COMPONENT-BASED PATH MODELING

Vincenzo ESPOSITO VINZI
ESSEC Business School, Paris
vinzi@essec.edu

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:


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.

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 Upsala Symposium on Psychological 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 August 1926 – 8 February 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 technique
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.
**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**

**SEM**

**Variance-based**  
**Composite-based**

**Component-based**

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)

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)
- Generalized Least Squares (GLS)
- Asymptotically Distribution Free (ADF)
- Unweighted Least Squares (ULS)

Component-based Methods

- PLS Path Modeling (PLS-PM or PLS-SEM)
- Generalized Maximum Entropy (GME)
- Generalized Structured Component Analysis (GSCA) method
- Regularized Generalized CanCorr Analysis (RGCCA)

LISREL-type methods

- LISREL model
  - Joreskog (1977)
  - Wold (1975), Tenenhaus et al. (2005)
  - Hwang & Takane (2004)
  - Tenenhaus & Tenenhaus (2011)

Regularized Generalized CanCorr Analysis (RGCCA)
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

Latent Variables (LV)

Manifest Variables (VM)

Unidirectional Path (cause-effect)

Bidirectional Path (correlation)

Feedback relation or reciprocal causation

Errors
Structural Equation models: notation

- **P manifest variables** (MVs) observed on \( n \) units
  - \( x_{pq} \) generic MV
- **Q latent variables** (LVs)
  - \( \xi_q \) generic LV
- **Q blocks** composed by each LV and the corresponding MVs
  - in each \( q \)-th block \( p_q \) manifest variables \( x_{pq} \), 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
"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

The structural model describes the relations among the latent variables:

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

where:
- \( \beta_{jq^*} \) 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 \( \xi_{q^*} \)
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:

\[ X_{pq} = \lambda_{pq} \xi_q + \varepsilon_{pq} \]

where:
- \( \lambda_{pq} \) 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:

\[ \xi_q = X_q w_q \]
PLS-PM Algorithm

PLS path modeling
Michel Tenenhaus*, Vincenzo Esposito Vinzi†++, Yves-Marie Chatelin*, Carlo Lauro8

*INSEAD, School of Management (ORSTOM), Fontainebleau, France
†Department of Mathematics and Statistics, University of Reading, UK
++Chair—Computational Statistics and Data Analysis
8DEA, Università degli Studi di Napoli "Federico II", Via Cintia—Complesso di Monte S. Angelo, 80126 Naples, Italy

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 offers some advantages. In this respect, some new improvements are proposed. Furthermore, PLS path modeling can be used for estimating multiple tables as to be related to some 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
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 $w_q$ 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:
   $$\hat{\xi}_q \propto X_q w_q$$

3) **Estimation of the path coefficients**
   Path coefficients are estimated as regression coefficients according to the structural model.

4) **Estimation of the loadings**
   Loadings are estimated as regression coefficients according to the measurement model.

---

PLS Path Model: the algorithm

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

$$\hat{\xi}_q \propto X_q w_q$$

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

\[ t_q \propto \sum_q w_{pq} x_{pq} = X_q w_q \]

- **Mode A** (for outwards directed links – reflective – principal factor model):
  \[ w_{pq} = (1/n) x_{pq}^T z_q \]
  - 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):
  \[ w_q = (X_q^T X_q)^{-1} X_q^T z_q \]
  - These indicators **should** **covary**
  - One multiple OLS regression (multicollinearity?)
  - Structural Predictions (higher \( R^2 \) values for endogenous LVs)
  - Multidimensionality (even partial, by sub-blocks)
  - Might incur in **unstable results** with ill-defined blocks

#### Latent or Emergent Constructs?

- **Latent Construct**
  - 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)

- **Emergent Construct**
  - 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.
A focus on Inner Estimation

**Inner Estimation**
Latent Variable proxies = weighted aggregates of connected LVs

\[ z_q \propto \sum_{q'} e_{qq'} t_{q'} \]

1. **Centroid scheme**: \( z_3 = e_{31} t_1 + e_{32} t_2 + e_{34} t_4 \) where \( e_{qq'} = \text{sign}(\text{cor}(t_q, t_{q'})) \)

2. **Factorial scheme**: \( z_3 = \text{cor}(t_3, t_1) t_1 + \text{cor}(t_3, t_2) t_2 + \text{cor}(t_3, t_4) t_4 \)

3. **Path weighting scheme**: \( z_3 = \hat{\gamma}_{31} t_1 + \hat{\gamma}_{32} t_2 + \text{cor}(t_3, t_4) t_4 \)

   Where the \( \gamma \)'s are the regression coefficients of the model: \( t_3 = \gamma_{31} t_1 + \gamma_{32} t_2 + \delta \)

The PLS Path Modeling algorithm

MV's are centered or standardized

- **Initial step**: \( t_q = \pm X_q w_q \)
- **Outer estimation**: Reiterate till Numerical Convergence
- **Inner Estimation**: \( t_q \) are also standardized

**Choice of weights**
- **Centroid**: correlation signs
- **Factorial**: correlations
- **Path weighting scheme**: OLS regression coefficients or simple linear correlations

**Update weights**

**Mode A**: \( w_q = (1/n)X_q z_q \)

**Mode B**: \( w_q = (X_q'X_q)^{-1}X_q'z_q \)
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:

\[
\text{arg max}_{w_q} \left\{ \sum_{q \neq q'} c_{qq'} g\left( \text{cor}\left( X_q w_q, X_{q'} w_{q'} \right) \right) \right\}
\]

s.t. \[X_{q}, w_{q}\] = 1

where:

\[c_{qq'} = \begin{cases} 
1 & \text{if } X_q \text{ and } X_{q'} \text{ is connected} \\
0 & \text{otherwise}
\end{cases}\]

\[g = \begin{cases} 
square & \text{(Factorial scheme)} \\
| & \text{(Centroid scheme)}
\end{cases}\]

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

\[
\arg \max_{w_q \neq 1} \left\{ \sum_{q \neq q'} c_{qq'} g \left( \text{cov} \left( X_q w_q, X_q' w_{q'} \right) \right) \right\}
\]

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

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

\[
\arg \max_{w_q} \left\{ \sum_{q \neq q'} c_{qq'} g \left( \text{cov} \left( X_q w_q, X_q' w_{q'} \right) \right) \right\} =
\]

\[
\arg \max_{w_q} \left\{ \sum_{q \neq q'} c_{qq'} g \left[ \text{cor} \left( X_q w_q, X_q' w_{q'} \right) \sqrt{\text{var} \left( X_q w_q \right) \text{var} \left( X_q' w_{q'} \right) } \right] \right\}
\]

s.t. \[||X_q w_q|| = 1 \text{ if Mode B for block } q\]
\[||w_q|| = 1 \text{ if New Mode A for block } q\]

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*

<table>
<thead>
<tr>
<th>SPSS results</th>
<th>XL-STAT graphical results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(principal components)</td>
<td></td>
</tr>
<tr>
<td>Component Matrix</td>
<td></td>
</tr>
<tr>
<td>VVLT1</td>
<td>.648</td>
</tr>
<tr>
<td>VVLT2</td>
<td>.729</td>
</tr>
<tr>
<td>VVLT3</td>
<td>.823</td>
</tr>
<tr>
<td>VVLT4</td>
<td>.830</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis
1 component extracted.

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

Two block case

Tucker Inter-batteries Analysis
(1st component)
\[
\arg \max \left\{ \text{cov}(X_1w_1, X_2w_2) \right\}
\]

 Canonical Correlation Analysis
(1st component)
\[
\arg \max \left\{ \text{cov}(X_1w_1, X_2w_2) \right\}
\]

 Redundancy Analysis
(1st component)
\[
\arg \max \left\{ \text{cov}(X_1w_1, X_2w_2) \right\}
\]
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

Mode B + Factorial
- Generalised Canonical Correlation Analysis (Carroll)
- Generalised CCA (Horst’s SUMCOR criterion)
- Multiple Factorial Analysis by Lohmöller’s Split PCA
- Mathes Algorithm

Mode B + Centroid
- Generalised CCA (Horst’s SUMCOR criterion)

Wine Tasting: Loire Wine Multiblock Dataset
(Pagès, Asselin, Morlat & Robichet 1987)

3 Appellations | 4 Soils | 21 observed wines

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity of attack</td>
<td>2.96, 2.04, 2.22</td>
</tr>
<tr>
<td>Astringency</td>
<td>2.43, 2.18</td>
</tr>
<tr>
<td>Alc.</td>
<td>2.50, 2.03, 2.64</td>
</tr>
<tr>
<td>Balance (Aid., Atr., Alco.)</td>
<td>3.25, 2.93, 3.32</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>2.73, 2.90, 2.68</td>
</tr>
<tr>
<td>Bitterness</td>
<td>1.93, 1.93, 2.00</td>
</tr>
<tr>
<td>Sculpting intensity in mouth</td>
<td>2.86, 2.89, 3.07</td>
</tr>
<tr>
<td>Aromatic intensity in mouth</td>
<td>3.14, 3.76, 3.14</td>
</tr>
</tbody>
</table>

Global quality
Illustrative variable
Parameters

X1 = Smell at rest (5), X2 = View (3), X3 = Smell after shaking (10), X4 = Tasting (9)
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

From Tenenhaus and Hanafi (2010)

ERCIM 2014 - Pisa, 5 December 2014
Vincenzo Esposito Vinzi – ESSEC Paris
Wine Tasting experience in 4 steps
Block Structure based on Prior Knowledge

Measurement Model Specification

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

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

Mode B (correlation-oriented) for both blocks

Double-headed arrow for interdependence
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)

---

Economic inequality and political instability

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

47 countries

<table>
<thead>
<tr>
<th>Gini</th>
<th>Farm</th>
<th>Rent</th>
<th>Gnpr</th>
<th>Labo</th>
<th>Inst</th>
<th>Ecks</th>
<th>Deat</th>
<th>Demo</th>
</tr>
</thead>
<tbody>
<tr>
<td>86.1</td>
<td>98.2</td>
<td>32.9</td>
<td>374</td>
<td>25</td>
<td>13.6</td>
<td>57</td>
<td>217</td>
<td>1</td>
</tr>
<tr>
<td>97.9</td>
<td>99.6</td>
<td>29.5</td>
<td>1215</td>
<td>14</td>
<td>11.3</td>
<td>0</td>
<td>1215</td>
<td>0</td>
</tr>
<tr>
<td>74.0</td>
<td>97.4</td>
<td>10.7</td>
<td>532</td>
<td>32</td>
<td>12.8</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>58.3</td>
<td>86.1</td>
<td>26.0</td>
<td>1046</td>
<td>26</td>
<td>16.3</td>
<td>46</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>43.7</td>
<td>79.8</td>
<td>0.0</td>
<td>297</td>
<td>67</td>
<td>0.0</td>
<td>9</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

The Demo variable is categorical (nominal) with:

1 = Stable democracy

2 = Unstable democracy

3 = Dictatorship
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 \( w \)
\[ w_{1q} = \text{cor}(x_{1q}, z_1) \]
\[ w_{2q} = \text{cor}(x_{2q}, z_2) \]
\[ w_{3q} = \text{cor}(x_{3q}, z_3) \]

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

Economic inequality and political instability

Results: Problems in the signs...

Weights (Dimension 1):

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Manifest variables</th>
<th>Outer weight</th>
<th>Outer weight (Standard error)</th>
<th>Standard error</th>
<th>Critical ratio (CR)</th>
<th>Lower bound (95%)</th>
<th>Upper bound (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Inequality</td>
<td>gini</td>
<td>0.515</td>
<td>0.044</td>
<td>12.489</td>
<td>0.325</td>
<td>0.845</td>
<td></td>
</tr>
<tr>
<td></td>
<td>farm</td>
<td>0.510</td>
<td>0.049</td>
<td>12.451</td>
<td>0.395</td>
<td>0.833</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rent</td>
<td>0.090</td>
<td>0.102</td>
<td>-0.275</td>
<td>0.137</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial Development</td>
<td>gini</td>
<td>-0.511</td>
<td>0.040</td>
<td>-12.489</td>
<td>0.325</td>
<td>0.845</td>
<td></td>
</tr>
<tr>
<td></td>
<td>farm</td>
<td>0.526</td>
<td>0.022</td>
<td>24.131</td>
<td>0.495</td>
<td>0.805</td>
<td></td>
</tr>
</tbody>
</table>

Correlations (Dimension 1):

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Manifest variables</th>
<th>Standardized loadings</th>
<th>Communalities</th>
<th>Redundancies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Inequality</td>
<td>gini</td>
<td>0.977</td>
<td>0.955</td>
<td></td>
</tr>
<tr>
<td></td>
<td>farm</td>
<td>0.886</td>
<td>0.972</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rent</td>
<td>0.516</td>
<td>0.266</td>
<td></td>
</tr>
<tr>
<td>Industrial Development</td>
<td>gini</td>
<td>0.950</td>
<td>0.903</td>
<td></td>
</tr>
<tr>
<td></td>
<td>farm</td>
<td>0.995</td>
<td>0.912</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rent</td>
<td>0.916</td>
<td>0.926</td>
<td></td>
</tr>
</tbody>
</table>

Loading = Regression coefficient of each MV on the corresponding LV
Correlation coefficient if MV’s are standardized
Economic inequality and political instability

Results: after sign inversion….

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Manifest variables</th>
<th>Outer weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Inequality</td>
<td>gini</td>
<td>0.460</td>
</tr>
<tr>
<td></td>
<td>farm</td>
<td>0.516</td>
</tr>
<tr>
<td></td>
<td>rent</td>
<td>0.081</td>
</tr>
<tr>
<td>Industrial Development</td>
<td>gnpr</td>
<td>0.511</td>
</tr>
<tr>
<td></td>
<td>labo</td>
<td>-0.538</td>
</tr>
<tr>
<td>Political Instability</td>
<td>inst</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>ecks</td>
<td>0.270</td>
</tr>
<tr>
<td></td>
<td>death</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>demostab</td>
<td>-0.336</td>
</tr>
<tr>
<td></td>
<td>demoinst</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>dictatur</td>
<td>0.285</td>
</tr>
</tbody>
</table>

Economic inequality and political instability
Economic inequality and political instability

PLS Results: Structural Equations (Inner Model)

Inner model (Dimension 1):

R² (Political Instability / 1):

<table>
<thead>
<tr>
<th></th>
<th>R²</th>
<th>R²(Bootstrap)</th>
<th>Std. deviation</th>
<th>Lower bound (95%)</th>
<th>Upper bound (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.622</td>
<td>0.632</td>
<td>0.063</td>
<td>0.532</td>
<td>0.762</td>
</tr>
</tbody>
</table>

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 |

<table>
<thead>
<tr>
<th>Value(Bootstrap)</th>
<th>Standard error(Bootstrap)</th>
<th>Lower bound (95%)</th>
<th>Upper bound (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.251</td>
<td>0.104</td>
<td>0.039</td>
<td>0.487</td>
</tr>
<tr>
<td>-0.673</td>
<td>-0.077</td>
<td>0.000</td>
<td>-0.466</td>
</tr>
</tbody>
</table>

PLS results: Standardized Latent Variable Scores

Latent variable scores (Dimension 1):

<table>
<thead>
<tr>
<th>Country</th>
<th>Agricultural Inequality</th>
<th>Industrial Development</th>
<th>Political Instability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>0.953</td>
<td>0.236</td>
<td>0.791</td>
</tr>
<tr>
<td>Australia</td>
<td>1.265</td>
<td>1.371</td>
<td>-1.601</td>
</tr>
<tr>
<td>Austria</td>
<td>0.404</td>
<td>0.253</td>
<td>-0.464</td>
</tr>
<tr>
<td>Belgium</td>
<td>-0.848</td>
<td>1.530</td>
<td>-0.881</td>
</tr>
<tr>
<td>Bolivia</td>
<td>1.115</td>
<td>-1.584</td>
<td>1.503</td>
</tr>
<tr>
<td>Brasil</td>
<td>0.789</td>
<td>-0.654</td>
<td>0.268</td>
</tr>
<tr>
<td>Canada</td>
<td>-1.539</td>
<td>1.690</td>
<td>-0.972</td>
</tr>
<tr>
<td>Chile</td>
<td>1.239</td>
<td>-0.324</td>
<td>0.016</td>
</tr>
<tr>
<td>Colombia</td>
<td>0.819</td>
<td>-0.443</td>
<td>0.810</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>0.939</td>
<td>-0.484</td>
<td>0.302</td>
</tr>
<tr>
<td>Cuba</td>
<td>0.734</td>
<td>-0.182</td>
<td>1.664</td>
</tr>
<tr>
<td>Denmark</td>
<td>-1.996</td>
<td>0.821</td>
<td>-1.528</td>
</tr>
<tr>
<td>Dominican Republic</td>
<td>0.720</td>
<td>-0.737</td>
<td>0.542</td>
</tr>
<tr>
<td>Ecuador</td>
<td>0.976</td>
<td>-0.690</td>
<td>0.966</td>
</tr>
<tr>
<td>Egypt</td>
<td>0.464</td>
<td>-1.086</td>
<td>0.865</td>
</tr>
<tr>
<td>Salvador</td>
<td>0.824</td>
<td>-0.718</td>
<td>0.426</td>
</tr>
<tr>
<td>Finland</td>
<td>-1.020</td>
<td>0.304</td>
<td>-0.262</td>
</tr>
<tr>
<td>France</td>
<td>-0.910</td>
<td>0.804</td>
<td>0.492</td>
</tr>
<tr>
<td>Guatemala</td>
<td>1.006</td>
<td>-0.959</td>
<td>1.099</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Economic inequality and political instability

Map of countries as mapped by LV scores

\( Y_1 = \) agricultural inequality, \( Y_2 = \) industrial development
Reliability

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

\[
\text{rel}(x_{pq}) = \frac{\lambda_p^2 \text{var}(\xi_q)}{\text{var}(x_{pq})} = \text{cor}^2(x_{pq}, \xi_q)
\]

\( \text{rel}(x_{pq}) \) can be interpreted as the variance of \( x_{pq} \) that is explained by \( \xi_q \)

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:

\[
\alpha_q = \frac{P_q}{(P_q - 1)P_q + \frac{1}{\lambda_{pq}}\sum_{x_{pq}} \text{cov}(x_{pq}, x_{pq}')}
\]

\[
\rho_q = \frac{\left(\sum_{x_{pq}} \lambda_{pq}^2\right) \times \text{var}(\xi_q)}{\left(\sum_{x_{pq}} \lambda_{pq}^2\right) \times \sum_{x_{pq}} \text{var}(\xi_q) + \sum_{x_{pq}} \text{var}(\varepsilon_{pq})}
\]

Where:
- \(x_{pq}\) is the p-th manifest variable in the block q,
- \(P_q\) is the number of manifest variables in the block,
- \(\lambda_{pq}\) is the component loading for \(x_{pq}\)
- 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)

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:

```
 x1 --> x
 x2 --> x
 .  
 .  
 xk --> x

\{ x1, x2, ..., xk \} --> x
```

- Split the multidimensional block into unidimensional sub-blocks
Model Assessment

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) \( \text{cor}(\xi_{q}; \xi_{q}') = 0 \), 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) \( \text{cor}(\xi_{q}; \xi_{q}') = 1 \), 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 \( \text{cor}^2(\xi_{q}; \xi_{q}') \), 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).

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.

\[
AVE_q = \frac{\sum_{n} \left( \lambda_{pq}^2 \text{var}(\xi_{q}) \right)}{\sum_{n} \left( \lambda_{pq}^2 \text{var}(\xi_{q}) \right) + \sum_{j} (1 - \lambda_{ij}^2)}
\]

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

$$\text{cor}^2 \left( x_{pq}, \xi_q \right) \gg \text{cor}^2 \left( x_{pq}, \xi'_q \right)$$

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.

$$H_0 : \text{cor} \left( \xi_q, \xi'_q \right) = 1$$

$$H_0 : \text{cor} \left( \xi_q, \xi'_q \right) = 0$$

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 validity only: \((\text{AVE}_q, \text{AVE}_{q'}) > \text{cor}^2 \left( \xi_q, \xi'_q \right)\) 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:

$$Com_{pq} = \text{cor}^2(x_{pq}, \xi_q)$$

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

$$Com_q = \frac{1}{P_q} \sum_{j=1}^{P_q} \text{cor}^2(x_{pq}, \xi_j)$$

(NB: if standardised MVs: $Com_q = \text{AVE}_q$)

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

$$\overline{Com} = \frac{\sum_{e \in E} (P_e \times Com_e)}{\sum_{e \in E} P_e}$$
Redundancy

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

\[ \text{Redundancy}_{q^*} = R^2 \left( \xi_{q^*} - \xi_{q^*} \rightarrow \xi_{q^*} \right) \times \text{Communality}_{q^*} \]

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)

\[ H_q = 1 - \frac{\sum_{q} \sum_{i} (x_{ij} - \bar{x}_{iq} - \hat{\lambda}_{pq} \xi_{iq} - \hat{\xi}_{q})^2}{\sum_{q} \sum_{i} (x_{ij} - \bar{x}_{iq})^2} \]

\[ F_q = 1 - \frac{\sum_{q} \sum_{i} (x_{ij} - \bar{x}_{iq} - \hat{\lambda}_{pq} \xi_{iq} - \hat{\xi}_{q})^2}{\sum_{q} \sum_{i} (x_{ij} - \bar{x}_{iq})^2} \]

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

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

Godness of Fit index

$$GoF = \frac{1}{\sqrt{\sum \sum \sum Cor^2(x_{pq}, \xi_q)}} \times \sqrt{\sum \sum R^2(\xi_{pq}, \xi_j, \text{explaining } \xi_{pq}^*)}$$

Validation of the outer model

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

Validation of the inner model

The validation of the inner model is obtained as average of the R² values of all the structural relationships.
Relative GoF

From **PCA** is well known that

\[
\sum_{j=1}^{p} \text{corr}^2(x_{pq}, \xi_j) = \lambda_{j}^{\text{PCA}}
\]

the largest eigenvalue of the \( X'_pX_p \)

From **CCA** we know that

\[
\sum_{q=1}^{Q} R^2(\xi_q, \xi_p, \text{explaining} \ \xi_p) = \rho_{pq}^2
\]

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

\[
\text{GoF} = \sqrt{\frac{1}{\sum_{q=1}^{Q} P_q} \sum_{q=1}^{Q} \text{corr}^2(x_{pq}, \xi_q) \times \frac{1}{G_p} \sum_{j=1}^{G_p} R^2(\xi_j, \xi_p, \text{explaining} \ \xi_p)}
\]

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

Let us consider this simple model:

\[
\begin{align*}
\xi_1 &= \beta_{13} \xi_3 \\
\xi_2 &= \beta_{12} \xi_1 + \xi_1 \\
\xi_3 &= \beta_{13} \xi_1 + \beta_{23} \xi_2 + \xi_2
\end{align*}
\]

Is this coefficient significant? 

\( H_0 : \beta_{23} = 0 \)

\( H_1 : \beta_{23} \neq 0 \)

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

**Step 2.** «Deflate» the MV block \( X_3 \) :

\[
\tilde{X}_3 = X_3 - X_2 \left( X_2'X_2 \right)^{-1} X_2' X_3
\]

**Step 3.** Compute GoF index for the «null» model, \( GoF_{H_0} \)

**Step 4.** Draw \( B \) samples from \( \hat{F}_{[X_1,X_2,A_1]} \), compute \( GoF_{H_0}^{(b)} \) and, for each \( b \), compute the ratio \( \frac{GoF_{H_0}^{(b)}}{GoF_{H_0}^{(b)}} \).

**Step 5.** By referring to the inverse cdf of \( \frac{GoF_{H_0}^{(b)}}{GoF_{H_0}^{(b)}} \), accept or reject \( H_0 \)
Let us consider the complete model:

Is this standardized coefficient significant?

\[ H_0: \beta_{\text{lm1,loy}} = 0 \]
\[ H_1: \beta_{\text{lm1,loy}} \neq 0 \]

\[ t = \frac{\text{GoF}_{\text{lm1,loy}} - \text{GoF}_{\text{fr} \beta_{\text{lm1,loy}} = 0}}{\text{GoF}_{\text{fr} \beta_{\text{lm1,loy}} = 0}} = 1.0037 \]

Thresholds, by bootstrap, for different confidence levels

\[ \Phi^B_i(0.1) = 1.00238 \]
\[ \Phi^B_i(0.05) = 1.00316 \]
\[ \Phi^B_i(0.025) = 1.00411 \]

We reject \( H_0 \)

\[ \text{p-value} = 0.035 \]

**Non parametric Bootstrap-based validation of all path coefficients**

\[ H_0 : \beta_{12} = \beta_{13} = \beta_{23} = 0 \]
\[ H_1 : \text{at least one of } \beta_{ij} (i = 1,2; j = 1,2,3; j > i) \neq 0 \]

**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):

\[ X_2^* = X_2 - X_1 (X_1' X_1)^{-1} X_1' X_2 \]
\[ X_3^* = X_3 - (X_1 | \hat{X}_2) \left( (X_1 | \hat{X}_2) (X_1 | \hat{X}_2) \right)^{-1} (X_1 | \hat{X}_2) X_3 \]

**Step 3.** Compute GoF index for the «null» model,

\[ \text{GoF}_{\beta_0} \]

**Step 4.** Draw \( B \) samples from \( \hat{F}_{X_1, \hat{X}_2, \hat{X}_3} \) and compute \( \text{GoF}_{\beta_0}^{(b)} \) for each sample \( b \)

**Step 5.** By referring to the inverse cdf of \( \text{GoF}_{\beta_0}^{(b)} \) accept or reject \( H_0 \)
Removing a link from the model.....

We can use the $t$-test to assess 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 R\(^2\) 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

\[
f^2 = \frac{R_{\text{included}}^2 - R_{\text{excluded}}^2}{1 - R_{\text{included}}^2}
\]

\(f^2 \geq 0.02 \rightarrow \text{small impact}\)
\(f^2 \geq 0.15 \rightarrow \text{medium impact}\)
\(f^2 \geq 0.35 \rightarrow \text{large impact}\)

Suppression effects and Redundancy

- When the path coefficient and the correlation between latent constructs do not have the same sign, \textit{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: \textit{non significant coefficients}
  - there are redundant variables that artificially change the signs; one or more redundant variables must be eliminated (multicollinearity)
  - \textit{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

- \textbf{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?

---

**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 indices

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Dimensions</th>
<th>Cronbach’s alpha</th>
<th>D.G. rho (PCA)</th>
<th>Condition number</th>
<th>Critical value</th>
<th>Eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>5</td>
<td>0.714</td>
<td>0.815</td>
<td>2.402</td>
<td>391.494</td>
<td>804.215</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>419.144</td>
<td>503.403</td>
</tr>
<tr>
<td>Expectation</td>
<td>5</td>
<td>0.433</td>
<td>0.717</td>
<td>1.631</td>
<td>422.484</td>
<td>616.441</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>419.144</td>
<td>503.403</td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>7</td>
<td>0.872</td>
<td>0.903</td>
<td>3.642</td>
<td>348.439</td>
<td>1400.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>320.960</td>
<td>1212.960</td>
</tr>
<tr>
<td>Perceived Value</td>
<td>2</td>
<td>0.817</td>
<td>0.921</td>
<td>2.430</td>
<td>503.782</td>
<td>861.605</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>682.995</td>
<td>1190.249</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>3</td>
<td>0.770</td>
<td>0.879</td>
<td>2.696</td>
<td>316.477</td>
<td>509.884</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>320.960</td>
<td>1212.960</td>
</tr>
<tr>
<td>Complaints</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loyalty</td>
<td>3</td>
<td>0.442</td>
<td>0.723</td>
<td>1.923</td>
<td>825.084</td>
<td>1190.249</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>983.354</td>
<td>1099.249</td>
</tr>
</tbody>
</table>
### XLSTAT Output: Block Homogeneity

Checking for block homogeneity by means of PCA

**Variables/Factors correlations (Expectation / 1):**

<table>
<thead>
<tr>
<th>Variables</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUEX1</td>
<td>0.571</td>
<td>0.432</td>
<td>0.701</td>
</tr>
<tr>
<td>CUEX2</td>
<td>0.506</td>
<td>0.759</td>
<td>-0.414</td>
</tr>
<tr>
<td>CUEX3</td>
<td>0.868</td>
<td>-0.491</td>
<td>-0.099</td>
</tr>
</tbody>
</table>

**Variables/Factors correlations (Loyalty / 1):**

<table>
<thead>
<tr>
<th>Variables</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUSL1</td>
<td>0.864</td>
<td>-0.365</td>
<td>-0.352</td>
</tr>
<tr>
<td>CUSL2</td>
<td>0.424</td>
<td>0.907</td>
<td>-0.033</td>
</tr>
<tr>
<td>CUSL3</td>
<td>0.774</td>
<td>-0.228</td>
<td>0.593</td>
</tr>
</tbody>
</table>

### XLSTAT Result: Cross-loadings

Detection of monofactorial MVs

<table>
<thead>
<tr>
<th>Cross-loadings (Monofactorial manifest variables / 1):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
</tr>
<tr>
<td>IMAG1</td>
</tr>
<tr>
<td>IMAG2</td>
</tr>
<tr>
<td>IMAG3</td>
</tr>
<tr>
<td>IMAG4</td>
</tr>
<tr>
<td>IMAG5</td>
</tr>
<tr>
<td>CUEX1</td>
</tr>
<tr>
<td>CUEX2</td>
</tr>
<tr>
<td>CUEX3</td>
</tr>
<tr>
<td>PERQ1</td>
</tr>
<tr>
<td>PERQ2</td>
</tr>
<tr>
<td>PERQ3</td>
</tr>
<tr>
<td>PERQ4</td>
</tr>
<tr>
<td>PERQ5</td>
</tr>
<tr>
<td>PERQ6</td>
</tr>
<tr>
<td>PERQ7</td>
</tr>
<tr>
<td>PERV1</td>
</tr>
<tr>
<td>PERV2</td>
</tr>
<tr>
<td>CUSA1</td>
</tr>
<tr>
<td>CUSA2</td>
</tr>
<tr>
<td>CUSA3</td>
</tr>
<tr>
<td>CUSL1</td>
</tr>
<tr>
<td>CUSL2</td>
</tr>
<tr>
<td>CUSL3</td>
</tr>
<tr>
<td>CUSCO</td>
</tr>
</tbody>
</table>
XLSTAT Output: Bivariate and Partial correlation between LVs

Checking for nomological and discriminant validity

Correlations (Latent variable) / Dimension (1):

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Expectation</th>
<th>Perceived Quality</th>
<th>Perceived Value</th>
<th>Satisfaction</th>
<th>Complaints</th>
<th>Loyalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>1.000</td>
<td>0.493</td>
<td>0.731</td>
<td>0.508</td>
<td>0.671</td>
<td>0.469</td>
<td>0.548</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.493</td>
<td>1.000</td>
<td>0.545</td>
<td>0.360</td>
<td>0.481</td>
<td>0.250</td>
<td>0.366</td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>0.731</td>
<td>0.545</td>
<td>1.000</td>
<td>0.576</td>
<td>0.791</td>
<td>0.537</td>
<td>0.524</td>
</tr>
<tr>
<td>Perceived Value</td>
<td>0.508</td>
<td>0.360</td>
<td>0.576</td>
<td>1.000</td>
<td>0.604</td>
<td>0.348</td>
<td>0.517</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>0.671</td>
<td>0.481</td>
<td>0.791</td>
<td>0.604</td>
<td>1.000</td>
<td>0.540</td>
<td>0.035</td>
</tr>
<tr>
<td>Complaints</td>
<td>0.469</td>
<td>0.250</td>
<td>0.537</td>
<td>0.348</td>
<td>0.540</td>
<td>1.000</td>
<td>0.401</td>
</tr>
<tr>
<td>Loyalty</td>
<td>0.548</td>
<td>0.366</td>
<td>0.524</td>
<td>0.517</td>
<td>0.635</td>
<td>0.401</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Expectation</th>
<th>Perceived Quality</th>
<th>Perceived Value</th>
<th>Satisfaction</th>
<th>Complaints</th>
<th>Loyalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>1.000</td>
<td>0.141</td>
<td>0.363</td>
<td>0.053</td>
<td>0.081</td>
<td>0.097</td>
<td>0.185</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.141</td>
<td>1.000</td>
<td>0.232</td>
<td>0.014</td>
<td>0.052</td>
<td>-0.102</td>
<td>0.052</td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>0.363</td>
<td>0.232</td>
<td>1.000</td>
<td>0.149</td>
<td>0.458</td>
<td>0.190</td>
<td>-0.103</td>
</tr>
<tr>
<td>Perceived Value</td>
<td>0.053</td>
<td>0.014</td>
<td>0.149</td>
<td>1.000</td>
<td>0.188</td>
<td>-0.031</td>
<td>0.107</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>0.081</td>
<td>0.052</td>
<td>0.458</td>
<td>0.188</td>
<td>1.000</td>
<td>0.170</td>
<td>0.314</td>
</tr>
<tr>
<td>Complaints</td>
<td>0.087</td>
<td>-0.102</td>
<td>0.150</td>
<td>-0.031</td>
<td>0.170</td>
<td>1.000</td>
<td>0.074</td>
</tr>
<tr>
<td>Loyalty</td>
<td>0.185</td>
<td>0.052</td>
<td>-0.103</td>
<td>0.197</td>
<td>0.314</td>
<td>0.074</td>
<td>1.000</td>
</tr>
</tbody>
</table>

XLSTAT Output: AVE

Checking for discriminant validity

→ (LV squared correlation < AVE)

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

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Expectation</th>
<th>Perceived Quality</th>
<th>Perceived Value</th>
<th>Satisfaction</th>
<th>Complaints</th>
<th>Loyalty</th>
<th>Mean Communalities (AVE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>0.240</td>
<td>0.240</td>
<td>0.266</td>
<td>0.236</td>
<td>0.266</td>
<td>0.236</td>
<td>0.266</td>
<td>0.240</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.240</td>
<td>1</td>
<td>0.297</td>
<td>0.134</td>
<td>0.231</td>
<td>0.063</td>
<td>0.134</td>
<td>0.240</td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>0.236</td>
<td>0.297</td>
<td>1</td>
<td>0.332</td>
<td>0.628</td>
<td>0.288</td>
<td>0.275</td>
<td>0.236</td>
</tr>
<tr>
<td>Perceived Value</td>
<td>0.266</td>
<td>0.134</td>
<td>0.332</td>
<td>1</td>
<td>0.365</td>
<td>0.121</td>
<td>0.247</td>
<td>0.266</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>0.236</td>
<td>0.628</td>
<td>0.288</td>
<td>0.365</td>
<td>1</td>
<td>0.292</td>
<td>0.403</td>
<td>0.236</td>
</tr>
<tr>
<td>Complaints</td>
<td>0.266</td>
<td>0.288</td>
<td>0.121</td>
<td>0.121</td>
<td>0.292</td>
<td>1</td>
<td>0.161</td>
<td>0.266</td>
</tr>
<tr>
<td>Loyalty</td>
<td>0.236</td>
<td>0.365</td>
<td>0.161</td>
<td>0.365</td>
<td>0.292</td>
<td>0.161</td>
<td>1</td>
<td>0.236</td>
</tr>
<tr>
<td>Mean Communalities (AVE)</td>
<td>0.240</td>
<td>0.240</td>
<td>0.266</td>
<td>0.236</td>
<td>0.266</td>
<td>0.236</td>
<td>0.266</td>
<td>0.240</td>
</tr>
</tbody>
</table>
# 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:)

### Latent variable

<table>
<thead>
<tr>
<th>Manifest variables</th>
<th>Outer weight (normalized)</th>
<th>Outer weight (Bootstrap)</th>
<th>Std. deviation</th>
<th>Lower bound (95%)</th>
<th>Upper bound (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUEX1</td>
<td>0.023</td>
<td>0.326</td>
<td>0.003</td>
<td>0.013</td>
<td>0.033</td>
</tr>
<tr>
<td>CUEX2</td>
<td>0.022</td>
<td>0.317</td>
<td>0.003</td>
<td>0.015</td>
<td>0.031</td>
</tr>
<tr>
<td>PERQ1</td>
<td>0.010</td>
<td>0.138</td>
<td>0.000</td>
<td>0.006</td>
<td>0.015</td>
</tr>
<tr>
<td>PERQ2</td>
<td>0.009</td>
<td>0.121</td>
<td>0.000</td>
<td>0.006</td>
<td>0.013</td>
</tr>
<tr>
<td>PERQ3</td>
<td>0.012</td>
<td>0.167</td>
<td>0.001</td>
<td>0.006</td>
<td>0.014</td>
</tr>
<tr>
<td>PERQ4</td>
<td>0.008</td>
<td>0.119</td>
<td>0.001</td>
<td>0.006</td>
<td>0.011</td>
</tr>
<tr>
<td>PERQ5</td>
<td>0.009</td>
<td>0.135</td>
<td>0.001</td>
<td>0.006</td>
<td>0.012</td>
</tr>
<tr>
<td>PERQ6</td>
<td>0.013</td>
<td>0.183</td>
<td>0.001</td>
<td>0.006</td>
<td>0.015</td>
</tr>
<tr>
<td>PERQ7</td>
<td>0.013</td>
<td>0.183</td>
<td>0.001</td>
<td>0.006</td>
<td>0.015</td>
</tr>
<tr>
<td>CUSL1</td>
<td>0.016</td>
<td>0.242</td>
<td>0.002</td>
<td>0.012</td>
<td>0.022</td>
</tr>
<tr>
<td>CUSL2</td>
<td>0.006</td>
<td>0.129</td>
<td>0.001</td>
<td>0.005</td>
<td>0.012</td>
</tr>
<tr>
<td>CUSL3</td>
<td>0.002</td>
<td>0.047</td>
<td>0.000</td>
<td>0.015</td>
<td>0.025</td>
</tr>
<tr>
<td>IMAG1</td>
<td>0.014</td>
<td>0.168</td>
<td>0.000</td>
<td>0.010</td>
<td>0.018</td>
</tr>
<tr>
<td>IMAG2</td>
<td>0.013</td>
<td>0.173</td>
<td>0.000</td>
<td>0.010</td>
<td>0.018</td>
</tr>
<tr>
<td>IMAG3</td>
<td>0.014</td>
<td>0.187</td>
<td>0.001</td>
<td>0.010</td>
<td>0.021</td>
</tr>
<tr>
<td>IMAG4</td>
<td>0.015</td>
<td>0.242</td>
<td>0.001</td>
<td>0.010</td>
<td>0.021</td>
</tr>
<tr>
<td>IMAG5</td>
<td>0.014</td>
<td>0.197</td>
<td>0.001</td>
<td>0.010</td>
<td>0.021</td>
</tr>
<tr>
<td>CUSL4</td>
<td>0.016</td>
<td>0.242</td>
<td>0.001</td>
<td>0.010</td>
<td>0.021</td>
</tr>
<tr>
<td>CUSL5</td>
<td>0.016</td>
<td>0.242</td>
<td>0.001</td>
<td>0.010</td>
<td>0.021</td>
</tr>
</tbody>
</table>

## Latent Variable Scores

### First 15 observations, out of 250

<table>
<thead>
<tr>
<th>Image</th>
<th>Perceived Value</th>
<th>Satisfaction</th>
<th>Complaints</th>
<th>Loyalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs1</td>
<td>46.673</td>
<td>16.762</td>
<td>52.174</td>
<td>66.667</td>
</tr>
<tr>
<td>Obs2</td>
<td>95.875</td>
<td>100.000</td>
<td>91.030</td>
<td>100.000</td>
</tr>
<tr>
<td>Obs3</td>
<td>58.719</td>
<td>66.667</td>
<td>69.380</td>
<td>55.556</td>
</tr>
<tr>
<td>Obs4</td>
<td>76.129</td>
<td>44.444</td>
<td>100.000</td>
<td>44.444</td>
</tr>
<tr>
<td>Obs5</td>
<td>82.013</td>
<td>55.556</td>
<td>83.164</td>
<td>44.444</td>
</tr>
<tr>
<td>Obs6</td>
<td>86.066</td>
<td>100.000</td>
<td>69.380</td>
<td>77.778</td>
</tr>
<tr>
<td>Obs7</td>
<td>56.389</td>
<td>55.746</td>
<td>77.778</td>
<td>66.667</td>
</tr>
<tr>
<td>Obs8</td>
<td>79.361</td>
<td>55.746</td>
<td>62.734</td>
<td>77.778</td>
</tr>
<tr>
<td>Obs9</td>
<td>67.949</td>
<td>55.746</td>
<td>66.667</td>
<td>77.778</td>
</tr>
<tr>
<td>Obs10</td>
<td>69.342</td>
<td>44.635</td>
<td>48.831</td>
<td>55.556</td>
</tr>
<tr>
<td>Obs11</td>
<td>72.460</td>
<td>66.667</td>
<td>70.813</td>
<td>66.667</td>
</tr>
<tr>
<td>Obs12</td>
<td>77.870</td>
<td>55.746</td>
<td>73.239</td>
<td>66.667</td>
</tr>
<tr>
<td>Obs13</td>
<td>77.515</td>
<td>50.095</td>
<td>73.239</td>
<td>77.778</td>
</tr>
<tr>
<td>Obs14</td>
<td>100.000</td>
<td>72.317</td>
<td>83.164</td>
<td>77.778</td>
</tr>
<tr>
<td>Obs15</td>
<td>72.878</td>
<td>28.063</td>
<td>65.427</td>
<td>66.667</td>
</tr>
</tbody>
</table>

### Example: Computation of the CS index:

\[
\text{CSI} = 0.01579 \times C_{\text{sat1}} + 0.02307 \times C_{\text{sat2}} + 0.02630 \times C_{\text{sat3}}
\]

\[
0.01579 + 0.02307 + 0.02630
\]
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):**

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Value</th>
<th>Standard Error</th>
<th>t</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.102</td>
<td>6.021</td>
<td>0.515</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>0.170</td>
<td>0.061</td>
<td>2.776</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expectation</td>
<td>0.040</td>
<td>0.048</td>
<td>0.837</td>
<td>0.403</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>0.586</td>
<td>0.064</td>
<td>9.128</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Value</td>
<td>0.149</td>
<td>0.034</td>
<td>4.401</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Path coefficients (Satisfaction / 1):**

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Value</th>
<th>Standard Error</th>
<th>Lower bound (95%)</th>
<th>Upper bound (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.102</td>
<td>6.021</td>
<td>0.515</td>
<td>0.000</td>
</tr>
<tr>
<td>Image</td>
<td>0.170</td>
<td>0.061</td>
<td>2.776</td>
<td>0.006</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.040</td>
<td>0.048</td>
<td>0.837</td>
<td>0.403</td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>0.586</td>
<td>0.064</td>
<td>9.128</td>
<td>0.000</td>
</tr>
<tr>
<td>Perceived Value</td>
<td>0.149</td>
<td>0.034</td>
<td>4.401</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Correlation * path coefficient:**

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>0.155</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.037</td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>0.544</td>
</tr>
<tr>
<td>Perceived Value</td>
<td>0.200</td>
</tr>
</tbody>
</table>

**Relative importance of Satisfaction predictors**

### XLSTAT Output: Contribution to R²

**Relative importance of Satisfaction predictors**

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

- **Correlation**: 0.791
- **Path coefficient**: 0.544
- **Contribution to R² (%)**: 64.134

**Cumulative %**: 64.134

**Relative importance:** 100.000

---

ERCIM 2014 - Pisa, 5 December 2014
Vincenzo Esposito Vinzi – ESSEC Paris
XLSTAT Output: MVs quality

**MV Loadings, Communalities and Redundancies**

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Manifest variables</th>
<th>Standardized loadings</th>
<th>Communalities</th>
<th>Redundancies</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUEX1</td>
<td></td>
<td>0.687</td>
<td>0.471</td>
<td>0.175</td>
</tr>
<tr>
<td>CUEX2</td>
<td></td>
<td>0.644</td>
<td>0.415</td>
<td>0.101</td>
</tr>
<tr>
<td>CUEX3</td>
<td></td>
<td>0.726</td>
<td>0.527</td>
<td>0.128</td>
</tr>
<tr>
<td>PERQ1</td>
<td></td>
<td>0.778</td>
<td>0.506</td>
<td>0.160</td>
</tr>
<tr>
<td>PERQ2</td>
<td></td>
<td>0.651</td>
<td>0.423</td>
<td>0.126</td>
</tr>
<tr>
<td>PERQ3</td>
<td></td>
<td>0.801</td>
<td>0.641</td>
<td>0.191</td>
</tr>
<tr>
<td>PERQ4</td>
<td></td>
<td>0.760</td>
<td>0.578</td>
<td>0.172</td>
</tr>
<tr>
<td>PERQ5</td>
<td></td>
<td>0.732</td>
<td>0.536</td>
<td>0.159</td>
</tr>
<tr>
<td>PERQ6</td>
<td></td>
<td>0.766</td>
<td>0.587</td>
<td>0.174</td>
</tr>
<tr>
<td>PERQ7</td>
<td></td>
<td>0.803</td>
<td>0.644</td>
<td>0.191</td>
</tr>
<tr>
<td>PERV1</td>
<td></td>
<td>0.933</td>
<td>0.870</td>
<td>0.291</td>
</tr>
<tr>
<td>PERV2</td>
<td></td>
<td>0.911</td>
<td>0.829</td>
<td>0.278</td>
</tr>
<tr>
<td>CUSA1</td>
<td></td>
<td>0.711</td>
<td>0.506</td>
<td>0.339</td>
</tr>
<tr>
<td>CUSA2</td>
<td></td>
<td>0.872</td>
<td>0.760</td>
<td>0.510</td>
</tr>
<tr>
<td>CUSA3</td>
<td></td>
<td>0.885</td>
<td>0.788</td>
<td>0.526</td>
</tr>
<tr>
<td>CUSL1</td>
<td></td>
<td>0.856</td>
<td>0.731</td>
<td>0.316</td>
</tr>
<tr>
<td>CUSL2</td>
<td></td>
<td>0.273</td>
<td>0.075</td>
<td>0.032</td>
</tr>
<tr>
<td>CUSL3</td>
<td></td>
<td>0.869</td>
<td>0.755</td>
<td>0.326</td>
</tr>
<tr>
<td>IMAG1</td>
<td></td>
<td>0.717</td>
<td>0.514</td>
<td></td>
</tr>
<tr>
<td>IMAG2</td>
<td></td>
<td>0.566</td>
<td>0.320</td>
<td></td>
</tr>
<tr>
<td>IMAG3</td>
<td></td>
<td>0.658</td>
<td>0.433</td>
<td></td>
</tr>
<tr>
<td>IMAG4</td>
<td></td>
<td>0.752</td>
<td>0.627</td>
<td></td>
</tr>
<tr>
<td>IMAG5</td>
<td></td>
<td>0.698</td>
<td>0.487</td>
<td></td>
</tr>
<tr>
<td>CUSCO</td>
<td></td>
<td>1.000</td>
<td>0.292</td>
<td></td>
</tr>
</tbody>
</table>

**XLSTAT Output: model assessment**

**Mean Communality and Redundancy**

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>R²</th>
<th>Adjusted R²</th>
<th>Mean Communalities</th>
<th>Mean Redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>0.243</td>
<td>0.243</td>
<td>0.471</td>
<td>0.115</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.297</td>
<td>0.297</td>
<td>0.574</td>
<td>0.170</td>
</tr>
<tr>
<td>Perceived Que</td>
<td>0.335</td>
<td>0.332</td>
<td>0.850</td>
<td>0.285</td>
</tr>
<tr>
<td>Perceived Valk</td>
<td>0.672</td>
<td>0.668</td>
<td>0.683</td>
<td>0.458</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>0.292</td>
<td>0.292</td>
<td>1.000</td>
<td>0.292</td>
</tr>
<tr>
<td>Complaints</td>
<td>0.432</td>
<td>0.427</td>
<td>0.520</td>
<td>0.225</td>
</tr>
<tr>
<td>Loyalty</td>
<td>0.376</td>
<td></td>
<td>0.570</td>
<td>0.257</td>
</tr>
<tr>
<td>Mean</td>
<td>0.376</td>
<td></td>
<td>0.570</td>
<td>0.257</td>
</tr>
</tbody>
</table>

GoF = \((0.3784^{*0.5702})^{1/2} = 0.4645\)
## XLSTAT Output: GoF

### Validation by Bootstrap

Goodness of fit index (1):

<table>
<thead>
<tr>
<th></th>
<th>GoF</th>
<th>GoF (Bootstrap)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute</td>
<td>0.4645</td>
<td>0.4702</td>
</tr>
<tr>
<td>Relative</td>
<td>0.9371</td>
<td>0.9110</td>
</tr>
<tr>
<td>Outer model</td>
<td>0.9957</td>
<td>0.9944</td>
</tr>
<tr>
<td>Inner model</td>
<td>0.9411</td>
<td>0.9162</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Lower bound (95%)</th>
<th>Upper bound (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute</td>
<td>0.4136</td>
<td>0.5262</td>
</tr>
<tr>
<td>Relative</td>
<td>0.8565</td>
<td>0.9448</td>
</tr>
<tr>
<td>Outer model</td>
<td>0.9884</td>
<td>0.9977</td>
</tr>
<tr>
<td>Inner model</td>
<td>0.8616</td>
<td>0.9482</td>
</tr>
</tbody>
</table>

## XLSTAT Output: Direct, Indirect and Total effects

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

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Expectation</th>
<th>Perceived Quality</th>
<th>Perceived Value</th>
<th>Satisfaction</th>
<th>Complaints</th>
<th>Loyalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>Expectation</td>
<td>0.507</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Perceived Quality</td>
<td>0.000</td>
<td>0.549</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Perceived Value</td>
<td>0.000</td>
<td>0.096</td>
<td>0.781</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Satisfaction</td>
<td>0.170</td>
<td>0.040</td>
<td>0.068</td>
<td>0.149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Complaints</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.889</td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Loyalty</td>
<td>0.328</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.645</td>
<td>0.042</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Expectation</th>
<th>Perceived Quality</th>
<th>Perceived Value</th>
<th>Satisfaction</th>
<th>Complaints</th>
<th>Loyalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>Expectation</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Perceived Quality</td>
<td>0.278</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Perceived Value</td>
<td>0.286</td>
<td>0.429</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Satisfaction</td>
<td>0.223</td>
<td>0.400</td>
<td>0.116</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Complaints</td>
<td>0.350</td>
<td>0.391</td>
<td>0.604</td>
<td>0.132</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Loyalty</td>
<td>0.286</td>
<td>0.300</td>
<td>0.479</td>
<td>0.102</td>
<td>0.037</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Expectation</th>
<th>Perceived Quality</th>
<th>Perceived Value</th>
<th>Satisfaction</th>
<th>Complaints</th>
<th>Loyalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>Expectation</td>
<td>0.507</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Perceived Quality</td>
<td>0.278</td>
<td>0.549</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Perceived Value</td>
<td>0.286</td>
<td>0.525</td>
<td>0.781</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Satisfaction</td>
<td>0.303</td>
<td>0.440</td>
<td>0.702</td>
<td>0.149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Complaints</td>
<td>0.350</td>
<td>0.391</td>
<td>0.604</td>
<td>0.132</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Loyalty</td>
<td>0.586</td>
<td>0.300</td>
<td>0.479</td>
<td>0.102</td>
<td>0.083</td>
<td>0.042</td>
</tr>
</tbody>
</table>
Blindfolding validation

<table>
<thead>
<tr>
<th></th>
<th>Communalities</th>
<th>Redundancies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>0.198</td>
<td></td>
</tr>
<tr>
<td>Expectation</td>
<td>-0.016</td>
<td>0.057</td>
</tr>
<tr>
<td>Perceived Quality</td>
<td>0.401</td>
<td>0.090</td>
</tr>
<tr>
<td>Perceived Value</td>
<td>0.452</td>
<td>0.240</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>0.388</td>
<td>0.465</td>
</tr>
<tr>
<td>Complaints</td>
<td></td>
<td>0.181</td>
</tr>
<tr>
<td>Loyalty</td>
<td>0.149</td>
<td>0.164</td>
</tr>
</tbody>
</table>

- 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:

\[
\begin{bmatrix}
  \mathbf{x}_i \\
  \varepsilon_i \\
  \xi_i \\
\end{bmatrix} = \begin{bmatrix}
  \Lambda \\
  \mathbf{B}
\end{bmatrix} \begin{bmatrix}
  \xi_i \\
  \varepsilon_i
\end{bmatrix} + \begin{bmatrix}
  \varepsilon_i \\
  \xi_i
\end{bmatrix}
\]

Since \( \xi_i = W \mathbf{x}_i \)

\[
\begin{bmatrix}
  \mathbf{I} \\
  W
\end{bmatrix} \mathbf{x}_i = \begin{bmatrix}
  0 & \Lambda \\
  0 & \mathbf{B}
\end{bmatrix} \begin{bmatrix}
  \mathbf{I} \\
  W
\end{bmatrix} \mathbf{x}_i + \begin{bmatrix}
  \varepsilon_i \\
  \xi_i
\end{bmatrix}
\]

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:

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

Under the constraint that latent variables are normalized, i.e. \( \sum_{i=1}^{N} \xi_i^2 = 1 \)

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

\( \text{ALS does not assure convergence in a global minimum, several starting values are needed} \)
GSCA optimizing function

Searching for standardized latent variables $\xi_q = X_q w_q$

minimizing:

$$\sum_{\text{Block } X_q \text{ reflective}} \left\| X_q - \xi_q \lambda_q \right\|^2 + \sum_{\xi_j \text{ endogenous, } \xi_j \text{ explaining } \xi_j} \left\| y_j - \sum_{q^*} b_{jq^*} \xi_{q^*} \right\|^2$$

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)

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):

\[ \arg \max_{w_q} \left\{ \sum_{q \neq q'} c_{qq'} g \left( \text{cov} (X_q w_q, X_{q'} w_{q'}) \right) \right\} \]

s.t. \( (1 - \tau_q) \text{var} (X_q w_q) + \tau_q \|w_q\|^2 = 1 \)

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

MV's are centered or standardized

Initial step

-vector times weight equals outer component

Reiterate till numerical convergence

Mode RIDGE

Update weights - \( n_q \) can be \( \leq p_q \), for \( \tau = 0 \)

Choice of weights \( e_{w_q} \):
- Horst: \( e_{w_q} = c_{w_q} \)
- Centroid: \( e_{w_q} = c_{w_q} \text{sign}(\text{Cor}(v_{w_q}, v_q)) \)
- Factorial: \( e_{w_q} = c_{w_q} \text{Cov}(v_{w_q}, v_q) \)

\( c_{w_q} = 1 \) 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

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

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

Research of $M$ (value chosen by cross-validation or defined by the user) orthogonal components $t_{mq} = X_q^a_{mq}$ which are as correlated as possible to $z_q$ (from the inner estimation step) and also explanatory of their own block $X_q$.

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

PLS1 (regression) Mode leads to a compromise between a multiple regression of $z_q$ on $X_q$ (Mode B) and a principal component analysis of $X_q$ (Mode A for a single block).

PLS Regression algorithm in PLS-PM

1. First PLS component $t_{1q}$ (with $x_{pq}$ standardized as well):

$$t_{1q} = X_q^a_{1q} = \frac{1}{\sqrt{\sum_p \text{cor}^2(z_q, x_{pq})}} \sum_p \text{cor}(z_q, x_{pq}) \times x_{pq}$$

2. Normalization of the vector $a_{1q} = (a_{11q}, \ldots, a_{1pq})$

3. Regression of $z_q$ on $t_{1q} = X_q^a_{1q}$ expressed in terms of $X_q$

4. Computation of the residuals $z_{q1}$ and $X_{q1}$ of the regressions of $z_q$ and $X_q$ on $t_{1q}$: $z_q = c_{1q} t_{1q} + z_{q1}$ and $X_q = t_{1q} p_{1q} + X_{q1}$

For successive components the procedure is iterated on residuals and assessed by means of cross-validation or stopped by the user.
Finally, the m-components PLS regression model yielding the weights for the outer estimate $v_q$:

$$z_q = c_{1q}t_{1q} + c_{2q}t_{2q} + \ldots + c_{mq}t_{mq} + \text{Residual}$$

$$= c_{1q}X_qa_1 + c_{2q}X_qa_2 + \ldots + c_{mq}X_qa_{m} + \text{Residual}$$

$$= X_q(c_{1q}a_{1q} + c_{2q}a^*_2q + \ldots + c_{mq}a^*_mq) + \text{Residual}$$

$$= w_{1q}X_{1q} + w_{2q}X_{2q} + \ldots + w_{pq}X_{pq} + \text{Residual}$$

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

**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**

<table>
<thead>
<tr>
<th>Outer weights (like in RGCCA)</th>
<th>Composite scores (like in PLS-PM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oriented to</strong></td>
<td></td>
</tr>
<tr>
<td>Covariances between LVs</td>
<td><strong>PLScow</strong></td>
</tr>
<tr>
<td>Correlations between LVs</td>
<td><strong>PLScore</strong></td>
</tr>
</tbody>
</table>

**PLScore Mode:**
PLS Mode oriented to maximizing **cor**relations between connected composites under normalization constraints on composite **scores**

**PLScow Mode:**
PLS Mode oriented to maximizing **cov**ariances between connected composites under normalization constraints on outer **weights**

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)

\[
m_y = 1 \\
\|X_q w_q\| = 1
\]

**PLScore Mode**

\[
m_y = P_y \\
\|X_q w_q\| = 1
\]

**Mode A**

[Esposito Vinzi et al., 2009]

Mode PLS in XLSTAT-PLSPM

**Mode B**

Correlation approach

PLS-R as an estimation method for measurement model in a modified PLS-PM

\[
m_y = 1 \\
\|w_q\| = 1
\]

**PLScow Mode**

\[
m_y = P_y \\
\|w_q\| = 1
\]

**New Mode A**

Covariance approach

**New Mode B**

**Mode Ridge (RGCCA)**

[Tenenhaus & Tenenhaus, 2011]
Hbat Model (Hair et al., 2010) with noisy variables

- Environmental Perceptions
- Job Satisfaction
- Staying intention
- Organizational Commitment
- Attitude toward Coworkers
- Job Satisfaction
- Organizational Commitment

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

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

PCA of the Organizational Commitment (OC + noisy data)

<table>
<thead>
<tr>
<th>PC</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>4.101</td>
<td>0.684</td>
<td>0.040</td>
</tr>
<tr>
<td>F2</td>
<td>0.475</td>
<td>0.220</td>
<td>0.077</td>
</tr>
<tr>
<td>F3</td>
<td>0.678</td>
<td>0.291</td>
<td>0.023</td>
</tr>
<tr>
<td>F4</td>
<td>0.387</td>
<td>0.203</td>
<td>0.034</td>
</tr>
<tr>
<td>F5</td>
<td>0.161</td>
<td>0.094</td>
<td>0.003</td>
</tr>
</tbody>
</table>

The real OC manifest variables appear only on the 3rd PC.
PLS Regression of the OC noisy data on Staying Intention (SI)

The noise variables are downweighted as they have no predictive power.

PLS Regression for the OC outer model in PLS-PM
A comparison between Modes PLScore, A and B: outer weights

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

Non-Metric PLS-PM
Steven’s measurement scale classification

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

- 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 practical 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.

<table>
<thead>
<tr>
<th>ID</th>
<th>$z$</th>
<th>$x$</th>
<th>$x_a$</th>
<th>$x_b$</th>
<th>$x_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>obs1</td>
<td>1</td>
<td>a</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>obs2</td>
<td>2</td>
<td>a</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>obs3</td>
<td>3</td>
<td>a</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>obs4</td>
<td>4</td>
<td>b</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>obs5</td>
<td>5</td>
<td>b</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>obs6</td>
<td>6</td>
<td>b</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>obs7</td>
<td>7</td>
<td>c</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>obs8</td>
<td>8</td>
<td>c</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>obs9</td>
<td>9</td>
<td>c</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$R^2 = 0.675$

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 \( \hat{x} \propto \hat{X}\phi \), where \( \phi = (\phi_1 \ldots \phi_K) \) is the vector of optimal scaling parameters and the matrix \( \hat{X} \) defines a space in which constraints imposed by the scaling level are respected.

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

Non-Metric PLS Path Modeling algorithm

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

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

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

- In standard PLS steps the outer weights are optimized for given scaling values.
- In the quantification step, instead, the scaling values are optimized for given outer weights: raw variables are properly transformed through scaling (quantification) functions \( Q() \)
NM-PLSPM algorithm iteration

Outer Estimation

\[ t_q = X_q w_q \]

Updating the outer weights (Normalization depending on the Mode)

Updating the inner weights

\[ \hat{x}_{pq} \propto Q \left( \tilde{X}_q w_q \right) \quad \text{Mode (new) A} \]

\[ \hat{x}_{pq} \propto Q \left( \tilde{X}_q w_q \right) \quad \text{Mode (new) B} \]

Quantification step

NM-PLSPM general criterion

\[
\arg \max_{w_q, \theta_q, \tilde{x}_q} \left\{ \sum_{q=1}^{Q} c_{qq} g \left[ \text{cor} \left( \hat{X}_q w_q, \tilde{X}_q w_q \right) \sqrt{\text{var} \left( \hat{X}_q w_q \right)} \sqrt{\text{var} \left( \tilde{X}_q w_q \right)} \right] \right\}
\]

s.t.  
\[ \left\| \tilde{x}_q \right\|^2 = \left\| \tilde{X}_q w_q \right\|^2 = n \]
\[ \left\| \tilde{x}_q \right\| = n \quad \text{if Mode B for block q} \]
\[ \left\| w_q \right\| = n \quad \text{if New Mode A for block q} \]

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:

Two types of quantification are currently allowed:

- **Nominal Scaling**, in which the following group constraint is considered:
  \[ x_i \sim x_j \Rightarrow \hat{x}_i = \hat{x}_j \]

- **Ordinal scaling**, in which a further order constraint is considered:
  \[ x_i^* \sim x_j^* \Rightarrow \hat{x}_i = \hat{x}_j \quad \text{and} \quad \hat{x}_i^* \prec \hat{x}_j^* \Rightarrow \hat{x}_i \leq \hat{x}_j \]

---

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

AGRI loadings
- 0.954
- 0.9588
- 0.6292

RUS data (1964): Quantifications

AGRI loadings
- 0.9634
- 0.9623

Gini farm

POLINS loadings
- 0.6254
- 0.8959
- 0.8995
- 0.8253

Russet data (1964): Model comparison

PLS Model:
- R² = 0.605
- GoF = 0.567

NM-PLS Model:
- R² = 0.793
- GoF = 0.772
### Application to mobile data

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

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Measure example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service (SB)</td>
<td>(a) It was not treated as it was easy and clear</td>
</tr>
<tr>
<td>(b) It is a mobile and handy maintained</td>
<td></td>
</tr>
<tr>
<td>(c) It has a social contribution for the society</td>
<td></td>
</tr>
<tr>
<td>(d) It is associated with the environment</td>
<td></td>
</tr>
<tr>
<td>Customer perception of the overall quality (Q)</td>
<td>(a) Dependence of the overall quality of &quot;your mobile phone provider&quot; to the actual you become satisfied of this provider</td>
</tr>
<tr>
<td>(b) Dependence of &quot;your mobile phone provider&quot; to provide product, and tendency to repeat your annual event</td>
<td></td>
</tr>
<tr>
<td>(c) Person’s satisfaction that things could be worse at &quot;your mobile phone provider&quot;</td>
<td></td>
</tr>
<tr>
<td>Served quality (LQ)</td>
<td>(a) Overall perception quality</td>
</tr>
<tr>
<td>(b) Technical quality of the network</td>
<td></td>
</tr>
<tr>
<td>(c) Customer service and general service offered</td>
<td></td>
</tr>
<tr>
<td>(d) Range of services and products offered</td>
<td></td>
</tr>
<tr>
<td>(e) Availability and accuracy of the provided and services provided</td>
<td></td>
</tr>
<tr>
<td>(f) Clarity and transparency of information provided</td>
<td></td>
</tr>
<tr>
<td>Served value (LV)</td>
<td>(a) Given the quality of the products and services offered by &quot;your mobile phone provider&quot; how would you rate the firm and place that competing for them</td>
</tr>
<tr>
<td>(b) Given the quality of the products and services offered by &quot;your mobile phone provider&quot; how would you rate the quality of the products and services offered</td>
<td></td>
</tr>
<tr>
<td>(c) Customer satisfaction (LS)</td>
<td>(a) Overall satisfaction</td>
</tr>
<tr>
<td>(b) Evaluation of expectations</td>
<td></td>
</tr>
<tr>
<td>(c) How would you rate &quot;your mobile phone provider&quot; compared with your usual mobile phone provider</td>
<td></td>
</tr>
<tr>
<td>Customer complaints (LC)</td>
<td>(a) How would you rate &quot;your mobile phone provider&quot; compared with your usual mobile phone provider</td>
</tr>
<tr>
<td>(b) You did not have complaint about &quot;your mobile phone provider&quot; but you have, because you have complaint about &quot;your mobile phone provider&quot;</td>
<td></td>
</tr>
<tr>
<td>(c) You did not have complaint about &quot;your mobile phone provider&quot; but you have, because you have complaint about &quot;your mobile phone provider&quot;</td>
<td></td>
</tr>
<tr>
<td>Customer loyalty (L)</td>
<td>(a) If you used to choose a new mobile phone provider how likely is that you would choose &quot;your provider&quot; again</td>
</tr>
<tr>
<td>(b) If you used to choose a new mobile phone provider how likely is that you would choose &quot;your provider&quot; again</td>
<td></td>
</tr>
<tr>
<td>(c) If you used to choose a new mobile phone provider how likely is that you would choose &quot;your provider&quot; again</td>
<td></td>
</tr>
</tbody>
</table>

#### Mobile data: Comparing model quality

<table>
<thead>
<tr>
<th>Linearity hypothesis</th>
<th>Monotonicity hypothesis</th>
<th>No hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>(No scaling)</td>
<td>(Ordinal Scaling)</td>
<td>(Nominal Scaling)</td>
</tr>
<tr>
<td>$GoF = 0.470$</td>
<td>$GoF = 0.526$</td>
<td>$GoF = 0.547$</td>
</tr>
<tr>
<td>$R^2 = 0.387$</td>
<td>$R^2 = 0.464$</td>
<td>$R^2 = 0.495$</td>
</tr>
<tr>
<td>$Com_M = 0.599$</td>
<td>$Com_M = 0.618$</td>
<td>$Com_M = 0.623$</td>
</tr>
<tr>
<td>$Red_M = 0.263$</td>
<td>$Red_M = 0.315$</td>
<td>$Red_M = 0.335$</td>
</tr>
</tbody>
</table>
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*

*Extract from W.W. Chin’s slide on mediation and moderation

Kind of Relationships between LVs

Review of Causal Relationships between Latent Variables

- Direct Causal Relationship
- Spurious Relationship
- Unanalyzed Relationship
- Indirect (Mediated) Causal Relationship
- Bidirectional Causal Relationship
- Moderated Causal Relationship

(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)

Equation

\[ Y = aX + bZ + cX \times Z + \delta \]

Equation

\[ Y = bZ + (a + cZ)X + \delta \]
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 \( P_1 \cdot P_2 \) product indicators.

2. These product indicators serve as indicators of the interaction term.

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 meaningful 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:

\[ Y = \beta_3 + \beta'_X + bM + \epsilon \]

**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.

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

Standard error of the mediated effect

Unobserved Heterogeneity in PLS-PM
A path model for BENETTON brand preference

[Ringle et al., 2005]

444 customers
10 manifest variables
3 variable blocks

PLS-PM estimates for BENETTON model

N=444

R²=.24

GoF=0.422
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

Global Model

“One model may not fit all”

Different sets of individuals with particular behaviors, i.e. with different sets of parameters
A priori unit segmentation

Global Model

Observed clustering variable:
c.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

A Priori Segmentation

Simultaneous Segmentation and Estimation

PLS-PM

PLS Typological Path Modeling

PATHMOX

REBUS-PLS

Finite Mixture PLS

Wald (1975), Tenenhaus et al. (2005)

Hahn et al. (2002)

Squillacciotti (2005)

Trinchera et al. (2006)

Sánchez and Aluja (2006)

Trinchera (2007), Esposito Vinzi et al. (2008)

Wold (1975), Tenenhaus et al. (2005)
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)

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

V. Esposito Vinzi1,*, 1, L. Tritchler2, S. Squillacciotti3 and M. Tenenhaus4

1ESSEC Business School of Paris, Avenue Bernard Hirsch - F-95044, Cergy Pontoise, France
2Uniwersytet Swietego Tomasza i Swietego Jodzdana w Krakowie, Kraków, Poland
3Universitá di Firenze, Dpto. di Statistica e Cianesi, 50134 Firenze, Italy
4INSA Paris, ENS, 1 Avenue Henry d'Orléans, 75011 Paris, France

SUMMARY

Structural equation models (SEM) 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 approach has been the most popular for estimating SEMs. However, this approach has severe limitations when 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 behaviors 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 unobserved heterogeneity an estimation 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).

$$CM_i = \frac{\sum_{i=1}^{Q} \sum_{p=1}^{P} e_{ipqk}^2 / \text{Com}(\xi_{qk}, x_{pq})}{\sum_{i=1}^{Q} \sum_{p=1}^{P} [f_{ijk}^2 / \text{R}^2(\xi_j, \xi_j's explaining \xi_j)]} \times \frac{\sum_{i=1}^{J} f_{ijk}^2 / \text{R}^2(\xi_j, \xi_j's explaining \xi_j)}{N - m_k - 1}$$

The Measurement residuals

The Structural residuals

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

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:

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

where: $$\hat{\xi}_{qk} = \sum_{p=1}^{P} w_{pqk} x_{pq}$$

$\lambda_{pqk}$ and $w_{pqk}$ are GROUP SPECIFIC!!

$P$ 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:

\[
\hat{\xi}_{ijk} = \sum_{p=1}^{n} w_{pq} \xi_{ipj}
\]

is obtained by using the outer weights estimated for the \( k \)-th group

GROUP SPECIFIC!!!

and

\[
y_{ijk} = \sum_{q=1}^{Q} \beta_{pq} \xi_{ijk}
\]

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!

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

\[
CM_{ik} = \frac{\sum_{q=1}^{Q} \xi_{iqk}^2 / R^2(\hat{\xi}_{iqk}, X_q) \sum_{j=1}^{J} f_{ijk}^2 / R^2(\xi_j, \xi_j \text{'s explaining } \hat{\xi}_i)}{\sum_{q=1}^{Q} \sum_{i=1}^{N} \xi_{iqk}^2 / R^2(\hat{\xi}_{iqk}, X_q) \sum_{j=1}^{J} \sum_{i=1}^{N} f_{ijk}^2 / R^2(\xi_j, \xi_j \text{'s explaining } \hat{\xi}_i)}
\]

\( (N - m_j - 1) \)

The Outer residuals
**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:

\[
g_{i\text{q}\text{k}} = \hat{\xi}_{i\text{q}\text{k}} - \hat{c}_{i\text{q}}
\]

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

\[
\hat{\xi}_{i\text{q}\text{k}} = \sum_{p=1}^{P} W_{p\text{q}\text{k}} x_{i\text{p}}
\]

Inner estimation of the latent variable

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

---

**REBUS-PLS algorithm**

Initial step

- Computation of the residuals
- Estimation of the global model

Cluster analysis on the residuals and definition of the number of classes $K$

Estimation of the $K$ local models

Computation of the $CM$ units/models

Allocation of the units to the closest local model

Reiterate until convergence on classes composition

Update the composition of the classes

REBUS-PLS is implemented in a R package (plspm) and in XLSTAT-PLSPM

---
BENETTON data: REBUS-PLS results

How many classes?

Three in REBUS-PLS

BENETTON data: REBUS-PLS results

The Global model results
N=444
GoF=0.422
GoF=0.455
GoF=0.676
GoF=0.417

Group 1 – 23% of the sample: Brand Preference impact more on Brand-usage than on Sympathy

Group 2 – 32% of the sample: Image-driven group

Group 3 – 45% of the sample: Character is more correlated with “Trends” than with the others.

Modernity
Style of living
Trust
Impression
Brand name
Fashion 2
Trends
Fashion 1

Imagery

Brand Preference

Sympathy

Brand usage

Character

GoF=0.26
R²=0.29
R²=0.67
R²=0.417

GoF=0.455
GoF=0.676
GoF=0.417

Character is more correlated with “Trends” than with the others.

30/11/2014
A real case example

Satisfaction survey on professional customers of a electricity provider

- 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

- 877 customers
  - \( R^2(\text{loy})=0.50 \)
  - GOF = 0.53

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

- 582 customers
  - \( R^2(\text{loy})=0.47 \)
  - 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

$$GQI_k = \left( \sum_{i=1}^{n_k} \frac{1}{P} \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[ 1 - \frac{\sum_{i=1}^{n_q} e^2_{ipq}}{\sum_{i=1}^{n_q} (x_{ipq} - \bar{x}_{ipq})^2} \right] \right) \times \frac{1}{J} \sum_{j=1}^{J} \left[ 1 - \frac{\sum_{i=1}^{n_j} f^2_{ipj}}{\sum_{i=1}^{n_j} (x_{ipj} - \bar{x}_{ipj})^2} \right]$$

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

BENETTON Data: evaluation of the partition

<table>
<thead>
<tr>
<th>Simple Statistics</th>
<th>GQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of observations</td>
<td>302</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.422</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.454</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>0.426</td>
</tr>
<tr>
<td>Median</td>
<td>0.428</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>0.431</td>
</tr>
<tr>
<td>Mean</td>
<td>0.429</td>
</tr>
<tr>
<td>Lower bound on mean (95%)</td>
<td>0.428</td>
</tr>
<tr>
<td>Upper bound on mean (95%)</td>
<td>0.429</td>
</tr>
</tbody>
</table>

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

Empirical distribution of the GQI values
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
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:

\[ H_0 : \beta_{ij}^{\text{Group 1}} = \beta_{ij}^{\text{Group 2}} \]

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):

\[ t = \frac{\text{Path}_{\text{sample 1}} - \text{Path}_{\text{sample 2}}}{\sqrt{\left(\frac{(m-1)^2}{m+n-2}\right) * S.E_{\text{sample 1}}^2 + \left(\frac{(n-1)^2}{m+n-2}\right) * S.E_{\text{sample 2}}^2}} \]

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

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

The t tests for path coefficients

If the variances are assumed different, a Smith-Satterthwait test can be applied to approximate a level \( \alpha \) test.

\[ t = \frac{\text{Path}_{\text{sample 1}} - \text{Path}_{\text{sample 2}}}{\sqrt{S.E_{\text{sample 1}}^2 + S.E_{\text{sample 2}}^2}} \]

\[ df = \text{round to nearest integer} \left[ \left( \frac{(S.E_{\text{sample 1}}^2 + S.E_{\text{sample 2}}^2)}{m+1} \right)^2 - \frac{S.E_{\text{sample 1}}^2}{m+1} \right] - 2 \]
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 bootstrap-based confidence intervals 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:

\[ H_0: \beta_{j*}^{\text{Group 1}} \geq \beta_{j*}^{\text{Group 2}} \]

\[ P(\hat{\beta}_{j*}^{\text{Group 1}} \geq \hat{\beta}_{j*}^{\text{Group 2}}) = 1 - \frac{1}{G^2} \sum_{g=1}^{G} \sum_{s=1}^{G} I(\hat{\beta}_{j*}^{\text{Group}(g)} < \hat{\beta}_{j*}^{\text{Group}(s)}) \]

where:

\[ I(\hat{\beta}_{j*}^{\text{Group}(g)} < \hat{\beta}_{j*}^{\text{Group}(s)}) = \begin{cases} 1 & \text{if } \hat{\beta}_{j*}^{\text{Group}(g)} < \hat{\beta}_{j*}^{\text{Group}(s)} \\ 0 & \text{otherwise} \end{cases} \]

Permutation tests (Chin, 2003)

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

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).


Permutation tests

**Principle:**

- Null hypothesis: “Difference between parameters equals 0”
- Used statistic: \( S = |\text{param}_{G1} - \text{param}_{G2}| \)

**Steps:**

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

Empirical p-value is defined as:

\[ P = \frac{1}{B-1} \sum_{i=1}^{B} I_{S_{int} < S_{perm}} \]

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:
Comparing structural coefficients across the 6 banks: t-test

<table>
<thead>
<tr>
<th>Groups</th>
<th>Difference</th>
<th>t (Observed value)</th>
<th>t (Critical value)</th>
<th>DF</th>
<th>p-value</th>
<th>alpha</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 vs 1</td>
<td>0.106</td>
<td>1.292</td>
<td>1.648</td>
<td>504</td>
<td>0.088</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>3 vs 1</td>
<td>0.106</td>
<td>1.104</td>
<td>1.648</td>
<td>500</td>
<td>0.135</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>3 vs 2</td>
<td>0.001</td>
<td>0.008</td>
<td>1.648</td>
<td>502</td>
<td>0.497</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>4 vs 1</td>
<td>0.124</td>
<td>1.396</td>
<td>1.648</td>
<td>501</td>
<td>0.082</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>4 vs 2</td>
<td>0.019</td>
<td>0.237</td>
<td>1.648</td>
<td>503</td>
<td>0.406</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>4 vs 3</td>
<td>0.018</td>
<td>0.192</td>
<td>1.648</td>
<td>499</td>
<td>0.424</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>5 vs 1</td>
<td>0.140</td>
<td>1.571</td>
<td>1.648</td>
<td>500</td>
<td>0.058</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>5 vs 2</td>
<td>0.034</td>
<td>0.436</td>
<td>1.648</td>
<td>502</td>
<td>0.331</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>5 vs 3</td>
<td>0.034</td>
<td>0.359</td>
<td>1.648</td>
<td>498</td>
<td>0.360</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>5 vs 4</td>
<td>0.016</td>
<td>0.182</td>
<td>1.648</td>
<td>499</td>
<td>0.428</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>6 vs 1</td>
<td>0.324</td>
<td>2.307</td>
<td>1.648</td>
<td>503</td>
<td>0.001</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td>6 vs 2</td>
<td>0.218</td>
<td>2.194</td>
<td>1.648</td>
<td>505</td>
<td>0.014</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td>6 vs 3</td>
<td>0.217</td>
<td>1.943</td>
<td>1.648</td>
<td>501</td>
<td>0.026</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td>6 vs 4</td>
<td>0.199</td>
<td>1.888</td>
<td>1.648</td>
<td>502</td>
<td>0.030</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td>6 vs 5</td>
<td>0.184</td>
<td>1.738</td>
<td>1.648</td>
<td>501</td>
<td>0.041</td>
<td>0.050</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Bank 6 is different from the other banks as concerns the relation between image and satisfaction.
Comparing structural coefficients: Permutation test

Comparing bank 1 to bank 6

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Difference</th>
<th>p-value</th>
<th>alpha</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image -&gt; QP</td>
<td>0.006</td>
<td>0.918</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>Image -&gt; sat</td>
<td>0.324</td>
<td>0.025</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td>QP -&gt; sat</td>
<td>0.263</td>
<td>0.042</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td>Image -&gt; fid</td>
<td>0.093</td>
<td>0.395</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>sat -&gt; fid</td>
<td>0.042</td>
<td>0.649</td>
<td>0.050</td>
<td>No</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Qualité du modèle (Variable latente)</th>
<th>Différence</th>
<th>p-value</th>
<th>alpha</th>
<th>Significatif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communalité (Image)</td>
<td>0,071</td>
<td>0,101</td>
<td>0,050</td>
<td>Non</td>
</tr>
<tr>
<td>Communalité (QP)</td>
<td>0,067</td>
<td>0,109</td>
<td>0,050</td>
<td>Non</td>
</tr>
<tr>
<td>Communalité (sat)</td>
<td>0,005</td>
<td>0,894</td>
<td>0,050</td>
<td>Non</td>
</tr>
<tr>
<td>Communalité (fid)</td>
<td>0,083</td>
<td>0,024</td>
<td>0,050</td>
<td>Oui</td>
</tr>
<tr>
<td>Redondance (Image)</td>
<td>Non défini</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redondance (QP)</td>
<td>0,035</td>
<td>0,595</td>
<td>0,050</td>
<td>Non</td>
</tr>
<tr>
<td>Redondance (sat)</td>
<td>0,067</td>
<td>0,375</td>
<td>0,050</td>
<td>Non</td>
</tr>
<tr>
<td>Redondance (fid)</td>
<td>0,111</td>
<td>0,062</td>
<td>0,050</td>
<td>Non</td>
</tr>
<tr>
<td>GoF</td>
<td>0,011</td>
<td>0,770</td>
<td>0,050</td>
<td>Non</td>
</tr>
</tbody>
</table>
Comparing bank 1 to bank 6 with the permutation test

**Outer model**

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Manifest variable</th>
<th>Difference</th>
<th>p-value</th>
<th>alpha</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>Q66S2 IMA</td>
<td>0.070</td>
<td>0.052</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Q66S3 IMA</td>
<td>0.417</td>
<td>0.003</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Q66S4 IMA</td>
<td>0.034</td>
<td>0.639</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Q66S5 IMA</td>
<td>0.052</td>
<td>0.344</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>QP</td>
<td>Q67S1 QP</td>
<td>0.005</td>
<td>0.955</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Q67S2 QP</td>
<td>0.001</td>
<td>0.984</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Q67S4 QP</td>
<td>0.231</td>
<td>0.060</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Q67S5 QP</td>
<td>0.044</td>
<td>0.512</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Q67S6 QP</td>
<td>0.013</td>
<td>0.700</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>sat</td>
<td>Q65 SAT</td>
<td>0.067</td>
<td>0.955</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Q81 SAT</td>
<td>0.067</td>
<td>0.038</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Q69 SAT</td>
<td>0.008</td>
<td>0.783</td>
<td>0.050</td>
<td>No</td>
</tr>
<tr>
<td>fid</td>
<td>Q73 FID</td>
<td>0.103</td>
<td>0.002</td>
<td>0.050</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Q80 FID</td>
<td>0.009</td>
<td>0.644</td>
<td>0.050</td>
<td>No</td>
</tr>
</tbody>
</table>

**Comparing latent variable scores (means)**

**Using the permutation test,** we have:

<table>
<thead>
<tr>
<th>Banque</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banque 1</td>
<td>0.06</td>
<td>0.12</td>
<td>0.76**</td>
<td>0.26</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>Banque 2</td>
<td>0.00</td>
<td>0.18</td>
<td>0.70**</td>
<td>0.20</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Banque 3</td>
<td>0.00</td>
<td>0.88**</td>
<td>0.36</td>
<td>0.49**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Banque 4</td>
<td>0.00</td>
<td>0.50**</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Banque 5</td>
<td>0.00</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Banque 6</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Differences between means with 1000 permutations**

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Banque 1</th>
<th>Banque 2</th>
<th>Banque 3</th>
<th>Banque 4</th>
<th>Banque 5</th>
<th>Banque 6</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>7,504</td>
<td>7,706</td>
<td>7,424</td>
<td>8,232</td>
<td>7,891</td>
<td>7,901</td>
<td>7,786</td>
</tr>
<tr>
<td>Qualité perçue</td>
<td>7,798</td>
<td>7,932</td>
<td>7,968</td>
<td>8,466</td>
<td>8,104</td>
<td>8,190</td>
<td>8,084</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>7,157</td>
<td>7,217</td>
<td>7,034</td>
<td>7,908</td>
<td>7,409</td>
<td>7,522</td>
<td>7,383</td>
</tr>
<tr>
<td>Fidélité</td>
<td>7,417</td>
<td>7,436</td>
<td>7,204</td>
<td>8,549</td>
<td>7,633</td>
<td>7,840</td>
<td>7,687</td>
</tr>
</tbody>
</table>

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...

“*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...”


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.
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

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”)
• PLS\textsuperscript{c} (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:
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 $\chi^2$ goodness of fit tests and classical inference

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.
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.

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)
Long Range Planning

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.
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? \( \rightarrow \) 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).

---

ERCIM 2014 - Pisa, 5 December 2014

Vincenzo Esposito Vinzi – ESSEC Paris
Main References 1/5


Main References 2/5


Main References 3/5


Main References 4/5


Main References 5/5


This presentation is made available through a Creative Commons Attribution-Noncommercial license. Details of the license and permitted uses are available at http://creativecommons.org/licenses/by-nc/4.0/

© 2014 V. Esposito Vinzi, L. Trinchera, G. Russolillo – PLS-PM
Title: Component-based Path Modeling
Attribution: Vincenzo Esposito Vinzi, Laura Trinchera, Giorgio Russolillo

For further information please contact Vincenzo Esposito Vinzi (vinzi@essec.edu).