A criterion-based PLS approach to SEM

Michel Tenenhaus (HEC Paris)

Arthur Tenenhaus (SUPELEC)

TRICAP 2009

ThRee-way methods In Chemistry And Psychology

Vall de Núria - Spain

Economic inequality and political instability Data from Russett (1964), in GIFI

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 labor force employed in agriculture

Political instability

INST : Instability of executive (45-61)

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

DEAT : Nb of people killed as a result of civic group violence (50-62)

D-STAB : Stable democracy

D-UNST : Unstable democracy

DICT : Dictatorship

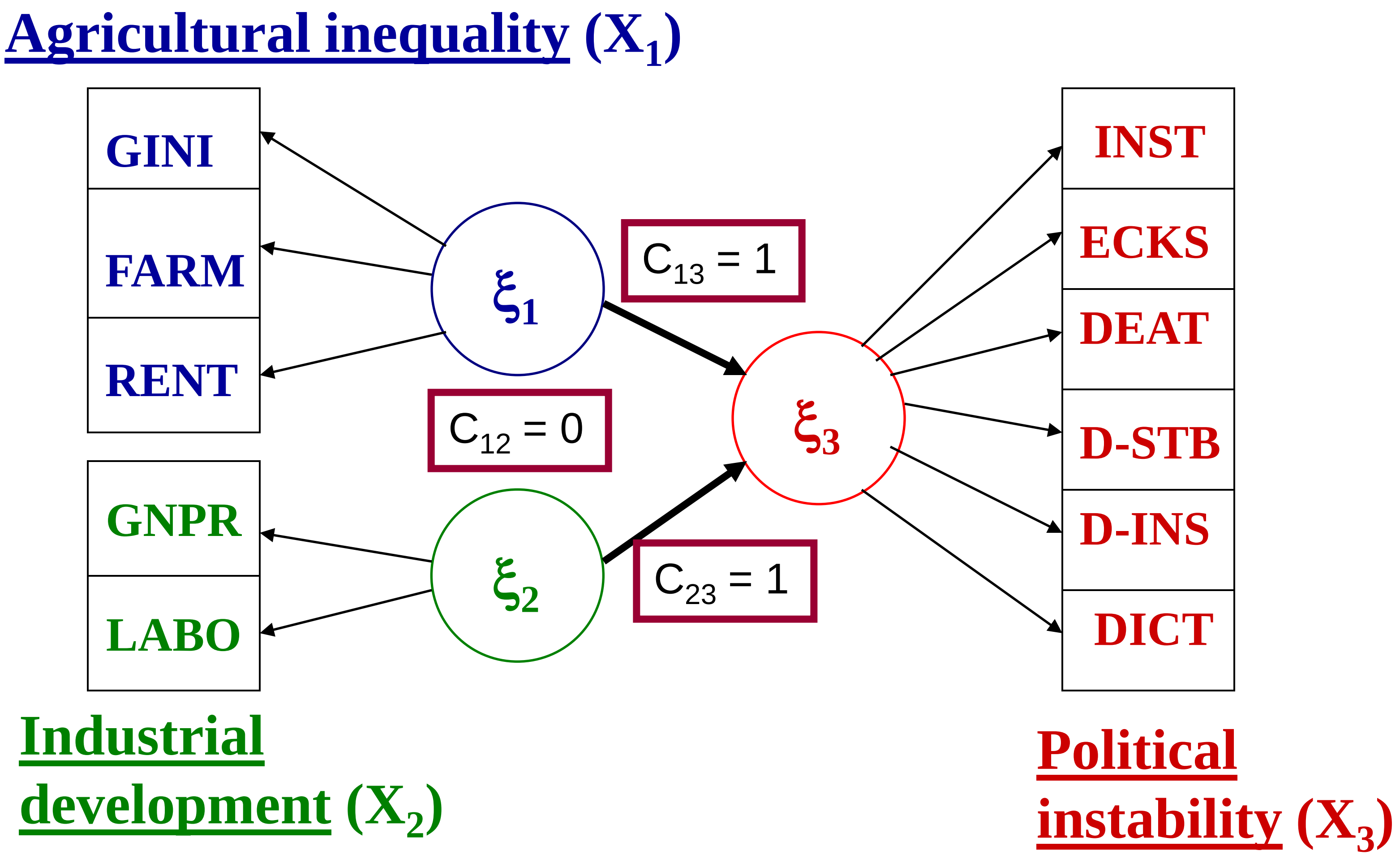
Economic inequality and political instability

(Data from Russett, 1964)

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

1 = Stable democracy
2 = Unstable democracy
3 = Dictatorship

A SEM model



Agricultural inequality (X_1)
GINI
FARM
RENT
GNPR
LABO
ξ_1
ξ_2
ξ_3
$C_{13} = 1$
$C_{12} = 0$
$C_{23} = 1$
INST
ECKS
DEAT
D-STB
D-INS
DICT
Industrial development (X_2)
Political instability (X_3)

Latent Variable outer estimation

$$Y_1 = X_1 w_1 = w_{11} GINI + w_{12} FARM + w_{13} RENT$$

$$Y_2 = X_2 w_2 = w_{21} GNPR + w_{22} LABO$$

$$\begin{aligned} Y_3 = X_3 w_3 = & \, w_{31} INST + w_{32} ECKS + w_{33} DEATH \\ & + w_{34} D\text{-}STB + w_{35} D\text{-}UNST \\ & + w_{36} DICT \end{aligned}$$

Some modified multi-block methods for SEM

c_{jk} = 1 if blocks are linked, 0 otherwise and c_{jj} = 0

SUMCOR (Horst, 1961)	$Max \sum_{j,k} c_{jk} Cor(X_j w_j, X_k w_k)$
S	$\sum_{j,k}$

GENERALIZED CANONICAL CORRELATION ANALYSIS

SABSCOR (Mathes, 1993, Hanafi, 2004)	$Max \sum_{j,k} c_{jk} \mid Cor(X_j w_j, X_k w_k) \mid$
MAXDIFF (Van de Geer, 1984) [SUMCOV]	$\underset{\text{All } \lVert w_j \rVert = 1}{Max} \sum_{j,k} c_{jk} Cov(X_j w_j, X_k w_k)$
MAXDIFF B (H [SSQCOV]	$\underset{\text{All } \lVert w_j \rVert = 1}{} \sum_{j,k} (X_j w_j, X_k w_k)$
SABSCOV (Krämer, 2007)	$\underset{\text{All } \lVert w_j \rVert = 1}{Max} \sum_{j,k} c_{jk} \left\lvert Cov(X_j w_j, X_k w_k) \right\rvert$

GENERALIZED PLS REGRESSION

Covariance-based criteria for SEM

c_{jk} = 1 if blocks are linked, 0 otherwise and c_{jj} = 0

SUMCOR-PLSPM	$\underset{\text{All } Var(X_j w_j)=1}{Max} \sum_{j,k} c_{jk} Cov(X_j w_j, X_k w_k)$
SSQCOR-PLSPM	$\underset{\text{All } Var(X_j w_j)=1}{Max} \sum_{j,k} c_{jk} Cov^2(X_j w_j, X_k w_k)$
SABSCOR-PLSPM	$\underset{\text{All } Var(X_j w_j)=1}{Max} \sum_{j,k} c_{jk} \lvert Cov(X_j w_j, X_k w_k) \rvert$
SUMCOV-PLSPM	$\underset{\text{All } \lVert w_j \rVert=1}{Max} \sum_{j,k} c_{jk} Cov(X_j w_j, X_k w_k)$
SSQCOV-PLSPM	$\underset{\text{All } \lVert w_j \rVert=1}{Max} \sum_{j,k} c_{jk} Cov^2(X_j w_j, X_k w_k)$
SABSCOV-PLSPM	$\underset{\text{All } \lVert w_j \rVert=1}{Max} \sum_{j,k} c_{jk} \lvert Cov(X_j w_j, X_k w_k) \rvert$

A continuum approach

$$Maximize \sum_{j<k} c_{jk}\, g(\mathrm{cov}(X_j w_j, X_k w_k))$$

subject to the constraints:

$$\tau_i \|w_i\|^2 + (1-\tau_i) Var(X_i w_i) = 1 \quad , \text{ with } \ 0 \le \tau_\mathrm{i} \le 1, \ i = 1, \ldots, J$$

where

$$g(x) = \begin{cases} x & \text{(Horst scheme)} \\ x^2 & \text{(Factorial scheme)} \\ |\mathrm{x}| & \text{(Centroid scheme)} \end{cases}$$

A general procedure to obtain critical points of the criteria

- Construct the Lagrangian function related to the optimization problem.
- Cancel the derivative of the Lagrangian function with respect to each w_i.
- Use the Wold's procedure to solve the stationary equations ($\neq$ Lohmöller's procedure).
- This procedure converges to a critical point of the criterion.
- The criterion increases at each step of the algorithm.

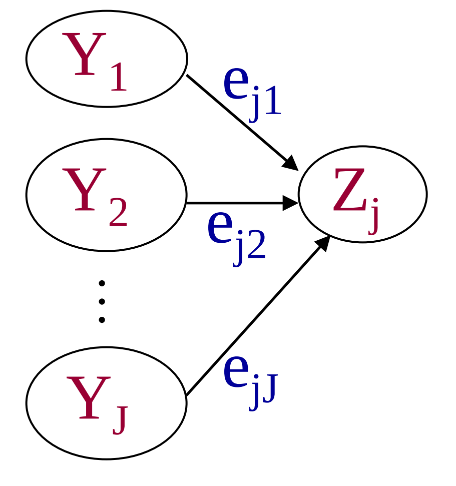
The general algorithm

$$Y_j = \boxed{\pm}\, X_j w_j$$

Initial step $\Rightarrow$ w_j

Outer Estimation

$$\tau_j \|w_j\|^2 + (1-\tau_j)Var(X_j w_j) = 1$$



Inner estimation

Iterate until convergence

$$w_j = \frac{[(\tau_j I + (1-\tau_j)\frac{1}{n}X_j'X_j]^{-1}X_j'Z_j}{\sqrt{Z_j'X_j[(\tau_j I + (1-\tau_j)\frac{1}{n}X_j'X_j]^{-1}X_j'Z_j}}$$

Choice of weights e_{jh}:

- Horst : $e_{jh} = c_{jh}$
- Centroid : $e_{jh} = c_{jh}\,\mathrm{sign}(Cor(Y_h , Y_j))$
- Factorial : $e_{jh} = c_{jh}\,Cov(Y_h , Y_j)$

c_{jh} = 1 if blocks are linked, 0 otherwise and c_{jj} = 0

Specific cases

$$\boxed{Maximize \sum_{j<k} c_{jk}\, g(\mathrm{cov}(X_j w_j, X_k w_k))}$$

subject to the constraints:

$$\tau_i \|w_i\|^2 + (1-\tau_i) Var(X_i w_i) = 1 \quad , \text{ with } \; 0 \le \tau_{\mathrm{i}} \le 1,\, i = 1,\ldots, J$$

Criterion	SUMCOR	SSQCOR	SABSCOR
Scheme	Horst (g(x) = x)	Factorial (g(x) = x^2)	Centroid (g(x) = |x|)
Value of τ_i	0	0	0

⇒ PLS Mode B

With usual PLS-PM constraints:

$$Var(X_i w_i) = 1$$

Specific cases

$$\boxed{Maximize \sum_{j<k} c_{jk}\, g(\text{cov}(X_j w_j, X_k w_k))}$$

subject to the constraints:

$$\boxed{\tau_i \|w_i\|^2 + (1-\tau_i) Var(X_i w_i) = 1 \quad , \text{ with } \; 0 \le \tau_i \le 1,\, i = 1, \ldots, J}$$

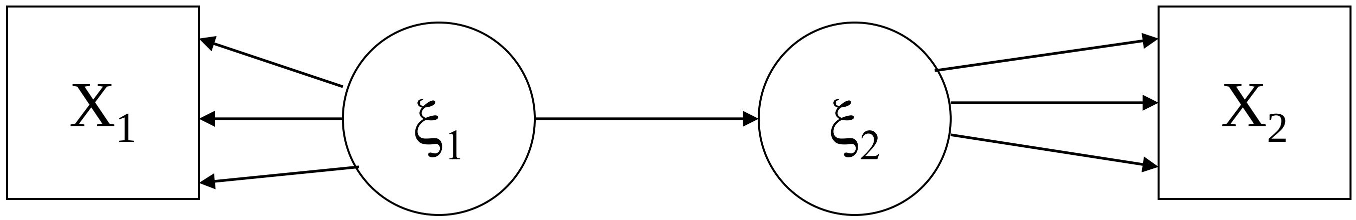
Criterion	SUMCOV	SSQCOV	SABSCOV
Scheme	Horst (g(x) = x)	Factorial (g(x) = x^2)	Centroid (g(x) = |x|)
Value of τ_i	1	1	1

→ PLS
New Mode A

With usual *PLS regression* constraint:

$$\|w_i\| = 1$$

I. PLS approach : 2 blocks



Mode for weight calculation

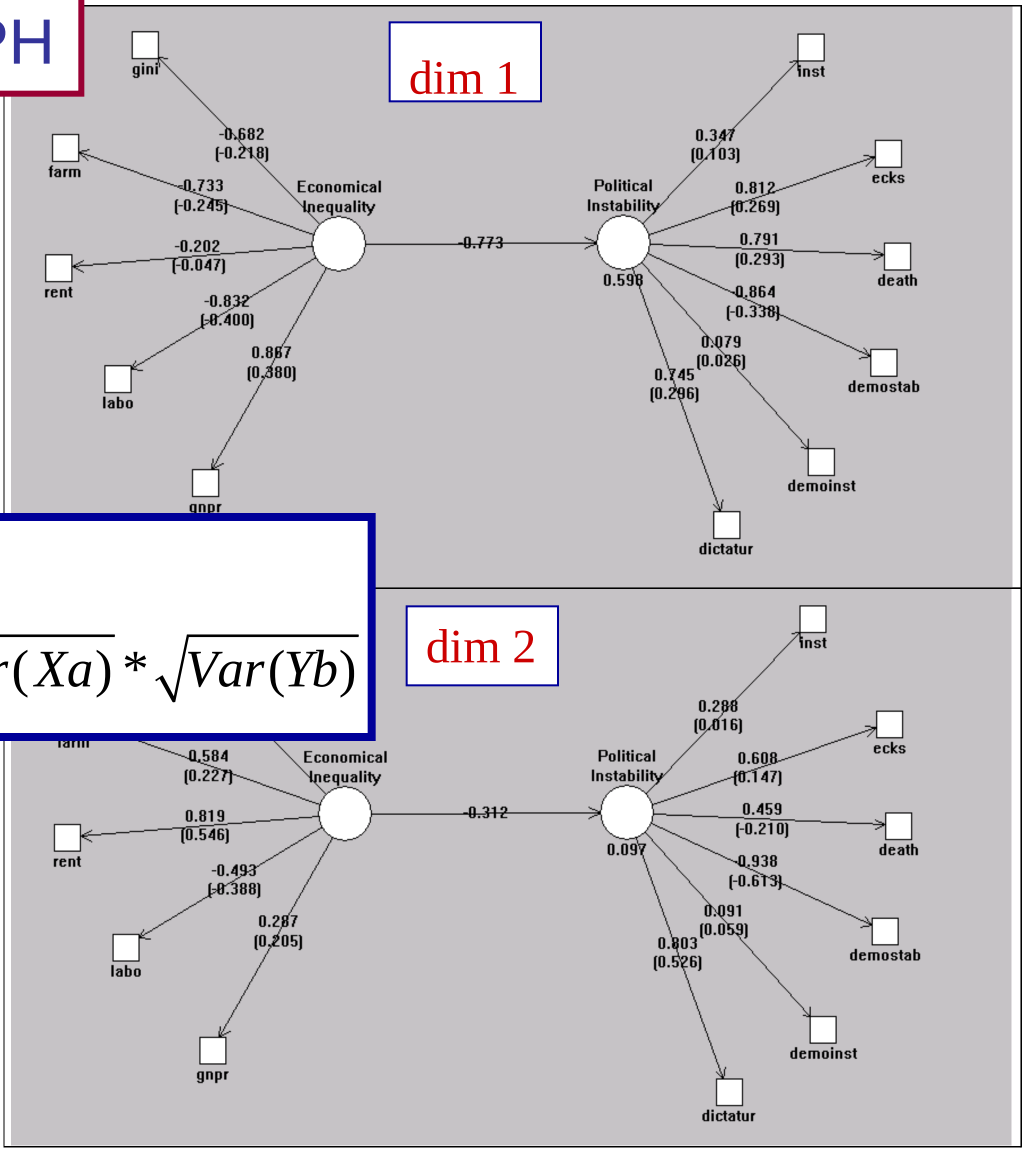
$Y_1 = X_1 w_1$	$Y_2 = X_2 w_2$	Method	Deflation (*)
A	A	PLS regression of X_2 on X_1	On X_1 only
B	A	Redundancy analysis of X_2 with respect to X_1	On X_1 only
A	A	Tucker Inter-Battery Factor Analysis	On X_1 and X_2
B	B	Canonical correlation Analysis	On X_1 and X_2

(*) Deflation: Working on residuals of the regression of X on the previous LV's in order to obtain orthogonal LV's.

PLS regression (2 components)

Use of PLS-GRAPH

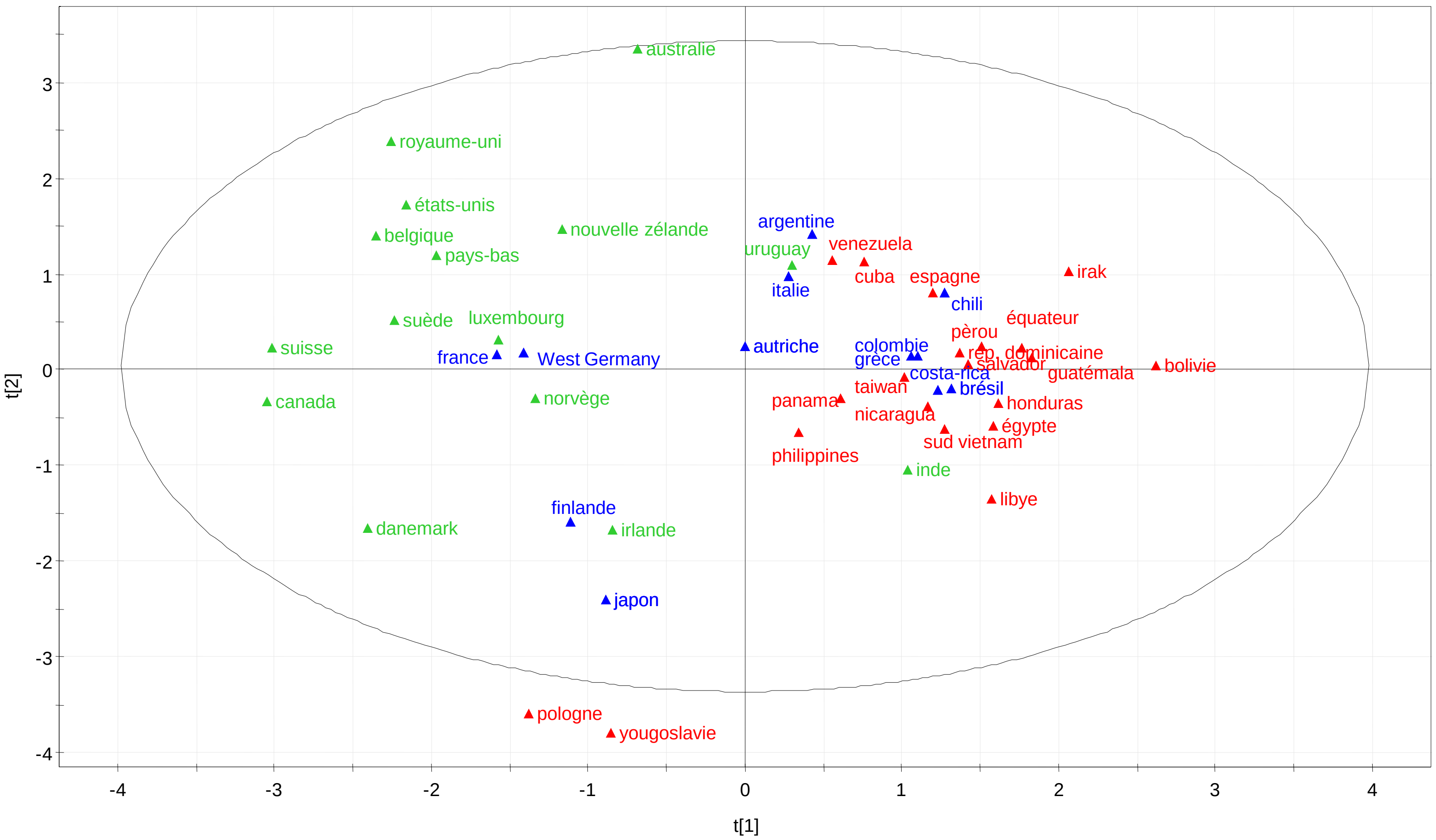
- Mode A for X
- Mode A for Y
- Deflate only X



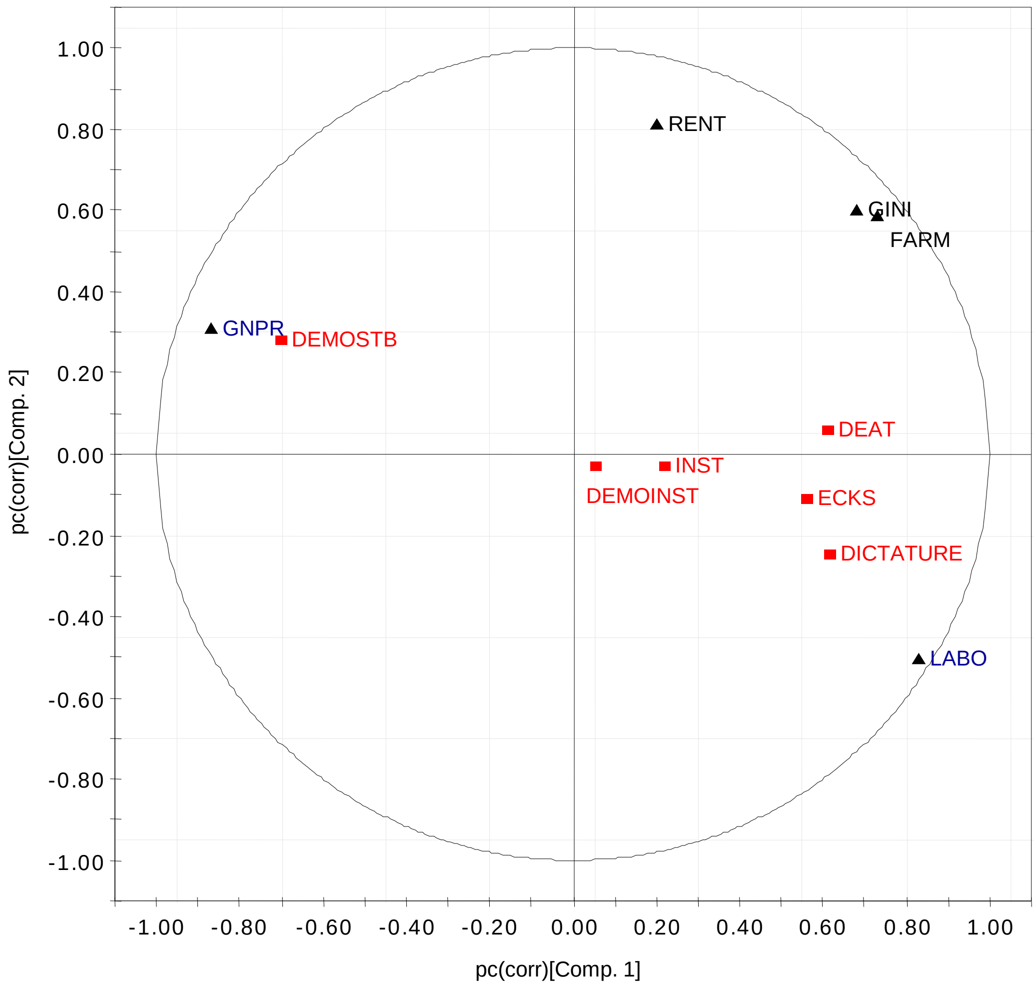
$$\max_{\|a\|=\|b\|=1} Cov(Xa, Yb)$$

$$= \max_{\|a\|=\|b\|=1} Cor(Xa, Yb) * \sqrt{Var(Xa)} * \sqrt{Var(Yb)}$$

PLS Regression in SIMCA-P : PLS Scores



Correlation loadings



Redundancy analysis of X on Y

(2 components)

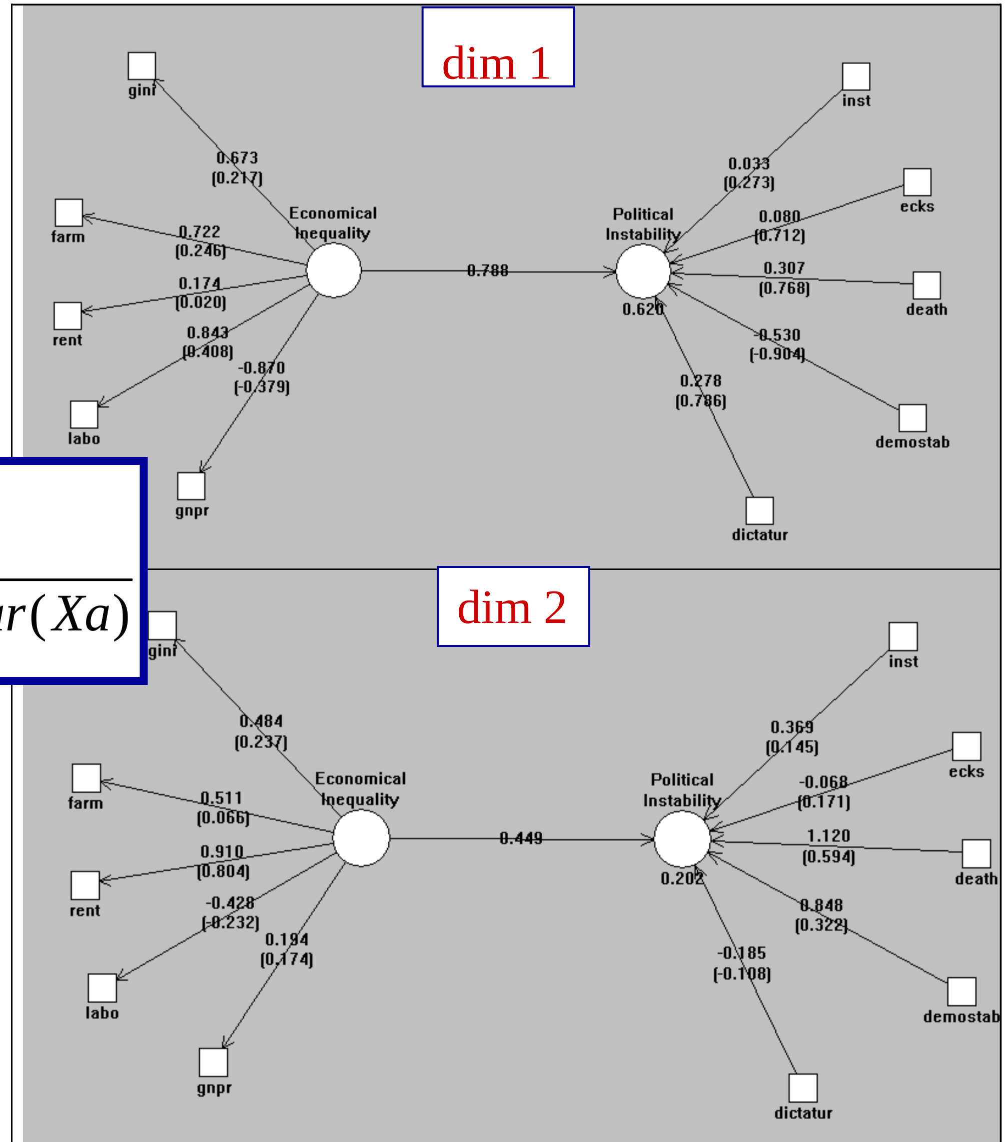
- Mode A for X
- Mode B for Y
- Deflate only X

$$\underset{\|a\|=Var(Yb)=1}{Max} \; Cov(Xa,Yb)$$

$$= \underset{\|a\|=Var(Yb)=1}{Max} \; Cor(Xa,Yb) * \sqrt{Var(Xa)}$$

$$\Updownarrow$$

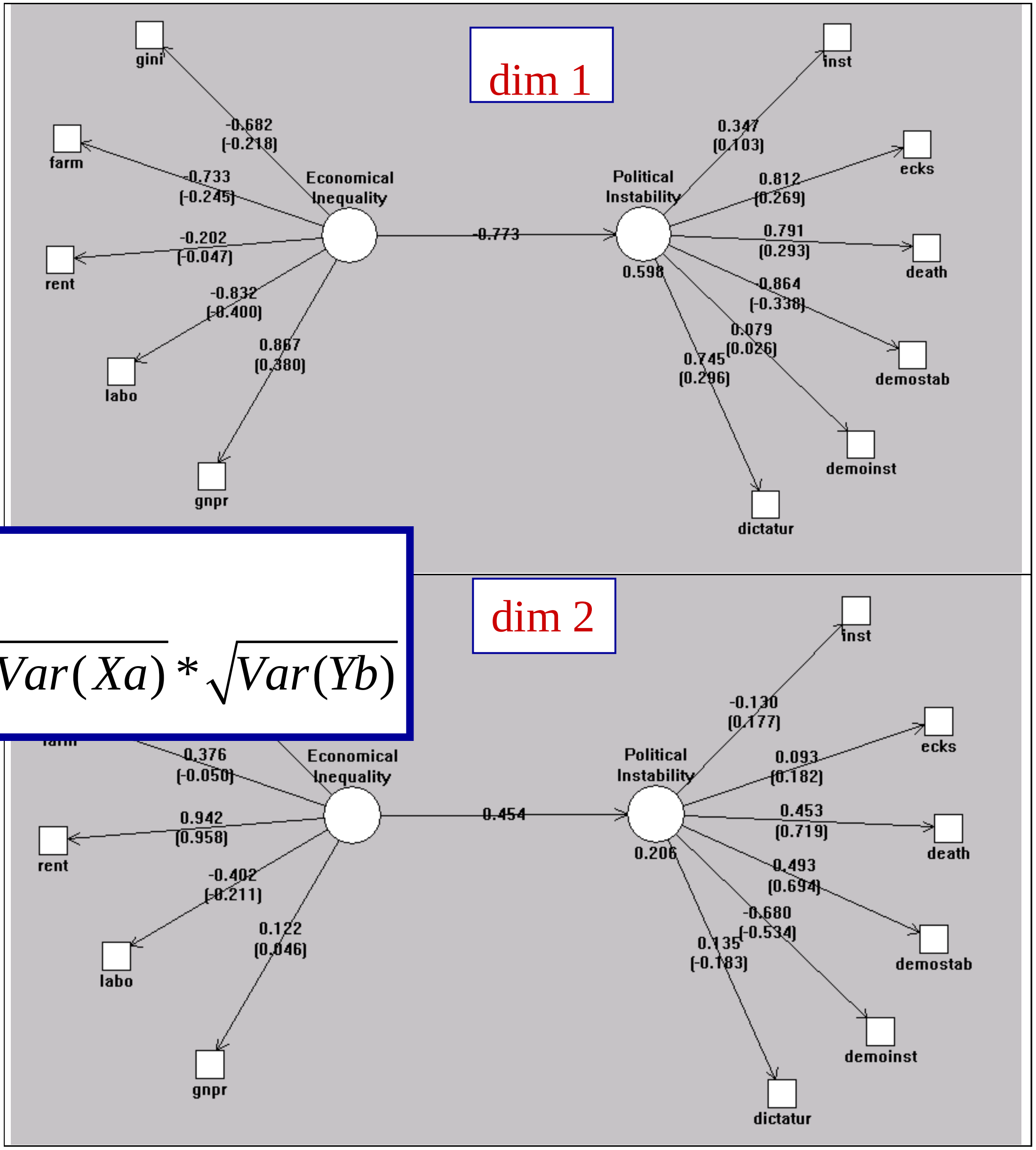
$$\underset{Var(Yb)=1}{Max} \; \sum_j Cor^2(x_j, Yb)$$



Inter-battery factor analysis (2 components)

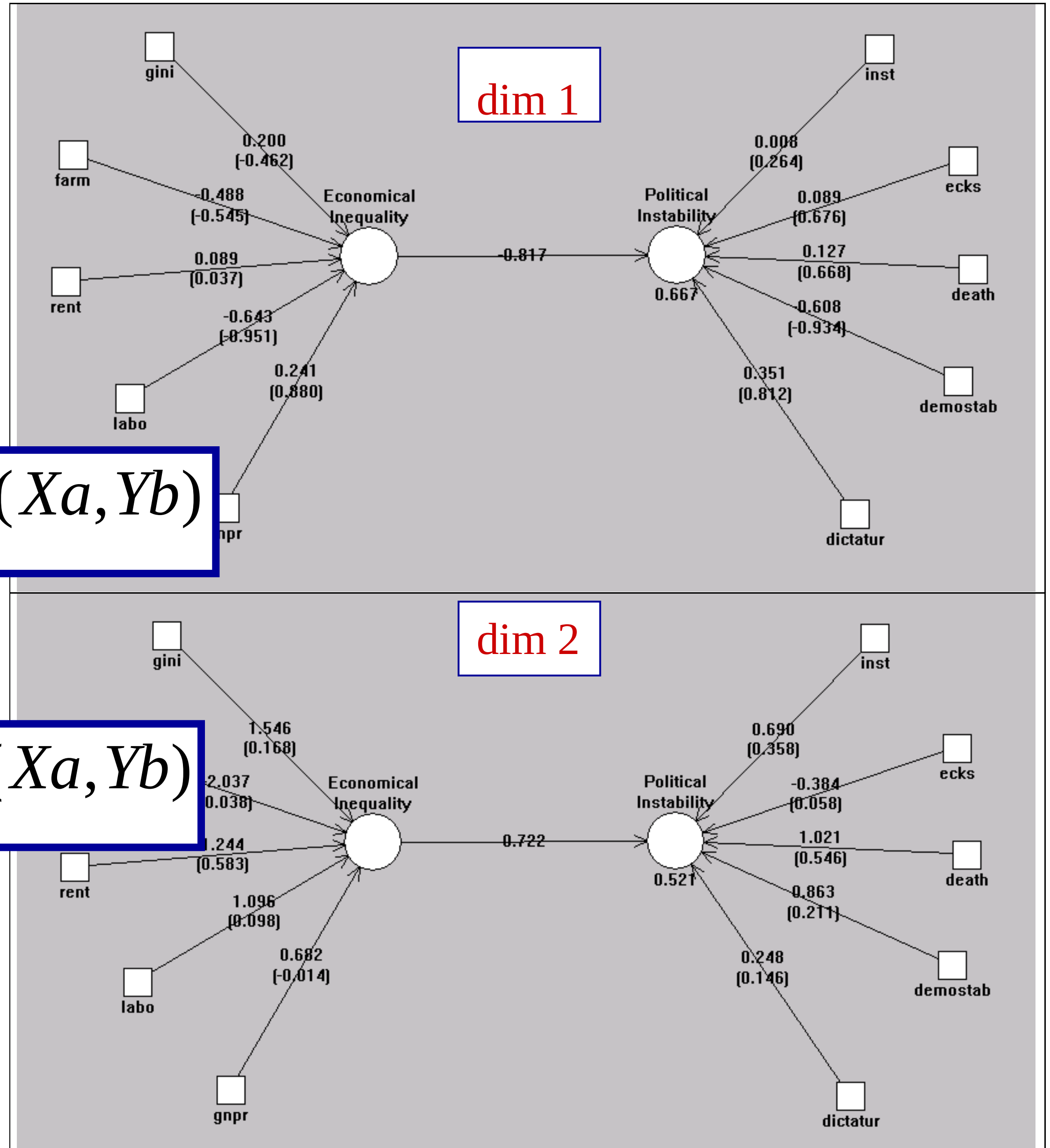
- Mode A for X
- Mode A for Y
- Deflate both X and Y

$$\underset{\|a\|=\|b\|=1}{Max}\ Cov(Xa,Yb)$$
$$=\underset{\|a\|=\|b\|=1}{Max}\ Cor(Xa,Yb)*\sqrt{Var(Xa)}*\sqrt{Var(Yb)}$$



Canonical correlation analysis (2 components)

- Mode B for X
- Mode B for Y
- Deflate both X and Y



$$\underset{Var(Xa)=Var(Yb)=1}{Max} Cov(Xa,Yb)$$

$$\underset{Var(Xa)=Var(Yb)=1}{Max} Cor(Xa,Yb)$$

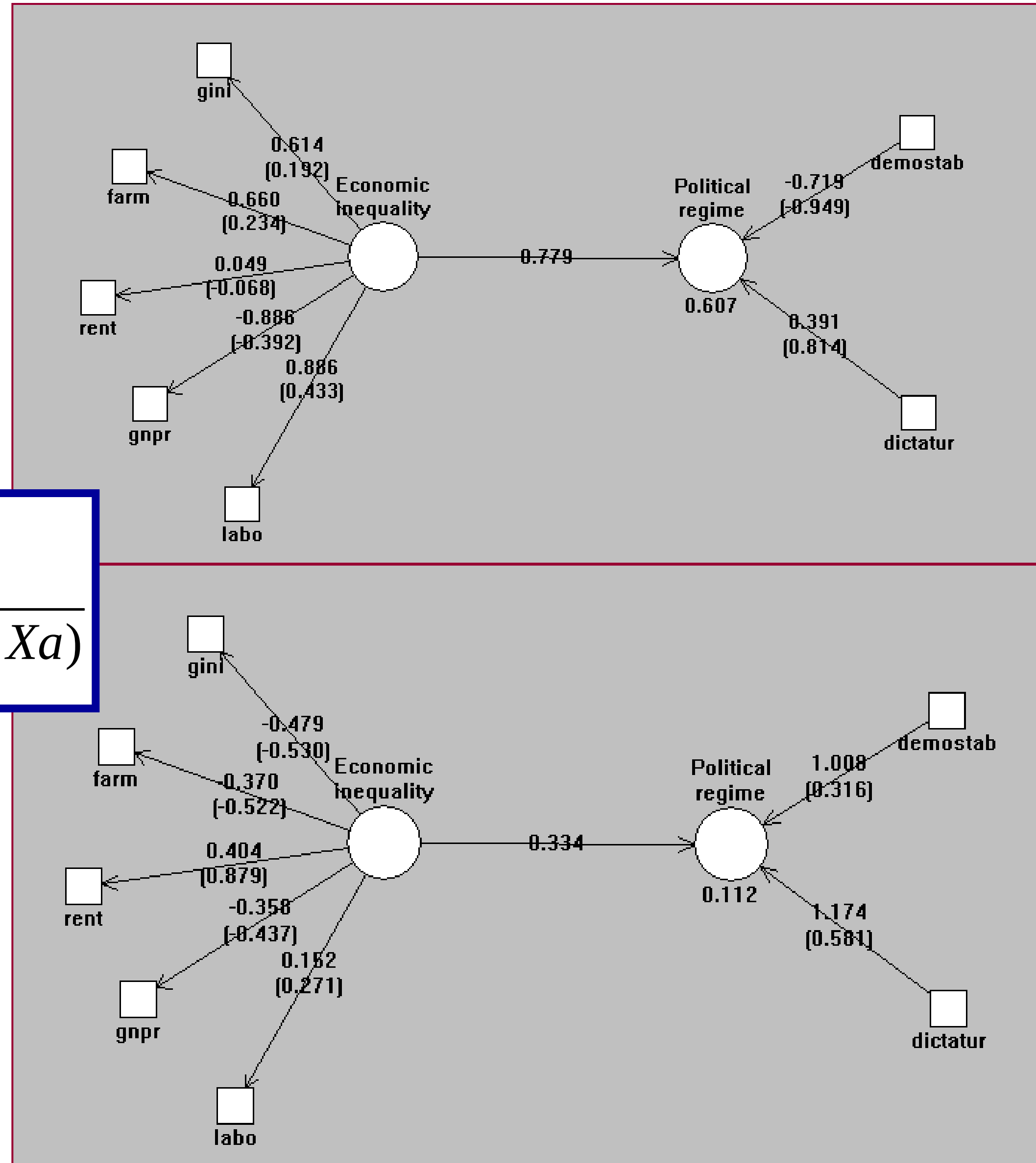
Barker & Rayens PLS DA

- Mode A for X
- Mode B for Y
- Deflate only on X

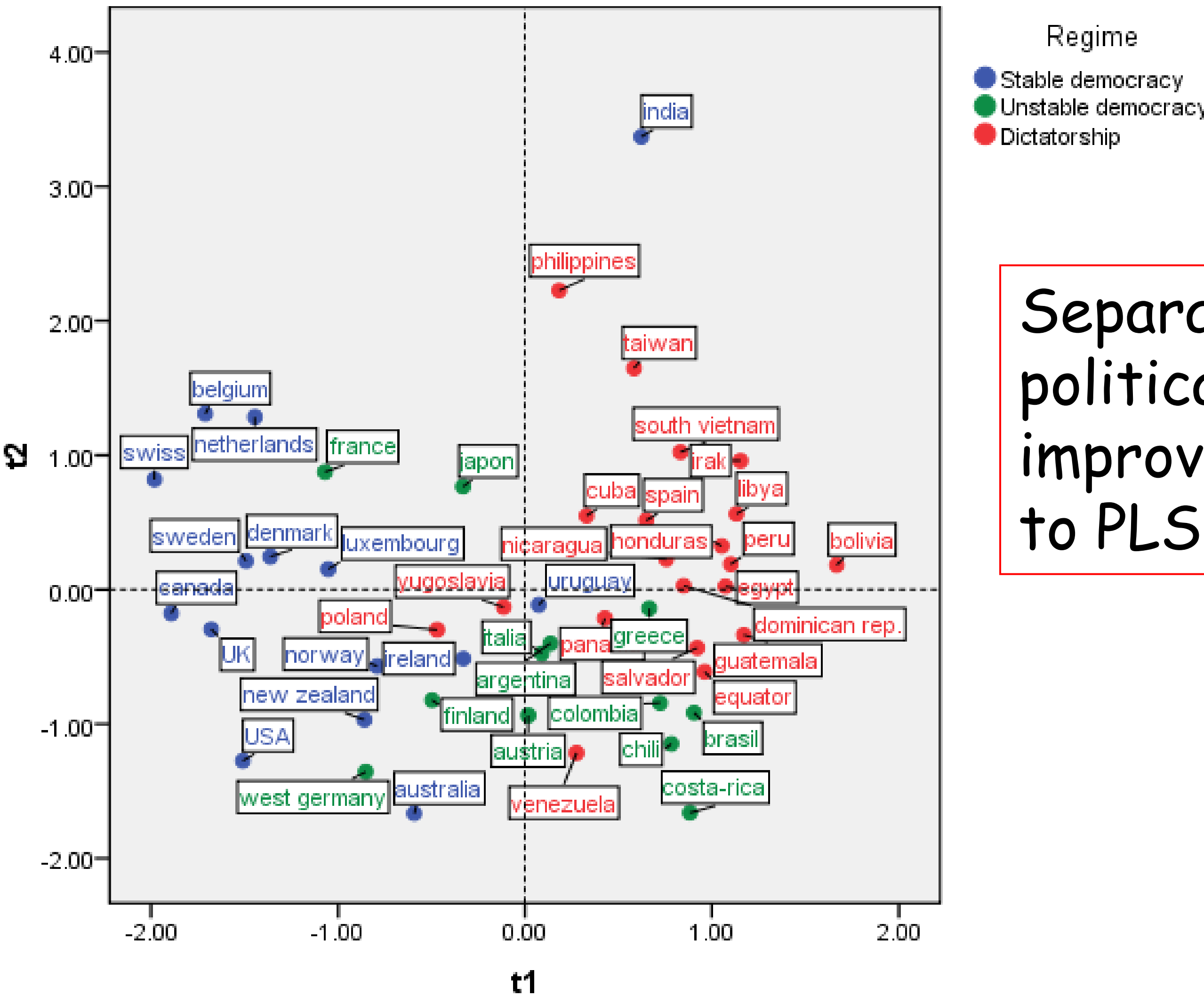
$$\underset{\|a\|=Var(Yb)=1}{Max} \quad Cov(Xa,Yb)$$
$$= \underset{\|a\|=Var(Yb)=1}{Max} \quad Cor(Xa,Yb)*\sqrt{Var(Xa)}$$

⇕

Redundancy analysis of X with respect to Y

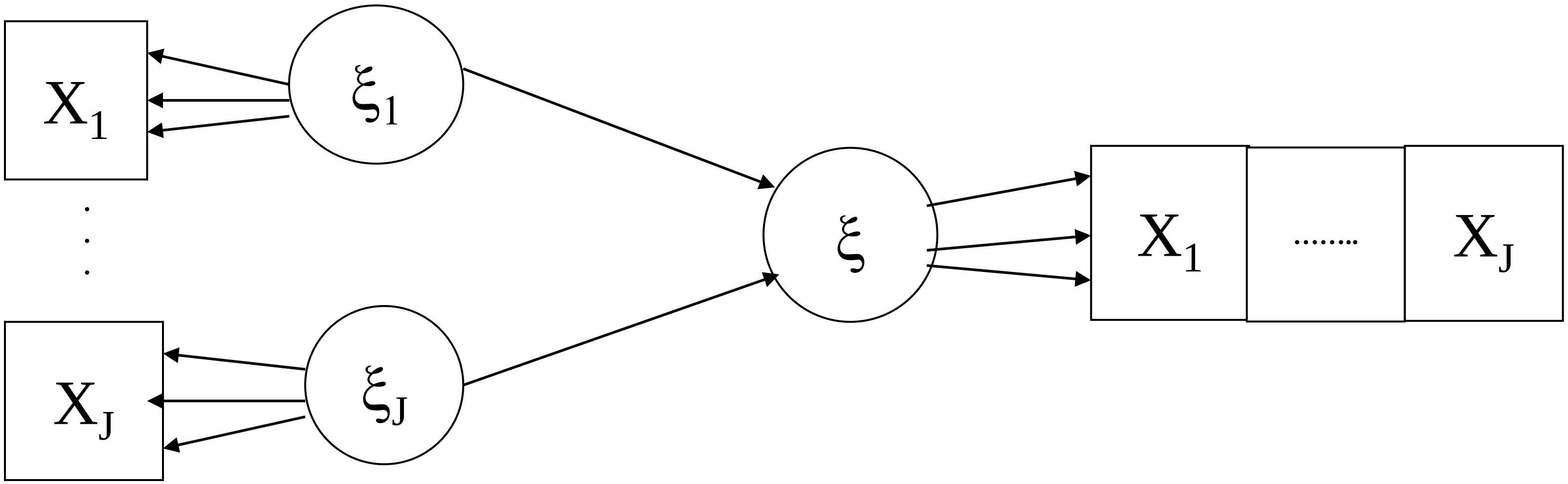


Barker & Rayens PLS DA
Economic inequality vs Political regime



Separation between political regimes is improved compared to PLS-DA.

II. Hierarchical model : J blocs



	Scheme for computation of the inner components Z_j		
Computation of outer weights w_j	Horst	Centroid	Factorial
Mode A	SUMCOV	SABSCOV	SSQCOV
Mode B	SUMCOR	SUMCOR	SSQCOR (Carroll GCCA)

Generalized PLS regression

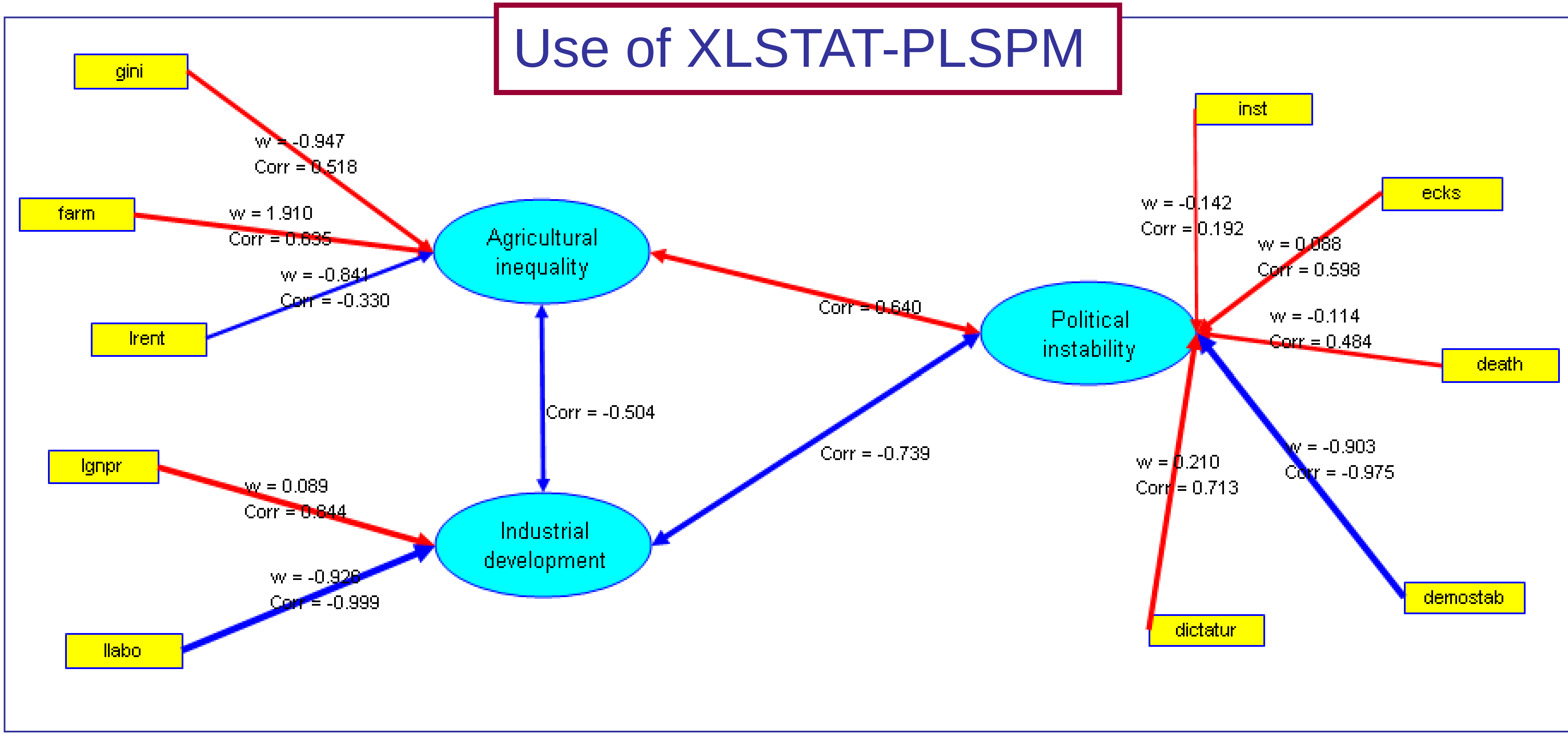
Generalized CCA

Deflation : *On original blocks and/or the super-block*

III. Multi-block data analysis

SABSCOR : PLS Mode B + Centroid scheme

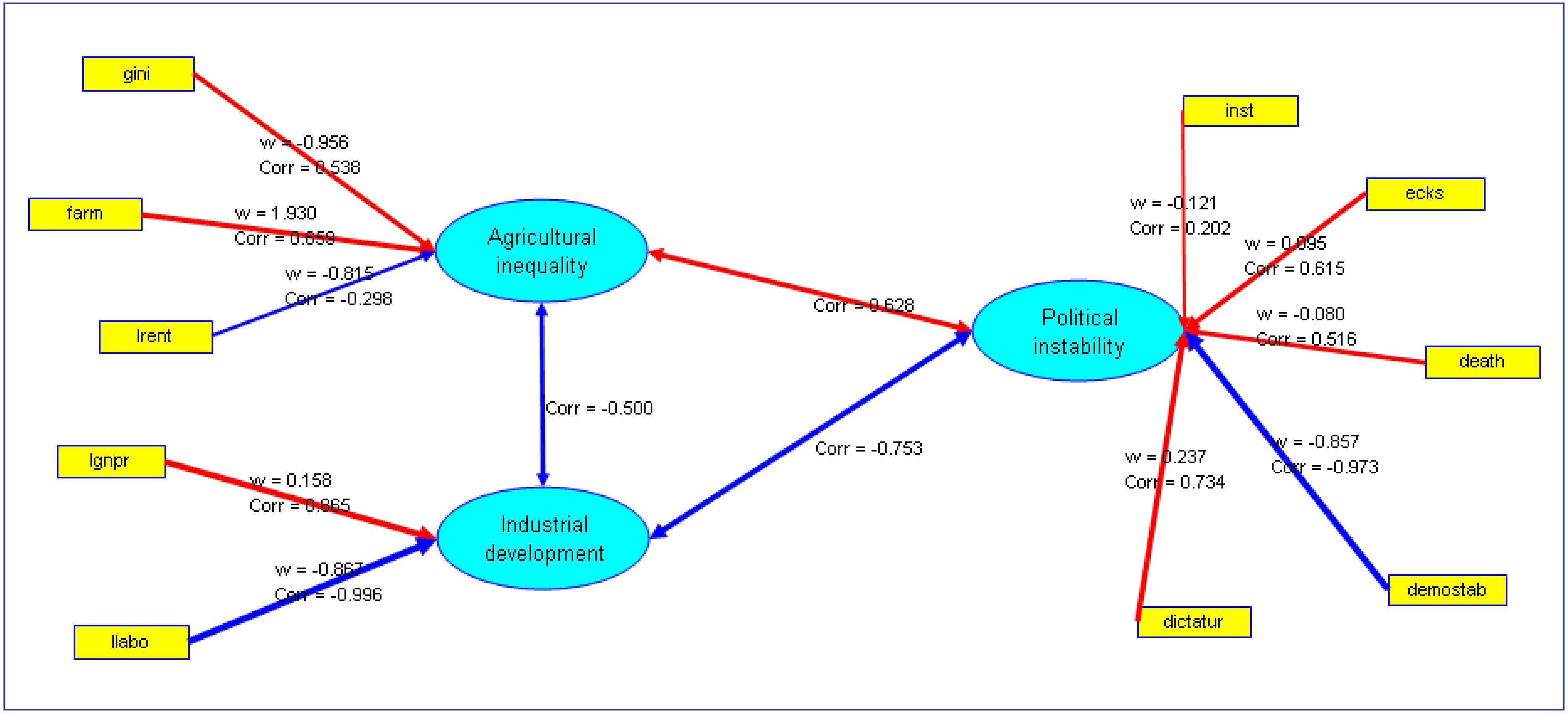
Use of XLSTAT-PLSPM



$$Max\left[\left|Cor(X_1w_1, X_2w_2)\right| + \left|Cor(X_1w_1, X_3w_3)\right| + \left|Cor(X_2w_2, X_3w_3)\right|\right]$$
$$= .504 + .640 + .739 = 1.883$$

Multiblock data analysis

SSQCOR : PLS Mode B + Factorial scheme



$$Max\left[Cor^2(X_1w_1, X_2w_2) + Cor^2(X_1w_1, X_3w_3) + Cor^2(X_2w_2, X_3w_3)\right]$$
$$= .500^2 + .628^2 + .753^2 = 1.211$$

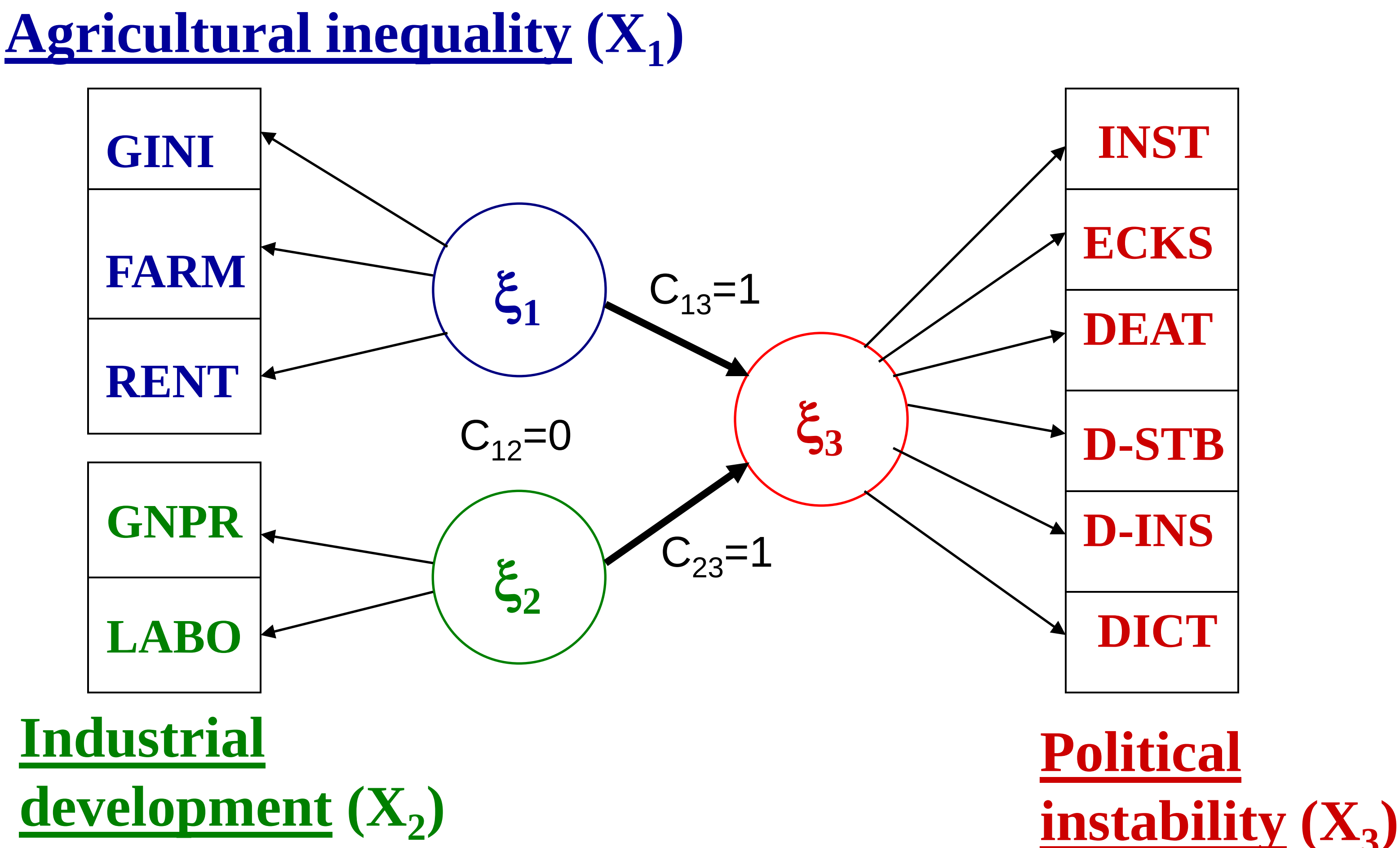
Practice supports "theory"

	Mode B + Centroid	Mode B + Factorial
Agricultural inequality <--> Industrial development Agricultural inequality <--> Political Instability Industrial development <--> Political Instability	-0.504 0.640 -0.739	-0.500 0.628 -0.753
SABSCOR SSQCOR	1.883 * 1.2097	1.881 1.211 *

* Criterion optimized by the method

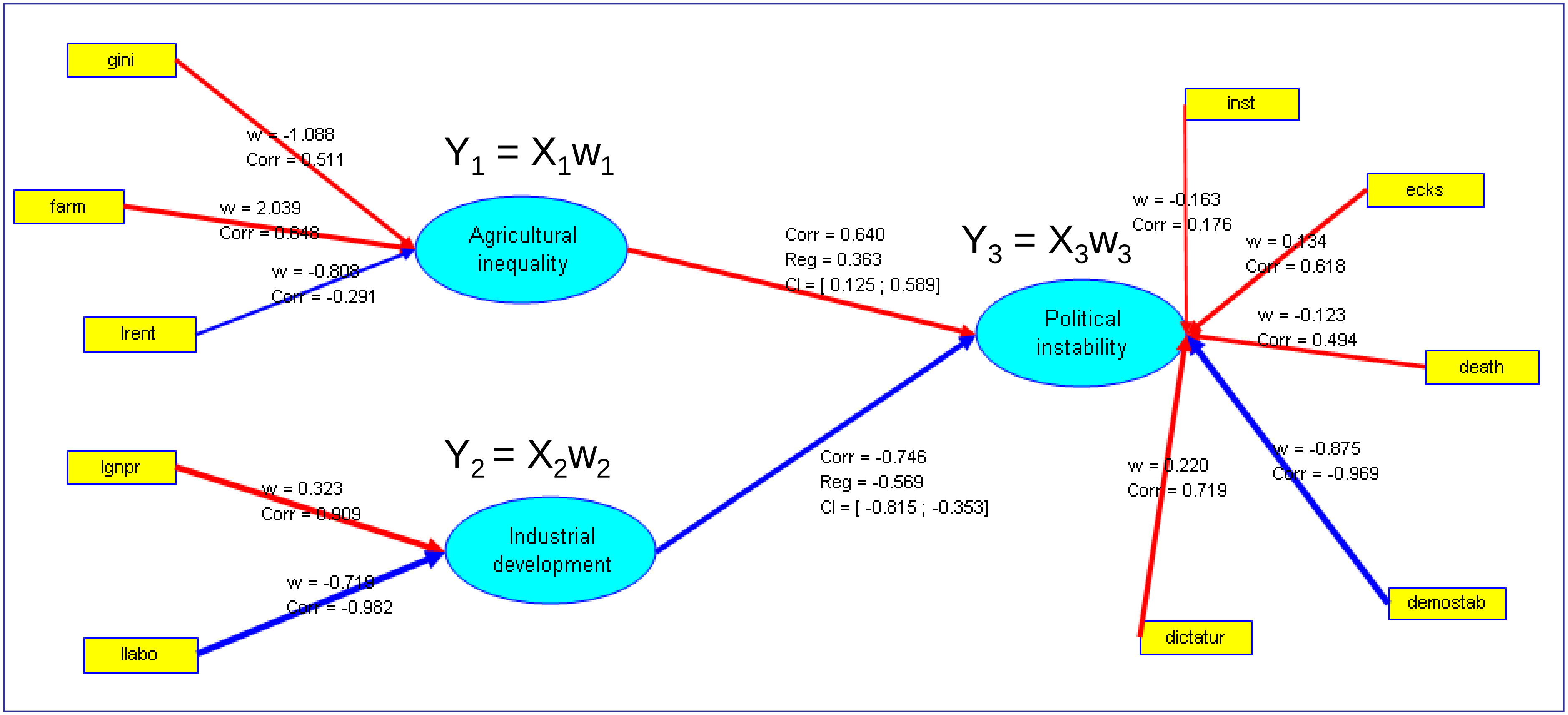
(checked on 50 000 random initial weights)

IV. Structural Equation Modeling



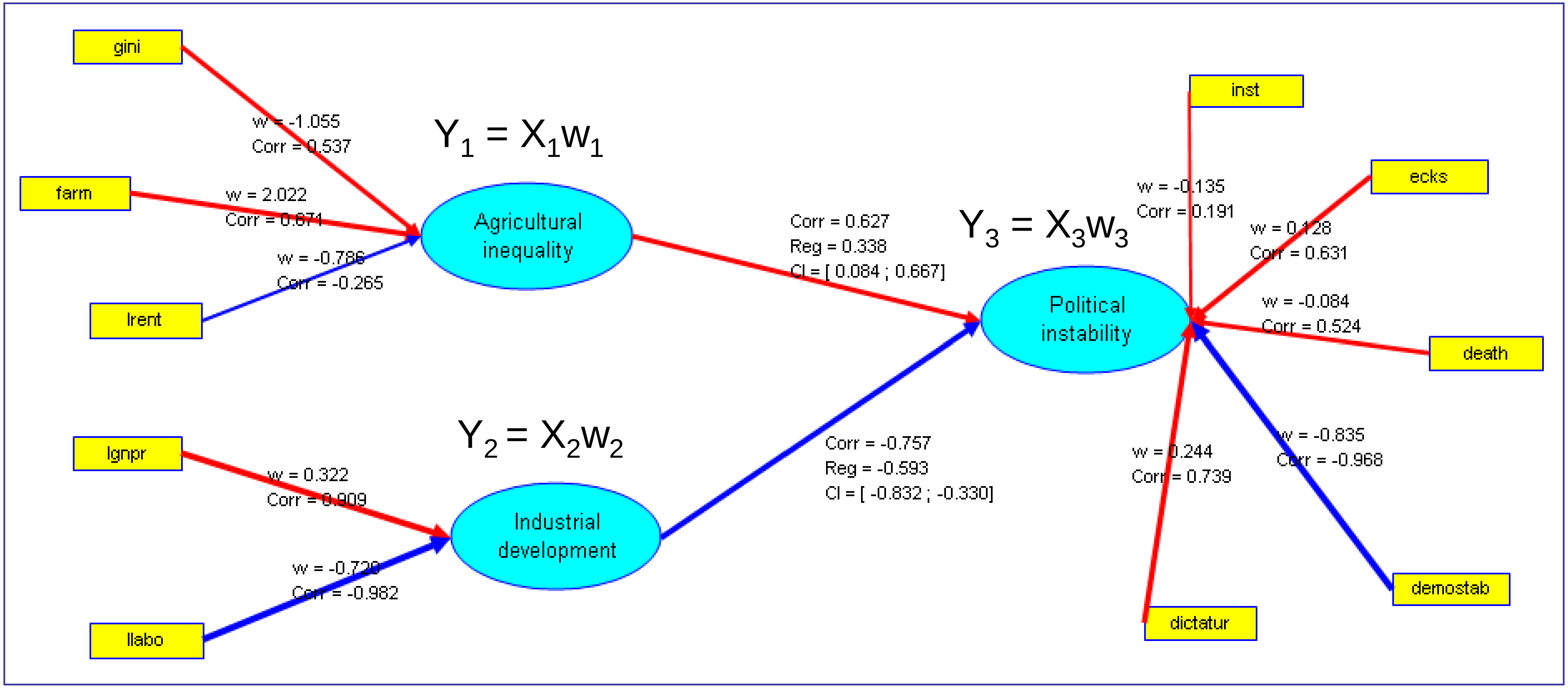
C_{ij} = 1 if ξ_i and ξ_j are connected
= otherwise

SABSCOR-PLSPM
Mode B + Centroid scheme



$$Max\left[\left|Cor(X_1w_1, X_3w_3)\right| + \left|Cor(X_2w_2, X_3w_3)\right|\right] = .640 + .746 = 1.386$$

SSQCOR-PLSPM
Mode B + Factorial scheme



$$Max\left[Cor^2(X_1w_1, X_3w_3)+Cor^2(X_2w_2, X_3w_3)\right]=.627^2+.757^2=.966178$$

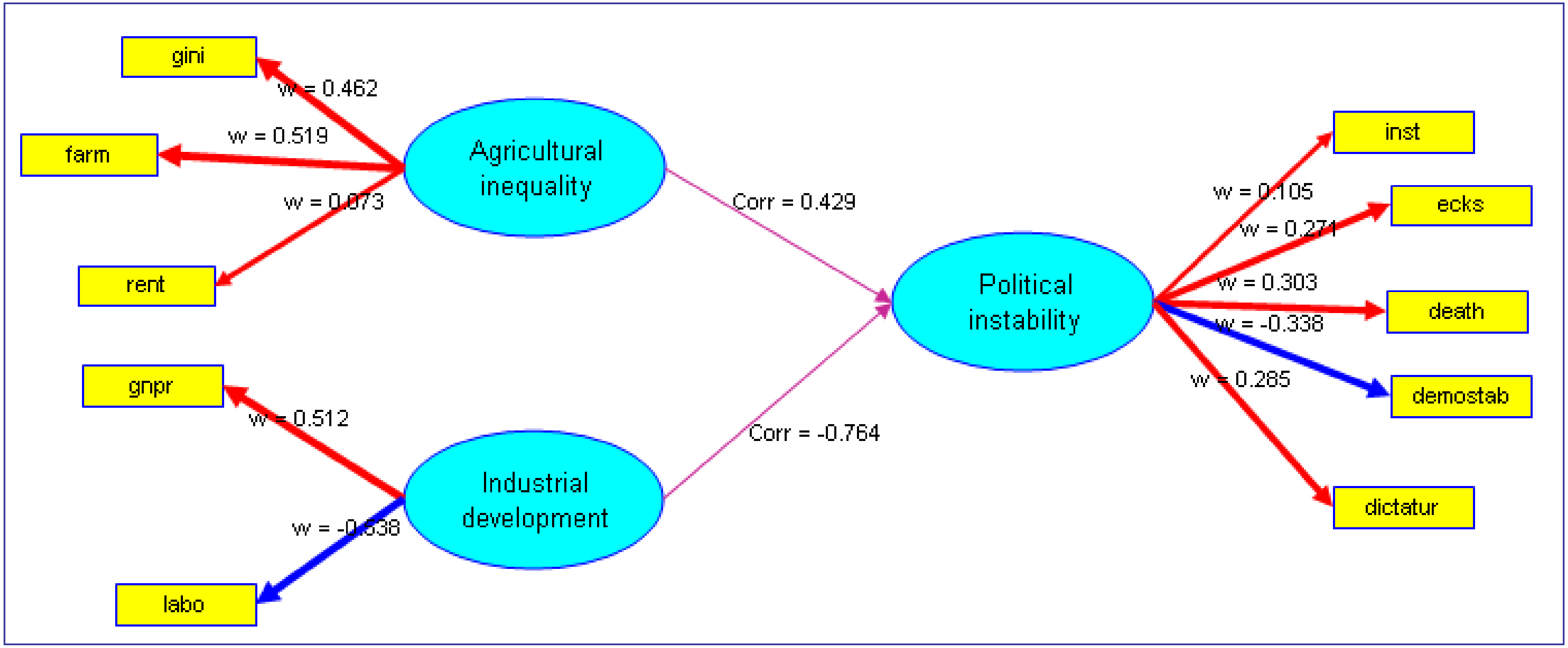
Comparison between methods

	Mode B + Centroid scheme	Mode B + Factorial scheme
$\lvert Cor(Y_1,Y_3)\rvert + \lvert Cor(Y_2,Y_3)\rvert$	1.386 *	1.384
$Cor^2(Y_1,Y_3) + Cor^2(Y_2,Y_3)$	.966116	.966178 *

* Criterion optimized by the method (checked on 50 000 random initial weights)

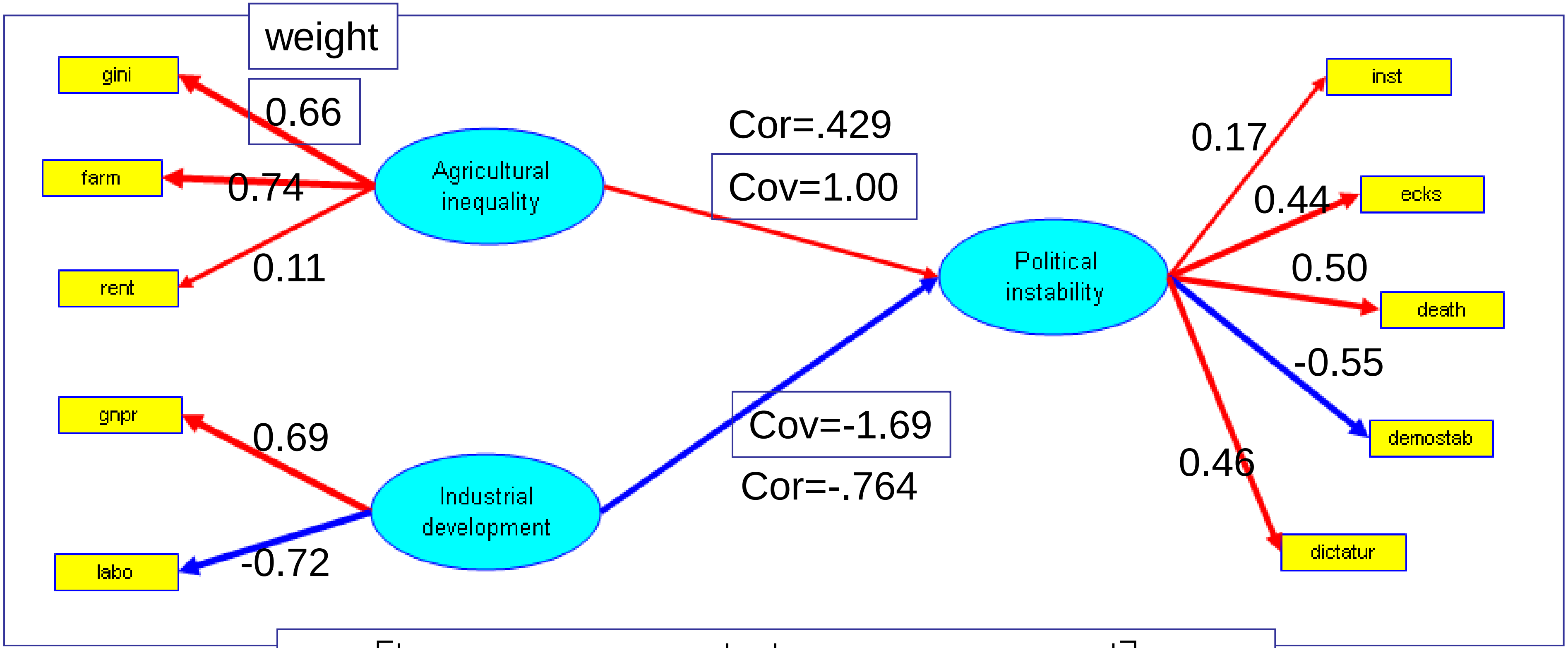
Practice supports "theory"

Mode A + Centroid scheme



The criterion optimized by the algorithm, if any, is unknown.

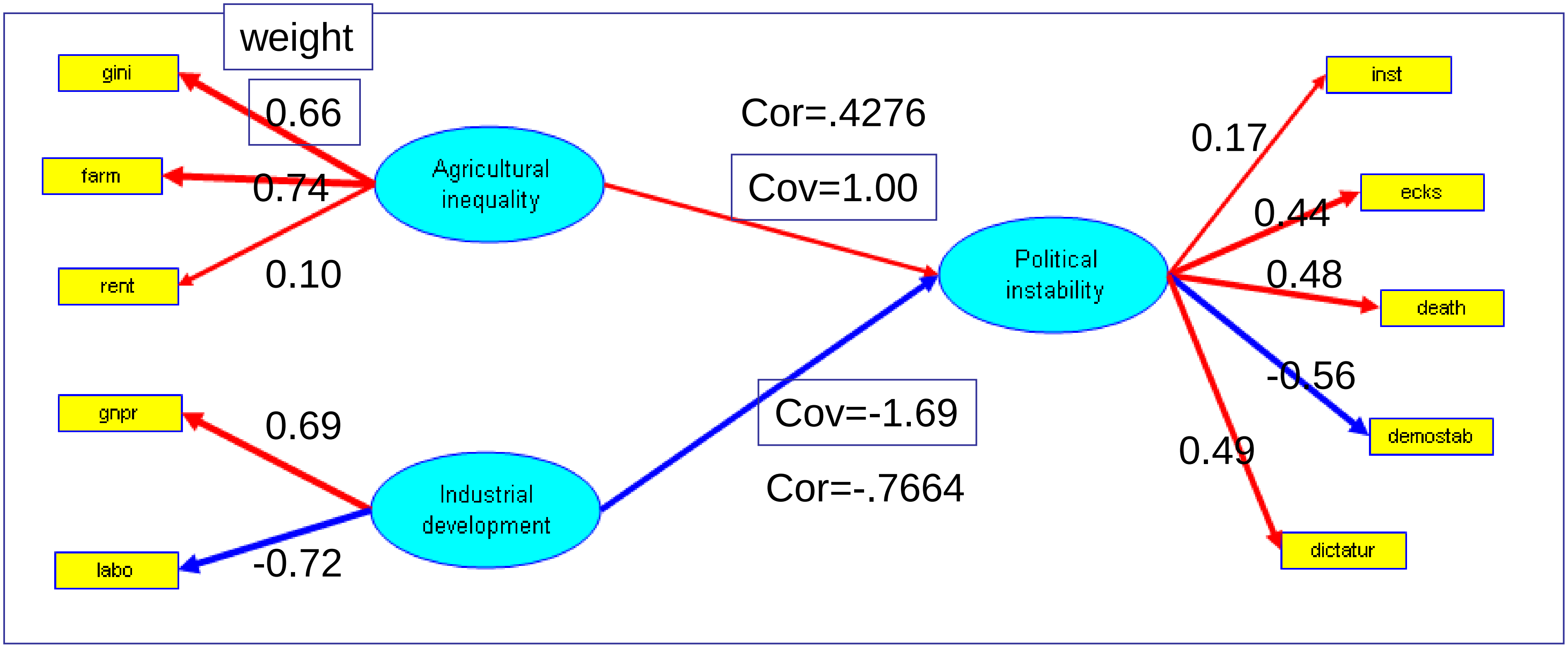
SABSCOV-PLSPM
New Mode A + Centroid scheme



$$Max\left[\left|Cov(X_1w_1, X_3w_3)\right| + \left|Cov(X_2w_2, X_3w_3)\right|\right] = 2.69$$

➔ One-step hierarchical PLS Regression

SABSCOV-PLSPM
New Mode A + Factorial scheme

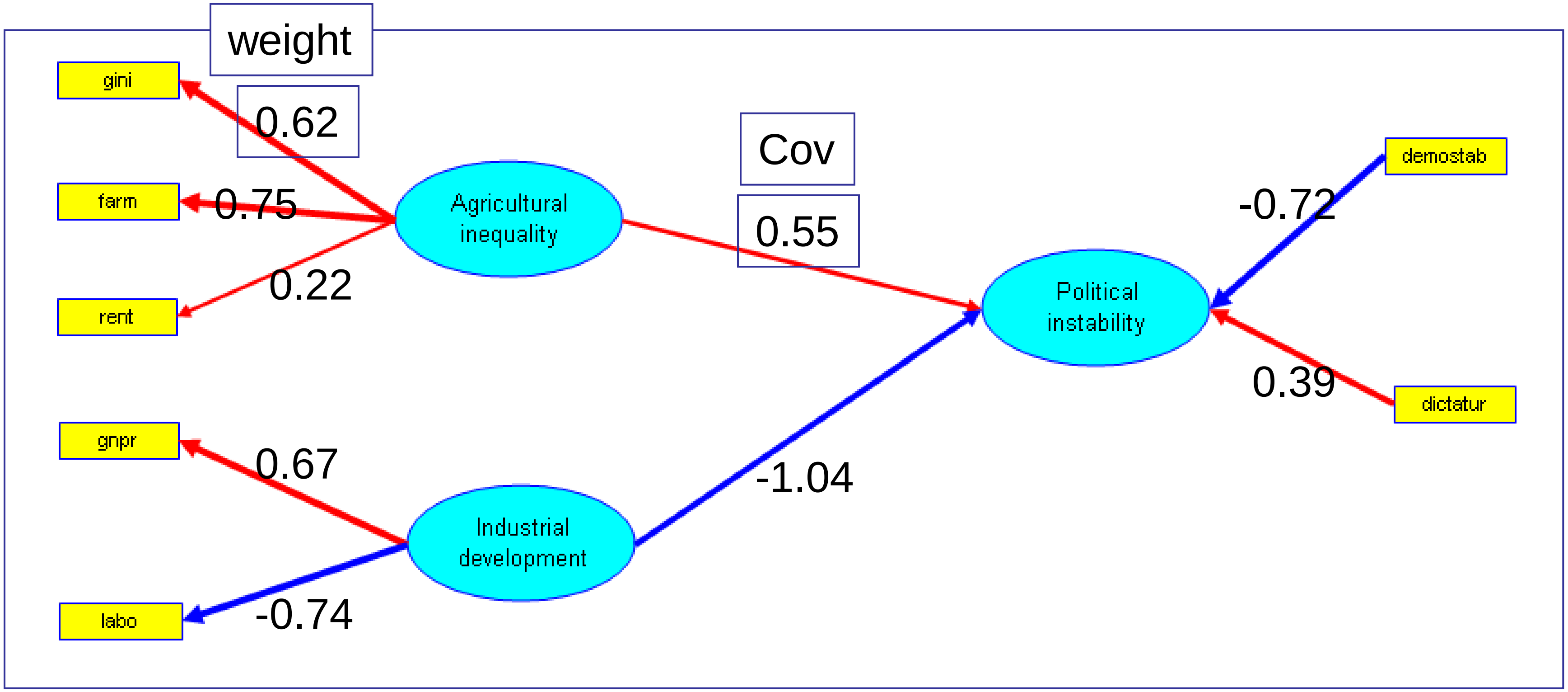


$$Max\left[Cov^2(X_1w_1, X_3w_3) + Cov^2(X_2w_2, X_3w_3)\right] = 3.86$$

➔ One-step hierarchical PLS Regression

Generalized Barker & Rayens PLS-DA SSQCOV-PLSPM

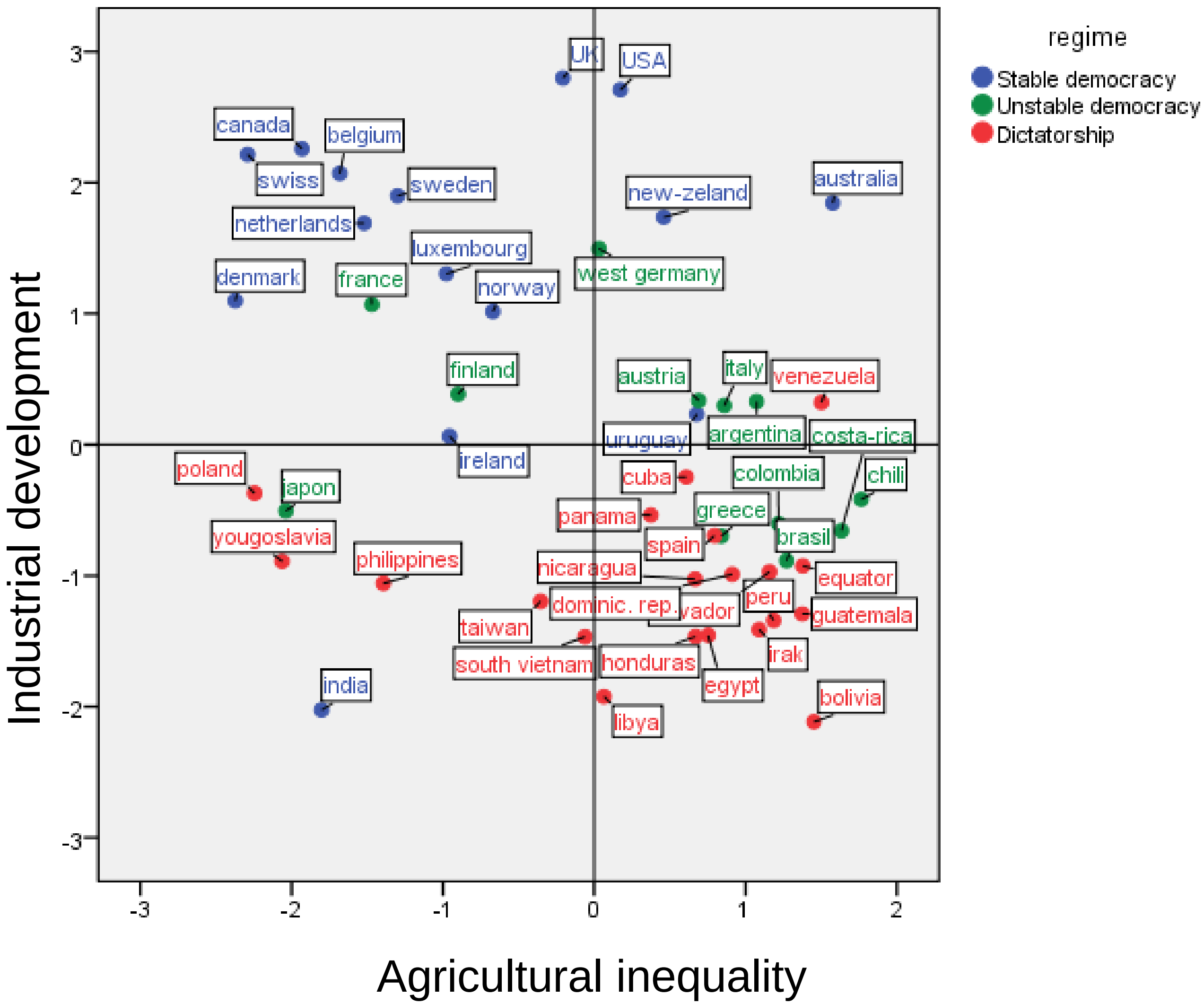
New Mode A for X_1 and X_2 and Mode B for Y



$$Max\left[Cor^2(X_1w_1, X_3w_3)*Var(X_1w_1)+Cor^2(X_2w_2, X_3w_3)*Var(X_2w_2)\right]=1.39$$

➔ One-step hierarchical B&R PLS-DA

Generalized Barker & Rayens PLS-DA



Conclusion

- In the PLS approach of Herman Wold, the constraint is:

$$\boxed{Var(X_j w_j) = 1}$$

- In the PLS regression of Svante Wold, the constraint is:

$$\boxed{\left\| w_j \right\| = 1}$$

- This presentation unifies both approaches.