COMPONENT-BASED PATH MODELING

Vincenzo ESPOSITO VINZI

ESSEC Business School, Paris

vinzi@essec.edu

ESSEC BUSINESS SCHOOL

In collaboration with:

L. Trinchera (NEOMA Business School, France)

G. Russolillo (CNAM, France)

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).

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

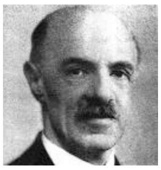
Path Analysis has been developed in the 20s by S. Wright to investigate genetic problems and to help his father in economic studies.

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

The main novelty 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 beginning of the 20th century to measure intelligence in an "objective" way.

The main idea is that intelligence is a MULTIDIMENSIONAL issue, thus the correlation observed among several 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 of 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 the "cause" of observed correlation among several set of manifest variables.

Louis Thurstone (29 May 1887–30 September 1955)

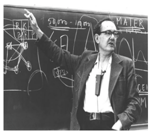
is the father of the Multiple Factor Analysis

Herman Wold (25 december 1908 – 16 february 1992)

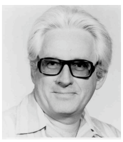
In the 50's he meets Thurstone. They decide to co-organize "the Upspsala Symposium on Psycological Factor Analysis". Since then, H. Wold started working on Latent Variable models.

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 in Sociology.

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

With the economist 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
is Professor at Uppsala University, Sweden

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

In the second half of the 60's, 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

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Soft Modeling and H. Wold

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

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 avoids restrictive hypothesis, i.e. multivariate normality and large samples, underlying maximum likelihood techniques.

It 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
Factor-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 aiming at validating a model (theory building)

SEM

Variance-based
Composite-based
Component-based
Methods

The aim is to provide latent variable scores (proxy, composites, factor scores) that are the most correlated to each other as possible (according to path diagram structure) and the most representative of their own block of manifest variables.
• it focuses on latent variable scores computation
• it focuses on explaining variances
• it is more an exploratory approach than a confirmatory one (operational model strategy)

Two families of methods

- Covariance-based Methods
  - Maximum Likelihood estimation (SEM-ML) — *Joreskog (1977)* — LISREL model
  - Generalized Least Squares (GLS)
  - Asymptotically Distribution Free (ADF)
  - Unweighted Least Squares (ULS) ....and others

LISREL-type methods

SEM

- Component-based Methods
  - PLS Path Modeling (PLS-PM or PLS-SEM) — *Wold (1975), Tenenhaus et al. (2005)*
  - Generalized Maximum Entropy (GME) — *Al-Nasser(2003)*
  - Generalized Structured Component Analysis (GSCA) method — *Hwang & Takane (2004)*
  - Regularized Generalized CanCorr Analysis (RGCCA) — *Tenenhaus & Tenenhaus (2011)*

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Two families of methods

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Structural Equation Modeling

Quite a **few (statistical) hypotheses** are usually needed

Important **theoretical knowledge** has to be available **for the model specification**:

- **Measurement** model
  (what manifest variables are exclusively measuring what concept)
- Direction of links between manifest and latent variables
  (**outwards or inwards**, i.e. reflective vs. formative)
- Network of **structural relationships**
  ("causality" direction, "predictive path", feedbacks, hidden?)

Confirmatory vs. Exploratory

Structural Equation Models: two approaches

In component-based SEM the latent variables are defined as components or weighted sums of the manifest variables

→ they are fixed variables (linear composites, scores)

In covariance- based SEMs the latent variables are equivalent to common factors

→ they are theoretical (and random) variables

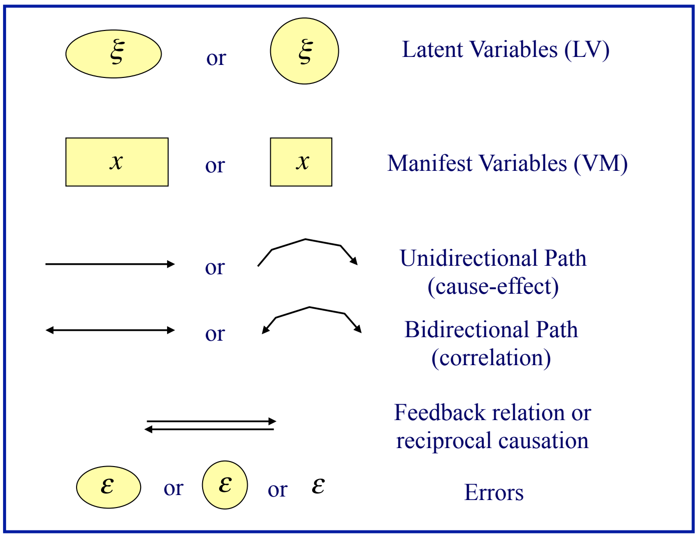
This leads to different parameters to estimate for latent variables, i.e.:

→ factor means and variances in covariance-based methods

→ weights in component based approaches

SEM: inner, outer and global model

SEM: drawing conventions



Structural Equation models: notation

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

  → $\mathbf{x}_{pq}$ generic MV

- Q **latent variables** (LVs)

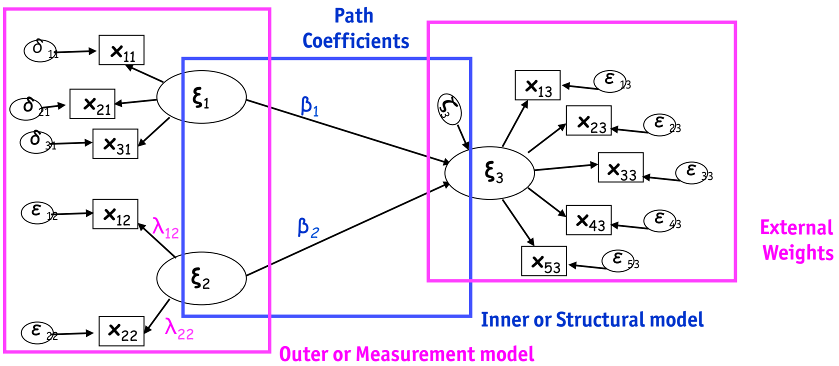
  → ξ_q generic LV

- 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

Structural Equation models: notation

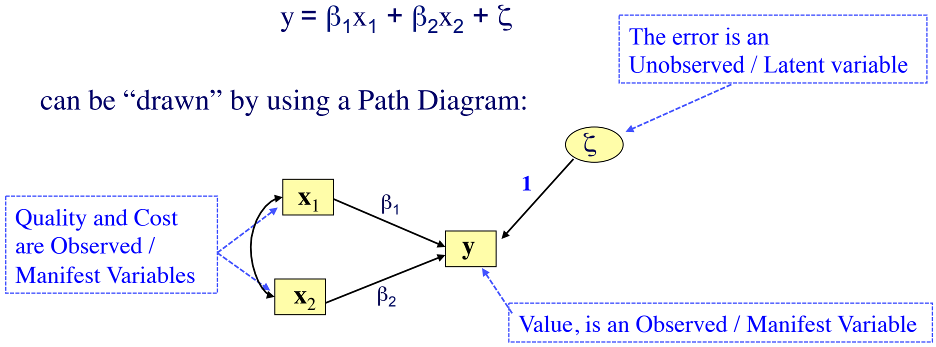


"Drawing" a regression model

The multiple regression model (on centred variables) :

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

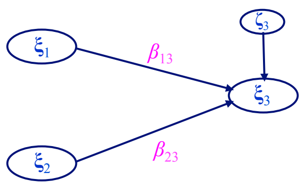
can be "drawn" by using a Path Diagram:



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

PLS Path Model Equations: inner model

The structural model describes the relations among the latent variables



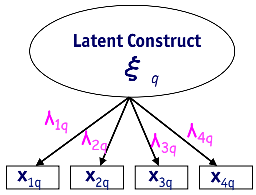
For each endogenous LV in the model it can be written as:

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

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:

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

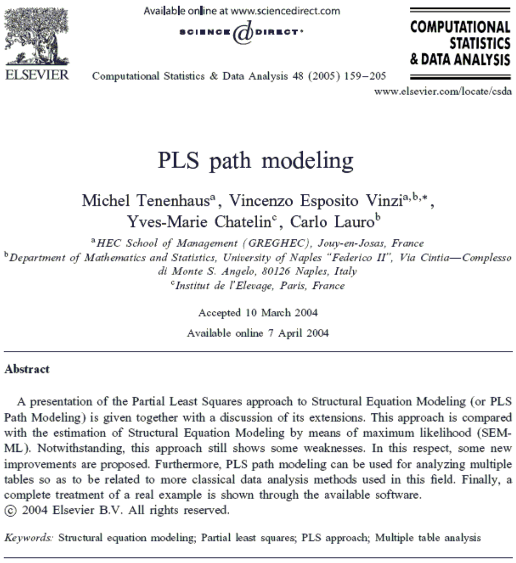
where:
- λ_{pq} is a loading term linking the *q-th* LV to the *p-th* MV

Weight relation – Linear composite

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

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

PLS-PM Algorithm


Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 159–205
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

PLS path modeling

Michel Tenenhaus[a], Vincenzo Esposito Vinzi[a,b,*], Yves-Marie Chatelin[c], Carlo Lauro[b]

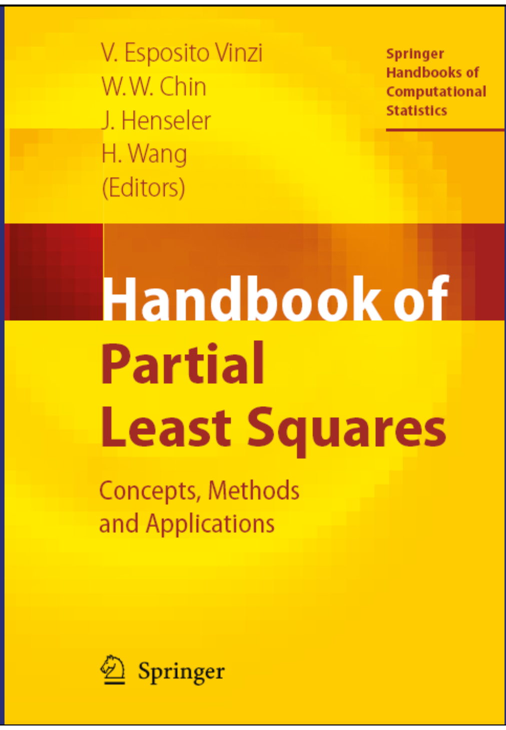
[a] HEC School of Management (GREGHEC), Jouy-en-Josas, France
[b] Department of Mathematics and Statistics, University of Naples "Federico II", Via Cintia—Complesso di Monte S. Angelo, 80126 Naples, Italy
[c] Institut de l'Elevage, Paris, France

Accepted 10 March 2004
Available online 7 April 2004

Abstract

A presentation of the Partial Least Squares approach to Structural Equation Modeling (or PLS Path Modeling) is given together with a discussion of its extensions. This approach is compared with the estimation of Structural Equation Modeling by means of maximum likelihood (SEM-ML). Notwithstanding, this approach still shows some weaknesses. In this respect, some new improvements are proposed. Furthermore, PLS path modeling can be used for analyzing multiple tables so as to be related to more classical data analysis methods used in this field. Finally, a complete treatment of a real example is shown through the available software.

Keywords: Structural equation modeling; Partial least squares; PLS approach; Multiple table analysis



Published in 2010

Partial LEAst Squares StrUctural Relationship Estimation

XLSTAT-PLSPM software (www.xlstat.com)

currently developed by:

Y.M. Chatelin, V. Esposito Vinzi, T. Fahmy,
E. Jakobowicz, M. Tenenhaus

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:

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

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

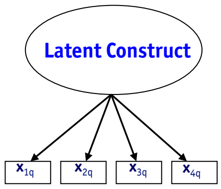
Composites = weighted aggregates of manifest variables

$$\mathbf{t}_q \propto \Sigma_q w_{pq}\mathbf{x}_{pq} = \mathbf{X}_q\mathbf{w}_q$$

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

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

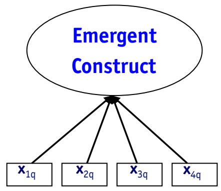
- ➔ These indicators **should covary**
- ➔ Several **simple OLS** regressions
- ➔ Explained Variance (higher **AVE**, communality)
- ➔ Internal Consistency
- ➔ **Stability of results** with well-defined blocks



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

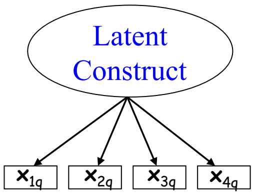
$\mathbf{w}_q = (\mathbf{X}_q^T\mathbf{X}_q)^{-1}\mathbf{X}_q^T\mathbf{z}_q$

- ➔ 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



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

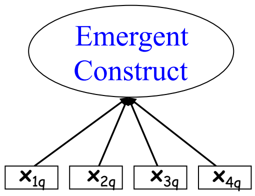
Latent or Emergent Constructs?



Reflective (or Effects) 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)



Formative (or Causal) Indicators

e.g. Social Status, Perceptions

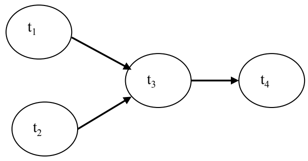
- Constructs are **combinations of observed variables(multidimensional)**
- Not designed to account for observed variables
- These indicators **need not covary**: changes in one indicator do not imply changes in the others.
- Measures of internal **consistency do not apply**.

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

A focus on Inner Estimation

Inner Estimation

Latent Variable proxies = weighted aggregates of connected LVs



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

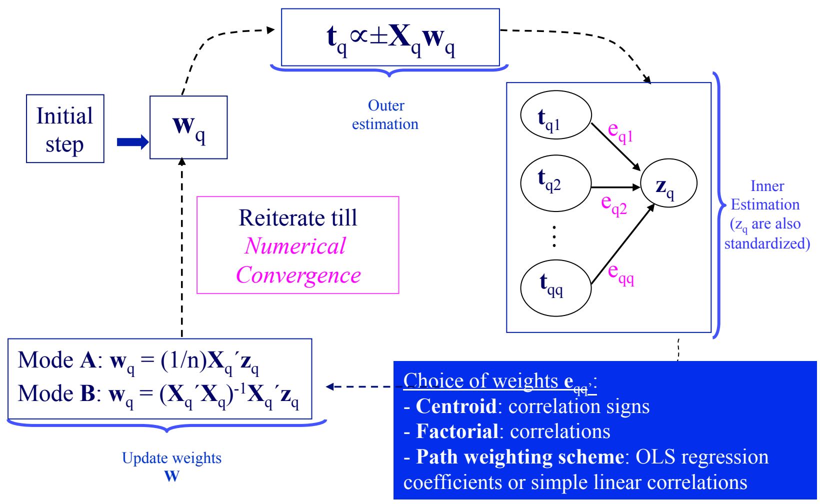
1. **Centroid scheme**: $\mathbf{z}_3 = e_{13}\mathbf{t}_1 + e_{23}\mathbf{t}_2 + e_{43}\mathbf{t}_4$ where $e_{qq'} = sign(cor(\mathbf{t}_q, \mathbf{t}_{q'}))$
2. **Factorial scheme**: $\mathbf{z}_3 = cor(\mathbf{t}_3, \mathbf{t}_1) * \mathbf{t}_1 + cor(\mathbf{t}_3, \mathbf{t}_2) * \mathbf{t}_2 + cor(\mathbf{t}_3, \mathbf{t}_4) * \mathbf{t}_4$
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 γ's are the regression coeddicients of the model: $\mathbf{t}_3 = \gamma_{31} \times \mathbf{t}_1 + \gamma_{32} \times \mathbf{t}_2 + \delta$

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

The PLS Path Modeling algorithm

MVs are centered or standardized



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

PLS-PM Optimizing 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:

$$\underset{\mathbf{w}_q}{\arg\max}\left\{\sum_{q\neq q'} c_{qq'}\, g\left(cor\left(\mathbf{X}_q\mathbf{w}_q, \mathbf{X}_{q'}\mathbf{w}_{q'}\right)\right)\right\}$$

s.t. $\left\|\mathbf{X}_q\mathbf{w}_q\right\| = 1$

where:

$$c_{qq'} = \begin{cases} 1 & \text{if } \mathbf{X}_q \text{ and } \mathbf{X}_{q'} \text{ is connected} \\ 0 & \text{otherwise} \end{cases} \qquad g = \begin{cases} \text{square} & \text{(Factorial scheme)} \\ \text{abolute value} & \text{(Centroid scheme)} \end{cases}$$

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:

$$\underset{\|\mathbf{w}_q\|=1}{\arg\max}\left\{\sum_{q\neq q'} c_{qq'}\, g\left(\operatorname{cov}\left(\mathbf{X}_q\mathbf{w}_q, \mathbf{X}_{q'}\mathbf{w}_{q'}\right)\right)\right\}$$

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:

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

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

s.t. $\|\mathbf{X}_q\mathbf{w}_q\| = 1$ if Mode B for block q

$\|\mathbf{w}_q\| = 1$ 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 « special » cases

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*

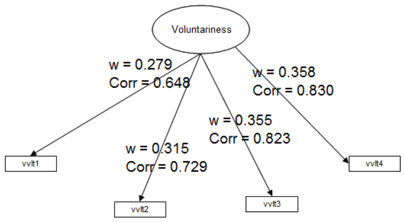
SPSS results (principal components)

Component Matrix[a]

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

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

XL-STAT graphical results

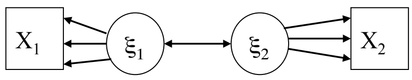


* Results from W.W. Chin slides on PLS-PM

Two block case

Tucker Inter-batteries Analysis
(1st component)

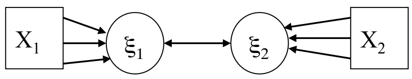
$$\underset{\|\mathbf{w}_1\|=\|\mathbf{w}_2\|=1}{\arg\max}\left\{\operatorname{cov}\left(\mathbf{X}_1\mathbf{w}_1,\mathbf{X}_2\mathbf{w}_2\right)\right\}$$



Mode A for X_1, Mode A for X_2

Canonical Correlation Analysis
(1st component)

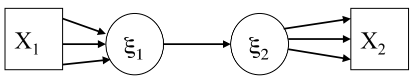
$$\underset{\operatorname{var}(\mathbf{X}_1\mathbf{w}_1)=\operatorname{var}(\mathbf{X}_2\mathbf{w}_2)=1}{\arg\max}\left\{\operatorname{cov}\left(\mathbf{X}_1\mathbf{w}_1,\mathbf{X}_2\mathbf{w}_2\right)\right\}$$



Mode B for X_1, Mode B for X_2

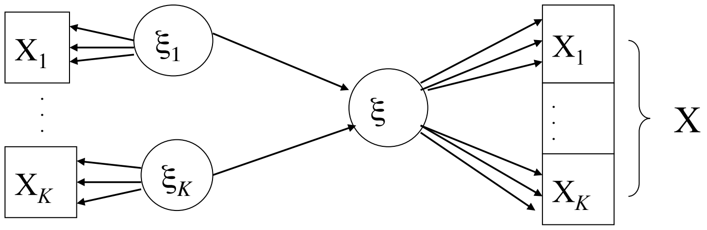
Redundancy Analysis
(1st component)

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



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

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

Mode B + Factorial

- Generalised Canonical Correlation Analysis (Carroll)

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

Mode B + Centroid

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

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

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Wine Tasting: Loire Wine Multiblock Dataset

(Pagès, Asselin, Morlat & Robichet 1987)

4 blocks of variables 3 Appellations 4 Soils **21 observed wines**

		2el (Saumur),1	1cha (Saumur),1	1fon (Bourgueil),1	1vau (Chinon),3	...	t1 (Saumur),4	t2 (Saumur),4
X_1	Smell intensity at rest	3.07	2.96	2.86	2.81	...	3.70	3.71
	Aromatic quality at rest	3.00	2.82	2.93	2.59	...	3.19	2.93
	Fruity note at rest	2.71	2.38	2.56	2.42	...	2.83	2.52
	Floral note at rest	2.28	2.28	1.96	1.91	...	1.83	2.04
	Spicy note at rest	1.96	1.68	2.08	2.16	...	2.38	2.67
X_2	Visual intensity	4.32	3.22	3.54	2.89	...	4.32	4.32
	Shading (orange to purple)	4.00	3.00	3.39	2.79	...	4.00	4.11
	Surface impression	3.27	2.81	3.00	2.54	...	3.33	3.26
X_3	Smell intensity after shaking	3.41	3.37	3.25	3.16	...	3.74	3.73
	Smell quality after shaking	3.31	3.00	2.93	2.88	...	3.08	2.88
	Fruity note after shaking	2.88	2.56	2.77	2.39	...	2.83	2.60
	Floral note after shaking	2.32	2.44	2.19	2.08	...	1.77	2.08
	Spicy note after shaking	1.84	1.74	2.25	2.17	...	2.44	2.61
	Vegetable note after shaking	2.00	2.00	1.75	2.30	...	2.29	2.17
	Phenolic note after shaking	1.65	1.38	1.25	1.48	...	1.57	1.65
	Aromatic intensity in mouth	3.26	2.96	3.08	2.54	...	3.44	3.10
	Aromatic persisitence in mouth	3.26	2.96	3.08	2.54	...	3.44	3.10
	Aromatic quality in mouth	3.26	2.96	3.08	2.54	...	3.44	3.10
X_4	Intensity of attack	2.96	3.04	3.22	2.70	...	2.96	3.33
	Acidity	2.11	2.11	2.18	3.18	...	2.41	2.57
	Astringency	2.43	2.18	2.25	2.18	...	2.64	2.67
	Alcohol	2.50	2.65	2.64	2.50	...	2.96	2.70
	Balance (Acid., Astr., Alco.)	3.25	2.93	3.32	2.33	...	2.57	2.77
	Mellowness	2.73	2.50	2.68	1.68	...	2.07	2.31
	Bitterness	1.93	1.93	2.00	1.96	...	2.22	2.67
	Ending intensity in mouth	2.86	2.89	3.07	2.46	...	3.04	3.33
	Harmony	3.14	2.96	3.14	2.04	...	2.74	3.00
	Global quality	3.39	3.21	3.54	2.46	...	2.64	2.85

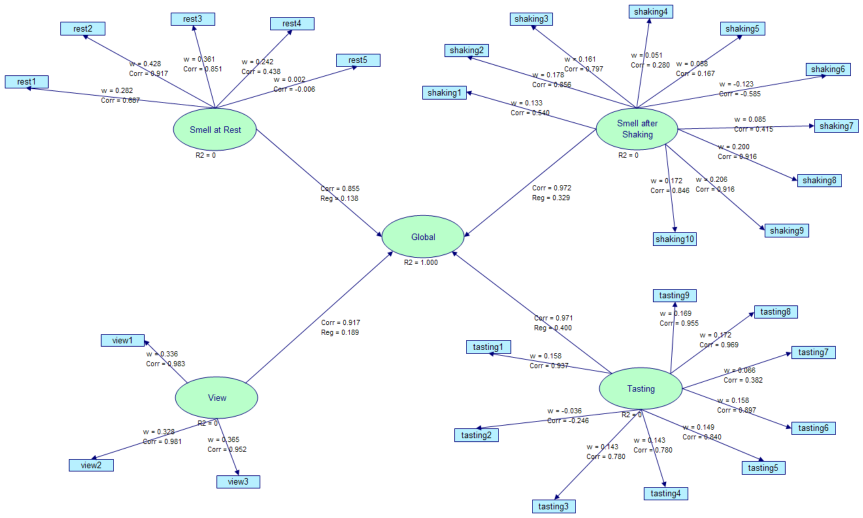
Illustrative variable

X_1 = Smell at rest (5), X_2 = View (3), X_3 = Smell after shaking (10), X_4 = Tasting (9)

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

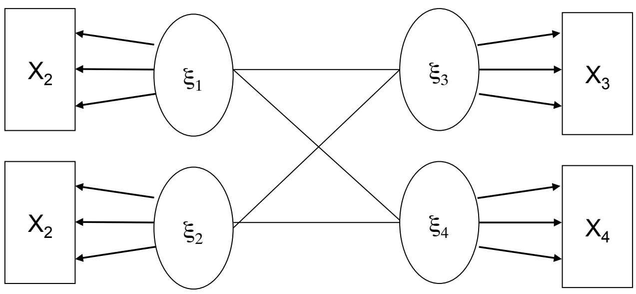
Hierarchical PLS Model: Wine data

Mode A + Path Weighting



‘Confirmatory’ PLS Model

Each LV is connected to all the others



PLS criteria for multiple table analysis

Method	Criterion $(F_k = \mathbf{X}_k \mathbf{w}_k, F = \mathbf{X}\mathbf{w})$	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)	$Max_{all \, \lVert w_j \rVert = 1} \sum_{j \neq k} Cov^2(X_j w_j, X_k w_k)$			

(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.

From Tenenhaus and Hanafi (2010)

PLS criteria for multiple table analysis

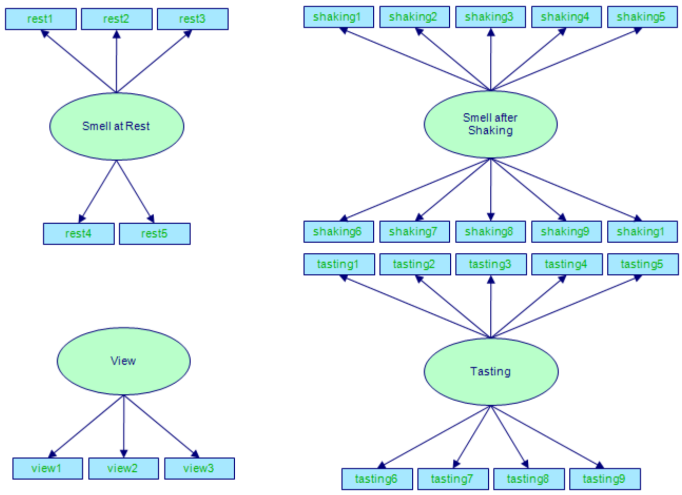
Method	Criterion	PLS path model	Mode	Scheme
(11) (Hanafi and Kiers 2006)	$Max_{all \, \lVert w_j \rVert = 1} \sum_{j \neq k} \lvert Cov(X_j w_j, X_k w_k)\rvert$			
(12) ACOM (Chessel and Hanafi 1996) or Split PCA (Lohmöller 1989)	$Max_{all \, \lVert w_j \rVert = 1} \sum_j Cov^2(X_j w_j, X_{j+1} w_{j+1})$ 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 theregression 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 and Hanafi (2010)

Wine Tasting experience in 4 steps
Block Structure based on Prior Knowledge

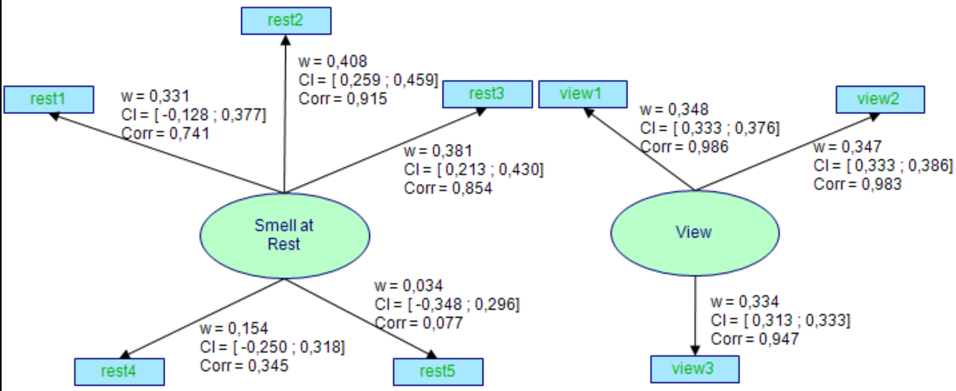
Measurement Model Specification



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

1-block Modeling: Principal Component Analysis
(1^{st} order solution – higher order components by deflation)

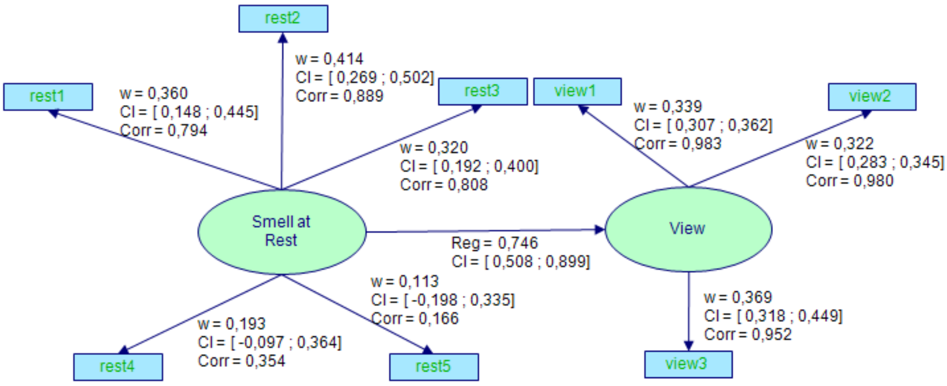
2 separate PCA:
Mode A (optimizing communalities), non connected blocks



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

2-block Modeling as a Regression Analysis
(1st order solution)

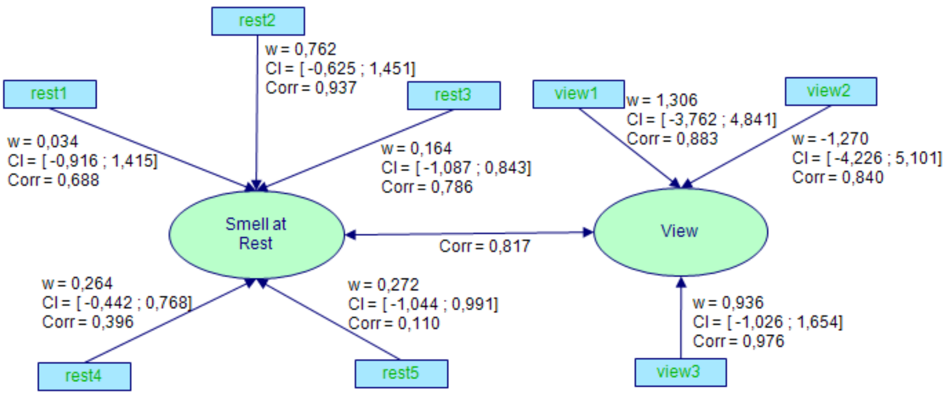
Mode A (covariance-oriented, p_j can be >> n) for both blocks
Directed arrow for dependence



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

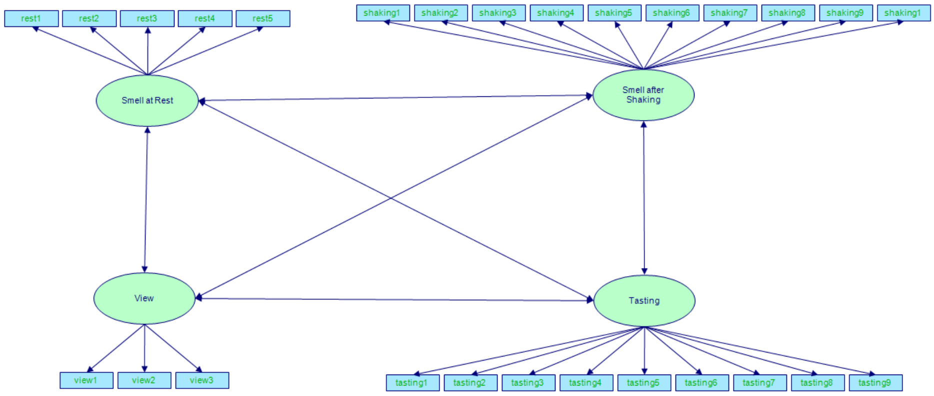
2-block Modeling as a Canonical Correlation Analysis
(1st order solution)

Mode B (correlation-oriented) for both blocks
Double-headed arrow for interdependence



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

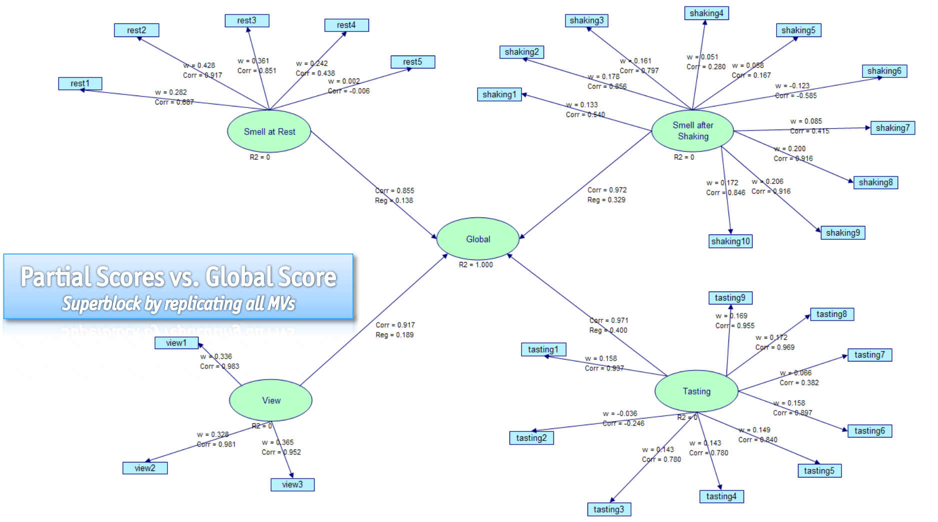
Wine Tasting: Multiple Table Analysis
(fully connected network: double-headed arrows)



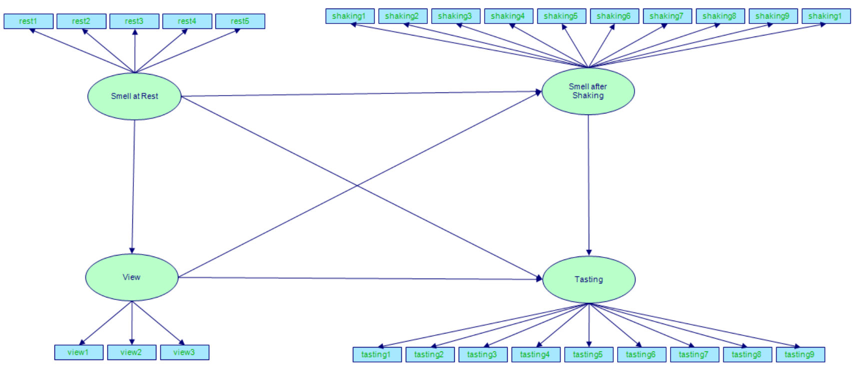
Blocks are connected to each other: interdependence

Direction of a link does not matter – Some links could be eventually removed

Wine Tasting: Hierarchical Model
(Second Order Construct in classical SEM or Auxiliary Variable in data analysis)



Wine Tasting: Predictive Path Model (or Structural Equation Model)



Network of "Predictive Relationships"
(yielding a DAG – Directed Acyclic Graph)

PLS-PM
a toy example for understanding the algorithm

Economic inequality and political instability

Economic inequality

Agricultural inequality

GINI : Inequality of land distributions

FARM : % farmers that own half of the land (> 50%)

RENT : % farmers that rent all their land

Industrial development

GNPR : Gross national product per capita ($ 1955)

LABO : % of labour force employed in agriculture

Political instability

INST : Instability of executive (1945-1961)

ECKS : Nb of violent internal war incidents ('46-'61)

DEAT : Nb of people killed as a result of civil war violence ('50-'62)

***DEMO* :**

D-STAB : Stable democracy

D-UNST : Unstable democracy

DICT : Dictatorship

Data from Russet (1964), with non linear transformations in Gifi (1981)

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Economic inequality and political instability

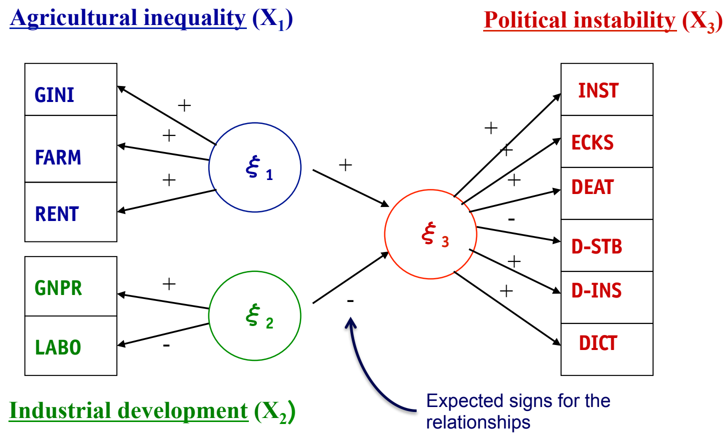
Original Data from Russet (1964), as presented in Tenenhaus (1998)
47 countries

	Gini	Farm	Rent	Gnpr	Labo	Inst	Ecks	Deat	Demo
Argentina	86.3	98.2	32.9	374	25	13.6	57	217	2
Australia	92.9	99.6	29.5	1215	14	11.3	0	0	1
Austria	74.0	97.4	10.7	532	32	12.8	4	0	2
⋮									
France	58.3	86.1	26.0	1046	26	16.3	46	1	2
⋮									
Yugoslavia	43.7	79.8	0.0	297	67	0.0	9	0	3

The Demo variable is categorical (nominal) with:
1 = Stable democracy
2 = Unstable democracy
3 = Dictatorship

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

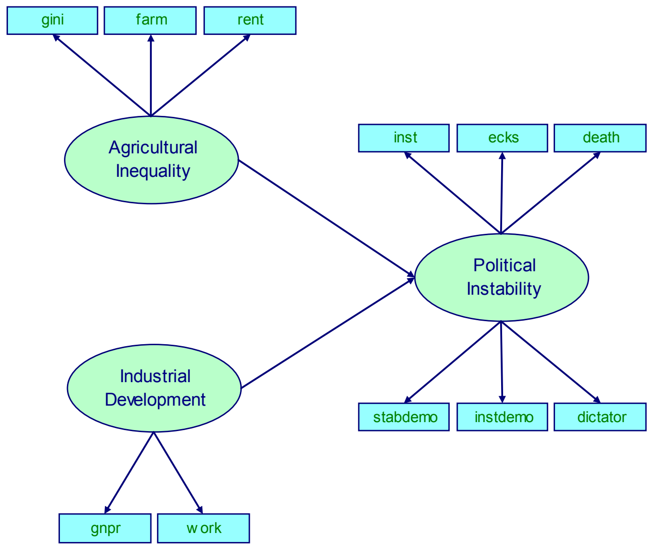
Economic inequality and political instability



There exists a structural relationship between latent variables :

Political inst. (ξ_3) = β_1×Agricultural ineq.(ξ_1) + β_2 × Industrial dev.(ξ_2) + residual

Use of the XLSTAT-PLSPM software



Estimation of Latent Variables in PLS-PM

An example with Mode A + Centroid Scheme

(1) External Estimates

$t_1 = X_1 w_1$

$t_2 = X_2 w_2$

$t_3 = X_3 w_3$

(2) Internal Estimates

$z_1 = t_3$

$z_2 = -t_3$

$z_3 = t_1 - t_2$

(3) Computation of w_j

$w_{1q} = cor(x_{1q}, z_1)$

$w_{2q} = cor(x_{2q}, z_2)$

$w_{3q} = cor(x_{3q}, z_3)$

Algorithm

- Start with arbitrary weights w_1, w_2, w_3. Here: $w_1 = (1,0,0)$
- Obtain the new weights w_j by means of steps from (1) to (3).
- Iterate the procedure till convergence.

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Economic inequality and political instability

Results: Problems in the signs…

Weights (Dimension 1):

Latent variable	Manifest variables	Outer weight	Outer weight (Bootstrap)	Standard error	Critical ratio (CR)	Lower bound (95%)	Upper bound (95%)
Agricultural Inequality	gini	0.460	0.453	0.046	9.889	0.325	0.545
	farm	0.516	0.510	0.049	10.451	0.395	0.653
	rent	0.081	0.080	0.152	0.535	-0.278	0.347
Industrial Development	**gnpr**	**-0.511**	-0.511	0.023	-22.112	-0.557	-0.448
	labo	0.538	0.536	0.022	24.111	0.498	0.606
Political Instability	**inst**	**-0.104**	-0.109	0.066	-1.579	-0.227	0.064
	ecks	-0.270	-0.265	0.041	-6.600	-0.340	-0.151
	death	-0.302	-0.296	0.039	-7.693	-0.379	-0.218
	demostab	**0.336**	0.326	0.037	9.157	0.257	0.421
	demoinst	-0.037	-0.031	0.064	-0.575	-0.156	0.123
	dictatur	-0.285	-0.281	0.036	-7.974	-0.357	-0.206

Correlations (Dimension 1):

Latent variable	Manifest variables	Standardized loadings	Communalities	Redundancies
Agricultural Inequality	gini	0.977	0.955	
	farm	0.986	0.972	
	rent	0.516	0.266	
Industrial Development	gnpr	-0.950	0.903	
	labo	0.955	0.912	
Political Instability	inst	-0.352	0.124	0.077
	ecks	-0.816	0.665	0.414
	death	-0.794	0.630	0.392
	demostab	0.866	0.749	0.466
	demoinst	-0.094	0.009	0.006
	dictatur	-0.733	0.537	0.334

Loading = Regression coefficient of each MV on the corresponding LV
= Correlation coefficient if MV's are standardized

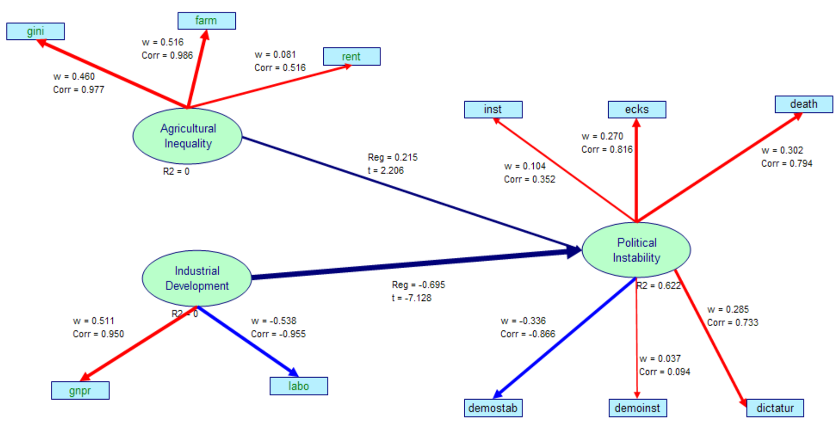
ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Economic inequality and political instability

Results: after sign inversion….

Latent variable	Manifest variables	Outer weight
Agricultural Inequality	gini	0.460
	farm	0.516
	rent	0.081
Industrial Development	gnpr	0.511
	labo	-0.538
Political Instability	inst	0.104
	ecks	0.270
	death	0.302
	demostab	-0.336
	demoinst	0.037
	dictatur	0.285

Economic inequality and political instability



Economic inequality and political instability

PLS Results: Structural Equations (Inner Model)

Inner model (Dimension 1):

R^2 (Political Instability / 1):

R²	R²(Bootstrap)	Std. deviation	Lower bound (95%)	Upper bound (95%)
0.622	0.632	0.063	0.532	0.762

Path coefficients (Political Instability / 1):

Latent variable	Value	Standard error	t	Pr > |t|
Intercept	0.000	0.092	0.000	0.000
Agricultural Inequality	0.215	0.097	2.206	0.033
Industrial Development	-0.695	0.097	-7.128	0.000

Value(Bootstrap)	Standard error(Bootstrap)	Lower bound (95%)	Upper bound (95%)
0.000	0.000	0.000	0.000
0.251	0.104	0.039	0.487
-0.673	0.077	-0.846	-0.466

Economic inequality and political instability

PLS results: Standardized Latent Variable Scores

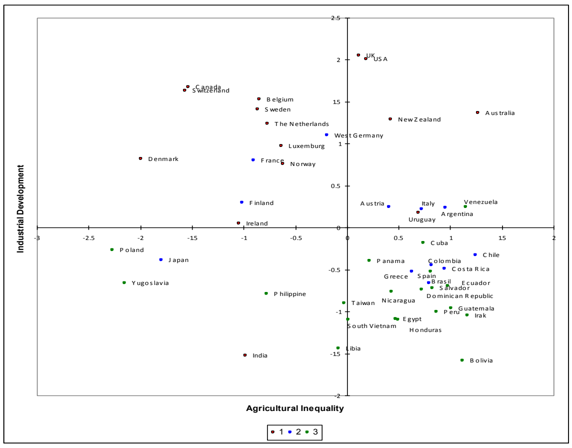
Latent variable scores (Dimension 1):

	Agricultural Inequality	Industrial Development	Political Instability
Argentina	0.953	0.238	0.751
Australia	1.265	1.371	-1.601
Austria	0.404	0.253	-0.464
Belgium	-0.848	1.530	-0.881
Bolivia	1.115	-1.584	1.503
Brasil	0.789	-0.654	0.268
Canada	-1.539	1.680	-0.972
Chile	1.239	-0.324	0.016
Colombia	0.819	-0.443	0.810
Costa Rica	0.939	-0.484	0.302
Cuba	0.734	-0.182	1.694
Denmark	-1.996	0.821	-1.528
Dominican Republic	0.720	-0.737	0.542
Ecuador	0.976	-0.690	0.956
Egypt	0.464	-1.086	0.865
Salvador	0.824	-0.718	0.426
Finland	-1.020	0.304	-0.262
France	-0.910	0.804	0.492
Guatemala	1.006	-0.959	1.099
…	…	…	…

Economic inequality and political instability

Map of countries as mapped by LV scores

Y_1 = agricultural inequality , Y_2 = industrial development



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's (or Jöreskog's) 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

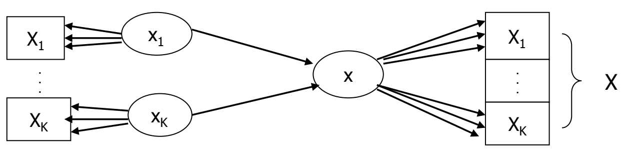
The manifest variables are reliable if these indices are at least 0.7
(0.6 to 0.8 according to exploratory vs. confirmatory purpose)

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

What if unidimensionality is rejected?

Four possible solutions:

- **Remove** manifest variables that are far from the model
- Change the measurement model into a **inwards-directed 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**

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

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 (convergent and monofactorial validity).

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

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

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

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

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Monofactorial MVs

A manifest variable needs to load significantly higher with the latent variables it is intended to measure than with the other latent variables:

$$\operatorname{cor}^2\left(x_{pq},\xi_q\right) >> \operatorname{cor}^2\left(x_{pq},\xi_{q'}\right)$$

Cross-loadings for checking proper reflection

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.

$$H_0 : \operatorname{cor}\left(\xi_q,\xi_{q'}\right) = 1 \qquad H_0 : \operatorname{cor}\left(\xi_q,\xi_{q'}\right) = 0$$

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**$_{q'}$) **>** $\operatorname{cor}^2\left(\hat{\xi}_q,\hat{\xi}_{q'}\right)$ 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 Quality

Communality

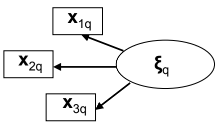
For each manifest variable x_{pq} the communality is a squared correlation:

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

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: $Com_q = AVE_q$)



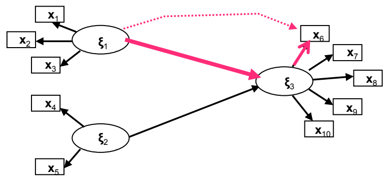
The communality of the whole model is the **Mean Communality**, obtained as:

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

Redundancy

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

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



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

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

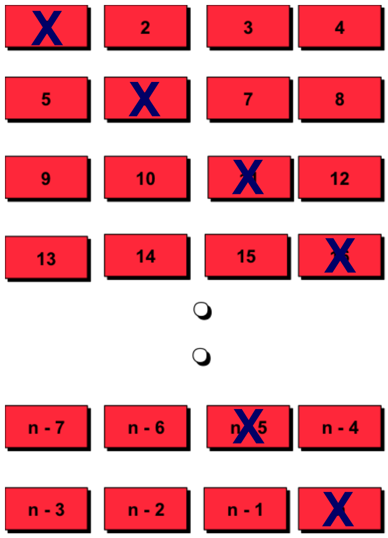
$$H_q^2 = 1 - \frac{\sum_q\sum_i (\mathrm{x}_{pqi} - \bar{\mathrm{x}}_{pq} - \hat{\lambda}_{pq(\text{-i})}\hat{\xi}_{q(\text{-i})})^2}{\sum_q\sum_i (\mathrm{x}_{pqi} - \bar{\mathrm{x}}_{pq})^2}$$

Redundancy Option (also called Q^2)

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

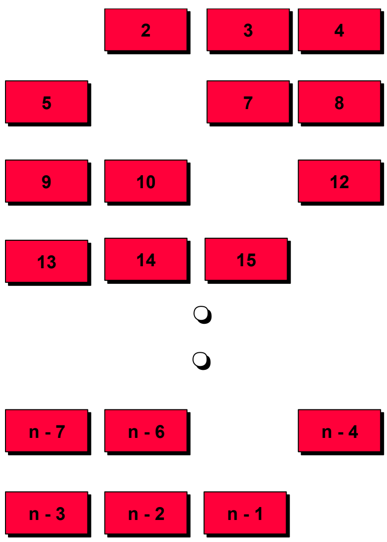
The mean of the CV-communality and the CV-redundancy (for endogenous blocks) indices can be used to measure the global quality of the measurement model if they are positive for all blocks (endogenous for redundancy).

Blindfolding procedure



From W.W. Chin's slides on PLS-PM

Blindfolding procedure



From W.W. Chin's slides on PLS-PM

A global prediction index for PLS-PM

- PLS-PM is blamed not to "optimize" a global scalar "fit" function….
- 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 "descriptive" index:
  - Compromise between outer and inner model performance
  - Allow a comparison of performances
  - Know its "real" maximum for a better interpretation
  - No claim of global quality assessment

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Godness of Fit index

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

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!

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

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Relative GoF

From **PCA** is well known that $\Rightarrow$ $\sum_{p=1}^{P_q} cor^2(\mathbf{x}_{pq}, \xi_q) \le \lambda_q^{PCA}$

the largest eigenvalue of the $\mathbf{X}_q{}'\mathbf{X}_q$

From **CCA** we know that $\Rightarrow$ $\sum_{q^*=1}^{Q^*} R^2(\xi_{q^*}, \xi_j \text{ explaining } \xi_{q^*}) \le \rho_{q^*}^2$

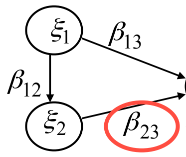
The square of the first canonical correlation

so relating each term of the GoF to the corresponding maximum:

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

Non parametric Bootstrap-based validation of one or more path coefficients

Let us consider this simple model:



$$\begin{cases} \xi_2 = \beta_{12}\xi_1 + \zeta_1 \\ \xi_3 = \beta_{13}\xi_1 + \beta_{23}\xi_2 + \zeta_2 \end{cases}$$

Is this coefficient significant ? $\Rightarrow$ $\mathrm{H}_0 : \beta_{23} = 0$, $\mathrm{H}_1 : \beta_{23} \neq 0$

Step 1. Estimate the complete model and compute the *GoF* index.

Step 2. «Deflate» the MV block X_3: $\widehat{X}_3 = X_3 - X_2\left(X_2{}'X_2\right)^{-1} X_2{}'X_3$

Step 3. Compute *Gof* index for the «null» model, GoF_{H_0}

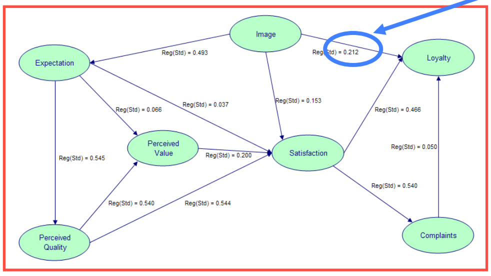
Step 4. Draw *B* samples from $\hat{F}_{[X_1, X_2, \hat{X}_3]}$, compute $GoF_{\mathrm{H}_0}^{(b)}$ and, for each *b*,

compute the ratio $\frac{GoF}{GoF_{\mathrm{H}_0}^{(b)}}$.

Step 5. By referring to the inverse cdf of $\frac{GoF}{GoF_{\mathrm{H}_0}^{(b)}}$ accept or reject H_0

The non parametric validation for one coefficient

Let us consider the complete model:



Is this standardized coefficient significant?

$$\mathrm{H}_0: \ \beta_{\mathrm{Ima,Loy}} = 0$$
$$\mathrm{H}_1: \ \beta_{\mathrm{Ima,Loy}} \neq 0$$

$$t = \frac{GoF}{GoF_{(\beta_{IMA,LOY}=0)}} = \frac{0.4645}{0.4628} = 1.0037$$

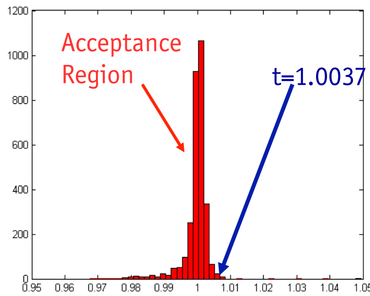
Thresholds, by *bootstrap*, for different confidence levels

$$\Phi_t^{(B)}(0.1) = 1.00238$$
$$\Phi_t^{(B)}(0.05) = 1.00316$$
$$\Phi_t^{(B)}(0.025) = 1.00411$$

$\boxed{t > 1.00316}$

We reject H_0

p-value= 0.035



Non parametric Bootstrap-based validation of all path coefficients

$$\mathrm{H}_0 : \beta_{12} = \beta_{13} = \beta_{23} = 0$$
$$\mathrm{H}_1 : \text{at least one of } \beta_{ij} (i = 1,2; \ j = 1,2,3; \ j > i) \neq 0$$

Step 1. Estimate the complete model and compute the *GoF* index.

Step 2. «Deflate» all MV blocks in accordance with the structural links (as we deal with recursive models, it is always possible to build blocks that verify the null hypothesis) :

$$X_2 : \hat{X}_2 = X_2 - X_1 (X_1' X_1)^{-1} X_1' X_2$$

$$X_3 : \hat{X}_3 = X_3 - \left(X_1 \mid \hat{X}_2\right)\left[\left(X_1 \mid \hat{X}_2\right)'\left(X_1 \mid \hat{X}_2\right)\right]^{-1}\left(X_1 \mid \hat{X}_2\right)' X_3$$

Step 3. Compute *Gof* index for the «null» model, GoF_{H_0}

Step 4. Draw *B* samples from $\hat{F}_{[X_1,\hat{X}_2,\hat{X}_3]}$ and compute $GoF_{\mathrm{H}_0}^{(b)}$ for each sample *b*

Step 5. By referring to the inverse cdf of $GoF_{\mathrm{H}_0}^{(b)}$, accept or reject H_0

Structural Model revision

Testing the path coefficients

Removing a link from the model…..

We can use the *t-test* to asses coefficient significance (for a large *n*)

H_0: $\beta = 0$

H_1: $\beta \neq 0$

$$T = \frac{B - 0}{s_B} \approx N$$

Decision rule:

We reject H_0 if: $|T| s_B > 2$

Contribution to the R^2

- Since the case values of LV 's are determined by the weight relations, the structural prediction may be assessed by looking at usual R^2's.

- The change in R2 can be explored to see whether the inclusion of a specific exogenous LV has a substantive impact (**effect size** f^2) on the predictive power

$$f^2 = \frac{R^2_{included} - R^2_{excluded}}{1 - R^2_{included}}$$

$f^2 \geq 0.02$ -> small impact
$f^2 \geq 0.15$ -> medium impact
$f^2 \geq 0.35$ -> large impact

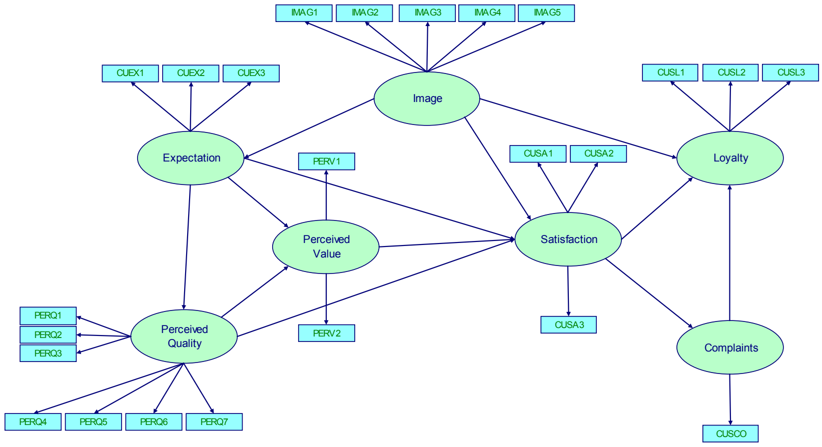
Suppression effects and Redundancy

- When the path coefficient and the correlation between latent constructs do not have the same sign, **the original relationship between the two has been suppressed**. This may be due to:
  - the **original relationship is so close to 0** that the difference in signs reflects random variation around 0: **non significant coefficients**
  - there are **redundant variables** that artificially change the signs; one or more redundant variables must be eliminated **(multicollinearity)**
  - **real suppression**: an important predictor variable, making the model correctly specified, suppresses the effect of another predictor variable. The correct sign interpretation is the one given by the path coefficient

- **How to detect** redundancy from real suppressor?
  - **Change model specification** (**remove a path**) and **check the change in R^2**. Redundancy does not provoke a decrease. Be cautious with **model trimming**.

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

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” ?

Study of the complete ECSI model

Fornell framework:

- Manifest variables are transformed from a scale "1-10" to a scale "0-100"

$$\tilde{x} = \frac{x-1}{9} \times 100$$

- Manifest Variables are left in the form of raw data
- "Classical" Estimation Options: Mode A + Centroid Scheme
- Each "latent variable" is estimated as a weighted average of its own manifest variables (keeps their same scale)

XLSTAT Output: Block Homogeneity

Checking for block homogeneity by means of classical reliability indeces

Composite reliability:

Latent variable	Dimensions	Cronbach's alpha	D.G. rho (PCA)	Condition number	Critical value	Eigenvalues
Image	5	0,714	0,815	2,302	397,494	954,332
						410,808
						238,904
						203,300
						180,126
Expectation	3	0,433	0,717	1,631	422,464	616,441
						419,132
						231,818
Perceived Quality	7	0,872	0,903	3,642	349,439	1400,080
						320,997
						212,997
						159,799
						127,368
						119,273
						105,558
Perceived Value	2	0,817	0,921	2,430	503,782	861,651
						145,912
Satisfaction	3	0,770	0,879	2,494	316,477	682,990
						156,596
						109,844
Complaints	1					
Loyalty	3	0,442	0,723	1,923	825,084	1190,249
						963,254
						321,749

XLSTAT Output: Block Homogeneity

Checking for block homogeneity by means of PCA

Variables/Factors correlations (Expectation / 1):

	F1	F2	F3
CUEX1	0.571	0.432	0.701
CUEX2	0.506	0.759	-0.414
CUEX3	0.868	-0.491	-0.099

Variables/Factors correlations (Loyalty / 1):

	F1	F2	F3
CUSL1	0.864	-0.365	-0.352
CUSL2	0.424	0.907	-0.033
CUSL3	0.774	-0.228	0.593

XLSTAT Result: Cross-loadings

Detection of monofactorial MVs

Cross-loadings (Monofactorial manifest variables / 1):

	Image	Expectation	Perceived Quality	Perceived Value	Satisfaction	Loyalty	Complaints
IMAG1	**0.717**	0.347	0.571	0.393	0.540	0.338	0.423
IMAG2	**0.566**	0.387	0.492	0.269	0.398	0.293	0.188
IMAG3	**0.658**	0.272	0.367	0.332	0.339	0.309	0.207
IMAG4	**0.792**	0.374	0.571	0.459	0.542	0.461	0.440
IMAG5	**0.698**	0.340	0.544	0.260	0.501	0.485	0.337
CUEX1	0.349	**0.687**	0.437	0.293	0.362	0.268	0.183
CUEX2	0.404	**0.644**	0.343	0.175	0.345	0.320	0.225
CUEX3	0.285	**0.726**	0.357	0.273	0.300	0.190	0.126
PERQ1	0.622	0.534	**0.778**	0.454	0.661	0.461	0.380
PERQ2	0.405	0.308	**0.651**	0.295	0.474	0.319	0.300
PERQ3	0.621	0.423	**0.801**	0.467	0.651	0.461	0.472
PERQ4	0.480	0.388	**0.760**	0.390	0.587	0.353	0.379
PERQ5	0.598	0.406	**0.732**	0.455	0.517	0.373	0.389
PERQ6	0.551	0.447	**0.766**	0.405	0.539	0.333	0.418
PERQ7	0.596	0.411	**0.803**	0.547	0.707	0.446	0.465
PERV1	0.405	0.314	0.477	**0.933**	0.495	0.435	0.287
PERV2	0.542	0.354	0.594	**0.911**	0.629	0.525	0.360
CUSA1	0.558	0.495	0.637	0.403	**0.711**	0.484	0.334
CUSA2	0.524	0.395	0.672	0.480	**0.872**	0.484	0.416
CUSA3	0.612	0.382	0.684	0.588	**0.885**	0.609	0.547
CUSL1	0.430	0.281	0.393	0.407	0.455	**0.855**	0.237
CUSL2	0.109	0.095	0.065	0.148	0.115	**0.273**	0.122
CUSL3	0.528	0.351	0.537	0.481	0.658	**0.869**	0.448
CUSCO	0.469	0.250	0.537	0.348	0.540	0.401	**1.000**

XLSTAT Output: Bivariate and Partial correlation between LVs

Checking for nomological and discriminant validity

Correlations (Latent variable) / Dimension (1):

	Image	Expectation	Perceived Quality	Perceived Value	Satisfaction	Complaints	Loyalty
Image	**1.000**	0.493	0.731	0.508	0.671	0.469	0.548
Expectation	0.493	**1.000**	0.545	0.360	0.481	0.250	0.366
Perceived Quality	0.731	0.545	**1.000**	0.576	0.791	0.537	0.524
Perceived Value	0.508	0.360	0.576	**1.000**	0.604	0.348	0.517
Satisfaction	0.671	0.481	0.791	0.604	**1.000**	0.540	0.635
Complaints	0.469	0.250	0.537	0.348	0.540	**1.000**	0.401
Loyalty	0.548	0.366	0.524	0.517	0.635	0.401	**1.000**

Partial correlations (Latent variable) / Dimension (1):

	Image	Expectation	Perceived Quality	Perceived Value	Satisfaction	Complaints	Loyalty
Image	**1,000**	0,141	0,363	0,053	0,081	0,087	0,185
Expectation	0,141	**1,000**	0,232	0,014	0,052	-0,102	0,052
Perceived Quality	0,363	0,232	**1,000**	0,149	0,458	0,180	-0,103
Perceived Value	0,053	0,014	0,149	**1,000**	0,188	-0,031	0,197
Satisfaction	0,081	0,052	0,458	0,188	**1,000**	0,170	0,314
Complaints	0,087	-0,102	0,180	-0,031	0,170	**1,000**	0,074
Loyalty	0,185	0,052	-0,103	0,197	0,314	0,074	**1,000**

XLSTAT Output: AVE

Checking for discriminant validity

→ (LV squared correlation < AVE)

Discriminant validity (Squared correlations < AVE) (Dimension 1):

	Image	Expectation	Perceived Quality	Perceived Value	Satisfaction	Complaints	Loyalty	Mean Communalities (AVE)
Image	**1**	0.243	0.535	0.258	0.450	0.220	0.300	0.476
Expectation	0.243	**1**	0.297	0.130	0.231	0.063	0.134	0.471
Perceived Quality	0.535	0.297	**1**	0.332	0.626	0.288	0.275	0.574
Perceived Value	0.258	0.130	0.332	**1**	0.365	0.121	0.267	0.850
Satisfaction	0.450	0.231	0.626	0.365	**1**	0.292	0.403	0.683
Complaints	0.220	0.063	0.288	0.121	0.292	**1**	0.161	
Loyalty	0.300	0.134	0.275	0.267	0.403	0.161	**1**	0.520
Mean Communalities (AVE)	0.476	0.471	0.574	0.850	0.683		0.520	**0**

XLSTAT Output: The Weights

- **Normalized outer weights** (sum up to 1) for computing raw scores
- **Bootstrap resampling** is used for statistical significance

Outer model (Dimension 1):

Weights (Dimension 1):

Latent variable	Manifest variables	Outer weight	Outer weight (normalized)	Outer weight (Bootstrap)	Std. deviation	Lower bound (95%)	Upper bound (95%)
Expectation	CUEX1	0.023	0.326	0.023	0.003	0.015	0.031
	CUEX2	0.022	0.317	0.022	0.005	0.013	0.033
	CUEX3	0.025	0.357	0.024	0.003	0.016	0.030
Perceived Quality	PERQ1	0.010	0.140	0.010	0.001	0.008	0.012
	PERQ2	0.009	0.121	0.009	0.001	0.006	0.013
	PERQ3	0.012	0.167	0.012	0.001	0.009	0.014
	PERQ4	0.009	0.134	0.009	0.001	0.008	0.011
	PERQ5	0.008	0.119	0.008	0.001	0.007	0.010
	PERQ6	0.009	0.135	0.009	0.001	0.008	0.012
	PERQ7	0.013	0.183	0.013	0.001	0.011	0.015
Perceived Value	PERV1	0.024	0.491	0.024	0.002	0.021	0.027
	PERV2	0.025	0.509	0.025	0.002	0.022	0.029
Satisfaction	CUSA1	0.016	0.242	0.016	0.002	0.012	0.019
	CUSA2	0.023	0.354	0.023	0.001	0.020	0.026
	CUSA3	0.026	0.404	0.026	0.002	0.022	0.029
Loyalty	CUSL1	0.018	0.393	0.018	0.001	0.016	0.021
	CUSL2	0.006	0.129	0.006	0.003	0.000	0.012
	CUSL3	0.022	0.477	0.022	0.002	0.019	0.026
Image	IMAG1	0.014	0.199	0.014	0.002	0.010	0.018
	IMAG2	0.013	0.174	0.013	0.002	0.009	0.018
	IMAG3	0.014	0.187	0.014	0.002	0.007	0.019
	IMAG4	0.018	0.242	0.017	0.002	0.014	0.021
	IMAG5	0.014	0.197	0.014	0.002	0.011	0.020
Complaints	CUSCO	0.040	1.000	0.040	0.002	0.036	0.043

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

XLSTAT Output: Latent Variable Scores

First 15 observations, out of 250

Latent variable scores (Dimension 1):

	Image	Perceived Value	Satisfaction	Complaints	Loyalty
Obs1	46,673	16,762	52,174	66,667	54,118
Obs2	95,875	100,000	91,030	100,000	88,497
Obs3	58,719	66,667	69,360	55,556	55,109
Obs4	76,129	44,444	100,000	44,444	91,373
Obs5	82,013	55,556	83,164	44,444	79,326
Obs6	86,066	100,000	69,360	77,778	89,935
Obs7	56,389	55,746	77,778	66,667	70,533
Obs8	79,361	55,746	62,734	77,778	63,234
Obs9	67,949	55,746	66,667	77,778	56,547
Obs10	69,342	44,635	49,831	55,556	70,155
Obs11	72,460	66,667	70,813	66,667	80,764
Obs12	77,870	55,746	73,293	66,667	40,131
Obs13	77,515	50,095	73,293	77,778	66,220
Obs14	100,000	72,317	83,164	77,778	91,373
Obs15	72,878	28,063	65,427	66,667	66,654

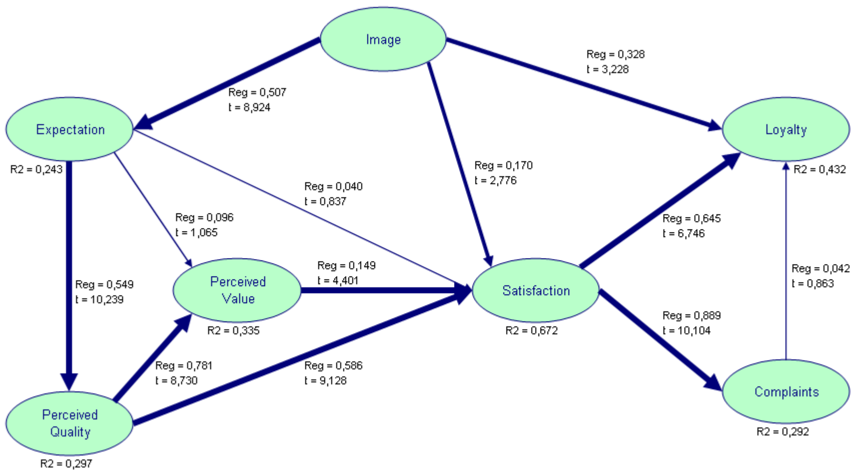
Example : Computation of the CS index:

$$CSI = \frac{0.01579 \times C\_sat1 + 0.02307 \times C\_sat2 + 0.02630 \times C\_sat3}{0.01579 + 0.02307 + 0.02630}$$

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

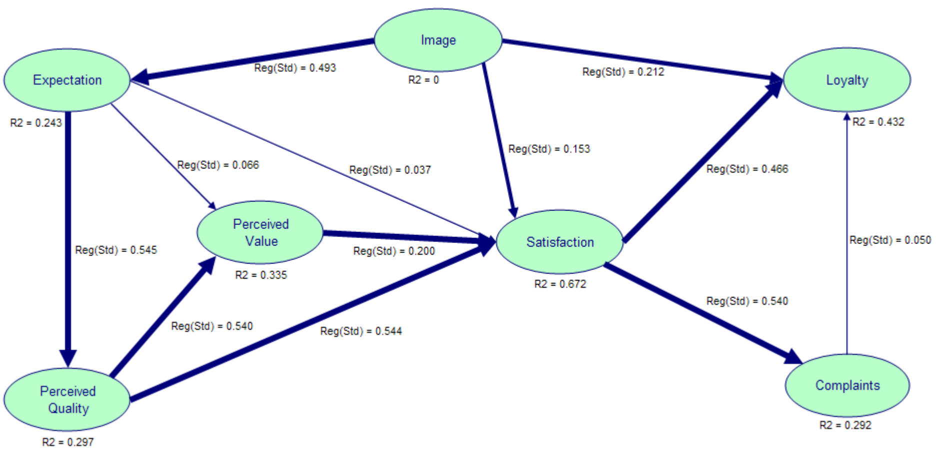
XLSTAT Output: Structural model

Path coefficients, value of the t-statistics and R^2



XLSTAT Output: Structural model

Standardized Path coefficients and R^2



XLSTAT Output: Structural model

Validation of Block « Satisfaction »

R² (Satisfaction / 1):

R²	R²(Bootstrap)	Std. deviation	Lower bound (95%)	Upper bound (95%)
0.672	0.673	0.047	0.567	0.749

Path coefficients (Satisfaction / 1):

Latent variable	Value	Standard error	t	Pr > |t|
Intercept	3.102	6.021	0.515	0.000
Image	0.170	0.061	2.776	0.006
Expectation	**0.040**	**0.048**	**0.837**	**0.403**
Perceived Quality	0.586	0.064	9.128	0.000
Perceived Value	0.149	0.034	4.401	0.000

Latent variable	Value(Bootstrap)	Standard error(Bootstrap)	Lower bound (95%)	Upper bound (95%)
Intercept	0.382	4.819	-8.894	11.973
Image	0.177	0.055	0.071	0.316
Expectation	0.039	0.055	**-0.063**	**0.154**
Perceived Quality	0.572	0.067	0.421	0.718
Perceived Value	0.152	0.054	0.059	0.244

Standardized coefficients (Satisfaction / 1):

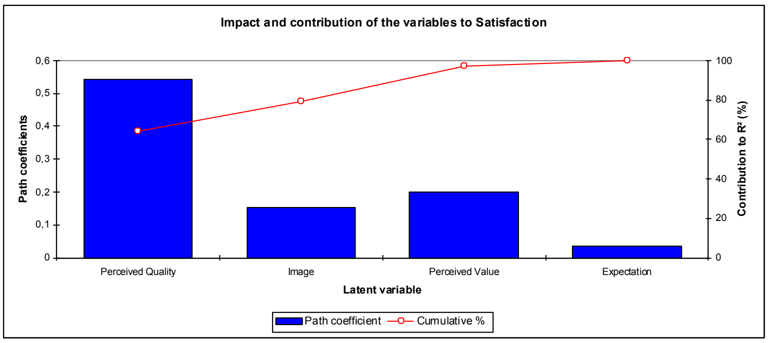
Latent variable	Value
Image	0.153
Expectation	0.037
Perceived Quality	0.544
Perceived Value	0.200

XLSTAT Output: Contribution to R^2

Relative importance of Satisfaction predictors

Impact and contribution of the variables to Satisfaction (Dimension 1):

	Perceived Quality	Image	Perceived Value	Expectation
Correlation	0,791	0,671	0,604	0,481
Path coefficient	0,544	0,153	0,200	0,037
Correlation * path coefficient	0,431	0,102	0,121	0,018
Contribution to R² (%)	64,134	15,242	17,968	2,656
Cumulative %	64,134	79,376	97,344	100,000



XLSTAT Output: MVs quality

MV Loadings, Communalities and Redundancies

Correlations (Dimension 1):

Latent variable	Manifest variables	Standardized loadings	Communalities	Redundancies
Expectation	CUEX1	0.687	0.471	0.115
	CUEX2	0.644	0.415	0.101
	CUEX3	0.726	0.527	0.128
Perceived Quality	PERQ1	0.778	0.605	0.180
	PERQ2	0.651	0.423	0.126
	PERQ3	0.801	0.641	0.191
	PERQ4	0.760	0.578	0.172
	PERQ5	0.732	0.536	0.159
	PERQ6	0.766	0.587	0.174
	PERQ7	0.803	0.644	0.191
Perceived Value	PERV1	0.933	0.870	0.291
	PERV2	0.911	0.829	0.278
Satisfaction	CUSA1	0.711	0.505	0.339
	CUSA2	0.872	0.760	0.510
	CUSA3	0.885	0.783	0.526
Loyalty	CUSL1	0.855	0.731	0.316
	CUSL2	0.273	0.075	0.032
	CUSL3	0.869	0.755	0.326
Image	IMAG1	0.717	0.514	
	IMAG2	0.566	0.320	
	IMAG3	0.658	0.433	
	IMAG4	0.792	0.627	
	IMAG5	0.698	0.487	
Complaints	CUSCO	1.000		0.292

XLSTAT Output: model assessment

Mean Communality and Redundancy

Model assessment (Dimension 1):

Latent variable	R²	Adjusted R²	Mean Communalities	Mean Redundancy
Image			0.476	
Expectation	0.243	0.243	0.471	0.115
Perceived Qua	0.297	0.297	0.574	0.170
Perceived Valu	0.335	0.332	0.850	0.285
Satisfaction	0.672	0.668	0.683	0.458
Complaints	0.292	0.292	1.000	0.292
Loyalty	0.432	0.427	0.520	0.225
Mean	0.378		0.570	0.257

$$GoF = (0.3784 * .5702)^{1/2} = 0.4645$$

XLSTAT Output: GoF

Validation by Bootstrap

Goodness of fit index (1):

	GoF	GoF (Bootstrap)
Absolute	0.4645	0.4702
Relative	0.9371	0.9110
Outer model	0.9957	0.9944
Inner model	0.9411	0.9162

Lower bound (95%)	Upper bound (95%)
0.4136	0.5262
0.8565	0.9448
0.9884	0.9977
0.8616	0.9482

XLSTAT Output:
Direct, Indirect and Total effects

Direct effects (Latent variable) / Dimension (1):

	Image	Expectation	Perceived Quality	Perceived Value	Satisfaction	Complaints	Loyalty
Image							
Expectation	0.507						
Perceived Quality	0.000	0.549					
Perceived Value	0.000	0.096	0.781				
Satisfaction	0.170	0.040	0.586	0.149			
Complaints	0.000	0.000	0.000	0.000	0.889		
Loyalty	0.328	0.000	0.000	0.000	0.645	0.042	

Indirect effects (Latent variable) / Dimension (1):

	Image	Expectation	Perceived Quality	Perceived Value	Satisfaction	Complaints	Loyalty
Image							
Expectation	0.000						
Perceived Quality	0.278	0.000					
Perceived Value	0.266	0.429	0.000				
Satisfaction	0.223	0.400	0.116	0.000			
Complaints	0.350	0.391	0.624	0.132	0.000		
Loyalty	0.268	0.300	0.479	0.102	0.037	0.000	

Total effects (Latent variable) / Dimension (1):

	Image	Expectation	Perceived Quality	Perceived Value	Satisfaction	Complaints	Loyalty
Image							
Expectation	0.507						
Perceived Quality	0.278	0.549					
Perceived Value	0.266	0.525	0.781				
Satisfaction	0.393	0.440	0.702	0.149			
Complaints	0.350	0.391	0.624	0.132	0.889		
Loyalty	0.596	0.300	0.479	0.102	0.683	0.042	

Blindfolding validation

	Communalities	Redundancies
Image	0,198	
Expectation	-0,016	0,057
Perceived Quality	0,401	0,090
Perceived Value	0,452	0,240
Satisfaction	0,388	0,465
Complaints		0,181
Loyalty	0,149	0,164

- Only the Customer Satisfaction block has an acceptable cross-validated redundancy index.
- Due to blindfolding, the cv-communality and the cv-redundancy measures may be negative.
- A negative value implies that the corresponding latent variable is badly estimated.

Other Component-based Approaches

Generalized Structured Component Analysis
Regularized Generalized CCA

Generalized Structured Component Analysis (GSCA)

GSCA (Hwang & Takane, Psychometrika 2004) integrates both the measurement and the structural model formulation in a unique equation:

$$\begin{bmatrix} \mathbf{x}_i \\ \xi_i \end{bmatrix} = \begin{bmatrix} \Lambda \\ \mathbf{B} \end{bmatrix} \xi_i + \begin{bmatrix} \varepsilon_i \\ \zeta_i \end{bmatrix}$$

Since $\xi_i = \mathbf{W}\mathbf{x}_i$

$$\underbrace{\begin{bmatrix} \mathbf{I} \\ \mathbf{W} \end{bmatrix} \mathbf{x}_i}_{\mathbf{u}_i} = \underbrace{\begin{bmatrix} 0 & \Lambda \\ 0 & \mathbf{B} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \mathbf{I} \\ \mathbf{W} \end{bmatrix} \mathbf{x}_i + \underbrace{\begin{bmatrix} \varepsilon_i \\ \zeta_i \end{bmatrix}}_{\mathbf{e}_i}$$

and so the model can be rewritten as: $\mathbf{u}_i = \mathbf{A}\mathbf{u}_i + \mathbf{e}_i$

GSCA optimizing function

Unlike PLS-PM, in GSCA a unique function is optimized:

$$\phi = \sum_{i=1}^{N} (\mathbf{u}_i - \mathbf{A}\mathbf{u}_i)^T (\mathbf{u}_i - \mathbf{A}\mathbf{u}_i) = SS(\mathbf{U} - \mathbf{U}\mathbf{A})$$

Under the constraint that latent variables are normalized, i.e. $\sum_{i=1}^{N} \xi_{iq}^2 = 1$

The unknown parameters of GSCA (W and A) are estimated so as to minimize the residual sum of squares by means of an ALS (Alternative Least Squares) algorithm

⚠ ALS does not assure convergence in a global minimum, several starting values are needed

GSCA optimizing function

Searching for standardized latent variables $\xi_q = \mathbf{X}_q \mathbf{w}_q$

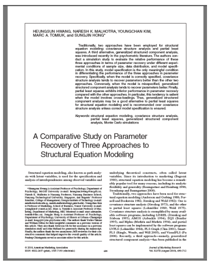
minimizing :

$$\underbrace{\sum_{\text{Block } \mathbf{X}_q \text{ reflective}} \left\| \mathbf{X}_q - \xi_q \lambda_q' \right\|^2}_{\text{PCA}} + \underbrace{\sum_{\substack{\xi_j \text{ endogenous,} \\ \xi_{q*} \text{ explaining } \xi_j}} \left\| \xi_j - \sum_{q*} b_{jq*} \xi_{q*} \right\|^2}_{\text{MSEV, Glang (1988)}}$$

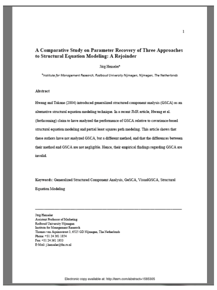
MSEV = Maximum Sum of Explained Variance

When the blocks are heterogeneous, GSCA might be trapped by PCA

GSCA vs. PLS-PM and Covariance-based SEM (simulated data)


A Comparative Study on Parameter Recovery of Three Approaches to Structural Equation Modeling

Hwang Heungsun *et al.*
A Comparative Study on Parameter Recovery of Three Approaches to Structural Equation Modeling.
Journal of marketing research, 2010, vol. 47, n.4, pp. 699-712


A Comparative Study on Parameter Recovery of Three Approaches to Structural Equation Modeling: A Rejoinder

Jörg Henseler.
A Comparative Study on Parameter Recovery of Three Approaches to Structural Equation Modeling: A Rejoinder (January 15, 2010).
Available at SSRN: http://ssrn.com/abstract=1585305

Hwang *et al.* (2010) tested the performance of PLS-PM, GSCA and the covariance-based approach to SEM under different simulation schemes:

→ When the model is correctly specified, covariance-based approach generally recovered parameters better than GSCA and PLS-PM (but with cross-loadings)

→When the model is misspecified, GSCA tends to recover parameters better than PLS-PM and covariance-based approach

However, Henseler (2010) claims:

"It seems like Hwang et al. (2010) did not apply GSCA at all, but only a reduced form of GSCA that ignores the structural model. This reduced form of GSCA provides estimates that differ substantially from those of GSCA. Consequently, these authors' empirical findings and conclusions are invalid and should be ignored."

Regularized Generalized Canonical Correlation Analysis (RGCCA)

A continuum between New Mode A and Mode B: RGCCA

Tenenhaus and Tenenhaus (Psychometrika, 2011) proposed a new framework, called Regularized Generalized Canonical Correlation Analysis (RGCCA) where a continuum is built between the covariance criterion (New Mode A) and the correlation criterion (Mode B) by means of a tuning parameter (Mode Ridge):

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

$$\text{s.t.} \quad \left(1-\tau_q\right)\operatorname{var}\left(\mathbf{X}_q \mathbf{w}_q\right) + \tau_q \left\|\mathbf{w}_q\right\|^2 = 1$$

Remarks:

- Choice of the tuning parameter for each block
- Matrix inversion (if tau different from 1)
- Interpretation of the composite scores

The PLS algorithm for RGCCA – Tenenhaus & Tenenhaus, 2011

MVs are centered or standardized

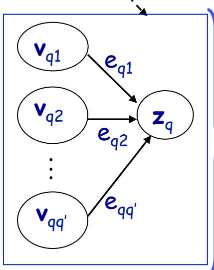
Initial step → $\mathbf{w}_q$

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

Outer component

v_q are unstandardized

$$\tau_q \|w_q\|^2 + (1-\tau_q) Var(X_q w_q) = 1$$



Inner components

z_q are unstandardized

Reiterate till *Numerical Convergence*

Mode RIDGE

$$w_q = \frac{[(\tau_q I + (1-\tau_q)\frac{1}{n} X_q^t X_q]^{-1} X_q^t z_q}{\sqrt{z_q^t X_q [(\tau_q I + (1-\tau_q)\frac{1}{n} X_q^t X_q]^{-1} X_q^t z_q}}$$

Update weights **W** - n_q can be $\leq p_q$, for $\tau_q > 0$

Choice of weights $e_{qq'}$:

Horst : $e_{qq'} = c_{qq'}$

Centroid: $e_{qq'} = c_{qq'}\ \text{sign}(\text{Cor}(v_q, v_{q'}))$

Factorial : $e_{qq'} = c_{qq'}\ \text{Cov}(v_q, v_{q'})$

$c_{qq'}$ = 1 if blocks are connected, 0 otherwise.

Multi-component estimation for Predictive PLS-PM

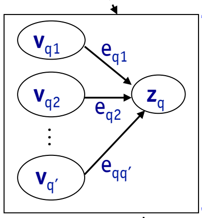
PLS Regression for outer model regularization in PLS-PM

Integrated PLS Regression-based Approach to PLS-PM algorithm

Initial step → $\mathbf{w}_q$

$\mathbf{v}_q \propto \pm\mathbf{X}_q\mathbf{w}_q$ if Mode PLScore
$\mathbf{v}_q = \pm\mathbf{X}_q\mathbf{w}_q$ if Mode PLScow

Outer estimation



Inner estimation
z_q are standardized if Mode PLScore
z_q are unstandardized if Mode PLScow

Choice of weights $e_{qq'}$:
- Centroid
- Factorial
- Path weighting scheme

Reiterate till *Numerical Convergence*

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

PLS Regression rationale

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

$$\boxed{\mathrm{Cov}^2(X_q a_{mq}, z_q) = \mathrm{Cor}^2(X_q a_{mq}, z_q) * \mathrm{Var}(X_q a_{mq})}$$

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 $\mathbf{t}_{1q}$ (with $\mathbf{x}_{pq}$ standardized as well):

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

2. **Normalization** of the vector $\mathbf{a}_{1q} = (a_{11q}, \dots, a_{1pq})$
3. **Regression** of $\mathbf{z}_q$ on $\mathbf{t}_{1q}=\mathbf{X}_q\mathbf{a}_{1q}$ 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 $\mathbf{t}_{1q}$: $\mathbf{z}_q = c_{1q}\mathbf{t}_{1q} + \mathbf{z}_{q1}$ and $\mathbf{X}_q = \mathbf{t}_{1q}\mathbf{p}'_{1q} + \mathbf{X}_{q1}$

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 $\mathbf{v}_q$:

$$\mathbf{z}_q = c_{1q}\mathbf{t}_{1q} + c_{2q}\mathbf{t}_{2q} + \dots + c_{mq}\mathbf{t}_{mq} + \text{Residual}$$

$$= c_{1q}\mathbf{X}_q\mathbf{a}_1 + c_2\mathbf{X}_{q1}\mathbf{a}_2 + \dots + c_m\mathbf{X}_{q(m-1)}\mathbf{a}_m + \text{Residual}$$

$$= \mathbf{X}_q(c_{1q}\mathbf{a}_{1q} + c_{2q}\mathbf{a}^*_{2q} + \dots + c_{mq}\mathbf{a}^*_{mq}) + \text{Residual}$$

$$= \underbrace{w_{1q}\mathbf{x}_{1q} + w_{2q}\mathbf{x}_{2q} + \dots + w_{pq}\mathbf{x}_{pq}}_{\mathbf{v}_q} + \text{Residual}$$

How should we normalize?

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

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:

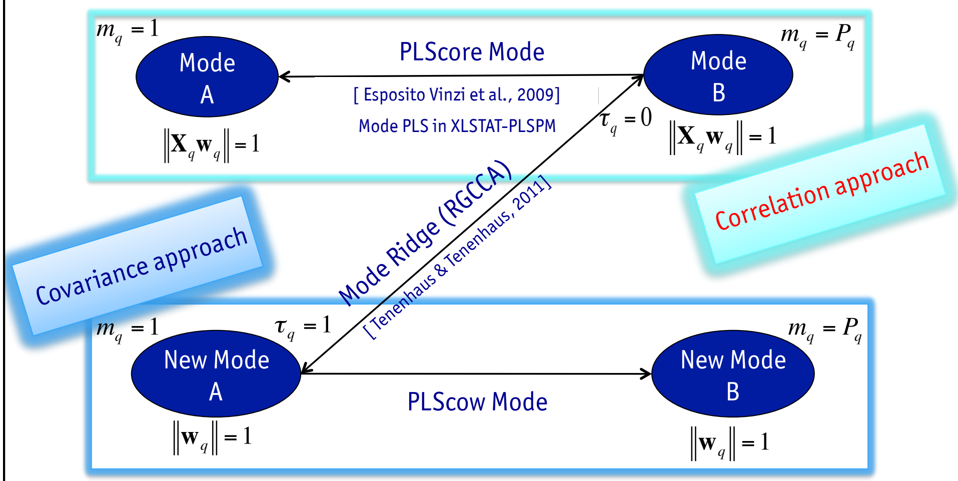
PLS Mode oriented to maximizing **cor**relations between connected composites under normalization constraints on composite s**core**s

PLScow Mode:

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

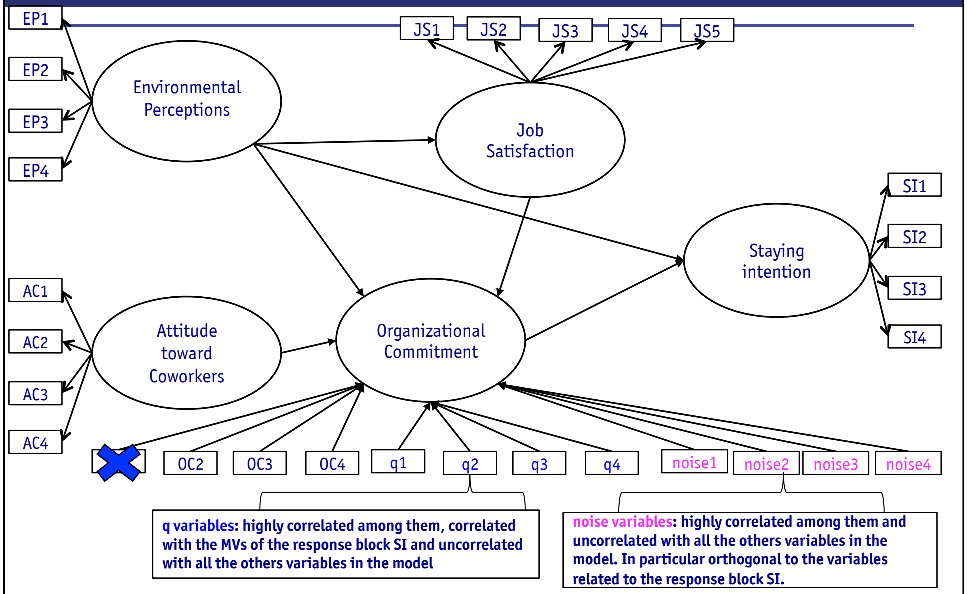
PLS regression Modes in PLS-PM and Ridge Mode in RGCCA

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



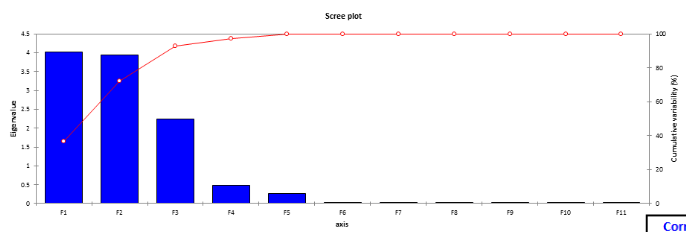
PLS-R as an estimation method for measurement model in a modified PLS-PM (normalization constraints on outer weights)

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



PCA of the Organizational Commitment (OC + noisy data)

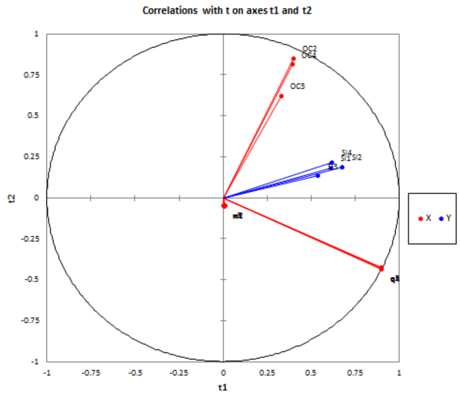
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
Eigenvalue	**4.021**	**3.947**	**2.246**	0.493	0.262	0.007	0.007	0.006	0.005	0.004	0.003
Variability (%)	36.555	35.884	20.416	4.478	2.380	0.066	0.061	0.054	0.042	0.038	0.026
Cumulative %	36.555	72.440	92.856	97.334	99.714	99.779	99.840	99.894	99.936	99.974	100.000



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)

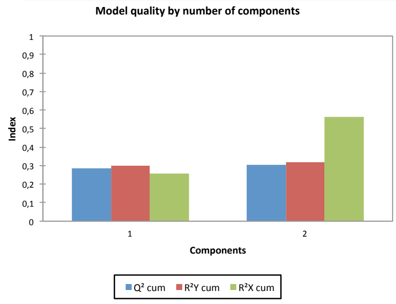


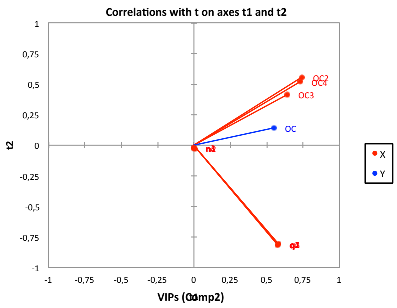
The noise variables are downweighted as they have no predictive power

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

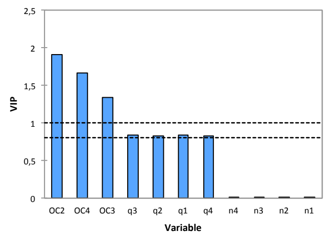
Variable	w*1	w*2
OC2	0.380	0.771
OC3	0.214	0.204
OC4	0.326	0.558
q1	0.416	-0.182
q2	0.420	-0.159
q3	0.423	-0.152
q4	0.417	-0.170
n1	0.000	-0.015
n2	0.003	-0.017
n3	-0.002	-0.015
n4	-0.001	-0.019

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 handled by using a 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..

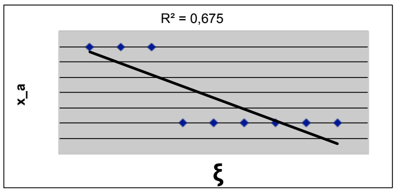
.. NOT to use boolean coding in PLS-PM

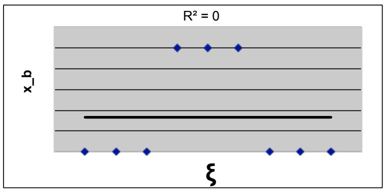
1) The number of categories affects the relative impact of categorical variables and generates sparse matrices.
2) It measures the impact of the single category, giving up the idea of the variable as a whole.
3) The importance of categories associated to central values of the LV distribution is systematically underestimated.

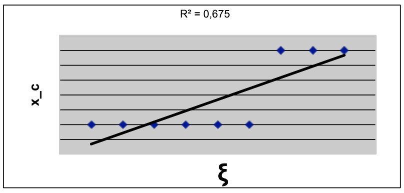
The relation between z_q and x_{pq}

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 had 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 variable 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 properties of the initial measurement scale that the researcher chooses 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 $\phi' = (\phi_1 \ldots \phi_K)$ 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

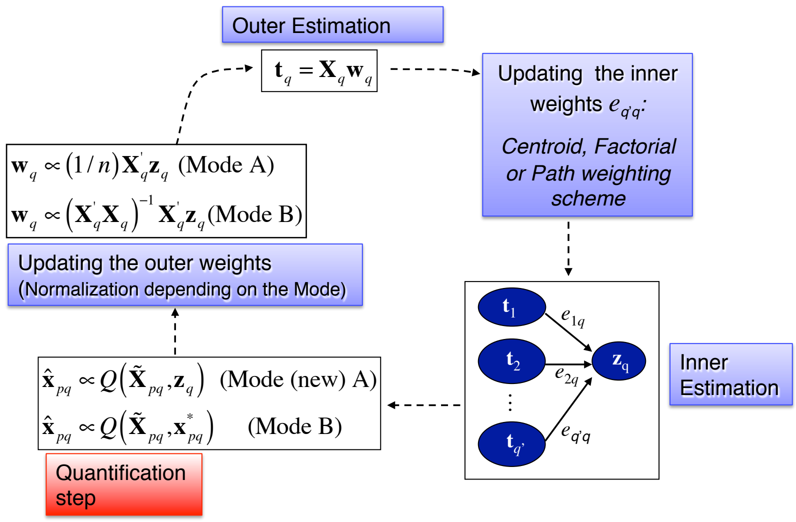
- It is coherent with the chosen scaling level;
- It optimizes the PLS criterion, if any.

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



NM-PLSPM general criterion

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

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

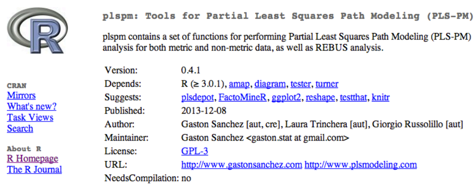
$$\left\|\hat{\mathbf{X}}_q \mathbf{w}_q\right\|^2 = n \quad \text{if Mode B for block } q$$

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

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

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

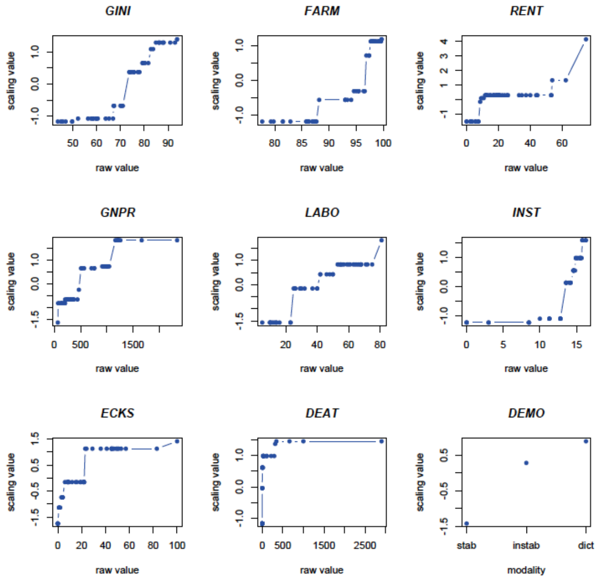
An application to the Russett data (1965)

- 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.

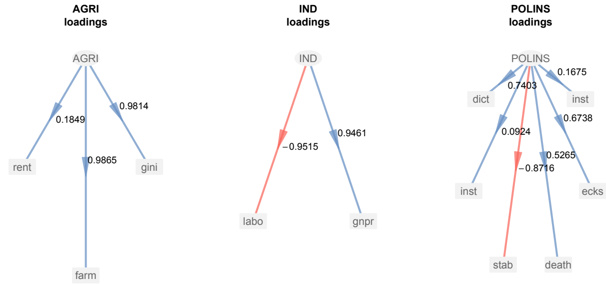
Russet data (1964): Quantifications



Russet data (1964): Model comparison

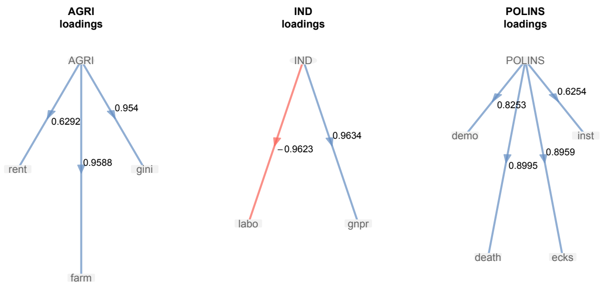
PLS Model:

- R2 = .605
- GoF = .567



NM-PLS Model:

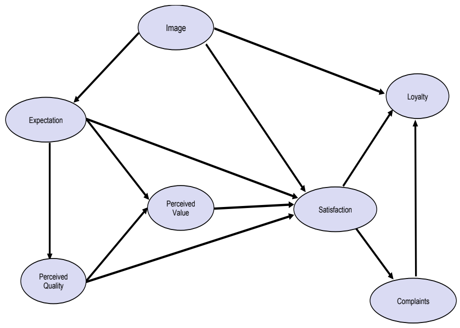
- R2 = .793
- GoF = .772



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)



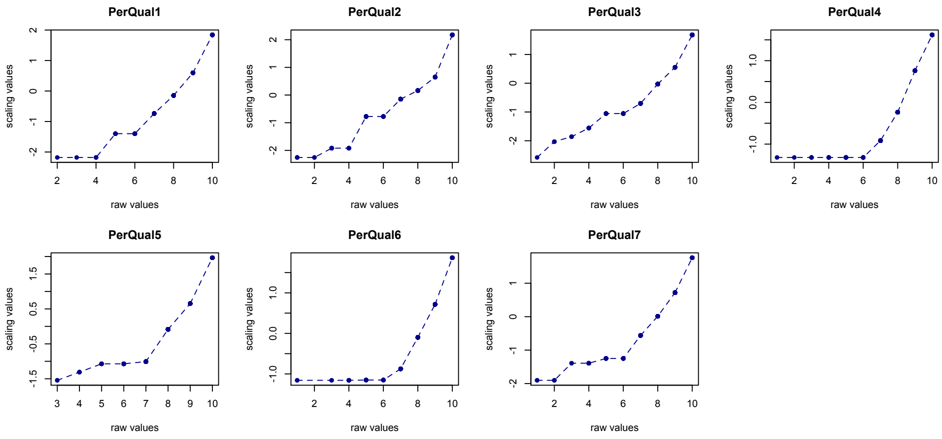
- Standardized MVs
- Centroid Scheme
- Mode A

Mobile data: Comparing model quality

Linearity hypothesis (No scaling)	Monotonicity hypothesis (Ordinal Scaling)	No hypothesis (Nominal Scaling)
$GoF = 0.470$	$GoF = 0.526$	$GoF = 0.547$
$\overline{R}^2 = 0.387$	$\overline{R}^2 = 0.464$	$\overline{R}^2 = 0.495$
$Com_M = 0.599$	$Com_M = 0.618$	$Com_M = 0.623$
$Red_M = 0.263$	$Red_M = 0.315$	$Red_M = 0.335$

Mobile data:
Ordinal quantification for perceived quality

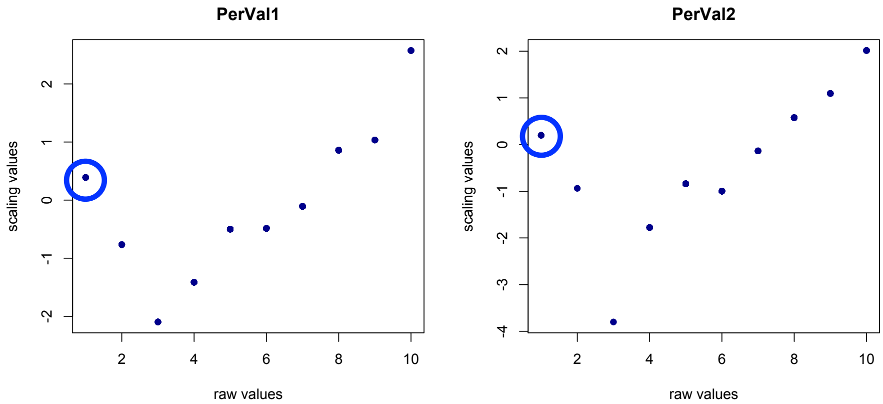
Perceived Quality Latent Variable: 7 indicators



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?

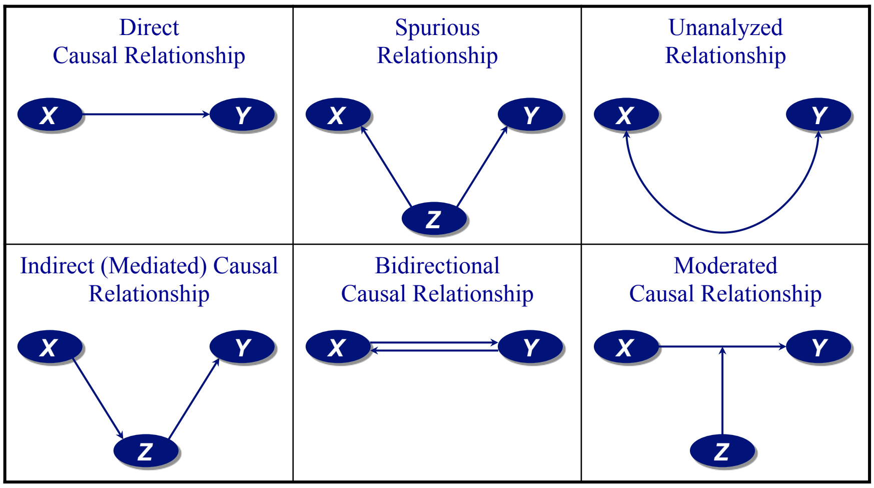


Mediation and Moderation*

•Exctract from W.W. Chin's slide on mediation and moderation

Kind of Relationships between LVs

Review of Causal Relationships between Latent Variables



(c.f. Jaccard/Turrisi, 2003)

Moderating effect

"In general terms, a moderator can be a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable"

Baron/Kenny, 1986, p. 1174

→ The effect of a moderator variable on the relation between two variables is called a "moderating effect" or an "interaction effect".

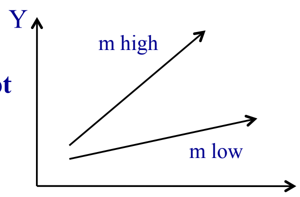
Example for moderated effect: *satisfaction-loyalty link*

- Moderated by socio-demographic variables (Homburg/Giering, 2001).
- Moderated by involvement (Bloemer/Kasper 1995)

Moderating effect

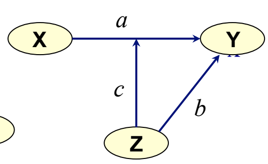
Example of moderator or interaction term

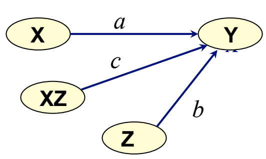
Graphical Plot



The slope of the exogenous variable *x* be no longer constant, but depend (linearly) on the moderator variable *z*.

Causal Diagram





Equation

$$Y = aX + bZ + cX * Z + \delta$$
$$= bZ + \left(a + cZ\right)X + \delta$$

Creating the Interaction Term

The way of creating the interaction term depends on the type of measurement models involved.

1. The measurement model of both the exogenous and the moderator variable is reflective
   - ⇨ Product-indicator approach
   - ⇨ Orthogonalizing approach

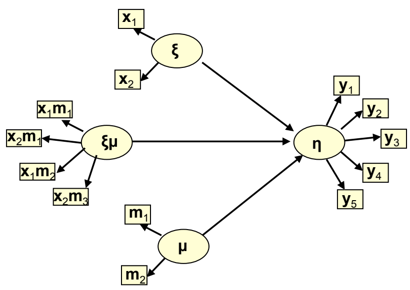
2. The measurement model of the exogenous or the moderator variable is reflective
   - ⇨ Two-stage approach

The Product-Indicator Approach

→ Mode A (arrows outwards) has to be used.

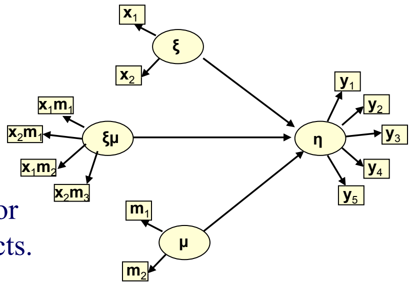
Idea:

1. Each of the P_1 indicators of the exogenous variable is elementwise multiplied with each of the P_2 indicators of the moderator variable, resulting in P_1*P_1 product indicators.
2. These product indicators serve as indicators of the interaction term.



Based on Kenny/Judd (1984) and Chin/Marcolin/ Newsted (1996, 2003).
Chin, Marcolin, & Newsted, 1996, Paper available at: **http://disc-nt.cba.uh.edu/chin/icis96.pdf**
Chin, Marcolin, & Newsted, 2003, *Information Systems Research,* 14 (2), 189-217.

The Product-Indicator Approach



Procedure:

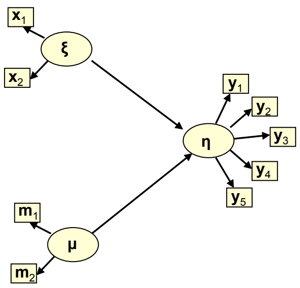
Step 1: Standardize or center indicators for the main and moderating constructs.

Step 2: Create all pair-wise product indicators where each indicator from the main construct is multiplied with each indicator from the moderating construct.

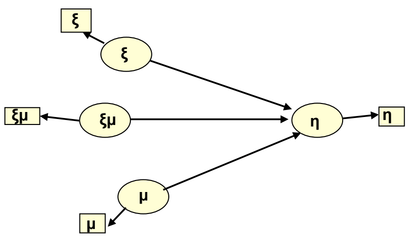
Step 3: Use the new product indicators to reflect the interaction construct.

Two-Stage Approach

1. Run the Main Effects Model



2. Estimating the Moderating Effect



Two-Stage Approach – formative MVs

Follow a two step construct score procedure:

Step 1: Use the formative indicators in conjunction with PLS to create underlying construct scores for the predictor and moderator variables.

Step 2: Take the single composite scores from PLS to create a single interaction term.

Attention: This approach has yet to be tested in a Monte Carlo simulation.

Remarks on Moderating effects

→ The path coefficients between latent variables no longer represent main effects but so-called "single effects".

→ A single effect expresses the strength of an effect when the moderator variable is zero.

→ It is indispensible to include all simple effects in the model.

→ Pay attention to the choices for Centering and Standardizing the Indicators

Centering or Standardizing the Indicators

Centering

→ Helps to circumvent problem of multicollinearity.

→ For prediction purposes, means have to be recovered.

Standardization

→ Same as for centering.

→ Beware of interpretational flaws!

→ A standardized interaction term is never interpretable.

In general, the elementwise product of two standardized manifest variables will not have a mean of zero and a standard deviation of one.

The interaction term's path coefficient as such is not interpretable.

Centering or Standardizing the Indicators

Correcting the Moderating Effect's Path Coefficient for Standardization Bias

One suggestion:

- Calculate the standard deviation of each product indicator.
- Create the weighted average of the standard deviations (weighted by the outer weights of the interaction term).
- Divide the interaction term's path coefficient by this weighted average.

Alternatively, the latent variable scores of the interaction term should be multiplied by the weighted average of the standard deviations of the product indicators, using the respective loadings of the product indicators as weights.

→ This corrected coefficient can be interpreted in relation to the moderated relation.

Determining the Strength of the Moderating Effect

Effect size

$$f^2 = \frac{R^2_{\text{model with moderator}} - R^2_{\text{model without moderator}}}{1 - R^2_{\text{model without moderator}}}$$

→ Effect sizes of 0.02/0.15/0.35 are regarded as weak/moderate/strong (Cohen, 1988).

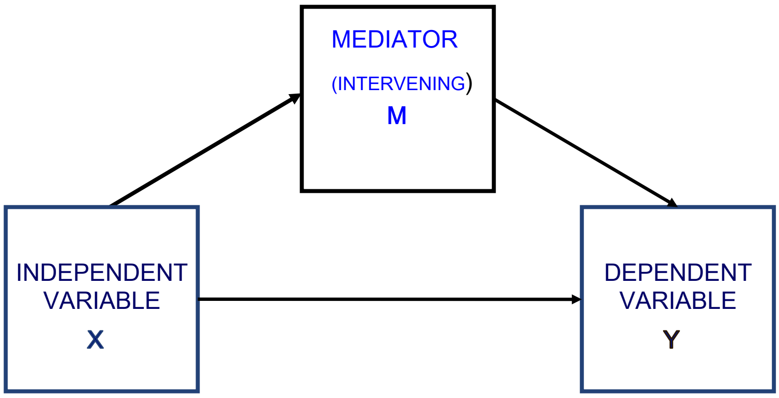
→ “Even a small interaction effect can be meaning-ful under extreme moderating conditions, if the resulting beta changes are meaningful, then it is important to take these conditions into account” (Chin/ Marcolin/Newsted 2003, p. 211).

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).

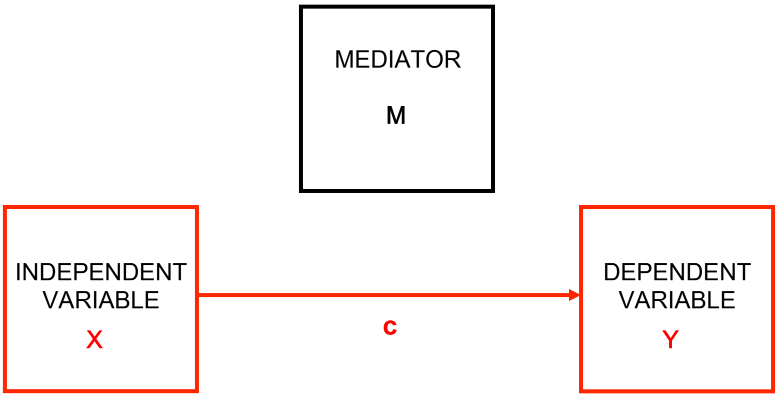
→ One of the most widely used methods to assess mediation in psychology.

→ 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:**

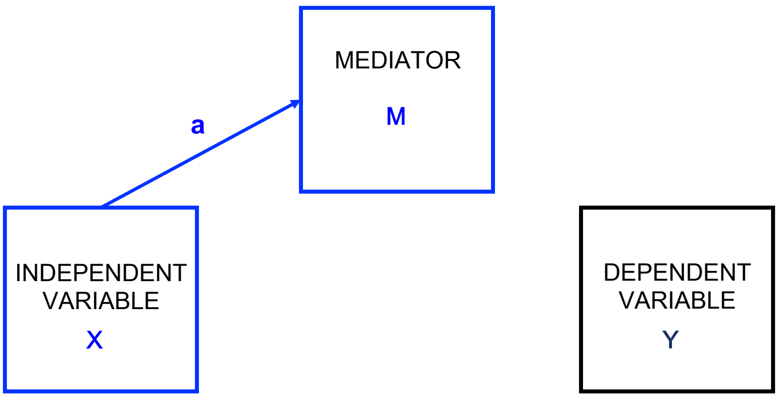
$$Y = i_1 + cX + e_1$$



Mediator model: *Direct effect of X on M*

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

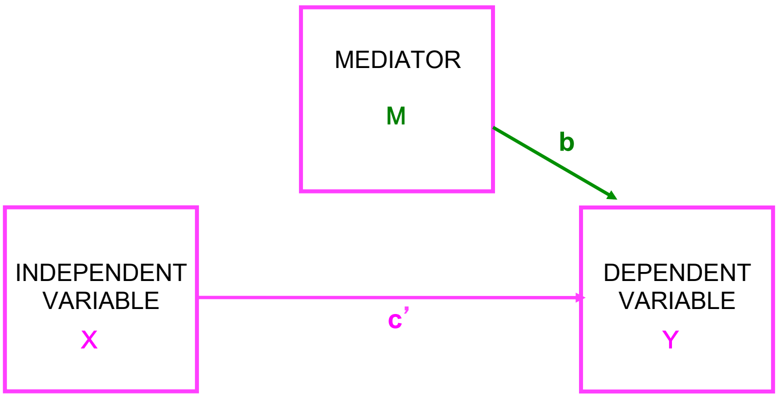
$$M = i_2 + aX + e_2$$



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

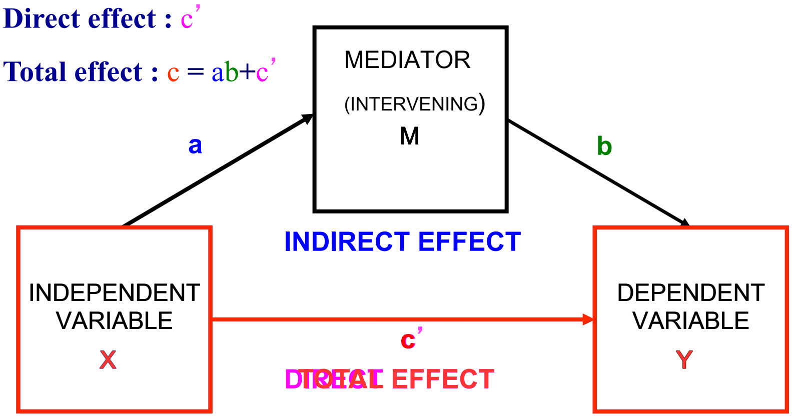
3. **The mediator causes the dependent variable controlling for the independent variable:**

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



Single mediator model

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



Test 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**.

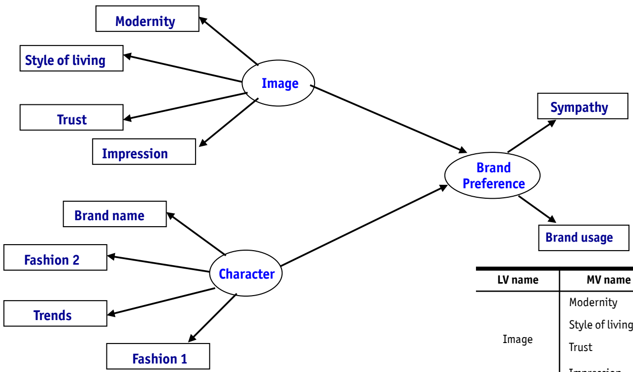
$$z = \frac{ab}{\sqrt{a^2 s_b^2 + b^2 s_a^2}}$$

Standandar error of the mediated effect

Unobserved Heterogeneity in PLS-PM

A path model for BENETTON brand preference

[Ringle et al., 2005]



LV name	MV name	Concept
Image	Modernity	It is modern and up to date
	Style of living	Represents a great style of life
	Trust	This brand can be trusted
	Impression	I have a clear impression of this brand
Character	Brand name	A brand name is very important to me
	Fashion2	I often talk about fashion
	Trends	I am interested in the latest trends
	Fashion1	Fashion is a way to express who I am
Brand Preference	Sympathy	Sympathy
	Brand Usage	Brand Usage

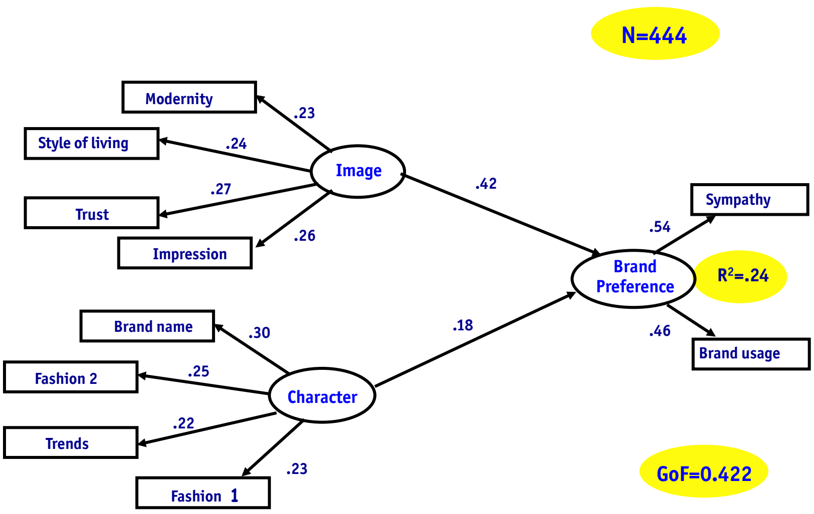
444 customers

10 manifest variables

3 variable blocks

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

PLS-PM estimates for BENETTON model



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Unobserved Heterogeneity in SEM

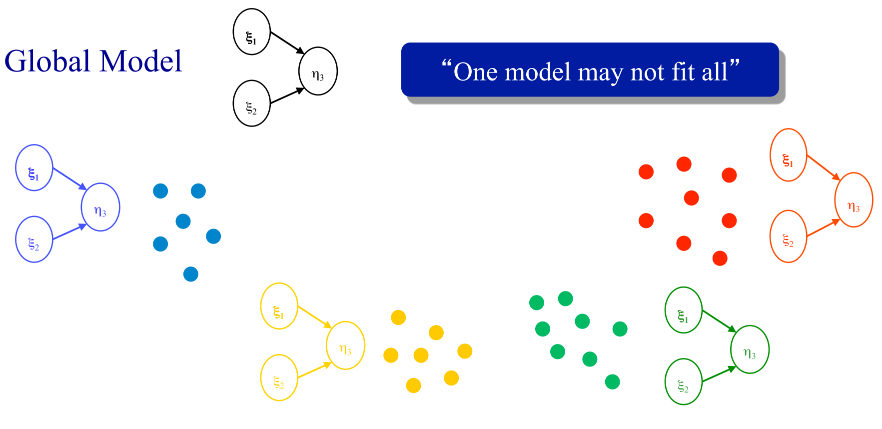
Traditional Structural Equation Models assume homogeneity across an entire population: data are treated as if they were collected from a single population

→ a unique model, i.e. the global model, is considered as explaining the behaviours of the whole set of units (the same set of parameter values applies to all individuals)

A unique model for all the observations may "hide" differences in consumer behaviours and may lead to biased results

ERCIM 2014 - Pisa, 5 December 2014 | Vincenzo Esposito Vinzi – ESSEC Paris

Unobserved Heterogeneity in SEM

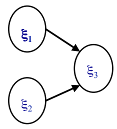


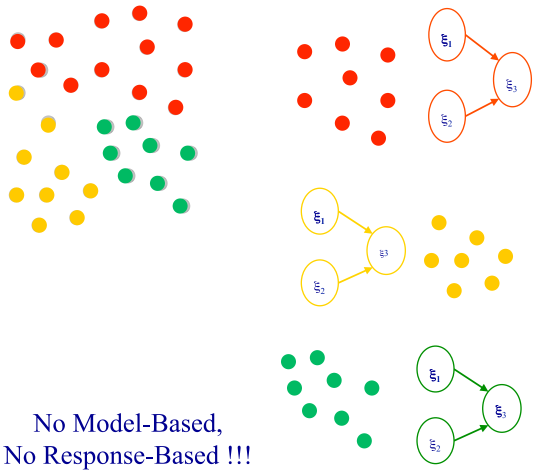
Different sets of individuals with particular behaviors, i.e. with different sets of parameters

ERCIM 2014 - Pisa, 5 December 2014 | Vincenzo Esposito Vinzi – ESSEC Paris

A priori unit segmentation

Global Model





Observed clustering variable:
e.g. **Education**

- Doctors
- Graduated
- Undergraduated

No Model-Based,
No Response-Based !!!

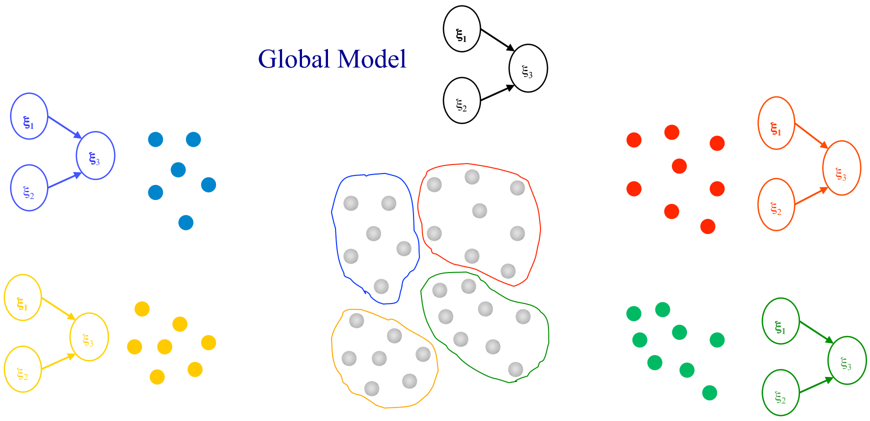
Unobserved Heterogeneity in PLS-PM

The need of searching for latent classes very often arises
in several business and marketing studies

Latent Classes are unobservable (latent) subgroups or segments of observations

Observations within the same latent class are homogeneous on certain criteria, while observations in different latent classes are dissimilar from each other in certain important ways

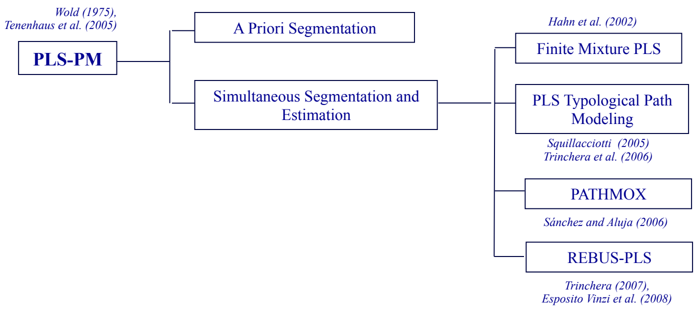
Unsupervised segmentation



How to detect latent classes??

Latent Class detection in component-based SEM

Several techniques to obtain unit clustering in Component-based Structural Equation Models



REsponse Based Unit Segmentation in PLS Path Modeling (REBUS-PLS): the idea behind

- Units showing similar performance (i.e. residuals!!) as regards the global model are considered to be represented by a unique model

- If a unit is assigned to the correct latent class, its performance (i.e. residuals!!) in the local model computed for that specific class will be better than the performance obtained by the same unit considered as supplementary in all the other local models

REBUS-PLS has a different implementation for outwards and inwards indicators

REBUS-PLS article in ASMBI (2008)

APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind. 2008; **24**:439–458
Published online 28 August 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.728

REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling

V. Esposito Vinzi[1,*,†], L. Trinchera[2], S. Squillacciotti[3] and M. Tenenhaus[4]

[1]*ESSEC Business School of Paris, Avenue Bernard Hirsch - B.P. 50105, 95021 Cergy-Pontoise, France*
[2]*DMS, University of Naples 'Federico II', Via Cintia 27, C.Monte Sant'Angelo, 80126 Naples, Italy*
[3]*Electricité de France, R&D, 1 Avenue du Général de Gaulle, 92141 Clamart, France*
[4]*HEC School of Management (GREGHEC), 1 Rue de la Libération, 78351 Jouy-en-Josas, France*

SUMMARY

Structural equation models (SEMs) make it possible to estimate the causal relationships, defined according to a theoretical model, linking two or more latent complex concepts, each measured through a number of observable indicators, usually called manifest variables. Traditionally, the component-based estimation of SEMs by means of partial least squares (PLS path modelling, PLS-PM) assumes homogeneity over the observed set of units: all units are supposed to be well represented by a unique model estimated on the overall data set. In many cases, however, it is reasonable to expect classes made of units showing heterogeneous behaviours to exist.

Two different kinds of heterogeneity could be affecting the data: observed and unobserved heterogeneity. The first refers to the case of *a priori* existing classes, whereas in unobserved heterogeneity no information is available either on the number of classes or on their composition.

REBUS-PLS: the closeness measure for PLS-PM with Outwards Indicators

Since the "distance" is a sum of squared residuals, it would be better defined as a measure of "closeness" of units to the model, i.e. a "closeness measure" (CM).

$$CM_{ik} = \sqrt{\frac{\sum_{q=1}^{Q}\sum_{p=1}^{P_q}\left[e^2_{ipqk}/Com\left(\xi_{qk},\mathbf{x}_{pq}\right)\right]}{\dfrac{\sum_{i=1}^{N}\sum_{q=1}^{Q}\sum_{p=1}^{P_q}\left[e^2_{ipqk}/Com\left(\xi_{qk},\mathbf{x}_{pq}\right)\right]}{(N-m_k-1)}} \times \frac{\sum_{j=1}^{J}\left[f^2_{ijk}/R^2\left(\xi_j,\left\{\xi_q\text{'s explaining }\xi_j\right\}\right)\right]}{\dfrac{\sum_{i=1}^{N}\sum_{j=1}^{J}\left[f^2_{ijk}/R^2\left(\xi_j,\xi_q\text{'s explaining }\xi_j\right)\right]}{(N-m_k-1)}}}$$

Number of extracted latent factors, so m_k=1 ALWAYS!

The Measurement residuals

The Structural residuals

REBUS-PLS: the Measurement Residuals for Outwards Indicators

For each *p-th* manifest variable in the q–th block…

…the Measurement Residual is obtained as the difference between the observed value of the *p-th* manifest variable and the corresponding estimated value, which is obtained by regression of x_{pq} on the *q-th* latent variable of the *k–th* group:

$$e_{ipqk} = x_{ipq} - \hat{x}_{ipqk}$$

Observed manifest variable

Estimated manifest variable:

$$\hat{x}_{ipqk} = \lambda_{pqk}\hat{\xi}_{iqk}$$

λ_{pqk} and $\mathbf{w}_{pqk}$ are **GROUP SPECIFIC!!**

where: $\hat{\xi}_{iqk} = \sum_{p=1}^{P_q} w_{pqk} x_{ipq}$

P Measurement Residuals are computed
for each *i-th* unit as regards the *k-th* group!

REBUS-PLS: the Structural Residuals

For each *j-th* endogenous latent variable…

…the Structural Residual is obtained as the difference between the *j-th* endogenous latent variable score and the inner estimation of the *j-th* latent variable, i.e. it is the residual of the multiple regressions of the endogenous latent variables on their explanatory latent variables:

$$f_{ijk} = \hat{\xi}_{ijk} - y_{ijk}$$

where:

$\hat{\xi}_{ijk} = \sum_{p=1}^{P_q} w_{pjk} x_{ipj}$ is obtained by using the **outer weights estimated for the *k-th* group**

GROUP SPECIFIC!!!

and $y_{ijk} = \sum_{q=1}^{Q\text{'s on } j} \beta_{qjk} \xi_{ijk}$ is obtained by using the **path coefficients estimated for the *k-th* group**

J Structural Residuals are computed
for each *i-th* unit as regards the *k-th* group!

ERCIM 2014 - Pisa, 5 December 2014 — Vincenzo Esposito Vinzi – ESSEC Paris

REBUS-PLS: the closeness measure for PLS-PM with Inwards Indicators

$$CM_{ik} = \sqrt{\frac{\sum_{q=1}^{Q}\left[g_{iqk}^2 / R^2\left(\xi_{qk}, X_q\right)\right]}{\frac{\sum_{i=1}^{N}\sum_{q=1}^{Q}\left[g_{iqk}^2 / R^2\left(\xi_{qk}, X_q\right)\right]}{(N - m_k - 1)}} \times \frac{\sum_{j=1}^{J}\left[f_{ijk}^2 / R^2\left(\xi_j, \left\{\xi_q\text{'s explaining } \xi_j\right\}\right)\right]}{\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\left[f_{ijk}^2 / R^2\left(\xi_j, \xi_q\text{'s explaining } \xi_j\right)\right]}{(N - m_k - 1)}}}$$

The Outer residuals

ERCIM 2014 - Pisa, 5 December 2014 — Vincenzo Esposito Vinzi – ESSEC Paris

REBUS-PLS: the Outer Residuals for Inwards Indicators

For each latent variable in the *q-th* reflective block…

…the Outer Residual is obtained as the difference between the latent variable score, obtained as linear combination of the manifest variable using the group-specific outer weights, and the last inner estimation of the latent variable:

$$g_{iqk} = \hat{\xi}_{iqk} - z_{iq}$$

Latent variable score obtained by using the **outer weights estimated for the *k-th* group**:

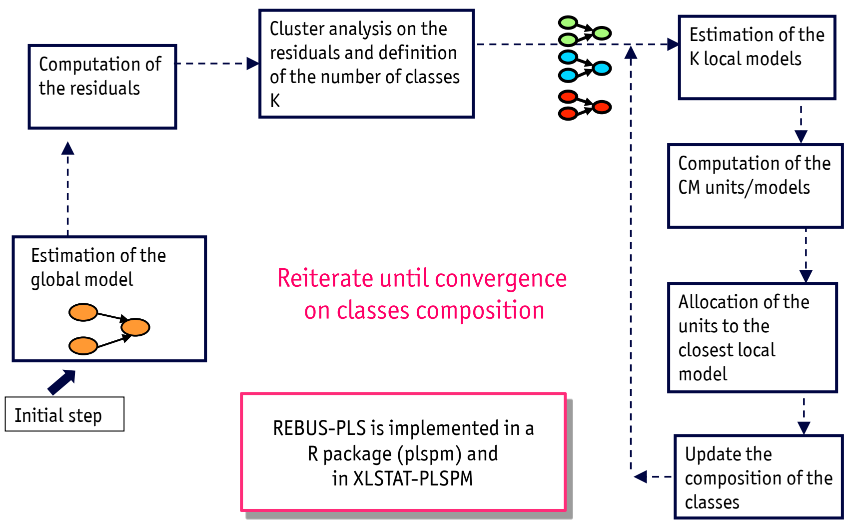
Inner estimation of the latent variable

$$\hat{\xi}_{iqk} = \sum_{p=1}^{P_q} w_{pqk} x_{ipq}$$

w_{pqk} is GROUP SPECIFIC!!

Q Structural Residuals are computed
for each *i*-th unit as regards the *k*-th group!

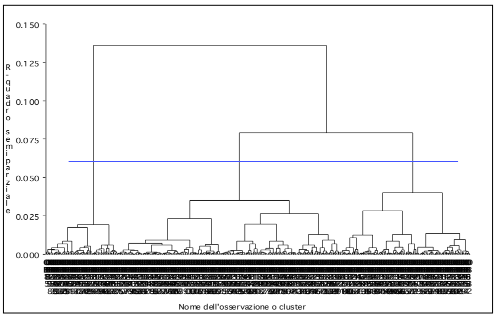
REBUS-PLS algorithm



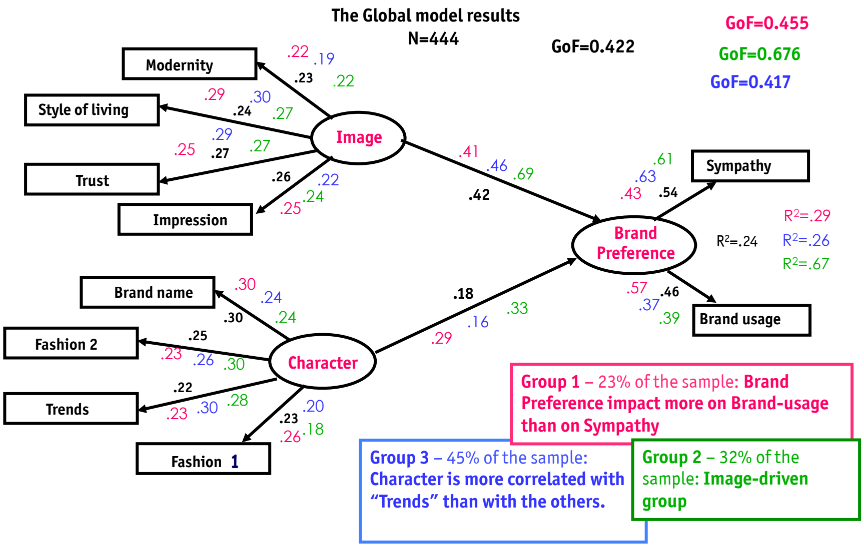
BENETTON data: REBUS-PLS results

How many classes?

Three in REBUS-PLS

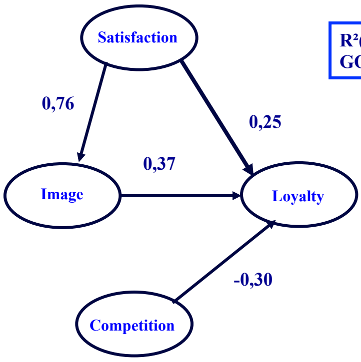


BENETTON data: REBUS-PLS results



A real case example

Satisfaction survey on professional customers of a electricity provider



R^2(loy)=0,47
GOF= 0,53

1459 customers

25 manifest variables

4 variable blocks

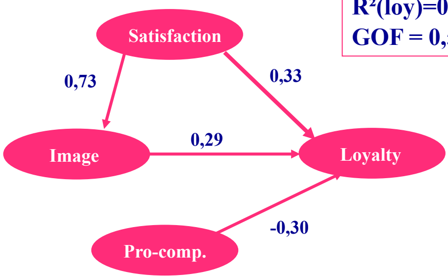
IMAGE with 13 manifest variables
SATISFACTION with 7 manifest variables
LOYALTY with 1 manifest variable
PRO-COMPETITION with 4 manifest variables

All blocks are FORMATIVE

→ Are there groups of customers showing **different behavioural models**?
→ Is the impact of the explanatory latent variables on loyalty the same for all customers?
→ Do customers show the same satisfaction model?

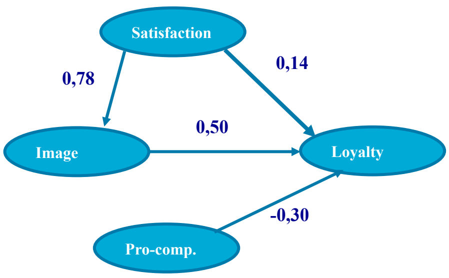
Electricity provider: the REBUS-PLS results

877customers
R^2(loy)=0,50
GOF = 0,53



Segment 1: 877 customers. **Satisfaction, Image** and **Pro-comp** have same impact **on Loyalty**

582 customers
R^2(loy)=0,47
GOF= 0,55



Segment 2: 582 customers. **Loyalty is very influenced by Image**; Satisfaction has a low effect on Loyalty

The Group Quality Index

The GQI → a new index based on residuals to assess if local models perform better than the global model

$$GQI_K = \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{P} \sum_{q=1}^{Q} \sum_{p=1}^{P_q} \left(1 - \frac{\sum_{i=1}^{n_k} e_{ipqk}^2}{\sum_{i=1}^{n_k} \left(x_{ipq} - \bar{x}_{pqk} \right)^2} \right) \times \frac{1}{J} \sum_{j=1}^{J} \left(1 - \frac{\sum_{i=1}^{n_k} f_{ijk}^2}{\sum_{i=1}^{n_k} \left(\xi_{ijk} - \bar{\xi}_{jk} \right)^2} \right) \right]}$$

GQI is a reformulation of the GoF index in a multi-group optic → If K=1 → GQI_1 = GoF of the global model!!

A **permutation test** can be performed to assess the quality of the detected partition

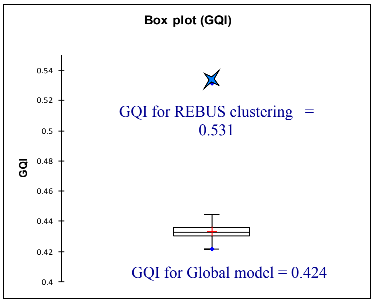
ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

BENETTON Data: evaluation of the partition

Simple Statistics	GQI
No. of observations	302
Minimum	0.422
Maximum	0.454
1° Quartile	0.426
Median	0.428
3° Quartile	0.431
Mean	0.429
Lower bound on mean (95%)	0.428
Upper bound on mean (95%)	0.429

300 random replications of the unit partition in 3 classes
+ the REBUS-based partition
+ the global model

Empirical distribution of the GQI values



ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Multi-group comparison

Multigroup comparison in PLS-PM

A form of PLS-PM analysis where two or more samples of respondents are compared using similar models.

Main question addressed:

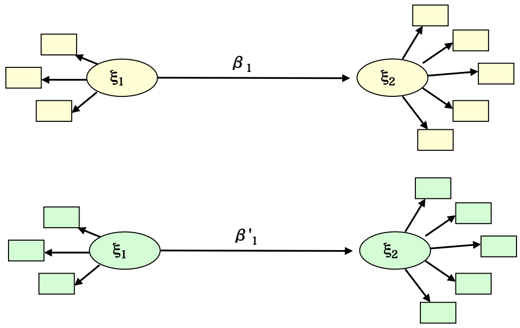
- Do values of model parameters vary across groups?
- Does group membership moderate the relations specified in the model?
- Is there an interaction between group membership and exogenous variables in effect on endogenous variables?

Local models can be compared according to differences in:

- Structural model parameters
- Measurement model parameters
- Latent variable scores
- Quality indexes

The **model structure** is considered **constant across the different classes**

Multigroup comparaison in PLS-PM



Are there significant differences between the parameters associated with each subsample?

Few methods are proposed in literature to compare groups by pairs:

- A t-test (classical parametric approach but based on bootstrap)
- Non-parametric resampling approach for confidence intervals
- Non-parametric approach for computing empirical p-value
- A permutation-based test
- Moderating variables → moderating effect = $\beta_1 - \beta'_1$

Detecting Moderating Effects through Group Comparisons

- Idea:
  - Split sample into two (or more) subsamples (categories).
  - Categorize observations according to the level of the moderator variable.
  - Estimate the path coefficients through PLS path modeling for each subsample.
  - Differences between path coefficients are interpreted as moderating effects.

The (pragmatic) t tests for path coefficients

The null hypothesis is :

$$H_0 : \beta_{ij}^{Group1} = \beta_{ij}^{Group2}$$

To compare structural (path) coefficients, the following t-test is used (with n units in sample 1 and m units in sample 2 and standard error estimates yielded by bootstrap re-samples):

$$t = \frac{Path_{sample\_1} - Path_{sample\_2}}{\left[\sqrt{\frac{(m-1)^2}{(m+n-2)} * S.E.^2_{sample1} + \frac{(n-1)^2}{(m+n-2)} * S.E.^2_{sample2}}\right] * \left[\sqrt{\frac{1}{m} + \frac{1}{n}}\right]}$$

Based on the use of S.E. estimates in a parametric sens, this would follow a t-distribution with m+n-2 degrees of freedom

(ref: http://disc-nt.cba.uh.edu/chin/plsfaq.htm)

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

The t tests for path coefficients

If the variances are assumed different, a Smith-Satterthwait test can be applied to approximate a level α test.

$$t = \frac{Path_{sample\_1} - Path_{sample\_2}}{\sqrt{S.E.^2_{sample1} + S.E.^2_{sample2}}}$$

$$df = \text{round to nearest integer}\left[\frac{(S.E.^2_{sample1} + S.E.^2_{sample2})^2}{\left(\frac{S.E.^2_{sample1}}{m+1} + \frac{S.E.^2_{sample2}}{n+1}\right)} - 2\right]$$

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

The t-test – Properties

→ Structural coefficients can be compared using bootstrap standard errors (i.e. standard error estimates from empirical bootstrap distributions).

→ It is a parametric-based test (mind the assumptions on independence of subsamples, normality and equal variances)

→ This approach works reasonably well if the two subsamples are not too far from normality and/or the two variances are not too different from one another

→ Empirical boostrap-based confidence intervls could be also built for the pair-wise differences between coefficients related to different groups.

Non parametric approach for an empirical p-value (Henseler and Fassot ,2009)

Four steps:

1. For each group, estimate the parameter and set the hypotheses
2. For each group, build G bootstrap samples and compute the G estimates for the parameter of interest
3. Build all the possible combinations (G^K) of the bootstrap parameters
4. Count how often, in the G^K combinations, the null hypothesis is rejected

Increasing the number of classes directly increases the number of bootstrap sample combinations to take into account

Non parametric approach for an empirical p-value (Henseler and Fassot ,2009)

Example:

$$H_o : \beta_{jj^*}^{Group1} \geq \beta_{jj^*}^{Group2}$$

$$P\left(\hat{\beta}_{jj^*}^{Group1} \geq \hat{\beta}_{jj^*}^{Group2}\right) = 1 - \frac{1}{G^2}\sum_{g=1}^{G}\sum_{s=1}^{G} I\left(\hat{\beta}_{jj^*}^{Group1(g)} < \hat{\beta}_{jj^*}^{Group2(s)}\right)$$

where:

$$I\left(\hat{\beta}_{jj^*}^{Group1(g)} < \hat{\beta}_{jj^*}^{Group2(s)}\right) = \begin{cases} 1 & \text{if } \hat{\beta}_{jj^*}^{Group1(g)} < \hat{\beta}_{jj^*}^{Group2(s)} \\ 0 & \text{otherwise} \end{cases}$$

Permutation tests (Chin, 2003)

You can compare:

- Structural (path) coefficients
- Outer weights
- Goodness of fit quality indexes
- Means of latent variable scores

Ref: Chin, W. W. (2003). A Permutation Procedure For Multi-Group Comparison Of PLS Models. Proceedings of the PLS'03 International symposium -- "*Focus on Customers*", Lisbon, September 15th to 17th, 2003, pp. 33-43.

These tests are based on the permutation of the elements of the sub-samples:

1. Merge the two sub-samples into a single one (case of K=2 groups);
2. Draw (at random) two new sub-samples of, respectively, size n and m from the overall sample;
3. Repeat the previous step B times (usually, B> 500);
4. Run the analysis on the B pairs of sub-samples (within the same draw);
5. Get the empirical distribution of the B differences between estimated parameters;
6. Compute empirical p-values (based on percentiles);
7. Draw conclusions on statistical significance of the difference.

Permutation tests

- Fisher "introduced the idea of permutation testing more as a theoretical argument supporting Student's t-test than as a useful statistical method in its own right." (Efron and Tibshirani, 1993)
- With modern computational power available for permutation tests to be used on a routine basis, the reliance on parametric tests as an approximation is no longer necessary.
- When samples are very large, decision based on parametric tests like the t and F tests usually agree with decisions based on the corresponding permutation test.
- But with small samples, "the parametric test will be preferable IF the assumptions of the parametric test are satisfied completely" (Good, 2000, p. 9). Otherwise, even for large samples, the permutation test is usually as powerful as the most powerful parametric test and may be more powerful when the test statistic does not follow the assumed distribution (Noreen, pp. 32-41).

Ref: Chin, W. W. (2003). A Permutation Procedure For Multi-Group Comparison Of PLS Models. Proceedings of the PLS'03 International symposium -- "*Focus on Customers*", Lisbon, September 15th to 17th, 2003, pp. 33-43.

Permutation tests

Principle:

- Null hypothesis: "Difference between parameters equals 0"
- Used statistic : $S = |param_{G1} - param_{G2}|$

Steps :

1. Select a statistic S and compute S_{init} on the original samples
2. Permute all elements of your sample and compute $S_{permut(i)}$ on the obtained subsamples.
3. Repeat step (2) B times (with B high)
4. Compare S_{init} and $S_{permut(i)}$ in order to obtain a probability (empirical p-value) showing if S_{init} is significantly different from $S_{permut(i)}$.

Permutation tests - Properties

Empirical p-value is defined as:

$$P = \frac{1}{B-1} \sum_{i=1}^{B} I_{S_{init} < S_{permut_i}}$$

If P<0.05, then the difference is significant.

Remarks:

- This test is a non parametric test well suited to PLS Path modeling.
- It has a very high computational cost (B has to be high). That is why this approach might take quite a long time if it is used to compare more than two groups.

Group Comparison: an application

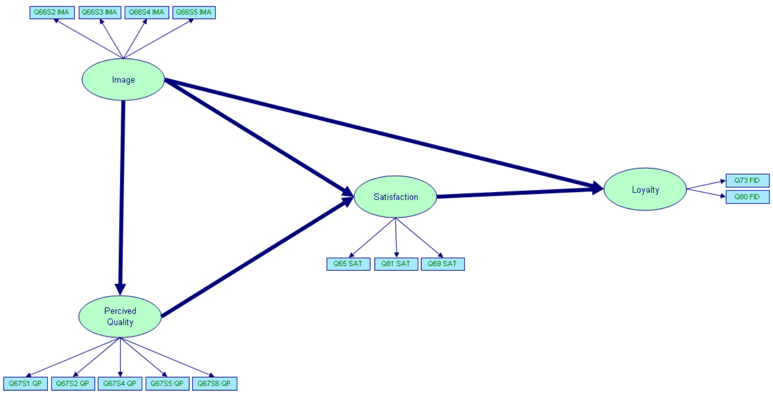
The banking sector in France (survey in 1999)

Number of manifest variables: 14

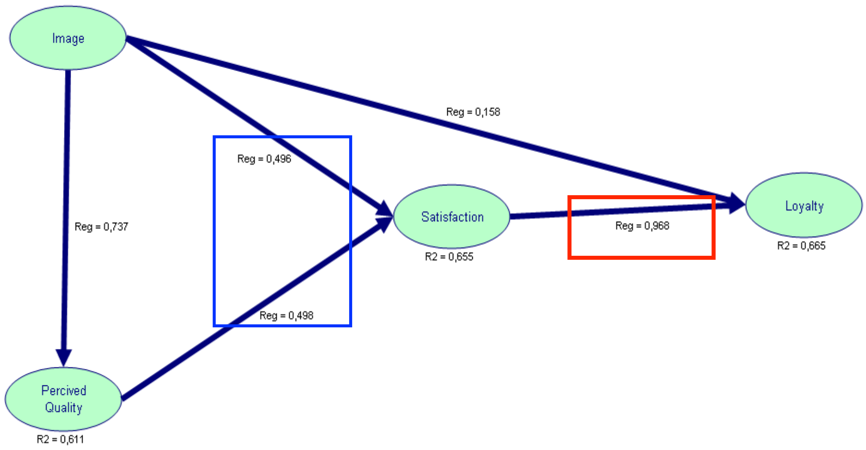
Overall sample size: 1510

Clients of 6 banks

Model:



Inner model parameters



Comparing structural coefficients across the 6 banks: t-test

Image – Satisfaction

Groups	Difference	t (Observed value)	t (Critical value)	DF	p-value	alpha	Significant
2 vs 1	0,106	1,292	1,648	504	0,098	0,050	No
3 vs 1	0,106	1,104	1,648	500	0,135	0,050	No
3 vs 2	0,001	0,008	1,648	502	0,497	0,050	No
4 vs 1	0,124	1,396	1,648	501	0,082	0,050	No
4 vs 2	0,019	0,237	1,648	503	0,406	0,050	No
4 vs 3	0,018	0,192	1,648	499	0,424	0,050	No
5 vs 1	0,140	1,571	1,648	500	0,058	0,050	No
5 vs 2	0,034	0,436	1,648	502	0,331	0,050	No
5 vs 3	0,034	0,359	1,648	498	0,360	0,050	No
5 vs 4	0,016	0,182	1,648	499	0,428	0,050	No
6 vs 1	0,323	2,997	1,648	503	0,001	0,050	Yes
6 vs 2	0,218	2,194	1,648	505	0,014	0,050	Yes
6 vs 3	0,217	1,943	1,648	501	0,026	0,050	Yes
6 vs 4	0,199	1,888	1,648	502	0,030	0,050	Yes
6 vs 5	0,184	1,738	1,648	501	0,041	0,050	Yes

Bank 6 is different from the other banks as concerns the relation between image and satisfaction.

Comparing structural coefficients : t-test

Perceived quality – Satisfaction

Groups	Difference	t (Observed value)	t (Critical value)	DF	p-value	alpha	Significant
2 vs 1	0,077	0,926	1,648	504	0,178	0,050	No
3 vs 1	0,071	0,717	1,648	500	0,237	0,050	No
3 vs 2	0,007	0,072	1,648	502	0,471	0,050	No
4 vs 1	0,110	1,325	1,648	501	0,093	0,050	No
4 vs 2	0,033	0,428	1,648	503	0,334	0,050	No
4 vs 3	0,040	0,426	1,648	499	0,335	0,050	No
5 vs 1	0,095	1,061	1,648	500	0,145	0,050	No
5 vs 2	0,017	0,207	1,648	502	0,418	0,050	No
5 vs 3	0,024	0,243	1,648	498	0,404	0,050	No
5 vs 4	0,016	0,189	1,648	499	0,425	0,050	No
6 vs 1	0,264	2,656	1,648	503	0,004	0,050	Yes
6 vs 2	0,187	1,984	1,648	505	0,024	0,050	Yes
6 vs 3	0,193	1,795	1,648	501	0,037	0,050	Yes
6 vs 4	0,154	1,638	1,648	502	0,051	0,050	No
6 vs 5	0,170	1,708	1,648	501	0,044	0,050	Yes

Bank 6 is different from the other banks (but Bank 4)
as concerns the relation between Perceived quality and Satisfaction.

Comparing structural coefficients: Permutation test

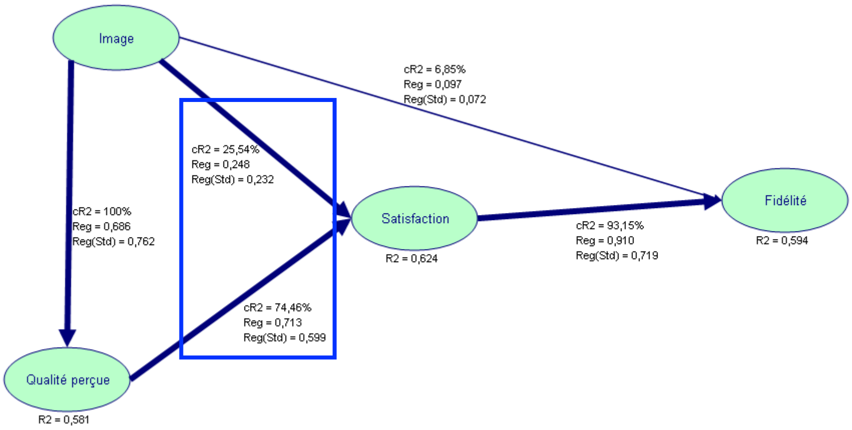
Comparing bank 1 to bank 6

Latent variables	Difference	p-value	alpha	Significant
Image -> QP	0,006	0,918	0,050	No
Image -> sat	0,324	0,025	0,050	Yes
QP -> sat	0,263	0,042	0,050	Yes
Image -> fid	0,093	0,395	0,050	No
sat -> fid	0,042	0,649	0,050	No

1000 permutations

Same conclusions as with the t-test

Results on bank 6



→ **Perceived quality has an high impact on satisfaction.**

Comparing bank 1 to bank 6 with the permutation test

Model quality

Qualité du modèle (Variable latente)	Différence	p-value	alpha	Significatif
Communalité (Image)	0,071	0,101	0,050	Non
Communalité (QP)	0,067	0,109	0,050	Non
Communalité (sat)	0,005	0,894	0,050	Non
Communalité (fid)	0,083	0,024	0,050	Oui
Redondance (Image)				Non défini
Redondance (QP)	0,035	0,595	0,050	Non
Redondance (sat)	0,057	0,375	0,050	Non
Redondance (fid)	0,111	0,092	0,050	Non
GoF	0,011	0,770	0,050	Non

Comparing bank 1 to bank 6 with the permutation test

Outer model

Latent variable	Manifest variable	Difference	p-value	alpha	Significant
Image	Q66S2 IMA	0,070	0,052	0,050	No
	Q66S3 IMA	0,417	0,003	0,050	Yes
	Q66S4 IMA	0,034	0,639	0,050	No
	Q66S5 IMA	0,052	0,344	0,050	No
QP	Q67S1 QP	0,005	0,955	0,050	No
	Q67S2 QP	0,001	0,984	0,050	No
	Q67S4 QP	0,231	0,060	0,050	No
	Q67S5 QP	0,044	0,512	0,050	No
	Q67S6 QP	0,013	0,700	0,050	No
sat	Q65 SAT	0,067	0,055	0,050	No
	Q81 SAT	0,067	0,038	0,050	Yes
	Q69 SAT	0,008	0,783	0,050	No
fid	Q73 FID	0,103	0,002	0,050	Yes
	Q80 FID	0,009	0,644	0,050	No

Comparing latent variable scores (means)

Using the permutation test, we have:

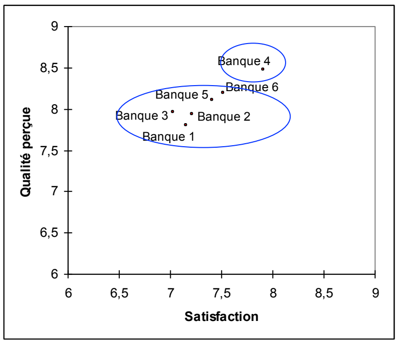
	1	2	3	4	5	6
Banque 1	0	0.06	0.12	0.76**	0.26	0.37
Banque 2		0	0.18	0.70**	0.20	0.31
Banque 3			0	0.88**	0.36	0.49**
Banque 4				0	0.50**	0.39
Banque 5					0	0.11
Banque 6						0

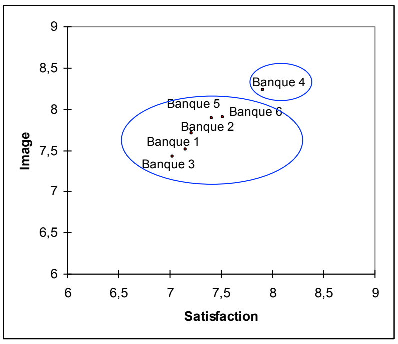
Differences between means with 1000 permutations

** : differences are significant ($P < 0.05$)

Results

Variable	Banque 1	Banque 2	Banque 3	Banque 4	Banque 5	Banque 6	Global
Image	7,504	7,706	7,424	8,232	7,891	7,901	7,786
Qualité perçue	7,798	7,932	7,958	8,468	8,104	8,190	8,084
Satisfaction	7,157	7,217	7,034	7,908	7,409	7,522	7,383
Fidélité	7,417	7,436	7,204	8,549	7,633	7,840	7,687





Concluding remarks

- ✓ The 2 methods yield very similar results.
- ✓ 2 banks have different behaviors :
  - Bank 6:
    - Satisfaction on bank 6 is mainly influenced by perceived quality.
    - Other parameters remain close to other banks
  - Bank 4:
    - The satisfaction score is significantly higher for customer of bank 4.
    - Other parameters remain close to other banks.

Advanced (recent) thoughts on PLS-PM

Component-based approaches to SEM

Quite a **few (statistical) hypotheses** are usually needed

Important **theoretical knowledge** has to be available **for the model specification**:

- **Measurement** model
  (what manifest variables are measuring what concept)
- Direction of links between manifest and latent variables
  (outwards or inwards, i.e. reflective vs. formative)
- Network of "causal" relationships
  ("causality" direction, "predictive path", feedbacks, hidden?)

Confirmatory vs. Exploratory

Comparisons between the two approaches

Comparisons between modeling methods:

- **Legitimacy of the comparison**
- **Differentially parameterized models**

Researchers in applied disciplines often seem to:

- **Overlook** some (if not most) of the **subsequent literature in Statistics**
- Strive to run **comparisons** and derive **properties by simulations…**

Comparisons between the two approaches

"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

Dolce and Lauro (Quality & Quantity, 2014) show that in SEM with **formative blocks**, the **effect of measurement model misspecification** is much larger on the SEM-ML estimates than PLS estimates.

For a **correctly specified** formative block, the **bias** and the **variability** of the estimates of the two approaches are differently affected by **correlation** among MVs and magnitude of the **disturbance's variance**.

Component-based method vs. Factor-based method

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**…"*

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

Component-based method vs. Factor-based method

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:

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

Component-based method vs. Factor-based method

How to assess the quality of the model?

- Covariance-based methods allow for goodness-of-fit tests
- PLS-PM lacks a probabilistic framework and an overall goodness-of-fit measure

However:

- Computational inference for empirical confidence intervals and hypothesis testing (Blindfolding, permutation and resampling techniques)
- On going work at UCLA (W. Huang, PhD Dissertation with P. Bentler) proposing a modified PLS-PM suitable for confirmatory research via χ^2 goodness of fit tests and classical inference

Component-based method vs. Factor-based method

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 (and random) 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.

Component-based method vs. Factor-based method

Prediction-oriented or 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**

PLS path modeling has strengths as a tool for prediction which have not been fully explored and appreciated.

Component-based method vs. Factor-based method

Out-of-sample vs. in-of-sample prediction

"Researchers applying PLS path modeling often assert the "predictive" nature of their research, though researchers often seem to mean nothing more than aiming to maximize R^2 for dependent variables" (Rigdon,2012)

What is good for out-of-sample prediction?

•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**

Component-based method vs. Factor-based method

"...PLS-SEM should retain its predictive character rather than fully subscribing to explanatory PAR QUAD modeling...
...further criteria and evaluation techniques for PLS-SEM need to be considered...
...The current guidelines for model evaluation have limited value in detecting model misspecification..."

Sarstedt et al. (2014)
On the Emancipation of PLS-SEM: A Commentary on Rigdon (2012)
Long Range Planning

Part I Part II Part III Part IV Part IV

Component-based method vs. Factor-based method

Oriented relations or symmetrical modeling ?

- Covariance-based approaches clearly consider the direction of the path in the inner model
- **PLS-PM** suffers from some deficiencies in terms of coherence with the direction of the links in the inner model. It **tends to amplify interdependence** and it misses to distinguish between the role of dependent and explanatory blocks in the inner model.

Dolce et al. (2014) proposed a more suitable non-symmetrical approach, (**NSCPM**), that aims at maximizing the explained variance of the dependent manifest variables.

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris

Component-based method vs. Factor-based method

Suitable for Big Data ?

- **Covariance-based** approach is **not easily** scalable to Big Data analysis
- **PLS-PM** is appropriate in the new research problems considering Big Data. It **can easily face 3 out of the Big Data's Fourth "V".**
  - **Volume**: classical PLS-PM is an easy and fast algorithm that is suitable for large N and/or large P problems (multicollinearity? → **PLS-R in PLS-PM**)
  - **Variety**: **Non-metric PLS-PM** (Russolillo, 2012) for mixed nature variables and non-linear relations.
  - **Veracity**: PLS-PM is more robust compared to SEM for modeling misspecification (Cassel et al., 1999).

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

Main References 4/5

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.
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. (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.
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.
Hwang, H. and Takane, Y. (2004). Generalized structured component analysis. Psychometrika, 69, 81-99.

Main References 5/5

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

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

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

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

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, 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.

ERCIM 2014 - Pisa, 5 December 2014 Vincenzo Esposito Vinzi – ESSEC Paris