

PLS path modeling and RGCCA for multi-block data analysis

Michel Tenenhaus

Sense 3 Appellations 4 Soils Loire Red Wines

4 blocks of variables

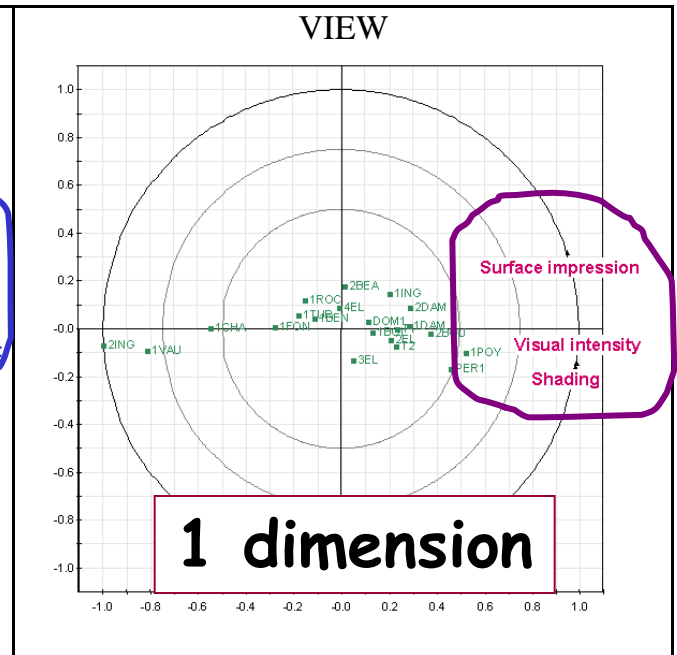
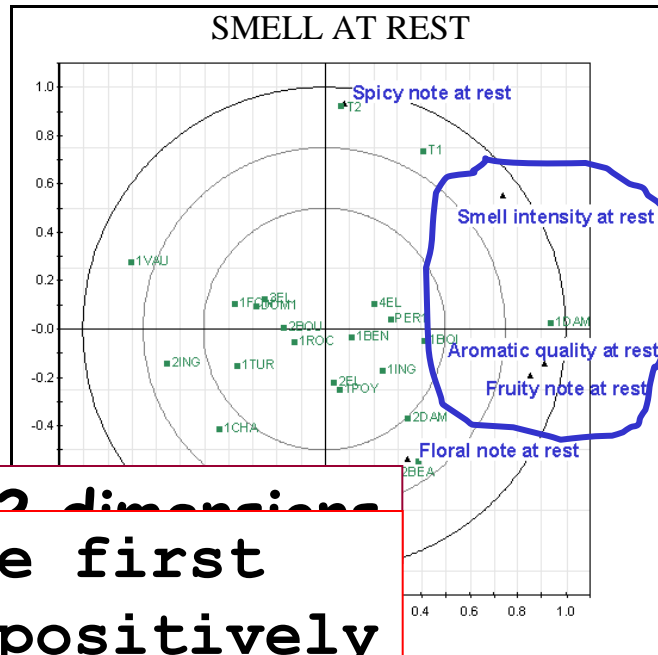
A famous example of Jérôme Pagès

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

Illustrative variable

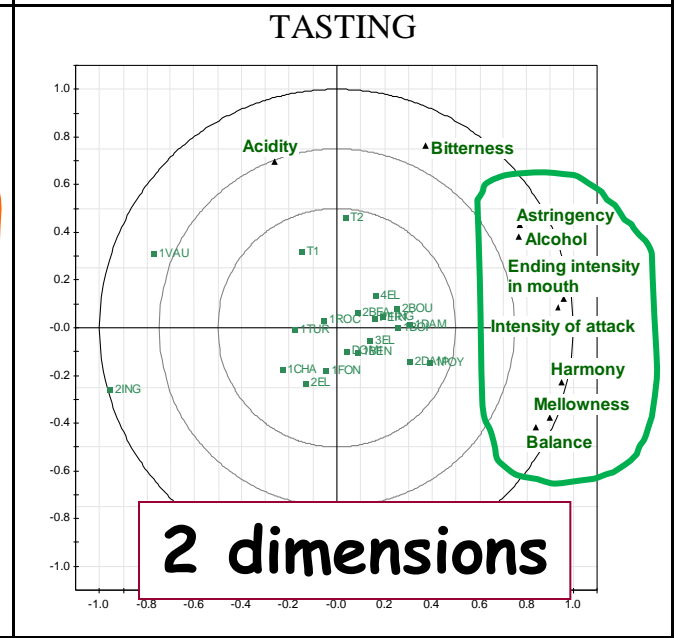
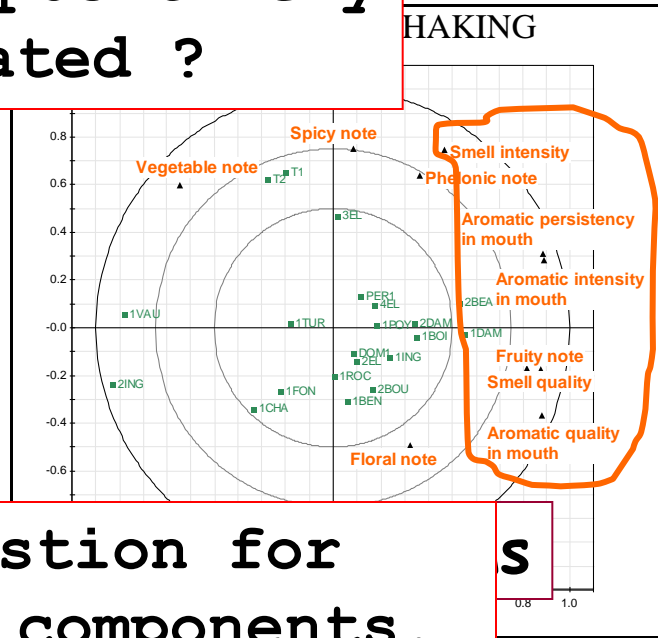
X_1 = Smell at rest, X_2 = View, X_3 = Smell after shaking, X_4 = Tasting

PCA of each block: Correlation loadings



2 dimensions

Are these first components positively correlated?

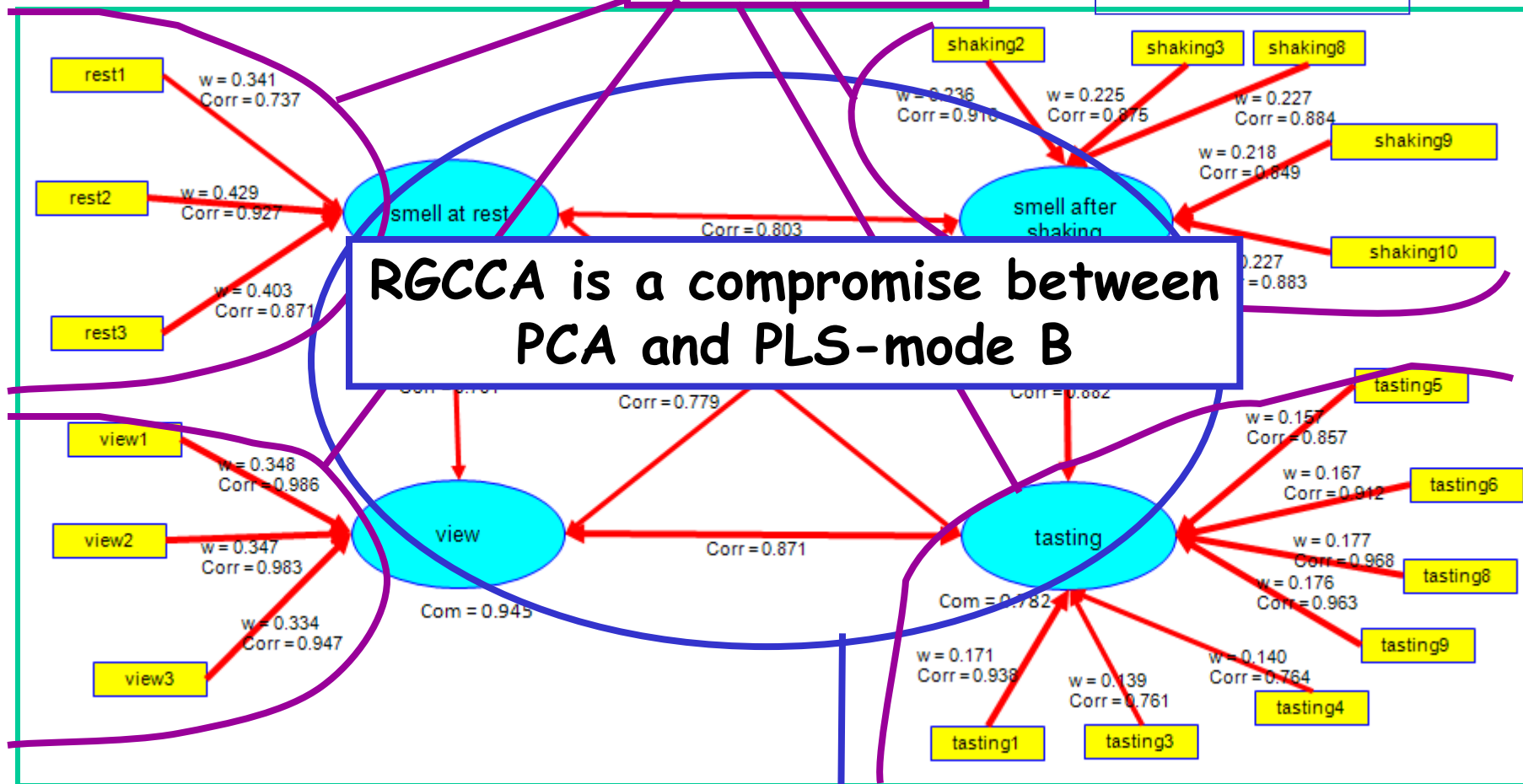


Same question for the second components.

Using XLSTAT-PSLPM / Mode PCA on variables more correlated to PC1 than to PC2

PCA optimizes the **Outer model**

Model 1



PLS-mode B optimizes the **Inner model**

Model 1 : PCA of each block

Correlations :

Latent variable	Manifest variables	Loadings	Critical ratio (CR)	Lower bound (95%)	Upper bound (95%)
smell at rest	rest1	0.737	3.946	0.206	0.923
	rest2	0.927	37.774	0.874	0.966
	rest3	0.871	17.618	0.749	0.949
view	view1	0.986	96.093	0.953	0.995
	view2	0.983	175.709	0.970	0.992
	view3	0.947	24.602	0.836	0.980
smell after shaking	shaking2	0.916	22.712	0.820	0.966
	shaking3	0.875	13.474	0.688	0.956
	shaking8	0.884	7.126	0.564	0.966
	shaking9	0.849	6.593	0.504	0.968
	shaking10	0.883	16.744	0.774	0.961
tasting	tasting1	0.938	25.860	0.822	0.977
	tasting3	0.761	4.365	0.133	0.906
	tasting4	0.764	3.272	-0.102	0.928
	tasting5	0.857	13.163	0.685	0.949
	tasting6	0.912	28.089	0.837	0.968
	tasting8	0.968	28.936	0.836	0.987
	tasting9	0.963	58.382	0.921	0.986

All loadings are significant (except one).

Model 1 : PCA of each block

Weights :					
Latent variable	Manifest variables	Outer weight	Critical ratio (CR)	Lower bound (95%)	Upper bound (95%)
smell at rest	rest1	0.341	4.633	0.117	0.387
	rest2	0.429	8.733	0.359	0.538
	rest3	0.403	9.149	0.347	0.512
view	view1	0.348	34.613	0.333	0.370
	view2	0.347	28.842	0.332	0.378
	view3	0.334	38.871	0.319	0.332
smell after shaking	shaking2	0.236	8.915	0.207	0.310
	shaking3	0.225	10.861	0.197	0.271
	shaking8	0.227	9.530	0.200	0.256
	shaking9	0.218	7.112	0.162	0.258
	shaking10	0.227	11.129	0.202	0.289
tasting	tasting1	0.171	10.123	0.150	0.215
	tasting3	0.139	5.973	0.059	0.149
	tasting4	0.140	3.916	0.009	0.155
	tasting5	0.157	11.121	0.141	0.198
	tasting6	0.167	8.810	0.149	0.224
	tasting8	0.177	9.757	0.153	0.223
	tasting9	0.176	7.937	0.152	0.242

All weights are significant.

Multi-Block Analysis is a factor analysis of tables :

$$X_1 = F_{11}p_{11}^t + \dots + \mathbf{F}_{1h}p_{1h}^t + \dots + F_{1m_1}p_{1m_1}^t + E_1$$

$$X_j = \mathbf{F}_{j1}p_{j1}^t + \dots + \mathbf{F}_{jh}p_{jh}^t + \dots + \mathbf{F}_{jm_j}p_{jm_j}^t + E_j$$

PLS-Mode B:
 $\mathbf{F}_{1h}, \dots, \mathbf{F}_{Jh}$
 optimize the
 inner model.

PCA: $\mathbf{F}_{j1}, \dots, \mathbf{F}_{jm_j}$
 optimize the
 outer model.

sub

(1) Factors (LV Scores Components) F_{1h}, \dots, F_{Jh} are well explained by $X_j a_{jm_j}$

**RGCCA gives a compromise
 between these two objectives.**

and/or

(2) Same order factors F_{1h}, \dots, F_{Jh} are well (*positively*) correlated
 (*to improve interpretation*).

PLS-mode B and RGCCA for Multi-Block data Analysis

- **Inner model:** connections between LV's
- **Outer model:** connections between MV's and their LV's.
- **Maximizing correlations for inner model:**
PLS-mode B (H. Wold, 1982 and Hanafi, 2007). *But, for each block, more observations than variables are needed.*
- **Maximizing correlations for inner model and explained variances for outer model:** *Regularized Generalized Canonical Correlation Analysis* (A. & M. Tenenhaus, 2011). *No constraints on block dimensions when the “shrinkage constants” are positive.*
- *PLS-mode B* is a special case of *RGCCA*.

PLS-mode B

$$\text{Maximize}_{a_1, \dots, a_J} \sum_{j, k=1, j \neq k}^J c_{jk} g(\text{Cor}(X_j a_j, X_k a_k))$$

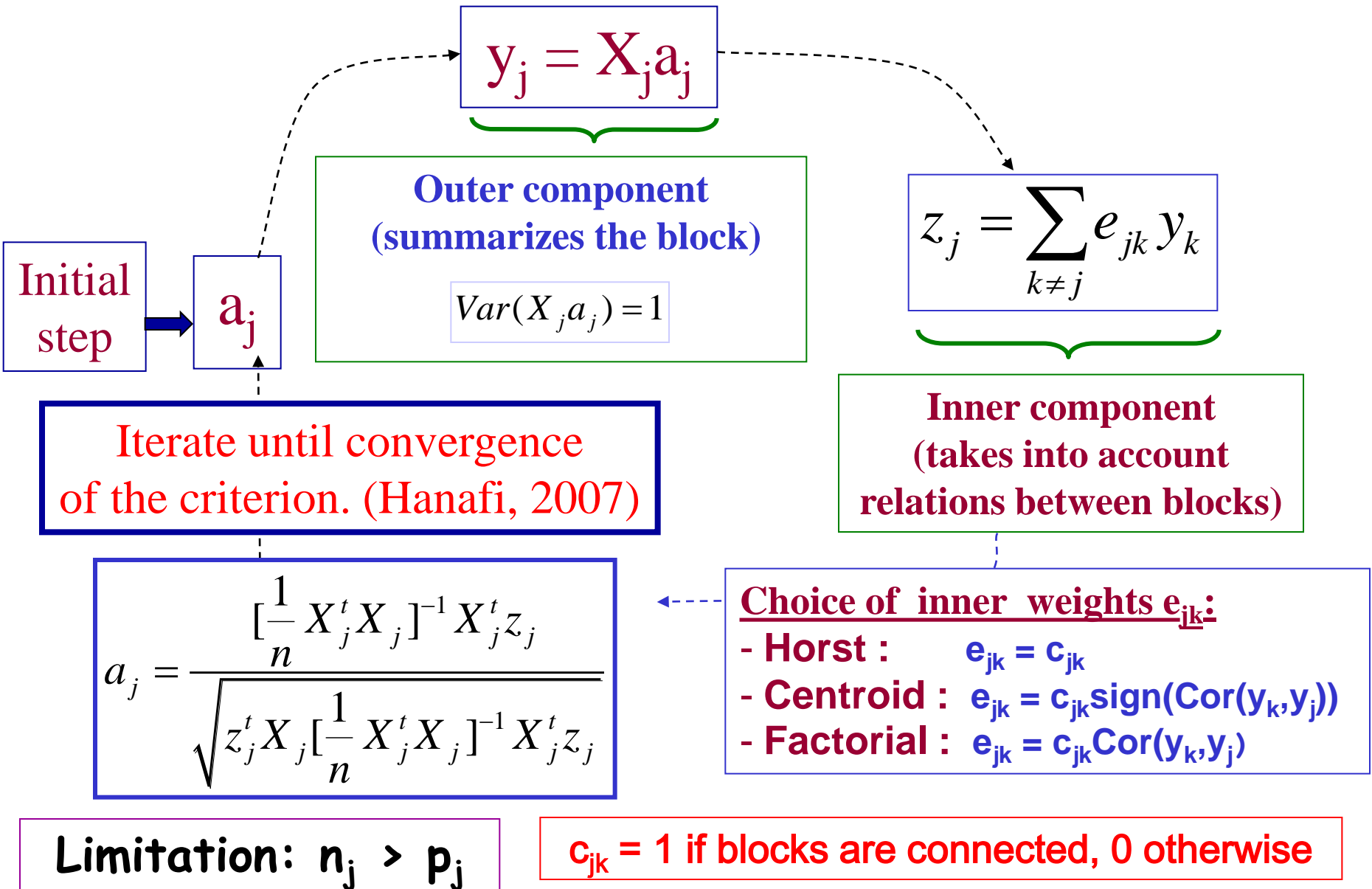
subject to the constraints $\text{Var}(X_j a_j) = 1, j = 1, \dots, J$

$$\text{where: } c_{jk} = \begin{cases} 1 & \text{if } X_j \text{ and } X_k \text{ are connected} \\ 0 & \text{otherwise} \end{cases}$$

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

H. Wold (1982) has described a monotone convergent algorithm related to this optimization problem. (Proof by Hanafi in 2007.)

Wold's algorithm: PLS-Mode B



Initial step

$$a_j$$

$$y_j = X_j a_j$$

Outer component
(summarizes the block)

$$Var(X_j a_j) = 1$$

$$z_j = \sum_{k \neq j} e_{jk} y_k$$

Inner component
(takes into account relations between blocks)

Iterate until convergence of the criterion. (Hanafi, 2007)

$$a_j = \frac{[\frac{1}{n} X_j^t X_j]^{-1} X_j^t z_j}{\sqrt{z_j^t X_j [\frac{1}{n} X_j^t X_j]^{-1} X_j^t z_j}}$$

- Choice of inner weights e_{jk} :**
- **Horst :** $e_{jk} = c_{jk}$
 - **Centroid :** $e_{jk} = c_{jk} \text{sign}(\text{Cor}(y_k, y_j))$
 - **Factorial :** $e_{jk} = c_{jk} \text{Cor}(y_k, y_j)$

Limitation: $n_j > p_j$

$c_{jk} = 1$ if blocks are connected, 0 otherwise

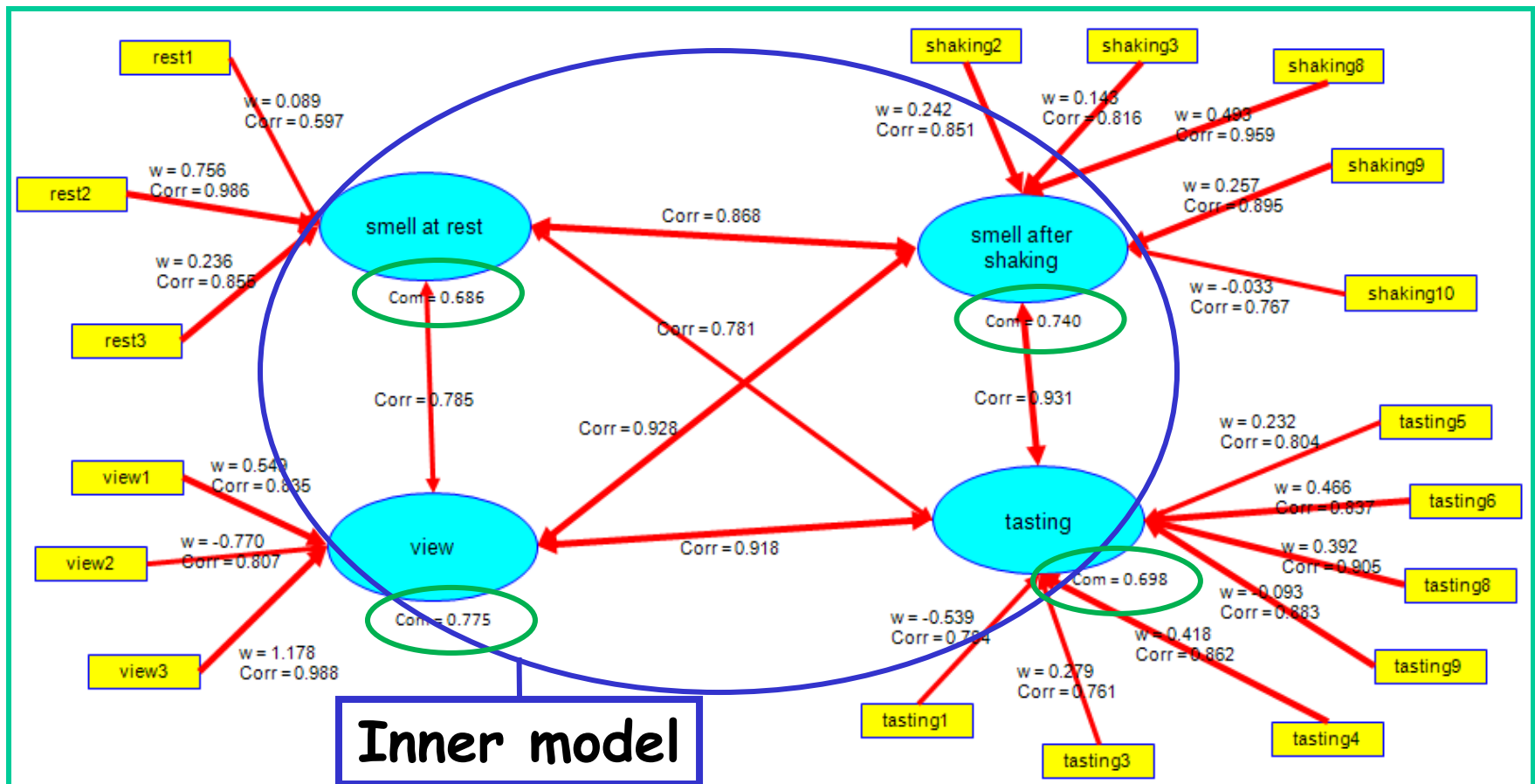
Optimizing the inner model (with XLSTAT)

PLS-mode B, *Centroid scheme*

$$\Leftrightarrow \text{Maximize } \sum_{j \neq k} | \text{cor}(X_j a_j, X_k a_k) |$$

$$p_j < n_j$$

Model 2



For outer model : Communalities for block $X_h = \text{Average}(\text{Cor}^2(x_{hj}, F_h))$

Optimizing the inner model (with XLSTAT)

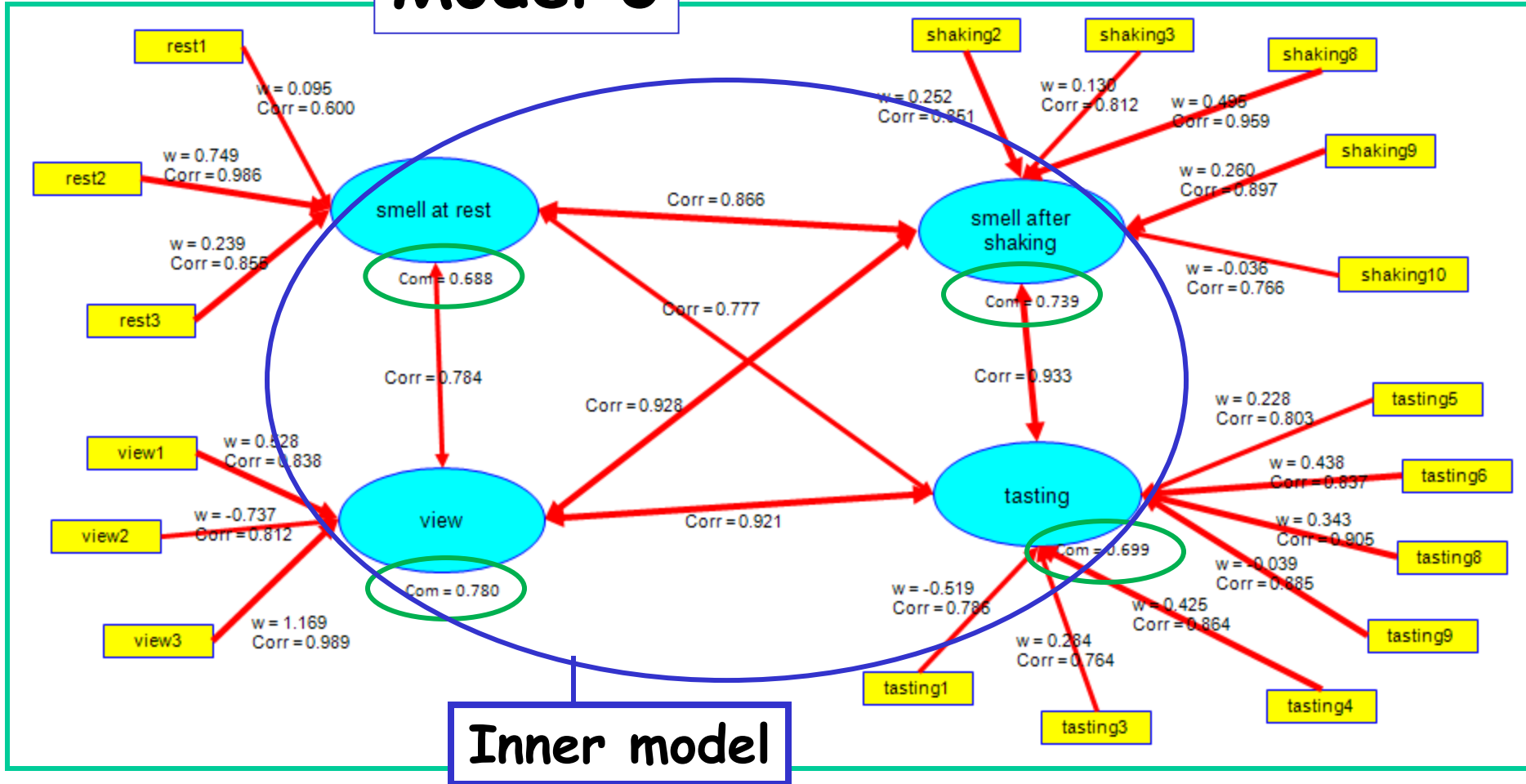
$$p_j < n_j$$

Mode B, *Factoriel*

\Leftrightarrow

$$\text{Maximize } \sum_{j \neq k} \text{cor}^2(X_j a_j, X_k a_k)$$

Model 3



For outer model : Communality for block $X_h = \text{Average}(\text{Cor}^2(x_{hj}, F_h))$

Model 3

Correlations :					
Latent variable	Manifest variables	Loadings	Critical ratio (CR)	Lower bound (95%)	Upper bound (95%)
smell at rest	<i>rest1</i>	<i>0.600</i>	<i>2.333</i>	<i>-0.092</i>	<i>0.934</i>
	rest2	0.986	11.917	0.731	0.998
	rest3	0.855	7.804	0.605	0.964
view	view1	0.838	4.390	0.221	0.977
	view2	0.812	4.423	0.216	0.958
	view3	0.989	6.795	0.710	0.999
smell after shaking	shaking2	0.851	7.165	0.516	0.952
	shaking3	0.812	6.677	0.531	0.940
	shaking8	0.959	9.835	0.674	0.988
	shaking9	0.897	7.298	0.512	0.978
	shaking10	0.766	4.497	0.308	0.947
tasting	tasting1	0.786	4.121	0.163	0.932
	<i>tasting3</i>	<i>0.764</i>	<i>3.048</i>	<i>-0.069</i>	<i>0.950</i>
	tasting4	0.864	6.824	0.486	0.952
	tasting5	0.803	3.931	0.212	0.972
	tasting6	0.837	4.890	0.326	0.962
	tasting8	0.905	6.864	0.439	0.968
	tasting9	0.885	5.905	0.404	0.967

Model 3

Weights :					
Latent variable	Manifest variables	Outer weight	Critical ratio (CR)	Lower bound (95%)	Upper bound (95%)
smell at rest	<i>rest1</i>	0.095	0.317	-0.484	0.663
	rest2	0.749	2.607	0.065	1.223
	<i>rest3</i>	0.239	0.806	-0.442	0.742
view	<i>view1</i>	0.528	0.658	-0.796	2.420
	<i>view2</i>	-0.737	-0.935	-2.465	0.642
	view3	1.169	3.127	0.380	1.636
smell after shaking	<i>shaking2</i>	0.252	1.146	-0.242	0.684
	<i>shaking3</i>	0.130	0.475	-0.405	0.680
	<i>shaking8</i>	0.495	1.912	-0.017	1.020
	<i>shaking9</i>	0.260	0.881	-0.360	0.763
	<i>shaking10</i>	-0.036	-0.155	-0.472	0.444
tasting	<i>tasting1</i>	-0.519	-0.754	-1.257	1.740
	<i>tasting3</i>	0.284	0.622	-0.971	0.875
	<i>tasting4</i>	0.425	1.077	-0.268	1.371
	<i>tasting5</i>	0.228	0.344	-1.148	1.745
	<i>tasting6</i>	0.438	0.519	-1.809	1.778
	<i>tasting8</i>	0.343	0.451	-1.315	1.952
	<i>tasting9</i>	-0.039	-0.045	-1.880	1.917

Conclusion

- Many weights are not significant !!!
- If you want the *butter* (good correlations for the inner and outer models) and the money of the butter (significant weights) , you must switch to Regularized Generalized Canonical Correlation Analysis (RGCCA).

Regularized generalized CCA

$$\text{Maximize}_{a_1, \dots, a_J} \sum_{j, k=1, j \neq k}^J c_{jk} g(\text{Cov}(X_j a_j, X_k a_k))$$

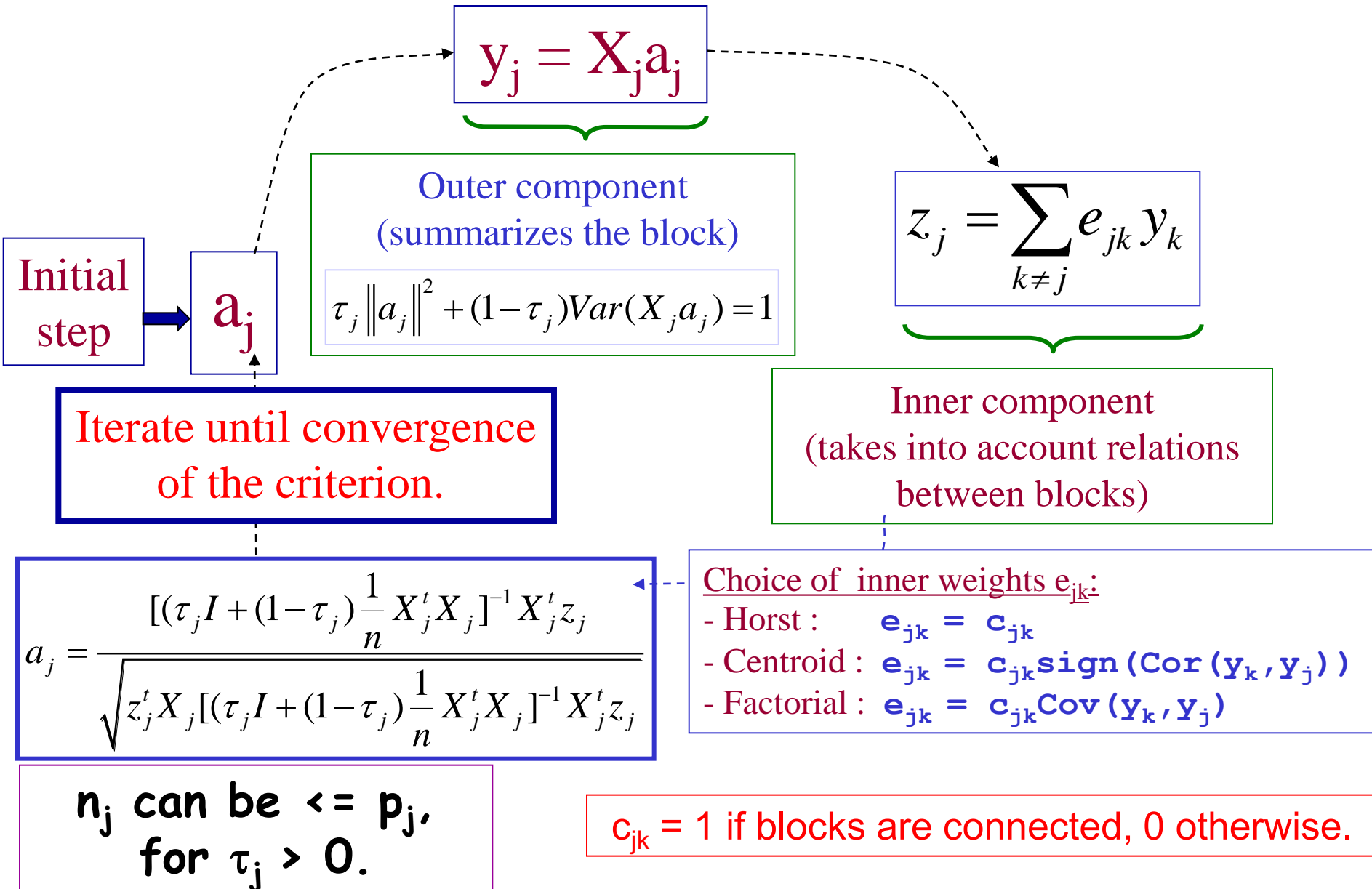
subject to the constraints $\tau_j \|a_j\|^2 + (1 - \tau_j) \text{Var}(X_j a_j) = 1, j = 1, \dots, J$

A monotone convergent algorithm related to this optimization problem is proposed (A. & M. Tenenhaus, 2011).

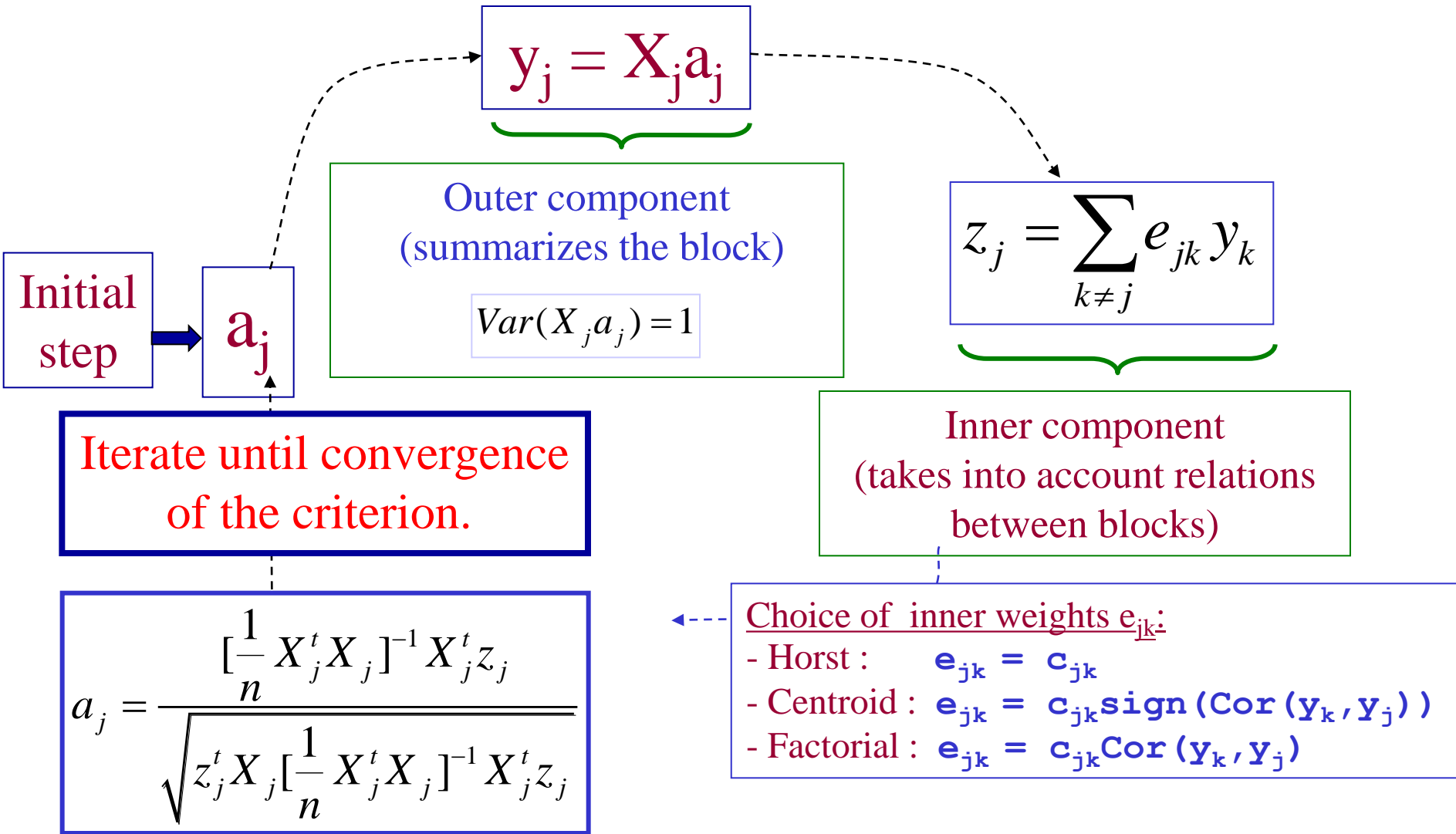
$$g = \begin{cases} \text{identity} & (\text{Horst scheme}) \\ \text{square} & (\text{Factorial scheme}) \\ \text{absolute value} & (\text{Centroid scheme}) \end{cases}$$

and: $\tau_j =$ Shrinkage constant between 0 and 1

The PLS algorithm for RGCCA

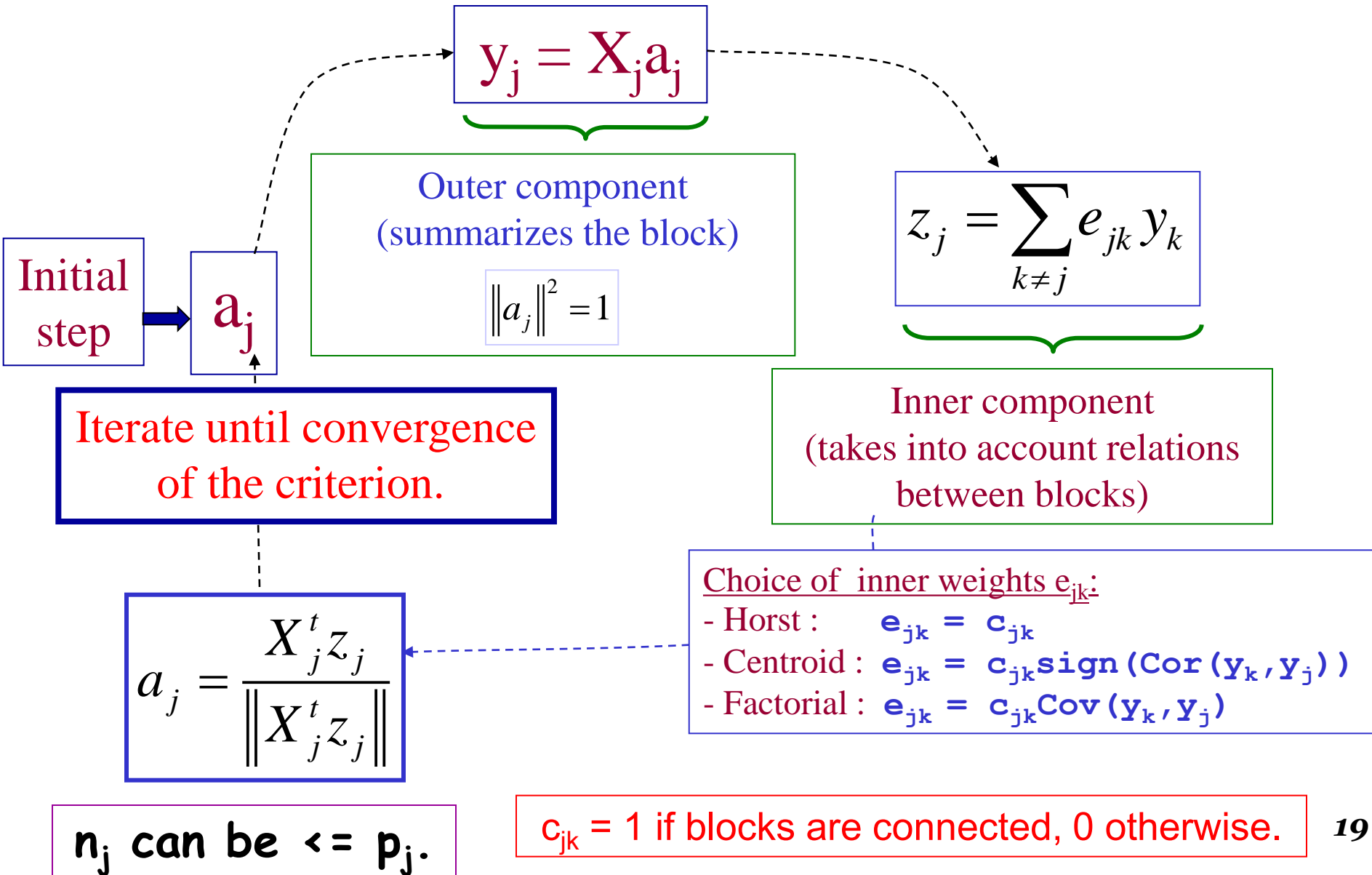


All $\tau_j = 0$, RGCCA = PLS-Mode B



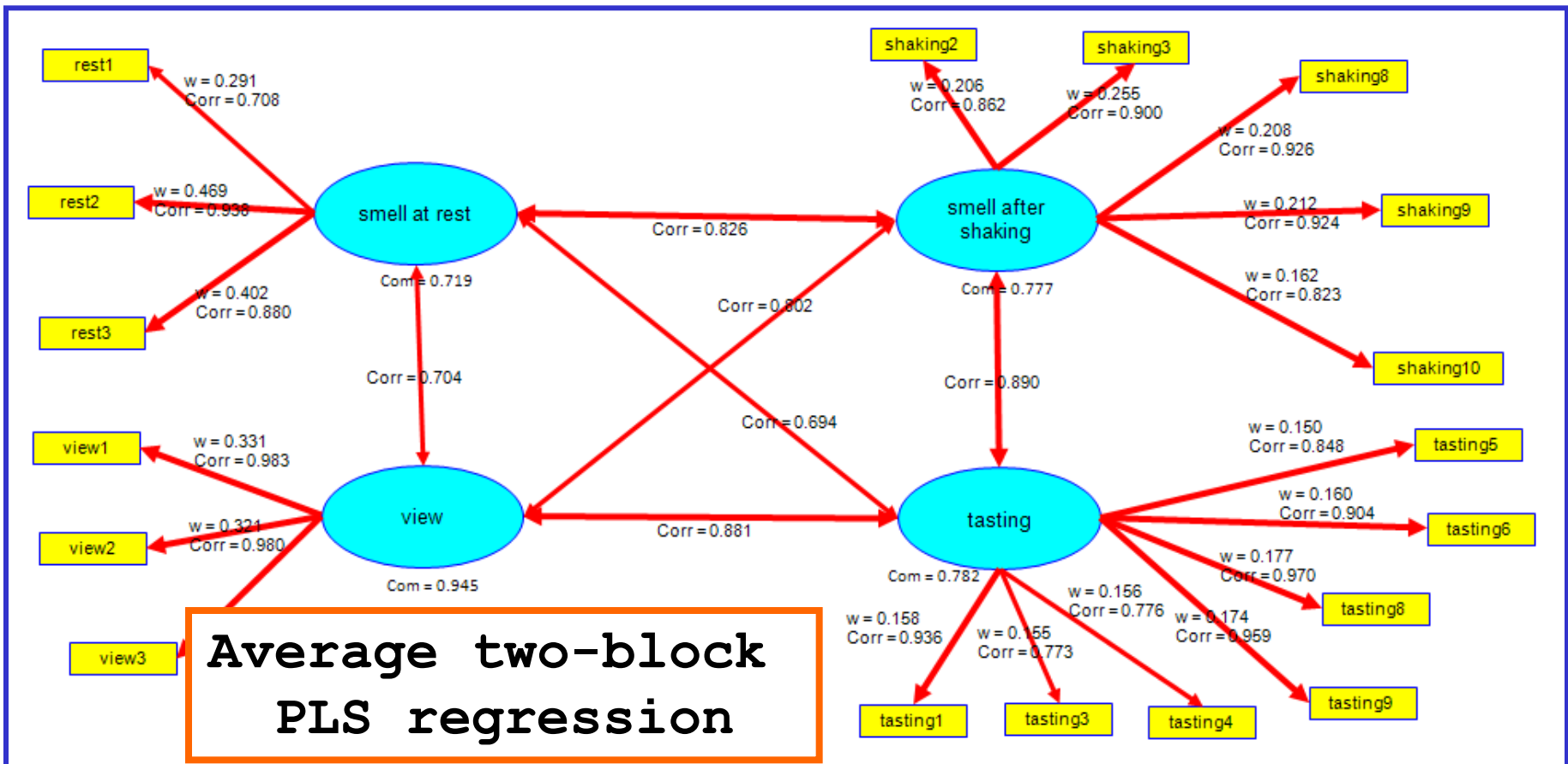
$c_{jk} = 1$ if blocks are connected, 0 otherwise.

All $\tau_j = 1$, RGCCA - Mode A



Model 4 : RGCCA, factorial scheme, mode A

$$\underbrace{\text{Maximize}}_{\|a_j\|=1, \forall j} \sum_{j \neq k} \text{cov}^2(X_j a_j, X_k a_k)$$



Latent variables have been afterwards standardized.

Model 4

Correlations :					
Latent variable	Manifest variables	Loadings	Critical ratio (CR)	Lower bound (95%)	Upper bound (95%)
smell at rest	rest1	0.708	3.301	0.113	0.924
	rest2	0.938	31.414	0.880	0.970
	rest3	0.880	20.537	0.776	0.954
view	view1	0.983	21.303	0.898	0.995
	view2	0.980	28.635	0.937	0.993
	view3	0.954	34.009	0.883	0.982
smell after shaking	shaking2	0.904	22.231	0.810	0.965
	shaking3	0.862	11.238	0.669	0.957
	shaking8	0.900	14.572	0.753	0.972
	shaking9	0.869	9.636	0.662	0.972
	shaking10	0.870	14.582	0.715	0.964
tasting	tasting1	0.936	14.373	0.693	0.978
	tasting3	0.773	5.563	0.423	0.911
	tasting4	0.776	4.980	0.361	0.930
	tasting5	0.848	9.118	0.582	0.958
	tasting6	0.904	17.424	0.767	0.972
	tasting8	0.970	30.511	0.879	0.989
	tasting9	0.959	31.991	0.864	0.986

All loadings are significant.

Model 4

Weights :					
Latent variable	Manifest variables	Outer weight	Critical ratio (CR)	Lower bound (95%)	Upper bound (95%)
smell at rest	rest1	0.291	3.077	0.041	0.381
	rest2	0.469	8.684	0.370	0.576
	rest3	0.402	8.266	0.321	0.520
view	view1	0.331	11.009	0.275	0.338
	view2	0.321	13.820	0.277	0.330
	view3	0.378	6.710	0.342	0.502
smell after shaking	shaking2	0.217	10.731	0.193	0.269
	shaking3	0.206	11.732	0.161	0.234
	shaking8	0.255	10.763	0.213	0.302
	shaking9	0.255	9.538	0.211	0.312
	shaking10	0.200	8.899	0.159	0.251
tasting	tasting1	0.158	10.451	0.126	0.180
	tasting3	0.155	4.659	0.107	0.223
	tasting4	0.156	6.010	0.119	0.225
	tasting5	0.150	6.293	0.105	0.204
	tasting6	0.160	8.660	0.140	0.215
	tasting8	0.177	6.874	0.151	0.257
	tasting9	0.174	8.554	0.152	0.230

All weights are also significant.

Model Comparison

		AVE(outer model)	AVE(inner model)
Using PCA		.79785	.62407
Using RGCCA (factorial scheme)	All tau = 0 (mode B)	.72179	.75817
	All tau = 0.2	.77630	.71860
	All tau = 0.4	.78785	.69217
	All tau = 0.6	.79320	.67285
	All tau = 0.8	.79588	.65745
	All tau = 1 (mode A)	.79615	.64456
	Optimal tau : (Schäfer & Strimmer, 2005) tau1 = .19, tau2 = .16 tau3 = .17, tau4 = .26	.77615	.72022

R-code
(Arthur T.)

New package RGCCA with initial version 1.0

Title: Regularized Generalized Canonical Correlation Analysis

Version: 1.0

Date: 2010-06-08

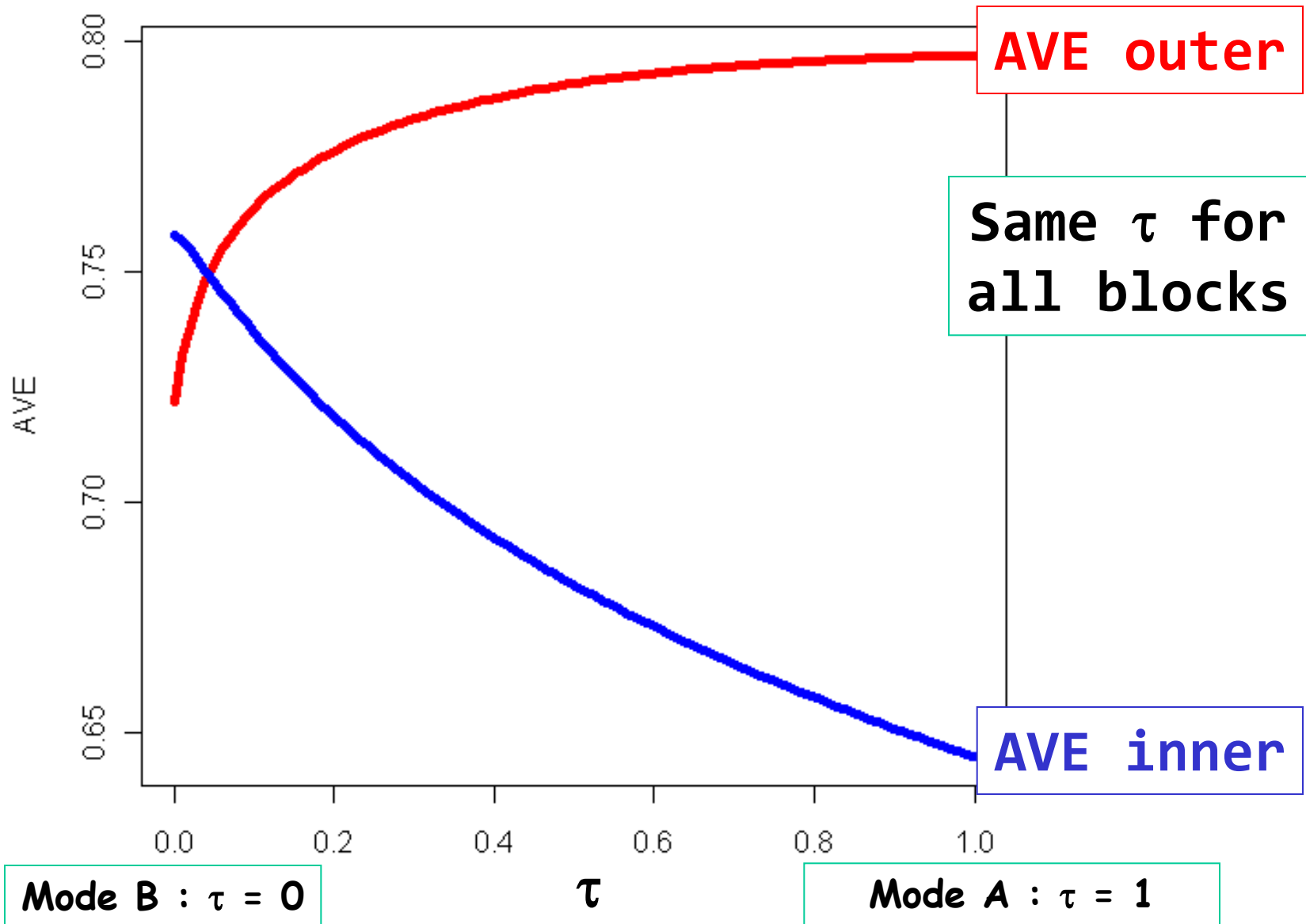
Author: Arthur Tenenhaus

Repository: CRAN

Date/Publication: 2010-10-15 14:58:02

[More information about RGCCA at CRAN](#)

Path: [/cran/new](#) | [permanent link](#)



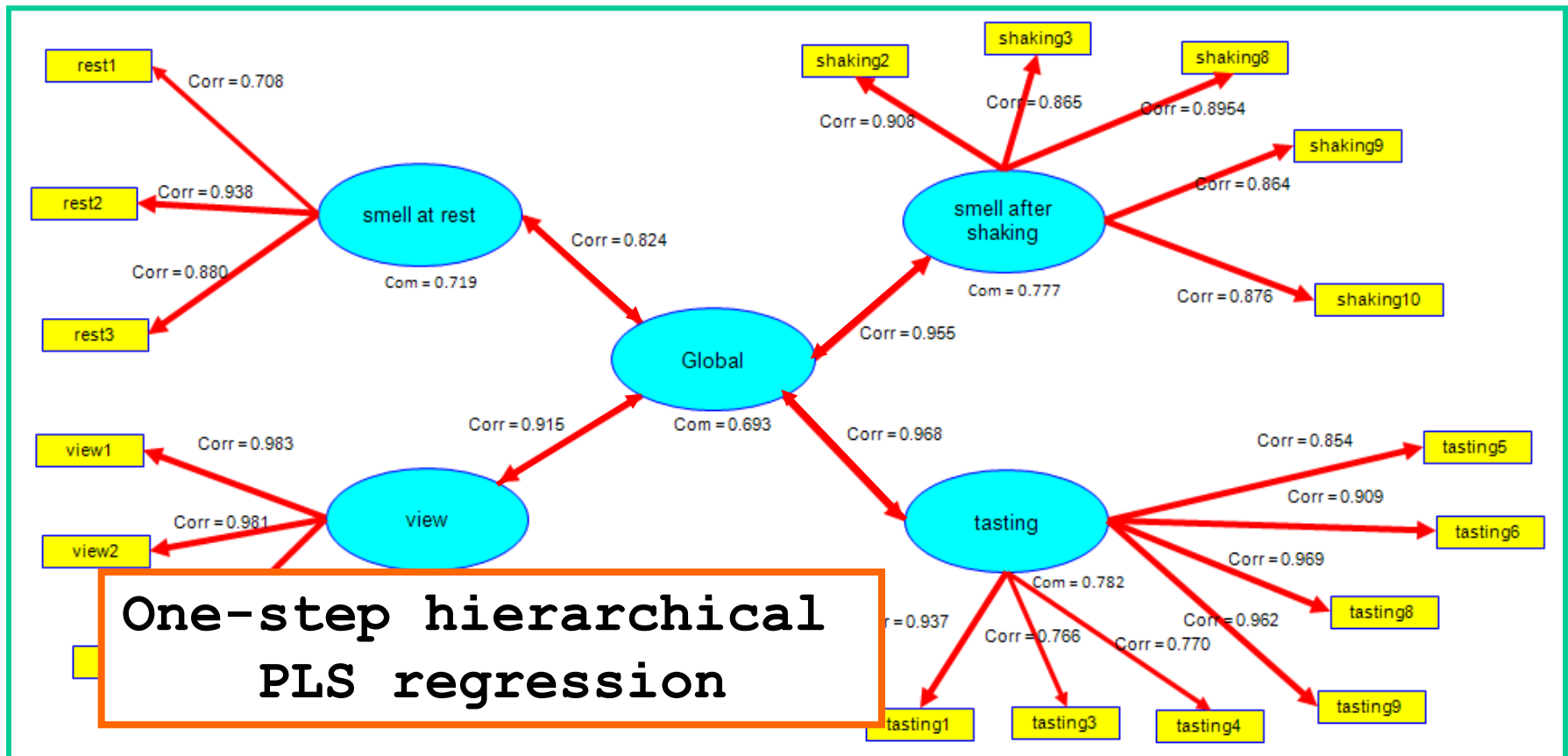
Mode A favors the outer model.
Mode B favors the inner model.

Hierarchical model for wine data: Model 5

Dimension 1

RGCCA:
Factorial,
Mode A

$$\underbrace{\text{Maximize}}_{\|a_j\|=1, \forall j} \sum_{j \neq 5} \text{cov}^2(X_j a_j, X_5 a_5)$$



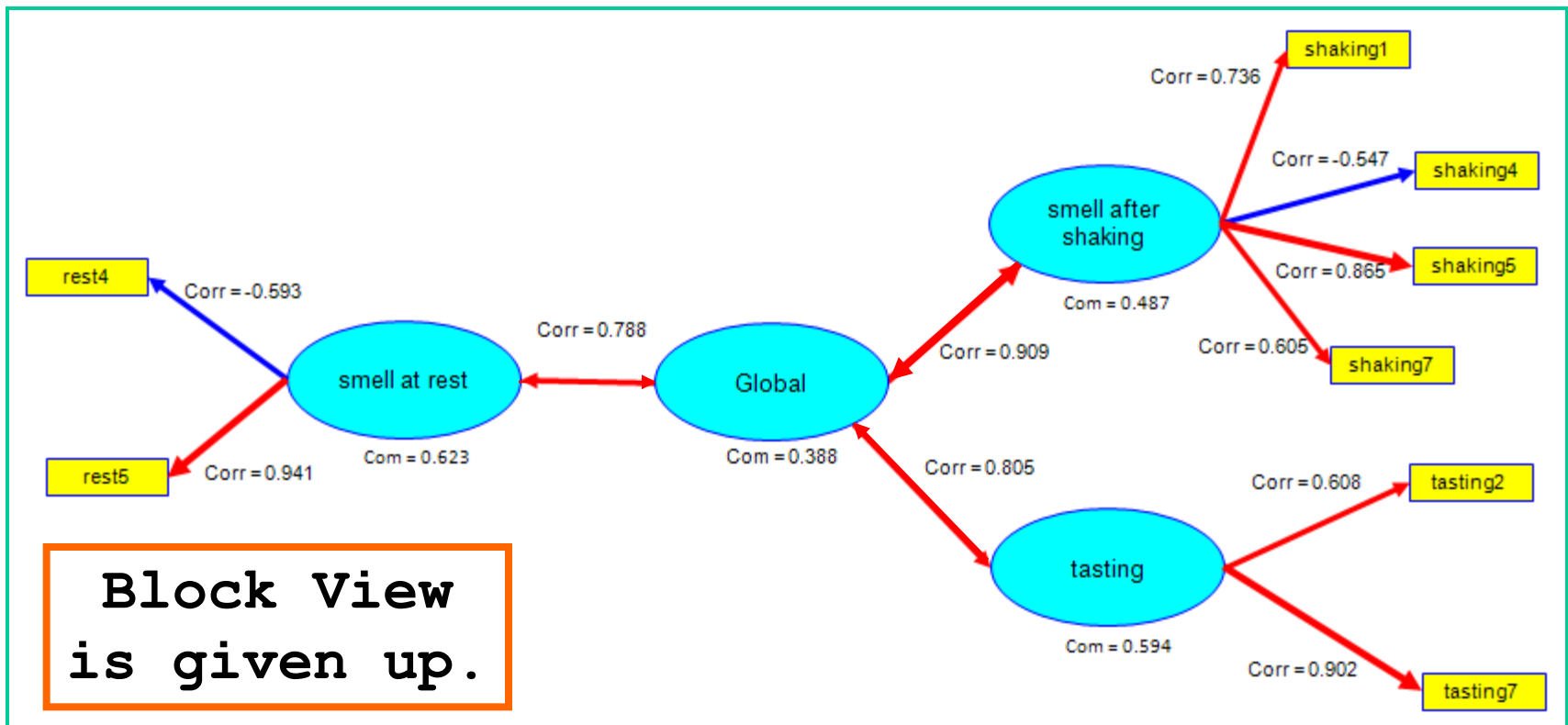
2nd order block "Global" contains all the MV's of the 1st order blocks

Hierarchical model for wine data: Model 6

Dimension 2

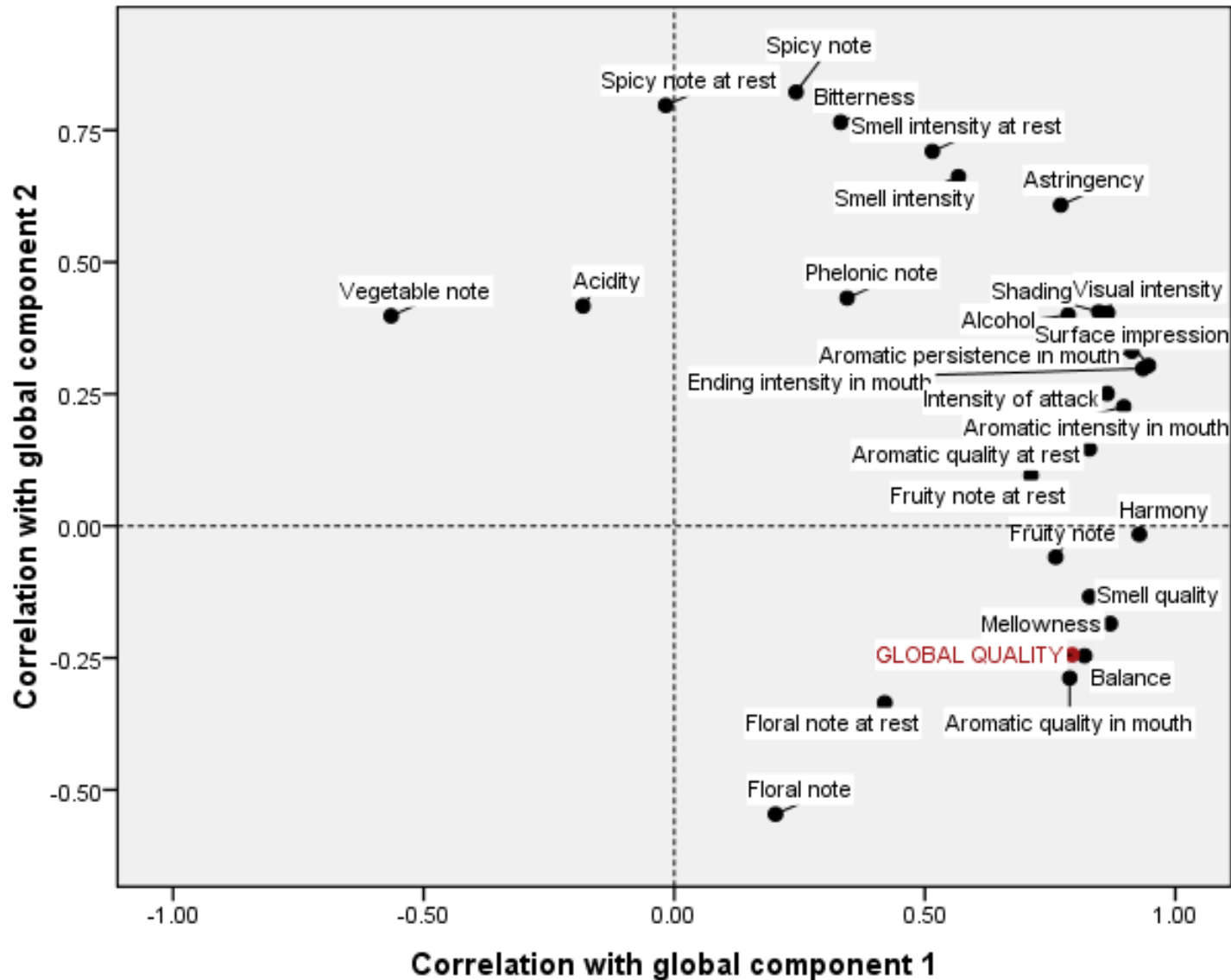
*RGCCA,
Factorial,
Mode A*

$$\underset{\|a_j\|=1, \forall j}{\text{Maximize}} \sum_{j \neq 4} \text{cov}^2(X_j a_j, X_4 a_4)$$



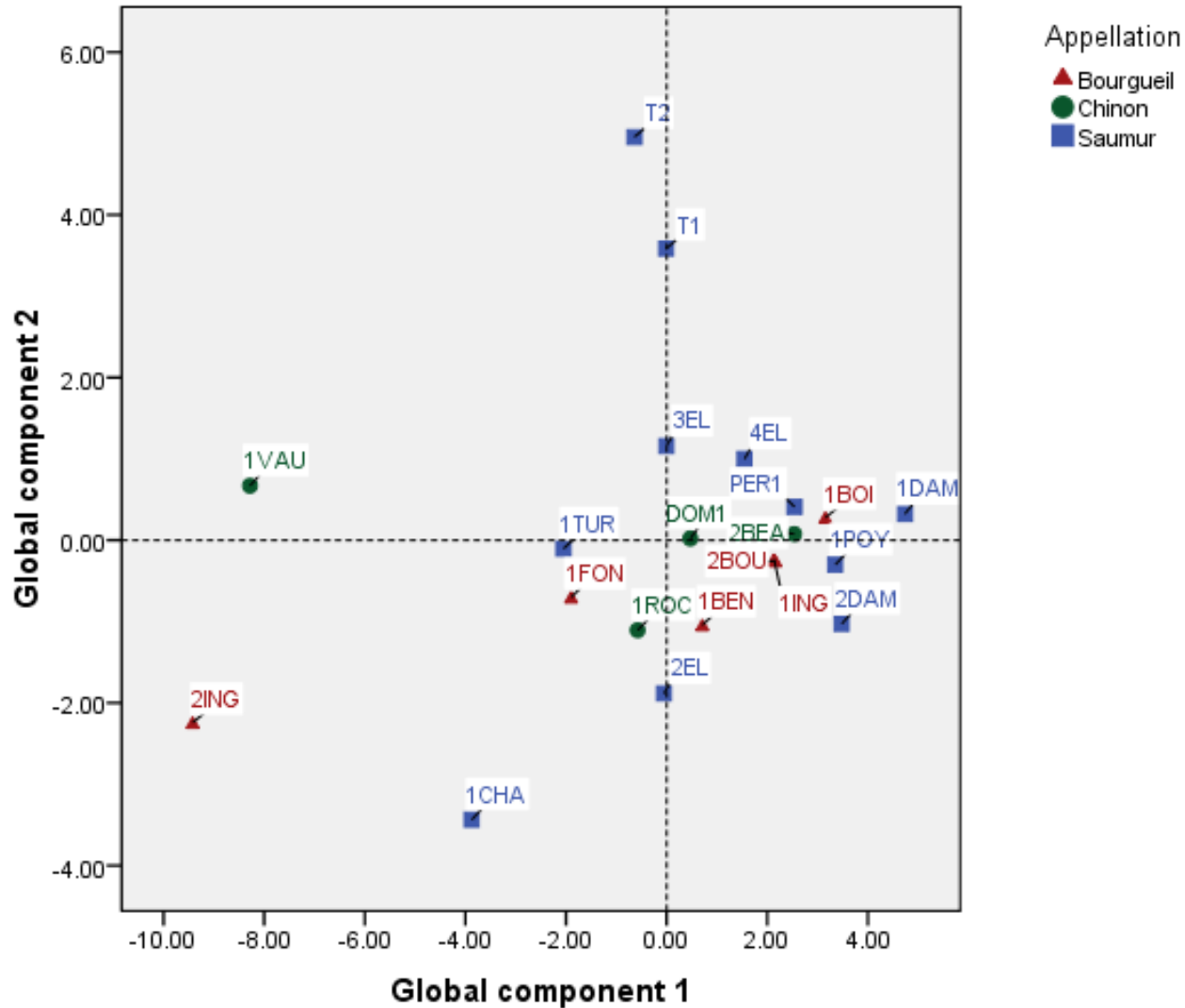
2nd order block "Global" contains all the MV's of the 1st order blocks

Mapping of the correlations with the global components



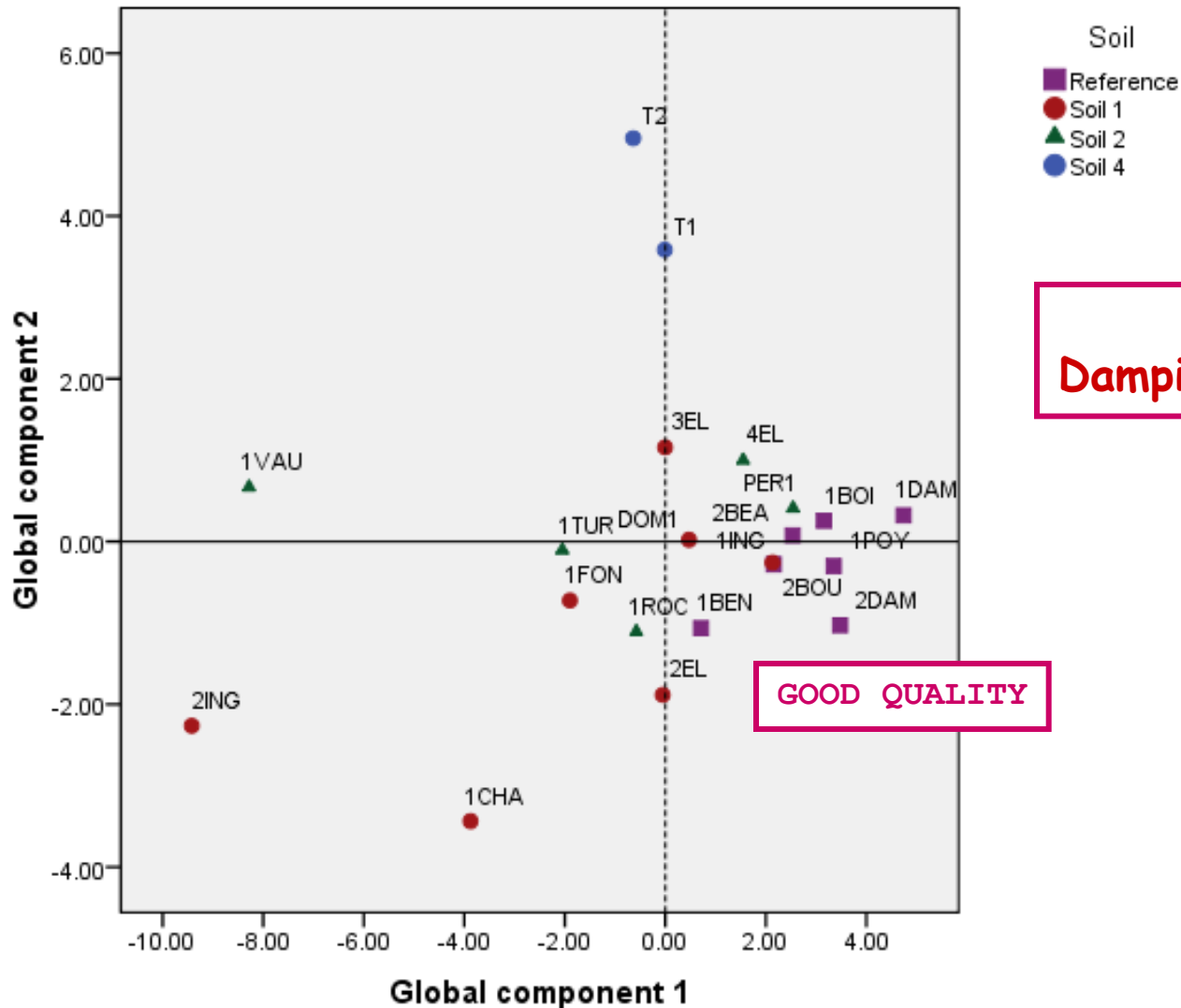
Wine visualization in the global component space

Wines marked by Appellation



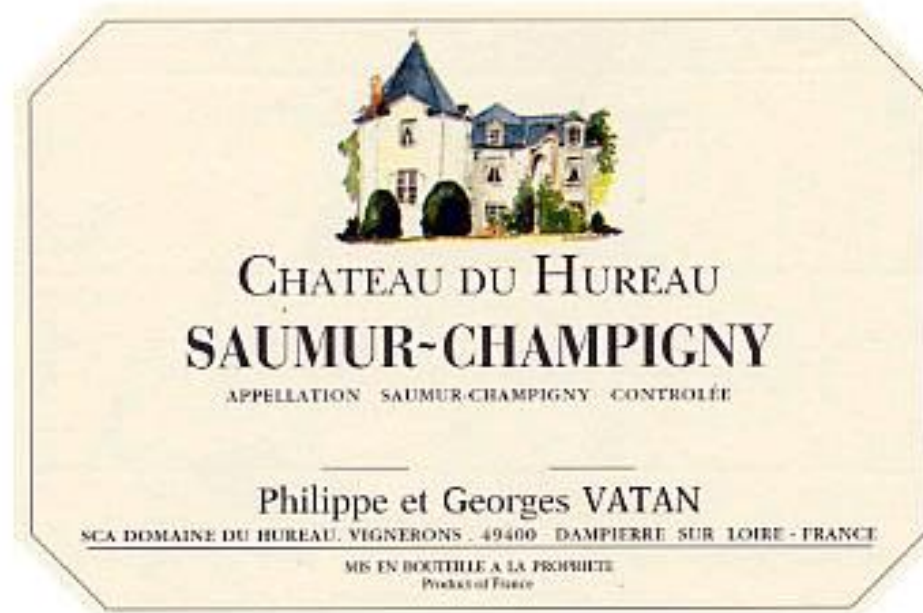
Wine visualization in the global component space

Wines marked by Soil



**DAM =
Dampierre-sur-Loire**

GOOD QUALITY



Cuvée Lisagathe 1995


A soft, warm, blackberry nose. A good core of fruit on the palate with quite well worked tannin and acidity on the finish; Good length and a lot of potential.

DECANTER (mai 1997)

(DECANTER AWARD ***** : Outstanding quality, a virtually perfect example)

References

Available online at www.sciencedirect.com

 **SCIENCE @ DIRECT®**

**COMPUTATIONAL
STATISTICS
& DATA ANALYSIS**

ELSEVIER Computational Statistics & Data Analysis 48 (2005) 159–205
www.elsevier.com/locate/cstda

PLS path modeling

Michel Tenenhaus^a, Vincenzo Esposito Vinzi^{a,b,*},
Yves-Marie Chatelin^c, Carlo Lauro^b

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

PSYCHOMETRIKA—VOL. 76, NO. 2, 257–284
APRIL 2011
DOI: 10.1007/s11336-011-9206-8

REGULARIZED GENERALIZED CANONICAL CORRELATION ANALYSIS

ARTHUR TENENHAUS
SUPELEC, GIF-SUR-YVETTE

MICHEL TENENHAUS
HEC PARIS, JOUY-EN-JOSAS

Final conclusion



All the proofs of a pudding are in the eating, but it will taste even better if you know the cooking.