Università della Calabria
Metodi Statistici per le Strategie Aziendali

CdLM in Data Science per le Strategie Aziendali

Relazione di laboratorio

1° Assignment

19/11/2025

Gruppo di lavoro:
Cristian Tedesco, tdscst02t24h579i@studenti.unical.it
Matricola 269279

Francesco De Nisi, dnsfnc00r09m208v@studenti.unical.it
Matricola 269762

Angela Karin Mancuso, mncnlk98m64i874m@studenti.unical.it
Matricola 268839

Nabil Larhram, lrhnbl98a11z330z@studenti.unical.it
Matricola 269686

Saveria Falvo, flvsvr02p64d086x@studenti.unical.it
Matricola 264008

Pierfrancesco Lindia, lndpfr00e20d086f@studenti.unical.it
Matricola 256641

Anno Accademico 2025-2026

Indice

1 Introduzione **2**

2 Analisi Descrittive **2**
2.1 Distribuzione delle Osservazioni 2
2.2 Distribuzione delle variabili di Composizione. 3

3 Analisi MDS Metrico del Dataset Wines **8**
3.1 Principi Teorici del Multidimensional Scaling (MDS) Metrico . . 8
3.2 Preparazione dei Dati 9
3.2.1 Calcolo della Matrice Q 10
3.3 Autovalori e Varianza Spiegata 10
3.3.1 Scree Plot degli Autovalori 11
3.4 Applicazione dell'MDS in 2 Dimensioni 11
3.4.1 Mappa Percettiva MDS 2D 12
3.5 Analisi della Varianza (ANOVA) delle Classi 14
3.6 Validità del Modello Metrico sui Centroidi e Analisi della Complessità Reale 15

4 Principal Component Analysis (PCA) **16**
4.1 Principi teorici della PCA 16
4.1.1 Il Biplot 17
4.2 Applicazione e Interpretazione della PCA 18

5 Confronto MDS e PCA **20**
5.1 Confronto Visivo delle Mappe Percettive 20
5.2 Confronto Numerico delle Coordinate e delle Performance 22
5.2.1 Analisi delle Performance 22

6 Analisi delle Corrispondenze **23**
6.1 La Distanza del Chi-Quadrato 24
6.2 Risultati dell'Analisi delle Corrispondenze 24
6.2.1 Analisi dei Profili Riga 24
6.2.2 Analisi dei Profili Colonna 26
6.3 Grafici 27
6.3.1 La Mappa Rowprincipal: Profili Riga in Coordinate Principali 27
6.3.2 La Mappa Colprincipal: Profili Colonna in Coordinate Principali 28
6.4 Digressione Teorica 28
6.4.1 La Mappa Simmetrica: Migliore Rappresentazione Simultanea 29
6.4.2 Qualità di Rappresentazione e Squared Correlations . . . 31

7 Conclusioni **31**

1 Introduzione

Il presente progetto affronta un'analisi multivariata su un dataset composto da 178 campioni di vino, ripartiti in 5 classi definite in base a caratteristiche chimico-composizionali. Gli obiettivi sono due:

- Rappresentare graficamente il grado di similarità/dissimilarità tra le classi di vino utilizzando esclusivamente le variabili di composizione;
- Indagare se specifiche caratteristiche dei sommelier (genere e fascia d'età) possano essere associate al consumo di particolari categorie di vino.

Il dataset comprende 19 variabili.

La Colonna 1 riporta l' identificativo della classe di appartenenza del vino.

Le colonne dalla 2 alla 14 contengono variabili chimiche specifiche dei vini (Alcohol, Malic.acid, Ash, Alcalinity.of.ash, Magnesium, Total.phenols, Flavanoids, Nonflavanoid.phenols, Proanthocyanins, Color.intensity, Hue, OD280/OD315 (dei vini diluiti), Proline). Le Colonne dalla 15 alla 19 presentano caratteristiche anagrafiche/descrittive dei sommelier (F, M, Eta18_25, Eta25_40, Eta_sup40).

Nei paragrafi successivi si costruisce la matrice di dissimilarità sulle variabili di composizione, si applica il Multidimensional Scaling (MDS) Metrico e si discutono dei risultati dell'analisi e della mappa percettiva, includendo un confronto mirato con l'Analisi in Componenti Principali (PCA) sui centroidi delle classi.

Dapprima viene condotta un'analisi descrittiva per delineare la struttura del dataset, individuare eventuali outlier e ottenere una visione d'insieme delle distribuzioni e delle relazioni tra variabili.

Nei paragrafi successivi si costruisce la matrice di dissimilarità sulle variabili di composizione, si applica il Multidimensional Scaling (MDS) Metrico e si discutono dei risultati dell'analisi e della mappa percettiva, includendo un confronto mirato con l'Analisi in Componenti Principali (PCA) sui centroidi delle classi.

Successivamente, si applica l'Analisi delle Corrispondenze (AC) per rappresentare simultaneamente le relazioni tra le classi di vino e le caratteristiche dei sommelier, si discutono i risultati e si riporta la mappa simultanea.

Dall'analisi condotta emergono risultati interessanti: l'MDS e la PCA restituiscono una struttura delle classi di vino coerente, distinguendo nettamente i prodotti tradizionali da quelli più innovativi. L'Analisi delle Corrispondenze evidenzia associazioni preferenziali per alcune classi di vino e segmenti di sommelier suddivisi per genere ed età.

2 Analisi Descrittive

2.1 Distribuzione delle Osservazioni

Il grafico riportato di seguito (Figura 1) mostra la distribuzione delle osservazioni all'interno delle cinque classi di vino presenti nel dataset, che fornisce una panoramica sulla composizione del campione in termini di frequenza per classe.

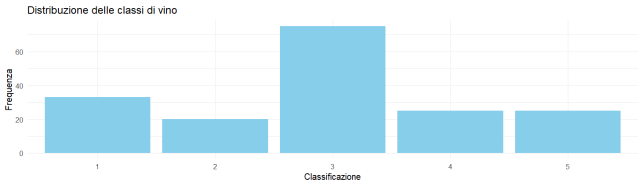


Figura 1: Diagramma a barre delle variabile *Classi*

La distribuzione delle 178 osservazioni non è uniforme: la classe 3 concentra la quota più ampia del campione, mentre le restanti quattro classi presentano numerosità sensibilmente inferiori e tra loro più ravvicinate. In particolare, la Classe 3 risulta essere il gruppo dominante, che potrebbe esercitare un'influenza significativa sulle statistiche descrittive globali (Media e Varianze complessive) e sulle relazioni tra variabili, specie laddove fossero presenti sottostrutture interne o valori anomali. Le Classi 4 e 5 appaiono pressoché equivalenti per frequenza, suggerendo una possibile prossimità anche dal punto di vista composizionale, mentre la Classe 2 rappresenta il segmento più contenuto del dataset; la Classe 1 si colloca in una posizione intermedia.

2.2 Distribuzione delle variabili di Composizione.

Si analizzano le variabili che descrivono la composizione chimica del vino, ovvero le colonne dalla 2 alla 14 del dataset. A tale scopo, si fa ricorso a un set di istogrammi che consente di osservare la forma delle distribuzioni, identificare eventuali asimmetrie, code lunghe e potenziali outliers. I grafici affiancano le statistiche di sintesi ottenute attraverso il comando *summary()* in R.

Dall'analisi congiunta emergono alcune considerazioni di rilievo:

* Diverse variabili come Alcohol, Ash, Hue, Nonflavanoid.phenols e Total.phenols, presentano distribuzione relativamente simmetriche o debolmente asimmetriche, con range contenuti e assenza di outlier marcati.
* Al contrario, Color.intensity, Proline e, in misura minore, Magnesium, mostrano una distribuzione fortemente asimmetrica a destra, con code lunghe e presenza di osservazioni isolate su valori elevati.
* Variabili come Malic.acid, Flavanoids, Proanthocyanins e OD280/OD315 of diluted wines rivelano una dispersione maggiore, con frequenze distribuite in modo non uniforme su un intervallo più ampio. Questo suggerisce una certa eterogeneità tra i campioni, confermata poi dalla variabilità intra-classe.
* La variabile Alcalinity of ash si distingue per una forma bimodale/multimodale, suggerendo la possibile presenza di sottogruppi all'interno di alcune classi.

In generale, si osserva che le variabili di composizione si distribuiscono su scale molto diverse, alcune con valori piccoli (es. Nonflavanoid.phenols da 0.2 a 0.6), altre con intervalli più ampi (es. Magnesium da 70 a 160, Proline da 270 a oltre 1500). Questo conferma la necessità della standardizzazione prima di calcolare distanze o applicare tecniche multivariate, al fine di evitare che le variabili con

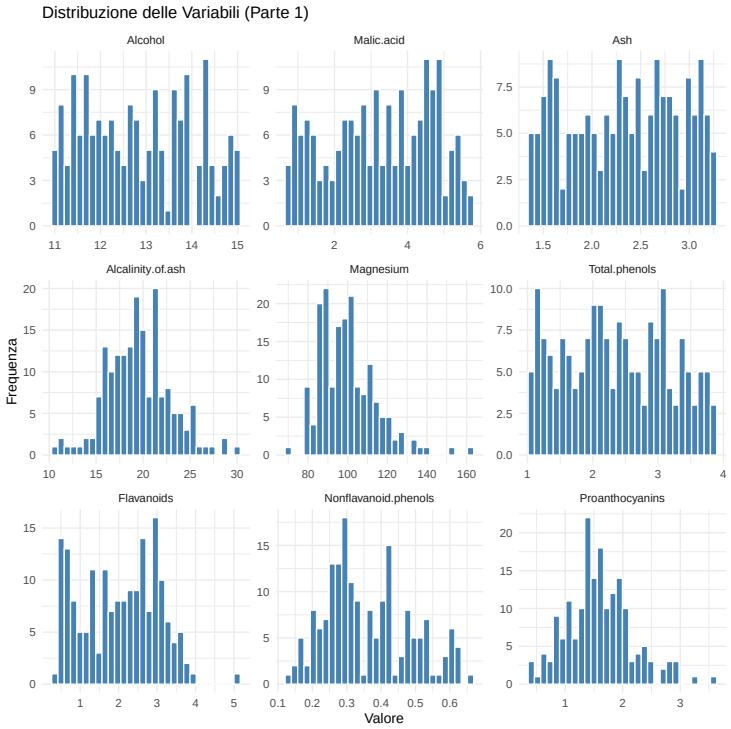


Figura 2: Istogramma delle variabili

range più esteso dominino le metriche di dissimilarità. In sintesi, questa analisi preliminare ha messo in luce la presenza di variabili fortemente eterogenee in termini di forma, dispersione e scala di misura, giustificando l'adozione di approcci robusti e tecniche di riduzione dimensionale nella fase successiva dell'elaborazione. Dopo aver osservato la forma delle distribuzioni tramite istogrammi possiamo ora raffinare l'analisi esaminando le statistiche descrittive sintetiche come minimo, quartili, mediana, media, massimo delle 13 variabili di composizione. La seguente tabella consente di quantificare, con maggiore precisione, le impressioni emerse visivamente e di individuare eventuali anomalie o squilibri che meritano attenzione nelle fasi successive.

VARIABLE	MIN	1ST QU.	MEDIAN	MEAN	3ST QU.	MAX
Alcohol	11.00	11.78	12.77	12.86	13.84	14.99
Malic.acid	0.73	2.13	3.35	3.25	4.50	5.73
Ash	1.38	1.83	2.35	2.33	2.79	3.25
Alcalinity.of.ash	10.60	17.20	19.50	19.49	21.50	30.00
Magnesium	70.00	88.00	98.00	99.74	107.00	162.00
Total.phenols	1.06	1.68	2.32	2.37	3.04	3.85

Tabella 1: Statistiche descrittive (prima parte).

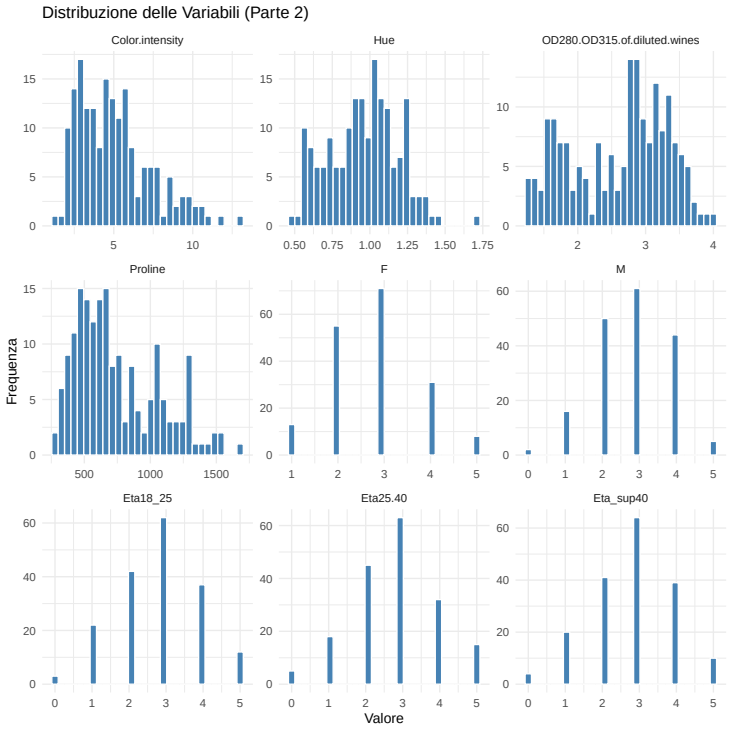


Figura 3: Istogramma delle variabili

VARIABLE	MIN	1ST QU.	MEDIAN	MEAN	3ST QU.	MAX
Flavanoids	0.34	1.21	2.14	2.03	2.88	5.08
Nonflavanoid.phenol	0.13	0.27	0.34	0.36	0.44	0.66
Proanthocyanins	0.41	1.25	1.56	1.59	1.95	3.58
Color.intensity	1.28	3.22	4.69	5.06	6.20	13.00
Hue	0.48	0.78	0.97	0.96	1.12	1.71
OD280.OD315	1.27	1.94	2.78	2.61	3.17	4.00
Proline	278.00	500.50	673.50	746.89	985.00	1680.00

Tabella 2: Statistiche descrittive (seconda parte).

Le informazioni riportate confermano alcune evidenze già rilevate nei grafici: Color.intensity e Proline presentano un range estremamente ampio: la prima passa da 1,28 a 13,00, mentre la seconda va da 278 a 1680. Entrambe mostrano valori massimi molto distanti dai rispettivi terzi quartili (6,20 e 985), segnalando la probabile presenza di outlier a destra. Questi valori elevati, se non correttamente gestiti, possono influenzare indebitamente la stima di distanze tra osservazioni, rendendo ancor più necessaria la standardizzazione preliminare. Le variabili Flavanoids, Total.phenols, OD280/OD315, e Hue risultano ben distribuite, con medie e mediane vicine e differenze contenute tra quartili. Questo indica una certa stabilità e regolarità nella distribuzione, utile per analisi comparative tra classi. Al contrario, Magnesium, Alcalinity.of.ash, e Malic.acid presentano una dispersione marcata, con differenze visibili tra i quartili e un ampio intervallo interquartile. In particolare, Magnesium spazia da 70 a 162 con una media elevata (99,7), riflettendo una variabilità che può tradursi in un

impatto rilevante nella determinazione delle dissimilarità se non normalizzato. Alcune variabili, come Nonflavanoid.phenols e Ash, si sviluppano su un range molto ristretto, suggerendo un'informazione meno discriminante, ma comunque potenzialmente utile nella definizione del profilo composizionale medio delle classi. Nel complesso, questa analisi numerica rafforza la motivazione alla standardizzazione delle variabili prima dell'applicazione di tecniche di tipo metrico come il Multidimensional Scaling. La presenza di variabili su scale differenti, con dispersioni e asimmetrie marcate, richiede di riportare tutte le misure a una scala comune per evitare che alcune dimensioni dominino la misura complessiva di dissimilarità. Dopo aver analizzato le distribuzioni univariate e le statistiche descrittive, proseguiamo con l'analisi delle relazioni tra le variabili, per comprendere se esistono strutture interne di dipendenze o ridondanza. La matrice

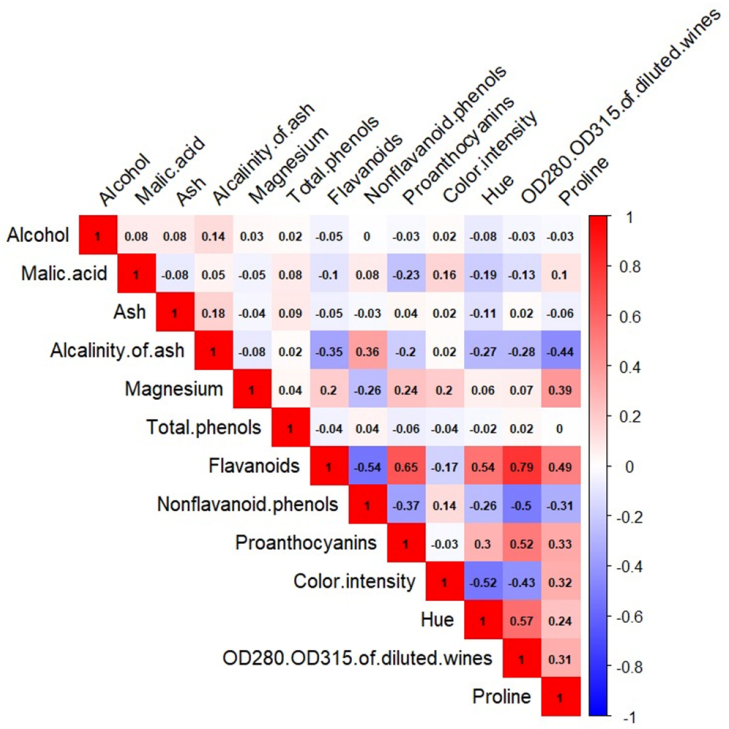


Figura 4: Heat Map

di correlazione riportata in Fig. 4 mostra le correlazioni di Pearson calcolare tra le 13 variabili composizionali. I valori sono codificati graficamente secondo una scala cromatica che varia dal blu (correlazioni negative) al rosso (correlazioni positive), con l'intensità proporzionale alla forza della relazione. Un primo blocco fortemente coeso è quello formato da Flavanoids, OD280/OD315, Total.phenols e Hue: queste variabili mostrano tra loro correlazioni elevate e positive, fino a 0.79. Si tratta di caratteristiche spesso associate alla qualità fenolica e all'assorbimento ottico del vino, ed è plausibile che rappresentino una dimensione comune nella descrizione del profilo composizionale.

Al contrario, Flavanoids è negativamente correlata con Nonflavanoid.phenols (-0.54), suggerendo una compensazione tra i due tipi di composti: vini con alta

concentrazione di flavonoidi tendono ad avere livelli più bassi di non−flavonoidi, e viceversa. Anche Color.intensity mostra una correlazione inversa con Hue (−0.52), confermando un'opposizione strutturale tra queste due misure ottiche. Proline presenta correlazioni moderate con Flavanoids (0.49), Color.intensity (0.32) e Magnesium (0.39), indicando una certa connessione con la struttura fenolica e minerale, ma senza appartenere a un blocco specifico. Variabili come Alcohol, Malic.acid e Ash risultano debolmente correlate con quasi tutte le altre (valori tra −0.1 e +0.2), il che indica una maggiore indipendenza statistica. Magnesium e Alcalinity.of.ash mostrano correlazioni significative solo in parte (es. −0.26 con Flavanoids, +0.36 con Hue), suggerendo una posizione intermedia tra i gruppi composizionali. L'osservazione di forti correlazioni tra alcune variabili conferma la presenza di strutture latenti nel dataset: dimensioni comuni che riassumono informazioni ridondanti. Questo rafforza la scelta di applicare metodi riduzione dimensionale, poiché è verosimile che una buona parte della variabilità tra le classi possa essere spiegata da pochi assi informativi ben costruiti. Inoltre, la presenza di correlazioni negative strutturare (es. Flavanoids vs Nonflavanoid.phenols) suggerisce che alcune opposizioni tra classi potrebbe emergere chiaramente nella mappa MDS.

Successivamente, è presente il Boxplot generale standardizzato di tutto il dataset: esso offre una panoramica complessiva delle distribuzioni delle variabili di composizione, senza distinzioni di classe, permettendo di valutare simmetria, dispersione e presenza di outliers in ciascuna variabile.

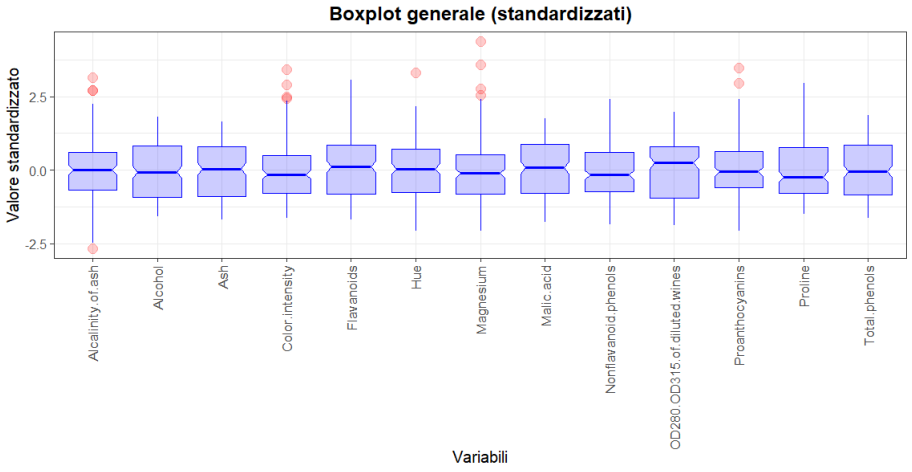


Figura 5: Box Plot

Le informazioni che emergono dal grafico sono:

- Simmetria: La maggior parte delle variabili mostra una distribuzione tendenzialmente simmetrica, come Alcalinity, Ash, Total.phenols. Alcune variabili, invece, risultano più asimmetriche, come OD280/OD315.of.diluted .wines, Nonflavanoid.phenols e Proline indicando una maggiore concentrazione dei valori verso una coda.

- Outliers: Sono presenti alcuni outliers visibili in variabili come Alcalinity.of.ash, Color.intensity, Hue, Magnesium e Proanthocyanins. Questi valori estremi si collocano oltre i limiti del box, evidenziando osservazioni potenzialmente atipiche.

La presenza di outliers suggerisce una certa eterogeneità interna al dataset, attribuibile sia alla naturale variabilità tra i vini (ad esempio nei contenuti di Magnesium o Color.intensity), sia a possibili sottogruppi latenti non esplicitamente rappresentati nelle classi considerate.

3 Analisi MDS Metrico del Dataset Wines

3.1 Principi Teorici del Multidimensional Scaling (MDS) Metrico

Il **Multidimensional Scaling (MDS)** è una tecnica utilizzata per l'analisi del posizionamento dal punto di vista statistico. Il posizionamento è l'attività di marketing orientata a creare un'immagine distintiva di un prodotto/servizio nella mente del consumatore. Tale strategia consente di determinare il corretto posizionamento di un prodotto nel mercato, individuando le caratteristiche d'interesse dei consumatori al fine di comunicarle per determinare un vantaggio competitivo. Inoltre, suggerisce eventuali riposizionamenti di prodotti esistenti e posizionamenti di nuovi prodotti. Le indagini condotte sui consumatori si concretizzano attraverso le mappe percettive, che dal punto di vista statistico vengono realizzate da tecniche come l'MDS.

L'obiettivo dell'MDS è, dunque, ottenere una rappresentazione geometrica delle unità statistiche, in uno spazio di dimensioni ridotto, che rifletta le dissimilarità tra gli oggetti, preservando il più possibile le distanze originali tra gli oggetti nella matrice di prossimità o distanza di partenza, ricavata dai dati originali.

Il modello Metrico presuppone che le misure di prossimità siano espresse almeno su scala a intervalli, la soluzione analitica è garantita dal teorema di Torgerson, che postula l'uguaglianza tra misure di prossimità e distanze $d_{ij} = f(\delta_{ij})$. La soluzione dell'MDS Metrico è data dalla matrice di coordinate delle n unità statistiche in p dimensioni (ottenuta a partire da una matrice di distanze euclidee), da cui si ricava la mappa percettiva che rappresenta le dissimilarità tra gli oggetti.

Nel caso specifico studiato, l'MDS consente di rappresentare, in uno spazio bidimensionale, le similarità/dissimilarità tra i profili chimici medi delle classi di vino. Partendo dalle 13 caratteristiche della variabile composizione di ciascuna classe, l'MDS calcola le distanze euclidee tra i centroidi (la media di ogni caratteristica di composizione per ognuna delle 5 classi di vino) e le proietta su un piano 2D, preservando il più fedelmente possibile le distanze originali. Questo permette di visualizzare quali classi sono chimicamente simili e quali diverse, orientando così strategie di posizionamento di mercato.

3.2 Preparazione dei Dati

Dapprima, l'MDS metrico è stato eseguito su tutte le osservazioni delle 5 classi, ma il modello risultante non è utilizzabile; in quanto, per ottenere una *discreta* bontà di adattamento occorre realizzare una rappresentazione geometrica dei vini in almeno 5 dimensioni, il che non è agevole. Dunque, vista la natura dei dati, si è deciso di considerare la media di ogni caratteristica di composizione (colonne 2-14) per ognuna delle 5 classi di vino, per applicare l'MDS ed ottenere la rappresentazione grafica degli oggetti. I centroidi, dunque, rappresentano il profilo chimico medio di ogni classe. Ad esempio, la Classe 1 ha media di *Flavanoids* = 2.895 (elevata complessità fenolica), mentre la Classe 4 ha *Flavanoids* = 0.934 (profilo semplice). Allo stesso modo, l'Alcalinity.of.ash varia da 17.5 (Classe 1) a 21.4 (Classe 4).

```
centroidi <- aggregate(composizione, by = list(classi), FUN
    = mean)
rownames(centroidi) <- paste("Classe", centroidi$Group.1)
```

Successivamente, per eliminare effetti di scala dovuti all'uso di unità di misura diverse, si è deciso di standardizzare i centroidi.

```
centroidi_scaled <- scale(centroidi[, -1])
```

Si calcola, dunque, la matrice delle distanze euclidee tra i 5 oggetti, i centroidi delle classi, per ottenere una matrice di dimensione (5×5) che contiene le distanze delle composizioni chimiche tra le classi. Le distanze sono state calcolate usando la *distanza euclidea* sui centroidi standardizzati:

$$d_{ij} = \left[\sum_{s=1}^{p} (x_{is} - x_{js})^2 \right]^{1/2} = \|x_i - x_j\| \tag{1}$$

dove x_{is} è il valore standardizzato della variabile s per il centroide della classe i. La matrice di distanze ottenuta sarà:

	Classe 1	**Classe 2**	**Classe 3**	**Classe 4**	**Classe 5**
Classe 1					
Classe 2	3.386293				
Classe 3	4.076366	4.777084			
Classe 4	5.996101	6.196618	4.551434		
Classe 5	6.163789	6.535622	5.149313	2.608987	

Tabella 3: Matrice delle Distanze Euclidee tra i Centroidi

La distanza minima rilevata è quella tra le Classi 4 e 5, che presentano profili chimici molto simili, mentre quella massima è quella tra le Classi 2 e 5, che sono chimicamente molto diverse.

3.2.1 Calcolo della Matrice Q

Q è una matrice simmetrica, di dimensione $(n \times n)$, che contiene i prodotti scalari tra le coordinate MDS. Gli elementi di Q sono del tipo:

$$q_{ij} = \sum_{s=1}^{p} x_{is} x_{js} \tag{2}$$

Ma una volta ottenuta la matrice delle distanze euclidee:

$$d_{ij}^2 = q_{ii} - q_{jj} + 2q_{ij} \tag{3}$$

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + \frac{2B}{n}\right) = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d^2) \tag{4}$$

con $B = tr(Q)$.

```
Q <- coord %*% t(coord)
```

	Classe 1	**Classe 2**	**Classe 3**	**Classe 4**	**Classe 5**
Classe 1	6.963175	7.778244	1.319173	-7.572107	-8.488486
Classe 2	7.778244	11.128085	-2.409486	-8.266774	-8.230070
Classe 3	1.319173	-2.409486	-6.431139	-1.739658	-3.601169
Classe 4	-7.572107	-8.266774	-1.739658	8.249351	9.329187
Classe 5	-8.488486	-8.230070	-3.601169	9.329187	10.990537

Tabella 4: Matrice Q dei Prodotti Scalari

Valori positivi elevati indicano vini posizionati nello stesso verso dello spazio (simili), mentre valori negativi indicano posizionamenti opposti (dissimili). Ad esempio, $Q(1,2) = 7.778$ (positivo e alto) riflette la similarità tra Classi 1-2, mentre $Q(1,4) = -7.572$ (negativo e basso) riflette la loro dissimilarità.

3.3 Autovalori e Varianza Spiegata

Dalla decomposizione spettrale di Q si ottengono le matrici $\Lambda(p \times p)$ e $A(n \times p)$ contenenti rispettivamente gli autovalori e gli autovettori di Q.

$$Q = A\Lambda A' \tag{5}$$

Gli autovalori calcolati dalla funzione *cmdscale* sono:

```
> (var_exp <- eig_pos / sum(eig_pos))
```

Dimensione	***Autovalore***	***Varianza Spiegata***	***Varianza Cumulata***
1	*34.49*	*66.33%*	*66.33%*
2	*9.27*	*17.83%*	*84.16%*
3	*5.44*	*10.46%*	*94.62%*
4	*2.80*	*5.38%*	*100.00%*
5	≈ 0	*0%*	*100.00%*

Tabella 5: Autovalori e Varianza Spiegata

Non è noto a priori se Q sia semi-definita positiva (non contenga autovalori negativi). Poiché solo l'ultimo autovalore è negativo, esso è trascurabile, e la soluzione MDS è valida.

3.3.1 Scree Plot degli Autovalori

Sulla base dei valori ottenuti per gli autovalori si può costruire uno screeplot con il seguente comando:

```
plot(seq_along(eig_pos), eig_pos, type = "b", pch = 19,
     xlab = "Dimensione", ylab = "Autovalore",
     main = "Scree_plot_autovalori_(Classical_MDS)",
     col="lightblue")
```

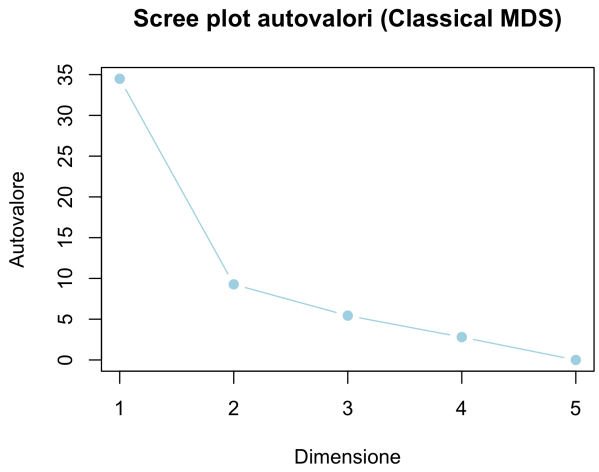


Figura 6: Caption

L'asse X rappresenta il numero di dimensioni e l'asse Y rappresenta il valore degli autovalori. Lo scree plot mostra un calo marcato dall'autovalore 1 (34.49) all'autovalore 2 (9.27), seguito da un decremento più graduale dalle dimensioni 3 e 4. Il "gomito" è evidente tra la dimensione 2 e 3, suggerendo che due dimensioni sono sufficienti per catturare la struttura principale dei dati (84.2% della varianza). L'aggiunta di una terza dimensione porterebbe solo il 10.46% di varianza aggiuntiva, non compensando il costo di complessità interpretativa. Dunque, si considerano soltanto le prime due componenti (autovalori).

3.4 Applicazione dell'MDS in 2 Dimensioni

L'obiettivo finale dei modelli MDS è la rappresentazione in un numero ridotto di dimensioni. Per tale motivo, si utilizza la funzione cmdscale, che prende in input la matrice di distanze, considerando uno spazio a $k = 2$ dimensioni, calcola gli autovalori e restituisce la matrice delle coordinate dei 5 centroidi. Se Q è semi definita positiva allora avrà autovalori non negativi:

$$Q = A\Lambda A' = A\Lambda^{1/2}\Lambda^{1/2}A' = XX' \tag{6}$$

Si ottiene cioè la matrice delle coordinate X per la rappresentazione delle n unità statistiche nello spazio a p dimensioni:

$$X = A\Lambda^{1/2} \tag{7}$$

```
mds_fit <- cmdscale(dist_matrix, k = 2, eig = TRUE)
```

Classe	Dimensione 1	Dimensione 2
Classe 1	-2.638	-0.045
Classe 2	-2.974	+1.511
Classe 3	-0.457	-2.494
Classe 4	+2.867	+0.172
Classe 5	+3.203	+0.856

Tabella 6: Coordinate delle Classi nello Spazio MDS 2D

Questa matrice permette di visualizzare il grafico in cui la distanza tra i punti riflette la distanza originale tra i profili chimici delle classi di vino. La qualità della riduzione dimensionale è quantificata dal **Goodness of Fit (GoF)**:

$$G = \frac{\sum_{i=1}^{2} \lambda_1 + \lambda_2}{\sum_{i=1}^{5} |\lambda_i|} = \frac{34.49 + 9.27}{34.49 + 9.27 + 5.44 + 2.80} = 0.842 \tag{8}$$

Cioè:

```
gof <- (mds_fit$eig[1] + mds_fit$eig[2]) / sum(mds_fit$eig[
    mds_fit$eig > 0])
```

Il valore indica che il modello in 2D replica l'84.2% della struttura delle distanze originali; ovvero spiega l'84.2% della variabilità totale. Un valore > 0.80 è considerato eccellente e garantisce la fedeltà della rappresentazione grafica alle distanze originali tra i centroidi.

3.4.1 Mappa Percettiva MDS 2D

Riprendendo le coordinate dei 5 centroidi ottenute precedentemente, si visualizza una rappresentazione geometrica dei punti bidimensionale. La mappa è stata costruita con il seguente comando:

```
plot(coord[, 1], coord[, 2], type = "n", asp = 1,
     main = "MDS_Metrico_delle_Classi_di_Vino_(basato_sulla_
         Composizione)",
     xlab = "Dimensione_1", ylab = "Dimensione_2")
text(coord[, 1], coord[, 2], labels = rownames(centroidi),
    col = "blue")
abline(v=0, h=0)
```

Analizzando la mappa MDS ottenuta dai dati chimici delle cinque classi di vino e guardando le relative statistiche descrittive, si osserva come le diverse tipologie di vino si distribuiscano secondo tre grandi segmenti che trovano, ipoteticamente, un riscontro, sia dal punto di vista chimico sia da quello del mercato. Le classi 1 e 2, che appaiono raggruppate sulla sinistra della mappa, rappresentano il mondo dei vini tradizionali, quelli più ricchi dal punto di vista fenolico e strutturale. Infatti, la media dei *Flavanoids* nella classe 1 è circa 2.90 e addirittura 3.10 nella classe 2, valori molto alti se confrontati col resto delle classi. Lo stesso vale per i valori di OD280/OD315 (circa 3.2 e 3.1 rispettivamente), che confermano il profilo fenolico accentuato di questi vini. Anche la Proline,

che riflette la struttura proteica e la robustezza del vino, raggiunge qui i suoi massimi: media di 1032 per la classe 1 e ben 1235 per la classe 2. Questi vini si collocano, non a caso, nel segmento "Heritage": quello dei vini per appassionati ed intenditori, pensati per chi ama la tradizione e cerca prodotti complessi. Il loro mercato ideale è quello premium, tra collezionisti, sommelier e clienti dei ristoranti stellati, mentre la comunicazione punta sull'autenticità, i metodi tradizionali e una lunga storia di produzione.

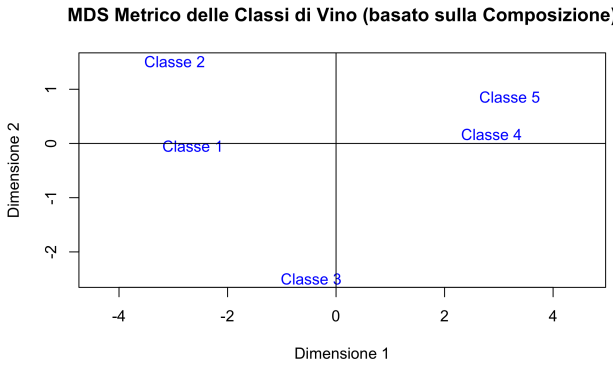


Figura 7: Caption

All'estremo opposto, sulla destra della mappa MDS, si ritrovano le classi 4 e 5, molto vicine anche nelle distanze numeriche (con una distanza di 2.61). Questi vini si distinguono soprattutto per la loro elevata alcalinità (media circa 21.4 in classe 4 e 21.6 in classe 5), ma anche per la bassissima presenza di flavonoidi (0.93 e 0.73 rispettivamente, i valori più bassi tra tutti). L'intensità di colore (8.6 nella classe 5) si alza molto, mentre la tonalità (Hue) scende fino a 0.66–0.71. Dal punto di vista del mercato, questi vini rappresentano l'innovazione: vini "moderni", con un profilo chimico diverso da quello classico, adatti a un pubblico giovane e attento alle novità, dai millennial ai frequentatori di wine bar ed enoteche digitali. Qui contano molto brand caratterizzati dalla grafica contemporanea, conta il racconto legato a tecniche di produzione sostenibile o biologica, e una comunicazione via social e che sfrutta gli influencer. Proprio perché la loro offerta rischia di assomigliarsi troppo, è importante differenziarsi bene su prezzo, packaging e canali distributivi per evitare che i prodotti di queste due classi si facciano troppa concorrenza tra loro.

La classe 3 occupa una zona del tutto particolare sulla mappa, più decentrata in basso. Dal punto di vista numerico, i suoi valori medi di *Flavanoids* (2.16), Hue (1.06) e OD280/OD315 (2.80) sono inferiori rispetto alle classi 1 e 2, e la Proline (573.9) è la più bassa in assoluto. Si tratta dunque di vini meno strutturati e meno "potenti", ma anche qui emerge una peculiarità interessante: questi prodotti sfuggono alle logiche dei due grossi cluster descritti prima, offrendo una proposta di equilibrio, moderazione e con molta variabilità interna. Non a caso nella classe 3 si trovano outlier (vini "anomali" rispetto al profilo medio), perché essendo la classe più numerosa (75 campioni), accoglie tanti stili e influenze produttive diverse. Questo segmento è perfetto per un posiziona-

mento più mainstream e sofisticato allo stesso tempo, pensato per un pubblico ampio, curioso e che apprezza il food pairing. La comunicazione qui funziona meglio se è diretta, accessibile e centrata sulla versatilità per tutte le occasioni. Guardando questi segmenti, le aziende possono pianificare diverse strategie.

Per le classi tradizionali (1-2), occorre mantenere una forte differenziazione, considerando la possibilità di fare qualche progetto comune per valorizzare ancora di più la tradizione. Per le classi moderne (4-5), serve inventiva su branding, marketing e dove vendere, perché i prodotti rischiano di essere troppo simili e concorrenti. Per la classe 3, la migliore mossa sarebbe renderla il ponte fra i due mondi: con prodotti di volume, facili da spiegare, e adatti a tante situazioni, senza troppi tecnicismi. Infine, la mappa aiuta a vedere anche dove manca qualcosa nel mercato. Nessuna classe, infatti, si trova al centro esatto, dove potrebbe collocarsi per una nuova tipologia di vino, pensata come prodotto "universale", non troppo strutturato, ma nemmeno troppo leggero, perfetto per chi cerca equilibrio fra tradizione e innovazione.

3.5 Analisi della Varianza (ANOVA) delle Classi

Per validare statisticamente queste separazioni visive, è stata condotta un'analisi della varianza (ANOVA) su ciascuna delle 13 variabili chimiche. La tabella seguente mostra i risultati dei test ANOVA a una via, che testa la differenza nelle medie di ciascuna variabile chimica tra le cinque classi di vino.

Variabile Chimica	**P-value**	**Significatività**
Flavanoids	< 0.001	***
OD280.OD315.of.diluted.wines	< 0.001	***
Hue	< 0.001	***
Proline	< 0.001	***
Color.intensity	< 0.001	***
Alcalinity.of.ash	< 0.001	***
Proanthocyanins	< 0.001	***
Nonflavanoid.phenols	< 0.001	***
Malic.acid	0.0055	**
Magnesium	0.0069	**
Alcohol	0.724	Non Significativo
Total.phenols	0.879	Non Significativo
Ash	0.950	Non Significativo

Tabella 7: Risultati dell'Analisi ANOVA

I risultati dell'ANOVA (p-values) confermano con forza la struttura osservata sulla mappa. Variabili chiave che definiscono gli assi, come **Flavanoids**, **Alcalinity.of.ash**, **OD280.OD315.of.diluted.wines**, **Hue**, **Proline** e **Color.intensity**, mostrano differenze estremamente significative tra le classi (tutte con **p-value** $<$ **0.001**). Questo conferma che la separazione tra i cluster non è casuale, ma è guidata da differenze chimiche profonde e statisticamente reali. È interessante notare che non tutte le variabili contribuiscono alla differenziazione. Variabili come **Alcohol (p = 0.724)**, **Ash (p = 0.95)** e **Total.phenols (p = 0.879)** mostrano p-value molto alti, indicando che le differenze osservate nelle loro medie non sono statisticamente significative.

3.6 Validità del Modello Metrico sui Centroidi e Analisi della Complessità Reale

La scelta dell'MDS Metrico come approccio principale per l'analisi dei centroidi è giustificata non solo dai risultati, ma anche dalla natura fondamentale dei dati. Le 13 variabili di composizione sono quantitative (misurate su scale a intervalli o a rapporti), permettendo il calcolo di distanze Euclidee metriche e significative. L'MDS Metrico è la tecnica d'elezione in questo scenario, poiché è progettata per preservare questi valori numerici, a differenza dell'MDS Non Metrico che, utilizzando solo i ranghi, trascurerebbe la ricchezza di queste informazioni metriche. Inoltre, il vantaggio pratico dell'approccio Metrico, attraverso la sua equivalenza con la PCA, è quello di fornire un'interpretazione analitica degli assi, associando la separazione delle classi a specifici driver chimici. Tuttavia, questa mappa dei centroidi, sebbene accurata nel suo scopo, rappresenta una forte semplificazione. La vera complessità del dataset è emersa tentando di mappare tutti i 178 vini individualmente (risultando in uno stress elevato) e quando si è analizzata la loro struttura interna (con il fallimento del modello INDSCAL). Ciò dimostra che la realtà dei dati è profondamente eterogenea e che le "regole" chimiche che definiscono ciascuna classe sono fondamentalmente diverse. Nonostante questa complessità di fondo, il modello MDS Metrico applicato ai centroidi si dimostra comunque un ottimo strumento di sintesi. Esso riesce a filtrare il rumore della variabilità interna ed estrarre l'informazione più importante, fornendo una mappa chiara, interpretabile e statisticamente robusta delle sole differenze *medie*, riassumendo visivamente il posizionamento strategico e le principali direttrici di differenziazione tra i profili delle classi di vino.

4 Principal Component Analysis (PCA)

4.1 Principi teorici della PCA

L'Analisi in Componenti Principali (PCA) consente di rappresentare un insieme di dati mediante un numero ridotto di variabili costruite come combinazioni lineari delle originali, trasformando variabili correlate in componenti mutuamente incorrelate che preservano la massima informazione possibile. Date n osservazioni descritte da p variabili, si ricerca la trasformazione lineare ottimale che massimizzi la varianza spiegata da ciascuna componente, ordinate in senso decrescente d'importanza.

1. La prima variabile y_1 è la combinazione lineare delle variabili originali che conserva la massima variabilità possibile.
2. La seconda y_2 spiega la quota maggiore della variabilità residua ed è non correlata con y_1.
3. Le successive componenti massimizzano la varianza non ancora spiegata.

La prima componente è:

$$y_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p \tag{9}$$

dove i coefficienti α_{1j} massimizzano la varianza di y_1 sotto il vincolo:

$$\alpha_1^T \alpha_1 = 1 \tag{10}$$

La varianza di y_1, combinazione lineare delle variabili di X, è:

$$Var(y_1) = \alpha_1^T S \alpha_1 \tag{11}$$

dove S è la matrice di covarianza $p \times p$. L'ottimizzazione con moltiplicatori di Lagrange porta alla condizione:

$$S\gamma = \lambda\gamma \quad \text{con} \quad \alpha_1 = \gamma_1 \tag{12}$$

ossia α_1 è l'autovettore associato al maggiore autovalore di S. In generale, la j-esima componente principale è:

$$y_j = \alpha_j^T x \tag{13}$$

con i vincoli:

$$\alpha_j^T \alpha_j = 1 \quad \text{e} \quad \alpha_j^T \alpha_i = 0; \forall (i < j) \tag{14}$$

e si ha:

$$\alpha_j = \gamma_j \tag{15}$$

La varianza della j-esima componente principale è data da:

$$Var(y_j) = \lambda_j \tag{16}$$

e la somma degli autovalori coincide con la varianza totale:

$$\sum_{i=1}^{p} \lambda_i = s_1^2 + s_2^2 + \cdots + s_p^2 = tr(S) \tag{17}$$

La proporzione di varianza spiegata dalla j-esima componente è:

$$P_j = \frac{\lambda_j}{tr(S)} \tag{18}$$

mentre le prime q componenti spiegano complessivamente:

$$P(q) = \frac{\sum_{j=1}^{q} \lambda_j}{tr(S)} \tag{19}$$

Infine, i valori proiettati nello spazio delle componenti principali si ottengono da:

$$y_{nq} = \alpha_q^T x_n \tag{20}$$

ottenendo n osservazioni descritte da q nuove variabili ($q < p$) che riassumono la parte più significativa dell'informazione originaria.

Le componenti principali sono ottenute come combinazioni lineari delle variabili originali, e i coefficienti α_{ij} rappresentano il contributo di ciascuna variabile x_i alla componente y_j. Tali coefficienti, opportunamente scalati, sono detti **carichi** (o *loadings*) e si calcolano come:

$$L_{ij} = \alpha_{ij} \sqrt{\lambda_j} \tag{21}$$

Quando i dati sono standardizzati (cioè ogni variabile ha media zero e varianza unitaria), i loadings coincidono con i coefficienti di correlazione tra x_i e y_j.

4.1.1 Il Biplot

Nel contesto della PCA, il Biplot rappresenta simultaneamente le osservazioni e le variabili originali nel piano delle prime componenti principali. Le coordinate dei punti corrispondono ai punteggi delle osservazioni sulle componenti, mentre i vettori (frecce) indicano le direzioni dei carichi fattoriali, cioè il contributo di ciascuna variabile alle componenti.

- La direzione di una freccia mostra la componente con cui la variabile è più correlata.
- La lunghezza riflette l'intensità della correlazione o la quota di varianza spiegata.
- L'angolo tra due vettori esprime la correlazione tra variabili (piccolo = positiva, 90° = indipendenza, 180° = negativa).
- La proiezione ortogonale di un punto su una freccia indica quanto quell'osservazione è associata a quella variabile.

Il Biplot permette così di interpretare in un'unica rappresentazione le relazioni tra variabili, individui e componenti principali.

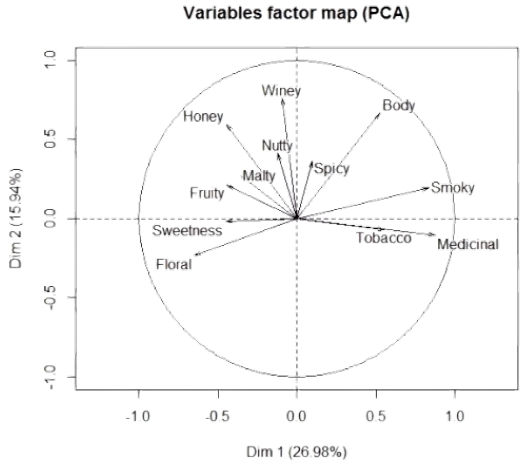


Figura 8: Esempio di Biplot delle variabili ottenuto da un'analisi in componenti principali.

4.2 Applicazione e Interpretazione della PCA

Dopo aver applicato l'MDS metrico al dataset oggetto di questo studio, si è ritenuto opportuno eseguire anche l'Analisi in Componenti Principali (PCA). Tale scelta è giustificata dal fatto che il dataset non contiene distanze pre-calcolate tra le osservazioni, bensì le singole valutazioni delle variabili, condizione che consente l'applicazione diretta della PCA. L'obiettivo è confrontare i risultati ottenuti dalle due tecniche e valutarne le analogie e le differenze nella rappresentazione dei dati.

```
> pca_fit <- prcomp(centroidi_scaled)
> print("Riepilogo della Varianza Spiegata (PCA):")
> summary(pca_fit)

Importance of components:
                          PC1    PC2    PC3     PC4       PC5
Standard deviation     2.9365 1.5224 1.1660 0.83659 3.054e-16
Proportion of Variance 0.6633 0.1783 0.1046 0.05384 0.000e+00
Cumulative Proportion  0.6633 0.8416 0.9462 1.00000 1.000e+00
```

Dall'applicazione della PCA ai centroidi standardizzati delle classi si osserva che la prima componente principale (PC1) spiega da sola circa il 66,3% della varianza totale, mentre la seconda (PC2) ne aggiunge un ulteriore 17,8%, portando la varianza cumulata all'84,2%. Le prime due componenti, quindi, catturano gran parte dell'informazione originaria del dataset, consentendo una rappresentazione bidimensionale efficace dei dati senza perdita significativa di informazione. Le componenti successive (PC3 e PC4) spiegano porzioni di varianza più ridotte (circa il 10,5% e il 5,4%), e risultano pertanto meno rilevanti ai fini dell'interpretazione globale. Questo conferma che la struttura principale dei dati può essere adeguatamente descritta dalle prime due componenti, rendendo la rappresentazione nel piano PC1–PC2 appropriata per il confronto con la configurazione ottenuta tramite MDS metrico.

```
> print("Loadings (Contributo delle variabili agli assi):")
> print(pca_fit$rotation)
```

	PC1	PC2
Alcohol	0.26840096	-0.001697032
Malic.acid	0.12574176	-0.555547465
Ash	0.28339167	0.007344377
Alcalinity.of.ash	0.32743705	0.174922122
Magnesium	-0.25175890	-0.433530011
Total.phenols	-0.09631104	0.176991477
Flavanoids	-0.34029892	0.003721283
Nonflavanoid.phenols	0.32412602	-0.049869567
Proanthocyanins	-0.32453376	-0.023880405
Color.intensity	0.17259520	-0.504757116
Hue	-0.32541572	0.171672784
OD280.OD315.of.diluted.wines	-0.33592781	0.100980207
Proline	-0.27793979	-0.379366318

L'analisi dei loadings (ovvero i contributi delle variabili alle componenti principali) consente di interpretare più nel dettaglio la struttura del piano PC1–PC2 mostrato nel biplot (Figura 9). La prima componente principale (**PC1**), che spiega oltre il 66% della varianza, è fortemente influenzata da variabili come *Alcalinity.of.ash*, *Ash*, *Alcohol* e *Nonflavanoid.phenols*, caratterizzate da coefficienti positivi elevati. Al contrario, variabili come *Flavanoids*, *OD280/OD315.of...*, *Proanthocyanins* e *Hue* presentano coefficienti negativi, suggerendo che la PC1 descrive un **contrasto tra vini con maggiore contenuto alcolico e alcalinità delle ceneri e quelli più ricchi di flavonoidi e pigmenti coloranti.** La seconda componente (**PC2**), che aggiunge circa il 18% di varianza, oppone invece le variabili *Malic.acid* e *Color.intensity* (con pesi negativi elevati) a *Total.phenols* e *Hue* (con pesi positivi). Essa sembra quindi **distinguere i vini in base all'intensità del colore e alla componente acida.**

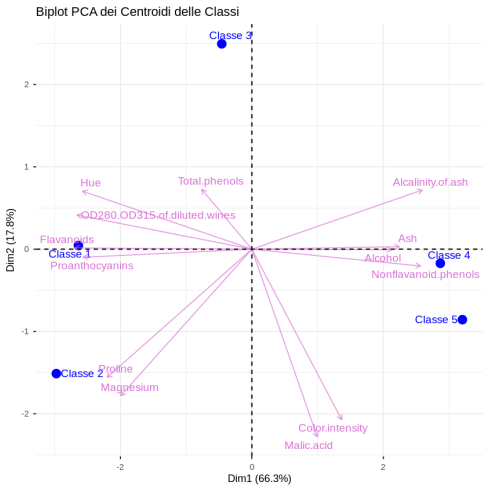


Figura 9: Biplot delle prime due componenti principali con i centroidi delle classi.

Nel biplot (Figura 9), le classi di vini risultano ben differenziate:

- le Classi 4 e 5 si collocano nel quadrante destro, in prossimità delle variabili *Alcohol* e *Ash*, suggerendo che siano caratterizzate da valori più elevati di tali componenti;
- la Classe 3 si distingue nettamente lungo PC2, associandosi a valori relativamente più alti di *Total.phenols* e *Hue*;
- le Classi 1 e 2 si dispongono invece nel quadrante sinistro, in relazione inversa alle variabili alcoliche e diretta a quelle fenoliche e colorimetriche.

La lunghezza dei vettori, inoltre, riflette l'intensità della loro influenza: i vettori più lunghi (*Alcohol*, *Alcalinity.of.ash*, *Flavanoids*) contribuiscono maggiormente alla definizione delle componenti principali. Nel complesso, la PCA fornisce una rappresentazione coerente con quella ottenuta dall'MDS metrico, confermando la buona separabilità delle classi e l'importanza di alcune variabili chimiche nella distinzione dei vini.

5 Confronto MDS e PCA

L'obiettivo di questa sezione è confrontare i risultati ottenuti tramite le due tecniche di riduzione dimensionale, MDS Metrico e PCA, per dimostrare la loro equivalenza concettuale e pratica nel contesto di questa analisi. Questo diagramma mostra come PCA e MDS semplifichino i dati. La differenza principale è il punto di partenza: la PCA sfrutta le correlazioni tra le variabili, mentre MDS partono dalle loro distanze. Entrambe le tecniche usano un motore matematico comune (decomposizione agli autovalori) per calcolare le coordinate per la rappresentazione geometrica dei punti in un grafico e la varianza spiegata da ciascun asse. In sostanza, raggiungono lo stesso obiettivo di visualizzazione attraverso percorsi leggermente diversi.

5.1 Confronto Visivo delle Mappe Percettive

L'ispezione visiva delle due mappe generate fornisce la prima, immediata evidenza di una profonda somiglianza strutturale.

* Mappa MDS Metrico: Posiziona le 5 classi di vino in uno spazio bidimensionale in cui la vicinanza tra i punti riflette la similarità dei loro profili chimici.
* Biplot PCA: Proietta le 5 classi sui primi due assi di massima varianza (PC1 e PC2).

Il grafico finale (Figura 10), che rappresenta le coordinate ottenute da entrambe le tecniche (in nero per l'MDS e in rosso per la PCA), illustra in modo inequivocabile la loro relazione: L'analisi visiva evidenzia che la configurazione dei punti della PCA è l'immagine speculare della configurazione MDS rispetto all'asse orizzontale. La struttura delle relazioni tra i punti è perfettamente conservata:

- Le Classi 4 e 5 formano un cluster coeso in entrambe le mappe.
- Le Classi 1 e 2 sono posizionate vicine tra loro.
- La Classe 3 risulta isolata in entrambe le rappresentazioni.

Poiché la rotazione o il ribaltamento di un grafico non ne alterano le distanze relative e poichè la soluzione MDS (secondo il teorema di Torgerson) è unica a meno di trasformazioni ortogonali, di conseguenza, l'interpretazione non cambia e possiamo concludere che le due mappe sono visivamente equivalenti.

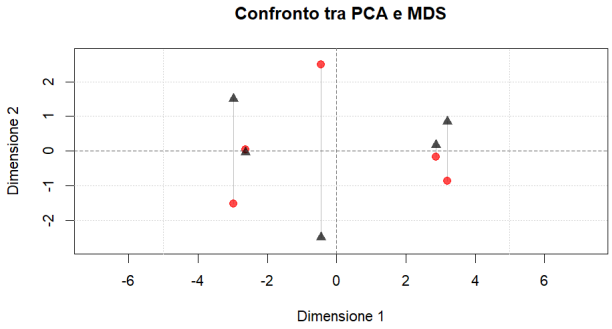


Figura 10: Confronto Visivo MDS(nero)-PCA(rosso)

5.2 Confronto Numerico delle Coordinate e delle Performance

L'equivalenza non è solo visiva, ma è rigorosamente confermata dai dati numerici prodotti dallo script. Dal confronto diretto dell'output, si osserva che:

- √ Le coordinate lungo la Dimensione 1 (MDS) sono identiche a quelle della Componente Principale 1 (PCA).
- √ Le coordinate lungo la Dimensione 2 (MDS) sono identiche in valore assoluto a quelle della Componente Principale 2 (PCA), ma di segno opposto.

```
Coordinate dei punti ottenute mediante MDS metrico

Classe 1 -2.638399 -0.04499941
Classe 2 -2.973869 1.51135289
Classe 3 -0.457447 -2.49437000
Classe 4 2.867035 0.17164295
Classe 5 3.202680 0.85637357

Coordinate dei punti ottenute mediante PCA
              PC1        PC2
Classe 1 -2.638399 0.04499941
Classe 2 -2.973869 -1.51135289
Classe 3 -0.457447 2.49437000
Classe 4 2.867035 -0.17164295
Classe 5 3.202680 -0.85637357
```

Questo conferma matematicamente che le due soluzioni sono l'una il riflesso dell'altra.

5.2.1 Analisi delle Performance

Il confronto tra gli indicatori di performance delle due tecniche fornisce la prova finale e più importante della loro equivalenza in questo contesto.

1. Per l'MDS, la Bontà di Adattamento (GoF), che misura la fedeltà con cui le distanze originali sono state riprodotte sulla mappa, è risultata pari a 0.842 (84.2%).
2. Per la PCA, la Varianza Cumulata Spiegata dalle prime due componenti, che indica la percentuale di informazione totale catturata dalla mappa, è risultata pari a 0.84158 (circa 84.2%).

```
> (test <- var_cumulata["PC2"] == round(gof, 5))

 TRUE
```

L'identità numerica di questi due indicatori chiave non è una coincidenza. Il test eseguito nello script sancisce in modo formale questa uguaglianza.

6 Analisi delle Corrispondenze

L'Analisi delle Corrispondenze (AC) è un metodo statistico di riduzione della dimensionalità che tratta dati qualitativi multidimensionali, originariamente introdotto nei lavori di R.A. Fisher (1940) sulle tabelle di contingenza. Tale metodologia è particolarmente adatta a rappresentare in uno spazio di dimensioni ridotte le relazioni di associazione tra modalità di due caratteri qualitativi.

Nel presente studio, l'AC è stata applicata ai dati di una tabella di contingenza 5×5 che incrocia le 5 classi di vini con i 5 profili di sommelier (genere e fasce di età).

La tabella di contingenza è stata creata selezionando la variabile Classificazione e le colonne di interesse, dalla 15 alla 19, relative alle caratteristiche dei sommelier; i dati sono stati aggregati calcolando la somma delle preferenze per ogni gruppo, raggruppando per Classificazione.

L'obiettivo è rappresentare graficamente le associazioni tra classi di vino e le specifiche caratteristiche di sommelier.

Classe	**F**	**M**	**Eta18_25**	**Eta25_40**	**Eta_sup40**
Classe 1	94	90	89	89	95
Classe 2	53	46	58	62	59
Classe 3	213	219	201	202	218
Classe 4	77	77	78	72	66
Classe 5	63	68	74	75	62

Tabella 8: Tabella di Contingenza

A partire dalla tabella di contingenza è stata ricavata la tabella delle frequenze relative congiunte che include le frequenze relative marginali (pesi), dei profili riga e dei profili colonna.

	F	**M**	**Eta18_25**	**Eta25.40**	**Eta_sup40**	**fi.**
Classe 1	0.0376	0.0360	0.0356	0.0356	0.0380	0.1828
Classe 2	0.0212	0.0184	0.0232	0.0248	0.0236	0.1112
Classe 3	0.0852	0.0876	0.0804	0.0808	0.0872	0.4212
Classe 4	0.0308	0.0308	0.0312	0.0288	0.0264	0.1480
Classe 5	0.0252	0.0272	0.0296	0.0300	0.0248	0.1368
f.j	0.2000	0.2000	0.2000	0.2000	0.2000	1.0000

Tabella 9: Tabella delle frequenze relative congiunte con marginali di riga e colonna

6.1 La Distanza del Chi-Quadrato

Contrariamente all'analisi fattoriale classica basata sulla metrica euclidea, l'AC utilizza la metrica del **chi-quadrato**, particolarmente adatta al trattamento di dati categorici e che considera i pesi dei profili (inverso delle frequenze relative marginali). Per il profilo riga i e il profilo riga i', la distanza al quadrato secondo la metrica del chi-quadrato è:

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \tag{22}$$

Analogamente, la distanza tra il profilo colonna j e il profilo colonna j' è:

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \tag{23}$$

Oltre alle consuete proprietà delle metriche, la metrica del chi-quadrato possiede la proprietà dell'**equivalenza distribuzionale**: se si aggregano due profili riga identici, la distanza tra gli elementi di colonna rimane invariata. Questa proprietà garantisce l'invarianza dei risultati rispetto alla codifica originaria delle variabili.

6.2 Risultati dell'Analisi delle Corrispondenze

L'analisi delle corrispondenze effettuata sulla tabella di contingenza evidenzia una forte struttura interna, spiegata efficacemente dalle prime due dimensioni, che spiegano rispettivamente il 57,6% e il 36,6% dell'inerzia totale, per un valore cumulativo pari al 94,2%. Questo risultato certifica che la maggior parte della variabilità nei profili delle classi e delle modalità è ben sintetizzata da queste due dimensioni principali.

	Dim 1	Dim 2	Dim 3	Dim 4
Autovalore	0.001867	0.001185	0.000189	0
Percentuale	57.61%	36.56%	5.83%	0%

Tabella 10: Inerzie principali (Autovalori) delle dimensioni

6.2.1 Analisi dei Profili Riga

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Mass	0.182800	0.111200	0.421200	0.148000	0.136800
ChiDist	0.028193	0.100848	0.036608	0.061036	0.078785
Inertia	0.000145	0.001131	0.000564	0.000551	0.000849
Dim. 1	-0.388	1.890	-0.799	-0.051	1.496
Dim. 2	0.587	1.691	0.259	-1.652	-1.169

Tabella 11: Statistiche sulle classi di vino (profili riga)

name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
Cls1	183	869	45	-17	354	28	20	514	63
Cls2	111	989	349	82	656	397	58	333	318
Cls3	421	947	174	-35	888	269	9	59	28
Cls4	148	869	170	-2	1	0	-57	868	404
Cls5	137	934	262	65	673	306	-40	261	187

Tabella 12: Riassunto dettagliato dei profili riga

Guardando ai risultati sui profili riga, si osserva come le cinque classi abbiano masse e inerzie differenti, a testimonianza di profili descrittivi distinti. La classe 2, pur avendo una massa marginale relativamente bassa (0,111), presenta l'inerzia più elevata (0,001131, pari al 34,9% dell'inerzia totale delle righe) e una qualità di rappresentazione quasi perfetta (qlt = 0,989), ad indicare come entrambi gli assi contribuiscano a determinare significativamente la sua posizione.. I valori di Squared Correlation (cor = 0,656 su asse 1; cor = 0,333 su asse 2) testimoniano che la posizione della classe 2 è determinata in modo dominante dal primo asse, mentre i contributi assoluti (ctr = 397 ‰ su asse 1; ctr = 318 ‰ su asse 2) indicano che la classe 2 ha un ruolo importante nella determinazione dell'inerzia di entrambi gli assi, sebbene più marcato per la prima dimensione. In particolare, essa è la classe che determina maggiormente l'inerzia del primo asse. La classe 3, che è la più frequente (massa = 0,421), ha un'inerzia inferiore (0,000564, circa 17,4%), ma è ben spiegata dall'asse orizzontale (cor = 0,888), a cui fornisce anche un contributo significativo nella determinazione dell'inerzia (ctr = 269 ‰). Inoltre, essa è la seconda classe con la qualità di rappresentazione più elevata, il che indica come entrambi gli assi contribuiscano a determinare significativamente la posizione di questo profilo riga. La classe 5 mostra anch'essa un peso elevato nella struttura complessiva. Con una massa di 0,137 e un'inerzia di 0,000849 (26,2%), risulta ben rappresentata (qlt = 0,934). Il valore di Squared Correlation (cor = 0,673 sull'asse 1 è il più elevato, ad indicare che l'asse orizzontale determina maggiormente la sua posizione. Il valore del contributo assoluto, ctr = 306 ‰) evidenzia come essa sia un ulteriore riferimento nella costruzione della prima dimensione; anche sulla seconda dimensione il contributo è rilevante (ctr = 187 ‰). La classe 4 si distingue in modo netto: massa 0,148, inerzia 0,000551 (17,0%) e una qualità di rappresentazione elevata, pari a quella della classe 1 (qlt = 0,869). Tuttavia, il valore di squared correlation per l'asse 2 (cor = 0,868) ed il corrispondente contributo assoluto (ctr = 404 ‰) rivelano che essa è quasi completamente rappresentata dalla seconda dimensione, di cui ne determina anche, più di tutte, l'inerzia. La classe 1, infine, si caratterizza per una massa di 0,183 e una inerzia ridotta (0,000145, a cui corrisponde un ruolo meno strutturante: la qualità di rappresentazione è pari a quella della classe 4, con la differenza che la sua posizione è determinata in maniera importante anche dalla prima dimensione, ma maggiormente dalla seconda. Essa è la classe che meno di tutte determina l'inerzia degli assi. Il profilo della classe 1 si avvicina quindi a quello medio.

6.2.2 Analisi dei Profili Colonna

	F	M	Eta18_25	Eta25_40	Eta_sup40
Mass	0.200000	0.200000	0.200000	0.200000	0.200000
ChiDist	0.039264	0.065391	0.050575	0.060693	0.064396
Inertia	0.000308	0.000855	0.000512	0.000737	0.000829
Dim. 1	-0.744	-1.160	0.986	1.383	-0.465
Dim. 2	-0.052	-1.126	-0.754	0.162	1.771

Tabella 13: Statistiche sui gruppi sommelier (profili colonna)

name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
F	200	672	95	-32	670	111	-2	2	1
M	200	939	264	-50	587	269	-39	351	253
Eta18_25	200	974	158	43	710	195	-26	264	114
Eta25_40	200	977	227	60	969	382	6	8	5
Eta_40	200	993	256	-20	97	43	61	896	627

Tabella 14: Riassunto dettagliato dei profili colonna

Analizzando i profili colonna, si nota che tutte le modalità hanno massa identica (0,200), ma le differenze emergono in termini di inerzia, qualità di rappresentazione e contributi. `Eta25_40` presenta l'inerzia più alta (0,000737) e una qualità di rappresentazione elevata (qlt = 0,977). Il contributo assoluto dei sommelier tra 25 e 40 anni alla dimensione 1 (ctr = 382 ‰) e il corrispettivo valore di squared correlation (cor = 0,969) indicano che questa fascia d'età è fondamentale nella determinazione della variabilità sull'asse orizzontale ed è anche il profilo meglio rappresentato dall'asse 1. `Eta_sup40` determina prevalentemente la variabilità sulla seconda dimensione (inerzia = 0,000829, qlt = 0,993, cor = 0,896, ctr = 627 ‰ su asse 2). `Eta18_25` è il secondo profilo colonna con la qualità di rappresentazione più elevata, presenta l'inerzia più bassa, la sua posizione è maggiormente determinata dall'asse 1, a cui contribuisce anche nella determinazione dell'inerzia, sebbene contribuisca anche sull'asse 2 (ctr = 195 ‰ su asse 1, ctr = 114 ‰ su asse 2)).

`F` e `M` presentano ruoli complementari ma meno incisivi; le loro inerzie sono contenute (F = 0,000308; M = 0,000855), e i contributi agli assi sono relativamente minori (F ctr = 111 ‰ su Dim. 1, M ctr = 269 ‰ su Dim. 1 e 253 ‰ su Dim. 2). Sono maggiormente i sommelier donna a determinare l'inerzia dell'asse 1, mentre i maschi determinano maggiormente quella dell'asse 2. Inoltre, sia sommelier donna che uomo sono maggiormente rappresentati dall'asse 1, ma gli uomini sono rappresentati bene anche dall'asse 2.

I risultati ottenuti mostrano dunque come la struttura dei dati sia costruita sulla forte opposizione tra i profili delle diverse classi e categorie di sommelier. Alcuni gruppi si allontanano significativamente dal profilo medio e svolgono un ruolo centrale nella spiegazione della variabilità, mentre altri seguono la struttura principale senza influenzarla in modo significativo.

6.3 Grafici

6.3.1 La Mappa Rowprincipal: Profili Riga in Coordinate Principali

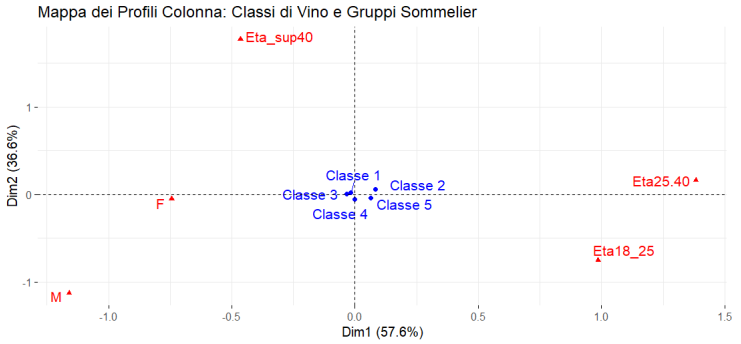


Figura 11: Mappa Profili Colonna.

La **mappa rowprincipal** rappresenta i profili riga all'interno della nuvola di punti dei profili colonna. Essa è una mappa baricentrica che, nel caso studiato, rivela come le cinque classi si distribuiscono nello spazio bidimensionale.

La Classe 2 emerge chiaramente come elemento estremo sul lato destro della Dimensione 1, confermando il suo ruolo di profilo maggiormente rappresentato. La Classe 3, pur prossima al baricentro per la sua natura centrale, mantiene una presenza significativa nel quadrante sinistro. La Classe 4 si posiziona nettamente nella parte inferiore della Dimensione 2, indicando una struttura radicalmente diversa dalle altre classi su questa dimensione. La Classe 5 si situa nel quadrante inferiore-destro, mentre la Classe 1 rimane nel quadrante superiore-sinistro. La classe 1 e la classe 4 rappresentano più delle altre il profilo medio, poichè si collocano quasi al baricentro della rappresentazione.

Dal lato sommelier, in questa rappresentazione i profili colonna servono principalmente da riferimento interpretativo. I sommelier della fascia E25-40 si posizionano verso destra, allineandosi con Classe 2, suggerendo una forte associazione. I sommelier con età superiore a 40 si situano nella parte superiore, distanti da Classe 4. Il genere M si posiziona leggermente a sinistra, mentre F rimane più centrale.

6.3.2 La Mappa Colprincipal: Profili Colonna in Coordinate Principali

La **mappa colprincipal** opera il contrario della rowprincipal: rappresenta i profili colonna nella nuvola dei punti dei profili riga. Questa rappresentazione enfatizza la struttura dei profili colonna, consentendo l'interpretazione delle distanze tra profili colonna come distanze chi-quadrato.

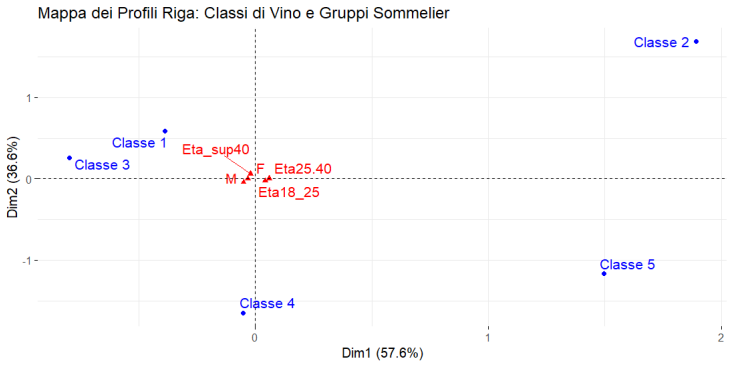


Figura 12: Mappa Profili Riga.

Dalla mappa emerge che i profili dei sommelier si dispongono secondo due poli principali: la fascia E25-40 si posiziona all'estrema destra della Dimensione 1, mentre la fascia E>40 si colloca nella parte superiore della Dimensione 2. Questo rivela che le preferenze dei sommelier si strutturano secondo due dimensioni indipendenti di differenziazione: la Dimensione 1 contrappone principalmente il gruppo E25-40 (a destra) alle altre fasce di età (a sinistra), mentre la Dimensione 2 oppone il gruppo E>40 (in alto) ai restanti profili (in basso).

Dal lato delle classi dei vini, in questa rappresentazione i profili riga forniscono il riferimento interpretativo. La Classe 2, posizionata all'estrema destra, risulta fortemente associata ai sommelier E25-40, confermando che questa classe esprime un profilo di preferenza specifico per il segmento giovane-adulto. La Classe 4, situata nella parte inferiore, si configura come elemento poco gradito ai sommelier E>40, suggerendo una polarizzazione su preferenze distinte. La Classe 3, rimasta centrale, testimonia il suo ruolo di prodotto a più ampia diffusione trasversale.

6.4 Digressione Teorica

Ma, da questa rappresentazione, come dalla precedente, non è possibile trarre conclusioni affidabili sulle relazioni tra i profili riga ed i profili colonna, poichè, in questa rappresentazione le coordinate sono prive del fattore di scala e dunque, le coordinate del generico profilo riga (colonna) su un asse nello spazio colprincipal (rowprincipal), sono una media pesata di tutti i profili colonna (riga) sull'analogo asse nello spazio rowprincipal (colprincipal).

Ogni profilo colonna (riga) nella mappa colprincipal (rowprincipal) è rappresentato da un punto che è il baricentro del profilo riga (colonna) nella mappa rowprincipal (colprincipal).

Inoltre, la nuvola di punti interna nei rispettivi spazi risulta essere troppo densa per un'interpretazione adeguata dei profili.

Dalle formule di transizione si ottengono i vettori delle coordinate di tutti i punti (e per i profili riga e per i profili colonna):

$$\begin{cases} \psi_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{D}_n^{-1} \mathbf{F} \varphi_\alpha \\ \varphi_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{D}_p^{-1} \mathbf{F}' \psi_\alpha \end{cases} \tag{24}$$

$$\begin{cases} \mathbf{D}_n^{-1} \mathbf{F} \varphi_\alpha = \sqrt{\lambda_\alpha} \psi_\alpha = \hat{\psi_\alpha} \\ \mathbf{D}_p^{-1} \mathbf{F}' \psi_\alpha = \sqrt{\lambda_\alpha} \varphi_\alpha = \hat{\varphi_\alpha} \end{cases} \tag{25}$$

In riferimento alla proprietà baricentrica descritta sopra, le formule di transizione esplicitate per le coordinate:

$$\begin{cases} \psi_{\alpha i} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} \varphi_{\alpha j} \\ \\ \varphi_{\alpha j} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}} \psi_{\alpha i} \end{cases} \tag{26}$$

Da cui è evidente la validità della precedente digressione.

Per la rappresentazione simultanea si cerca infatti una rappresentazione β-baricentrica prossima ad 1 (rappresentazione simultanea perfetta):

$$\begin{cases} \psi = \mathbf{D}_n^{-1} \mathbf{F} \, \varphi \\ \varphi = \mathbf{D}_p^{-1} \mathbf{F}' \psi \end{cases} \tag{27}$$

$$\begin{cases} \psi = \beta \mathbf{D}_n^{-1} \mathbf{F} \, \varphi \\ \varphi = \beta \mathbf{D}_p^{-1} \mathbf{F}' \psi \end{cases} \tag{28}$$

6.4.1 La Mappa Simmetrica: Migliore Rappresentazione Simultanea

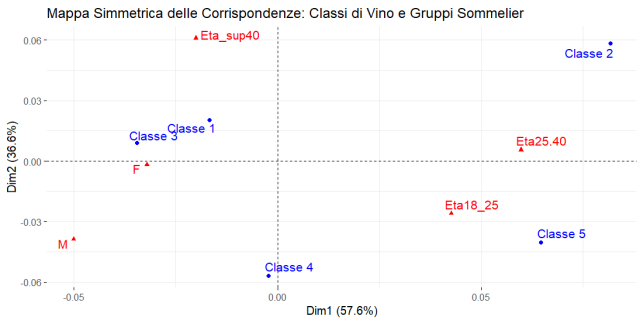


Figura 13: Mappa Simmetrica.

La **mappa simmetrica** rappresenta sia i profili riga che i profili colonna utilizzando le loro coordinate principali, opportunamente scalate per garantire la comparabilità e l'interpretabilità delle distanze.

Dal punto di vista teorico, la mappa simmetrica è la rappresentazione **1-baricentrica ottimale** secondo la metrica del chi-quadrato. Essa garantisce che ogni profilo riga è il baricentro ponderato dei profili colonna (e viceversa) rispetto alle coordinate visualizzate. Di conseguenza, la vicinanza tra un profilo riga e un profilo colonna indica una forte associazione positiva tra quella classe di vino e quel profilo sommelier, mentre la distanza segnala una associazione negativa o indifferenza.

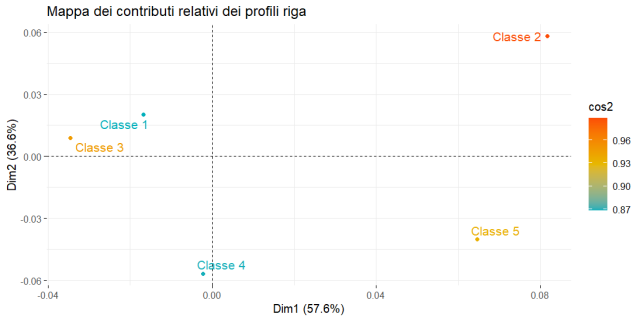
Analizzando la mappa prodotta, emergono alcune considerazioni di interesse. La prossimità tra Classe 2 e il profilo E25-40 nel primo quadrante (destra-centro) testimonia una forte attrazione: i sommelier della fascia 25-40 anni mostrano un'elevata preferenza per Classe 2, che si era detto facesse parte del gruppo dei vini tradizionali, ad indicare inoltre, che questa categoria di vini è maggiormente rivolta ai sommelier tra i 25 e i 40 anni. Questa associazione è supportata dal fatto che la coordinata di Classe 2 sulla Dimensione 1 è positiva e marcata, così come la coordinata di E25-40 sulla medesima dimensione.

La distanza considerevole tra Classe 4 (situata nella parte inferiore) e il profilo E>40 (nella parte superiore) indica un'associazione negativa: i sommelier della fascia d'età superiore a 40 anni mostrano scarsa o nulla preferenza per Classe 4 oppure essa non attrae questa categoria di sommelier, il che è evidenziato dalla posizione opposta lungo la Dimensione 2. La classe 4, in realtà, sembra non essere preferita da nessuna categoria di sommelier. Infatti, essa è molto distante da tutte le categorie, sebbene sia leggermente più vicina agli uomini ed alla fascia di età 18-25. Tuttavia, questa fascia d'età (18-25) risulta molto prossima alla classe 5, che faceva parte dei vini moderni, il che è confermato dalla coordinata positiva sulla prima dimensione di entrambi i profili e che indica una preferenza dei sommelier abbastanza marcata. Le Classi 1 e 3 sono molto vicine in questa mappa, cosa che non si evidenziava invece nella mappa MDS, in cui la classe 3 formava un gruppo a parte, quello dei vini sofisticati. In particolare, i sommelier donna preferiscono la classe 3 più di quanto preferiscano la classe 1, che invece trova un maggiore accordo con i sommelier di età superiore a 40, fascia che dunque preferisce i vicini tradizionali. I sommelier uomini appaiono avere un'associazione negativa con tutte le classi, le distanze minori si registrano con le classi 3 e 4, sebbene quest'ultima si trovi in una posizione poco favorevole alla determinazione della vicinanza in termini di asse. Inoltre, le classi 1 e 3 sono quelle più vicine al baricentro, il che potrebbe indicare che esse rappresentano una scelta di equilibrio per tutte le categorie di sommelier. In generale, emerge che le differenze di genere influenzano le preferenze in modo minore rispetto alle differenze di fascia d'età.

6.4.2 Qualità di Rappresentazione e Squared Correlations

La **mappa dei contributi relativi** fornisce un'ulteriore prospettiva sulla qualità della rappresentazione bidimensionale. Ogni punto è colorato secondo il valore del suo squared correlation totale $\text{QLT}_i = \text{cr}_1(i) + \text{cr}_2(i)$, con gradiente che varia dal blu (rappresentazione di bassa qualità) al rosso (rappresentazione di elevata qualità).

In questo contesto, tutti i profili riga manifestano valori di QLT elevati (superiori a 0.86), indicato dalla predominanza del colore arancione e rosso nella mappa. La Classe 2 si conferma come il profilo riga meglio rappresentato, a cui segue la classe 3. Le classi 1, 4 e 5 presentano qualità della rappresentazione meno elevate, ma comunque superiori all'85%, garantendo l'interpretabilità delle rispettive posizioni.



Si è deciso di produrre questa mappa per verificare se per qualche profilo riga la mappa simmetrica fosse fuorviante. Dunque, la mappa conferma la validità dell'interpretazione delle distanze relativamente ai profili riga e cattura adeguatamente la struttura di variabilità di ogni classe di vino.

7 Conclusioni

Il presente progetto ha consentito di esplorare le classi di vino attraverso MDS metrico e di metterli in relazione a specifiche categorie di sommelier attraverso l'Analisi delle Corrispondenze. Queste tecniche hanno permesso di evidenziare chiaramente le differenze e le similitudini tra le classi di vino e per specifiche categorie, offrendo una rappresentazione multidimensionale che integra le prospettive sia degli esperti (sommelier) sia delle caratteristiche intrinseche dei vini. Il progetto offre così una base chiara sia per migliorare la segmentazione del mercato dei vini, sia per supportare strategie di marketing mirate in funzione dei differenti gruppi di sommelier e delle loro preferenze.