The European Commission’s science and knowledge service

Joint Research Centre
Step 7: Statistical coherence (II)
PCA, Exploratory Factor Analysis, Cronbach alpha

Hedvig Norlén

COIN 2017 - 15th JRC Annual Training on Composite Indicators & Scoreboards
06-08/11/2017, Ispra (IT)
Step 7: Statistical coherence (II)

• The CI measures a multifaceted phenomenon - a combination of different aspects (SUB-PILLARS/PILLARS) (all related to the phenomenon but different from each other)

• Each aspect can be measured by a set of observable variables (INDICATORS)
PCA, FA are used to verify the internal consistency

• within each SUB-PILLAR (across INDICATORS)
PCA, FA are used to verify the internal consistency

- within each **SUB-PILLAR** (across **INDICATORS**)
- within each **PILLAR** (across **SUB-PILLARS**)

Framework of the Global Innovation Index 2017
PCA, FA are used to verify the internal consistency

- within each **SUB-PILLAR** (across **INDICATORS**)
- within each **PILLAR** (across **SUB-PILLARS**)
- within the **CI** (across **PILLARS**)

Search for **1-DIMENSIONALITY** within each subset
Step 7: Statistical coherence (II)

Want to answer the following questions by using a multivariate method (PCA, FA,...)

• Is the hierarchical structure of our composite indicator statistically well-defined?

• Is the set of available indicators sufficient/appropriate to describe the pillars (dimensions) and the sub-pillars (sub-dimensions) of the composite indicator?
Multivariate data

The type of analysis to be performed depends on the data of the indicators.

- **Quantitative**
  - PCA/FA

- **Qualitative**
  - disPCA/FA,(M)CA
  - FAMD/MFA

- **Mixed/structured**
  - FAMD Factor Analysis of Mixed Data
  - MFA Multiple Factor Analysis

---

PCA  Principal Component Analysis
FA  Factor Analysis

Note: PCA for discrete data (based on polychoric correlations)
(M)CA  (Multiple) Correspondence Analysis
Data matrix of indicators

\[
X = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1p} \\
  x_{21} & x_{22} & \cdots & x_{2p} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{np}
\end{bmatrix}
\]

\[n = \text{number of objects (countries, ...)}\]

\[p = \text{number of variables (indicators)}\]

The previous steps in building our CI

- **Step 3** Outlier detection and missing data estimation

- **Step 4** Normalisation/Standardisation

\(X\) correlated data set (without outliers and (without or a few) missing values)

PCA, FA are based on correlations
Exploratory FA

• Two main types of FA: Exploratory (EFA) and Confirmatory (CFA)

• **EFA** examining the relationships among the variables/indicators and do not have an “à priori” fixed number of factors

• **CFA** a clear idea about the number of factors you will encounter, and about which variables will most likely load onto each factor
Is PCA = Exploratory FA?

- Similar in many ways but an important difference that has effects on how to use them
  1) Both are data reduction techniques - they allow you to capture the variance in variables in a smaller set
  2) Same (or similar) procedures used to run in many software (same steps extraction, interpretation, rotation, choosing the number of components/factors)
Is PCA = Exploratory FA? NO!

• Similar in many ways but an important difference that has effects on how to use them

1) Both are data reduction techniques - they allow you to capture the variance in variables in a smaller set

2) Same (or similar) procedures used to run in many software (same steps extraction, interpretation, choosing the number of components/factors)

• Difference:

  PCA is a linear combination of variables

  FA is a measurement model of a latent variable (statistical model)
Is PCA = Exploratory FA? NO!

PCA

Observed indicators are reduced into components

- Item 1 X₁
- Item 2 X₂
- ...  
- Item p Xₚ

Curved arrows from each item to a component indicate that observed indicators are reduced into components using weights w₁, w₂, ..., wₚ.

Component

PC₁ = w₁X₁ + w₂X₂ + ... + wₚXₚ

PCA summarizes information of all indicators and reduces it into a fewer number of components.

Each PCᵢ is a new variable computed as a linear combination of the original variables.

FA

Latent factors drive the observed variables

Factor

- Item 1 X₁
- Item 2 X₂
- ...  
- Item p Xₚ

Latent factors λ₁, λ₂, ..., λₚ are associated with each observed variable. Error terms ε₁, ε₂, ..., εₚ are associated with the observed variables.

Error terms

Model of the measurement of a latent variable. This latent variable cannot be directly measured with a single variable. Seen through the relationships it causes in a set of X variables.

X₁ = λ₁F₁ + ε₁
Xᵢ = λᵢF₁ + εᵢ
Xₚ = λₚF₁ + εₚ
Some historical notes

- **PCA** probably the oldest of multivariate analysis methods.
- **Pearson (1901)** said that his methods “can be easily applied to numerical problems” and although he says that the calculations become “cumbersome” for four or more variables he suggests that they are “still quite feasible”
- **Hotelling (1933)** chose his “components” so as to maximize their successive contributions to the total of the variances of the original variables, “principal components”
- **FA** was developed by **Spearman (1904)** in the field of psychology (discovered that school children's scores on a wide variety of seemingly unrelated subjects were positively correlated). **Cattell (1952)** expanded on Spearman's idea of a two-factor theory of intelligence
First steps in PCA, FA

- Check the **correlation structure** of the data and do 1) and 2)

1) **Bartlett’s sphericity test**

The Bartlett’s test checks if the observed correlation matrix $R$ diverges significantly from the identity matrix.

$H_0 : |R| = 1 , \ H_1 : |R| \neq 1$

**Test statistic**

$$\chi^2 = -\left( n - 1 - \frac{2p + 5}{6} \right) \log|R|$$

$n = \text{sample size} \quad p = \text{number of indicators}$

Under $H_0$, it follows a $\chi^2$ distribution with a $p(p-1)/2$ degrees of freedom.

Want to reject $H_0$ to be able to do PCA, FA! (check $n/p<5$) (Bartlett (1937))
First steps in PCA FA

• Check the **correlation structure** of the data and do 1) and 2)

2) **Kaiser-Meyer-Olkin (KMO) Measure for Sampling Adequacy**

Indicators are more or less correlated, but the correlation between two indicators can be influenced by the others. Use **partial correlation** in order to measure the relation between two indicators by removing the effect of the remaining indicators (**controlling for the confounding variable**)

\[
KMO = \frac{\sum \sum r_{ij}^2}{\sum \sum r_{ij}^2 + \sum \sum a_{ij}^2}
\]

\(r_{ij}\) Pearson correlation coefficient
\(a_{ij}\) partial correlation coefficient

Want KMO close to 1 to be able to do PCA, FA! KMO>0.6 ok

First steps in \(\text{PCA} \quad \text{FA}\)

- Check the **correlation structure** of the data and do 1) and 2)

2) Kaiser-Meyer-Olkin (KMO) Measure for Sampling Adequacy

Kaiser put the following values on the results:

- 0.00 to 0.49 unacceptable
- 0.50 to 0.59 miserable
- 0.60 to 0.69 mediocre
- 0.70 to 0.79 middling
- 0.80 to 0.89 meritorious
- 0.90 to 1.00 marvelous

1) Bartlett’s sphericity test and
2) KMO test provide info whether it is possible to do PCA, FA but do not give info of the “magic number” - how components/factors are needed
How to get the PCs

• Eigenvalue decomposition of a data correlation matrix (case of CI), (or covariance matrix)

• Singular Value Decomposition (SVD) of a data matrix after mean centering (normalizing) the data matrix
Finding the “magic number”- determining how many components in PCA and factors in FA

Many methods exist. The 3 most common are:

1) **Kaiser–Guttman ‘Eigenvalues greater than one’ criterion**
   (Guttman (1954), Kaiser (1960)). Select all components with eigenvalues over 1 (or 0.9)

2) **Cattell’s scree test**
   (Cattell (1966)) “Above the elbow” approach

3) **Certain percentage of explained variance**
   e.g., >2/3, 75%, 80%,...
Example PCA: The Global Talent Competitiveness Index 2017 (GTCI 2017)

1. Enable

- Use PCA to explore the dimensionality of the first sub-pillar “Regulatory Landscape”
- The 5 variables should be correlated. Supports the assumption that they are all measuring the same concept!
GTCL 2017 Sub-pillar “Regulatory Landscape”

1) Bartlett’s sphericity test

\[ \chi^2 - \text{test statistic } 588, \text{ df } = p(p-1)/2=10 \]
\[ p\text{-value } <<< 0.01 \text{ Reject } H_0 \]

2) KMO Measure for Sampling Adequacy

Overall MSA = 0.86  \text{ KMO}>0.6

MSA for each variable
Government Effectiveness = 0.79
Business-Government Relations = 0.96
Political Stability = 0.94
Regulatory Quality = 0.87
Corruption = 0.84

Check the correlation structure
## GTCI 2017 Sub-pillar “Regulatory Landscape” PCA results

### Total variance explained

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
<th>Variance</th>
<th>Cumulative variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.85</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
<td>0.13</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>0.31</td>
<td>0.06</td>
<td>0.96</td>
</tr>
<tr>
<td>4</td>
<td>0.12</td>
<td>0.02</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>0.06</td>
<td>0.01</td>
<td>1.00</td>
</tr>
<tr>
<td>Sum</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### Component loadings

$$3.85 = 0.96^2 + 0.67^2 + 0.84^2 + 0.94^2 + 0.95^2$$

### Pearson correlation coefficients between pillar and principal component

<table>
<thead>
<tr>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.96</td>
<td>-0.09</td>
<td>-0.17</td>
<td>-0.06</td>
<td>-0.2</td>
</tr>
<tr>
<td>0.67</td>
<td>0.73</td>
<td>0.14</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.84</td>
<td>-0.31</td>
<td>0.44</td>
<td>0.06</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.94</td>
<td>-0.05</td>
<td>-0.23</td>
<td>0.24</td>
<td>0.08</td>
</tr>
<tr>
<td>0.95</td>
<td>-0.10</td>
<td>-0.09</td>
<td>-0.24</td>
<td>0.13</td>
</tr>
</tbody>
</table>

### Stopping criterion

Eigenvalue > 0.90 (or cumulative variance over 0.70)

$$PC_1 = 0.96X_{GE} + 0.67X_{BGR} + 0.84X_{PS} + 0.94X_{RQ} + 0.95X_C$$

One dimension verified!
Example 2 PCA: The Global Talent Competitiveness Index 2017 (GTCI 2017)

- Use PCA to explore the dimensionality of the sub-pillar “Employability”
- 4 variables correlated:
  1. Ease of finding skilled employees
  2. Rel of education system to the economy
  3. Availability of scientists and engineers
  4. Skills gap as major constraint
1) Bartlett’s sphericity test

\( \chi^2 \) - test statistic 264, df = \( p(p-1)/2 = 6 \)

p-value <<< 0.01  \textbf{Reject H}_0

2) KMO Measure for Sampling Adequacy

Overall MSA = 0.73 \textbf{KMO} > 0.6

MSA for each variable
1. Ease of finding skilled employees = 0.67
2. Relev of education system to the eco = 0.80
3. Availability of scientists and engineers = 0.74
4. Skills gap as major constraint = 0.68

Check the correlation structure
GTČI 2017 Sub-pillar “Employability” PCA results

### Total variance explained

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
<th>Variance</th>
<th>Cumulative variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,66</td>
<td>0,66</td>
<td>0,66</td>
</tr>
<tr>
<td>2</td>
<td>0,91</td>
<td>0,23</td>
<td>0,89</td>
</tr>
<tr>
<td>3</td>
<td>0,28</td>
<td>0,07</td>
<td>0,96</td>
</tr>
<tr>
<td>4</td>
<td>0,15</td>
<td>0,04</td>
<td>1,00</td>
</tr>
</tbody>
</table>

### Pearson correlation coefficients between pillar and principal component

<table>
<thead>
<tr>
<th>Indicator</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind 1</td>
<td>0,91</td>
<td>-0,28</td>
<td>-0,08</td>
<td>0,30</td>
</tr>
<tr>
<td>Ind 2</td>
<td>0,90</td>
<td>-0,09</td>
<td>0,41</td>
<td>-0,12</td>
</tr>
<tr>
<td>Ind 3</td>
<td>0,92</td>
<td>-0,04</td>
<td>-0,33</td>
<td>-0,21</td>
</tr>
<tr>
<td>Ind 4</td>
<td>0,42</td>
<td>0,91</td>
<td>0,00</td>
<td>0,07</td>
</tr>
<tr>
<td>SS</td>
<td>2,66</td>
<td>0,91</td>
<td>0,28</td>
<td>0,15</td>
</tr>
</tbody>
</table>

**Stopping criterion Eigenvalue>0,90 (or cumulative variance over 0,70)**

\[
PC_1 = 0,91X_1 + 0,90X_2 + 0,92X_3 + 0,42X_4
\]

\[
PC_2 = -0,28X_1 - 0,09X_2 - 0,04X_3 + 0,91X_4
\]

2 PCs retrieved but Indicator 4 “Skills gap as major constraint” correlates considerably better with the PC2 ...
**Indicator 4 “Skills gap as major constraint” problematic**

**JRC Audit report 2017**: “In the 2017 data set, the indicator “Skills gap as major constraint” has a very low impact on the GTCI (Index)...not found to be important at the overall index level.... It is suggested that the GTCI development team reconsider the inclusion of this variable (or replace it with other variable/s) in next year’s release.”

**GTCI 2018 report (to be published, January 2018)**: “Indicator “Skills gap as a major constraint” has been deleted entirely from the framework, as pointed out by the JRC last year.”
Cronbach’s $\alpha$

Cronbach’s $\alpha$ is measure of internal consistency and reliability

Based upon classical test theory. Most common estimate of reliability, Cronbach (1951)

Cronbach’s $\alpha$ is defined as Ratio of the Total Covariance in the “test” to the Total Variance in the “test”. By “test” we mean the set of indicators constituting a sub-pillar/ pillar/CI.

Let $\text{Var}(X_i)$ be the variance of indicator $i$ and $\text{Var}\left(\sum_{i=1}^{p} X_i\right)$ total variance in sub-pillar/pillar/CI

Average covariance of an indicator with any other indicator is

$$\alpha = \frac{\text{Var}\left(\sum_{i=1}^{p} X_i\right) - \sum_{i=1}^{p} \text{Var}(X_i)}{\text{Var}\left(\sum_{i=1}^{p} X_i\right) / p^2} = \frac{p(p-1)}{p-1} \left(\frac{\text{Var}\left(\sum_{i=1}^{p} X_i\right) - \sum_{i=1}^{p} \text{Var}(X_i)}{\text{Var}\left(\sum_{i=1}^{p} X_i\right)} \right)$$

$p = \text{number of indicators}$
Cronbach’s $\alpha$

Cronbach’s $\alpha$ is a measure of internal consistency and reliability.

$$\alpha = \frac{p}{p-1} \left( 1 - \frac{\sum_{i=1}^{p} Var(X_i)}{Var(\sum_{i=1}^{p} X_i)} \right)$$

$p = \text{number of indicators}$

$\alpha$ coefficient ranges from 0 to 1. If all indicators are entirely independent from one another (i.e. are not correlated or share no covariance) $\alpha = 0$. If all of the items have high covariances, then $\alpha$ will approach 1.

Caution! $\alpha$ increases when the number of indicators increases ...

Cronbach (1951)
Cronbach’s α is measure of internal consistency and reliability

\[
\alpha = \frac{p}{p-1} \left( 1 - \frac{\sum_{i=1}^{p} \text{Var}(X_i)}{\text{Var}\left(\sum_{i=1}^{p} X_i\right)} \right)
\]

<table>
<thead>
<tr>
<th>Cronbach's alpha</th>
<th>Internal consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9 ≤ α</td>
<td>Excellent</td>
</tr>
<tr>
<td>0.8 ≤ α &lt; 0.9</td>
<td>Good</td>
</tr>
<tr>
<td>0.7 ≤ α &lt; 0.8</td>
<td>Acceptable</td>
</tr>
<tr>
<td>0.6 ≤ α &lt; 0.7</td>
<td>Questionable</td>
</tr>
<tr>
<td>0.5 ≤ α &lt; 0.6</td>
<td>Poor</td>
</tr>
<tr>
<td>α &lt; 0.5</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>
Cronbach’s $\alpha$ is a measure of internal consistency and reliability.

### Cronbach’s $\alpha$ Table

<table>
<thead>
<tr>
<th>GTCI Sub-pillar</th>
<th>Cronbach's Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Regulatory Landscape&quot;</td>
<td>0.92</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach’s Alpha if indicator is dropped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind 1. Government Effectiveness</td>
<td>0.93</td>
<td>0.88</td>
</tr>
<tr>
<td>Ind 2. Business-Government Relations</td>
<td>0.55</td>
<td>0.95</td>
</tr>
<tr>
<td>Ind 3. Political Stability</td>
<td>0.75</td>
<td>0.92</td>
</tr>
<tr>
<td>Ind 4. Regulatory Quality</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>Ind 5. Corruption</td>
<td>0.92</td>
<td>0.88</td>
</tr>
</tbody>
</table>

#### Cronbach's alpha

- $0.9 \leq \alpha$: Excellent
- $0.8 \leq \alpha < 0.9$: Good
- $0.7 \leq \alpha < 0.8$: Acceptable
- $0.6 \leq \alpha < 0.7$: Questionable
- $0.5 \leq \alpha < 0.6$: Poor
- $\alpha < 0.5$: Unacceptable
Cronbach’s $\alpha$ is measure of internal consistency and reliability

<table>
<thead>
<tr>
<th>GTCI Sub-pillar</th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Employability&quot;</td>
<td>0.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach’s Alpha if indicator is dropped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind 1. Ease of finding skilled employees</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>Ind 2. Relev of educ system to the economy</td>
<td>0.67</td>
<td>0.58</td>
</tr>
<tr>
<td>Ind 3. Availability of scientists and engineers</td>
<td>0.75</td>
<td>0.58</td>
</tr>
<tr>
<td>Ind 4. Skills gap as major constraint</td>
<td>0.23</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Cronbach's alpha  | Internal consistency |
-------------------|----------------------|
0.9 ≤ $\alpha$    | Excellent            |
0.8 ≤ $\alpha$ < 0.9 | Good                |
0.7 ≤ $\alpha$ < 0.8 | Acceptable          |
0.6 ≤ $\alpha$ < 0.7 | Questionable        |
0.5 ≤ $\alpha$ < 0.6 | Poor                |
$\alpha$ < 0.5     | Unacceptable         |

Ind 4 problematic and Cronbach’s $\alpha$ may hence be used as another measure to include/exclude indicators!
A practically important violation of the normality assumption underlying the PCA occurs when the data are discrete!

Several kinds of discrete data (binary, ordinal, count, nominal...)

Different ways on how to proceed:

1) Simple approach, use the discrete x’s as if they were continuous in the PCA but instead of using the Pearson moment correlation coefficients use Spearman or Kendall rank correlation coefficients.

2) PCA using polychoric correlations. Pearson (1901) founder of this approach (as well!), developed further by Olsson (1979) and Jöreskog (2004)
The polychoric correlation (PCC) coefficient is a measure of association for ordinal variables which rests upon an assumption of an underlying joint continuous normal distribution. (Later generalizations of different continuous distributions).

PCC coefficient is MLE of the product-moment correlation between the underlying normal variables.
References

Books


• Leandre, Fabrigar, Duane, “Exploratory Factor Analysis”, Oxford (2011)


• Tucker, MacCallum, “Exploratory Factor Analysis” (1997)
THANK YOU

Welcome to email us at: jrc-coin@ec.europa.eu

COIN in the EU Science Hub

COIN tools are available at:
https://composite-indicators.jrc.ec.europa.eu/