SOSEWEST

FISCO E SERVIZI

Regression models for panel data: Some recent developments

Franco Peracchi

Georgetown University, EIEF and University of Rome Tor Vergata

SOSE Workshop, September 28, 2018

Outline

Why panel data?

Panel (or longitudinal) data consist of repeated observations on a set of units.

My focus is on panels where the units are elements of a well defined population and observations are repeated through time.

This kind of panels combine features of both cross-section and time-series data:

- as for cross-sections, issues of sample design, sample selection and measurement error may affect representativeness of the underlying population;
- as for time-series, the data are naturally ordered by the value of the time index and usually display some regularity or persistence over time.

Advantages of panel data:

- they simplify the analysis of a variety of economic problems that would be much harder to study using a pure cross-section;
- unlike macroeconomic time-series, they allows us to study behavior at the level of the individual microeconomic units, while controlling for time invariant unit-specific unobserved heterogeneity in a flexible manner.

Examples of panel data

Bank of Italy:

- Survey on Household Income and Wealth (SHIW), nationally-representative (almost) biennial survey of households, with a panel component;
- Industrial and Service Firms ("Indagine sulle Imprese Industriali e dei Servizi"), nationally-representative annual panel of firms, stratified by industry and firm size.

SOSE:

- ISA project (replaces the "Studi di Settore" project), nationally-representative annual panel of firms, stratified by industry;
- "Fabbisogni Standard" and "Capacità Fiscale" projects, annual panels of (almost) all units comprising three different levels of government (Municipalities, Provinces and Metropolitan Cities, and Regions).

Issues with panel data

Data collection issues:

- survey design and survey process;
- sample design;
- missing data due to either nonresponse or unbalanced panel design;
- measurement error.

Modeling issues:

- conditions for identifiability of the parameter(s) of interest;
- unobserved heterogeneity;
- nonlinearity.

The standard linear model for balanced panel data

Given T observations $\{(\boldsymbol{X}_{it}, Y_{it})\}$ on an outcome Y and k regressors $\boldsymbol{X}$ for n units (households, firms, municipalities, etc.), the standard linear model for balanced panel data assumes

$$Y_{it} = \alpha_i + \boldsymbol{X}_{it}^\top \boldsymbol{\beta} + U_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \tag{1}$$

where α_i is an unknown unit-specific intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is the vector of parameters of interest, and U_{it} is an unobserved error term.

The "individual effect" α_i represents time-invariant unobserved heterogeneity, i.e. omitted time-invariant determinants of Y.

Different assumptions about the relations between the α_i's, the $\boldsymbol{X}_{it}$'s and the U_{it}'s result in different versions of the standard model.

Conventional linear estimators of $\boldsymbol{\beta}$

Large menu of estimators to choose from:

- Ordinary LS (OLS): Regress Y_{it} on $\boldsymbol{X}_{it}$.
- First differencing (FD): Regress ΔY_{it} on $\Delta \boldsymbol{X}_{it}$.
- Fixed effects (FE) or "within" estimator: Regress $Y_{it} - \overline{Y}_i$ on $\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i$.
- "Between" estimator: Regress $\overline{Y}_i$ on $\overline{\boldsymbol{X}}_i$.
- Generalized least squares (GLS) and feasible GLS (FGLS) estimators: Regress $Y_{it} - \psi \overline{Y}_i$ on $\boldsymbol{X}_{it} - \psi \overline{\boldsymbol{X}}_i$, with ψ known (GLS) or estimated (FGLS) by $\hat{\psi}$.
- Mundlak's correlated RE estimator: Regress Y_{it} on $\boldsymbol{X}_{it}$ and $\overline{\boldsymbol{X}}_i$.
- Chamberlain's correlated RE estimator: Regress Y_{it} on $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$.

Note: $\Delta \boldsymbol{X}_{it} = \boldsymbol{X}_{it} - \boldsymbol{X}_{i,t-1}$, $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$, $\overline{\boldsymbol{X}}_i = n^{-1} \sum_{i=1}^{T} \boldsymbol{X}_{it}$, and $\overline{Y}_i = n^{-1} \sum_{i=1}^{T} Y_{it}$.

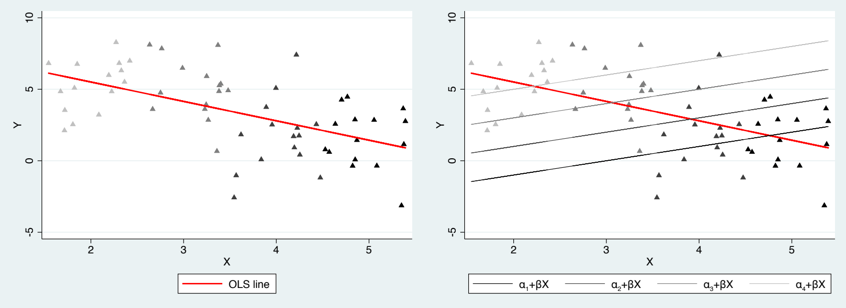
Remarks

The GLS estimator is a matrix weighted average of the FE and "between" estimators.

The GLS and FGLS estimators converge to the OLS estimator when $\mathbb{V}(\alpha_i) \to 0$, and to the FE estimator when $T \to \infty$.

Given a set of instruments, the class of linear estimators of β may be enlarged by considering instrumental variable (IV) versions of all the estimators I mentioned.

Example



Choosing among estimators

Conventional linear estimators are asymptotically normal (Gaussian) under mild regularity conditions, but require different exogeneity assumptions for identification of β. For example: FD and FE only require mean independence of any U_{it} from $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$, while "between", GLS and FGLS also require mean independence of α_i from $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$.

Consequences:

- robustness-efficiency tradeoffs;
- different treatment of time-invariant regressors;
- differences in what one can predict.

With IV procedures an additional issue arises, namely validity (i.e., exogeneity and relevance) of the proposed instruments. In the case of survey data, the characteristics of the interview process and the interviewers provide arguably valid instruments (Nicoletti & Peracchi 2005).

Remarks:

- The Law of Decreasing Credibility (Manski 2003): The credibility of inference decreases with the strength of the assumptions maintained.
- Pre-testing issues arise when using Hausman-type tests of exogeneity assumptions as model selection devices.

Correlated errors

The standard linear model assumes that the errors in (1) are uncorrelated both within and between units. This assumption can easily be weakened to cover cases where the errors are either serially correlated within units or cross-correlated between units.

The second case has become quite relevant as it includes settings that are increasingly common in empirical work:

- clustered samples;
- spatial panel data;
- network panel data.

In all these cases, consistency (or lack thereof) and asymptotic normality of conventional linear estimators of $\boldsymbol{\beta}$ are unaffected.

However, inference is less straightforward because of the more complex nature of the asymptotic variance matrix of the estimators of $\boldsymbol{\beta}$ (Moulton 1986). For this reason, jackknife or bootstrap methods are increasingly used.

Nearly-singular panel designs

The FE estimator, although often preferred because of its weaker identification assumptions, also requires nonsingularity of the second moment matrix

$$\mathbf{S}_{XX} = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} (\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i)^\top$$

or, more precisely, requires the smallest eigenvalue of $\mathbf{S}_{XX}$ to be sufficiently far from zero.

Low longitudinal variation of the $\boldsymbol{X}$'s causes failure of this condition and creates both numerical problems and problems for conventional inference, as the usual normal approximation may not be appropriate (Hahn, Ham & Moon 2011)

Possible solution: shrinkage methods to reduce variability of the estimates of $\boldsymbol{\alpha}$ and improve precision of the estimates of $\boldsymbol{\beta}$. Examples:

- quadratic penalty;
- least absolute penalty (Koenker 2004).

Heterogeneous slopes

Why should unobserved heterogeneity be confined to the model intercept?

Replacing $\boldsymbol{\beta}$ in (1) by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ dramatically increases the number of model parameters from $n + k$ in model to nk.

Mixed models, a parsimonious way of treating heterogeneity in the $\boldsymbol{\beta}_i$, essentially generalize the RE model.

Key assumptions:

- random sample of units;
- mean independence of $\boldsymbol{\beta}_i$ from $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$.

Time-varying unobserved heterogeneity

Why should unobserved heterogeneity be time-invariant?

One possibility is to replace α_i in (1) by $\alpha_{i1}, \ldots, \alpha_{iT}$, where the α_{it}'s follow a unit-specific process, typically a stationary or nonstationary ARMA process, e.g. a stationary AR(1) or a random walk. This is equivalent to assuming that the omitted variables in (1) are time-varying instead of time-invariant.

Testing the null hypothesis of time-invariant unobserved heterogeneity in model (1) may be based on the comparison of the FE and FD estimators of $\boldsymbol{\beta}$ (Bartolucci, Belotti & Peracchi 2017).

Further extensions: replace $\boldsymbol{\beta}$ in (1) by $\boldsymbol{\beta}_{it}$.

Time effects and heterogeneous time trends

How about controlling for time-varying “macro” effects?

This just amounts to adding a “time effect” γ_t to (1).

The time effects $\gamma_1, \ldots, \gamma_T$ may be modeled in a completely unrestricted way through a sequence of time dummies.

Alternatively, one may assume a low-order polynomial time trend, e.g., a linear one ($\gamma_t = \gamma t$).

Replacing $\alpha_i + \gamma_t$ by a unit-specific linear trend, i.e. by $\alpha_i + \gamma_i t$, is a simple way of modeling heterogeneous time trends. The resulting model is easily estimated via Frisch-Waugh-Lovell, which amounts to linear detrending of both Y_{it} and $\boldsymbol{X}_{it}$.

Failures of exogeneity

The crucial problem when employing model (1) in empirical work is potential failure of the exogeneity assumptions, which may arise from a variety of (not mutually exclusive) reasons:

- omitted variables;
- measurement error;
- simultaneity;
- missing data;
- sample selection.

In what follows I focus on failure of the last two assumptions.

This is not because failure of the first three is less frequent or less relevant, but mainly because failure of the last two leads naturally to nonlinear models, often with an appealing latent linear structure for unobservables plus a nonlinear observation equation mapping unobservables into observables.

Missing data

Common sources of missing data are item and unit nonresponse. Both are widespread and increasingly frequent in sample surveys (Meyer, Mok & Sullivan 2015, Bollinger et al. 2018).

Other sources, specific to panel data and leading to unbalanced panels (i.e. T varying across units), are:

- attrition (monotone or not);
- new entry.

The real issue is not how to allow for unit-specific T's (all conventional estimators allow this), but whether missingness can cause bias.

Classification of missing data mechanisms (Rubin 1976; Little & Rubin 2002):

- missing completely at random (MCAR);
- missing at random (MAR);
- missing not at random (MNAR).

While MCAR and MAR only lead to inefficient estimation of the parameters of interest, MNAR causes bias.

Approaches to MCAR and MAR

Three possibilities:

- Complete-case analysis (not recommended).
- Imputation-based approaches for missing X's:
  - the fill-in approach;
  - missing-indicators methods (Little 1992; Dardanoni, Modica & Peracchi 2011; Dardanoni et al. 2015);
  - multiple imputations (Rubin 1987, 1996).
- Re-weighting approaches for missing Y's:
  - the Horvitz-Thompson method (Horvitz & Thompson 1952);
  - generalizations via inverse probability weighting (Wooldridge 2007).

Approaches to MNAR and sample selection

Two possibilities:

- Achieve point identification of the parameters of interest by explicitly modeling the sample selection process:
  - Heckman framework (Heckman 1979), based on a model of the form (1) for a latent outcome Y_{it}^*, the observability condition $Y_{it} = Y_{it}^*$ if $S_{it} = 1$, and a model for the observability indicator S_{it};
  - more general Tobit models (Amemiya 1984; Vella 1998).
- Do not insist on point identification and only impose "minimal" assumptions that still allow to set identify the parameters of interest, e.g. to assert that $\underline{\mu}_{it} \leq \mathbb{E}[Y_{it}] \leq \overline{\mu}_{it}$:
  - Manski's bounds (Manski 1989) on $\mathbb{E}[Y_{it}]$ for the case when a binary Y_{it} is only observed if $S_{it} = 1$;
  - improving upon Manski's bounds, e.g. by using information on re-entering units when attrition is nonmonotone (Arpino, De Cao & Peracchi 2015).

Linear vs. nonlinear models

Nonlinear models are important when the range of Y_{it} is restricted (e.g. Y_{it} is binary, discrete, categorical, censored or truncated), when data are MNAR, or in the presence of sample selection.

Linear models may still be employed as simple and useful "best approximations" to nonlinear models (Angrist & Pischke 2009).

Key elements of a nonlinear panel data model:

- definition of exogeneity;
- relationship between unobserved heterogeneity and observed regressors;
- temporal dependence among the unobservables.

Difficulties with nonlinear models:

- more complicated identification conditions;
- incidental parameter problem with the FE approach when T is small;
- distinction between model parameters, partial effects at the average, and average derivatives (or average partial effects);
- computational issues;
- need stronger exogeneity assumptions on the distribution of the unobservables.

Nonlinear models with strictly exogeneity

Generalized linear models (McCullagh & Nelder 1989) as a tractable class that includes the Gaussian linear model as special case.

Popular panel data examples include:

- binary logit and probit models (Andersen 1970; Chamberlain 1980);
- ordered logit and probit models;
- Poisson and negative binomial models;
- exponential models.

Approaches to estimation (Liang & Zeger 1986):

- nonlinear LS and nonlinear weighted LS;
- maximum likelihood (ML) and conditional ML (CML).

Nonlinear models with weak exogeneity

The key purpose of models with lagged endogenous variables is to separate true state dependence from spurious state dependence due to unobserved heterogeneity (Heckman 1991).

Role of initial conditions when T is small (Blundell & Bond 1998).

Example: dynamic logit model (Honoré & Kiriazidou 2000; Bartolucci & Nigro 2010, 2012).

Big data

What is "big-data"?

"The leapfrogging of the discourse on big data to more popular outlets implies that a coherent understanding of the concept and its nomenclature is yet to develop" (Gandomi & Haider 2015).

A popular definition is: high-volume, high-velocity, high-variety data (the "Three V" definition). For example: "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information" (TechAmerica Foundation 2012)

I prefer a narrower definition based on two dimensions of data size: length (n) and width (k). Big data are, at the same time:

- "long": large n or, in the panel case, large n and T (crucial for validity of asymptotic approximations);
- "wide" (high-dimensional): $k \gg n$, because of either a large number of available variables, or transformations or interactions of the underlying set of variables.

Machine learning methods

This term is often used to denote a rather heterogeneous set of methods employed for analyzing big data.

Their main characteristics include:

- focus on classification or prediction, rather than model fitting and parameter estimation;
- emphasis on flexibility;
- trade-off between bias and variance;
- importance of Bayesian ideas;
- interest in the whole distribution – not just a few features such as mean or variance.

Characterizing a distribution

What characterization of a distribution should one focus on?

The answer partly depends on the nature of Y_{it}. Possibilities include:

- the density function – only defined when Y_{it} is continuously distributed but easily generalized to multivariate outcomes;
- the quantile function – requires no mass points in the distribution of Y_{it} and cannot be generalized to multivariate outcomes;
- the distribution function – can accomodate mass points in the distribution of Y_{it} and is easily generalized to multivariate outcomes.

The last two characterizations are the basis of two alternative methods:

- quantile regression (Koenker & Basset 1978; Koenker 2005, 2017);
- distribution regression (Foresi & Peracchi 1995; Chernozhukhov, Fernandéz-Val & Melly 2013).

Classes of methods

Shrinkage or "regularization" methods (often with a Bayesian justification), including:

- ridge regression (Hoerl & Kennard 1970);
- smoothing splines (Silverman 1985);
- least absolute shrinkage (Frank & Friedman 1993), LAR (Efron et al. 2004) and variants.

Semi-parametric methods, including:

- projection pursuit (Friedman & Tukey 1974; Huber 1985);
- additive modeling (Hastie & Tibshirani 1990);
- neural networks (Ripley 1996) and variants ("deep-learning algorithms");
- local fitting (Tibshirani & Hastie 1987; Fan & Gijbels 1996).

Nonparametric methods, including:

- kernel methods;
- CART (Breiman et al. 1984) and other tree-based methods (bagging, boosting, random forests, etc.).

Inference and model selection

Construction of confidence intervals:

- bootstrap methods (Efron 1980, Efron & Tibshirani 1993): nonparametric and parametric bootstrap;
- subsampling (Politis, Romano & Wolff 1999).

Model selection:

- Mallows C_p (Mallows 1973);
- information criteria: AIC (Akaike 1973), BIC (Schwarz 1978) and variants;
- cross-validation (Stone 1974, 1977);
- the problem of post-selection inference.

Shrinkage and model selection via the LASSO (Tibshirani 1996).

Compromise estimators:

- Bayesian model averaging (Leamer 1978; Raftery, Madigan & Hoeting 1997);
- frequentist model averaging (Hjort & Claeskens 2003; Hansen 2007);
- Bayesian-frequentist fusions, e.g. WALS (Magnus, Prüfer & Powell 2010).

Data infrastructures

Research in the era of big data requires proper infrastructures, such as open data (OD) initiatives and Research Data Centers (RDCs).

Examples of OD initiatives:

- Bank of Italy Remote Access to Micro Data (BIRD) system (https://www.bancaditalia.it/statistiche/basi-dati/bird/).
- VisitINPS Research Program (http://www.inps.it/NuovoportaleINPS/default.aspx?itemdir=47212&lang=IT).
- SOSE Open Civitas portal (https://www.opencivitas.it).

RDCs are partnerships between statistical agencies and leading research institutions. They are secure facilities providing authorized access to restricted-use microdata for statistical purposes only.

An example from the U.S. are the Federal Statistical RDCs, a system of 28 open RDCs locations partnering with over 50 research organizations such as universities (including Georgetown University), non-profit research institutions, and government agencies.

The experience of the Georgetown RDC

The Georgetown RDC, opened in 2017 and located at the Massive Data Institute at Georgetown University's McCourt School of Public Policy, provides secure access to qualified researchers examining a wide range of social and economic issues.

Individuals wishing to conduct research at the Georgetown RDC must submit a research proposal to the Center for Economic Studies (CES) at the U.S. Census Bureau (http://www.census.gov/ces/).

After a researcher has developed a proposal with RDC administrators, the proposal is submitted to the CES for Census review. (Depending on the data sets requested, other agencies might also review the proposal.) Researchers on approved projects must also complete a background investigation. These steps may take months to complete.

See http://mccourt.georgetown.edu/massive-data-institute/RDC?).

La compliance fiscale nell'era dei big data: considerazioni sul caso italiano

Santoro, A

Roma, SOSE, 28 Settembre 2018.
Workshop Economico Statistico Tecnologico.

Sommario

Big data e politiche pubbliche

Caratteristiche dei big data rilevanti per l'analisi economica (Einav and Levin (2013)):

- + osservazioni e/o + variabili;
- + veloci (tempo reale);
- + eterogenei (network);
- meno strutturati ("non rettangolari").

> *High volume and highly structured administrative data possibly combined with structured and unstructured real time data to design and evaluate evidence-based public policies. Mergel et al. (2016).*

Big data e previsione dei comportamenti

We know from experience that it is all too easy to construct a predictor that works well in-sample but fails miserably out-of-sample: e.g. n predictors fit perfectly n observations (Varian 2014).

- piccoli dataset e modelli lineari: poco bias, ma molta varianza;
- grandi dataset: test di validazione incrociata di modelli lineari e non lineari per minimizzare PE.

Con grandi dataset, il modello di analisi + appropriato risulta *dai* dati, non viene *imposto ai* dati .

La previsione dei comportamenti economici

La moderna economia comportamentale (behavioural economics) ha identificato molte ragioni per cui l'impatto delle politiche pubbliche non è quello atteso sulla base dei modelli economici convenzionali.

> *Many models used for policymaking assume that people will quickly recognise and respond to a change in their financial incentives in the way that the policymaker intends. In reality, this may not happen. Tax policy provides some good examples (...) (Behavioural Intelligence Unit, UK Govt., 2018) .*

La previsione dei comportamenti fiscali

La possibilità di predire correttamente i comportamenti fiscali è fondamentale nell'implementazione del *nuovo fisco* (Oecd, 2017):

- approccio tradizionale alla compliance fiscale:
  - pochi dati interni;
  - politiche repressive (controlli) post dichiarative;
- nuovo rapporto tra fisco e contribuente:
  - dati massivi ma individuali, sia interni sia esterni, disponibili in tempi ravvicinati;
  - politiche sia post sia pre-dichiarative (c.d pre-filing);
  - politiche sia repressive sia di incentivo alla compliance (*nudging*) fino al *no filing*.

Cosa serve al nuovo fisco? Nuove competenze per nuovi dati.

> *Advanced analytics is the process of applying statistical and machine-learning techniques to uncover insights from data (Oecd, 2017).*

- *predictive analytics* per anticipare i problemi guardando a ció che è accaduto in passato: quali parti delle dichiarazioni sono + frequentemente fraintese e/o mal compilate? quali anomalie dichiarative si ripetono?
- *prescriptive analytics* per comprendere i nessi causali fra policy e comportamenti: un certo tipo di comunicazione ai contribuenti è stata efficace a ridurre le compilazioni tardive? quale policy ha effettivamente aumentato la compliance?

NON SONO DOMANDE NECESSARIAMENTE NUOVE, MA NUOVO È IL MODO DI RISPONDERE A QUESTE DOMANDE.

Esempi dal mondo/1

Nel 2016 Oecd ha pubblicato uno studio sull'uso dell'advanced analytics (Oecd, 2016a)a cui hanno partecipato 16 amministrazioni fiscali (non l'Italia). L'utilizzo di *advanced analytics* avviene

- nel 94% dei casi per la selezione dei contribuenti da controllare;
- nel 75% dei casi per la gestione del debito fiscale;
- nel 69% dei casi per la gestione delle tempistiche dei pagamenti;
- nel 50% dei casi per i servizi ai contribuenti.

Esempi dal mondo/2

Modelli predittivi e prescrittivi sono stati sviluppati per affrontare problemi specifici (Oecd, 2016a).

a Australia: analisi del rischio a livello di tax consultant;

b Canada: previsione del rischio di omessa dichiarazione;

c Singapore: *text mining* per prevedere le richieste di assistenza dei contribuenti.

d Cina: modello CGE per previsione e validazione degli impatti della riforma della tassazione delle imprese.

Cosa serve al nuovo fisco? Modelli organizzativi e sfide culturali.

Due modelli organizzativi:

- diffuso per progetti: ogni unità operativa sviluppa in modo indipendente i propri progetti basati su tecniche avanzate di analisi dei dati, utilizzando le risorse umane e tecnologiche necessarie anche coordinandosi con la o le unità che le possiedono;
- centralizzato: una unità centrale di analisi che raggruppa e gestisce tutti i progetti di interesse delle diverse unitá operative.

Conflitto culturale *operativi* vs. *analisti* :

- background/linguaggio giuridico-amministrativo vs statistico-informatico;
- esperienza e istinto vs. analisi dei dati;
- "astrattezza" vs. "superficialità".

Esempi dal mondo

- Irlanda: creazione del RAG (Research Analytics Group) che decide quali progetti vanno implementati e riunisce sia le competenze analitiche sia quelle IT;corsi interni e inserimento di un insegnamento di analytics nel corso seguito dal personale addetto alle verifiche fiscali.
- Canada: modello c.d *hub and spoke*: *hub* è il gruppo che si dedica a monitorare la qualità dei processi e gestisce solo i più complessi, *hub* è il gruppo che applica l'approccio analitico ai diversi livelli operativi; i dirigenti di questi sviluppano specifici programmi per diffondere la fiducia degli operativi e de-misterizzare l'approccio analitico.

In an era of persistently reduced budgets, the use of data analytics has become more important than ever to drive innovation, risk management, and decision making across the agency (Jeff Butler, IRS).

Cosa serve al nuovo fisco? Il problema del rispetto della privacy.

Negli USA, dove IRS usa attivamente i dati dei social media incrociati con quelli fiscali (caso Wilson), si parla apertamente di *sistema di sorveglianza fiscale*:

> *A substantial part of the research agenda related to taxation and surveillance should be dedicated to determining how to gather and analyze tax-relevant information without losing the* **public good** *of privacy (Hatfield, 2015)*

- Uso regolato da norme generali o principio dello scambio volontario?

Il caso italiano: il report dell'Oecd/1

Efforts to increase taxpayers compliance and make it easier to comply have followed a path of constant improvement since the creation of the agencies (Oecd, 2016b, section 79).

Segmentation and modern risk-assessment practices have been introduced over time by the Revenue Agency to work more efficiently (section 81).

There are still large margins for improvement and certain key issues need to be addressed with determination (section 83)

Il caso italiano: il report dell'Oecd/2

Each institution in charge manages tax compliance independently and **there is no strategic process in place for jointly identifying key compliance risks** *and priorities, how these risks will be addressed, and how resources will be allocated across the board (section 84).*

The case of **Sose** *is rather emblematic: it produces benchmark analyses for different business sectors (sector studies) which are discussed and agreed with representatives of business associations and the agencies, as well as risk analyses based on them. While there is a working group (...), this is at the operational level and is not replicated at management level (...) shortcomings such as the lack of access to certain data needed to carry out data analyses and the lack of feedback on the use of them (...). focus on certain sectors may well be decided* **ad hoc at the operational level** *rather than as part of an overall compliance strategy (section 85).*

Il caso italiano: il report dell'Imf

Audit case selection is largely decentralized (...). This is out of step with international trends, which have seen a strong move to centralized audit case selection. There is little point in investing time and effort into creating sophisticated risk modelling processes at the central level unless they are used to drive the national audit case selection process. Risk assessment and case selection are complex processes in a modern tax administration and require high level of expertise. **Advanced analytics** *have been proven to deliver better audit case selection and higher tax revenue yields. This level of expertise simply cannot be developed or maintained across a large number of very small distributed units. Centralizing the function also reduces the risk of inappropriate case selections. Significant savings and improvements in effectiveness could be achieved by consolidating the audit case selection function to at least the regional level, as a first step (Imf, 2015, sections 66, 67 and 68).*

Il caso italiano: riforme recenti dell'Agenzia delle entrate

- La creazione della Direzione ICT;
- la nuova organizzazione a livello centrale:
  - due divisioni (contribuenti e servizi);
  - nella divisione contribuenti, creazione del settore analisi del rischio;
  - all'interno della divisione contribuenti, direzioni suddivise per tipologia di contribuente (grandi contribuenti, piccole imprese, persone fisiche e lavoratori autonomi) con ulteriori funzioni di analisi del rischio.
- la nuova organizzazione a livello regionale: analisi del rischio territoriale.

Il caso italiano: questioni aperte

- Su quali competenze punterà l'Ade per l'analisi del rischio a livello locale?
- Come verrà implementato l'approccio analitico nei servizi al contribuente, a livello locale e a livello centrale?
- La frammentazione identificata nei report di Oecd e Imf è stata affrontata? Il modello di gestione della filiera dei dati utilizzato è ancora preferibile?.
- Fino a che punto il garante della privacy è il "vero" ministro delle finanze?[la vicenda dell'Archivio dei Rapporti]

Il caso italiano: la riforma degli studi di settore

Elementi di coerenza con nuovo fisco:

- transizione da SdS a ISA segna passaggio da logica repressiva a preventiva e premiale;
- negli ISA utilizzati anche dati non da SdS.

Aspetti da valutare:

- gli ISA potrebbero diventare lo strumento su cui fa perno una *strategia condivisa* di induzione alla compliance?
- è ancora necessario un modello da stimare o possiamo "far parlare i dati" senza una struttura predefinita?;
- cosa capiscono i contribuenti del percorso logico che porta al calcolo dell'indicatore?

Riferimenti bibliografici

- Einav, L. and Levin, J., 2013, "The Data Revolution and Economic Analysis", NBER Working Paper 19035, May 2013.
- Hatfield, M., 2015 "Taxation Surveillance: an Agenda", Yale Journal of Law and Technology, 17, 319.
- Imf, 2015, "Italy. Enhancing Government and Effectiveness of Fiscal Agencies", Fiscal Affairs Department, December.
- Mergel, I., Rethemeyer R.K. and Isett, K., 2016, "Big Data in Public Affairs", Public Administration Review, 76,6.
- Oecd, 2016a,"Italy's Tax Administration. A Review of Institutional and Governance Aspects", January.
- Oecd, 2016b,"Advanced Analytics for Better Tax Administration. Putting Data to Work", Oecd publishing, Paris.
- Oecd, 2017,"Tax Administration 2017. Comparative information on OECD and other advanced and emerging countries", Oecd Publishing, Paris.
- Varian, H.R., 2014, "Big Data: New Tricks for Econometrics", Journal of Economic Perspectives, 28,2.

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

<u>Advantages</u>:

- 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

<u>Pitfalls</u>:

- 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

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

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

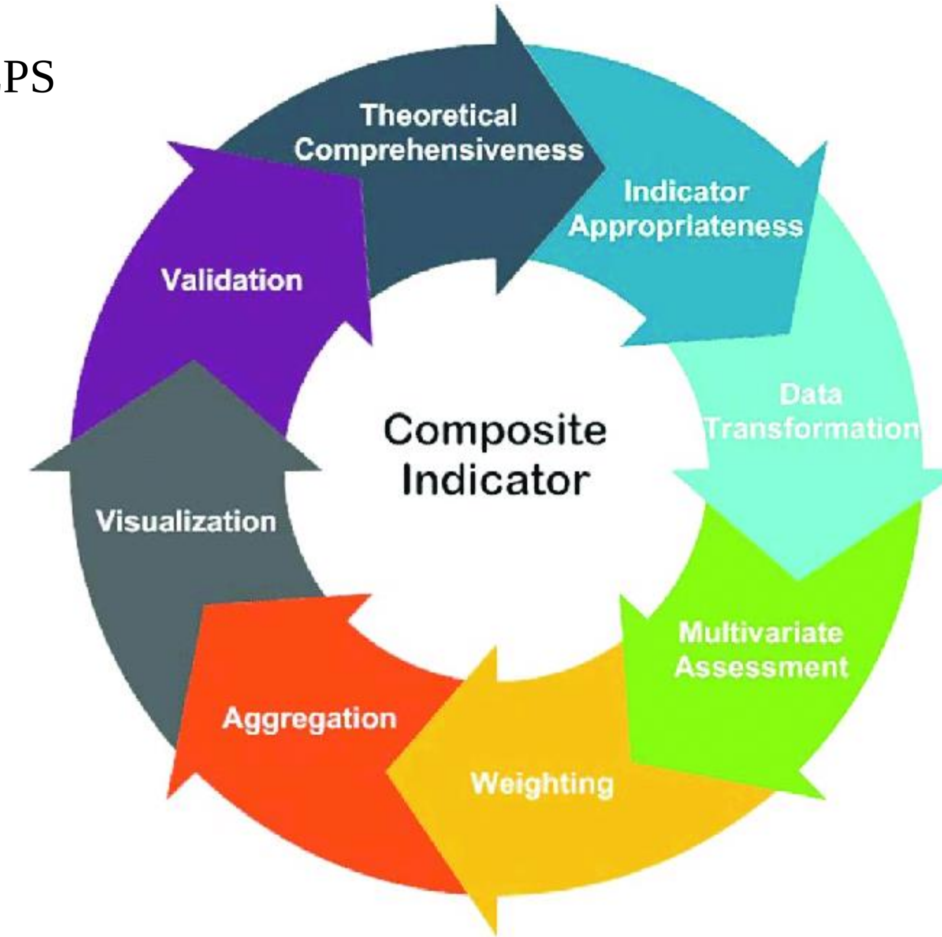
European Commission

Ingredients for Constructing Composite indicators

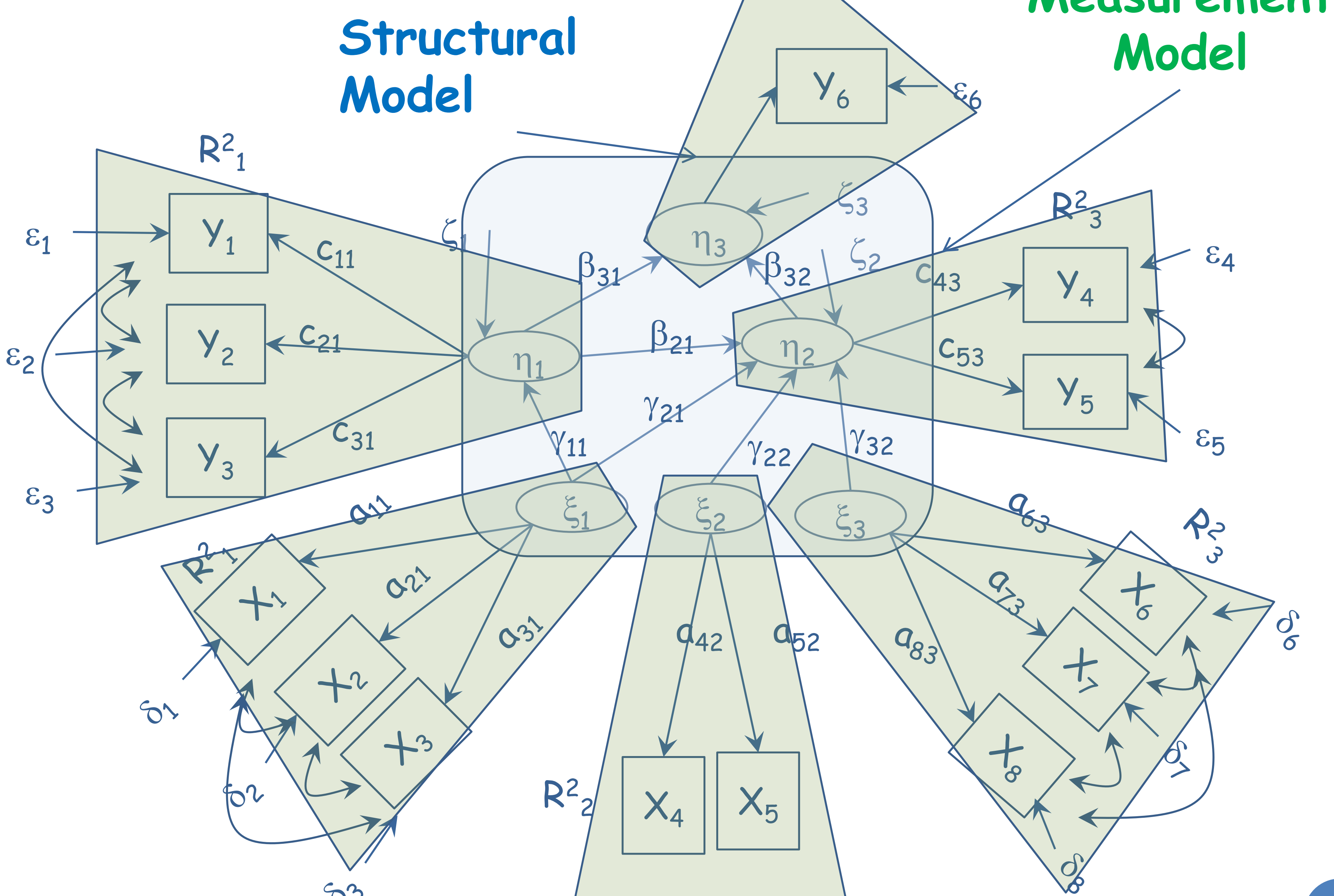
COMPOSTE INDICATOR CONSTRUCTION

HANDBOOK 2008

STEPS

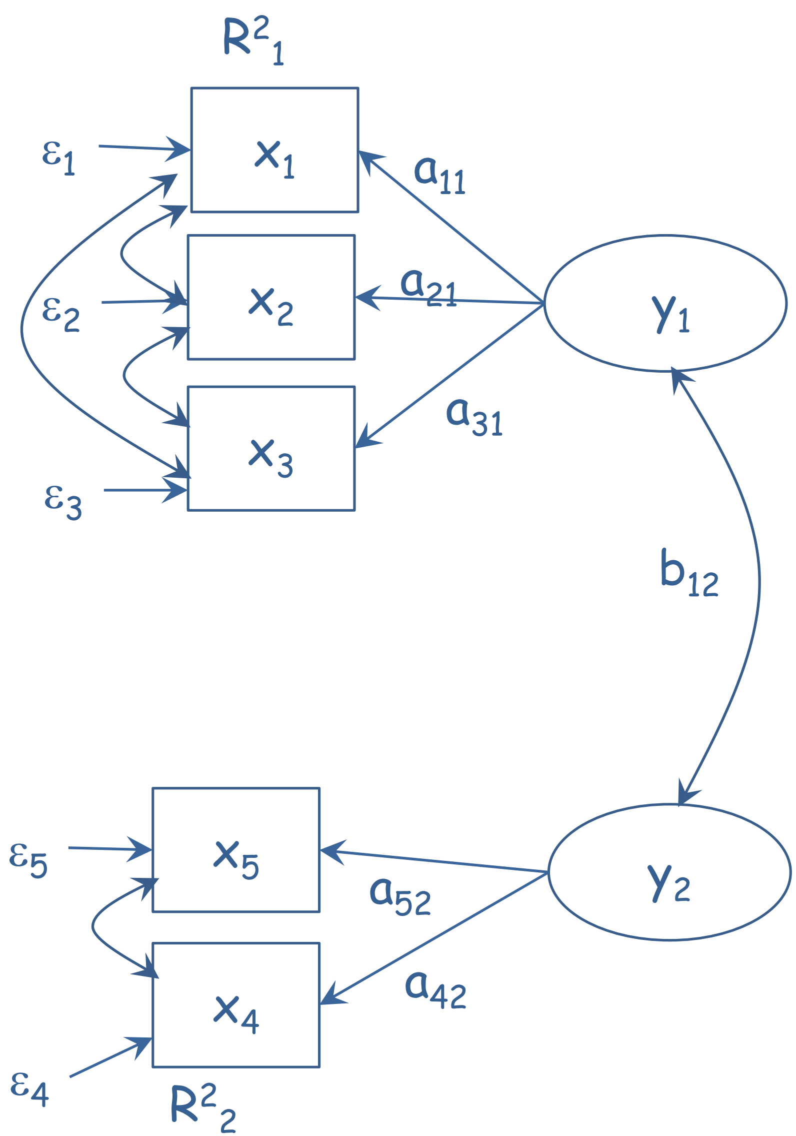


SEM Model



Structural Model
Measurement Model
R^2_1
R^2_2
R^2_3
y_1
y_2
y_3
y_4
y_5
y_6
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
η_1
η_2
η_3
ξ_1
ξ_2
ξ_3
ζ_1
ζ_2
ζ_3
ε_1
ε_2
ε_3
ε_4
ε_5
ε_6
δ_1
δ_2
δ_3
δ_6
δ_7
δ_8
c_{11}
c_{21}
c_{31}
c_{43}
c_{53}
β_{31}
β_{32}
β_{21}
γ_{11}
γ_{21}
γ_{22}
γ_{32}
a_{11}
a_{21}
a_{31}
a_{42}
a_{52}
a_{63}
a_{73}
a_{83}

Path diagram Model of Confirmatory Factor Analysis



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

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

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

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

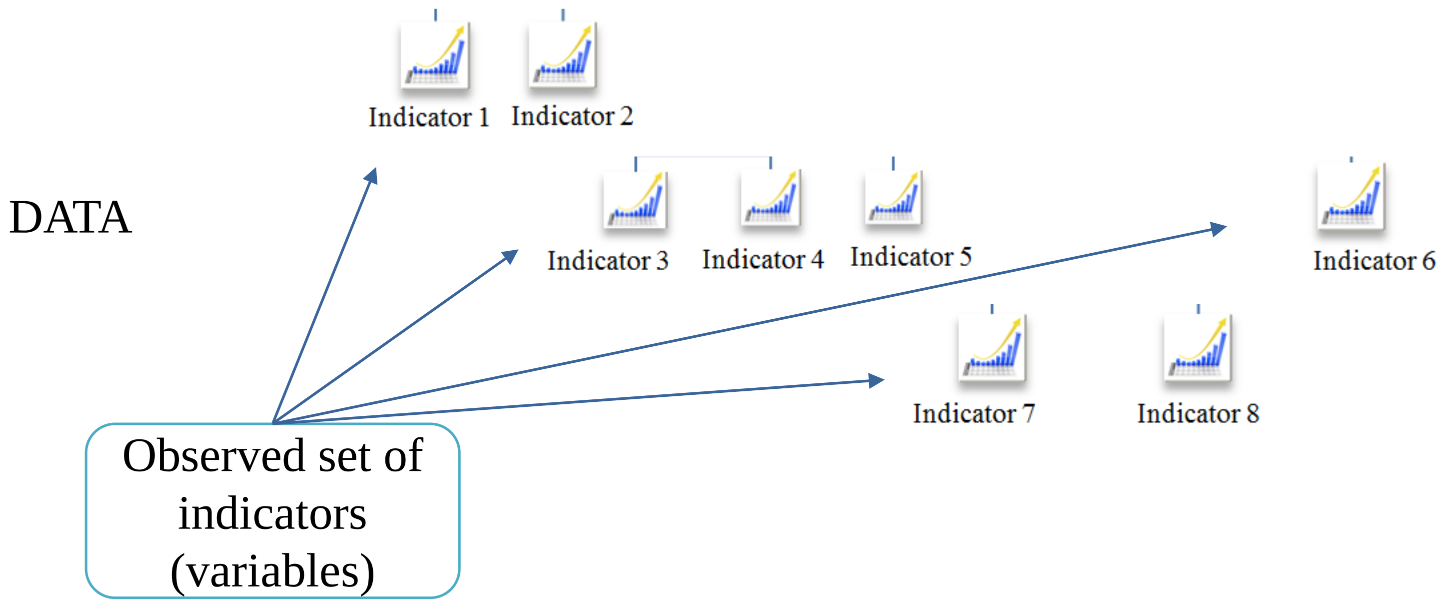
$$x_5 = a_{52}y_2 + \varepsilon_5$$

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

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

Hierarchical Model for Composite Indicators

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



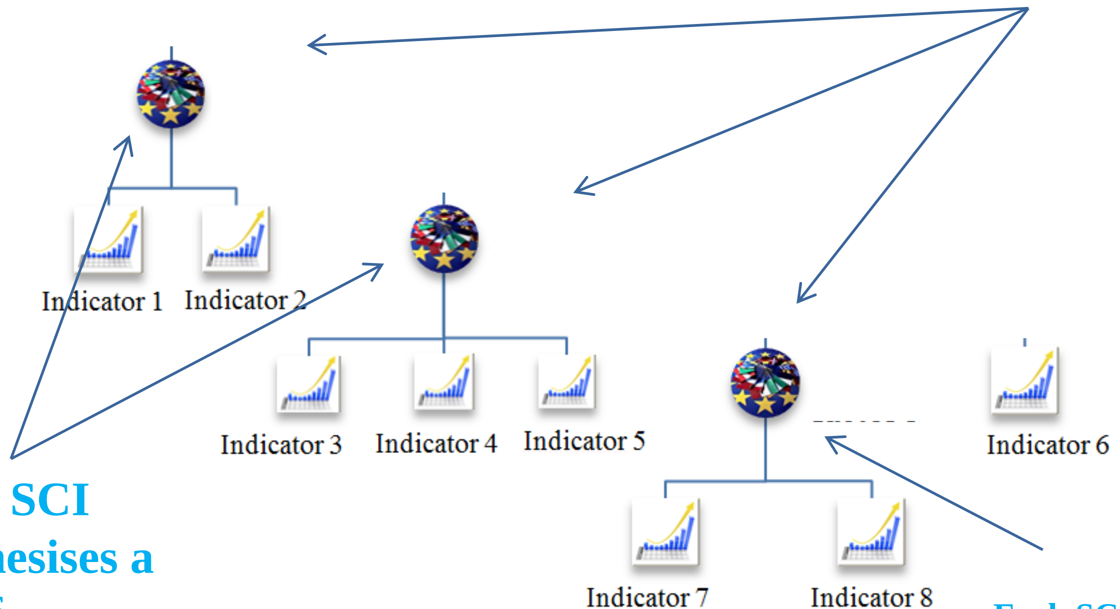
Specific Composite Indicators for detecting “specific concepts” (2/4)

First Level Synthesis

First Order Specific Composite Indicators

COMPOSITE INDICATOR

DATA

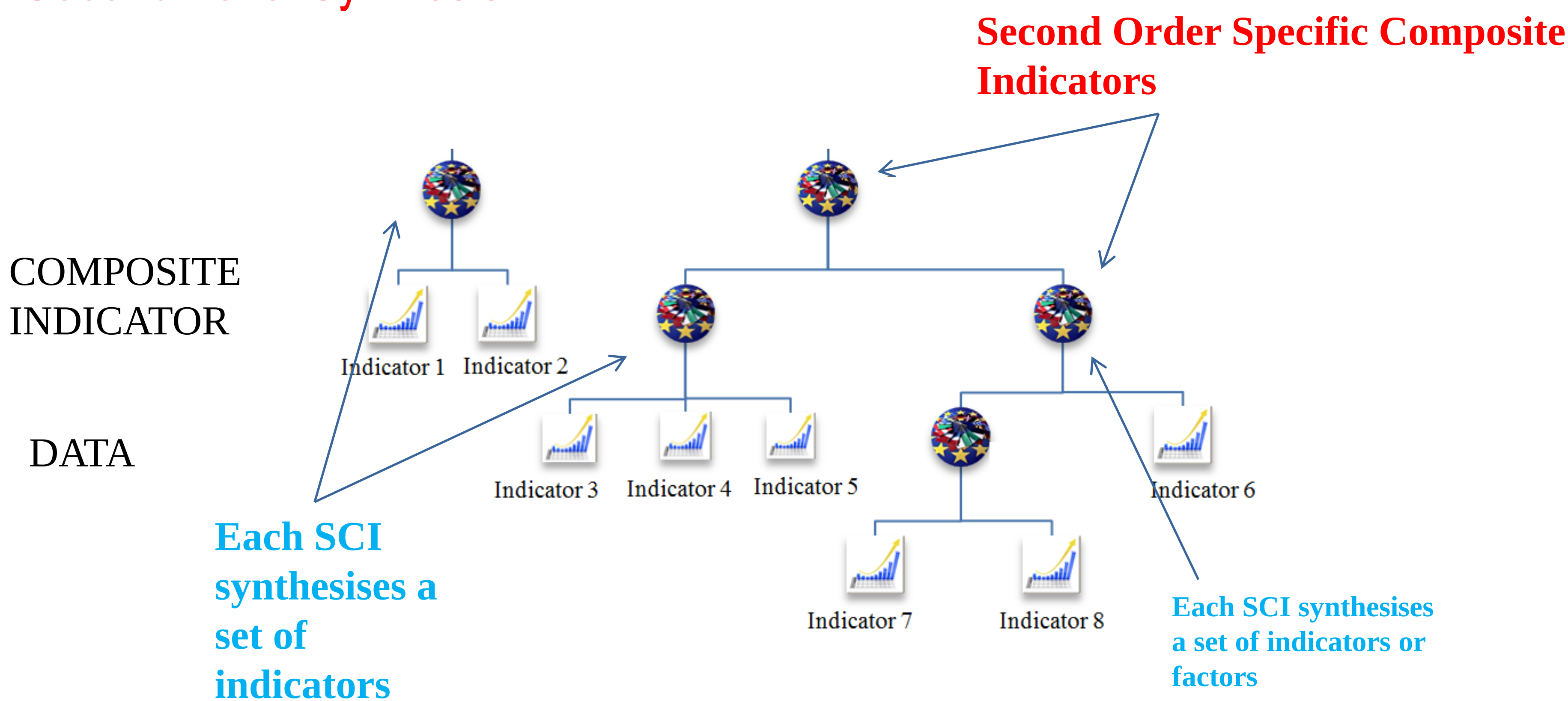


Each SCI synthesises a set of indicators

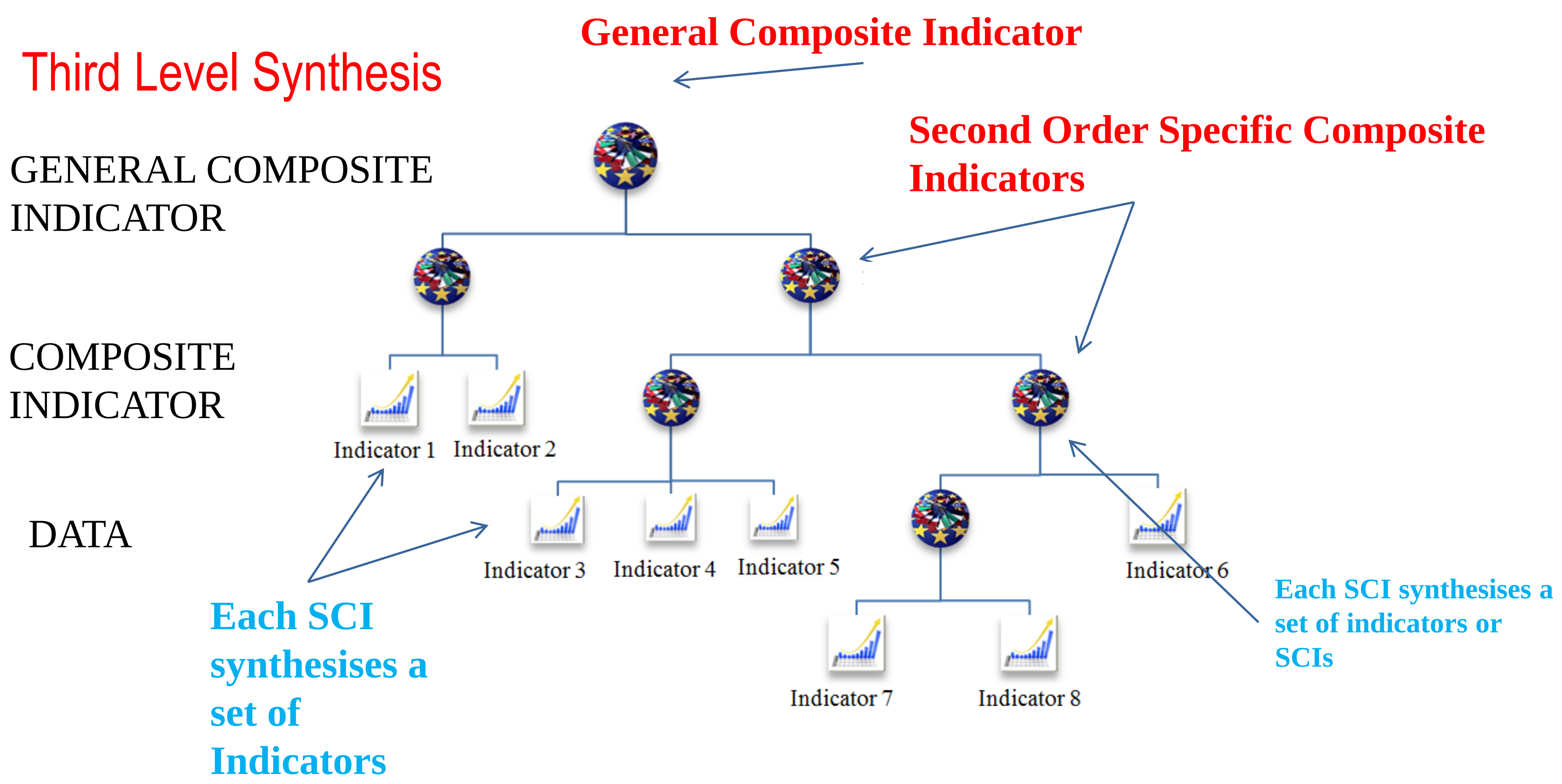
Each SCI synthesises a set of indicators

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

Second Level Synthesis

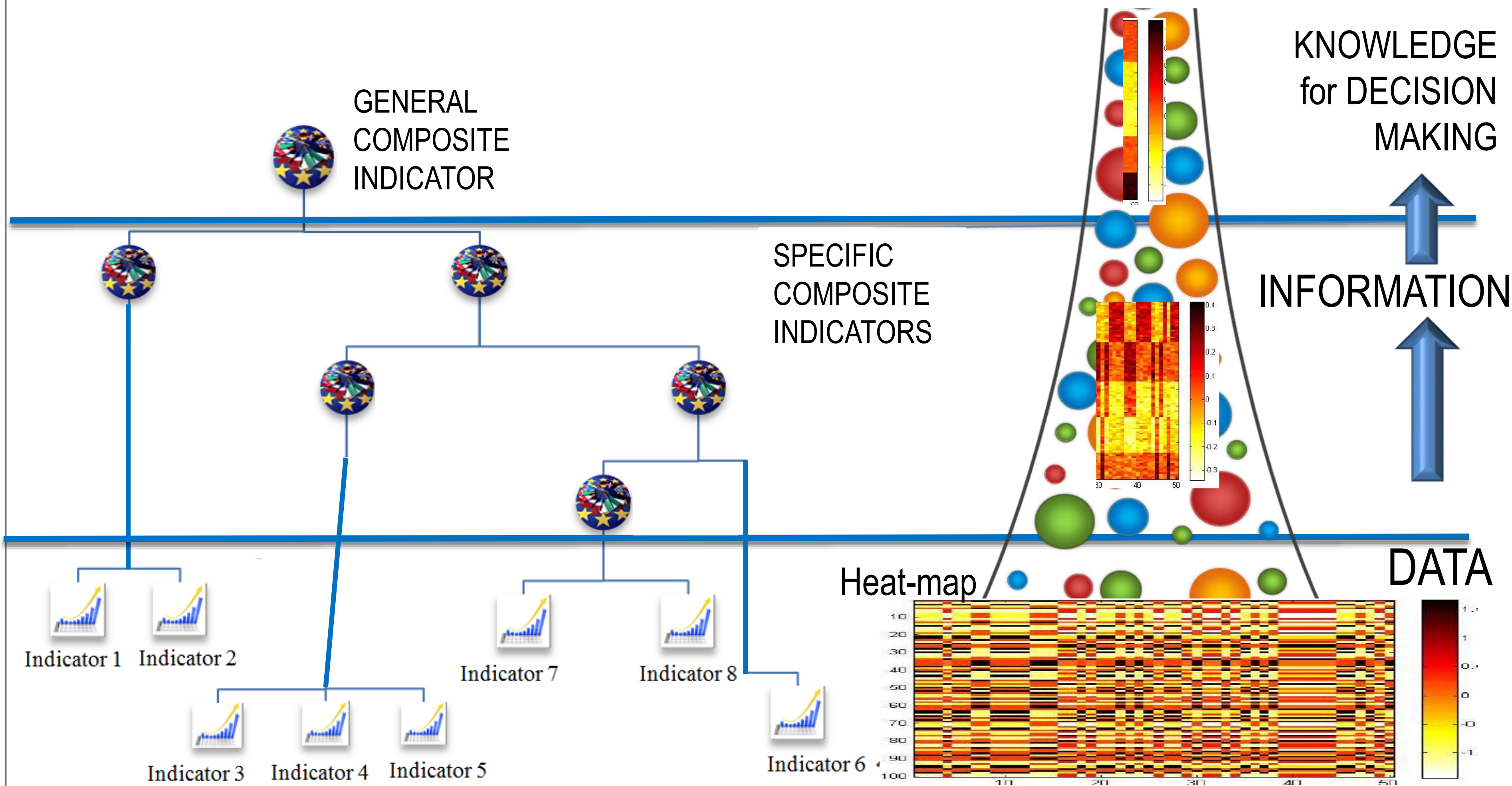


General Composite Indicator (for decision making) (4/4)



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:

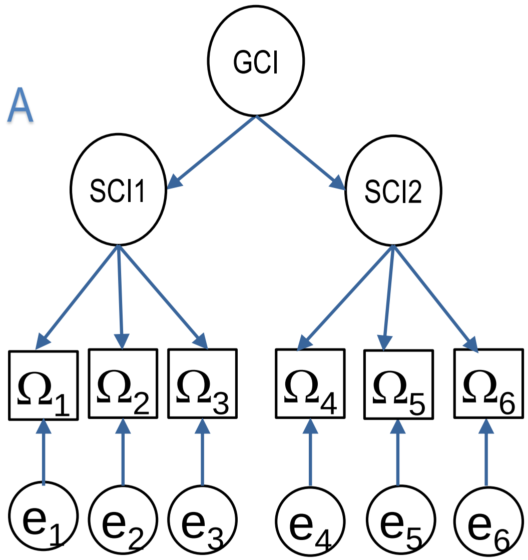
- **Confirmatory –** a Scientific Theory (ST) is assumed and has to be confirmed by the observed indicators;
- **Exploratory –** no clear ST is known, thus, regularities are searched in the data;
- **Mixed Confirmatory & Exploratory –** part of the ST is known, but it is not completely known

Which typology of relations between indicators:

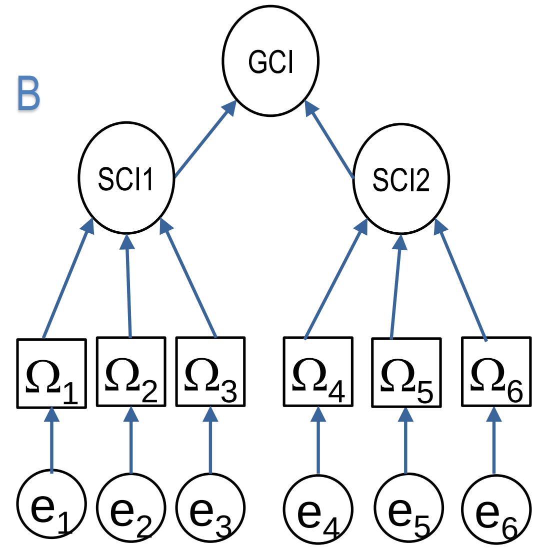
- **Reflective**
- **Formative**

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

A) Reflective
B) Formative



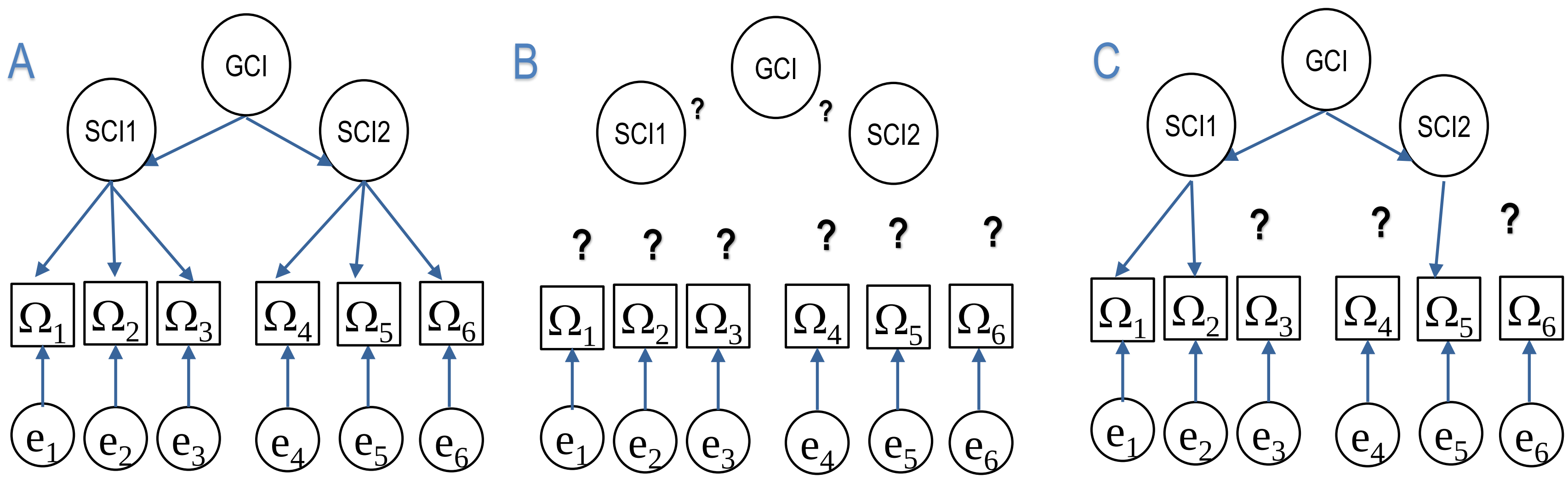
The General Composite Indicator is a determinant (causes) the Specific Composite Indicators & these last are determinant (causes) of the Manifest Indicators, i.e.,
The GCI reconstructs the SCIs that reconstruct the MI



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

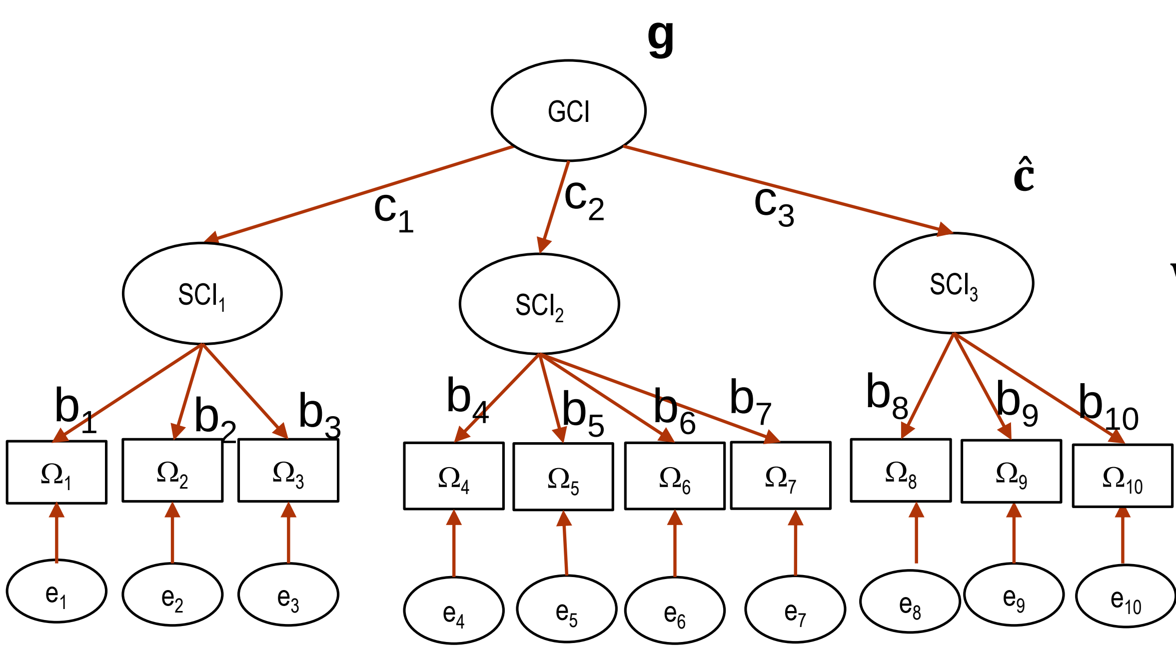
Hierarchical Composite Indicator (HCI)

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 MODEL ERROR

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



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

Weights for variables

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

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

Two-Leval Hierarchical Disjoint Factor Analysis

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

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

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 **Ψ**

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**

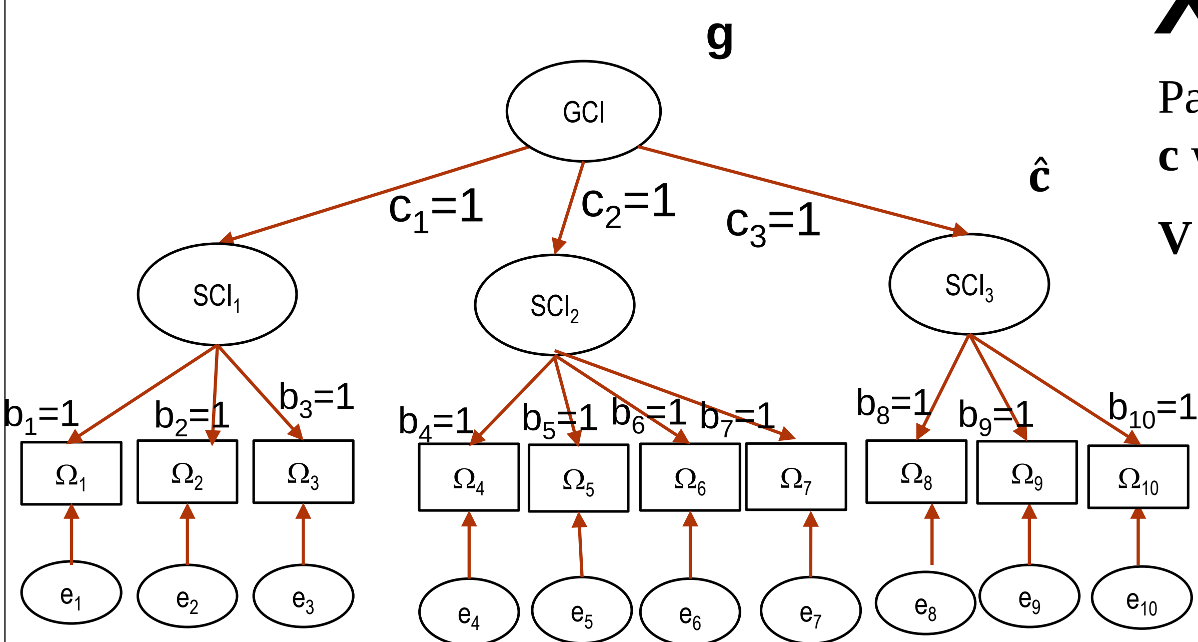
Special cases of HDFA(1/2)

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

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

Data MODEL ERROR

$$\mathbf{X} = \underbrace{\mathbf{gc}'\mathbf{V}'\mathbf{B}}_{\text{MODEL}} + \mathbf{E}$$



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 variables

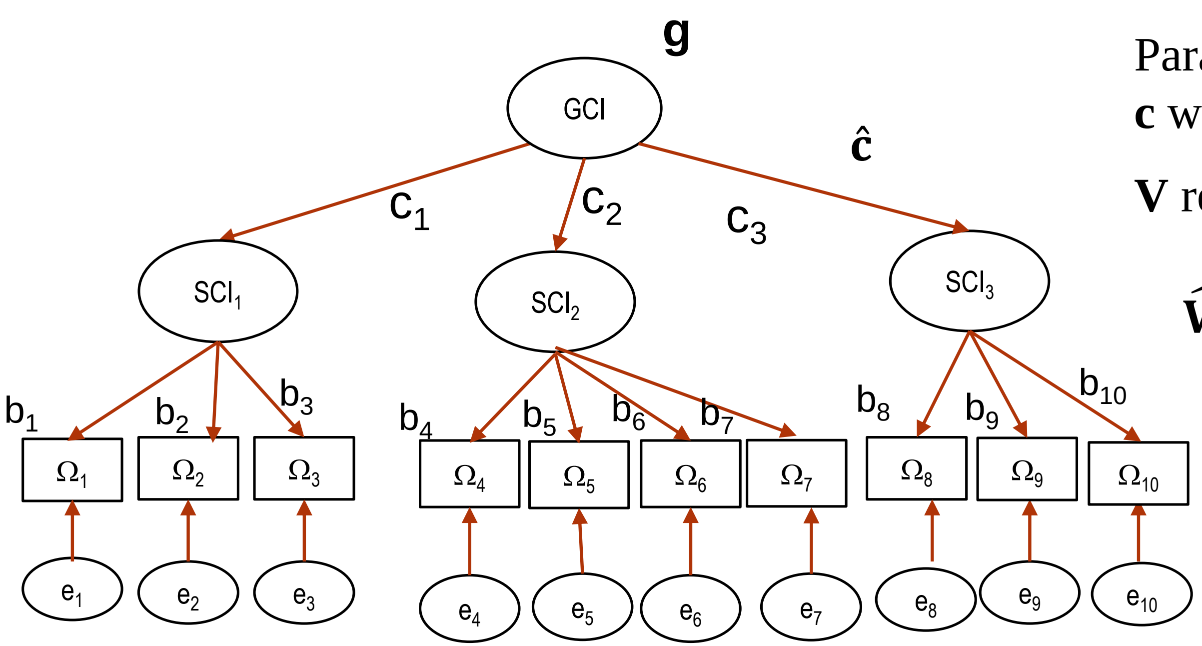
$$\hat{\mathbf{B}} = diag(\mathbf{b})$$

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

Special cases of HCI (2/2)

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

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



Data MODEL ERROR

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

Parameters of the model

c weights for SCIs

V relations between MIs & SCIs

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

Weights for MIs

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

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

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

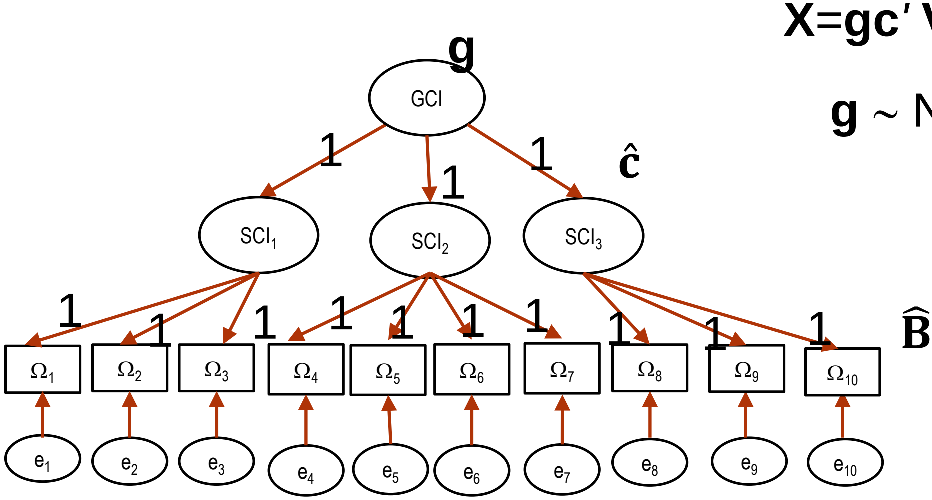
$-2\log Ol(\theta, \pi) + 2d$

BIC

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

Example 1 : Assessment of the Model-Based CI

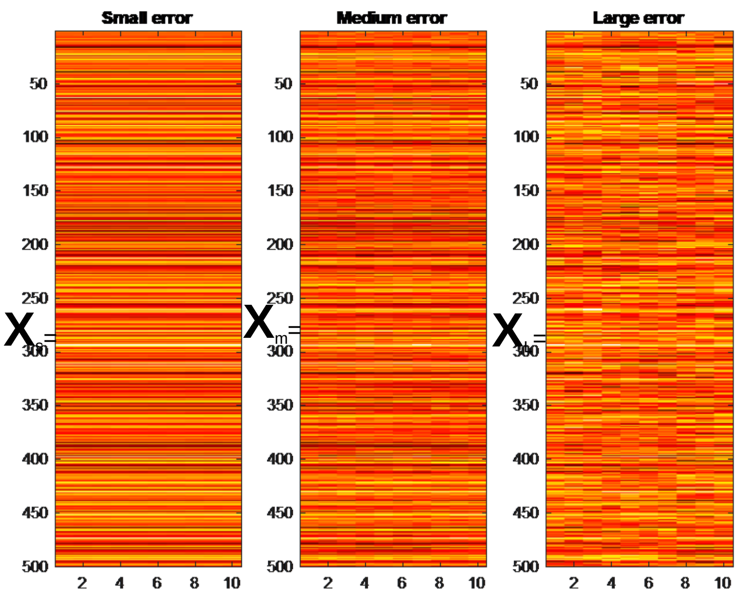
Case of ARITHMETIC MEAN



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

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

$\mathbf{X}_s=$ $\mathbf{X}_m=$ $\mathbf{X}_L=$



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

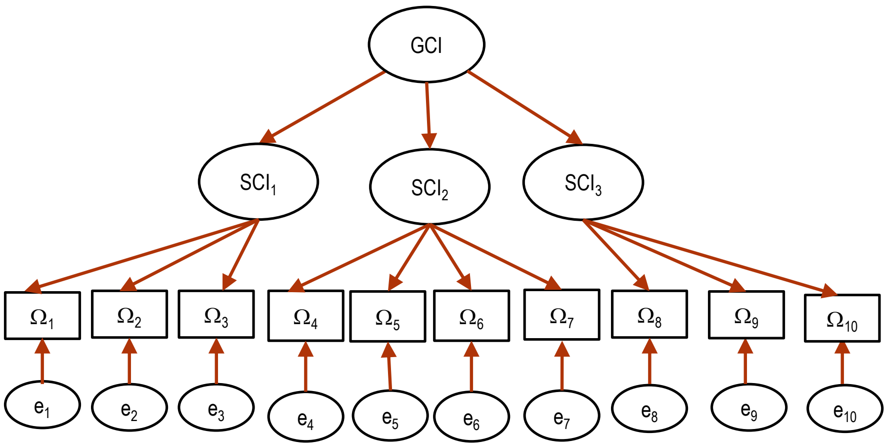
$$\hat{\mathbf{V}}'=\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

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

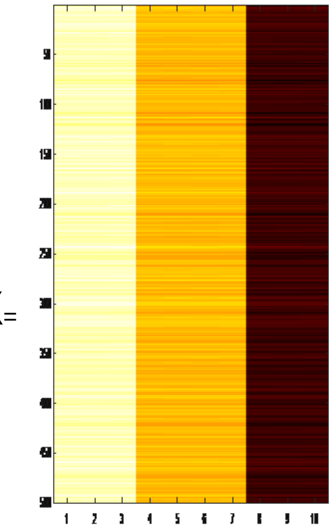
Arithmetic mean is a good GCI only when the MIs are very similar

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

Case of ARITHMETIC MEAN



X=



X divided in three blocks

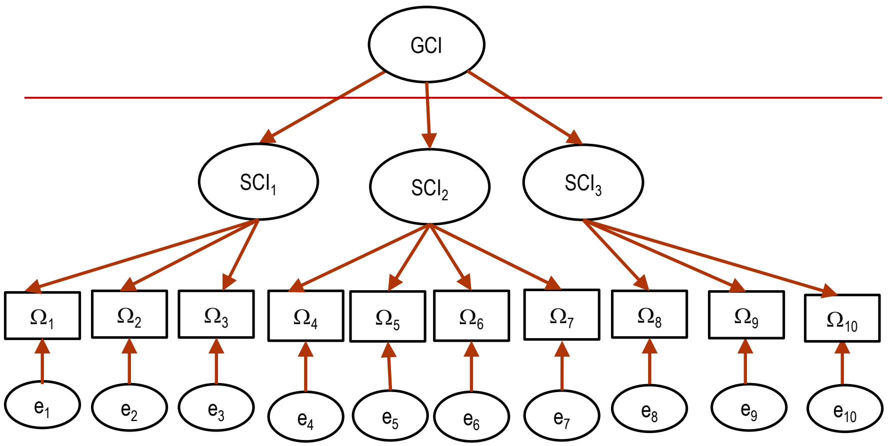
$$R^2_{\text{GCI}} = \frac{SS_{mod}}{SS_{tot}} = \mathbf{0.548}$$

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

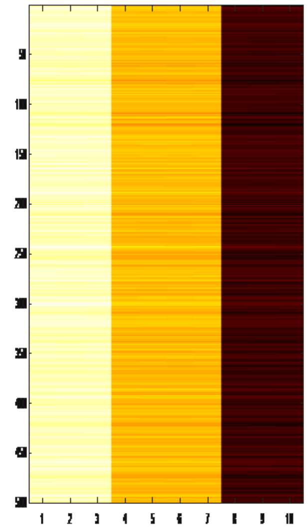
In a situation like this is better to stop at an intermediate level of synthesis (i.e., SCIs level)
because a GCI built as the arithmetic mean of MIs is not a good representation of
the phenomenon to describe

Example 2 : Assessment of the Model-Based CI

Case of ARITHMETIC MEAN



X=



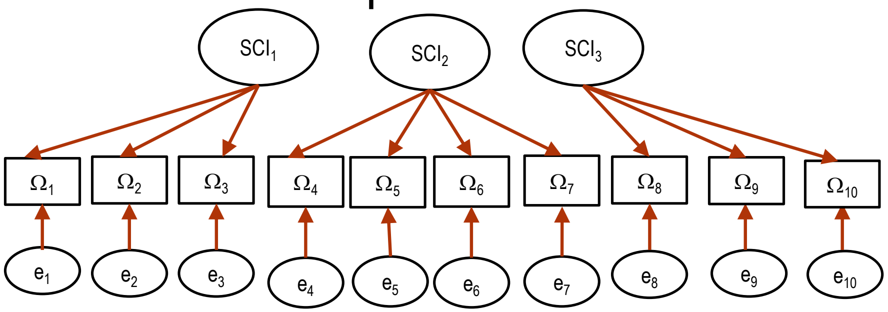
$$R^2_{\mathrm{GCI}} = \frac{SS_{mod}}{SS_{tot}} = \mathbf{0.548}$$

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

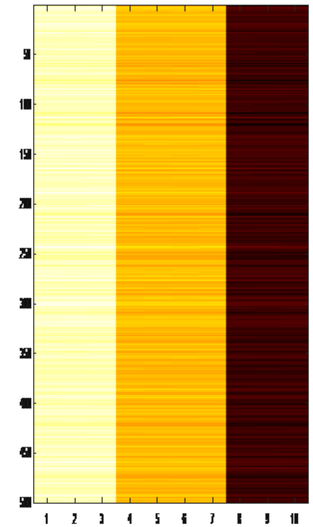
In a situation like this is better to stop at an intermediate level of synthesis (i.e., SCIs level)
because a GCI built as the arithmetic mean of MIs is not a good representation of
the phenomenon to describe

Example 2 : 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

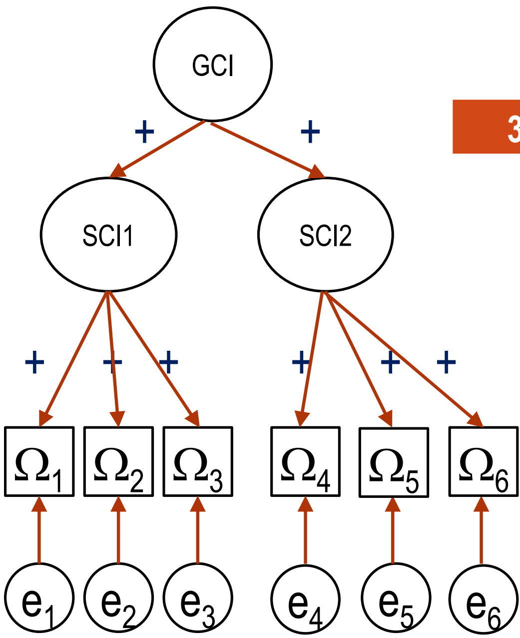
Data are normalized in order to allow the comparison and the combination of the MIs into the SCIs and GCI.

- Standardization $\mathbf{Z} = \mathbf{JX}\mathrm{diag}(\mathrm{dg}(\mathbf{\Sigma_X}))^{-1/2}$ 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 $\mathbf{Z} = \mathbf{JX}\mathrm{diag}(\boldsymbol{\mu}_\mathbf{X})^{-1}$ 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.



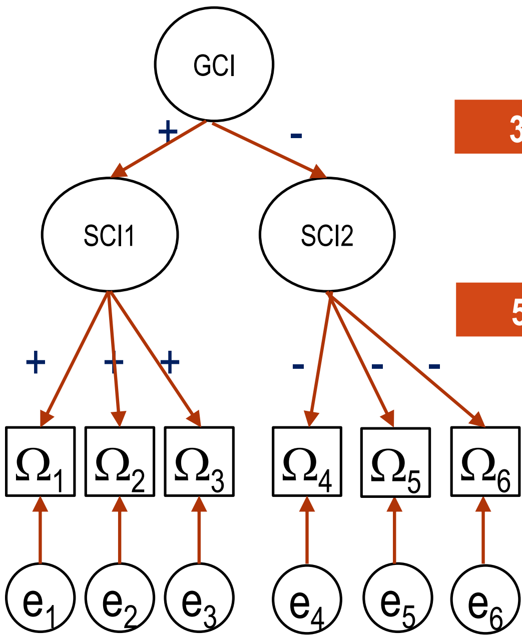
Ranking of the not compensated model:

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

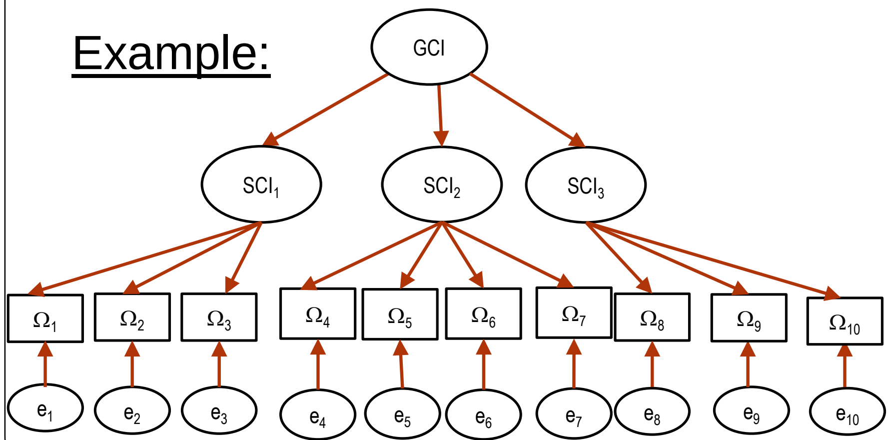
So **non-negativity** and non-compensability are strictly connected.

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.

<u>Example:</u>



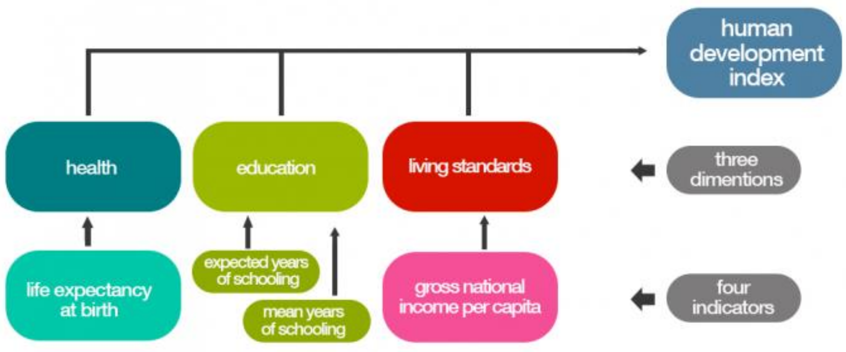
	Factor 1	Factor 2
Unidimensionality	2.737	0.556
Reliability	0.526	0.476

	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):



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

Based on the above informations:

- **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= $\sqrt[3]{LEI * EI * II}$

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

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.

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

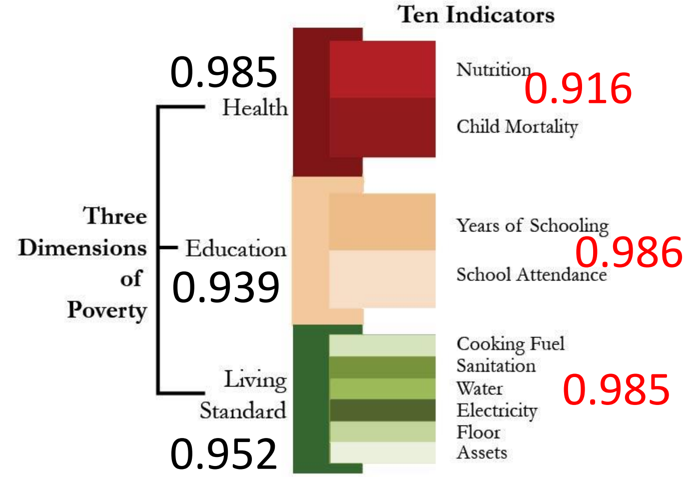
The increase of the 27% of R^2_{HDI} with respect to $R^2_{\text{MinMax\_HDI}}$ has to be imputed to the use of different normalisations. Therefore, it is important to understand that different normalizations of the MIs must be strongly motivated.

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

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

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

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

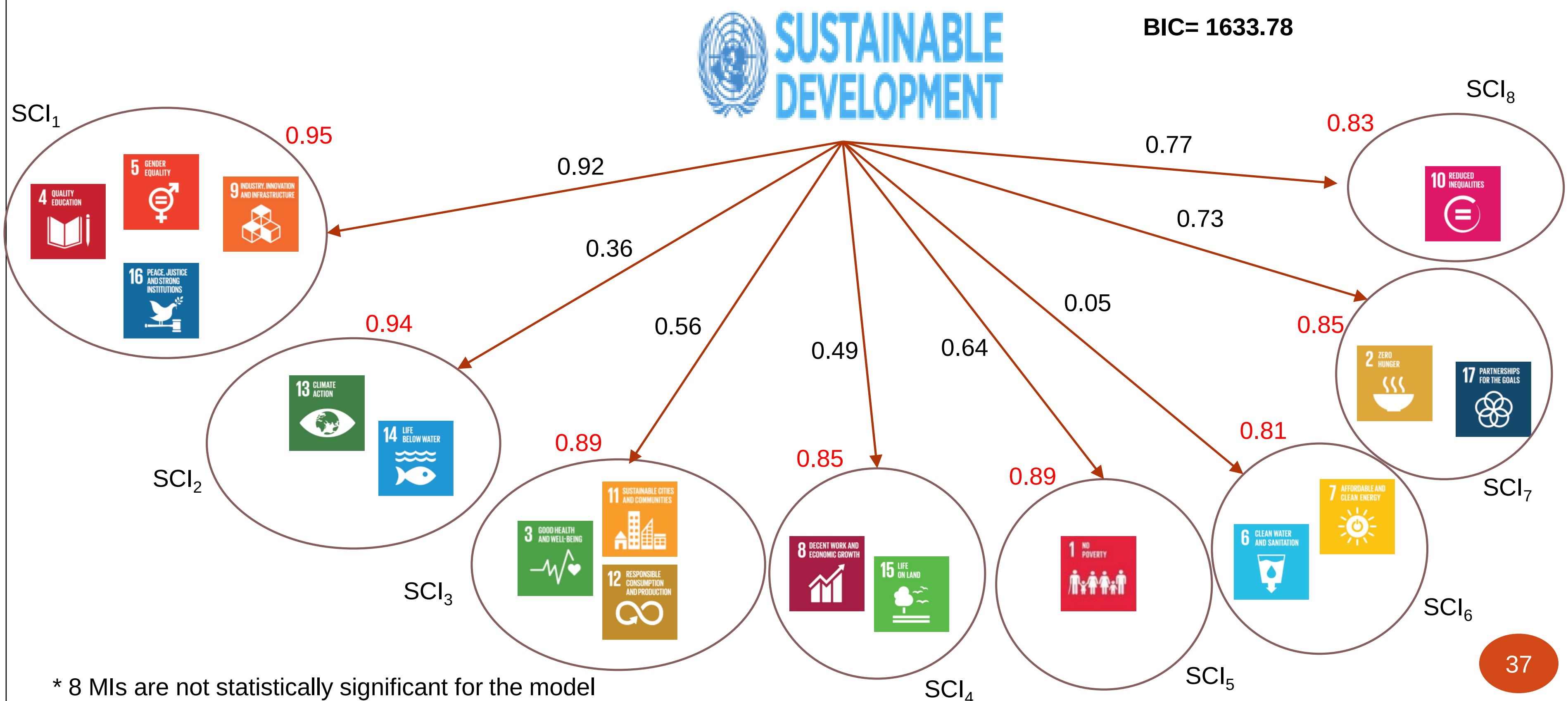
ASSESSMENT of HCI model: 17 goals

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

$$\mathbf{X} = \mathbf{U}_1\mathbf{M}_{11}\mathbf{V}_1'\mathbf{B}_1 + \mathbf{E}_1,$$

$$\mathbf{X} = \mathbf{U}_2\mathbf{M}_{22}\mathbf{V}_2'\mathbf{B}_2 + \mathbf{E}_2,$$

$$\ldots\ \ldots\ \ldots$$

$$\mathbf{X} = \mathbf{U}_Q\mathbf{M}_{QQ}\mathbf{V}_Q'\mathbf{B}_Q + \mathbf{E}_Q,$$

$$\ldots\ \ldots\ \ldots$$

$$\mathbf{X} = \mathbf{U}_K\mathbf{M}_{KQ}\mathbf{V}_Q'\mathbf{B}_Q + \mathbf{E}_K,$$

$$\ldots\ \ldots\ \ldots$$

$$\mathbf{X} = \mathbf{U}_n\mathbf{M}_{nJ}\,\mathbf{V}_J'\,\mathbf{B}_Q + \mathbf{E}_n, \qquad (1)$$

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

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

$\mathbf{V}_q = [v_{jhk} \in \{0, 1\} : j=1,\ldots,J,\ p=1,\ldots,q]\ ,\quad Q=2,\ldots,J-1$ 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 $\mathbf{V}$.

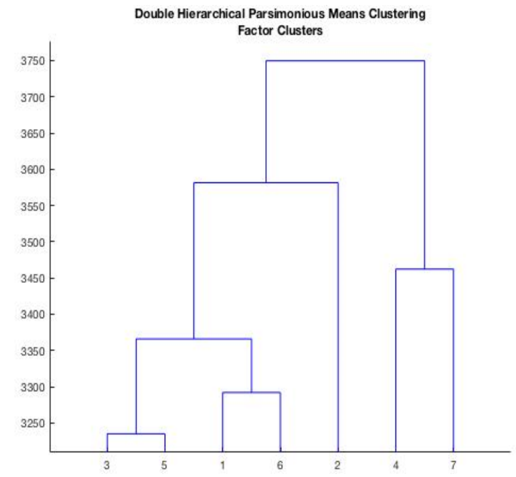
Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (2/2)

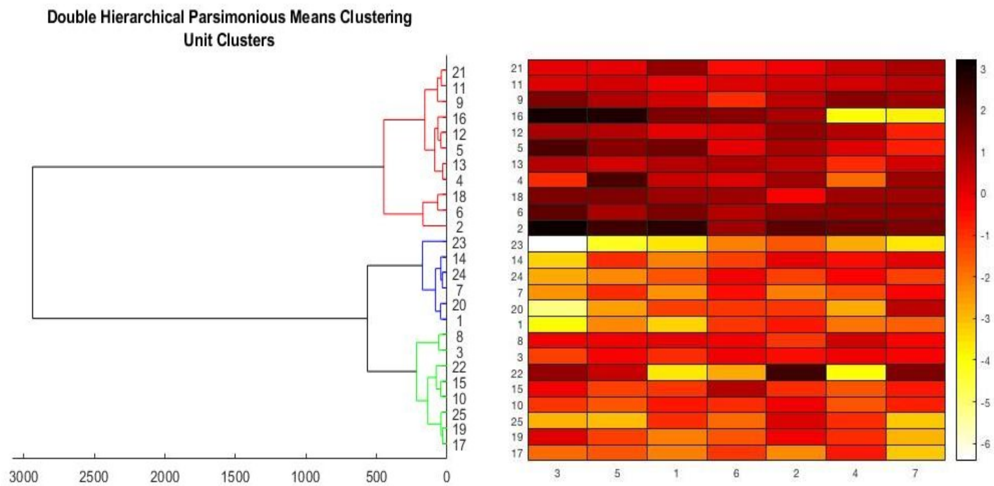
APPLICATION (ECSI DATA)

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).





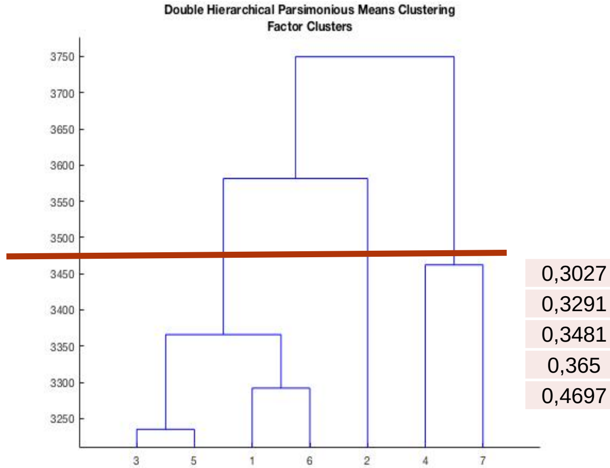
Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (2/2)

APPLICATION (ECSI DATA)

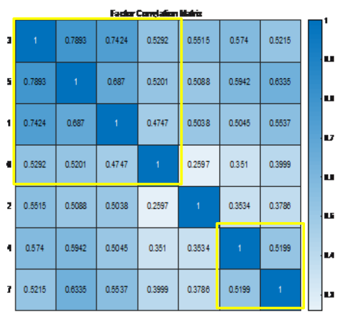
European Consumer Satisfaction Index: ECSI approach in mobile phone industry.
The dataset contains 250 units and 24 variables.
We supposed to have 7 interrelated latent variables, as follows:

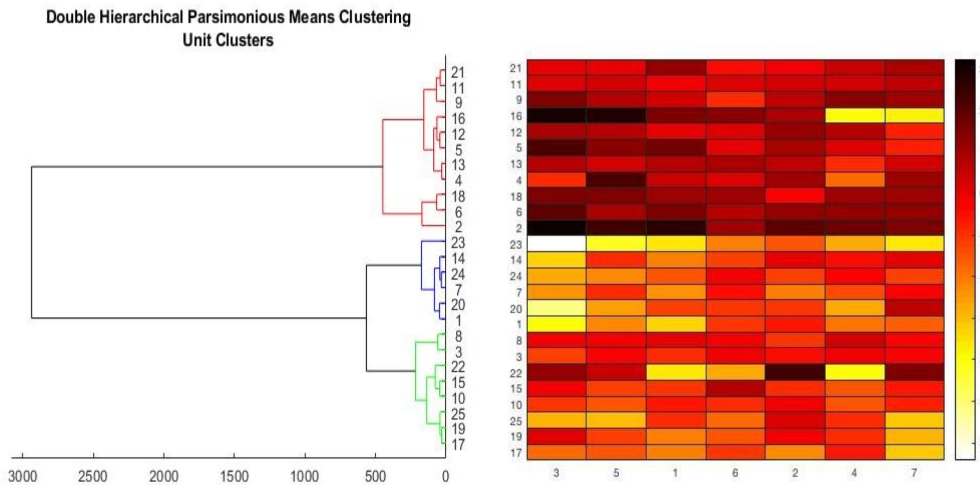
1. **Image** related to manifest variables from 1 to 5.
2. **Expectations** related to manifest variables from 6 to 8.
3. **Perceived Quality** related to manifest variables from 9 to 15.
4. **Perceived Value** related to manifest variables 16 and 17.
5. **Satisfaction** related to manifest variables 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).



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





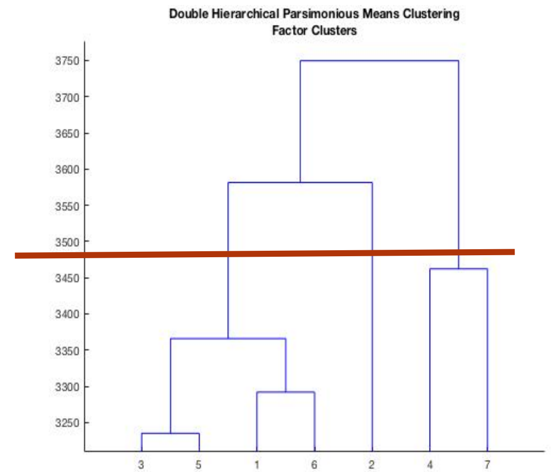
Double Hierarchical Means Clustering Model (Cavicchia, Vichi, Zaccaria, 2018) (2/2)

APPLICATION (ECSI DATA)

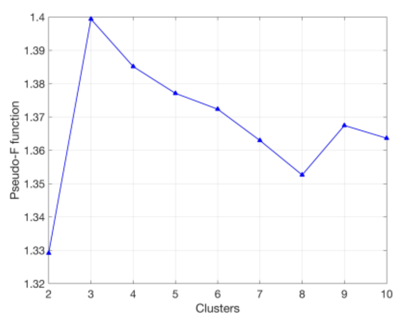
European Consumer Satisfaction Index: ECSI approach in mobile phone industry.
The dataset contains 250 units and 24 variables.
We supposed to have 7 interrelated latent variables, as follows:

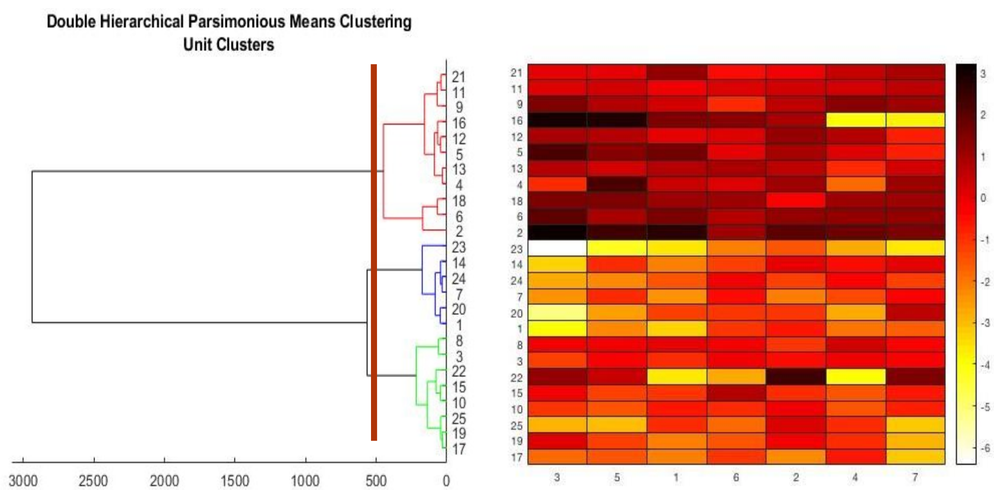
1. **Image** related to manifest variables from 1 to 5.
2. **Expectations** related to manifest variables from 6 to 8.
3. **Perceived Quality** related to manifest variables from 9 to 15.
4. **Perceived Value** related to manifest variables 16 and 17.
5. **Satisfaction** related to manifest variables 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