UNIVERSITÀ
DEGLI STUDI
FIRENZE
Da un secolo, oltre.

Scuola di
Economia e Management

Corso di Laurea in
Statistica

Un'analisi statistica degli acidi grassi a catena corta: studio multi-malattia su due campioni paralleli

Relatore

Prof. Francesco Claudio Stingo

Candidato

Cecilia Boni

Anno Accademico 2024/2025

Indice

Capitolo 1 - Introduzione

In questa tesi si effettua un'analisi statistica sugli acidi grassi a catena corta, rilevati in due campioni distinti, plasma e feci, appartenenti a 4 gruppi clinici di pazienti, di cui uno di controllo; i dati sono stati acquisiti grazie al dipartimento di Medicina Sperimentale e Clinica dell'Università degli Studi di Firenze. Gli obiettivi di tale analisi sono confrontare come si distribuiscono le concentrazioni di sette acidi grassi a catena corta in due campioni, fecale e plasmatico, tra pazienti affetti da diverse patologie, verificare se i livelli degli acidi grassi sono in grado di discriminare tra le diverse malattie e valutare la presenza di una relazione tra le concentrazioni plasmatiche e fecali degli acidi, per indagare l'eventualità di un pattern tra i due campioni in ciascun gruppo di pazienti.

Gli acidi grassi sono acidi carbossilici saturi con una catena alifatica e possono essere classificati in base alla lunghezza della loro catena alifatica, alla presenza o assenza di doppi legami e alla posizione dei doppi legami; tra di essi, in particolare, gli acidi grassi a catena corta (Short Chain Fatty Acids, SCFA) presentano una catena alifatica costituita da 1 a 5 atomi di carbonio. Il 95% degli SCFA intestinali è rappresentato dagli acidi **Acetico**, **Propionico** e **Butirrico**; in quantità inferiori sono presenti altri acidi come **IsoButirrico**, **IsoValerico**, **2MetilButirrico** e **Valerico**. Per ognuno dei pazienti selezionati per lo studio, è stata misurata la concentrazione degli acidi grassi sopra citati nei relativi campioni plasmatico e fecale. La scelta di tali campioni è legata al fatto che gli SCFA si trovano in maggior quantità nelle feci e nel colon, dove possono essere assorbiti dall'epitelio o entrare nel flusso sanguigno.

L'adeguata produzione di SCFA da parte del microbiota intestinale è fondamentale per mantenere la normale fisiologia intestinale e le funzioni metaboliche dell'uomo. La fabbricazione degli SCFA ha origine da alcuni componenti della fibra alimentare che non vengono digeriti e vengono fermentati dalla flora batterica intestinale. Dalla fermentazione, insieme agli acidi grassi a catena corta, si generano gas intestinale e energia. Gli SCFA sono solitamente considerati una semplice fonte di energia, ma in realtà svolgono anche altre funzioni: mantenere l'integrità intestinale, regolare il metabolismo del glucosio e dei lipidi e modulare la regolazione immunitaria. Patologie come il cancro e le malattie autoimmuni, infettive o cardiovascolari provocano disbiosi intestinale, ovvero uno squilibrio del microbiota nella sua composizione e funzione, con la conseguente modifica delle concentrazioni degli acidi grassi nelle feci e nel sangue. Per questi motivi, per questo studio sono stati selezionati pazienti affetti da celiachia, da tumore del colon-retto e da obesità.

La **celiachia (CD)** è una patologia autoimmune dell'intestino, scatenata dall'ingestione di glutine in individui geneticamente predisposti. In tali soggetti l'esposizione al glutine provoca una serie di sintomatologie associate a disbiosi intestinale, con conseguente riduzione di batteri benefici, tra cui quelli produttori degli SCFA. Quest'ultimi contribuiscono a stimolare la produzione di muco, a mantenere l'integrità della barriera epiteliale e a modulare la risposta immunitaria locale, svolgendo funzione antinfiammatoria sulla mucosa intestinale. Per riequilibrare il microbiota, i pazienti celiaci, oltre a seguire una dieta priva di glutine, sono trattati con sostanze probiotiche specifiche che favoriscono la colonizzazione dei batteri capaci di produrre SCFA.

Il **cancro del colon-retto (CRC)** è una patologia il cui sviluppo è influenzato da complesse interazioni tra fattori genetici, esposizione ambientale e alterazione del microbiota intestinale. In pazienti affetti da CRC si riscontrano squilibri nel microbioma intestinale con conseguente riduzione di batteri benefici, tra cui i produttori degli SCFA. Da alcuni studi scientifici è emerso che in individui ad alto rischio di CRC le concentrazioni di alcuni acidi grassi a catena corta sono inferiori rispetto ai livelli degli individui sani.

L'**obesità** è una patologia multifattoriale che si caratterizza per un eccessivo accumulo di tessuto adiposo, con conseguenze importanti per la salute e la qualità della vita. Il microbiota intestinale svolge un ruolo critico nell'aumentare il rischio di insorgenza di tale patologia: la rottura del suo equilibrio e l'alterazione delle sue vie metaboliche possono influenzare il metabolismo del soggetto e aumentare l'accumulo di adipe. Un cambiamento nell'attività metabolica del microbiota intestinale può quindi contribuire allo sviluppo dell'obesità, ma nessuno studio ha identificato con precisione alcun gruppo di microrganismi intestinali che causino o contribuiscano all'insorgere dell'obesità; vi è certamente una relazione, seppur complessa e non ben nota.

1.1 I dati

Il dataset utilizzato per l'analisi è costituito da 18 variabili e 193 unità. I soggetti sono separati in gruppi in base alle categorie della variabile Malattia, la quale identifica la malattia da cui sono affetti: sono considerate tre malattie e il gruppo dei controlli sani. Si riportano nella Tabella 1.1 le malattie prese in esame e le numerosità dei gruppi.

Tra le 18 variabili, 14 sono le concentrazioni in percentuale dei 7 acidi grassi a catena corta rilevati per ogni soggetto sia dal campione plasmatico che dal campione fecale. Nel dataset il campione di provenienza è indicato con il nome dell'SCFA seguito da _stool (per il campione

fecale) e da _plasma (per il campione plasmatico). Nella Tabella 1.2 si osservano i valori della mediana e dei quartili degli acidi grassi separati per campione di provenienza.

Tabella 1.1: Numerosità dei gruppi

MALATTIA	NUMERO DI PAZIENTI
Celiachia (CD)	35
Cancro del colon-retto (CRC)	30
Controlli sani (HC)	95
Obesi	33

Tabella 1.2: Mediana, 1° e 3° quartile degli acidi grassi per campione

	PLASMATICO			FECALE		
ACIDI GRASSI	Q1	MEDIANA	Q3	Q1	MEDIANA	Q3
Acetico	56.841	66.415	86.727	49.482	55.455	61.639
Propionico	1.563	4.861	7.815	14.618	17.093	20.277
Butirrico	0.373	6.334	21.402	10.670	15.771	20.220
IsoButirrico	3.063	4.930	7.091	1.791	2.962	4.454
IsoValerico	0.821	1.327	2.910	0.831	1.679	2.877
2MetilButirrico	0.9997	1.685	2.439	0.885	1.773	2.895
Valerico	0.116	0.503	0.976	2.097	2.838	4.060

Le restanti 3 variabili sono Età, Sesso e ID che contengono rispettivamente l'età dei pazienti, con un range da 16 a 88 anni, il genere, a due livelli "M" (maschio) e "F" (femmina), e un codice identificativo, univoco per ogni paziente. Nella Tabella 1.3 sono riportati l'età media per malattie e numero di pazienti per malattia diviso per sesso.

Tabella 1.3: Età media e numero di pazienti maschi e femmine per ogni malattia

MALATTIA	ETÀ MEDIA	NUMERO FEMMINE	NUMERO MASCHI
Celiachia (CD)	35.26	24	11
Cancro del colon-retto (CRC)	72.97	10	20
Controlli sani (HC)	49.40	61	34
Obesi	51.30	17	16
		112	81

Durante il corso delle analisi si è reso necessario creare due dataset distinti chiamati dati_plasma e dati_feci. Tali datasets contengono entrambi tutte le 193 unità e 11 variabili, ovvero: Età,

Sesso, ID, Malattia e le concentrazioni dei 7 SCFA rilevate nel campione plasmatico per il dataset dati_plasma e nel campione fecale per il dataset dati_feci.

Quando le analisi vengono condotte separatamente per i due campioni, si utilizzano i datasets dati_plasma e dati_stool; mentre quando si considerano congiuntamente i valori degli acidi grassi relativi a ciascun paziente, l'analisi viene svolta utilizzando il dataset completo.

Capitolo 2 - Analisi esplorative

In questo capitolo vengono eseguite una serie di analisi esplorative sui dati, volte a esplorare e conoscere il dataset oggetto di studio in modo approfondito. Tramite l'utilizzo di statistiche descrittive e grafici si esaminano la struttura dei dati, la distribuzione delle variabili e le relazioni tra quest'ultime, con l'obiettivo di individuare associazioni e differenze nella composizione degli acidi tra i gruppi clinici esaminati.

2.1 Analisi descrittive

La parte iniziale dell'esplorazione dei dati si sofferma su rappresentazioni grafiche che permettono di ottenere una visione d'insieme della struttura e del comportamento delle variabili.

I *box plot* nella Figura 2.1 e nella Figura 2.2 mostrano la distribuzione delle concentrazioni degli acidi grassi a catena corta nei due campioni considerati: è evidente che l'acido presente in percentuale maggiore è l'acido Acetico, anche se nel campione plasmatico mostra maggiore variabilità. Per concentrazione segue in entrambi i casi l'acido Butirrico, mostrando nel plasma una distribuzione ampia con outlier verso l'alto, ma più simmetrica nelle feci. L'acido Propionico, terzo per concentrazione, presenta valori più elevati e stabili nel campione fecale

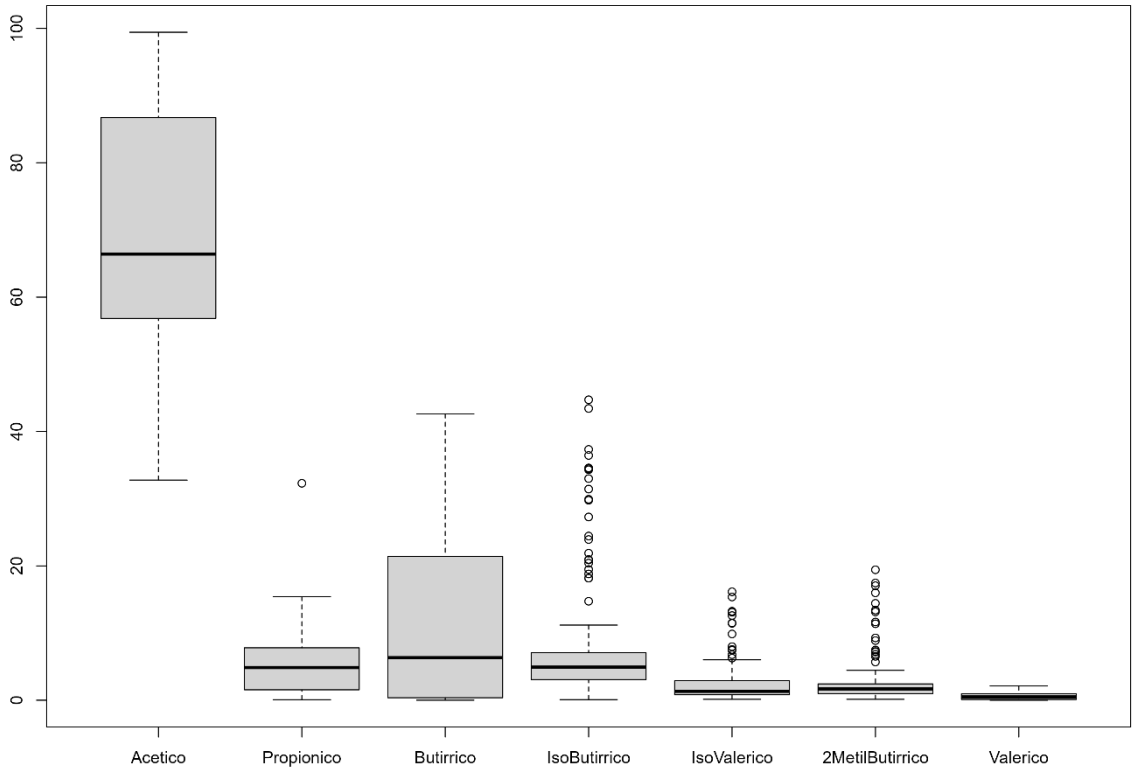

Figura 2.1: Box plot percentuali per ogni acido dal campione plasmatico

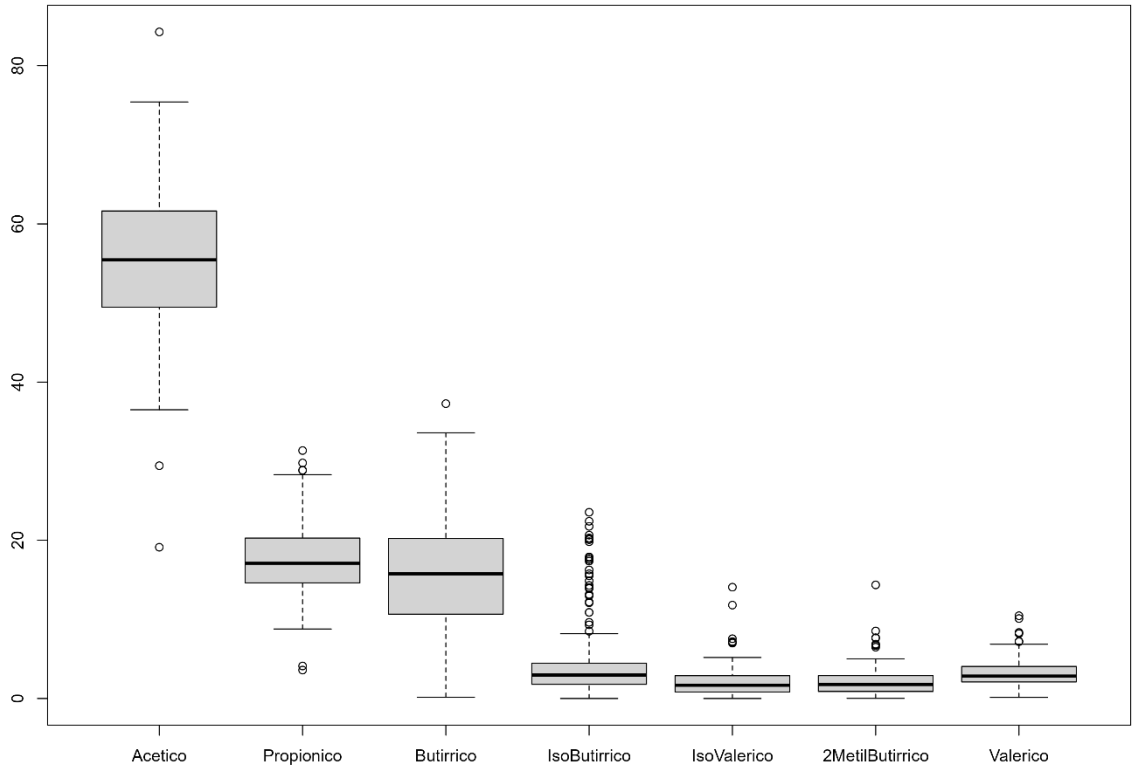

Figura 2.2: Box plot percentuali per ogni acido dal campione fecale

Nella Figura 2.3 e nella Figura 2.4 sono raffigurati i *violinplot* (diagrammi a violino) delle concentrazioni degli acidi grassi nei due campioni, stratificati per gruppo clinico. Questi grafici forniscono una rappresentazione della distribuzioni di densità e simultaneamente della variabilità dei dati.

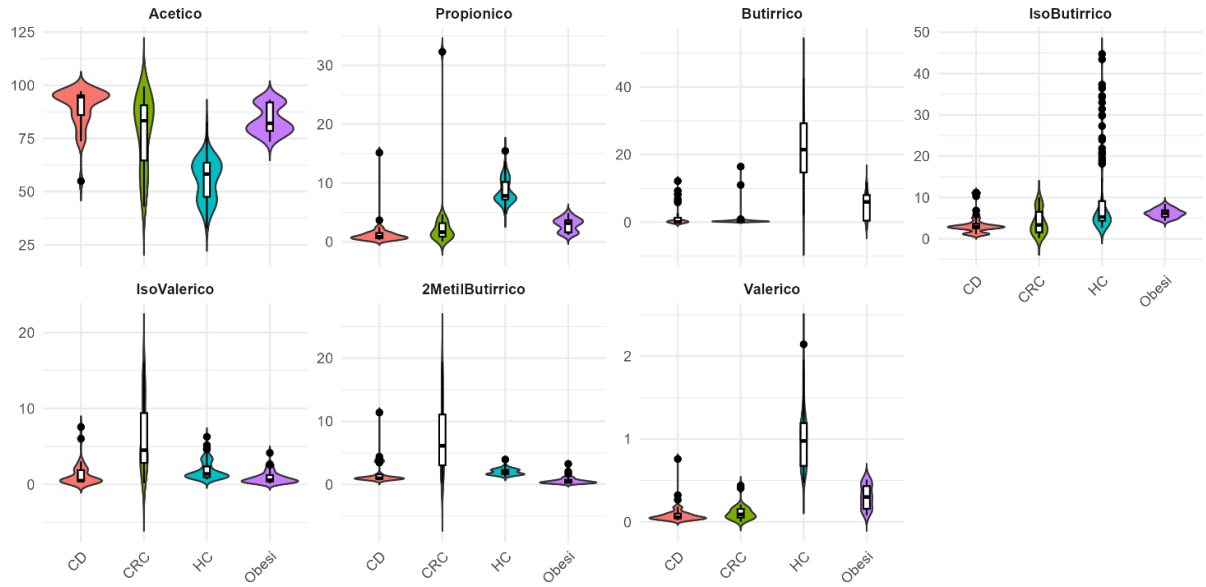

Figura 2.3: Violinplot per ogni acido stratificati per malattia dal campione plasmatico

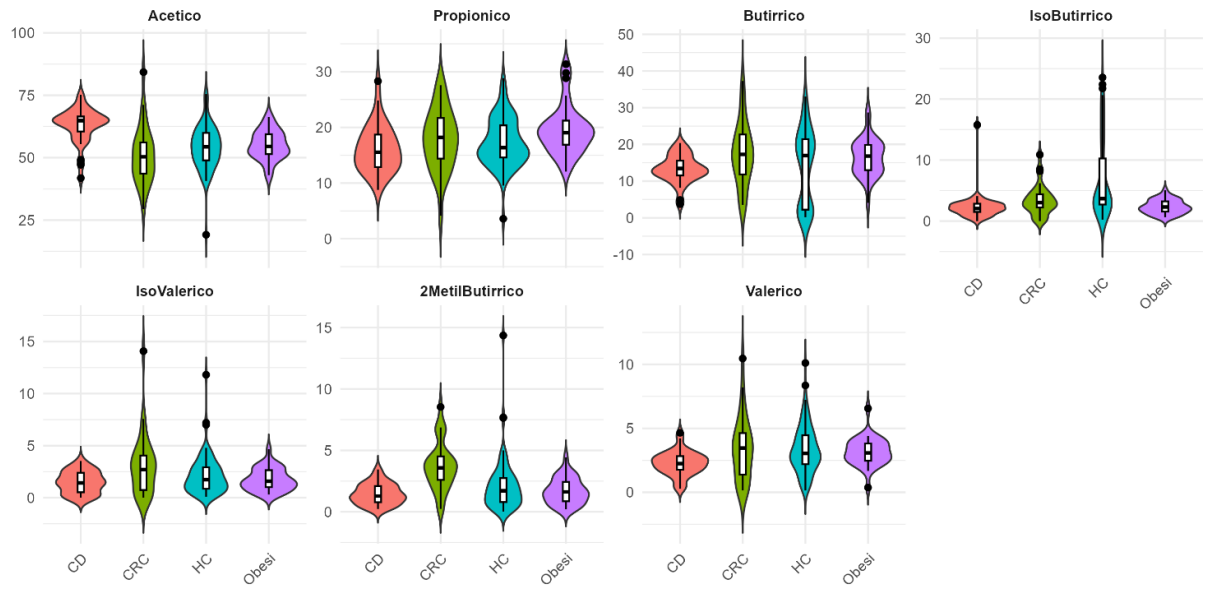

Figura 2.4: Violinplot per ogni acido stratificati per malattia dal campione fecale

L'acido Acetico si conferma predominante in tutti i gruppi: nel plasma le tre patologie mostrano concentrazioni mediamente più elevate rispetto ai controlli sani; nelle feci i valori risultano più bassi e con distribuzioni più variabili, con code verso l'alto per il gruppo CRC e verso il basso per i gruppi CD e HC.

L'acido Propionico presenta nel plasma distribuzioni molto più irregolari, con concentrazioni maggiori per CRC e HC e distribuzioni più strette e vicine allo zero per gli altri gruppi; nelle feci invece tutti i gruppi mostrano percentuali più elevate e distribuzioni più ampie; nei gruppi HC e CRC sono presenti anche qualche outlier verso il basso, mentre nei gruppi CD e Obesi verso l'alto.

L'acido Butirrico nel plasma registra valori modesti per le tre patologie e distribuzioni con code allungate con presenza di outlier; il gruppo HC mostra valori decisamente più elevati e una distribuzione notevolmente più variabile. Nelle feci le distribuzioni sono più ampie e omogenee per i gruppi CD e Obesi, mentre i gruppi HC e CRC mantengono una maggiore dispersione.

L'acido IsoButirrico, sia nel plasma che nelle feci, presenta per il gruppo HC distribuzioni con code molto lunghe e svariati outlier, con valori più elevati rispetto alle altre patologie, che al contrario mostrano valori bassi e distribuzioni compatte; anche i soggetti celiaci mostrano qualche outlier verso l'alto, rimanendo comunque inferiori rispetto ai valori dei controlli sani.

L'acido IsoValerico e l'acido 2MetilButirrico presentano distribuzioni piuttosto simili nei due campioni: nel plasma le distribuzioni sono particolarmente allungate e variabili nel gruppo CRC, rispetto a quelle compatte e ridotte negli altri gruppi, per i quali però si osservano qualche

outlier verso l'alto; nelle feci le distribuzioni sono più dense, ma l'acido IsoValerico, nel gruppo CRC, mostra comunque una variabilità maggiore rispetto agli altri gruppi clinici, con alcuni outlier, mentre l'acido 2MetilButirrico presenta tali caratteristiche nel gruppo HC.

L'acido Valerico nel plasma è quello con le concentrazioni più basse, per tutti i gruppi ha distribuzioni piatte e valori bassi, con eccezione per il gruppo HC nel quale registra valori poco più alti e maggiore variabilità; nelle feci invece le distribuzioni sono più compatte e dense, nel gruppo CRC si osserva maggiore variabilità con code più lunghe, seguito dai controlli sani.

In generale dai grafici emerge che, nel campione fecale, i gruppi CRC e HC spesso tendono a mostrare distribuzioni di concentrazione molto simili per la maggior parte degli acidi, al contrario le distribuzioni degli acidi per tali gruppi si differenziano notevolmente nel campione plasmatico nel quale spesso, quando uno dei due gruppi presenta distribuzione ampia e variabile, per l'altro è compatta e ridotta. Inoltre, nel campione plasmatico, le distribuzioni di quasi tutti gli acidi sono caratterizzate da maggiore variabilità e da una più frequente presenza di outlier; mentre, per il campione fecale, le distribuzioni appaiono generalmente più stabili e dense, seppur mantenendo visibili le differenze tra gruppi.

Nei *barplot* nella Figura 2.5 e nella Figura 2.6 si osservano le differenze nella composizione percentuale degli acidi per paziente tra le diverse malattie e divisi per campione.

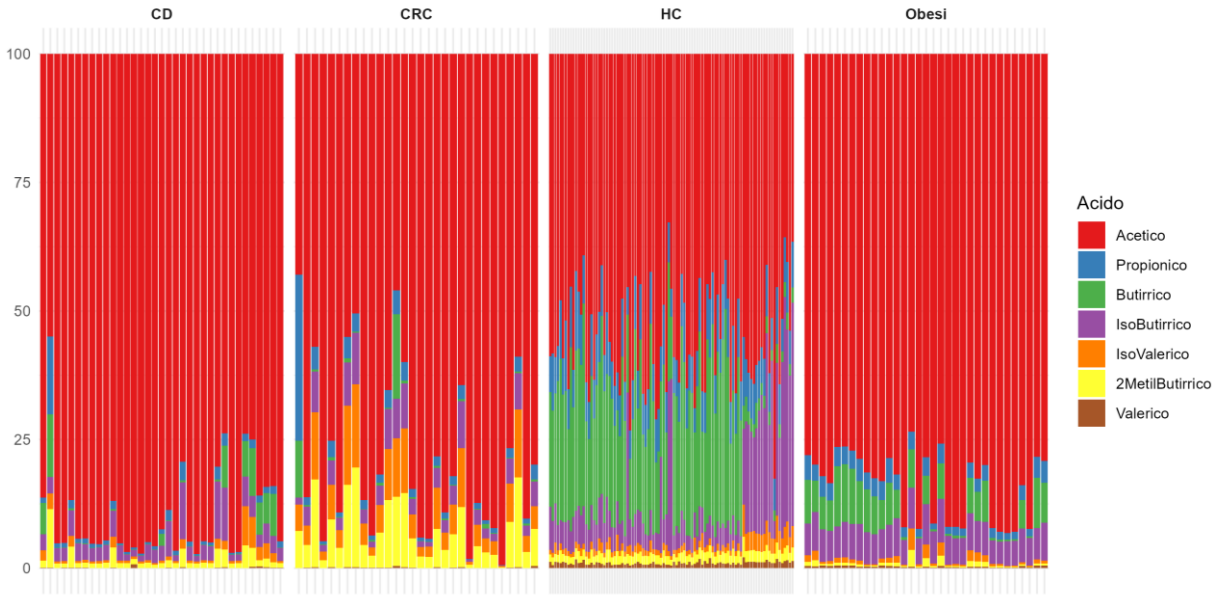

Figura 2.5: Barplot per ogni paziente nel campione plasmatico

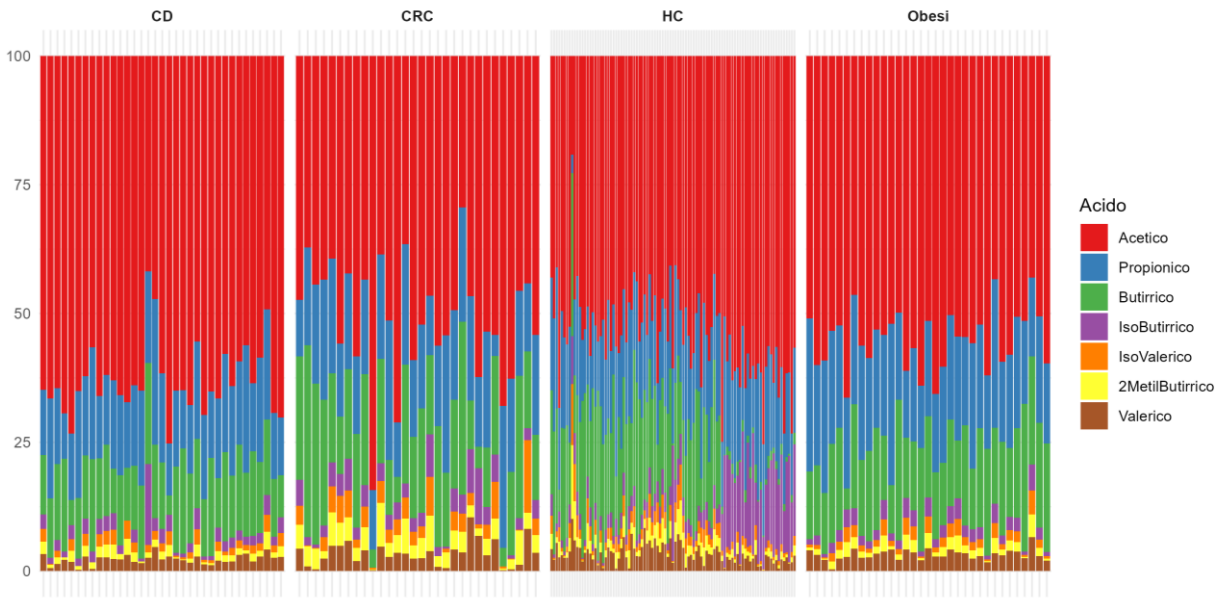

Figura 2.6: Barplot per ogni paziente nel campione fecale

Ancora una volta dai grafici si evidenzia la predominanza dell'acido Acetico, presente in concentrazioni elevate soprattutto nei campioni plasmatici. Si notano anche le concentrazioni degli acidi Propionico, Butirrico e Valerico più elevate nel campione fecale, rispetto al campione plasmatico.

Nei campioni fecali, come già osservato dai *violinplot*, emerge in modo ancora più evidente una similitudine per composizione percentuale degli SCFA tra i gruppi CD e Obesi e tra i gruppi CRC e HC. Tuttavia questa analogia si perde nei rispettivi campioni provenienti dal plasma: mentre il gruppo CRC mostra una maggiore presenza degli acidi IsoValerico e 2MetilButirrico, nel gruppo HC questi acidi sono meno rappresentati, a favore di una maggiore abbondanza degli acidi Propionico, Butirrico e IsoButirrico.

Emerge con coerenza che il gruppo HC, sia nel campione fecale che nel plasmatico, risulta il gruppo con la composizione più eterogenea e varia degli SCFA, suggerendo una possibile associazione tra diversità degli acidi grassi e salute intestinale.

Nella Figura 2.7 e nella Figura 2.8 sono raffigurati gli *scatterplot* dei dati sul piano delle prime due componenti principali, separatamente per i due campioni di raccolta. Tramite l'analisi in componenti principali (PCA) si riduce la dimensionalità del dataset mantenendo una buona rappresentatività dell'informazione originaria, utile in fase di analisi esplorativa dei dati per visualizzare strutture non osservabili. In questo caso l'obiettivo è quello di rappresentare le 193 osservazioni di entrambi i campioni, descritte originariamente da 7 variabili (gli SCFA), all'interno di uno spazio a 2 dimensioni.

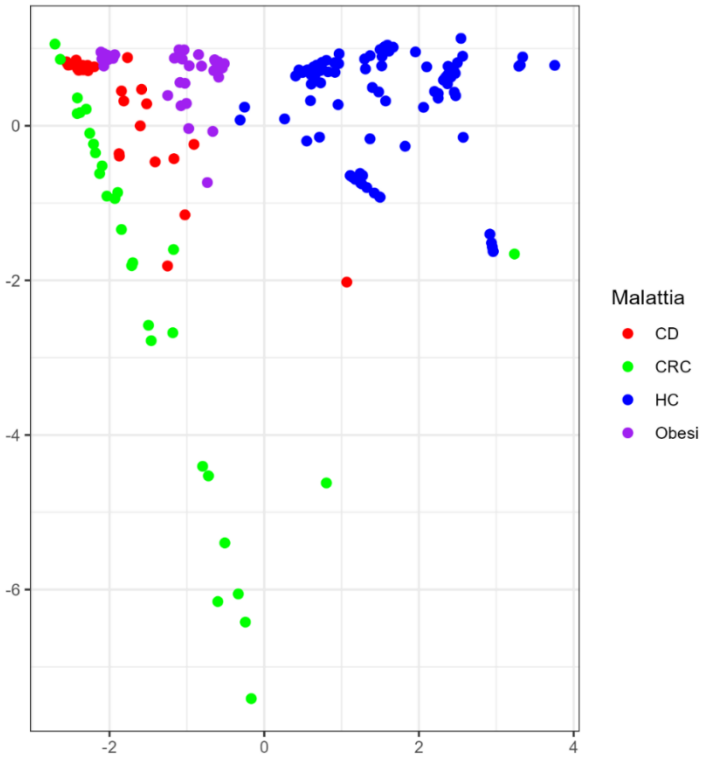

Figura 2.7: Scatterplot sul piano delle prime due componenti principali nel campione plasmatico

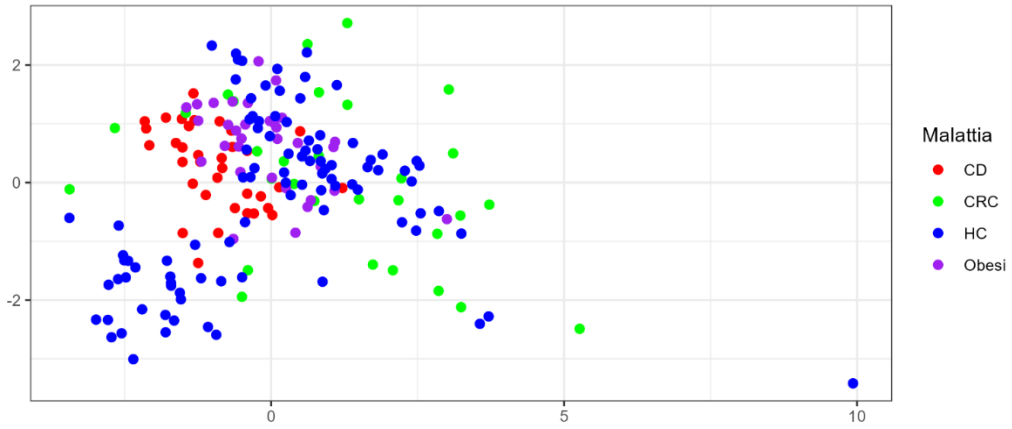

Figura 2.8: Scatterplot sul piano delle prime due componenti principali nel campione fecale

Nel campione plasmatico si osserva una separazione piuttosto netta tra i gruppi, suggerendo una buona capacità discriminativa della PCA in tale campione. Si nota in particolare che il gruppo HC si distingue bene dagli altri, compatto lungo entrambe le componenti; Obesi e CD mostrano una parziale sovrapposizione; il gruppo CRC presenta la maggiore dispersione, indicando variabilità nei valori degli SCFA per i soggetti di tale gruppo.

Nel campione fecale i gruppi appaiono maggiormente sovrapposti, solo HC tende in parte a discostarsi ma senza una separazione netta e il gruppo CRC è ancora quello con più ampia dispersione.

Nella Figura 2.9 e nella Figura 2.10 sono presentate le *heatmap* delle matrici di correlazione, separatamente per campione, sia quella generale (per tutti i pazienti indistintamente dal gruppo di appartenenza), sia quelle specifiche per ogni malattia.

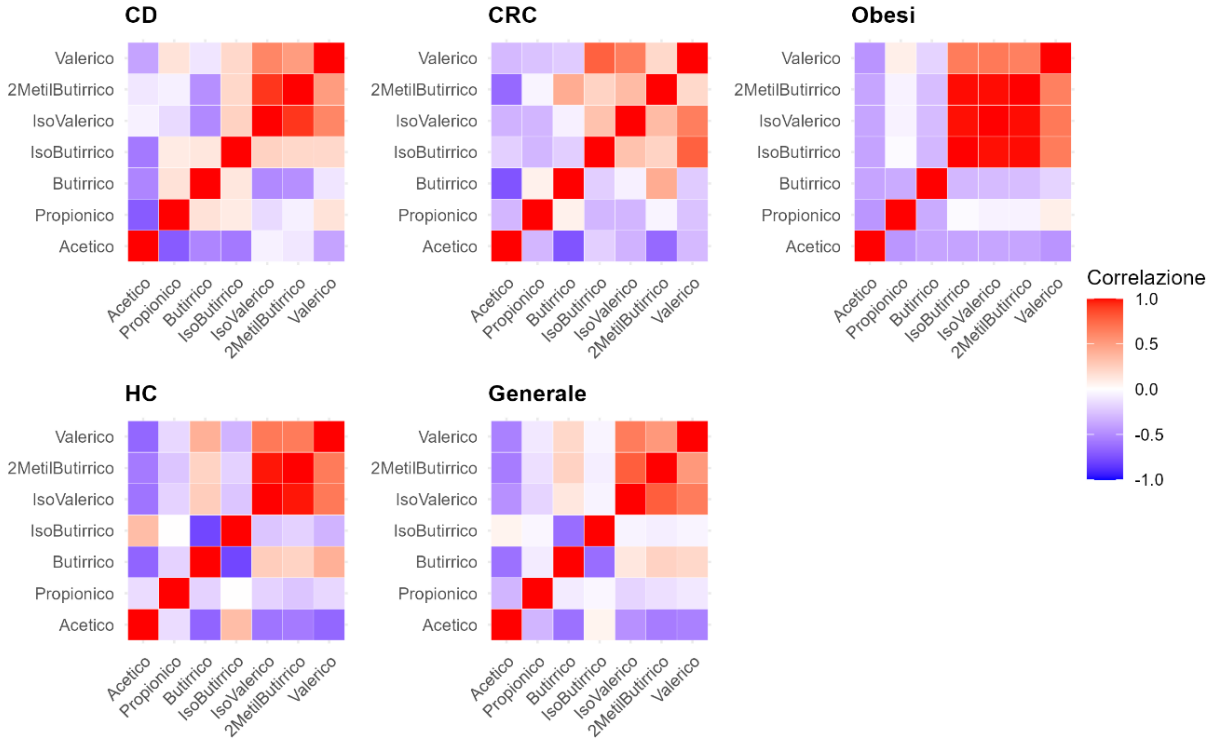

Figura 2.9: Heatmap generale e per ogni malattia per il campione fecale

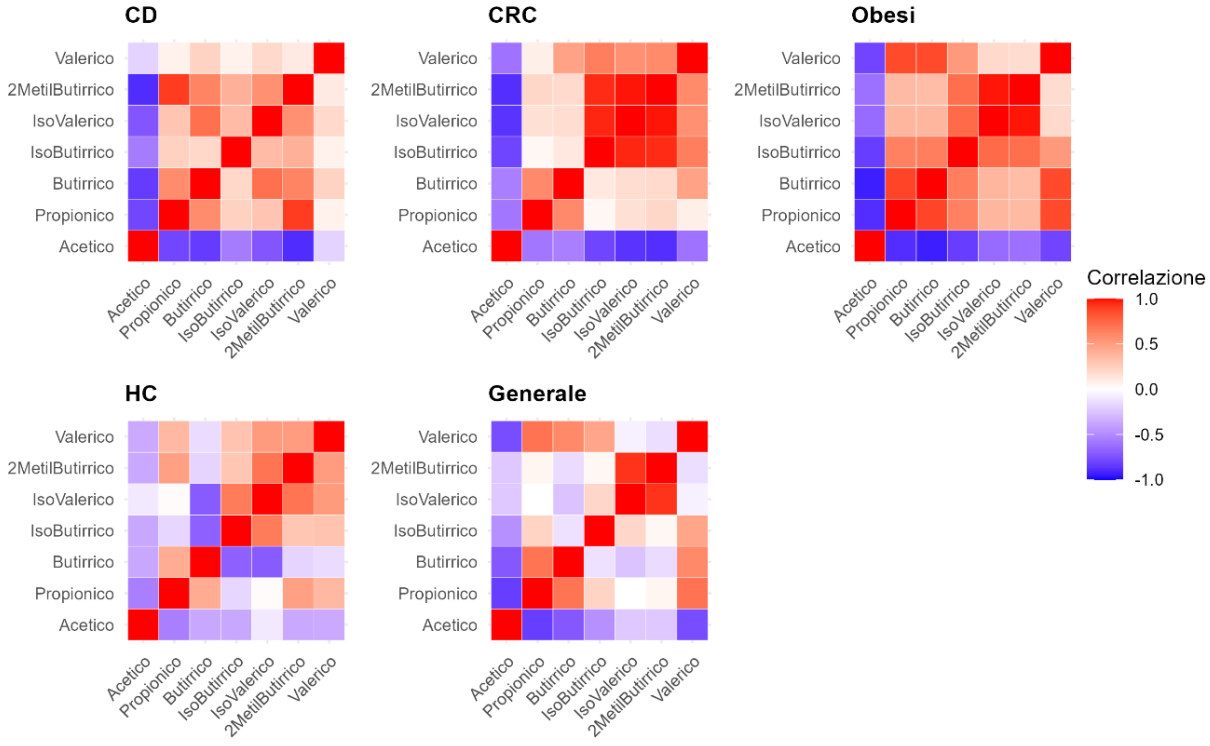

Figura 2.10: Heatmap generale e per ogni malattia per il campione plasmatico

Tramite le *heatmap* è possibile visualizzare in modo immediato le associazioni positive o negative tra gli SCFA tramite i colori che indicano la forza e la direzione.

Osservando prima le *heatmap* generali, emerge che nel campione fecale l'acido Acetico ha correlazioni debolmente negative con quasi tutti gli altri acidi, mentre nel campione plasmatico le correlazioni negative diventano più forti, in modo particolare con gli acidi Propionico, Butirrico e Valerico. Inoltre nel campione fecale si osserva che gli acidi IsoValerico, Valerico e 2MetilButirrico tendono ad avere correlazioni positive abbastanza forti tra loro; mentre gli acidi IsoButirrico e Butirrico hanno una forte correlazione negativa solo tra di loro. Nel campione plasmatico invece sono presenti correlazioni positive tra gli acidi Propionico e Butirrico, tra Propionico e Valerico e, più debolmente, tra Propionico e IsoButirrico e tra Valerico e sia Butirrico che IsoButirrico; molto forte è la correlazione positiva tra gli acidi IsoValerico e 2MetilButirrico.

Si prosegue analizzando le *heatmap* per i gruppi clinici, cominciando dal campione fecale. Nel gruppo CD l'acido Acetico è correlato negativamente quasi con tutti gli altri acidi, soprattutto con gli acidi Propionico, Butirrico, IsoButirrico e Valerico; gli altri acidi sono correlati debolmente tra loro, tranne gli acidi IsoValerico, 2MetilButirrico e Valerico correlati tra loro positivamente in modo più forte. Nel gruppo CRC le correlazioni tra gli acidi sono deboli o assenti, ad eccezione degli acidi Acetico e Butirrico e di Acetico e 2MetilButirrico tra i quali sono presenti correlazioni fortemente negative, mentre tra gli acidi Valerico e IsoButirrico e tra Valerico e IsoValerico si osservano correlazioni positive. Nel gruppo Obesi l'acido Acetico è correlato negativamente con tutti gli altri acidi, allo stesso modo anche l'acido Butirrico è correlati negativamente, seppur un po' più debolmente, con gli altri acidi. Al contrario, gli acidi Valerico, 2MetilButirrico, IsoValerico e IsoButirrico sono fortemente correlati tra loro con associazione positiva. Nel gruppo dei controlli sani l'acido Acetico ha correlazioni negative solo con gli acidi Valerico, 2MetilButirrico, IsoValerico e Butirrico,; è presente una forte correlazione tra gli acidi Butirrico e IsoButirrico. Le altre correlazioni sono molto deboli, quasi assenti, anche se rimane una correlazione positiva tra gli acidi Valerico, 2MetilButirrico e IsoValerico.

Successivamente si osservano le *heatmap* del campione plasmatico per i diversi gruppi. In CD l'acido Acetico è fortemente correlato in modo negativo con tutti gli altri acidi tranne che con l'acido Valerico; tutti gli altri acidi sono debolmente correlati in modo positivo tra loro, tranne gli acidi 2MetilButirrico e Propionico e gli acidi IsoValerico e Butirrico. Nel gruppo CRC: l'acido Acetico è fortemente correlato in modo negativo con tutti gli altri acidi soprattutto con

gli acidi 2MetilButirrico, IsoValerico e IsoButirrico; sono invece fortemente correlati in modo positivo gli acidi IsoButirrico, IsoValerico e 2MetilButirrico, ma anche gli acidi Butirrico e Propionico. L'acido Valerico ha correlazioni positive con tutti gli altri acidi tranne che con gli acidi Valerico e Propionico. Anche nel gruppo Obesi l'acido Acetico è fortemente correlato in modo negativo con tutti gli altri acidi; gli altri acidi hanno tutti correlazioni positive tra loro, in modo particolare tra Valerico e Propionico, tra Valerico e Butirrico, tra 2MetilButirrico e IsoValerico e tra Butirrico e Propionico. Nei controlli sani le correlazioni tra gli acidi sono presenti sia positive che negative ma in modo debole, tranne che tra gli acidi IsoValerico e Butirrico e tra IsoButirrico e Butirrico, tra i quali è presente correlazione negativa.

Si prosegue con un analisi delle correlazioni tra gli acidi grassi, separatamente per ogni gruppo clinico, per esplorare l'associazione tra le concentrazioni degli SCFA nei campioni plasmatico e fecale. Le correlazioni vengono calcolate utilizzando due coefficienti di correlazione, quello di Pearson e quello di Spearman. Nella Tabella 2.1, nella Tabella 2.2, nella Tabella 2.3 e nella Tabella 2.4 sono riportati i risultati ottenuti per ogni gruppo.

Tabella 2.1: Coefficienti di correlazione per il gruppo CD

	Acetico	Propionico	Butirrico	IsoButirrico	IsoValerico	2MetilBut	Valerico
PEARSON	-0.0258	0.0764	0.0352	-0.0044	0.0317	-0.0964	-0.0424
SPEARMAN	-0.0403	-0.1409	-0.2070	0.1625	0.2365	0.1515	0.0448

Tabella 2.2: Coefficienti di correlazione per il gruppo CRC

	Acetico	Propionico	Butirrico	IsoButirrico	IsoValerico	2MetilBut	Valerico
PEARSON	0.1665	-0.2183	-0.1419	-0.1433	-0.1460	0.3139	-0.1156
SPEARMAN	0.1662	-0.0474	0.2236	-0.0719	-0.0514	0.3268	-0.1190

Tabella 2.3: Coefficienti di correlazione per il gruppo Obesi

	Acetico	Propionico	Butirrico	IsoButirrico	IsoValerico	2MetilBut.	Valerico
PEARSON	-0.0596	**0.4668**	-0.0324	-0.1895	0.0679	0.0789	-0.3121
SPEARMAN	-0.0338	**0.4459**	-0.0531	-0.2600	-0.0849	-0.0702	-0.2807

Tabella 2.4: Coefficienti di correlazione per il gruppo HC

	Acetico	Propionico	Butirrico	IsoButirrico	IsoValerico	2MetilBut.	Valerico
PEARSON	0.0730	0.1599	**0.5446**	**0.5427**	-0.3347	-0.2561	-0.2041
SPEARMAN	0.0303	0.1832	**0.4491**	0.2329	-0.3343	-0.1670	-0.1887

Nei gruppi CD, CRC e in parte anche Obesi, i coefficienti di correlazione risultano prossimi allo zero o si discostano di poco. Questa scarsa correlazione tra i due campioni è probabilmente

dovuta a modifiche nel microbiota intestinale o disfunzioni del metabolismo dovute alla patologia. I valori di correlazione più elevati in valore assoluto si osservano nel gruppo dei controlli sani, suggerendo una maggiore coerenza tra le concentrazioni plasmatiche e fecali degli acidi grassi. Questo potrebbe riflettere una migliore corrispondenza tra gli SCFA prodotti a livello intestinale e quelli che poi vengono assorbiti e circolano nel plasma. Tuttavia in realtà nessuna di queste correlazioni raggiunge valori elevati in valore assoluto da poter affermare l'esistenza di una relazione significativa tra i due campioni. Tali risultati quindi indicano l'assenza di una chiara relazione biunivoca tra le concentrazioni degli acidi grassi nei due campioni.

2.2 Analisi dei cluster

L'obiettivo dell'analisi dei cluster, o *Cluster Analysis*, è quello di suddividere le osservazioni in gruppi che siano il più possibile omogenei al loro interno e il più possibile eterogenei tra loro. In questo caso si valuta se i gruppi individuati con tale analisi, sulla base delle concentrazioni degli SCFA, corrispondono ai gruppi generati dalla variabile Malattia. Se la corrispondenza è buona, significa che gli acidi grassi sono informativi e riescono a discriminare tra le varie malattie.

In fase di analisi sono stati realizzati diversi dendrogrammi (rappresentazioni grafiche delle aggregazioni progressive) applicando vari metodi agglomerativi. Il più capace tra essi a discriminare in gruppi è il metodo di Ward. Si mostra quindi in Figura 2.11 il dendrogramma ottenuto con il metodo di Ward applicato ai dati congiunti, con l'aggiunta della divisione a un'altezza in modo tale da formare quattro gruppi, uno per ogni malattia. Inoltre si riporta in Tabella 2.5 la rispettiva matrice di confusione, la quale contiene sulla diagonale principale le osservazioni correttamente classificate e fuori dalla diagonale le classificazioni errate. L' analisi è affiancata dai criteri di performance - Specificità, Sensibilità, Valore Predittivo Positivo (PPV) e Valore Predittivo Negativo (NPV) - in Tabella 2.6, i quali aiutano a comprendere se la classificazione è avvenuta correttamente.

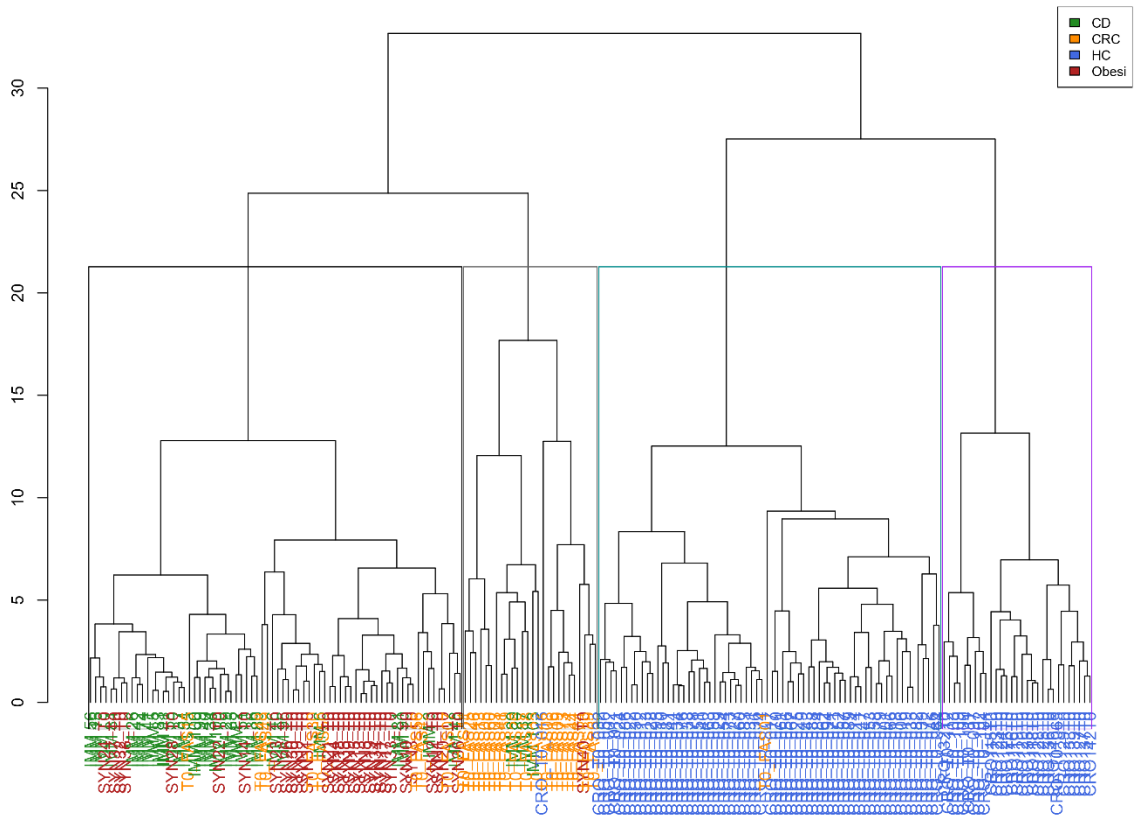

Figura 2.11: Dendrogramma ottenuto con il metodo di Ward

Tabella 2.5: Matrice di confusione ottenuta con il metodo di Ward

		Gruppo osservato			
		CD	CRC	HC	Obesi
Gruppo assegnato con il metodo di Ward	CD	1	1	64	0
	CRC	0	24	1	1
	HC	0	0	29	0
	Obesi	34	5	1	32

Tabella 2.6: Criteri di performance ottenuti con il metodo di Ward

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.029	0.800	0.305	0.970
SPECIFICITÀ	0.589	0.988	1.000	0.750
POS PRED VALUE	0.015	0.923	1.000	0.444
NEG PRED VALUE	0.732	0.946	0.598	0.992

Il gruppo meglio classificato è CRC con 24 su 30 soggetti correttamente assegnati; anche il gruppo Obesi è identificato bene con 32 su 33 corretti, nonostante presenti qualche falso positivo provenienti dal gruppo CD. Tali gruppi infatti mostrano performance molto buone con

sensibilità elevate; tuttavia per Obesi si osserva un PPV basso, sottolineando come già notato che tra i soggetti assegnati a tale gruppo sono incluse anche molte unità appartenenti ad altri gruppi, soprattutto provenienti da CD. I gruppi HC e CD sono invece riconosciuti molto peggio. Si osserva, infatti, che il gruppo CD presenta valori di sensibilità e PPV molto bassi, indicando che i soggetti appartenenti a CD vengono quasi sempre classificati erroneamente, venendo attribuiti principalmente a Obesi. Anche il gruppo HC mostra una bassa sensibilità, ma al contrario specificità e PPV perfetti, il che significa che i pochi identificati come appartenenti a tale gruppo sono effettivamente corretti.

Nel complesso, il metodo di Ward applicato ai dati provenienti da entrambi i campioni, produce risultati contrastanti, risultando discretamente buono per alcuni gruppi, come CRC e Obesi, ma sbagliato e inadeguato per altri, come CD.

Si esegue nuovamente l'analisi dei gruppi separatamente per i campioni fecale e plasmatico.

In Figura 2.12 si osserva il dendrogramma ottenuto con il metodo di Ward per il campione plasmatico, con l'aggiunta della divisione a un'altezza da formare quattro gruppi, uno per ogni malattia; nella Tabella 2.7 è riportata la matrice di correlazione, mentre in Tabella 2.8 si trovano i rispettivi criteri di performance.

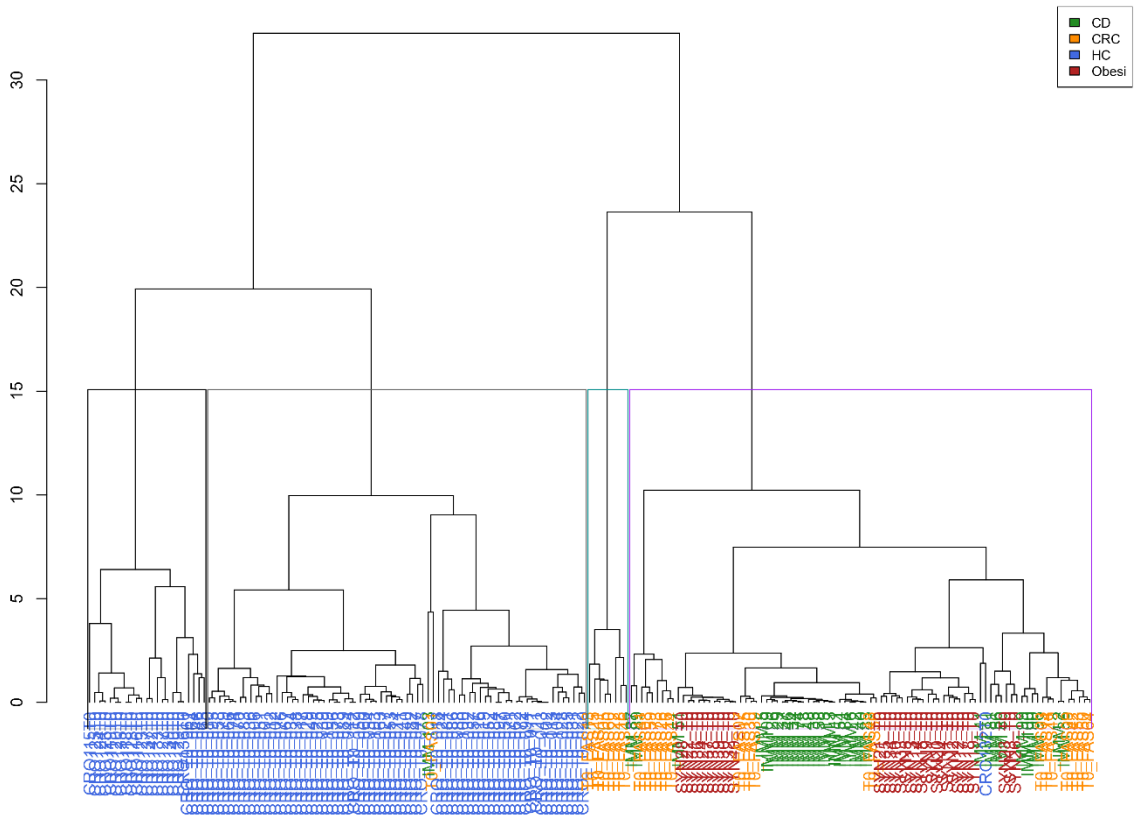

Figura 2.12: Dendrogramma ottenuto con il metodo di Ward dal campione plasmatico

Tabella 2.7: Matrice di confusione ottenuta con il metodo di Ward dal campione plasmatico

		Gruppo osservato			
		CD	CRC	HC	Obesi
Gruppo assegnato con il metodo di Ward	CD	34	21	1	33
	CRC	0	8	0	0
	HC	1	1	71	0
	Obesi	0	0	23	0

Tabella 2.8: Criteri di performance ottenuti con il metodo di Ward dal campione plasmatico

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.971	0.267	0.747	0.000
SPECIFICITÀ	0.652	1.000	0.980	0.856
POS PRED VALUE	0.382	1.000	0.973	0.000
NEG PRED VALUE	0.990	0.881	0.800	0.806

Il gruppo HC mostra una buona identificazione, con 71 soggetti correttamente assegnati, i criteri di performance registrano infatti buoni risultati per tale gruppo, indicando che è quello meglio riconosciuto dal metodo. Il gruppo CD ha 34 assegnazioni corrette su 89, ma viene fortemente confuso con Obesi e CRC; ha sensibilità molto alta suggerendo che i soggetti di quel gruppo sono quasi sempre assegnati ma non sempre al gruppo esatto. Per il gruppo Obesi il modello ha pessime prestazioni: nessun soggetto viene identificato correttamente, né alcun soggetto viene assegnato a tale gruppo, quasi tutti i soggetti appartenenti a quel gruppo vengono assegnati erroneamente a CD. Il campione plasmatico non è sufficiente a discriminare correttamente le patologie considerate.

In Figura 2.13 si osserva il dendrogramma ottenuto con il metodo di Ward per il campione fecale, con l'aggiunta della divisione a un'altezza da formare quattro gruppi, uno per ogni malattia; nella Tabella 2.9 è riportata la matrice di correlazione, mentre in Tabella 2.10 si trovano i rispettivi criteri di performance.

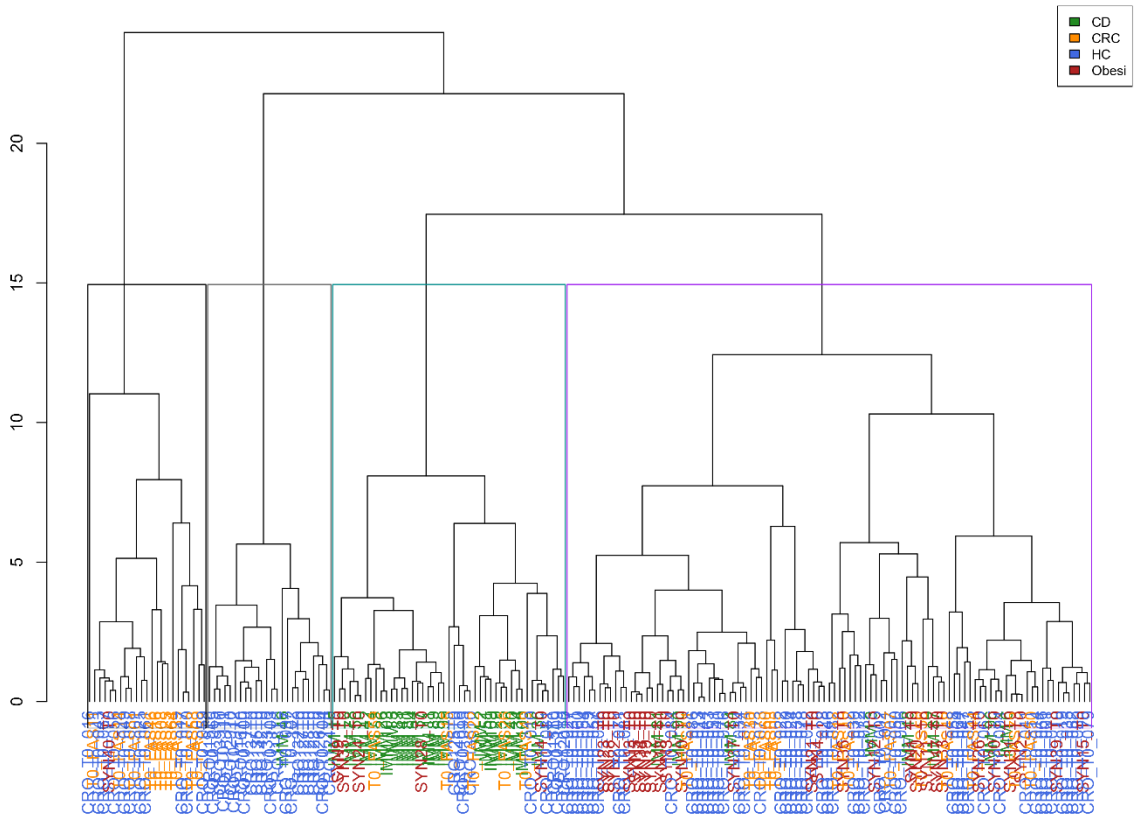

Figura 2.13: Dendrogramma ottenuto con il metodo di Ward dal campione fecale

Tabella 2.9: Matrice di confusione ottenuta con il metodo di Ward dal campione fecale

		Gruppo osservato			
		CD	CRC	HC	Obesi
Gruppo assegnato con il metodo di Ward	CD	26	5	9	5
	CRC	0	11	11	1
	HC	1	0	23	0
	Obesi	8	14	52	27

Tabella 2.10: Criteri di performance ottenuti con il metodo di Ward dal campione fecale

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.743	0.367	0.242	0.818
SPECIFICITÀ	0.880	0.926	0.990	0.537
POS PRED VALUE	0.578	0.478	0.958	0.267
NEG PRED VALUE	0.939	0.888	0.574	0.935

Nel caso del campione fecale il metodo di Ward riesce a ottenere per il gruppo Obesi una buona identificazione, con 27 soggetti correttamente assegnati: ha una sensibilità buona, indicando che le unità di tale gruppo sono ben riconosciute, anche se in tale gruppo il metodo inserisce

anche molte unità non provenienti dal gruppo Obesi. Il gruppo HC ha 23 soggetti classificati correttamente ma con notevole dispersione e ampiamente confusi con Obesi: quando il metodo identifica un soggetto come HC è quasi sempre corretto, ma ciò accade solamente con poche unità poiché allo stesso tempo si ha anche un numero elevato di falsi negativi. Il gruppo CD mostra sovrapposizione in particolare con i gruppi HC e Obesi, anche se la maggior parte dei veri soggetti viene riconosciuta ed è affidabile anche nel dire chi non appartiene a quel gruppo.

Le tabelle confermano quindi che dati fecali hanno una capacità discriminante diversa da quelli plasmatici, ma che nessuno dei due campioni è ottimale per distinguere tutti i gruppi.

2.3 Conclusioni analisi esplorative

Le analisi esplorative hanno evidenziato importanti differenze sia nella composizione che nella distribuzione degli acidi grassi a catena corta tra i gruppi clinici. Nel campione fecale i gruppi CRC e HC mostrano distribuzioni simili per la maggior parte degli acidi, mentre nei campioni plasmatici si osserva spesso un comportamento opposto dei due gruppi: quando uno dei due mostra una distribuzione ampia, l'altro presenta concentrazioni più compatte. In generale nel plasma si nota una maggiore variabilità complessiva, mentre le distribuzioni nei campioni fecali risultano più regolari.

I due campioni differiscono anche per le correlazioni tra gli SCFA, infatti ad esempio in quello fecale l'acido Acetico presenta correlazioni deboli con gli altri acidi, mentre nel plasma mostra forti correlazioni negative, in particolare con il Propionico, il Butirrico e il Valerico. In generale le associazioni tra acidi sono consistenti all'interno di ciascun campione, anche se i due variano per quali sono gli acidi maggiormente correlati. Complessivamente nel gruppo dei soggetti sani le correlazioni risultano più forti, suggerendo una maggiore regolarità nelle concentrazioni degli SCFA, al contrario nei soggetti affetti dalle patologie le correlazioni tendono ad essere più deboli.

L'analisi dei cluster condotta con il metodo di Ward non ha dato buoni risultati: ha mostrato una discreta capacità di individuare alcuni gruppi, ma inadeguatezza per altri; inoltre nessuno dei due campioni singolarmente si è rivelato pienamente efficace nel discriminare tutti i gruppi clinici. È tuttavia emerso chiaramente che plasma e feci forniscono informazioni complementari, suggerendo che la capacità discriminante degli SCFA varia tra i due campioni.

Capitolo 3 - Tecniche statistiche

In questo capitolo vengono applicate alcune tecniche statistiche inferenziali e predittive con l'obiettivo di approfondire l'analisi dei dati e valutare la presenza di differenze significative tra i gruppi clinici e la capacità discriminante degli acidi grassi a catena corta.

3.1 Test di Kruskal-Wallis e Test di Dunn

Il test di Kruskal-Wallis serve per confrontare le distribuzioni delle concentrazioni degli SCFA nei quattro gruppi clinici. Nel caso in cui le variabili sono continue, ma non è possibile assumere distribuzione normale, si utilizza tale test come procedura alternativa al test F per l'analisi della varianza, per il quale invece la condizione di normalità sarebbe necessaria.

Con questo test, separatamente per i due campioni plasmatico e fecale, per ogni acido grasso si verifica l'ipotesi nulla che i quattro gruppi siano identici, contro l'ipotesi alternativa che almeno un gruppo differisca in media per distribuzione delle concentrazioni degli SCFA dagli altri. I risultati del test sono riportati nella Tabella 3.1.

Tabella 3.1: Risultati del test di Kruskal-Wallis per campioni

	PLASMATICO		FECALE	
SCFA	STATISTICA TEST	*p-value*	STATISTICA TEST	*p-value*
Acetico	121.96	**2.92e-26** ***	38.78	**1.93e-08** ***
Propionico	139.19	**5.64e-30** ***	10.61	**0.014** *
Butirrico	118.90	**1.33e-25** ***	6.89	0.075
IsoButirrico	45.60	**6.89e-10** ***	34.03	**1.96e-07** ***
IsoValerico	68.93	**7.24e-15** ***	8.05	**0.045** *
2MetilButirrico	97.13	**6.43e-21** ***	30.04	**1.35e-06** ***
Valerico	152.04	**9.54e-33** ***	13.31	**0.004** **

Per il campione plasmatico il test è altamente significativo per tutti gli acidi grassi, concludendo che le distribuzioni delle concentrazioni degli SCFA nel campione plasmatico variano significativamente tra i gruppi clinici. Questi risultati indicano una forte capacità discriminante degli acidi grassi a catena corta nel plasma. I risultati per il campione fecale mostrano una significatività complessivamente inferiore rispetto al plasma, in particolare per l'acido Butirrico il test non risulta neanche significativo, indicando che per tale campione le

differenze tra i gruppi non sono sufficientemente marcate e che quindi gli SCFA hanno una ridotta capacità discriminante.

Il test però ha la limitazione di non indicare di fatto quali gruppi differiscono tra loro, ma solo che almeno uno dei gruppi presenta una distribuzione significativamente diversa rispetto agli altri. Un *p-value* significativo in realtà non garantisce che tutti i gruppi differiscano tra loro, ma può derivare da una differenza marcata di un solo gruppo. Per questo motivo, nel caso in cui il test di Kruskal-Wallis risulta significativo, si procede con il test di Dunn che permette di determinare quali sono i gruppi diversi, confrontando a coppie ciascun gruppo.

Nel caso del campione plasmatico il test risulta significativo per tutti gli acidi, per cui il test di Dunn viene eseguito per tutti gli SCFA, per il campione fecale invece viene eseguito per tutti escluso l'acido Butirrico per il quale il testi di Kruskall-Wallis non risulta significativo. Il test viene eseguito per ciascuno dei due campioni e i risultati sono riportati nella Tabella 3.2 e nella Tabella 3.3.

Tabella 3.2: Risultati del test di Dunn per il campione plasmatico

SCFA	Coppia di confronto	Statistica test Z	*p-value* non aggiustato	*p-value* aggiustato
Acetico	CD - CRC	2.849	**0.004** **	**0.026** *
	CD - HC	9.596	**8.34e-22** ***	**5.00e-21** ***
	CRC - HC	5.675	**1.38e-08** ***	**8.30e-08** ***
	CD - Obesi	1.784	0.074	0.446
	CRC - Obesi	-1.093	0.274	1.000
	HC - Obesi	-7.247	**4.25e-13** ***	**2.55e-12** ***
Propionico	CD - CRC	-1.415	0.157	0.943
	CD - HC	-9.997	**1.57e-23** ***	**9.41e-23** ***
	CRC - HC	-7.758	**8.62e-15** ***	**5.17e-14** ***
	CD - Obesi	-2.598	**0.009** **	0.056
	CRC - Obesi	-1.104	0.270	1.000
	HC - Obesi	6.663	**2.69e-11** ***	**1.613e-10** ***
Butirrico	CD - CRC	-0.269	0.788	1.000
	CD - HC	-8.952	**3.49e-19** ***	**2.09e-18** ***
	CRC - HC	-8.133	**4.20e-16** ***	**2.52e-15** ***
	CD - Obesi	-3.245	**0.001** **	**0.007** **
	CRC - Obesi	-2.856	**0.004** **	**0.026** *
	HC - Obesi	4.863	**1.15e-06** ***	**6.92e-06** ***

IsoButirrico	CD - CRC	-1.046	0.296	1.000
	CD - HC	-5.459	**4.79e-08** ***	**2.87e-07** ***
	CRC - HC	-3.912	**9.16e-05** ***	**5.50e-04** ***
	CD - Obesi	-5.205	**1.94e-07** ***	**1.16e-06** ***
	CRC - Obesi	-3.975	**7.04e-05** ***	**4.22e-04** ***
	HC - Obesi	-0.908	0.364	1.000
IsoValerico	CD - CRC	-6.649	**2.95e-11** ***	**1.77e-10** ***
	CD - HC	-3.579	**3.45e-04** ***	**0.003** **
	CRC - HC	4.520	**6.18e-06** ***	**3.78e-05** ***
	CD - Obesi	0.898	0.369	1.000
	CRC - Obesi	7.421	**1.16e-13** ***	**6.95e-13** ***
	HC - Obesi	4.580	**4.64e-06** ***	**2.78e-05** ***
2MetilButirrico	CD - CRC	-6.167	**6.97e-10** ***	**4.18e-09** ***
	CD - HC	-3.251	**0.001** **	**0.009** **
	CRC - HC	4.257	**2.07e-05** ***	**1.25e-04** ***
	CD - Obesi	3.291	**9.99e-04** ***	**0.006** **
	CRC - Obesi	9.247	**2.30e-20** ***	**1.38e-19** ***
	HC - Obesi	7.133	**9.84e-13** ***	**5.90e-12** ***
Valerico	CD - CRC	-0.536	0.592	1.000
	CD - HC	-10.145	**3.49e-24** ***	**2.09e-23** ***
	CRC - HC	-8.942	**3.83e-19** ***	**2.30e-18** ***
	CD - Obesi	-3.108	**0.002** **	**0.011** *
	CRC - Obesi	-2.461	**0.014** *	0.083
	HC - Obesi	6.196	**5.80e-10** ***	**3.48e-09** ***

Tabella 3.3: Risultati del test di Dunn per il campione fecale

SCFA	Coppia di confronto	Statistica test Z	*p-value* non aggiustato	*p-value* aggiustato
Acetico	CD - CRC	5.679	**1.35e-08** ***	**8.12e-08** ***
	CD - HC	5.354	**8.61e-08** ***	**5.16e-07** ***
	CRC - HC	-1.692	0.091	0.544
	CD - Obesi	3.993	**6.54e-05** ***	**3.92e-04** ***
	CRC - Obesi	-1.761	0.078	0.469
	HC - Obesi	-0.445	0.656	1.000

PROPIONICO	CD - CRC	-1.752	0.080	0.478
	CD - HC	-1.597	0.110	0.661
	CRC - HC	0.573	0.566	1.000
	CD - Obesi	-3.200	**0.001****	**0.008****
	CRC - Obesi	-1.350	0.177	1.000
	HC - Obesi	-2.279	**0.023**	0.136
ISOBUTIRRICO	CD - CRC	-2.189	**0.029***	0.172
	CD - HC	-4.921	**8.61e-07*****	**5.17e-06*****
	CRC - HC	-2.045	**0.041***	0.245
	CD - Obesi	-0.421	0.674	1.000
	CRC - Obesi	1.755	0.079	0.476
	HC - Obesi	4.311	**1.63e-05*****	**9.77e-05*****
ISOVALERICO	CD - CRC	-2.760	**0.006****	**0.035****
	CD - HC	-1.772	0.076	0.459
	CRC - HC	1.606	0.108	0.649
	CD - Obesi	-0.897	0.370	1.000
	CRC - Obesi	1.859	0.063	0.378
	HC - Obesi	0.657	0.511	1.000
2METILBUTIRRICO	CD - CRC	-4.991	**6.01e-07*****	**3.61e-06*****
	CD - HC	-1.356	0.175	1.000
	CRC - HC	4.649	**3.33e-06*****	**2.00e-05*****
	CD - Obesi	-0.637	0.524	1.000
	CRC - Obesi	4.310	**1.63e-05*****	**9.79e-05*****
	HC - Obesi	0.562	0.574	1.000
VALERICO	CD - CRC	-2.599	**0.009****	0.056
	CD - HC	-3.483	**4.95e-04*****	**0.003****
	CRC - HC	-0.202	0.840	1.000
	CD - Obesi	-2.832	**0.005****	**0.028***
	CRC - Obesi	-0.161	0.872	1.000
	HC - Obesi	0.008	0.994	1.000

Nel campione plasmatico il gruppo HC si distingue in modo statisticamente significativo sia dal gruppo CD che dal gruppo CRC per tutti gli acidi grassi a catena corta, e da Obesi per tutti gli SCFA ad eccezione dell'Acido IsoButirrico. Questo suggerisce che le concentrazioni nel campione plasmatico consentono una buona discriminazione i soggetti sani dai gruppi clinici.

Al contrario, la capacità di riconoscere le patologie risulta più variabile a seconda dell'acido grasso considerato: in particolare, i gruppi Obesi e CD mostrano le maggiori sovrapposizioni con gli altri gruppi. L'Acido 2MetilButirrico è l'unico tra gli SCFA ad evidenziare differenze significative in tutte le coppie di confronto.

Nel campione fecale le differenze tra gruppi sono notevolmente meno significative rispetto a quelle osservate nel plasma. L'Acido Acetico riesce bene a discriminare il gruppo CD dagli altri, mentre l'Acido 2MetilButirrico ha capacità nel separare il gruppo dei CRC dagli altri; l'Acido IsoButirrico è significativo nel riconoscere HC da CD e Obesi ma lo confonde con CRC. Gli altri acidi sono significativi solo per qualche coppia e in modo più debole.

Nonostante il campione plasmatico sia il più informativo per discriminare tra i diversi gruppi, soprattutto per riconoscere i controlli sani rispetto agli altri casi clinici, anche il campione fecale mostra differenze tra gruppi rilevanti per specifici acidi.

3.2 Modello di regressione logistica multinomiale

La regressione logistica multinomiale viene utilizzata quando la variabile di risposta è categoriale con più di due livelli, i quali suddividono le osservazioni in gruppi distinti. Nel caso in esame, il modello permette di stimare la probabilità che un'unità appartenga a uno dei gruppi definiti dalla variabile Malattia: HC, CRC, CD e Obesi.

Le variabili indipendenti (o predittive), in funzione delle quali viene stimata la probabilità, sono: Età, Sesso e i livelli di concentrazione dei sette acidi grassi. La classe della variabile di risposta Malattia base-line è HC, per ciascun gruppo $j \neq \text{HC}$ stima la seguente equazione logit:

$$\log\left(\frac{P(Y_i = j|X_{i1} = x_{i1}, \dots, X_{i9} = x_{i9})}{P(Y_i = HC|X_{i1} = x_{i1}, \dots, X_{i9} = x_{i9})}\right) = \beta_{0j} + \beta_{1j}X_{i1} + \beta_{2j}X_{i2} + \cdots + \beta_{9j}X_{i9}$$

dove Y_i rappresenta la variabile Malattia per l'osservazione i-esima e $X_{i1}, \dots, X_{i9}$ sono le variabili predittive associate al soggetto i-esimo. I coefficienti $\beta_{0j}, \beta_{1j}, \cdots, \beta_{9j}$ associati alle variabili indipendenti X_i descrivono l'effetto di ogni predittore sulla probabilità di appartenere a uno specifico gruppo clinico *j* rispetto al gruppo dei controlli sani; essi indicano in che misura una variazione unitaria della variabile associata, al netto delle altre, influenza il log-odds di appartenenza al gruppo *j* rispetto al gruppo base-line HC.

Tali coefficienti sono riportati nella Tabella 3.4 e nella Tabella 3.5, rispettivamente per il modello sul campione plasmatico e quello sul campione fecale. Ogni coefficiente è affiancato

dal *p-value* ad esso associato, il quale indica se la variabile indipendente *h* contribuisce in modo significativo a distinguere il gruppo *j* dal gruppo dei controlli sani HC.

Tabella 3.4: Risultati regressione logistica multinomiale per il campione plasmatico

	j= CD			*j*= Obesi			*j*= CRC		
h	$\hat{\beta}_{hj}$	$OR = e^{\hat{\beta}_{hj}}$	*p-value*	$\hat{\beta}_{hj}$	$OR = e^{\hat{\beta}_{hj}}$	*p-value*	$\hat{\beta}_{hj}$	$OR = e^{\hat{\beta}_{hj}}$	*p-value*
Intercept	-2.74	0.07	0.663	-3.63	0.03	0.086	-2.14	0.12	0.822
Età	1.56	4.74	0.973	0.41	1.51	0.981	2.75	15.57	0.947
SessoM	11.89	1.46e+05	0.938	-34.69	8.57e-16	0.470	40.50	3.88e+17	0.807
Acetico	2.35	1.05	0.974	2.45	11.58	0.968	1.54	4.66	0.982
Propion	-12.83	2.68e-06	0.957	7.39	162	0.973	-16.24	8.87e-08	0.936
Butir	-0.60	0.55	0.998	-3.81	0.02	0.992	-7.42	5.99e-04	0.967
IsoBut	-20.44	1.33e-09	0.864	2.61	13.56	0.978	-38.08	2.90e-17	0.704
IsoVal	20.11	5.44e+08	0.956	63.18	2.75e+27	0.868	32.36	1.13e+14	0.904
2MetilB	20.11	5.41e+08	0.984	-92.03	1.08e-40	0.671	31.05	3.04e+13	0.976
Valerico	-282.29	2.54e-123	**0.000**	-342.57	1.67e-149	**0.000**	-217.33	4.10e-95	**0.001**

Tabella 3.5: Risultati regressione logistica multinomiale per il campione fecale

	j = CD			j = Obesi			j = CRC		
h	$\hat{\beta}_{hj}$	$OR = e^{\hat{\beta}_{hj}}$	*p-value*	$\hat{\beta}_{hj}$	$OR = e^{\hat{\beta}_{hj}}$	*p-value*	$\hat{\beta}_{hj}$	$OR = e^{\hat{\beta}_{hj}}$	*p-value*
Intercept	-0.01	0.99	**0.003**	-0.01	0.9901	**0.006**	-0.02	0.98	**0.000**
Età	-0.05	0.95	**0.003**	0.02	1.02	0.229	0.23	1.25	**0.000**
SessoM	-0.67	0.51	0.281	0.55	1.73	0.242	1.31	3.70	0.054
Acetico	0.11	1.11	**0.000**	0.01	1.00	0.932	-0.17	0.85	**0.000**
Propion	-0.01	0.99	0.842	0.08	1.08	0.063	-0.04	0.96	0.442
Butir	-0.12	0.89	**0.009**	-0.10	0.90	**0.006**	-0.22	0.80	**0.000**
IsoBut	-0.53	0.59	**0.001**	-0.54	0.58	**0.003**	-0.58	0.56	**0.035**
IsoVal	0.32	1.37	0.412	-0.02	0.98	0.941	-0.14	0.87	0.559
2MetilB	-0.31	0.73	0.493	-0.18	0.84	0.554	0.28	1.33	0.256
Valerico	-0.35	0.70	0.228	0.04	1.04	0.817	-0.23	0.79	0.407

Si analizzano innanzitutto i risultati della regressione logistica multinomiale per i dati del campione plasmatico. È importante che il *p-value* associato al coefficiente sia significativo poiché indica che l'effetto del predittore è rilevante per distinguere quel gruppo rispetto alla base-line, si commentano quindi solo gli effetti significativi.

In questo campione l'acido Valerico è l'unica variabile indipendente per ogni patologia ad avere coefficienti significativi, con valori fortemente negativi in tutti e tre i casi: ciò indica che un aumento della sua concentrazione riduce fortemente la probabilità di appartenenza ai gruppi clinici rispetto ai controlli sani.

Si prosegue osservando i risultati della regressione logistica multinomiale per i dati del campione fecale. L'effetto della variabile Età è significativo per i gruppi CD e CRC, con coefficiente negativo per il gruppo CD e positivo per CRC: l'aumento dell'età è associato ad un aumento del rischio di appartenenza a CRC piuttosto che ai controlli sani e a una diminuzione della probabilità di appartenenza a CD, è invece non significativo per il gruppo Obesi. Il coefficiente dell'acido Acetico risulta significativo e positivo nel gruppo CD, mentre nel gruppo CRC è significativo e negativo.

I coefficienti dell'acido Butirrico sono negativi e significativi per tutte le malattie, indicando che maggior presenza di acido Butirrico è associata a un minor rischio di appartenenza ai gruppi CD, CRC e Obesi piuttosto che ai controlli sani HC. L'acido IsoButirrico ha un comportamento analogo, ma solo nei gruppi CD e Obesi, registrando solo una debole significatività per il gruppo CRC.

Per un esempio pratico di interpretazione, si considera il coefficiente del gruppo CD associato all'acido Acetico $\left(\hat{\beta}_{\text{Acetico,CD}} = 0.107\right)$: l'odds ratio è $OR = \exp(0.107) = 1.113$, quindi per ogni incremento unitario della concentrazione percentuale di Acido Acetico si ha un aumento di 11.3% nelle probabilità relative, ovvero le odds che un soggetto appartenga al gruppo CD piuttosto che a HC aumentano dell'11.3%.

3.3 Cross-Validation

Per effettuare le previsioni con il modello di regressione logistica multinomiale, ovvero la riclassificazione delle unità, e valutare la capacità discriminante degli acidi grassi a catena corta, è stata utilizzata una strategia di validazione chiamata *cross-validation 10-fold*.

Questo approccio prevede di creare 10 sottoinsiemi del dataset di pari in modo casuale, dette *fold*. Ad ogni iterazione, si considera un *fold* come *test set* e le unità appartenenti ai restanti nove *fold* come *train set*. Il modello viene stimato sulle unità del *train set* e successivamente è impiegato per prevedere la classe di appartenenza di ciascuna unità nel *test set*. In questo modo, ogni osservazione viene utilizzata una sola volta per l'addestramento e nove volte per la validazione classificazione.

Alla fine del processo, tutte le predizioni ottenute nei 10 *fold* vengono aggregate per costruire una matrice di confusione complessiva, dalla quale è possibile ricavare i criteri di performance.

Questa strategia consente di ridurre il rischio di *overfitting* poiché ogni previsione viene effettuata su dati indipendenti da quelli usati per l'addestramento del modello; allo stesso tempo garantisce che tutte le unità del campione vengano utilizzate per l'addestramento del modello e successivamente per la previsione. L'intero processo è stato eseguito separatamente per i dati plasmatici e per quelli fecali, al fine di valutare distintamente la capacità predittiva dei due campioni.

Si confrontano le classi predette con le etichette reali tramite le matrici di confusione e l'analisi dei relativi criteri di performance, nelle Tabella 3.6 e nella Tabella 3.7 per il campione plasmatico e nella Tabella 3.8 e nella Tabella 3.9 per il campione fecale.

Tabella 3.6: Matrice di confusione del campione plasmatico

		Gruppo osservato			
		CD	CRC	HC	Obesi
Gruppo assegnato dal modello di regressione	CD	30	3	0	1
	CRC	3	26	1	0
	HC	1	1	92	1
	Obesi	1	0	2	31

Tabella 3.7: Criteri di performance del modello di regressione nel campione plasmatico

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.8571	0.8667	0.9684	0.9394
SPECIFICITÀ	0.9747	0.9755	0.9694	0.9812
POS PRED VALUE	0.8824	0.8667	0.9684	0.9118
NEG PRED VALUE	0.9686	0.9755	0.9694	0.9874

Nel campione plasmatico per tutti i gruppi la classificazione è complessivamente molto accurata, con quasi tutte le unità riconosciute correttamente e solo pochi soggetti classificati erroneamente; anche osservando i criteri di *performace* si hanno prestazioni molto elevate. I risultati suggeriscono che i livelli di concentrazione degli acidi grassi nel plasma sono informativi e ben discriminatori tra i gruppi.

Tabella 3.8: Matrice di confusione del campione fecale

		Gruppo osservato			
		CD	CRC	HC	Obesi
Gruppo assegnato dal modello di regressione	CD	27	1	7	4
	CRC	2	20	9	2
	HC	4	7	75	19
	Obesi	2	2	4	8

Tabella 3.9: Criteri di performance del modello di regressione nel campione fecale

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.7714	0.6667	0.7895	0.2424
SPECIFICITÀ	0.9241	0.9202	0.6939	0.9500
POS PRED VALUE	0.6923	0.6061	0.7143	0.5000
NEG PRED VALUE	0.9481	0.9375	0.7727	0.8588

Per il campione fecale emerge che il gruppo meglio classificato è CD con 27 su 35 soggetti correttamente assegnati; anche il gruppo HC mostra una buona classificazione con 75 corretti, ma 19 soggetti obesi vengono assegnati a HC erroneamente, suggerendo sovrapposizione tra soggetti sani e Obesi. Quest'ultimo è infatti il gruppo con peggiore riclassificazione con solo 8 soggetti su 33 correttamente classificati.

Per valutare la capacità discriminante complessiva degli acidi grassi a catena corta è stato applicato il processo di *cross-validation* 10-*fold* descritto in precedenza anche sul dataset contenente complessivamente i valori relativi ai due campioni. Nella Tabella 3.10 e nella Tabella 3.11 sono presentate rispettivamente la matrice di confusione aggregata e i criteri di performance.

Tabella 3.10: Matrice di confusione dei dati combinati

		Gruppo osservato			
		CD	CRC	HC	Obesi
Gruppo assegnato dal modello di regressione	CD	27	3	0	2
	CRC	3	26	1	0
	HC	1	1	93	0
	Obesi	4	0	1	31

Tabella 3.11: Criteri di performance del modello di regressione dei dati combinati

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.7714	0.8667	0.9789	0.9394
SPECIFICITÀ	0.9684	0.9755	0.9796	0.9688
POS PRED VALUE	0.8438	0.8667	0.9789	0.8611
NEG PRED VALUE	0.9503	0.9755	0.9796	0.9873

Il gruppo HC è quello meglio classificato con 93 su 95 soggetti correttamente identificati, ma anche i gruppi Obesi, CD e CRC sono classificati con ottima precisione, con solo lievi confusioni. Nel complesso l'utilizzo dei dati combinati aumenta la capacità discriminante del modello rispetto all'utilizzo dei singoli comparti. Ciò si può notare anche dai criteri di performance che sono nel complesso molto alti, superiori rispetto a quelle ottenute con i soli dati plasmatici o fecali.

3.4 Analisi discriminante lineare di Fisher

L'analisi discriminante lineare di Fisher (LDA) è una tecnica per classificare le osservazioni in gruppi, usata quando i gruppi di appartenenza sono noti a priori. Tale analisi si basa sull'identificazione di una combinazione lineare degli SCFA (variabili predittive), chiamata funzione discriminante, che massimizzi la separazione tra gruppi. Poiché le unità sono divise in 4 gruppi, si generano al massimo 3 funzioni discriminanti, chiamate LD1, LD2 e LD3, ordinate in base alla quantità di varianza tra gruppi che spiegano. I coefficienti associati a ciascun SCFA nelle funzioni discriminanti rappresentano i pesi con cui ogni acido grasso contribuisce alla funzione. Se il segno del coefficiente è positivo allora all'aumentare della concentrazione dell'acido grasso associato aumenta anche il valore della funzione, viceversa se negativo. Più il coefficiente ha un valore elevato in valore assoluto, maggiore è il contributo dell'acido grasso nella distinzione tra i gruppi. I coefficienti di tali funzioni sono riportati nella Tabella 3.12 sia per il campione plasmatico che per quello fecale.

Tabella 3.12: Coefficienti delle funzioni discriminanti

	Plasma			Feci		
	LD1	LD2	LD3	LD1	LD2	LD3
Acetico	-0.0417	0.0077	-0.0043	-0.0754	-0.0073	0.0285
Propionico	0.0303	-0.0255	0.1255	-0.0206	0.0370	-0.1488
Butirrico	0.0562	-0.0066	0.0133	0.0209	0.0006	-0.0089
IsoButirrico	0.0345	0.0327	-0.0169	**0.1859**	-0.0549	0.0699
IsoValerico	-0.0228	-0.2161	1.1834	-0.0120	-0.0366	0.0544
2MetilButirrico	-0.2669	-0.2043	-1.0184	-0.0557	**0.6665**	**0.2222**
Valerico	**2.2041**	**-0.6103**	**-1.5254**	0.0963	-0.2619	**-0.2656**

Per il campione plasmatico l'acido Valerico è quello offre il maggiore contributo a tutte e tre le funzioni discriminanti, con un valore del coefficiente superiore in valore assoluto rispetto al coefficiente associato agli altri acidi. Per il campione fecale, LD1 è dominata principalmente dall'acido IsoButirrico, LD2 è invece maggiormente influenzata dall'acido 2MetilButirrico, mentre la funzione LD3 in modo più lieve da 2MetilButirrico e Valerico.

Le proporzioni di varianza spiegata, dicono quanto ogni funzione discriminante contribuisce alla separazione delle classi, maggiore è la proporzione di varianza spiegata per una funzione discriminate e maggiore è il suo potere discriminante. Nella Tabella 3.13 sono riportate tali proporzioni di varianza spiegata per il campione plasmatico e per quello fecale.

Tabella 3.13: Proporzioni di varianza spiegata per ogni funzione discriminante

	LD1	LD2	LD3
Plasma	0.897	0.097	0.006
Feci	0.538	0.320	0.142

La prima funzione discriminante nel campione plasmatico spiega circa il 90% della varianza: quasi tutta l'informazione discriminante nel plasma è concentrata in questa dimensione. Questo rende il contributo dell'acido Valerico, già evidenziato come predominante in termini di coefficiente, ancora più rilevante, confermandolo come il principale discriminante tra i gruppi a livello plasmatico. Per il campione fecale invece la separazione tra i gruppi è più distribuita tra le componenti ed è quindi necessario per interpretare i risultati considerare tutte e tre le funzioni discriminanti.

Nella Figura 3.1 e nella Figura 3.2 sono mostrate le rappresentazioni grafiche dell'analisi discriminante usando combinazioni a due a due delle funzioni discriminanti, rispettivamente per il campione plasmatico e per quello fecale. Si utilizza una coppia di funzioni discriminanti perché nel plasma, anche se LD1 spiega quasi il 90% della varianza, e quindi anche da sola sarebbe molto informativa, affiancarla a LD2 consente di migliorare la visualizzazione; nelle feci LD1 spiega circa il 54% e LD2 il 32%, quindi insieme racchiudono oltre l'85% dell'informazione discriminante, rendendo la proiezione su due assi una sintesi efficace e interpretabile del fenomeno.

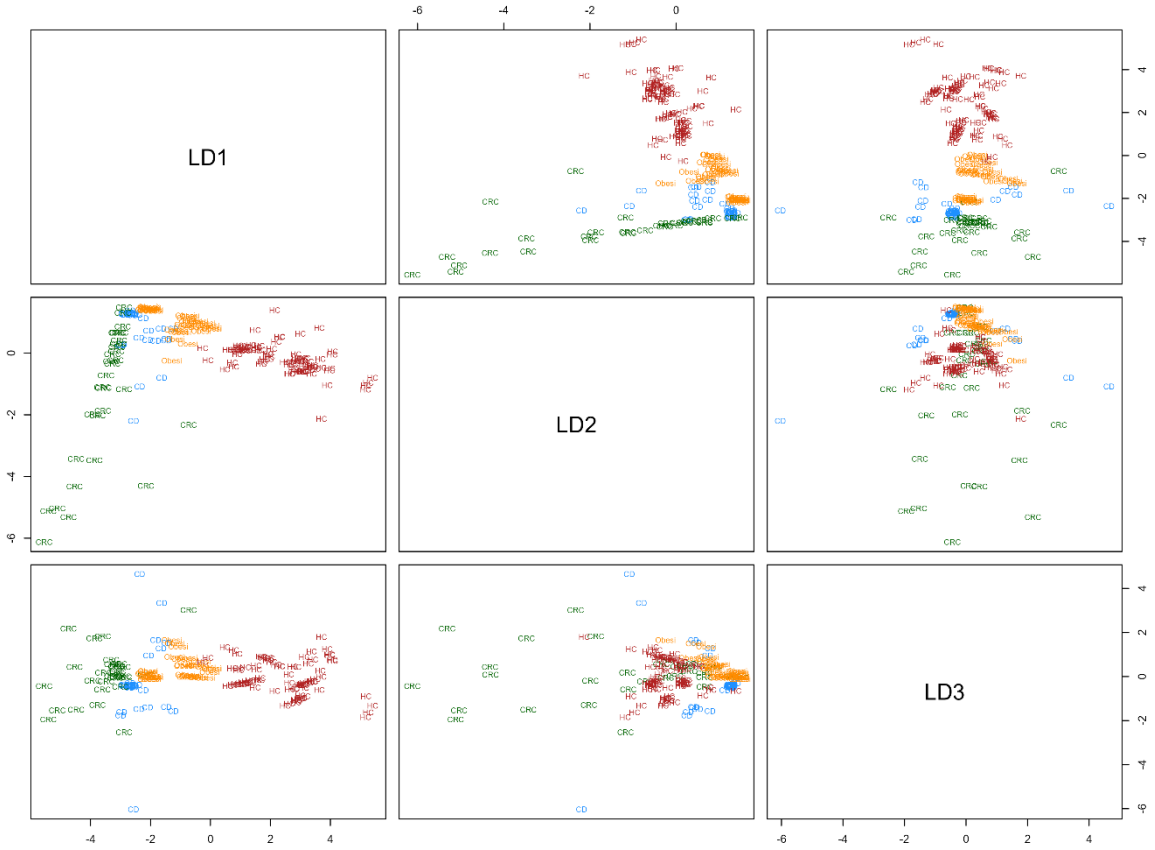

Figura 3.1: Scatterplot delle funzioni discriminanti ottenute tramite LDA per il campione plasmatico

Per quanto riguarda il campione plasmatico, lo *scatterplot* ottenuto dalla combinazione delle LD1 e LD2 distingue in modo efficace i gruppi HC, CRC e Obesi, mentre i soggetti appartenenti al gruppo CD risultano più dispersi e meno distinti. La combinazione delle LD1 e LD3 evidenzia una separazione ancora più netta tra CRC e HC, ma mostra maggiore sovrapposizione tra i gruppi Obesi e CD.

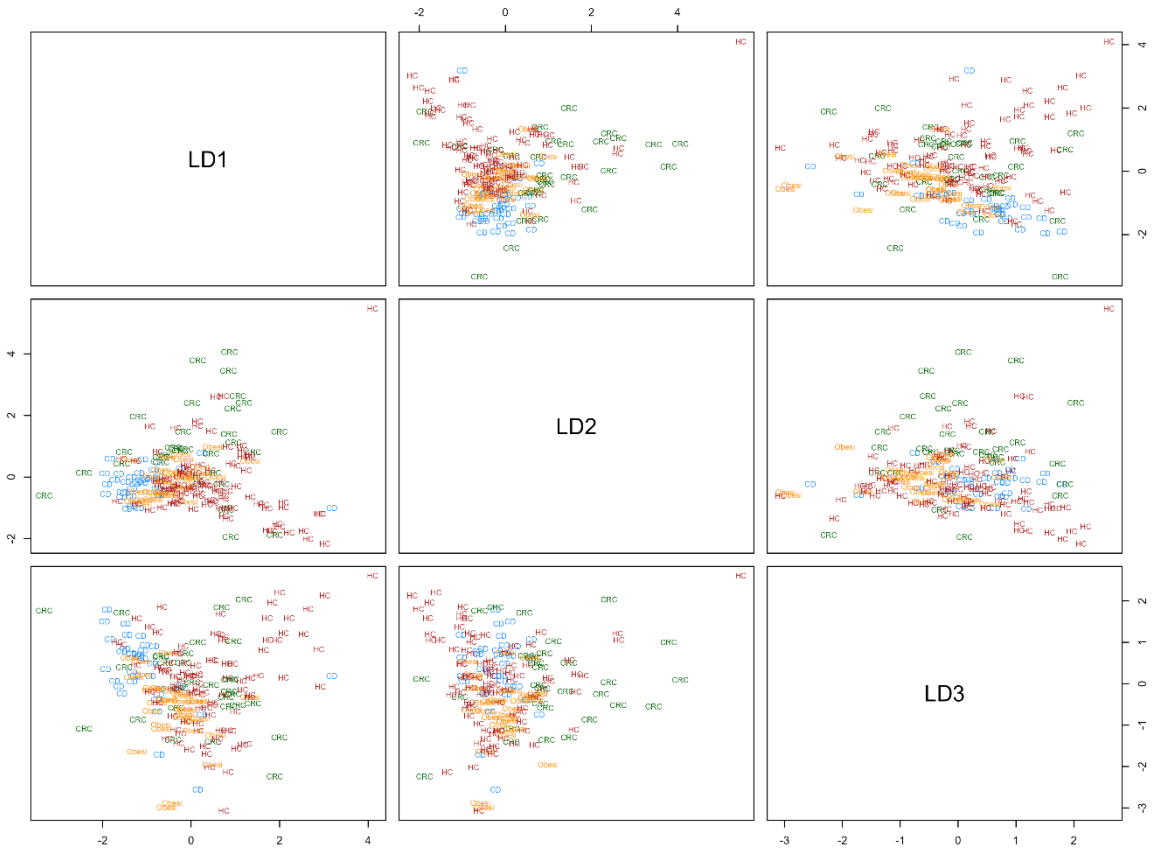

Figura 3.2: Scatterplot delle funzioni discriminanti ottenute tramite LDA per il campione fecale

Nel caso del campione fecale, si osserva che la funzione discriminante LD1 è quella che contribuisce maggiormente alla separazione tra i gruppi, ma non risulta sufficiente a garantire una discriminazione chiara. La combinazione migliore è quella delle prime due funzioni discriminanti, ma anche in questo caso il gruppo CD si distribuisce in modo eterogeneo, tendendo a sovrapporsi parzialmente agli altri gruppi clinici.

Per valutare l'efficacia dell'analisi discriminante lineare di Fisher nel classificare correttamente le unità nei quattro gruppi, si osservano nella Tabella 3.14, nella Tabella 3.15, nella Tabella 3.16 e nella Tabella 3.17 le matrici di confusione e i rispettivi criteri di performance per il campione plasmatico e fecale. Questa valutazione consente di confrontare la capacità di discriminare le osservazioni nei quattro gruppi dei due campioni e in generale l'efficacia di tale modello nel distinguere i soggetti.

Per il campione plasmatico i risultati ottenuti suggeriscono prestazioni complessivamente ottime, con il gruppo HC classificato quasi perfettamente; anche per il gruppo CD emergono buone performance. Per i gruppi CRC e Obesi la riclassificazione risulta più debole, ma comunque soddisfacente, concludendo che i valori degli acidi nel plasma hanno capacità discriminanti affidabili. Anche per il campione fecale, il gruppo HC rimane quello meglio

classificato, ma presenta una specificità molto bassa, suggerendo che il modello tende a inserire erroneamente soggetti di altri gruppi clinici tra i controlli sani. Per i gruppi CRC e Obesi le prestazioni del modello sono invece più critiche, suggerendo difficoltà per il modello nel riconoscere correttamente le unità appartenenti a tali gruppi clinici.

Tabella 3.14: Matrice di confusione ottenuta con LDA per il campione plasmatico

		Gruppo osservato			
		CD	CRC	HC	Obesi
Gruppo assegnato dall'LDA	CD	28	12	0	11
	CRC	2	18	0	0
	HC	0	0	93	0
	Obesi	5	0	2	22

Tabella 3.15: Criteri di performance dell'LDA per il campione plasmatico

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.8000	0.6000	0.9789	0.6667
SPECIFICITÀ	0.8544	0.9877	1.0000	0.9563
POS PRED VALUE	0.8490	0.9000	1.0000	0.7586
NEG PRED VALUE	0.9507	0.9306	0.9800	0.9329

Tabella 3.16: Matrice di confusione ottenuta con LDA per il campione fecale

		Malattia vera			
		CD	CRC	HC	Obesi
Gruppo assegnato dall'LDA	CD	25	4	5	4
	CRC	0	9	8	0
	HC	8	16	79	24
	Obesi	2	1	2	5

Tabella 3.17: Criteri di performance dell'LDA per il campione fecale

	CD	CRC	HC	Obesi
SENSIBILITÀ	0.7143	0.3000	0.8316	0.1515
SPECIFICITÀ	0.9114	0.9509	0.5102	0.9688
POS PRED VALUE	0.6410	0.5294	0.6220	0.5000
NEG PRED VALUE	0.9351	0.8807	0.7576	0.8470

3.5 Conclusioni tecniche statistiche

Il test di Kruskal-Wallis ha permesso di confrontare la capacità discriminatoria delle concentrazioni degli SCFA nei due campioni concludendo che quella del plasmatico è più forte, con variazioni significative degli SCFA tra gruppi, rispetto a quella del fecale, per il quale emergono differenze meno marcate tra i gruppi. Approfondendo quali siano effettivamente i gruppi significativamente diversi con il test di Dunn, è possibile affermare che il campione plasmatico è il più informativo per discriminare tra i diversi gruppi, soprattutto per riconoscere i controlli sani rispetto agli altri casi clinici, ma anche il campione fecale mostra differenze tra gruppi rilevanti per specifici acidi.

Con il modello di regressione logistica multinomiale è stato possibile notare che nel campione plasmatico, per ogni patologia, l'unico acido sempre significativamente associato alla probabilità di appartenenza a un gruppo clinico piuttosto che ai controlli sani è l'acido Valerico: i suoi coefficienti, fortemente negativi in modo costante, indicano che un aumento della sua concentrazione è associato a una riduzione della probabilità di appartenenza ai gruppi clinici. Nel campione fecale invece solo alcuni acidi hanno effetti significativi e solo su determinate patologie nel distinguersi dai controlli sani. Utilizzando il modello dei dati plasmatici per effettuare previsioni emerge che la riclassificazione è complessivamente accurata, segno che i livelli di concentrazione degli SCFA nel plasma sono informativi e ben discriminanti tra i gruppi. Invece il modello del campione fecale, utilizzato su tale campione, riesce a discriminare bene solo alcuni gruppi clinici, come HC, ma risulta meno efficace per altri, come Obesi. Utilizzando i dati combinando i due campioni, si ottiene un modello con il quale le classificazioni risultano ottime e precise, con solo lievi confusioni: dimostra che l'utilizzo dei dati combinati crea un modello con buona capacità discriminante, ma con nessun miglioramento importante rispetto all'utilizzo del campione plasmatico singolarmente.

Dall'utilizzo dell'LDA, si osserva che nel campione plasmatico l'acido che più influenza la prima funzione discriminante, la quale singolarmente spiega quasi il 90% della varianza è il Valerico, indicando nuovamente che è il principale SCFA a determinare la separazione tra i gruppi in questo campione; nel campione fecale invece i principali sono gli acidi IsoButirrico, 2MetilButirrico e Valerico, rispettivamente per LD1, LD2 e LD3, con percentuale di varianza spiegata più distribuita tra le componenti, quindi i tre acidi contribuiscono più equamente alla capacità discriminante di questo campione. Dalle matrici di confusione e dai criteri di performance si conclude che l'LDA ha prestazioni complessivamente ottime per il campione plasmatico, mentre in quello fecale il modello ha maggiori difficoltà nel classificare le unità.

Capitolo 4 - Conclusioni

Le analisi esplorative hanno evidenziato differenze rilevanti sia nella composizione che nella distribuzione degli acidi grassi a catena corta tra i gruppi clinici. I due campioni differiscono anche per le strutture di correlazione tra gli SCFA: nonostante siano presenti associazioni consistenti all'interno di ciascun campione, gli acidi maggiormente correlati variano tra plasma e feci. Nel gruppo dei soggetti sani le correlazioni risultano generalmente più forti, suggerendo una maggiore regolarità nelle concentrazioni degli SCFA, al contrario nei soggetti affetti da patologie le correlazioni tendono ad essere più deboli. L'analisi dei cluster condotta con il metodo di Ward ha già messo in luce che plasma e feci forniscono informazioni complementari nel distinguere tra i gruppi, suggerendo che il potere discriminante degli SCFA varia tra i due campioni. I test d'ipotesi mostrano che nel campione plasmatico le concentrazioni degli SCFA variano significativamente tra i gruppi clinici, indicando una forte capacità discriminante degli acidi grassi a catena corta nel plasma. In particolare l'acido 2MetilButirrico si distingue per la sua significatività in tutti i confronti a coppie tra i gruppi, mentre gli altri acidi permettono di distinguere bene tra i controlli sani e i gruppi patologici. Nel campione fecale il test evidenzia invece differenze meno marcate tra i gruppi, suggerendo quindi che gli SCFA mostrano una minore capacità discriminante in tale campione, sebbene, analizzando in modo più approfondito, alcuni acidi considerati singolarmente sono in grado di distinguere specifici gruppi clinici. Nel campione plasmatico, per tutte le patologie considerate, l'unico acido sempre significativamente associato alla probabilità di appartenenza a un gruppo clinico piuttosto che ai controlli sani è l'acido Valerico. Nel campione fecale invece solo alcuni SCFA mostrano effetti significativi e solo per determinati gruppi. Il modello di regressione logistica multinomiale costruito sui dati plasmatici restituisce una riclassificazione complessivamente accurata per tale campione, mentre il modello basato sui dati del campione fecale, utilizzato su tale campione, riesce a discriminare bene solo alcuni gruppi clinici. Quando i dati dei due campioni vengono combinati si ottiene un modello con prestazioni ottime e precise, con solo lievi confusioni: il modello di regressione logistica multinomiali ha buone capacità discriminanti se si utilizzano congiuntamente i dati dei due campioni, ma non superiori a quelle del modello costruito per il campione plasmatico.

Anche l'LDA conferma questi risultati: nel campione plasmatico l'acido che maggiormente contribuisce a spiegare quasi il 90% della varianza è il Valerico, mentre nel campione fecale gli acidi IsoButirrico, 2MetilButirrico e Valerico contribuiscono più equamente alla capacità discriminante. Le matrici di confusione e i criteri di performance mostrano che l'LDA ha

prestazioni complessivamente ottime per il campione plasmatico, mentre nel campione fecale il modello ha difficoltà a classificare correttamente le unità nei rispettivi gruppi clinici.

Si conclude che dall'analisi è emerso che gli SCFA differiscono significativamente tra i campioni plasmatici e fecali, sia per composizione che per potere discriminante. In particolare il campione plasmatico si è rivelato più variabile e informativo, con una maggiore capacità di distinguere tra i gruppi. Inoltre è emersa una mancanza di associazione chiara e univoca tra le concentrazioni dello stesso acido nei due campioni, indicando che la conoscenza delle concentrazioni degli SCFA in uno dei due non permette di prevedere quelle nell'altro campione.

È però importante esplorare e considerare anche alcune possibili limitazioni dello studio che possono aver influito sulla qualità dell'analisi e di conseguenza sui risultati ottenuti.

Una prima limitazione risiede nella numerosità campionaria non elevata: seppur sufficiente per un'analisi statistica, risulta comunque contenuta, non permettendo di generalizzare i risultati ad una popolazione più ampia e con la possibilità che abbia potuto far emergere associazioni deboli ma potenzialmente rilevanti. La diversità nella numerosità dei gruppi, potrebbe aver limitato la potenza statistica di alcune analisi, influenzando la stabilità delle stime nei modelli multivariati. Inoltre non sono stati considerati possibili fattori di confondimento come dieta, attività fisica o altri parametri metabolici, limitando l'esplorazione di legami causali e associazioni. Per migliorare lo studio sarebbe quindi utile poter creare gruppi più omogenei all'interno oppure stratificare i gruppi per caratteristiche rilevanti; un'altra possibilità sarebbe quella di inserire variabili aggiuntive che possano fungere da variabili di controllo e migliorare la qualità dei risultati.

Potrebbe risultare interessante per lo studio sapere se i soggetti in questione sono ben distinti nelle malattie o ci sono casi in cui un paziente soffre contemporaneamente di più patologie: in questo caso la loro presenza potrebbe causare sovrapposizioni tra gruppi, per cui potrebbe essere utile identificarli e isolarli dalle altre unità così da analizzarli separatamente e evitare che confondano i confronti tra gruppi.

Anche conoscere i criteri con cui è stato creato il gruppo dei controlli sani potrebbe migliorare le analisi, in quanto potrebbe essere definito sano un soggetto solo perché non presenta nessuna delle tre patologie in esame, ma in realtà potrebbero essere presenti altre condizioni cliniche che alterano l'equilibrio intestinale e influenzano quindi le concentrazioni degli SCFA, sollevando subbi su una possibile variabilità anche all'interno del gruppo di controllo.

Appendice teorica

A.1 Analisi in componenti principali (PCA)

L'analisi in componenti principali, o PCA, è una tecnica di analisi multivariata che mira alla riduzione delle dimensioni del dataset, rappresentando le osservazioni in uno spazio di dimensioni ridotte, ma mantenendo nel miglior modo possibile le informazioni originali. Ciò avviene trasformando le variabili originarie in un insieme più piccolo di variabili, chiamate componenti principali, generate come combinazioni lineari delle variabili originali (Karl Pearson 1901, Hotelling 1933). Questa tecnica è utilizzata in fase esplorativa dei dati poiché permette di esplorare il dataset estraendo le caratteristiche più informative e le strutture latenti. La PCA inoltre riduce problemi come la multicollinearità, dovuta a due o più variabili indipendenti altamente correlate, e l'*overfitting*. Per trovare le componenti principali si usa una media ponderata, detta combinazione lineare standardizzata, definita come segue:

$$\boldsymbol{W} = \delta^T \boldsymbol{X} = \sum_{j=1}^{p} \delta_j \boldsymbol{X}_j\,, \qquad \text{dove } \sum_{j=1}^{p} \delta_j^2 = 1.$$

Il vettore dei pesi $\delta = \left(\delta_1, \dots, \delta_p\right)^T$ è ottimizzato in modo da rilevare le caratteristiche dei dati originali. Si sceglie la combinazione lineare che massimizza la varianza di $\boldsymbol{W}$ ($Var[\boldsymbol{W}] = \delta^T \Sigma \delta$), dove Σ è la matrice di covarianza $Var[\boldsymbol{X}]$.

Per stabilire il numero di componenti principali da utilizzare, si considera la proporzione di varianza spiegata da ciascuna componente: si analizza la varianza cumulativa e si individua il numero di componenti che spiega cumulativamente una quota sufficiente di varianza totale, solitamente almeno pari all'80-90%. Come supporto a tale scelta viene utilizzato anche il grafico degli autovalori, chiamato *scree plot*, e si individua il punto in cui la curva mostra un piegamento "a gomito": tale punto corrisponde al numero ottimale di componenti principali.

La prima componente principale (PC1) è $y_1 = \gamma_1^T X$, dove il vettore γ_1 è l'autovettore normalizzato associato all'autovalore più grande, λ_1, della matrice Σ, mentre la seconda componente principale (PC2) è data da $\gamma_2^T X$, dove il vettore γ_2 è l'autovettore normalizzato associato al secondo più grande autovalore, λ_2, della matrice di covarianza Σ.

La prima componente principale è quella che riesce a spiegare la percentuale più alta di varianza, maggiore è tale percentuale e maggiore è l'informazione conservata dal dataset

originale; la seconda componente principale rappresenta la varianza più elevata successiva. Le due componenti principali sono per costruzione ortogonali tra loro, quindi non correlate. Per mostrare la relazione tra le due componenti di un set di dati viene utilizzato un grafico a dispersione, dove le PC sono gli assi perpendicolari.

A.2 Coefficienti di correlazione di Pearson e di Spearman

La correlazione fa riferimento alla relazione reciproca tra due variabili, senza quindi nessun riferimento a legami causali, ma solo ad un'associazione tra le due, alla tendenza a cambiare a seguito di un cambiamento dell'altra. La correlazione è lineare quando all'aumentare di una variabile anche l'altra aumenta o l'altra diminuisce.

Il *coefficiente di correlazione di Pearson* r, o coefficiente di correlazione lineare, è un indice che esprime un'eventuale relazione di linearità tra due variabili X e Y. Tale coefficiente è calcolato, partendo dalla serie doppia di osservazioni delle due variabili $(x_1, y_1), \dots, (x_N, y_N)$, come media aritmetica delle quantità $c_1, c_2, \dots, c_N$, dove $c_i = \frac{x_i - \mu_X}{\sigma_X} \cdot \frac{y_i - \mu_Y}{\sigma_Y}$.
Il coefficiente di correlazione di Pearson è quindi definito come:

$$r = \rho_{XY} = \frac{1}{N}\sum_{i=1}^{N} c_i = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i - \mu_X}{\sigma_X} \cdot \frac{y_i - \mu_Y}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

dove N è il numero di osservazioni, $\mu_X \; e \; \mu_Y$ le medie di X e Y, σ_X e σ_Y le deviazioni standard e σ_{XY} la covarianza tra X e Y. Tale coefficiente varia nell'intervallo $[-1,1]$, con -1 e 1 assunti quando è presente perfetta relazione lineare, rispettivamente negativa e positiva, tra le due variabili.

Il *coefficiente di correlazione di Spearman* r_s è una variante del coefficiente dei correlazione di Pearson: si calcola r non sulle coppie $(x_1, y_1), \dots, (x_n, y_n)$, ma sulla serie doppia dei ranghi che si ottiene sostituendo a x_i e y_i i rispettivi posti, p_{x_i} e p_{y_i}, nelle graduatorie non decrescenti dei valori di X e Y. Il coefficiente è poi calcolato come somma dei quadrati delle differenze tra ranghi corrispondenti, come segue:

$$r_s = 1 - \frac{6\sum_{i=1}^{N}\left(p_{x_i} - p_{y_i}\right)^2}{N(N^2 - 1)}$$

Anche in questo caso il coefficiente è compreso tra -1 e 1 che esprimono rispettivamente presenza di massima discordanza o massima concordanza.

A.3 Matrice di confusione e Criteri di performance

Prima di utilizzare un modello predittivo su nuovi soggetti, viene sottoposto ad una fase di validazione per valutarne la capacità classificatoria. A tal fine, si costruisce la matrice di confusione, da cui si ricavano alcune metriche che misurano la qualità della classificazione.

La matrice di confusione è una tabella di frequenza che incrocia la vera classificazione con quella effettuata dal modello predittivo:

		Realtà		
		G_0	G_1	
Previsione	G_0	n_{00}	n_{01}	$n_{0.}$
	G_1	n_{10}	n_{11}	$n_{1.}$
		$n_{.0}$	$n_{.1}$	n

dove:

- n_{11} sono i "veri positivi" (TP): soggetti appartenenti al gruppo G_1 e classificati correttamente al gruppo G_1;
- n_{00} sono i "veri negativi" (VN): soggetti non appartenenti al gruppo G_1 e non classificati nel gruppo G_1;
- n_{10} sono i "falsi positivi" (FP): soggetti non appartenenti al gruppo G_1, ma classificati in tale gruppo;
- n_{01} sono i "falsi negativi" (FN): soggetti appartenenti al gruppo G_1, ma non assegnati a tale gruppo.

Le metriche che permettono di valutare la capacità predittiva del modello sono dette *criteri di performance* e sono: Specificità, Sensibilità, Valore Predittivo Positivo (PPV) e Valore Predittivo Negativo (NPV).

Specificità e Sensibilità esprimono la proporzione di classificazioni corrette per quanto riguarda, rispettivamente, le unità negative (nel gruppo G_0) e le unità positive (nel gruppo G_1).

$$Sensibilità = \frac{TP}{TP + FN} \qquad Specificità = \frac{TN}{TN + FP}$$

Il PPV è la proporzione di unità correttamente assegnate al gruppo sul totale delle unità assegnate dal modello a tale gruppo, mentre il NPV è la proporzione delle unità correttamente escluse dal gruppo sul totale delle unità non assegnate al gruppo dal modello.

$$PPV = \frac{TP}{TP + FP} \qquad NPV = \frac{TN}{TN + FN}$$

A.4 Cluster Analysis e Metodo di Ward

L'obiettivo dell'analisi dei gruppi (*Cluster Analysis*) è quello di ottenere, a partire da una matrice di dati contenente misurazioni di p variabili su n unità, gruppi (cluster) di unità che siano il più possibile omogenei al loro interno e il più possibile separati tra loro.

I metodi di raggruppamento (clustering) si distinguono in due categorie principali: metodi gerarchici (agglomerativo/ divisivo) e metodi non gerarchici.

Nei metodi gerarchici agglomerativi, inizialmente ogni unità rappresenta un singolo gruppo, ad ogni passo si raggruppano le unità più simili fino ad avere un solo gruppo contenente tutte le unità; si dice gerarchico perché il processo costruisce una gerarchia di gruppi, rappresentata tramite un dendrogramma.

A partire dai dati si costruisce la matrice di distanze tra coppie di unità, si uniscono in un gruppo le due unità più vicine, si calcola la matrice di distanze tra gli $n-1$ gruppi, si uniscono i gruppi più vicini e si ripete fino a che tutte le unità sono contenute in un solo gruppo. La principale differenza tra i metodi gerarchici agglomerativi sta nel criterio con cui viene calcolata la distanza tra gruppi. I metodi più comuni sono: legame singolo, legame completo, legame medio, centroide e metodo di Ward. In particolare, il metodo di Ward si basa sull'idea di unire a due a due i gruppi in modo tale da ottenere il minore incremento di inerzia, misura dell'eterogeneità all'interno del gruppo (variabilità), utilizzando come misura di distanza la distanza euclidea.

A.5 Test di Kruskal−Wallis

Nelle situazioni in cui non è possibile assumere normalità è necessaria una procedura alternativa al test *F* nell'analisi della varianza, non dipendente da tale assunzione; la formalizzazione di una procedura di questo tipo è dovuta a Kruskal e Wallis (1952).

Il test di Kruskal-Wallis è usato per verificare l'ipotesi nulla che tutti i gruppi provengano dalla stessa distribuzione, contro l'ipotesi alternativa che almeno un gruppo differisce dagli altri. Poiché la procedura è stata costruita per verificare eventuali differenze nelle medie di popolazione, conviene considerare il test di Kruskal-Wallis come un test per l'eguaglianza delle medie di trattamento; si tratta di un test non parametrico alternativo all'abituale analisi della varianza. I test non parametrici hanno prestazioni migliori quando i dati non sono distribuiti normalmente e sono adatti soprattutto nei casi in cui la dimensione dei dati è piccola.

Si suppone di avere *k* gruppi, ognuno dei quali contiene un insieme di valori. Per eseguire il test di Kruskal-Wallis, si deve prima ordinare in ordine crescente le osservazioni y_{ij}, *j*-esima osservazione dell'*i*-esimo gruppo. Ogni osservazione viene sostituita con il suo rango R_{ij}, assegnando alla più piccola osservazione rango 1. Nel caso di ex aequo (osservazioni coincidenti) si assegna il rango medio ad ognuna delle osservazioni uguali.

Sia $R_{i.}$ la somma dei ranghi nell'*i*-esimo gruppo; la statistica test è:

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{k}\frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right]$$

dove n_i è il numero di osservazioni nell'*i*-esimo gruppo, *N* è il numero totale di osservazioni, e

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right]$$

Si noti che S^2 non è altro che la varianza dei ranghi; in assenza di ex aequo $S^2 = N(N+1)/12$ e quindi la statistica test si semplifica in:

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_{i.}^2}{n_i} - 3(N+1)$$

Quando il numero di ex aequo è limitato, la differenza tra le due equazioni per H è minima e quindi si può usare quest'ultima, essendo una formulazione più semplice. Se i valori numerici degli n_i sono abbastanza grandi ($n_i \geq 5$) la statistica H sotto l'ipotesi nulla è distribuita approssimativamente come χ^2_{a-1}. Pertanto se $H > \chi^2_{\alpha,a-1}$ l'ipotesi nulla è rifiutata. Si può usare anche l'approccio del p −value.

A.6 Test di Dunn

Il test di Dunn è un test non parametrico per confronti a seguito di significatività del test di Kruskal-Wallis, infatti il test di Dunn permette di effettuare confronti a coppie tra ciascun gruppo per capire quali sono quelli significativamente diversi.

Per la procedura sono utilizzati i ranghi comuni dell'intero dataset, ovvero non riclassificati a ogni confronto, infatti il test si basa sulle differenze tra le medie di tali ranghi. La statistica test per il confronto tra due generici gruppi A e B è:

$$z_i = \frac{y_i}{\sigma_i}$$

dove i, con $i = 1, \dots, m$, è uno degli m confronti, $y_i = \bar{W}_A - \bar{W}_B$ sono le differenze tra le medie dei ranghi dei due gruppi e σ_i è la deviazione standard di y_i, ottenuta come:

$$\sigma_i = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_A} + \frac{1}{n_B}\right) - C}$$

dove N è il numero totale di osservazioni, n_A e n_B sono le dimensioni campionarie dei gruppi e C è un termine correttivo per l'eventuale presenza di osservazioni con lo stesso valore.

Quando si effettuano numerosi test di ipotesi contemporaneamente è importante controllare il tasso di errore per famiglia aggiustando i valori dei *p-values*. Uno dei metodi per regolare i *p-values* è tramite l'aggiustamento di Bonferroni, che moltiplica il *p-value* grezzo per il numero di confronti effettuati.

A.7 Regressione logistica multinomiale (MLR)

Il modello di regressione logistica multinomiale (MLR) rappresenta un'estensione del modello logistico binario e si basa, come quest'ultimo, sull'analisi logit, ovvero sulla regressione logistica. Quando la variabile risposta è di tipo discreto, il modello di regressione lineare non risulta più applicabile, rendendo l'analisi logit un'alternativa efficace. Per molti aspetti, essa può essere considerata il naturale complemento della regressione lineare quando la variabile risposta non è quantitativa ma qualitativa.

Il modello logit binario è concepito per variabili dipendenti con due modalità, tuttavia può essere esteso a situazioni in cui la variabile risposta assume più di due categorie, prive di un ordine naturale. In questi casi si utilizza il modello di regressione logistica multinomiale, che applica versioni generalizzate dei metodi utilizzati nei modelli dicotomici.

Nel modello multinomiale si ha una variabile risposta Y, qualitativa con k categorie, e un insieme di p variabili esplicative $X_1, X_2, \dots, X_p$. La costruzione del modello richiede la scelta di una categoria di riferimento, chiamata base-line, rispetto alla quale vengono calcolati i log-odds delle restanti categorie. Poiché le categorie non hanno un ordine naturale, ognuna di esse può essere la base-line.

Il modello stima la variazione del logaritmo degli odds della probabilità di appartenenza a ciascuna categoria rispetto alla base-line, in funzione delle variabili esplicative. Le stime dei parametri si ottengono tramite il metodo della massima verosimiglianza, dopo aver trasformato la variabile dipendente in logit.

Sia $\pi_{ij} = P(Y_i = j | X_{i1} = x_{i1}, \dots, X_{ip} = x_{ip})$ la probabilità che l'osservazione i-esima venga classificata nel gruppo j-esimo, con $j \neq m$ e m la categoria base-line, si ha $\pi_{im} = 1 - \sum_{j \neq m} \pi_{ij}$.

Il modello di regressione logistica multinomiale è definito come:

$$\log\left(\frac{\pi_{ij}}{\pi_{im}}\right) = \log\left(\frac{P(Y_i = j | X_{i1} = x_{i1}, \dots, X_{ip} = x_{ip})}{P(Y_i = m | X_{i1} = x_{i1}, \dots, X_{ip} = x_{ip})}\right) = \beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \cdots + \beta_{pj}x_{ip}$$

dove $i = 1,2, \dots, n$ e $j \neq m$. Ogni equazione è un modello logit binario tra la categoria j e la categoria di riferimento m.

I coefficienti β_{kj} indicano l'effetto della variabile X_k, al netto delle altre, sui log-odds della categoria j rispetto alla categoria di riferimento m: un valore positivo indica che, all'aumentare di X_k, aumenta la probabilità relativa di appartenere alla categoria j rispetto alla base-line m, all'opposto per valori negativi. Esplicitando $OR_{kj} = \exp(\beta_{kj})$, questo rappresenta il fattore di variazione nelle odds di appartenere a j anziché a m per ogni incremento unitario di X_k.

La probabilità stimata dal modello di regressione logistica multinomiale per ciascuna categoria $j \neq m$ è data da:

$$\pi_{ij} = P(Y_i = j | X_{i1} = x_{i1}, \dots, X_{ik} = x_{ik}) = \frac{\exp(\beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \cdots + \beta_{pj}x_{ip})}{1 + \sum_{h \neq m} \exp(\beta_{0h} + \beta_{1h}x_{i1} + \beta_{2h}x_{i2} + \cdots + \beta_{ph}x_{ip})}$$

A.8 Analisi discriminante di Fisher (LDA)

L'analisi lineare discriminante, o LDA, è una tecnica statistica multivariata introdotta da R. A. Fisher (1936) usata nelle situazioni in cui i gruppi sono noti a priori. Lo scopo è classificare una o più osservazioni all'interno di questi gruppi noti, tramite una combinazione lineare di variabili quantitative. Il fine può essere sia descrittivo, per indagare la struttura dei gruppi, che classificatorio, in caso di nuove osservazioni da assegnare ai gruppi.

In caso di J popolazioni $\Pi_1, \Pi_2, \dots, \Pi_J$, descritte da funzioni di massa o di densità $f_1, f_2, \dots, f_J$ con supporto in R^p, una regola discriminante è una partizione dello spazio delle osservazioni in insiemi $\mathrm{R}_1, \mathrm{R}_2, \dots, \mathrm{R}_J$ tali che se un'osservazione nuova $x \in R_j$, essa viene assegnata alla popolazione Π_j.

L'dea alla base di tale procedimento è quella di proiettare i dati in uno spazio di dimensione ridotta, tramite combinazioni lineari delle variabili originali, massimizzando il rapporto tra la

varianza tra gruppi e la varianza entro gruppi, ovvero mantenendo il più possibile la separazione tra i gruppi.

Sia $\mathcal{X} = \begin{pmatrix} \mathcal{X}_1 \\ \vdots \\ \mathcal{X}_j \end{pmatrix}$ dove, per $j \in \{1, \dots, J\}$, $\mathcal{X}_j$ è una matrice di dati di dimensione $n_j \times p$ dalla popolazione Π_j. Si può definire la devianza complessiva dei dati come:

$$\mathcal{T} = \mathcal{X}'H\mathcal{X} = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$$

dove x_i è l'$i-$eima riga della matrice $\mathcal{X}$, ovvero l'$i-$esima osservazione e $\bar{x}$ è il vettore delle medie campionarie. Si può scrivere la devianza totale come $\mathcal{T} = \mathcal{W} + \mathcal{B}$, dove $\mathcal{W}$ e $\mathcal{B}$ sono la devianza $between$ (variabilità interna ai gruppi) e la devianza $within$ (variabilità tra gruppi).

Considerando la proiezione dei dati $\mathcal{Y} = \mathcal{X}\boldsymbol{a}$, allora la devianza dei dati in $\mathcal{Y}$ è

$$\mathcal{Y}'H\mathcal{Y} = \boldsymbol{a}'\mathcal{X}'H\mathcal{X}\boldsymbol{a} = \boldsymbol{a}'\mathcal{T}\boldsymbol{a} = \boldsymbol{a}'\mathcal{W}\boldsymbol{a} + \boldsymbol{a}'\mathcal{B}\boldsymbol{a}$$

L'obiettivo è selezionare un vettore $\boldsymbol{a} \in \mathrm{R}^p$ tale che la proiezione $\mathcal{Y}$ massimizzi il potere discriminante tra gruppi, ovvero che massimizzi la quantità di informazione relativa alla variabilità tra i gruppi rispetto a quella interna. Per fare ciò si utilizza il rapporto $\frac{\boldsymbol{a}'\mathcal{B}\boldsymbol{a}}{\boldsymbol{a}'\mathcal{W}\boldsymbol{a}}$, criterio di Fisher per la discriminazione, e si cerca il vettore $\boldsymbol{a}_*$ che massimizzi il rapporto sotto il vincolo $\boldsymbol{a}'\boldsymbol{a} = 1$. La soluzione è l'autovettore associato al più grande autovalore di $\mathcal{W}^{-1}\mathcal{B}$.
A questo punto la regola discriminate per classificare una nuova osservazione x consiste nell'assegnare x a Π_k per se cui $a_*(x - \bar{x}_j)$ è minima, ovvero se k è tale che

$$k = \underset{j \in \{1,\dots,J\}}{\operatorname{argmin}} \left| \boldsymbol{a}_*'(\boldsymbol{x} - \bar{\boldsymbol{x}}_j) \right|.$$

Con più di due gruppi ($J > 2$) il numero massimo di funzioni discriminanti è pari a $s = \min(p, J - 1)$, ciascuna ottenuta come autovettore associato a uno degli autovalori di $\mathcal{W}^{-1}\mathcal{B}$, ordinati in ordine decrescente.

Per utilizzare l'LDA è necessario poter assumere variabili quantitative distribuite normalmente all'interno dei gruppi, omoschedasticità e indipendenza tra le osservazioni.

Bibliografia

1. Baldi, S., Menicatti, M., Nannini, G., Niccolai, E., Russo, E., Ricci, F., Pallecchi, M., Romano, F., Pedone, M., Poli, G., et al. (2021). Free fatty acids signature in human intestinal disorders: Significant association between butyric acid and celiac disease. *Nutrients, 13*(3), 742. https://doi.org/10.3390/nu13030742
2. Solomons, T. W. G. (2001). *Chimica organica* (2ª ed.). Bologna: Zanichelli.
3. Baldi, S. (n.d.). *WP2 - Workshop: Functional metagenomics* [Presentazione PowerPoint].
4. Hou, H., Meng, Z., Zhao, W., & Wang, S. (2022). Gut microbiota-derived short-chain fatty acids and colorectal cancer: Ready for clinical translation? *Cancer Letters, 526*, 225-235. https://doi.org/10.1016/j.canlet.2022.02.026
5. Alvandi, E., Xu, Y., Kong, L., Liu, Y., Liu, L., Wang, H., ... & Zhang, L. (2022). Short-chain fatty acid concentrations in the incidence and risk-stratification of colorectal cancer: A systematic review and meta-analysis. *BMC Medicine, 20*(1), 323. https://doi.org/10.1186/s12916-022-02542-w
6. Geng, J., Song, Q., Tang, X., Liang, X., Fan, H., & Li, X. (2022). The links between gut microbiota and obesity and obesity-related diseases. *Biomedicine & Pharmacotherapy, 147*, 112678. https://doi.org/10.1016/j.biopha.2022.112678
7. Cicchitelli, G., D'Urso, P., & Minozzo, M. (2022). *Statistica: Principi e metodi* (4ª ed.). Milano: Pearson.
8. Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences, 19*(3), 303-342. https://doi.org/10.1016/0098-3004(93)90002-7
9. Alaimo, L. (2020). Analisi di correlazione. In *Ragionando di sviluppo locale: una lettura nuova di tematiche antiche* (pp. 387-395). Milano: Franco Angeli.
10. Montgomery, D. C. (2021). *Introduzione all'analisi statistica dei dati sperimentali* (pp. 132-133). Milano: Apogeo Education.
11. Marchetti, G. M. (2019). *Introduzione ai modelli statistici* (Rev. 1). Firenze: Dipartimento di Statistica, Informatica, Applicazioni.
12. El-Habil, A. M. (2012). An application on multinomial logistic regression model. *Pakistan Journal of Statistics and Operation Research, 8*(2), 271-291. https://doi.org/10.18187/pjsor.v8i2.299
13. Pang, Y. et al. (2014) Learning Regularized LDA by Clustering. *IEEE Transactions on Neural Networks and Learning Systems*. [Online] 25 (12),.
14. Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal, 15*(1), 292-300. https://doi.org/10.1177/1536867X1501500117