UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Corso di Laurea Magistrale
in Management

Tesi di Laurea

ANALISI DELLE COMPONENTI PRINCIPALI LINEARE E NONLINEARE: STUDIO METODOLOGICO ED EMPIRICO

Relatore: Chiar.ma Prof.ssa Paola Zuccolotto

Correlatore: Chiar.ma Prof.ssa Marica Manisera

Laureanda:
Francesca Ghiglia

Matricola n. 723739

Anno Accademico 2022/2023

Ai miei cari

INDICE

INDICE DELLE FIGURE

INDICE DELLE TABELLE

INTRODUZIONE

Nel mondo in continua evoluzione, una delle maggiori sfide che le aziende si trovano ad affrontare consiste nell'acquisire un vantaggio competitivo lavorando sui *big data*. Al giorno d'oggi si registra un'elevata pervasività di questa tipologia di dato in innumerevoli ambiti, soprattutto grazie alla presenza di tecnologie sempre più innovative che consentono l'archiviazione e l'elaborazione di enormi moli di dati. Tuttavia, ciò che è veramente rilevante non è tanto la quantità di dati, quanto ciò che l'azienda è in grado di realizzare con essi: per poterne realmente usufruire, i dati devono essere analizzati per ricavare delle informazioni di valore che aiutino a sviluppare decisioni aziendali migliori e strategie di business più efficaci.

In tale contesto, la riduzione della dimensionalità assume un ruolo fondamentale poiché facilita l'elaborazione di dati ad alta dimensionalità, ovvero di dataset con molte caratteristiche. Difatti le informazioni più importanti sono sommerse da un insieme di dati complessi e potenzialmente trascurabili, pertanto è necessaria un'appropriata metodologia al fine di isolarle ed estrarle.
La riduzione della dimensionalità può essere definita come una procedura statistica che consente l'analisi di grandi quantità di dati, permettendo di ricavare caratteristiche informative utili da dataset ad alta dimensionalità, che è il principale obiettivo della *data science*. Tale processo riduce la complessità, semplifica l'analisi e migliora l'efficienza computazionale delle attività ad alta intensità di dati.
La riduzione della dimensionalità, inoltre, aiuta i *data scientists* nel visualizzare relazioni complesse tra le variabili e nell'interpretazione e comprensione dei dati.
A tal fine sono state sviluppate diverse tecniche, tra cui l'Analisi delle Componenti Principali Lineare (PCA), l'Analisi delle Componenti Principali Nonlineare (NLPCA), l'Analisi Discriminante Lineare (LDA), l'Analisi delle Componenti Indipendenti (ICA), il Multidimensional Scaling (MDS).
La riduzione della dimensionalità viene comunemente utilizzata nei campi che trattano un gran numero di osservazioni o un gran numero di variabili, di conseguenza trova applicazione in una vasta gamma di discipline scientifiche, quali il marketing, l'informatica, la biologia, le scienze psico-sociali.

Il presente elaborato si configura come una tesi sperimentale nella quale verrà svolto un importante lavoro di studio e ricerca con particolare riferimento alla PCA e NLPCA, partendo dall'approfondimento della letteratura ed elaborando una propria osservazione originale.

La PCA e la NLPCA sono due tecniche di riduzione della dimensionalità molto simili, ma dispongono di una fondamentale differenza: la prima opera esclusivamente con variabili di tipo quantitativo, mentre la seconda tratta le variabili di natura qualitativa. Dunque, in primo luogo si presenterà una panoramica esaustiva dei fondamenti teorici relativi alla PCA, con una revisione della letteratura internazionale. Seguirà una breve esposizione riguardante la NLPCA. Questa prima parte fornirà una base solida per comprendere il contesto in cui si inserisce la ricerca, e quindi proseguire con l'indagine empirica.

In seguito, si procederà con lo studio tramite l'applicazione delle due tecniche a un caso studio, effettuando così l'esperimento pratico oggetto della tesi. I dati necessari per lo svolgimento dell'analisi sono stati ricavati tramite la somministrazione di un questionario ai visitatori del Museo di Santa Giulia nell'ambito del progetto "Data Science for Brescia – Arts and Cultural Places" (DS4BS), svolto dall'Università degli Studi di Brescia in collaborazione con il Comune di Brescia e la Fondazione Cariplo. Tale ricerca è volta ad analizzare la *visitor experience* nei luoghi culturali della città di Brescia, al fine di elaborare strategie di marketing utili a migliorare l'esperienza di visita degli utenti. Nel dettaglio, è risultata fondamentale la domanda relativa al grado di accordo del rispondente, in quanto tale quesito si configura come una batteria di item in scala di Likert a 5 punti, da cui si originano variabili qualitative in scala ordinale. Dunque, si applicheranno le tecniche NLPCA e PCA al medesimo dataset, che in questo caso richiederebbe l'utilizzo della NLPCA per una corretta elaborazione, tuttavia tramite questo esperimento si vogliono indagare empiricamente i risultati che si ottengono applicando sia la NLPCA (come è corretto fare) sia la PCA lineare (trascurando la natura non quantitativa delle variabili), e successivamente effettuare un confronto.

Gli obiettivi perseguiti tramite questa tesi sono molteplici: comprendere veramente a fondo il funzionamento della PCA, grazie a un importante approfondimento teorico,

esplorare i tratti metodologici fondamentali della NLPCA e, su tutto, lavorare su una batteria di item in scala di Likert comparando gli esiti ottenuti applicando le due tecniche. Inoltre, si confronterà quanto emerge dall'applicazione concreta con le evidenze teoriche, le quali sono ampiamente note grazie alle pubblicazioni di letteratura scientifica.
Complessivamente, quindi, si vuole investigare tramite un esperimento i limiti della PCA lineare; d'altra parte, lo studio permetterà di verificare l'effettivo miglioramento dell'analisi conseguibile tramite l'applicazione della NLPCA.

La tesi si compone di quattro capitoli. Nel primo capitolo si introdurrà il concetto di riduzione della dimensionalità, esplorandone il significato, l'utilità e le considerazioni che è importante ricordare quando si opera tale procedura. A tal fine, si farà riferimento alla letteratura in campo internazionale della materia oggetto di studio. Inoltre, ad anticipare tale argomento verrà effettuata una breve disamina sull'importanza attuale dei big data e della data science.

Nel secondo capitolo verrà ampiamente esplorata la PCA, analizzando innanzitutto il processo di derivazione algebrica della PCA, e, dunque, esponendo i concetti principali di algebra delle matrici. In seguito, si proseguirà con l'approfondimento teorico: anche in questo caso la letteratura scientifica offre un importante contributo per la trattazione dell'argomento. Tale approfondimento teorico è utile per comprendere completamente il motivo per cui la PCA è intrinsecamente connessa all'uso di variabili quantitative. Infine, si presenterà un breve excursus sulla NLPCA, soffermandosi sui punti di diversità dalla PCA che la rendono idonea a trattare anche le variabili qualitative.

Nel terzo capitolo verrà effettuato l'esperimento pratico, che rappresenta il corpo centrale della tesi di ricerca. In particolare, si esporrà l'applicazione della NLPCA e PCA al caso studio, con le relative risultanze e conseguenti osservazioni. Il confronto tra NLPCA e PCA avverrà sui dati di una scala di Likert.

Nel quarto e ultimo capitolo si realizzerà un riepilogo di ciò che è stato svolto tramite l'esperimento e si discuteranno gli esiti dell'applicazione delle due tecniche, elaborando

così le conclusioni finali. Lo studio condotto non necessariamente deve portare a risultati innovativi ma può confermare o arricchire ricerche originali della materia in esame.

CAPITOLO 1

LA RIDUZIONE DI DIMENSIONALITÀ

1.1 Big data e data science

Al giorno d'oggi, con il progresso sempre più rapido delle tecnologie moderne, vengono generate e archiviate enormi quantità di dati.

Quando tali dati soddisfano determinati requisiti in termini di volume, varietà e velocità di aggiornamento, essi vengono indicati con il termine inglese *big data* (in italiano *megadati*). Secondo la definizione proposta dal McKinsey Global Institute, i big data si riferiscono a dataset le cui dimensioni eccedono la capacità dei tipici software di database di acquisire, archiviare, gestire e analizzare dati[1]. In altre parole, big data è un termine che si riferisce ai set di dati troppo grandi o troppo complessi per essere gestiti dalle tradizionali tecniche di elaborazione e gestione dei dati.

Ad oggi, i big data sono risorse dal potenziale enorme, in quanto offrono notevoli opportunità e benefici per le imprese, quali l'aumento dell'efficienza operativa e la riduzione dei costi. In tal senso, la data science occupa un ruolo fondamentale, qualificandosi come mezzo principale per poter scoprire e sfruttare quel potenziale.

La *data science* è la scienza dei dati, intesa come studio scientifico dei dati per ottenerne conoscenza. È un approccio multidisciplinare che combina matematica e statistica, programmazione specializzata, analytics avanzata, intelligenza artificiale e machine learning al fine di estrarre informazioni di valore da set di dati allo scopo di prendere decisioni informate[2]. Sostanzialmente, la data science raccoglie tutte quelle discipline volte alla preparazione, elaborazione e analisi dei dati.

L'importanza della data science sta aumentando rapidamente man mano che la quantità di dati cresce in modo esponenziale; allo stesso tempo, le aziende dipendono maggiormente dall'analisi dei dati per favorire la propria produttività e innovazione.

---

[1] McKinsey Global Institute (2011), *Big data: The next frontier for innovation, competition, and productivity*

[2] https://www.ibm.com/it-it/topics/data-science

Ad esempio, si consideri come le interazioni tra azienda e cliente stiano diventando sempre più digitali; in questo modo, viene creato un maggior quantitativo di dati, il che comporta nuove opportunità per ricavare informazioni utili su come personalizzare al meglio la *customer experience*, migliorare la soddisfazione di servizi e clienti, sviluppare prodotti nuovi e, in ultima istanza, aumentare le vendite.

I big data stanno diventando sempre più un fattore chiave della competizione, poiché il loro impiego consapevole permette di conseguire un importante vantaggio competitivo per le imprese. A conferma di ciò, da tempo si afferma che "i dati sono il nuovo petrolio", andando così a rimarcare la sempre maggiore centralità e rilevanza dell'utilizzo dei dati per condurre un'efficace gestione aziendale.
La rivista britannica *The Economist*, già anni fa, ha attribuito particolare importanza all'argomento, dedicandovi la copertina del numero del 6 maggio 2017.

Figura 1.1: Copertina del The Economist, numero del 6 maggio 2017
(Fonte: The Economist)

Come mostrato in Figura 1.1, secondo il *The Economist*, oggi i dati sono la risorsa più preziosa del mondo[3].

[3] The Economist (2017). *The World's Most Valuable Resource Is No Longer Oil, but Data*

Esistono diverse modalità per trarre profitto dai dati, tuttavia è fondamentale considerare le molteplici difficoltà che possono essere riscontrate approcciando l'analisi dei dati. Ogni azienda necessita di una strategia sui dati, che sia completamente integrata con la propria strategia aziendale complessiva. La strategia dovrebbe abbracciare la posizione industriale attuale e desiderata dell'azienda, il panorama competitivo, come e dove l'azienda intende competere con i dati, i dati proprietari, il personale e la tolleranza rispetto al rischio d'impresa[4].

L'auspicio dei *data scientists* è che la data science continui a evolversi, con metodi sempre più perfezionati per analizzare i dati, estrarre le informazioni essenziali, interpretare, presentare e riassumere i risultati, al fine di condurre un'inferenza statistica più corretta. Questi metodi di analisi richiedono, a loro volta, nuove tecniche per raccogliere, manipolare, archiviare e ordinare i dati.
A tal fine, risulta di particolare importanza il continuo sviluppo dell'intelligenza artificiale, del machine learning e del deep learning.

In conclusione, è importante considerare che, nonostante l'utilità di questi strumenti, molte imprese e i rispettivi manager mostrano ancora un certo scetticismo sui concreti benefici che tali tecnologie possono apportare.
I big data sono destinati a cambiare irrevocabilmente il modo di essere e di fare di un'impresa, dunque è necessario che le organizzazioni si strutturino adeguatamente non solo per poterne beneficiare, ma anche per non perdere competitività nei confronti della concorrenza.

1.1.1 Il mercato italiano dei big data

Secondo le stime contenute nel report annuale 2022 dell'Osservatorio Big Data & Business Analytics della School of Management del Politecnico di Milano, il mercato italiano dei big data ha raggiunto un valore pari a 2,41 miliardi di euro, in crescita del +20% rispetto al 2021 (Figura 1.2). Le tensioni geopolitiche e gli alti tassi di inflazione dovuti alla guerra tra Russia e Ucraina, dunque, non hanno posto gravi ostacoli

[4] Kenett, R. S., Redman, T. C., Manzi, G., & Salini, S. (2021). *La data science nella realtà: come trasformare i dati in informazioni, decisioni migliori e organizzazioni più forti*. Giappichelli.

all'evoluzione del mercato Data Management e Analytics in Italia, che, al contrario, si è ripreso dopo la battuta d'arresto dovuta alla pandemia.

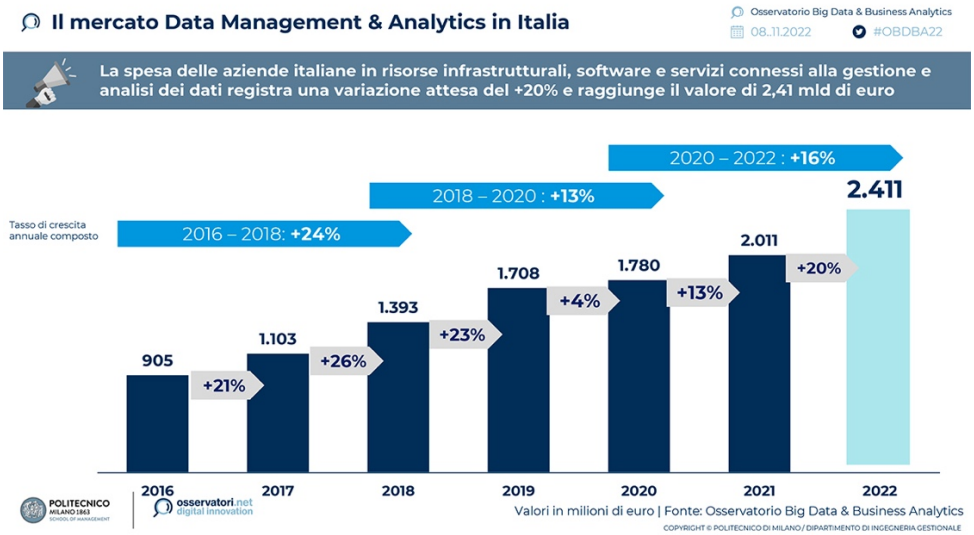


Figura 1.2: Il mercato Data Management & Analytics in Italia
(Fonte: Osservatorio Big Data & Business Analytics – Politecnico di Milano)

A trainare la crescita è soprattutto la componente software, che vale il 54% del mercato e risulta in aumento del +25% rispetto al 2021. Nelle grandi aziende, si registra un incremento al 65% della sperimentazione di *advanced analytics* (rispetto al 54% del 2021); inoltre, figure di Data Scientist sono presenti nel 49% delle grandi organizzazioni, Data Engineer nel 59% e Data Analyst nel 76%[5].

La ricerca conferma quindi il persistere delle difficoltà a inserire ruoli professionali specializzati nell'ambito della gestione e analisi dei dati.

Per quanto riguarda le PMI, il 55% ha dichiarato di aver condotto investimenti in ambito *data management & analytics,* o comunque prevede di farlo nel prossimo futuro. Si tratta di una quota in crescita rispetto al 2021, ma che non mostra importanti accelerazioni rispetto a quanto registrato nel triennio precedente. Inoltre, quattro aziende su dieci affermano di non avere alcuna figura dedicata, neanche parzialmente, all'analisi dei dati. Sul fronte della data science si è conseguito un aumento delle imprese che hanno avviato almeno una sperimentazione in ambito *advanced analytics* (con una quota pari al 65%,

[5] Osservatorio Big Data & Analytics (2022). *Data-driven culture: connettere algoritmi e persone*. Politecnico di Milano.

rispetto al 54% del 2021). Le funzioni aziendali in cui la data science trova maggiore applicazione sono Marketing, Vendite e Produzione.

Contestualmente alla pubblicazione di tale resoconto, Carlo Vercellis, Responsabile Scientifico dell'Osservatorio Big data & Business Analytics, ha dichiarato che "nonostante le difficoltà dello scenario globale, nel 2022 le imprese italiane continuano a mostrare grande interesse per gli Analytics. Cresce la maturità delle organizzazioni verso una cultura data-science-driven e insieme il mercato, che ha lasciato alle spalle il periodo nero. Ma la sfida di chi ha avviato sperimentazioni o progetti di Advanced Analytics ora è quella dell'industrializzazione dei processi per garantire efficienza e governance dei dati in tutti i livelli"[6].

Alessandro Piva, Responsabile della Ricerca dell'Osservatorio Big Data & Business Analytics, ha affermato che "la spesa delle aziende italiane è tornata stabilmente a crescere, ancor più velocemente per le realtà più in ritardo, mentre si consolidano i progetti delle aziende più mature. Ma il forte interesse per le soluzioni di analytics non corrisponde sempre a un cambio di rotta complessivo: sono ancora una minoranza le organizzazioni con una Data Strategy di livello corporate. Ora è necessario trasformare le organizzazioni nel profondo, creando ponti tecnologici, organizzativi e culturali tra le opportunità di analisi avanzate, le applicazioni intelligenti e le competenze e attività quotidiane dei lavoratori"[7].

A conclusione della ricerca, l'Osservatorio ha creato un indice di maturità complessivo denominato *Data Strategy Index*, che indica come solo il 15% delle grandi aziende può dirsi "avanzato", mentre il 30% "intraprendenti", il 22% "prudenti" e il 33% "immature" o "ai primi passi".

1.1.2 Il mercato globale dei big data

A livello mondiale, secondo il report del Market Data Forecast, pubblicato nel mese di marzo 2023, si prevede che la dimensione del mercato globale dei big data passerà dai

[6] https://www.osservatori.net/it/ricerche/comunicati-stampa//mercato-big-data-crescita

[7] https://www.osservatori.net/it/ricerche/comunicati-stampa//mercato-big-data-crescita

157,9 miliardi di dollari del 2020 ai 268,4 miliardi di dollari entro il 2027, con un CAGR[8] pari a circa il 12% nel periodo di previsione[9].

Si consideri che questa stima evidenzia un aumento significativo rispetto alle analoghe previsioni elaborate da Statista nel 2018. Come rappresentato in Figura 1.3, infatti, si era stimato un valore del mercato globale dei big data che, dai 42 miliardi di dollari del 2018, avrebbe raggiunto i 103 miliardi di dollari del 2027[10].
Pertanto, le stime per il 2027 del Data Market Forecast sono di quasi tre volte superiori rispetto a quelle elaborate da Statista cinque anni in precedenza.

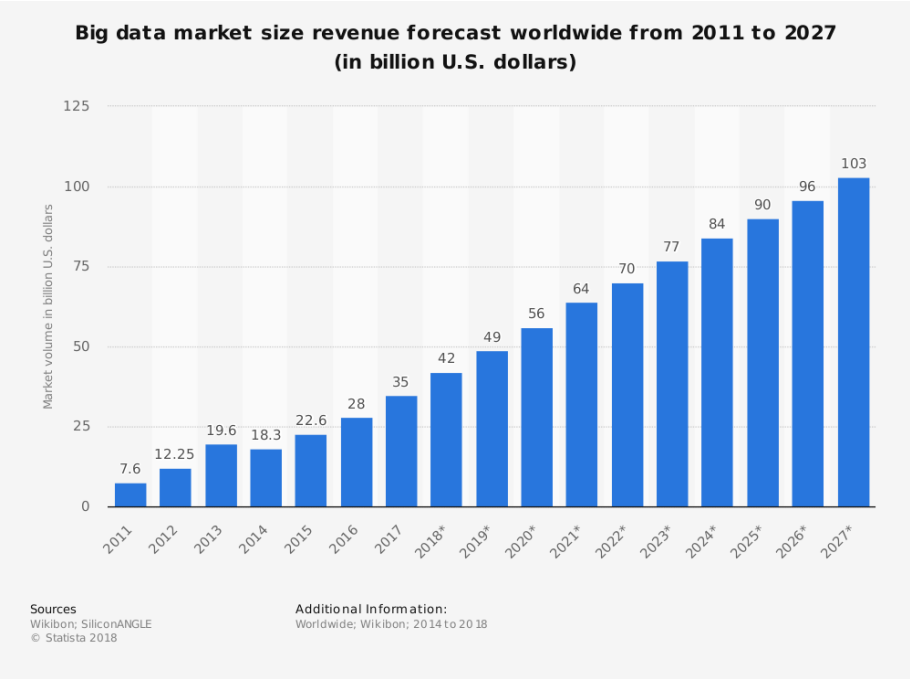


Figura 1.3: Previsioni sulle entrate relative al mercato dei big data in tutto il mondo dal 2011 al 2027 (Fonte: Statista)

[8] Il tasso annuo di crescita composto, più comunemente noto come CAGR dall'acronimo anglosassone Compounded Average Growth Rate, rappresenta la crescita percentuale media di una grandezza in un lasso di tempo.

[9] https://www.marketdataforecast.com/market-reports/big-data-market

[10] https://www.statista.com/statistics/254266/global-big-data-market-forecast/

Questo dimostra come negli ultimi anni vi sia stato un incremento notevole nell'ecosistema complessivo dei big data. Si tratta, infatti, di un mercato in continua evoluzione, che deve essere monitorato e sfruttato sia da aziende grandi sia da quelle più piccole.
L'implementazione dei big data consente di trovare il giusto equilibrio tra costi operativi, velocità, flessibilità e qualità. Molte imprese, oggi, applicano soluzioni e servizi di questo tipo per valutare i propri processi interni e migliorare le proprie operazioni.

1.2 I dati ad alta dimensionalità

Oggi la crescita e la velocità di aggiornamento dei dataset stanno accelerando e i dati si stanno sviluppando in una direzione altamente dimensionale e non strutturata.
I dataset con molte caratteristiche sono chiamati dati ad alta dimensionalità (Jia *et al*., 2022). Nel contesto dell'analisi dei dati, situazioni che contemplano dati ad alta dimensionalità possono verificarsi frequentemente e possono portare a numerose sfide nell'ambito della data science (Clarke *et al*., 2008).

Per definire il concetto di dati ad alta dimensionalità si consideri che, in generale, ogni dataset è caratterizzato da due parametri. Il primo è il numero di osservazioni statistiche, ovvero il numero di unità statistiche: si indichi questo parametro con N. Il secondo è il numero delle caratteristiche/variabili descrittive dei dati: si denoti questo parametro come p. In particolare, p può essere interpretato come il numero di tratti o attributi correlati al fenomeno oggetto dell'analisi, mentre N riflette il numero di volte in cui ogni tratto o attributo è stato osservato durante lo studio (Oskolkov, 2022).
In molte applicazioni il numero di variabili p è molto grande, e può anche essere maggiore del numero di unità statistiche osservate N. Una tale elevata dimensionalità e la particolare configurazione caratterizzata da un grande p e un piccolo N hanno posto nuove sfide ai metodi di analisi statistica (Li *et al*., 2009).
Secondo quanto è emerso da alcune ricerche, se N risulta troppo piccolo, aumentare p può comportare delle singolarità e divergenze nelle equazioni matematiche alla base dell'analisi statistica tradizionale (Altman & Krzywinski, 2018). In particolare, per l'ineguaglianza $p \gg N$ (che si legge come "p molto maggiore di N", ovvero p è almeno

~ 10 volte maggiore di N) l'assunzione di distribuzione normale, che è tipica della statistica matematica classica, perde validità e può portare a conclusioni scientifiche fuorvianti.

In altre parole, la difficoltà nel lavorare con dati ad alta dimensionalità è principalmente riflessa nello scarso rendimento dei modelli statistici comuni a causa delle violazioni delle assunzioni fondamentali dei modelli: questo problema è tradizionalmente noto come *The Curse of Dimensionality* (in italiano *La Maledizione della Dimensionalità*), e appare come conseguenza dell'espansione delle dimensioni sempre più rapida e su larga scala (Altman & Krzywinski, 2018).
Spesso, i dati sono caratterizzati da elevati volume e complessità, includendo un gran numero di variabili; tuttavia, non tutte le caratteristiche sono necessarie o addirittura utili per risolvere un determinato problema. Ad esempio, le caratteristiche possono essere equivalenti, ridondanti (ovvero contenere informazioni derivabili da altre), oppure alcune possono risultare addirittura irrilevanti per la soluzione (Mainali *et al*., 2021).

Quanto illustrato precedentemente, dunque, si può riassumere osservando come i dati voluminosi e complessi contengono molte informazioni importanti, ma allo stesso tempo aumenta anche la difficoltà nell'utilizzarli in modo efficace. L'elevata dimensionalità dei big data, infatti, rende particolarmente impegnativo recuperare le informazioni utili dai dati a disposizione.

Trovare un modo per superare *The Curse of Dimensionality,* che si verifica nei dati ad alta dimensionalità, è oggi una delle sfide più importanti che il campo della data science si trova ad affrontare. La riduzione di dimensionalità è un metodo utilizzato al fine di mitigare questo problema, in quanto consente il pre-processamento di dati ad alta dimensionalità (Oskolkov, 2022).
La riduzione di dimensionalità è una tecnica standard dell'Analisi Esplorativa dei Dati (EDA), che è tipicamente considerata essere il primo passo nell'analisi dei dati.

L'Analisi Esplorativa dei Dati (EDA) è un approccio all'analisi dei dati che utilizza una varietà di tecniche per riassumere le principali caratteristiche di un dataset, spesso utilizzando metodi visivi (Nanaware *et al*., 2018).

L'EDA è utile per una serie di scopi, i più importanti dei quali sono: massimizzare la comprensione di un insieme di dati, mappare la struttura sottostante dei dati, identificare variabili utili, rilevare *outliers* e anomalie, e testare ipotesi.

1.2.1 The Curse of Dimensionality

Il presente sottoparagrafo è mirato a presentare un breve approfondimento su *The Curse of Dimensionality*, per poi proseguire con la trattazione della riduzione di dimensionalità.

The Curse of Dimensionality (ovvero *La Maledizione della Dimensionalità*) è un'espressione coniata dal matematico statunitense Richard E. Bellman, per descrivere l'aumento di volume dello spazio euclideo associato all'aggiunta di dimensioni extra, nell'area della programmazione dinamica. Oggi, questo fenomeno si osserva in campi come l'apprendimento automatico, l'analisi dei dati e il data mining, tra gli altri.
Nell'analisi dei dati, il termine si riferisce alla difficoltà di trovare strutture nascoste quando il numero di variabili è elevato; in altre parole, tale problema si verifica durante l'analisi e l'organizzazione dei dati nello spazio ad alta dimensione. Infatti, un incremento delle dimensioni può, in teoria, aggiungere più informazioni ai dati migliorandone così la qualità, ma in pratica raramente aiuta a causa della maggiore possibilità di rumore e ridondanza nei dati del mondo reale (Venkat, 2018).
Nel *data mining*, ossia il processo di estrazione di informazioni utili da grandi quantità di dati, la Maledizione della Dimensionalità si riferisce a un set di dati con troppe caratteristiche. In tale contesto si può dimostrare che, all'aumentare del numero di variabili esplicative, il problema della scoperta della struttura diventa più difficile (Banks & Fienberg, 2003).

Per superare la Maledizione della Dimensionalità, la tecnica più comunemente utilizzata è la riduzione di dimensionalità, che permette di mantenere le dimensioni più importanti ed eliminare quelle irrilevanti.
Come anticipato in precedenza, la riduzione di dimensionalità è una delle pratiche maggiormente impiegate nell'analisi dei dati.

1.3 La riduzione di dimensionalità: una revisione della letteratura

I dati del mondo reale, come segnali vocali e fotografie digitali, solitamente hanno una dimensione elevata. Per gestire adeguatamente tali dati, è necessario ridurre la loro dimensionalità.

La riduzione della dimensionalità (DR) è la trasformazione dei dati da uno spazio ad alta dimensione in uno spazio di dimensione inferiore, in modo che quest'ultimo mantenga alcune proprietà significative dei dati originali (van der Maaten *et al.,* 2009). Idealmente, la rappresentazione ridotta dovrebbe avere una dimensionalità che corrisponda alla dimensionalità intrinseca dei dati, dove questa è il numero minimo di parametri necessari per spiegare le proprietà osservate dei dati (Fukunaga, 1990).

La riduzione della dimensionalità è importante in molti ambiti, poiché attenua la Maledizione della Dimensionalità e altre proprietà indesiderate degli spazi ad alta dimensionalità, facilitando la classificazione, la visualizzazione e la compressione dei dati ad alta dimensionalità.

Tale tecnica viene comunemente utilizzata nei campi che trattano un gran numero di osservazioni o un gran numero di variabili, come l'elaborazione del segnale, il riconoscimento vocale, la neuroinformatica e la bioinformatica.

La riduzione della dimensionalità rappresenta uno step fondamentale al fine di trasformare grandi dataset in informazioni utili, che è il principale obiettivo della data science. Le informazioni significative sono sommerse da dati complessi, rendendo difficile scoprirne le caratteristiche rilevanti. Pertanto, la riduzione di dimensionalità risulta assolutamente imprescindibile nell'ambito della scienza dei dati. Questo è in particolar modo vero per i big data, in quanto a causa della loro complessità è spesso necessario utilizzare tecniche di riduzione delle dimensioni prima di tentare di condurre inferenze statistiche o risolvere un problema.

La riduzione della dimensionalità è il processo di riduzione di un dataset ad alta dimensionalità in una rappresentazione di dimensione inferiore che conserva la maggior parte della sua struttura importante (Brehmer *et al.,* 2014). Dunque, l'obiettivo è ridurre la dimensione dei dati conservando la maggior parte dell'informazione.

In generale, lo scopo preminente è trovare rappresentazioni dei dati di dimensione inferiore che conservino le loro proprietà chiave per un dato problema. Sostanzialmente,

la riduzione della dimensionalità permette di creare degli indicatori di sintesi; per fare una sintesi servono delle variabili che parzialmente portano la stessa informazione, dunque tale processo di riduzione risulta più efficace per i big data con un elevato numero di variabili input che sono correlate tra loro (Deng *et al*., 2022).

La riduzione della dimensionalità facilita l'analisi di enormi quantità di informazioni e permette di estrarre caratteristiche informative utili dai dati ad alta dimensionalità, nonché consente di eliminare l'influenza di fattori correlati o ripetitivi (Jia *et al*., 2022). Il principio base della riduzione di dimensionalità delle variabili è quello di mappare un campione di dati da uno spazio ad alta dimensionalità a uno spazio a dimensionalità relativamente bassa. In altre parole, il compito fondamentale è ottenere una mappatura che consenta di acquisire un'efficace struttura a bassa dimensione, che è nascosta nei dati osservabili ad alta dimensione.

La finalità della riduzione della dimensionalità è identificare e selezionare o derivare le caratteristiche più utili in un dato corpus di dati per un determinato problema riguardante una popolazione. Tuttavia, si tratta di un obiettivo difficile da raggiungere in astratto poiché un certo numero di problemi diversi possono essere posti sugli stessi dati, possibilmente anche con scopi differenti (Mainali *et al*., 2021).

1.3.1 Le tecniche di riduzione di dimensionalità: lineari e non lineari

Le tecniche di riduzione della dimensionalità sono varie e differiscono sostanzialmente per il tipo di dati da analizzare (osservazioni di variabili quantitative, osservazioni di variabili quantitative, graduatorie, preferenze, ecc.).
Generalmente, i diversi metodi sono suddivisi in due categorie principali: metodi lineari e non lineari.

Le tecniche di riduzione della dimensionalità lineari si basano tipicamente sull'algebra lineare, dispongono di solide fondamenta matematiche e incorporano la cosiddetta *fattorizzazione di matrici*, nota anche come *decomposizione di matrici* (Stein-O'Brien *et al*., 2018). L'idea alla base di questo approccio è approssimare una matrice di dati tramite il prodotto di almeno altre due matrici, una delle quali rappresenta una nuova

configurazione della matrice di dati iniziale ma con un parametro p di dimensione ridotta.
Tra tutti i metodi di riduzione di dimensionalità, le tecniche più ampiamente utilizzate sono le proiezioni ortogonali, le quali cercano di proiettare i dati in uno spazio di dimensioni inferiori in modo lineare. Questi metodi producono una mappatura lineare a bassa dimensionalità degli originali dati ad alta dimensionalità, che preserva alcune caratteristiche di interesse nei dati (Cunningham & Ghahramani, 2015).
Le tecniche di riduzione della dimensionalità lineari devono la loro popolarità, in parte, alla loro semplice interpretazione geometrica come una visione a bassa dimensionalità dei dati ad alta dimensionalità. Alcuni esempi includono tecniche quali l'Analisi delle Componenti Principali (PCA), il Multidimensional Scaling (MDS), l'Analisi Discriminante Lineare (LDA) e l'Analisi delle Componenti Indipendenti (ICA), tra gli altri.

Tradizionalmente, la riduzione della dimensionalità veniva eseguita utilizzando tecniche lineari come la PCA, tuttavia questi metodi non possono gestire in modo adeguato dati complessi e non lineari. Di conseguenza, sono state proposte diverse tecniche di riduzione di dimensionalità non lineari che mirano a superare i limiti delle tecniche lineari tradizionali.
A differenza delle tecniche lineari tradizionali, le tecniche non lineari hanno la capacità di gestire dati complessi e non lineari. In particolare, per i dati del mondo reale, le tecniche non lineari possono offrire un vantaggio, poiché è probabile che i dati provenienti dal mondo reale formino una varietà altamente non lineare.
Le tecniche non lineari di riduzione della dimensionalità cercano di catturare relazioni non lineari tra le variabili. Alcuni esempi sono t-SNE (t-Distributed Stochastic Neighbor Embedding), UMAP (Uniform Manifold Approximation and Projection), Isomap, NLPCA (Nonlinear Principal Component Analysis).

Ogni metodo ha i suoi vantaggi, svantaggi e ambiti di applicazione. La scelta di quale tecnica adottare dipende dalla natura dei dati, dagli obiettivi dell'analisi e dal contesto specifico in cui si opera.

1.3.2 Il principio della riduzione di dimensionalità delle variabili

Negli anni i ricercatori hanno proposto vari algoritmi di riduzione della dimensionalità. Nel presente sottoparagrafo si intende presentare il principio della riduzione di dimensionalità delle variabili.

La riduzione di dimensionalità delle variabili utilizza i parametri delle variabili esistenti per formare uno spazio di variabili a bassa dimensionalità e permette di superare gli effetti dell'informazione ridondante o irrilevante, in modo da mappare l'informazione importante contenuta nelle variabili originali su un numero inferiore di variabili.

In termini matematici, si supponga che esista un vettore X n-dimensionale e un vettore Y m-dimensionale:

$$X = [x_1, x_2, \dots, x_n]^T$$
$$Y = [Y_1, Y_2, \dots, Y_m]^T$$

e

$$m \ll n$$

Il vettore X è mappato sul vettore Y, attraverso una mappa f. Matematicamente, la funzione di mappatura può essere espressa come:

$$Y = f(X)$$

Quanto esposto è il processo di estrazione e selezione delle variabili. La mappatura f è l'algoritmo ricercato per la riduzione delle variabili, e la scelta del tipo di mappatura varia a seconda del problema in esame.

1.3.3 Le tecniche di riduzione della dimensionalità: estrazione e selezione delle variabili

I diversi approcci di riduzione della dimensionalità possono anche essere suddivisi in selezione ed estrazione delle variabili.

Selezione delle variabili

La selezione delle variabili è conosciuta anche come *selezione di sottoinsiemi di variabili*. Le tecniche appartenenti a questa tipologia mirano a selezionare sottoinsiemi

delle variabili input (note anche come *caratteristiche* o *attributi* o *feature*). L'impiego di questi metodi permette di conseguire alcuni vantaggi, quali semplificare il modello al fine di renderlo più facilmente interpretabile dai ricercatori e/o utenti, ridurre il tempo di esecuzione, evitare fenomeni di Maledizione della Dimensionalità (Jia *et al.,* 2022).
Vi è un importante prerequisito che il dataset del problema considerato deve possedere al fine di impiegare le tecniche di selezione delle variabili: i dati devono contenere molte caratteristiche ridondanti o correlate che possono essere eliminate senza perdere molta informazione. Si consideri che quando si parla di attributi ridondanti e attributi correlati si tratta di due concetti diversi, in quanto una caratteristica correlata può risultare ridondante in presenza di altre caratteristiche correlate a essa strettamente correlate.
Generalmente, la selezione delle variabili viene utilizzata in aree caratterizzate da molte variabili e relativamente pochi campioni.

Estrazione delle variabili
L'estrazione delle variabili, nota anche come *proiezione delle variabili*, opera tramite la trasformazione dei dati dallo spazio ad alta dimensione a uno spazio di dimensioni inferiori. Tale trasformazione può essere lineare, come nell'Analisi delle Componenti Principali (PCA), oppure può avvenire tramite tecniche non lineari di riduzione della dimensionalità.
L'estrazione delle variabili genera nuovi attributi a partire dalle variabili originali, il che significa che i nuovi attributi sono una mappatura delle variabili originali. L'impiego di questo approccio implica sia vantaggi sia svantaggi: se da un lato la compressione delle nuove variabili è più efficiente, dall'altro, quando il set di variabili originali ha un ovvio significato fisico, i nuovi attributi potrebbero perdere significato.

Storicamente, la tecnica classica impiegata per la riduzione della dimensionalità è l'Analisi delle Componenti Principali (PCA), basata sull'estrazione delle variabili tramite la combinazione delle variabili fondamentali in modo lineare, con l'obiettivo di creare variabili incorrelate tra loro, che contengono una quota rilevante dell'informazione complessiva (Mainali *et al.,* 2021).
Come citato in precedenza, nel tempo sono state sviluppate delle alternative non lineari, in quanto in presenza di dati complessi i metodi lineari possono presentare rilevanti lacune.

Allo stesso modo, sono state introdotte altre tecniche che operano attraverso la selezione delle caratteristiche/variabili, anziché l'estrazione delle stesse.
In generale, la scelta tra questi due approcci dipende principalmente dal problema che si sta risolvendo.

1.3.4 Considerazioni sulla riduzione di dimensionalità

Come già anticipato, esistono diverse tecniche di riduzione della dimensionalità a seconda della tipologia di dati del problema considerato. In questo paragrafo faremo riferimento al caso specifico di dati provenienti dall'osservazione di variabili quantitative, il che è propedeutico alla trattazione della PCA nel capitolo successivo, in quanto essa opera esclusivamente con variabili quantitative.

Dati N soggetti su cui sono state osservate p variabili quantitative, si chiama riduzione della dimensionalità la procedura secondo la quale si determinano q nuove variabili (con $q \ll p$), che contengono gran parte dell'informazione che era contenuta nelle p variabili iniziali. Si vuole, quindi, effettuare una buona sintesi della variabile p-dimensionale in una variabile di dimensione q, (molto) inferiore a p.

Innanzitutto, si consideri che ridurre la dimensionalità significa proiettare da uno spazio a un altro di dimensione inferiore: nell'ambito della PCA, la riduzione della dimensionalità si ottiene attraverso le proiezioni ortogonali. Questo procedimento comporta inevitabilmente una perdita di informazione: il processo di mappatura dei dati ad alta dimensionalità nello spazio a bassa dimensionalità porta necessariamente la perdita di una parte dell'informazione contenuta nel dataset originale.
In altre parole, anche la miglior proiezione comunque implica informazione perduta. Pertanto, è necessario scegliere la proiezione che traduca l'informazione il più correttamente possibile, riducendo al minimo la perdita informativa, ovvero massimizzando la varianza spiegata[11].

[11] Bassi F., Ingrassia S. (2022), *Statistica per analisi di mercato: metodi e strumenti*. Pearson, Milano-Torino.

Si ha una corretta traduzione dell'informazione se punti distanti (vicini) nello spazio p-dimensionale vengono proiettati in posizioni distanti (vicine) nello spazio q-dimensionale.

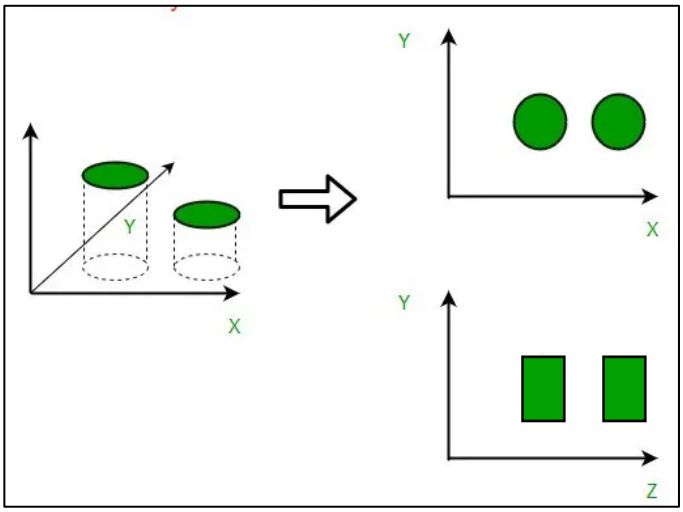


Figura 1.4: Esempio di riduzione di dimensionalità su un oggetto cilindrico (Fonte: https://medium.com/@aakriti.sharma18/data-science-interview-entry-level-questions-b6f436759c98)

Come si può vedere in Figura 1.4, la riduzione dimensionale di un oggetto caratterizzato da tre dimensioni X, Y e Z implica inevitabilmente una perdita informativa. Infatti, per esempio, rimuovendo la dimensione Z si perde l'informazione riguardo la forma cilindrica e l'altezza del solido. Analogamente, rimuovendo la dimensione X vengono persi dati di valore sulla forma originale dell'oggetto.
Da questo esempio, applicato a un solido, si può concludere come trovare la rappresentazione ottimale in uno spazio di dimensione inferiore è fondamentale per preservare una quota rilevante di informazione, considerando però che una parte di essa verrà sempre persa.

Tra tante possibili proiezioni ne esiste una che massimizza l'informazione trattenuta, minimizzando quindi quella perduta. Tra tutte le proiezioni, dunque, si sceglie la migliore possibile.

L'informazione trattenuta dalla proiezione coincide con la variabilità o varianza dei punti proiettati nel sottospazio, mentre l'informazione perduta con la proiezione coincide con la variabilità o varianza dei punti da proiettare attorno al sottospazio.

Quanto appena esposto viene ora esemplificato con riferimento alle nubi di punti in spazi euclidei (che è il contesto in cui opera la PCA): consideriamo il caso particolare di riduzione di dimensionalità da uno spazio bidimensionale a uno spazio unidimensionale, supponendo di voler rappresentare in una sola dimensione una nube di punti bidimensionale. Quindi, si opera con $p = 2$ e $q = 1$.

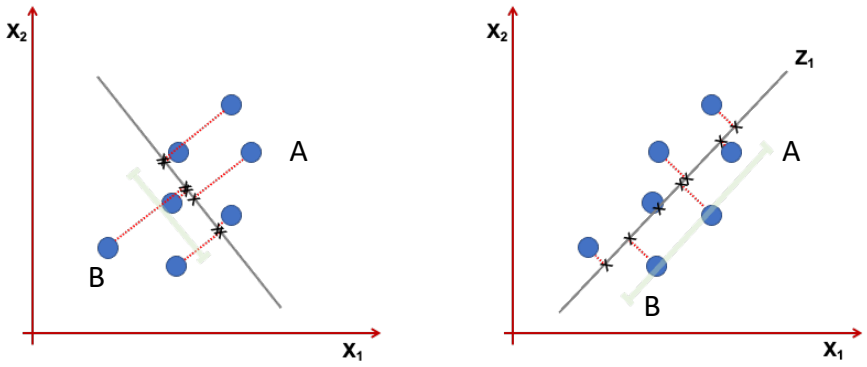


Figura 1.5: Riduzione di dimensionalità da uno spazio bidimensionale a uno spazio unidimensionale (retta)
(Fonte: https://bookdown.org/tpinto_home/Unsupervised-learning/principal-components-analysis.html)

Come esemplificato nella Figura 1.5, quando si applica la riduzione di dimensionalità, tra le infinite proiezioni possibili, si deve scegliere quella che massimizza l'informazione trattenuta e minimizza l'informazione perduta.
Nel caso di sinistra, proiettando i punti sulla retta, i punti che in origine (ovvero nello spazio a due dimensioni) erano distanti (ad esempio A e B), sulla retta sono vicini. Ciò significa che la proiezione ha distorto l'informazione dello spazio bidimensionale. Nel caso di destra, invece, l'informazione viene tradotta correttamente.
Da questo esempio si può dedurre come non tutte le proiezioni siano corrette, dunque si deve trovare un sottospazio che si adatta il meglio possibile alla nube che si sta proiettando. In questo caso specifico, la qualità della riduzione di dimensionalità dipende

dalla posizione della retta e, in particolare, la retta migliore è l'interpolante della nube di punti.

I punti vanno proiettati in una dimensione inferiore, scelta in modo tale che la variabilità tra i punti proiettati sia massima. Questo garantisce che l'immagine dei punti mantenga la maggior parte possibile di informazione perché il sottospazio scelto in questo modo è quello che "dista" meno dai punti da proiettare.

A questo punto, però, è importante osservare che ci sono casi in cui la proiezione è in grado di dare una descrizione abbastanza fedele dell'immagine iniziale, ma esistono anche casi in cui questo non è possibile. Infatti, non sempre si può individuare una proiezione che porta con sé l'informazione più importante. A seconda del fenomeno oggetto di analisi, potrebbe capitare che la riduzione di dimensionalità sia realizzabile solamente rinunciando a una parte significativa di informazione.

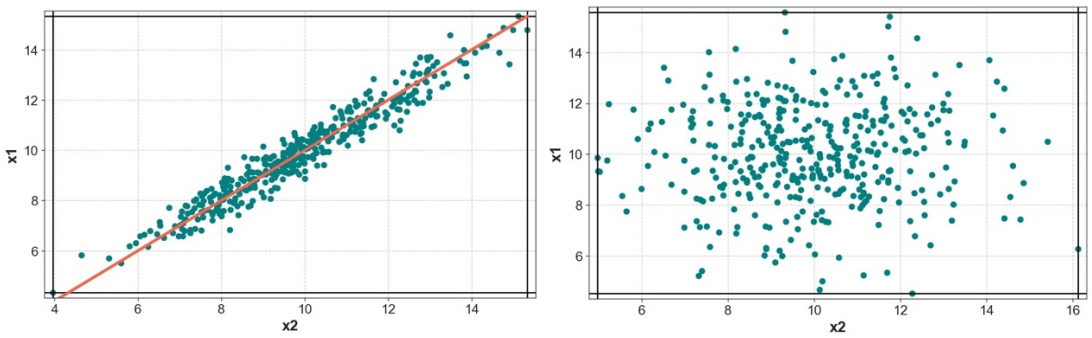


Figura 1.6: Scatterplot (nube di punti) con diversa dispersione dei punti (Fonte: https://medium.com/analytics-vidhya/principal-component-analysis-pca-part-1-fundamentals-and-applications-8a9fd9de7596)

Facendo riferimento alla rappresentazione a sinistra della Figura 1.6, quando i punti sono meno dispersi attorno alla retta su cui si effettua la proiezione, ovvero al crescere della correlazione tra X_1 e X_2 (misurata tramite il *coefficiente di correlazione lineare di Pearson* ρ), si perde una quantità sempre inferiore di informazione, poiché l'informazione che va perduta con la proiezione è costituita da quanto i punti sono dispersi attorno alla retta.

Al crescere di ρ in valore assoluto, la perdita di informazione che si avrebbe dalla proiezione sarebbe sempre più bassa.
Al limite, quando la correlazione tra X_1 e X_2 è massima, cioè quando i punti sono disposti esattamente su una retta crescente ($\rho = +1$) o decrescente ($\rho = -1$), la proiezione mantiene tutta l'informazione sulla nube di punti e non vi è alcuna perdita di informazione. Si consideri, però, che quest'ultimo è un caso puramente teorico.

Si può concludere, quindi, che i risultati della riduzione sono tanto migliori quanto maggiore è la correlazione lineare tra le variabili in gioco: più c'è correlazione lineare, meno informazione viene perduta. Al contrario, anche con la migliore proiezione possibile, si avrà un'importante perdita di informazione (dal punto di vista statistico questo accade quando la correlazione tra le due variabili è molto bassa). Queste considerazioni aiutano anche a chiarire meglio il significato di riduzione della dimensionalità lineare (di cui la PCA è un esempio), menzionati in precedenza in contrapposizione ai metodi non lineari: i metodi lineari di riduzione della dimensionalità sono adatti in caso di elevata correlazione lineare tra le variabili, mentre quelli non lineari sono in grado di trattare anche i casi in cui tra le variabili vi sono legami non lineari.

In altre parole, la perdita di informazione che si subisce nel ridurre la dimensionalità di una nube di punti con una proiezione ortogonale su un sottospazio lineare è tanto minore quanto più le variabili sono linearmente correlate tra loro.

CAPITOLO 2

PCA E NLPCA

2.1 Introduzione alla PCA

Oggi, i dataset di grandi dimensioni sono sempre più diffusi in molte discipline e, a causa della loro elevata dimensionalità[12], spesso sono difficili da interpretare.
L'obiettivo è quello di diminuire la dimensionalità dei dati in modo tale da renderli interpretabili, e allo stesso tempo preservare la maggior parte di informazione presente nei dati stessi, minimizzandone la perdita.

Molte tecniche sono state sviluppate a questo scopo, ma l'Analisi delle Componenti Principali (PCA) è una delle più ampiamente utilizzate. L'idea che sta alla base della PCA è semplice: ridurre la dimensionalità di un dataset, mantenendo quanta più "variabilità" (cioè informazione statistica) possibile delle variabili di partenza.
Ciò si traduce nella ricerca di nuove variabili che siano funzioni lineari di quelle del dataset originale, che massimizzino la varianza e che siano incorrelate tra loro. Trovare tali nuove variabili, le cosiddette *componenti principali* (PC), comporta la risoluzione di un problema di autovalori/autovettori (Joliffe & Cadima, 2016).

Considerando quanto appena esposto, appare evidente che lo studio del funzionamento e dell'applicazione della PCA richiede la preliminare conoscenza di alcuni elementi fondamentali di statistica e di algebra lineare, al fine di conseguire una piena comprensione del meccanismo della PCA.
A tale proposito, nella prima parte di questo capitolo si presenteranno brevemente alcuni concetti principali di algebra delle matrici, prima di proseguire con l'approfondimento teorico sulla PCA: le trasformazioni lineari, le proiezioni ortogonali di vettori in sottospazi, gli autovalori e gli autovettori di una matrice.

[12] Con il termine dimensionalità si indica il numero di variabili casuali considerate per descrivere un dato esperimento/fenomeno.

2.1.1 Le trasformazioni lineari

Si consideri una variabile p-dimensionale X, ovvero una variabile composta da p variabili unidimensionali:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

Siano noti il vettore delle medie e la matrice di varianze e covarianze del vettore X, così come di seguito:

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

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

La matrice di varianze e covarianze è una matrice quadrata che raccoglie le varianze nella diagonale principale e le covarianze negli elementi esterni alla diagonale principale.

La variabile q-dimensionale Y:

$$Y = CX$$

dove $C_{q \times p}$ è una matrice di parametri:

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{q1} & \cdots & c_{qp} \end{bmatrix}$$

è detta *trasformazione lineare* di X attraverso la matrice C.

Ogni variabile unidimensionale Y_j che compone la variabile q-dimensionale Y è una combinazione lineare delle p variabili unidimensionali X_j che compongono la variabile p-dimensionale X:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix} = \begin{bmatrix} c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \\ \vdots \\ c_{q1}X_1 + c_{q2}X_2 + \cdots + c_{qp}X_p \end{bmatrix}$$

È possibile ricavare vettore delle medie e matrice di varianze e covarianze di Y avendo a disposizione quelli della variabile X:

$$\bar{Y} = \begin{bmatrix} \bar{Y}_1 \\ \vdots \\ \bar{Y}_p \end{bmatrix} = C\bar{X}$$

$$\Sigma_Y = C\Sigma_X C'$$

Le osservazioni x_i della variabile X sono punti nello spazio p-dimensionale. Le osservazioni y_i della variabile Y sono punti nello spazio q-dimensionale.

Da ciò deriva un'importante conclusione: con la trasformazione lineare si trasferiscono i punti da uno spazio di p dimensioni a uno di q dimensioni, che è esattamente quello che accade applicando la PCA. Infatti, la PCA avviene tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano (si noti che il complessivo funzionamento della tecnica verrà ampiamente approfondito successivamente, nel presente capitolo).
Nell'ambito del presente elaborato, si considera $q \leq p$ poiché ha un senso dal punto di vista geometrico. Infatti, se $q > p$ il passaggio a una dimensione superiore è un'operazione fittizia, poiché i punti continuerebbero a giacere su un sottospazio delle dimensioni iniziali. Al contrario, l'obiettivo della riduzione di dimensionalità, e, allo stesso modo, della PCA, è quello di passare da uno spazio di dimensione superiore a uno spazio di dimensione inferiore.

Sintetizzando, sia data X, una variabile p-dimensionale con vettore delle medie $\bar{X}$ e matrice di varianze e covarianze Σ_X. Inoltre, sia data C, una matrice di dimensioni $q \times p$ (con $q \leq p$).
Allora, la variabile q-dimensionale $Y = CX$ è una trasformazione lineare di X attraverso C e ha vettore delle medie $\bar{Y} = C\bar{X}$ e matrice di varianze e covarianze $\Sigma_Y = C\Sigma_X C'$.

2.1.2 Le proiezioni ortogonali

Al fine di comprendere a fondo il meccanismo della PCA, risulta importante approfondire come effettuare proiezioni ortogonali di punti da uno spazio di dimensione p a uno spazio di dimensione q. La PCA, infatti, opera in un modo analogo, effettuando proiezioni ortogonali da spazi di dimensione p a spazi di dimensione inferiore, scelti in modo che la sintesi che scaturisce dalla riduzione della dimensionalità porti con sé la maggior parte possibile dell'informazione che era contenuta nei dati iniziali.

Di seguito, vengono presentati alcuni concetti rilevanti con riferimento alle proiezioni ortogonali.

In generale, dato un vettore $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{bmatrix}$ si definisce *norma* del vettore lo scalare:

$$||x|| = \sqrt{x'x} = \sqrt{\sum_{i=1}^{p} x_1^2}$$

che rappresenta, intuitivamente, la "lunghezza" del vettore.

Inoltre, si definisce *versore* un vettore u di norma unitaria, per cui vale:

$$||u|| = \sqrt{u'u} = \sqrt{\sum_{i=1}^{p} u_1^2} = 1$$

Date tali premesse, si vogliono ricavare le coordinate dei punti nel nuovo spazio.

Si consideri, inizialmente, una proiezione dal piano su una retta, ovvero una proiezione da uno spazio bidimensionale a uno spazio unidimensionale.
Data una variabile bidimensionale X, e una sua osservazione definita dal generico vettore $x_i' = [x_{i1} \quad x_{i2}]$, si vogliono proiettare ortogonalmente i punti (vettori) nel piano su una retta passante per l'origine.
Innanzitutto, è necessario ottenere un versore che "rappresenta" la retta, ovvero un vettore di norma unitaria che si sovrappone esattamente alla retta:

$$u' = [u_1 \quad u_2]$$

La proiezione ortogonale y_i del vettore $x_i' = [x_{i1} \quad x_{i2}]$ sulla retta passante per l'origine, rappresentata dal versore u, è data da:

$$y_i = u'x_i = [u_1 \quad u_2] \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}$$

Di seguito viene presentato un esempio numerico con relative rappresentazioni grafiche. Siano date tre osservazioni di una variabile bidimensionale X (Figura 2.1 a sinistra):

$x_1' = [0.5 \quad 0]$ $x_2' = [0 \quad 0.5]$ $x_3' = [-1 \quad -1]$

A ogni osservazione corrisponde un punto nel piano: si vogliono proiettare ortogonalmente i punti (vettori) su una retta passante per l'origine (Figura 2.1 a destra).

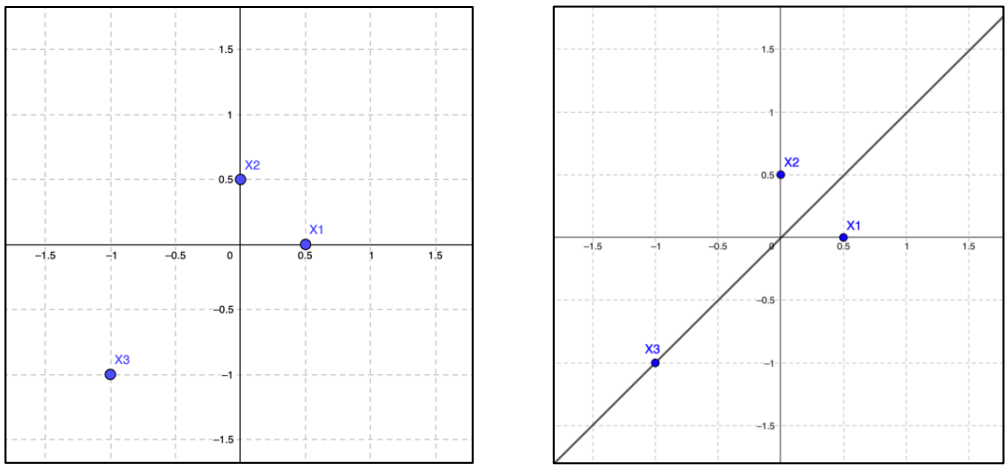


Figura 2.1: Rappresentazione grafica delle tre osservazioni sul piano
(Fonte: nostre elaborazioni)

Bisogna prima di tutto ricavare un versore u che "rappresenta" la retta, cioè un vettore di norma unitaria che si sovrappone esattamente alla retta, e successivamente effettuare la proiezione ortogonale dei punti dal piano sulla retta (Figura 2.2 e 2.3).

$$u = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

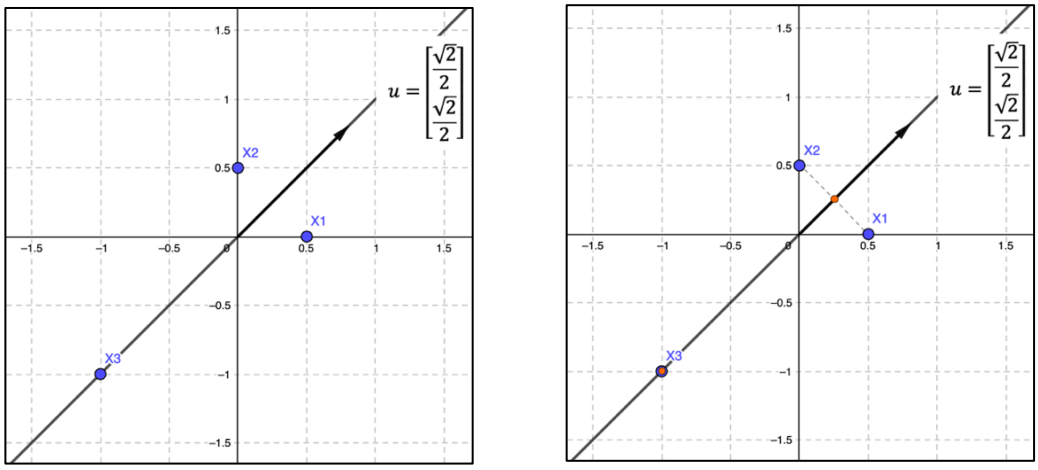


Figura 2.2: Individuazione del versore u ai fini della proiezione ortogonale
(Fonte: nostre elaborazioni)

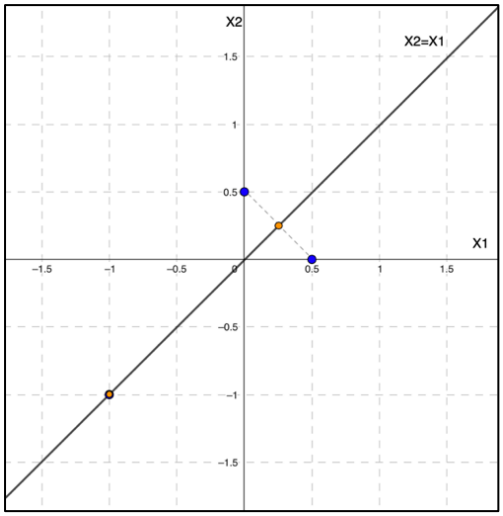


Figura 2.3: Proiezione ortogonale dei punti sulla retta
(Fonte: nostre elaborazioni)

A questo punto, si vogliono ricavare le coordinate dei punti nel nuovo spazio, nel caso specifico si tratta di uno spazio unidimensionale (Figura 2.4).

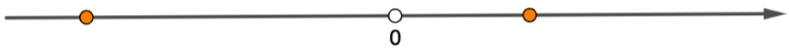


Figura 2.4: Punti proiettati nel nuovo spazio unidimensionale
(Fonte: nostre elaborazioni)

La proiezione ortogonale y_i del vettore $x_1' = [x_{i1} \quad x_{i2}]$ sulla retta passante per l'origine, rappresentata dal versore u, è data da:

$$y_i = u'x_i = [u_1 \quad u_2] \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}$$

ovvero:

- proiezione di $x_1' = [0.5 \quad 0]$

$$y_1 = u'x_1 = \left[\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2}\right] \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} = \frac{\sqrt{2}}{4}$$

- proiezione di $x_2' = [0 \quad 0.5]$

$$y_2 = u'x_2 = \left[\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2}\right] \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} = \frac{\sqrt{2}}{4}$$

- proiezione di $x_3' = [-1 \quad -1]$

$$y_3 = u'x_3 = \left[\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2}\right] \begin{bmatrix} -1 \\ -1 \end{bmatrix} = -\sqrt{2}$$

Nella Figura 2.5 vengono rappresentate le coordinate dei punti nel nuovo spazio di una dimensione.

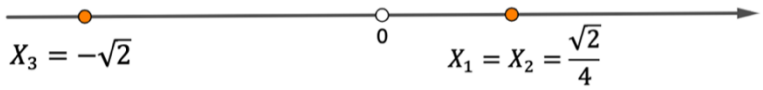


Figura 2.5: Coordinate dei punti proiettati nel nuovo spazio unidimensionale (Fonte: nostre elaborazioni)

Talvolta, può accadere che la retta su cui si vogliono proiettare i punti non passi per l'origine (Figura 2.6). In tali casi, è necessario effettuare preliminarmente una traslazione della retta in modo da forzarla a passare per l'origine (sostanzialmente, viene eliminata l'intercetta della retta). Le posizioni relative dei punti proiettati non cambiano a causa delle traslazioni dello spazio su cui si proietta. Successivamente, si procede esattamente come visto in precedenza.

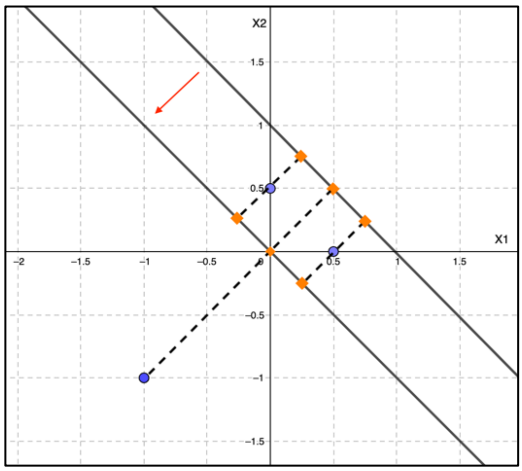


Figura 2.6: Traslazione della retta non passante per l'origine
(Fonte: nostre elaborazioni)

Allo stesso modo, si può pensare di proiettare dallo spazio tridimensionale su un piano (bidimensionale). Il ragionamento è esattamente analogo:

1) In primo luogo, il piano deve passare per l'origine, altrimenti è necessario traslarlo in modo opportuno.
2) In secondo luogo, bisogna individuare due versori u_1 e u_2, che siano ortogonali e che "rappresentino" il piano:

$$u'_1 = [u_{11} \quad u_{12} \quad u_{13}]$$
$$u'_2 = [u_{21} \quad u_{22} \quad u_{23}]$$

3) I due versori vengono riuniti in una matrice U:

$$U = \begin{bmatrix} u'_1 \\ u'_2 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{bmatrix}$$

Si noti che $UU' = I$.

4) La proiezione y_i del generico vettore $x'_i = [x_{i1} \quad x_{i2} \quad x_{i3}]$ sul piano si ricava tramite l'operazione:

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} = UX_i = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{bmatrix}$$

A questo punto, è possibile generalizzare e definire la proiezione ortogonale da uno spazio p-dimensionale a uno spazio q-dimensionale (con $q < p$). I punti rappresentati in spazi di p dimensioni sono osservazioni di una variabile p-dimensionale X. Al fine di

ricavare la proiezione ortogonale, il ragionamento è un'estensione di quello mostrato in precedenza:

1) Innanzitutto, lo spazio q-dimensionale deve contenere l'origine, altrimenti bisogna effettuare un'opportuna traslazione.
2) In seguito, è necessario individuare q versori, che siano ortogonali a due a due e che "rappresentino" lo spazio su cui proiettare:

$$u'_1 = [u_{11} \quad u_{12} \quad \dots \quad u_{1p}] \quad \dots \quad u'_q = [u_{q1} \quad u_{q2} \quad \dots \quad u_{qp}]$$

3) I q versori vengono riuniti in una matrice $U_{q\times p}$ (per cui vale sempre $UU' = I$):

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

4) La proiezione y_i del generico vettore $x'_i = [x_{i1} \quad x_{i2} \quad \dots \quad x_{ip}]$ sullo spazio q-dimensionale si ricava con l'operazione:

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iq} \end{bmatrix} = UX_i = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ \vdots & \vdots & \dots & \vdots \\ u_{q1} & u_{q2} & \dots & u_{qp} \end{bmatrix} \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$$

Pertanto, la proiezione ortogonale dei punti in un sottospazio di dimensione q equivale a creare una variabile q-dimensionale Y, che è una *trasformazione lineare di X attraverso la matrice U:*

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix} = UX = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ \vdots & \vdots & \dots & \vdots \\ u_{q1} & u_{q2} & \dots & u_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

Questo significa che dalla proiezione ortogonale in un sottospazio si origina una variabile Y, con vettore delle medie e matrice di varianze e covarianze noti e dati da:

$$\bar{Y} = \begin{bmatrix} \bar{Y}_1 \\ \vdots \\ \bar{Y}_p \end{bmatrix} = U\bar{X}$$

$$\Sigma_Y = U\Sigma_X U'$$

Di seguito si riporta la sintesi delle evidenze più significative ottenute, necessarie per la successiva trattazione della PCA.

Dato un vettore x, osservazione di una variabile p-dimensionale X, è possibile proiettarlo ortogonalmente dallo spazio p-dimensionale a un sottospazio q-dimensionale passante per l'origine (con $q < p$). Lo strumento che consente di effettuare la proiezione è una matrice U che contiene q versori a due a due ortogonali, che rappresentano lo spazio q-dimensionale su cui si proietta.
Dunque, la proiezione ortogonale è una trasformazione lineare della variabile p-dimensionale X attraverso la matrice U.

2.1.3 Le rotazioni

Finora, è stato considerato il caso con $q < p$. Tuttavia, esiste anche il caso con $q = p$, che si riferisce a una trasformazione lineare da uno spazio a un altro della medesima dimensione:

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

In tale circostanza, però, non si ha una proiezione dei punti, bensì una *rotazione* dello spazio p-dimensionale. Considerato questo elemento di differenziazione, il procedimento per ricavare le nuove coordinate (nel nuovo spazio di dimensione $q = p$) è del tutto analogo a quello visto precedentemente con $q < p$.

In sintesi, la rotazione è una trasformazione lineare della variabile p-dimensionale X attraverso la matrice U; in questo caso particolare con $q = p$, la matrice U contiene p versori a due a due ortogonali che inducono una rotazione dello spazio p-dimensionale di partenza.

Al fine di una migliore comprensione, viene ora presentato un esempio numerico con relativo grafico di una rotazione di uno spazio bidimensionale.
Siano date tre osservazioni di una variabile bidimensionale X:
$x_1' = [1 \quad 1]$ $x_2' = [-1 \quad 1]$ $x_3' = [-1 \quad -1]$
A ogni osservazione corrisponde un punto nel piano: si vuole effettuare una rotazione del piano e trasferire in tal modo i punti nel nuovo spazio bidimensionale, rappresentato dai due versori u_i (Figura 2.7):

$u_1' = [\sqrt{2}/2 \quad \sqrt{2}/2]$ $u_2' = [-\sqrt{2}/2 \quad \sqrt{2}/2]$

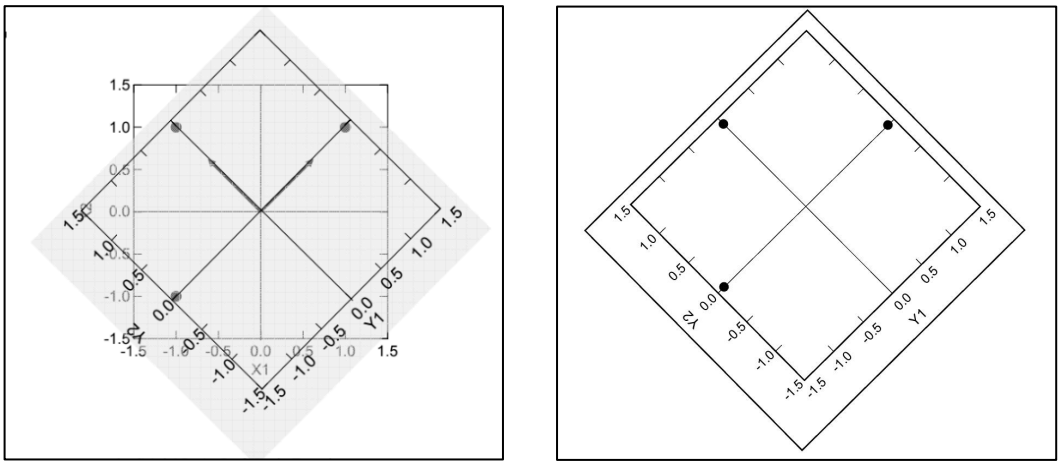


Figura 2.7: Rotazione del piano nel nuovo spazio bidimensionale con individuazione dei versori u_i (Fonte: corso di "Analisi dei Dati" A.A. 2010/2011 – Prof.ssa Paola Zuccolotto)

Si può verificare che u_1 e u_2 sono due versori ortogonali poiché la matrice U è una matrice ortogonale:

$$U = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$

$$UU' = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

A questo punto si procede con il calcolo delle nuove coordinate dei punti:

- rotazione di $x_1' = [1 \quad 1]$

$$y_1 = Ux_1 = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}$$

- rotazione di $x_2' = [-1 \quad 1]$

$$y_2 = Ux_2 = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix}$$

- rotazione di $x_3' = [-1 \quad -1]$

$$y_3 = Ux_3 = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} -\sqrt{2} \\ 0 \end{bmatrix}$$

In Figura 2.8 vengono rappresentate le coordinate dei punti nel nuovo spazio bidimensionale, a seguito della rotazione del piano.

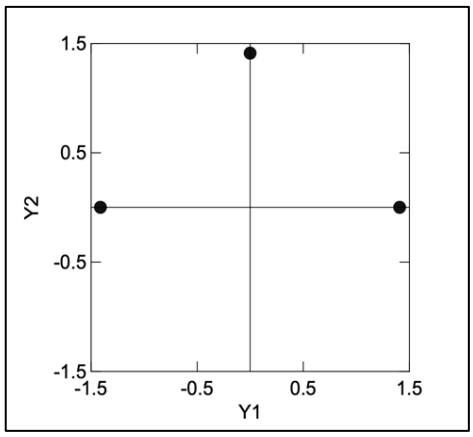


Figura 2.8: Nuovo spazio bidimensionale dopo la rotazione
(Fonte: corso di "Analisi dei Dati" A.A. 2010/2011 – Prof.ssa Paola Zuccolotto)

2.1.4 Gli autovalori e gli autovettori

Data una matrice quadrata $A_{p\times p}$, uno scalare λ e un vettore $h_{p\times 1}$, si dice che λ è un *autovalore* per A e che h è il suo corrispondente *autovettore* se vale la seguente condizione:

$$Ah = \lambda h$$

Il calcolo di autovalori e autovettori procede attraverso i seguenti passaggi:

1) I p autovalori della matrice $A_{p\times p}$ sono le radici dell'equazione caratteristica, di grado p, per cui vale:

   $$Det(A - \lambda I) = 0$$

   dove I denota la matrice identità. In pratica, dopo aver calcolato il determinante, si ricavano le radici dell'equazione caratteristica, che sono gli autovalori della matrice A.

2) Ad ogni autovalore λ_j corrispondono infiniti autovettori, dati dalle infinite soluzioni del sistema:

   $$\left(A - \lambda_j I\right)h_j = 0$$

Generalmente, tra gli infiniti autovettori così determinati ne viene scelto uno di norma unitaria, che viene chiamato autovettore normalizzato corrispondente all'autovalore λ_j.

Di seguito viene proposto un esempio di calcolo di autovalori e autovettori di una matrice 2×2.

Sia data la matrice $A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$. Nel caso in esame, si tratta di una matrice quadrata $A_{2\times2}$, simmetrica a valori reali. Si ricava il determinante:

$$\det(A - \lambda I) = \det\begin{bmatrix} 1-\lambda & 3 \\ 3 & 1-\lambda \end{bmatrix}$$

L'equazione caratteristica corrispondente risulta:

$$\lambda^2 - 2\lambda - 8 = 0$$

e ha come radici $\lambda_1 = -2$ e $\lambda_2 = 4$, che sono gli autovalori della matrice A.

A ogni autovalore λ_j corrispondono infiniti autovettori, dati dalle infinite soluzioni del sistema $(A - \lambda_j I)h_j = 0$. Dunque, per $\lambda_1 = -2$ si ha:

$$\begin{bmatrix} 1-\lambda_1 & 3 \\ 3 & 1-\lambda_1 \end{bmatrix} \cdot \begin{bmatrix} h_{11} \\ h_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix} \cdot \begin{bmatrix} h_{11} \\ h_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Si tratta di un sistema lineare con infinite soluzioni perché la matrice dei coefficienti ha determinante nullo. Gli autovettori sono dunque le soluzioni non nulle del sistema lineare, date da:

$$h_1 = \begin{bmatrix} h_{11} \\ h_{12} \end{bmatrix} = \begin{bmatrix} h \\ -h \end{bmatrix}$$

In genere tra gli infiniti autovettori così determinati se ne sceglie uno di norma unitaria, che viene detto autovettore normalizzato corrispondente all'autovalore λ_1:

$$||h_1|| = \sqrt{h^2 + h^2} = 1$$

$$h = \pm\frac{1}{\sqrt{2}} = \pm\frac{\sqrt{2}}{2}$$

Scegliendo, ad esempio, il segno positivo, si ha l'autovettore normalizzato corrispondente a λ_1:

$$h_1 = \begin{bmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{bmatrix}$$

Per $\lambda_2 = 4$ si ha:

$$\begin{bmatrix} 1-\lambda_2 & 3 \\ 3 & 1-\lambda_2 \end{bmatrix} \cdot \begin{bmatrix} h_{21} \\ h_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -3 & 3 \\ 3 & -3 \end{bmatrix} \cdot \begin{bmatrix} h_{21} \\ h_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Come in precedenza, si tratta di un sistema lineare con infinite soluzioni perché la matrice dei coefficienti ha determinante nullo. Quindi, le soluzioni sono date da:

$$h_2 = \begin{bmatrix} h_{21} \\ h_{22} \end{bmatrix} = \begin{bmatrix} h \\ h \end{bmatrix}$$

Procedendo analogamente a prima, si determina un autovettore di norma unitaria:

$$||h_2|| = \sqrt{h^2 + h^2} = 1$$

$$h = \pm\frac{1}{\sqrt{2}} = \pm\frac{\sqrt{2}}{2}$$

Scegliendo, ad esempio, il segno positivo, si ha l'autovettore normalizzato corrispondente a λ_2:

$$h_2 = \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$$

Ora, si consideri il caso particolare in cui *A* è una matrice simmetrica a valori reali, come nell'esempio appena esposto. Sotto questa ipotesi valgono determinate proprietà:

a) Gli autovettori sono a due a due ortogonali.

   Si ricordi che due vettori a e b sono ortogonali se $a'b = 0$. Dall'esempio di calcolo precedente, si può verificare che i due autovettori normalizzati h_1 e h_2 sono ortogonali:

$$h_1' h_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{-\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} = 0$$

b) Detta H la matrice le cui righe sono costituite dagli autovettori di A normalizzati, si ha:

$$HAH' = D$$

   dove D è una matrice diagonale che contiene sulla diagonale principale gli autovalori λ_j.

   In sintesi: i due autovettori sono ortogonali, pertanto si possono utilizzare per diagonalizzare A, ovvero per calcolare la matrice D.

c) Gli autovalori sono valori reali.

A questo punto, considerando la premessa teorica appena presentata, si procede a contestualizzare con riferimento al caso di nostro interesse.

Sia X una variabile p-dimensionale con matrice di varianze e covarianze Σ_X. Sia H la matrice degli autovettori normalizzati di Σ_X. Poiché Σ_X è una matrice simmetrica a valori reali, valgono per H le proprietà prima enunciate.
Si consideri una variabile Y data da $Y = HX$, dove Y è una trasformazione lineare di X attraverso H.
H gode delle proprietà prima enunciate, in base alle quali è possibile trarre alcune conclusioni su Y, così come di seguito:

1) Gli autovettori normalizzati contenuti in H sono a due a due ortogonali. Dunque, la matrice degli autovettori normalizzati $H_{p\times p}$ contiene dei versori a due a due ortogonali, che possono rappresentare uno spazio p-dimensionale.

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

Pertanto, la variabile Y (che è una trasformazione lineare di X attraverso H, da uno spazio p-dimensionale a uno spazio della stessa dimensione) corrisponde a una rotazione dello spazio di partenza.

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

2) $H\Sigma_X H'$ è uguale a una matrice diagonale contenente gli autovalori di Σ_X. Si consideri che, in base alle proprietà delle trasformazioni lineari, la matrice di varianze e covarianze di $Y = HX$ è data proprio da $\Sigma_Y = H\Sigma_X H'$. Quindi:

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

Dunque, la matrice di varianze e covarianze di Y è una matrice diagonale contenente gli autovalori di Σ_X. Di conseguenza, Y è una variabile p-dimensionale composta da variabili Y_j con le seguenti caratteristiche:

a) sono incorrelate tra loro (poiché la matrice Σ_Y è diagonale);
b) hanno varianze pari agli autovalori della matrice Σ_X.

In sintesi, se X è una variabile p-dimensionale con matrice di varianze e covarianze Σ_X e H è la matrice degli autovettori normalizzati di Σ_X, allora la variabile $Y = HX$ viene ottenuta attraverso una trasformazione lineare che induce una rotazione dello spazio di partenza. Inoltre, Y è una variabile p-dimensionale composta da variabili incorrelate e con varianze pari agli autovalori di Σ_X.

2.2 L'Analisi delle Componenti Principali (PCA)

Spesso, si rende necessario realizzare delle mappe di posizionamento per confrontare marche, prodotti e/o aziende sul mercato.

Il posizionamento "oggettivo" fa riferimento ad analisi effettuate sulla base di caratteristiche tecniche, o comunque oggettivamente rilevabili, delle unità statistiche in esame. Per valutare il posizionamento delle unità statistiche nella mappa di posizionamento oggettivo, si fa ricorso a una particolare tecnica di statistica multivariata: l'Analisi delle Componenti Principali (PCA).

A ogni unità statistica osservata si può quindi associare un vettore p-dimensionale: in altre parole, i soggetti sono punti in spazi di p dimensioni con $p =$ numero di variabili. Questo fa sì che ogni unità statistica sia rappresentabile mediante un punto in uno spazio di p dimensioni. In particolare, quando $p = 2$ o $p = 3$ il grafico che si ottiene rappresentando tutte le unità statistiche nel piano o nello spazio è detto *nube di punti* o *scatterplot*. Tuttavia, quando si opera in una dimensione superiore allo spazio (tridimensionale), non è possibile rappresentare graficamente la mappa a causa dell'elevata dimensione dello spazio, quindi si ricorre alla PCA.

Pertanto, la PCA si applica in questo contesto: quando i prodotti, le marche, i consumatori o altre unità statistiche che si stanno analizzando vengono pensati come punti in iperspazi p-dimensionali (in genere p è piuttosto elevato), definiti da un insieme di variabili statistiche di interesse per il posizionamento oggettivo o per la segmentazione.

La PCA si utilizza per ridurre la dimensionalità di una variabile p-dimensionale quantitativa. In questo modo, nel contesto del posizionamento oggettivo, diviene possibile rappresentare lo scatterplot dei punti, in seguito alla riduzione della dimensione dello spazio originario.

La PCA opera esclusivamente con variabili di tipo quantitativo, tuttavia è comune la sua applicazione anche con variabili qualitative ordinali, espresse su scala a intervalli o su scala di rapporto, che derivano da quesiti posti nella forma di batterie di item in scala di

Likert[13]. In quest'ultima fattispecie, a rigore non si potrebbe applicare la PCA. Per questo motivo esiste una tecnica analoga, adatta a situazioni di questo tipo: si parla di PCA per variabili categoriali o PCA non lineare. Tuttavia, quando la scala in cui sono espressi i punteggi è sufficientemente ampia (da 1 a 10 e non da 1 a 4, per esempio), la forzatura che si opera applicando la PCA non comporta, in genere, distorsioni gravi.

Ai fini del posizionamento oggettivo delle unità statistiche, le valutazioni espresse dagli intervistati vengono sottoposte alla PCA per individuare le dimensioni latenti (componenti principali), che sintetizzano tali valutazioni, dove ciascuna componente riassume un insieme di variabili originarie tra loro correlate.

Spesso risulta utile confrontare le unità statistiche rispetto a coppie di componenti principali (PC) e rappresentarle nel piano costituito dalle PC: in questo modo è possibile individuare prodotti o marche tra loro in concorrenza e/o spazi vuoti (nicchie), che rivelano l'assenza di prodotti o marche in grado di soddisfare particolari combinazioni di variabili[14].

2.2.1 Definizione di PCA

La PCA è una delle tecniche maggiormente utilizzate per la riduzione del numero di variabili[15]. 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 p variabili iniziali, effettuando una buona sintesi della variabile p-dimensionale in una dimensione inferiore. Lo scopo della PCA è ridurre il numero più o meno elevato di variabili che descrivono un dataset a un numero minore di variabili latenti, limitando il più possibile la perdita di informazioni.

Tale metodologia consente di pervenire a una trasformazione proiettiva, che sostituisce un insieme di variabili originarie con un numero inferiore di componenti principali,

[13] Tale tecnica consiste nel definire un certo numero di affermazioni, gli item, che esprimono un atteggiamento positivo e negativo rispetto a uno specifico oggetto. Per ogni item si presenta una scala di accordo/disaccordo o soddisfazione/insoddisfazione. Ai rispondenti si chiede di indicare su di esse il loro grado di accordo/disaccordo o soddisfazione/insoddisfazione con quanto espresso dall'affermazione.

[14] De Luca A. (2016). *Modelli di marketing: statistica per le analisi di mercato: segmentazione, posizionamento, comunicazione, innovazione, customer satisfaction*. FrancoAngeli.

[15] De Luca A. (2010). *Le applicazioni dei metodi statistici alle analisi di mercato. Manuale di ricerche di mercato*. FrancoAngeli.

ottenute come combinazione lineare di tali variabili, senza che ciò comporti una perdita significativa dell'informazione (variabilità) contenuta nei dati di partenza.

La PCA consiste nella proiezione ortogonale di una nube di punti p-dimensionale in un sottospazio q-dimensionale (con $q < p$), scelto in modo da massimizzare la variabilità dei punti proiettati.
Dunque, sono valide tutte le osservazioni già citate nel capitolo precedente a proposito della riduzione di dimensionalità:

1) La riduzione di dimensionalità si ottiene attraverso una proiezione ortogonale, cui è connessa inevitabilmente una perdita di informazione.
2) Per minimizzare tale perdita bisogna fare in modo che sia massima la variabilità dei punti proiettati nel sottospazio e minima la variabilità dei punti attorno al sottospazio.
3) Vi sono casi in cui la conformazione della nube di punti comporta una consistente perdita di informazione e casi in cui ciò non avviene.
4) Una riduzione della dimensionalità con piccola perdita di informazione si ha quando le variabili sono linearmente correlate. La perdita di informazione è tanto minore quanto più le variabili sono linearmente correlate tra loro.

2.2.2 Procedura algebrica della PCA

Se la PCA consiste in una proiezione ortogonale, dal punto di vista algebrico essa si effettua trasformando la variabile p-dimensionale X in una variabile q-dimensionale Y attraverso la trasformazione lineare $Y = UX$, dove U è una matrice ortogonale che consente la proiezione ortogonale.
Tutto il problema si riduce alla determinazione di U, in modo da assicurare che la proiezione ortogonale sia la migliore possibile in termini di informazione trattenuta (sia, cioè, la proiezione che massimizza la variabilità dei punti proiettati).

Per capire come U possa essere determinata, si procede ad analizzare il problema da un punto di vista nuovo, che attiene al trattamento dell'informazione che i dati portano con sé. Dunque, si consideri che in un dataset sia contenuto un certo quantitativo di informazione, dove ognuna delle p variabili osservate si occupa di portarne una parte.

Vi sono due modi in cui le p variabili possono portare informazione: con ridondanza di informazione oppure senza ridondanza di informazione. La ridondanza di informazione è dovuta alla correlazione esistente tra le variabili: in presenza di ridondanza di informazione la dimensione dello spazio in cui è definita la variabile osservata viene utilizzata in modo inefficiente.
A questo punto, si potrebbe effettuare una trasformazione della variabile X, in modo da eliminare la ridondanza di informazione. Una volta redistribuita l'informazione in maniera più efficiente, si può decidere di tralasciare qualche variabile con una piccola perdita di informazione; in alternativa, si può accettare una perdita maggiore mantenendo un numero ancora inferiore di variabili, a seconda degli scopi dell'analisi.

Le variabili Y, ottenute trasformando le variabili X con il fine di redistribuire l'informazione in modo più efficiente, si chiamano *componenti principali* (o, un po' impropriamente, *fattori*) di X.
Come citato in precedenza, la ridondanza di informazione è indotta dalla correlazione tra le variabili, quindi per ottenere le Y è necessario cercare di determinare variabili incorrelate tra loro, poiché questo assicura l'assenza di ridondanza. Questa è un'operazione già conosciuta, in quanto è stata trattata nel paragrafo "Gli autovalori e gli autovettori" (si veda la sintesi a pag. 48). Si ricordi, infatti, che se X è una variabile p-dimensionale con matrice di varianze e covarianze Σ_X e H è la matrice degli autovettori normalizzati di Σ_X, allora la variabile $Y = HX$ è una variabile p-dimensionale composta da variabili incorrelate e con varianze pari agli autovalori di Σ_X.

Nel contesto della PCA, si può affermare quanto segue: se X è una variabile p-dimensionale con matrice di varianze e covarianze Σ_X e H è la matrice degli autovettori normalizzati di Σ_X, allora la variabile $Y = HX$ è composta da p variabili incorrelate, ovvero le cosiddette componenti principali della variabile X. Pertanto, la redistribuzione efficiente dell'informazione si effettua attraverso una rotazione dello spazio p-dimensionale in cui è definita X.
Si parte da X in cui è presente correlazione e si arriva a Y, in cui la correlazione è assente.

Si può dimostrare che la matrice H che consente la redistribuzione ottimale dell'informazione contiene la matrice di proiezione U che genera la proiezione ortogonale ottimale dal punto di vista dell'informazione trattenuta (variabilità dei punti proiettati nel sottospazio). In particolare, se si proietta in un sottospazio di dimensione q, U è costituita dalle prime q righe di H.

Quando la scala e/o l'unità di misura delle variabili sono molto diverse tra loro, conviene utilizzare le variabili standardizzate. Tuttavia, con la PCA non serve calcolare effettivamente la corrispondente variabile standardizzata Z perché $\Sigma_Z = \mathrm{P}_X$, dove:

$$\mathrm{P}_X = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \cdots & 1 \end{bmatrix}$$

è la matrice di correlazione di X, cioè la matrice che contiene tutti 1 sulla diagonale principale e il coefficiente di correlazione lineare tra la variabile i-esima e la variabile j-esima alla posizione ij. In altre parole, la matrice di varianze e covarianze delle variabili standardizzate è uguale alla matrice di correlazione delle variabili originali e ciò significa che calcolare autovalori e autovettori della matrice di varianze e covarianze della variabile standardizzata Z equivale esattamente a calcolare autovalori e autovettori della matrice di correlazione della variabile originaria X.

Si consideri, dunque, la matrice di correlazione di X come il punto di partenza. Si ricavino, poi, gli autovalori della matrice di correlazione di X e gli autovettori della matrice di correlazione di X, corrispondenti agli autovalori ordinati (si ricordi la matrice degli autovettori normalizzati $H_{p\times p}$).
Le componenti principali Y si ottengono attraverso la trasformazione lineare (rotazione) $Y = HX$. Matematicamente si ottengono p variabili Y come combinazione lineare delle variabili X originali, con pesi opportunamente definiti:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$
$$Y_3 = a_{31}X_1 + a_{32}X_2 + \cdots + a_{3p}X_p$$
$$\vdots$$

I pesi a_{ij} determinano la posizione esatta del sottospazio in cui si proiettano i punti e vengono calcolati tramite un algoritmo di ottimizzazione. Vi sono, in tutto, tante componenti principali quante sono le variabili osservate e ognuna si ottiene come combinazione lineare delle variabili di partenza, sotto il vincolo di non correlazione con tutte le precedenti.

Calcolando le componenti principali Y è stata effettuata la redistribuzione ottimale dell'informazione. Tuttavia, il risultato di tale redistribuzione ottimale può essere di vario genere. Una buona redistribuzione:

- si ha quando l'informazione è molto concentrata su poche variabili;
- significa che la nube di punti p-dimensionale ha una conformazione che si presta bene alla riduzione di dimensionalità;
- si verifica quando la correlazione tra le variabili iniziali è elevata.

Una pessima redistribuzione:

- si ha quando l'informazione è equidistribuita su tutte le variabili;
- significa che la nube di punti p-dimensionale ha una conformazione che non si presta alla riduzione di dimensionalità;
- si verifica quando la correlazione tra le variabili iniziali è ridotta.

Quando ogni variabile porta con sé una quota di informazione uguale a tutte le altre, tralasciare alcune variabili comporta la massima perdita di informazione possibile.

La percentuale di informazione contenuta in ogni componente principale è $\frac{1}{p} \times 100$. Quindi, nel peggiore dei casi, la riduzione di dimensionalità a uno spazio di q dimensioni sarà in grado di portare con sé una percentuale di informazione pari a $q \times \left(\frac{1}{p} \times 100\right)$. Tale valore deve essere considerato come punto di riferimento: dati p e q, è la performance peggiore in assoluto.

Poiché le componenti principali sono ottenute con una particolare proiezione (ovvero una rotazione), i valori che esse assumono corrispondono alle proiezioni dei punti originali sullo spazio da esse definito.
Per valutare la performance ottenuta dall'operazione di riduzione di dimensionalità, si ricordi che l'informazione trattenuta da ciascuna componente principale è data dalla variabilità dei punti nel sottospazio unidimensionale da essa definito. Dunque, la varianza delle variabili Y (le componenti principali) può indicare quanta informazione è contenuta in esse.

A questo punto, si rende necessario calcolare la varianza delle componenti principali, ricordando che la variabile $Y = HX$ è una variabile p-dimensionale avente varianze pari agli autovalori di Σ_X.
Ora, la matrice di varianze e covarianze della variabile p-dimensionale Y è data da:

$$\Sigma_Y = H\Sigma_Z H' = H\mathrm{P}_X H' = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_{p-1} & 0 \\ 0 & 0 & \dots & 0 & \lambda_p \end{bmatrix}$$

Il passo successivo è considerare gli autovalori λ_p della matrice di correlazione di X, che sono stati calcolati assieme al calcolo degli autovettori. Poiché gli autovalori corrispondono alle varianze delle componenti principali, ogni autovalore si può interpretare come la quantità di informazione contenuta in ogni singola componente principale.
Un'importante osservazione da rilevare è che quando gli autovalori sono calcolati sulla matrice di correlazione (cioè sulla matrice di varianze e covarianze delle variabili standardizzate), la loro somma è pari a p.
In seguito, dividendo ciascun autovalore per p e moltiplicando per 100, si ottiene la percentuale di informazione trattenuta da ogni componente principale. Poi, si calcola la cumulata, che indica la percentuale di informazione trattenuta dalle prime q componenti principali. La proiezione si effettua decidendo, attraverso opportuni indicatori, di considerare solo un numero ridotto (q) di componenti principali, che costituiranno gli assi dello spazio q-dimensionale sul quale si proietta. Le componenti principali

selezionate costituiscono la miglior sintesi (secondo criteri statistici) delle p variabili iniziali, con ciò intendendo che contengono la maggior parte possibile dell'informazione che può essere trasportata in una dimensione inferiore. Da questo punto in avanti, le componenti principali scelte saranno utilizzate per condurre le analisi al posto delle p variabili iniziali.

2.2.3 La qualità della sintesi

La PCA, dunque, pone un primo problema: si tratta di stabilire la qualità della sintesi, intesa come percentuale di informazione trattenuta dalla proiezione, e decidere se è opportuno o meno effettuare la proiezione. In caso affermativo, si procede a interpretare le componenti principali, in quanto esse sono le nuove variabili che verranno analizzate, quindi bisogna attribuire loro un corretto significato. Questo, a sua volta, permette di interpretare il fenomeno oggetto di studio.
Per stabilire la qualità della sintesi, si procederà all'esame degli autovalori e dello Scree Plot. In seguito, per l'interpretazione delle PC, si procederà all'analisi del Factor Loadings Plot.

Per decidere la dimensione ottimale, ovvero stabilire la qualità della sintesi, vanno valutati congiuntamente tre criteri:

1) La % di informazione trattenuta dalle componenti principali (cumulata). In questo step, per valutare la bontà del risultato, è anche possibile effettuare un confronto con la peggior performance in assoluto, e quindi con l'informazione trattenuta nel peggiore dei casi, pari a $q \times \left(\frac{1}{p} \times 100\right)$.

2) Il numero di autovalori superiori a 1 (criterio di Guttman-Keiser).
   Nel caso peggiore (puramente teorico), si avrebbero tutti gli autovalori pari a 1. Per questo motivo il valore 1 è considerato una sorta di "spartiacque": le componenti principali con un autovalore superiore a 1 hanno accumulato informazione. Dunque, un altro criterio per decidere la dimensione ottimale per la riduzione della dimensionalità è scegliere il numero di componenti principali che presentano un autovalore maggiore di 1.

3) Lo Scree Plot (la spezzata degli autovalori): si tratta di un utile strumento grafico per decidere la dimensione ottimale del sottospazio in cui proiettare. Sull'asse *X* vengono indicate le componenti principali (*Number of Factors*), e, in corrispondenza di ognuna di esse, si rappresenta il corrispondente autovalore (*Eigenvalue*) sull'asse *Y*. Il grafico segnala un'analisi di buona qualità quando si osserva una caduta brusca della spezzata, con ciò significando che l'operazione di concentrazione dell'informazione nelle prime PC per la proiezione in sottospazi di dimensioni inferiori è avvenuta con successo, ovvero la ripartizione dell'informazione è buona. In seguito, va identificato il cosiddetto gomito della spezzata, in quanto la dimensione ottimale per la proiezione è quella subito prima del gomito (Figura 2.9).

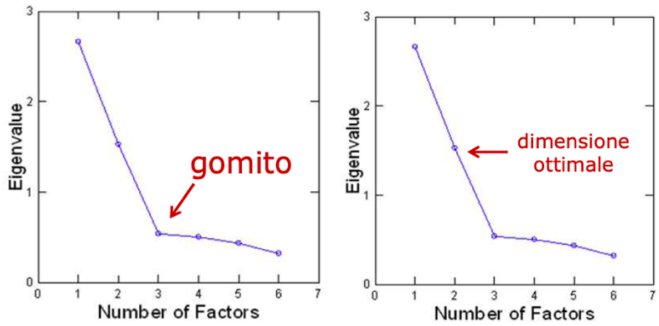


Figura 2.9: Esempio di Scree Plot con indicazione del gomito e della dimensione ottimale (Fonte: corso di "Statistica per il marketing" A.A. 2022/2023 – Prof.ssa Paola Zuccolotto)

Se nella spezzata si osserva una brusca caduta, allora la quantità di informazione portata dalle prime PC è molto elevata e si è in presenza di una buona proiezione. Al contrario, Scree Plot che decrescono gradualmente segnalano consistenti perdite di informazione dovute alla proiezione, ovvero situazioni di cattiva ripartizione dell'informazione (Figura 2.10).

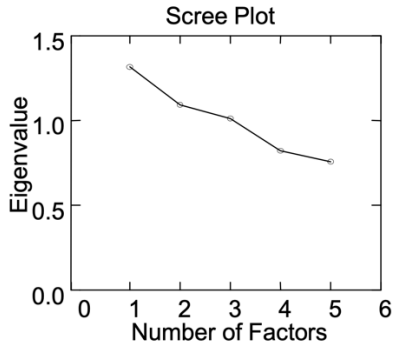


Figura 2.10: Esempio di Scree Plot derivante da una sintesi di scarsa qualità
(Fonte: corso di "Analisi dei Dati" A.A. 2010/2011 – Prof.ssa Paola Zuccolotto)

Nel caso peggiore (puramente teorico) si avrebbe uno Scree Plot perfettamente orizzontale, con tutti gli autovalori pari a 1 (Figura 2.11).

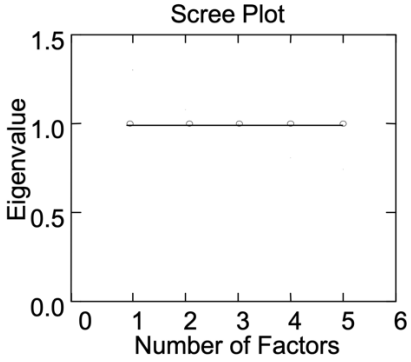


Figura 2.11: Scree Plot orizzontale
(Fonte: corso di "Analisi dei Dati" A.A. 2010/2011 – Prof.ssa Paola Zuccolotto)

È importante considerare che non sempre i tre criteri portano la stessa informazione, pertanto vanno analizzati unitamente.

In generale, si predilige $q = 2$ o al massimo $q = 3$. Tuttavia, mantenere tre componenti principali è una scelta "scomoda" poiché dal punto di vista della facilità di rappresentazione è meglio disporre di una variabile bidimensionale. In certi casi, invece, è opportuno scegliere tre dimensioni, tuttavia è una decisione che va valutata attentamente, anche in funzione dell'interpretazione del significato delle componenti principali.

Ridurre a uno spazio bidimensionale (piano) significa scegliere di mantenere le prime due componenti principali, calcolate mediante l'operazione matriciale:

$$Y_{1:2} = H_{1:2}X$$

dove $H_{1:2}$ è la matrice di dimensione $(2 \times p)$ formata dalle prime due righe di H.
La matrice H che consente la redistribuzione ottimale dell'informazione contiene la matrice di proiezione U che genera la proiezione ortogonale ottimale dal punto di vista dell'informazione trattenuta. In particolare, se si proietta in un sottospazio di dimensione q, U è costituita dalle prime q righe di H. Nel caso di riduzione a uno spazio bidimensionale, U corrisponde a $H_{1:2}$. Perciò, quando si scelgono le prime q componenti principali e si selezionano, quindi, le prime q righe della matrice H, da un punto di vista geometrico si proietta ortogonalmente nel sottospazio q-dimensionale definito dai primi q versori ortogonali di H.

2.2.4 Interpretazione delle componenti principali

Una volta selezionato il numero q di componenti principali da mantenere, che possono essere analizzate in sostituzione delle p variabili iniziali contenute in X, si pone il problema dell'attribuzione di un significato alle stesse q componenti principali.
Il problema dell'interpretazione delle componenti principali è molto delicato: occorre dare un significato appropriato alle variabili $Y_1, Y_2, ..., Y_q$, altrimenti l'intera analisi perde di valore. Le componenti principali sono combinazioni lineari delle variabili di partenza, quindi, in genere, hanno un significato complesso, spesso più astratto o generale rispetto a quello delle singole variabili osservate, che può essere individuato soltanto con una buona conoscenza della realtà che si sta studiando.
Lo strumento che consente di interpretare ogni singola componente principale sono i cosiddetti *loadings* (punteggi) attribuiti a ognuna delle p variabili iniziali per esprimere il contributo che esse conferiscono al significato complessivo della PC. Per ogni PC, si ottengono tanti loadings quante sono le variabili inserite nell'analisi.

I loadings sono coefficienti di correlazione lineare tra le variabili originarie (X_p) e le componenti principali: essi indicano quanto ciascuna PC è correlata con quelle variabili, ossia quanto forte è la correlazione lineare, positiva o negativa, tra una variabile e la PC. Per un dato fattore Y_j, i loadings si calcolano moltiplicando il j-esimo autovettore (la j-

esima riga di H, dove H è la matrice degli autovettori normalizzati) per la radice quadrata dell'autovalore corrispondente.

Da questo punto in poi, per un'efficace interpretazione dei risultati ottenuti, si considererà $q = 2$ (che permette una maggior facilità di rappresentazione di una variabile bidimensionale).

Siano dati i loadings delle q componenti principali selezionate, con $q = 2$.

Innanzitutto, per ogni variabile X_p, si osservi con quale componente principale ha il coefficiente di correlazione più elevato in valore assoluto. Questo permette di capire quali componenti principali sintetizzano quali variabili originarie. La comprensione di questo aspetto viene agevolata dal *Factor Loadings Plot* (grafico dei punteggi delle PC), rappresentazione grafica in cui ogni variabile originaria è rappresentata con un punto nel piano (Y_1, Y_2), di coordinate pari ai loadings che essa ha con ognuna delle due componenti principali. Ogni punto viene poi unito all'origine degli assi tracciando un segmento; in seguito, ogni variabile viene "assegnata" all'asse con cui il suo segmento forma l'angolo più piccolo.

In questo modo, si individua un gruppo di variabili correlate tra loro e con la prima componente principale e incorrelate con quelle dell'altro gruppo e con la seconda componente principale, e viceversa (Figura 2.12).

In altre parole, si tratta di due gruppi di variabili dal significato simile, poiché le variabili di entrambi portano informazione parzialmente sovrapposta tra loro, ma non sovrapposta a quella portata dalle variabili dell'altro gruppo.

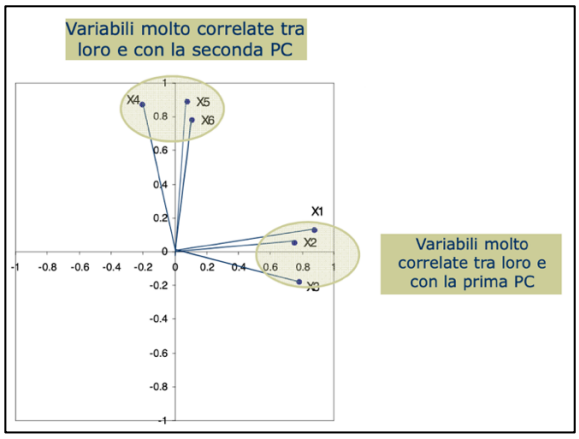


Figura 2.12: Gruppi di variabili individuati nel Factor Loadings Plot
(Fonte: corso di "Statistica per il marketing" A.A. 2022/2023 – Prof.ssa Paola Zuccolotto)

Nella Figura 2.13 viene rappresentato un esempio di ottimo Factor Loadings Plot, nel quale:

1) le variabili tendono a raggrupparsi (ciò indica la presenza di ridondanza di informazione);
2) le variabili appartenenti a gruppi diversi hanno segmenti pressoché perpendicolari tra loro;
3) i segmenti sono molto vicini a uno dei due assi fattoriali.

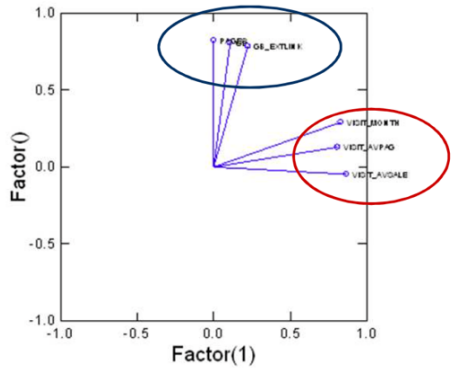


Figura 2.13: Esempio di ottimo Factor Loadings Plot
(Fonte: corso di "Statistica per il marketing" A.A. 2022/2023 – Prof.ssa Paola Zuccolotto)

Nella Figura 2.14, anche se vi sono variabili isolate, il grafico può andare bene ugualmente, purché valgano le condizioni 2 e 3 precedentemente citate: dunque, le variabili con ridondanza di informazione non sono solo quelle vicine, ma in generale tutte quelle che sono correlate (positivamente o negativamente) con lo stesso fattore. Si dovrà poi prestare attenzione all'interpretazione dei fattori.

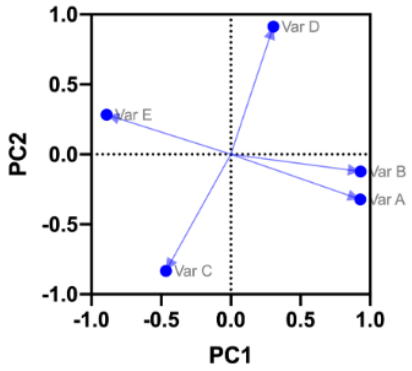


Figura 2.14: Esempio di Factor Loadings Plot
(Fonte: https://www.graphpad.com/guides/prism/latest/statistics/stat_pca_example_loadings_plot.htm)

Di seguito si presentano altri particolari tipi di Factor Loadings Plot e relative osservazioni.

La presenza di una variabile con segmento molto corto indica la necessità di includere un'altra dimensione (Figura 2.15, grafico a sinistra). Anche la presenza di variabili che contribuiscono in misura pressoché uguale alle due componenti principali può indicare, ma non è certo, la necessità di includere un'altra dimensione (Figura 2.15, grafico al centro). Il grafico a destra della Figura 2.15 mostra una situazione in cui non c'era ridondanza di informazione tra le variabili originarie, dunque tutte le variabili incidono debolmente su entrambi i fattori. Questo significa che si tratta di un caso di pessima redistribuzione dell'informazione.

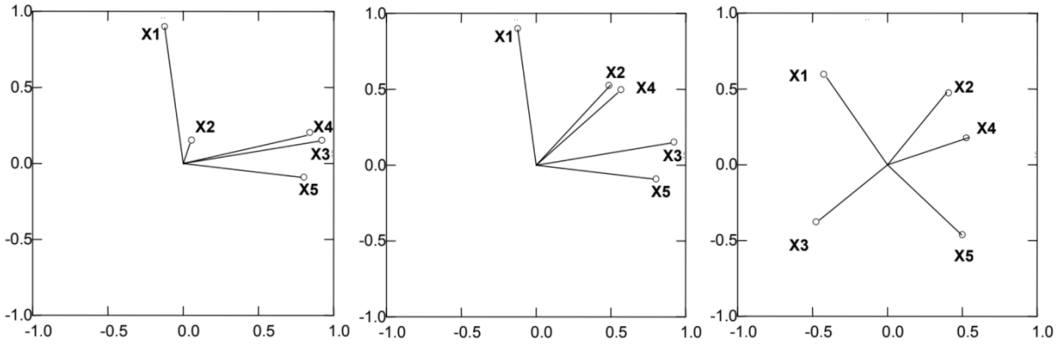


Figura 2.15: Altri esempi di Factor Loadings Plot
(Fonte: corso di "Analisi dei Dati" A.A. 2010/2011 – Prof.ssa Paola Zuccolotto)

Il passaggio successivo consiste nell'interpretazione delle componenti principali Y_1 e Y_2, considerando le variabili originarie X_p, alle quali sono correlate.

In seguito, si procede a rappresentare lo scatterplot di Y_1 e Y_2 ricordando che la nube di punti risultante è la proiezione ortogonale sul piano della variabile originaria definita in p dimensioni. Si tratta di una proiezione che mantiene una complessiva percentuale di informazione pari a quella trattenuta dalle prime q componenti principali selezionate. Normalmente, però, lo scatterplot è poco significativo in assenza di qualche informazione aggiuntiva. Nel caso in cui nel dataset sia presente una variabile qualitativa (che quindi non ha potuto essere inserita nell'analisi), essa può essere impiegata per migliorare la comprensione dello scatterplot (Figura 2.16).

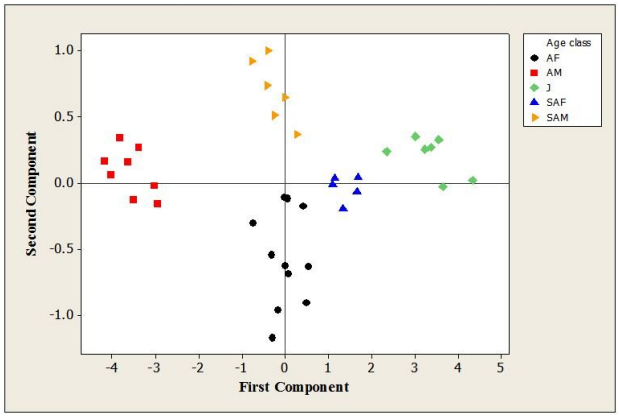


Figura 2.16: Esempio di scatterplot di due componenti principali, con presenza di una variabile qualitativa
(Fonte: https://www.researchgate.net/publication/320280481_Maturity_Stage_Categorization_of_Endemic_Lizard_Calotes_nigrilabris_in_the_Grasslands_of_HPNP)

2.2.5 La rotazione Varimax

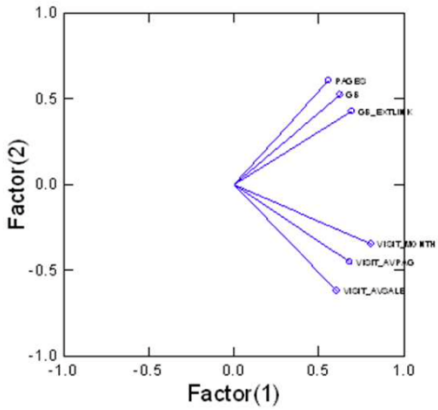


Figura 2.17: Esempio di Factor Loadings Plot non ottimo
(Fonte: corso di "Statistica per il marketing" A.A. 2022/2023 – Prof.ssa Paola Zuccolotto)

Osservando la Figura 2.17, si nota che nel Factor Loadings Plot le variabili tendono a raggrupparsi e le variabili appartenenti a gruppi diversi hanno segmenti praticamente perpendicolari tra loro; tuttavia, risulta difficile assegnare i segmenti a uno dei due assi fattoriali. Si tratta di un caso di difficoltà di interpretazione dei loadings, pertanto è conveniente procedere all'utilizzo dell'operazione di rotazione del piano. Si nota, infatti, che il grafico in esame non è molto diverso dal grafico definito ottimo, ma è solamente ruotato.

In particolare, si applicherà la rotazione Varimax, che permette di ruotare la soluzione bidimensionale (Figura 2.18). La percentuale di informazione spiegata da ogni fattore cambia, ma ovviamente non cambia il totale dell'informazione contenuta nella rappresentazione bidimensionale. La rotazione Varimax significa imporre il miglior bilanciamento possibile tra le due PC, mantenendo invariata la quantità di informazione complessiva.

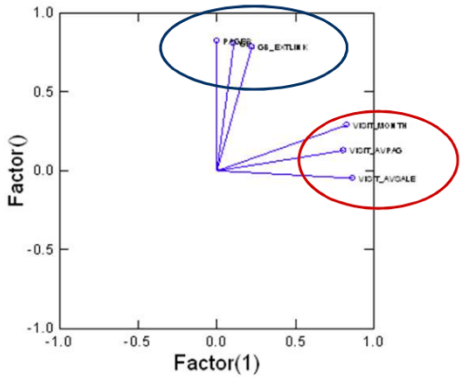


Figura 2.18: Factor Loadings Plot dopo la rotazione Varimax
(Fonte: corso di "Statistica per il marketing" A.A. 2022/2023 – Prof.ssa Paola Zuccolotto)

Il metodo Varimax minimizza, per ogni fattore, il numero di variabili che hanno elevato loading su di esso, permettendo la redistribuzione dell'informazione nel modo più efficiente possibile.

Infine, eseguita la rotazione, si procede alla realizzazione dello scatterplot dei punti proiettati nel sottospazio bidimensionale di Y_1 e Y_2 ruotate.

Risulta importante citare il caso in cui la dimensione ottimale scelta sia $q = 3$. Nella Figura 2.19 si rappresenta un esempio di Factor Loadings Plot e di scatterplot tridimensionale.

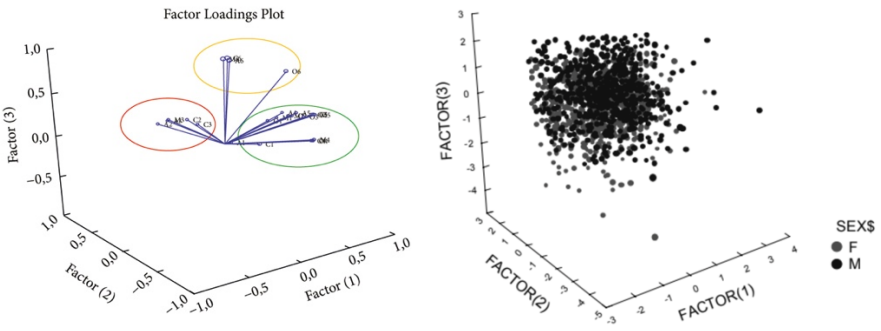


Figura 2.19: Factor Loadings Plot e scatterplot tridimensionali
(Fonte: corso di "Analisi dei Dati" 2010/2011 – Prof.ssa Paola Zuccolotto)

A questo punto, si possono distinguere due situazioni:

1) le unità statistiche sono in numero limitato, quindi il problema dello scatterplot tridimensionale si può risolvere rappresentando una scatterplot matrix;
2) le unità statistiche sono in numero molto alto, dunque lo scatterplot tridimensionale dà luogo a una rappresentazione scarsamente informativa, per due ordini di motivi: la difficoltà di lettura di un grafico tridimensionale e l'elevata numerosità dei soggetti.

2.2.6 Conclusioni sulla PCA

In sintesi, la PCA è una tecnica di statistica multidimensionale, la quale, partendo da una matrice di dati di dimensioni $n \times p$ (con $n =$ numero di unità statistiche osservate, $p =$ numero di variabili quantitative rilevate), è in grado di sostituire le p variabili tra loro correlate con un nuovo set di variabili composite, denominate componenti principali, le quali hanno le seguenti proprietà:

- sono incorrelate tra di loro (ortogonali);
- sono elencate in ordine decrescente rispetto alla quota di varianza totale da esse spiegata.

La PCA è una tecnica di riduzione della dimensionalità lineare che converte un insieme di variabili correlate nello spazio ad alta dimensionalità in una serie di variabili (componenti principali) non correlate nello spazio a bassa dimensionalità. La PCA è una

trasformazione lineare ortogonale, il che significa che tutte le componenti principali sono perpendicolari tra loro.

La PCA può essere definita come una tecnica di riduzione e interpretazione dei dati, in quanto consente di:

1) ridurre il numero di variabili da considerare, escludendo determinate componenti principali, laddove si ritenga trascurabile il loro contributo alla spiegazione della variabilità osservata;
2) interpretare il fenomeno oggetto di studio, mediante un'opportuna interpretazione delle componenti principali che sono state mantenute nell'analisi.

L'idea alla base della PCA è di ridurre p variabili in un numero inferiore di fattori latenti, preservando quanta più variabilità possibile dei dati iniziali (Ringnér, 2008). Pertanto, la PCA può essere spiegata come una procedura in cui la ricerca delle direzioni ortogonali di massima variabilità (varianza) dei dati è l'obiettivo principale. Matematicamente, queste direzioni sono dimostrate essere equivalenti agli autovettori nell'algebra lineare e vengono ricavate attraverso il cosiddetto *problema di decomposizione degli autovalori* (Stein-O'Brien *et al.,* 2018), tale per cui si ottiene una certa frazione della variabilità dei dati di partenza associata a ciascuna componente principale ortogonale.

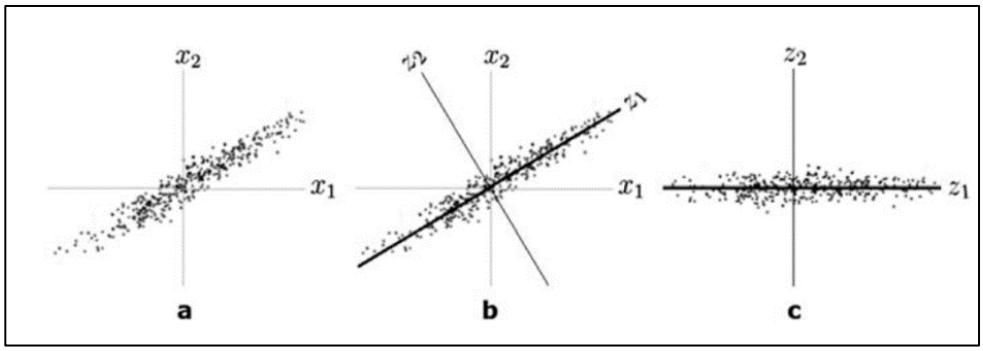


Figura 2.20: PCA con due componenti principali z_1 e z_2
(Fonte: Deng, L. Y., Garzon, M., & Kumar, N. (2022). What Is Dimensionality Reduction (DR)? In Dimensionality Reduction in Data Science (pp. 67-77). Cham: Springer International Publishing)

La prima componente principale è la direzione che massimizza la varianza di una proiezione dei dati; la seconda componente principale può essere considerata come una

direzione ortogonale alla prima componente principale, che massimizza la varianza delle componenti rimanenti dei dati proiettati, come illustrato in Figura 2.20.
Iterando il processo, la PCA calcola la terza componente principale in modo ortogonale alle prime due PC.
La PCA riduce la dimensionalità proiettando ciascun dato solo sulle prime poche componenti principali per ottenere rappresentazioni dei dati a dimensionalità inferiore, preservando al massimo la variabilità dei dati.
In generale, la PCA equivale a un cambio di base degli assi di coordinate nello spazio delle variabili, mediante rotazioni appropriate (come rappresentato in Figura 2.20), al fine di catturare la massima variabilità dei dati lungo i nuovi assi. Pertanto, la PCA estrae variabili che conservano la maggior quantità di variabilità (varianza) nei dati ad alta dimensionalità.

La forza della PCA è quella di condensare in q macro-gruppi le p variabili iniziali, creando degli indicatori di sintesi. Questo avviene individuando gruppi di variabili originarie, correlate tra loro, ovvero che portano informazione simile (parzialmente sovrapposta). Vengono così determinate le componenti principali, le quali vengono mantenute in una quantità inferiore rispetto alle variabili iniziali: questo comporta inevitabilmente una perdita di informazione. L'obiettivo è conservare il minor numero possibile di PC, rinunciando a una quota trascurabile (ovvero che non comprometta la qualità dell'analisi) di informazione contenuta nei dati di partenza.

2.2.7 I limiti della PCA

Sebbene, in generale, la PCA sia considerata una tecnica adeguata al fine di eseguire la riduzione della dimensionalità dei dati (Fabrigar *et al.,* 1999), è necessario considerare che tale metodo soffre di due importanti limiti. Innanzitutto, si assume che le relazioni tra le variabili siano lineari; in secondo luogo, l'interpretazione della PCA è esclusivamente valida in presenza di variabili di carattere quantitativo, e, in particolare, richiede che vi sia equidistanza tra i valori assunti dalla variabile quantitativa.
Nell'ambito delle scienze sociali e comportamentali, tali condizioni spesso non sono soddisfatte, pertanto la PCA può non risultare il metodo di analisi più appropriato.

Per superare questi limiti, è stata sviluppata un'alternativa alla PCA, denominata NLPCA o *nonlinear* PCA (*Nonlinear Principal Component Analysis*), la quale può portare a un apprezzabile miglioramento dell'analisi. Tale tecnica è l'equivalente *nonlineare* della PCA classica. Pertanto, la NLPCA ha i medesimi obiettivi della PCA tradizionale, ma consente di trattare variabili statistiche di diversa natura (nominali, ordinali e numeriche) ed è in grado di identificare le relazioni non lineari esistenti tra le stesse.

Nel caso di nostro interesse, in particolare, si farà riferimento a variabili qualitative in scala ordinale, poiché esse si qualificano come le variabili che vengono originate da quesiti caratterizzati da batterie di item in scala di Likert.

2.3 L'Analisi delle Componenti Principali Nonlineare (NLPCA)

La NLPCA si definisce come una tecnica di analisi multivariata che si utilizza per ridurre la dimensionalità di una variabile p-dimensionale anche in presenza di variabili qualitative. Dunque, dati N soggetti su cui sono state osservate p variabili qualitative, si vogliono determinare q nuove variabili ($q < p$) che contengano gran parte dell'informazione che era contenuta nelle p variabili iniziali, ottenendo una buona sintesi della variabile p-dimensionale in una dimensione inferiore. Da tale definizione emerge come la NLPCA sia una variante della PCA classica: mentre quest'ultima permette di svolgere analisi solamente su variabili quantitative, la NLPCA risponde all'esigenza di trattare le variabili qualitative.

Data la complessità che caratterizza tale metodo, nel presente elaborato ci si limiterà a esporre i principali concetti, necessari per comprendere l'applicazione della tecnica al dataset oggetto di indagine (per l'applicazione e i risultati ottenuti si rimanda al capitolo successivo).

2.3.1 La misurazione della customer satisfaction attraverso la NLPCA

Operando all'interno di ambienti sempre più altamente competitivi, le imprese devono adottare comportamenti *customer-oriented*, poiché la soddisfazione del cliente è un elemento fondamentale per l'acquisizione e il consolidamento nel lungo periodo di vantaggi competitivi rispetto ai concorrenti.

Una tecnica ampiamente utilizzata per l'analisi della soddisfazione dei clienti (*customer satisfaction* – CS) è basata su un approccio di riduzione della dimensionalità: si tratta della NLPCA. Tramite la NLPCA, vi è la possibilità di misurare la CS mediante la costruzione di indicatori sintetici che tengano conto di diverse caratteristiche del concetto latente, quali la multidimensionalità e la non misurabilità concreta. Infatti, la CS per un servizio o prodotto non è né facilmente né direttamente quantificabile, ma può essere valutata attraverso l'osservazione di alcune variabili collegate a diversi aspetti del prodotto/servizio in esame. Generalmente, i dati sulla CS vengono raccolti tramite un questionario in cui agli intervistati viene chiesto di dichiarare il proprio grado di soddisfazione rispetto a diversi aspetti del prodotto o servizio. In particolare, la misurazione della CS prevede l'impiego di questionari con batterie di item, che permettono di cogliere la multidimensionalità degli aspetti oggetto di analisi. Si consideri, inoltre, che ogni item corrisponde a una variabile, dunque risulta fondamentale l'impiego di una tecnica che operi una riduzione di dimensionalità per consentire una corretta interpretazione dei dati.

La soddisfazione complessiva si definisce multidimensionale, in quanto si può essere soddisfatti in modo diverso dei differenti aspetti caratterizzanti il prodotto o servizio. Si parla, invece, di non misurabilità concreta poiché le risposte alle batterie di item derivano da percezioni e sentimenti dei rispondenti, dunque si tratta di valutazioni puramente soggettive.

Pertanto, la valutazione quantitativa della CS richiede una certa accortezza considerato che coinvolge variabili psicologiche, le quali per loro natura non sono direttamente misurabili (*variabili latenti*).

Dalle batterie di item vengono generate variabili qualitative ordinali, poiché le risposte sono rilevate su scale con modalità ordinate, ad esempio in scala di Likert (solitamente da 1 a 5 o da 1 a 7). Dunque, trattandosi di una scala qualitativa ordinale e non quantitativa, i giudizi rilevati non hanno proprietà lineari: non si può assumere che vi sia equidistanza tra una modalità di risposta e l'altra, ovvero tra un giudizio di soddisfazione e l'altro.

L'assunto di base della NLPCA è che vi sia equidistanza tra i valori assunti da una variabile quantitativa, ma non necessariamente esista equidistanza tra le categorie di una variabile qualitativa.

Concettualmente, la NLPCA è simile alla PCA classica, salvo il fatto che prevede la trasformazione delle modalità qualitative in valori numerici (*quantificazioni*) secondo un determinato criterio di ottimalità: questo è effettuato tramite una tecnica nota come *optimal scaling*. Tale metodo assegna delle quantificazioni q_i alle modalità di giudizio della scala di Likert e, successivamente, su queste quantificazioni verrà applicata l'Analisi delle Componenti Principali. Quindi, la NLPCA simultaneamente trasforma le variabili qualitative in variabili quantitative e riduce la dimensionalità dei dati, mediante lo svolgimento della PCA sulle quantificazioni. Naturalmente, se le variabili su cui viene applicata la NLPCA sono numeriche, essa fornirà in output una soluzione esattamente identica a quella della PCA classica, poiché il processo di ottimizzazione non sarà necessario e le variabili saranno meramente standardizzate.
Il problema di ottimizzazione vincolata trova soluzione mediante un metodo iterativo chiamato *Alternating Least Squares* (ALS), che consente di svolgere il processo di ottimizzazione delle variabili in quantificazioni e la riduzione della dimensionalità in contemporanea. L'algoritmo ALS funziona nel seguente modo: date le q componenti principali che si vogliono estrarre, si definiscono delle quantificazioni in modo arbitrario, sulle quali viene svolta la PCA classica, calcolando la varianza spiegata trattenuta dalle prime q componenti principali. L'algoritmo procede modificando le quantificazioni al fine di aumentare la varianza spiegata. Tale processo viene iterato fino a convergenza, la quale si ha quando si ottiene la massima varianza spiegata, ovvero finché, per la dimensione q considerata, non è più possibile conseguire un miglioramento della varianza rispetto a quella ottenuta nell'iterazione precedente. L'algoritmo, convergendo nel punto stazionario in corrispondenza del quale le quantificazioni non vengono più modificate, fornisce le quantificazioni ottimali e le componenti principali con massima varianza spiegata in q dimensioni. Le quantificazioni ottimali garantiscono la miglior PCA possibile.

Per l'attribuzione delle quantificazioni alle categorie è necessario imporre dei vincoli (*scaling level*) all'algoritmo, grazie ai quali è possibile scegliere la trasformazione che consente di ottenere le quantificazioni tra alcune alternative:

- Nominal: impone il raggruppamento in categorie, dunque alle unità statistiche che hanno fornito la stessa risposta deve essere attribuita la medesima quantificazione. Si tratta del vincolo meno stringente di tutti.

- Ordinal: impone, oltre al raggruppamento in categorie, anche l'ordine delle stesse. Pertanto, vengono attribuite quantificazioni crescenti a categorie crescenti in modo tale da mantenere l'ordine delle categorie originarie.
- Numerical: impone il raggruppamento in categorie, l'ordine e l'equispazialità delle stesse. Dunque, oltre a quanto previsto dagli scaling precedenti, le categorie devono avere tra loro la stessa distanza. Utilizzando lo scaling Numerical, dall'applicazione della NLPCA si otterrebbero gli stessi risultati che si avrebbero procedendo con la PCA classica.

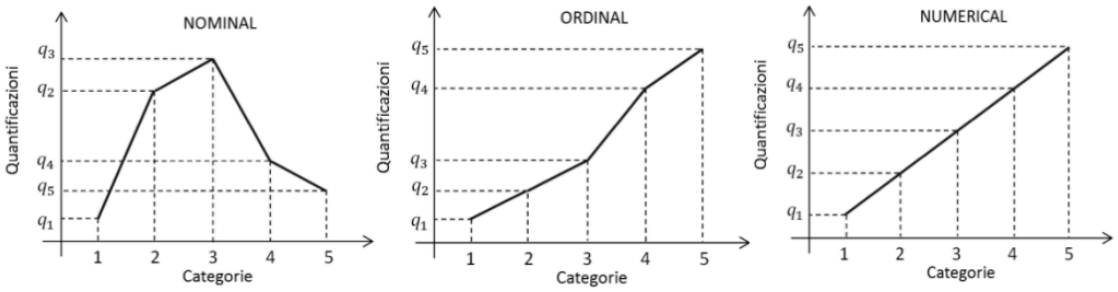


Figura 2.21: Transformation Plot per i diversi scaling level
(Fonte: corso di "Statistica per il marketing" A.A. 2022/2023 – Prof.ssa Paola Zuccolotto)

Come rappresentato nella Figura 2.21, a ciascun scaling level è associato un *Transformation Plot* (grafico di trasformazione) che mostra, per ogni variabile, le quantificazioni delle categorie. Il grafico Nominal rispetta il raggruppamento in categorie, ma non l'ordine delle stesse: si nota, infatti, che le quantificazioni q_i non sono vincolate a un ordine crescente. Di conseguenza, il transformation plot è costituito da una spezzata, che disegna un profilo delle quantificazioni avente una forma arbitraria. Il grafico Ordinal, pur rispettando il raggruppamento in categorie e l'ordine delle stesse, non rispetta l'equispazialità delle stesse; si noti, ad esempio, che la distanza tra le quantificazioni q_1 e q_2 è diversa rispetto alla distanza tra le quantificazioni q_3 e q_4. In questo caso, il Transformation Plot è dato da una spezzata non decrescente: il profilo delle quantificazioni sarà non decrescente, rispettando il vincolo di ordine. D'altra parte, mancando il vincolo di equispaziatura, le quantificazioni rispettano l'idea che i giudizi non siano necessariamente equispaziati tra loro. Infine, il grafico Numerical rispetta tutti

e tre i vincoli: raggruppamento in categorie, ordine originario delle categorie ed equispazialità delle stesse. Il Transformation Plot è crescente e lineare: una retta.

In sintesi, utilizzando i primi due scaling level si andrà a svolgere la NLPCA, al contrario scegliendo il Numerical si svolgerà la PCA classica.
Nel presente elaborato, come verrà applicato nel capitolo successivo, si opterà per trasformare le variabili con una trasformazione di tipo Ordinal, al fine di mantenere anche nelle variabili trasformate l'ordine delle categorie originarie, ottenendo in tal modo delle trasformazioni non decrescenti. In particolare, in questa tesi verrà esplorata su un caso di studio pratico la differenza tra imporre uno scaling level Ordinal (coerentemente con la natura delle variabili) e forzare uno scaling level Numerical (trattando le variabili come se fossero quantitative e imponendo pertanto l'equispaziatura).

L'applicazione della NLPCA comporta la risoluzione di due problemi fondamentali, che sostanzialmente sono gli stessi che si affrontano nell'ambito PCA classica:

1) valutare la qualità della rappresentazione nello spazio a dimensione ridotta;
2) interpretare il significato delle componenti principali, attribuendo loro un corretto significato.

Per quanto riguarda il primo aspetto, ovvero la valutazione della qualità della sintesi, anche con la NLPCA si ricorre all'esame degli *autovalori* e dello *Scree Plot*, calcolando la percentuale di informazione trattenuta dalle componenti principali nel modo già noto dalla PCA classica. In particolare, al fine di scegliere la dimensione ottimale di minori dimensioni, si valutano congiuntamene tre criteri: la percentuale di informazione trattenuta dalle componenti principali (la varianza spiegata cumulata), il numero di autovalori maggiori di 1 e lo Scree Plot. Gli autovalori indicano la percentuale di informazione trattenuta da ciascuna componente principale, mentre lo Scree Plot è la rappresentazione grafica degli autovalori (anche nota come *spezzata degli autovalori*).
Tuttavia, a differenza della PCA classica, è importante considerare che le soluzioni della NLPCA sono *non nested* (non annidate), cioè variano a seconda della dimensione dello spazio in cui si sceglie di proiettare la nube di punti. In pratica, ciò significa che, ad esempio, le prime due PC di una soluzione a tre dimensioni sono differenti dalle due PC di una soluzione a due dimensioni; ugualmente, gli Scree Plot differiscono a seconda

della dimensione scelta. Per questo motivo, si effettua l'analisi *full*, ovvero si proiettano i punti in uno spazio della stessa dimensione di quello di partenza; in sostanza, l'analisi full non riduce la dimensionalità ma calcola solamente le quantificazioni ottimali. In questo modo, sulla base degli autovalori ottenuti dall'analisi full, si derivano alcune indicazioni utili per stabilire la dimensione ottimale dello spazio in cui proiettare la nube di punti.

Successivamente, si effettua la riduzione della dimensionalità. Gli autovalori calcolati con la soluzione a dimensioni ridotte saranno diversi da quelli della soluzione full (diversamente da quanto accade con la PCA classica), tuttavia normalmente si consegue un miglioramento dei risultati rispetto a quelli ottenuti con l'analisi full, e quindi un incremento della varianza spiegata.

Se tutte le variabili iniziali sono altamente correlate, una singola componente principale è sufficiente per descrivere i dati; al contrario, se le variabili formano due o più insiemi e le correlazioni sono elevate all'interno dei gruppi e basse tra i diversi gruppi, è necessaria una seconda o terza componente principale per riassumere le variabili. In caso di soluzioni multidimensionali, le componenti principali sono ordinate in base ai loro autovalori: la prima PC è associata all'autovalore più elevato e rappresenta la maggior parte della varianza, la seconda PC rappresenta quanto più possibile della varianza rimanente, e così via.

Il secondo problema attiene alla corretta interpretazione del significato delle componenti principali, in quanto esse sostituiscono le variabili originarie nelle analisi successive. Analogamente alla PCA classica, gli strumenti utilizzati sono i *loadings* e l'esame del *Factor Loadings Plot*. I loadings sono coefficienti di correlazione lineare tra le variabili originarie e le componenti principali, mentre il Factor Loadings Plot è il grafico dei punteggi delle componenti principali. Si ricordi che i loadings assumono valori tra -1 e $+1$: tanto più il valore è vicino (in termini assoluti) all'unità, maggiore è il contributo che la variabile originaria attribuisce al significato della PC.

Anche nell'ambito della NLPCA, spesso vi è la necessità di effettuare una rotazione della soluzione per facilitare l'interpretazione del Factor Loadings Plot, applicando la rotazione Varimax.

Infine, altri strumenti grafici utili ai fini dell'analisi sono l'*Object Scores Plot* e il *Transformation Plot*.

Tramite l'Object Scores Plot è possibile rappresentare le singole unità statistiche nello spazio di dimensione ridotta, nel quale tali unità vengono posizionate in base ai valori assunti da ciascuna di esse sulle componenti principali (Figura 2.22).

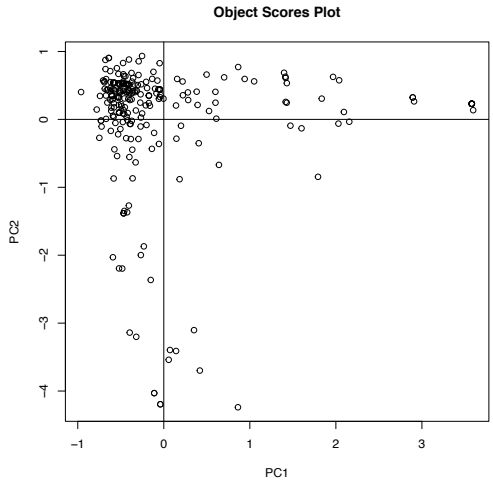


Figura 2.22: Esempio di Object Scores Plot
(Fonte: nostre elaborazioni)

I Transformation Plot sono grafici che, per ogni variabile, mostrano le quantificazioni delle categorie.

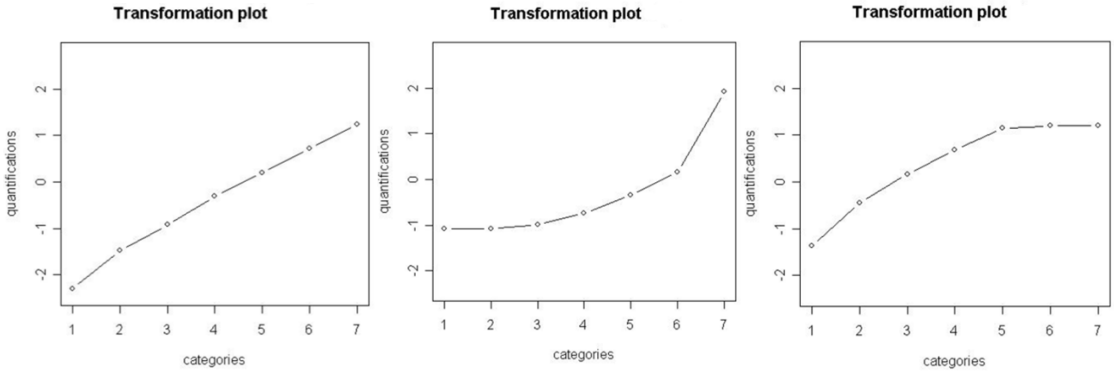


Figura 2.23: Esempi di Transformation Plot con uno scaling level di tipo Ordinal
(Fonte: corso di "Statistica per il marketing" A.A. 2022/2023 – Prof.ssa Paola Zuccolotto)

Nel caso dello scaling level di tipo Ordinal, la trasformazione può essere:

- approssimativamente lineare: ciò significa che le categorie originarie sono approssimativamente equispaziate (Figura 2.23, a sinistra);

- nonlineare e approssimare una funzione convessa: quando è presente una minore distinzione tra le categorie di insoddisfazione e maggiore distinzione tra le categorie di elevata soddisfazione (Figura 2.23, al centro);
- nonlineare e approssimare una funzione concava: quando si ha una maggiore distinzione tra le categorie di insoddisfazione e minore distinzione tra le categorie di elevata soddisfazione (Figura 2.23, a destra).

2.3.2 Conclusioni sulla NLPCA

Complessivamente, la PCA e la NLPCA sono tecniche molto simili in termini di obiettivi, metodo, risultati e interpretazione. Vi è, però, una cruciale differenza: se nella PCA le variabili osservate vengono poi direttamente analizzate, nella NLPCA le variabili osservate sono sottoposte a una trasformazione, dunque le variabili effettivamente analizzate vengono quantificate durante l'analisi stessa (salvo quando le variabili iniziali siano di tipo numerico). Un'altra rilevante dissomiglianza riguarda la cosiddetta *non nestedness* delle soluzioni della NLPCA, per la quale le componenti principali corrispondenti nelle dimensioni q e $q + 1$ sono differenti.
Quindi, quando si è in dubbio circa la scelta della dimensione q ottimale, è consigliabile procedere al confronto con le soluzioni con $q - 1$ e $q + 1$ componenti principali.

I vantaggi più significativi della NLPCA rispetto alla PCA classica sono la possibilità di trattare variabili di natura qualitativa (nominali e ordinali), e la capacità di scoprire relazioni non lineari tra le stesse. Inoltre, la NLPCA può gestire le variabili in base alla loro scala di misurazione: ad esempio, può trattare le scale di tipo Likert ordinalmente anziché numericamente.
Un ulteriore beneficio della NLPCA è che può essere in grado di tenere conto di una maggiore varianza spiegata nei dati rispetto alla PCA lineare, trattandosi di analisi di variabili nominali o ordinali (come dimostrato da alcuni studi e prove empiriche effettuati dai ricercatori). Infatti, per le variabili di tipo non numerico, il metodo è meno restrittivo e può quindi raggiungere una percentuale maggiore di varianza spiegata. Sebbene tale aumento sembri vantaggioso, nei casi in cui le variabili abbiano relazioni solo leggermente non lineari tra loro (ciò spesso si verifica quando vengono misurate scale di Likert), la PCA non lineare non comporterà un rilevante miglioramento della soluzione lineare. Inoltre, si dovrebbe anche considerare se il guadagno in varianza

supera il possibile aumento della complessità della soluzione: se la soluzione diventa molto più difficile da interpretare, si potrebbe voler utilizzare la PCA lineare indipendentemente da qualsiasi incremento di varianza (Meulman et al., 2007).

CAPITOLO 3

APPLICAZIONE DI PCA E NLPCA AL CASO STUDIO

3.1 Progetto DS4BS: Data Science for Brescia – Arts and Cultural Places

Il Laboratorio Big and Open Data Innovation (BODaI-Lab) dell'Università degli Studi di Brescia ha avviato, nel settembre 2021, il progetto denominato "Data Science for Brescia – Arts and Cultural Places" (DS4BS), conclusosi al termine del mese di giugno 2023[16]. L'attività è stata finanziata da Fondazione Cariplo e ha ricevuto il patrocinio del Comune di Brescia (ITIS – Settore Informatica, Innovazione e Statistica) e della Fondazione Brescia Musei, e rientra nelle attività di ricerca realizzate dall'Università degli Studi di Brescia a sostegno dell'evento "Bergamo e Brescia Capitale Italiana della Cultura 2023". L'obiettivo principale del progetto è quello di aumentare la conoscenza sulle modalità di visita e fruizione dei luoghi dell'arte e della cultura, quali musei, teatri, monumenti ed edifici storici della città di Brescia, al fine di supportare le decisioni strategiche di istituzioni e *decision maker*. In particolare, la ricerca è stata mirata alla profilazione dell'utenza dei poli museali della città, individuando caratteristiche, preferenze e attitudini dei visitatori, in modo da elaborare strategie di marketing e di policy utili a migliorare la *visitor experience* e a valorizzare l'offerta museale e il territorio bresciano in generale.

I siti culturali coinvolti in tale progetto sono il Museo di Santa Giulia, il Parco Archeologico di Brescia romana e la Pinacoteca Tosio Martinengo, nei quali l'indagine è stata condotta con uno specifico focus sulla sperimentazione di nuove modalità di *public engagement*, esplorando atteggiamenti e percezioni culturali e nuove forme di accessibilità alla cultura, anche con riferimento al turismo culturale[17].

Il progetto è stato svolto utilizzando un approccio di Data Science, che prevede l'uso combinato di big data e nuove tecnologie con metodi statistici complessi e strumenti predittivi. Tale approccio si articola in due linee di ricerca integrate, così come di seguito:

---

[16] This work has been supported by Fondazione Cariplo, grant n. 2020-4334, project Data Science for Brescia – Arts and Cultural Places (DS4BS).

[17] https://bodai.unibs.it/ds4bs/

a) Linea di ricerca 1: orientata al monitoraggio geo-referenziato e dinamico dei big data provenienti dai segnali di telefonia mobile, allo scopo di rilevare le visite ai siti culturali della città di Brescia. A tal fine viene utilizzato un approccio multistadio per dati ad alta dimensione, che consente di stimare il numero di utenti telefonici su diversi giorni e zone della città tramite l'algoritmo *Histogram of Oriented Gradients* per la riduzione della dimensionalità dei dati, e un mix di k-medie e metodi di *Clustering Model-Based* per l'analisi funzionale dei dati per la profilazione di periodi di tempo. Il focus dell'analisi è la rilevazione dei movimenti delle persone nei luoghi della cultura e lungo gli itinerari culturali. Tale linea di ricerca restituisce informazioni di tipo quantitativo sui flussi di visitatori, grazie alle quali è possibile trarre importanti considerazioni sul grado di apprezzamento e attrattività dei poli museali bresciani.

b) Linea di ricerca 2: orientata alla valutazione dell'esperienza multisensoriale delle persone durante la visita di luoghi artistici e culturali, anche alla luce delle nuove iniziative digitali promosse dai musei a causa dell'emergenza sanitaria da COVID-19. Il focus, dunque, è l'esperienza sensoriale dei visitatori dei siti museali. Per condurre questa linea di ricerca, si è proceduto alla somministrazione di questionari di indagine tramite app su smartphone; i dati raccolti vengono impiegati per quantificare l'esperienza sensoriale dei visitatori, facendo riferimento a fenomeni quali la sinestesia e l'ideastesia, per i quali la visita a un museo o la contemplazione di un quadro possono evocare sensazioni sensoriali (visive, gustative, olfattive, tattili, uditive). In tale contesto, si parla del cosiddetto *sensory museology*. Segue l'elaborazione dei dati tramite l'applicazione di metodi statistici di analisi multivariata: si tratta di metodi idonei a trattare i dati di indagine, in particolar modo i dati generati da quesiti posti nella forma di batteria di item in scala di Likert o scala a differenziale semantico. La valutazione dell'esperienza sensoriale dei visitatori verrà opportunatamente integrata con altre informazioni riguardanti i movimenti dei visitatori stessi lungo i diversi itinerari culturali bresciani, i contenuti multimediali delle varie opere d'arte e il loro adattamento alle diverse categorie di visitatori. Tutto ciò contribuirà a creare una visione olistica delle opere d'arte come *oggetti intelligenti*, ovvero come oggetti digitali fisici, autonomi, potenziati, con

capacità di rilevamento e arricchiti con un'enorme varietà di contenuti multimediali.

Nell'ambito del progetto DS4BS, la rappresentazione di oggetti intelligenti verrà successivamente impiegata per supportare gli esperti competenti del museo nella preparazione di nuovi itinerari culturali per i visitatori (senza che debbano disporre di specifiche capacità tecnologiche), e altresì per offrire ai visitatori un'esplorazione personalizzata del patrimonio culturale sul proprio smartphone.

All'interno del progetto DS4BS la Data Science svolge un ruolo fondamentale, in quanto l'integrazione dei big data provenienti dalle reti di antenne e dalle app installate sugli smartphone dei visitatori dei siti culturali/museali permette di produrre nuova conoscenza, utile poi per definire nuove politiche e processi decisionali al fine di migliorare la qualità dell'offerta e dei servizi. Questo, a sua volta, porterà benefici ai cittadini, al turismo culturale, alla società e all'economia in generale.

3.2 L'indagine statistica e il questionario

Come citato in precedenza, per implementare la seconda linea di ricerca del progetto DS4BS sono stati elaborati dei questionari, rivolti ai visitatori del Parco Archeologico di Brescia romana, del Museo di Santa Giulia e della Pinacoteca Tosio Martinengo. Naturalmente, si tratta di differenti questionari per i diversi siti culturali coinvolti nell'indagine.

Nel presente elaborato, si farà particolare riferimento al questionario relativo al Museo di Santa Giulia, in quanto la parte sperimentale di questa tesi di ricerca verterà su una specifica domanda contenuta in tale questionario.

L'indagine sul campo è stata condotta da aprile a luglio 2022, periodo durante il quale sono state raccolte 665 risposte al questionario somministrato ai visitatori del museo. I questionari sono stati sottoposti al termine delle visite.

L'obiettivo principale è quello di indagare la *visitor experience*, valutando sia la soddisfazione complessiva degli utenti, sia la soddisfazione relativa a specifici aspetti del museo (quest'ultimi verranno successivamente approfonditi nella descrizione del questionario).

Lo scopo dell'indagine sull'esperienza vissuta dagli utenti è quello di identificare dei segmenti nel pubblico di riferimento, sulla base delle preferenze, degli interessi e delle attitudini dal punto di vista culturale dei visitatori, al fine di supportare le decisioni strategiche della direzione museale finalizzate a creare percorsi di visita personalizzati in base al profilo del visitatore. Ciò, a sua volta, consente di migliorare la soddisfazione degli utenti nonché la qualità complessiva della visitor experience, valorizzando l'offerta culturale del Museo di Santa Giulia.

3.2.1 Descrizione del questionario

Il questionario è composto complessivamente da 38 domande, di cui le prime 27 sono strettamente relative all'esperienza vissuta durante la visita del Museo di Santa Giulia, mentre le rimanenti 11 domande riguardano i dati anagrafici dei rispondenti.
Il questionario somministrato è visionabile nella sezione Appendice del presente elaborato.

Le prime tre domande sono volte a indagare le fonti informative che hanno portato i rispondenti a conoscere il Museo di Santa Giulia, così da valutare l'efficacia dei canali di comunicazione impiegati dal museo. In particolare, la prima è una domanda chiusa a risposta multipla in cui viene chiesto di indicare il mezzo tramite cui si è venuti a conoscenza del museo. In seguito, il secondo e terzo quesito sono domande aperte in cui viene chiesto di esplicitare, nel caso in cui il soggetto abbia selezionato l'alternativa "altri social media" o "altri siti web" nella domanda precedente, quale altro social media o quale altro sito web.

Segue una serie di domande relative alla visita presso il museo. Vi sono alcune domande finalizzate a indagare le motivazioni che hanno spinto l'utente a visitare il museo. In particolare, la quarta domanda è di tipo *graduatoria*: questa chiede di mettere in ordine di importanza cinque differenti motivi che possono aver portato il rispondente a visitare il museo, secondo il grado di importanza che egli vi abbia attribuito (dal più importante al meno importante). La quinta domanda è volta a distinguere i rispondenti che per la prima volta hanno visitato il museo dai rispondenti che invece già lo avevano visitato. Si tratta di una domanda filtro in quanto per gli utenti che hanno selezionato la seconda opzione viene sottoposta una specifica domanda (domanda 6) volta a indagare il motivo

del ritorno. La settima domanda è rivolta a coloro i quali nel precedente quesito abbiano selezionato l'opzione "una mostra temporanea", e invita a specificare quale mostra temporanea abbia spinto l'utente a tornare a visitare il museo.
Altri aspetti che vengono investigati nelle successive domande sono con chi è stata condivisa l'esperienza (domanda 8), il mezzo con il quale si è raggiunto il museo (domanda 9), se nella visita è previsto o meno il pernottamento a Brescia (domanda 10), ed eventualmente quante notti sono previste (domanda 11).

A questo punto, vengono sottoposte alcune domande volte a indagare il grado di soddisfazione del visitatore. Lo strumento principale impiegato a tale scopo è la batteria di item in scala di Likert.
La domanda 12 è posta tramite una batteria di item in scala di Likert a 5 punti, in cui il rispondente è invitato a dichiarare la propria soddisfazione in relazione a 9 item, in una scala comprendente 5 diversi giudizi: "molto insoddisfatto", "insoddisfatto", "né soddisfatto né insoddisfatto", "soddisfatto" e "molto soddisfatto". Gli item sono: orari di apertura, facilità di raggiungimento del museo, cortesia e competenza del personale, orientamento nei percorsi, cura e pulizia degli ambienti, accessibilità per gli utenti con ridotta capacità motoria, materiali informativi, servizi di accoglienza, prezzo del biglietto. Segue la domanda 13, in cui al rispondente viene chiesto di indicare il proprio grado di soddisfazione complessiva relativamente alla visita, attribuendo un punteggio da 1 (molto insoddisfatto) a 10 (pienamente soddisfatto).
La domanda 14 è anch'essa posta nella forma di una batteria di item in scala di Likert a 5 punti, ma in questo caso le modalità sono "per nulla", "poco", "abbastanza", "molto", "moltissimo". In particolare, tale quesito ha l'obiettivo di indagare il grado di accordo o disaccordo del visitatore relativamente a 7 item, chiedendo al rispondente di esprimere, su una scala da "per nulla" a "moltissimo", quanto si ritenga in accordo con le seguenti 7 affermazioni:

- 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.

La domanda 15 è formulata tramite una batteria di item in scala a differenziale semantico, in cui l'intervistato può scegliere tra 7 modalità che vanno gradualmente da un aggettivo a quello di significato opposto, al fine di investigare la sua percezione relativa alla visita del museo. Il rispondente deve indicare la sensazione vissuta durante la visita, relativamente a 5 coppie di aggettivi opposti: da "noiosa" a "piacevole", da "banale" a "interessante", da "difficile" a "agevole", da "insignificante" a "coinvolgente", da "ordinaria" a "straordinaria".

La domanda 16 è posta nella forma di una batteria di item in scala di Likert a 5 punti con modalità "per nulla", "poco", "abbastanza", "molto", "moltissimo", tramite la quale si richiede al visitatore di esprimere, su una scala da "per nulla" a "moltissimo", quanto gli elementi di seguito elencati abbiano contribuito a rendere unica l'esperienza di visita del museo (*visitor experience*):

- il susseguirsi di avvenimenti storici che hanno influito sullo sviluppo del complesso monastico;
- la maestosità delle architetture monumentali dell'antica Brixia;
- la storia e le peculiarità del complesso monastico;
- le forme, le geometrie e i colori dell'arte figurativa presente nel museo;
- l'osservazione dei diversi sistemi di rappresentazione grafica che si sono susseguiti nei secoli;
- la possibilità di conoscere le abitudini e gli oggetti della quotidianità nei diversi periodi storici;
- la dimensione estetica associata alla visita, connessa al desiderio di conoscere opere e reperti importanti;
- la dimensione edonistica rispondente al mio desiderio di trascorrere un momento personale piacevole.

Analogamente, la domanda 17 è presentata tramite una batteria di item in scala di Likert a 5 punti, in cui si richiede all'utente di esprimere, su una scala da "per nulla" a "moltissimo", quanto una serie di elementi aggiuntivi abbiano contribuito a rendere unica l'esperienza museale. Gli item sono:

- la calma e la tranquillità del giardino esterno (Viridarium);
- la possibilità di utilizzare gli ArtGlass (occhiali multimediali);

- la presenza di un percorso tattile;
- la possibilità di sperimentare una forma nuova di conoscenza attraverso l'utilizzo delle tecnologie interattive presenti nel museo;
- l'offerta del bookshop;
- la presenza di contenuti digitali di elevata qualità (foto, video, ricostruzioni 3D, musiche, supporti audio ecc.).

Per questa domanda gli intervistati avevano la possibilità di selezionare una sesta opzione, ovvero "non applicabile (elemento non sperimentato)", poiché tra le alternative proposte vi sono dei servizi non strettamente legati alla visita del museo, dunque non sperimentati da tutti i rispondenti.

Le domande dalla 18 alla 26 sono volte a indagare eventuali nuove proposte o miglioramenti, al fine di potenziare l'offerta per migliorare ulteriormente la visita, o ancora l'effettiva conoscenza da parte dei rispondenti delle attuali proposte (ad esempio, le attività predisposte per bambini e ragazzi, gli strumenti a disposizione per le famiglie in visita autonoma al museo, il calendario di iniziative rivolte al pubblico organizzate da Fondazione Brescia Musei).
La domanda 27, non obbligatoria, è rivolta ai visitatori che intendano iscriversi alla newsletter di Fondazione Brescia Musei.

Infine, l'ultima sezione del questionario comprende una serie di domande (dalla 28 alla 38), relative ai dati anagrafici dei rispondenti: sesso, età, titolo di studio (eventuale titolo di laurea/post-laurea), residenza e professione. Questa tipologia di dati è fondamentale per la profilazione dell'utenza del Museo di Santa Giulia, sia a livello anagrafico sia sociodemografico, risultando necessaria poi per svolgere le opportune analisi in merito alla visitor experience.

Come accennato già precedentemente, uno specifico quesito sarà l'oggetto di indagine del presente elaborato: si tratta della domanda 14, una batteria di item in scala di Likert volta a indagare il grado di accordo del rispondente in relazione a una serie di affermazioni concernenti la qualità dell'offerta museale.

3.3 Obiettivo e metodologia della ricerca

Il presente capitolo è finalizzato a presentare l'applicazione, con le relative risultanze, delle tecniche di riduzione della dimensionalità PCA e NLPCA a un caso pratico.

I dati oggetto del presente lavoro sono stati preliminarmente raccolti nell'ambito del progetto di ricerca DS4BS, relativo all'utilizzo della Data Science per i luoghi della cultura, tramite la somministrazione di un questionario ai visitatori del Museo di Santa Giulia. In particolare, le due tecniche verranno applicate alla domanda 14 (relativa al grado di accordo del rispondente), approfondita nel paragrafo precedente. Si ricordi, infatti, che tale quesito si configura come una batteria di item in scala di Likert, da cui vengono originate variabili qualitative in scala ordinale. Di conseguenza, l'elaborazione di tali variabili richiede, di norma, l'applicazione della NLPCA.

Lo scopo di questa tesi è confrontare i risultati che si ottengono applicando la NLPCA (come è corretto fare) e la PCA lineare (il che è una forzatura, trascurando la natura non quantitativa delle variabili). Si tratta, dunque, di un lavoro metodologico con l'obiettivo di indagare empiricamente i limiti della PCA lineare, universalmente riconosciuti dai ricercatori; d'altra parte, lo studio permetterà di verificare l'effettivo miglioramento dell'analisi che è possibile conseguire tramite l'applicazione della PCA non lineare.

L'applicazione sia della PCA sia della NLPCA comporta, nel seguente ordine:

1) l'esame dello scree plot e degli autovalori per determinare la dimensione q in cui effettuare la riduzione della dimensionalità;
2) l'analisi del Factor Loadings Plot per effettuare l'interpretazione delle q componenti principali.

Dato l'obiettivo del presente elaborato, ovvero il confronto dei risultati che si ottengono applicando la PCA e la NLPCA, si procederà a presentare innanzitutto il punto 1, per entrambi i metodi, con un primo confronto tra le due tecniche. In seguito, si ripeterà lo stesso iter per il punto 2, realizzando un secondo confronto, anche alla luce di quanto emerso nello step precedente.

Infine, considerando tutte le risultanze ottenute, si potranno derivare alcune importanti conclusioni in merito all'impiego della PCA e NLPCA[18].

Dal punto di vista operativo, la PCA e la NLPCA verranno svolte mediante R, software per calcolo e grafica statistica (versione 4.3.1). In primo luogo, è necessario convertire il dataset in formato excel in un dataset in formato .txt (testo con valori delimitati da tabulazioni). Le variabili (gli item) sono 7, i soggetti rispondenti sono 665, quindi il dataset avrà 7 colonne × 665 righe.

3.3.1 Scelta della dimensione ottimale *q* nella NLPCA

Il primo passo per l'applicazione della NLPCA è scaricare su R il pacchetto "homals". In seguito, è possibile aprire lo script NLPCA.r e procedere a modificare i comandi. Tale script proviene dal materiale didattico del corso di "Statistica per il marketing" erogato nella LM in Management dell'Università degli Studi di Brescia.
I parametri sono stati settati così come di seguito:

```
dir <- "C:\\Users\\Administrator\\Desktop\\domanda tesi"     # settare
la directory di lavoro

miss <- "0"                    # etichetta assegnata ai missing values

K <- 5                         # numero categorie

scalev <- "ordinal"            # scaling level ("nominal", "ordinal",
"numerical")

colstart <- 2                  # colonna in cui si trova la prima
variabile su cui effettuare l'analisi
colend <- 8                    # colonna in cui si trova l'ultima
variabile su cui effettuare l'analisi

dim <- 0               # dimensione dello spazio in cui proiettare
                       # scrivere 0 se non Ë noto per svolgere solo
analisi full preliminare

rot <- "No"            # perform varimax rotation ? ("No" - "Yes")
```

[18] I dati oggetto di analisi derivano da una domanda nella forma di batteria di item in scala di Likert, quindi la NLPCA è la tecnica naturalmente adatta a tale contesto. In questo lavoro di tesi, si vuole effettuare un confronto applicando anche la PCA, nonostante non sia la tecnica adeguata, in quanto l'obiettivo è indagare quanto l'applicazione di tale tecnica generi risultati "vicini" o "lontani" da quelli che si ottengono applicando fin da subito la procedura corretta (NLPCA).

```
groups <- "No"                  # vuoi rappresentare grafici con una
variabile di raggruppamento ? ("No" - "Yes")
varg <- 14                      # colonna in cui si trova la variabile
di raggruppamento

datafile <- "dom accordo.txt"  # nome file che contiene i dati
```

Per prima cosa, si effettua l'analisi full a 7 dimensioni, ponendo `dim` pari a 0 nell'algoritmo. Con l'analisi full la dimensione rimane quella di partenza, per cui uguale a 7, e non viene operata la riduzione della dimensionalità. Il software restituisce autovalori, varianza spiegata e varianza spiegata cumulata (Tabella 3.1).

Tabella 3.1: Valutazione della qualità dell'analisi full – NLPCA
(Fonte: nostre elaborazioni)

Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
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%
Totale	7	100,00%	

La somma degli autovalori è pari alla dimensione dello spazio di partenza p, dato dal numero degli aspetti (item) su cui gli intervistati hanno espresso il loro grado di accordo. La varianza spiegata è stata calcolata dividendo ciascun autovalore per la somma degli stessi, mentre la varianza spiegata cumulata è stata ottenuta sommando progressivamente le varianze spiegate.

Alla prima nonlinear PC è associato un autovalore pari a 2,80 e una varianza spiegata pari al 40,01%. La seconda e la terza nonlinear PC hanno degli autovalori abbastanza vicini tra loro e di poco inferiori all'unità, rispettivamente 0,93 e 0,87. Alle restanti nonlinear PC sono invece associati degli autovalori nettamente inferiori a 1. In sintesi, in questo caso vi è un solo autovalore maggiore di 1, dunque il criterio di Guttman-Keiser suggerirebbe di mantenere solo la prima PC. In corrispondenza di tale dimensione, la varianza spiegata è pari al 40,01% del massimo teorico: si tratta di una percentuale buona ma non altissima, indicando una proiezione di medio-bassa qualità. Mantenendo anche la seconda PC, invece, si otterrebbe una migliore proiezione, con una

quantità di informazione trattenuta dalle prime due PC pari al 53,25%. La riduzione della dimensionalità in uno spazio bidimensionale, dunque, comporta un incremento significativo dell'informazione trattenuta.

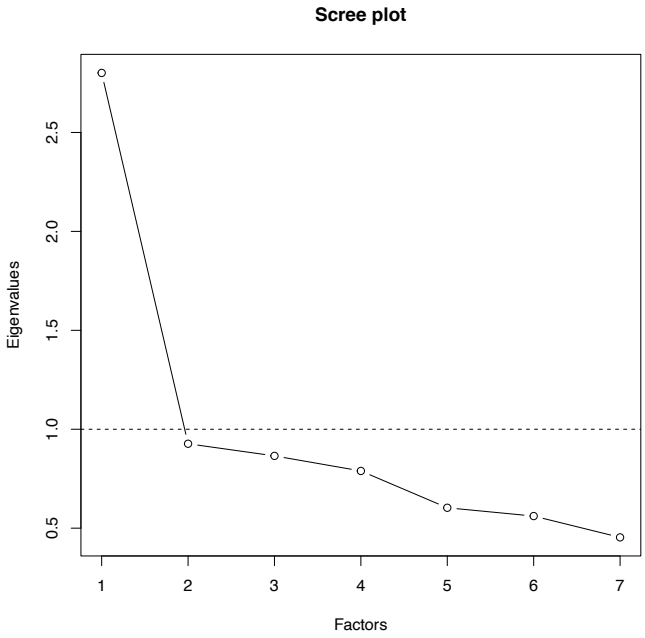


Figura 3.1: Scree Plot relativo all'analisi full – NLPCA
(Fonte: nostre elaborazioni)

Dall'analisi dello Scree Plot, si nota come il cambio di pendenza (il cosiddetto gomito) avvenga in corrispondenza della seconda PC (Figura 3.1). Poiché la dimensione ottimale è quella che si colloca immediatamente prima del gomito, lo Scree Plot indica che la dimensione ottimale del sottospazio in cui proiettare la nube di punti è 1.

Risultando necessarie ulteriori indicazioni per determinare la dimensione ottimale q, si procede a effettuare l'analisi a dimensione ridotta ponendo `dim` pari a 2.

Tabella 3.2: Valutazione della qualità dell'analisi a due dimensioni senza rotazione – NLPCA (Fonte: nostre elaborazioni)

Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
1	2,88	41,15%	41,15%
2	1,01	14,44%	55,59%

Effettuando la riduzione della dimensionalità con 2 dimensioni è possibile osservare i nuovi autovalori e le nuove varianze spiegate (Tabella 3.2). Si rileva un miglioramento dei risultati rispetto a quelli ottenuti con l'analisi full, come era prevedibile poiché le soluzioni della NLPCA sono *non nested* (cioè variano a seconda della dimensione dello spazio). In particolare, il secondo autovalore ora supera l'unità e si consegue una varianza spiegata cumulata pari al 55,59%.

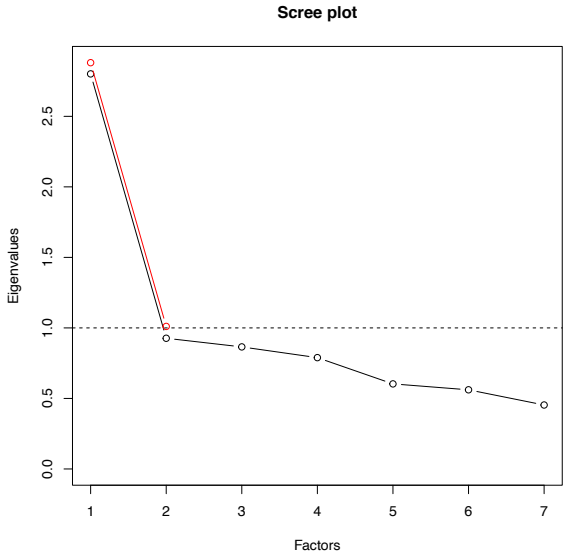


Figura 3.2: Scree Plot relativo all'analisi in due dimensioni – NLPCA (Fonte: nostre elaborazioni)

La Figura 3.2 rappresenta lo Scree Plot generato dall'analisi a due dimensioni, nel quale è possibile notare l'aumento degli autovalori corrispondenti alle prime due PC (spezzata rossa), rispetto agli autovalori dell'analisi full (spezzata nera).

A questo punto è necessario esporre alcune considerazioni, che chiariscano e supportino la scelta della dimensione ottimale q in cui operare la riduzione della dimensionalità. Tale decisione emerge dalla valutazione congiunta dei tre criteri precedentemente

approfonditi. Nel caso in esame, con riferimento all'analisi full, lo Scree Plot suggerirebbe una dimensione, il criterio della varianza spiegata suggerirebbe due dimensioni, e il criterio di Guttmann-Keiser oscilla tra una e due dimensioni. Effettuando l'analisi in due dimensioni (l'analisi ottimizzata), anche il secondo autovalore risulta maggiore di 1. Dunque, dopo una ponderata valutazione, si conclude che sia una scelta migliore tenere la seconda dimensione, operando la riduzione della dimensionalità in due dimensioni. Si è giunti a tale conclusione considerando anche che la varianza spiegata da una sola dimensione non è molto alta: dato che due dimensioni è comunque un'ottima riduzione della dimensionalità, conviene garantirsi la quota di informazione aggiuntiva garantita dalla seconda dimensione. Dunque, l'aggiunta della seconda componente principale è stata decisa dopo aver valutato il trade-off esistente tra il maggior livello di complessità dell'analisi, causato dall'aggiunta della stessa, e il miglioramento della varianza spiegata (cumulata).
Concludendo, si preferisce optare per due dimensioni, anche per rendere maggiormente interessante il confronto che si vuole fare con la PCA lineare, che si ricorda essere il fine ultimo del presente elaborato.

3.3.2 Scelta della dimensione ottimale *q* nella PCA

Per applicare la PCA, in primo luogo, si procede ad aprire lo script PCA.r sul software R. Tale script proviene dal materiale didattico del corso di "Statistica per il marketing" erogato nella LM in Management dell'Università degli Studi di Brescia.
Per implementare l'algoritmo, i parametri sono stati settati così come di seguito:

```
dir <- "C:\\Users\\Administrator\\Desktop\\pca tesi"       #   settare
la directory di lavoro

miss <- "0"                    # etichetta assegnata ai missing values

colstart <- 2                  #  colonna  in  cui  si  trova  la  prima
variabile su cui effettuare l'analisi
colend <- 8              # colonna in cui si trova l'ultima variabile su
cui effettuare l'analisi
colnames <- 1            # colonna in cui si trovano i nomi dei soggetti
- se non disponibile scrivere 0
dim <- 0                 # dimensione dello spazio in cui proiettare
```

```
                        # scrivere 0 se non È noto per svolgere solo
analisi preliminare

rot <- "No"             # perform varimax rotation ? ("No" - "Yes")

groups <- "No"                # vuoi rappresentare il grafico con una
variabile di raggruppamento ? ("No" - "Yes")
varg <- 8                     # colonna in cui si trova la variabile di
raggruppamento

datafile <- "dom accordo.txt"       # nome file che contiene i dati
```

Inizialmente, poiché non è nota la dimensione del sottospazio in cui proiettare la nube di punti, si pone `dim` pari a 0 per lo svolgimento dell'analisi preliminare. Il software restituisce autovalori, varianza spiegata e varianza spiegata cumulata (Tabella 3.3).

Tabella 3.3: Valutazione della qualità dell'analisi preliminare – PCA
(Fonte: nostre elaborazioni)

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%
Totale	7	100,00%	

Alla prima PC lineare è associato un autovalore pari a 2,75 e una varianza spiegata pari al 39,32%. La seconda e la terza PC lineari hanno degli autovalori leggermente inferiori all'unità, rispettivamente 0,97 e 0,85: tra i due, risulta particolarmente rilevante 0,97, praticamente pari all'unità. Alle restanti PC lineari sono invece associati degli autovalori nettamente inferiori a 1. Vi è un solo autovalore pienamente maggiore di 1, dunque secondo il criterio di Guttman-Keiser (se applicato in modo rigoroso) si dovrebbe mantenere solo la prima PC. In corrispondenza di tale dimensione, la varianza spiegata è pari al 39,32%: si tratta di una percentuale ben al di sotto del massimo teorico, indicando una proiezione di medio-bassa qualità. Considerando anche la seconda PC,

invece, si otterrebbe una proiezione di qualità ben superiore, con una percentuale di informazione trattenuta pari al 53,23%.

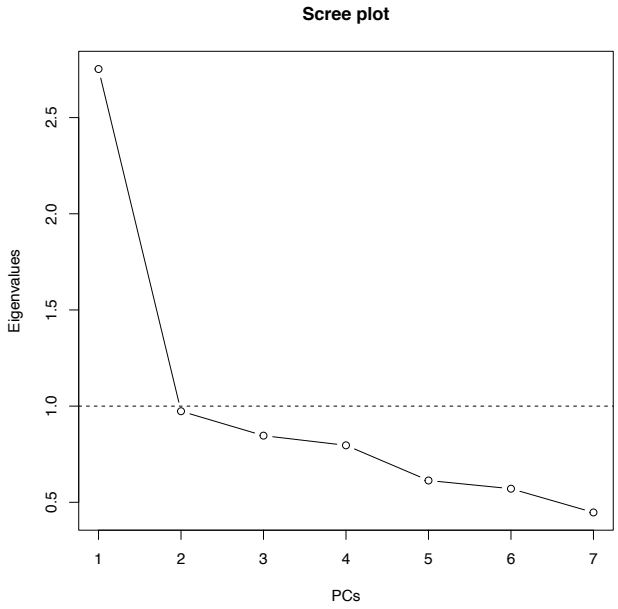


Figura 3.3: Scree Plot relativo all'analisi preliminare – PCA
(Fonte: nostre elaborazioni)

Dall'analisi dello Scree Plot si individua il gomito, ovvero la caduta brusca della spezzata, in prossimità della seconda PC (Figura 3.3). La dimensionale ottimale per la proiezione è quella precedente al gomito, pertanto lo Scree Plot indicherebbe come scelta ottimale $q = 1$.

La definizione della dimensione ottimale q deriva dalla valutazione congiunta dei tre criteri. Parimenti a quanto visto con la NLPCA, lo Scree Plot suggerirebbe una dimensione, il criterio della varianza spiegata suggerirebbe invece due dimensioni, e il criterio di Guttmann-Keiser oscilla tra una e due dimensioni.
Complessivamente, risulta conveniente ridurre la dimensionalità a due dimensioni, mantenendo le prime due PC. Si noti, infatti, che il secondo autovalore è quasi identico all'unità (0,97); inoltre, la varianza spiegata da una sola dimensione non è sufficientemente elevata per garantire un'analisi di buona qualità. Quindi, dato che due dimensioni è un'ottima riduzione della dimensionalità, è opportuno trattenere la percentuale di informazione aggiuntiva garantita dalla seconda dimensione.

A garanzia della bontà di tale decisione, nel caso in cui fosse stata svolta esclusivamente la PCA (quindi senza dover realizzare un confronto con la NLPCA sulla base di q) si avrebbe comunque optato per una riduzione a due dimensioni.

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

A questo punto, dopo aver applicato il punto 1 per entrambe le tecniche, si procede a presentare un primo confronto tra le risultanze emerse.

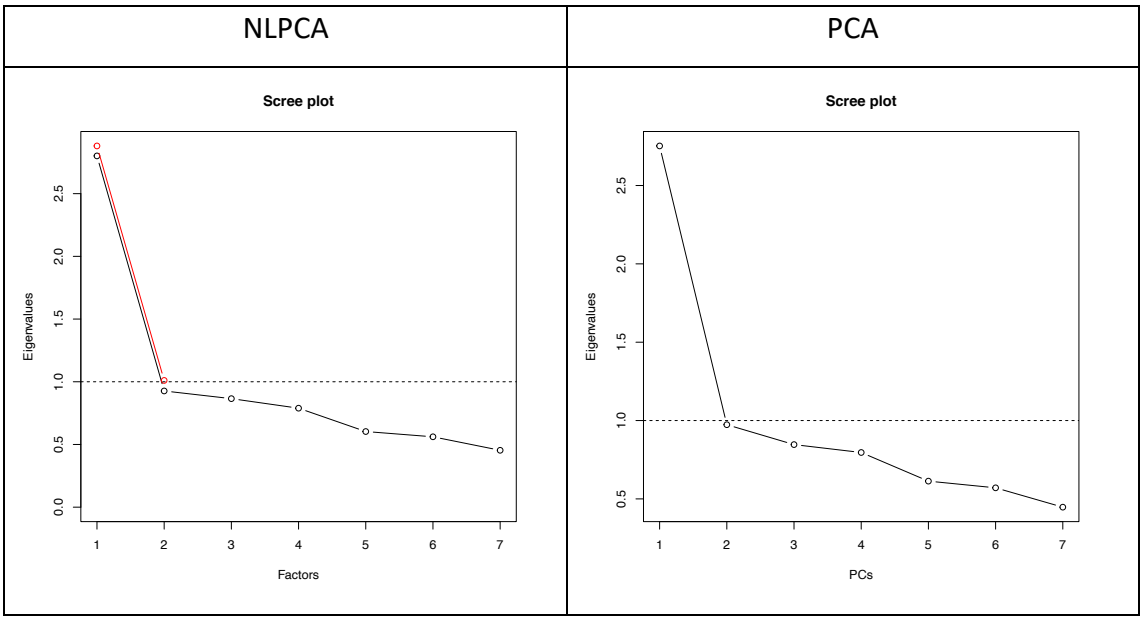


Figura 3.4: Confronto tra gli Scree Plot di NLPCA e PCA
(Fonte: nostre elaborazioni)

Confrontando i due Scree Plot (Figura 3.4), appare evidente quanto i due grafici siano simili. In entrambi i casi si registra un cambio di pendenza significativo in prossimità della seconda PC, in virtù del quale l'informazione portata dalle prime PC è elevata. In particolare, emerge chiaramente come il maggior contributo all'informazione complessiva sia dato essenzialmente dalla prima PC.

La Tabella 3.4 mostra il confronto tra NLPCA e PCA, con riferimento ad autovalori, varianza spiegata, varianza spiegata cumulata e conseguente scelta del numero q di dimensioni in cui effettuare la riduzione della dimensionalità.

Tabella 3.4: Confronto tra NLPCA e PCA
(Fonte: nostre elaborazioni)

	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. Pertanto, la scelta della dimensione $q = 2$ è un primo risultato in comune.

Dalla comparazione, emerge che la NLPCA consente una varianza spiegata leggermente superiore rispetto alla PCA, come è nelle attese. Difatti, la NLPCA è fatta in modo da massimizzare la varianza spiegata, ed effettivamente questo si riscontra nei risultati ottenuti. In particolare, nella NLPCA l'informazione trattenuta è pari al 55,59%, mentre nella PCA risulta pari al 53,23%.

3.3.4 Interpretazione delle q componenti principali nella NLPCA

Il passo successivo è attribuire un significato appropriato alle q componenti principali. A tal fine, si ricorre all'esame del Factor Loadings Plot (Figura 3.5).

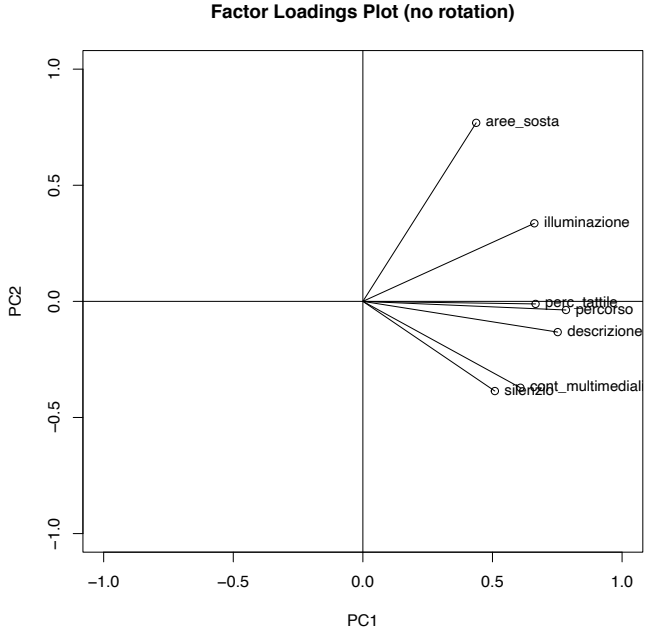


Figura 3.5: Factor Loadings Plot senza rotazione – NLPCA
(Fonte: nostre elaborazioni)

Poiché risulta difficile attribuire le variabili alle componenti principali ottenute, è necessario effettuare la rotazione Varimax, in modo da facilitare la comprensione della rappresentazione grafica.

Tabella 3.5: Valutazione della qualità dell'analisi a due dimensioni con rotazione – NLPCA
(Fonte: nostre elaborazioni)

Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
1	2,31	32,99%	32,99%
2	1,58	22,59%	55,59%

A seguito della rotazione Varimax, si può notare come l'informazione complessivamente trattenuta si ridistribuisca tra le due PC (Tabella 3.5). Naturalmente, la quantità di informazione complessiva rimane invariata, mentre la quota di informazione portata da ciascuna PC si è modificata. In particolare, si rileva che la varianza spiegata della prima PC si è ridotta (32,99%), mentre è aumentata quella associata alla seconda PC (22,59%).

Si riportano, di seguito, il Factor Loadings Plot ruotato (Figura 3.6) e la tabella dei loadings (Tabella 3.6), i quali consentiranno di comprendere il contributo di ciascuna variabile alla determinazione del significato delle due componenti principali e, quindi, di attribuire un titolo a queste ultime.

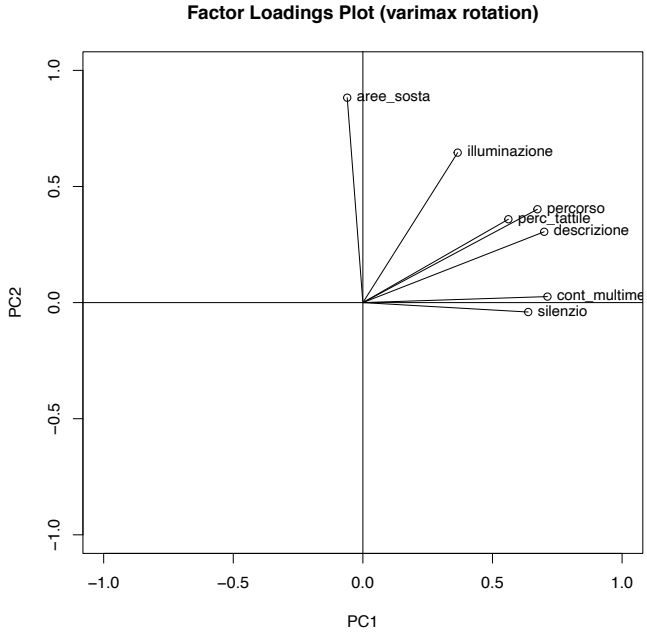


Figura 3.6: Factor Loadings Plot con rotazione – NLPCA
(Fonte: nostre elaborazioni)

Generalmente, in un Factor Loadings Plot ideale si desidera che gli item siano completamente dedicati a una o all'altra delle PC, ma può anche accadere che ci siano due PC e vari item che contribuiscono a entrambe. L'intensità della correlazione degli item "di mezzo" con le due PC può essere coerente con il significato attribuito ad esse.
Nel caso in esame, dal Factor Loadings Plot ruotato emerge una prima PC collegata all'*apprezzamento per gli aspetti di contenuto forniti nei percorsi* e una seconda PC collegata all'*apprezzamento per il modo in cui sono progettati (progettazione)*.
Gli item sono così distribuiti:

- contenuti multimediali sta totalmente con la prima PC;
- il gruppetto da tre item poco sopra (descrizione, percorso tattile, percorso) è relativo ad aspetti di contenuto ma anche un po' progettuali, e infatti è correlato molto con la prima PC ma anche un po' con la seconda PC;

- l'illuminazione è in gran parte progettuale, ma incide anche sul contenuto perché la luce può cambiare molto quello che vediamo dell'opera;
- le aree di sosta sono esclusivamente legate alla parte progettuale;
- l'item silenzio è un caso molto interessante: non è né un aspetto di contenuto né un aspetto progettuale, tuttavia qui va a posizionarsi esclusivamente con il contenuto. Questo sembra indicare che i contenuti stessi non vengono percepiti se si è in mezzo al caos e quindi un ambiente tranquillo e silenzioso rende più belle anche le opere stesse.

I loadings, coefficienti di correlazione lineare tra la variabile iniziale e le componenti principali, confermano quanto appena descritto. Si ricordi, infatti, che un valore alto (in valore assoluto) significa una forte correlazione della variabile iniziale con la componente principale: questo accade solo con tre item, ovvero "silenzio", "aree sosta" e "contenuti multimediali". I loadings dei rimanenti item avvalorano la loro posizione "di mezzo" rispetto alle due componenti principali, con ciò denotando che tali item contribuiscono al significato di entrambe le componenti principali.

Tabella 3.6: Loadings delle variabili – NLPCA
(Fonte: nostre elaborazioni)

Variabili	PC 1	PC 2
Illuminazione	0,3657	0,6458
Percorso	0,6736	0,4019
Silenzio	0,6378	-0,0406
Aree sosta	-0,0604	0,8824
Descrizione	0,6996	0,3050
Percorso tattile	0,5614	0,3590
Contenuti multimediali	0,7121	0,0259

Avendo attribuito un nome alle due componenti principali PC 1 e PC 2, si riporta il Factor Loadings Plot nel quale sono state trascritte tali denominazioni (Figura 3.7).

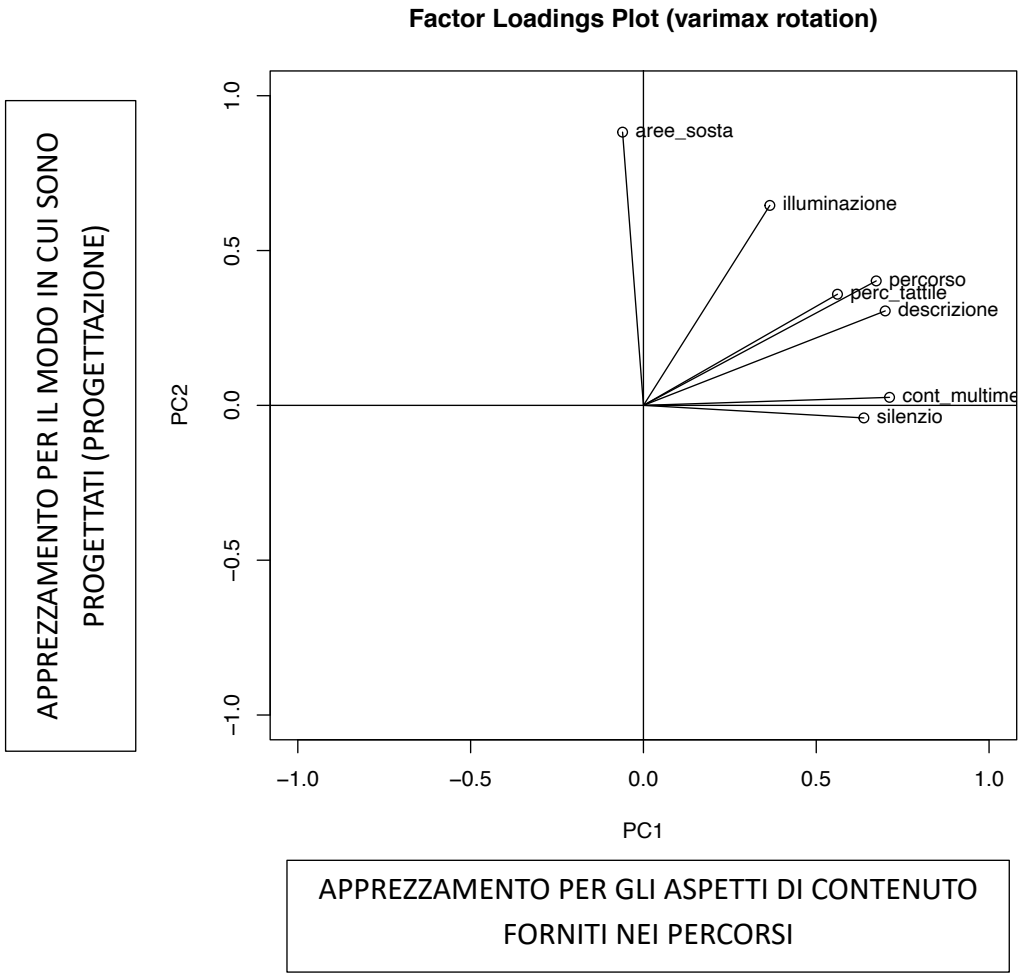


Figura 3.7: Factor Loadings Plot con le PC rinominate – NLPCA
(Fonte: nostre elaborazioni)

Il software restituisce anche i Transformation Plot, ossia i grafici che mostrano le quantificazioni delle categorie per ogni item. Nel commento di tali grafici si tenga conto che:

- $q1$ è la quantificazione della modalità 1 “per nulla”;
- $q2$ è la quantificazione della modalità 2 “poco”;
- $q3$ è la quantificazione della modalità 3 “abbastanza”;
- $q4$ è la quantificazione della modalità 4 “molto”;
- $q5$ è la quantificazione della modalità 5 “moltissimo”.

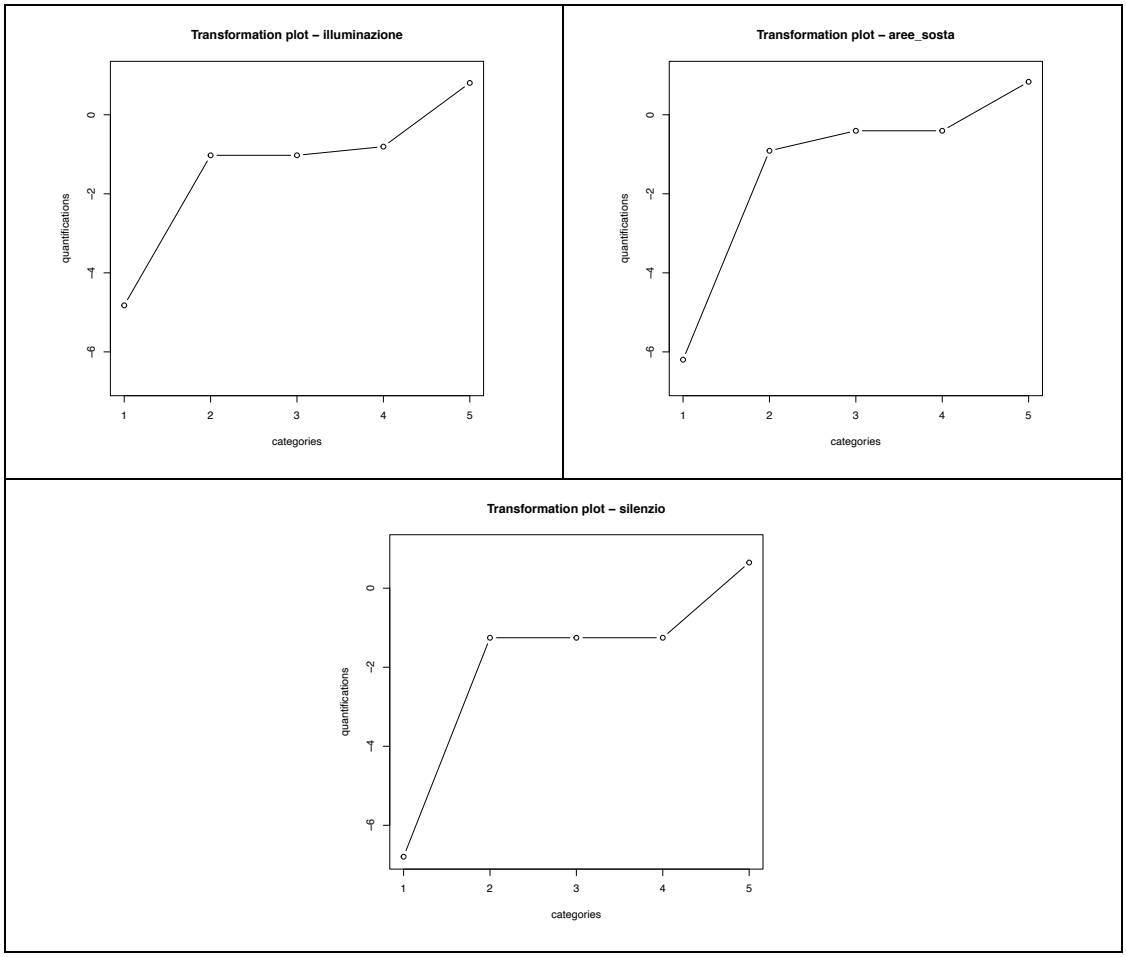


Figura 3.8: Transformation Plot degli item "illuminazione", "aree sosta" e "silenzio" – NLPCA (Fonte: nostre elaborazioni)

Per gli item "illuminazione", "aree sosta" e "silenzio" la trasformazione è non lineare e il grafico è dato da una spezzata non decrescente che approssima una funzione a "S" (Figura 3.8). Nei tre grafici si rileva una significativa crescita della $q2$ rispetto alla $q1$ e un ulteriore incremento, sebbene minore, della $q5$ rispetto alla $q4$. Al contrario, nel tratto tra la $q2$ e la $q4$ i grafici presentano un differente andamento. In particolare, per l'item "illuminazione" si individua una sostanziale crescita con un lieve aumento nella $q4$. Nel grafico di "aree sosta" vi è un leggero incremento della $q3$ rispetto alla $q2$, e poi una costanza dalla $q3$ alla $q4$. Nel grafico di "silenzio" si rileva una costanza dalla $q2$ alla $q4$. Ciò implica una maggiore distinzione tra le modalità di basso grado di accordo e di elevato grado di accordo, mentre vi è minore distinzione tra le categorie intermedie.

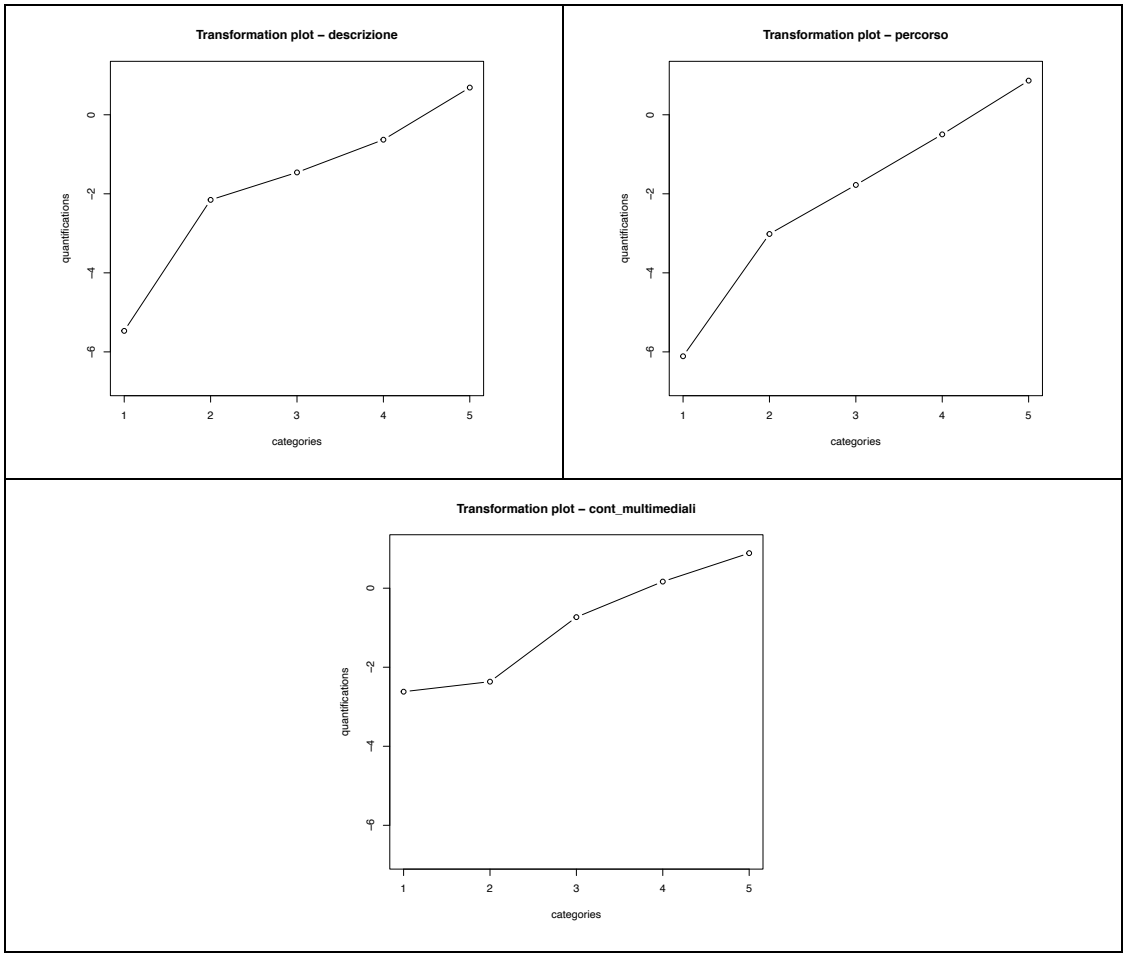


Figura 3.9: Transformation Plot degli item "descrizione", "percorso" e "contenuti multimediali" – NLPCA
(Fonte: nostre elaborazioni)

Per gli item "descrizione", "percorso" e "contenuti multimediali" la trasformazione è approssimativamente lineare, con qualche lieve differenza tra i tre grafici (Figura 3.9). Nel Transformation Plot di "descrizione" e "percorso" si rileva un considerevole incremento della $q2$ rispetto alla $q1$, seguito poi sostanzialmente da una crescita lineare fino alla $q5$. Nel Transformation Plot di "contenuti multimediali", invece, si individua un lievissimo incremento della $q2$ rispetto alla $q1$, poi vi è una crescita più netta nella $q3$, e infine un incremento fondamentalmente lineare fino alla $q5$. Complessivamente, nei tre grafici le quantificazioni sono perlopiù tendenzialmente lineari, con ciò indicando che le categorie originarie sono approssimativamente equispaziate.

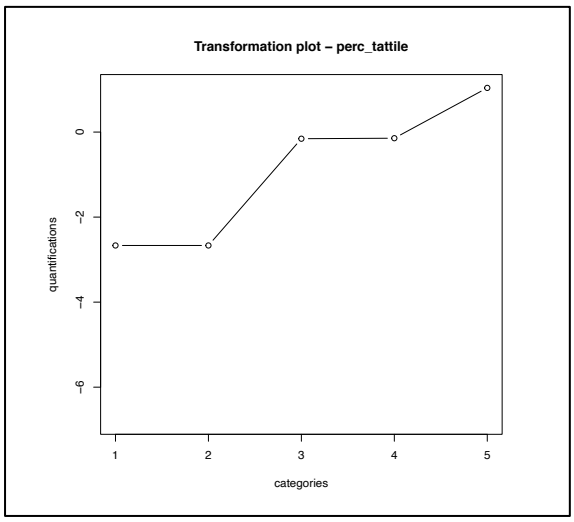


Figura 3.10: Transformation Plot dell'item "percorso tattile" – NLPCA
(Fonte: nostre elaborazioni)

Infine, l'item "percorso tattile" dispone di un grafico di trasformazione con andamento decisamente non lineare, che individua una spezzata non decrescente (Figura 3.10). Nel dettaglio, non si rilevano differenze tra la $q1$ e la $q2$: in tale tratto, la trasformazione è piatta. Analogamente, si rileva una sostanzialmente costanza tra la $q3$ e la $q4$. Al contrario, passando dalla $q2$ alla $q3$ emerge un elevato incremento nelle quantificazioni, così come accade tra la $q4$ e la $q5$ (in quest'ultimo caso la crescita è però inferiore).

Il software fornisce in output anche l'Object Scores Plot (Figura 3.11), il quale rappresenta come si posizionano le 665 unità statistiche rispetto alle due componenti principali.

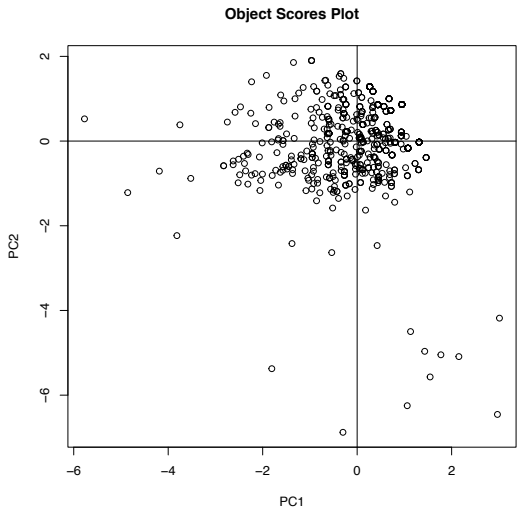


Figura 3.11: Object Scores Plot
(Fonte: nostre elaborazioni)

Le due linee in corrispondenza degli 0 indicano la media delle PC (queste, essendo standardizzate, hanno media esattamente pari a 0). Dunque, il grafico va interpretato in questo modo: il quadrante in alto a destra presenta i soggetti con un grado di accordo sopra la media per entrambe le PC; il quadrante in alto a sinistra contiene i soggetti con un apprezzamento per gli aspetti di contenuto forniti nei percorsi sotto la media, e un apprezzamento per il modo in cui sono progettati sopra la media; il quadrante in basso a sinistra indica le unità statistiche che hanno un grado di accordo sotto la media relativamente a entrambe le PC; il quadrante in basso a destra individua i soggetti che hanno un apprezzamento per gli aspetti di contenuto forniti nei percorsi sopra la media, e un apprezzamento per il modo in cui sono progettati sotto la media.

3.3.5 Interpretazione delle q componenti principali nella PCA

Analogamente a quanto svolto nel paragrafo precedente, si prosegue l'applicazione della PCA con l'interpretazione delle q componenti principali. Dunque, di seguito viene presentato il Factor Loadings Plot (Figura 3.12).

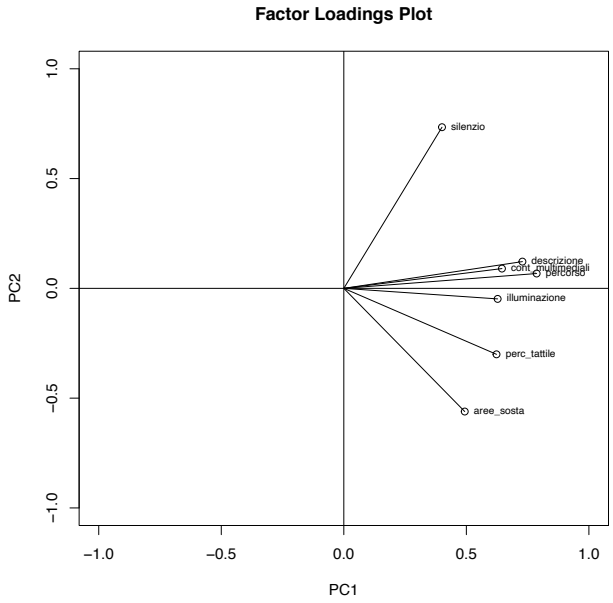


Figura 3.12: Factor Loadings Plot senza rotazione – PCA
(Fonte: nostre elaborazioni)

Al fine di una migliore comprensione del grafico, risulta necessario applicare la rotazione Varimax.

Tabella 3.7: Valutazione della qualità dell'analisi a due dimensioni con rotazione – PCA
(Fonte: nostre elaborazioni)

Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
1	2,14	30,51%	30,51%
2	1,59	22,72%	53,23%

In virtù della rotazione effettuata, l'informazione complessivamente trattenuta viene ridistribuita tra le due componenti principali, mantenendo invariata la quantità di informazione complessiva (Tabella 3.7).
Si riporta, di seguito, il Factor Loadings Plot ruotato (Figura 3.13).

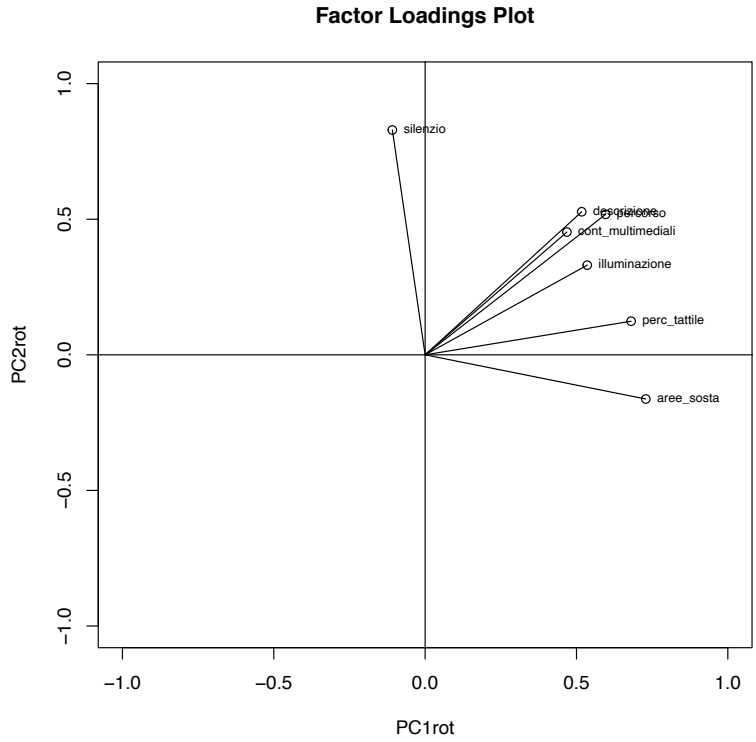


Figura 3.13: Factor Loadings Plot con rotazione – PCA
(Fonte: nostre elaborazioni)

Osservando il Factor Loadings Plot, la PC 1 risulta essere la sintesi di ben 7 elementi ("aree sosta", "percorso tattile", "illuminazione", "contenuti multimediali", "descrizione", "percorso"). La PC 2, invece, è costituita solo dalla variabile "silenzio", dunque il suo significato è esclusivamente formato da tale item.

Giunti a questo punto, non è necessario procedere a una puntuale interpretazione delle due PC, con conseguente assegnazione di un titolo. Questo poiché, considerato il fine ultimo della presente ricerca, è sufficiente limitarsi al confronto tra le risultanze emerse, sottolineando (nella fattispecie) la differenza del significato delle PC delle due tecniche. A tal fine, il successivo paragrafo sarà dedicato all'approfondimento del suddetto confronto.

Per completezza, si presenta la tabella dei loadings (Tabella 3.8).

Tabella 3.8: Loadings delle variabili – PCA
(Fonte: nostre elaborazioni)

Variabili	PC 1	PC 2
Illuminazione	0,5356	0,3307
Percorso	0,5962	0,5178
Silenzio	-0,1083	0,8291
Aree sosta	0,7291	-0,1630
Descrizione	0,5173	0,5275
Percorso tattile	0,6805	0,1239
Contenuti multimediali	0,4685	0,4527

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

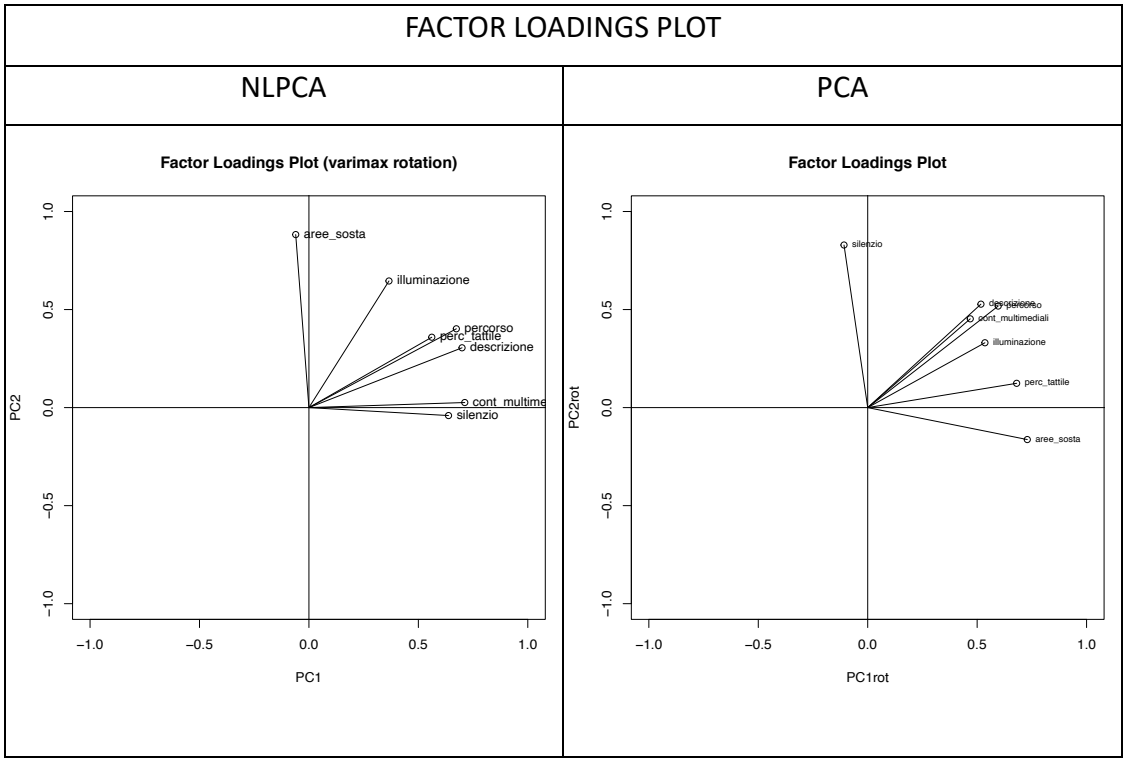


Figura 3.14: Confronto tra i Factor Loadings Plot
(Fonte: nostre elaborazioni)

Dal confronto, emerge chiaramente che applicando la NLPCA e la PCA si ottengono 2 PC differenti, dal significato diverso, pertanto non sono confrontabili. I Factor Loadings Plot (Figura 3.14) sono completamente diversi tra loro e le PC che si otterrebbero da essi sono difformi. Mantenendo come riferimento il Factor Loadings Plot della NLPCA (che è la tecnica corretta), nel grafico della PCA le variabili "aree sosta" e "silenzio" appaiono

perfettamente ribaltate, mentre le restanti risultano diversamente distribuite, in modo vario, senza uno schema preciso. Quindi, non è nemmeno possibile affermare che le PC si siano semplicemente scambiate di posizione, in quanto hanno dei significati estremamente differenti.
Si tratta, sicuramente, di un risultato molto interessante, anche abbastanza inaspettato in termini della decisamente netta differenza dei risultati ottenuti.
Dalla letteratura è risaputo che, in questo caso, la PCA è la tecnica errata, mentre l'applicazione della NLPCA consente di trattare in modo appropriato le variabili in esame. Pertanto, è importante sottolineare che se non si tiene conto del fatto che le variabili sono ordinali si ottengono risultati totalmente diversi, nemmeno vicini o simili. Addirittura, le PC individuate nella NLPCA vengono mascherate completamente nella PCA.
Il caso di studio presentato ha rilevato un effetto molto importante: talvolta si ritiene che classificare le variabili come qualitative ordinali e non quantitative, e di conseguenza applicare la NLPCA piuttosto che la PCA, sia un eccesso di precisione e dettaglio, ma l'analisi empirica effettuata ha chiaramente dimostrato il contrario.
In conclusione, la scelta di ridurre la dimensionalità in due dimensioni è certamente corretta, tuttavia l'applicazione dell'una o dell'altra tecnica origina risultati diversi, in particolare le componenti principali assumono un significato diverso a seconda della tecnica impiegata.

3.3.7 La NLPCA in una dimensione

Nel paragrafo precedente si è potuto verificare che, nel caso in esame, la riduzione della dimensionalità in due dimensioni operata dalla NLPCA e PCA al medesimo dataset genera due componenti principali la cui interpretazione è totalmente differente. Quindi, nell'ambito della presente ricerca, il passo successivo è obbligare le due analisi ad avere una componente principale con lo stesso risultato: ciò si può ottenere operando la riduzione in una sola dimensione. In questo modo, tutte le variabili saranno sicuramente correlate con quell'unica componente principale, che può essere denominata il *grado complessivo di accordo (del rispondente)*.

Di seguito si riporta l'applicazione della NLPCA ponendo `dim` pari a 1.

Tabella 3.9: Valutazione della qualità dell'analisi a una dimensione – NLPCA
(Fonte: nostre elaborazioni)

Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
1	2,95	42,20%	42,20%

Con l'analisi a una dimensione l'autovalore corrispondente alla prima PC assume valore pari a 2,95, con una varianza spiegata del 42,20% (Tabella 3.9). Quindi, sintetizzando tutte le variabili in un'unica dimensione si ha complessivamente un'informazione trattenuta pari al 42,20%, che è una percentuale non altissima. Naturalmente, si rileva un miglioramento della varianza spiegata dalla prima PC, anche se di lieve entità, rispetto alla corrispettiva ottenuta nell'analisi a due dimensioni (in tal caso la varianza spiegata dalla prima PC era pari al 41,15%).

Nella Tabella 3.10 sono riportati i loadings delle variabili, che risultano avere tutte una correlazione medio-alta con la componente principale estratta.

Tabella 3.10: Loadings delle variabili – NLPCA
(Fonte: nostre elaborazioni)

Variabili	PC 1
Illuminazione	0,6592
Percorso	0,7876
Silenzio	0,4967
Aree sosta	0,5069
Descrizione	0,7432
Percorso tattile	0,6753
Contenuti multimediali	0,6224

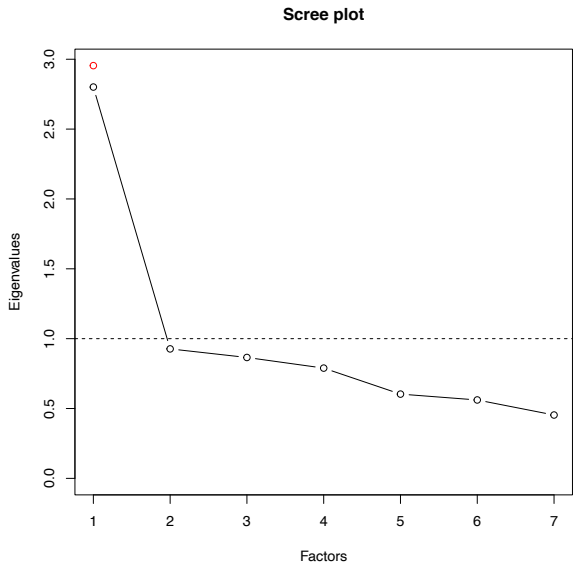


Figura 3.15: Scree Plot relativo all'analisi in una dimensione – NLPCA
(Fonte: nostre elaborazioni)

La Figura 3.15 rappresenta lo Scree Plot originato dall'analisi a una dimensione, nel quale è possibile notare l'incremento dell'autovalore corrispondente alla prima PC (pallino rosso), rispetto agli autovalori dell'analisi full (spezzata nera).
Di seguito vengono riportati i nuovi Transformation Plot per ogni variabile.

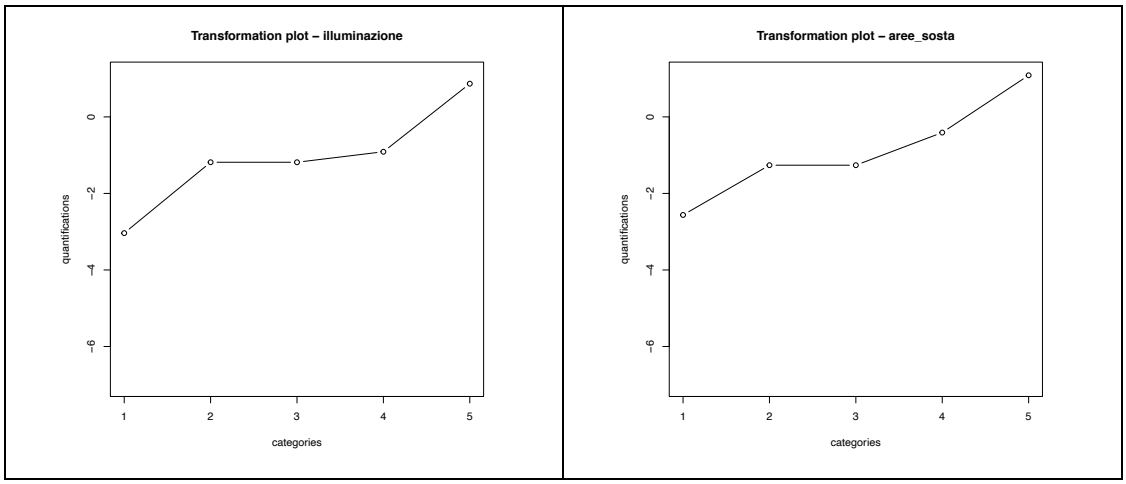


Figura 3.16: Transformation Plot degli item "illuminazione" e "aree sosta"
(Fonte: nostre elaborazioni)

In entrambi i grafici di trasformazione rappresentati in Figura 3.16 si rileva una crescita passando dalla quantificazione della modalità 1 alla quantificazione della modalità 2, poi

vi è una sostanziale costanza passando alla $q3$, e da qui in poi ancora un graduale incremento nella $q4$ e soprattutto nella $q5$.

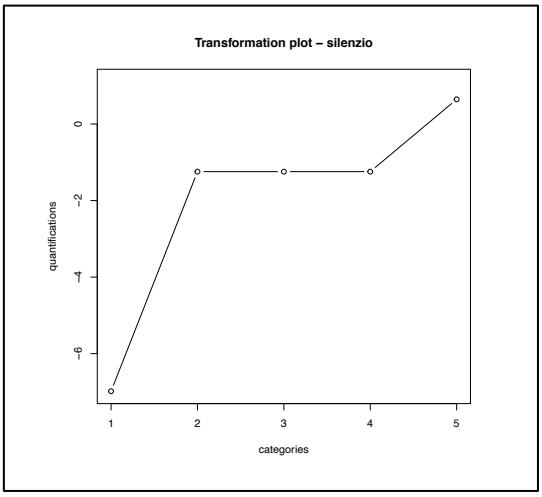


Figura 3.17: Transformation Plot dell'item "silenzio"
(Fonte: nostre elaborazioni)

Nel grafico della Figura 3.17 si registra una notevole crescita tra le quantificazioni delle modalità 1 e 2, poi vi è una sostanziale costanza passando per $q3$ e $q4$, e infine vi è un importante incremento nella $q5$.

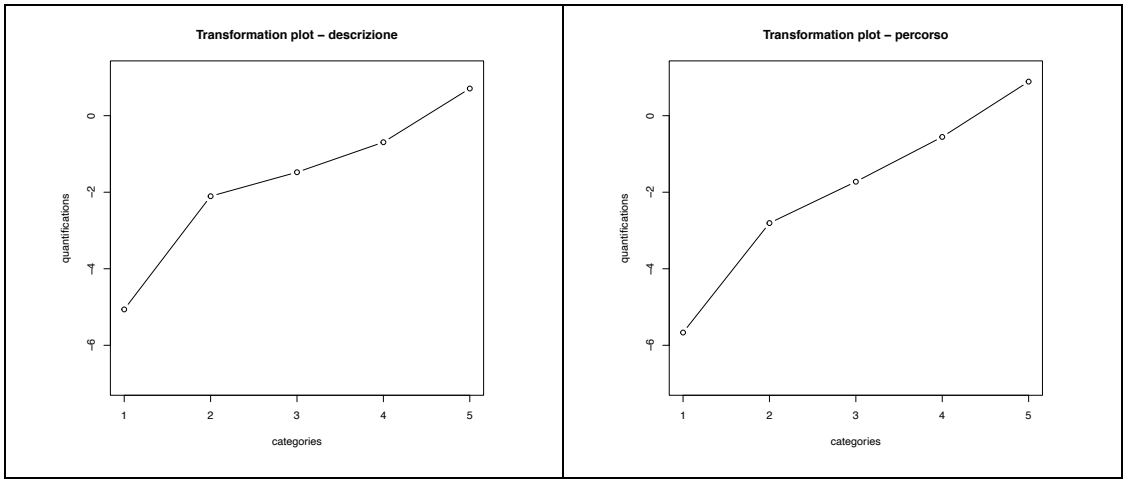


Figura 3.18: Transformation Plot degli item "descrizione" e "percorso"
(Fonte: nostre elaborazioni)

Nei grafici della Figura 3.18, si rileva una trasformazione che approssima una funzione lineare. L'unica eccezione in tale andamento quasi lineare è data dal tratto tra le quantificazioni delle modalità 1 e 2, in cui si nota un incremento maggiore rispetto alla crescita che si ha in $q3$, $q4$ e $q5$.

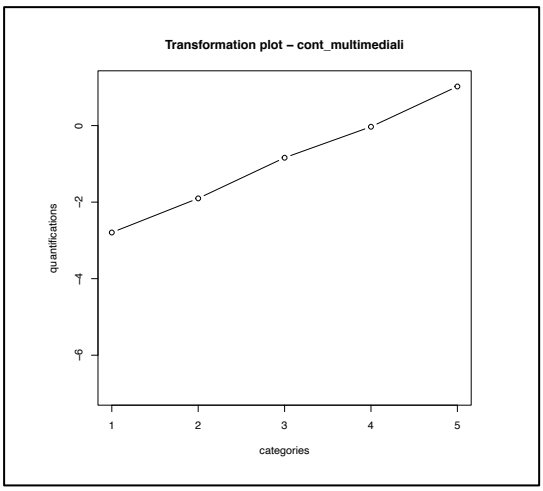


Figura 3.19: Transformation Plot dell'item "contenuti multimediali"
(Fonte: nostre elaborazioni)

Nel grafico dell'item "contenuti multimediali" si ha, sostanzialmente, una trasformazione lineare (Figura 3.19). L'incremento che si ha passando da una quantificazione alla successiva è costante, con ciò significando che le categorie originarie sono praticamente equispaziate.

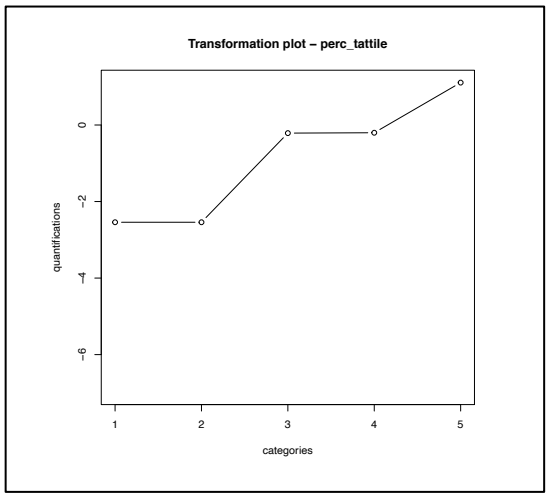


Figura 3.20: Transformation Plot dell'item "percorso tattile"
(Fonte: nostre elaborazioni)

In Figura 3.20 si rileva una netta costanza tra le quantificazioni delle modalità 1 e 2; poi segue una crescita passando alla $q3$, da qui in avanti ancora una sostanziale costanza nella $q4$, e infine un discreto incremento nella $q5$.

3.3.8 La PCA in una dimensione

Analogamente a quanto svolto precedentemente, si presentano i risultati derivanti dall'applicazione della PCA ponendo `dim` pari a 1. Il software R restituisce i risultati già visti con l'analisi in due dimensioni (Tabella 3.11).

Tabella 3.11: Valutazione della qualità dell'analisi a una dimensione – PCA
(Fonte: nostre elaborazioni)

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%
Totale	7	100,00%	

Mantenendo solo una dimensione si consegue una varianza spiegata, quindi un'informazione complessivamente trattenuta, pari al 39,32%.

Nella Tabella 3.12 sono riportati i loadings delle variabili che, anche in questo caso, esibiscono tutte correlazione medio-alta con la componente principale estratta.

Tabella 3.12: Loadings delle variabili – PCA
(Fonte: nostre elaborazioni)

Variabili	PC 1
Illuminazione	0,6277
Percorso	0,7868
Silenzio	0,4006
Aree sosta	0,4934
Descrizione	0,7288
Percorso tattile	0,6230
Contenuti multimediali	0,6452

3.3.9 Confronto tra NLPCA e PCA in una dimensione

Ricapitolando quanto effettuato nei paragrafi 3.3.7 e 3.3.8, sono state applicate le due tecniche di riduzione della dimensionalità NLPCA e PCA in una dimensione, forzando così le due analisi ad avere un'unica componente principale dallo stesso significato. Dunque, tale componente principale risulta essere la sintesi delle 7 variabili oggetto di analisi, che per questo motivo è nominata il *grado complessivo di accordo*.
Con la NLPCA si ottiene una varianza spiegata pari al 42,20%, mentre con la PCA risulta pari al 39,32%. Come prevedibile, la NLPCA consente di trattenere una maggiore informazione nell'analisi a dimensione ridotta. A ogni modo, in entrambi i casi si tratta di una percentuale non elevata, dovuta alla ridotta complessità dell'analisi a una sola dimensione.

A questo punto, avendo una componente principale con significato identico, è possibile eseguire un confronto tra la PC 1 della NLPCA e la PC 1 della PCA. Ciò permette di verificare quanto almeno su una componente principale di accordo complessivo le due PC 1 sono concordi. A tal fine, si è realizzato uno scatterplot (grafico a dispersione), dove sull'asse X sono rappresentati gli object scores corrispondenti alla PC 1 della PCA, mentre sull'asse Y sono riportati gli object scores corrispondenti alla PC 1 della NLPCA[19]. Al grafico è stata aggiunta una linea di tendenza lineare (a 45°).

[19] A ogni rispondente corrisponde un object score, fornito in output dal software R in seguito alla riduzione della dimensionalità in una dimensione (PC 1). Il questionario è stato somministrato a 665 utenti, quindi, nel grafico sono presenti 665 coppie di punti.

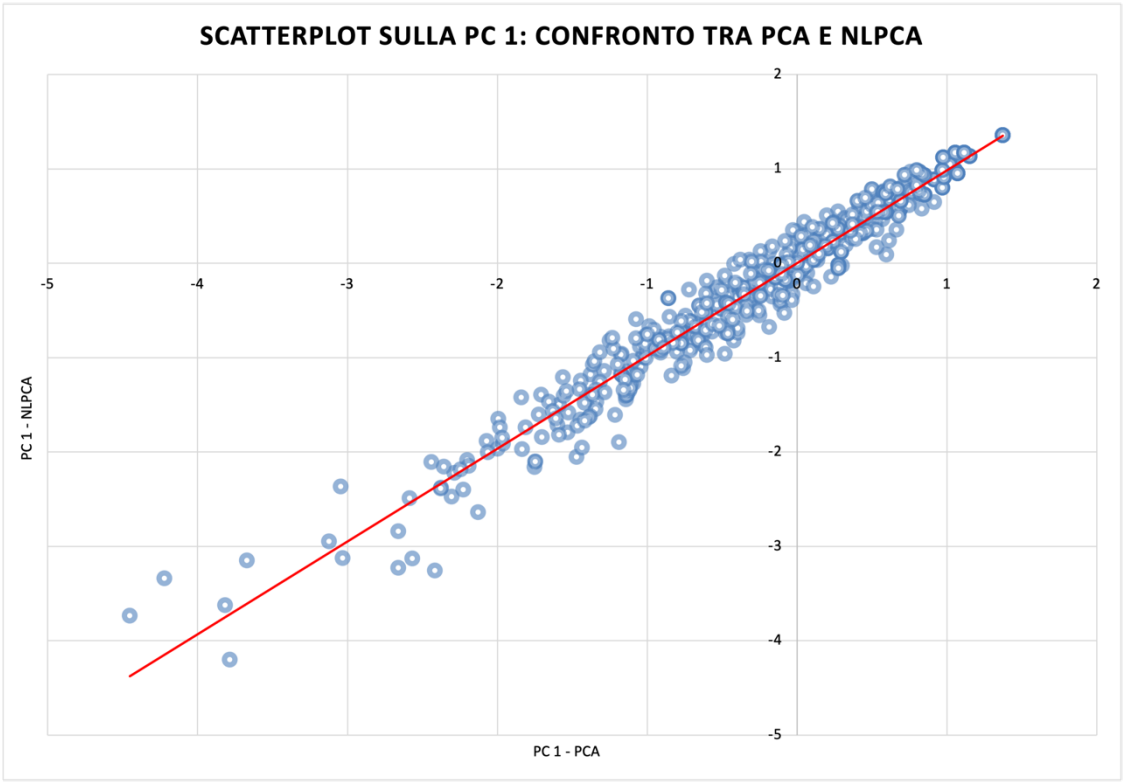


Figura 3.21: Scatterplot sulla PC 1 – confronto tra PCA e NLPCA
(Fonte: nostre elaborazioni)

Date le due variabili PC 1 per la PCA e PC 1 per la NLPCA, lo scatterplot è l'insieme dei punti del piano identificati utilizzando come prima coordinata i valori di PC 1 – PCA e come seconda coordinata i valori di PC 1 – NLPCA, in modo tale che, all'aumentare delle osservazioni, venga a comporsi una nube ("nuvola") di punti che si orienta nel piano cartesiano seguendo una certa linea di tendenza.

Nella Figura 3.21 è possibile notare come i punti dello scatterplot tendono a distribuirsi lungo una linea retta crescente, che si muove da sinistra verso destra e dal basso verso l'alto. In particolare, si individua chiaramente una nube di punti molto concentrata. Questa tendenza significa che tra le variabili esiste una relazione lineare positiva: dunque, al crescere (diminuire) dei valori di una variabile, ovvero PC 1 – PCA, si osserverà una crescita (diminuzione), in media, dei valori dell'altra, ovvero PC 1 – NLPCA.

Dal punto di vista matematico tale relazione viene esplicata tramite il coefficiente di correlazione lineare di Pearson, che in questo caso risulta pari a 0,9830, il quale esprime la relazione esistente tra le due variabili in termini di entità e direzione. Nella fattispecie,

l'andamento dei punti suggerisce una relazione con direzione positiva, nota anche come *concordanza*. Inoltre, si registra l'esistenza di una quantità estremamente limitata di *outliers* (valori anomali), ovvero di punti (dati) che non si adattano alla tendenza generale del resto dei dati.
Il grafico a dispersione, quindi, illustra il grado di correlazione tra la PC 1 della PCA e la PC 1 della NLPCA: nel caso in esame si può affermare che, complessivamente, l'applicazione delle due tecniche per la riduzione in una dimensione origina delle PC 1 tra loro concordi.

3.3.10 La NLPCA in due dimensioni con scaling level Nominal

Nell'ambito del presente studio metodologico, appare di interesse approfondire la ricerca con l'applicazione della NLPCA in due dimensioni, modificando lo scaling level. In particolare, lo scaling Ordinal viene sostituito con il Nominal. Lo scaling level Nominal è il vincolo meno stringente di tutti, in quanto impone esclusivamente il raggruppamento in categorie e non l'ordine delle stesse.
Tale analisi risulterà poi utile per comparare i risultati della NLPCA Nominal e della NLPCA Ordinal, da cui ricavarne le opportune considerazioni.

Nello script NLPCA.r i comandi sono stati settati così come di seguito:

```
dir <- "C:\\Users\\Administrator\\Desktop\\domanda nominal"  # settare
la directory di lavoro

miss <- "0"                    # etichetta assegnata ai missing values

K <- 5                         # numero categorie

scalev <- "nominal"            # scaling level ("nominal", "ordinal",
"numerical")

colstart <- 2                  # colonna in cui si trova la prima
variabile su cui effettuare l'analisi
colend <- 8                    # colonna in cui si trova l'ultima
variabile su cui effettuare l'analisi

dim <- 2               # dimensione dello spazio in cui proiettare
                       # scrivere 0 se non Ë noto per svolgere solo
analisi full preliminare

rot <- "No"            # perform varimax rotation ? ("No" - "Yes")

groups <- "No"                 # vuoi rappresentare grafici con una
variabile di raggruppamento ? ("No" - "Yes")
```

```
varg <- 14                          # colonna in cui si trova la variabile
di raggruppamento

datafile <- "dom accordo.txt"  # nome file che contiene i dati
```

Il software restituisce autovalori, varianza spiegata e varianza spiegata cumulata (Tabella 3.13).

Tabella 3.13: Valutazione della qualità dell'analisi a due dimensioni – NLPCA Nominal (Fonte: nostre elaborazioni)

Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
1	2,87	40,93%	40,93%
2	1,04	14,84%	55,77%

Come era nelle attese, con lo scaling level Nominal si ottiene una varianza spiegata maggiore rispetto all'analisi Ordinal, sebbene si tratti di un lievissimo incremento. Nel dettaglio, con il Nominal si trattiene il 55,77% dell'informazione complessiva, mentre con l'Ordinal viene mantenuto il 55,59% dell'informazione.

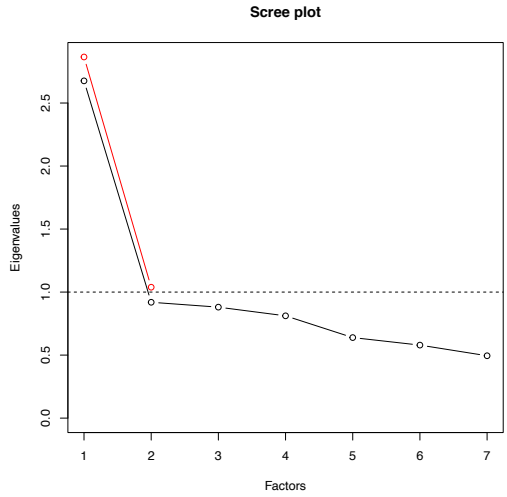


Figura 3.22: Scree Plot relativo all'analisi in due dimensioni – NLPCA Nominal (Fonte: nostre elaborazioni)

Nella Figura 3.22 viene rappresentato lo Scree Plot che confronta l'analisi full (spezzata nera) e l'analisi a due dimensioni (spezzata rossa). Naturalmente, come è accaduto per lo scaling Ordinal, anche con il Nominal la riduzione della dimensionalità consente un miglioramento dei risultati, in termini sia di autovalori sia di varianza spiegata.

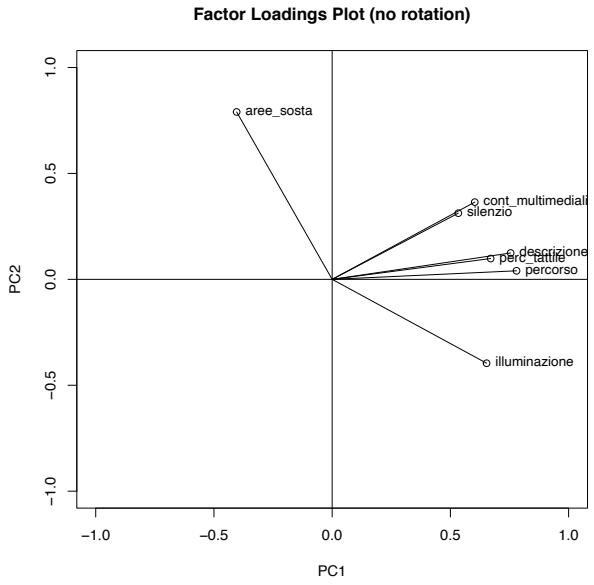


Figura 3.23: Factor Loadings Plot senza rotazione - NLPCA Nominal
(Fonte: nostre elaborazioni)

Osservando la Figura 3.23, il Factor Loadings Plot risulta difficilmente interpretabile, pertanto si effettua la rotazione Varimax.

Tabella 3.14: Valutazione della qualità dell'analisi a due dimensioni con rotazione – NLPCA Nominal
(Fonte: nostre elaborazioni)

Dimensione	Autovalori	Varianza spiegata	Varianza spiegata cumulata
1	2,42	34,56%	34,56%
2	1,48	21,21%	55,77%

A seguito della rotazione Varimax, l'informazione complessivamente trattenuta viene ridistribuita in modo più equilibrato tra le due PC (Tabella 3.14). In particolare, la varianza spiegata della prima PC si è ridotta (34,56%), mentre è aumentata quella associata alla seconda PC (21,21%).
Si riportano, di seguito, il Factor Loadings Plot ruotato (Figura 3.24) e la tabella dei loadings (Tabella 3.15).

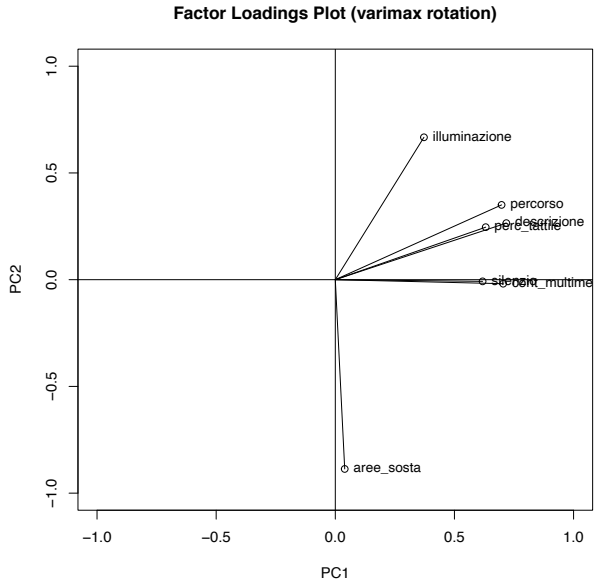


Figura 3.24: Factor Loadings Plot con rotazione – NLPCA Nominal
(Fonte: nostre elaborazioni)

Tabella 3.15: Loadings delle variabili – NLPCA Nominal
(Fonte: nostre elaborazioni)

Variabili	PC 1	PC 2
Illuminazione	0,3720	0,6670
Percorso	0,6978	0,3503
Silenzio	0,6182	-0,0079
Aree sosta	0,0395	-0,8866
Descrizione	0,7177	0,2649
Percorso tattile	0,6313	0,2460
Contenuti multimediali	0,7045	-0,0189

Considerato che nel presente elaborato non è strettamente rilevante la discussione a sé stante dei risultati della NLPCA Nominal, si è deciso di proseguire l'analisi con il confronto tra NLPCA Ordinal e NLPCA Nominal, con riferimento particolare ai Factor Loadings Plot e ai Transformation Plot.

3.3.11 Confronto tra NLPCA Ordinal e NLPCA Nominal in due dimensioni

L'obiettivo del presente paragrafo è comparare le risultanze emerse applicando la NLPCA con scaling level Ordinal e la NLPCA con scaling level Nominal.

La NLPCA si svolge sia con l'Ordinal sia con il Nominal, tuttavia, generalmente, si opta per una trasformazione di tipo Ordinal per mantenere anche nelle variabili trasformate

l'ordine delle categorie originarie: in tal caso, si ottengono delle trasformazioni non decrescenti.
Nel presente elaborato, dunque, è interessante procedere al confronto tra le due analisi, per individuare eventuali differenze nei risultati (Factor Loadings Plot e Transformation Plot) quando si opera la NLPCA in assenza del vincolo di ordine delle categorie.

La Figura 3.25 rappresenta il raffronto tra i Factor Loadings Plot delle analisi NLPCA con scaling Ordinal e Nominal.

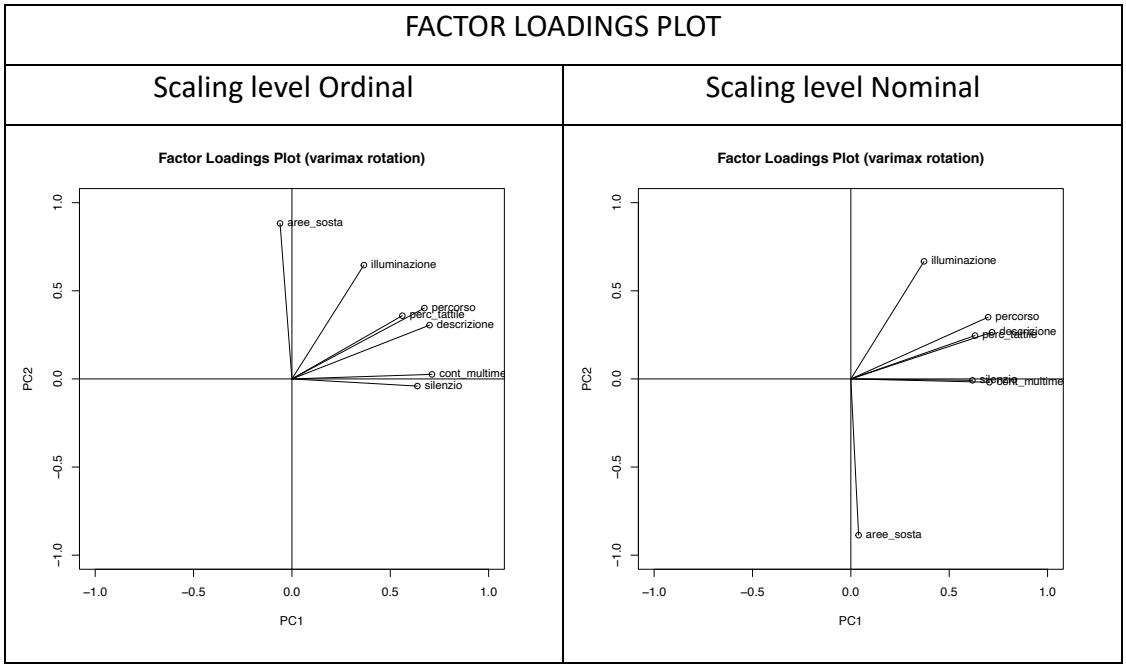


Figura 3.25: Confronto tra Factor Loadings Plot di NLPCA Ordinal e NLPCA Nominal (Fonte: nostre elaborazioni)

Dall'osservazione del Factor Loadings Plot individuato dalla NLPCA Ordinal era emersa una prima PC collegata all'*apprezzamento per gli aspetti di contenuto forniti nei percorsi* e una seconda PC collegata all'*apprezzamento per il modo in cui sono progettati (progettazione)*. Ora, confrontando tale grafico con il Factor Loadings Plot generato dalla NLPCA Nominal, le PC risultanti sono praticamente le stesse; la posizione di "aree sosta" è solo apparentemente stravolta nel grafico, e in realtà ha lo stesso significato che aveva nell'analisi Ordinal. Difatti, questa variabile è attribuita alla PC 2, solo che in questo caso è posizionata sulla parte negativa dell'asse. Come si vedrà, la sua

correlazione negativa con la seconda componente principale si spiega perfettamente alla luce del risultante Transformation Plot.

Si procede ora a confrontare gli Object Scores Plot derivanti dall'applicazione della NLPCA con i due differenti scaling level (Figura 3.26).

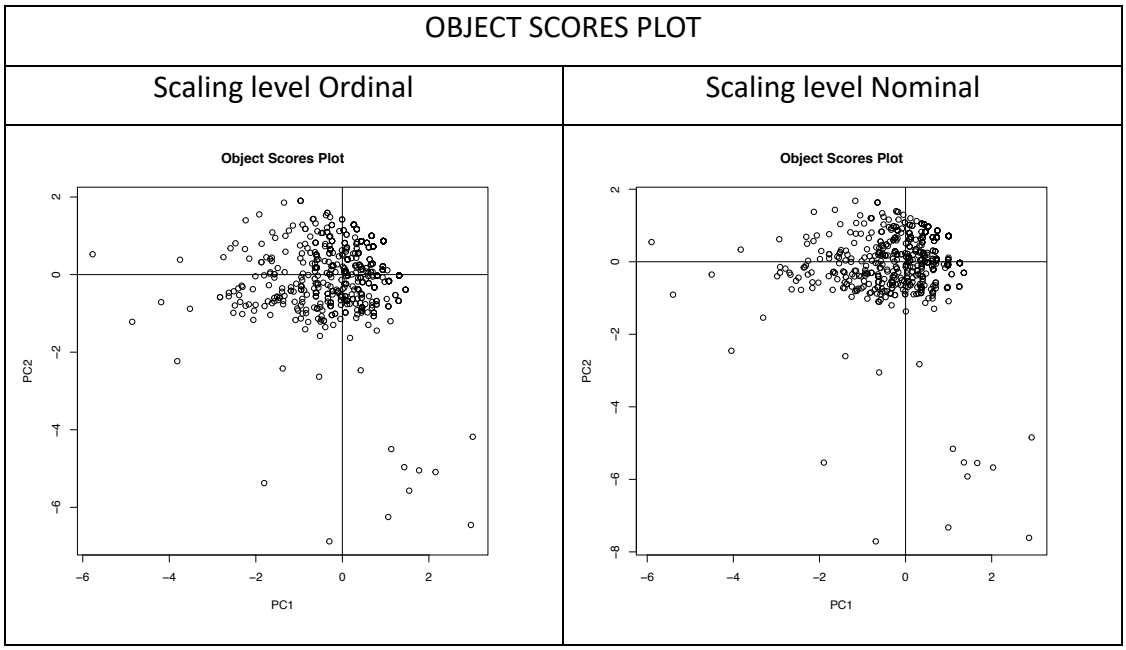
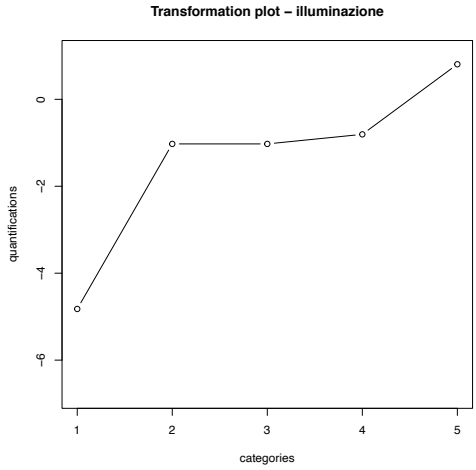
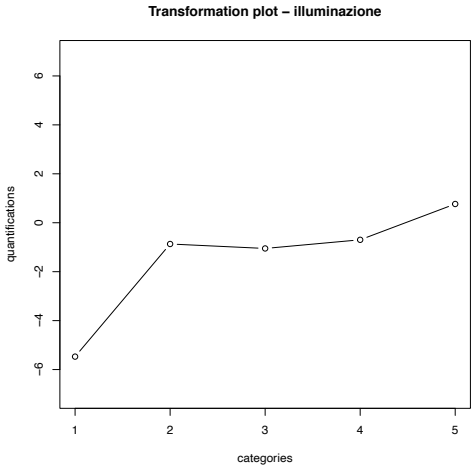
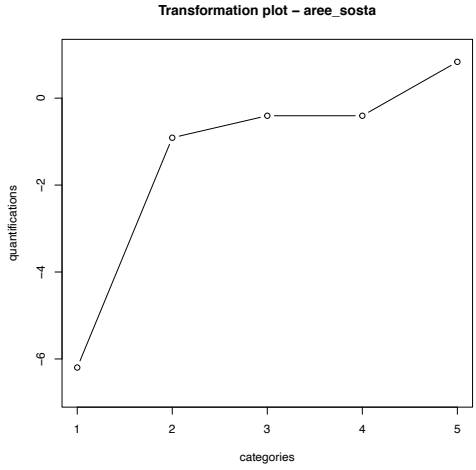
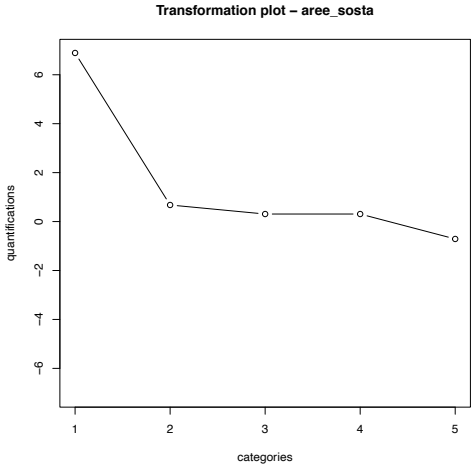
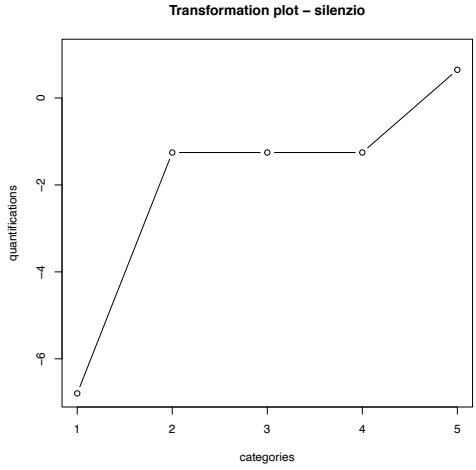
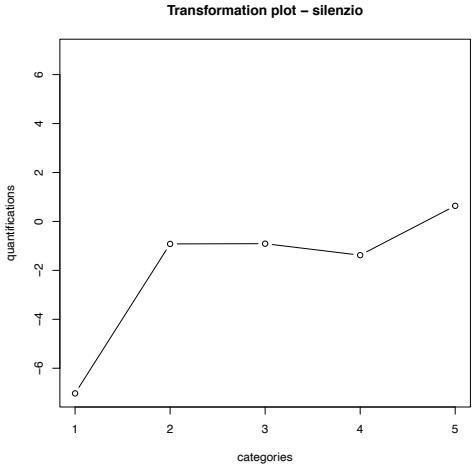
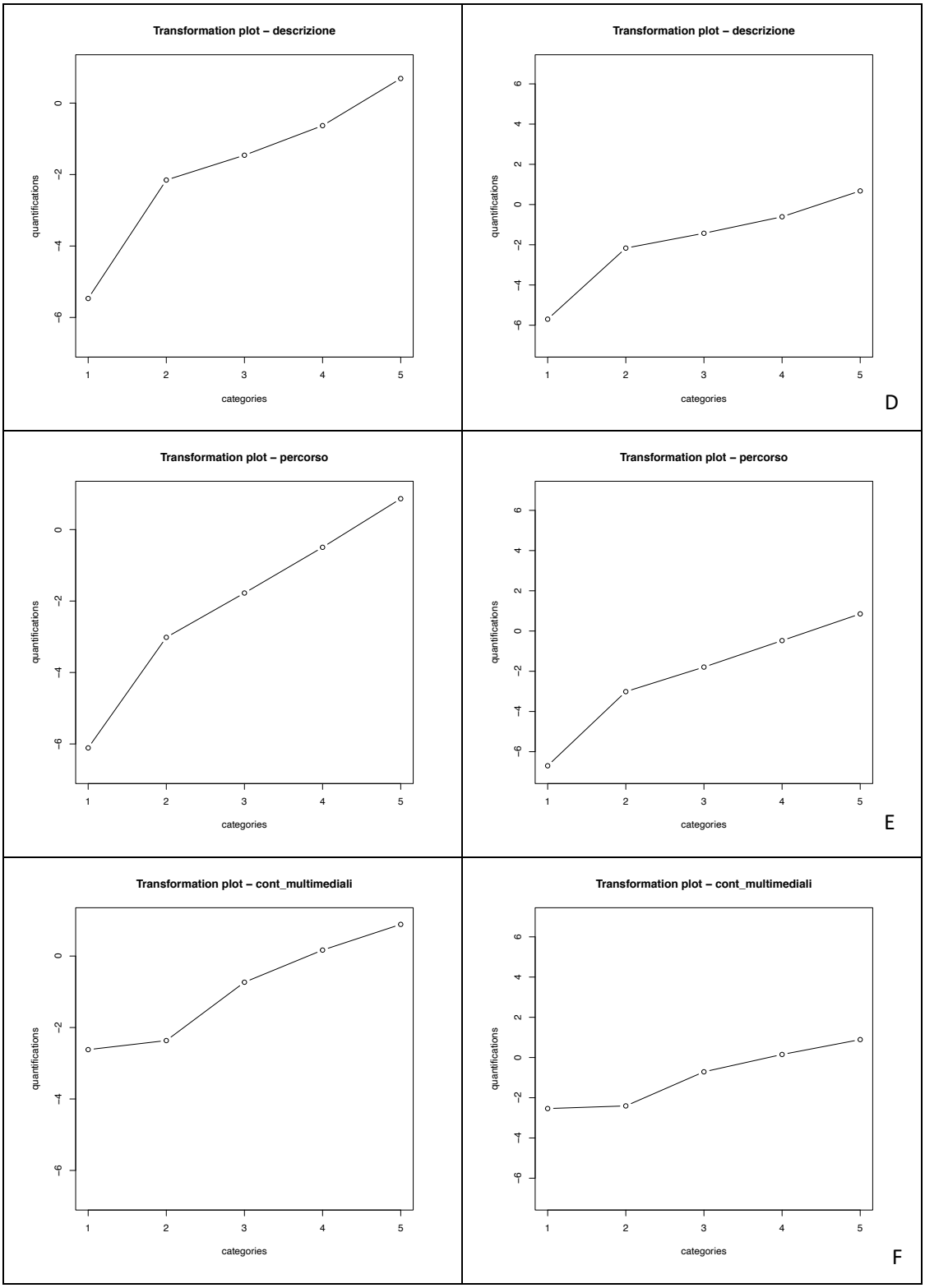


Figura 3.26: Confronto tra Object Scores Plot di NLPCA Ordinal e NLPCA Nominal (Fonte: nostre elaborazioni)

L'interpretazione dell'Object Scores Plot con scaling Nominal è la medesima di quella descritta per lo scaling Ordinal (si veda pag. 101 per la spiegazione dettagliata). Rispetto all'analogo Ordinal, nel grafico Nominal si osserva una concentrazione più intensa delle unità statistiche attorno ai valori medi per entrambe le PC.

Di seguito viene mostrato il confronto tra i Transformation Plot dell'analisi Ordinal e quelli dell'analisi Nominal (Figura 3.27).

TRANSFORMATION PLOT	
Scaling level Ordinal	Scaling level Nominal
	 A
	 B
	 C


Transformation plot – descrizione
quantifications
categories
Transformation plot – descrizione
quantifications
categories
D
Transformation plot – percorso
quantifications
categories
Transformation plot – percorso
quantifications
categories
E
Transformation plot – cont_multimediali
quantifications
categories
Transformation plot – cont_multimediali
quantifications
categories
F

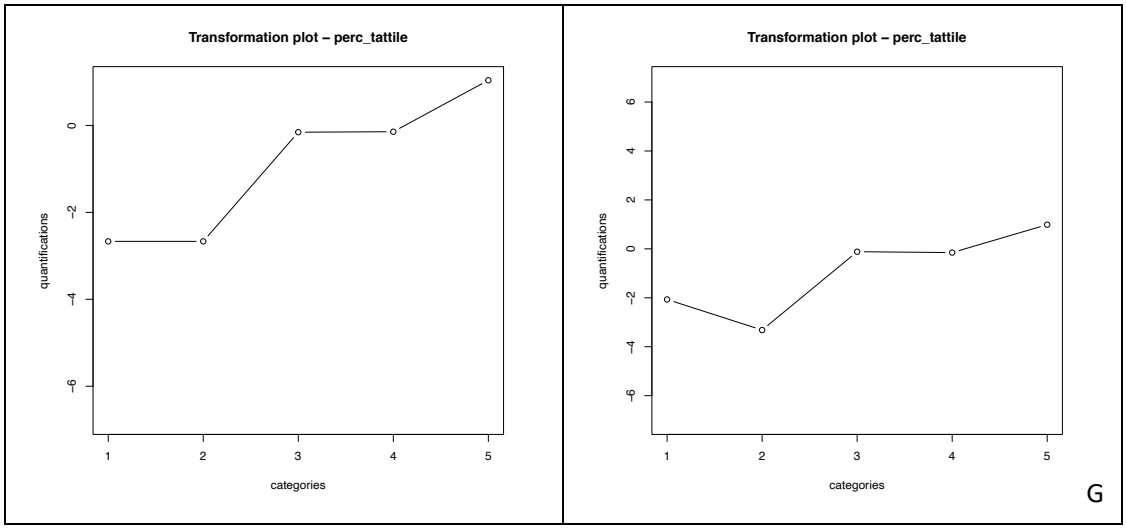


Figura 3.27: Confronto tra Transformation Plot di NLPCA Ordinal e NLPCA Nominal (Fonte: nostre elaborazioni)

Complessivamente, i risultati sono molto interessanti poiché confermano che le categorie sono sostanzialmente ordinate, quindi le risposte sono state date con una coerenza di fondo. In particolare, quasi tutte le coppie di Transformation Plot presentano un andamento analogo tra loro (la descrizione dei Transformation Plot relativi alla NLPCA Ordinal è stata approfondita nel paragrafo 3.3.4).

Alcuni commenti a parte merita il Transformation Plot di "aree sosta" (grafico B). Esso è decrescente e si rileva una consistente decrescita della $q2$ rispetto alla $q1$, poi vi è un'ulteriore ma lievissima diminuzione della $q3$, segue una sostanziale costanza nella $q4$ e infine una leggera decrescita nella $q5$. Tuttavia, la forma del Transformation Plot è grosso modo simmetrica a quella del Transformation Plot crescente indotto dallo scaling level Ordinal. Questo significa che non vi è, in sostanza, differenza di impatto dell'item sui risultati dell'analisi, se non per il fatto che – avendo in pratica semplicemente "ribaltato" le quantificazioni delle categorie – a valutazioni negative si associano valori elevati ed è esattamente questo che giustifica la correlazione negativa dell'item con la seconda componente principale. In termini interpretativi, dunque, l'item ha lo stesso significato nelle due analisi.

Infine, per le due analisi NLPCA Ordinal e NLPCA Nominal è possibile realizzare un confronto con scatterplot dei valori delle due componenti principali estratte (rispettivamente PC 1 e PC 2).

Nel primo scatterplot (Figura 3.28), sull'asse X sono riportati gli object scores corrispondenti alla PC 1 della NLPCA Nominal, mentre sull'asse Y sono identificati gli object scores corrispondenti alla PC 1 della NLPCA Ordinal.

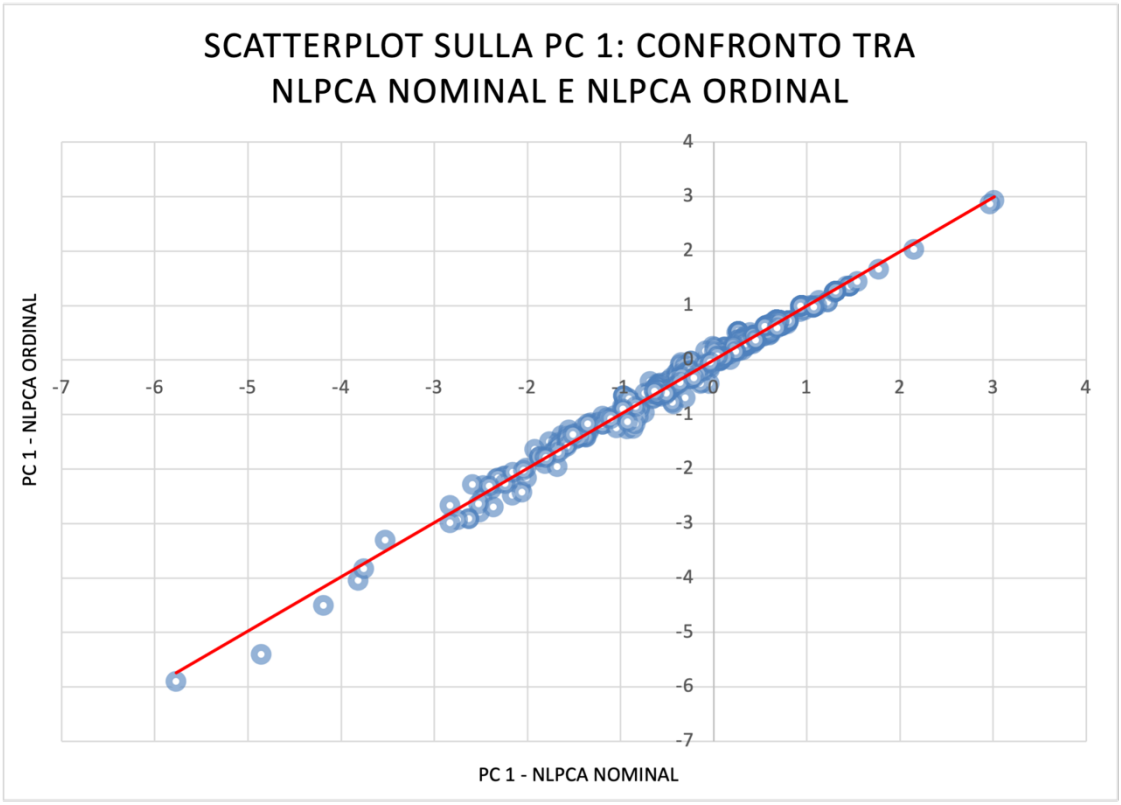


Figura 3.28: Scatterplot sulla PC 1 – confronto tra NLPCA Nominal e NLPCA Ordinal (Fonte: nostre elaborazioni)

Le 665 osservazioni individuano una nube di punti che risulta essere molto concentrata, poiché sostanzialmente tutti i punti tendono a disporsi lungo la linea di tendenza a 45°. Tale tendenza significa che tra le variabili esiste una relazione lineare positiva: dunque, al crescere o diminuire dei valori di PC 1 – NLPCA Nominal si osserverà una crescita o diminuzione dei valori di PC 1 – NLPCA Ordinal. Questa relazione viene matematicamente esplicata tramite il coefficiente di correlazione lineare di Pearson, che

in questo caso assume valore pari a 0,9939: osservando il grafico, l'andamento dei punti suggerisce una correlazione positiva (concordanza) tra le PC 1. Inoltre, si registra l'assenza di *outliers* (valori anomali), poiché tutti i punti si adattano alla tendenza lineare.

Il grafico a dispersione, quindi, illustra il grado di correlazione tra la PC 1 della PCA Nominal e la PC 1 della PCA Ordinal: in questo caso si può affermare che la scelta dello scaling Ordinal piuttosto che lo scaling Nominal non comporta rilevanti differenze in termini di PC 1: in altre parole, si generano PC 1 tra loro concordi.

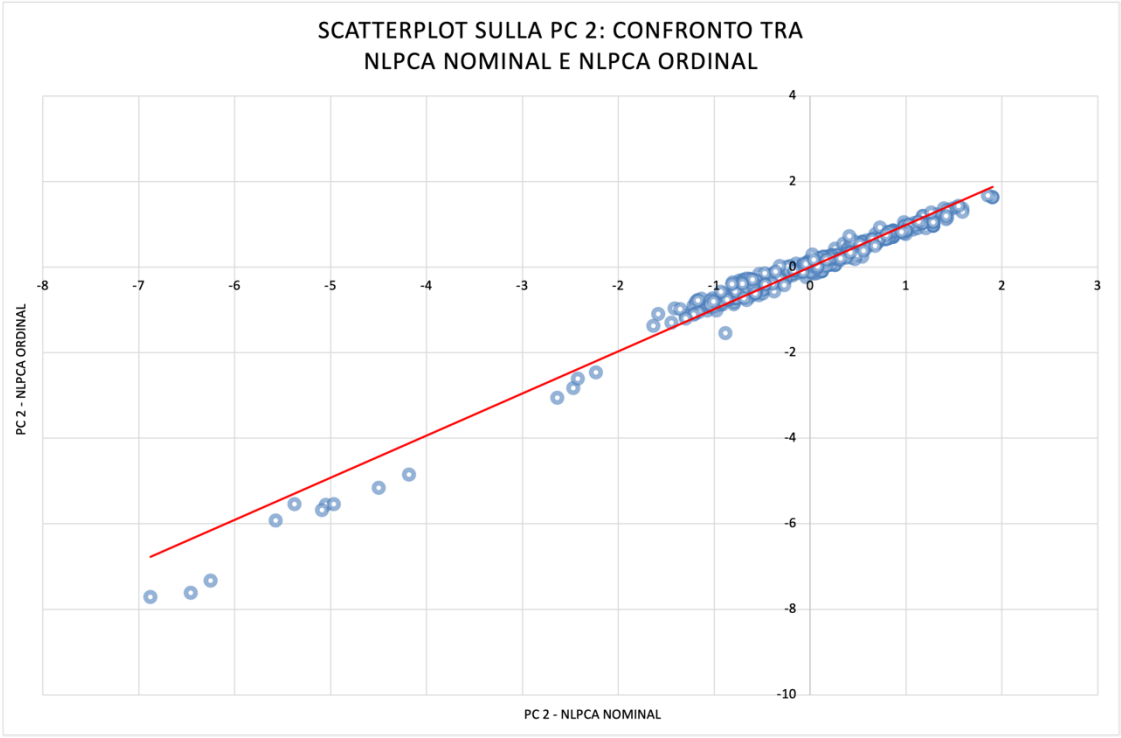


Figura 3.29: Scatterplot sulla PC 2 – confronto tra NLPCA Nominal e NLPCA Ordinal
(Fonte: nostre elaborazioni)

Nel secondo scatterplot (Figura 3.29), sull'asse X sono riportati gli object scores corrispondenti alla PC 2 della NLPCA Nominal, mentre sull'asse Y sono identificati gli object scores corrispondenti alla PC 2 della NLPCA Ordinal.

Anche in questo grafico, le osservazioni complessivamente identificano una nube di punti concentrata, tuttavia rispetto allo scatterplot sulla PC 1 si rileva una maggiore quantità di *outliers* che non si adattano alla tendenza generale dei dati.

La tendenza individuata dalla nube di punti indica che, analogamente al grafico precedente, tra le variabili esiste una relazione lineare positiva: al crescere (diminuire) dei valori di PC 2 – NLPCA Nominal si registrerà una crescita (diminuzione) dei valori di PC 2 – NLPCA Ordinal. Il coefficiente di correlazione lineare di Pearson, che rappresenta a livello matematico tale relazione, è pari a 0,9847: osservando il grafico, l'andamento dei punti suggerisce una correlazione positiva (concordanza) tra le PC 2.
Il grafico a dispersione, quindi, illustra il grado di correlazione tra la PC 2 della NLPCA Nominal e la PC 2 della NLPCA Ordinal: salvo un limitato numero di *outliers*, ai quali viene attribuito un valore della seconda componente principale leggermente inferiore utilizzando lo scaling level Nominal, in generale la scelta dello scaling Ordinal o dello scaling Nominal non comporta importanti differenze nemmeno in termini di PC 2, originando PC 2 tra loro concordi.

CAPITOLO 4

CONCLUSIONI

4.1 Discussione dei risultati

Il presente capitolo è finalizzato a presentare un riepilogo complessivo del lavoro di studio e ricerca che è stato svolto, insieme alle risultanze ottenute.
Le analisi che sono state descritte nel capitolo precedente hanno permesso di comprendere al meglio il funzionamento dell'applicazione di NLPCA e PCA su una batteria di item in scala di Likert, anche tramite lo studio comparativo dei risultati che sono emersi applicando le due tecniche.

Poiché l'obiettivo principale di questo elaborato è condurre un'analisi sperimentale sulle tecniche PCA e NLPCA, in primo luogo si è proceduto a effettuare un approfondimento sulla riduzione della dimensionalità, in termini sia di approccio all'analisi dei dati sia di effettiva procedura statistica di riduzione di dataset ad alta dimensionalità. In seguito, si è realizzata un'importante parte compilativa e teorica volta a esplorare a fondo la PCA, dalla quale sono emerse diverse considerazioni ed elaborazioni frutto del lavoro di analisi delle fonti. L'approfondita ricerca dal punto di vista teorico è risultata fondamentale per poter poi proseguire con l'indagine empirica: difatti, il confronto tra gli esiti dell'applicazione al caso studio e le evidenze teoriche (derivanti dalla letteratura e dagli studi di ricercatori) ha consentito di elaborare le opportune osservazioni.

L'esperimento pratico è stato effettuato su un dataset generato da una domanda nella forma di batteria di item in scala di Likert, contenuta in un questionario creato e somministrato nell'ambito del progetto DS4BS. Il fine ultimo dell'analisi empirica compiuta in questa tesi è quello di 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 si traduce nell'applicazione al dataset di due diverse tecniche di riduzione della dimensionalità: NLPCA e PCA. In particolare, la NLPCA è la tecnica appropriata per trattare variabili di tipo qualitativo, quali sono le variabili originate dalla

batteria di item in scala di Likert; contrariamente, la PCA è la tecnica adeguata ad analizzare le variabili di tipo quantitativo. Pertanto, lo studio metodologico è stato condotto con lo scopo di indagare su un caso pratico i limiti che si riscontrano nella PCA; d'altra parte, allo stesso tempo è stato possibile verificare empiricamente i vantaggi, in termini di miglioramento dell'analisi, che può apportare l'impiego della NLPCA.

L'applicazione sia della NLPCA sia della PCA ha comportato lo stesso iter di esecuzione, con il rispetto dei seguenti passaggi: determinare la dimensione q in cui eseguire la riduzione della dimensionalità e procedere all'interpretazione delle q componenti principali. Ciò consente di realizzare una buona sintesi della variabile p-dimensionale in una dimensione inferiore, creando degli indicatori di sintesi che contengono gran parte dell'informazione che era contenuta nelle p variabili iniziali.

Il lavoro di ricerca è iniziato con lo svolgimento della NLPCA nella sua prima parte, ovvero la determinazione della dimensione ottimale per la riduzione della dimensionalità. In seguito, lo stesso è stato effettuato applicando la PCA. Nel caso studio si è deciso di optare per $q = 2$ come dimensione ottimale di riduzione in entrambe le applicazioni, riassumendo le 7 variabili iniziali in 2 componenti principali. Tuttavia, a fronte di questa similarità, si registra una differenza che appare importante ricordare: tramite la NLPCA si ottiene una varianza spiegata leggermente superiore rispetto alla PCA.

La ricerca è proseguita sviluppando la seconda parte delle due tecniche, vale a dire l'attribuzione di un appropriato significato alle due componenti principali. Analogamente a quanto svolto nello step precedente, inizialmente è stata applicata la NLPCA: a primo impatto, dall'analisi del Factor Loadings Plot non è stato possibile individuare due distinti gruppi di variabili, come nei casi ideali, tuttavia una differente prospettiva di interpretazione di tale grafico, congiuntamente alla lettura dei loadings delle variabili, ha permesso di identificare una prima PC collegata all'*apprezzamento per gli aspetti di contenuto forniti nei percorsi* e una seconda PC collegata all'*apprezzamento per il modo in cui sono progettati (progettazione)*.

Successivamente, l'operazione di interpretazione delle PC è stata ripetuta nell'ambito della PCA. A questo punto, l'evidenza più significativa che è emersa dalla comparazione tra le due tecniche consiste nella sostanziale disuguaglianza tra le rispettive coppie di PC ottenute: indubbiamente, le PC hanno dei significati estremamente differenti, non confrontabili tra loro.
Dunque, concludendo questa prima parte dell'analisi sperimentale, è possibile affermare che fintanto che si è trattato di discutere la decisione sulla riduzione della dimensionalità l'applicazione delle due tecniche NLPCA e PCA ha condotto a degli esiti concordi tra loro, con ciò significando che la scelta della dimensione ottimale in cui effettuare la riduzione è risultata essere la stessa in ambedue i casi. In seguito, quando si è reso necessario approfondire e ricercare il significato delle nuove variabili, l'applicazione dell'una o dell'altra tecnica ha portato a delle componenti principali completamente differenti.
Riassumendo, a fronte di una riduzione in una eguale dimensione, si ottengono però due componenti principali con diverso significato e, quindi, in ultima istanza, a dei risultati complessivamente diversi tra loro.

Infine è rilevante sottolineare che, come citato in precedenza, la NLPCA permette di trattenere una quantità di informazione superiore rispetto a quella della PCA nella riduzione in una dimensione inferiore, come era preliminarmente pronosticabile. Difatti, dalla letteratura è risaputo che per trattare in modo adeguato variabili di tipo qualitativo è necessario applicare la NLPCA, mentre la PCA in questo caso è la tecnica errata. Pertanto, alla luce degli esiti dell'analisi empirica, è possibile affermare che se non si considera la natura delle variabili, e di conseguenza queste non vengono trattate in modo coerente, l'analisi può condurre a risultati distorti rispetto a quelli che si otterrebbero applicando la tecnica corretta.
Dunque, la scelta della tecnica da utilizzare è di notevole importanza, e da essa dipende fortemente la qualità complessiva dell'analisi.

A questo punto, poiché nella precedente indagine non si è potuto procedere a un confronto tra NLPCA e PCA su due componenti principali equivalenti, si è deciso di operare la riduzione in una sola dimensione, in questo modo forzando le due analisi ad avere un'unica componente principale dallo stesso significato, denominata il *grado*

complessivo di accordo. Parimenti all'applicazione in due dimensioni, si osserva che anche in questo caso la NLPCA consente di mantenere una maggiore informazione nell'analisi a dimensione ridotta.
In seguito, si è realizzato uno scatterplot al fine di verificare quanto, almeno su quest'unica componente principale di accordo complessivo, le due componenti principali (quella della NLPCA e quella della PCA) fossero concordi. Rappresentando le 665 coppie di punti si è delineata una nube di punti, distribuita lungo una linea di tendenza lineare, con coefficiente di correlazione lineare pari a 0,9830: si è rilevata una correlazione positiva (concordanza) tra la PC 1 della PCA e la PC 1 della NLPCA, con le due PC 1 tra loro concordi.

A conclusione della parte sperimentale, nell'ambito della presente tesi di ricerca si è voluto approfondire lo studio tramite il confronto dei risultati della NLPCA in due dimensioni utilizzando due diversi scaling level: Ordinal e Nominal. Generalmente, la NLPCA si svolge applicando lo scaling Ordinal, per mantenere anche nelle variabili trasformate l'ordine delle categorie originarie; in alternativa è possibile utilizzare lo scaling Nominal, che è il vincolo meno stringente di tutti in quanto impone il solo raggruppamento in categorie (quindi, rispetto all'Ordinal, manca il vincolo di ordine delle categorie).
Dalla comparazione delle due analisi è emerso che con lo scaling level Nominal si ottiene una varianza spiegata maggiore rispetto all'analisi Ordinal, come era nelle attese (benché si tratti di un lievissimo incremento). In seguito, per quanto riguarda l'interpretazione delle componenti principali, si è rilevata l'esistenza di due coppie di PC praticamente coincidenti, con medesimo significato. Pertanto, l'impiego di uno scaling rispetto all'altro, in questo caso, non conduce a rilevanti differenze nell'ambito della riduzione della dimensionalità, se non in termini di una lieve diversità nella quantità di informazione trattenuta nella dimensione inferiore.

Infine, si è deciso di effettuare un ulteriore confronto tramite scatterplot dei valori delle due componenti principali estratte (PC 1 e PC 2 delle analisi Ordinal e Nominal). In entrambi i grafici si è potuta constatare una correlazione positiva (concordanza) tra le variabili, con ciò indicando che a prescindere dalla scelta dello scaling level la NLPCA genera sia PC 1 sia PC 2 tra loro concordi.

In conclusione, lo studio metodologico che è stato condotto è servito a verificare, almeno su questo caso studio, una delle più importanti proprietà riconosciute in letteratura quando si parla di NLPCA e PCA nel campo della riduzione di dimensionalità. Difatti, l'analisi empirica svolta ha rilevato un risultato di fondamentale importanza: talvolta viene trascurata la natura delle variabili, e di conseguenza viene applicata l'una o l'altra tecnica con una certa superficialità, senza considerare che possano così delinearsi degli effetti distorti. Tale leggerezza può condurre all'errore, come è stato dimostrato nell'applicazione al caso in esame, nel quale emerge chiaramente che l'impiego della PCA ai fini dell'elaborazione di variabili qualitative porti a dei risultati totalmente differenti rispetto al caso in cui venga applicata la NLPCA, che è la tecnica corretta, la quale conduce a una migliore e più approfondita comprensione dei dati analizzati.
Dunque, la tecnica da utilizzare per trattare determinate variabili deve essere scelta in modo coerente con esse, non trascurandone la natura, poiché un'applicazione non rigorosa può produrre delle analisi oggettivamente imprecise o inesatte.

Alla luce di quanto esposto in precedenza, la ricerca condotta nel presente elaborato ha portato a risultati che sostanzialmente hanno confermato le ricerche originali riguardanti la NLPCA e la PCA nell'ambito della riduzione della dimensionalità. Tuttavia, è necessario sottolineare che tale evidenza empirica appare circoscritta allo specifico caso studio che è stato impiegato per questa tesi di ricerca, per cui un limite è rappresentato dall'impossibilità di generalizzazione dei risultati ottenuti.
Non si esclude, quindi, che in futuro ulteriori ricerche possano condurre a degli esiti discordanti rispetto a quanto emerso in questa tesi, o addirittura all'identificazione di *pattern* o strutture nascoste nei dati.

APPENDICE

Gentile visitatore, l'Università degli Studi di Brescia con Fondazione Brescia Musei e Comune di Brescia e il supporto di Fondazione Cariplo sta conducendo un'indagine sull'esperienza vissuta dagli utenti durante la visita del Museo di Santa Giulia. Il fine è creare percorsi personalizzati sulla base delle preferenze dei visitatori, anche tramite l'utilizzo di app che forniscano, per esempio, mappe interattive per la visita del Museo. Le chiedo pochi minuti del suo tempo per la compilazione totalmente anonima del questionario.

***Campo obbligatorio**

1. Come è venuto a conoscenza del Museo di Santa Giulia? (è possibile più di una risposta) *
   - ☐ Parenti/amici e passaparola
   - ☐ Sito web di Fondazione Brescia Musei
   - ☐ Altri siti web
   - ☐ Social media di Fondazione Brescia Musei
   - ☐ Trasmissioni TV/radio
   - ☐ Altri social media
   - ☐ Riviste/giornali quotidiani
   - ☐ Infopoint (di Brescia centro)
   - ☐ Cartelli stradali (segnaletica)
   - ☐ Altro……

2. Se è venuto a conoscenza del Museo tramite "altri social media", indichi quali
   _____
   _____

3. Se è venuto a conoscenza del Museo tramite "altri siti web", indichi quali
   _____
   _____

4. Come nella graduatoria di una gara, metta in ordine di importanza da 1 (il più importante) a 5 (il meno importante) i seguenti motivi che l'hanno spinta a visitare il Museo di Santa Giulia: *
   1) Approfondire la conoscenza storica della città di Brescia
   2) Ammirare, dal punto di vista architettonico, il complesso monastico sede del museo di Santa Giulia
   3) Studiare e approfondire come si svolgeva la vita quotidiana nelle differenti epoche del passato
   4) Osservare l'arte figurativa e scultorea presente nel complesso museale

5) Partecipare attivamente all'esperienza museale attraverso la realtà virtuale e la realtà aumentata

5. È la prima volta che visita il Museo di Santa Giulia? *
   - ☐ Sì (passa alla domanda 8)
   - ☐ No (passa alla domanda 6)

6. Cosa l'ha spinta a tornare? (è possibile più di una risposta) *
   - ☐ Una mostra temporanea
   - ☐ Per accompagnare amici/parenti
   - ☐ Per rivivere l'esperienza a distanza di tempo
   - ☐ Per accrescere le mie conoscenze
   - ☐ Un'attività didattica
   - ☐ Per completare la visita
   - ☐ Altro _____

7. Se è tornato per visitare "una mostra temporanea", indichi quale
   _____
   _____
   _____

8. Con chi ha condiviso questa esperienza? *
   - ☐ Amici
   - ☐ Famiglia
   - ☐ Gruppo organizzato
   - ☐ Partner
   - ☐ Nessuno, sono venuto/a da solo/a
   - ☐ Altro_____

9. Come è arrivato qui? (è possibile più di una risposta) *
   - ☐ A piedi
   - ☐ In bicicletta
   - ☐ In auto
   - ☐ In autobus
   - ☐ In metro
   - ☐ In treno
   - ☐ Altro_____

10. Ha pernottato o pernotterà a Brescia? *
    - ☐ Sì (passa alla domanda 11)
    - ☐ No (passa alla domanda 12)

11. Quante notti si ferma a Brescia? *
    _____

12. In generale, quanto si ritiene soddisfatto/a dei seguenti aspetti? *

	Molto insoddisfatto	Insoddisfatto	Né soddisfatto né insoddisfatto	Soddisfatto	Molto soddisfatto
Orari di apertura					
Facilità di raggiungimento (indicazioni stradali, parcheggi, mezzi pubblici)					
Cortesia e competenza del personale					
Orientamento nei percorsi					
Cura e pulizia degli ambienti					
Accessibilità per gli utenti con ridotta capacità motoria					
Materiali informativi (schede, pannelli, didascalie)					
Servizi di accoglienza					
Prezzo del biglietto					

13. Quanto è complessivamente soddisfatto/a della visita? (Da 1, molto insoddisfatto, a 10, pienamente soddisfatto) *

_____

14. 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					

15. Di seguito troverà delle coppie di aggettivi di significato opposto: selezioni la casella che, tra i due aggettivi estremi, meglio corrisponde alla Sua percezione relativa alla visita del museo di Santa Giulia: *

Noiosa	O	O	O	O	O	O	O	Piacevole
Banale	O	O	O	O	O	O	O	Interessante
Difficile	O	O	O	O	O	O	O	Agevole
Insignificante	O	O	O	O	O	O	O	Coinvolgente
Ordinaria	O	O	O	O	O	O	O	Sorprendente

16. Quanto hanno contribuito, su una scala da “Per nulla” a “Moltissimo”, i seguenti elementi nel rendere unica la sua esperienza nel Museo di Santa Giulia? *

	Per nulla	Poco	Abbastanza	Molto	Moltissimo
Il susseguirsi di avvenimenti storici che hanno influito sullo sviluppo del complesso monastico					
La maestosità delle architetture monumentali dell’antica Brixia					
La storia e le peculiarità del complesso monastico					
Le forme, le geometrie e i colori dell’arte figurativa presente nel museo					
L’osservazione dei diversi sistemi di rappresentazione grafica che si sono susseguiti nei secoli					
La possibilità di conoscere le abitudini e gli oggetti della quotidianità nei diversi periodi storici					
La dimensione estetica associata alla visita, connessa al desiderio di conoscere opere e reperti importanti					
La dimensione edonistica rispondente al mio desiderio di trascorrere un momento personale piacevole					

17. Quanto hanno contribuito, su una scala da “Per nulla” a “Moltissimo”, i seguenti elementi aggiuntivi nel rendere unica la sua esperienza nel Museo di Santa Giulia? *

	Per nulla	Poco	Abbastanza	Molto	Moltissimo	Non applicabile (elemento non sperimentato)
La calma e la tranquillità del giardino esterno (Viridarium)						
La possibilità di utilizzare gli ArtGlass (occhiali multimediali)						
La presenza di un percorso tattile						
La possibilità di sperimentare una forma nuova di conoscenza attraverso l’utilizzo delle tecnologie interattive presenti nel museo						
L’offerta del bookshop						
La presenza di contenuti digitali di elevata qualità (foto, video, ricostruzioni 3D, musiche, supporti audio ecc.)						

18. Il Museo sta valutando la possibilità di potenziare alcune proposte: tra le seguenti cosa Le piacerebbe trovare? Indichi 3 alternative e le ordini per importanza (come nella classifica di una gara: 1= la più importante; 3= la meno importante)*

- ☐ Maggiore multimedialità (esempio: video, musica, …)
- ☐ Più materiali informativi (schede, pannelli, didascalie, brochure)
- ☐ Attività interattive (per adulti e per bambini)

- ☐ Audio-guide
- ☐ Maggiori esperienze sensoriali (esempio: percorso tattile, auditivo, olfattivo, ...)
- ☐ Maggiore realtà aumentata (esempio: occhiali ArtGlass)
- ☐ Maggiore e più diversificata offerta del bookshop

19. Quali altri musei intende visitare nel prossimo futuro? *
    - ☐ Brixia. Parco Archeologico di Brescia romana
    - ☐ Pinacoteca Tosio Martinengo
    - ☐ Museo delle Armi "Luigi Marzoli"
    - ☐ Altro _____

20. Ha figli in età scolare? *
    - ☐ Sì ☐ No (passa alla domanda 27)

21. Sa che Fondazione Brescia Musei predispone molte attività per bambini e ragazzi? *
    - ☐ Sì ☐ No

22. Conosce i seguenti strumenti a disposizione per le famiglie in visita autonoma ai Musei? *

    App game Geronimo Stilton Brescia Musei adventures
    - ☐ Sì ☐ No

    Activity books
    - ☐ Sì ☐ No

23. È a conoscenza del ricco calendario di attività rivolte al pubblico organizzate da Fondazione Brescia Musei? *
    - ☐ Decisamente no
    - ☐ Più no che sì
    - ☐ Più sì che no
    - ☐ Decisamente sì

24. Quali dei seguenti programmi conosce? *

    Museo e Scuola
    - ☐ Sì ☐ No

    Museo per Tutti
    - ☐ Sì ☐ No

25. Se conosce altri programmi, indichi quali

    _____
    _____

26. Desidera esprimere qualche suggerimento/critica/consiglio?

_____
_____
_____
_____

27. Se vuole essere informato sulle iniziative di Fondazione Brescia Musei, può iscriversi alla newsletter, lasciando qui il suo indirizzo email o sul web alla pagina https://www.bresciamusei.com/iscriviti-alla-newsletter, dove troverà anche l'informativa privacy.

_____
_____

Dati anagrafici

28. Sesso: * ☐ Femmina ☐ Maschio ☐ Altro

29. Età in anni compiuti: * ____

30. Titolo di studio: *
- ☐ Licenza elementare o diploma terza media (passa alla domanda 38)
- ☐ Diploma di scuola media superiore (passa alla domanda 38)
- ☐ Laurea triennale (passa alla domanda 35)
- ☐ Laurea magistrale (passa alla domanda 36)
- ☐ Titolo post-laurea (passa alla domanda 37)

31. Quale Laurea triennale? *

_____
_____

32. Quale Laurea magistrale? *

_____
_____

33. Quale Titolo post-laurea? *

_____
_____

34. Residenza: *
- ☐ Brescia (passa alla domanda 42)
- ☐ Provincia di Brescia (passa alla domanda 39)
- ☐ Italia (passa alla domanda 40)
- ☐ Estero (passa alla domanda 41)

35. In quale comune? *

_____
_____

36. In quale città italiana? *

_____
_____

37. In quale Stato? *

_____
_____

38. Professione: *

☐ Impiegato/a ☐ Studente ☐ Pensionato/a ☐ Imprenditore/ libero professionista

☐ Casalingo/a ☐ Operaio/a ☐ Insegnante ☐ Altro_____

Grazie per la collaborazione!

BIBLIOGRAFIA

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality this-month. In *Nature Methods* (Vol. 15, Issue 6, pp. 399–400). Nature Publishing Group. https://doi.org/10.1038/s41592-018-0019-x

Banks, D. L., & Fienberg, S. E. (2003). Data Mining, Statistics. *Encyclopedia of Physical Science and Technology*, 247–261. https://doi.org/10.1016/B0-12-227410-5/00164-2

Bassi, F., & Ingrassia, S. (2022). *Statistica per analisi di mercato. Metodi e strumenti*. Pearson.

Box, P. O., Van Der Maaten, L., Postma, E., & Van Den Herik, J. (2009). *Tilburg centre for Creative Computing Dimensionality Reduction: A Comparative Review Dimensionality Reduction: A Comparative Review*. http://www.uvt.nl/ticc

Brehmer, M., Sedlmair, M., Ingram, S., & Munzner, T. (2014). Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. *ACM International Conference Proceeding Series*, *10-November-2015*, 1–8. https://doi.org/10.1145/2669557.2669559

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., & Wang, Y. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. In *Nature Reviews Cancer* (Vol. 8, Issue 1, pp. 37–49). https://doi.org/10.1038/nrc2294

Cunningham, J. P., & Ghahramani, Z. (2015). Linear Dimensionality Reduction: Survey, Insights, and Generalizations. In *Journal of Machine Learning Research* (Vol. 16).

De Luca, A. (2010). *Le applicazioni dei metodi statistici alle analisi di mercato. Manuale di ricerche di mercato*. FrancoAngeli.

De Luca, A. (2016). *Modelli di marketing: statistica per le analisi di mercato: segmentazione, posizionamento, comunicazione, innovazione, customer satisfaction*. FrancoAngeli.

Deng, L. Y., Garzon, M., & Kumar, N. (2022). What is dimensionality reduction (dr)? In *Dimensionality Reduction in Data Science* (pp. 67–77). Springer International Publishing. https://doi.org/10.1007/978-3-031-05371-9_3

Fabrigar, L. R., Wegener, D. T., Maccallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. In *Psychological Methods* (Vol. 4, Issue 3).

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press Professional, Inc., San Diego, CA, USA.

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex and Intelligent Systems*, *8*(3), 2663–2693. https://doi.org/10.1007/s40747-021-00637-x

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Kenett, R. S., Redman, T. C., Manzi, G., & Salini, S. (2021). *La data science nella realtà: come trasformare i dati in informazioni, decisioni migliori e organizzazioni più forti*. Giappichelli.

Li, L. (2010). Dimension reduction for high-dimensional data. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 620, pp. 417–434). https://doi.org/10.1007/978-1-60761-580-4_14

Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007). Nonlinear Principal Components Analysis: Introduction and Application. *Psychological Methods*, *12*(3), 336–358. https://doi.org/10.1037/1082-989X.12.3.336

Mainali, S., Garzon, M., Venugopal, D., Jana, K., Yang, C. C., Kumar, N., Bowman, D., & Deng, L. Y. (2021). An Information-theoretic approach to dimensionality reduction in data science. *International Journal of Data Science and Analytics*, *12*(3), 185–203. https://doi.org/10.1007/s41060-021-00272-2

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Nanaware, T., Mahajan, P., Chandak, R., Deshpande, P., & Patil, M. (2018). Exploratory Data Analysis Using Dimension Reduction. In *IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org ISSN* (Vol. 2). www.iosrjen.org

Oskolkov, N. (2022). Dimensionality Reduction: Overview, Technical Details, and Some Applications. In *Tourism on the Verge: Vol. Part F1051* (pp. 151–167). Springer Nature. https://doi.org/10.1007/978-3-030-88389-8_9

Osservatorio Big Data & Analytics. (2022). *Data-driven culture: connettere algoritmi e persone*.

Ringnér, M. (2008). What is principal component analysis? In *NATURE BIOTECHNOLOGY* (Vol. 26). http://www.nature.com/naturebiotechnology

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., Xu, Y., & Fertig, E. J. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. In *Trends in Genetics* (Vol. 34, Issue 10, pp. 790–805). Elsevier Ltd. https://doi.org/10.1016/j.tig.2018.07.003

The Economist. (2017, May). *The World's Most Valuable Resource Is No Longer Oil, but Data*.

Venkat, N. (2018). *The Curse of Dimensionality: Inside Out*.

SITOGRAFIA

IBM, *Cos'è la data science?*
https://www.ibm.com/it-it/topics/data-science

Osservatorio Big Data & Analytics, *Il mercato italiano dei Big Data vale 2,41 miliardi di euro, +20%*
https://www.osservatori.net/it/ricerche/comunicati-stampa//mercato-big-data-crescita

Market Data Forecast, *Big Data Market*
https://www.marketdataforecast.com/market-reports/big-data-market

Statista, *Big data market size revenue forecast worldwide from 2011 to 2027*
https://www.statista.com/statistics/254266/global-big-data-market-forecast/

Medium, *Data Science Interview Entry Level Questions*
https://medium.com/@aakriti.sharma18/data-science-interview-entry-level-questions-b6f436759c98

The University of Sidney, *Principal Components Analysis*
https://bookdown.org/tpinto_home/Unsupervised-learning/principal-components-analysis.html

Medium, *Principal Component Analysis (PCA) – Part 1 – Fundamentals and Applications*
https://medium.com/analytics-vidhya/principal-component-analysis-pca-part-1-fundamentals-and-applications-8a9fd9de7596

GraphPad, *Principal Component Analysis*
https://www.graphpad.com/guides/prism/latest/statistics/stat_pca_example_loadings_plot.htm

ResearchGate, *Maturity Stage Categorization of Endemic Lizard (Calotes nigrilabris) in the Grasslands of HPNP*
https://www.researchgate.net/publication/320280481_Maturity_Stage_Categorization_of_Endemic_Lizard_Calotes_nigrilabris_in_the_Grasslands_of_HPNP

BODaI-Lab (Big and Open Data Innovation Laboratory dell'Università degli Studi di Brescia), *Data Science for Brescia – Arts and Cultural Places*
https://bodai.unibs.it/ds4bs/

RINGRAZIAMENTI

Scrivere i ringraziamenti al termine di questa tesi di Laurea Magistrale è un po' come chiudere un cerchio, prendendo concretamente consapevolezza della fine del mio percorso di studi universitari e l'inizio di un qualcosa di nuovo. Sono stati anni importanti, sfidanti, ma soprattutto formativi dal punto di vista professionale e personale. Raggiungere questo traguardo non sarebbe stato possibile senza il prezioso aiuto di alcune persone, a cui va il mio più sentito ringraziamento.

Desidero innanzitutto ringraziare la Professoressa Paola Zuccolotto, per il sostegno, la disponibilità e la gentilezza che mi ha riservato durante tutto il periodo di stesura della tesi. Grazie per avermi guidato in tutte le fasi della ricerca e per avermi dato la possibilità di affrontare una realtà del tutto nuova per me, quale è la realizzazione di una tesi sperimentale. Oltre a essere una professionista estremamente competente, la Professoressa Zuccolotto ha dimostrato particolare sensibilità e comprensione, doti che ho veramente apprezzato.
Ringrazio altresì la Professoressa Marica Manisera, per la cortesia e l'importante contributo che ha offerto nell'ambito del progetto di ricerca.

Un immenso grazie va alla mia famiglia, in particolar modo a mia mamma, mio papà e mia sorella, per il costante supporto in questo percorso di Laurea Magistrale. Grazie per avermi spronato a dare sempre il meglio, specialmente nei momenti per me più impegnativi e moralmente difficili, e per avermi dato la possibilità di conseguire questo titolo. Questo risultato è anche merito vostro.

Vorrei inoltre ringraziare le mie amiche, in particolare Clarissa e Sofia: ci conosciamo oramai da anni, stiamo vivendo fasi differenti nelle nostre vite, ma mai come in quest'ultimo periodo ho sentito la vostra vicinanza e il vostro affetto. Grazie per esserci sempre state, e per avermi ascoltato e incoraggiato quando ne avevo più bisogno.

Infine, ma non meno importanti, desidero ringraziare le mie compagne di studio, che in quest'ultimo anno sono state fondamentali: grazie per aver condiviso con me gioie, risate, ma anche preoccupazioni e paure. Un pensiero speciale va a Nicole, che mi ha

accolto e supportato in un momento difficile, e insieme abbiamo navigato il mio ultimo anno accademico: grazie per la tua sensibilità e amicizia.

Sono grata per aver compiuto questo percorso, che è stato stimolante e ricco di sfide, ma soprattutto mi ha regalato molte soddisfazioni, che finalmente oggi vorrei celebrare con voi.

Grazie.