SEM: historical corner

Structural Equation Modeling (SEM)

- Structural Equation Models (SEM) are complex models allowing us to study real world complexity by taking into account a whole number of causal relationships among latent concepts (i.e. the Latent Variables, LVs), each measured by several observed indicators usually defined as Manifest Variables (MVs).
- Factor analysis, path analysis and regression are special cases of SEM.
- SEM is a largely confirmatory, rather than exploratory, technique. It is used more to determine whether a model is valid than to find a suitable model. But some exploratory elements are allowed

Key concepts:

Latent variables (unobservable by a direct way): abstract psychological variables like «intelligence», «attitude toward the brand», «satisfaction», «social status», «ability», «trust».

Manifest variables are used to measure latent concepts and they contain sizable measurement errors to be taken into account: multiple measures are allowed to be associated with a single construct.

Measurement is recognized as difficult and error-prone: the **measurement error is explicitly modeled** seeking to derive unbiased estimates for the relationships between latent constructs.

Structural Equation Modeling (SEM)

Several fields played a role in developing Structural Equation Models :

- From **Psychology**, comes the belief that the measurement of a valid construct **cannot rely on a single measure**.
- From **Economics** comes the conviction that **strong theoretical specification** is necessary for the estimation of parameters.
- From **Sociology** comes the notion of **ordering theoretical variables** and decomposing types of effects.

Sewall Wright and Path Analysis

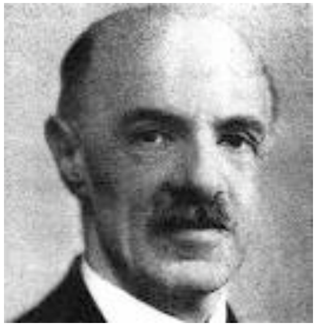
Sewall Wright (21 December 1889 –3 Mars 1988)
American Geneticist, son of the economist Philip Wright

Path Analysis has been developed in the 20s by S. Wright to investigate genetic problems and to help his father in economic studies.

Path Analysis aims to study cause-effect relations among several variables by looking to the correlation matrix among them.

The main newness is the introduction of a new tool to investigate cause-effect relations: the path diagram

Factor Analysis and the idea of Latent Variable



Charles Edward Spearman (10 September 1863 – 17 September 1945)

English psychologist

C. Spearman proposed Factor Analysis (FA) at the begin of the '900s to measure intelligence in a "objective" way.

The main idea is that intelligence is measured by several variables, but the correlation observed among the variables should be explained by a unique underlying "factor".

The most important input from Factor Analysis is the introduction of the concept of "factor", in other words the concept of Latent Variable

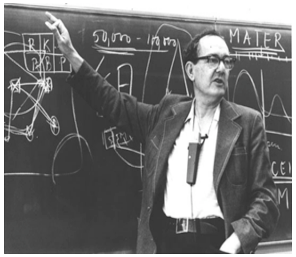
Thurstone and Multiple Factor Analysis

Spearman approach has been modified in the following 40 years in order to consider more than one factor as "cause" of observed correlation among several set of manifest variables

Louis Thurstone (29 May 1887–30 September 1955)
Psychometricien

the father of the Multiple Factor Analysis

Causal models rediscovered



Herbert Simon (June, 15 1916 – February 9, 2001)
Economist – Nobel Prize for economic in 1978

In 1954 presents a paper proving that "*under certain assumptions correlation is an index of causality*"

Hubert M. Blalock (23 Augut 1926 – 8 Febrary 1991)
Sociologist

In 1964 published the book "Causal Inference in Nonexperimental Research", in which he defines methods able to make causal inference starting from the observed covariance matrix. He faces the problem of assessing relations among variables by means of the inferential method.

They developed the SIMON-BLALOCK techinque

Path analysis and Causal models

Otis D. Duncan

(December, 2 1921– November, 16 2004)

He was one of the leading sociologists in the world. He introduces the Path Analysis of Wright's in Sociology.

In the mid-60's comes to the conclusion that there is no difference between the Path Analysis of Wright and the Simon-Blalock model.

With the economist (and econometricien) Arthur Goldberger he comes to the conclusion that **there is no difference between what was known in sociology as Path Analysis and simultaneous equations models commonly used in econometrics.**

Along with Goldberger he organizes a conference in 1970 in Madison (USA) where he invited Karl Jöreskog.

Covariance Structure Analysis and K. Jöreskog

Karl Jöreskog

Statistician, Professor Emeritus at Uppsala University, Sweden

In the late 50s, he started working with Herman Wold.
He discussed a thesis on Factor Analysis.

In the second half of the 60s, he started collaborating with O.D. Duncan and A. Goldberger. This collaboration represents a meeting between Factor Analysis (and the concept of latent variable) and Path Analysis (i.e. the idea behind causal models).

In 1970, at a conference organized by Duncan and Goldberger, Jöreskog presented the Covariance Structure Analysis (CSA) for estimating a linear structural equation system, later known as LISREL

Soft Modeling and H. Wold

Herman Wold (December 25, 1908 – February 16, 1992)
Econometrician and Statistician

In the 50's Thurston meets Herman Wold meets Louis Thurstone. They decide to co-organize "the Upspsala Symposium on Psycological Factor Analysis". Since then, H. Wold started working on Latent Variables models.

In 1975, H. Wold extended the basic principles of an iterative algorithm aimed to the estimation of the PCs (NIPALS) to a more general procedure for the estimation of relations among several blocks of variables linked by a network of relations specified by a path diagram.

The PLS Path Modeling was proposed to estimate Structural Equation Models (SEM) parameters, as a Soft Modeling alternative to Jöreskog's Covariance Structure Analysis

Two families of methods

Covariance-based Methods

SEM

Component-based Methods

The aim is to reproduce the sample covariance matrix of the manifest variables by means of the model parameters:

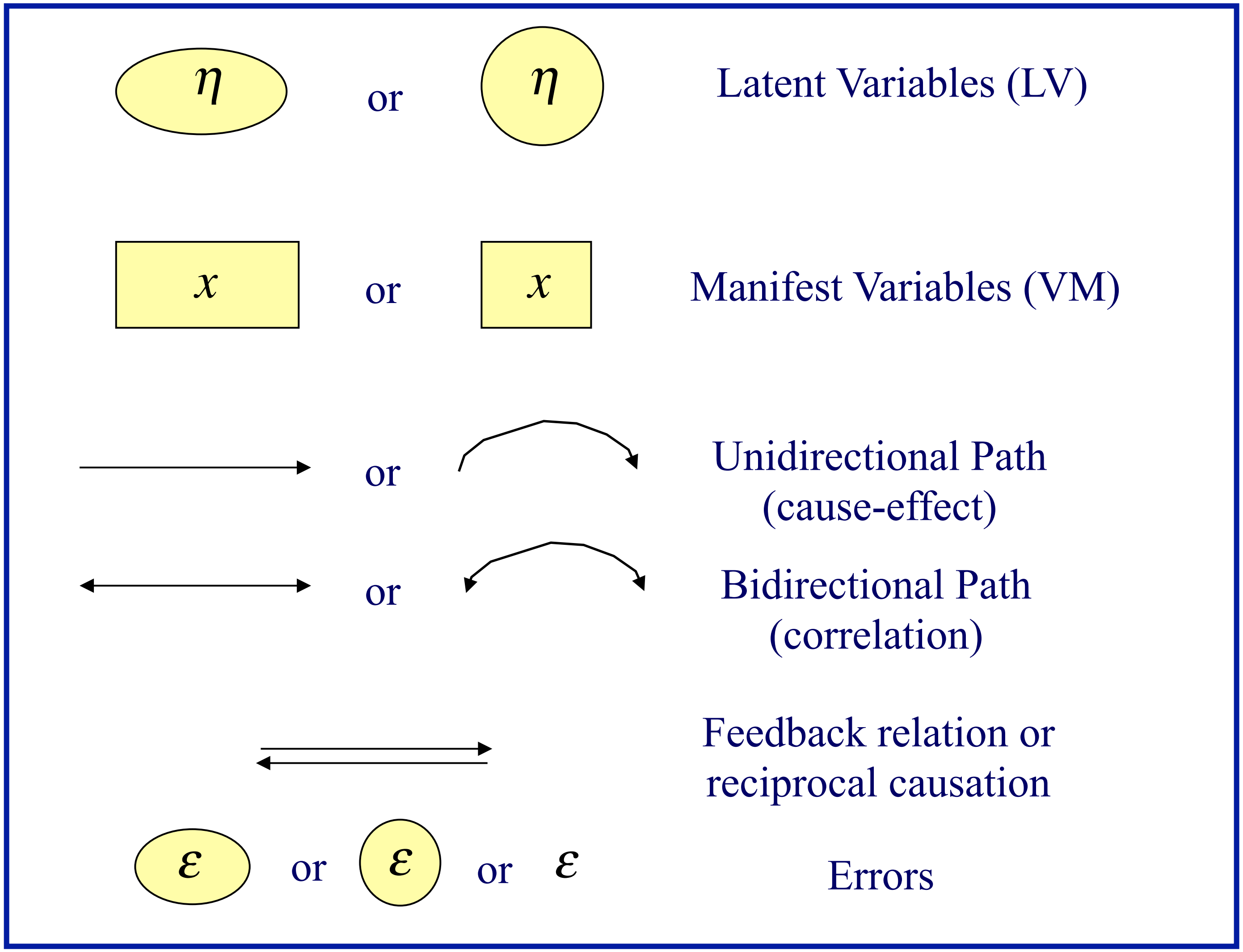
- the implied covariance matrix of the manifest variables is a function of the model parameters
- it is a confirmatory approach aimed at validating a model (theory building)

The aim is to provide an estimate of the latent variable scores in such a way that they are the most correlated with one another (according to path diagram structure) and the most representative of each corresponding block of manifest variables.

- latent variable score estimation plays a main role
- it is more an exploratory approach, than a confirmatory one (operational model strategy)

From Path Analysis to SEM

SEM: drawing conventions



Structural Equation models: notation

Greek characters are used to refer to Latent Variables:

$$\xi_i = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_J \end{bmatrix}'$$

J= # exogenous Latent Variables (LV)

$$\eta_i = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_M \end{bmatrix}'$$

M= # endogenous Latent Variables (LV)

Latin characters refer to Manifest Variables

$$\mathbf{x}_i = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}'$$

P= # exogenous MV

$$\mathbf{y}_i = \begin{bmatrix} y_1 \\ \vdots \\ y_Q \end{bmatrix}'$$

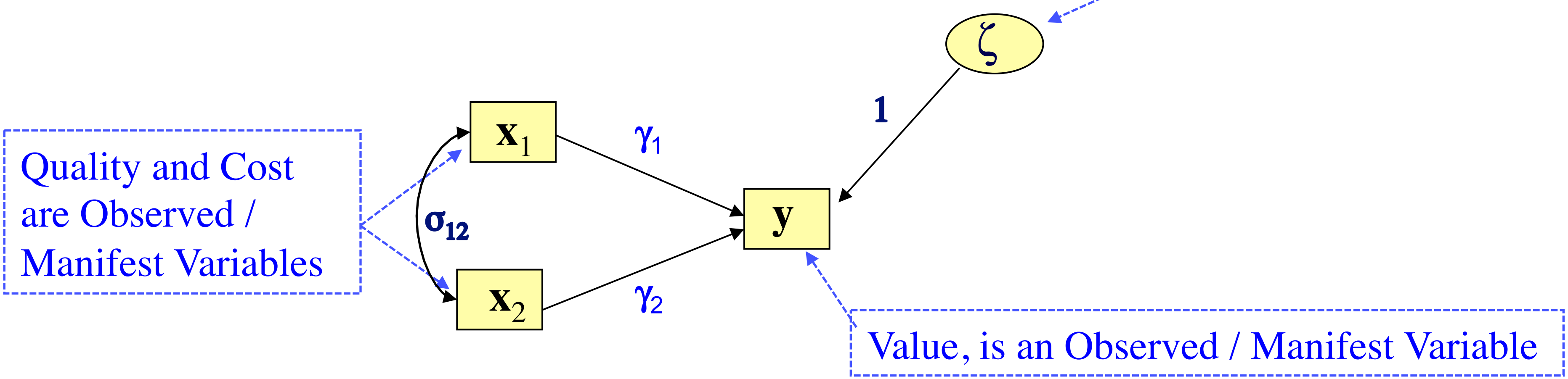
Q= # endogenous MV

“Drawing” a regression model

The multiple regression model (on centred variables) :

$$y = \beta_1 x_1 + \beta_2 x_2 + \zeta$$

can be “drawn” by using a Path Diagram:



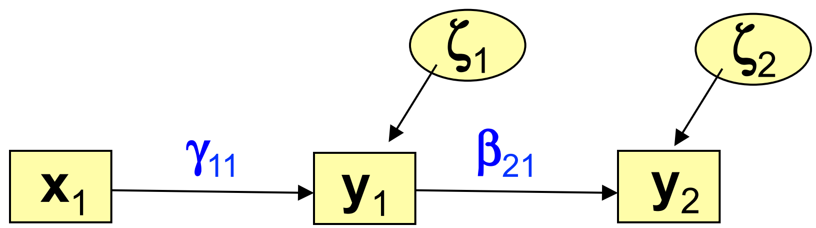
Example: The Value for a brand in terms of Quality and Cost

Path Models with Manifest Variables

The multiple regression model can be generalized to paths where endogenous variables are on their turn causative of others endogenous variables.

$$y_1 = \gamma_{11}\, x_1 + \zeta_1$$

$$y_2 = \beta_{21}\, y_1 + \zeta_2$$



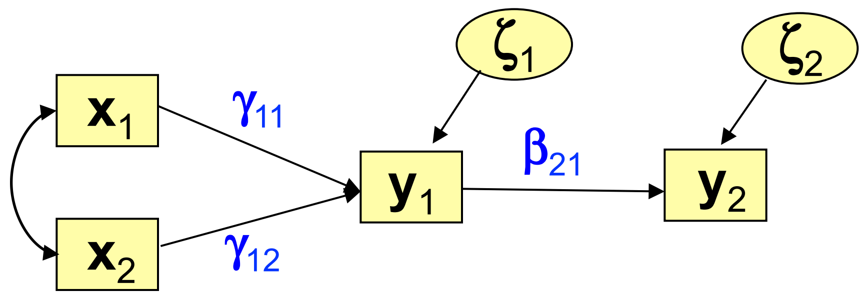
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad \Gamma = \begin{bmatrix} \gamma_{11} \\ 0 \end{bmatrix} \quad \mathbf{x} = [x_1] \quad \mathrm{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \quad \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

endo # endo # exo # exo # endo # endo # endo

$$\mathbf{y} = \Gamma\mathbf{x} + \mathrm{B}\mathbf{y} + \zeta$$

Path Models with Manifest Variables

$$y_1 = \gamma_{11}\, x_1 + \gamma_{12}\, x_2 + \zeta_1$$
$$y_2 = \beta_{21}\, y_1 + \zeta_2$$

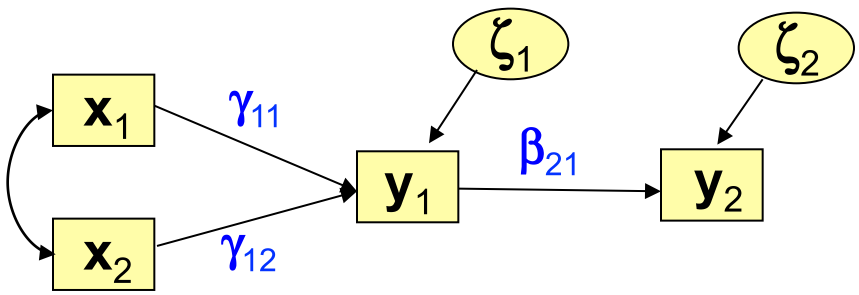


$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad \Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & 0 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathrm{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \quad \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

endo # endo # exo # exo # endo # endo # endo

$$\mathbf{y} = \Gamma\mathbf{x} + \mathbf{B}\mathbf{y} + \zeta \Leftrightarrow \mathbf{y} = \left(\mathbf{I} - \mathbf{B}\right)^{-1}\left(\Gamma\mathbf{x} + \zeta\right)$$

Analysing covariance structures of Path models



Population Covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Assuming that:

i) the MVs are centered

ii) Two structural errors do not covariate

iv) The covariance between structural error and exogenous MVs is equal to zero

We can write the covariance matrix among the MVs in terms of model parameters (**implied covariance matrix**):

$$C = \Sigma(\Omega) = \Sigma(\Gamma, \mathbf{B}, \Psi)$$

Path Coefficients

Structural Error Covariance

Path model implied covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Population Covariance matrix

$$C \approx S \approx \Sigma$$

$$S = \begin{bmatrix} s_{xx} & \\ s_{yx} & s_{yy} \end{bmatrix}$$

Empirical covariance matrix

$$C = \Sigma(\Omega) = \left[\begin{array}{c:c} \Sigma_{xx} & \\ \hdashline (\mathbf{I}-\mathbf{B})^{-1}\Gamma\Sigma_{xx} & (\mathbf{I}-\mathbf{B})^{-1}(\Gamma\Sigma_{xx}\Gamma' + \Psi)(\mathbf{I}-\mathbf{B})^{-1\prime} \end{array}\right]$$

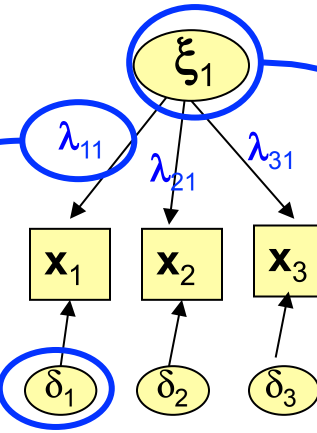
"Implied" covariance matrix

Confirmatory Single Factor Model

$$x_1 = \lambda_{11}\,\xi_1 + \delta_1$$
$$x_2 = \lambda_{21}\,\xi_1 + \delta_2$$
$$x_3 = \lambda_{31}\,\xi_1 + \delta_3$$

Factor Loading

Unique Factor



Common Factor

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \Lambda_x = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \end{bmatrix} \quad \xi = [\xi_1] \quad \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

$$\boxed{\mathbf{x} = \Lambda_x \xi + \delta}$$

Multiple Confirmatory Factor Model

$x_1 = \lambda_{11}\,\xi_1 + \delta_1$,

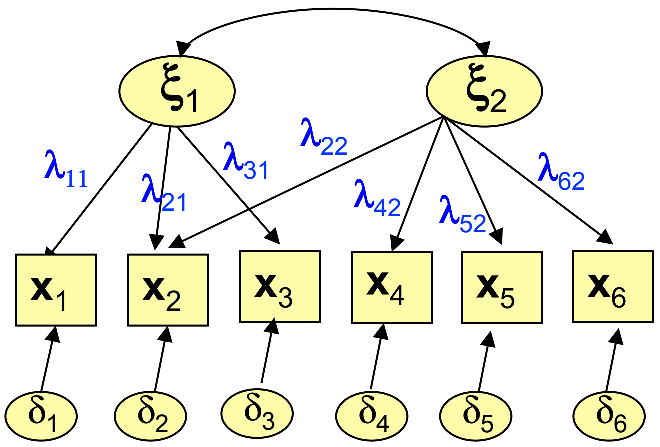
$x_2 = \lambda_{21}\,\xi_1 + \lambda_{22}\,\xi_2 + \delta_2$

$x_3 = \lambda_{31}\,\xi_1 + \delta_3$

$x_4 = \lambda_{42}\,\xi_2 + \delta_4$

$x_5 = \lambda_{52}\,\xi_2 + \delta_5$

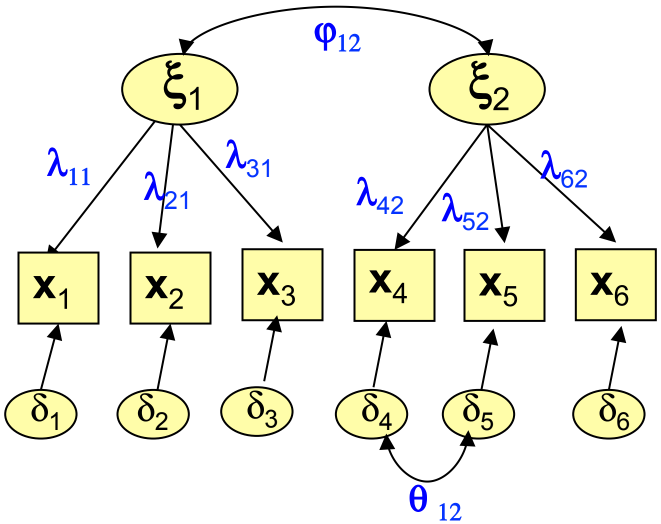
$x_6 = \lambda_{62}\,\xi_2 + \delta_6$



$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_6 \end{bmatrix} \quad \Lambda_x = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \quad \xi = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \quad \delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_6 \end{bmatrix}$$

$$\mathbf{x} = \Lambda_x \xi + \delta$$

Analysing covariance structures in CF models



Assuming that:

i) the MVs, the LV and the errors are centered

ii) Measurement errors and LVs do not covariate

We can write the covariance matrix among the MVs in terms of model parameters (**implied covariance matrix**):

$$C = \Sigma(\Omega) = \Sigma(\Lambda, \Phi, \Theta)$$

Loadings LV Covariance Measurement Error Covariance

Analysing covariance structures in CF models

Confirmative factor model :

$$\mathbf{x} = \Lambda_x \xi + \delta$$

The covariance matrix of the MVs can be rewritten in terms of model parameters:

$$\Sigma(\Omega) = \mathrm{E}(\mathbf{xx}') = \mathrm{E}\left[(\Lambda\xi + \delta)(\Lambda\xi + \delta)'\right] =$$

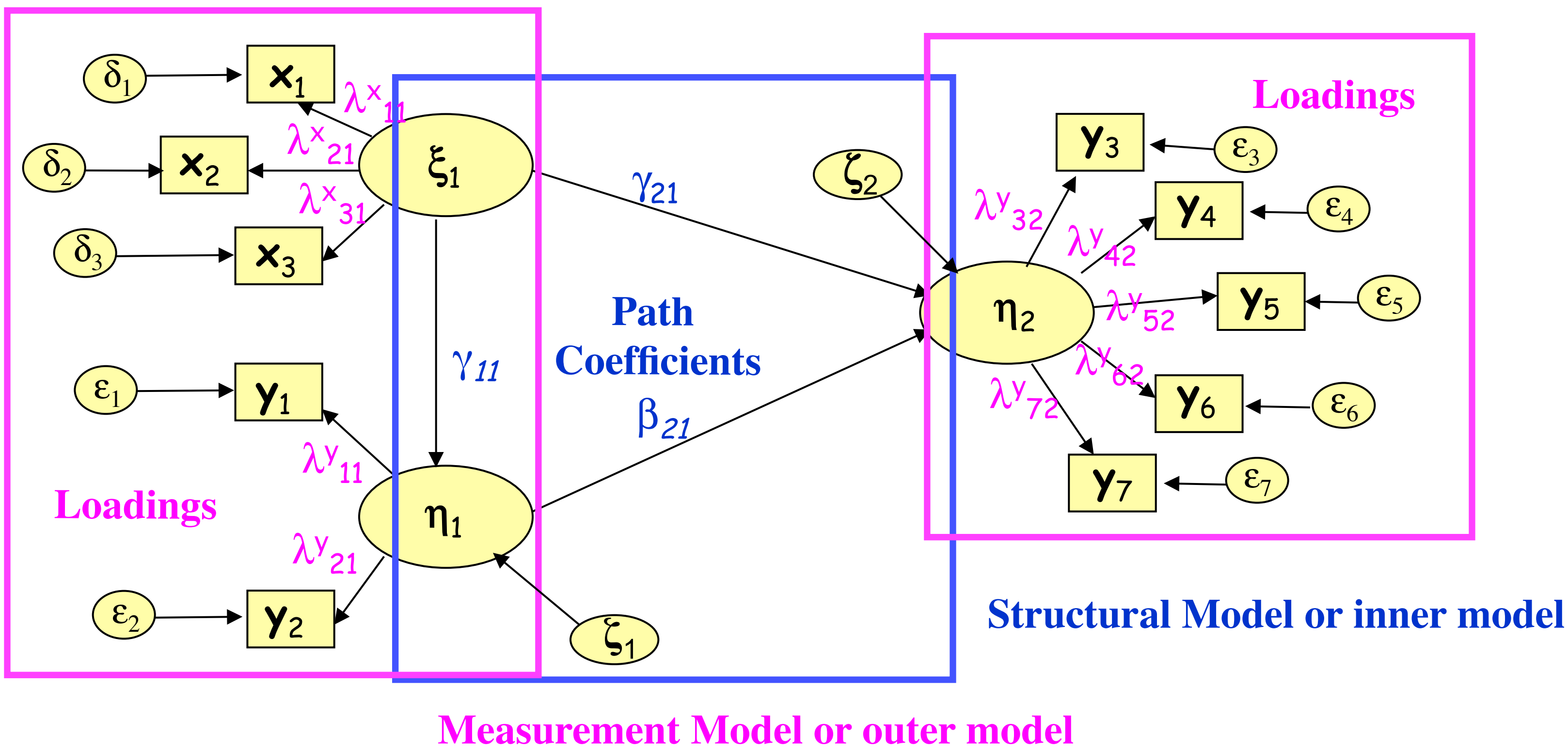
$$= \mathrm{E}\left[(\Lambda\xi + \delta)(\xi'\Lambda' + \delta')\right] =$$

$$= \Lambda\mathrm{E}(\xi\xi')\Lambda' + \Lambda\ \mathrm{E}(\xi\delta') + \mathrm{E}(\delta\xi')\Lambda' + \mathrm{E}(\delta\delta')$$

let $E(\xi\xi') = \Phi$; $E(\delta\delta') = \Theta$ $\Rightarrow$

$$\boxed{\Sigma(\Omega) = \Lambda\Phi\Lambda' + \Theta}$$

Path model with latent variables



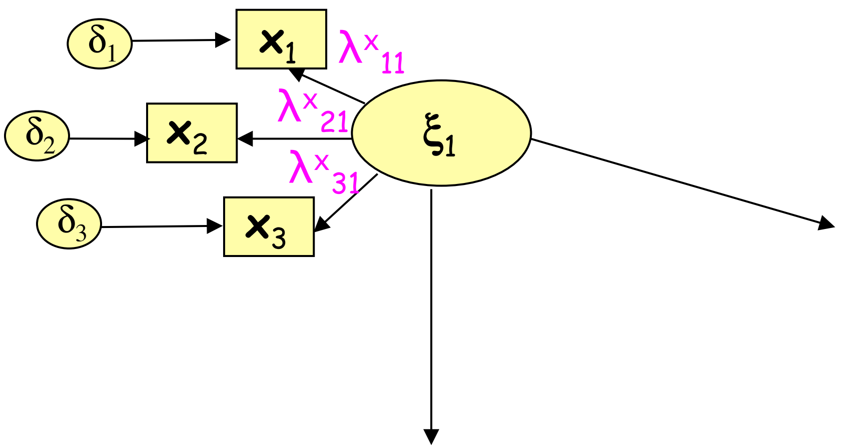
The measurement (outer) model

For the exogenous MVs

$x_1 = \lambda^x_{11}\,\xi_1 + \delta_1$

$x_2 = \lambda^x_{21}\,\xi_1 + \delta_2$

$x_3 = \lambda^x_{31}\,\xi_1 + \delta_3$



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \Lambda_x = \begin{bmatrix} \lambda^x_{11} \\ \lambda^x_{21} \\ \lambda^x_{31} \end{bmatrix} \quad \xi = [\xi_1] \quad \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

$$\boxed{\mathbf{x} = \Lambda_x \xi + \delta}$$

The measurement (outer) model

For the exogenous MVs (A reflective scheme must hold)

$$\mathbf{x} = \Lambda_x \xi + \delta$$

Vector of the exogenous MVs

RECTANGULAR matrix with the loadings linking each exogenous MV to the corresponding exogenous LV

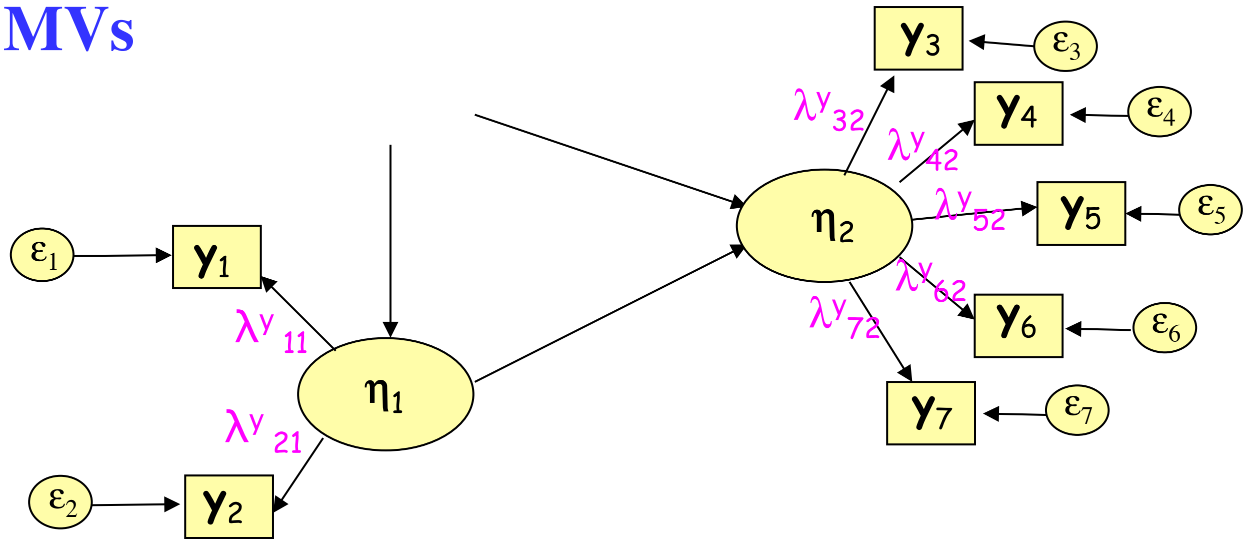
Vector of the exogenous LVs

Vector of the measurement errors associated to the exogenous MVs

The measurement (outer) model

For the endogenous MVs

$$y_1 = \lambda^{y_{11}}\eta_1 + \varepsilon_1$$
$$y_2 = \lambda^{y_{21}}\eta_1 + \varepsilon_2$$
$$y_3 = \lambda^{y_{32}}\eta_2 + \varepsilon_3$$
$$y_4 = \lambda^{y_{42}}\eta_2 + \varepsilon_4$$
$$y_5 = \lambda^{y_{52}}\eta_2 + \varepsilon_5$$
$$y_6 = \lambda^{y_{62}}\eta_2 + \varepsilon_6$$
$$y_7 = \lambda^{y_{72}}\eta_2 + \varepsilon_7$$



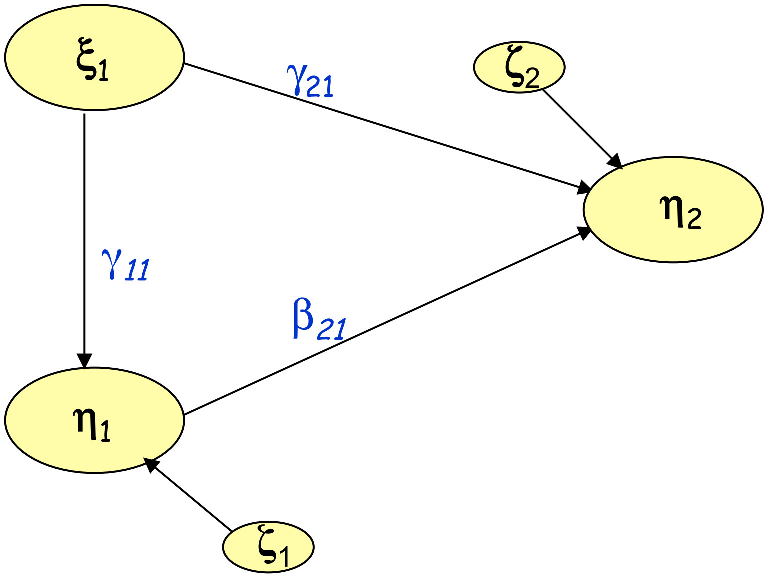
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} \qquad \Lambda_y = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ 0 & \lambda_{32} \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \\ 0 & \lambda_{72} \end{pmatrix} \qquad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

$$\mathbf{y} = \Lambda_y \eta + \varepsilon$$

The structural (inner) model

$\eta_1 = \gamma_{11}\,\xi_1 + \zeta_1$

$\eta_2 = \gamma_{21}\,\xi_1 + \beta_{21}\eta_{1} + \zeta_2$



$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \quad \xi = \begin{bmatrix} \xi_1 \end{bmatrix} \quad \Gamma = \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} \quad \mathrm{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \quad \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

$$\boxed{\eta = \Gamma\xi + \mathrm{B}\eta + \zeta}$$

The structural (inner) model

Vector of the endogenous LVs

Vector of the errors associated to the endogenous LVs

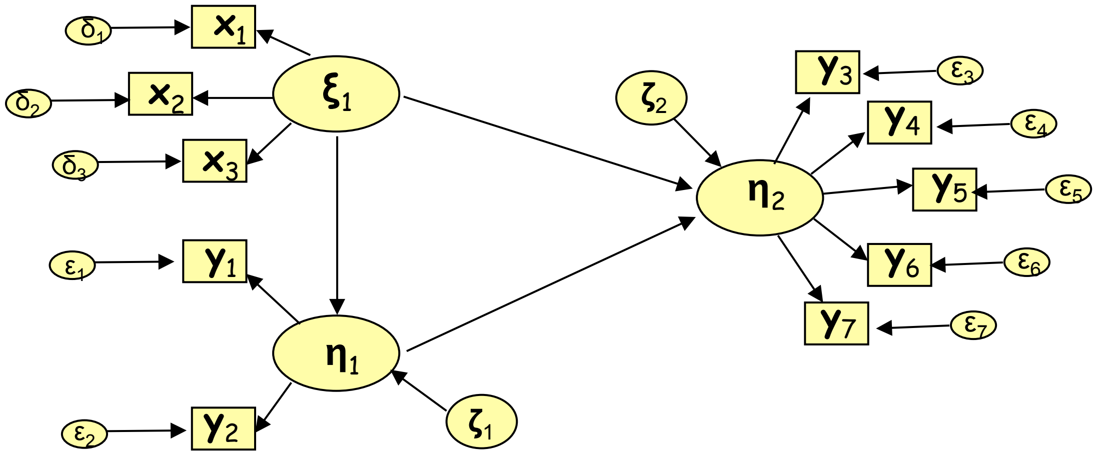
$$\eta = \Gamma\xi + \mathrm{B}\eta + \zeta$$

Vector of the exogenous LVs

Path-coefficients Matrix: it is a RECTANGULAR matrix with the path coefficients of the exogenous LVs.

Path-coefficients Matrix: it is a SQUARED matrix with the path coefficients of the endogenous LVs. On the diagonal there are only zeros

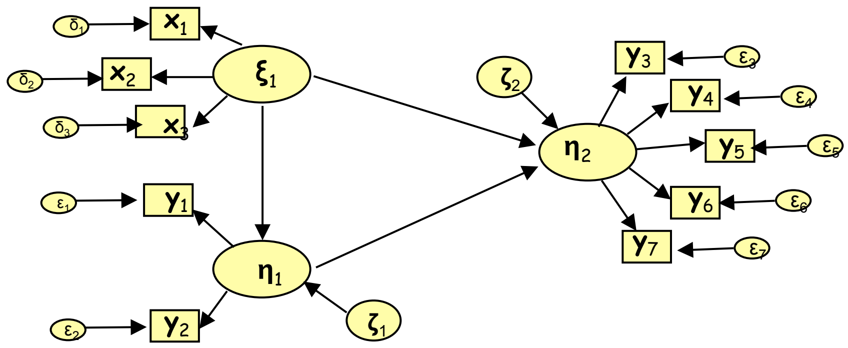
The Structural Equation Model



$$\left.\begin{aligned} \mathbf{x} &= \Lambda_x \xi + \delta \\ \mathbf{y} &= \Lambda_y \eta + \varepsilon \end{aligned}\right\} \text{Measurement Models}$$

$$\overbrace{\eta = \Gamma\xi + \mathrm{B}\eta + \zeta \;\Leftrightarrow\; \eta = (\mathbf{I} - \mathbf{B})^{-1}(\Gamma\xi + \zeta)}^{\text{Structural Model}}$$

Model assumptions



$$\mathbf{x} = \Lambda_x \xi + \delta$$

$$\mathbf{y} = \Lambda_y \eta + \varepsilon$$

$$\eta = (\mathbf{I} - \mathbf{B})^{-1}(\Gamma \xi + \zeta)$$

Assuming that:

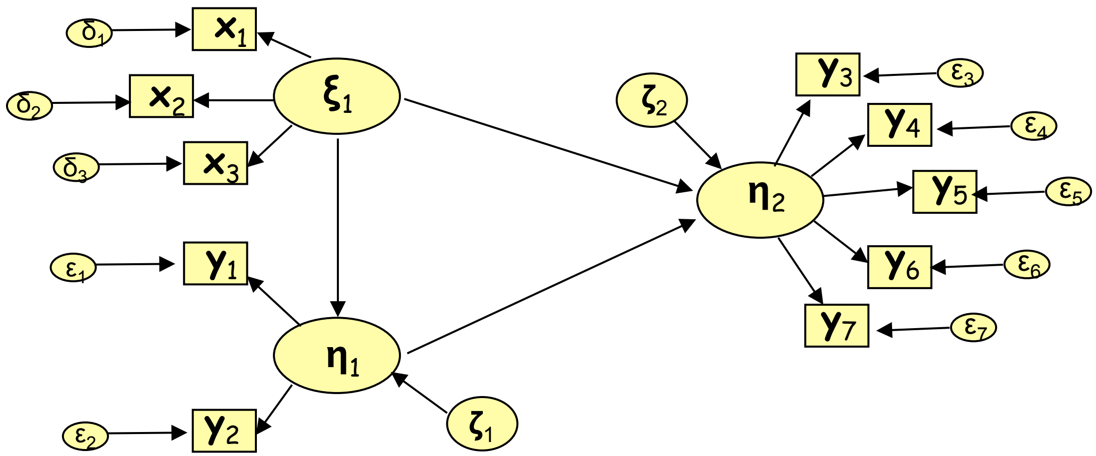
i) the MVs, the LV and the errors (both in structural and measurement models) are centered

ii) Two errors of different type (structural, exogenous measurement and endogenous measurement) do not covariate

iii) Measurement errors and LVs do not covariate

iv) The covariance between structural error and exogenous LVs is equal to zero

Model assumptions



$$\mathbf{x} = \Lambda_x \xi + \delta$$
$$\mathbf{y} = \Lambda_y \eta + \varepsilon$$
$$\eta = (\mathbf{I} - \mathbf{B})^{-1}(\Gamma\xi + \zeta)$$

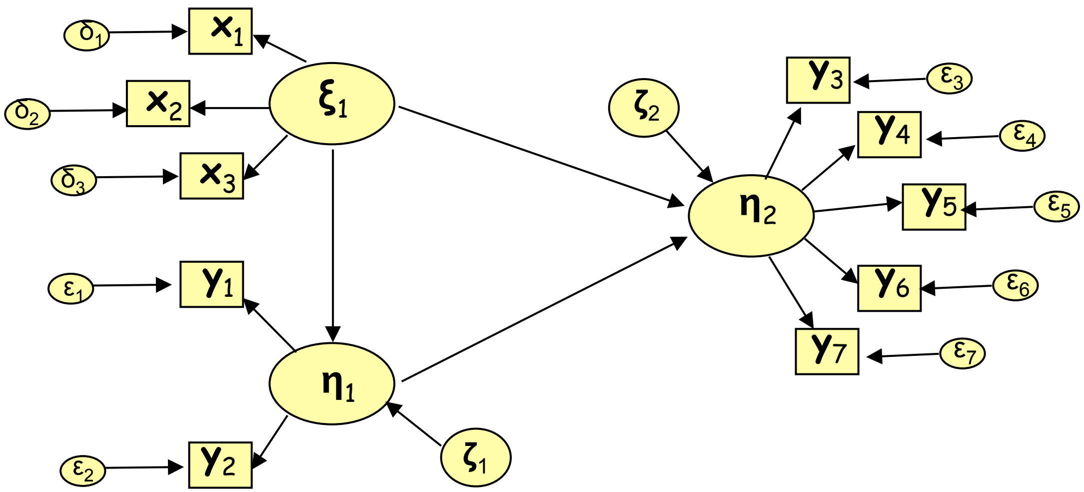
Hypotheses on Expectations

$$E(\mathbf{x}) = 0,\ E(\mathbf{y}) = 0,\ E(\xi) = 0,\ E(\eta) = 0,\ E(\delta) = 0,\ E(\varepsilon) = 0,\ E(\zeta) = 0$$

$$E(\delta\varepsilon') = 0,\ E((\zeta\delta')) = 0,\ E(\zeta\varepsilon') = 0$$
$$E(\delta\xi') = 0,\ E(\delta\eta') = 0,\ E(\varepsilon\xi') = 0,\ E(\varepsilon\eta') = 0$$
$$E(\zeta\xi') = 0$$

Hypotheses on Correlations

Analysing the covariance



Population Covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

We can write the covariance matrix among the MVs in terms of model parameters (**implied covariance matrix**)

$$C = \Sigma(\Omega) = \Sigma\left(\Gamma, \mathbf{B}, \Lambda_x, \Lambda_y, \Phi, \Psi, \Theta_\delta, \Theta_\varepsilon\right)$$

Path Coefficients

Loadings

Exog. LV Covariance

Structural Error Covariance

Measurement Error Covariance

Implied Covariance matrix Σ(Ω)

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Covariance matrix of the population

It can be rewritten as a **function of model parameters**

$$\Sigma(\Omega) = \left[\begin{array}{c|c} \Lambda_x \Phi \Lambda'_x + \Theta_\delta & \\ \hline \Lambda_y (\mathbf{I}-\mathbf{B})^{-1} \Gamma \Phi' \Lambda'_x & \Lambda_y \left[(\mathbf{I}-\mathbf{B})^{-1} (\Gamma \Phi \Gamma' + \Psi) (\mathbf{I}-\mathbf{B})^{-1\prime} \right] \Lambda'_y + \Theta_\varepsilon \end{array}\right]$$

Covariance matrix of the population implied by the model

Analysing the covariance

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Population Covariance matrix

$$C \approx S \approx \Sigma$$

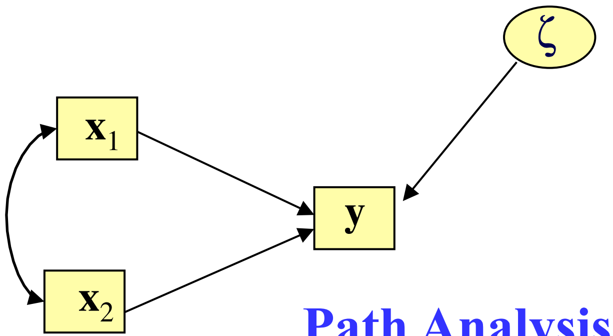
$$S = \begin{bmatrix} s_{xx} & \\ s_{yx} & s_{yy} \end{bmatrix}$$

Empirical covariance matrix

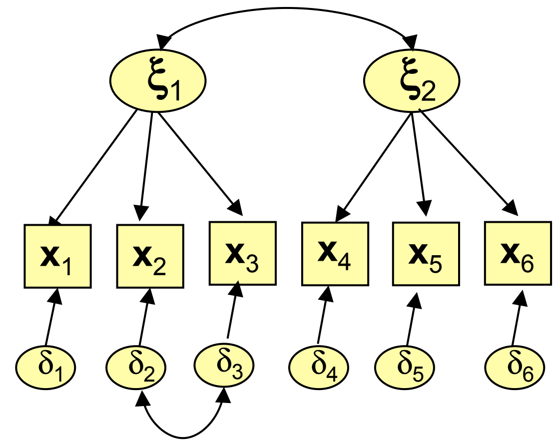
"Implied" covariance matrix

$$C = \Sigma(\Omega) = \left[\begin{array}{c:c} \Lambda_x \Phi \Lambda'_x + \Theta_\delta & \\ \hdashline \Lambda_y (\mathbf{I}-\mathbf{B})^{-1} \Gamma \Phi' \Lambda'_x & \Lambda_y \left[(\mathbf{I}-\mathbf{B})^{-1} (\Gamma \Phi \Gamma' + \Psi)(\mathbf{I}-\mathbf{B})^{-1'} \right] \Lambda'_y + \Theta_\varepsilon \end{array}\right]$$

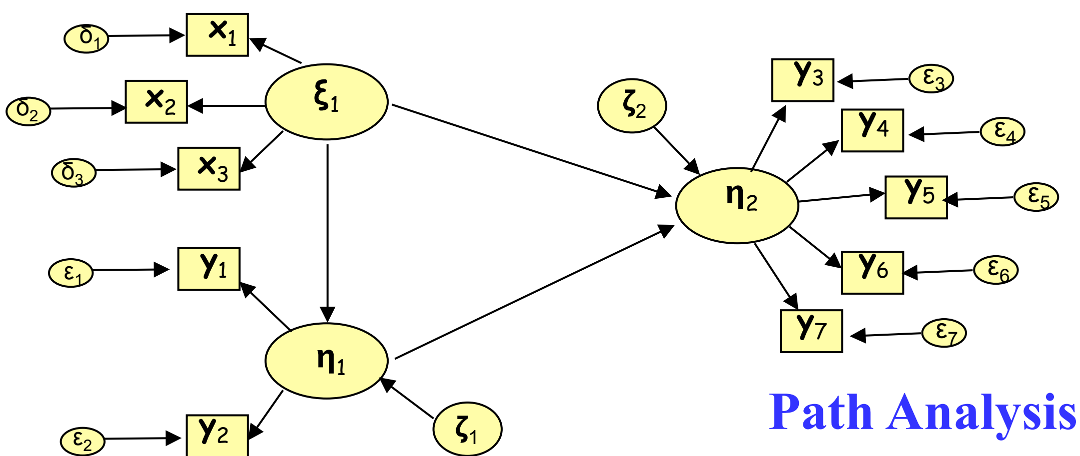
To summarize



Path Analysis with Manifest Variables



Confirmatory Factor Model



Path Analysis with Latent Variables (Structural Equation Model)

Discrepancy function

Estimation minimizes some **discrepancy function between the implied covariance matrix and the observed one**.

$$F = f\left(S - \Sigma\left(\hat{\Omega}\right)\right)$$

Estimated covariance matrix

$$\Sigma\left(\hat{\Omega}\right) = \mathbf{C} = \left[\begin{array}{c|c} \hat{\Lambda}_x \hat{\Phi} \hat{\Lambda}'_x + \hat{\Theta}_\delta & \\ \hline \hat{\Lambda}_y \left(\mathbf{I} - \hat{\mathbf{B}}\right)^{-1} \hat{\Gamma} \hat{\Phi}' \hat{\Lambda}'_x & \hat{\Lambda}_y \left[\left(\mathbf{I} - \hat{\mathbf{B}}\right)^{-1} \left(\hat{\Gamma}\hat{\Phi}\hat{\Gamma}' + \hat{\Psi}\right)\left(\mathbf{I} - \hat{\mathbf{B}}\right)^{-1'}\right] \hat{\Lambda}'_y + \hat{\Theta}_\varepsilon \end{array}\right]$$

Main discrepancy function for estimation

Assumptions:

→ data are multinormal

→ S follows a Wishart distribution

→ Both S and C are positive-definite

Maximum Likelihood

$$F_{ML} = \log\left|\mathbf{C}\right| + tr\left(\mathbf{SC}^{-1}\right) - \log\left|\mathbf{S}\right| - \left(P + Q\right)$$

Properties of the ML estimators:

- Asymptotically unbiased
- Consistent
- Asymptotically efficient
- The distribution of the ML estimators approximates a normal distribution as sample size increases

Alternative discrepancy functions

Unweighted Least Squares

$$F_{ULS} = \frac{1}{2} tr\left[(\mathbf{S} - \mathbf{C})^2 \right]$$

Generalised Least Squares

$$F_{GLS} = \frac{1}{2} tr\left[\mathbf{W}^{-1} (\mathbf{S} - \mathbf{C})^2 \right]$$

Asymptotically Distribution Free

$$F_{ADF/WLS} = (\underline{\mathbf{s}} - \underline{\mathbf{c}})^{\mathrm{T}} \mathbf{W}^{-1} (\underline{\mathbf{s}} - \underline{\mathbf{c}})$$

Model identification

Model identifiability and Degrees of freedom

Identifiability

A model is identifiable if its parameters are uniquely determined.

Degrees of freedom (DF)

DF = # equations (knowns) - # parameters to be estimated (unknowns)

Model identification condition:

$DF \geq 0$

This is a necessary **(but not sufficient)** condition

Some consideration on model identification

Perfect Identification:

→ DF (# equations - # parameters) = 0

A perfectly identified model yields a trivially perfect fit, making the test of fit uninteresting.

Overidentification:

→ DF (# equations - # parameters) > 0

A model is overidentified if there are **more knowns than unknowns**. Overidentified models **may not fit well and this is their interesting feature**.

Model identification in SEM

T-rule: A SEM is identified if the **covariance matrix may be uniquely decomposed in function of the model parameters**

→ if its **DF ≥ 0**, the number of covariances is larger than the number of parameters to be estimated: the model is **potentially identifiable**

$$DF = \left[\frac{1}{2}(P+Q)(P+Q+1) - t\right]$$

\# of MVs in the model:
P = # of exogenous MVs
Q = # of endogenous MVs

\# of parameters to be estimated

$\frac{1}{2}(P+Q)(P+Q+1)$ ➡ Number of unique elements in the Σ matrix

Necessary (but not sufficient) condition

Model fit and validation

Overall model fit measures

Chi-square Test - Global Validation Tests

$H_0 : \Sigma = C$ ➔ Good fit

$H_1 : \Sigma \neq C$

Test Statistic

$$(N-1)F \sim \chi^2_{DF}$$

Decision Rule:

The model is accepted if p-value $\geq$ 0.05 (We cannot reject the hypothesis H_0) or if Chi-square/DF ≤ 2 (or other thresholds such as 3 or 5)

N.B.: For a fixed level of differences in covariance matrices, the estimate of the Chi-square increases with N

- The power (i.e. the probability of rejecting a false H_0) depends on the sample size. **If the sample size is important, this test may lead to reject the model even if the data fit well the model**!
- **We cannot use this test to compare model estimated on different sample size**

Indices based on a baseline model

Model Comparison

The SATURATED Model:

This model contains as many parameter estimates as there are available degrees of freedom or inputs into the analysis. Therefore, this model shows 0 degrees of freedom. [This is the least restricted model possible]

The INDEPENDENCE Model:

This model contains estimates of the variance of the observed variables only. In other words, it assumes all relationships between the observed variables are zero (uncorrelated), no theoretical relationships. Therefore, this model shows the maximum number of degrees of freedom. [This is the most restrictive model possible and ANY TEST SHALL ALWAYS LEAD TO ITS REJECTION]

Independence vs Saturated model

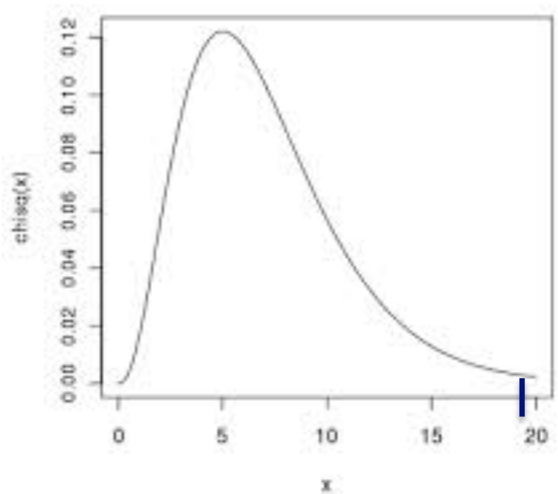
INDEPENDENCE Model: → SATURATED Model:

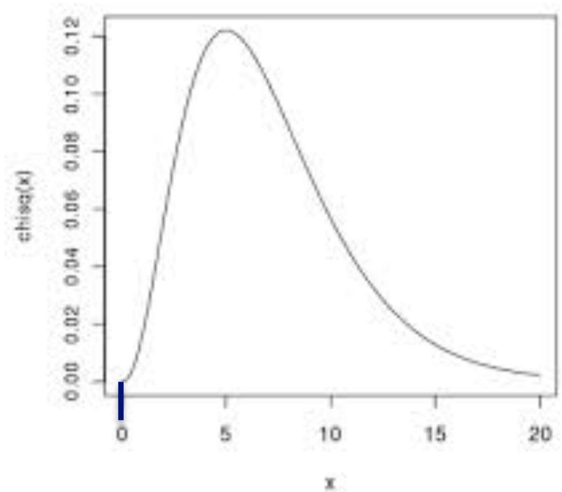
INDEPENDENCE Model:

FIT = min
DF = max
N_PAR = min
N_CONSTR = max
CHI2 = max
P-value = min

SATURATED Model:

FIT = max
DF = 0
N_PAR = max
N_CONSTR = min
CHI2 = 0
P-value = max





Indices based on a baseline model

Goodness-of-Fit Index (GFI)

This index was initially devised by Joreskog and Sorbom (1984) for ML and ULS estimation. It has then been generalised to other estimation criteria.

$$GFI = 1 - \frac{F}{F_{IND}}$$

Fit function that would results if all parameters were zero
(fit function of the INDEPENDENCE Model)

If a model is able to explain any true covariance between the observed variables, then F/F_{IND} would be 0 → GFI=1

The model is accepted if GFI is at least 0.9

PLS-Path Modeling

PLS Path Modeling: notations

- P **manifest variables** (MVs)observed on n units

→ $\mathbf{x}_{pq}$ generic MV

- Q **latent variables** (LVs)

→ ξ_q generic LV

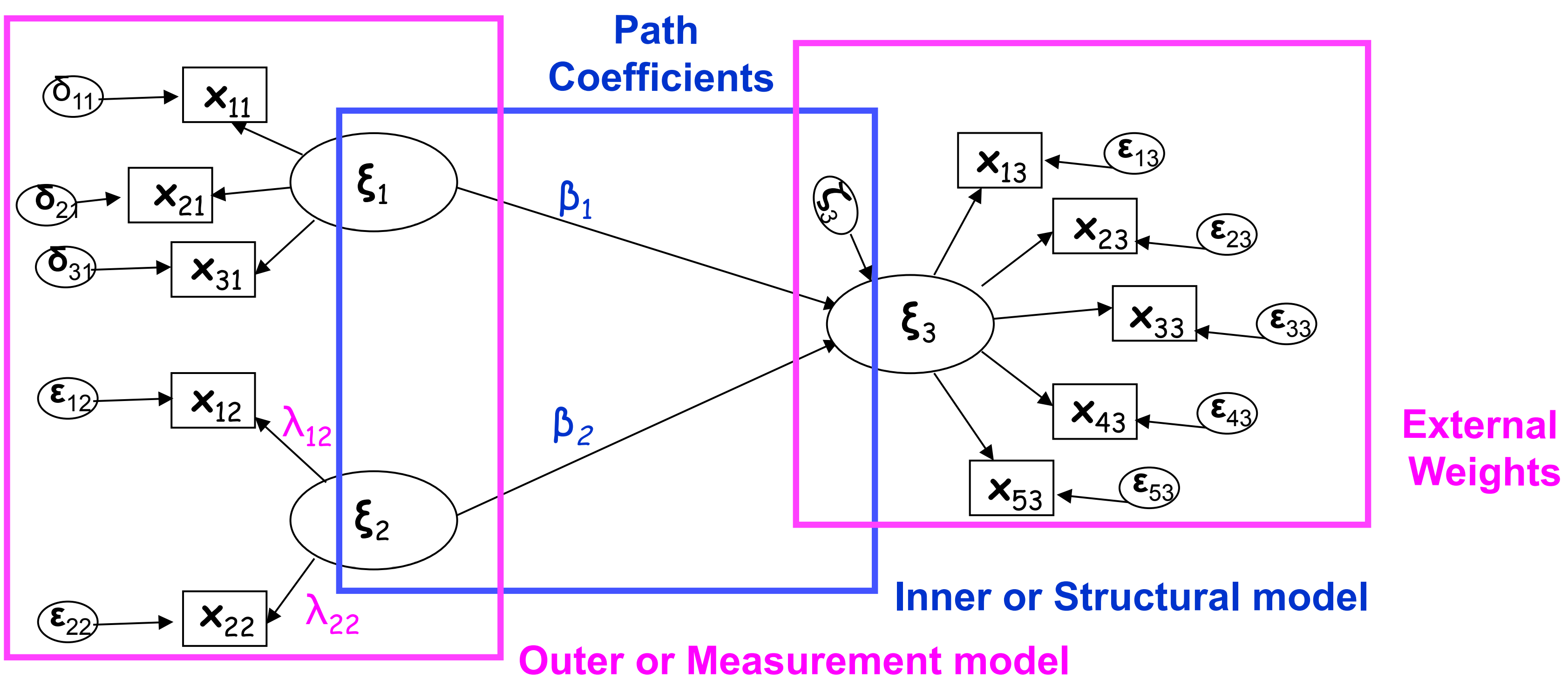
- Q **blocks** composed by each LV and the corresponding MVs

→ in each *q-th* block p_q manifest variables $\mathbf{x}_{pq}$, with $\sum_{q=1}^{Q} p_q = P$

N.B. Greek characters are used to refer to Latent Variables

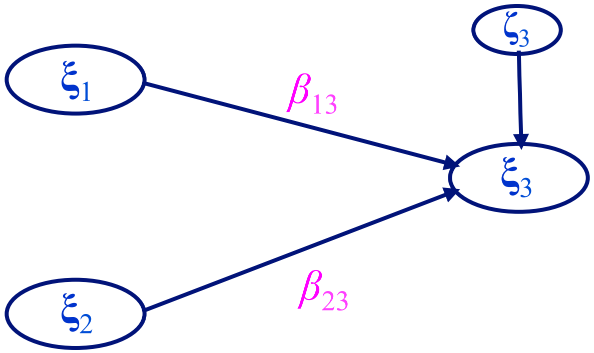
Latin characters refer to Manifest Variables

PLS Path Modeling: notation



PLS Path Model Equations: inner model

The structural model describes the relations among the latent variables



For each endogenous LV in the model it can be written as:

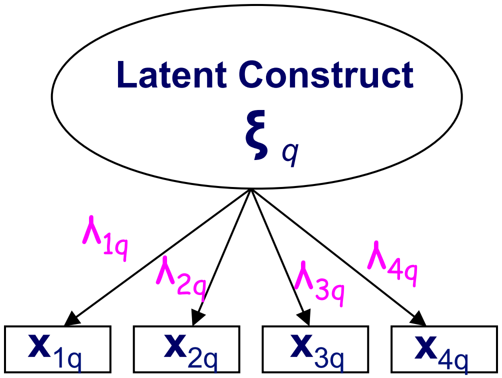
$$\xi_{q*} = \sum_{j=1}^{J} \beta_{jq*} \xi_j + \zeta_{q*}$$

where:

- B_{jq*} is the path-coefficient linking the *j-th* LV to the *q*-th* endogenous LV
- J is the number of the explanatory LVs impacting on ξ_{q*}

PLS Path Model Equations: inner model

The measurement model describes the relations among the manifest variables and the corresponding latent variable.



For each MV in the model it can be written as:

$$\mathbf{x}_{pq} = \lambda_{pq}\xi_q + \varepsilon_{pq}$$

where:

- l_{pq} is a loading term linking the *q-th* LV to the *p-th* MV

Weight relation – Linear composite

In component-based approach a weight relation defines each latent variable score as a weighted aggregate of its own MVs:

$$\boldsymbol{\xi}_q = \mathbf{X}_q \mathbf{w}_q$$

PLS-PM Algorithm

PLS-PM approach in 4 steps

1) Computation of the outer weights

Outer weights $\mathbf{w}_q$ are obtained by means of an iterative algorithm based on alternating LV estimations in the structural and in the measurement models

2) Computation of the LV scores (composites)

Latent variable scores are obtained as weighted aggregates of their own MVs:

$$\hat{\xi}_q \propto \mathbf{X}_q \mathbf{w}_q$$

3) Estimation of the path coefficients

Path coefficients are estimated as regression coefficients according to the structural model

4) Estimation of the loadings

Loadings are estimated as regression coefficients according to the measurement model

PLS Path Model: the algorithm

The aim of the PLS-PM algorithm is to define a system of weights to be applied at each block of MVs in order to estimate the corresponding LV, according to the weight relation:

$$\hat{\xi}_q \propto \mathbf{X}_q \mathbf{w}_q$$

This goal is achieved by means of an iterative algorithm based on two main steps:

- the outer estimation step
  → Latent Variable proxies = weighted aggregates of MVs
- the inner estimation step
  → Latent Variable proxies = weighted aggregates of connected LVs

A focus on the Outer Estimation

External (Outer) Estimation

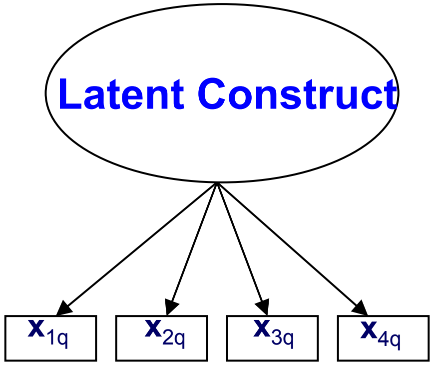
Composites = weighted aggregates of manifest variables

$$\mathbf{t}_q = \mathbf{X}_q\mathbf{w}_q$$

Mode A (for outwards directed links – **reflective** – principal factor model):

$w_{pq} = (1/n)\, \mathbf{x}_{pq}^T \mathbf{z}_q$

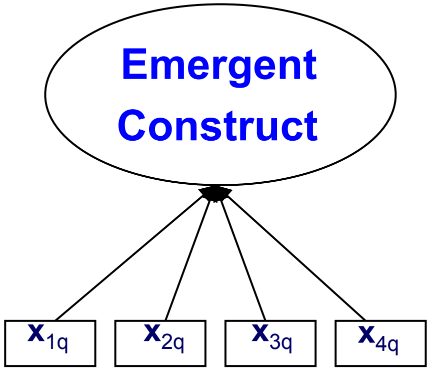
- ➔ These indicators **should covary**
- ➔ Several **simple OLS** regressions
- ➔ Explained Variance (higher **AVE**, communality)
- ➔ Internal Consistency
- ➔ **Stability of results** with well-defined blocks



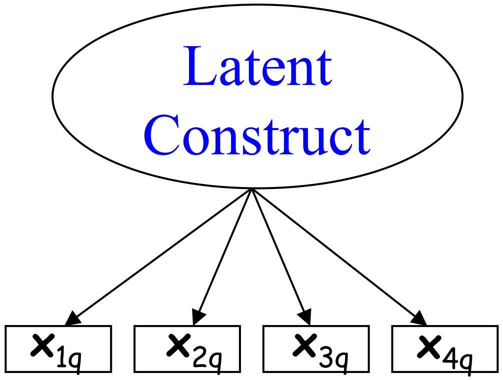
Mode B (for inwards directed links – **formative** – composite LV):

$\mathbf{w}_q = (\mathbf{X}_q^T\mathbf{X}_q)^{-1}\mathbf{X}_q^T\mathbf{z}_q$

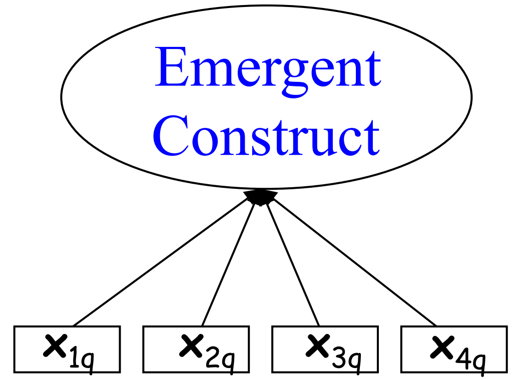
- ➔ These indicators **should covary**
- ➔ One **multiple OLS** regression (multicollinearity?)
- ➔ Structural Predictions (higher **R^2 values** for endogenous LVs)
- ➔ Multidimensionality (even partial, by sub-blocks)
- ➔ Might incur in **unstable results** with ill-defined blocks



Latent or Emergent Constructs?



Reflective (or Effects) Indicators



Formative (or Causal) Indicators

e.g. Consumer's attitudes, feelings

- Constructs **give rise** to observed variables (unique cause→ unidimensional)
- Aim at **accounting for observed variances** or covariances
- These indicators **should covary**: changes in one indicator imply changes in the others.
- **Internal consistency** is measured (es. Cronbach's alpha)

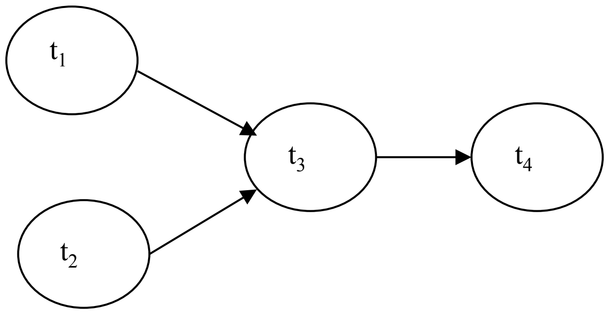
e.g. Social Status, Perceptions

- Constructs are **combinations of observed variables(multidimensional**)
- Not designed to account for observed variables
- These indicators **need not covary**: changes in one indicator do not imply changes in the others.
- Measures of internal **consistency do not apply**.

A focus on Inner Estimation

Inner Estimation

Latent Variable proxies = weighted aggregates of connected LVs



$$\mathbf{z}_q \propto \sum_{q'} e_{qq'} \mathbf{t}_{q'}$$

1. **Centroid scheme**: $\mathbf{z}_3 = e_{13}\mathbf{t}_1 + e_{23}\mathbf{t}_2 + e_{43}\mathbf{t}_4$ where $\mathrm{e}_{qq'} = sign(cor(\mathbf{t}_q, \mathbf{t}_{q'}))$

2. **Factorial scheme**: $\mathbf{z}_3 = cor(\mathbf{t}_3, \mathbf{t}_1) * \mathbf{t}_1 + cor(\mathbf{t}_3, \mathbf{t}_2) * \mathbf{t}_2 + cor(\mathbf{t}_3, \mathbf{t}_4) * \mathbf{t}_4$

3. **Path weighting scheme** : $\mathbf{z}_3 = \hat{\gamma}_{31} \times \mathbf{t}_1 + \hat{\gamma}_{32} \times \mathbf{t}_2 + cor(\mathbf{t}_3, \mathbf{t}_4) \times \mathbf{t}_4$

   Where the betas are the regression coeddicients of the model: $\mathbf{t}_3 = \gamma_{31} \times \mathbf{t}_1 + \gamma_{32} \times \mathbf{t}_2 + \delta$

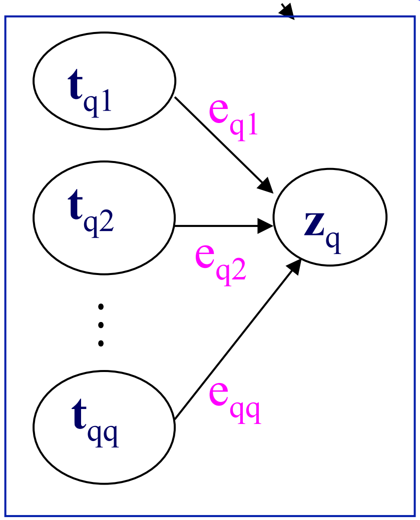
The PLS Path Modeling algorithm

MVs are centered or standardized

Initial step → $\mathbf{w}_q$

$$\mathbf{t}_q = \pm \mathbf{X}_q \mathbf{w}_q$$

Outer estimation



Inner estimation

Reiterate till *Numerical Convergence*

Mode **A**: $\mathbf{w}_q = (1/n)\mathbf{X}_q{}'\mathbf{z}_q$

Mode **B**: $\mathbf{w}_q = (\mathbf{X}_q{}'\mathbf{X}_q)^{-1}\mathbf{X}_q{}'\mathbf{z}_q$

Update weights **w**

Choice of weights $\mathbf{e}_{qq'}$:
- **Centroid**: correlation signs
- **Factorial**: correlations
- **Path weighting scheme**: multiple regression coefficients or correlations

PLS-PM Criteria

Optimization Criteria behind the PLS-PM

Full Mode B PLS-PM

Glang (1988) and Mathes (1993) showed that the stationary equation of a "full mode B" PLS-PM solves this optimization criterion:

$$\arg\max_{\mathbf{w}_q} \left\{ \sum_{q \neq q'} c_{qq'} g\left(\operatorname{cov}\left(\mathbf{X}_q \mathbf{w}_q, \mathbf{X}_{q'} \mathbf{w}_{q'}\right)\right) \right\} \quad \text{s.t. } \left\|\mathbf{X}_q \mathbf{w}_q\right\|^2 = n$$

where:

$$c_{qq'} = \begin{cases} 1 & \text{if } \mathbf{X}_q \text{ and } \mathbf{X}_{q'} \text{ is connected} \\ 0 & \text{otherwise} \end{cases}$$

$$g = \begin{cases} \text{square} & \text{(Factorial scheme)} \\ \text{abolute value} & \text{(Centroid scheme)} \end{cases}$$

Hanafi (2007) proved that PLS-PM iterative algorithm is monotonically convergent to these criteria

Optimization Criteria behind PLS-PM

Full Mode A PLS-PM

Kramer (2007) showed that "full Mode A" PLS-PM algorithm is not based on a stationary equation related to the optimization of a twice differentiable function

Full NEW Mode A PLS-PM

In 2007 Kramer showed also that a slightly adjusted PLS-PM iterative algorithm (in which a normalization constraint is put on outer weights rather than latent variable scores) we obtain a stationary point of the following optimization problem:

$$\underset{\|\mathbf{w}_q\|^2=n}{\arg\max}\left\{\sum_{q\neq q'} c_{qq'}\, g\left(\operatorname{cov}\left(\mathbf{X}_q\mathbf{w}_q, \mathbf{X}_{q'}\mathbf{w}_{q'}\right)\right)\right\}$$

Tenenhaus and Tenenhaus (2011) proved that the modified algorithm proposed by Kramer is monotonically convergent to this criterion.

Optimization Criteria behind PLS-PM

A general criterion for PLS-PM, in which (New) Mode A and B are mixed, can be written as follows:

$$\arg\max_{\mathbf{w}_q} \left\{ \sum_{q \neq q'} c_{qq'} g\left(\text{cov}\left(\mathbf{X}_q \mathbf{w}_q, \mathbf{X}_{q'} \mathbf{w}_{q'} \right) \right) \right\} =$$

$$\arg\max_{\mathbf{w}_q} \left\{ \sum_{q \neq q'} c_{qq'} g\left[cor\left(\mathbf{X}_q \mathbf{w}_q, \mathbf{X}_{q'} \mathbf{w}_{q'} \right) \sqrt{\text{var}\left(\mathbf{X}_q \mathbf{w}_q \right)} \sqrt{\text{var}\left(\mathbf{X}_{q'} \mathbf{w}_{q'} \right)} \right] \right\}$$

s.t. $\left\| \mathbf{X}_q \mathbf{w}_q \right\|^2 = n$ if Mode B for block q

$\left\| \mathbf{w}_q \right\|^2 = n$ if New Mode A for block q

The empirical evidence shows that Mode A (unknown) criterion is approximated by the New Mode A criterion

PLS-PM as a general framework for data analysis

PLS-PM SPECIAL CASES

- Principal component analysis
- Multiple factor analysis
- Canonical correlation analysis
- Redundancy analysis
- PLS Regression
- Generalized canonical correlation analysis (Horst)
- Generalized canonical correlation analysis (Carroll)
- Multiple Co-inertia Analysis (MCOA) (Chessel & Hanafi, 1996)

One block case

Principal Component Analysis through PLS-PM*

SPSS results (principal components)

Component Matrix[a]

	Component 1
VVLT1	.648
VVLT2	.729
VVLT3	.823
VVLT4	.830

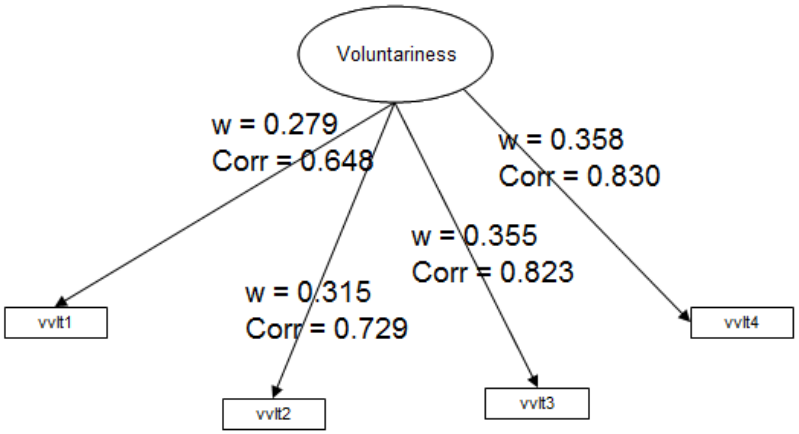
Extraction Method: Principal Component Analysis.
a. 1 components extracted.

Component Matrix[a]

	Component 1
VCPT1	.869
VCPT2	.919
VCPT3	.938
VCPT4	.920

Principal Component Analysis.
a. 1 components extracted.

XL-STAT graphical results

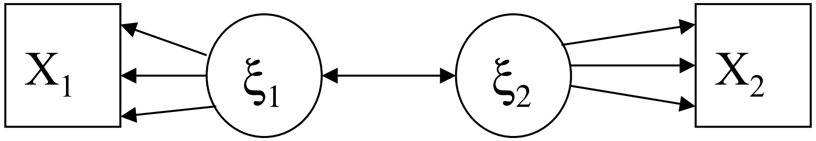


* Results from W.W. Chin slides on PLS-PM

Two block case

Tucker Inter-batteries Analysis
(1st component)

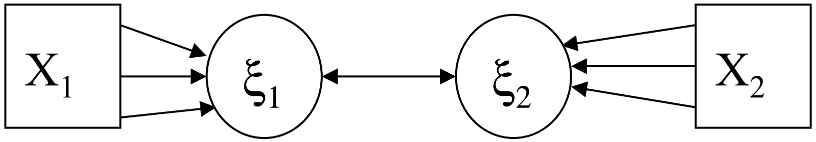
$$\underset{\|\mathbf{w}_1\|=\|\mathbf{w}_2\|=1}{\arg\max}\left\{\operatorname{cov}\left(\mathbf{X}_1\mathbf{w}_1,\mathbf{X}_2\mathbf{w}_2\right)\right\}$$



Mode A for X_1, Mode A for X_2

Canonical Correlation Analysis
(1st component)

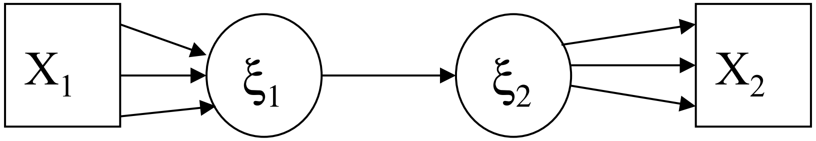
$$\underset{\operatorname{var}(\mathbf{X}_1\mathbf{w}_1)=\operatorname{var}(\mathbf{X}_2\mathbf{w}_2)=1}{\arg\max}\left\{\operatorname{cov}\left(\mathbf{X}_1\mathbf{w}_1,\mathbf{X}_2\mathbf{w}_2\right)\right\}$$



Mode B for X_1, Mode B for X_2

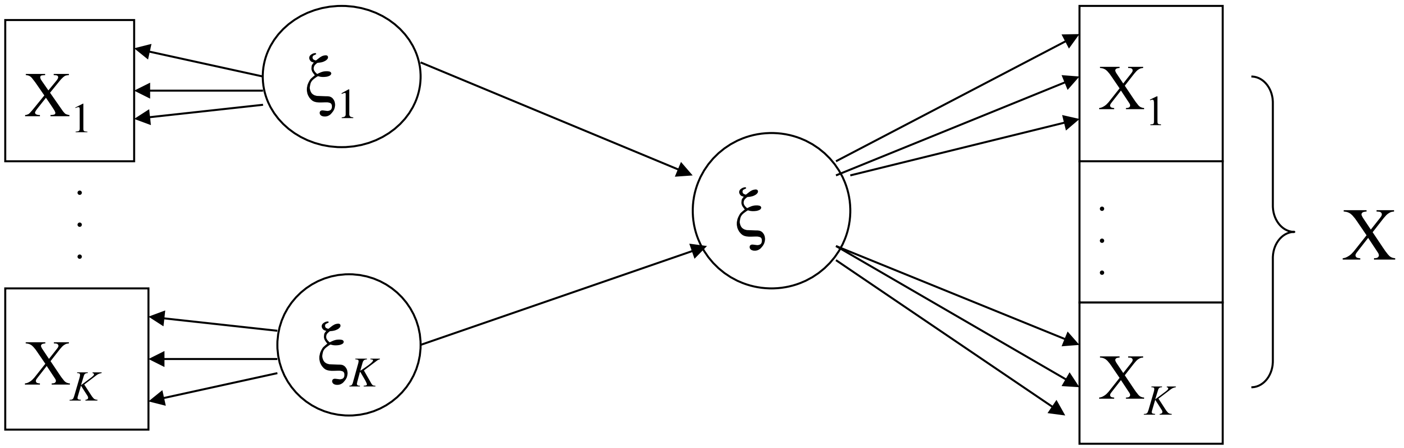
Redundancy Analysis
(1st component)

$$\underset{\operatorname{var}(\mathbf{X}_1\mathbf{w}_1)=\|\mathbf{w}_2\|=1}{\arg\max}\left\{\operatorname{cov}\left(\mathbf{X}_1\mathbf{w}_1,\mathbf{X}_2\mathbf{w}_2\right)\right\}$$



Mode B for X_1, Mode A for X_2

Hierarchical Models



Mode A + Path Weighting

- Lohmöller's Split PCA
- Multiple Factorial Analysis by Escofier and Pagès
- Horst's Maximum Variance Algorithm
- Multiple Co-Inertia Analysis (ACOM) by Chessel and Hanafi

$$\underset{\mathrm{var}(\mathbf{X}_k\mathbf{w}_k)=1,\mathbf{Xw}=\sum_k \mathbf{X}_k\mathbf{w}_k}{\arg\max}\left\{\sum_k \mathrm{cov}^2\left(\mathbf{X}_k\mathbf{w}_k,\mathbf{Xw}\right)\right\}$$

Mode B + Factorial

- Generalised Canonical Correlation Analysis (Carroll)

$$\underset{\mathrm{var}(\mathbf{X}_k\mathbf{w}_k)=1,\mathbf{Xw}=\sum_k \mathbf{X}_k\mathbf{w}_k}{\arg\max}\left\{\sum_k cor\left(\mathbf{X}_k\mathbf{w}_k,\mathbf{Xw}\right)\right\}$$

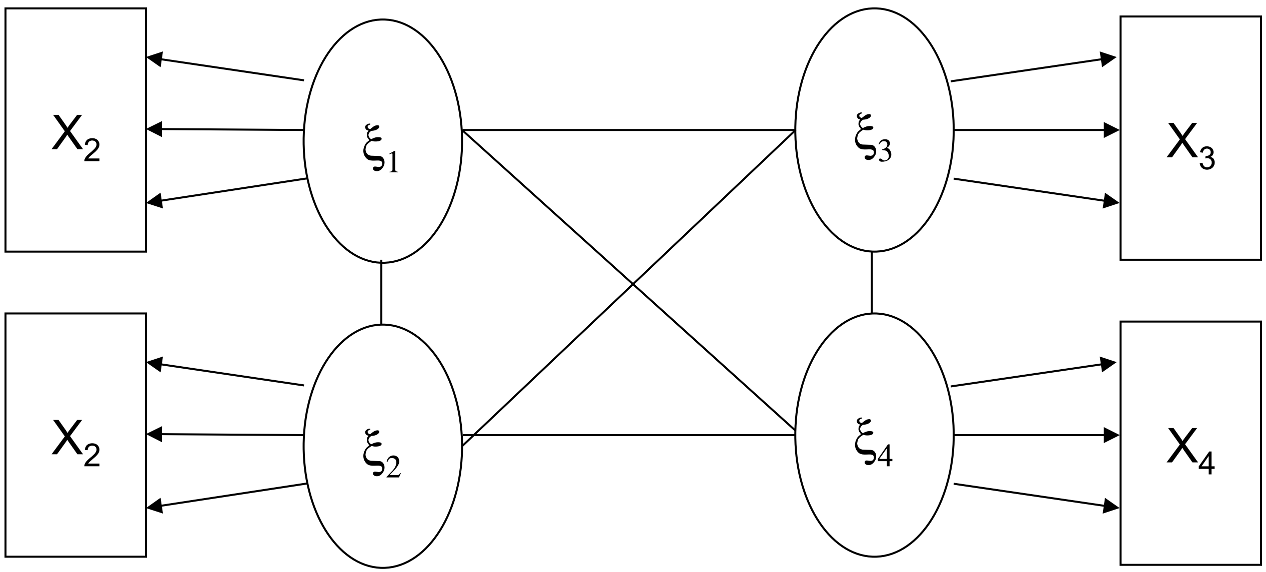
Mode B + Centroid

- Generalised CCA (Horst's SUMCOR criterion)
- Mathes (1993) & Hanafi (2004)

$$\underset{\mathrm{var}(\mathbf{X}_k\mathbf{w}_k)=1,\mathbf{Xw}=\sum_k \mathbf{X}_k\mathbf{w}_k}{\arg\max}\left\{\sum_k cor^2\left(\mathbf{X}_k\mathbf{w}_k,\mathbf{Xw}\right)\right\}$$

‘Confirmatory’ PLS Model

Each LV is connected to all the others



PLS criteria for multiple table analysis

Method	Criterion $\left(F_k = \mathbf{X}_k \mathbf{w}_k, F = \mathbf{X}\mathbf{w}\right)$	PLS path model	Mode	Scheme
(1) SUMCOR (Horst 1961)	$Max \sum_{j,k} Cor(F_j, F_k)$ or $Max \sum_j Cor(F_j, \sum_k F_k)$	Hierarchical	B	Centroid
(2) MAXVAR (Horst 1961) or GCCA (Carroll 1968)	$Max \{\lambda_{first}[Cor(F_j, F_k)]\}$ (a) or $Max \sum_j Cor^2(F_j, F_{j+1})$	Hierarchical	B	Factorial
(3) SsqCor (Kettenring 1971)	$Max \sum_{j,k} Cor^2(F_j, F_k)$	Confirmatory	B	Factorial
(4) GenVar (Kettenring 1971)	$Min \{\det[Cor(F_j, F_k)]\}$			
(5) MINVAR (Kettenring 1971)	$Min \{\lambda_{last}[Cor(F_j, F_k)]\}$ (b)			
(6) Lafosse (1989)	$Max \sum_j Cor^2(F_j, \sum_k F_k)$			
(7) Mathes (1993) or Hanafi (2005)	$Max \sum_{j,k} \lvert Cor(F_j, F_k)\rvert$	Confirmatory	B	Centroid
(8) MAXDIFF (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all\ \lVert w_j \rVert = 1} \sum_{j \neq k} Cov(X_j w_j, X_k w_k)$			
(9) MAXBET (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all\ \lVert w_j \rVert = 1} \sum_{j,k} Cov(X_j w_j, X_k w_k)$			
(10) MAXDIFF B (Hanafi and Kiers 2006)	$Max_{all\ \lVert w_j \rVert = 1} \sum_{j \neq k} Cov^2(X_j w_j, X_k w_k)$			

From Tenenhaus et Hanafi (2010)

(a) $\lambda_{first}\left[Cor(F_j, F_k)\right]$ is the first eigenvalue of block LV correlation matrix.
(b) $\lambda_{last}\left[Cor(F_j, F_k)\right]$ is the last eigenvalue of block LV correlation matrix.

PLS criteria for multiple table analysis

Method	Criterion	PLS path model	Mode	Scheme
(11) (Hanafi and Kiers 2006)	$Max_{all\ \lVert w_j \rVert=1} \sum_{j \neq k} \lvert Cov(X_j w_j, X_k w_k) \rvert$			
(12) ACOM (Chessel and Hanafi 1996) or Split PCA (Lohmöller 1989)	$Max_{all\ \lVert w_j \rVert=1} \sum_j Cov^2(X_j w_j, X_{j+1} w_{j+1})$ or $Min_{F,p_j} \sum_j \left\lVert X_j - F p_j^T \right\rVert^2$	Hierarchical	A	Path-weighting
(13) CCSWA (Hanafi et al., 2006) or HPCA (Wold et al., 1996)	$Max_{all\ \lVert w_j \rVert=1, Var(F)=1} \sum_j Cov^4(X_j w_j, F)$ or $Min_{\lVert F \rVert=1} \sum_j \left\lVert X_j X_j^T - \lambda_j F F^T \right\rVert^2$			
(14) Generalized PCA (Casin 2001)	$Max \sum_j R^2(F, X_j) \sum_h Cor^2\left(x_{jh}, \hat{F}_j\right)$ (c)			
(15) MFA (Escofier and Pagès 1994)	$Min_{F,p_j} \sum_j \left\lVert \frac{1}{\sqrt{\lambda_{first}\left[Cor(x_{jh}, x_{jl})\right]}} X_j - F p_j^T \right\rVert^2$	Hierarchical (applied to the reduced X_j) (d)	A	Path-weighting
(16) Oblique maximum variance method (Horst 1965)	$Min_{F,p_j} \sum_j \left\lVert X_j \left(\frac{1}{n} X_j^T X_j\right)^{-1/2} - F p_j^T \right\rVert^2$	Hierarchical (applied to the transformed X_j) (e)	A	Path-weighting

(c) $\widehat{F}_j$ is the prediction of F in the regression of F on block X_j.

(d) The reduced block number j is obtained by dividing the block X_j by the square root of $\lambda_{first}\left[Cor(x_{jh}, x_{j\ell})\right]$.

(e) The transformed block number j is computed as $X_j[(1/n)X_j^T X_j]^{-1/2}$.

From Tenenhaus et Hanafi (2010)

Model Assessment

Reliability

The reliability rel(x_{pq}) of a measure x_{pq} of a true score ξ_q modeled as $x_{pq} = \lambda_p \xi_q + \delta_{pq}$ is defined as:

$$rel\left(x_{pq}\right) = \frac{\lambda_{pq}^2 \operatorname{var}\left(\xi_q\right)}{\operatorname{var}\left(x_{pq}\right)} = cor^2\left(x_{pq}, \xi_q\right)$$

***rel*(x_{pq}) can be interpreted as the variance of x_{pq} that is explained by ξ_q**

Measuring the Reliability

Question:

How to measure the overall reliability of the measurement tool ?
In other words, how to measure the homogeneity level of a block X_q of positively correlated variables?

Answer:

The composite reliability (**internal consistency**) of manifest variables can be checked using:

- the Cronbach's Alpha
- the Dillon Goldstein rho

Composite reliability

The measurement model (in a **reflective** scheme) assumes that each group of manifest variables is homogeneous and unidimensional (related to a single variable). The composite reliability (**internal consistency or homogeneity of a block**) of manifest variables is **measured** by either of the following indices:

$$\alpha_q = \frac{P_q}{(P_q - 1)} \frac{\sum_{p \neq p'} \operatorname{cov}(x_{pq}, x_{p'q})}{P_q + \sum_{p \neq p'} \operatorname{cov}(x_{pq}, x_{p'q})}$$

$$\rho_q = \frac{\left(\sum_p \lambda_{pq}\right)^2 \times \operatorname{var}(\xi_q)}{\left(\sum_p \lambda_{pq}\right)^2 \times \operatorname{var}(\xi_q) + \sum_p \operatorname{var}(\varepsilon_{pq})}$$

Where:

- $\mathbf{x}_{pq}$ is the p-th manifest variable in the block q,
- P_q is the number of manifest variables in the block,
- λ_{pq} is the component loading for $\mathbf{x}_{pq}$
- $\operatorname{var}(\varepsilon_{pq})$ is the variance of the measurement error
- MVs are standardized

Cronbach's alpha assumes lambda-equivalence (parallelity) and is a lower bound estimate of reliability

The manifest variables are reliable if these indices are at least 0.7

(0.6 to 0.8 according to exploratory vs. confirmatory purpose)

Average Variance Extracted (AVE)

The goodness of measurement model (**reliability of latent variables**) is evaluated by the amount of variance that a LV captures from its indicators (**average communality**) relative to the amount due to measurement error.

Average Variance Extracted

$$AVE_q = \frac{\sum_p \left[\lambda_{pq}^2 \operatorname{var}\left(\xi_q\right)\right]}{\sum_p \left[\lambda_{pq}^2 \operatorname{var}\left(\xi_q\right)\right] + \sum_q \left(1 - \lambda_{pq}^2\right)}$$

- The convergent validity holds if AVE is >0.5
- Consider also **standardised loadings >0.707**

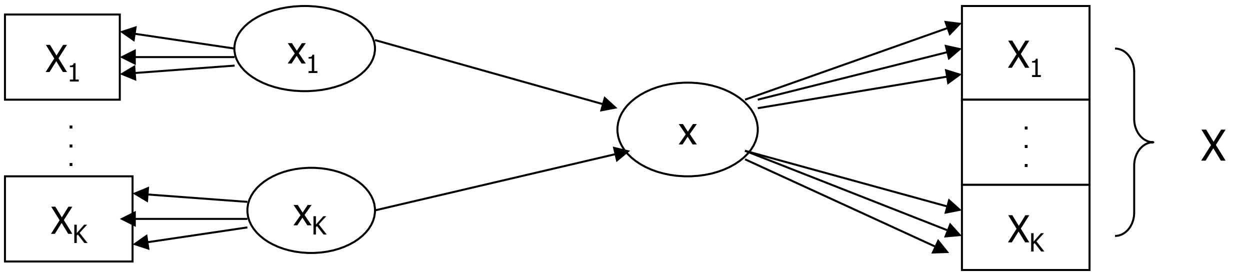
What if unidimensionality is rejected?

Four possible solutions:

- **Remove** manifest variables that are far from the model

- Change the measurement model into a **formative model** (eventual multicollinearity problems -> via PLS Regression)

- Use the auxiliary variable in the **multiple table analysis** of unidimensional sub-blocks:



- Split the multidimensional block into **unidimensional sub-blocks**

Discriminant and Nomological Validity

The latent variables shall be correlated (nomological validity) but they need to measure different concepts (discriminant validity). It must be possible to discriminate between latent variables if they are meant to refer to distinct concepts.

$$H_0 : \mathrm{cor}\left(\xi_q, \xi_{q'}\right) = 1$$

$$H_0 : \mathrm{cor}\left(\xi_q, \xi_{q'}\right) = 0$$

The correlation between two latent variables is tested to be significantly lower than 1 (discriminant validity) and significantly higher than 0 (nomological validity):

Decision Rules:

The null hypotheses are rejected if:

1. **95% confidence interval** for the mentioned correlation does not comprise 1 and 0, respectively (bootstrap/jackknife empirical confidence intervals);
2. For discriminant validty only: (**AVE**$_q$ and **AVE**$_{q'}$) **>** $\mathrm{cor}^2\left(\hat{\xi}_q, \hat{\xi}_{q'}\right)$ which indicates that more variance is shared between the LV and its block of indicators than with another LV representing a different block of indicators.

Model Assessement

Since PLS-PM is a Soft Modeling approach, model validation regards only the way relations are modeled, in both the structural and the measurement model; in particular, the following null hypotheses should be **rejected**:

a) $\lambda_{pq} = 0$, as each MV is supposed to be correlated to its corresponding LV;

b) $w_{pq} = 0$, as each LV is supposed to be affected by all the MVs of its block;

c) $\beta_{qq'} = 0$, as each latent predictor is assumed to be explanatory with respect to its latent response;

d) $R^2_{q*} = 0$, as each endogenous LV is assumed to be explained by its latent predictors;

e) $cor(\xi_q; \xi_{q'}) = 0$, as LVs are assumed to be connected by a statistically significant correlation. Rejecting this hypothesis means assessing the Nomological Validity of the PLS Path Model;

f) $cor(\xi_q; \xi_{q'}) = 1$, as LVs are assumed to measure concepts that are different from one another. Rejecting this hypothesis means assessing the Discriminant Validity of the PLS Path Model;

g) Both AVE_q and $AVE_{q'}$ smaller than $cor^2(\xi_q; \xi_{q'})$, as a LV should be related more strongly with its block of indicators than with another LV representing a different block of indicators.

Model Quality

Communality

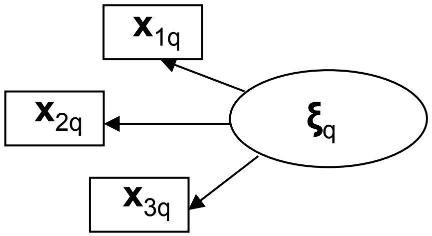
For each manifest variable x_{pq} the communality is a squared correlation:

$$Com_{pq} = cor^2\left(\mathbf{x}_{pq}, \xi_q\right)$$

The communality of a block is the mean of the communalities of its MVs

$$Com_q = \frac{1}{p_q}\sum_{p=1}^{p_q} cor^2\left(\mathbf{x}_{pq}, \xi_q\right)$$

(NB: if standardised MVs: $Com_q = AVE_q$)



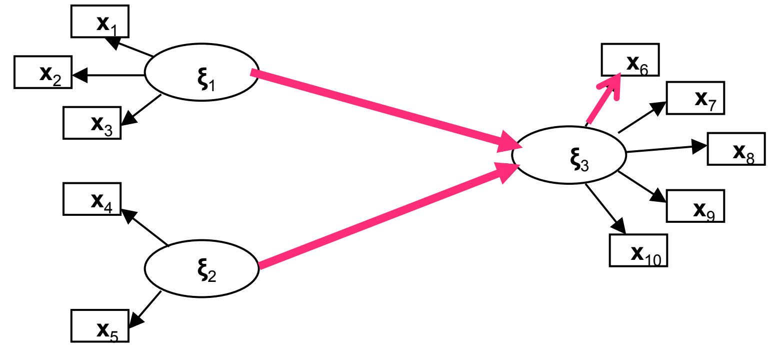
The communality of the whole model is the **Mean Communality**, obtained as:

$$\overline{Com} = \frac{\sum_{q:P_q>1}\left(p_q \times Com_q\right)}{\sum_{q:P_q>1} P_q}$$

Redundancy

Redundancy is the average variance of the MVs set, related to the J* endogenous LVs, explained by the exogenous LVs:

$$RED_{x_{pq^*}} = \frac{Var\left[\beta_{qq^*}\xi_q\right]}{Var\left[x_{pq^*}\right]}\lambda_{pq^*}^2$$



$$\text{Redundancy}_{q^*} = \text{R}^2\left(\xi_{q^*}, \xi_{q:\ \xi_q \to \xi_{q^*}}\right) \times \text{Communality}_{q^*}$$

CV-communality and redundancy

The **Stone-Geisser** test follows a **blindfolding procedure**: repeated (for all data points) omission of a **part of the data** matrix (by row and column, where **jackknife** proceeds exclusively by row) while estimating parameters, and then reconstruction of the omitted part by the estimated parameters.

This procedure results in:
- a generalized cross-validation measure that, in case of a negative value, implies a bad estimation of the related block
- « jackknife standard deviations » of parameters (but most often these standard deviations are very small and lead to significant parameters)

Communality Option

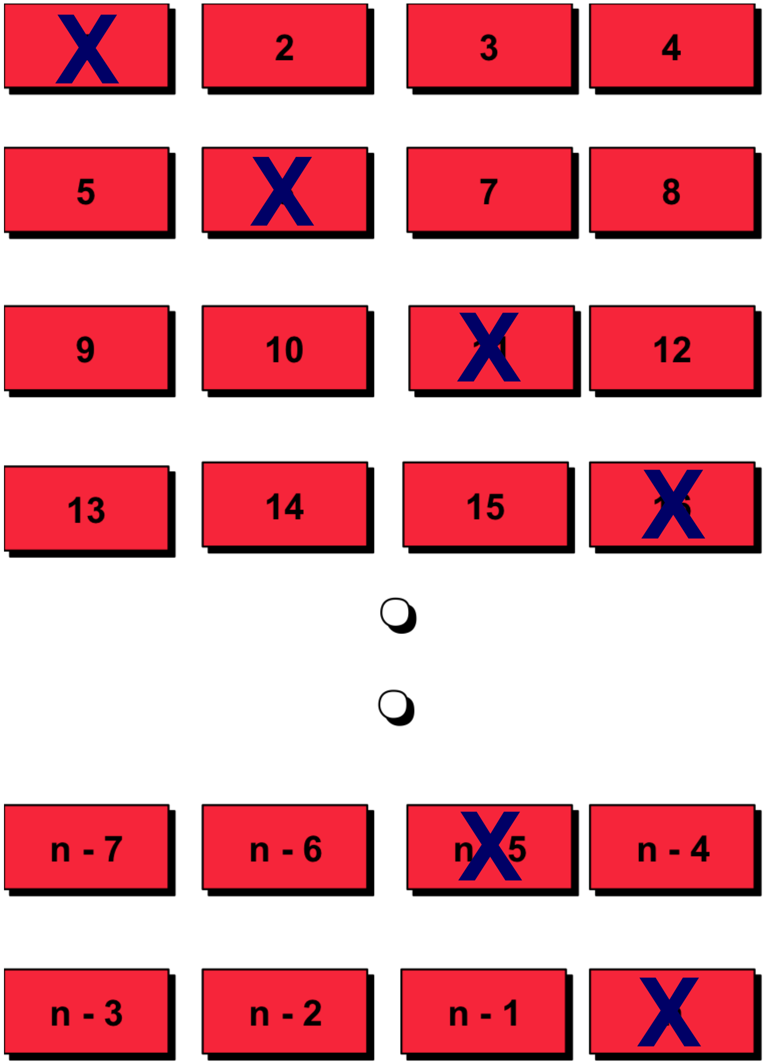
$$H_q^2 = 1 - \frac{\sum_q \sum_i (x_{pqi} - \bar{x}_{pq} - \hat{\lambda}_{pq(-i)} \hat{\xi}_{q(-i)})^2}{\sum_q \sum_i (x_{pqi} - \bar{x}_{pq})^2}$$

Redundancy Option (also called Q^2)

$$F_q^2 = 1 - \frac{\sum_q \sum_i (x_{pqi} - \bar{x}_{pq} - \hat{\lambda}_{pq(-i)} \mathrm{Pred}(\hat{\xi}_{q(-i)}))^2}{\sum_q \sum_i (x_{pqi} - \bar{x}_{pq})^2}$$

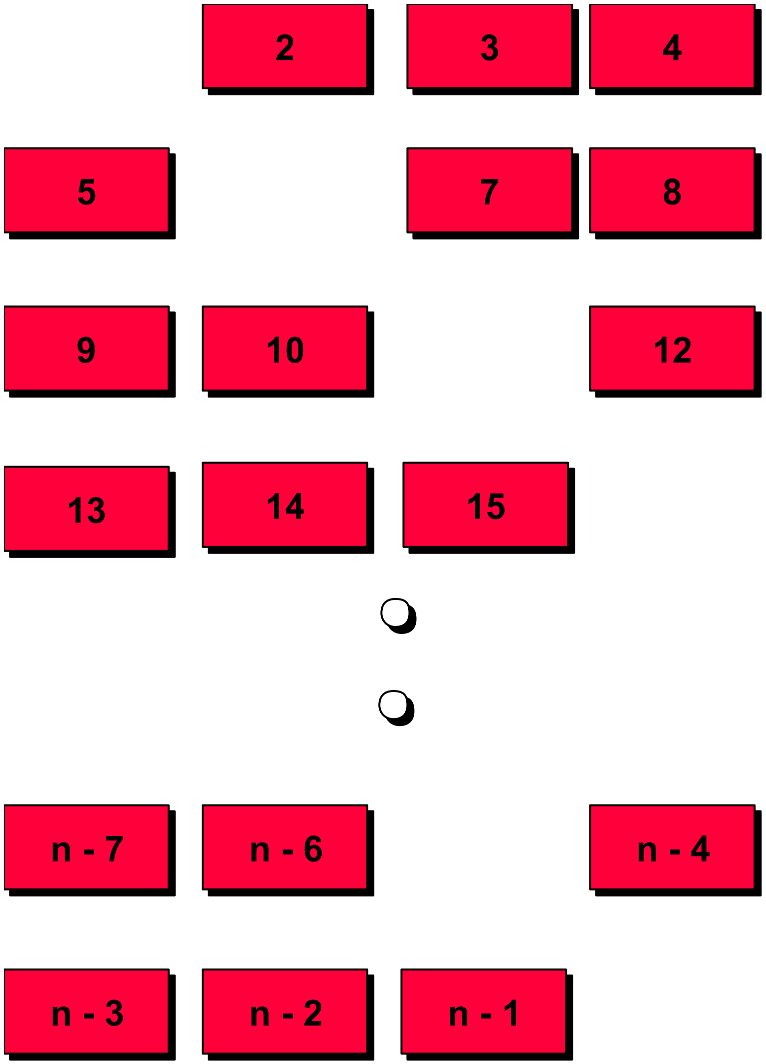
The mean of the CV-communality and the CV-redundancy (for endogenous blocks) indices can be used to measure the global quality of the measurement model if they are positive for all blocks (endogenous for redundancy).

Blindfolding procedure



From W.W. Chin’s slides on PLS-PM

Blindfolding procedure



From W.W. Chin’s slides on PLS-PM

A global quality index for PLS-PM

- PLS-PM does not optimize one single criterion, instead it is very flexible as it can optimize several criteria according to the user's choices for the estimation modes, schemes and normalization constraints.
- Users and researchers often feel uncomfortable especially as compared to the traditional covariance-based SEM.
- Features of a global index:
  - compromise between outer and inner model performance;
  - bounded between a maximum and a maximum

Godness of Fit index

$$GoF = \sqrt{\frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \sum_{p=1}^{P_q} Cor^2\left(\mathbf{x}_{pq}, \xi_q\right)} \times \sqrt{\frac{1}{Q^*} \sum_{q^*=1}^{Q^*} R^2\left(\xi_{q^*}, \xi_j \text{ explaining } \xi_{q^*}\right)}$$

Validation of the outer model

The validation of the outer model is obtained as average of the squared correlations between each manifest variables and the corresponding latent variable, i.e. the average communality!

Validation of the inner model

The validation of the inner model is obtained as average of the R^2 values of all the structural relationships.

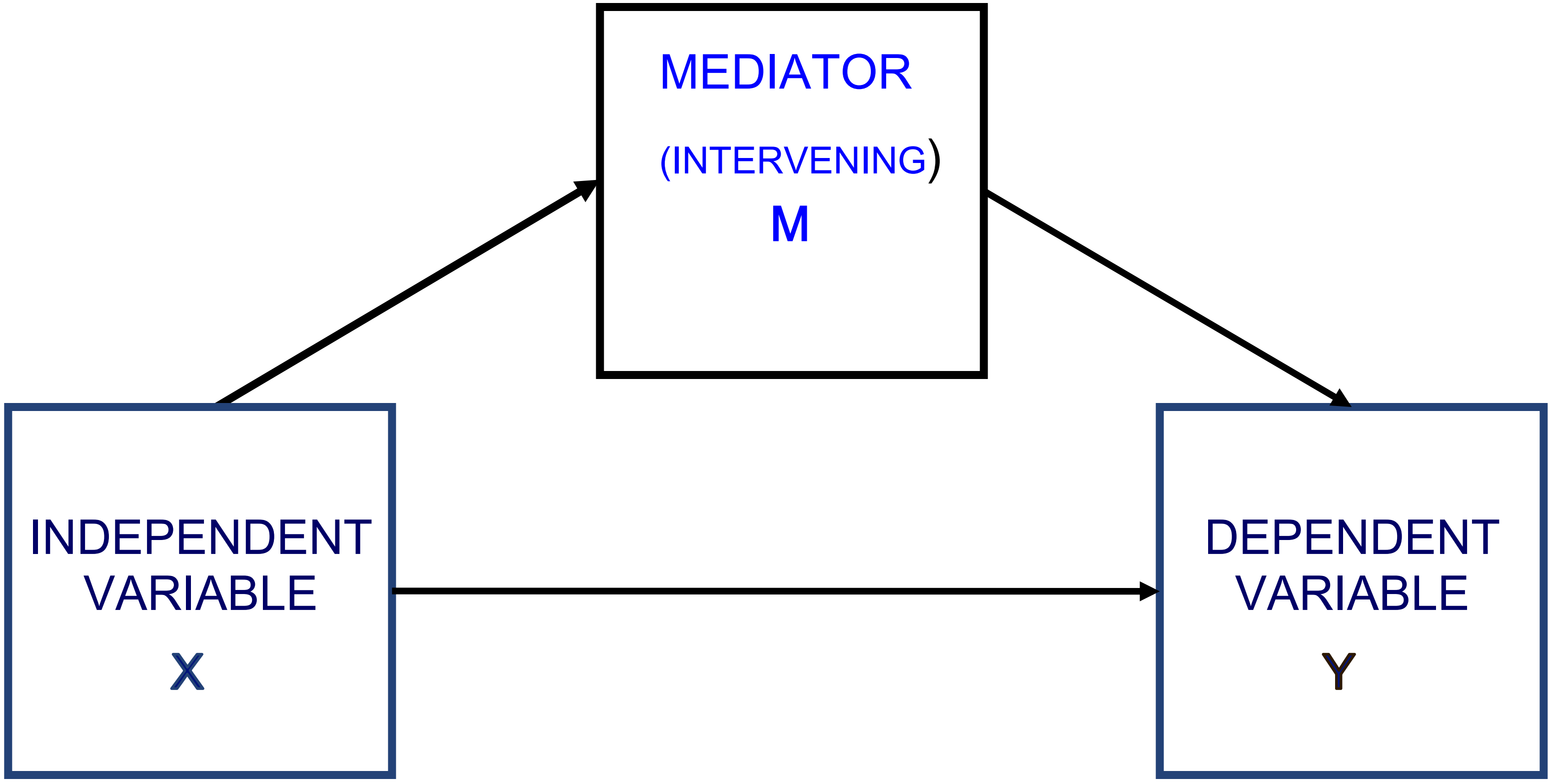
Mediation

Mediated effect

Mediator: a variable that is intermediate in the causal process relating an independent to a dependent variable.

- A mediator is a variable in a chain whereby an independent variable causes the mediator which in turn causes the outcome variable (Sobel, 1990)
- The generative mechanism through which the focal independent variable is able to influence the dependent variable (Baron & Kenny, 1986)
- A variable that occurs in a causal pathway from an independent variable to a dependent variable. It causes variation in the dependent variable and itself is caused to vary by the independent variable (Last, 1988)

Single mediator model



Mediation Causal Steps Test

→ Series of steps described in Judd & Kenny (1981) and Baron & Kenny (1986).

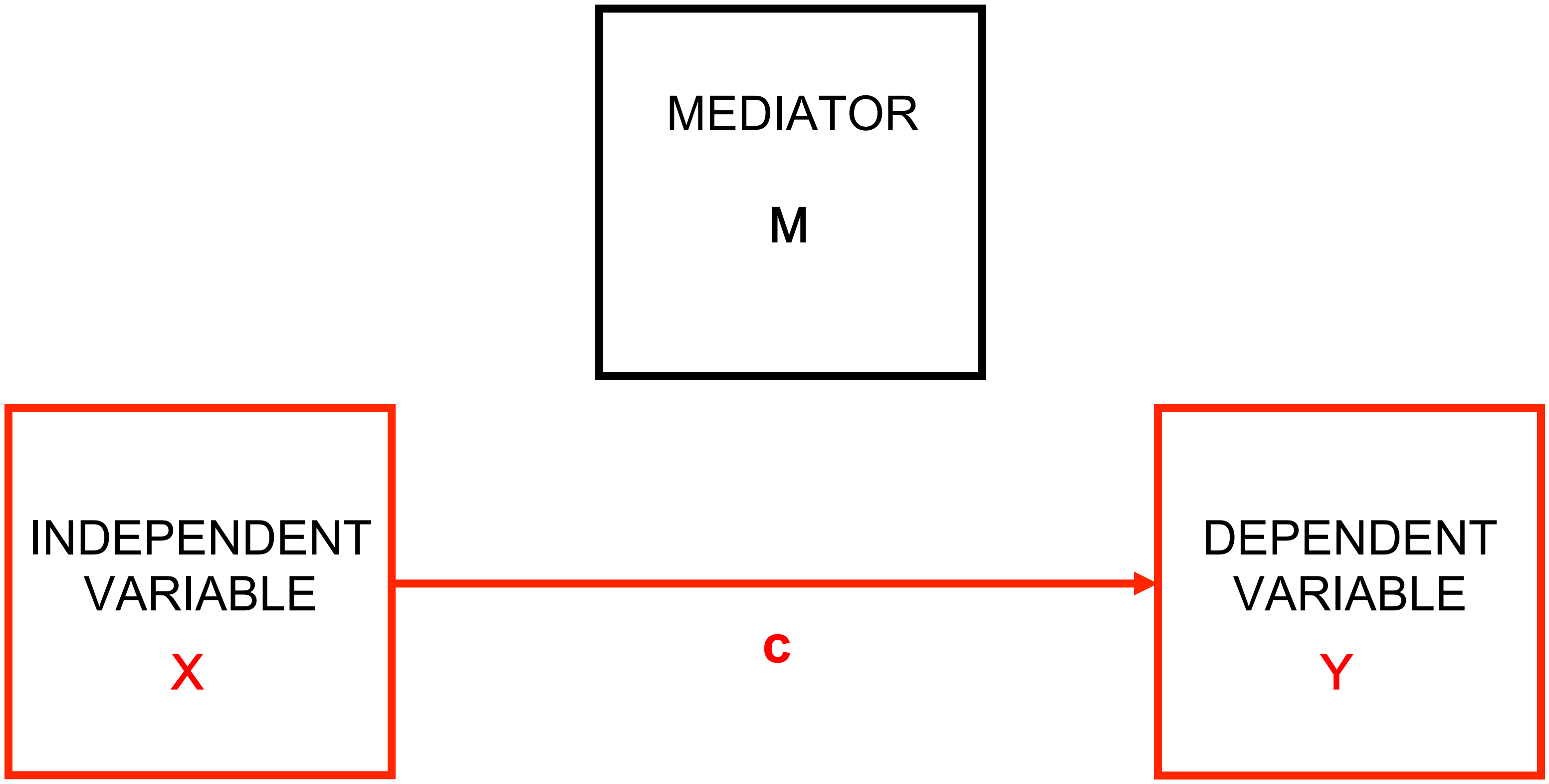
→ One of the most widely used methods to assess mediation in psychology.

→ Consists of a series of tests required for mediation as shown in the next slides.

Mediator model: *Total Effect*

1. The independent variable causes the dependent variable:

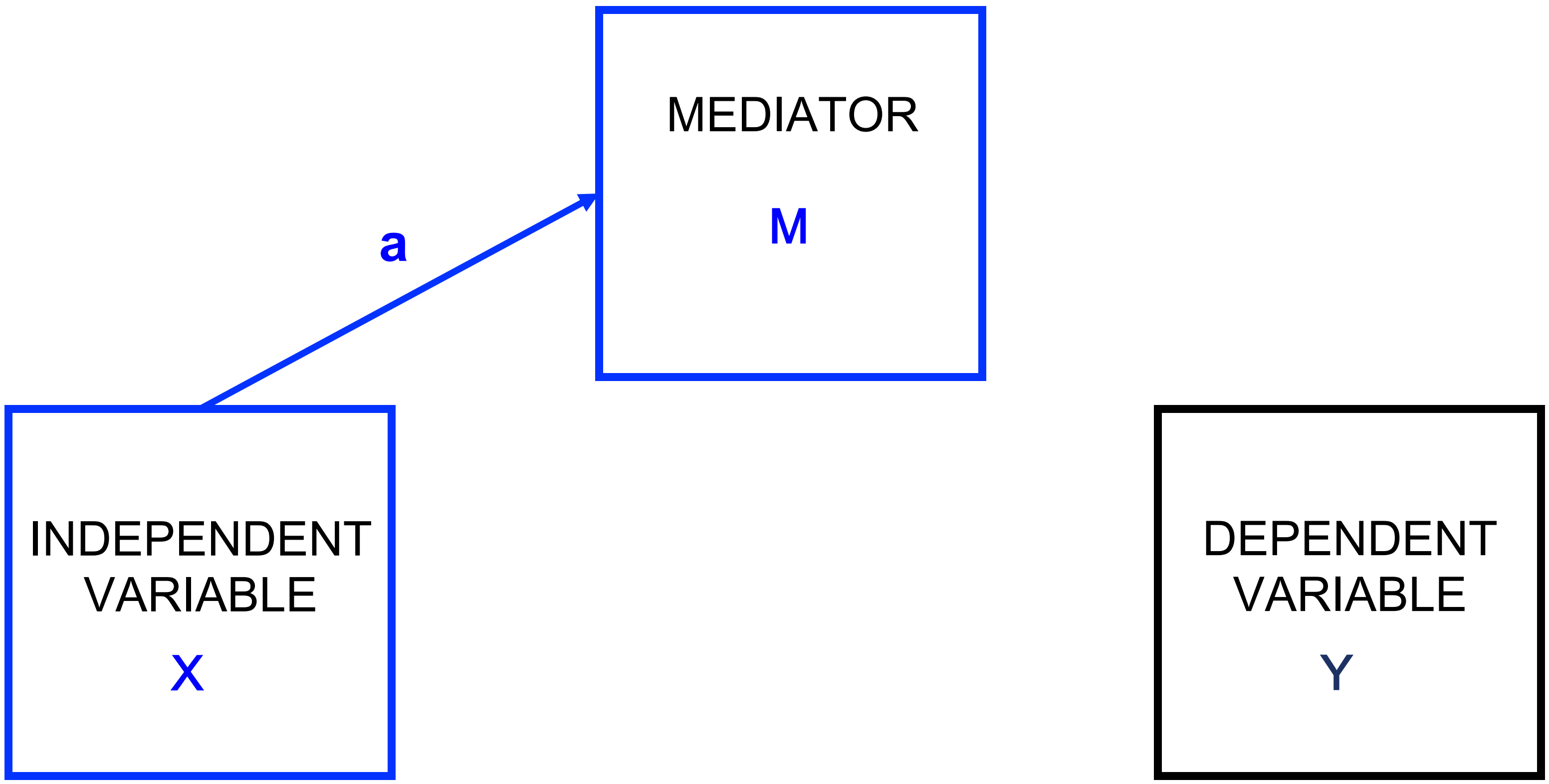
$$Y = i_1 + cX + e_1$$



Mediator model: *Direct effect of X on M*

2. **The independent variable causes the potential mediator:**

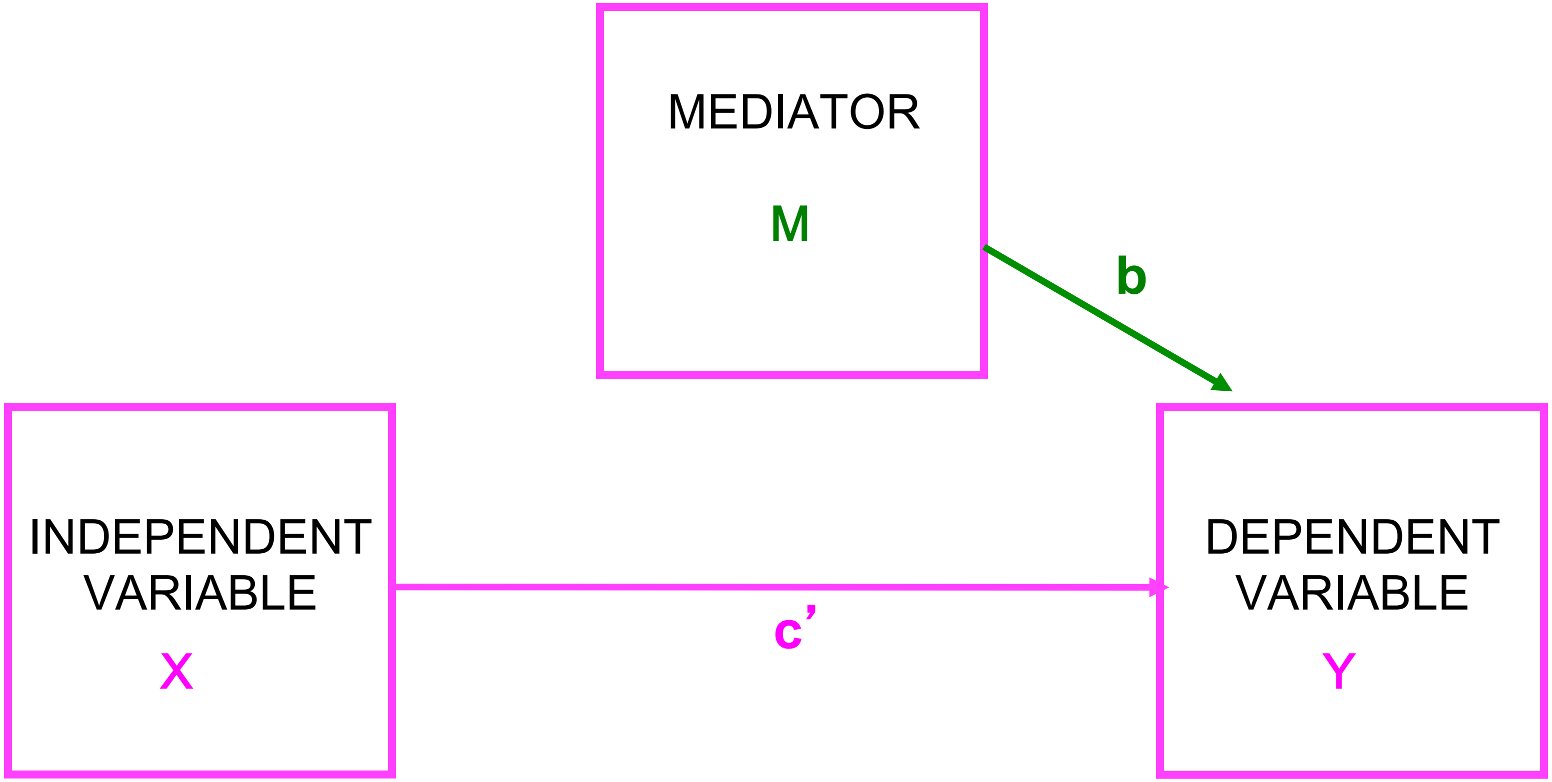
$$M = i_2 + aX + e_2$$



Mediator model: *Direct Effects of M and X on Y*

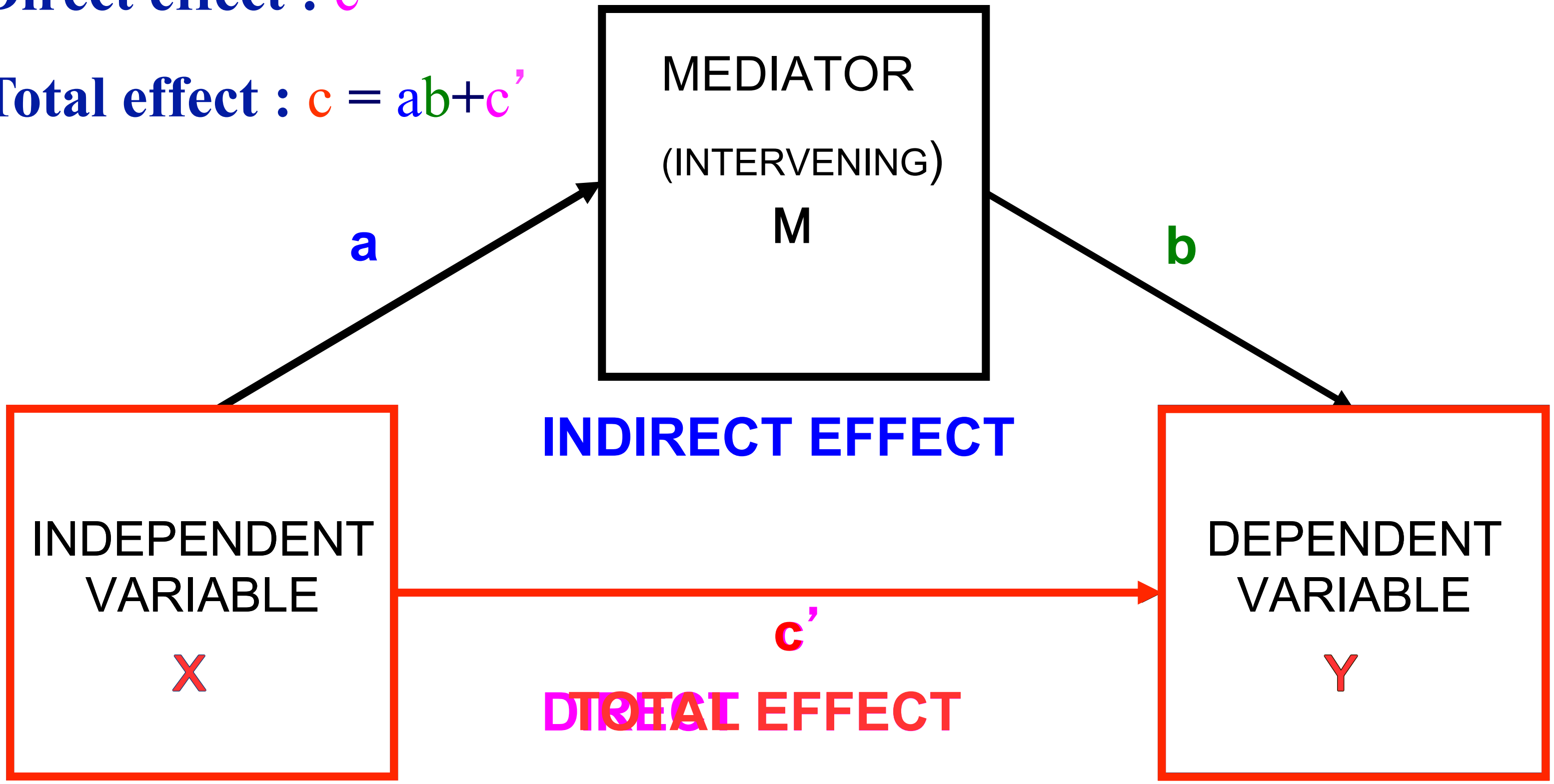
3. The mediator causes the dependent variable controlling for the independent variable:

$$Y = i_3 + c' X + bM + e_3$$



Single mediator model

- **Mediated (Indirect) effect :** ab
- **Direct effect :** c’
- **Total effect :** c = ab+c’



Testing for significant mediation

M is a full (partial) mediator if the following conditions are satisfied:

→ **c is significant**

→ **c'** is not significant (still significant **but less than c**)

→ **Indirect effect ab is significant**:

1. Sobel Test: $$z = \frac{ab}{\sqrt{a^2 s_b^2 + b^2 s_a^2}}$$ → Standard error of the mediated effect

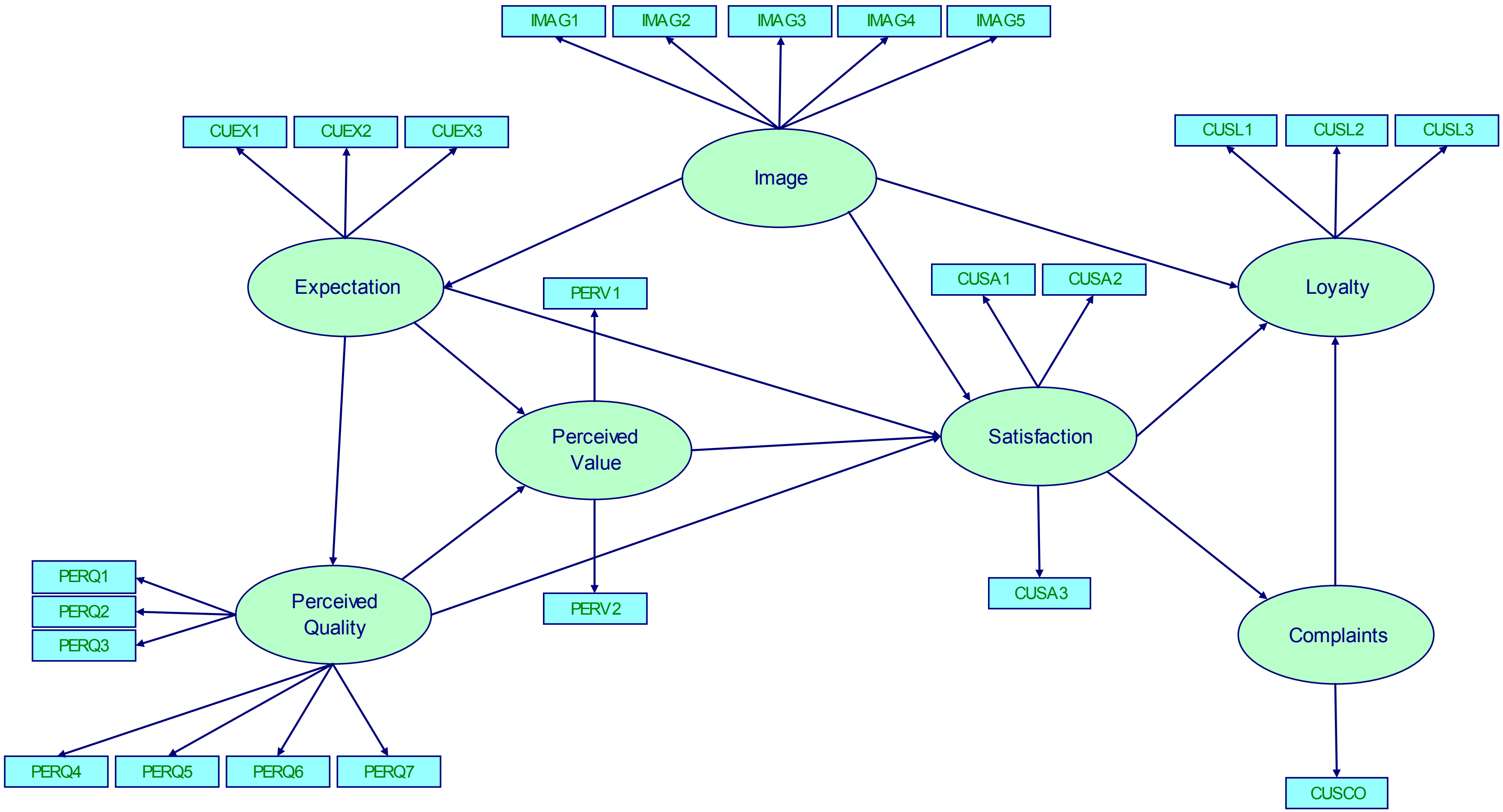
2. Bootstrap confidence interval

PLS-PM

an example for measuring Customer Satisfaction

European Customer Satisfaction Index (ECSI) Model

Perceptions of consumers on one brand, product or service

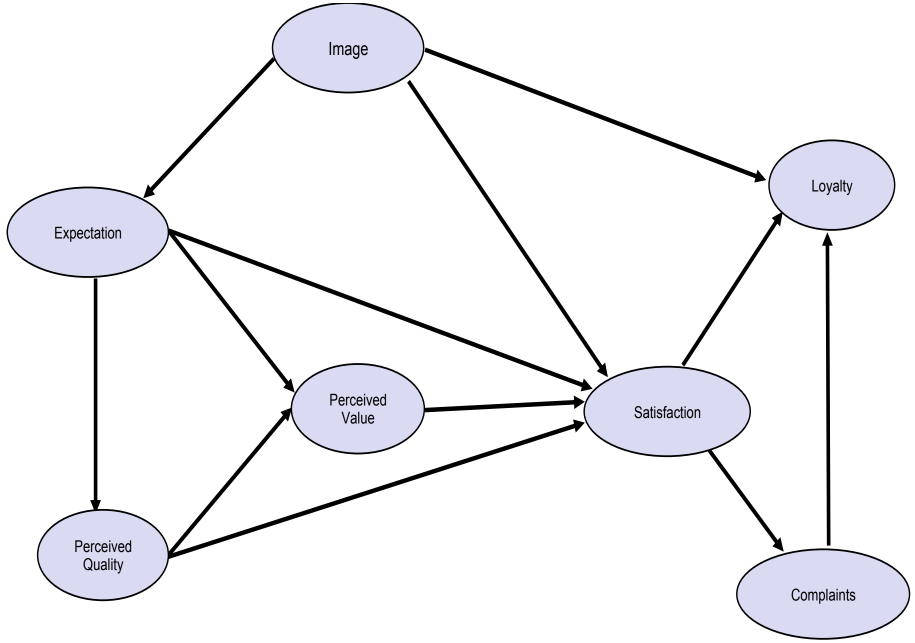


- ECSI is an economic indicator describing how the satisfaction of a customer is modeled
- It is an adaptation of the « Swedish Customer Satisfaction Barometer » and of the « American Customer Satisfaction Index (ACSI) proposed by Claes Fornell

Application to mobile data

Latent variables	Manifest variables
Image (ξ_1)	(a) It can be trusted in what it says and does (b) It is stable and firmly established (c) It has a social contribution for the society (d) It is concerned with customers (e) It is innovative and forward looking
Customer expectations of the overall quality (ξ_2)	(a) Expectations for the overall quality of "your mobile phone provider" at the moment you became customer of this provider (b) Expectations for "your mobile phone provider" to provide products and services to meet your personal need (c) How often did you expect that things could go wrong at "your mobile phone provider"
Perceived quality (ξ_3)	(a) Overall perceived quality (b) Technical quality of the network (c) Customer service and personal advice offered (d) Quality of the services you use (e) Range of services and products offered (f) Reliability and accuracy of the products and services provided (g) Clarity and transparency of information provided
Perceived value (ξ_4)	(a) Given the quality of the products and services offered by "your mobile phone provider" how would you rate the fees and prices that you pay for them? (b) Given the fees and prices that you pay for "your mobile phone provider" how would you rate the quality of the products and services offered by "your mobile phone provider"?
Customer satisfaction (ξ_5)	(a) Overall satisfaction (b) Fulfillment of expectations (c) How well do you think "your mobile phone provider" compares with your ideal mobile phone provider?
Customer complaints (ξ_6)	(a) You complained about "your mobile phone provider" last year. How well, or poorly, was your most recent complaint handled or (b) You did not complain about "your mobile phone provider" last year. Imagine you have to complain to "your mobile phone provider" because of a bad quality of service or product. To what extent do you think that "your mobile phone provider" will care about your complaint?
Customer loyalty (ξ_7)	(a) If you would need to choose a new mobile phone provider how likely is it that you would choose "your provider" again? (b) Let us now suppose that other mobile phone providers decide to lower their fees and prices, but "your mobile phone provider" stays at the same level as today. At which level of difference (in %) would you choose another mobile phone provider? (c) If a friend or colleague asks you for advice, how likely is it that you would recommend "your mobile phone provider"?

All the items measured on a Likert scale from 1 (very negative point of view on the service) to 10 (vey positive point of view on the service)



- Standardized MVs
- Centroid Scheme
- Mode A

Examples of Manifest Variables

Customer expectation

1. Expectations for the overall quality of "your mobile phone provider" at the moment you became customer of this provider.
2. Expectations for "your mobile phone provider" to provide products and services to meet your personal need.
3. How often did you expect that things could go wrong at "your mobile phone provider" ?

Customer satisfaction

1. Overall satisfaction
2. Fulfilment of expectations
3. How well do you think "your mobile phone provider" compares with your ideal mobile phone provider ?

Examples of Manifest Variables

Customer loyalty

1. If you would need to choose a new mobile phone provider how likely is it that you would choose “your provider” again ?

2. Let us now suppose that other mobile phone providers decide to lower fees and prices, but “your mobile phone provider” stays at the same level as today. At which level of difference (in %) would you choose another phone provider ?

3. If a friend or colleague asks you for advice, how likely is it that you would recommend “your mobile phone provider” ?

Final thoughts about PLS and SEM

Component-based methods vs. Factor-based methods

Latent variable or linear composite?

- In component-based SEM the "latent variables" are defined as linear composites or weighted sums of the manifest variables. They are fixed variables (scores)
- In covariance- based SEMs the latent variables are equivalent to common factors. They are theoretical variables

This leads to different parameters to estimate for latent variables, i.e.:

- factor means and variances in covariance-based methods
- weights in component based approaches

Casewise scores are essential in several applications where observations count…
PLS-PM is a component-based method, and we should see this character as a strength.

Prediction-oriented of confirmatory approach?

Reproducing model parameters is not the same thing as making valid predictions about individual observations.

"Factor-based methods are fundamentally unsuitable for prediction, especially for prediction outside the dataset used to estimate the factor model, because of factor indeterminacy" (Rigdon, 2014)

PLS is a prediction-oriented method

Using an **inwards-directed measurement model** in PLS-PM produces higher R^2 values for proxies of endogenous construct. It provides most accurate **in-of-sample prediction**

Using an **outwards-directed measurement model** in PLS-PM produces higher R^2 values in regression with observed variables. It delivers better prediction on **out-of-sample data**

PLS as a SEM estimator

Could we consider PLS-PM as a SEM estimator?

NO, because:

- Lack of **unbiasedness** and **consistency**

YES, because:

- **Consistency at large**, i.e. large number of cases and of indicators for each latent variable ("finite item bias")

- **PLSc** (Dijkstra and Henseler, 2015), **PLS algorithm yield** all the ingredients for obtaining **CAN** (**consistent** and **asymptotically normal**) **estimations** of loadings and LVs squared correlations of a 'clean' second order factor model.

The correction factor for weights is equal to:

$$\hat{c}_q := \sqrt{\frac{\hat{\mathbf{w}}_q'(\mathbf{S}_q - diag(\mathbf{S}_q))\hat{\mathbf{w}}_q}{\hat{\mathbf{w}}_q'(\hat{\mathbf{w}}_q\hat{\mathbf{w}}_q' - diag(\hat{\mathbf{w}}_q\hat{\mathbf{w}}_q'))\hat{\mathbf{w}}_q}}$$

PLS as a SEM estimator: recent standpoints

*"PLS path modeling should separate itself from factor-based SEM and renounce entirely all mechanisms, frameworks and jargon associated with factor models…
Without rejecting rigor, but defining rigor in composite terms…"*

Ed Rigdon (2012)
Rethinking PLSPM: In Praise of Simple Methods
Long Range Planning, 341-358

"I wish to maintain the double-sided nature of PLS that characterized it from the very start. In the family of a structural equations estimators PLS, when properly adjusted, can be a valuable member as well…"

"Our task is to find out which approach works best in which circumstances...Let us establish empirically where each works best. For problems in well-established fields highly structured approaches like mainstream SEM may be appropriate, other fields will be well served by highly efficient means of extracting information from high dimensional data..."

Dijkstra (2014)
PLS' Janus Face – Response to Professor Rigdon's 'Rethinking Partial Least Squares Modeling: In Praise of Simple Methods'
Long Range Planning

Multi-component estimation for Predictive PLS-PM

PLS Regression for outer model regularization in PLS-PM

Integrated PLS Regression-based Approach to PLS-PM algorithm

MVs are centered or standardized

$\mathbf{t}_q$ prop to $\pm\mathbf{X}_q\mathbf{w}_q$ if Mode PLScore

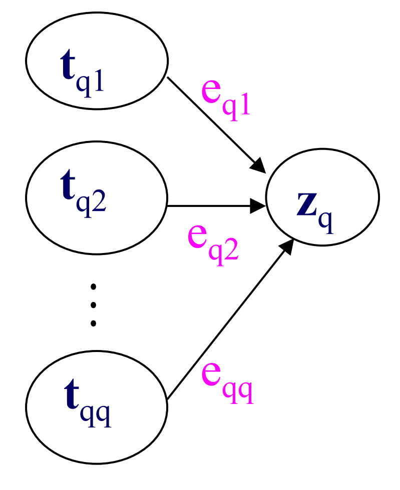
$\mathbf{t}_q = \pm\mathbf{X}_q\mathbf{w}_q$ if Mode PLScow

Outer estimation

Initial step

$\mathbf{w}_q$

Reiterate till *Numerical Convergence*



Inner estimation

Mode PLScore (inwards directed links): **PLS Regression** under the classical PLS-PM **constraints of unitary variance of the composite scores**

Mode PLScow (outwards directed links): **PLS Regression** under the **constraints of normalized outer weights**

Update weights **W**

Choice of weights $\mathbf{e}_{qq'}$:

- **Centroid**: correlation signs
- **Factorial**: correlations
- **Path weighting scheme**: multiple regression coefficients or correlations

PLS Regression rationale

Research of **m** (value chosen by **cross-validation** or **defined by the user**) **orthogonal** components $\mathbf{v}_{mq} = \mathbf{X}_q \mathbf{a}_{mq}$ which are as **correlated** as possible to $\mathbf{z}_q$ (from the inner estimation step) and **also explanatory** of their own block $\mathbf{X}_q$.

$$\mathrm{Cov}^2(\mathbf{X}_q\mathbf{a}_{mq}, \mathbf{z}_q) = \mathrm{Cor}^2(\mathbf{X}_q\mathbf{a}_{mq}, \mathbf{z}_q) * \mathrm{Var}(\mathbf{X}_q\mathbf{a}_{mq})$$

PLS1 (regression) Mode leads to a **compromise** between a **multiple regression** of $\mathbf{z}_q$ on $\mathbf{X}_q$ **(Mode B)** and a **principal component analysis** of $\mathbf{X}_q$ **(Mode A for a single block)**

PLS Regression algorithm in PLS-PM

1. **First** PLS component $\mathbf{v}_{1q}$ (with $\mathbf{x}_{pq}$ standardized as well):

$$\mathbf{v}_{1q} = \mathbf{X}_q \mathbf{a}_{1q} = \frac{1}{\sqrt{\sum_p cor^2\left(\mathbf{z}_q, \mathbf{x}_{pq}\right)}} \sum_p cor\left(\mathbf{z}_q, \mathbf{x}_{pq}\right) \times \mathbf{x}_{pq}$$

2. **Normalization** of the vector $\mathbf{a}_{1q} = (a_{11q}, \ldots, a_{1pq})$
3. **Regression** of $\mathbf{z}_q$ on $\mathbf{v}_{1q} = \mathbf{X}_q \mathbf{a}_{1q}$ expressed in terms of $\mathbf{X}_q$
4. Computation of the **residuals** $\mathbf{z}_{q1}$ and $\mathbf{X}_{q1}$ of the **regressions** of $\mathbf{z}_q$ and $\mathbf{X}_q$ on $\mathbf{v}_{1q}$: $\mathbf{z}_q = \mathbf{c}_{1q}\mathbf{v}_{1q} + \mathbf{z}_{q1}$ and $\mathbf{X}_q = \mathbf{v}_{1q}\mathbf{p}'_{1q} + \mathbf{X}_{q1}$

For successive components the procedure is **iterated** on **residuals** and assessed by means of **cross-validation** or stopped by the user

PLS Regression algorithm in PLS-PM

Finally, the m-components PLS regression model yielding the weights for the outer estimate, as each component can be expressed as a function of **X** :

$$
\begin{aligned}
\mathbf{z}_q &= c_{1q}\mathbf{v}_{1q} + c_{2q}\mathbf{v}_{2q} + \ldots + c_{mq}\mathbf{v}_{mq} + res \\
&= c_{1q}\mathbf{X}_q\mathbf{a}_{1q} + c_{2q}\mathbf{X}_{1q}\mathbf{a}_{2q} + \ldots + c_{mq}\mathbf{X}_{m-1q}\mathbf{a}_{mq} + res \\
&= c_{1q}\mathbf{X}\mathbf{a}_{1q} + c_{2q}\mathbf{X}\mathbf{a}^*_{2q} + \ldots + c_{mq}\mathbf{X}\mathbf{a}^*_{mq} + res \\
&= \mathbf{X}_q\left(c_{1q}\mathbf{a}_{1q} + c_{2q}\mathbf{a}^*_{2q} + \ldots + c_{mq}\mathbf{a}^*_{mq}\right) + res \\
&= \underbrace{\mathbf{X}_q\mathbf{w}_q + res = w_{1q}\mathbf{x}_{1q} + w_{2q}\mathbf{x}_{2q} + \ldots + w_{pq}\mathbf{x}_{pq}}_{\mathbf{t}_q} + res
\end{aligned}
$$

Further transformed so as to satisfy the classical normalization constraint: Var(t_q)=1

Features of the integrated PLS approach

- **No need to invert $X_q'X_q$** (i.e. takes full advantage of the NIPALS algorithmic approach)
- Decomposition into common (explanatory) and distinctive dimensions
- **Criterion of fairness** across blocks, i.e. takes into account heterogeneous levels of noise
- **Number of dimensions** in each block chosen in coherence with a **prediction purpose**
- **Choosing** a **different number of dimensions** per block **does not affect normalization constraints**

Two possible normalization constraints for PLS regression Modes

		Normalization constraints on	
		Outer weights (like in RGCCA)	Composite scores (like in PLS-PM)
Oriented to	Covariances between LVs	**PLScow**	
	Correlations between LVs		**PLScore**

PLScore Mode:

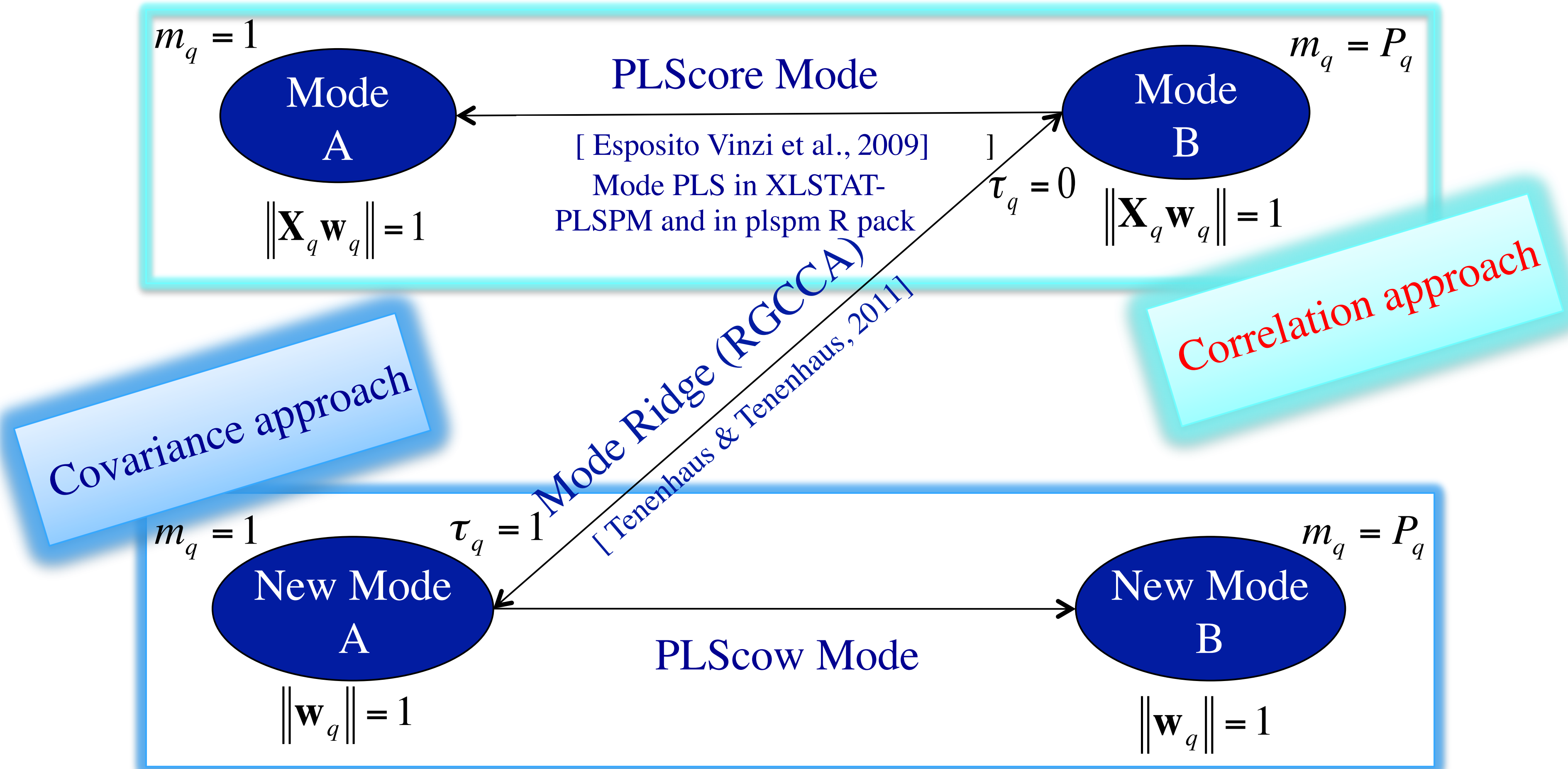
PLS Mode oriented to maximizing **cor**relations between connected composites under normalization constraints on composite s**core**s

PLScow Mode:

PLS Mode oriented to maximizing **cov**ariances between connected composites under normalization constraints on outer **w**eights

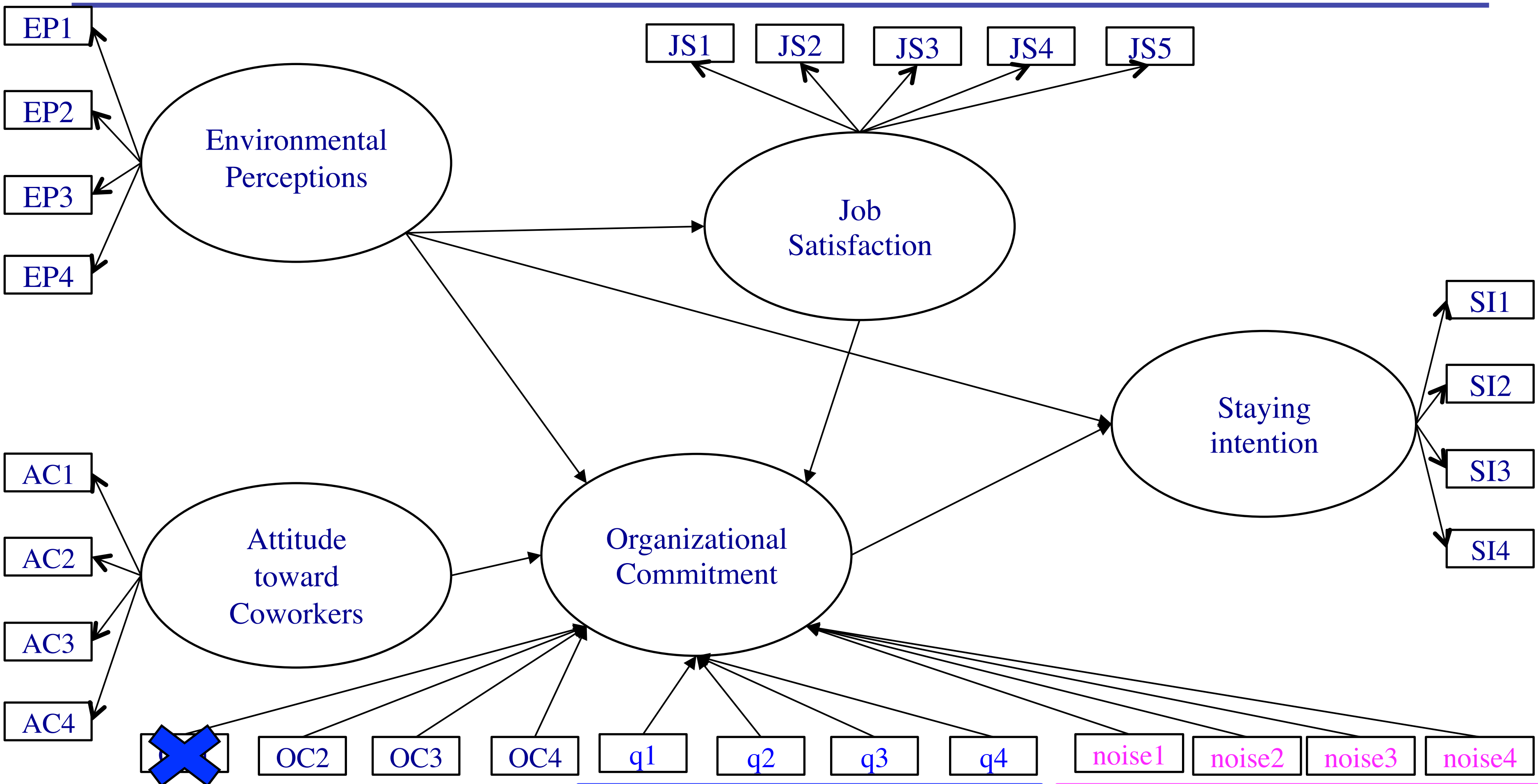
PLS regression Modes in PLS-PM and Ridge Mode in RGCCA

PLS-R as an estimation method for measurement model in standard PLS-PM (normalization constraints on composite scores)



PLS-R as an estimation method for measurement model in a modified PLS-PM (normalization constraints on outer weights)

Hbat Model (Hair et al., 2010) with noisy variables

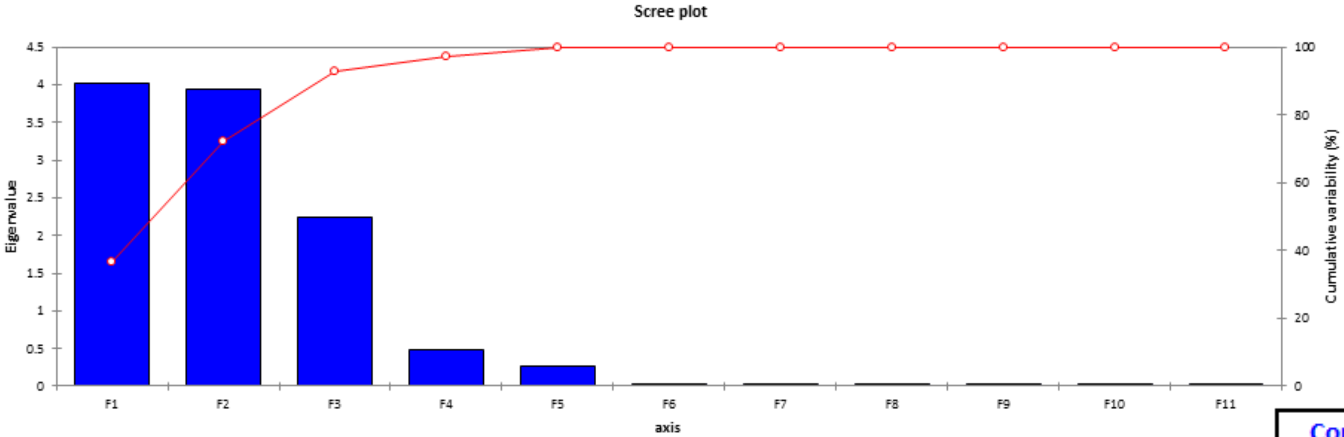


q variables: highly correlated among them, correlated with the MVs of the response block SI and uncorrelated with all the others variables in the model

noise variables: highly correlated among them and uncorrelated with all the others variables in the model. In particular orthogonal to the variables related to the response block SI.

PCA of the Org. Commitment (OC + noisy data)

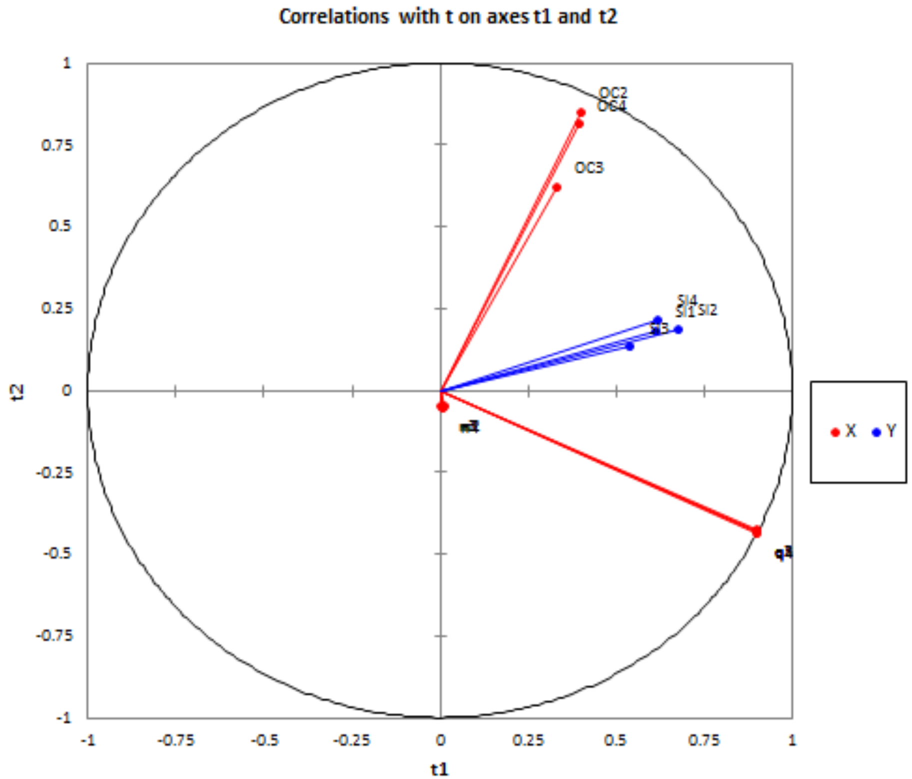
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
Eigenvalue	**4.021**	**3.947**	**2.246**	0.493	0.262	0.007	0.007	0.006	0.005	0.004	0.003
Variability (%)	36.555	35.884	20.416	4.478	2.380	0.066	0.061	0.054	0.042	0.038	0.026
Cumulative %	36.555	72.440	92.856	97.334	99.714	99.779	99.840	99.894	99.936	99.974	100.000



The real OC manifest variables appear only on the 3rd PC

Correlation with factors	F1	F2	F3
OC2	0.000	0.006	**0.892**
OC3	0.000	0.004	**0.806**
OC4	0.000	0.006	**0.895**
q1	0.676	**0.733**	0.000
q2	0.665	**0.743**	-0.004
q3	0.672	**0.737**	0.002
q4	0.670	**0.740**	-0.008
noise1	**0.745**	-0.665	0.001
noise2	**0.749**	-0.660	0.005
noise3	**0.742**	-0.668	0.003
noise4	**0.745**	-0.664	0.000

PLS Regression of the OC noisy data on Staying Intention (SI)

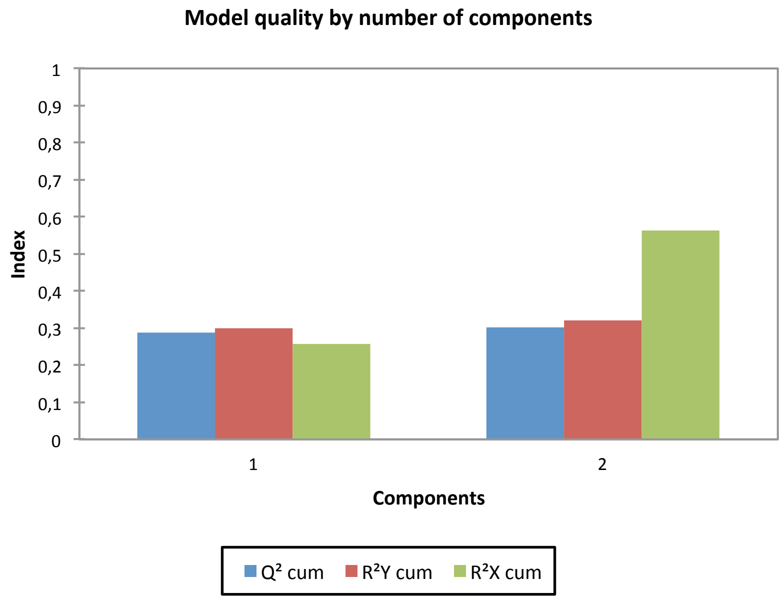


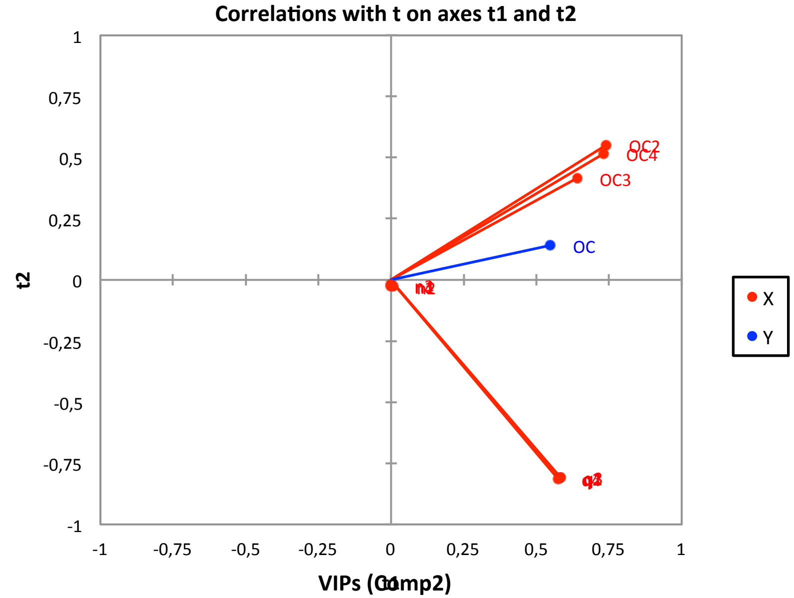
The noise variables are downweighted as they have no predictive power

Variable	t1	t2
OC2	0.400	0.850
OC3	0.330	0.623
OC4	0.394	0.814
q1	0.900	-0.428
q2	0.898	-0.432
q3	0.901	-0.427
q4	0.896	-0.436
n1	**0.006**	**-0.049**
n2	**0.013**	**-0.049**
n3	**0.002**	**-0.046**
n4	**0.006**	**-0.051**

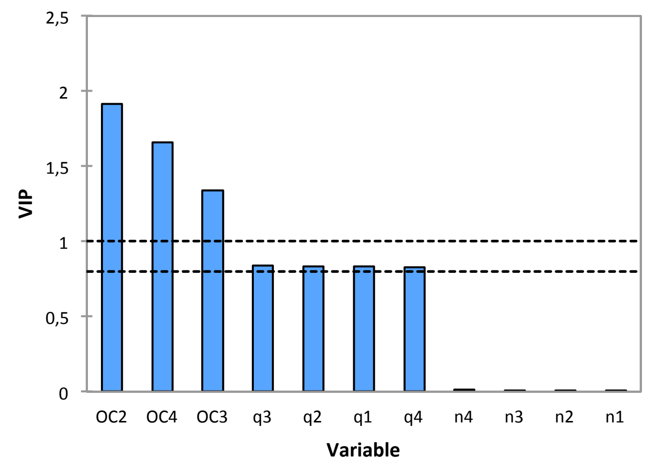
Variable	w*1	w*2
OC2	0.380	0.771
OC3	0.214	0.204
OC4	0.326	0.558
q1	0.416	-0.182
q2	0.420	-0.159
q3	0.423	-0.152
q4	0.417	-0.170
n1	**0.000**	**-0.015**
n2	**0.003**	**-0.017**
n3	**-0.002**	**-0.015**
n4	**-0.001**	**-0.019**

PLS Regression for the OC outer model in PLS-PM





Variable	t1	t2
OC2	0.740	0.552
OC3	0.640	0.415
OC4	0.731	0.516
q1	0.581	-0.808
q2	0.577	-0.811
q3	0.583	-0.808
q4	0.575	-0.814
n1	0.001	-0.024
n2	0.007	-0.028
n3	-0.001	-0.020
n4	0.000	-0.026
OC	0.548	0.141



A comparison between Modes PLScore, A and B: outer weights

	Mode PLScore	Mode A	Mode B
OC2	0.435	0.361	-0.655
OC3	0.277	0.258	-0.156
OC4	0.358	0.317	-0.222
q1	0.088	0.144	0.656
q2	0.090	0.145	-0.228
q3	0.092	0.147	-0.225
q4	0.090	0.144	-0.563
n1	0.000	0.000	-0.589
n2	0.000	0.001	-0.093
n3	-0.002	-0.001	0.374
n4	-0.003	-0.002	0.304

Non-Metric PLS-PM

Steven's measurement scale classification

Scale	Basic empirical operations	Mathematical group structure	Permissible statistics
NOMINAL	Determination of equality	Permutation group	mode, chi square
ORDINAL	Determination of greater or less	Isotonic group	median, percentile
INTERVAL	Determination of equality of intervals or differences	General linear group	mean, standard deviation, product moment and rank order correlations
RATIO	Determination of equality of ratios	Similarity group	geometric mean, harmonic mean, coefficient of variation

- Interval and Ratio scales are METRIC structures, i.e. sets where notion of distance (metric) between elements of the set is defined.
- Nominal and Ordinal scales are NON-METRIC structures (unordered and ordered sets).
- Statistical analyses based on Pearson's correlation should be performed only on metric variables.

Ordinal vs Nominal variables

Nominal and ordinal variables are categorical variables, i.e. variables that associate each observation to one of the m groups defined by their categories.

From the mathematical point of view, they are similar:

- Both are not continuous variables
- Both have no metric properties
- Both do have no origin or units of measurements

The only difference between nominal and ordinal variables is that groups defined by categories of an ordinal variable can be conceptually ordered.

PLS-PM assumptions

Two basic **assumptions** underlying PLS models:

- Each variable is measured on a **interval (or ratio) scale**.
- Relationships between variables and latent constructs are **linear** and, consequently, **monotonic**.

However, in practice:

- Nominal variables are handle using boolean coding
- Ordinal variables (e.g. likert scale items) are coded by numerals (1,2,3..)
- Linearity is almost never checked

Three good pratical reasons..

.. To NOT use boolean coding in PLS-PM

1) The numbers of categories affects the relative impact of categorical variables and generates sparse matrices.

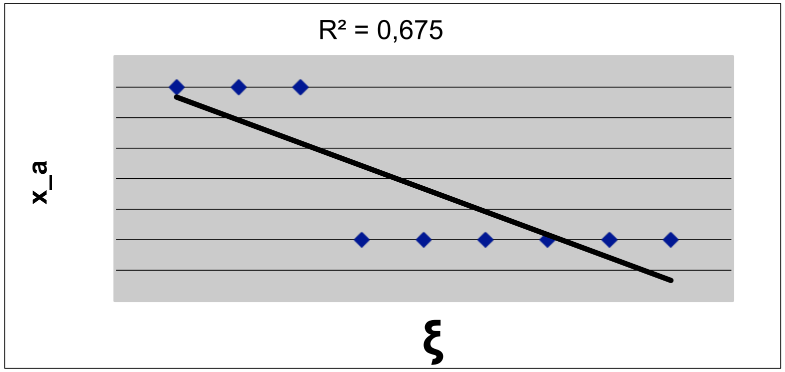
2) It measures the impact of the single category, giving up the idea of the variable as a whole

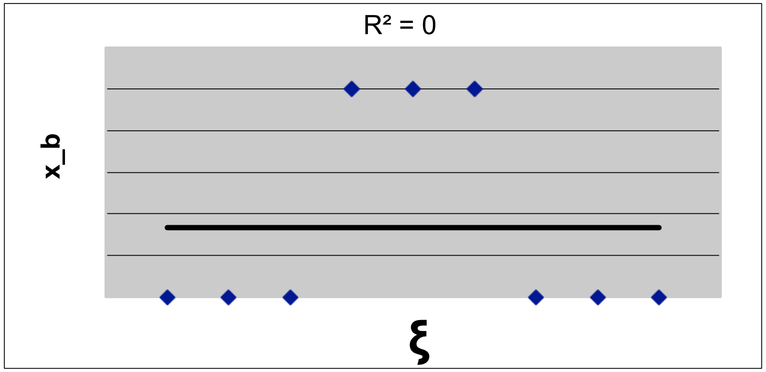
3) The importance of categories associated to central values of the LV distribution is systematically underestimated.

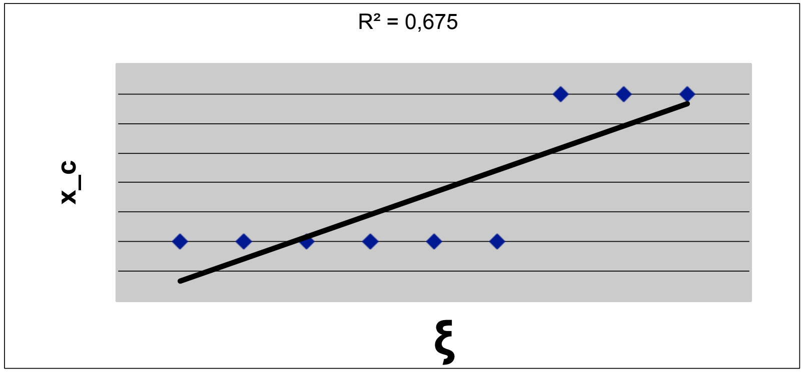
The relation between z_q and x_{pq}

The weight of a MV depends on the linear relation between the MV and the LV inner estimate

ID	z	x	x_a	x_b	x_c
obs1	1	a	1	0	0
obs2	2	a	1	0	0
obs3	3	a	1	0	0
obs4	4	b	0	1	0
obs5	5	b	0	1	0
obs6	6	b	0	1	0
obs7	7	c	0	0	1
obs8	8	c	0	0	1
obs9	9	c	0	0	1







Ordinal variables in linear models

- “Ordinal variables are not continuous variables and should not be treated as if they were”.
- “It is common practice to treat scores 1,2,3…. assigned to categories as if they have metric properties but this is wrong.”
- “Ordinal variables do not have origins or units of measurements”
- “To use ordinal variables in SEM requires other techniques than those that are traditionally used with continuous variables”

Jöreskog (1994) speaking about covariance-base SEM

These statements are valid in PLS-PM framework too!

Scaling

- Scaling a variables means providing the variable with a metric: each observed category (or distinct value) of the raw (i.e. to be scaled) variable is replaced by a numerical value.

- The new scale is an interval scale, independently of the properties of the initial measurement scale.

- Scaling techniques are generally used to convert a WEAKER measurement scale INTO A STRONGER measurement scale..

- However, it can be useful to RE-SCALE a metric variable by providing it with a DIFFERENT metric..

Scaling Level

- We don't need to retain all of the properties of the initial measurement scale of the variable.
- The scaling level is defined by the the properties of the initial measurement scale that the reseacher choose to retain in the new measurement scale

Optimal Scaling (OS)

To define a scaling process as optimal, the scaling parameter estimates must be:

→ Suitable, as it must respect the constraints defined by the scaling level

→ Optimal, as it must optimize the same criterion of the analysis in which the OS process is involved.

Non-Metric Partial Least Squares

Electronic Journal of Statistics
Vol. 6 (2012) 1641–1669
ISSN: 1935-7524
DOI: 10.1214/12-EJS724

Non-Metric Partial Least Squares

Giorgio Russolillo

Laboratoire CEDRIC,
Conservatoire National des Arts et Métiers
292 rue Saint Martin 75141 Paris - France
e-mail: giorgio.russolillo@cnam.fr

Non-Metric Partial Least Squares

The OS principle, applied to PLS-PM, allows us:

- Handling numerical, ordinal and nominal variables in the same model
- Checking and/or adjusting the data for non-linearity and non-monotonicity
- Dealing with outliers
- Suggesting a discretization process

- Each raw variable is transformed as $\hat{\mathbf{x}} \propto \tilde{\mathbf{X}}\phi$, where $\phi' = (\phi_1 \dots \phi_K)$ is the vector of optimal scaling parameters and the matrix $\tilde{\mathbf{X}}$ defines a space in which constraints imposed by the scaling level are respected.

- Optimal quantification are calculated by means of a PLS-based iterative algorithm

Non-Metric PLS Path Modeling algorithm

A **new PLS algorithm** which works (also) as an optimal scaling algorithm: NM-PLSPM assigns a scaling (numeric) value to each category (or distinct value) k ($k = 1 \ldots K \leq N$) of raw variables x, such that

- It is coherent with the chosen scaling level;
- It optimizes the PLS criterion, if any.

Outer weights and scaling parameters are alternately optimized in a modified PLS loop where a quantification step is added.

→ In standard PLS steps the outer weights are optimized for given scaling values.

→ In the quantification step, instead, the scaling values are optimized for given outer weights: raw variables are properly transformed through scaling (quantification) functions *Q()*

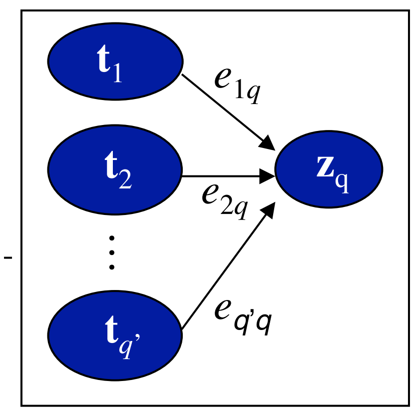
NM-PLSPM algorithm iteration

Outer Estimation

$$\mathbf{t}_q = \mathbf{X}_q \mathbf{w}_q$$

Updating the inner weights $e_{q'q}$:

Centroid, Factorial or Path weighting scheme



Inner Estimation

$$\hat{\mathbf{x}}_{pq} \propto Q\left(\tilde{\mathbf{X}}_{pq}, \mathbf{z}_q\right) \quad \text{(Mode (new) A)}$$

$$\hat{\mathbf{x}}_{pq} \propto Q\left(\tilde{\mathbf{X}}_{pq}, \mathbf{x}^*_{pq}\right) \quad \text{(Mode B)}$$

Quantification step

$$\mathbf{w}_q \propto (1/n)\mathbf{X}'_q \mathbf{z}_q \quad \text{(Mode A)}$$

$$\mathbf{w}_q \propto \left(\mathbf{X}'_q \mathbf{X}_q\right)^{-1} \mathbf{X}'_q \mathbf{z}_q \quad \text{(Mode B)}$$

Updating the outer weights
(Normalization depending on the Mode)

NM-PLSPM general criterion

$$\underset{\mathbf{w}_q,\,\phi_{pq},\,\tilde{\mathbf{X}}_{pq}}{\arg\max}\left\{\sum_{q\neq q'} c_{qq'}\, g\left[cor\left(\hat{\mathbf{X}}_q\mathbf{w}_q, \hat{\mathbf{X}}_{q'}\mathbf{w}_{q'}\right)\sqrt{\mathrm{var}\left(\hat{\mathbf{X}}_q\mathbf{w}_q\right)}\sqrt{\mathrm{var}\left(\hat{\mathbf{X}}_{q'}\mathbf{w}_{q'}\right)}\right]\right\}$$

$$\text{s.t.}\quad \left\|\hat{\mathbf{x}}_{pq}\right\|^2 = \left\|\tilde{\mathbf{X}}_{pq}\phi_{pq}\right\|^2 = n$$

$$\left\|\hat{\mathbf{X}}_q\mathbf{w}_q\right\|^2 = n \quad \text{if Mode B for block } q$$

$$\left\|\mathbf{w}_q\right\|^2 = n \quad \text{if New Mode A for block } q$$

Each time the PLS-PM algorithm converges to a criterion, the corresponding Non-Metric version converge to the same criterion

PLSPM R-package

The NN-PLSPM algorithm is implemented in the R-package **plspm**:

plspm: Tools for Partial Least Squares Path Modeling (PLS-PM)

plspm contains a set of functions for performing Partial Least Squares Path Modeling (PLS-PM) analysis for both metric and non-metric data, as well as REBUS analysis.

Version:	0.4.1
Depends:	R (≥ 3.0.1), amap, diagram, tester, turner
Suggests:	plsdepot, FactoMineR, ggplot2, reshape, testthat, knitr
Published:	2013-12-08
Author:	Gaston Sanchez [aut, cre], Laura Trinchera [aut], Giorgio Russolillo [aut]
Maintainer:	Gaston Sanchez <gaston.stat at gmail.com>
License:	GPL-3
URL:	http://www.gastonsanchez.com http://www.plsmodeling.com
NeedsCompilation:	no

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Two types of quantification are currently allowed:

- Nominal Scaling, in which the following group constraint is considered:

$$\left(x_i \sim x_{i'}\right) \Rightarrow \left(\hat{x}_i = \hat{x}_{i'}\right)$$

- Ordinal scaling, in which a further order constraint is considered:

$$\left(x_i^* \sim x_{i'}^*\right) \Rightarrow \left(\hat{x}_i = \hat{x}_{i'}\right) \quad \text{and} \quad \left(x_i^* \prec x_{i'}^*\right) \Rightarrow \left(\hat{x}_i \leq \hat{x}_{i'}\right)$$

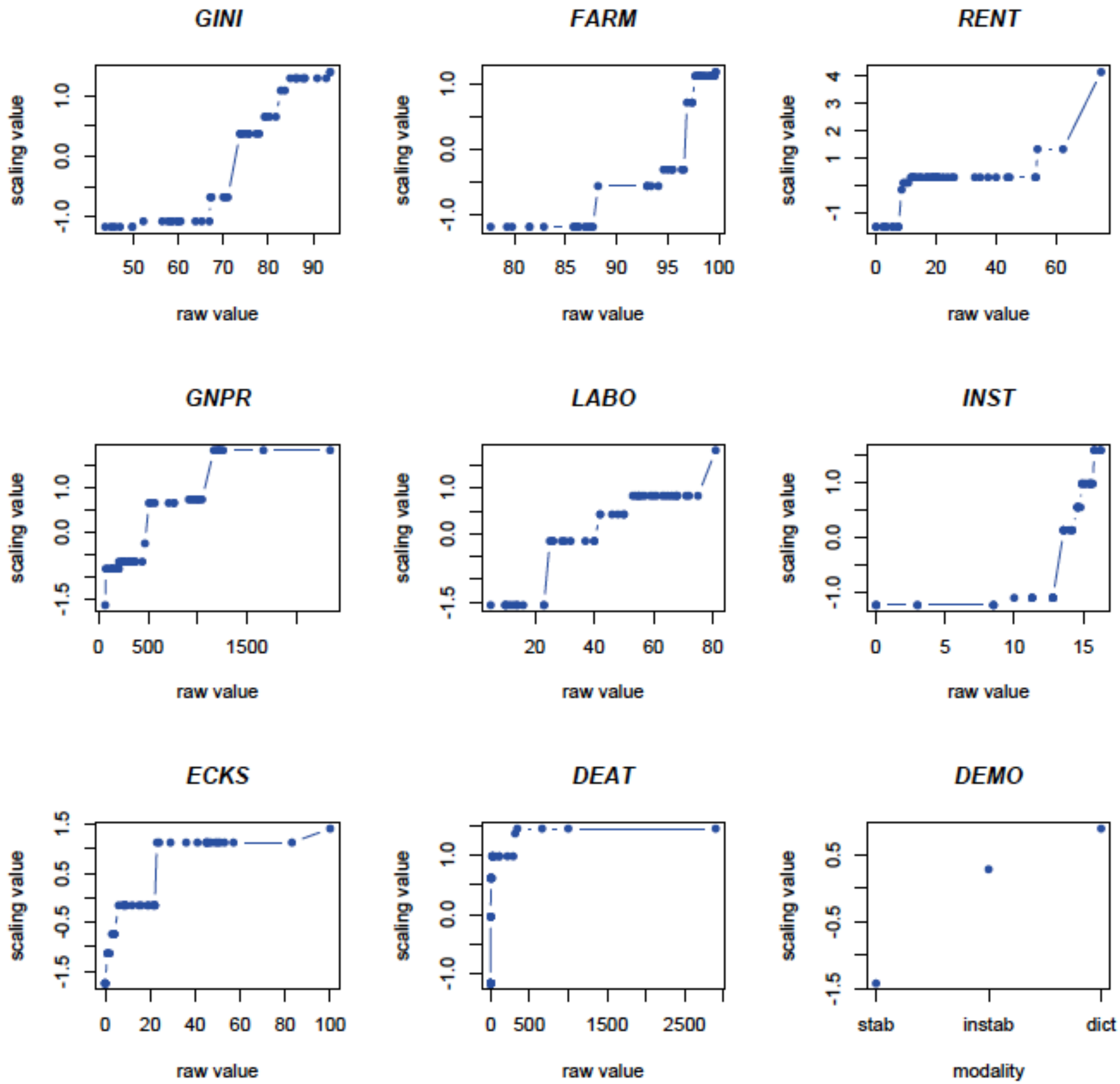
An application to the Russett data (1965)

- gini: Gini's index of concentration;
- farm: complement of the percentage of farmers that own half of the lands, starting with the smallest ones. Thus if farm is 90%, then 10% of the farmers own half of the lands;
- rent: percentage of farm households that rent all their land.

- gnpr: gross national product pro capite (in U.S. dollars) in 1955;
- labo: the percentage of labor force employed in agriculture.

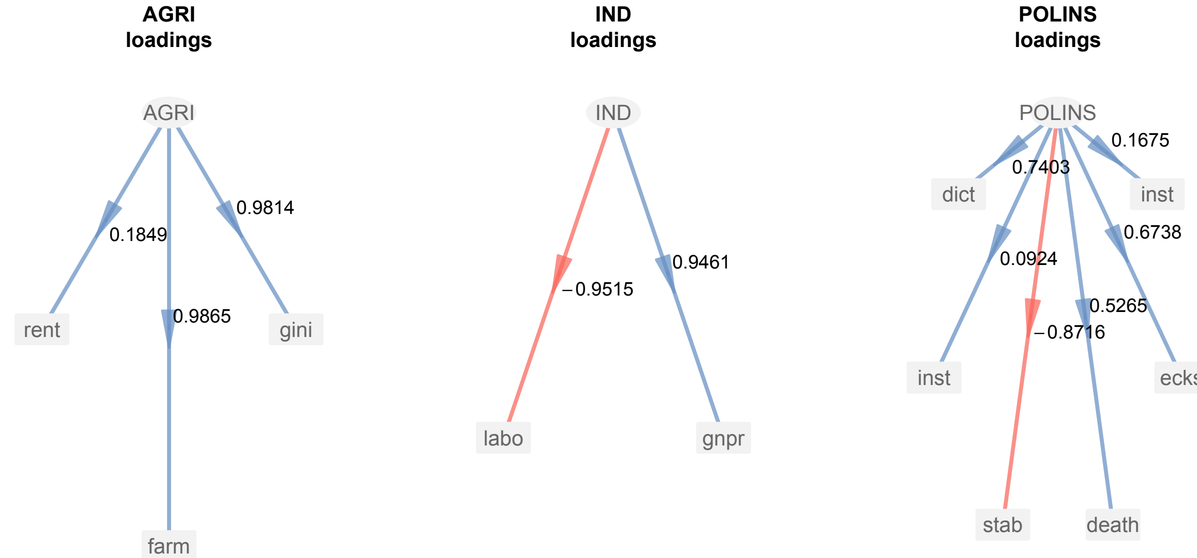
- inst: an index, bounded from 0 (stability) to 17 (instability), calculated as a function of the number of the chiefs of the executive and of the number of years of independence of the country during the period 1946-1961;
- ecks: the Eckstein's index, which measures the number of violent internal war incidents during the same period;
- death: number of people killed as a result of violent manifestations during the period 1950-1962;
- demo: a categorical variable that classifies countries in three groups: stable democracy, unstable democracy and dictatorship.

Russet data (1964): Quantifications

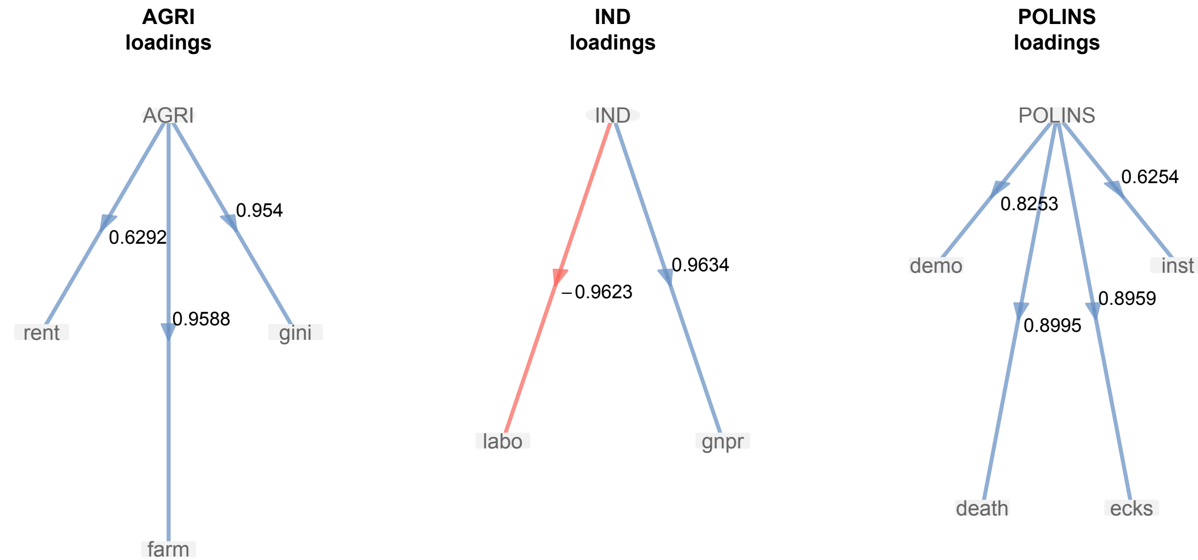


Russet data (1964): Model comparison

PLS Model:
- R2 = .605
- GoF = .567



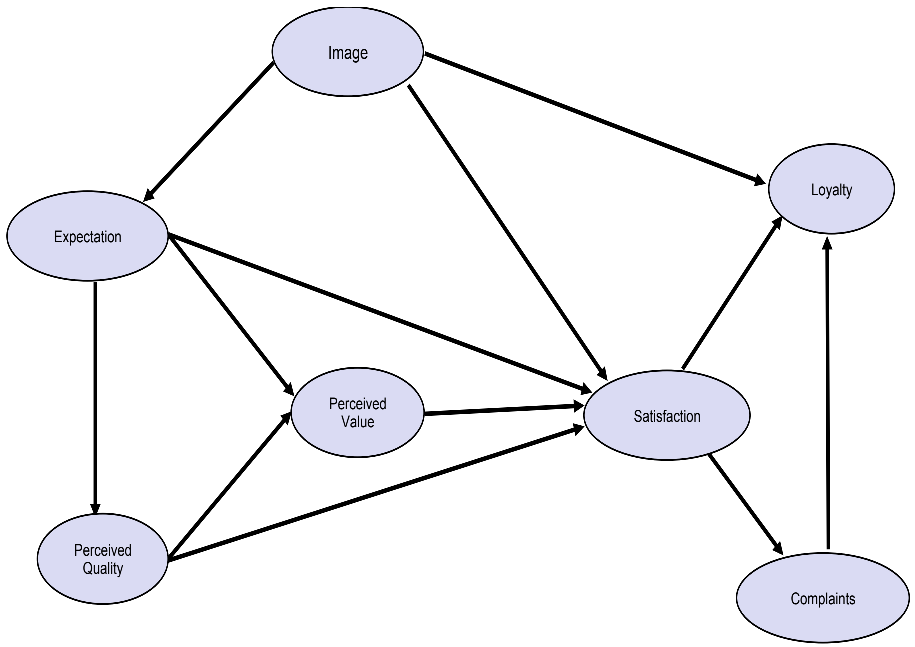
NM-PLS Model:
- R2 = .793
- GoF = .772



Application to mobile data

Latent variables	Manifest variables
Image (ξ_1)	(a) It can be trusted in what it says and does (b) It is stable and firmly established (c) It has a social contribution for the society (d) It is concerned with customers (e) It is innovative and forward looking
Customer expectations of the overall quality (ξ_2)	(a) Expectations for the overall quality of "your mobile phone provider" at the moment you became customer of this provider (b) Expectations for "your mobile phone provider" to provide products and services to meet your personal need (c) How often did you expect that things could go wrong at "your mobile phone provider"
Perceived quality (ξ_3)	(a) Overall perceived quality (b) Technical quality of the network (c) Customer service and personal advice offered (d) Quality of the services you use (e) Range of services and products offered (f) Reliability and accuracy of the products and services provided (g) Clarity and transparency of information provided
Perceived value (ξ_4)	(a) Given the quality of the products and services offered by "your mobile phone provider" how would you rate the fees and prices that you pay for them? (b) Given the fees and prices that you pay for "your mobile phone provider" how would you rate the quality of the products and services offered by "your mobile phone provider"?
Customer satisfaction (ξ_5)	(a) Overall satisfaction (b) Fulfillment of expectations (c) How well do you think "your mobile phone provider" compares with your ideal mobile phone provider?
Customer complaints (ξ_6)	(a) You complained about "your mobile phone provider" last year. How well, or poorly, was your most recent complaint handled or (b) You did not complain about "your mobile phone provider" last year. Imagine you have to complain to "your mobile phone provider" because of a bad quality of service or product. To what extent do you think that "your mobile phone provider" will care about your complaint?
Customer loyalty (ξ_7)	(a) If you would need to choose a new mobile phone provider how likely is it that you would choose "your provider" again? (b) Let us now suppose that other mobile phone providers decide to lower their fees and prices, but "your mobile phone provider" stays at the same level as today. At which level of difference (in %) would you choose another mobile phone provider? (c) If a friend or colleague asks you for advice, how likely is it that you would recommend "your mobile phone provider"?

All the items measured on a Likert scale from 1 (very negative point of view on the service) to 10 (vey positive point of view on the service)



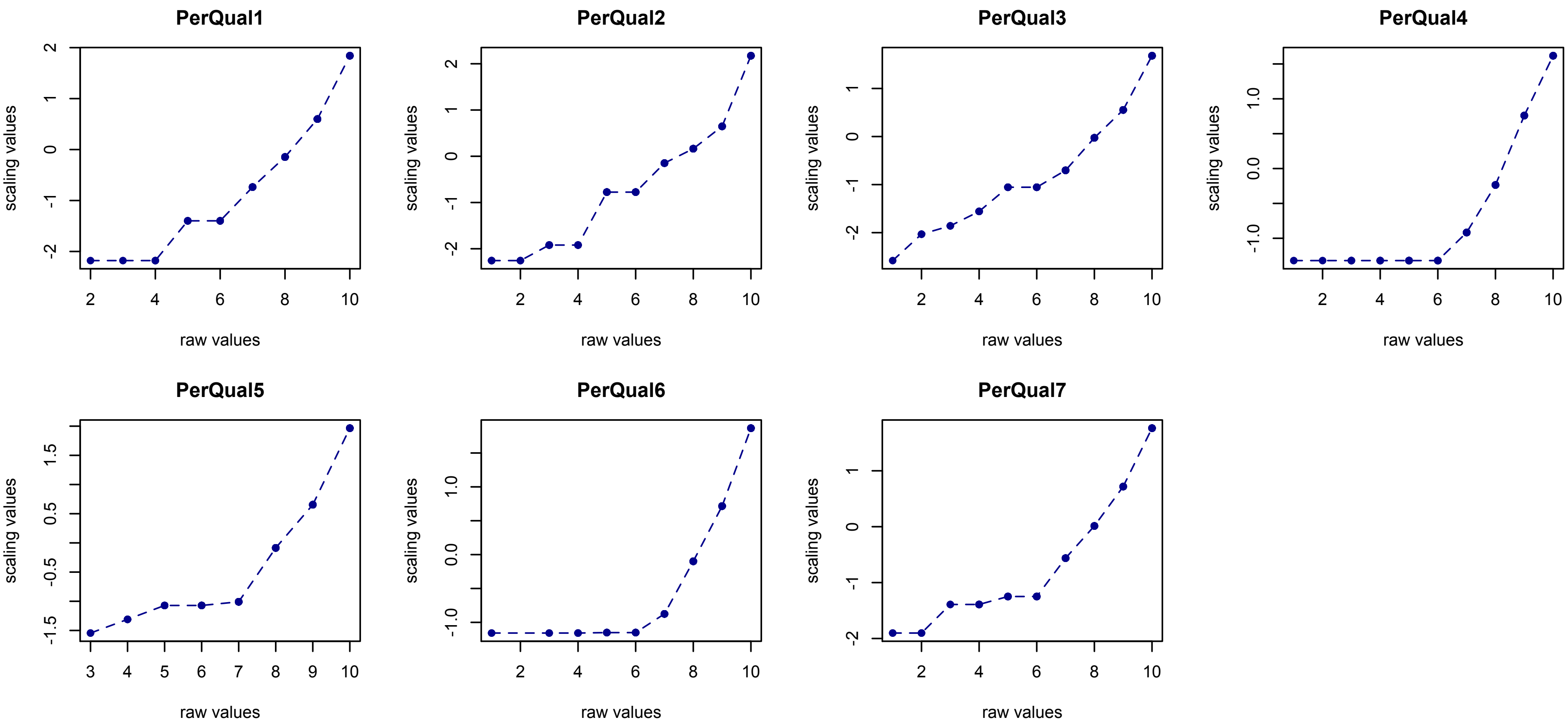
- Standardized MVs
- Centroid Scheme
- Mode A

Mobile data: Comparing model quality

Linearity hypothesis (No scaling)	Monotonicity hypothesis (Ordinal Scaling)	No hypothesis (Nominal Scaling)
$GoF = 0.471$	$GoF = 0.526$	$GoF = 0.547$
$\overline{R}^2 = 0.387$	$\overline{R}^2 = 0.464$	$\overline{R}^2 = 0.495$
$Com_M = 0.599$	$Com_M = 0.618$	$Com_M = 0.623$
$Red_M = 0.263$	$Red_M = 0.315$	$Red_M = 0.335$

Mobile data:
Ordinal quantification for perceived quality

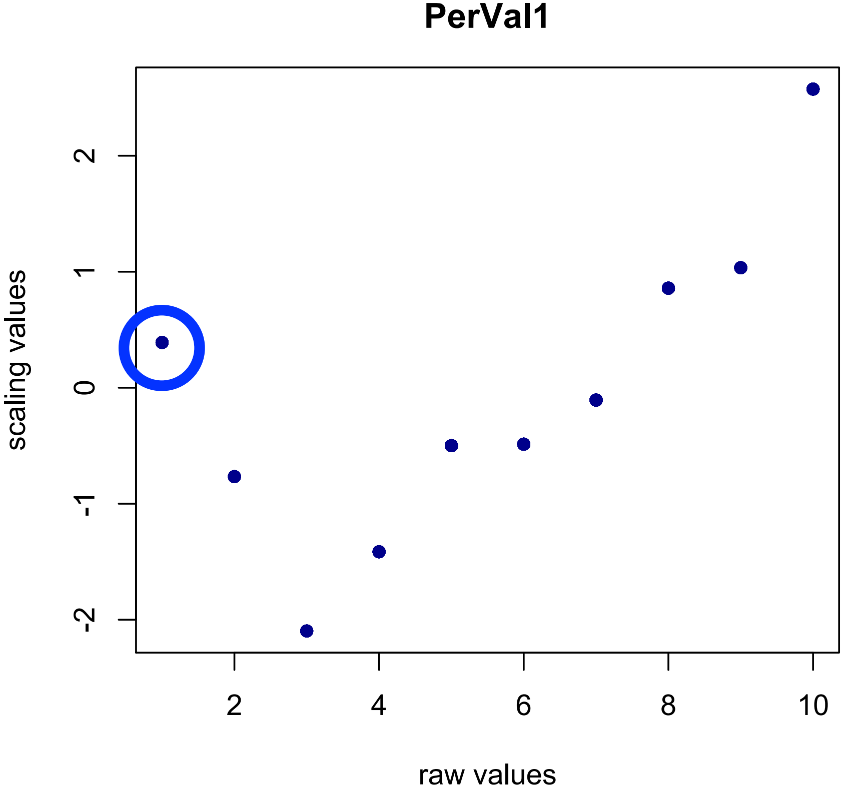
Perceived Quality Latent Variable: 7 indicators

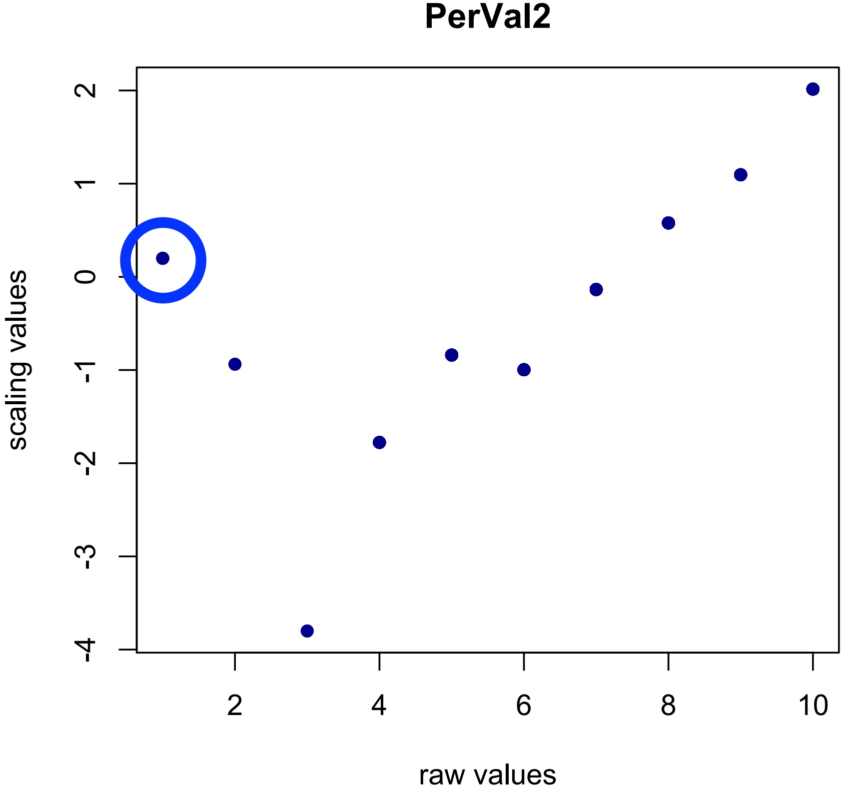


Mobile data: Nominal quantification for perceived quality

Perceived Value: 2 manifest variables

- **PerVal1:** Given the quality of the product and services offered by your mobile phone provider, how would you rate the fees and the price that you pay for them?
- **PerVal2:** Given the fees and the price of the product and services offered by your mobile phone provider, how would you rate the quality of the products and services offered by your mobile phone provider?





Main References 1/5

Baron, R.M.; Kenny, D.A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology, 51(6), 1173–1182

Bentler, P.M., Huang, W., (2014). On components, latent variables, PLS and simple methods: reactions to Ridgon's rethinking of PLS. Long Range Plann

Bookstein F.L. (1982). *Data Analysis by Partial Least Squares*, in: C. Fornell (ed.), A Second Generation of Multivariate Analysis, Praeger, New York, 348-366, 1982

Chatelin, Y.M.; Esposito Vinzi, V.; Tenenhaus, M. (2002). State-of-Art on PLS Path Modeling Through the Available Software, Working Paper, Jouy-en-Josas.

Chin, W.W.; Marcolin, B.L.; Newsted, P.N. (1996). A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects. Results from a Monte Carlo Simulation Study and Voice Mail Emotion/Adoption Study. In DeGross, J.I.; Jarvenpaa, S.; Srinivasan, A. (eds.) Proceedings of the 17th International Conference on Information Systems, pp. 21–41, Cleveland, OH.

Chin, W.W. (1998). *The Partial Least Squares Approach to Structural Equation Modeling,* in: G.A. Marcoulides (ed.), Modern Methods for Business Research, Lawrence Erlbaum Associates, New Jersey, 295-336.

Chin, W.W. (2003). A permutation procedure for multi-group comparison of PLS models, in M. Vilares, M. Tenenhaus, P. Coelho, V. Esposito Vinzi A. Morineau (eds.) PLS and related methods - Proceedings of the International Symposium PLS'03, DECISIA, pp. 33-43

Chin, W.W.; Marcolin, B.L.; Newsted, P.N. (2003). A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and an Electronic-Mail Emotion/Adoption Study. Information Systems Research, 14(2), 189–217.

Cortina, J.M. (1993). Interaction, Nonlinearity, and Multicollinearity. Implications for Multiple Regression. Journal of Management, 19, 915–922.

Dijkstra T. (1983). Some comments on maximum likelihood and Partial Least Squares Methods, *Journal of Econometrics*, 22, 67-90.

Dijkstra, T.K. (2014). PLS' Janus face: response to Professor Rigdon's 'Rethinking Partial Least Squares Modeling: in praise of simple methods'. Long Range Plann.

Dijkstra, T. K. And Henseler J. (2015). Consistent and asymptotically normal PLS estimators for linear structural equations. Computational Statistics and Data Analysis, 81, 10–23.

Main References 2/5

Esposito Vinzi V., Trinchera L., Squillacciotti S. & Tenenhaus M. (2008). REBUS-PLS: A Response–Based Procedure for detecting Unit Segments in PLS Path modeling. To appear on Applied Stochastic Models in Business and Industry.

Fornell, C. (1992). A National Customer Satisfaction Barometer. The Swedish Experience. Journal of Marketing, 56(1), 6–21.

Fornell, C. and Bookstein, F.L. (1982). A Comparative Analysis of Two Structural Equation Models: LISREL and PLS Applied to Market Data, in: C. Fornell (ed.), A Second Generation of Multivariate Analysis, Praeger, New York, 289-324, 1982

Fornell C. and Bookstein F.L. (1982). Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory, Journal of Marketing Research, 19, 440-452.

Fornell C. and Cha J. (1994). Partial Least Squares, in: R.P. Bagozzi (ed.), Advanced Methods of Marketing Research, Blackwell Business, 52-78, 1994

Hahn, C., Johnson, M., Herrmann, A., Huber, F. (2002). Capturing Customer Heterogeneity using a Finite Mixture PLS Approach. Schmalenbach Business Review, 54, 243-269.

Hensler, J. and Fassott, G. (2010). Testing moderating effects in PLS path models: An illustration of available procedure, in V. Esposito Vinzi, W. Chin, J. Henseler H. Wang (eds.) Handbook of Partial Least Squares - Concepts, Methods and Applications, Springer, Berlin, Heidelberg, New York.

Hoyle, R. H. & Kenny, D. A. (1999). Sample size, reliability,and tests of statistical mediation. In R. H. Hoyle (Ed.), Statistical strategies for small sample research (195-222). Thousand Oaks, CA: Sage.

Hwang, H. and Takane, Y. (**2004).** Generalized Structured Component Analysis, Psychometrika, 69, 81-99.

Hwang, H., De Sarbo, W. & Takane, Y. **(2007)**. Fuzzy clusterwise generalized structured component analysis. Psychometrika 72, 181-198.

Huang, W. **(2013).** PLSe: Efficient Estimators and Tests for Partial Least Squares. UCLA. PhD Dissertation.

Hulland J.: Use of Partial Least Squares (PLS) in Strategic Management Research: A Review of Four Recent Studies, Strategic Management Journal, 20, 195-204, 1999

Jaccard, J.; Turrisi, R. (2003). Interaction Effects in Multiple Regression, 2nd ed. Sage Publications, Thousand Oaks.

Jagpal, H.S. **(1982).** Multicollinearity in Structural Equation Models with Unobservable Variables, Journal of Marketing Research, (19), 431-439.

Jarvis, C.B., MacKenzie, S.B. and Podsakoff, P.M. **(2003).** A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research, Journal of Consumer Research, (30), 199-218, 2003

Main References 3/5

Jedidi, K., Harshanjeet, S. J. & De Sarbo W.S. (1997). STEMM: A General Finite Mixture Structural Equation Model. Journal of Classification, 14, 23-50.

Jöreskog K.G. (1971). Simultaneous Factor Analysis in Several Populations. Psychometrika, 36, 409-426.

Jöreskog KG (1977). Structural equation models in the social sciences: Specification, estimation and testing. In: Krishnaiah PR (ed) Applications of statistics. North-Holland Publishing Co. Amsterdam, pp 265–287

Judd, C. M. & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. Evaluation Review, 5, 602-219.

Kenny, D.A.; Judd, C.M. (1984). Estimating the Nonlinear and Interactive Effects of Latent Variables. Psychological Bulletin, 96, 201–210

Last, J. M. (1988). A dictionary of epidemiology (2nd ed.). New York: Oxford University Press.

Lohmöller J.B. (1987). LVPLS 1.8 Program Manual: Latent variable path analysis with partial least squares estimation, Universitaet zu Koehn, Zentralarchiv fuer Empirische Sozialforschung, Köln.

Lohmöller J.B. (1989). Latent variable path modeling with partial least squares, Physica-Verlag, Heidelberg.

MacKinnon, D. P., Lockwood, C.M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. Psychological Methods, 7, 83-104.

MacKinnon, D. P., Lockwood, C.M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Research, 39, 99-128.

MacKinnon, Warsi, D. P., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. Multivariate Behavioral Research, 30, 41-62.

Ringle C.M., Wende S., Will A. (2005). Customer Segmentation with FIMIX-PLS. In: Proceedings of PLS-05, International Symposium, SPAD Test&go, France.

Russolillo, G. (2012). Non-Metric Partial Least Squares. *Electronic Journal of Statistics*, *6*, 1641-1669.

Sánchez G., Aluja T. (2006). PATHMOX: a PLS-PM segmentation algorithm. In: V. Esposito Vinzi, C. Lauro, A. Braverma, H. Kiers & M. G.Schmiek, eds, `Proceedings of KNEMO 2006', number ISBN 88-89744-00-6, Tilapia, Anacapri, p. 69.

Sanchez, G. (2013) PLS Path Modeling with R Trowchez Editions, Berkeley. http://www.gastonsanchez.com/PLS Path Modeling with R.pdf

Sarstedt, M., Ringle, C.M., Henseler, J., Hair, J.F., (2014). On the emancipation of PLS-SEM: a commentary on Rigdon (2012). Long Range Plann

Sobel, M. W. (1990). Effect analysis and causation in linear structural equation models, Psychometrika, 55, 495-515.

Tenenhaus, M. (1998). *La Régression PLS*, Editions Technip, Paris.

modelling. In `Proceedings of the XLII SIS Scientific Meeting', Vol. Contributed Papers, CLEUP, Padova, pp. 739-742.

Main References 4/5

Tenenhaus, M. (1999). *L'approche PLS*, Revue de Statistique Appliquée, 47 (2), 5-40.

Tenenhaus, M., Amato, S. and Esposito Vinzi, V. (2004). *A global goodness-of-fit index for PLS structural equation*

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis* 48, 159-205.

Tenenhaus, M., Esposito Vinzi, V. (2005). PLS regression, PLS path modeling and generalized Procrustean analysis: a combined approach for multiblock analysis, *Journal of Chemometrics*, 19, 145-153, John Wiley & Sons.

Tenenhaus, M., Hanafi, M. (2010) *A bridge between PLS path modelling and multi-block data analysis*, in: Handbook of Partial Least Squares (PLS): Concepts, methods and applications, (V. Esposito Vinzi, J. Henseler, W. Chin, H. Wang, Eds), Volume II in the series of the Handbooks of Computational Statistics, Springer, 2010.

Tenenhaus, M. and Gonzalez P.L. (2001). Comparaison entre les approches PLS et LISREL en modélisation d 'équations structurelles: Application à la mesure de la satisfaction clientèle, 8ème Congrès de la Société Francophone de Classification, Guadeloupe, 2001.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. Psychometrika, 76, 257–284.

Trinchera L. (2007). *Unobserved Heterogeneity in Structural Equation Models: a new approach in latent class detection in PLS Path Modeling*, PhD thesis, DMS, University of Naples.

Williams, J. & MacKinnon, D. P. (2008). Resampling and distribution of the product methods for testing indirect effects in complex models. Structural Equation Modeling, 15, 23-51.

Wold, H. (1966). *Estimation of principal component and related models by iterative least squares*. In: Krishnaiah P.R. (Eds.), Multivariate Analysis, Academic Press, New York, 391-420.

Wold, H. (1975). *Modelling in complex situations with soft infromation*. In "Third World Congress of Econometric Society", Toronto, Canada.

Wold, H. (1975). Soft modeling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. In: Gani, J. (Ed.), Perspectives in Probability and Statistics: Papers, in Honour of M.S. Bartlett on the Occasion of his Sixty-8fth Birthday. Applied Probability Trust, Academic, London, pp. 117–142.

Wold, H. (1981). *The Fix-Point Approach to Interdependent Systems: Review and Current Outlook*, in: H. Wold (ed.), The Fix-Point Approach to Interdependent Systems, North-Holland, Amsterdam.

Main References 5/5

Wold, H. (1982). *Soft modeling: the basic design and some extensions*. In: Joreskog, K.G. and Wold, H. (Eds.), Systems under Indirect Observation, North-Holland, Amsterdam Part 2, 1-54.

Wold, H. (1983). *Quantitative Systems Analysis: the Pedigree and Broad Scope of PLS (Partial Least Squares) Soft Modeling*, in: H. Martens and H. Russwurm Jr. (eds.) Food Research and Data Analysis, Applied Science Publisher Ltd.

Wold, H. (1985). *Partial Least Squares*, in: Vol.6 of S. Kotz & N.L. Johnson (eds.), Encyclopedia of Statistical Sciences, John Wiley & Sons, New York, 581-591. **Huang, W. (2013).** PLSe: Efficient Estimators and Tests for Partial Least Squares. UCLA. PhD Dissertation.

Wold, S. Martens, H. and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the pls method. In Matrix Pencils, E Kagstrom, B. and Ruhe, A. (Eds.) Springer Berlin Heidelberg, 973, 286–293.

Wollenberg, A. L. (1977) Redundancy analysis an alternative for canonical correlation analysis. Psychometrika, 42, 207–219.