UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

Analisi delle Componenti Principali Lineare e Nonlineare: studio metodologico ed empirico

---

Francesca Ghiglia

Dipartimento di Economia e Management

Corso di LM in Management

Relatore

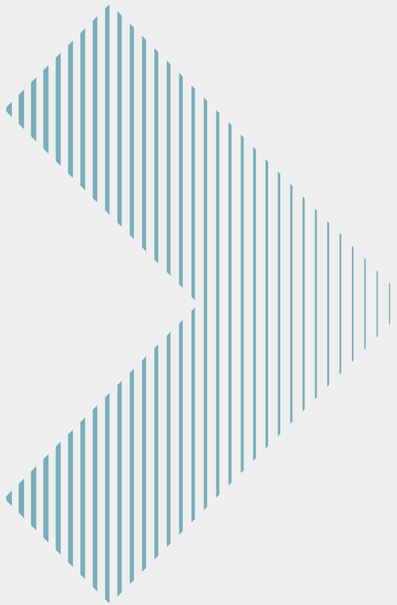
Chiar.ma Prof.ssa Paola Zuccolotto

Correlatore

Chiar.ma Prof.ssa Marica Manisera

INDICE DEI CONTENUTI

I. Riduzione di dimensionalità



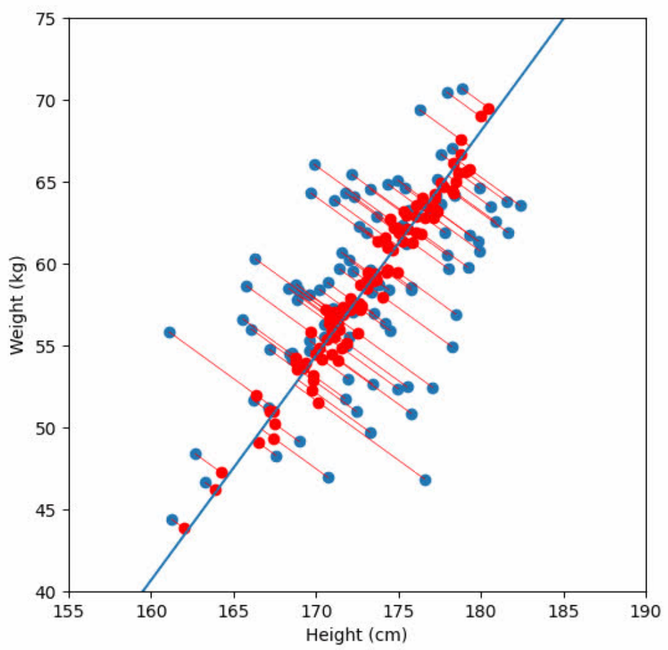
Cos’è la riduzione di dimensionalità (RD)?

È una procedura statistica che permette di ricavare caratteristiche informative utili da **dataset ad alta dimensionalità.** Si tratta di un **processo di riduzione** di un dataset ad alta dimensionalità in una rappresentazione di dimensione inferiore che conserva la maggior parte dell’informazione.

Perché è importante?

Presenza pervasiva dei **Big Data**: dataset le cui dimensioni eccedono la capacità dei tipici software di database di acquisire, archiviare, gestire e analizzare dati.

Cosa significa ridurre la dimensionalità?



In molti metodi di statistica multivariata ridurre la dimensionalità significa **proiettare** da uno spazio a un altro di dimensione inferiore.

La proiezione comporta inevitabilmente una **perdita di informazione.**

Tra tante possibili proiezioni ne esiste una che massimizza l'informazione trattenuta minimizzando quindi quella perduta.

Tecniche Lineari: PCA, MDS, LDA, ICA.

Tecniche Non Lineari: NLPCA, t–SNE, UMAP, ISOMAP.

II. Analisi delle Componenti Principali (PCA)

$$U = \begin{bmatrix} u'_1 \\ \vdots \\ u'_q \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ \vdots & \vdots & \dots & \vdots \\ u_{q1} & u_{q2} & \dots & u_{qp} \end{bmatrix}$$

Matrice di proiezione (ortogonale)

$$H = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ \vdots & \vdots & \dots & \vdots \\ u_{p1} & u_{p2} & \dots & u_{pp} \end{bmatrix}$$

Matrice degli autovettori normalizzati

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} = HX = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ \vdots & \vdots & \dots & \vdots \\ u_{p1} & u_{p2} & \dots & u_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

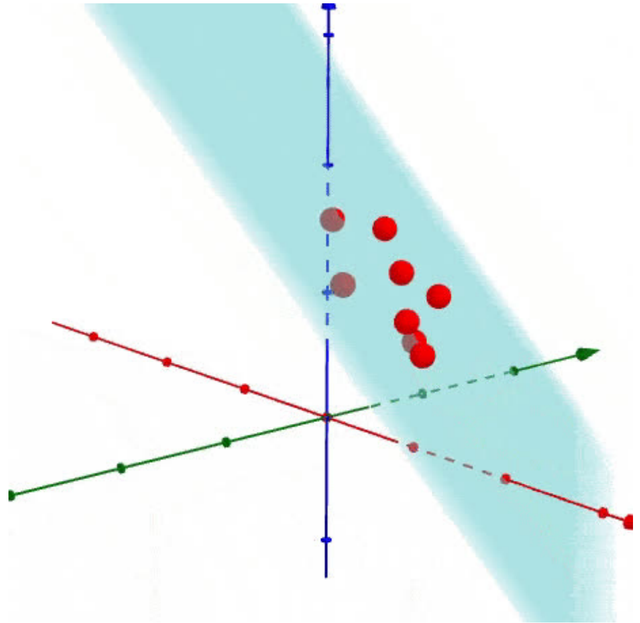
$$\Sigma_X = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \dots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \dots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \dots & \sigma_p^2 \end{bmatrix}$$

Matrice di varianze e covarianze di X

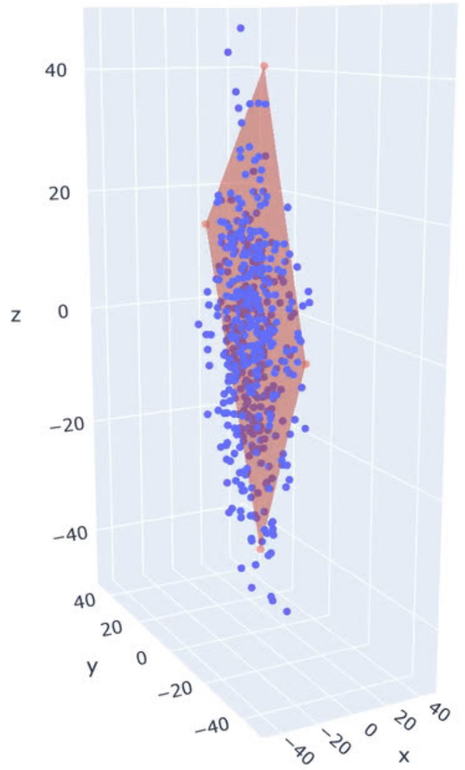
$$\Sigma_Y = H\Sigma_X H' = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Matrice di varianze e covarianze di Y

La PCA è una tecnica di RD lineare che converte un insieme di variabili correlate nello spazio ad alta dimensionalità in una serie di **variabili (PC) incorrelate** nello spazio a bassa dimensionalità. La riduzione della dimensionalità si ottiene attraverso le **proiezioni ortogonali**.



Svolgimento della PCA



Dati N soggetti su cui sono state osservate **p** variabili **quantitative**, l'obiettivo è determinare **q** nuove variabili (q≪p), che contengono gran parte dell'informazione che era contenuta nelle variabili iniziali.

L'applicazione della PCA comporta la risoluzione di due problemi fondamentali:

1. **valutare la qualità** della rappresentazione nello spazio a dimensione ridotta (scelta della dimensione ottimale in cui effettuare la riduzione della dimensionalità);

- **VARIANZA SPIEGATA**
- **AUTOVALORI**
- **SCREE PLOT**

2. **interpretare il significato** delle componenti principali.

- **FACTOR LOADINGS PLOT**
- **LOADINGS**

III. Analisi delle Componenti Principali Nonlineare (NLPCA)

Tecnica di analisi multivariata che si utilizza per ridurre la dimensionalità di una variabile p-dimensionale anche in presenza di **variabili qualitative.** Prevede la trasformazione delle modalità qualitative in valori numerici (quantificazioni) secondo un determinato criterio di ottimalità.

La NLPCA viene impiegata per misurare le **percezioni** (ad es. customer satisfaction), i cui dati vengono raccolti tramite la somministrazione di questionari con **batterie di item in scala di Likert.**

"Quanto si ritiene soddisfatto dei seguenti aspetti?"
“Quanto è d’accordo con le seguenti affermazioni?”

↓

VARIABILI QUALITATIVE IN SCALA ORDINALE

Osservazioni sulla NLPCA

SCALING LEVEL

- Nominal
- Ordinal
- Numerical

NON NESTEDNESS

Le soluzioni della NLPCA sono **non nested** (non annidate), cioè variano a seconda della dimensione dello spazio in cui si proietta la nube di punti.

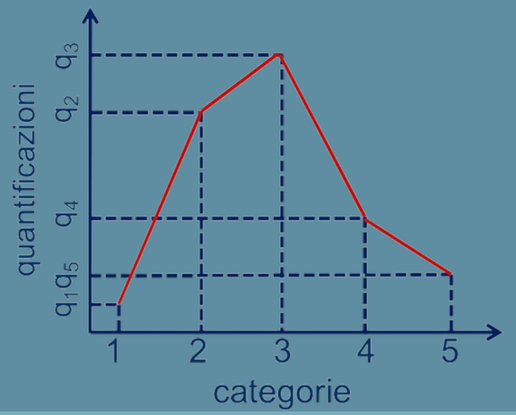
Preliminarmente si effettua l'**analisi full**, che non riduce la dimensionalità ma calcola solamente le quantificazioni ottimali.

Successivamente viene operata la riduzione della dimensionalità, eseguendo l'**analisi a dimensione ridotta.**

Osservazioni sulla NLPCA

SCALING LEVEL

- Nominal
- Ordinal
- Numerical



NON NESTEDNESS

Le soluzioni della NLPCA sono **non nested** (non annidate), cioè variano a seconda della dimensione dello spazio in cui si proietta la nube di punti.

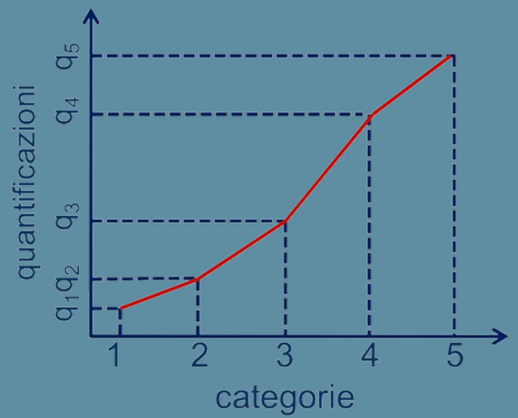
Preliminarmente si effettua l'**analisi full**, che non riduce la dimensionalità ma calcola solamente le quantificazioni ottimali.

Successivamente viene operata la riduzione della dimensionalità, eseguendo l'**analisi a dimensione ridotta.**

Osservazioni sulla NLPCA

SCALING LEVEL

- Nominal
- Ordinal
- Numerical



NON NESTEDNESS

Le soluzioni della NLPCA sono **non nested** (non annidate), cioè variano a seconda della dimensione dello spazio in cui si proietta la nube di punti.

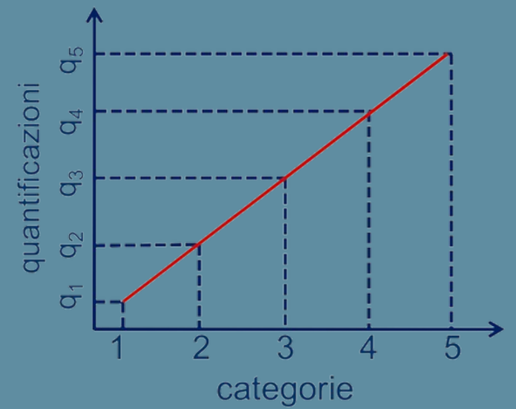
Preliminarmente si effettua l'**analisi full**, che non riduce la dimensionalità ma calcola solamente le quantificazioni ottimali.

Successivamente viene operata la riduzione della dimensionalità, eseguendo l'**analisi a dimensione ridotta.**

Osservazioni sulla NLPCA

SCALING LEVEL

- Nominal
- Ordinal
- Numerical



NON NESTEDNESS

Le soluzioni della NLPCA sono **non nested** (non annidate), cioè variano a seconda della dimensione dello spazio in cui si proietta la nube di punti.

Preliminarmente si effettua l'**analisi full**, che non riduce la dimensionalità ma calcola solamente le quantificazioni ottimali.

Successivamente viene operata la riduzione della dimensionalità, eseguendo l'**analisi a dimensione ridotta.**

IV. Esperimento

OBIETTIVO

Verificare l'eventuale differenza nelle risultanze ottenute quando si trattano le variabili in modo coerente con la loro **natura qualitativa** rispetto al caso in cui vengono forzate a una **natura quantitativa** che in realtà non hanno.

Nella fattispecie, ciò significa applicare al medesimo dataset la **NLPCA** (come è corretto fare) e la **PCA** (trascurando quindi la natura non quantitativa delle variabili).

APPLICAZIONE

Batteria di item in **scala di Likert a 5 punti**, contenuta nel questionario rivolto ai visitatori del **Museo di Santa Giulia – Progetto DS4BS** (BODaI-Lab), in cui veniva chiesto al rispondente di esprimere il proprio grado di accordo relativamente a 7 affermazioni (item) su una scala da "Per nulla" a "Moltissimo".

Sono stati raccolti **665 questionari.**

Con riferimento al Museo di Santa Giulia, quanto è d'accordo con le seguenti affermazioni?

	Per nulla	Poco	Abbastanza	Molto	Moltissimo
L'illuminazione valorizza le opere					
Il percorso espositivo è funzionale alla valorizzazione delle opere					
Il silenzio consente di riflettere e ammirare					
La presenza di aree di sosta (sedie, panche) consente di apprezzare meglio le opere					
La descrizione delle opere è precisa ed interessante					
La presenza di un percorso tattile permette di valorizzare le opere					
I contenuti multimediali sono coinvolgenti e aiutano capire i temi trattati					

Batteria di item in scala di Likert

Applicazione: parte 1

1

Scelta della dimensione ottimale q nella NLPCA

Scelta della dimensione ottimale q nella PCA

2

Scelta della dimensione ottimale q: **confronto tra NLPCA e PCA**

3

Interpretazione delle q componenti principali nella NLPCA

Interpretazione delle q componenti principali nella PCA

4

Interpretazione delle q componenti principali: **confronto tra NLPCA e PCA**

Scelta della dimensione ottimale q : confronto tra NLPCA e PCA

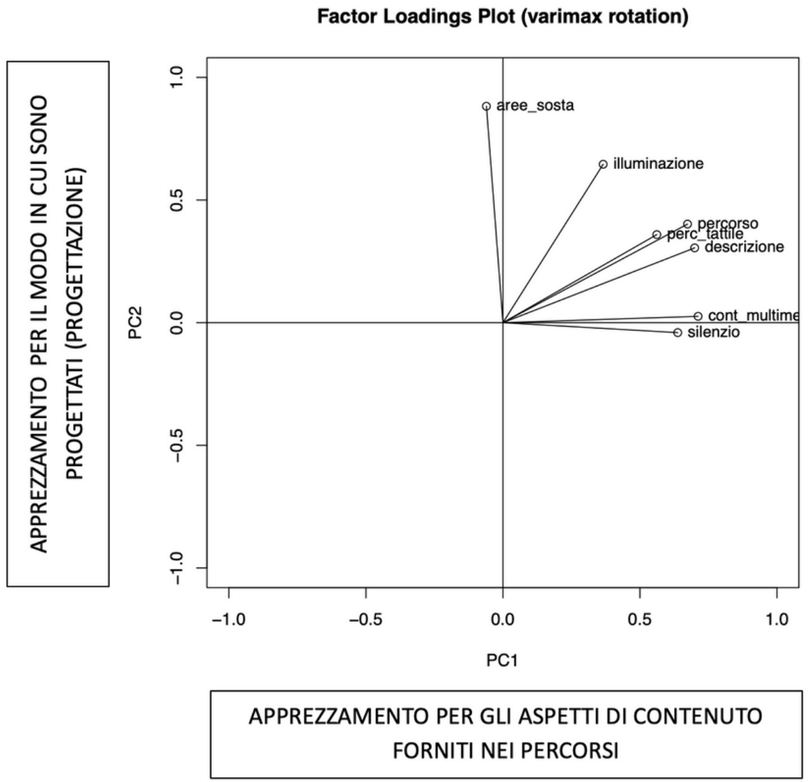
	Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
NLPCA (analisi full)	1	2,80	40,01%	40,01%
	2	0,93	13,24%	53,25%
	3	0,87	12,36%	65,61%
	4	0,79	11,27%	76,89%
	5	0,60	8,62%	85,50%
	6	0,56	8,02%	93,52%
	7	0,45	6,48%	100,00%
NLPCA (dim = 2)	Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
	1	2,88	41,15%	41,15%
	2	1,01	14,44%	55,59%
PCA	Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
	1	2,75	39,32%	39,32%
	2	0,97	13,91%	53,23%
	3	0,85	12,09%	65,32%
	4	0,80	11,37%	76,70%
	5	0,61	8,76%	85,46%
	6	0,57	8,15%	93,61%
	7	0,45	6,39%	100,00%

Sia per la NLPCA sia per la PCA, si è deciso di ridurre la dimensionalità a **2 dimensioni**, riassumendo i 7 item oggetto di indagine in 2 componenti principali (PC).

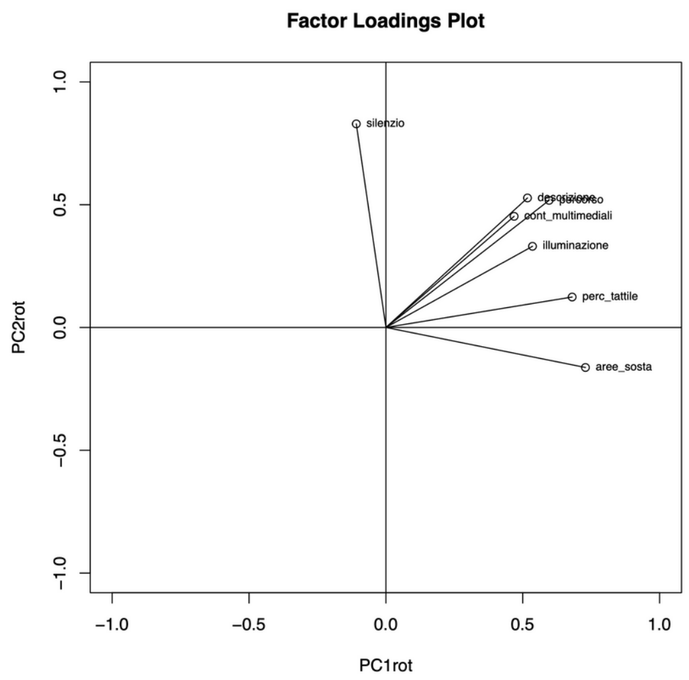
La NLPCA consente una varianza spiegata leggermente superiore rispetto alla PCA, come è nelle attese.

Interpretazione delle q componenti principali: confronto tra NLPCA e PCA

NLPCA



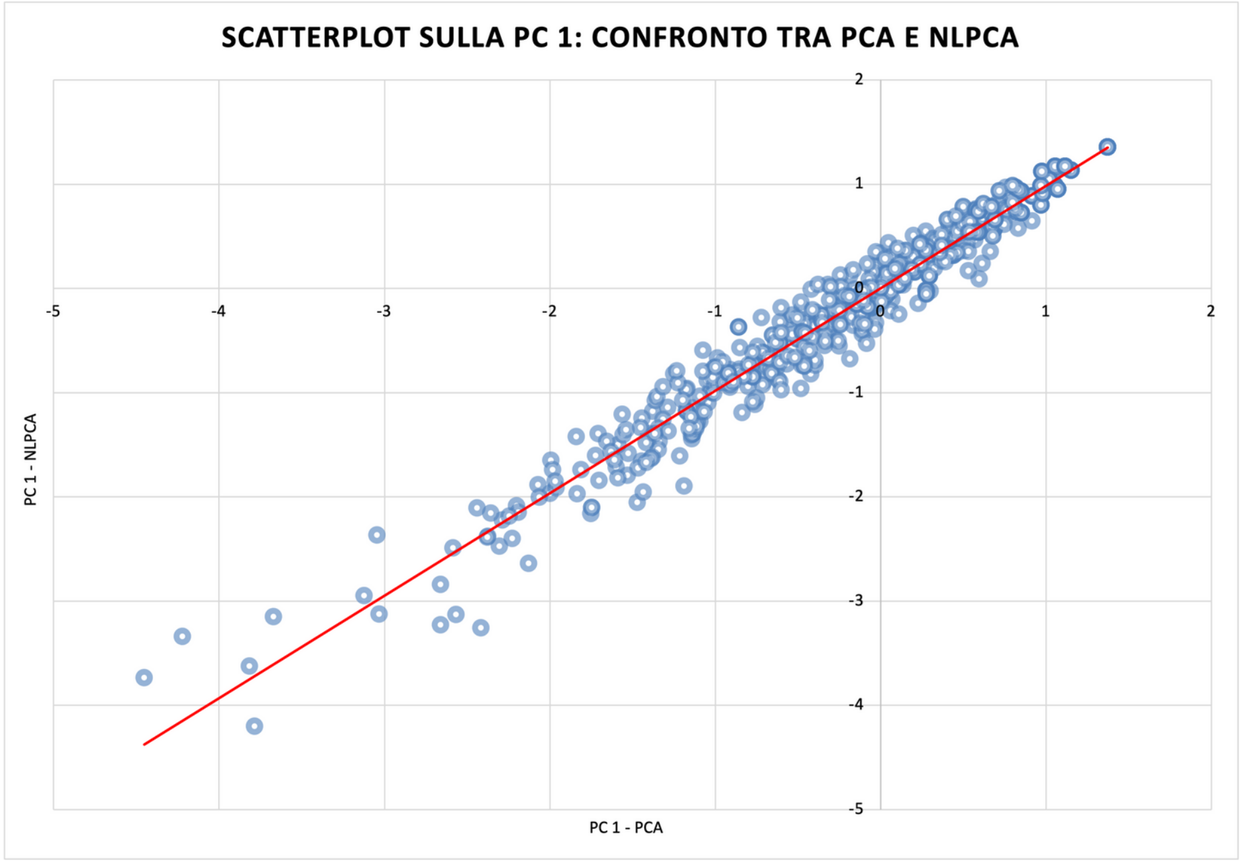
PCA



Dal confronto, emerge che applicando la NLPCA e la PCA si ottengono **2 PC differenti,** con significato diverso.

Parte 2: NLPCA e PCA in una dimensione

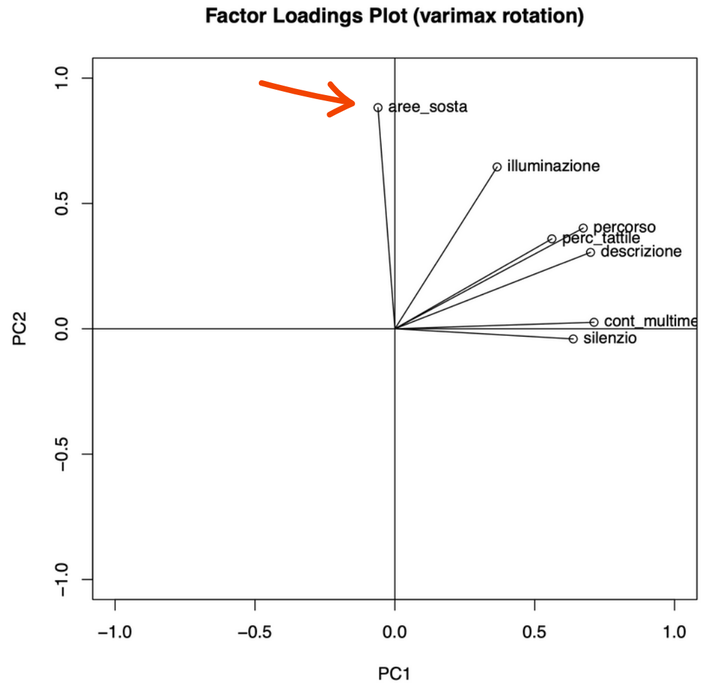
Successivamente, si è deciso di operare la riduzione in **una sola dimensione,** in questo modo forzando le due analisi ad avere un'unica componente principale con uguale significato, denominata il **grado complessivo di accordo.**



Coefficiente di correlazione lineare di Pearson ρ = 0,9830.

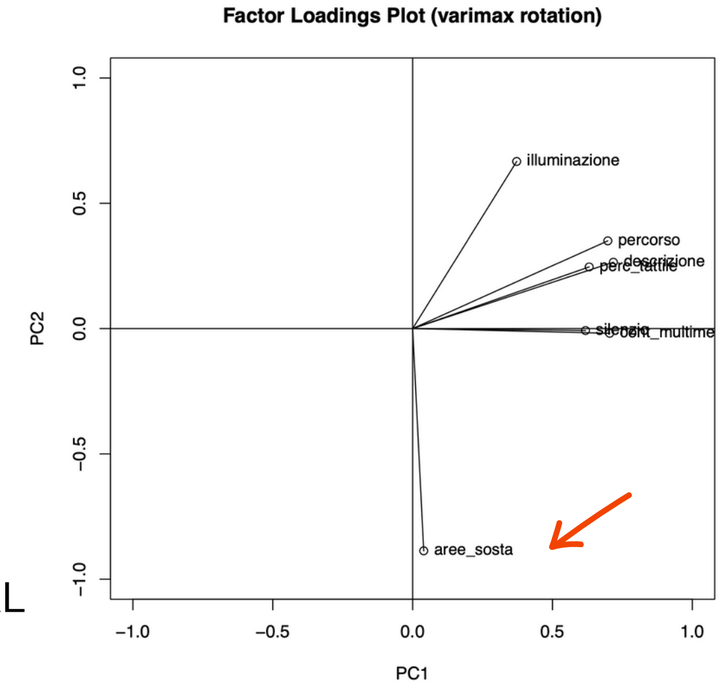
Parte 3: Confronto tra NLPCA in dim = 2 con scaling level Ordinal e Nominal

Le PC risultanti sono praticamente le stesse; a livello grafico, la posizione di "aree sosta" nel FLP Nominal è solo apparentemente ribaltata, e in realtà ha lo stesso significato che aveva nell'analisi Ordinal.

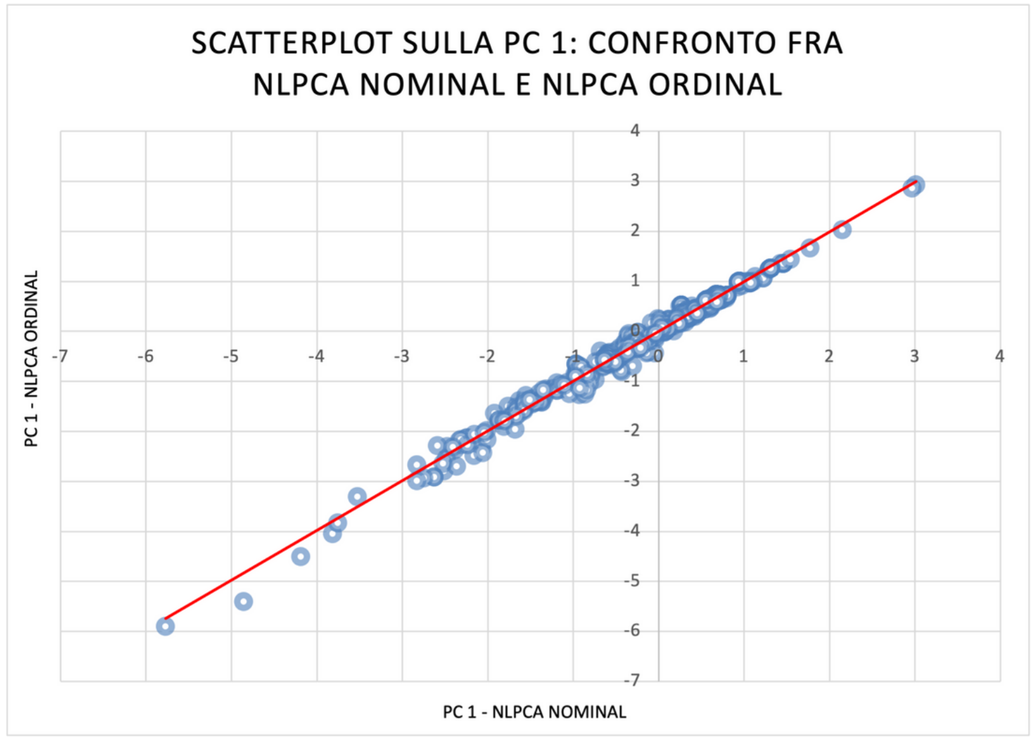


ORDINAL

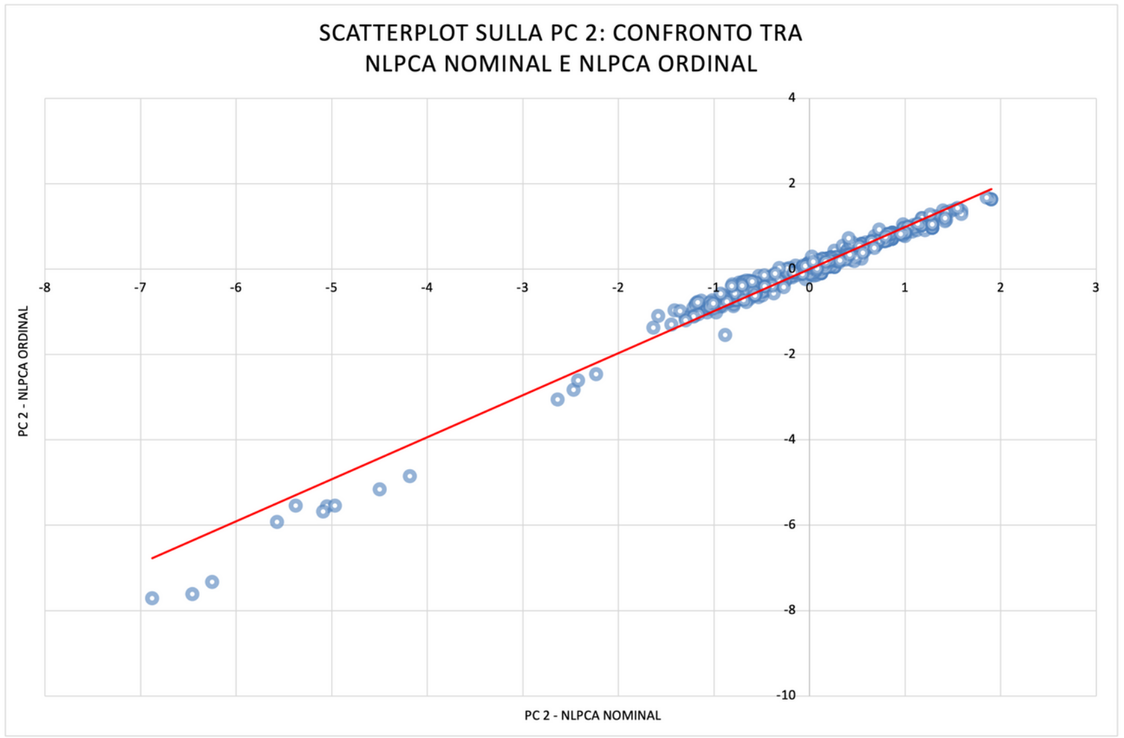
NOMINAL



Scatterplot dei valori delle 2 PC estratte con scaling Ordinal e Nominal: si generano PC 1 e PC 2 tra loro concordi.

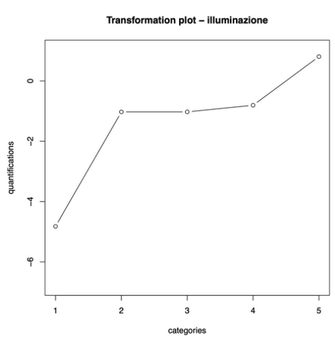


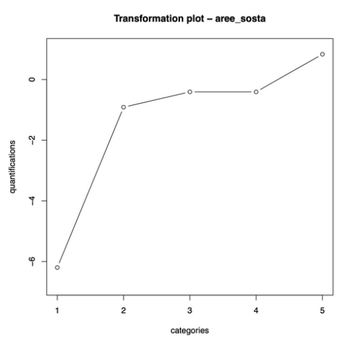
ρ = 0,9939

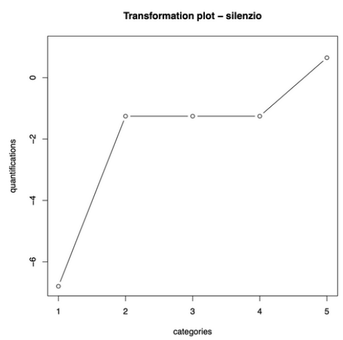


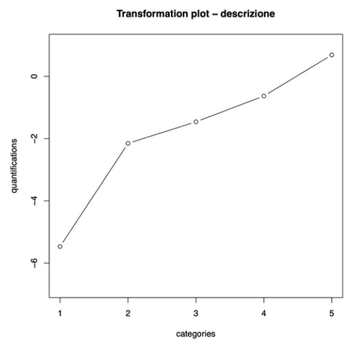
ρ = 0,9847

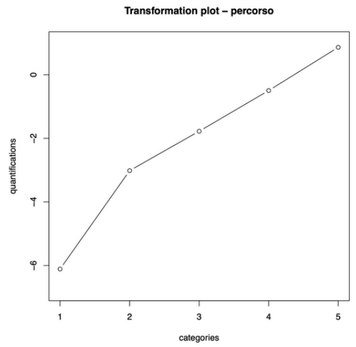
Transformation Plot (scaling Nominal): i risultati sono molto interessanti poiché confermano che le categorie sono sostanzialmente ordinate, quindi le risposte sono state date con una coerenza di fondo.

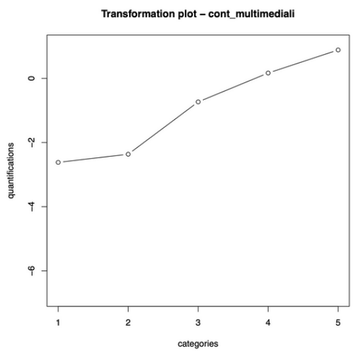


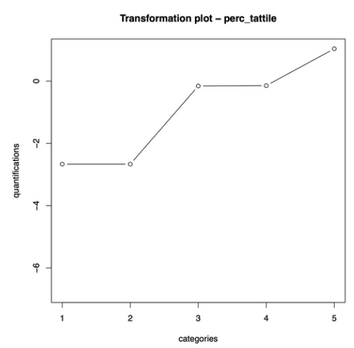




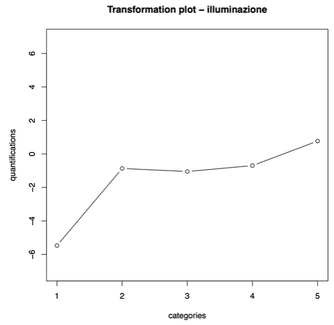


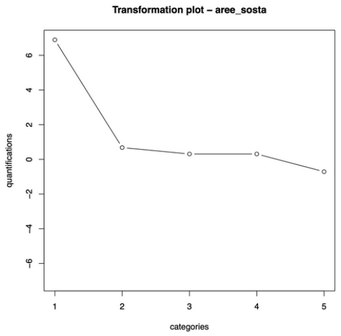


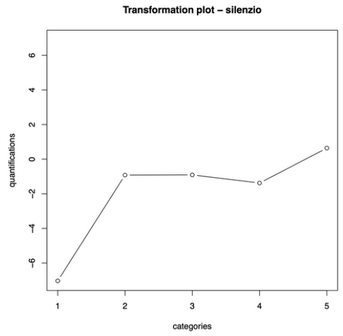


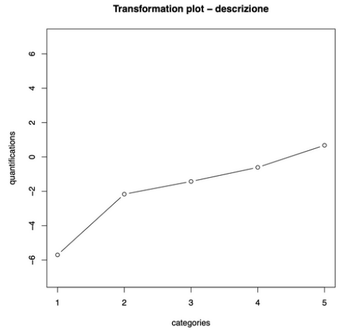


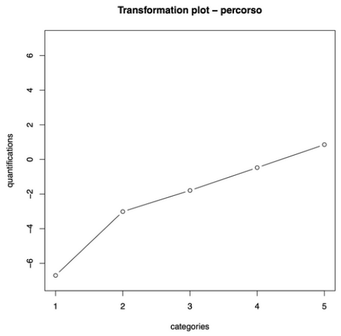
ORDINAL

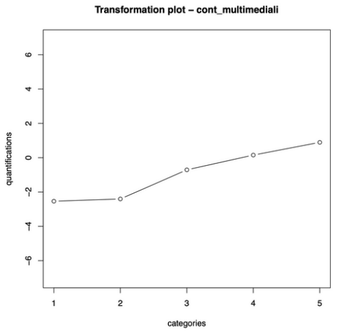


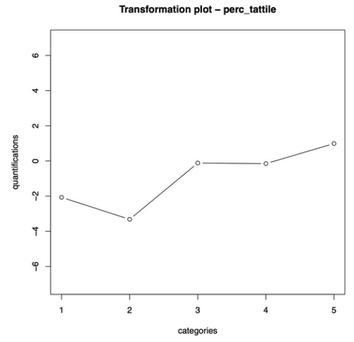












NOMINAL

V. Osservazioni conclusive

PCA O NLPCA?

La tecnica da utilizzare per trattare determinate variabili deve essere **scelta in modo coerente** con esse. Un'applicazione non rigorosa può produrre delle analisi imprecise o inesatte.

LETTERATURA VS PROVA EMPIRICA

La ricerca condotta ha portato a risultati che hanno **confermato le ricerche originali** riguardanti la NLPCA e la PCA nell'ambito della riduzione della dimensionalità.

LIMITI DELLA RICERCA

Questo esperimento mostra solo un esempio, i cui **risultati non sono generalizzabili.** Altri casi di studio potrebbero portare evidenze differenti.

Grazie per l’attenzione!

Francesca Ghiglia

Università degli Studi di Brescia