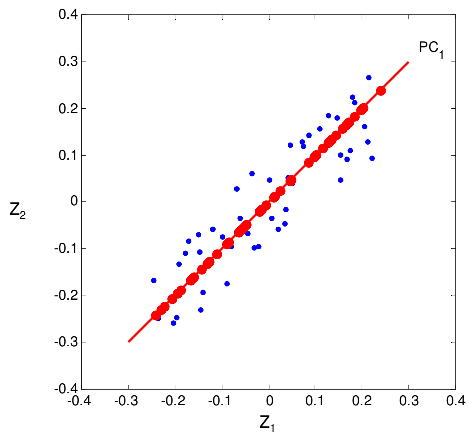
**Z**: standardized scores of **X**, $n^{-1}\mathbf{Z}'\mathbf{Z}=\mathbf{R}$

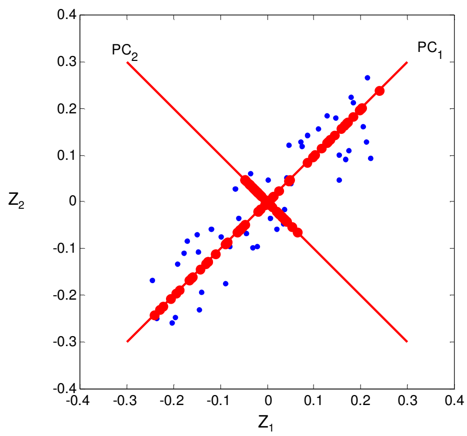
Principal Components (PCs):
standardized linear combinations of **Z**

$$\left.\begin{array}{l}\mathbf{f}_1=\mathbf{Zb}_1\\ \mathbf{f}_2=\mathbf{Zb}_2\\ \ldots\\ \mathbf{f}_Q=\mathbf{Zb}_Q\end{array}\right\}\ \mathbf{F},\ \text{with } n^{-1}\mathbf{F}'\mathbf{F}=\mathbf{I}$$

such that $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_Q$ successively explain as much of the variance of **Z** as possible





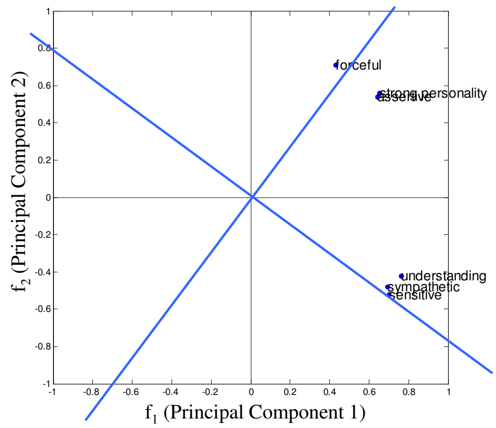
# Interpreting PCs

$\mathbf{A}=n^{-1}\mathbf{Z}'\mathbf{F}$ : loading matrix, correlations between variables and components

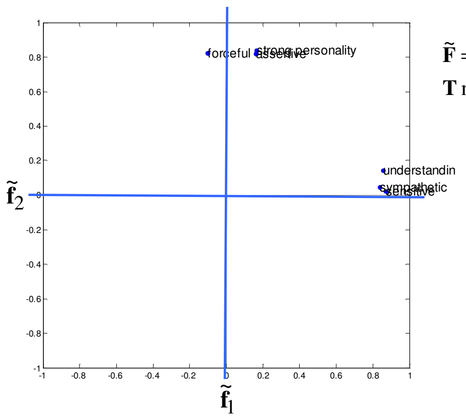
*Example:* Loading matrix of BEM Sex Role Inventory ($n$=681)

| | $\mathbf{f}_1$ | $\mathbf{f}_2$ |
|---|---|---|
| sympathetic | .69 | -.48 |
| sensitive to needs of others | .70 | -.52 |
| understanding | .76 | -.42 |
| assertive | .64 | .54 |
| strong personality | .65 | .56 |
| forceful | .43 | .71 |








$\tilde{\mathbf{F}} = \mathbf{FT}$

$\mathbf{T}$ non - singular



# Rotated Loading matrix

| | $\tilde{\mathbf{f}}_1$ | $\tilde{\mathbf{f}}_2$ |
|---|---|---|
| sympathetic | **.84** | .05 |
| sensitive to needs of others | **.87** | .02 |
| understanding | **.85** | .14 |
| assertive | .16 | **.82** |
| strong personality | .16 | **.84** |
| forceful | .10 | **.82** |



# Effects of Rotation

- Rotation
  - does not alter total explained variance:

$$\hat{\mathbf{Z}} = \mathbf{FA}'$$
$$\hat{\mathbf{Z}} = \mathbf{FT}(\mathbf{T}^{-1})\mathbf{A}'$$

  - alters explained variance per component: no longer PCs
    - information on nature of dominant dimensions is lost



# Aims of Rotation

- offer better interpretable components: simple structure (Thurstone, 1947)
  - using rotation criterion, e.g., Varimax, Orthomax, Oblimin, …
- enable optimal comparison to other loading pattern (hypothesized, previous research, …)
  - using Procrustes rotation

# Orthogonal/oblique rotation

$\tilde{\mathbf{F}} = \mathbf{FT}$ $\qquad$ $\hat{\mathbf{Z}} = \mathbf{FT}(\mathbf{T}^{-1})\mathbf{A}'$

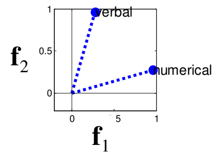
- Orthogonal rotation: **T** orthonormal
  - rotated components still uncorrelated
  - rotated loadings still correlations between variables and components
- Oblique rotation: **T** non-singular
  - rotated components correlated
  - rotated loadings no longer correlations between variables and components
    - structure matrix **S**: correlations between variables and components
    - pattern matrix **P**: weights to estimate **Z** from **FT**



# Use of Oblique rotation

- 'Artificial' orthogonality:

$$\mathbf{A} = \begin{array}{c} \textit{verb.}\ \mathbf{z}_1 \\ \textit{verb.}\ \mathbf{z}_2 \\ \textit{num.}\ \mathbf{z}_3 \\ \textit{num.}\ \mathbf{z}_4 \end{array} \overset{\begin{array}{cc} \mathbf{f}_1 & \mathbf{f}_2 \end{array}}{\begin{bmatrix} .28 & .96 \\ .28 & .96 \\ .96 & .28 \\ .96 & .28 \end{bmatrix}}$$



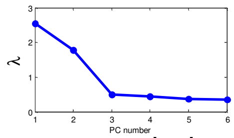

- After oblique rotation:

$$\tilde{\mathbf{S}} = \begin{array}{c} \textit{verb.}\ \mathbf{z}_1 \\ \textit{verb.}\ \mathbf{z}_2 \\ \textit{num.}\ \mathbf{z}_3 \\ \textit{num.}\ \mathbf{z}_4 \end{array} \overset{\begin{array}{cc} \tilde{\mathbf{f}}_1 & \tilde{\mathbf{f}}_2 \end{array}}{\begin{bmatrix} .54 & 1 \\ .54 & 1 \\ 1 & .54 \\ 1 & .54 \end{bmatrix}} \qquad \tilde{\mathbf{P}} = \begin{array}{c} \textit{verb.}\ \mathbf{z}_1 \\ \textit{verb.}\ \mathbf{z}_2 \\ \textit{num.}\ \mathbf{z}_3 \\ \textit{num.}\ \mathbf{z}_4 \end{array} \overset{\begin{array}{cc} \tilde{\mathbf{f}}_1 & \tilde{\mathbf{f}}_2 \end{array}}{\begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}}$$



# How many PCs to retain?

- Interpretability
- Cattell's Scree plot
  - Balancing explained variance and number of components




- Kaiser's eigenvalue-over-one criterion
  - only those PCs that explain more variance than each observed variable in **Z**
- …



# Inference in PCA

- Personality research
  - Individuals scoring a series of adjectives
  - Single individual repeatedly scores a series of adjectives for one month, on a daily basis
- Parameters of interest:
  Loadings, but what are the population loadings?
  - A. Principal Component loadings
  - B. Procrustes rotated loadings to external structure
  - C. Loadings after rotation to simple structure using particular, prespecified criterion
  - D. Rotation to 'best interpretable simple structure', no prespecified criterion



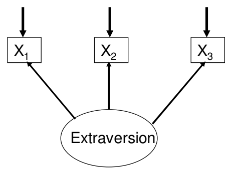
# Confidence intervals in PCA

- Based on distributional results
  - mostly asymptotic
  - often assuming multivariate normality
  - only for PC loadings, and rotated loadings using prespecified criterion
- Bootstrap confidence intervals
  - for all four types of population loadings
- Bootstrap favorite (Timmerman, Kiers & Smilde, 2007)
  - much more flexible
  - more efficient



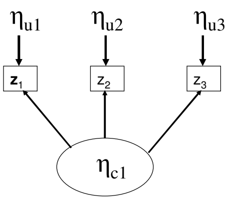
# Should we have used Common Factor Analysis (CFA)?

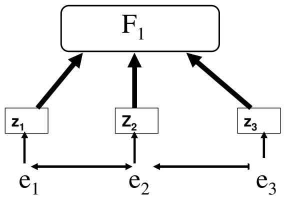
- Use [of PCA] in Psychology:
  - identifying different dimensions (e.g., of personality traits, impulsive behavior)
- Latent traits, reflective models?



# CFA versus PCA

$z = P\eta_c + U\eta_u$ $\quad$ $z = Af + e$




19

---

- CFA population – CFA parameter estimates

  $z = P\eta_c + U\eta_u$

- 'PCA population' – PCA parameter estimates

  $z = Af + e$

20

---

- 'PCA population' – CFA parameter estimates

  $z = Af + e$ $\quad$ $z = P\eta_c + U\eta_u$

  - CFA Parameter estimates?!
    Depend on number of estimated factors and estimation procedure
    (e.g., Maximum Likelihood FA, Minres)
    - non-perfect model fit

21

---

- CFA population – PCA parameters

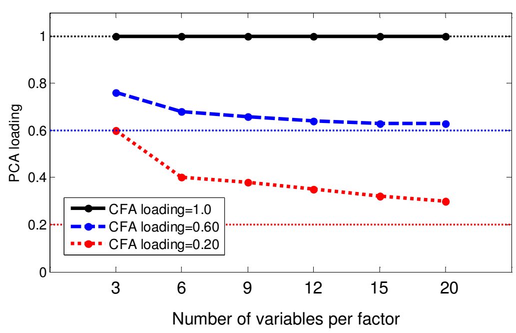
  $z = P\eta_c + U\eta_u$ $\quad$ $z = Af + e$

  - PCA loadings upwards biased (a.o. Widaman, 1993)
    - Bias reduces with
      - increasing communalities
      - increasing number of variables per factor
    - Example: simple structure

$$\mathbf{P} = \begin{bmatrix} p & 0 \\ p & 0 \\ p & 0 \\ 0 & p \\ 0 & p \\ 0 & p \end{bmatrix}$$

22

---




Based on Widaman (2007)

23

---

# PCA versus CFA in practice

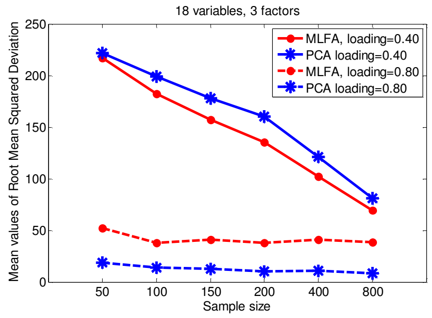
- Effects of model error
  - presence of minor factors

  $$z = P_{major}\,\eta_{c,major} + P_{minor}\,\eta_{c,minor} + U\eta_u$$

  - violations of distributional assumptions

24

- Effects of sample size
  - Recovery of population pattern with PCA and MLFA (From: Velicer & Fava, 1998)



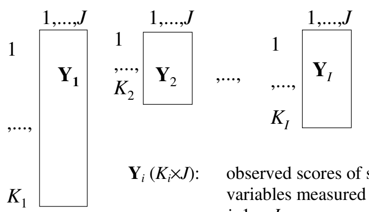

25

# Generalizations of PCA

26

# Multilevel Component Analysis

(MLCA; Timmerman, Kiers and others, 2003, 2006, 2008, in press)




$\mathbf{Y}_i$ ($K_i \times J$): observed scores of subject $i$ on $J$ variables measured at $K_i$ occasions, $i$=1,...,$I$

27

grand mean | between part | within part

$$\begin{bmatrix} \\ \mathbf{Y}_i \\ \\ \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \\ \dots \\ \mathbf{m} \end{bmatrix} + \begin{bmatrix} \mathbf{y}_{ib}' \\ \mathbf{y}_{ib} \\ \dots \\ \mathbf{y}_{ib}' \end{bmatrix} + \begin{bmatrix} \\ \mathbf{Y}_{iw} \\ \\ \end{bmatrix}$$

$i$=1,…,$I$

28

grand mean | between part model | within part model | residuals

$$\begin{bmatrix} \\ \mathbf{Y}_i \\ \\ \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \\ \dots \\ \mathbf{m} \end{bmatrix} + \begin{bmatrix} \mathbf{f}_{ib}' \\ \mathbf{f}_{ib}' \\ \dots \\ \mathbf{f}_{ib}' \end{bmatrix} \mathbf{B}_b' + \begin{bmatrix} \\ \mathbf{F}_{iw} \\ \\ \end{bmatrix} \mathbf{B}_w' + \begin{bmatrix} \\ \mathbf{E}_i \\ \\ \end{bmatrix}$$

with $\mathbf{1}_{K_i}'\mathbf{F}_{iw} = \mathbf{0}_{Q_w}'$

29

# MLCA to unravel trait and state effects

- between model
  - major dimensions of interindividual differences
  - traits
- within model
  - major dimensions of intraindividual differences
  - states

30

# MLCA to unravel trait and state effects

- 12 subjects scored PANAS daily (range 53-71 days)
- PANAS (Positive And Negative Affect Schedule):
  - Originally designed to describe individual differences in traits
    - Positive Affect
    - Negative Affect

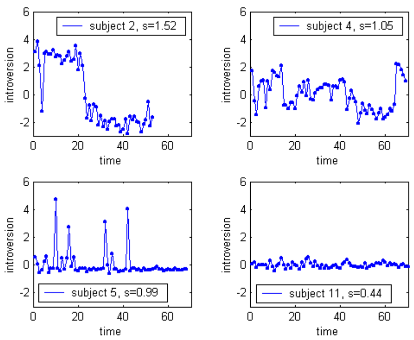
31

## Between and within loading matrices

| Total fit: 61.5 | Between (Fit: 38.3) Positive Affect | Between (Fit: 38.3) Negative Affect | Within (Fit: 23.2) Extraversion | Within (Fit: 23.2) Emotional Instability |
|---|---|---|---|---|
| Jittery | **0.58** | -0.06 | **-0.46** | **0.34** |
| Distressed | **0.61** | -0.19 | 0.01 | **0.48** |
| Upset | **0.67** | -0.19 | -0.00 | **0.52** |
| Afraid | **0.63** | -0.04 | **-0.44** | 0.00 |
| Scared | **0.60** | -0.10 | **-0.46** | -0.04 |
| Hostile | **0.64** | -0.11 | **-0.30** | **0.37** |
| Irritable | **0.71** | -0.21 | 0.03 | **0.48** |
| Guilty | **0.50** | -0.22 | 0.07 | **0.50** |
| Ashamed | **0.50** | -0.08 | **-0.31** | 0.18 |
| Nervous | **0.59** | -0.08 | **-0.43** | **0.33** |
| Inspired | -0.05 | **0.55** | **0.57** | 0.06 |
| Excited | 0.02 | **0.62** | 0.02 | -0.14 |
| Determined | -0.23 | **0.70** | 0.17 | -0.27 |
| Interested | -0.08 | **0.55** | **0.60** | -0.17 |
| Enthusiastic | -0.18 | **0.58** | **0.53** | -0.20 |
| Attentive | -0.09 | **0.58** | **0.55** | -0.17 |
| Proud | -0.25 | **0.63** | -0.02 | -0.10 |
| Strong | -0.20 | **0.63** | 0.04 | **-0.34** |
| Active | -0.07 | **0.54** | **0.47** | **-0.36** |
| Alert | -0.18 | **0.56** | **0.51** | -0.22 |

32

## Within-component scores across time: Introversion




33

# MLCA to unravel effects resulting from an experimental design

(Timmerman, Kiers, Smilde, Ceulemans & Stouten, in press)

- Social dilemma research: public good dilemma game
- Scores obtained on 20 emotion terms, of 420 participants, randomly allocated to 6 conditions

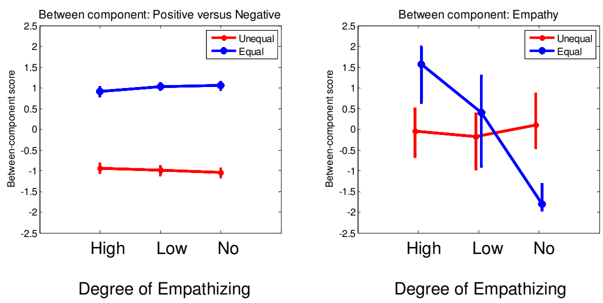
| | | *Amount of Bonus* | |
|---|---|---|---|
| | | Equal | Unequal |
| *Degree of Empathizing* | High | 70 | 70 |
| | Low | 70 | 70 |
| | No | 70 | 70 |

34

| | *Between components* | |
|---|---|---|
| Emotion terms | Positive-versus-Negative-Affect | Empathy |
| hurt | **-.79** | -.04 |
| irritated | **-.89** | -.03 |
| elated | **.82** | -.02 |
| satisfied | **.91** | -.02 |
| relieved | **.66** | -.05 |
| warm | **.79** | -.01 |
| compassionate | -.03 | **.36** |
| concerned | **-.25** | **.27** |
| tenderhearted | -.04 | .12 |
| … | … | … |

Loadings significantly deviating from 0 (α=0.05) in bold face

35




36

# Redundancy analysis (RDA = external PCA)

- RDA: Component Analysis with regression
  - Find Components that optimally predict a set of dependent variables
- Principal Response Curve Analysis
  (Ter Braak, Van den Brink, Timmerman)
  - Special type of RDA, to summarize variation in longitudinal data resulting from experimental design

37

- Effects of insecticide in ditches on invertebrate community?
  - 4 dose levels (0 µg/L to 44 µg/L), with 2-4 replicates per level
  - Samples taken at -4 up to 24 weeks post-treatment

38

- Multivariate responses
  - Samples contained 189 invertebrate taxa
  - Examples of taxa
    - cloeon dipterum
    - daphnia galeata
    - nauplius
  - Frequency of each taxum per sample



39

$t$=1

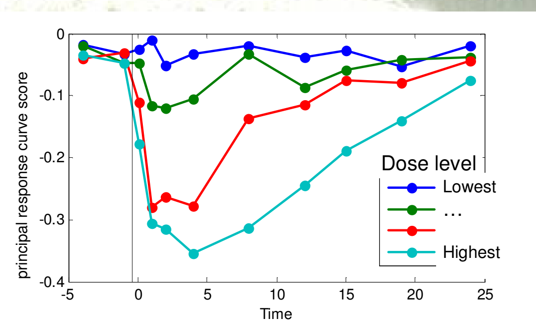
| dose level ($d$=0,…,$D$) | replicate ($i_d$=1,…,$I_d$) | taxa | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | … | 189 |
| $d$=0 (control) | $i_0$=1 | | | | |
| | ,…, | | | | |
| | $I_0$=4 | | | | |
| $d$=$D$=1 | $i_1$=1 | | | | |
| | $I_1$=2 | | | | |
| … | … | | | … | |
| $d$=$D$=4 | $i_4$=1 | | | | |
| | $I_4$=2 | | | | |

…

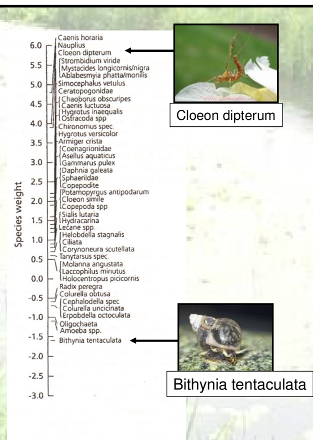
$T$=11

40

## First PRCs of Invertebrate data




41




42

## Component analysis

- Flexible Approach to Summarize Data
- Weak modeling, mild assumptions

43