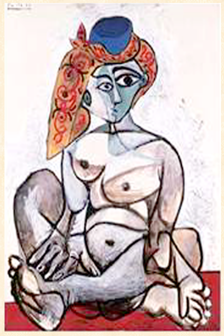
The PLS approach to Generalised Linear Models and Causal Path Modeling:

Algorithms and Applications



IASC Session
INTERFACE Meeting
Montreal (Canada)
April 19th, 2002

Vincenzo Esposito Vinzi
Dipartimento di Matematica e Statistica
Università degli Studi di Napoli "Federico II"
vincenzo.espositovinzi@unina.it

1

PLS1 Regression - Single y

Research of **m** (value chosen by **cross-validation**) **orthogonal** components $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h$ which are as **correlated** to **y** as possible and **also explanatory** of their own group.

$$Cov^2(Xw_h, y) = Cor^2(Xw_h, y) * Var(Xw_h)$$

PLS1 regression leads to a **compromise** between
a **multiple regression** of **y** on **X**
and a **principal component analysis** of **X**.

2

A new presentation of PLS1 in terms of OLS simple and multiple regressions

1. The m-components **PLS regression model** (**non linear** in the parameters) may be written as:

$$\mathbf{y} = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \right) + residual$$

with the **orthogonality** constraints on the PLS **components**.

2. The **first PLS component** is defined as:

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}^2\left(\mathbf{y}, \mathbf{x}_j\right)}} \sum_{j=1}^{p} \mathrm{cov}\left(\mathbf{y}, \mathbf{x}_j\right) \times \mathbf{x}_j$$

A new presentation of PLS1 in terms of OLS simple and multiple regressions

3. The covariance is also the regression coefficient (a_{1j}) in the **OLS simple regression** between **y** and $\mathbf{x}_j/\mathrm{var}(\mathbf{x}_j)$:

$$\mathbf{y} = a_{0j} + a_{1j}\left(\mathbf{x}_j / \mathrm{var}\left(\mathbf{x}_j\right)\right) + \boldsymbol{\varepsilon}$$

In fact:

$$a_{1j} = \frac{\mathrm{cov}\left(\mathbf{y}, \frac{1}{\mathrm{var}\left(\mathbf{x}_j\right)}\mathbf{x}_j\right)}{\mathrm{var}\left(\frac{1}{\mathrm{var}\left(\mathbf{x}_j\right)}\mathbf{x}_j\right)} = \mathrm{cov}\left(\mathbf{y}, \mathbf{x}_j\right)$$

4. **Atest on the regression coefficient** (a_{1j}) evaluates the importance of variable $\mathbf{x}_j$ in building up $\mathbf{t}_1$.

Non significant covariances are set to 0.

A new presentation of PLS1 in terms of OLS simple and multiple regressions

5. For the computation of the **second PLS component**, we first deflate **y** and $\mathbf{x}_j$'s with respect to $\mathbf{t}_1$:

$$\mathbf{y} = \mathbf{c}_1\mathbf{t}_1 + \mathbf{y}_1$$
$$\mathbf{x}_j = \mathbf{p}_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}$$

and then we define $\mathbf{t}_2$ as:

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p}\text{cov}^2(\mathbf{y}_1, \mathbf{x}_{1j})}}\sum_{j=1}^{p}\text{cov}(\mathbf{y}_1, \mathbf{x}_{1j}) \times \mathbf{x}_{1j}$$

6. Because of the **orthogonality between residual $\mathbf{x}_{1j}$ and component $\mathbf{t}_1$**, the covariance is now the regression coefficient in the following **OLS multiple regression**:

$$\mathbf{y} = c_1\mathbf{t}_1 + a_{2j}\left(\mathbf{x}_{1j} / \text{var}(\mathbf{x}_{1j})\right) + residual$$

A new presentation of PLS1 in terms of OLS simple and multiple regressions

7. **Partial correlation** between **y** and $\mathbf{x}_j$ conditioned to $\mathbf{t}_1$ is defined as the correlation between residuals $\mathbf{y}_1$ and $\mathbf{x}_{1j}$. The same applies to partial covariance:

$$\text{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1) = \text{cov}(\mathbf{y}_1, \mathbf{x}_{1j})$$

leading to:

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p}\text{cov}^2(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1)}}\sum_{j=1}^{p}\text{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1) \times \mathbf{x}_{1j}$$

8. Since ($\mathbf{t}_1$,$\mathbf{x}_{1j}$) and ($\mathbf{t}_1$,$\mathbf{x}_j$) span the same space, the contribution of variable $\mathbf{x}_j$ to the construction of $\mathbf{t}_2$ is finally tested by means of the following **OLS multiple regression**:

$$\mathbf{y} = d_{0j} + d_{1j}\mathbf{t}_1 + d_{2j}\mathbf{x}_j + \boldsymbol{\varepsilon}$$

Non significant covariances are set to 0.

A new presentation of PLS1 in terms of OLS simple and multiple regressions

9. The second PLS compoent $\mathbf{t}_2$ may be well expressed as a **function of the original variables** (namely, those retained for $\mathbf{t}_1$ and those significant for $\mathbf{t}_2$) because the **residuals $\mathbf{x}_{1j}$** are expressed as functions of the original variable $\mathbf{x}_j$:

$$\mathbf{x}_{1j} = \mathbf{x}_j - p_{1j}\mathbf{t}_1$$

10. The procedure **STOP**s when **all partial covariances become non significant**.

7

PLS for Logistic Regression

Bordeaux Wine Dataset

<u>Variables observed in 34 years (1924 - 1957)</u>

Meteorological Variables (covariates) - *standardised*

- TEMPERATURE : Sum of daily mean temperatures (°C)
- SUNSHINE : Duration of sunshine (hours)
- HEAT : Number of very warm days
- RAIN : Rain height (mm)

Ordinal Response Variable (three categories)

- QUALITY of WINE: 1=Good, 2=Average, 3=Poor

8

The Dataset
Bordeaux Wine

Obs	Year	Temperature	Sunshine	Heat	Rain	Quality
1	1924	3064	1201	10	361	2
2	1925	3000	1053	11	338	3
3	1926	3155	1133	19	393	2
4	1927	3085	970	4	467	3
5	1928	3245	1258	36	294	1
6	1929	3267	1386	35	225	1
7	1930	3080	966	13	417	3
8	1931	2974	1189	12	488	3
9	1932	3038	1103	14	677	3
10	1933	3318	1310	29	427	2
11	1934	3317	1362	25	326	1
12	1935	3182	1171	28	326	3
13	1936	2998	1102	9	349	3
14	1937	3221	1424	21	382	1
15	1938	3019	1230	16	275	2
16	1939	3022	1285	9	303	2
17	1940	3094	1329	11	339	2
18	1941	3009	1210	15	536	3
19	1942	3227	1331	21	414	2
20	1943	3308	1366	24	282	1
21	1944	3212	1289	17	302	2
22	1945	3361	1444	25	253	1
23	1946	3061	1175	12	261	2
24	1947	3478	1317	42	259	1
25	1948	3126	1248	11	315	2
26	1949	3458	1508	43	286	1
27	1950	3252	1361	26	346	2
28	1951	3052	1186	14	443	3
29	1952	3270	1399	24	306	1
30	1953	3198	1259	20	367	1
31	1954	2904	1164	6	311	3
32	1955	3247	1277	19	375	1
33	1956	3083	1195	5	441	3
34	1957	3043	1208	14	371	3

9

Classical Ordinal Logistic Regression

y = Quality : Good (1), Average (2), Poor (3)

Proportional Odds Ratio Model

PROB(y ≤ l) =

$$\frac{e^{\alpha_\ell + \beta_1 Temperature + \beta_2 Sunshine + \beta_3 Heat + \beta_4 Rain}}{1 + e^{\alpha_\ell + \beta_1 Temperature + \beta_2 Sunshine + \beta_3 Heat + \beta_4 Rain}}$$

10

Ordinal Logistic Regression
SAS Results (Proc LOGISTIC)

Score Test for the Proportional Odds Assumption

Chi-Square = 2.9159 with 4 DF (p=0.5720)

Model with equal slopes is acceptable

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.6638	0.9266	8.2641	0.0040
INTERCP2	1	2.2941	0.9782	5.4998	0.0190
TEMPERA	1	3.4268	1.8029	3.6125	0.0573
SUN	1	1.7462	1.0760	2.6335	0.1046
HEAT	1	-0.8891	1.1949	0.5536	0.4568
RAIN	1	-2.3668	1.1292	4.3931	0.0361

Significant at 10% risk level

Uncoherent Sign

11

Ordinal Logistic Regression
Model Prediction Performance

OBSERVED QUALITY / PREDICTION

Frequency	1	2	3	Total
1	8	3	0	11
2	2	8	1	11
3	0	1	11	12
Total	10	12	12	34

Result: 7 (20.6%) years are misclassified

12

Ordinal Logistic Regression
Problems for Interpretation

- **Not significant coefficients** for some covariates that are known to be influent
- **Uncoherent signs** for some coefficients
- High percentage of **misclassified observations**

↑

Multicollinearity between covariates

13

Covariates
Correlation Matrix

	Temperature	Sunshine	Heat	Rain
Temperature	1.00000	0.71235	0.86510	-0.40962
Sunshine	0.71235		0.64645	-0.47340
Heat	0.86510	0.64645	1.00000	-0.40114
Rain	-0.40962	-0.47340	-0.40114	1.00000

Quite **strong correlations** between Temperature, Heat and Sunshine

14

Use of PLS Discriminant Analysis

PLS Regression of y_1, y_2, y_3 on X

The PLS Procedure
Cross Validation for the Number of Latent Variables

Number of Latent Variables	Root Mean PRESS	Prob > PRESS
0	1.0313	0
1	0.8304	1.0000
2	0.8313	0.4990
3	0.8375	0.4450
4	0.8472	0.3500

(Test for larger residuals than minimum)

Minimum Root Mean PRESS = 0.830422 for 1 latent variable
Smallest model with p-value > 0.1: 1 latent

TABLE OF QUALITY BY PREDICTION

QUALITY / PREDICTION: Frequency	1	3	Total
1	11	0	11
2	4	7	11
3	1	11	12
Total	16	18	34

Result:
t_1: 12 (35.3%) years are misclassified
t_1 & t_2: 7 years are misclassified

15

PLS Logistic Regression with *variable selection*

Step 1 : Research of *m* orthogonal components **t_h = Xw_h** which are good predictors of **y** and explanatory of the **X** xariables.
-> *m* is the number of significant components based on p-values.

Step 2 : Logistic regression of **y** on the **t_h** components.

Step 3 : Express the logistic regression equation as a function of **X**.

16

PLS Logistic Regression
1^{st} order solution - t_1

1. Simple Logistic Regressions of **y** on each $\mathbf{x}_j$: regression coefficients w_{1j}
   The non significant coefficients w_{1j} are set to 0
   -> only **significant variables** contribute to $\mathbf{t}_1$

2. Normalization of $\mathbf{w}_1 = (w_{11},\ldots,w_{1k})$

3. Simple Logistic Regression of **y** on $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ expressed in terms of **X**

17

Bordeaux wines
Step 1: 1^{st} order solution - t_1

Four simple logistic regressions:

	Coefficient	p-value
Temperature	3.0117	.0002
Sunshine	3.3401	.0002
Heat	2.1445	.0004
Rain	-1.7906	.0016

PLS component t_1 :

$$t_1 = \frac{3.0117\ \text{Température} + 3.3401\ \text{Soleil} + 2.1445\ \text{Chaleur} - 1.7906\ \text{Pluie}}{\sqrt{(3.0117)^2 + (3.3401)^2 + (2.1445)^2 + (-1.7906)^2}}$$

$$= 0.5688\ \text{Température} + 0.6309\ \text{Soleil} + 0.4050\ \text{Chaleur} - 0.3382\ \text{Pluie}$$

18

Bordeaux wine
Step 2: Logistic Regression on t_1

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-2.2650	0.8644	6.8662	0.0088
Intercept2	1	2.2991	0.8480	7.3497	0.0067
t1	1	2.6900	0.7155	14.1336	0.0002

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	9	2	0	11
2	2	8	1	11
3	0	1	11	12
Total	11	11	12	34

6 misclassified years

Bordeaux wine
Step 3: Logistic Regression in terms of X

$$\text{Prob}(Y = 1) = \frac{e^{-2.265+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}{1+e^{-2.5265+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}$$

and

$$\text{Prob}(Y \leq 2) = \frac{e^{2.2991+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}{1+e^{2.2991+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}$$

Comment: This model outperforms the classical ordinal logistic regression model with respect to:

1) coherence of regression coefficients;
2) misclassification rate.

PLS Logistic Regression
2^{nd} order solution - t_2

1. Multiple Logistic Regressions of **y** on $\mathbf{t}_1$ and each $\mathbf{x}_j$
   -> retain the **significant** predictors
2. Calculation of the residuals $\mathbf{x}_{1j}$ related to simple regressions of retained variables on $\mathbf{t}_1$
3. Multiple Logistic Regression of **y** on $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ and each residual $\mathbf{x}_{1i}$ of retained variables -> regression coefficients $\mathbf{w}_{2j}$ of $\mathbf{x}_{1i}$
4. Normalization of $\mathbf{w}_2 = (w_{21},\ldots,w_{2k})$
5. Calculation of $\mathbf{w}^*_2$ such that $\mathbf{t}_2 = \mathbf{X}_1\mathbf{w}_2 = \mathbf{X}\mathbf{w}^*_2$
6. Multiple Logistic Regression of **y** on $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ and $\mathbf{t}_2 = \mathbf{X}\mathbf{w}^*_2$ both expressed as a function of **X**

Bordeaux wine
Selection of Variables contributing to t_2

Multiple Logistic Regressions of Quality on $\mathbf{t}_1$ and each $\mathbf{x}_j$

	Coefficient	p-value
Temperature	-.6309	.6765
Sunshine	.6459	.6027
Heat	-1.9407	.0983
Rain	-.9798	.2544

Comment:
All coefficients are non significant at a level of 5%

-> only the first PLS component is retained

PLS Logistic Regression

The Regression Equation for a binary *y*

$$\widehat{\log\left(\frac{\pi}{1-\pi}\right)} = c_1\mathbf{t}_1 + \ldots + c_h\mathbf{t}_h$$

$$= c_1\mathbf{X}\mathbf{w}_1^* + \ldots + c_h\mathbf{X}\mathbf{w}_h^* = \mathbf{X}\mathbf{b}$$

$$\boxed{\mathbf{b} = c_1\mathbf{w}_1^* + \ldots + c_h\mathbf{w}_h^*}$$

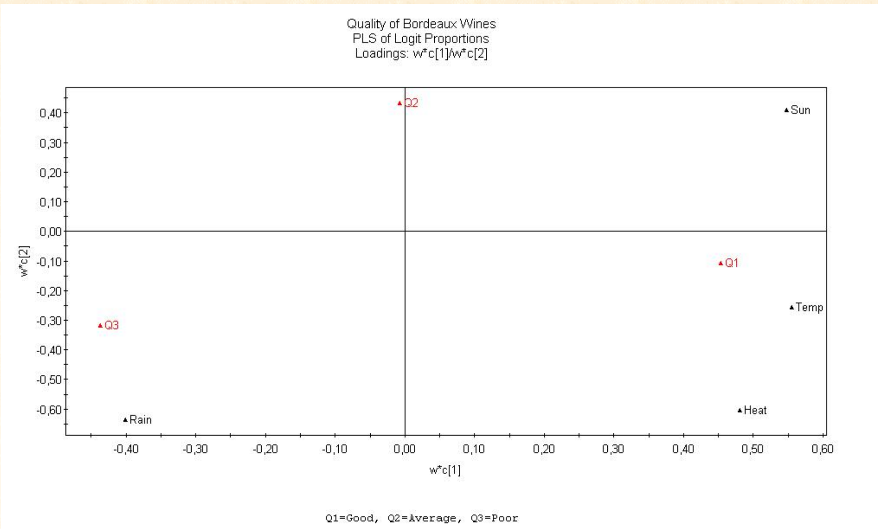
Graphical Representations as in PLSR

"Data Analysis Approach"

23

PLS Logistic Regression

A Graphical Representation of the Decomposition of *b*



24

Logistic Regression on PLS components
Second Algorithm

(1) **PLS regression** of the binary variables describing the categories of **y** on **X** variables.

(2) **Logistic regression** of **y** on the **X**-PLS components.

25

Logistic Regression on PLS components
Results

- **Temperature of year 1924** is supposed to be unknown (missing)
- PLS regression of {Good, Average, Poor} on {Temperature, Sunshine, Heat, Rain} leads to one PLS component $\mathbf{t}_1$ (cross validation result):

$$\mathbf{t}_1 = 0.55\times Temperature + 0.55\times Sun + 0.48\times Heat - 0.40\times Rain$$

$$\mathbf{t}_{11} = (0.55\times Sun + 0.48\times Heat - 0.40\times Rain)/0.69 = -0.90285 \text{ for year 1924}$$

26

Logistic Regression on PLS component t_1

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.1492	0.8279	6.7391	0.0094
INTERCP2	1	2.2845	0.8351	7.4841	0.0062
t1	1	2.6592	0.7028	14.3182	0.0002

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	9	2	0	11
2	2	8	1	11
3	0	1	11	12
Total	11	11	12	34

6 misclassified years

27

Logistic Regression on PLS component The Model

Prob (Y≤ i)=

$$= \frac{e^{-2.15\times Bon + 2.28\times Moyen + 2.66\times t_1}}{1 + e^{-2.15\times Bon + 2.28\times Moyen + 2.66\times t_1}}$$

$$= \frac{e^{-2.15\times \mathrm{Bon} + 2.28\times \mathrm{Moyen} + 1.47\times \mathrm{Temp.} + 1.46\times \mathrm{Soleil} + 1.28\times \mathrm{Chaleur} - 1.07\times \mathrm{Pluie}}}{1 + e^{-2.15\times \mathrm{Bon} + 2.28\times \mathrm{Moyen} + 1.47\times \mathrm{Temp.} + 1.46\times \mathrm{Soleil} + 1.28\times \mathrm{Chaleur} - 1.07\times \mathrm{Pluie}}}$$

28

Algorithm 3 (Grouped Data)
PLS Regression of the response **logit** on the predictors

Example: Job satisfaction

(Models for discrete data, D. Zelterman, Oxford Press,1999)

- 9949 employees in the 'craft' job within a company
- Response : Satisfied/Dissatisfied
- Demographic Factors : Sex, Race (White/Nonwhite), Age (<35, 35-44, >44), Region (Northeast, Mid-Atlantic, Southern, Midwest, Northwest, Southwest, Pacific)
- Objective: Explain Job satisfaction by means of: **all main effects** (factors) **and** 2^{nd} **order interactions**.

Job Satisfaction :
First PLS component t_1
Variables contributing to the construction of t_1

Logistic Regression of Job Satisfaction on:
- each **factor**, taken one at a time (**simple** regressions);
- **interactions with main effects** (**multiple** regressions).

Variable	Wald	p-value
Race	2.687	.1012
Age	51.4856	**<.0001**
Sex	20.8241	**<.0001**
Region	33.9109	**<.0001**
Race*Age	1.0578	.5893
Race*Sex	10.77	**.001**
Race*Region	3.4125	.7556
Age*Sex	7.9389	**.0189**
Age*Region	7.8771	.7947
Sex*Region	4.1857	.6516

Job Satisfaction:
First PLS component t_1

$$t_1 =$$

$$\beta_0 + \begin{matrix}\text{Non-Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}\beta_1\\ -\beta_1\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}\beta_2\\ \beta_3\\ -\beta_2-\beta_3\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}\beta_4\\ -\beta_4\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid-Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}\beta_5\\ \beta_6\\ \beta_7\\ \beta_8\\ \beta_9\\ \beta_{10}\\ -\beta_5-\ldots-\beta_{10}\end{bmatrix}$$

$$+ \begin{matrix}\text{Non-Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}\beta_{11} & -\beta_{11}\\ -\beta_{11} & \beta_{11}\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}\beta_{12} & -\beta_{12}\\ \beta_{13} & -\beta_{13}\\ -\beta_{12}-\beta_{13} & \beta_{12}+\beta_{13}\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

Job Satisfaction:
First PLS component t_1

The first PLS component $\mathbf{t_1}$ is yielded by a PLS regression of **logit[Prob(Satisfied)]** on the variables:

- Non white - White
- $\text{Age}_{<35}$ - $\text{Age}_{>44}$

...

- $(\text{Age}_{35\text{-}44} - \text{Age}_{>44})*(\text{Male} - \text{Female})$

Job Satisfaction:
First PLS component t_1

$$t_1 =$$

$$-.01 + \begin{matrix}\text{Non-Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.01\\ +.01\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}-.78\\ -.43\\ +1.21\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}+.49\\ -.49\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid-Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}-.49\\ +.35\\ +.14\\ -.26\\ -.21\\ -.03\\ +.50\end{bmatrix}$$

$$+ \begin{matrix}\text{Non-Blanc}\\ \text{Blanc}\\ \\ \end{matrix}\begin{bmatrix}+.43 & -.43\\ -.43 & +.43\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\\ \\ \end{matrix}\begin{bmatrix}-.02 & +.02\\ -.13 & +.13\\ +.15 & -.15\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

Job Satisfaction:
Logistic Regression of Satisfaction on t_1

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.6227	0.0216	830.6539	<.0001
t1	1	0.1989	0.0212	88.0183	<.0001

Job Satisfaction:
Logistic Regression of Satisfaction on t_1 expressed as a function of X

$$\text{Logit(Prob(Satisfait))} =$$

$$0.62 + \begin{matrix}\text{Non-Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.002\\ +.002\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}-.16\\ -.09\\ +.25\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}+.10\\ -.10\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid-Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}-.097\\ +.070\\ +.028\\ -.053\\ -.041\\ -.007\\ +.100\end{bmatrix}$$

$$+ \begin{matrix}\text{Non-Blanc}\\ \text{Blanc}\\ \\ \end{matrix}\begin{bmatrix}+.09 & -.09\\ -.09 & +.09\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\\ \\ \end{matrix}\begin{bmatrix}-.003 & +.003\\ -.025 & +.025\\ +.028 & -.028\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

35

Job Satisfaction :
Second PLS component t_2
Variables contributing to the construction of t_2

Multiple Logistic Regression of Job Satisfaction on:
- **t_1** and each **factor**, taken one at a time;
- **t_1** and **interactions with main effects**.

Variables	Wald	p-value
Race	.20	.66
Age	12.81	**.00**
Sex	4.39	**.04**
Region	16.28	**.01**
Race*Age	.71	.70
Race*Sex	.44	.51
Race*Region	4.05	.67
Age*Sex	7.23	**.03**
Age*Region	7.86	.80
Sex*Region	3.19	.78

36

Job Satisfaction:
Second PLS component t_2

The second PLS component $\mathbf{t_2}$ is yielded by a PLS regression of **logit[Prob(Satisfied)]** on the **residuals** from regressions of the variables:

- Non white - White
- $Age_{<35}$ - $Age_{>44}$
- …
- $(Age_{35\text{-}44} - Age_{>44})*(Male - Female)$

on the first PLS component $\mathbf{t_1}$.

37

Job Satisfaction:
Second PLS component t_2

$$t_2 =$$

$$0.004 + \begin{matrix} \text{Non-Blanc} \\ \text{Blanc} \end{matrix}\begin{bmatrix} -.008 \\ +.008 \end{bmatrix} + \begin{matrix} <35 \\ 35-44 \\ >44 \end{matrix}\begin{bmatrix} -.12 \\ +.85 \\ -.73 \end{bmatrix} + \begin{matrix} \text{Homme} \\ \text{Femme} \end{matrix}\begin{bmatrix} +.61 \\ -.61 \end{bmatrix} + \begin{matrix} \text{Northeast} \\ \text{Mid-Atlantic} \\ \text{Southern} \\ \text{Midwest} \\ \text{Northwest} \\ \text{Southwest} \\ \text{Pacific} \end{matrix}\begin{bmatrix} -.56 \\ +1.34 \\ +.93 \\ -.01 \\ +.11 \\ +.56 \\ -2.37 \end{bmatrix}$$

$$+ \begin{matrix} \text{Non-Blanc} \\ \text{Blanc} \\ \end{matrix}\begin{bmatrix} +.30 & -.30 \\ -.30 & +.30 \\ \text{Homme} & \text{Femme} \end{bmatrix} + \begin{matrix} <35 \\ 35-44 \\ >44 \\ \end{matrix}\begin{bmatrix} +.14 & -.14 \\ -.07 & +.07 \\ -.07 & +.07 \\ \text{Homme} & \text{Femme} \end{bmatrix}$$

38

Job Satisfaction:
Logistic Regression of Satisfaction on t_1 and t_2

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.6172	0.0217	809.8129	<.0001
t1	1	0.2075	0.0214	93.7883	<.0001
t2	1	0.0486	0.0187	6.7525	0.0094

39

Job Satisfaction:
Logistic Regression of Satisfaction on t_1 and t_2 expressed as a function of X

$$\text{Logit(Prob(Satisfait))} =$$

$$0.62 + \begin{array}{r}\text{Non-Blanc}\\ \text{Blanc}\end{array}\begin{bmatrix}-.003\\ +.003\end{bmatrix} + \begin{array}{r}<35\\ 35-44\\ >44\end{array}\begin{bmatrix}-.17\\ -.05\\ +.22\end{bmatrix} + \begin{array}{r}\text{Homme}\\ \text{Femme}\end{array}\begin{bmatrix}+.13\\ -.13\end{bmatrix} + \begin{array}{c}\text{Northeast}\\ \text{Mid-Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{array}\begin{bmatrix}-.13\\ +.14\\ +.07\\ -.06\\ -.04\\ +.02\\ +.00\end{bmatrix}$$

$$+ \begin{array}{r}\text{Non-Blanc}\\ \text{Blanc}\\ \\ \end{array}\begin{bmatrix}+.10 & -.10\\ -.10 & +.10\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{array}{r}<35\\ 35-44\\ >44\\ \\ \end{array}\begin{bmatrix}+.003 & -.003\\ -.030 & +.030\\ +.027 & -.027\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

40

Job Satisfaction:
Logistic Regression of Satisfaction on t_1, t_2 and t_3

Model based on three PLS components:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.6502	0.0240	732.8875	<.0001
t1	1	0.2193	0.0217	102.1492	<.0001
t2	1	0.0369	0.0193	3.6493	0.0561
t3	1	0.0476	0.0145	10.8368	0.0010

41

Job Satisfaction:
Logistic Regression of Satisfaction on t_1, t_2 and t_3 expressed as a function of X

$$
\text{Logit(Prob(Satisfait))} =
$$

$$
0.62 + \begin{array}{r} \text{Non-Blanc} \\ \text{Blanc} \end{array}\begin{bmatrix} -.003 \\ +.003 \end{bmatrix} + \begin{array}{r} <35 \\ 35-44 \\ >44 \end{array}\begin{bmatrix} -.17 \\ -.05 \\ +.22 \end{bmatrix} + \begin{array}{r} \text{Homme} \\ \text{Femme} \end{array}\begin{bmatrix} +.13 \\ -.13 \end{bmatrix} + \begin{array}{r} \text{Northeast} \\ \text{Mid-Atlantic} \\ \text{Southern} \\ \text{Midwest} \\ \text{Northwest} \\ \text{Southwest} \\ \text{Pacific} \end{array}\begin{bmatrix} -.13 \\ +.14 \\ +.07 \\ -.06 \\ -.04 \\ +.02 \\ +.00 \end{bmatrix}
$$

$$
+ \begin{array}{r} \text{Non-Blanc} \\ \text{Blanc} \\ \\ \end{array}\begin{bmatrix} +.10 & -.10 \\ -.10 & +.10 \\ \text{Homme} & \text{Femme} \end{bmatrix} + \begin{array}{r} <35 \\ 35-44 \\ >44 \\ \\ \end{array}\begin{bmatrix} +.003 & -.003 \\ -.030 & +.030 \\ +.027 & -.027 \\ \text{Homme} & \text{Femme} \end{bmatrix}
$$

42

Job Satisfaction :
Fourth PLS component t_4
Variables contributing to the construction of t_4

Multiple Logistic Regression of Job Satisfaction on:
- **t_1, t_2, t_3**, and each **factor**, taken one at a time;
- **t_1, t_2, t_3**, and **interactions with main effects**.

Variables	Wald	p-value
Race	.22	.64
Age	.77	.68
Sex	1.63	.20
Region	8.60	.20
Race*Age	.74	.69
Race*Sex	.23	.63
Race*Region	4.64	.59
Age*Sex	3.66	.16
Age*Region	7.75	.80
Sex*Region	3.05	.80

All p-values >0.10

Conclusion: The fourth PLS component is not significant.
The model is built on 3 components.

43

A more **Exploratory** Approach

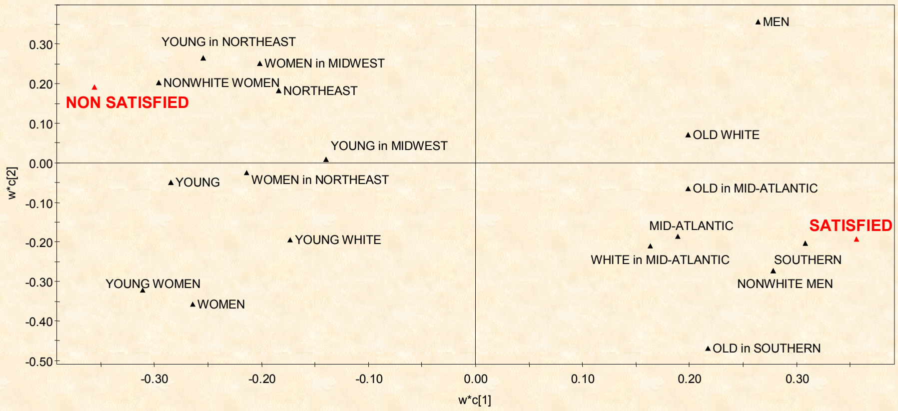
(1) **PLS Regression** of:
Y_1 = Logit(proportion of satisfied people)
Y_2 = Logit(proportion of non satisfied people)
on the 4 factors and all interactions;

(2) Iterative elimination of predictors with **small VIP**, verifying an increase of Q^2(cum);

(3) **Map** of the finally retained variables.

44

Graphical Representation of **PLS Regression of Logits**



Y_1 = Logit (Proportion of Satisfied)
Y_2 = Logit (Proportion of Non Satisfied)
X = Explanatory variables kept after elimination of **small VIP terms**

45

Considerations on PLS Logistic Regression

- The **« principles » of PLS regression** have been extended to logistic regression (**qualitative**);
- Algortihm 1 and Algorithm 2 show **comparable** results and performances;
- Logistic regression on PLS components is immediate at the **implementation** level (SIMCA + SAS or SPSS);
- Algorithm 3 is specifically developed for **grouped data** where logit can be computed;

46

Hints for Further Research

- Further **applications and simulation studies** are needed for better evaluating performances and for studying properties + **optimisation criteria**;
- Extensions to the linear modeling of a:
  - **transformation g(π)** of the pdf of **y** as a function of **X** (PROC LOGISTIC and PROC CATMOD in SAS);
  - **transformation g(μ)** of the mean of **y** as a function of **X** (PROC GENMOD in SAS);
- **G**eneralised **LI**near **M**odel **(Bastien & Tenenhaus 2001).**

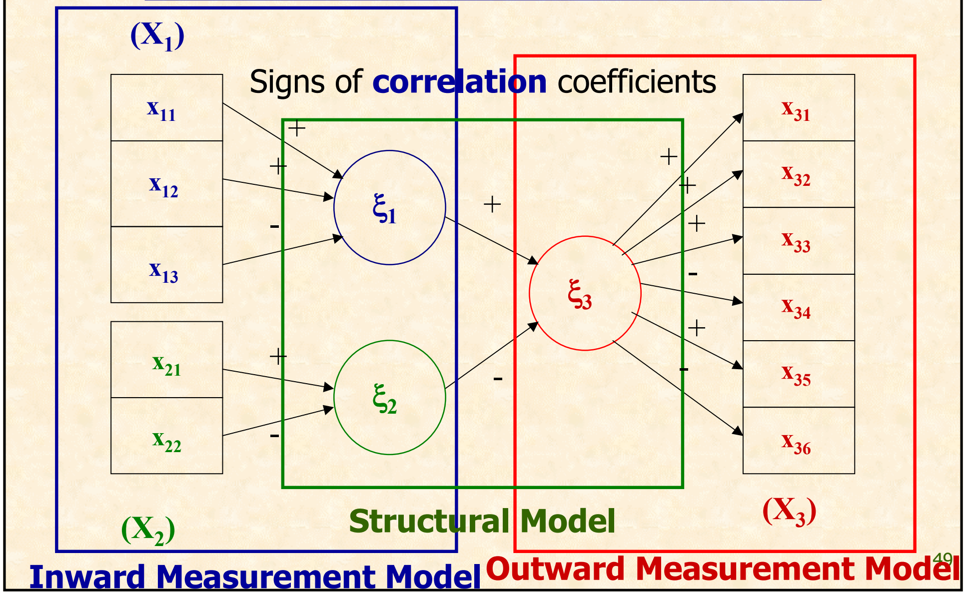
47

"PLS Path Modeling"
The PLS Approach (NILES or NIPALS) of Herman WOLD to Structural Equations Modeling

- Study of a system of **linear relationships** between latent variables **by solving blocks** (combinations of theoretical constructs and measurements) **one at a time** (partial) by use of **interdependent OLS regressions**: no global scalar function for optimization but **fixed-point** (FP) constraint.
- The overall diagram is partitioned into the designated blocks and an **initial estimate of the composite or latent variable** is established whose **scores are constrained to unitary variance**.
- LVPLS is **never underidentified** -> no constraints are needed on any of the parameters in the model as it is the case in SEM.
- The Least Squares criterion is applied on the **residuals** of both manifest and latent variables (here, with a **preference for the estimation of latent variables from their manifest ones** as the theory is softer than the empirical observations).
- **Predictions and parameter accuracy may not be jointly optimised**: optimizing the prediction of composite scores requires deemphasizing parameter estimation between latent variables.

48

"PLS Path Modeling"
The PLS Approach (NILES or NIPALS) of Herman WOLD to Structural Equations Modeling



Model Equations

- Each (reflective) **manifest variable** is written as (**outer-directed** measurement model):

Loadings

$$\mathbf{x}_{jh} = \lambda_{jh}\xi_h + \varepsilon^x_{jh} \qquad (\xi_h = \text{exogenous variables})$$

$$\mathbf{y}_{lk} = \lambda_{lk}\eta_k + \varepsilon^y_{jk} \qquad (\eta_k = \text{endogenous variables})$$

- Each (formative) **manifest variable** may contribute (**inner-directed** measurement model) to the corresponding latent variable:

Weights

$$\xi_h = \Sigma \pi_{jh}\mathbf{x}_{jh} + \delta^{\xi}{}_h$$

$$\eta_k = \Sigma \pi_{lk}\mathbf{y}_{lk} + \delta^{\eta}{}_k$$

Linear Conditional Expectation

- There is a **structural relationship** among the latent variables (**structural** model):

Path Coefficients

$$\eta_k = \Sigma_{k'\text{->}k}\beta_{k'}\eta_{k'} + \Sigma_{h\text{->}k}\gamma_h\xi_h + \zeta_k = E(\eta_k|\eta_{k'},\xi_h) + \zeta_k$$

50

Estimation Options of PLS Path Modeling

External Estimation

weighted aggregate of MV's

$$\mathbf{v}_h \propto \sum_j w_{jh}\mathbf{x}_{jh} = \mathbf{X}_h\mathbf{w}_h$$

Mode Centroid:
$w_{jh} = \text{sign}[\text{cor}(\mathbf{x}_{jh}, \mathbf{z}_h)]$

Mode A
(for reflective/endogenous vars.):
$w_{jh} = \text{cor}(\mathbf{x}_{jh}, \mathbf{z}_h)$
-> first PLS regression comp.

Mode B
(for formative/exogenous vars.):
$\mathbf{w}_h = (\mathbf{X}_h'\mathbf{X}_h)^{-1}\mathbf{X}_h'\mathbf{z}_h$
-> multiple regression = all PLS regression components

Mode PLS: intermediate

Internal Estimation

weighted aggregate of adjacent LV's

$$\mathbf{z}_h \propto \sum e_{hh'}\mathbf{v}_{h'}$$

Centroid Scheme (Wold's original):
$e_{hh'} = \text{sign}[\text{cor}(\mathbf{v}_h, \mathbf{v}_{h'})]$
-> problems with correlations ≈ 0.

Factorial Scheme (PLS, Lohmoller):
$e_{hh'} = r_{hh'} = \text{cor}(\mathbf{v}_h, \mathbf{v}_{h'})$

Structural Scheme (Path Weighting):
$e_{hh'}$ = multiple regression coefficient of $\mathbf{v}_h$ on $\mathbf{v}_{h'}$ if $\xi_{h'}$ is explicative of ξ_h
$e_{hh'} = r_{hh'}$ if ξ_h explicative of $\xi_{h'}$

Mode PLS: intermediate
Mode LISREL: take LISREL estimates

51

Computation of Estimates

An example with Mode A + Centroid Scheme

(1) External Estimates

$\mathbf{v}_1 = \mathbf{X}_1\mathbf{w}_1$

$\mathbf{v}_2 = \mathbf{X}_2\mathbf{w}_2$

$\mathbf{v}_3 = \mathbf{X}_3\mathbf{w}_3$

(2) Internal Estimates

$\mathbf{z}_1 = \mathbf{v}_3$

$\mathbf{z}_2 = -\mathbf{v}_3$

$\mathbf{z}_3 = \mathbf{v}_1 - \mathbf{v}_2$

(3) Computation of w_h

$w_{1j} = \text{cor}(\mathbf{x}_{1j}, \mathbf{z}_1)$

$w_{2j} = \text{cor}(\mathbf{x}_{2j}, \mathbf{z}_2)$

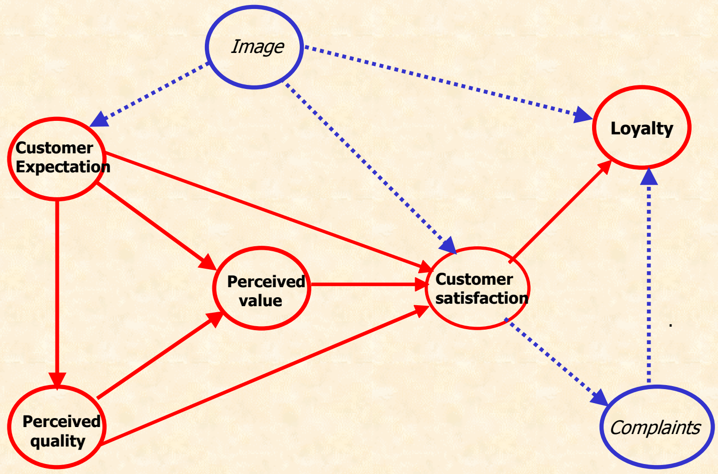
$w_{3j} = \text{cor}(\mathbf{x}_{3j}, \mathbf{z}_3)$

Algorithm

- Start with **arbitrary weights** $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$. $w_1 = (1,0,\dots,0)$
- Obtain the **new weights** w_h by means of steps from (1) to (3).
- **Iterate** the procedure till **convergence** (guaranteed only for 2 blocks but encountered in practice also for more than 2 blocks).

52

Path model describing causes and consequences of ECSI (European Customer Satisfaction Index)



Full model in red and blue, Reduced model in red

53

Computation of the latent variables
The Fornell Mode

Example : Customer Satisfaction Index

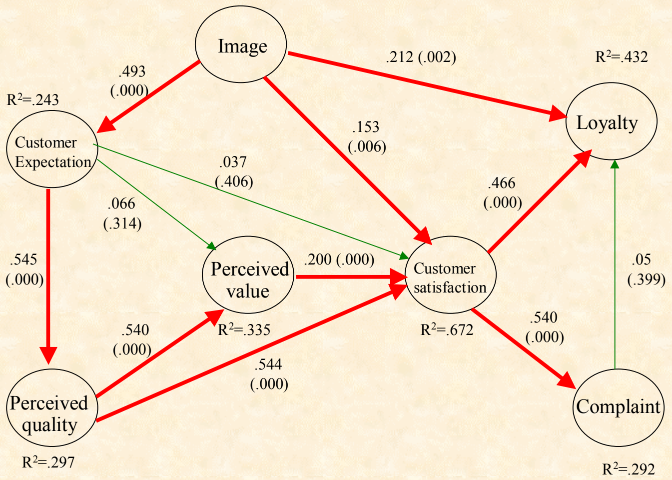
$$\text{CSI} = \frac{0.0158 \times \text{C\_sat1} + 0.0231 \times \text{C\_sat2} + 0.0264 \times \text{C\_sat3}}{0.0158 + 0.0231 + 0.0264}$$

Mean and standard deviation of the latent variables

	N	Minimum	Maximum	Mean	Std. Deviation
IMAGE	250	26.49	100.00	72.6878	13.7660
CUSTOMER EXPECTATION	250	25.85	100.00	72.3198	14.1259
PERCEIVED QUALITY	250	23.95	100.00	74.5765	14.2573
PERCEIVED VALUE	250	.00	100.00	61.5887	20.5987
CUSTOMER SATISFACTION	250	23.68	100.00	71.2876	15.3417
COMPLAINT	250	.00	100.00	67.4704	25.2684
LOYALTY	250	1.29	100.00	69.1757	21.2668

54

ECSI Path model for a "Mobile phone provider" (Regression on standardized variables)



55

PLS vs. LISREL

PLS is related to LISREL as PCA is related to FACTOR ANALYSIS

PLS

- Oriented to Prediction of MV's and LV's (variance-based)
- Reflective + Formative MV's
- **Distribution free + Predictor Specification**
- Observations may be dependent
- **Each latent variable is a linear combination of its own manifest variables**
- Consistency "at large"
- Optimal prediction accuracy
- Evaluation of the predictive performance by means of jackknife -> Q^2
- **N=10, p=28**
- Better **measurement model** because latent variables are constrained in the **X-space**

LISREL

- Oriented to parameter estimation (modeling covariances)
- Typically Reflective LV's
- Distributional Assumptions
- **Observations need to be independent**
- **Factor Indeterminacy**
- Indirect estimation of the latent variables built with the whole set of manifest variables
- **Consistent estimates**
- **Optimal parameter accuracy**
- Model evaluation by means of hypothesis testing so that N is required to be big enough
- **Sooner or later the model will be refused by chi-square -> RMSEA**
- Better **structural model** because latent variables are **space-free**

56

Main References for PLS Logistic and GLM

- Bastien, P. & Tenenhaus, M. (2001) : PLS generalized linear regression. Application to the analysis of life time data, Proceedings of the 2nd International Symposium on PLS and Related Methods, (Capri, October 1-3, 2001), Paris: CISIA-CERESTA.
- Esposito Vinzi, V. & Tenenhaus, M. (2001) : PLS logistic regression, Proceedings of the 2nd International Symposium on PLS and Related Methods, (Capri, October 1-3, 2001), Paris: CISIA-CERESTA.
- Esposito Vinzi, V. & Tenenhaus, M. (2002) : PLS logistic regression: recent developments with variable selection and grouped data features, Club PLS, (Jouy-en-Josas, March 14, 2002).
- Marx, B.D. (1996) : Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. Technometrics, vol. 38, n°4, pp. 374-381.
- Tenenhaus, M. (1998) : La régression PLS. Paris : Technip.
- Tobias, R.D. (1996) : An introduction to Partial Least Squares Regression. SAS Institute Inc., Cary, NC.
- Wold S., Ruhe A., Wold H. & Dunn III, W. J. (1984) : The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput., vol. 5, n° 3, pp. 735-743.

Main References for PLS Path Modeling

M.P. Bayol, A. de la Foye, C. Tellier, M. Tenenhaus :
Use of PLS Path Modeling to Estimate the European Consumer Satisfaction Index (ECSI) Model, Statistica Applicata - Italian Journal of Applied Statistics, (12), 3, 361-375, 2000

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

C. Lauro, V. Esposito Vinzi :
Some contributions to PLS Path Modeling and a system for the European Customer Satisfaction, Italian Statistical Society Meeting, 2002

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

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

H. Wold :
Soft modeling. The basic design and some extensions, in: Vol.II of Jöreskog-Wold (eds.), Systems under indirects observation, North-Holland, Amsterdam, 1982