



Role and treatment of categorical variables in PLS Path Models for Composite Indicators

Laura Trincherà^{1,2} & Giorgio Russolillo²

¹Dipartimento di Studi sullo Sviluppo Economico, Università degli Studi di Macerata

²Dipartimento di Matematica e Statistica, Università degli Studi di Napoli "Federico II"



Talk outline

- ② **Part I:** Using Structural Equation Models and PLS Path Modeling to build systems of Composite Indicators
- ② **Part II:** Categorical Variables as moderating variables
 - ✘ Manifest Moderating variables
 - ✘ Latent Moderating variables
- ② **Part III:** Categorical Indicators as manifest variables
 - ✘ Modified PLS Path Modeling algorithm
- ② **Part IV:** An example with RUSSET data
- ② **Conclusion and Perspectives**

Introduction

Composite Indicators (CIs) are mathematical combinations of single quantitative indicators representing different dimensions of the concept to be measured (Saisana et al., 2002)



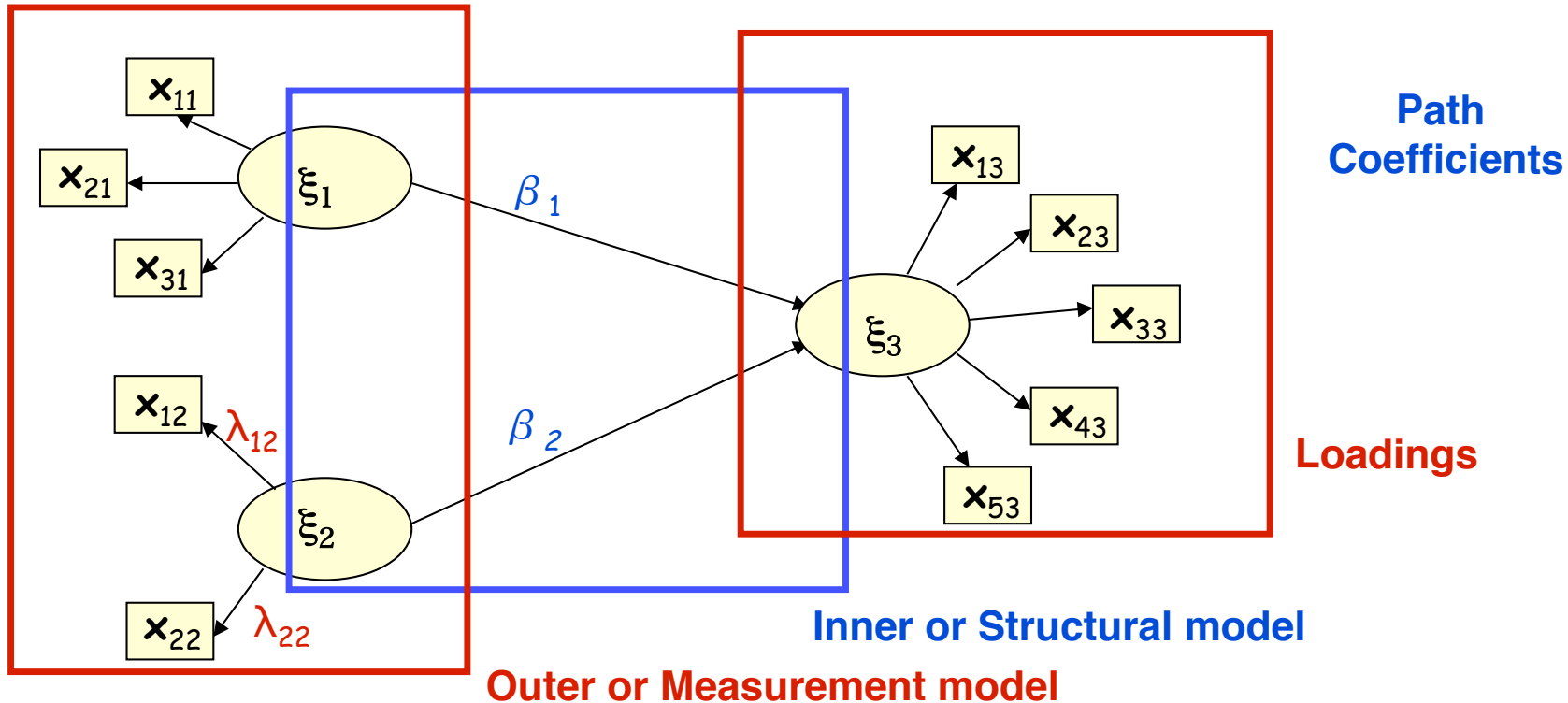
The main feature of a **CI** is that it summarizes complex and multidimensional issues

Structural Equation Models (SEMs) are complex models allowing the study of real world complexity by taking into account a number of causal relationships among latent concepts (i.e. the **CIs**), each measured by several observed indicators usually defined as Manifest Variables.

Each **CI** is not only a composite indicators but also a complex indicator due to the causal relations with the other CIs

Part I

Structural Equation Models



- **P manifest variables** or indicators (MVs) observed on N units $\rightarrow x_{pq}$ generic MV
- **Q latent variables** or Composite Indicators (LVs) $\rightarrow \xi_q$ generic LV
- **Q blocks** composed by each LV and the corresponding MVs
 \rightarrow in each q -th block p_q manifest variables x_{pq} , with $\sum_{q=1}^Q p_q = P$

PLS Path Modeling (PLS-PM)

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

It is an **iterative algorithm** that allows us to estimate the LV scores through a **system of interdependent linear equations** modeling the relations among the MVs and their corresponding LV and among the LVs of the model

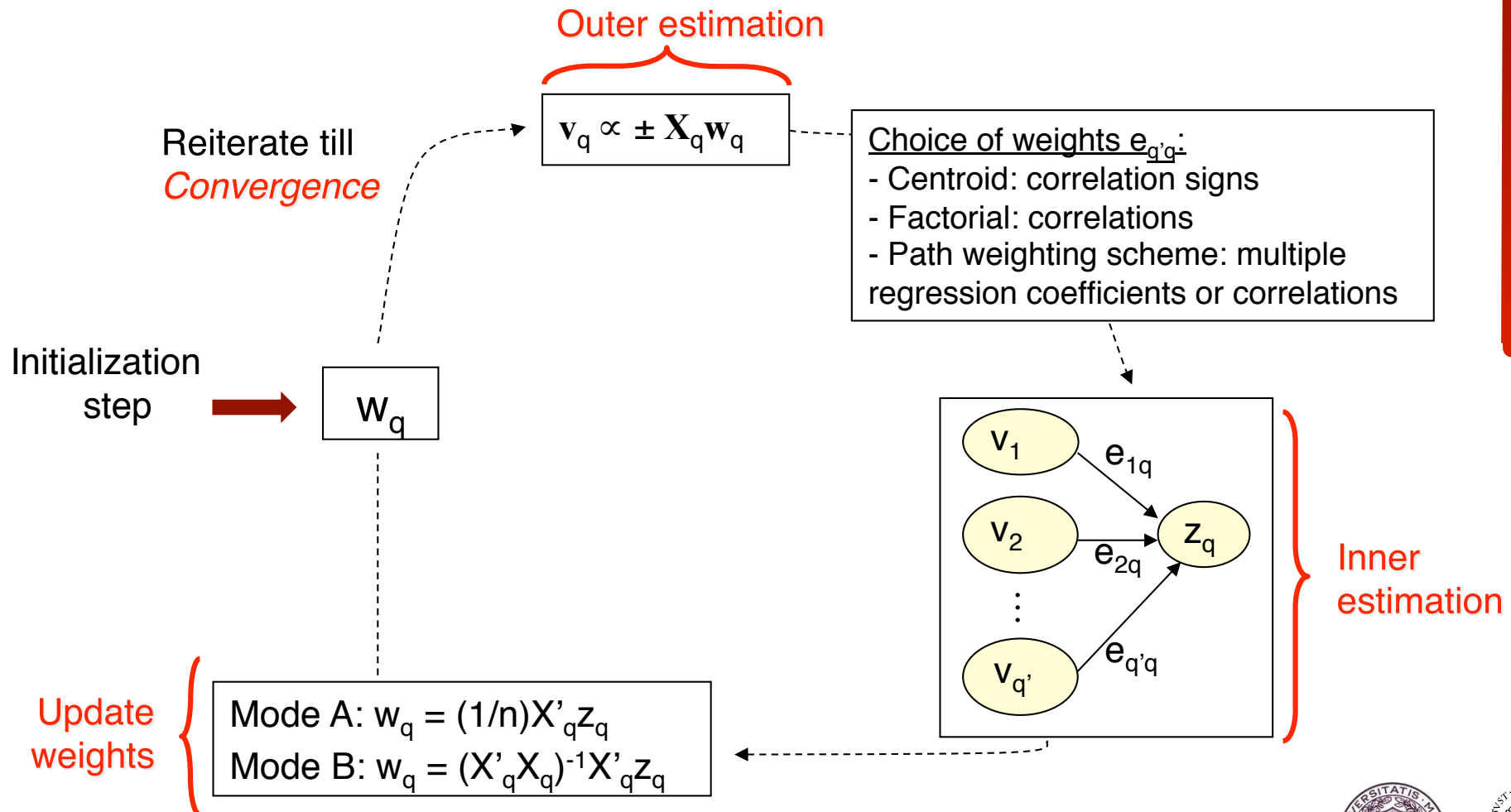
The LV scores (i.e. the **CI**s) are obtained so as to be **the most representative** of each block of **indicators** and **the most correlated** with **one another** (according to path diagram)

PLS-PM pro and cons :

- it is always identified
- it is a **distribution-free** technique
- the LV scores are “directly” obtained
- PLS-PM convergence is assured in practice, but it is not proved
- PLS-PM does not maximizes a unique function

PLS Path Modeling (PLS-PM)

A schematic representation of the PLS-PM algorithm



Categorical Variables in PLS-PM

Categorical Variables can **play two different roles** in a PLS Path Model:

Moderating categorical variables

--> are variables **influencing the relations**, in terms of strength and/or direction, between an exogenous and an endogenous variable



The moderating effect can be seen as the effect obtained by considering **several groups of units**

Active categorical variables

--> are variables directly participating in computing LV scores



Categorical variables are **indicators (MVs)** in a PLS Path Model

Moderating Categorical Variables

The moderating effect can be seen as the effect obtained by considering **several groups of units** defined by the **categories of the moderating categorical variable**

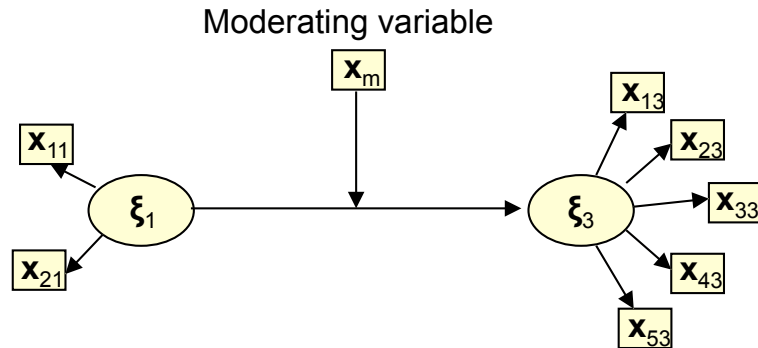
(a) **Manifest** moderating categorical variables

--> An observer categorical variable defines *a priori* the **groups** of units to be considered (*e.g.*: gender-specific indexes like GDI)

(b) **Latent** moderating categorical variables

--> A latent moderating categorical variable is a variable defining **latent classes** of units

Manifest Moderating Categorical Variables



x_m is a **categorical variables** defining classes, for instance the gender variable

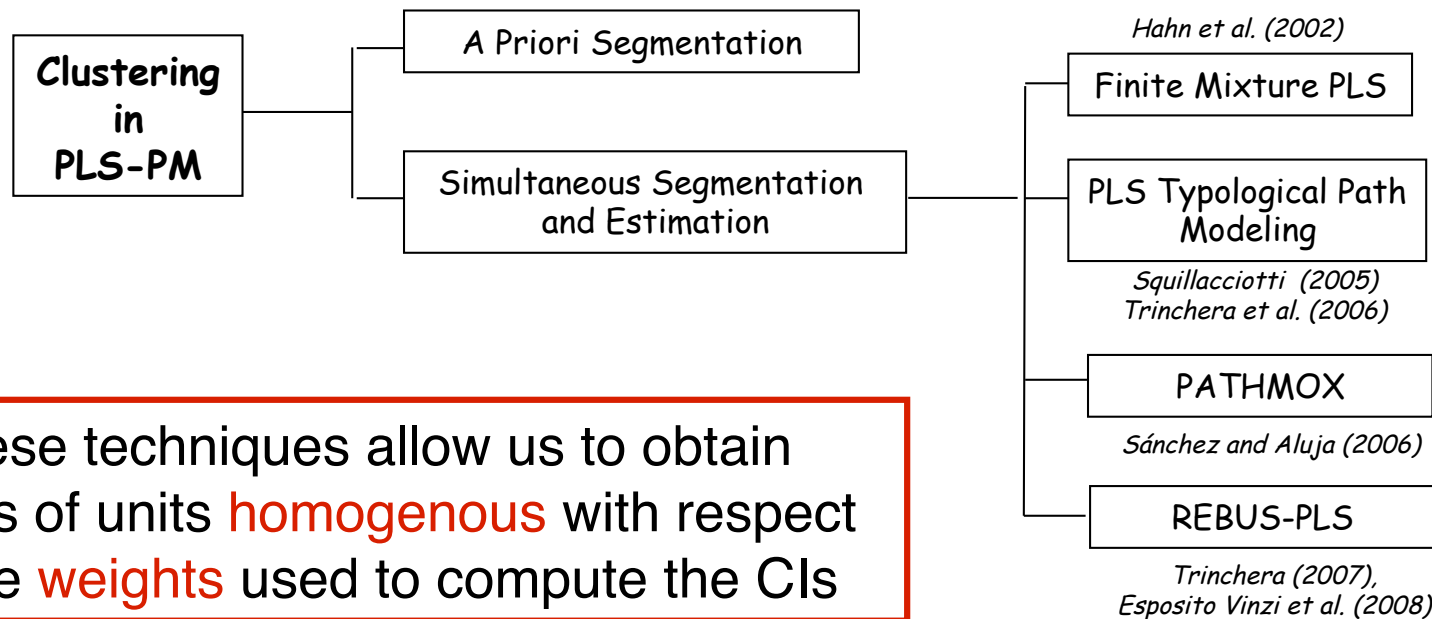
Several techniques has been proposed to consider x_m in a PLS Path Model:

- > by including **interaction term as the product of the indicators** linked to the exogenous LV and the categories of the moderating variable (Chin *et. al*, 2003)
- > by including **interaction term trough a two steps procedure** (first define the LV scores and then used these variables to obtain interaction terms) (Henseler *et. al*, 2009)
- > by considering the interaction term in the sense of Chin *et al.* (2003) but removing the redundant information (Tenenhaus *et. al*, 2009)

Latent Moderating Categorical Variables

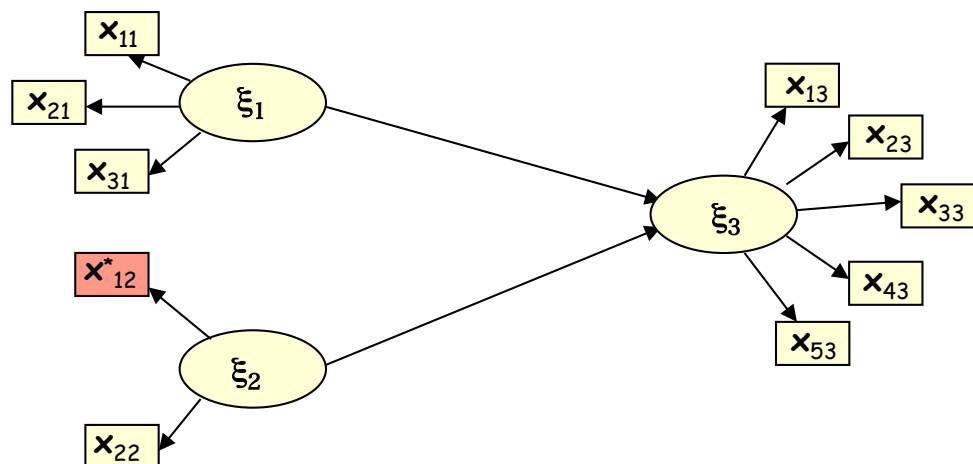
We search for **latent classes** showing different models, i.e. **local models**

Several techniques have been proposed to obtain **response-based unit clustering** in PLS-PM :



These techniques allow us to obtain groups of units **homogenous** with respect to the **weights** used to compute the CIs

Categorical Variables as MVs



x_{12}^* is a **categorical indicator**,
e.g. type of government

Usually x_{12}^* is **replaced** by the corresponding **dummy-matrix** \tilde{X}_{12} , but:

--> The **number of indicators increases** (we have as many indicators as the categories of categorical indicator are)

--> The impact of each categories is measured, but **no information about the role of the categorical indicator as a whole**

➔ **Optimal scaling** of categorical indicators

Categorical Variables as MVs

We extend **PLS-CAP** regression algorithm by Russolillo (2008) to PLS-PM

Each categorical indicator \mathbf{x}_{pq}^* is quantified in such a way that its weight in building the corresponding LV score is a function of the LV variance explained by \mathbf{x}_{pq}^* categories, in particular:

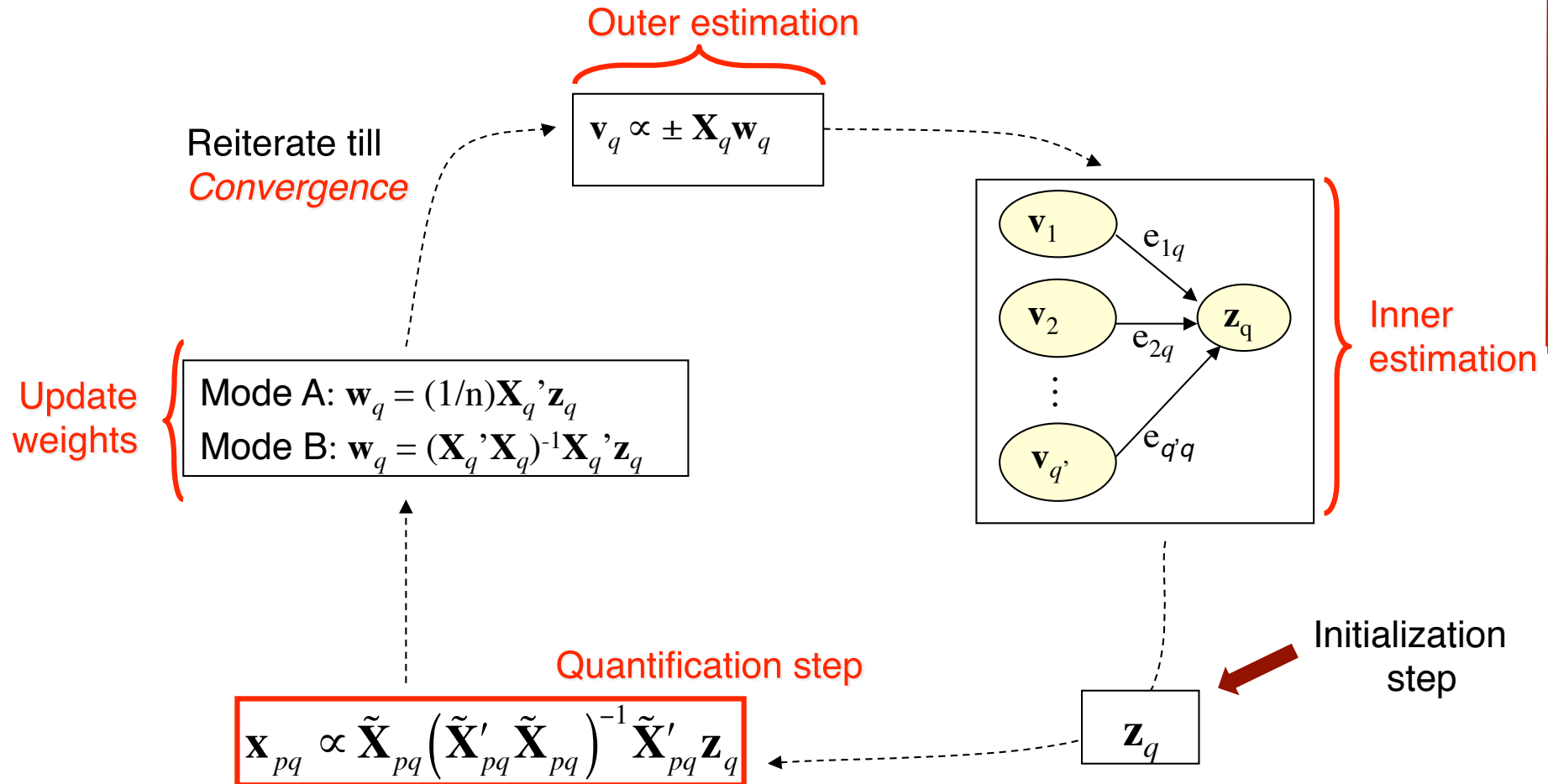
$$\text{cor}(\mathbf{x}_{pq}, \mathbf{z}_q) = \eta_{x_{pq}, z_q}^*$$

where \mathbf{x}_{pq} is the quantified indicator obtained as the normalized orthogonal projection of the inner estimate of the LV (\mathbf{z}_q) on the space spanned by the columns of $\tilde{\mathbf{X}}_{pq}$:

$$\mathbf{x}_{pq} \propto \tilde{\mathbf{X}}_{pq} \left(\tilde{\mathbf{X}}_{pq}' \tilde{\mathbf{X}}_{pq} \right)^{-1} \tilde{\mathbf{X}}_{pq}' \mathbf{z}_q$$

Categorical Variables as MVs

Modified PLS-PM algorithm to handle qualitative indicators



Real case example: data from RUSSET (1964)

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

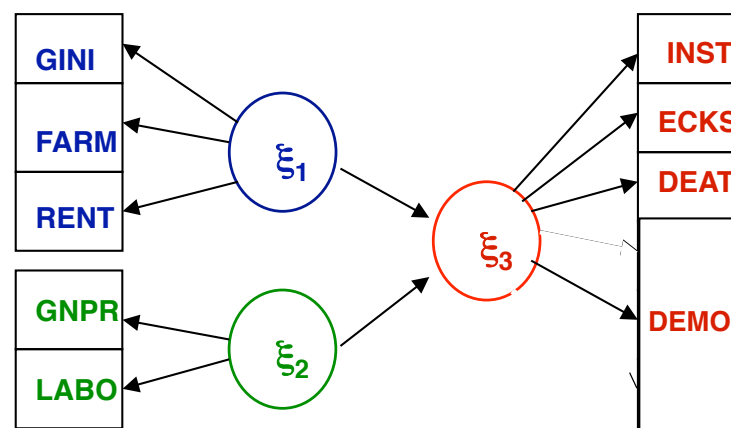
INST : Instability of executive (1945-1961)

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

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

DEMO : categorical variable with three levels

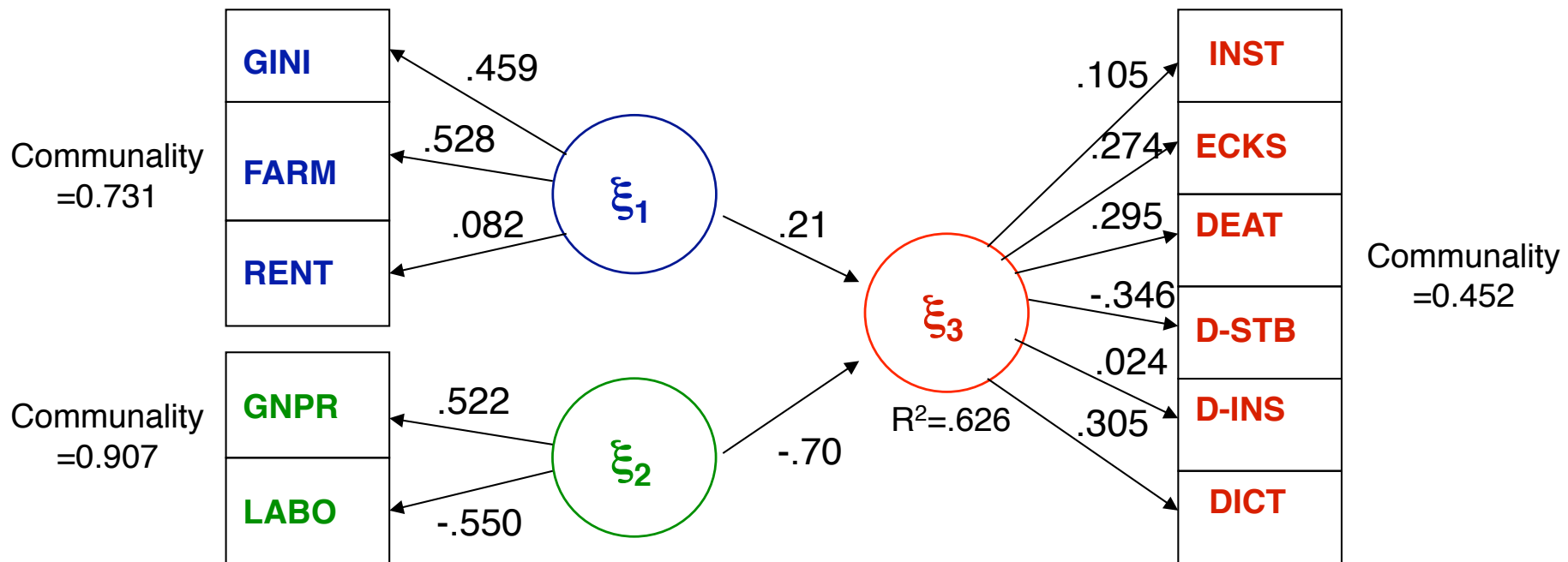
- **D-STAB** : Stable democracy
- **D-UNST** : Unstable democracy
- **DICT** : Dictatorship



Real case example: results for the model with dummy variables

GoF = 0.618

Agricultural Inequality



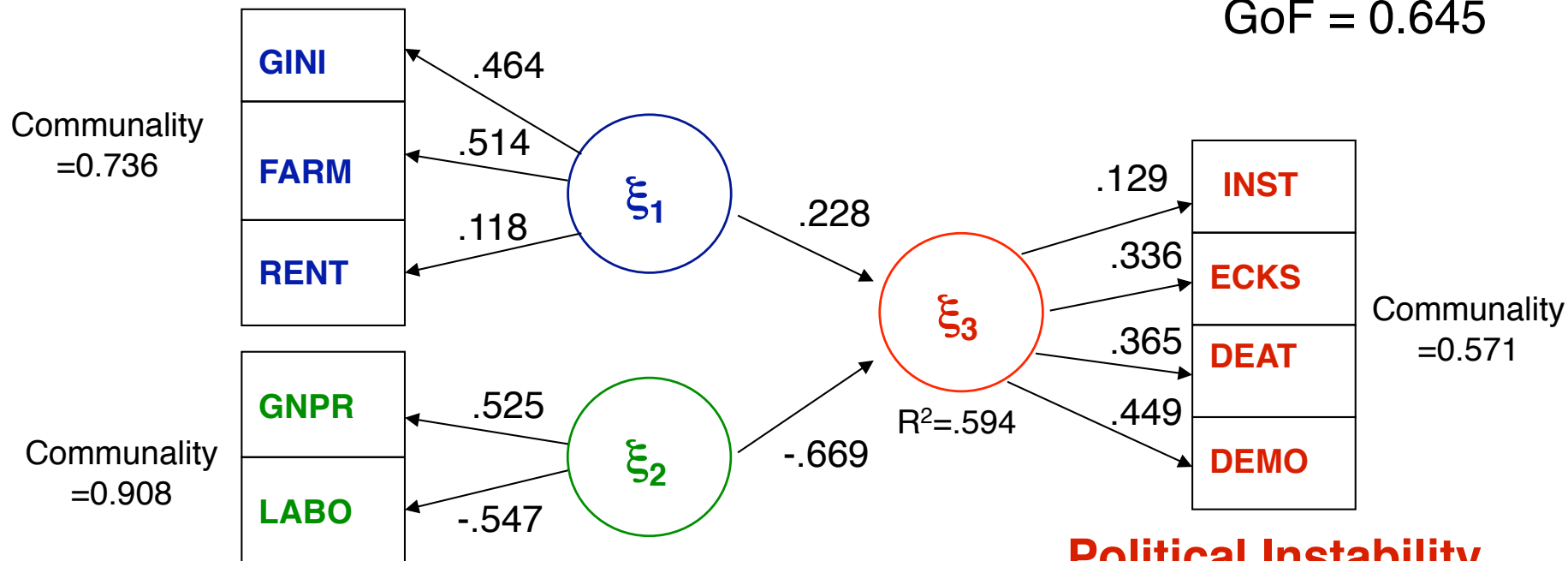
Industrial Development

Political Instability



Real case example: results for the model with the quantified variable

Agricultural Inequality



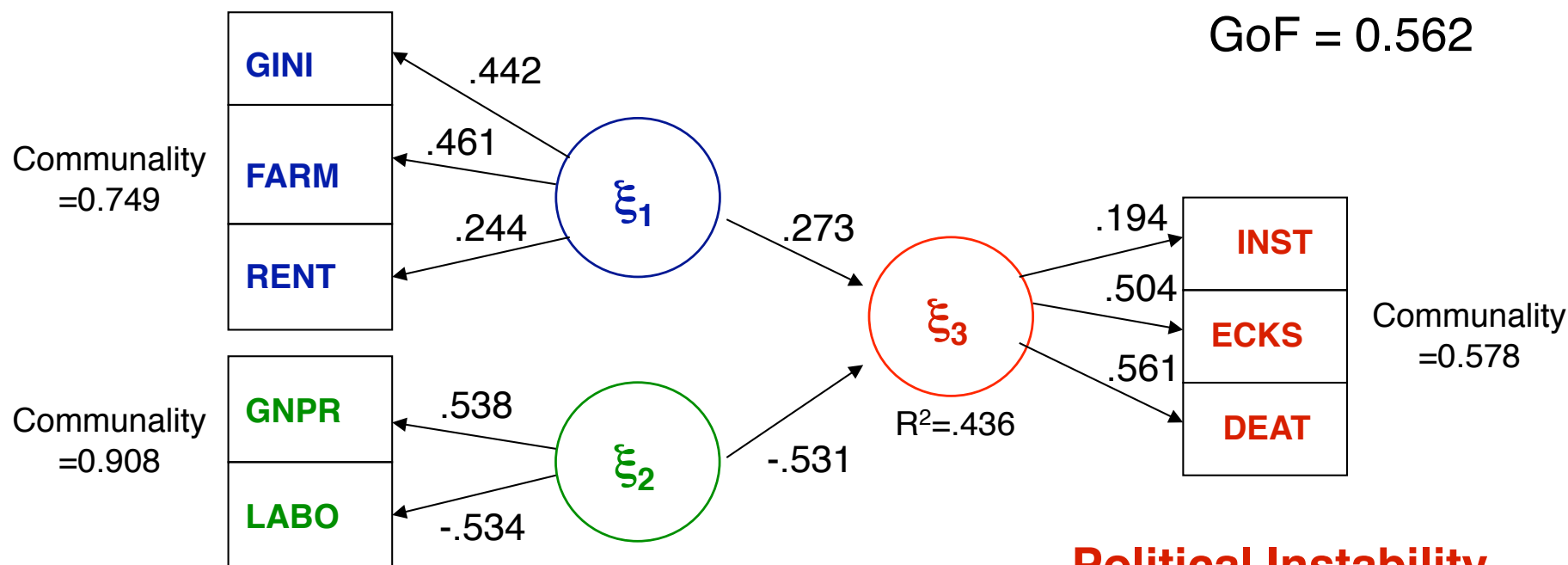
Industrial Development

Political Instability



Real case example: results for the model without DEMO

Agricultural Inequality



Industrial Development

Political Instability

Real case example: Conclusions

According to the results obtained through the modified PLS-PM algorithm

--> The **LV scores, i.e. the Cis**, are obtained as linear combination of the corresponding indicators, e.g. for the CI "Political Instability":

$$\text{Political instability } (\xi_3) = 0.13 \times \text{INST} + 0.34 \times \text{ECKS} + 0.37 \times \text{DEAT} + 0.45 \times \text{DEMO}$$

--> However, each CI can be seen also as a **complex indicator**, obtained according to the others Cis in the model, in particular each endogenous LV can be predicted also through the exogenous LVs and the path coefficients:

$$\text{Political instability } (\hat{\xi}_3) = 0.23 \times \text{Agricultural ineq.} (\xi_1) - 0.67 \times \text{Industrial dev.} (\xi_2)$$

To conclude, **categorical indicators** can be used as MVs in a reflective PLS-PM by means of our modified PLS-PM algorithm



Conclusions and perspectives

- ② The effect of the number of categories of a categorical indicator on the quantification need to be study
- ② The identification of the compromise model need to be further investigate
- ② Constrained PLS-PM should be developed in order to include a *priori* information on the weights defining composite indicators

Main references

1. **Baron R.M. and Kenny D.A.**, The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations, *Journal of Personality and Social Psychology*, 51 (6), 1173-1182 (1986).
2. **Bollen K. A.**, *Structural equations with latent variables*, Wiley, New York (1989).
3. **Chin W.W.**, A permutation procedure for multi-group comparison of PLS models, in *PLS and related methods - Proceedings of the International Symposium PLS'03*, M. Vilarés et al. (eds), DECISIA, 33-43 (2003).
4. **Escofier B. and Pagés J.**, Multiple factor analysis (AFMULT package), *Computational Statistics and Data Analysis*, 18,121-140 (1994).
5. **Esposito Vinzi V., Trinchera L., Squillacciotti S. and Tenenhaus M.**, REBUS-PLS: A Response-Based Procedure for detecting Unit Segments in PLS Path modeling, *Applied Stochastic Models in Business and Industry*, (2008).
6. **Hahn C., Johnson M., Herrmann A. and Huber F.**, Capturing Customer Heterogeneity using a Finite Mixture PLS Approach, *Schmalenbach Business Review*, 54, 243-269 (2002).
7. **Hensler J. and Fassott G.**, Testing moderating effects in PLS path models: An illustration of available procedure, in *Handbook of Partial Least Squares - Concepts, Methods and Applications*, V. Esposito Vinzi et al. (eds), Springer, Berlin, Heidelberg, New York (2009).
8. **Russolillo G.**, A proposal for handling categorical predictors in PLS regression framework. in *First joint meeting of the Socit Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society. Book of short papers*, Edizioni Scientifiche Italiane, 401 (2008).
9. **Saisana M. and Tarantola S.** *State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development*, EUR 20408 EN, European Commission-JRC: Italy (2002).
10. **Tenenhaus M., Esposito Vinzi V., Chatelin Y.-M. and Lauro C.**, PLS Path Modeling, *Computational Statistics and Data Analysis*, 48, 159-205 (2005).
11. **Trinchera L.**, Unobserved Heterogeneity in Structural Equation Models: a new approach in latent class detection in PLS Path Modeling, *PhD thesis, DMS, University of Naples* (2007).

