SOSEWEST

Workshop Economico, Statistico e Tecnologico

sose ROMA - 28 SETTEMBRE 2018

NUOVE FRONTIERE SULL'UTILIZZO DELLE INFORMAZIONI

PER FISCO, ENTI LOCALI E PMI

#SOSEWEST18

Misure di valutazione mediante indicatori compositi

Maurizio Vichi

Dipartimento di Scienze Statistiche

Sapienza Università di Roma

SAPIENZA

UNIVERSITÀ DI ROMA

Composite Indicator

DEFINITION (OECD, 2004) A Composite Indicator (CI) is formed when **observed** (manifest) **indicators** (MIs) are **compiled** into a **single index**, on the basis of an underlying **model** of the multi-dimensional **concept** that is being measured.

CONCEPT: is the first notion considered to characterize a CI. It can be measured only **indirectly** and it **relates** to a **fact**, in general to a phenomenon that, for its complexity and multidimensionality, is **not sufficiently described** by a **single indicator**

EXAMPLES: Concepts such as for example **poverty**, **satisfaction**, human development, **gender equality**, well-being, **intelligence** cannot be satisfactorily represented by individual indicators and therefore need to be described by several variables

MODEL: is the second notion considered to characterize a CI. It is required to simplify and synthesize the complexity of the reality by means of a mathematically-formalized **reconstruction** of the **observed data** and their main relations.

In the **statistical development process used to specify the appropriate** CI model for the studied phenomenon, **three integral parts** are needed:
variable selection, to properly characterize phenomenon object of study,
model selection from a set of candidate models and
model assessment to evaluate the performances of the CI.

STRUCTURE of the MODEL: a **hierarchical structure** that goes from the original MIs to the final **General Composite Indicator** (GCI), passing through a reduced set of Specific Composite Indicators (SCIs), i.e., dimensions, which measure specific concepts describing the main components of the phenomenon under study

PROS and CONS by JRC

Composite Indicators

Advantages:

- Support decision makers by summarizing complex or multi-dimensional issues
- Provide the "big picture", highlight common trends
- Measure a latent phenomenon that is not directly measureable
- Attract public interest by benchmarking

The development **process** helps
- Better understand how a system functions
- Identify latent dimensions, overlaps, redundancies or trade-offs between components

Pitfalls:

- Offer misleading, non-robust policy messages if they are poorly constructed or misinterpreted
- May invite politicians to draw simplistic policy conclusions
- Easier to "manipulate" than individual indicators; the selection of sub-indicators and weights could be the target of political challenge

Assessing their **quality and validity** is particularly relevant

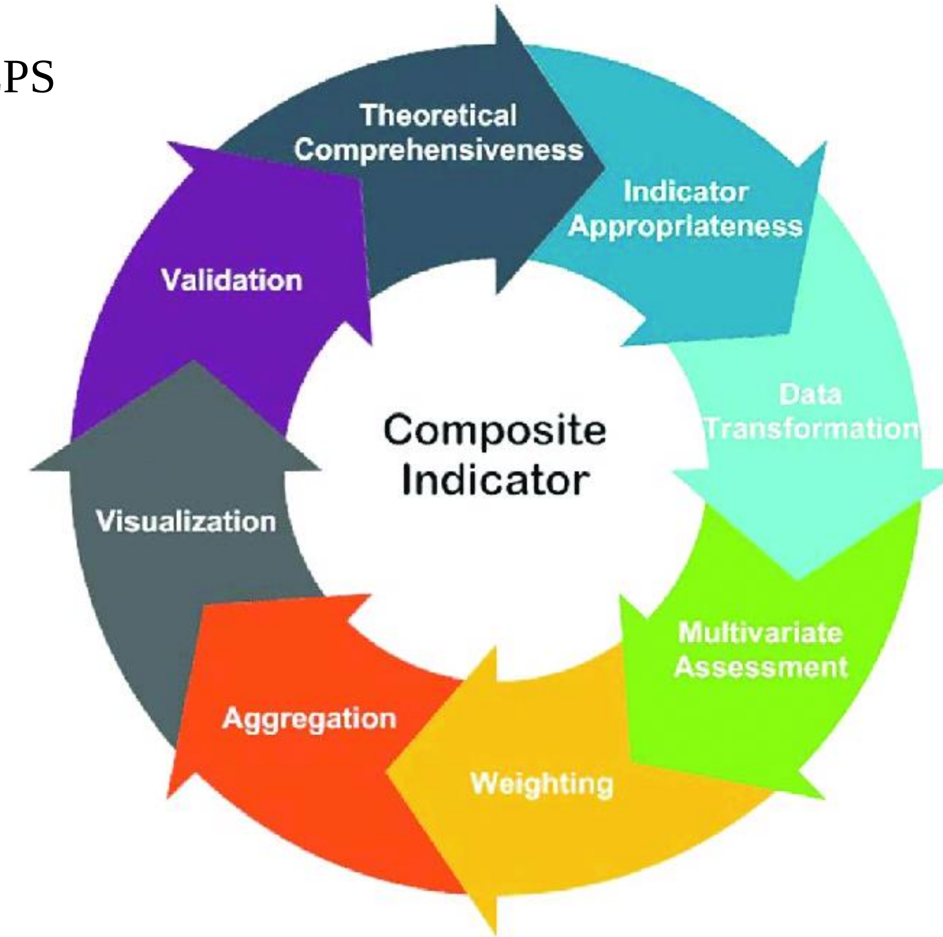
European Commission

Ingredients for Constructing Composite indicators

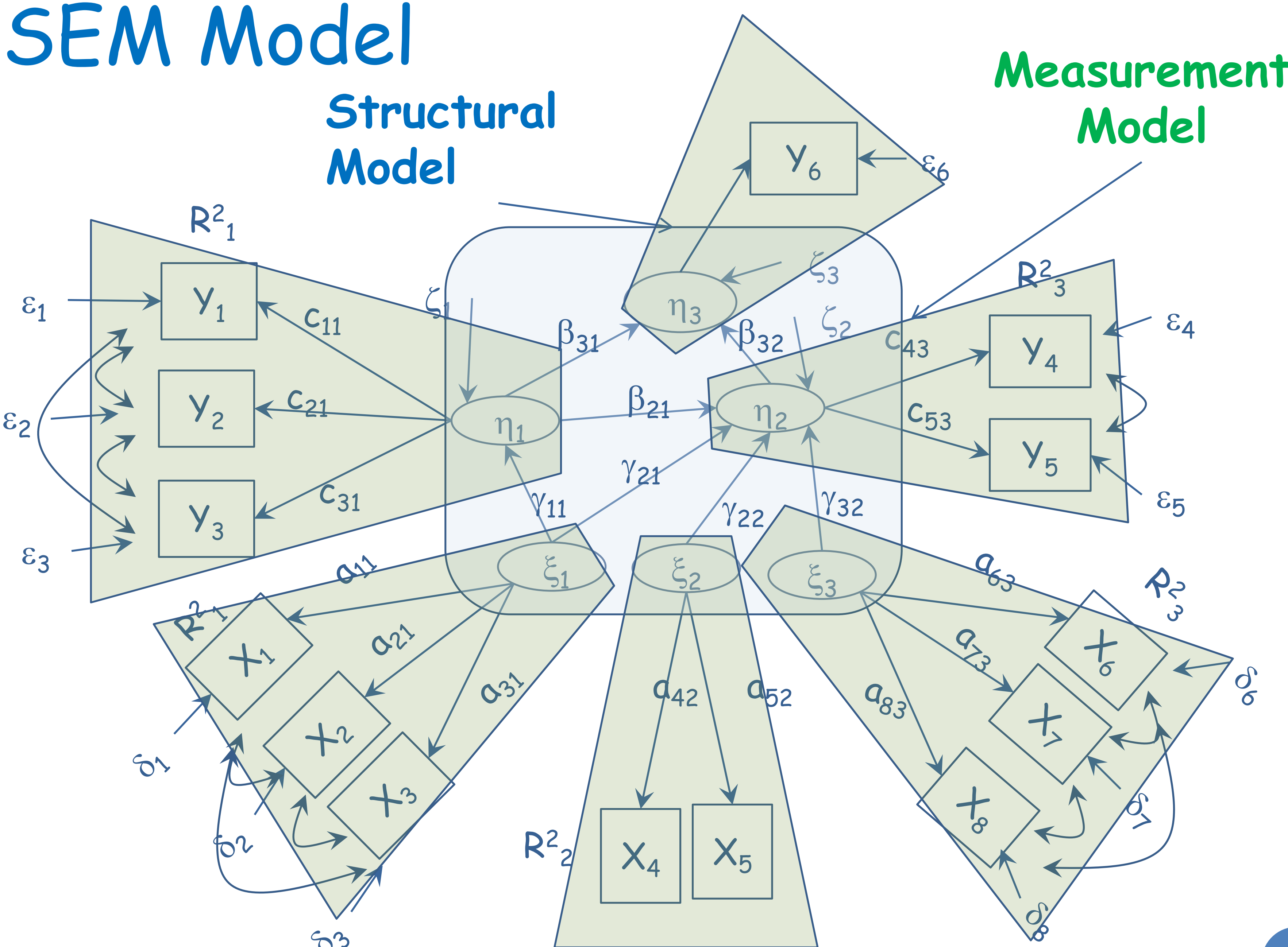
COMPOSTE INDICATOR CONSTRUCTION

HANDBOOK 2008

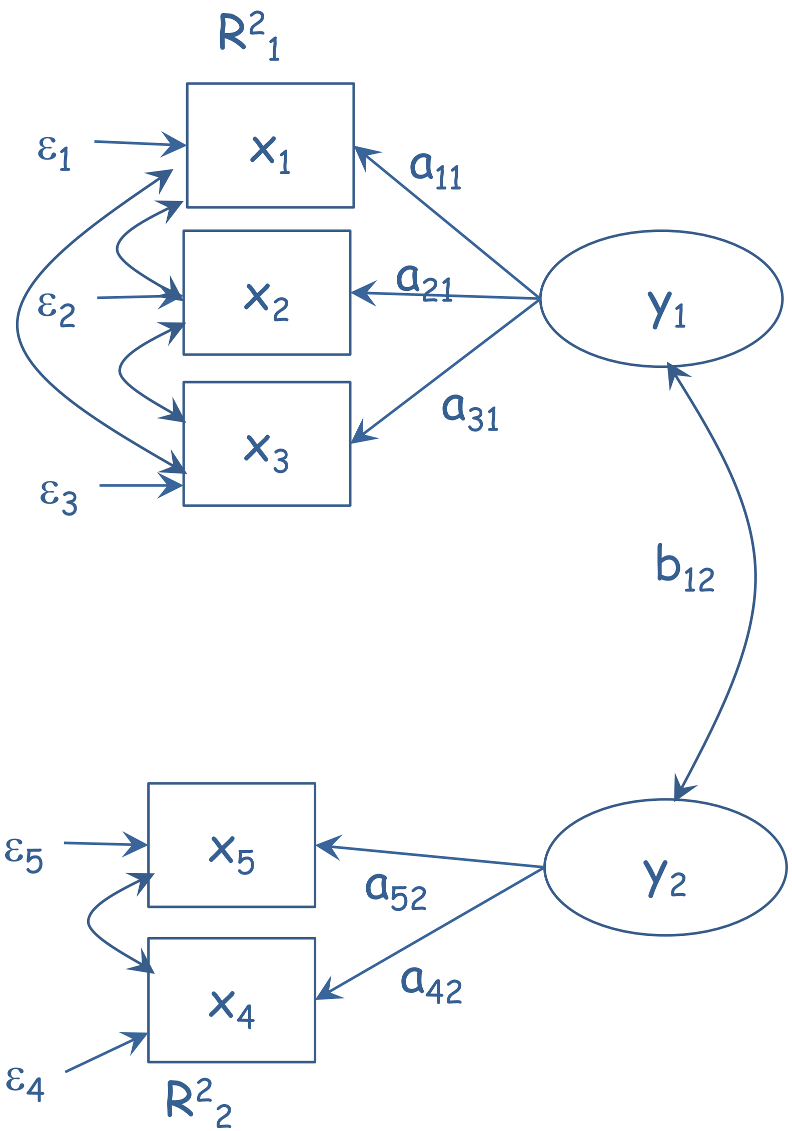
STEPS



SEM Model


Structural Model
Measurement Model
R^2_1
R^2_3
R^2_2
y_1
y_2
y_3
y_4
y_5
y_6
ε_1
ε_2
ε_3
ε_4
ε_5
ε_6
c_{11}
c_{21}
c_{31}
c_{43}
c_{53}
ζ_1
ζ_2
ζ_3
η_1
η_2
η_3
β_{31}
β_{32}
β_{21}
γ_{11}
γ_{21}
γ_{22}
γ_{32}
ξ_1
ξ_2
ξ_3
a_{11}
a_{21}
a_{31}
a_{42}
a_{52}
a_{63}
a_{73}
a_{83}
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
δ_1
δ_2
δ_3
δ_6
δ_7
δ_8

Path diagram Model of Confirmatory Factor Analysis



$$x_1 = a_{11}y_1 + \varepsilon_1$$

$$x_2 = a_{21}y_1 + \varepsilon_2$$

$$x_3 = a_{31}y_1 + \varepsilon_3$$

$$x_4 = a_{42}y_2 + \varepsilon_4$$

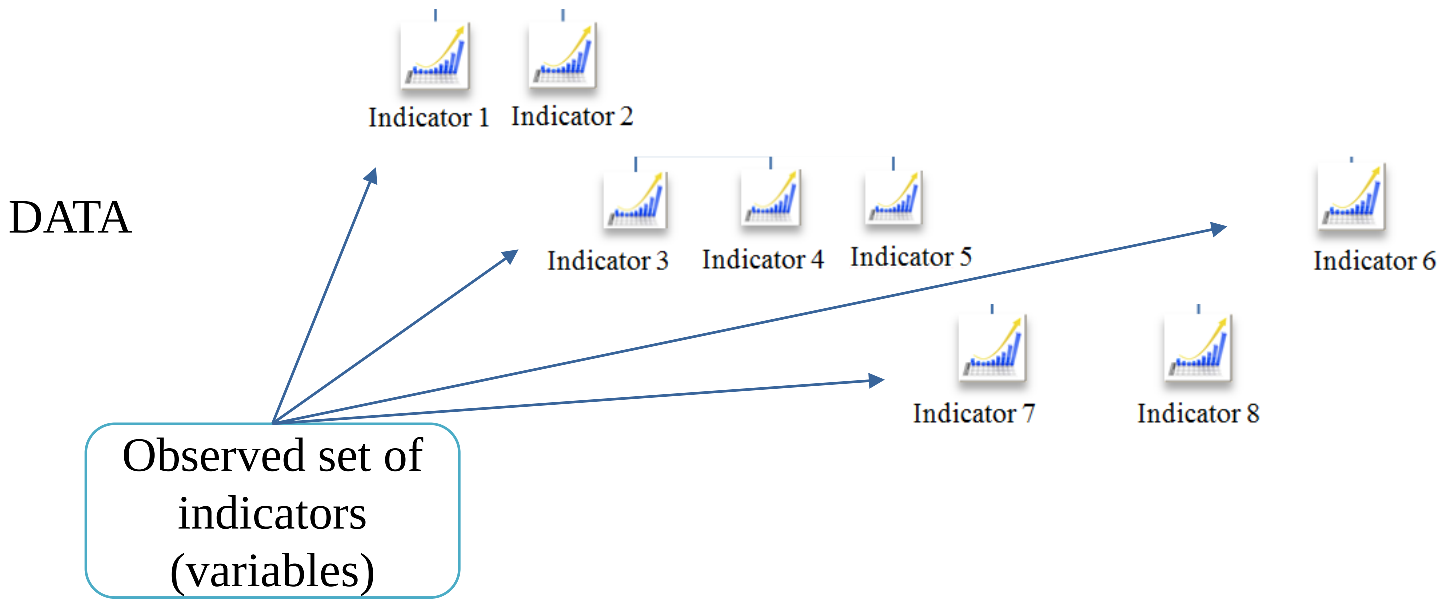
$$x_5 = a_{52}y_2 + \varepsilon_5$$

$$[x_1, x_2, x_3, x_4, x_5] = [y_1, y_2]A' + [\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5]$$

$$\mathbf{X} = \mathbf{Y}A' + \mathbf{E}$$

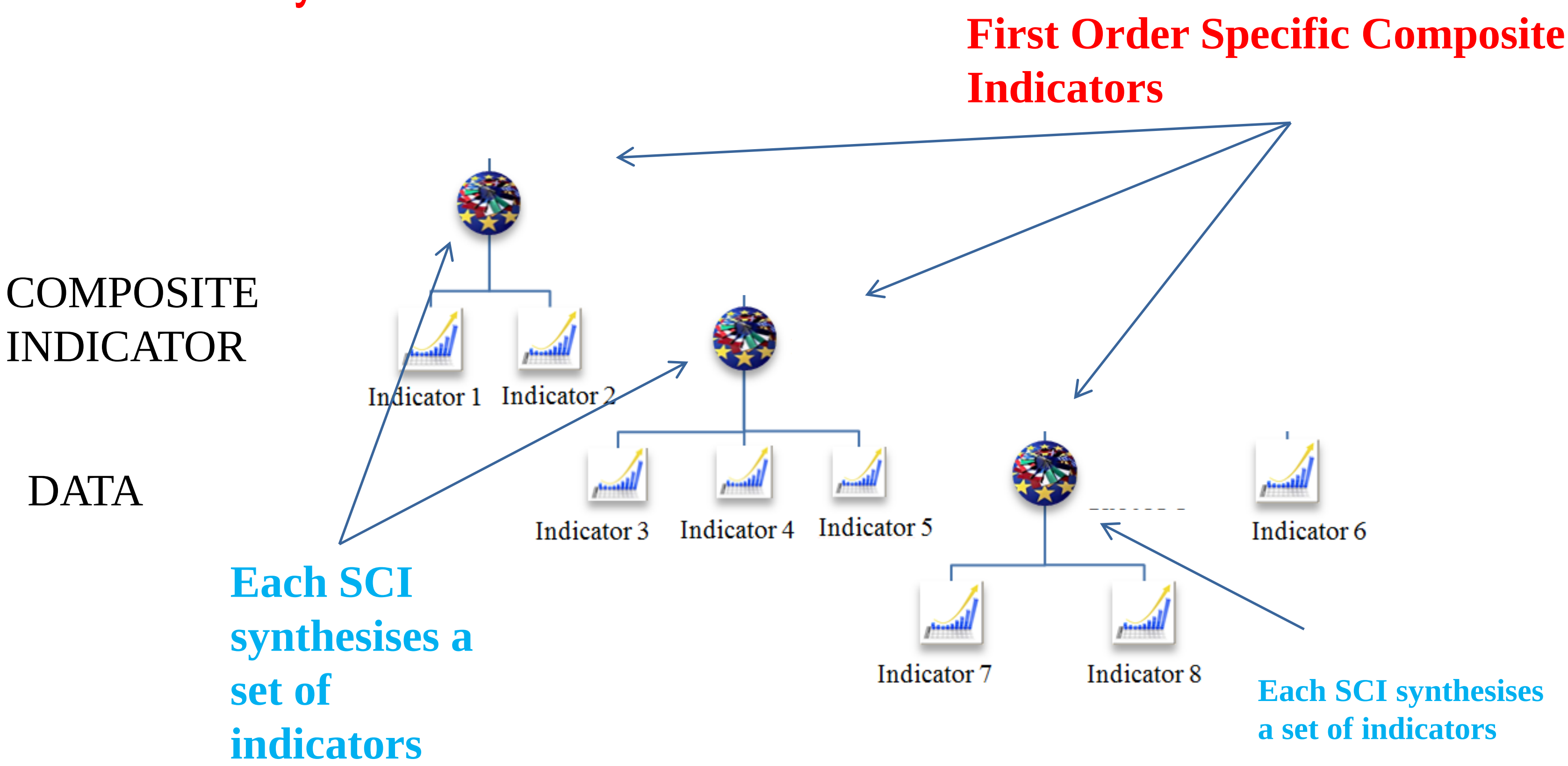
Hierarchical Model for Composite Indicators

Observed Data (Manifest Indicators) on the multidimensional concept (1/4)



Specific Composite Indicators for detecting "specific concepts" (2/4)

First Level Synthesis



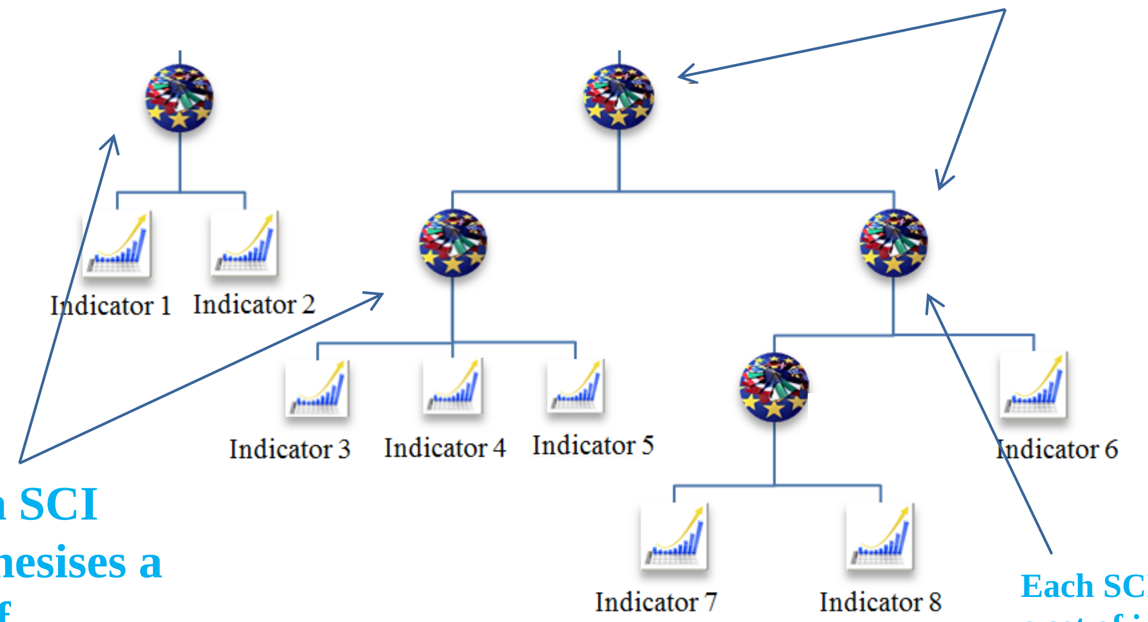
Specific Composite Indicators for detecting the “main concepts” (3/4)

Second Level Synthesis

Second Order Specific Composite Indicators

COMPOSITE INDICATOR

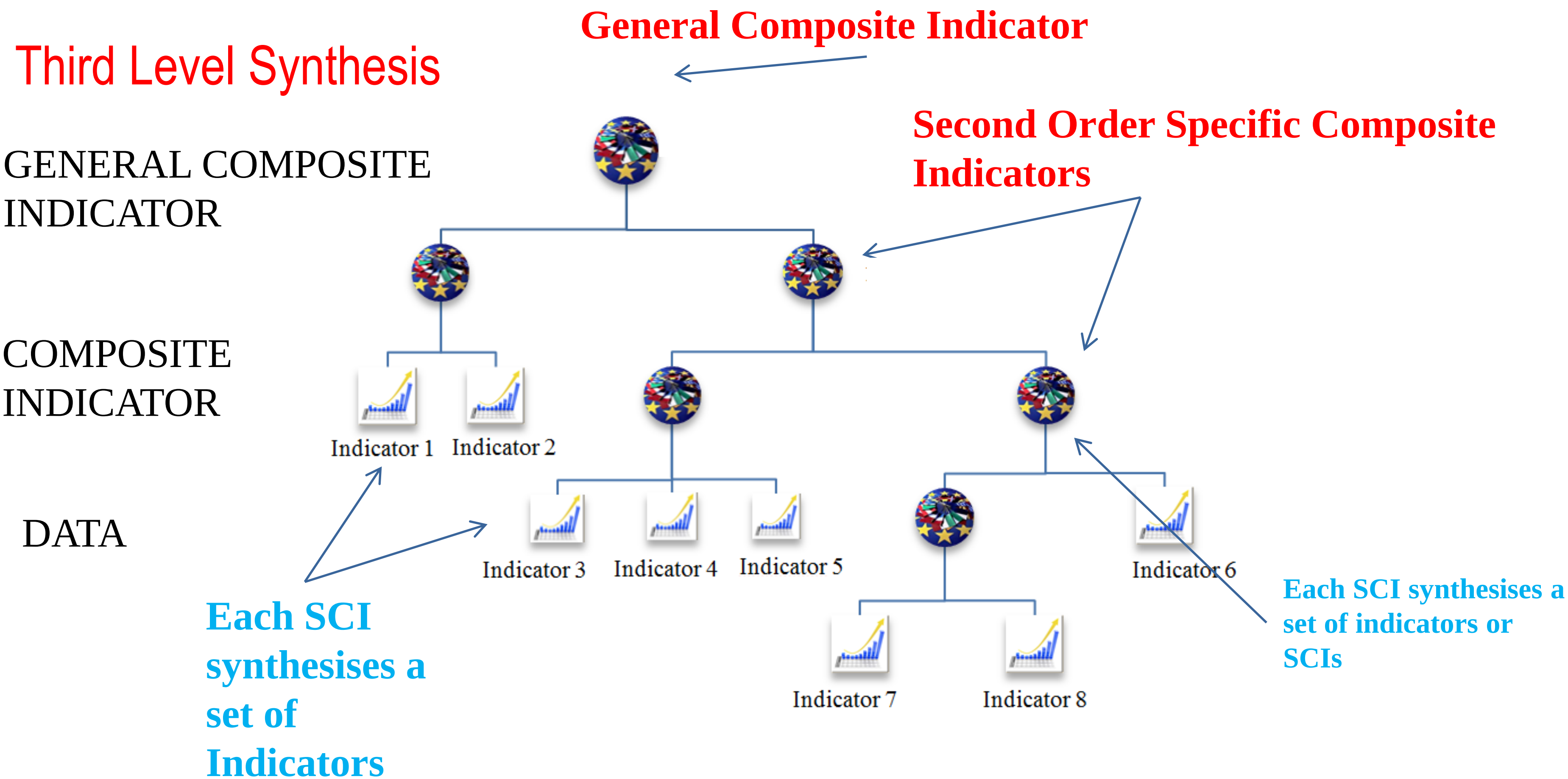
DATA



Each SCI synthesises a set of indicators

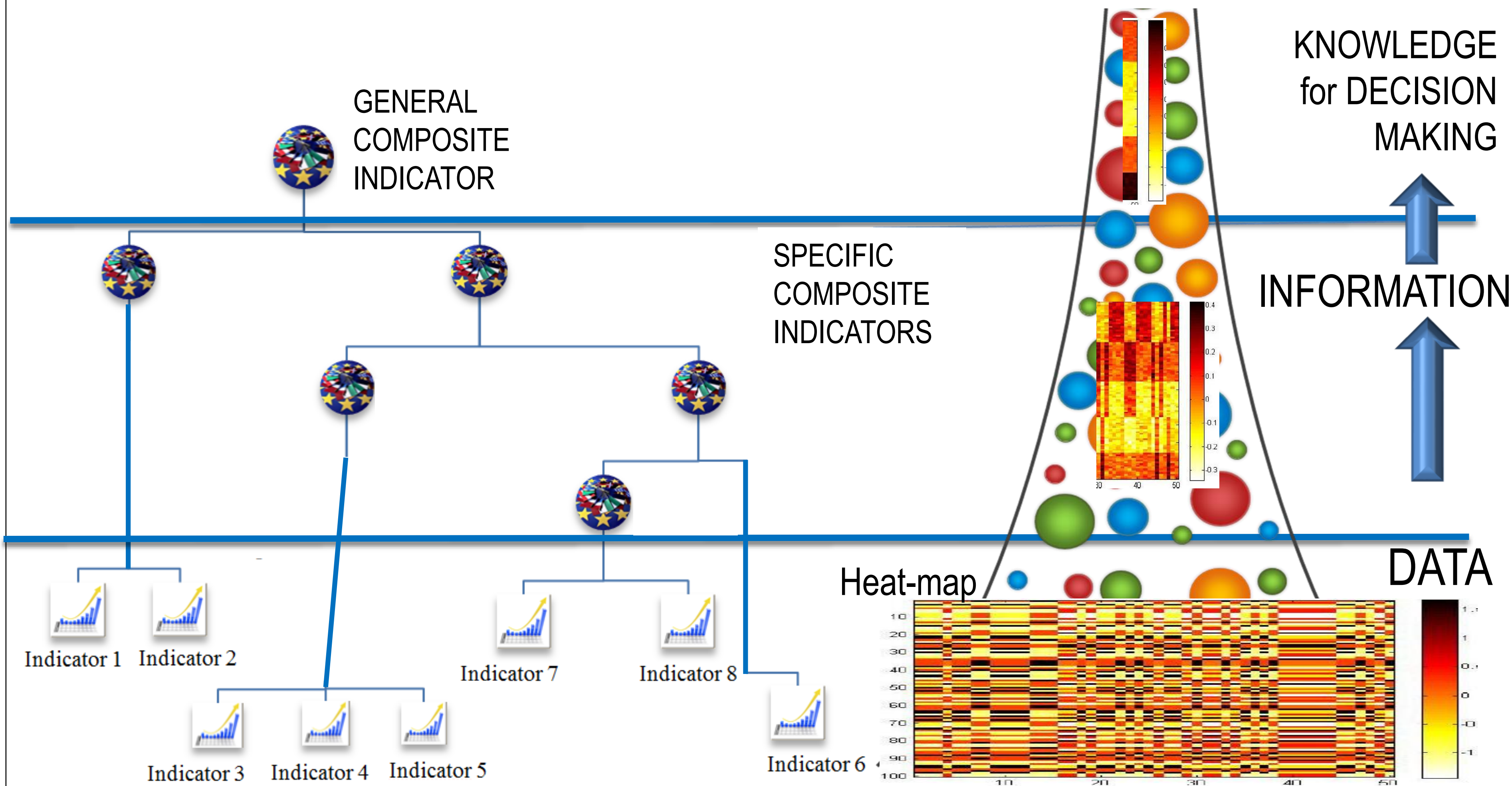
Each SCI synthesises a set of indicators or factors

General Composite Indicator (for decision making) (4/4)



Hierarchical CI model to move from DATA to KNWOLEDGE

From DATA to KNOWLEDGE via Data Dimensionality Reduction Composite Indicators



Statistical Model: Hierarchical CI

Model-based CI &
its statistical estimation (i.e., non-normative):

Data = Hierarchical CI model + error

Manifest Indicators

Measurement error + residual

Advantages
Statistical estimation (LS, MLE, …)
Validation: Goodness of Fit (to confirm the model)
Inference on the weights, GoF, …

Which typology of constructive approach:

- **Confirmatory** – a Scientific Theory (ST) is assumed and has to be confirmed by the observed indicators;
- **Exploratory** – no clear ST is known, thus, regularities are searched in the data;
- **Mixed Confirmatory & Exploratory** – part of the ST is known, but it is not completely known

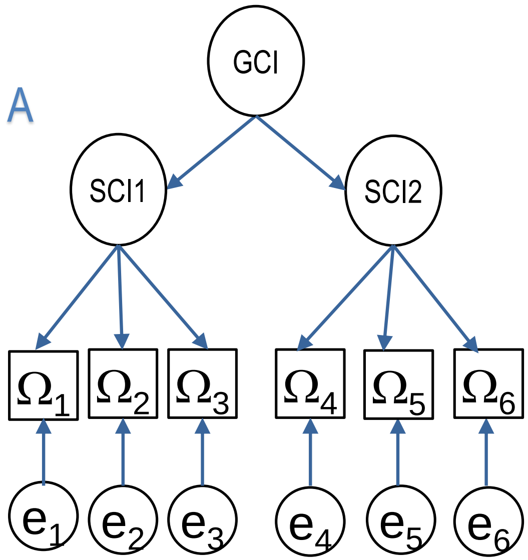
Which typology of relations between indicators:

- **Reflective**
- **Formative**

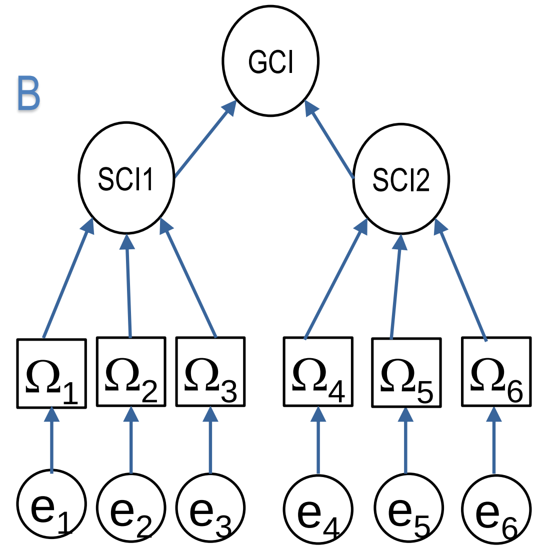
Relations between Composite Indicators (GCI & SCIs) and Manifest Indicators

A) Reflective

B) Formative



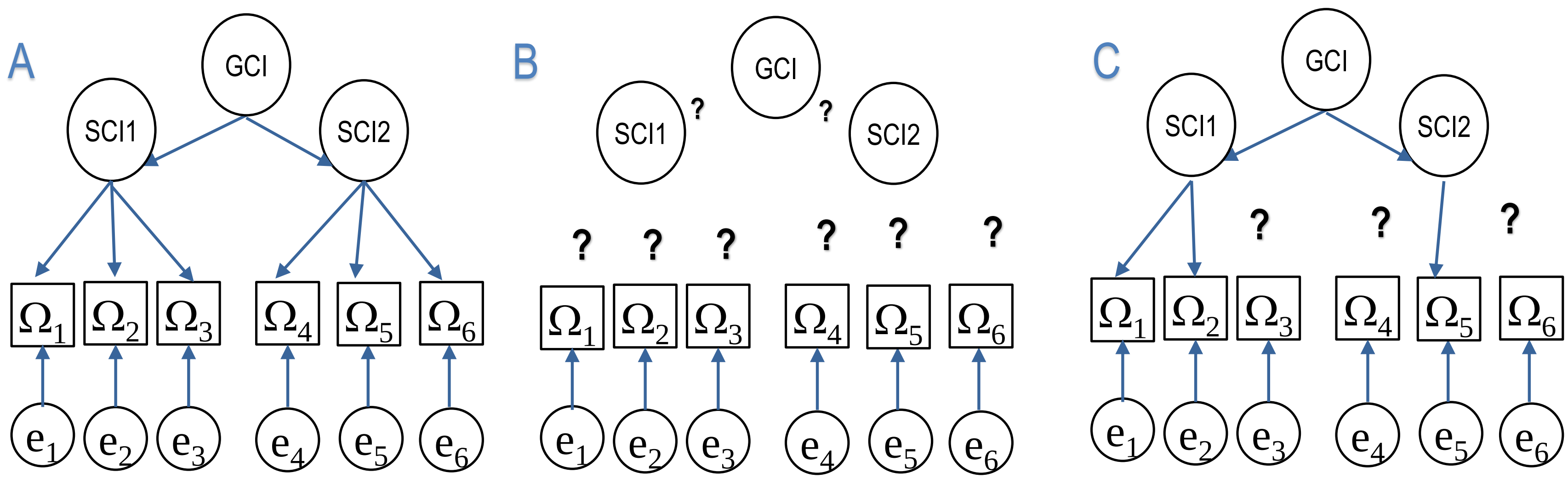
The General Composite Indicator is a determinant (causes) the Specific Composite Indicators & these last are determinant (causes) of the Manifest Indicators, i.e.,
The GCI reconstructs the SCIs that reconstruct the MI



Independent Manifest Indicators are determinant (cause, explain) of independent Specific Composite indicators that are determinant of the General Composite Indicator)

Confirmatory, Exploratory, Mixed-Confirmatory/Exploratory

- **Confirmatory** model: if a theory on the model of the CI is available, i.e., **all relationships** between manifest variables and latent variables **are** and a **priori known**;
- **Exploratory** model: **all relationships** between manifest variables and latent variables **are not a priori known**;
- **Mixed-confirmatory/exploratory** : **some relationships** are **known** according to a theory and **some** are **unknown** and must be achieved by exploratory analysis.



The Special Case of two level Hierarchical Composite Indicator

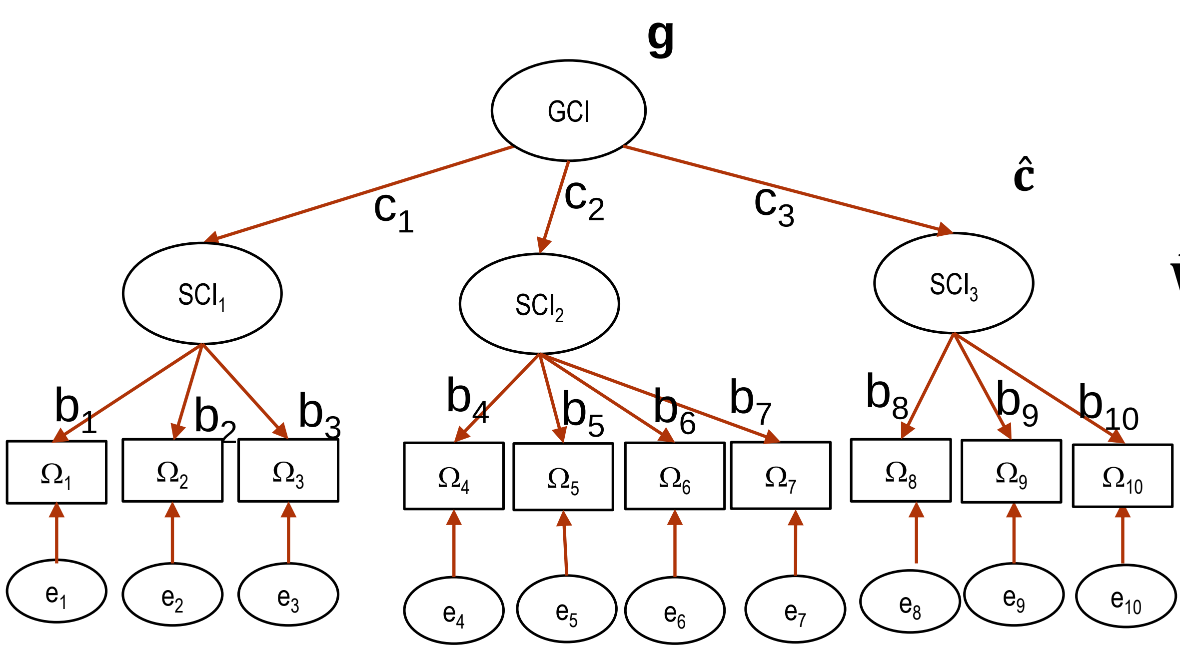
Hierarchical Composite Indicator (HCI)

a model to identify the latent Hierarchical Composite Indicator and the set of specific Composite Indicators that best reconstruct the observed data

SOME METODOLOGICAL CONSIDERATIONS

Data ERROR

$$\mathbf{X} = \overbrace{\mathbf{g}\mathbf{c}'\,\mathbf{V}'\,\mathbf{B}}^{\text{MODEL}} + \mathbf{E}$$



$$\hat{\mathbf{V}}' = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Weights for variables

$$\widehat{\mathbf{B}} = diag(\mathbf{b})$$

$$b_1, b_2, \ldots, b_{10}$$

Two-Leval Hierarchical Disjoint Factor Analysis

$$\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} = \mathbf{A}\mathbf{y} + \mathbf{e}_{\mathbf{x}}, \quad (\mathbf{y} \text{ Specific factors}) \tag{1}$$

$$\mathbf{y} = \mathbf{c}g + \mathbf{e}_{\mathbf{y}}, \quad (g \text{ General factor}) \tag{2}$$

Let include model (2) into model (1) the loading matrix **A** is restricted to the product **A**=**BV**, thus the **2-HDFA model** is defined

$$\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} = \mathbf{BV}(\mathbf{c}g + \mathbf{e}_{\mathbf{y}}) + \mathbf{e}_{\mathbf{x}} = \mathbf{BVc}g + \mathbf{BVe}_{\mathbf{y}} + \mathbf{e}_{\mathbf{x}}. \tag{3}$$

Let rewrite the model in matrix form

$$\mathbf{X} = \mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B} + \mathbf{E}_{\mathbf{x}}. \tag{4}$$

with

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{BVc}\frac{1}{n}(\mathbf{g}'\mathbf{g})\mathbf{c}'\mathbf{V}'\mathbf{B} + \boldsymbol{\Psi}_{\mathbf{x}}, \tag{5}$$

where $\boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{c}\frac{1}{n}(\mathbf{g}'\mathbf{g})\mathbf{c}' + \boldsymbol{\Psi}_{\mathbf{y}}.$ (6)

such that

$\mathbf{V} = [v_{jh} : \forall v_{jh} \in \{0,1\}]$	(binary)	(7)
$\mathbf{V}\mathbf{1}_H = \mathbf{1}_J$	(row stochastic)	(8)
$\mathbf{B} = diag(b_1, \ldots, b_J)$ with $b_j^2 > 0$	(diagonal, non-null)	(9)
$\mathbf{V}'\mathbf{BBV} = diag(b_{.1}^2, \ldots, b_{.H}^2)$, with $b_{.h}^2 = \sum_{j=1}^{J} b_{jh}^2 > 0$	(orthogonal, non-empty)	(10)

Estimation of 2-HDFA

Minimization of the **discrepancy functions** w.r.t. **B, V, U, $\bar{\mathbf{Y}}$** and **$\boldsymbol{\Psi}$**

Least-Squares Estimation

$$LSE(\mathbf{B},\mathbf{V},\boldsymbol{\Psi}) = \|\mathbf{S} - \mathbf{B}\mathbf{V}\tfrac{1}{n}(\mathbf{g}'\mathbf{g}))\mathbf{V}'\mathbf{B} - \boldsymbol{\Psi}_{\mathbf{x}}\|^2 \rightarrow \min_{\mathbf{B},\mathbf{V},\boldsymbol{\Psi},\mathbf{U},\bar{\mathbf{Y}}} \qquad 11)$$

Maximum likelihood Estimation

$$MLE(\mathbf{B},\mathbf{V},\boldsymbol{\Psi}) = ln\left|\mathbf{B}\mathbf{V}\tfrac{1}{n}(\mathbf{g}'\mathbf{g})\mathbf{V}'\mathbf{B}+\boldsymbol{\Psi}\right| - ln|\mathbf{S}| + tr\left(\left(\mathbf{B}\mathbf{V}\tfrac{1}{n}(\mathbf{g}'\mathbf{g})\mathbf{V}'\mathbf{B}+\boldsymbol{\Psi}\right)^{-1}\mathbf{S}\right) - J \rightarrow \min_{\mathbf{B},\mathbf{V},\boldsymbol{\Psi},\mathbf{U},\bar{\mathbf{Y}}} \qquad (12)$$

Generalised Least-Squares Estimation

$$GLSE(\mathbf{B},\mathbf{V},\boldsymbol{\Psi}) = \|(\mathbf{S} - \mathbf{B}\mathbf{V}\tfrac{1}{n}(\mathbf{g}'\mathbf{g})\mathbf{V}'\mathbf{B} - \boldsymbol{\Psi}_{\mathbf{x}})\mathbf{S}^{-1/2}\|^2 \rightarrow \min_{\mathbf{B},\mathbf{V},\boldsymbol{\Psi},\mathbf{U},\bar{\mathbf{Y}}} \qquad (13)$$

such that

$\mathbf{V} = [\, v_{jh} : \forall\, v_{jh} \in \{0,1\}]$ (binary) (14)

$\mathbf{V}\mathbf{1}_H = \mathbf{1}_J$ (row stochastic) (15)

$\mathbf{B} = diag(b_1, \ldots, b_J)$ with $b_j^2 > 0$ (diagonal, non-null) (16)

$\mathbf{V}'\mathbf{B}\mathbf{B}\mathbf{V} = diag(b_{.1}^2, \ldots, b_{.H}^2)$, with $b_{.h}^2 = \sum_{j=1}^{J} b_{jh}^2 > 0$ (orthogonal, non-empty) (17)

A coordinated descendent algorithm has been developed this problem.

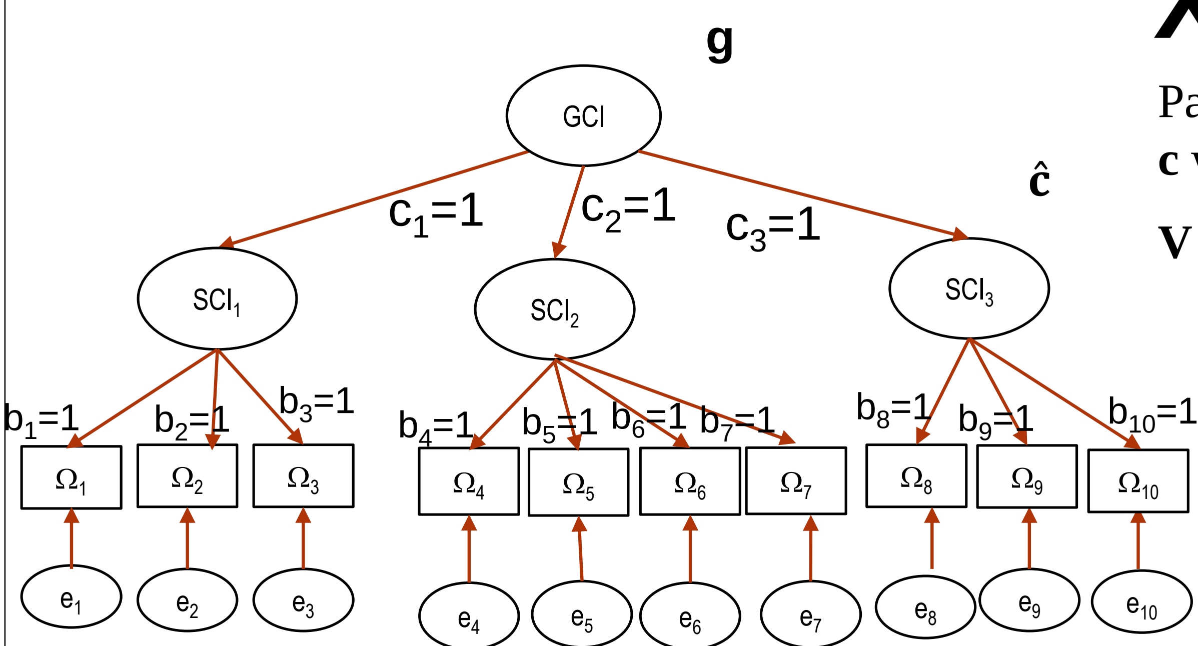
NOTE: This is a discrete and continuous problem that cannot be solved by a quasi-Newton type algorithm

Special cases of HDFA(1/2)

g=arithmetic mean of MIs if : $c_1 = c_2 = \ldots = c_Q = 1$; $b_1 = b_2 = \ldots b_J = 1$ (equal weights)

$$\hat{\mathbf{g}}_M = \mathbf{X}(\mathbf{1}'_H\hat{\mathbf{V}}')^+ = \mathbf{X}\mathbf{1}_J'^{+} = \frac{1}{J}(\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_J),$$

$$\overset{\text{Data}}{\mathbf{X}} = \overbrace{\mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B}}^{\text{MODEL}} + \overset{\text{ERROR}}{\mathbf{E}}$$



Parameters of the model

c weights for SCIs

V relations between MIs & SCIs

$$\hat{\mathbf{V}}' = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Weights for variables

$$\hat{\mathbf{B}} = diag(\mathbf{b})$$

$$b_1, b_2, \ldots, b_{10}$$

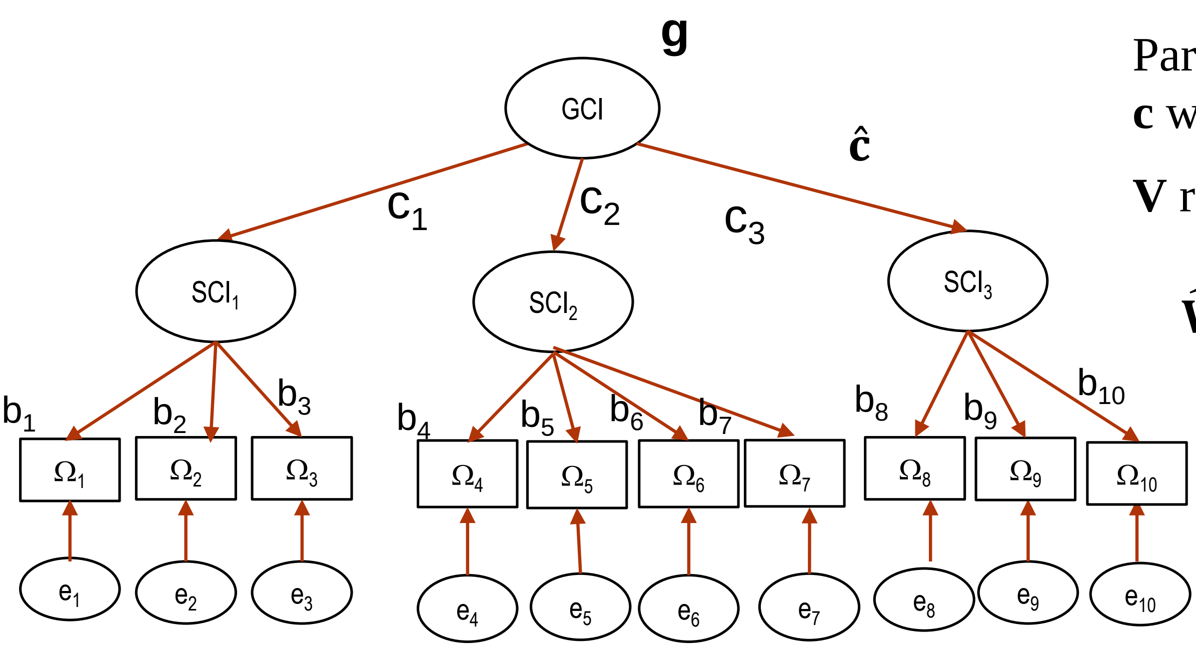
Special cases of HCI (2/2)

g = weighted arithmetic mean of MIs (i.e., different weights)

$$\hat{\mathbf{g}}_{WM} = \mathbf{X}(\hat{\mathbf{B}}\hat{\mathbf{V}}\hat{\mathbf{c}})(\hat{\mathbf{c}}'\hat{\mathbf{V}}'\hat{\mathbf{B}}\mathbf{1}_J)^{-1}$$

Data MODEL ERROR

$$\mathbf{X}=\mathbf{g}\overbrace{\mathbf{c}'\,\mathbf{V}'\,\mathbf{B}}^{\text{MODEL}}+\mathbf{E}$$



Parameters of the model

c weights for SCIs

V relations between MIs & SCIs

$$\hat{\mathbf{V}}'=\begin{pmatrix}1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1\end{pmatrix}$$

Weights for MIs

$$\hat{\mathbf{B}} = diag(\mathbf{b})$$

$$b_1, b_2, \ldots, b_{10}$$

MODEL ASSESSMENT

The ***goodness of fit*** of the CI model:

$$R^2_{\mathrm{GCI}} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{tr(\mathbf{X}'\mathbf{X}) \; - \; tr\left(\left(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}}\right)\hat{\mathbf{g}}'\hat{\mathbf{g}}(\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})\right)}{tr(\mathbf{X}'\mathbf{X})}$$

$$R^2_{\mathrm{SCI}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$= 1 - \frac{tr(\mathbf{X}'\mathbf{X}) \; - \; tr(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}})}{tr(\mathbf{X}'\mathbf{X})}$$

$$R^2_{\mathrm{SCI}_h} = 1 - \frac{SS_{res_{Yh}}}{SS_{tot_h}} = 1 - \frac{tr(\mathbf{X}_h{}'\mathbf{X}_h) \; - \; tr(\widehat{\mathbf{B}}_h\hat{\mathbf{v}}_h\hat{\mathbf{y}}_h{}'\hat{\mathbf{y}}_h\hat{\mathbf{v}}_{\boldsymbol{h}}'\widehat{\mathbf{B}}_h)}{tr(\mathbf{X}_h{}'\mathbf{X}_h)}$$

The ***Information criteria***

AIC

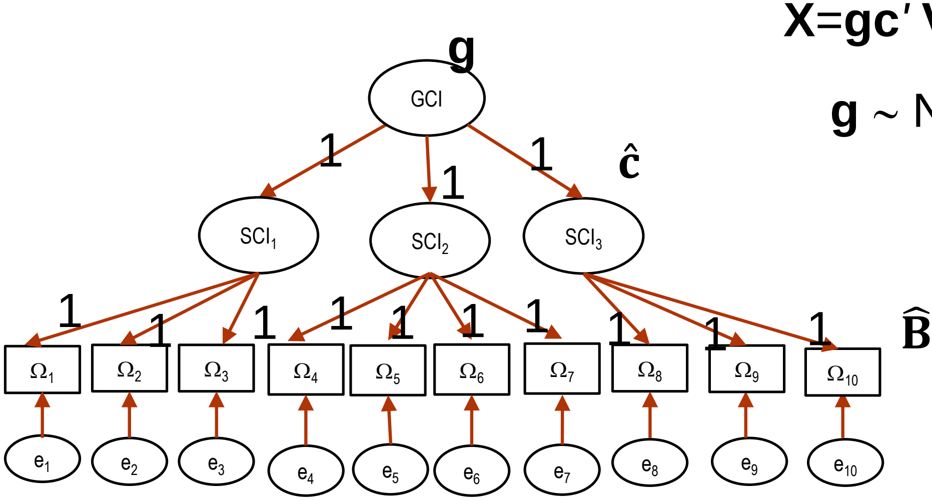
-2log $Ol(\theta, \pi) + 2d$

BIC

-2log $Ol(\theta, \pi) + d \log n$

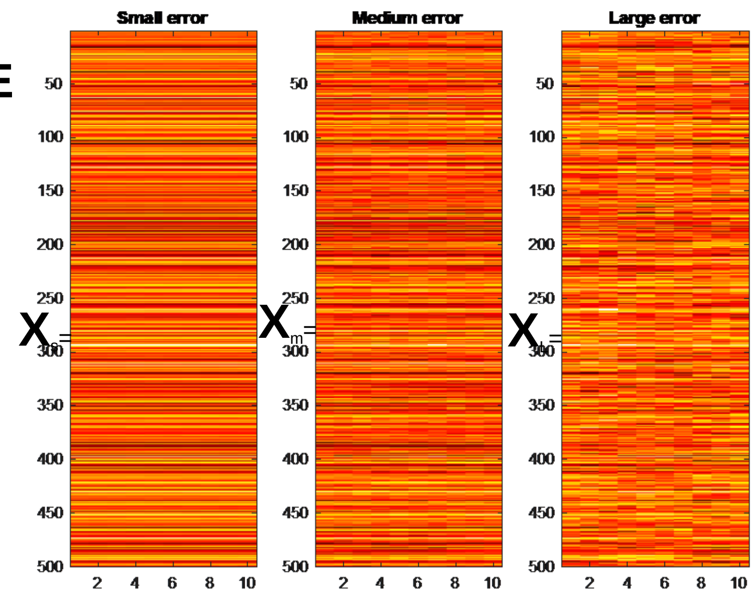
Example 1 : Assessment of the Model-Based CI

Case of ARITHMETIC MEAN



$$\mathbf{X}=\mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B}+\mathbf{E}$$

$$\mathbf{g} \sim N(0,1)$$



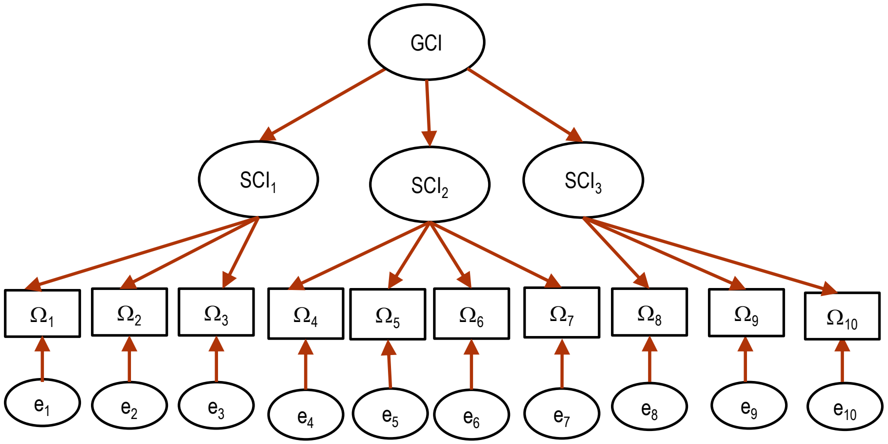
if $\hat{\mathbf{c}}=\mathbf{1}_3$ and $\widehat{\mathbf{B}}=\mathbf{I}_{10}$,

$$\widehat{\mathbf{V}}'=\begin{pmatrix}1&1&1&0&0&0&0&0&0&0\\0&0&0&1&1&1&1&0&0&0\\0&0&0&0&0&0&0&1&1&1\end{pmatrix}$$

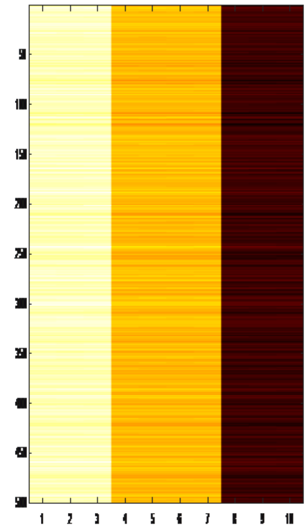
Error:	R^2_{GCI}	$R^2_{SCI_1}$	$R^2_{SCI_2}$	$R^2_{SCI_3}$
$\mathbf{X}_s$ Small	**0.974**	0.988	0.988	0.989
$\mathbf{X}_m$ Medium	**0.622**	0.778	0.837	0.855
$\mathbf{X}_L$ Large	**0.131**	0.624	0.539	0.672

Arithmetic mean is a good GCI only when the MIs are very similar

Example 2 : Assessment of the Model-Based CI
Case of ARITHMETIC MEAN



X=



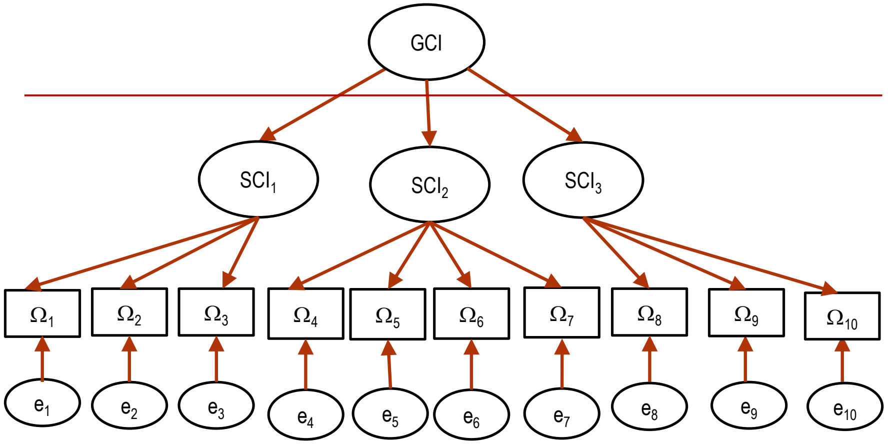
X divided in three blocks

$$R^2_{\mathrm{GCI}} = \frac{SS_{mod}}{SS_{tot}} = \mathbf{0.548}$$

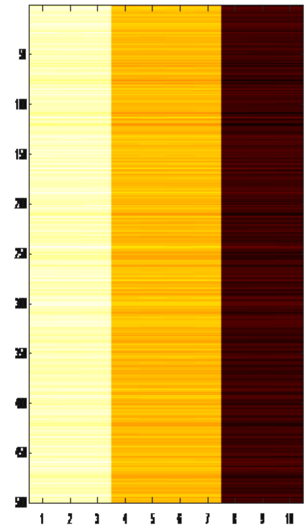
$R^2_{\mathrm{SCI}_1}$	$R^2_{\mathrm{SCI}_2}$	$R^2_{\mathrm{SCI}_3}$
0.999	0.999	1

In a situation like this is better to stop at an intermediate level of synthesis (i.e., SCIs level)
because a GCI built as the arithmetic mean of MIs is not a good representation of
the phenomenon to describe

Example 2 : Assessment of the Model-Based CI
Case of ARITHMETIC MEAN



X=



$$R^2_{\mathrm{GCI}} = \frac{SS_{mod}}{SS_{tot}} = \mathbf{0.548}$$

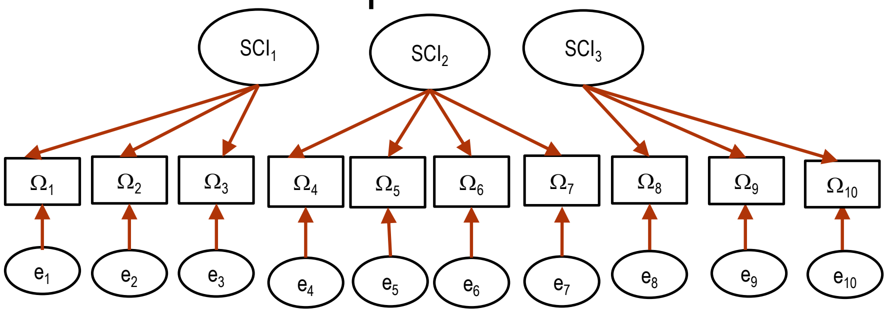
$R^2_{\mathrm{SCI}_1}$	$R^2_{\mathrm{SCI}_2}$	$R^2_{\mathrm{SCI}_3}$
0.999	0.999	1

In a situation like this is better to stop at an intermediate level of synthesis
(i.e., SCIs level)
because a GCI built as the arithmetic mean of MIs is not a good representation of
the phenomenon to describe

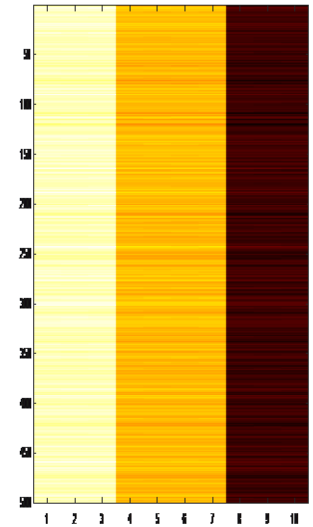
<u>Example 2</u> : Assessment of the Model-Based CI

Case of ARITHMETIC MEAN

Final correct description



X=



$$R^2_{\mathrm{GCI}} = \frac{SS_{mod}}{SS_{tot}} = \mathbf{0.548}$$

$R^2_{\mathrm{SCI}_1}$	$R^2_{\mathrm{SCI}_2}$	$R^2_{\mathrm{SCI}_3}$
0.999	0.999	1

In a situation like this is better to stop at an intermediate level of synthesis (i.e., SCIs level)
because a GCI built as the arithmetic mean of MIs is not a good representation of
the phenomenon to describe

PROPERTIES of CI

Scale-invariance

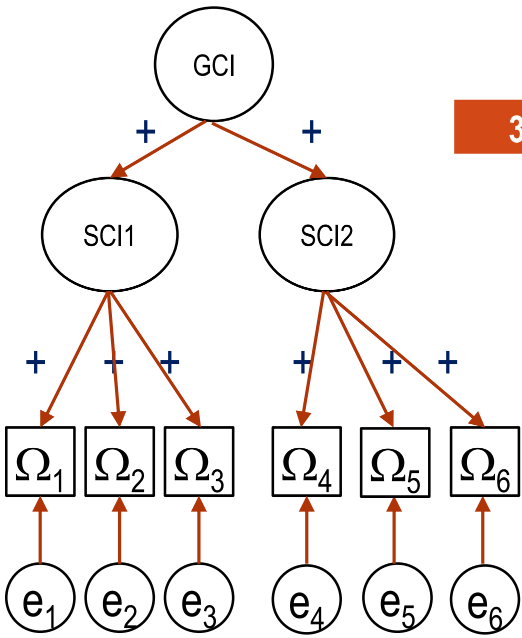
Data are normalized in order to allow the comparison and the combination of the MIs into the SCIs and GCI.

- Standardization $\mathbf{Z} = \mathbf{JX}\text{diag}(\text{dg}(\mathbf{\Sigma_X}))^{-1/2}$ with $\mathbf{J}=\mathbf{I}_n-(1/n)\,\mathbf{1}_n\mathbf{1}'_n$
- Min-max normalization $\mathbf{Z} = \mathbf{X} - \mathbf{1}_n\min\mathbf{X}./(\mathbf{1}_n\max\mathbf{X}-\mathbf{1}_n\min\mathbf{X})$
- Normalized dispersion $\mathbf{Z} = \mathbf{JX}\text{diag}(\boldsymbol{\mu}_\mathbf{X})^{-1}$ with $\mathbf{J}=\mathbf{I}_n-(1/n)\,\mathbf{1}_n\mathbf{1}'_n$

A scale-invariant CI is a latent Indicator that is not sensitive to linear transformations such as normalization methods.

Non-Compensability & Non-Negativity.

The CI satisfies the non-compensability property if its relationships with latent and/or MIs are all positives. Thus, the effect of the SCIs and/or MIs do not compensate each other.



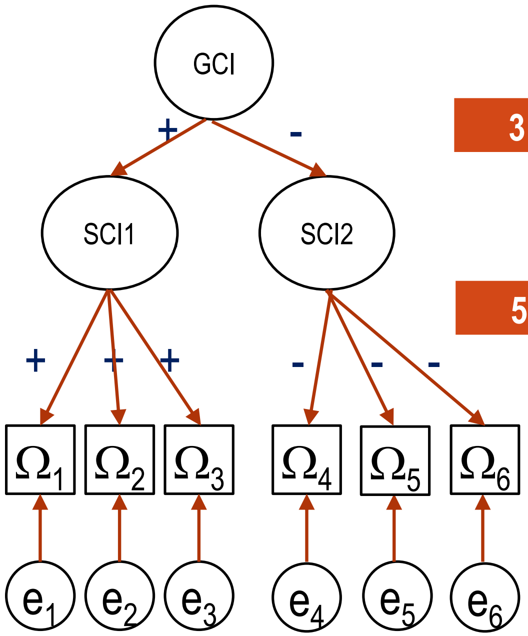
Ranking of the not compensated model:

3	10	1	7	5	9	4	2	8	6

So **non-negativity** and non-compensability are strictly connected.

Non-Compensability & Non-Negativity.

The CI satisfies the non-compensability property if its relationships with latent and/or MIs are all positives. Thus, the effect of the SCIs and/or MIs do not compensate each other.



Ranking of the not compensated model:

3	10	1	7	5	9	4	2	8	6

Ranking of the compensated model:

5	9	3	7	1	6	8	4	2	10

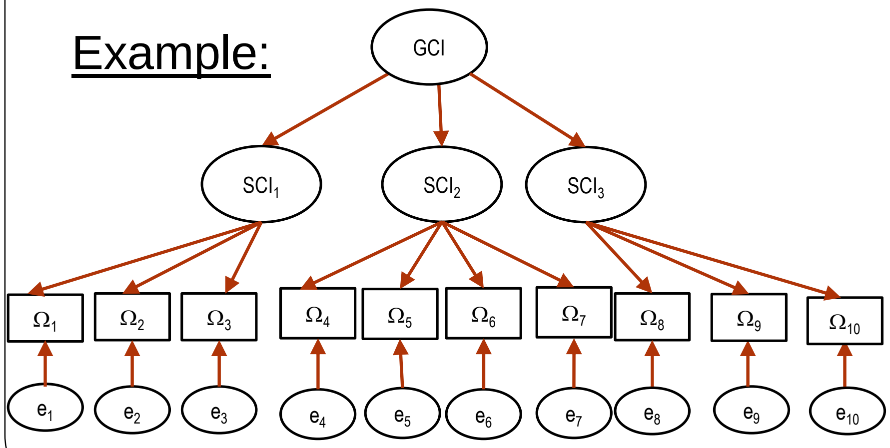
So **non-negativity** and non-compensability are strictly connected.

Reliability, Unidimensionality & General Factor

Reliability of a CI is the global consistency of MIs based on the correlations between different MIs on the same CI.
It is frequently called internal consistency and it is usually measured with Cronbach's alpha (Cronbach, 1951)

Unidimensionality evaluates to which extend a single latent indicator, generally a SCI, has been measured with a set of MIs.
Unidimensionality is more realistic for SCIs, while Revelle and Zinbarg, (2009) hypothesize that there is a **general factor**, i.e., a GCI that can be tested by nested confirmatory SCIs.
A measure of unidimensionality for each SCI might be the variance of the second component of the set of MIs explained by the related SCI.

Example:



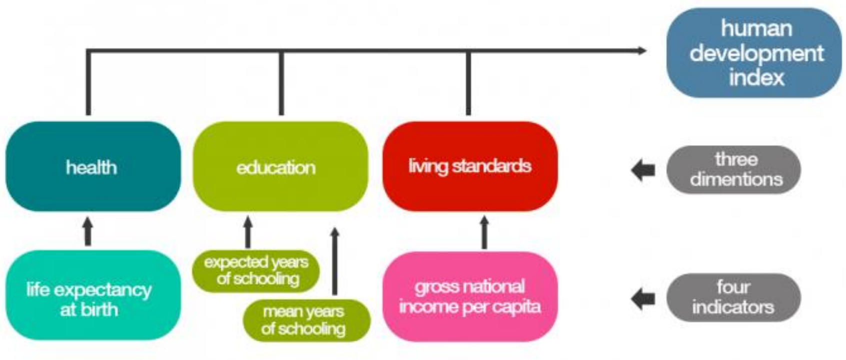
	Factor 1	Factor 2
Unidimensionality	2.737	0.556
Reliability	0.526	0.476

	Factor 1	Factor 2	Factor 3
Unidimensionality	0.400	0.556	0.618
Reliability	0.781	0.794	0.781

APPLICATIONS

Human Development Index - HDI

The HDI is the geometric mean of three normalized indices:
Life Expectancy Index (LEI), Education Index (EI) and Income Index (II)
we can measure the goodness of fit of the HDI by considering that the logarithm of the geometric mean is equal to the arithmetic mean of the logarithm of MIs. Each dimension is represented by a specific index(normalized with a own method):



<u>Based on the above informations:</u>

- **Life Expectancy Index (LEI)** = Actual LE – 20/(85-20)
- **Income Index (II)** = {ln(GNI pc)- ln(100)}/{ln(75,000) – ln(100)}
- **Education Index (EI) =** MYSI+EYSI / 2
- **Mean Years of Schooling Index (MYSI)** = MYS-0 / 15-0
- **Expected Years of Schooling Index (EYSI)** = EYS-0 / 18-0

Now, HDI is the geometric mean of previous three indices i.e.
HDI= $\sqrt[3]{LEI * EI * II}$

Let us consider: $\widehat{\mathbf{B}} = \widehat{\mathbf{V}} = \mathbf{I}_3$
$\hat{\mathbf{c}} = \mathbf{1}_3$

$$R^2_{\mathrm{HDI}} = \frac{SS_{mod}}{SS_{tot}} = \frac{tr\left(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}}\, log(\hat{\mathbf{g}}_{\mathbf{HDI}})'log(\hat{\mathbf{g}}_{\mathbf{HDI}})\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\right)}{tr((log(\mathbf{X}))'(log(\mathbf{X})))} = \mathbf{0.901}$$

where $log(\mathbf{X})$ is a matrix where each column is the logarithmic transformation of the respective column of $\mathbf{X}$.

Thus, everything is perfect?
HOWEVER … we have different and specific normalisations of the three indices

It's important to see how the three indices are normalized and how these transformations have a role on the goodness of HDI.

- Life Expectancy Index (LEI) is normalized according to the formula: $Z = (X - 20)/65$, where X is "life expectancy at birth".
- Education Index (EI) is the composition (i.e. the arithmetic mean) of two variables: Expected years of schooling (X_1) and Mean years of schooling (X_2), where the first one is normalized by the formula:
- $Z_1 = min(X_1, 18)/18$ and the second one according to the formula: $Z_2 = X_2/15$.
- Thus, the Education Index is calculated by: $Z = \frac{Z_1+Z_2}{2}$.
- So, Income Index (II) is normalized according to: $Z = \frac{l\,n(X)-ln(100)}{l\,n(75000)-ln(100)}$, where X is "GNI per capita".

Let us see what is the assessment of the HDI if we use a unique normalization for Min-max.

$$R^2_{\mathrm{NN\_HDI}} = \frac{SS_{mod}}{SS_{tot}} = \frac{tr\left(\hat{\mathbf{B}}\hat{\mathbf{V}}\hat{\mathbf{c}}\, log\left(\hat{\mathbf{g}}_{\mathbf{MinMax\_HDI}}\right)' log\left(\hat{\mathbf{g}}_{\mathbf{MinMax\_HDI}}\right)\hat{\mathbf{c}}'\hat{\mathbf{V}}'\hat{\mathbf{B}}\right)}{tr\left((log(\mathbf{X}))'(log(\mathbf{X}))\right)} = 0.632$$

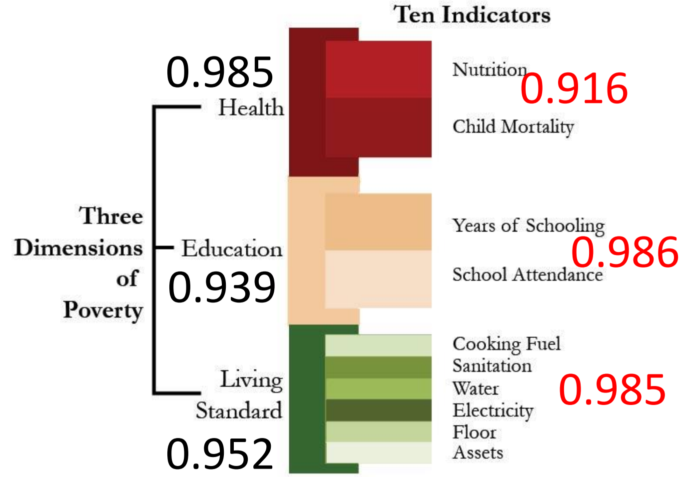
The increase of the 27% of R^2_{HDI} with respect to $R^2_{\mathrm{MinMax\_HDI}}$ has to be imputed to the use of different normalisations. Therefore, it is important to understand that different normalizations of the MIs must be strongly motivated.

Correlation	LEI	EI	II
HDI	0.90	0.95	0.94

Is useful to create another indicator that provides little more information than some traditional indicator like GNI? (McGillivray, 1991)

Multidimensional Poverty Index- MPI

The global Multidimensional Poverty Index (MPI) is an international measure of acute poverty covering over 100 developing countries developed by OPHI and the United Nations Development Programme. The index uses the same three dimensions as the Human Development Index: health, education, and standard of living. These are measured using ten indicators divided in three dimensions.



Let us consider: $\widehat{\mathbf{B}} = diag(\frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6})$

$$\widehat{\mathbf{V}}' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \qquad \widehat{\mathbf{c}}' = (\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3})$$

$$R^2_{\mathrm{MPI}} = \frac{SS_{mod}}{SS_{tot}} = \frac{tr((\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}})^{-1}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}})\hat{\mathbf{g}}_{\mathbf{MPI}}{}'\hat{\mathbf{g}}_{\mathbf{MPI}}(\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})(\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}})^{-1})}{tr(\mathbf{X}'\mathbf{X})} = \mathbf{0.515}$$

If matrices **B**, **V** and **c** are estimated

$$\widehat{\mathbf{B}} = diag(0.71\ 0.71\ 0.37\ 1\ 0.39\ 0.38\ 0.38\ 0.37\ 0.39\ 0.36)$$

$$\widehat{\mathbf{V}}' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \qquad \widehat{\mathbf{c}}' = (0.85\ \ 0.43\ 0.30)$$

$$R^2 = \frac{SS_{mod}}{SS_{tot}} = \frac{tr((\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}})^{-1}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}})\hat{\mathbf{g}}'\hat{\mathbf{g}}(\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})(\hat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\hat{\mathbf{c}})^{-1})}{tr(\mathbf{Z}'\mathbf{Z})} = \mathbf{0.884}$$

where $\mathbf{Z}$ is a matrix where each column is the standardized column of $\mathbf{X}$, respectively.

Application to Sustainable Development Goals

SDGs Europe: 100 Indicators, 17 Goals

Goal1:
1- People at risk of poverty or social exclusion 01.11
2- People at risk of poverty after social transfers 01.12
3- Severely materially deprived people 01.13
4- People living in households with very low work intensity 01.14
5- Housing cost overburden rate 01.21
6- Share of total population living in a dwelling with a leaking roof, damp walls, floors or foundation, or rot in window frames or floor 01.22

Goal2:
7- Obesity rate 02.11
8- Agricultural factor income per annual work unit (AWU) 02.21
9- Government support to agricultural research and development 02.26
10- Area under organic farming 02.31
11- Ammonia emissions from agriculture 02.52
12- Gross nutrient balance on agricultural land 02.54

Goal3:
13- Life expectancy at birth 03.11
14- Self-perceived health 03.14
15- Death rate due to chronic diseases 03.25
16- Suicide death rate 03.31
17- Smoking prevalence 03.36
18- Self-reported unmet need for medical examination and care 03.41

Goal4:
19- Early childhood education and care 04.10
20- Early leavers from education and training 04.20
21- Tertiary educational attainment 04.30
22- Employment rate of recent graduates 04.31
23- Adult participation in learning 04.40
24- Underachievement in reading, maths and science 04.50

Goal5:
25- Gender pay gap 05.10
26- Gender employment gap 05.12
27- Proportion of seats held by women in national parliaments and local government 05.20
28- Proportion of women in senior management positions 05.21
29- Physical and sexual violence by a partner or a non-partner 05.33
30- Inactivity rates due to caring responsibilities 05.44

Goal6:
31- Share of total population having neither a bath, nor a shower, nor indoor flushing toilet in their household 06.11
32- Population connected to urban wastewater treatment with at least secondary treatment 06.13
33- Biochemical oxygen demand in rivers 06.21
34- Nitrate in groundwater 06.24
35- Phosphate in rivers 06.26
36- Water exploitation index (WEI) 06.41

Goal7:
37- Percentage of people affected by fuel poverty (inability to keep home adequately warm) 07.10
38- Share of renewable energy in gross final energy consumption 07.20
39- Primary energy consumption; final energy consumption by sector 07.30
40- Final energy consumption in households per capita 07.32
41- Energy dependence 07.33
42- Energy productivity 07.35

Goal8:
43- Real GDP per capita - growth rate 08.10
44- Young people neither in employment nor in education and training 08.20
45- Total employment rate 08.30
46- Long-term unemployment rate 08.31
47- Involuntary temporary employment 08.35
48- Fatal accidents at work by sex (NACE Rev. 2, A, C-N) - Unstandardised incidence rate 08.60

Goal9:
49- Gross domestic expenditure on R&D 09.10
50- Employment in high- and medium-high technology manufacturing sectors and knowledge intensive service sectors 09.11
51- Total R&D personnel 09.13
52- Patent applications to the European Patent Office (EPO) 09.14
53- Share of collective transport modes in total passenger land transport 09.40
54- Share of rail and inland waterways activity in total freight transport 09.41

Goal11:
61- Overcrowding rate by degree of urbanisation 11.12
62- Distribution of population by level of difficulty in accessing public transport 11.21
63- People killed in road accidents 11.25
64- Urban population exposure to air pollution by particulate matter 11.31
65- Proportion of population living in households considering that they suffer from noise 11.3
66- Recycling rate of municipal waste 11.52

Goal13:
73- Greenhouse gas emissions (indexed totals and per capita) 13.11
74- Greenhouse gas emissions intensity of energy consumption 13.14
75- Global (and European) near surface average temperature 13.21
76- Economic losses caused by climate extremes (consider climatological, hydrological, meteorological) 13.45
77- Contribution to the 100bn international commitment on climate related expending (public finance) 13.51
78- Share of EU population covered by the new Covenant of Mayors for Climate and Energy (integrating mitigation, adaptation, and access to clean and affordable energy) 13.63

Goal15:
84- Forest area as a proportion of total land area 15.11
85- Artificial land cover per capita 15.11
86- Change in artificial land cover per year 15.24
87- Common bird index 15.31
88- Sufficiency of terrestrial sites designated under the EU habitats directive 15.32
89- Estimated soil erosion by water 15.41

Goal17:
96- Official development assistance as share of gross national income 17.10
97- EU financing for developing countries 17.11
98- EU Imports from developing countries 17.12
99- General government gross debt 17.13
100- Shares of environmental and labour taxes in total tax revenues 17.19

Goal10:
55- GDP per capita in PPS 10.10
56- Real adjusted gross disposable income of households per capita in PPS 10.11
57- Relative median at-risk-of-poverty gap 10.22
58- Gini coefficient of equivalised disposable income 10.24
59- Income growth of the bottom 40 per cent of the population and the total population 10.25
60- Number of first time asylum applications (total and accepted) per capita 10.31

Goal12:
67- Generation of waste excluding major mineral wastes 12.10
68- Recycling and landfill rate of waste excluding major mineral wastes 12.11
69- Consumption of toxic chemicals 12.30
70- Resource productivity 12.40
71- Average CO2 emissions per km from new passenger cars 12.51
72- Volume of freight transport relative to GDP 12.54

Goal14:
79- Bathing water quality 14.13
80- Sufficiency of marine sites designated under the EU habitats directive 14.21
81- Ocean acidification (CLIM 043) 14.31
82- Catches in major fishing areas 14.41
83- Assessed fish stocks exceeding fishing mortality at maximum sustainable yield (Fmsy) 14.43

Goal16:
90- Death due to homicide, assault, by sex 16.10 (tps00146)
91- Share of population which reported occurrence of crime, violence or vandalism in their area 16.19
92- General government total expenditure on law courts 16.32
93- Corruption Perception Index 16.50
94- Perceived independece of the justice system 16.61
95- Level of citizens' confidence in EU institutions 16.62

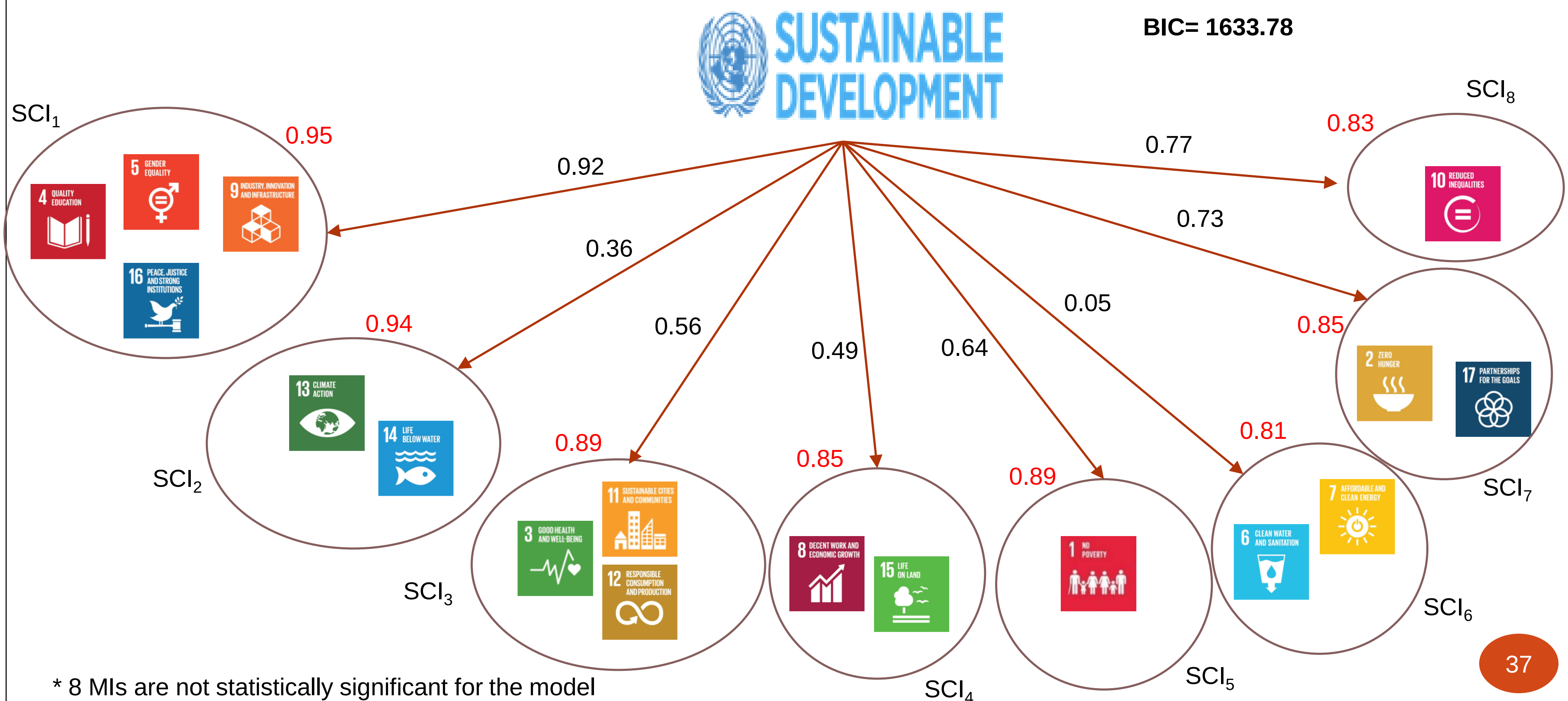
ASSESSMENT of HCI model: 17 goals

100 Manifest Indicators
6 for each goal

- **BIC= 2472.65**
- Polarity: **38** MIs need to change polarity
- **33** MIs are not statistically significant for the model (correlation ≈ 0)
- (They are STATISTICS, but not INDICATORS)
- Reliability: 8 goals are not reliable (low Cronbach's alpha)

- Unidimensionality: only the goal 14 is unidimensional

Exploratory model: 8 factors



* 8 MIs are not statistically significant for the model

Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (1/2)

The *Double Hierarchical Means Clustering* (DHMC) is specified by the following system of equations

$$
\begin{aligned}
&\mathbf{X} = \mathbf{U}_1\mathbf{M}_{11}\mathbf{V}_1'\mathbf{B}_1 + \mathbf{E}_1,\\
&\mathbf{X} = \mathbf{U}_2\mathbf{M}_{22}\mathbf{V}_2'\mathbf{B}_2 + \mathbf{E}_2,\\
&\ldots\ \ldots\ \ldots\\
&\mathbf{X} = \mathbf{U}_Q\mathbf{M}_{QQ}\mathbf{V}_Q'\mathbf{B}_Q + \mathbf{E}_Q,\\
&\ldots\ \ldots\ \ldots\\
&\mathbf{X} = \mathbf{U}_K\mathbf{M}_{KQ}\mathbf{V}_Q'\mathbf{B}_Q + \mathbf{E}_K,\\
&\ldots\ \ldots\ \ldots\\
&\mathbf{X} = \mathbf{U}_n\mathbf{M}_{nJ}\,\mathbf{V}_J'\,\mathbf{B}_Q + \mathbf{E}_n,
\end{aligned}
\qquad (1)
$$

Double Parsimonious tree, Hierarchical partition in Q clusters for units and variables

Double Parsimonious tree, Hierarchical partition in K clusters for units and Q clusters variables

Double Hierarchical Means Classification, Complete tree

Each equation corresponds to a K, Q Clustering & Disjoint PCA (Vichi Saporta 2009)

subject to

$\mathbf{U}_k = [u_{ihk} \in \{0, 1\} : i=1,\ldots,n,\ h=1,\ldots,k]$, $k=2,\ldots,n-1$ — binary, (2)

$\mathbf{U}_k\mathbf{1}_k = \mathbf{1}_n$, $\mathbf{V}_q\mathbf{1}_q = \mathbf{1}_J$, — row stochastic, (3)

$\mathbf{U}_k=[\mathbf{U}_{k-1}\backslash\mathbf{u}_{k-1,k-1},\ \mathbf{u}_{k-1,k},\ \mathbf{u}_{k,k}]$,with $\mathbf{u}_{k-1,k-1} = \mathbf{u}_{k-1,k} + \mathbf{u}_{k,k}$ $k = 3, \ldots, n-1$, nested partitions (4)

$\mathbf{V}_q = [v_{jhk} \in \{0, 1\} : j=1,\ldots,J,\ p=1,\ldots,q]$, $Q=2,\ldots,J-1$ — binary, (5)

$\mathbf{V}_q=[\mathbf{V}_{q-1}\backslash\mathbf{v}_{q-1,q-1},\ \mathbf{v}_{q-1,q},\ \mathbf{v}_{q,q}]$,with $\mathbf{v}_{q-1,q-1} = \mathbf{v}_{q-1,q} + \mathbf{v}_{q,q}$ $q = 3, \ldots, J-1$, nested partitions (7)

Matrix $\mathbf{U}_k$, for $k = 3, \ldots, n$, has k-2 columns equal to $\mathbf{U}_{k-1}$, w.l.o.g.. The last column of $\mathbf{U}_{k-1}$, $\mathbf{u}_{k-1,k-1}$, is equal to the sum of the last two columns of $\mathbf{U}_k$, $\mathbf{u}_{k-1,k}$, $\mathbf{u}_{k,k}$, for $k = 3, \ldots, n-1$. The same considerations apply to matrix $\mathbf{V}$.

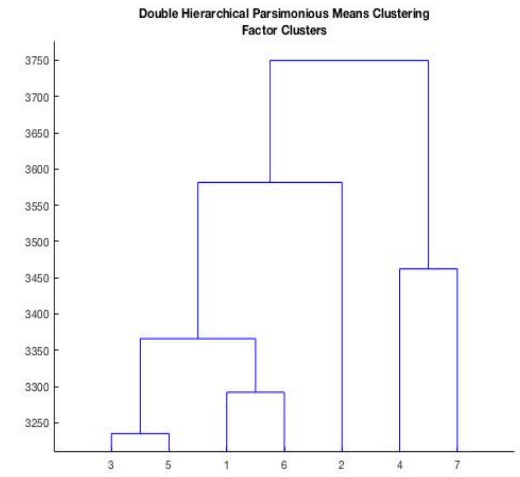
Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (2/2)

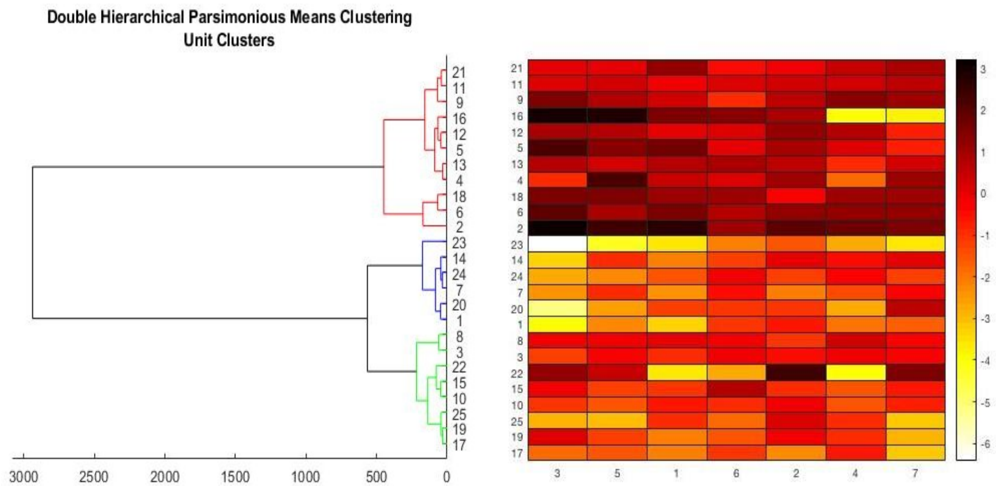
APPLICATION (ECSI DATA)

European Consumer Satisfaction Index: ECSI approach in mobile phone industry.
The dataset contains 250 units and 24 variables.
We supposed to have 7 interrelated latent variables, as follows:

1. **Image** related to manifest variables from 1 to 5.
2. **Expectations** related to manifest variables from 6 to 8.
3. **Perceived Quality** related to manifest variables from 9 to 15.
4. **Perceived Value** related to manifest variables 16 and 17.
5. **Satisfaction** related to manifest variables from 18 to 20.
6. **Complaints** related to manifest variables 21.
7. **Loyalty** related to manifest variables from 22 to 24.

Hierarchical representation of unit and factor clusters and the heatmap computed on the latent scores (obtained by CDPCA).





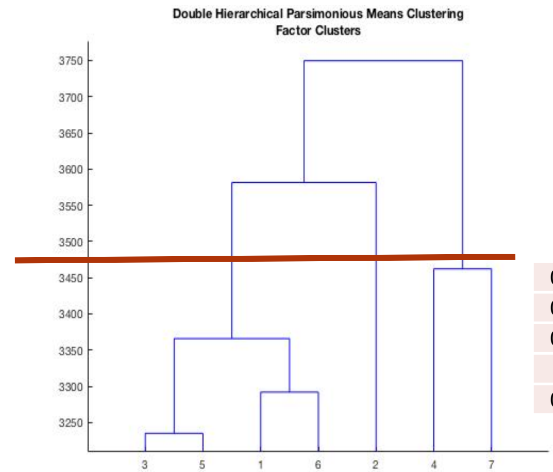
Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (2/2)

APPLICATION (ECSI DATA)

European Consumer Satisfaction Index: ECSI approach in mobile phone industry.
The dataset contains 250 units and 24 variables.
We supposed to have 7 interrelated latent variables, as follows:

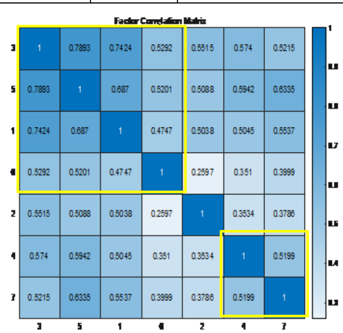
1. **Image** related to manifest variables from 1 to 5.
2. **Expectations** related to manifest variables from 6 to 8.
3. **Perceived Quality** related to manifest variables from 9 to 15.
4. **Perceived Value** related to manifest variables 16 and 17.
5. **Satisfaction** related to manifest variables from 18 to 20.
6. **Complaints** related to manifest variables 21.
7. **Loyalty** related to manifest variables from 22 to 24.

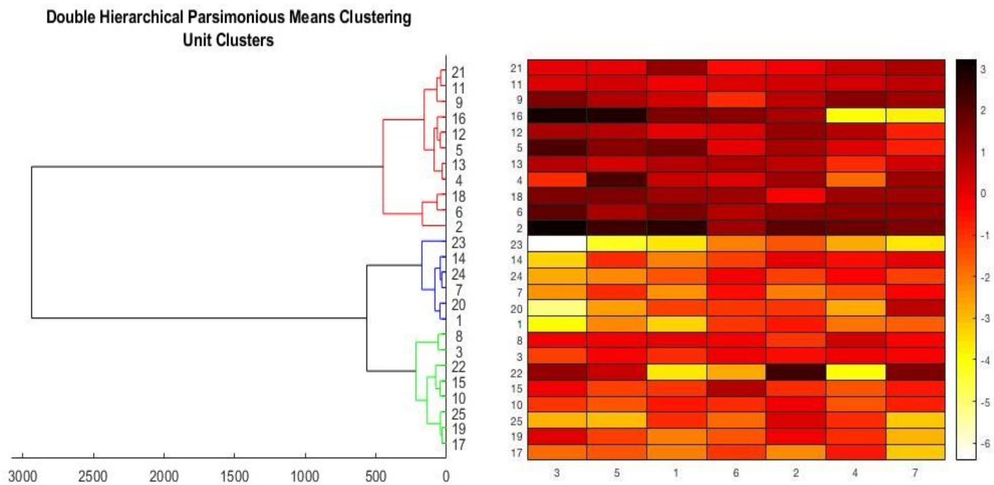
Hierarchical representation of unit and factor clusters and the heatmap computed on the latent scores (obtained by CDPCA).



0,3027
0,3291
0,3481
0,365
0,4697

Hierarchical Level – Factor Clusters	GOF	R^2 specific for each hierarchical level	Cronbach's alpha
1	0,9393	0	0.723
2	0,9371	0,2037	0.452
3	0,9339	0,3027	0.877
4	0,9306	0,3291	0.824
5	0,9331	0,3481	0.779
6	0,9360	0,3650	1.000
7	0,9375	0,4697	0.472



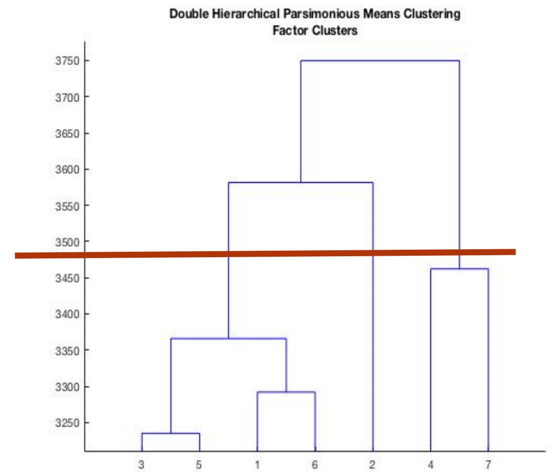


Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (2/2)

APPLICATION (ECSI DATA)

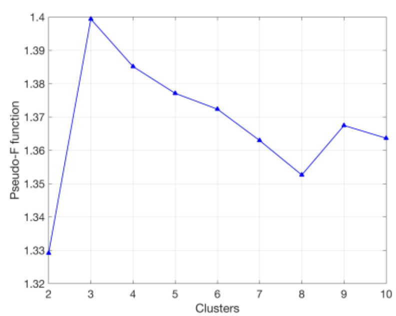
European Consumer Satisfaction Index: ECSI approach in mobile phone industry.
The dataset contains 250 units and 24 variables.
We supposed to have 7 interrelated latent variables, as follows:

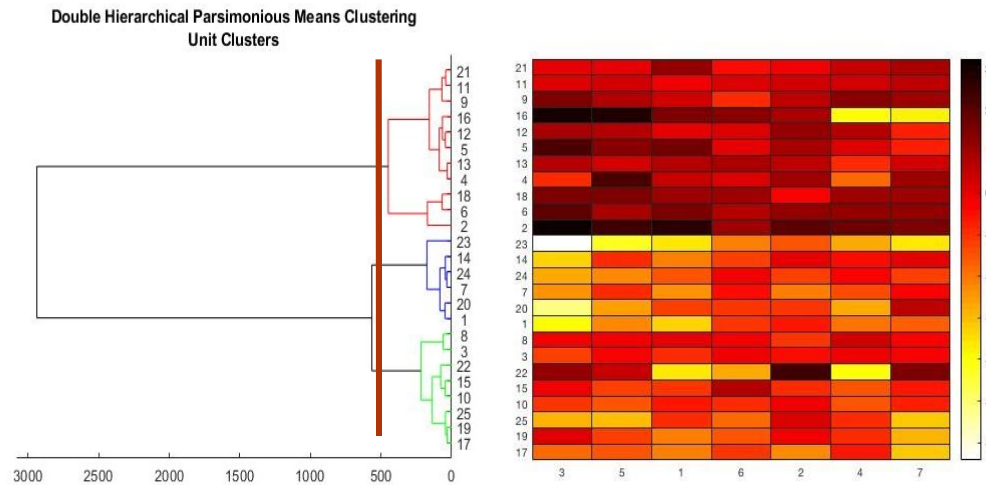
1. **Image** related to manifest variables from 1 to 5.
2. **Expectations** related to manifest variables from 6 to 8.
3. **Perceived Quality** related to manifest variables from 9 to 15.
4. **Perceived Value** related to manifest variables 16 and 17.
5. **Satisfaction** related to manifest variables from 18 to 20.
6. **Complaints** related to manifest variables 21.
7. **Loyalty** related to manifest variables from 22 to 24.



Hierarchical representation of unit and factor clusters and the heatmap computed on the latent scores (obtained by CDPCA).

Fordellone Vichi 2018
Gap method Pseudo-F





Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (2/2)

STATISTICS on Clusters

***Group 1: n* = 137 Satisfied**

Stat/Factors	1	2	3	4	5	6	7
Min	0,452	0,168	0,545	0	0,492	0	0,034
Q1	0,663	0,626	0,705	0,625	0,647	0,667	0,760
Median	0,753	0,714	0,787	0,727	0,738	0,778	0,844
Mean	0,752	0,723	0,788	0,714	0,746	0,781	0,812
Q3	0,828	0,814	0,864	0,798	0,816	1	0,920
Max	1	1	1	1	1	1	1

***Group 2: n* = 82 Medially Satisfied**

Stat/Factors	1	2	3	4	5	6	7
Min	0,125	0,098	0,279	0	0,061	0	0
Q1	0,481	0,446	0,546	0,444	0,430	0,444	0,479
Median	0,545	0,532	0,612	0,565	0,538	0,667	0,609
Mean	0,541	0,521	0,598	0,535	0,512	0,576	0,567
Q3	0,611	0,608	0,648	0,667	0,600	0,667	0,681
Max	0,780	1	0,782	0,879	0,783	1	0,955

***Group 3: n* = 31 Lowly Satisfied**

Stat/Factors	1	2	3	4	5	6	7
Min	0	0	0	0	0	0	0
Q1	0,287	0,375	0,269	0,333	0,247	0,333	0,414
Median	0,397	0,473	0,334	0,444	0,354	0,556	0,539
Mean	0,372	0,459	0,337	0,417	0,346	0,462	0,539
Q3	0,470	0,562	0,439	0,543	0,446	0,667	0,701
Max	0,678	0,806	0,594	1	0,692	0,889	1

***Group 1: n* = 58 Very Satisfied**

Stat/Factors	1	2	3	4	5	6	7
Min	0,640	0,168	0,672	0,444	0,568	0,667	0,726
Q1	0,751	0,644	0,798	0,727	0,754	0,778	0,854
Median	0,824	0,766	0,845	0,778	0,801	1	0,909
Mean	0,830	0,754	0,860	0,810	0,818	0,906	0,901
Q3	0,908	0,895	0,904	0,889	0,907	1	0,945
Max	1	1	1	1	1	1	1

***Group 2: n* = 79 Satisfied**

Stat/Factors	1	2	3	4	5	6	7
Min	0,452	0,397	0,545	0	0,492	0	0,034
Q1	0,627	0,626	0,673	0,565	0,616	0,556	0,664
Median	0,697	0,696	0,716	0,667	0,692	0,667	0,787
Mean	0,694	0,701	0,736	0,644	0,694	0,689	0,746
Q3	0,780	0,775	0,786	0,741	0,765	0,778	0,851
Max	1	1	0,957	1	1	1	1

GRAZIE PER L'ATTENZIONE