POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Matematica

Tesi di Laurea Magistrale

Metodi statistici per la ricerca di biomarcatori in metabolomica Il caso del tumore alla prostata

Relatore
prof. Mauro Gasparini

Laureando
Andrea Capra

Anno accademico 2017 – 2018

Ringraziamenti

Ringrazio la mia famiglia per avermi incoraggiato e aiutato nei momenti difficili.
Ringrazio gli amici con i quali in questi anni di universitá ho condiviso momenti bellissimi.
Desidero inoltre ringraziare il professor Mauro Gasparini e la Dottoressa Lidia Sacchetto, per la grande disponibilitá e cortesia dimostratemi e per tutto l'aiuto fornito durante il periodo di tesi.
Ringrazio il laboratorio di Farmacogenomica dei Tumori, Fondazione Edo e Elvo Tempia di Biella per avermi gentilmente concesso i dati utilizzati nelle analisi.
Infine ringrazio Marco e Federica per essermi stati vicino in questo periodo particolarmente impegnativo.

Sommario

L'obiettivo del seguente elaborato é l'esplorazione di nuovi possibili biomarcatori non invasivi che permettano di individuare la presenza del carcinoma alla prostata. Attualmente, per tale malattia, i metodi di screening per la diagnosi precoce prevedo l'utilizzo del PSA, un enzima prodotto dalla prostata, il quale peró porta a risultati insoddisfacenti con alte percentuali di falsi positivi e falsi negativi. La ricerca di nuovi possibili biomarcatori é stata effettuata nel campo della metabolomica, una scienza recente e ancora abbastanza inesplorata che studia i processi cellulari dell'organismo umano, misurando la concentrazione di molecole molto piccole (i metaboliti) per valutare le condizioni di salute di una persona. In particolare, sono state condotte indagini statistiche su un dataset reale gentilmente fornito dal laboratorio di Farmacogenomica dei Tumori, Fondazione Edo e Elvo Tempia di Biella.
L'elaborato é cosí strutturato:

- Nel capitolo 1 sono stati introdotti alcuni concetti di biologia, con particolare enfasi sullo strumento che ha portato allo sviluppo della metabolomica, permettendo l'analisi su larga scala dei derivati del metabolismo cellulare. É stata data una definizione di biomarcatore e sono state discusse le diverse tipologie e possibili utilizzi in biomedicina. Infine sono state presentate alcune nozioni generali sul tumore alla prostata.

- Nel capitolo 2 sono stati analizzati i principali metodi statistici utilizzati nella ricerca di biomarcatori. Sono stati proposti metodi di analisi univariata utili a individuare variabili che, prese singolarmente, sono associate con un certo stato fisiologico; metodi di analisi multivariata che permetto di selezionare piú variabili insieme. Si é discusso dei principali indici statistici utilizzati per valutare la qualitá di un biomarcatore, con particolare riferimento alla curva ROC. Infine sono state presentate le principali problematiche specifiche legate ai dataset di metabolomica: i valori mancati e le eterogeneitá presenti tra i diversi campioni dovute a fattori non biologici.

- Nel capitolo 3 si é analizzato un dataset reale; in particolare sono stati applicati i metodi statistici descritti nel capitolo 2, con lo scopo di discriminare due

gruppi di soggetti: quelli affetti da carcinoma alla prostata e quelli che soffrono di iperplasia prostatica benigna.

L'analisi ha consentito di individuare alcuni metaboliti che, in combinazione tra loro e con l'etá sembrano avere un potere discriminante superiore al PSA. Tuttavia le prestazioni dei modelli analizzati risultano ancora troppo limitate per poter includere questi metaboliti come nuovi biomarcatori diagnostici del tumore alla prostata.

Indice

Capitolo 1

Introduzione

1.1 La metabolomica e il metabolismo

Il metabolismo é l'insieme delle reazione chimiche che avvengono in un organismo vivente e che gli consentono di crescere e riprodursi. I composti chimici che intervengono nel metabolismo sono chiamati metaboliti. Il metaboloma umano é l'insieme completo di tutti i metaboliti presenti in un biofluido (plasma, siero, urina...) in un certo istante di tempo e ne costituisce, dunque, la componente chimica, escluse le macro-molecole (le proteine e gli acidi nucleici).
Il termine metaboloma é stato coniato in analogia con il termine genoma, ma a differenza di quest'ultimo, é un'entitá estremamente dinamica, in grado di cambiare da secondo a secondo e molto variabile da persona a persona; questo perché il metaboloma é il risultato dell'interazione dell'espressione genica con l'ambiente in cui viviamo.
Le "piccole" molecole che costituiscono il metaboloma includono: peptidi, lipidi, amminoacidi, carboidrati, acidi organici, vitamine, minerali, additivi dei cibi, medicine, droghe, tossine, inquinanti, e ogni altro composto chimico (con peso molecolare < 2000 Da) con cui l'essere umano entra in contatto.
Gli esperimenti di metabolomica si suddividono in due approcci complementari: l'approccio *targeted* in cui, essendo noti a priori i metaboliti da analizzare, i campioni vengono preparati in modo tale che l'esperimento metta in evidenza tali metaboliti; l'approccio *untargeted* il cui scopo é misurare, almeno idealmente, tutti i metaboliti presenti nel campione. Inoltre, la metabolomica ha un duplice utilizzo: puó essere utilizzata per la comprensione dei processi biologici e per identificare nuovi biomarcatori.

1.2 I biomarcatori

In generale, i biomarcatori sono un indicatore dello stato di salute di un organismo. A seconda del loro utilizzo distinguiamo tra biomarcatori prognostici e biomarcatori predittivi.
Un biomarcatore prognostico permette di discriminare i pazienti in base al grado di rischio di insorgenza della malattia e fornisce informazioni sul decorso naturale della patologia, in assenza di un intervento terapeutico. Fanno parte di questa categoria i biomarcatori diagnostici, che consentono di individuare in un soggetto un particolare stato fisiologico o una malattia.
Ad esempio, l'antigene carboidrato 19-9 (Ca19-9) é un marker tumorale che viene ricercato nel sangue ed é associato alla presenza del carcinoma al pancreas. Questo test non ha un alto valore diagnostico a causa della scarsa specificitá e della presenza di falsi negativi. Tuttavia, una volta diagnosticato il cancro, la concentrazione dei livelli di Ca 19-9 ha un elevato valore prognostico, in quanto una rapida riduzione dei livelli in seguito alla terapia chirurgica correla con un buon grado di resezione di neoplasia attiva. Analogamente, l'aumento del Ca 19-9 in corso di follow-up post operatorio correla con recidiva di neoplasia pancreatica o con la presenza di metastasi.
I marcatori predittivi permettono invece di predire se un paziente risponderá o meno ad una determinata terapia e consentono dunque di definire il miglior trattamento medico da somministrare. Permettono inoltre di stabilire il dosaggio di una terapia e di valutarne l'efficacia.
Un classico esempio di marcatore predittivo é il gene ERBB2 per il carcinoma della mammella: le pazienti che presentano un'amplificazione di tale gene beneficiano del trattamento con un determinato farmaco (il trastuzumab), mentre le pazienti in cui il recettore per gli estrogeni é espresso dal tumore rispondono al trattamento con un altro farmaco (il tamoxifen).
Chiaramente non é detto che un biomarcatore appartenga ad un unica classe; ne esistono alcuni che hanno un duplice utilizzo (prognostico e predittivo).
A seconda della loro natura, possiamo distinguere tra biomarcatori genetici, biochimici e biologici.
In generale un buon biomarcatore é caratterizzato dalla non-invasivitá, dalla semplicitá di analisi, da elevati valori di sensitivitá e specificitá (nel caso di marcatori prognostici la predizione del biomarcatore deve infatti rispecchiare il vero stato del paziente), da un alto grado di interazione con la terapia (nel caso di biomarcatori predittivi). Un altro requisito fondamentale é la riproducibilitá: la misurazione di un biomarcatore, se ripetuta piú volte, deve mantenere i propri valori inalterati, fornendo lo stesso risultato.

1.3 Lo strumento per le analisi di laboratorio

Il recente sviluppo della metabolomica é stato favorito dalla nascita di tecniche strumentali, quali la risonanza magnetica nucleare (NMR) e la spettrometria di massa (MS), in grado di misurare, attraverso un unico esperimento, migliaia di metaboliti contemporaneamente (approccio omico).
In particolare la MS, preferita nelle analisi *untargeted* per la sua maggiore sensibilitá, si basa sul principio che per una molecola ionizzata si puó facilmente misurare la sua massa: é sufficiente immergerla in un campo magnetico e osservare il suo moto; il processo di ionizzazione risulta dunque cruciale. Ionizzare molecole biologiche non é peró un processo semplice e viene prodotto molto rumore. L'identificazione della molecola, ovvero risalire al composto chimico nota la massa, é un procedimento complicato, poiché esistono composti diversi con stessa massa (gli isomeri).

1.3.1 Alcune definizioni utili

Nel contesto della spettrometria di massa bisogna fare attenzione al concetto di massa di un composto/metabolita. É necessario distinguere tra il numero di massa e la massa monoisotopica.
Il **numero di massa** é pari alla somma di protoni e neutroni presenti in un atomo. É un numero intero.
La **massa molecolare** é la massa di un certo composto chimico espressa in senso relativo, rispetto alla dodicesima parte della massa del piú importante e abbondante isotopo naturale del carbonio, il ^{12}C. L' unitá di misura della massa molecolare é il Dalton (Da), che si riferisce dunque alla dodicesima parte della massa del ^{12}C
La **massa monoisotopica** di una molecola é la somma delle masse degli atomi che la compongono, considerando, per ogni atomo, la massa non approssimata dell'isotopo piú presente in natura.
Nella MS si usa il termine massa riferendosi alla massa monoisotopica.
Infine, un altro concetto importante é il **rapporto massa-carica (mass-to-charge ratio)**. Si tratta di una quantitá fisica utilizzata in elettrodinamica, in quanto due molecole aventi lo stesso rapporto massa-carica seguono la stessa traiettoria nel vuoto, se immerse nelle stesso campo elettromagnetico.
In presenza di ioni con lo stesso stato di carica, si ha una corrispondenza uno-a-uno tra massa monoisotopica e rapporto massa-carica.

1.3.2 La spettrometria di massa

Esistono diverse tipologie di MS. Tutte sono, peró, costituite da tre parti: lo ionizzatore, l'analizzatore e il rilevatore.

Lo ionizzatore converte composti chimici elettricamente neutri in ioni carichi. Esistono diversi tipi di ionizzatori: il piú usato in metabolomica é l'elettrospray

(ESI), poiché é in grado di ionizzare un'ampia gamma di molecole con diversa massa molecolare e polaritá, lasciando le molecole intatte quindi piú facilmente identificabili.

L'analizzatore separa gli ioni, elettricamente carichi, secondo il loro rapporto massa-carica utilizzando un campo elettromagnetico.
Il principio di funzionamento delle varie tipologie di analizzatore é lo stesso, nonostante la loro risoluzione (ovvero la precisione della misura) possa variare significativamente. In particolare l'accelerazione di uno ione risulta determinata proprio dal suo rapporto massa-carica (m/z).

Il rilevatore ha lo scopo di quantificare una determinata specie ionica. Esso converte l'abbondanza di uno ione in un segnale elettrico, registrandone la corrente prodotta quando questo passa attraverso il rilevatore.

Sebbene le moderne spettrometrie di massa abbiano una risoluzione elevatissima, non é semplice ricondurre un certo m/z al metabolita corrispondente. Ció é dovuto al fatto che metaboliti con formula chimica diversa possono presentare masse simili; inoltre esistono metaboliti aventi uguale formula chimica, ma diversa struttura. Questi metaboliti vengono definiti isomeri e possono essere di due tipi:

Gli isomeri strutturali: in cui gli atomi e i gruppi funzionali sono gli stessi, ma legati in maniera differente (differenti legami chimici).

Gli stereoisomeri: in cui gli atomi e i legami chimici coincidono, ma cambia la posizione geometrica.

Gli isomeri strutturali hanno masse esatte differenti, ma osservabile solo con strumenti ad altissima risoluzione; gli stereoisomeri hanno masse perfettamente coincidenti. Un esempio di isomeri strutturali sono la cotinina e la serotonina, aventi massa coincidente arrotondata alla quinta cifra decimale, ma differente formula di struttura e ruolo biologico.

1.3.3 La spettrometria di massa accoppiata con la cromatografia liquida

Quando si analizza un miscuglio complesso come il sangue, la MS é spesso preceduta dalla cromatografia liquida (LC).
Il campione da analizzare é sciolto in un liquido/solvente chiamato "fase mobile". Tale liquido viene spinto lungo una colonna di vetro da forti pressioni. All'interno della colonna é presente un materiale poroso ("fase stazionaria"), che "ostacola" il passaggio dei vari composti chimici presenti nel solvente. Ogni composto, dunque, eluisce (ovvero viene espulso dalla colonna) in un tempo diverso, indicato con *retention time* (RT). Il RT é soggetto a variazioni significative dipendenti dalle condizioni dell'esperimento; solo pochi metaboliti hanno un RT uguale o simile in un

esperimento (fenomeno chiamato coeluzione). Dunque la LC riduce la complessitá del campione e diminuisce il rumore di sottofondo nel processo di rilevamento dell'MS.
Poiché i composti presenti nel campione sono separati sia attraverso la LC sia attraverso la MS, i dati ottenuti attraverso la tecnica LC-MS generano un segnale tridimensionale (come mostrato in Figura 1.1). Una dimensione del segnale é il RT, la seconda é l'm/z e la terza fornisce un quantitativo del metabolita, la concentrazione o intensitá (a seconda se si pone maggiore attenzione al significato biologico o al significato fisico, intensitá di corrente).

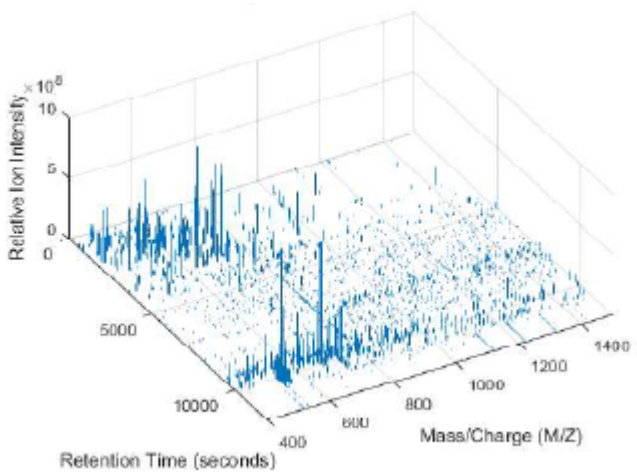


Figura 1.1: Dati grezzi generati dalla LC-MS.

Per predisporre i dati a successive analisi statistiche, i dati grezzi devono essere processati. Il pre-processamento include varie fasi. Nella *peak detection* i dati grezzi continui sono convertiti in dati discreti, in modo tale che ogni ione sia rappresentato da un picco. Tale trasformazione offre due vantaggi:

- viene rimosso parte del rumore presente nei dati grezzi;
- la dimensione dei dati viene ridotta senza perdita di informazioni.

Ogni picco rappresenta quindi un metabolita identificato da una terna di valori: m/z, RT, concentrazione.
Successivamente, il *peak aligment* rende possibile confrontare i dati tra i vari campioni, assegnando lo stesso RT ai medesimi metaboliti nei diversi campioni (il RT dei vari metaboliti puó infatti variare da campione a campione). La Tabella 1.1 é un esempio di sintesi delle fasi di *peak detection* e *peak aligment* per un set di campioni: ogni riga rappresenta un metabolita (ma lo stesso metabolita puó essere

Tabella 1.1: Dati in forma tabulare, pronti per le analisi statistiche.

m/z	RT	Campione 1	Campione 1	...	Campione N
167	1870	2997876	4066690	...	2336552
170	1439	10299087	2950550	...	4275303
184	1850	3098962	2145263	...	1173226
186	1842	3638807	1482374	...	1614796
⋮	⋮	⋮	⋮	⋮	⋮
202	1849	2392494	1346477	...	1220198

rappresentato in piú righe), identificato tramite le prime due colonne (m/z, RT), mentre le colonne successive rappresentano il valore di concentrazione per i vari campioni.

1.4 Il carcinoma della prostata

Il carcinoma prostatico (PCa) é un tumore maligno che colpisce le cellule epiteliali della prostata, una ghiandola dell'apparato genitale maschile. In Italia, nel 2012, il PCa é stato il tumore maligno piú frequentemente diagnosticato, seguito dal tumore ai polmoni e al colon retto. Inoltre costituisce la terza causa di morte tra i morti di tumore (Figura 1.2).

I sintomi dell'insorgenza del PCa sono difficoltá a iniziare la minzione e a mantene-

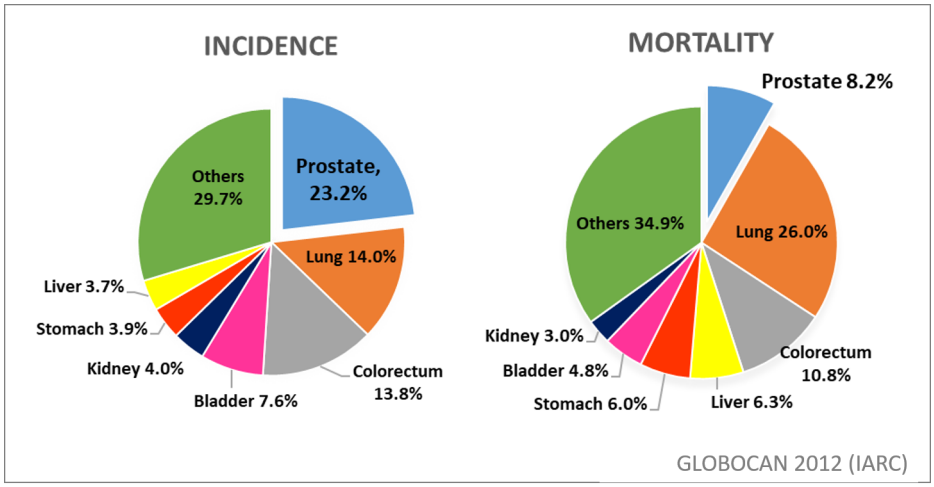


Figura 1.2: Mortalitá e incidenza dei principali tumori maligni (2012).

re un getto costante, presenza di sangue nelle urine, minzioni frequenti in cui viene espulsa piccola quantitá di urina. Tali sintomi sono simili a quelli che insorgono in

caso di iperplasia prostatica benigna (BPH), che é una condizione caratterizzata dall'aumento di volume della ghiandola prostatica, ma non é una neoplasia maligna.
Attualmente, solo la biopsia, ovvero l'asportazione di un frammento di tessuto, puó confermare pienamente la presenza di un PCa. Le scelte diagnostiche di screening, utili per individuare tumori in fase precoce, nel caso del tumore della prostata comprendono l'esame rettale (*digital rectal examination*, DRE) e il dosaggio dell'antigene prostatico specifico (PSA).
Il DRE permette di valutare le dimensioni, la forma e la consistenza della prostata: zone irregolari, dure o bozzolute devono essere sottoposte a ulteriori valutazioni, perché potrebbero indicare la presenza di tumore. Tale esame risulta efficace poiché in genere le irregolaritá nella prostata dovute al PCa si differenziano da quelle generate da BPH.
L'alterazione del dosaggio ematico del PSA é un altro segnale della possibile insorgenza del tumore prostatico. Il PSA é infatti un enzima prodotto dalla prostata; livelli di PSA sotto 4 ng/mL sono generalmente considerati normali, mentre livelli sopra i 4 ng/mL indicano un maggiore rischio di tumore. In ogni caso circa 1/3 dei pazienti affetti da PCa non presenta valori alterati del dosaggio di PSA e, d'altra parte, ci sono soggetti con valori di PSA molto elevati che non presentano PCa. Il PSA non risulta quindi essere un buon indicatore diagnostico per il carcinoma alla prostata, poiché produce un elevato numero di falsi positivi (soggetti sani indicati come malati dal test del PSA e quindi costretti a biopsie invasive) e falsi negativi (soggetti malati indicati come sani dal test).

1.5 Propositi della tesi

Vista la scarsa utilitá del PSA, si é alla ricerca di biomarcatori alternativi che abbiano prestazioni migliori a livello prognostico e diagnostico per il tumore alla prostata.
Il recente sviluppo di strumenti di analisi che consentono di misurare contemporaneamente migliaia di variabili biologiche (*high-throughput experiments*) ha ampliato e velocizzato le possibilitá di individuare nuovi biomarcatori.
Lo scopo di questa tesi é l'esplorazione di possibili nuovi biomarcatori nel campo della metabolomica. L'analisi dei metaboliti é infatti promettente in tal senso: numerosi studi scientifici hanno individuato alcune di queste variabili biologiche come biomarcatori per il tumore al pancreas e al colon.
La nostra indagine é stata condotta su un dataset reale con soggetti affetti da tumore prostatico e iperplasia benigna.

Capitolo 2

Tecniche di analisi statistica

Le analisi statistiche che faremo hanno lo scopo di individuare variabili che, prese singolarmente o in combinazione, permettano di discriminare tra due gruppi, il gruppo dei "sani" e il gruppo dei "malati". Tratteremo, cioé, metodi per la classificazione binaria.
L'estensione di tali metodi a casistiche in cui sono presenti piú di due gruppi puó non essere immediata.
Presenteremo le tecniche piú utilizzate per valutare la qualitá di un classificatore.
Infine tratteremo le principali tematiche legate al pre-processamento di dati acquisiti tramite LC-MS.

2.1 Analisi univariata

L'analisi univariata consiste nell'analizzare le singole variabili separatamente, al fine di individuare quelle differenzialmente espresse tra i vari gruppi.
Quando la distribuzione di una certa variabile per i diversi gruppi ha un andamento normale, si utilizza un test parametrico (*t-test*).
In caso di non normalitá dei dati, se tale condizione continua a valere anche dopo aver applicato la trasformazione di *Box-Cox*, si utilizza il test dei ranghi di *Mann-Whitney*, basato su un approccio non parametrico. Per valutare la normalitá dei dati si puó utilizzare, ad esempio, il test di *Shapiro-Wilk*.

2.1.1 Il *t-test* per dati normali

Verificare se le variabili sono differenzialmente espresse nei due gruppi, in un' ottica parametrica, dove le distribuzioni delle variabili nei due gruppi si assumo normali, vuol dire andare a valutare se le medie di tali distribuzioni posso essere assunte identiche. A seconda che la varianza sia nota, incognita ma si possa assumere identica nei due gruppi, oppure incognita e diversa nei gruppi, esistono diversi test.

Siano dunque $X = (x_1, \ldots, x_n) \sim \mathcal{N}(\mu_x, \sigma_x^2)$ e $Y = (y_1, \ldots, y_m) \sim \mathcal{N}(\mu_y, \sigma_y^2)$, realizzazioni indipendenti e identicamente distribuite (iid) provenienti dalle due variabili aleatorie; si vuole verificare:

$$H_0 : \mu_x = \mu_y \qquad \text{contro} \qquad H_1 : \mu_x \neq \mu_y \,. \tag{2.1}$$

Nel caso di varianze incognite e identiche, si definisce la statistica T:

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{1/n + 1/m}} \tag{2.2}$$

$$\text{con} \quad S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n + m - 2} \tag{2.3}$$

dove $\bar{X}$, S_x e n sono la media campionaria, la deviazione standard campionaria e la dimensione del campione del gruppo 1, analogamente per il gruppo 2.
Sotto l'ipotesi di normalitá T segue una distribuzione t di Student con $\nu = m+n-2$ gradi di libertá:

$$T \sim t_{n+m-2} \,.$$

Si fissa, quindi, un livello di significativitá α (ovvero la probabilitá di rifiutare l'ipotesi nulla (H_0) quando questa invece é vera) e si costruisce una regione di rifiuto del test. Valori tipici per α sono 0.05, 0.01.
In particolare, possiamo verificare l'ipotesi $\mu_x = \mu_y$ come segue:

$$\text{si rifiuta } H_0 \text{ se} \quad |T| > t_{\frac{\alpha}{2}, n+m-2}$$

$$\text{si accetta } H_0 \text{ se} \quad |T| \leq t_{\frac{\alpha}{2}, n+m-2}$$

Per decidere se accettare o meno un test si utilizza il *p-value*. Il *p-value* é cosí definito:

$$p\text{-}value = \mathbb{P}(|T| > t_{oss} | H_0 \quad \text{é vera}) \tag{2.4}$$

dove t_{oss} é il valore osservato per la statistica T ed é una misura di quanto i dati supportano l'ipotesi nulla H_0. *p-value* maggiori del livello di significativitá α suggeriscono di accettare l'ipotesi nulla, diversamente si rifiuta l'ipotesi nulla e il test viene detto significativo.
In biologia spesso si utilizza una variante del *t-test* chiamata *moderate t-test*, la cui differenza principale sta nel calcolo della varianza campionaria. Avendo a disposizione piú metaboliti, nel *moderate t-test*, la varianza di ogni metabolita viene stimata con una sorta di media ponderata tra la varianza del metabolita in questione e la varianza complessiva dei vari metaboliti. Si assume che la varianza per

il metabolita g, S_g^2 segua una distribuzione distribuzione chi-quadro con d_g gradi di libertá:

$$S_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2\,, \tag{2.5}$$

dove σ_g^2 é la varianza del metabolita. Si assume che σ_g^2 abbia come distribuzione a priori una chi-quadro inversa con d_0 gradi di libertá:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0\, s_0^2}\chi_{d_0}^2\,.$$

Sotto tali ipotesi, la media a posteriori di σ_g^2, dato S_g^2 é cosí determinata:

$$\tilde{S}_g^2 = \frac{d_o s_0^2 + d_g S_g^2}{d_0 + d_g}.$$

I termini d_0 e s_0^2 vengono stimate dai dati.
$\tilde{S}_g$ é sostituita a S_p, definita in (2.3), per il calcolo della statistica T:

$$\tilde{T} = \frac{(\bar{X} - \bar{Y})}{\tilde{S}_g\sqrt{2/n}} \tag{2.6}$$

dove n é il numero di soggetti appartenenti a ciascun gruppo e $d_g = 2n - 2$ (stiamo considerando il caso di esperimenti bilanciati).
Sotto l'ipotesi nulla ($H_0 : \mu_x = \mu_y$) la statistica $\tilde{T}$ segue una distribuzione t con $d_g + d_0$ gradi di libertá. Utilizzare la varianza "moderata", ha un duplice effetto sul *t-test*: viene modificato il termine S^2 e cambia il numero dei gradi di libertá della distribuzione t associata alla statistica $\tilde{T}$. Quindi, per metaboliti con varianza campionaria elevata la potenza statistica risulta incrementata (la varianza campionaria "moderata" é minore rispetto alla varianza campionaria propria di quel metabolita e il numero di gradi di libertá é maggiore), per i metaboliti con piccola varianza campionaria, la potenza statistica é ridotta dall'utilizzo della varianza moderata e incrementata dall'utilizzo di una statistica T a maggior numero di gradi di libertá.

2.1.2 Il test dei ranghi di *Mann-Whitney*

Il test dei ranghi é un test non parametrico che verifica se due gruppi di campioni statistici provengono dalla stessa popolazione. Viene utilizzato per dati quantitativi quando non sussiste l'ipotesi di normalitá. Siano $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_m$ i campioni provenienti dalle due popolazioni e F_X, F_Y le rispettive funzioni di ripartizione. Si vuole verificare l'ipotesi nulla $H_0 : F_X = F_Y$.
Il test dei ranghi é cosí costruito: si ordinano dal minore al maggiore le $n + m$ osservazioni, si assegna a ciascuna osservazione la propria posizione nell'ordinamento

appena ottenuto, si denota con R_i la posizione (il rango) dell'osservazione x_i. La statistica utilizzata nel test é la somma dei ranghi delle osservazioni di F_X:

$$T = \sum_{i=1}^{n} R_i \,.$$

Se la statistica T assume valori troppo estremali, tali da escludere una deviazione casuale, si rifiuta l'ipotesi nulla. In particolare, sia α livello di significativitá stabilito e t il valore di T; si rifiuta l'ipotesi nulla se:

$$\mathbb{P}(T \leq t) < \frac{\alpha}{2} \quad \text{o} \quad \mathbb{P}(T \geq t) < \frac{\alpha}{2} \,.$$

Siccome T assume valori interi,

$$\mathbb{P}(T \geq t) = 1 - \mathbb{P}(T \leq t-1) \,;$$

ció ci consente di affermare che H_0 va rifiutata se:

$$\mathbb{P}(T \leq t) < \frac{\alpha}{2} \quad \text{o} \quad \mathbb{P}(T \leq t-1) > 1 - \frac{\alpha}{2} \,.$$

É necessario dunque definire la funzione di ripartizione di T sotto l'ipotesi che H_0 sia vera.
Sia $\mathbb{P}(n, m, K)$ la probabilitá, condizionata ad H_0, dell'evento $\{T \leq K\}$, quando i campioni hanno numerositá n ed m. $\mathbb{P}(n, m, K)$ puó essere definita ricorsivamente dalla seguente formula:

$$\mathbb{P}(n, m, K) = \frac{n}{m+m}\mathbb{P}(n-1, m, K-n-m) + \frac{m}{n+m}\mathbb{P}(n, m-1, K)$$

con condizioni al contorno:

$$\mathbb{P}(1{,}0, K) = \begin{cases} 0 & K \leq 0 \\ 1 & K > 0 \end{cases} \quad \mathbb{P}(0{,}1, K) = \begin{cases} 0 & K < 0 \\ 1 & K \geq 0 \end{cases}$$

Tale formula permette di ricavare le probabilitá necessarie al test:

$$\mathbb{P}(T \leq t) = \mathbb{P}(n, m, t) \quad \text{e} \quad \mathbb{P}(T \leq t-1) = \mathbb{P}(n, m, t-1)$$

2.1.3 Il problema dei test multipli

Quando un test di ipotesi viene effettuato simultaneamente per piú variabili i livelli di significativitá normalmente usati non sono piú adatti e sono necessarie ulteriori considerazioni.
Ricordiamo che, se la statistica T, per una singola variabile, é maggiore di un certo valore soglia t_α (che dipende dal livello di significativitá α scelto), la variabile in

questione é differenzialmente espressa. Tale conclusione potrebbe peró derivare solo da effetti casuali (e questo capita con una probabilitá α). Se la variabile é invece effettivamente differenzialmente espressa, si giunge alla conclusione corretta. Ció avviene con una probabilitá:

$$\mathbb{P}(\text{"conclusione corretta"}) = 1 - \alpha.$$

Ora, nel caso in cui k variabili debbano essere testate, la probabilitá di arrivare alla conclusione corretta per tutte le variabili é data da:

$$\mathbb{P}(\text{"tutte le conclusioni sono corrette"}) = (1-\alpha)\dots(1-\alpha) = (1-\alpha)^k,$$

mentre la probabilitá di trarre almeno una conclusione errata é:

$$\mathbb{P}(\text{"almeno una conclusione é errata"}) = 1 - (1-\alpha)^k.$$

Tale valore puó essere visto come il livello di significativitá dell'intero esperimento (per l'intera famiglia di test). Ad esempio, un esprimento in cui vengono testate 20 variabili, ha una probabilitá di avere almeno un falso positivo (rifiuto di almeno un'ipotesi H_0 corretta) dell'87%, se il livello di significativitá per il singolo test é fissato a 0.1. In presenza di test multipli, il classico approccio ai test di ipotesi risulta quindi inadeguato.
Esistono dei metodi per controllare il livello di significativitá globale dell'esperimento. Il metodo *Bonferroni* consiste nel ridurre il livello di significativitá di ogni singolo test, dividendolo per il numero di test eseguiti.
Siano $H_1, \dots, H_k$ una famiglia di test di ipotesi e $p_1 \dots p_k$ i corrispettivi p-value. Supponiamo che per k_0 di questi k test l'ipotesi nulla sia corretta. Il *familywise error rate* (FWER) é la probabilitá di rifiutare, errando, almeno un'ipotesi nulla H_i, cioé la probabilitá di commettere almeno un errore di 1^o tipo. Rifiutare ogni singola ipotesi nulla H_i avente $p_i \leq \frac{\alpha}{m}$ permette di controllare il $FWER$ a un livello α.

$$FWER = \mathbb{P}\{\cup_{i=1}^{k_0}(p_i \leq \frac{\alpha}{k})\} \leq \sum_{i=1}^{k_0}\mathbb{P}\{(p_i \leq \frac{\alpha}{k})\} = k_0\frac{\alpha}{k} \leq \frac{\alpha}{k}k = \alpha\,. \tag{2.7}$$

La correzione di Bonferroni é molto restrittiva e comporta un aumento della probabilitá di commettere errore di 2^o tipo riducendo la potenza del test.
In alternativa si puó adottare la procedura di *Benjamini-Hochberg*. Mentre il metodo Bonferroni permette di controllare il $FWER$, l'approccio di Benjamini-Hochberg si basa sul concetto di *false discovery rate* (FDR). Il FDR é cosí definito:

$$FDR = \mathbb{E}\left[\frac{\text{numero di errori di } 1^o \text{ tipo}}{\text{numero di rifiuti dell'ipotesi nulla}}\right]$$

L'idea é quella di mantenere il FDR sotto una certa soglia α. A tal fine, si ordinano i p-value in ordine crescente$(p_{(1)}, \dots, p_{(m)})$. Si trova il piú grande k t.c $p_{(k)} \leq \frac{k}{m}\alpha$. Si rifiutano tutte le ipotesi nulle H_i con $i = 1 : k$. Tale approccio risulta meno stringente del metodo Bonferroni sul controllo dell'errore di 1^o tipo.

2.1.4 Strumenti per l'analisi grafica

Un primo strumento per visualizzare l'espressione di una certa variabile nei diversi gruppi é il *box-plot*. Il *box-plot* permette di visualizzare la mediana, il 3^o e il 1^o quartile per una certa distribuzione. L'analisi visiva deve poi essere avvalorata da test di ipotesi.
Per visualizzare invece, il risultato di molteplici test si puó usare il *volcano plot*, un tipo di *scatter-plot*, in cui ogni punto rappresenta una variabile. Questo grafico permette di valutare, per ogni variabile, il livello di significativitá e il grado di differenziazione nei due gruppi. Nello specifico, il *volcano plot* si costruisce plottando il logaritmo (di solito in base 10) del *p-value* cambiato di segno sull'asse y. Piú una variabile si allontana dall'origine lungo l'asse y, maggiore é il suo livello di significativitá. Sull'asse x si rappresenta invece il logaritmo (in base 2) del *fold change* per una determinata variabile (il *fold change* é il rapporto tra le medie campionarie dei due gruppi).

2.2 Analisi multivariata

2.2.1 Il modello logistico

Il modello logistico é un caso particolare di modello lineare generalizzato (Generalized Linear Model GLM).
Prima di descrivere il modello logistico, presentiamo brevemente i modelli lineari. In un modello lineare, si vuole stabilire se una variabile $\boldsymbol{y}$, detta variabile risposta, é esprimibile come combinazione lineare delle variabili $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1}$, dette predittori, piú un residuo. In termini di ciascuna unitá sperimentale N, ció vuol dire:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots \beta_{p-1} x_{i,p-1} + \epsilon_i \quad \text{per } i = 1, \ldots, N$$

esprimibile anche in forma matriciale come:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.8}$$

con $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ avente come colonne i vettori $\mathbf{1} = diag(\boldsymbol{I}_N), \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1}$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^T$ e $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma \boldsymbol{I}_N)$.
L'equazione (2.8) mette in evidenza come i valori attesi $\boldsymbol{\mu}$ della variabile risposta siano espressi come combinazione lineare dei predittori:

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}.$$

Si vuole trovare quel valore $\hat{\boldsymbol{\beta}}$ t.c $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2$ assuma il valore minimo.
In un modello lineare generalizzato, la relazione tra $\boldsymbol{\mu}$ e $\boldsymbol{X}\boldsymbol{\beta}$ non é lineare, ma é espressa da una funzione g invertibile, detta *link function*:

$$g(\mathbb{E}(\boldsymbol{y})) = \boldsymbol{X}\boldsymbol{\beta}.$$

In particolare, nel modello logistico le variabili risposta $y_1, \ldots, y_N$, assunte indipendenti, sono originate da una distribuzione Bernoulli appartenente alla famiglia esponenziale

$$y_i \sim Bernoulli(\pi_i)$$

in cui π_i é il valore atteso delle variabili risposta

$$\pi_i = \mathbb{E}(y_i) = \mathbb{P}(y_i = 1)$$

e dipende da un set di predittori secondo la seguente relazione:

$$g(\mathbb{E}(y_i)) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i\,\boldsymbol{\beta} \tag{2.9}$$

dove $\mathbf{x}_i \in \mathbb{R}^p$ é una generica riga della matrice $\boldsymbol{X}$ e il *log-odds* ($\log \frac{\pi_i}{1-\pi_i}$) risulta una funzione lineare dei predittori.
Dall'equazione (2.9) otteniamo:

$$\pi_i = \pi_i(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}} \quad . \tag{2.10}$$

Osserviamo che $0 < \pi_i < 1$.
Nel modello logistico i coefficienti $\boldsymbol{\beta}$ sono stimati con il metodo di massima verosimiglianza; diversamente nella regressione lineare si utilizza il metodo dei minimi quadrati.
L'adeguatezza di un modello lineare si verifica tramite la somma dei quadrati dei residui (RSS), mentre nel modello logistico si usa invece la devianza, definita come:

$$\text{Devianza} = -2\log\frac{L(\boldsymbol{\beta})}{L(\boldsymbol{\beta}_{max})} = 2(l(\boldsymbol{\beta}_{max}) - l(\boldsymbol{\beta}))$$

dove con $l(\boldsymbol{\beta})$ si indica il logaritmo della verosimiglianza $L(\boldsymbol{\beta})$. La devianza é dunque il doppio della differenza tra la *log-verosimigliaza* del modello massimale e la *log-verosimiglianza* del modello utilizzato con p predittori.
Il modello massimale ha un numero di parametri uguale al numero di righe linearmente indipendenti della matrice $\boldsymbol{X}$; se non ci sono repliche tale numero é uguale al numero di unitá sperimentali N. Se la matrice X ha m righe linearmente indipendenti, con $p < m \leq N$, é possibile costruire un modello con m parametri stimabili, cioé costruire una matrice $\boldsymbol{X_{max}}$ con m colonne linearmente indipendenti aggiungendo alla matrice $\boldsymbol{X}$ $m - p$ colonne ottenute come funzioni non lineari delle p colonne originali.
L'adattamento dei dati al modello logistico é tanto migliore, quanto minore é il valore assunto dalla devianza.

2.2.2 La selezione delle variabili tramite la regressione logistica penalizzata

Quando il numero di variabili p é elevato, é necessario selezionarne un sottoinsieme da utilizzare per la costruzione del modello, andando a rimuovere le variabili irrilevanti nella predizione, per evitare modelli sovradeterminati.
Tale selezione ha un duplice vantaggio:

- migliora l'accuratezza del modello, riducendo la varianza;
- rende il modello piú facilmente interpretabile.

Esistono diversi metodi di selezione delle variabili; i principali sono:

***Subset Selection*:** si costruisce il modello identificando un sottoinsieme di predittori. Tale selezione puó essere effettuata impiegando i seguenti algoritmi:

  - *Best subset selection*: prevede di testare tutte le possibili combinazioni di predittori; fra tutte le possibili combinazione viene scelta quella con accuratezza piú alta.
  - *Forward stepwise selection*: algoritmo *greedy* che costruisce un modello iniziale contenente solo l'intercetta e aggiunge ad ogni passo la variabile che produce il maggior miglioramento nel modello, finché tutte le variabili non vengono incluse.
  - *Backward stepwise selection*: algoritmo *greedy* che costruisce un modello iniziale contenente tutte le variabili e rimuove ad ogni passo quella meno significativa.

***Shrinkage*:** il modello viene costruito utilizzando tutti i predittori, ma rispetto ai modelli classici, i coefficienti vengono vincolati ad assumere valori attorno allo zero; appartengono a tale categoria gli algoritmi *Ridge Regression* e *Lasso*.

***Dimension Reduction*:** i p predittori originali vengono proiettati in un sottospazio M dimensionale con $M < p$; tali proiezioni vengono poi usate come nuovi predittori per costruire il modello. Appartengono a tale gruppo i metodi *Principal Component Regression (PCR)* e *Partial Least Square (PLS)*.

Presentiamo ora, nel dettaglio la *Ridge Regression* e il *Lasso*, nel caso della regressione lineare. Analoghe considerazioni possono essere applicate al modello logistico.

La Ridge Regression

La *Ridge Regression* riduce in valore assoluto i coefficienti di regressione imponendo una penalitá sulla loro dimensione.

I coefficienti minimizzano una versione modificata della somma dei quadrati dei residui (RSS):

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.11}$$

dove $\lambda \geq 0$ é il parametro che permette di scegliere il livello di penalitá. Maggiore é λ, maggiore é la penalitá sui coefficienti.
L'equazione (2.11) puó essere vista anche come un problema di ottimizzazione vincolata:

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$
$$\text{soggetta al vincolo} \quad \sum_{j=1}^{p} \beta_j^2 \leq t\,.$$

C'é una corrispondenza uno-a-uno tra t e λ.
É necessario standardizzare le variabili prima di risolvere l'equazione (2.11) perché il valore del generico coefficiente β_i non é invariante per cambiamenti di scala.
É bene notare che il termine relativo all'intercetta non viene penalizzato, perché ció renderebbe il problema dipendente dall'origine scelta per $\boldsymbol{y}$. La soluzione di (2.11) si ottiene prima calcolando il coefficiente dell'intercetta

$$\beta_0 = \frac{1}{N} \sum_{i=1}^{N} y_i$$

e, successivamente stimando i rimanenti coefficienti utilizzando gli input centrati $(x_{ij} - \frac{1}{N}\sum_{i=1}^{n} x_{ij})$.
Nel seguito tratteremo $\boldsymbol{X}$ come una matrice centrata (ogni colonna ha media zero), di dimensione $N \times (p-1)$, in cui la prima colonna $diag(\boldsymbol{I}_N)$ é stata eliminata.
Riscrivendo perció l'equazione (2.11) in forma matriciale abbiamo:

$$RSS(\lambda) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta}$$

e la soluzione della *Ridge Regression* risulta:

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{ridge}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

dove $\boldsymbol{I}$ é la matrice identitá di dimensione $(p-1)\times(p-1)$. Rispetto alla regressione lineare viene aggiunta una costante positiva alla diagonale della matrice $\boldsymbol{X}^T\boldsymbol{X}$ per rendere la matrice complessiva non-singolare.
La scomposizione in valori singolari (SVD) della matrice di input $\boldsymbol{X}$ ci da una maggiore conoscenza della natura della *Ridge Regression*.
La SVD della matrice $\boldsymbol{X} \in \mathbb{R}^{N\times(p-1)}$ ha la forma:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}}$$

dove $\boldsymbol{U}$ e $\boldsymbol{V}$ sono matrici ortogonali rispettivamente di dimensione $N \times (p-1)$ e $(p-1) \times (p-1)$, $\boldsymbol{D}$ é una matrice diagonale che ha come valori $d_1 \geq d_2 \ldots d_{p-1} \geq 0$ chiamati valori singolari di $\boldsymbol{X}$. Se $d_j = 0$ per almeno un j, allora $\boldsymbol{X}$ é una matrice singolare. Applicando alla regressione lineare la scomposizione in valori singolari, la stima del vettore dei minimi quadrati $\hat{\boldsymbol{y}}$ risulta:

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\boldsymbol{ls}} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} = \boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{y}$$

Nel caso della *Ridge Regression* si ha:

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\boldsymbol{ridge}} &= \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^{\boldsymbol{2}} + \lambda\boldsymbol{I})^{-\boldsymbol{1}}\boldsymbol{D}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{y} = \\ &= \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j^T \boldsymbol{y}\end{aligned}$$

dove $\boldsymbol{u_j}$ sono le colonne di $\boldsymbol{U}$. Come la regressione lineare, la *Ridge Regression* calcola le componenti di $\boldsymbol{y}$ rispetto a $\boldsymbol{U}$, ma restringe tali componenti di un fattore

$$\frac{d_j^2}{\lambda + d_j^2} \quad ; \tag{2.12}$$

pertanto la penalizzazione applicata sará tanto maggiore quanto d_j é minore.
Inoltre, la SVD della matrice $\boldsymbol{X}$ é un altro modo di rappresentare le componenti principali della variabile $\boldsymbol{X}$. La matrice di covarianza, se $\boldsymbol{X}$ é centrata, é data da $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}/N$, da cui abbiamo:

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^{\boldsymbol{2}}\boldsymbol{V}^{\boldsymbol{T}}.$$

Gli autovettori $\boldsymbol{v}_j$ sono anche chiamati componenti principali di $\boldsymbol{X}$. La prima componente principale $\boldsymbol{z_1} = \boldsymbol{X}\boldsymbol{v_1}$ ha la piú grande varianza fra tutte le combinazioni lineari normalizzate delle colonne di $\boldsymbol{X}$

$$\mathrm{Var}(\boldsymbol{z_1}) = \mathrm{Var}(\boldsymbol{X}\boldsymbol{v_1}) = \frac{d_1^2}{N}\,.$$

Le successive componenti principali z_j hanno varianza $\frac{d_j^2}{N}$ e sono soggette ad essere ortogonali alle precedenti. L'ultima componente principale ha varianza minima. Le componenti principali che spiegano "poca" varianza, sono associate a un d_j piccolo; la *Ridge Regression* penalizzerá maggiormente queste direzioni come risulta da (2.12).

Il Lasso

Il *Lasso* é un metodo di shrinkage, in cui la stima dei coefficienti si ottiene nel seguente modo:

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.13}$$

Equivalentemente:

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$
$$\text{soggetta al vincolo} \quad \sum_{j=1}^{p} |\beta_j| \leq t.$$

Confrontando le equazioni (2.11) e (2.13) si constata che *Lasso* e *Ridge Regression* presentano una formulazione del tutto analoga; l'unica differenza si ha nel termine che penalizza i coefficienti: nella *Ridge Regression* si usa la norma L_2 ($\sum_{j=1}^{p} \beta_j^2$), mentre nel *Lasso* la norma L_1 ($(\sum_{j=1}^{p} |\beta_j|)$. Quest'ultima norma rende la soluzione non lineare in $\boldsymbol{y}$; non esiste quindi una soluzione in forma chiusa del *Lasso*, ed é necessario usare un algoritmo di programmazione quadratica.
Nel caso in cui la matrice $\boldsymbol{X}$ sia ortonormale, sia *Lasso* che *Ridge Regression* applicano una trasformazione ai $\hat{\boldsymbol{\beta}}$ stimati tramite minimi quadrati. In particolare la *Ridge Regression* applica la trasformazione

$$\frac{\hat{\beta}_j}{1+\lambda},$$

mentre *Lasso* la trasformazione

$$\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+ .$$

Nel caso in cui siano presenti solo due predittori, l'RSS ha le curve di isovalore sul piano (β_1, β_2) a forma di ellissi centrati sulla stima dei minimi quadrati. I vincoli definiscono, su tale piano, una regione che nel caso della *Ridge Regression* é un cerchio, mentre nel caso del *Lasso* é un quadrato (Figura 2.1). Si osservi come il Lasso non solo regolarizzi i coefficienti, ma ne vincoli alcuni al valore nullo, effettuando una selezione intrinseca delle variabili.
La scelta del coefficiente λ avviene, sia per la *Ridge Regression* che per *il Lasso*, tramite la valutazione del λ che produce l'errore minore su un set di osservazioni non impiegate nella costruzione del modello, seguendo un approccio basato sulla *k-fold cross validation*, che tratteremo nel prossimo paragrafo.

Come detto precedentemente, tali approcci *shrinkage* posso essere applicati anche per la regressione logistica; nel caso del *Lasso* la penalitá sui coefficienti, stimati tramite massima verosimiglianza, viene cosí imposta:

$$\max_{\boldsymbol{\beta}} \left\{ L(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

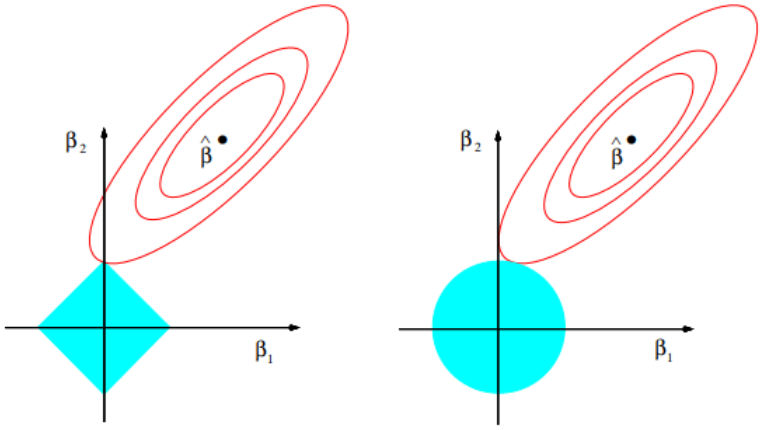


Figura 2.1: Stima dei coefficienti di *Lasso* e *Ridge Regression* nel caso in cui $p = 2$.

2.3 Le prestazioni del modello

La valutazione delle prestazioni sia nell'analisi univariata, sia nell'analisi multivariata viene effettuata in modo analogo.
In particolare, nell'analisi univariata, per le variabili significative, la regola di decisione si basa sulla definizione di un *cut-off*: un soggetto viene classificato sano/malato a seconda che la variabile in considerazione assuma valori superiori/inferiori a tale cut-off.
Nell'analisi multivariata, una volta calcolati i parametri del modello questi vengono moltiplicati per le variabili associate, in modo tale da definire uno score:

$$\text{score}_i = \beta_1 \times x_{i1} + \ldots \beta_m \times x_{im} \quad \text{i=1,...,N}$$

Tale score puó essere interpretato come una nuova variabile e si procede nuovamente alla definizione del cut-off che fornirá la regola di decisione, analogamente all'analisi univariata.

2.3.1 La matrice di confusione: accuratezza, sensibilitá e specificitá

La valutazione di un classificatore viene effettuata confrontando il risultato predetto con il vero (noto) risultato.
La matrice di confusione riassume la performance di un classificatore e permette di calcolare i principali indici di bontá. Un esempio di matrice di confusione é mostrato in Figura 2.2. L'indice piú semplice e *naive* consiste nel considerare la percentuale di classificazioni corrette. Tale indicatore prende il nome di accuratezza di un

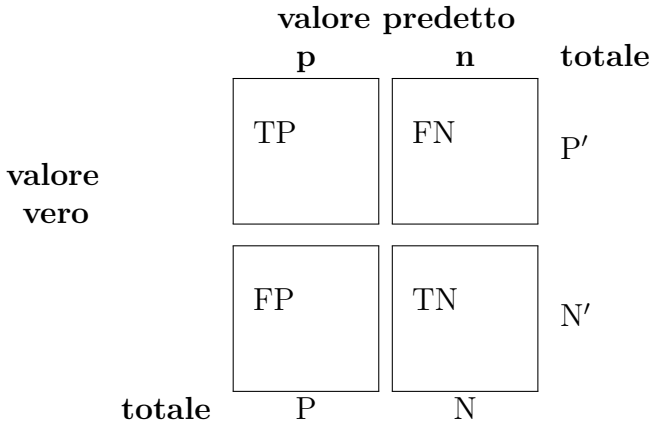


Figura 2.2: Matrice di confusione
TP(veri positivi): numero di soggetti malati correttamente classificati come malati;
TN(falsi negativi): numero di soggetti sani correttamente classificati come sani;
FN(falsi negativi): numero di soggetti malati erroneamente classificati come sani;
FP(falsi positivi): numero di soggetti sani erroneamente classificati come malati.

classificatore. Facendo riferimento alla tabella 2.2, l'accuratezza é cosi definita:

$$\text{accuratezza} = \frac{\text{TP} \ + \text{TN}}{\text{TP} \ + \text{TN} + \text{FN} + \text{FP}} .$$

L'utilizzo dell'accuratezza come metrica di valutazione é sconsigliato, soprattutto in esperimenti non bilanciati, ovvero quando il numero di soggetti appartenenti alle differenti classi varia notevolmente. Supponendo di avere 95 soggetti sani e 5 soggetti malati, un classificatore che etichetta ogni individuo come sano, indipendentemente dal valore assunto dai vari predittori, avrá un'accuratezza del 95%, pur essendo un pessimo classificatore.
Si predilige, dunque, l'uso di due metriche: la sensibilitá e la specificitá cosí definite:

$$\text{Sensibilitá} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.14}$$

$$\text{Specificitá} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.15}$$

La sensibilitá puó quindi essere considerata come la probabilitá che il test (la predizione) sia positivo, dato che il soggetto é realmente positivo. La specificitá, invece, é la probabilitá che un test dia un risultato negativo, dato che il soggetto appartiene alla classe dei negativi. La specificitá e la sensibilitá di un biomarcatore variano a seconda del cut-off che viene scelto per classificare i soggetti come positivi e negativi. Cambiare il cut-off puó aumentare la sensitivitá e ridurre la specificitá, o viceversa.

Facendo riferimento alla Figura 2.3, indicando con $F_+(c)$ (rispettivamente con $F_-(c)$) la distribuzione della generica variabile per il gruppo dei "malati" (rispettivamente dei "sani") si ha:

$$\text{sensibilitá} = \text{Se} = \mathbb{P}(S > c \mid \text{soggetto malato}) = 1 - F_+(c) \tag{2.16}$$

$$\text{specificitá} = \text{Sp} = \mathbb{P}(S \leq c \mid \text{soggetto sano}) = F_-(c) \tag{2.17}$$

dove S é il valore della variabile di cui si vuole valutare il potere discriminatore e c il cut-off. Si é inoltre assunto che un soggetto venga classificato "malato" se la variabile in considerazione assume valori superiori al cut-off, seguendo l'impostazione della Figura 2.3.

2.3.2 La curva ROC

La curva ROC *(Receiver Operating Characteristic)* mostra come sensibilitá e specificitá variano al variare del valore di cut-off.
In particolare, la curva ROC é una curva parametrica costruita con i punti di coordinate (1-specificitá, sensibilitá):

$$ROC = \{(1 - Sp(c), Se(c)) \text{ tale che } -\infty < c < +\infty\}$$

Assumendo che la funzione inversa della distribuzione F (F^{-1}) esista sia per la

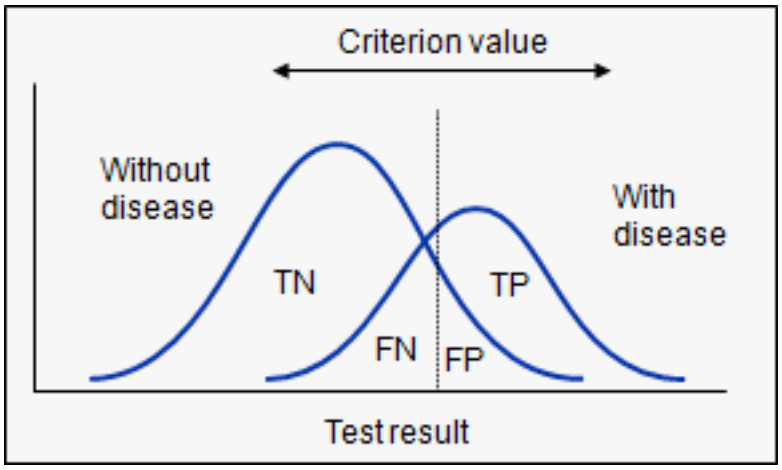


Figura 2.3: Distribuzione di una generica variabile nel gruppo dei "sani" (a destra) e dei "malati" (a sinistra). L'impostazione di un cut-off, linea verticale tratteggiata, permette di definire i TN, FN, FP, TP.

popolazione dei "malati" sia per quella dei "sani", é possibile ridefinire la curva ROC in forma esplicita, eliminando c dalla definizione. Sia $x \in (0,1)$ il valore di (1-Sp(c)),

$$x = 1 - F_-(c)$$

si ha che

$$c = F_{-}^{-1}(1 - x)$$

La curva ROC puó essere espressa dalla seguente relazione:

$$y = 1 - F_{+}(c) = 1 - F_{+}(F_{-}^{-1}(1 - x)) = ROC(x)\,. \tag{2.18}$$

A differenza dell'accuratezza, la curva ROC non risente della prevalenza di soggetti appartenenti a una determinata classe.
Per valutare la qualitá di una curva ROC, e quindi del relativo classificatore, si utilizza l'area sottesa dalla curva ROC (*Area Under the Curve (AUC)*):

$$\text{AUC} = \int_0^1 ROC(q)\,dq\,. \tag{2.19}$$

Tale indice permette di svincolarsi, nella valutazione del classificatore, dall'arbitrarietá della scelta di un unico valore di cut-off. L'AUC é interpretabile come la probabilitá che un valore estratto dalla distribuzione dei "malati" sia maggiore di un valore estratto dalla distribuzione dei "sani", come mostrano i seguenti passaggi:

$$\mathbb{P}(Y_{+} > Y_{-}) = \mathbb{E}\Big(\mathbb{P}(Y_{+} > Y_{-}|Y_{-})\Big) = \int_{-\infty}^{+\infty} \mathbb{P}(Y_{+} > y) f_{-}(y)\,dy =$$
$$\int_{-\infty}^{+\infty} (1 - F_{+}(y)) f_{-}(y)\,dy = -\int_1^0 (1 - F_{+}(F_{-}^{-1}(1 - q)))\,dq = AUC$$

dove f_{-} é la densitá di probabilitá associata a F_{-} e Y_{-}, Y_{+} sono due realizzazioni indipendenti di F_{+} e F_{-}.
Le curve ROC passano per i punti (0,0) e (1,1). Ci sono due particolari ROC che rappresentano casi limite:

- una taglia il grafico a 45°, passando per l'origine. Questa retta rappresenta il caso del classificatore casuale (linea di nessun beneficio), e l'AUC é pari a 0.5. É il caso in cui la distribuzione di una certa variabile per il gruppo dei sani si sovrappone alla distribuzione dei malati.

- l'altra é rappresentata dal segmento che dall'origine sale al punto (0,1) e da quello che congiunge il punto (0,1) a (1,1), ha un'area sottesa di valore pari a 1. É il classificatore perfetto, in cui le distribuzioni nei due gruppi, per la variabile test, sono completamente disgiunte.

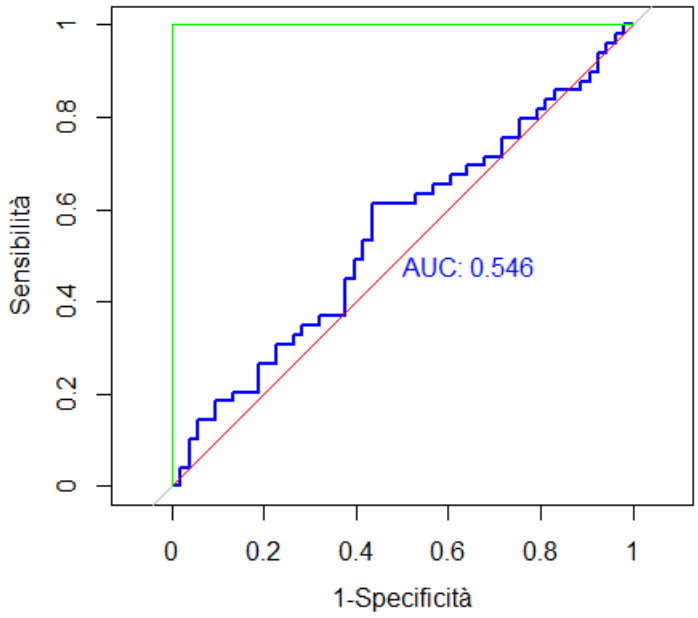


Figura 2.4: Esempio di curve ROC: in blu una generica curva ROC, in verde la curva ROC associata a un classificatore perfetto, in rosso la curva ROC associata a un classificatore casuale.

In Figura 2.4 é mostrato un esempio di curva ROC. Una volta costruita la curva ROC per un generico modello, si passa alla scelta del cut-off ottimale. In generale il punto di cut-off ottimale é quello che corrisponde al punto (Sp=1,Se=1) nella curva ROC. Raramente peró una curva ROC attraversa tale punto. Tipicamente quindi, la scelta del cut-off viene effettuata sulla base dell'indice $Youden$, cosí definito:

$$\text{Youden} = \arg\max_{c}\{Se(c) + Sp(c) - 1\} \tag{2.20}$$

Un' altra scelta tipica é definire il cut-off ottimale come quello corrispondente al punto sulla curva ROC con distanza minima da (Sp=1,Se=1).

2.3.3 La *cross validation*

Indipendentemente dagli indici utilizzati per valutare la qualitá di un classificatore, é necessario che essi vengano calcolati su un insieme di osservazioni che non sono state utilizzate nella creazione del modello.
I dati utilizzati per la creazione del modello (quindi per la definizione dei parametri) costituiscono il *training set* e il relativo errore *training error*; i dati utilizzati per la valutazione del modello vengono definiti *test set* e il relativo errore *test error*.
La valutazione di un modello deve basarsi sul valore del *test error*, che ci da una

visione piú generale di come il modello creato permetta di fare previsioni su nuovi dati. Il *test error* assume valori sempre maggiori del *training error*; tale differenza aumenta all'aumentare della complessitá del modello, quindi all'aumentare delle variabili utilizzate nella costruzione del modello.
Tipicamente, si divide il data-set a disposizione in due parti: una parte dei dati viene utilizzata come *training set* per la costruzione del modello, i rimanenti vengono utilizzati per valutare il modello creato.
Quando il dataset a disposizione é di piccole dimensioni, seguire tale approccio risulta peró impraticabile e le stime prodotte dipendono molto della divisione del dataset. Si segue dunque una procedura chiamata *k-Fold Cross Validation*, che consiste nel suddividere il dataset in k sottoinsiemi di uguale dimensione. Il modello viene costruito k volte, usando ogni volta $k-1$ sottoinsiemi per la costruzione del modello e la restante parte come *test set*. L'errore totale é determinato sommando gli errori dei k modelli costruiti. Tipicamente, l'intera popolazione di campioni viene suddivisa in 5-10 sottoinsiemi ($k = 5, 10$). Tale scelta per i valori di k ha lo scopo di ottimizzare il *trade off* tra varianza ed errore: scegliere k troppo piccolo implica l'utilizzo di una ridotta parte per del dataset per la costruzione del modello rendendolo poco attendibile; diversamente scegliere valori di k molto alti (prossimi a N) riduce il numero di osservazioni su cui testare ciascuno dei k modelli.

2.4 Problemi tipici in metabolomica

Nel seguito presenteremo le principali tematiche legate al pre-processamento dei dati in metabolomica. Descriveremo metodi per la rimozione delle variabiltá non biologiche presenti tra i campioni e metodi per uniformare le concentrazioni dei vari metaboliti. Infine, poiché le tecniche statistiche precedentemente descritte richiedono che i dati non presentino valori mancanti, tratteremo i principali metodi per l'imputazione di dati mancanti.

2.4.1 Metodi di normalizzazione e rimozione degli effetti *batch*

Una delle maggiori problematiche che insorgono negli esperimenti biologici, in cui vengono misurate contemporaneamente migliaia di variabili, é la presenza di forti variazioni tra i diversi campioni, dovute a cause non biologiche. Tali variazioni rendono i campioni difficilmente confrontabili. In metabolomica, queste fonti di variabilitá possono avere origine da una diversa preparazione dei campioni, dalle condizioni dello strumento di analisi, dalle condizioni ambientali, dall'ordine di processamento dei campioni e da proprietá intrinseche del campione analizzato (per esempio campioni a diverso pH).
Queste fonti di eterogeneitá possono ridurre l'accuratezza di un modello statistico.

Sono stati perció sviluppati diversi metodi per correggere tali situazioni nel campo della genomica; diversamente, per la piú recente metabolomica, non esistono algoritmi specifici.
Un primo approccio per ridurre tali variabilitá consiste nello standardizzare le distribuzioni di valori per i diversi campioni. Presentiamo due tra i metodi di normalizzazione piú comuni.

- La normalizzazione *quantile* ha lo scopo di rendere le distribuzioni identiche. A ogni valore, per ciascun campione, viene assegnato un rango, cioé la posizione nell'ordinamento. Per ogni rango viene calcolato il valore medio e sostituito nel dataset originale. Chiariamo il procedimento con un esempio: sia dunque la Tabella 2.1 la tabella originale che deve essere normalizzata, avente sulle colonne i campioni, sulle righe i metaboliti. Per ogni colonna, a ogni valore

Tabella 2.1: Tabella originale.

	campione 1	campione 2	campione 3
var1	6.213	6.471	6.124
var2	6.684	6.009	6.197
var3	6.215	6.116	5.771
var4	6.287	5.923	5.928

viene associato il proprio rango (Tabella 2.2).

Tabella 2.2: Tabella dei ranghi.

	campione 1	campione 2	campione 3
var1	i	iv	iii
var2	iv	ii	iv
var3	ii	iii	i
var4	iii	i	ii

Si calcola il valore medio per ogni rango:

$$\begin{aligned} \text{i} &= (6.213 + 5.923 + 5.771)/3 = 5.969 \\ \text{ii} &= (6.215 + 6.009 + 5.928)/3 = 6.050 \\ \text{iii} &= (6.287 + 6.116 + 6.124)/3 = 6.175 \\ \text{iv} &= (6.684 + 6.471 + 6.197)/3 = 6.450 \end{aligned}$$

Tali valori medi vengono poi sostituiti nella tabella originale (Tabella 2.3).

Tabella 2.3: Tabella normalizzata.

	campione 1	campione 2	campione 3
var1	5.969	6.450	6.175
var2	6.450	6.050	6.450
var3	6.050	6.175	5.969
var4	6.175	5.969	6.050

É evidente come le colonne (i campioni) presentino ora la medesima distribuzione.

- La normalizzazione *scale* modifica i valori di concentrazione in modo tale che le distribuzioni presentino un' identica deviazione media assoluta (MAD *mean absolute deviation*).
  Indicando con $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ la generica distribuzione, la MAD é cosí definita:
  $$\text{MAD} = \text{median}(|x_i - \text{median}(\boldsymbol{x})|)$$
  La MAD é una misura di dispersione. É uno stimatore piú robusto della semplice deviazione standard, poiché ha una minore sensibilitá agli outliers.

Talvolta tuttavia, tali normalizzazioni non sono sufficienti a rendere i campioni completamente confrontabili; in particolare ció si verifica quando si analizzano soggetti provenienti da esperimenti diversi, in cui le variabilitá sono estremamente accentuate. Risulta pertanto necessario introdurre un'altra classe di metodi statistici (gli algoritmi per la rimozione degli effetti *batch*) il cui obiettivo é rimuovere le diversitá presenti nei diversi esperimenti. Un primo algoritmo per la rimozione degli effetti *batch* assume che la generica concentrazione y_{ijg} del metabolita g per il j-esimo soggetto appartenente all' i-esimo *batch* possa essere cosí determinata:

$$y_{ijg} = \alpha_g + \boldsymbol{X}\boldsymbol{\beta}_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg} \tag{2.21}$$

dove α_g é la concentrazione media per il metabolita g, $\boldsymbol{X}$ la matrice del piano sperimentale rappresentante le condizioni del campione, $\boldsymbol{\beta}_g$ é il vettore dei coefficienti corrispondenti a $\boldsymbol{X}$. ϵ_{ijg} rappresenta il termine di errore e, per ipotesi, ha una distribuzione normale con valore atteso 0 e varianza σ_g^2. I termini γ_{ig} e δ_{ig} sono rispettivamente la componente additiva e moltiplicativa associata all'effetto *batch* per il *batch* i e per il metabolita g. Il valore "aggiustato" y^*_{ijg} é dato dalla seguente formula:

$$y^*_{ijg} = \frac{y_{ijg} - \hat{\alpha}_g - \boldsymbol{X}\hat{\boldsymbol{\beta}}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha_g} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_g$$

dove $\hat{\alpha}_g$, $\hat{\beta}_g$, $\hat{\gamma}_{ig}$ e $\hat{\delta}_{ig}$ sono le stime dei rispettivi parametri. Tale modello funziona bene solo se il numero di campioni per ogni *batch* é sufficientemente grande (> 25).

In alternativa si puó quindi adottare un metodo empirico bayesiano, che non risente delle dimensioni del *batch*. Tale metodo, simile a quello precedente descritto, assume che le concentrazioni siano state normalizzate e che i vari metaboliti presentino stessa media e varianza. Si suppone che il dataset sia costituito da k *batch* e che l'i-esimo *batch* contenga n_i campioni. Si assume che valga il modello descritto precedentemente nell'equazione (2.21); si stimano i coefficienti $\hat{\alpha}_g$, $\hat{\beta}_g$ e $\hat{\gamma}_{ig}$ tramite minimi quadrati. Successivamente si definisce:

$$\hat{\sigma}_g^2 = \frac{1}{n}\sum_{i,j}^{n}(y_{ijg} - \hat{\alpha}_g - \boldsymbol{X}\hat{\boldsymbol{\beta}}_g - \hat{\gamma}_{ig})^2.$$

I dati standardizzati z_{ijg} sono cosí calcolati:

$$z_{ijg} = \frac{y_{ijg} - \hat{\alpha}_g - \boldsymbol{X}\hat{\boldsymbol{\beta}}_g}{\hat{\sigma}_g},$$

e si assume che seguano una distribuzione normale:

$$z_{ijg} \sim \mathcal{N}(\gamma_{ig}, \delta_{ig}^2),$$

dove il parametro γ_{ig} non é lo stesso che compare in (2.21). In una prospettiva Bayesiana, si suppone che i parametri γ_{ig} e δ_{ig}^2 abbiano le seguenti distribuzioni a priori:

$$\gamma_{ig} \sim \mathcal{N}(\gamma_i, \tau_i^2) \quad \text{e} \quad \delta_{ig}^2 \sim \text{Gamma Inversa}(\lambda_i, \theta_i)$$

I parametri γ_i, τ_i^2, λ_i e θ_i sono stimati empiricamente tramite il metodo dei momenti. Queste distribuzioni a priori sono state selezionate per la loro proprietá di coniugatezza con la distribuzione Normale.
Le stime di γ_{ig} e δ_{ig}^2 sono date rispettivamente da:

$$\gamma_{ig}^* = \frac{n_i\bar{\tau}_i^2\hat{\gamma}_{ig} + \delta_{ig}^{2*}\bar{\gamma}_i}{n_i\bar{\tau}_i^2 + \delta_{ig}^{2*}} \tag{2.22}$$

$$\delta_{ig}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j(z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_j}{2} + \bar{\lambda}_i - 1} \tag{2.23}$$

Calcolati i coefficienti γ_{ig}^* e δ_{ig}^{2*}, si possono determinare i valori "aggiustati", a cui é stato rimosso l'effetto *batch*:

$$y_{ijg}^* = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}(z_{ijg} - \hat{\gamma}_{ig}^*) + \hat{\alpha}_g + \boldsymbol{X}\hat{\boldsymbol{\beta}}_g.$$

2.4.2 Metodi di pre-elaborazione dei dati

La pre-elaborazione dei dati é utile quando si analizzano dataset di grandi dimensioni, con molte variabili, che possono differire tra loro di diversi ordini di grandezza

e presentare distribuzioni dissimili.
Tale situazione é tipica quando si ha a che fare con dati di metabolomica e piú in generale con dati biologici. Ovviamente l'importanza di una variabile non é data dall'ordine di grandezza che essa assume, ma molti metodi statistici risentono di tali differenze. Inoltre i dataset di metabolomica sono caratterizzati da una grande variabilitá. Alcuni composti chimici, che ricoprono un ruolo fondamentale nel metabolismo, presentano piccolissime variazioni tra campione e campione; altri, con un ruolo secondario, posso variare molto nei diversi soggetti. Alcune tecniche di analisi possono risentire di tale variabilitá.
La scelta di quale tecnica utilizzare per pre-processare i dati dipende dallo scopo delle analisi statistiche (dalle domande a cui il ricercatore vuole rispondere), dalle caratteristiche specifiche del dataset in esame e dalle proprietá dei diversi algoritmi statistici.
Si posso identificare tre classi di metodi di pre-elaborazione (nel seguito indicheremo con $\bar{x}_l$ e s_l la media e la deviazione standard della generica variabile l):

Classe I, Metodi di *centering*: convertono tutte le concentrazioni in fluttuazioni intorno allo zero, invece che intorno al valore medio, focalizzandosi quindi sulla variazione tra campioni e rimuovendo le differenze dei valori assoluti. In formule:

$$\tilde{x}_{jl} = x_{jl} - \bar{x}_l \, .$$

Solitamente questi metodi, il cui effetto é limitato nel caso di dati eteroschedastici, sono applicati in combinazione con quelli delle altre due classi.

Classe II, Metodi di *scaling*: dividono ogni variabile per un fattore (detto fattore di scalamento) differente per ogni variabile. Si suddividono in due gruppi: i metodi che usano una misura di dispersione (quale ad esempio la deviazione standard) come fattore di scala, e quelli che invece usano una misura della concentrazione della variabile (ad esempio la media). Fanno parte del primo gruppo di metodi i metodi di *autoscaling*, *pareto* e *vast scaling*.

***Autoscaling*:** é la tecnica di riscalamento piú frequentemente usata e utilizza la deviazione standard come fattore di riscalamento. Dopo aver applicato l'*autoscaling* e il *centering* le variabili presentano media nulla e varianza unitaria. Ogni metabolita assume uguale importanza. In formule:

$$\tilde{x}_{jl} = \frac{x_{jl} - \bar{x}_l}{s_l}.$$

***Pareto scaling*:** simile a *autoscaling*, ma utilizza la radice quadrata della deviazione standard come fattore di riscalamento. Ció implica una minor dominanza delle variabili con una maggiore varianza e, a differenza dell'autoscaling, i dati mantengono parzialmente intatta la loro struttura. In

formule:

$$\tilde{x}_{jl} = \frac{x_{jl} - \bar{x}_l}{\sqrt{s_l}}.$$

***Vast scaling*:** si focalizza sulle variabili che non presentano grandi variazioni. Come fattore di scala viene utilizzato il rapporto tra la varianza e il valore medio.

$$\tilde{x}_{jl} = \frac{x_{jl} - \bar{x}_l}{s_l^2} \ \bar{x}_l.$$

***Level scaling*:** utilizza come fattore di scalamento il valore medio o in alternativa la mediana. É frequentemente utilizzato per l'identificazione di biomarcatori, nel caso in cui grandi differenze rispetto ai valori medi siano di interesse.

$$\tilde{x}_{jl} = \frac{x_{jl} - \bar{x}_l}{\bar{x}_l}.$$

Classe III, trasformazioni: si applicano funzioni non lineari per rimuovere l'eteroschedasticitá, convertire errori moltiplicativi in errori additivi e rendere le distribuzioni piú simmetriche. Esempi di tali trasformazioni sono la trasformazione logaritmo e la radice quadrata, che riduco maggiormente i valori elevati piuttosto che i valori bassi e per tale motivo hanno pure un effetto di scalamento. Tale effetto peró non riduce completamente le differenze tra le variabili ed é quindi consigliata l'applicazione dei metodi di classe II precedentemente descritti.

2.4.3 La gestione dei valori mancanti

Nei data-set di metabolomica, acquisiti tramite LC-MS, i valori mancanti (NA) sono largamente diffusi. In generale, gli NA posso essere suddivisi in tre tipologie, a seconda della loro origine.

MCAR (*missing complete at random*): i valori mancanti non dipendono né dalla variabile né dal valore assunto dalla variabile. Nel contesto della MS, tali NA derivano da un errore casuale nel processo di acquisizione (ionizzazione incompleta).

MAR (*missing at random*): i valori mancanti dipendono dalla variabile, ma non dal valore assunto da tale variabile (ci sono variabili piú soggette a essere mancanti). In MS tali NA sono dovuti a un pre-processamento non ottimale (errori in fase di *peak detection* o fenomeno di coeluzione).

MNAR (*missing not at random*): i valori mancanti dipendono dal valore assunto dalle variabili. In MS, tali NA sono sostanzialmente dovuti a concentrazioni inferiori a una certa soglia, sotto la quale lo strumento non é in grado di misurare. Si tratta quindi di dati censurati a sinistra.

Non é semplice distinguere tra MCAR e MAR: alcuni metodi di imputazione possono quindi essere usati per entrambi.
Prima di descrivere le varie tecniche di imputazione, ricordiamo che innanzitutto si applica "la regola dell' 80%", ovvero solo i metaboliti presenti in almeno l'80% dei campioni vengono analizzati, i restanti vengono rimossi. Per ridurre il rischio di perdere metaboliti espressi in maniera differente tra i due gruppi ("sani", "malati") tale regola puó essere modificata, andando a rimuovere solo i metaboliti assenti in almeno l'80% dei campioni per entrambi i gruppi.

Classici algoritmi per l'imputazione

A seconda della tipologia di NA, esistono diversi metodi di imputazione. Mentre i metodi per imputare gli NA di tipo MCAR e MAR sono piuttosto generali, questo non accade per i MNAR, per i quali, a seconda del contesto, esistono algoritmi specifici.
Come detto precedentemente, in MS, i MNAR sono dovuti a valori sotto una soglia limite: detta X^* la concentrazione di un generico metabolita per un generico soggetto e d la soglia limite di rilevamento, si osserva la quantitá X cosí definita:

$$X = \begin{cases} X^* & \text{se } X^* \geq d \\ 0 & \text{se } X^* < d \end{cases} \tag{2.24}$$

In [8], vengono confrontati piu metodi di imputazione sia per gli NA MCAR/MAR sia per quelli MNAR. Per quanto riguarda i MCAR/MAR vengono analizzati i seguenti metodi:

Media: i valori mancanti sono sostituiti con la media dei valori non mancanti per la variabile corrispondente.

Mediana: i valori mancanti vengono sostituiti con la mediana dei valori non mancanti per la variabile corrispondente.

kNN *(k Nearest Neighbors)*: viene definita una distanza euclidea tra campioni, basata sulle variabili che non presentano NA; si sostituisce poi il valore mancante mediando sul valore di tale variabile nei k campioni piu vicini al campione che presenta NA. Prima di applicare tale metodo la matrice dei dati deve essere normalizzata, in quanto le variabili devono avere la stessa scala.

***Random Forest*:** sia $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ una matrice $n \times p$. Per un' arbitraria variabile $\boldsymbol{X}_s$ che presenta valori mancanti per i campioni $\boldsymbol{i}_{mis}^{(s)} \subseteq \{1, \ldots, n\}$, il data-set viene separato in quattro parti:

1. i valori osservati (non mancanti) di $\boldsymbol{X}_s$, indicati con $\boldsymbol{y}^{(s)}{}_{obs}$;
2. i valori mancanti per la variabile $\boldsymbol{X}_s$, indicati con $\boldsymbol{y}_{mis}^{(s)}$;

3. i valori delle altre variabili, esclusa $\boldsymbol{X}_s$, corrispondenti ai valori non mancanti di $\boldsymbol{X}_s$, indicati con $\boldsymbol{x}_{obs}^{(s)}$;
4. i valori delle altre variabili, esclusa $\boldsymbol{X}_s$, corrispondenti ai valori mancanti di $\boldsymbol{X}_s$, indicati con $\boldsymbol{x}_{mis}^{(s)}$;

Come primo step viene assegnato un valore iniziale ai valori mancanti (si puó ad esempio usare la media); tale inizializzazione é necessaria in quanto in generale $\boldsymbol{x}_{obs}^{(s)}$ non é privo di missing. Si ordinano poi le variabili in ordine crescente di NA. Per ogni variabile $\boldsymbol{X}_s$, partendo da quella che presenta una minore percentuale di valori mancanti, si costruisce un *Random Forest* utilizzando come variabile risposta $\boldsymbol{y}_{obs}^{(s)}$ e come predittori $\boldsymbol{x}_{obs}^{(s)}$; tale modello viene infine usato per predire i valori mancanti $\boldsymbol{y}_{mis}^{(s)}$ della variabile $\boldsymbol{X}_s$ applicando il precedente *Random Forest* sui predittori $\boldsymbol{x}_{mis}^{(s)}$.

SVD *(Singular Value Decomposition)*: gli NA vengono inizializzati a 0, poi vengono stimati con una combinazione lineare delle prime k componenti principali, costruite applicando la decomposizione a valori singolari alla matrice dei dati. Prima di applicare tale metodo, la matrice dei dati deve essere centrata e riscalata. Un valore tipico per il numero di componenti é $k = 5$.

Per gli MNAR vengono valutati, oltre a kNN, SVD e *Random Forest*, i seguenti metodi:

QRILC *(Quantile Regression Imputation for Left Censored data)*: é un metodo specifico per dati censurati a sinistra e utilizza un modello basato sulla regressione quantile. Diversamente dalla regressione costruita tramite minimi quadrati, che approssima il valore atteso della variabile risposta dati i predittori, la regressione quantile approssima la mediana della variabile risposta, o piú in generale un generico quantile. La stima dei coefficienti richiede un maggiore sforzo computazionale, ma rispetto alla classica regressione si adatta meglio a situazioni in cui la variabile risposta presenta una distribuzione non normale (in particolare risente meno della presenza di *outliers*).
Prima di descrivere come avviene la stima dei parametri nella regressione quantile, é bene ricordare una caratterizzazione del quantile in generale. Il $\tau-$quantile della generica distribuzione di valori $\{y_t\}_{t=1}^n$, puó essere definito come soluzione del seguente problema di minimo:

$$\min_{b\in\mathbb{R}} \left[\sum_{t\in\{t:y_t\geq b\}} \tau|y_t - b| + \sum_{t\in\{t:y_t<b\}} (1-\tau)|y_t - b| \right] \tag{2.25}$$

con $\tau \in (0,1)$; per $\tau = 1/2$, b assume il valore di mediana.
La (2.25) ci aiuta nella definizione della regressione quantile: siano $\{(y_t, \boldsymbol{x}_t) \in$

$\mathbb{R}^{1+p}\}_{t=1}^{n}$ la variabile risposta e i p predittori per la generica unitá sperimentale t; la stima dei coefficienti della regressione quantile si ottiene:

$$\min_{\boldsymbol{b}\in\mathbb{R}^p}\Big[\sum_{t\in\{t:y_t\geq \boldsymbol{b}\cdot\boldsymbol{x_t}\}}\tau|y_t-\boldsymbol{b}\cdot\boldsymbol{x_t}|+\sum_{t\in\{t:y_t<\boldsymbol{b}\cdot\boldsymbol{x_t}\}}(1-\tau)|y_t-\boldsymbol{b}\cdot\boldsymbol{x_t}|\Big].$$

I valori mancanti vengono quindi sostituiti con elementi estratti a caso da una distribuzione stimata con tale regressione quantile. Prima di utilizzare questo metodo é bene applicare una trasformazione logaritmica ai dati per migliorarne l'accuratezza.

Zero: questo metodo sostituisce gli NA con il valore zero.

HM *(Half Minimum)*: questo metodo sostituisce gli NA con la metá del valore minimo assunto dalla corrispondente variabile nei vari campioni.

Per valutare il metodo piú accurato, nel caso di MCAR/MAR, si utilizza il NRMSE (*Normalized Root Mean Square Error*):

$$NRMSE=\sqrt{\frac{mean(x^{(true)}-x^{(imp)})}{var(x^{true})}};\tag{2.26}$$

vengono generati casualmente valori mancanti da un dataset completo, e si valuta quanto il valore imputato differisce dal valore originale, secondo l'$NRMSE$.
Per i MNAR, si utilizza invece il $NRMSE-based\,sum\,of\,rank\,(SOR)$, che é cosí definito:

$$SOR=\sum_{i=1}^{q}Rank_i(NRMSE),\tag{2.27}$$

dove q é il numero delle variabili che presentano valori mancanti e $Rank_i(NRMSE)$ é il rango del metodo considerato per la variabile i. Dato un dataset completo si selezionano casualmente alcune variabili e si impongono le concentrazioni inferiori a un certo cut-off uguali a NA.
Secondo quanto investigato in [8], per il caso MCAR/MAR, l'algoritmo piú efficiente risulta essere il *Random Forest*, seguito da SVD e kNN (per il quale le performance calano all'aumentare della proporzione di NA). Per i MNAR invece l'algoritmo piú performante é la QRILC, seguito da HM.

Capitolo 3

Analisi dei dati reali

In questo capitolo verranno presentati i risultati delle analisi condotte su un dataset reale, gentilmente concesso dal laboratorio di Farmacogenomica dei Tumori, Fondazione Edo e Elvo Tempia di Biella.
Il codice *R* utilizzato é riportato in appendice.

3.1 Descrizione del *dataset*

Nel dataset analizzato sono presenti 102 campioni per i quali sono stati misurati, mediante la spettrometria di massa accoppiata con la cromatografia liquida (LC-MS), le concentrazioni di 254 metaboliti. Inoltre per ogni campione sono disponibili anche il valore dell'antigene prostatico specifico (PSA) e l'etá.
Di questi 102 pazienti, 53 sono affetti da iperplasia prostatica benigna (BPH), mentre i restanti hanno contratto il carcinoma della prostata (PCa). Nel dataset tale informazione é identificata nella variabile "Type".
La tabella 3.1 mostra un estratto del dataset: lungo le righe sono disposti i campioni mentre nelle colonne ci sono i metaboliti, l'etá , il PSA e il "Type".
Lo scopo delle analisi é quello di individuare dei metaboliti che presi singolarmente o in combinazione permettano di discriminare i due gruppi di pazienti (PCa e BPH). Oltre ai metaboliti abbiamo incluso nelle analisi, come predittori, anche l'etá e il PSA (che, come descritto nell'introduzione, puó presentare valori alterati sia nel caso di carcinoma, sia nel caso di iperplasia).
Per quanto riguarda i metaboliti, nel nostro dataset essi vengono identificati da una coppia di numeri reali, la massa e il *retention time*, piú un codice progressivo:

$$\mathrm{ID}x\_m\_RT$$

dove IDx é il codice progressivo con x da 1 a 254, m é la massa del metabolita e RT il relativo *retention time*.
Alcuni campioni presentano dei valori mancanti; le possibili cause possono essere:

- il metabolita risulta assente per tali campioni;
- la concentrazione risulta sotto un valore soglia;
- sono occorsi degli errori nella fase di misurazione.

Nel nostro caso, per la maggior parte dei metaboliti, si tratta di valori mancanti per concentrazioni inferiori al limite di rilevabilitá. Complessivamente il dataset presenta 52 valori mancanti.

Tabella 3.1: Estratto del dataset originale

	ID1_167_1870	...	ID254_402_1465	Etá	PSA	Type
Campione 1	2997876	...	2083192	64	5.5	PCa
Campione 2	4066690	...	4658797	78	4.92	BPH
Campione 3	2336552	...	6709999	78	5.8	PCa
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Campione 102	3651409	...	3886580	80	6.2	PCa

3.2 Analisi preliminare del *dataset*

I soggetti analizzati hanno un'età compresa tra 46 e 84 anni e, come si osserva nel *box-plot* in Figura 3.1, la mediana dell'età del gruppo dei PCa risulta maggiore di quella del gruppo dei BPH. Per quanto riguarda il PSA, facendo riferimento al *box-plot* in Figura 3.2, notiamo che tutti i soggetti presentano un valore superiore a 4 (limite di normalitá), le due distribuzioni sono confrontabili e le mediane si trovano allo stesso livello; tuttavia nel gruppo dei BPH sono presenti alcuni *outlier* (con PSA supera a 13). Da tale *box-plot* e dalla Figura 3.3, che mostra la distribuzione di densitá del PSA nei due gruppi, risulta evidente che il potere discriminatore del PSA nel nostro dataset é basso.
Procediamo ora con l'analisi preliminare dei metaboliti. Come detto precedentemente sono presenti 52 valori mancanti (NA). La Figura 3.4 é una fotografia del dataset in cui sulle colonne ci sono i campioni e sulle righe i metaboliti: i valori mancanti sono colorati in rosso. I metaboliti e i campioni sono ordinati in base alla percentuale di valori mancati. Tale matrice mostra come sia presente un metabolita mancante in 35 campioni. In particolare, esso é assente in 19 campioni del gruppo BPH e in 16 campioni del gruppo PCa; risulta dunque mancante per piú dell'80% dei campioni in entrambi i gruppi e, secondo la regola descritta nel paragrafo 2.4.3, deve essere rimosso dal dataset.
Applichiamo poi la trasformazione logaritmica (in base 10) alle concentrazioni dei vari metaboliti. Tale trasformazione ha lo scopo di rendere le distribuzioni dei

singoli metaboliti piú simmetriche. Puó essere interessante (almeno in fase esplorativa) rappresentare, per alcuni metaboliti, le concentrazione nei vari campioni (Figura 3.5): ad esempio, per i primi 4 metaboliti sembra evidenziarsi la presenza di due gruppi. Tali gruppi non sono determinati dalla variabile "Type" ma piuttosto dall'ordine dei campioni nel dataset. La presenza di tali gruppi é confermata dalla PCA riportata nella Figura 3.6. Rappresentando le prime due componenti principali, vediamo come la prima componente principale (asse x) separi i campioni in due gruppi distinti. L'elaborazione della PCA richiede che il *dataset* sia completo, ovvero privo di valori mancanti. In tale fase abbiamo applicato il metodo media (descritto nel paragrafo 2.4.3), in quanto non invalida i risultati della PCA.
Inoltre, si evidenzia la natura di tali gruppi di campioni visualizzando i *box-plot* per campione (Figura 3.7). Ogni *box-plot* rappresenta un campione e riassume il valore dei vari metaboliti: ci si attende che tali *box-plot* siano molto simili tra loro, in quanto alterazioni sarebbero indice di grandi variazioni per piú metaboliti. In particolare, si osserva come i primi 52 campioni (il 52^o é attraversato da una linea) assumano valori maggiori rispetto ai restanti 50. La presenza di tali due gruppi puó essere ricondotta all'analisi strumentale ed all'algoritmo di estrazione utilizzato (il quale prevede la suddivisione dell'intero dataset in due gruppi casuali di campioni). Si tratta quindi di un effetto *batch* che deve essere eliminato. Nel prossimo paragrafo si cercherá di omogenizzare l'intero dataset.

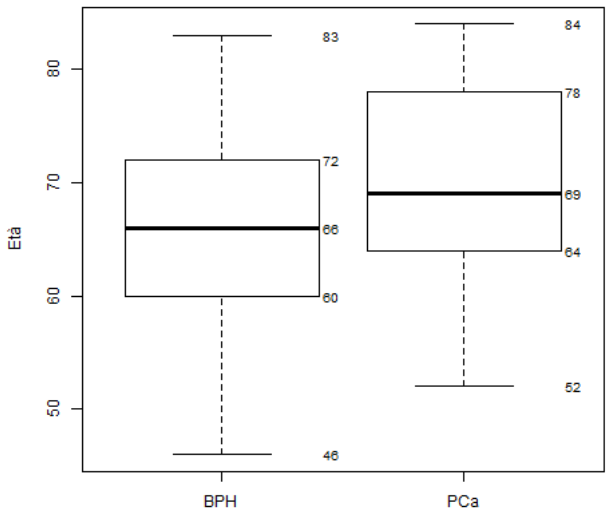


Figura 3.1: *Box-plot* delle etá nei due gruppi (PCa e BPH).

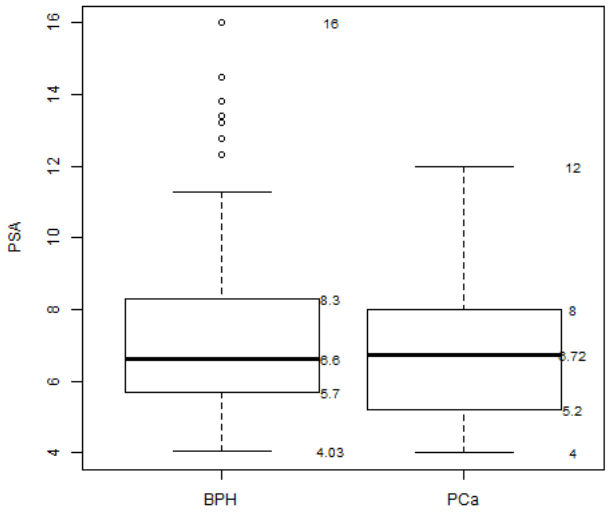


Figura 3.2: *Box-plot* del PSA nei due gruppi (PCa e BPH).

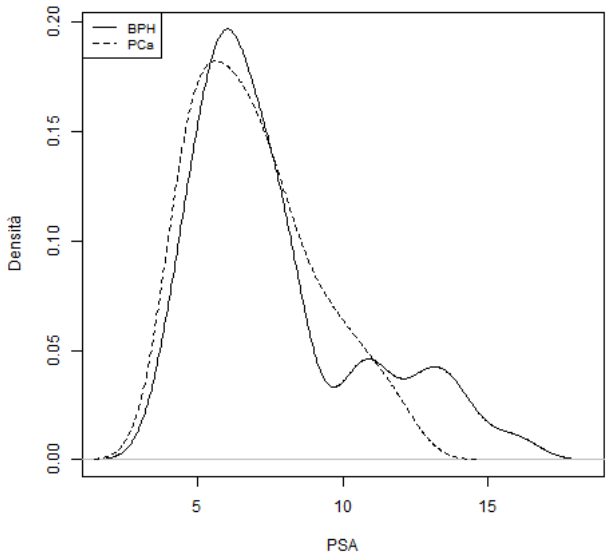


Figura 3.3: Distribuzione del PSA nei due gruppi (PCa e BPH).

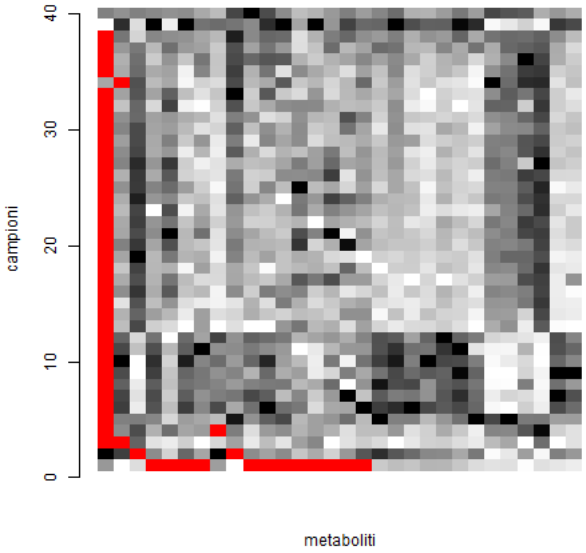
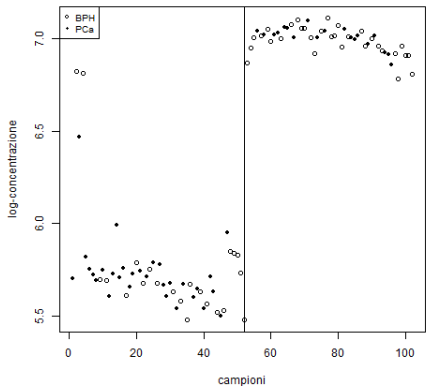
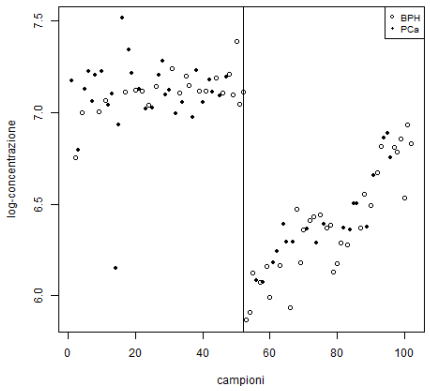


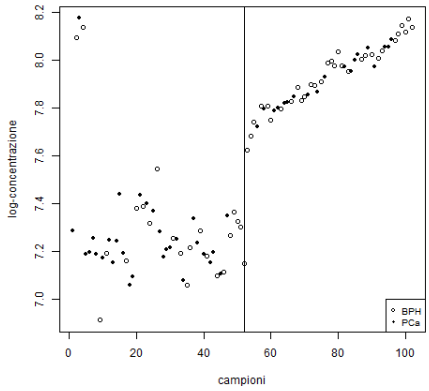
Figura 3.4: *Shadow matrix* (ridotta) con la distribuzione dei valori mancanti.



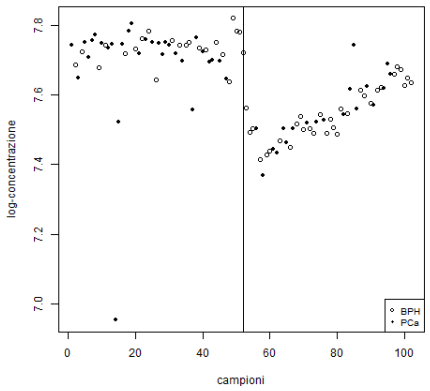
(a) Metabolita ID11_312_734



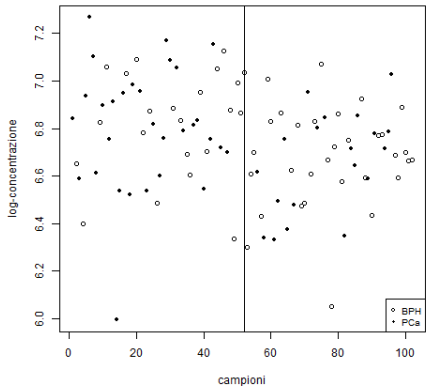
(b) Metabolita ID12_313_1469



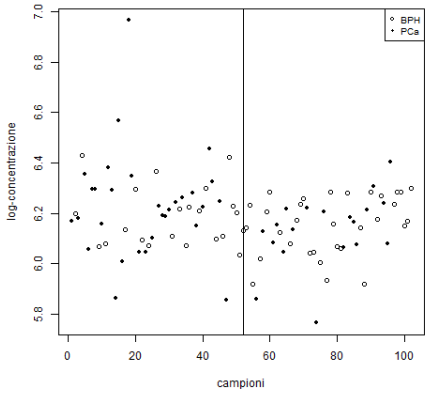
(c) Metabolita ID21_343_1069



(d) Metabolita ID25_401_1257



(e) Metabolita ID125_197_1771.803936



(f) Metabolita ID185_403_1462

Figura 3.5: Concentrazioni di alcuni metaboliti per i vari campioni.

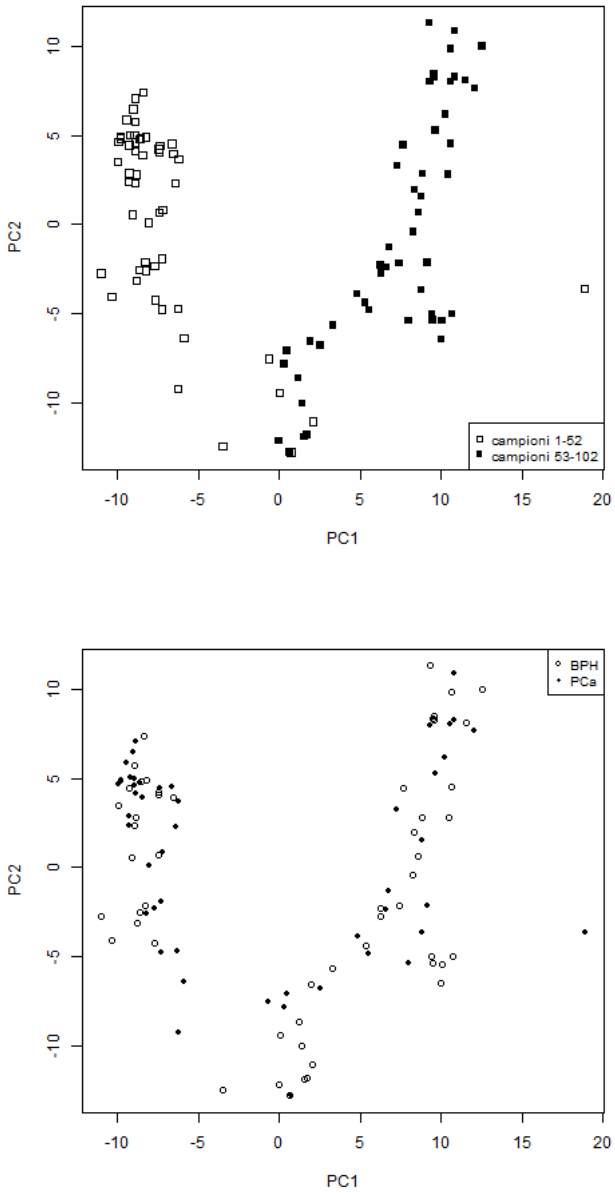


Figura 3.6: Proiezione dei vari campioni sulle prime due componenti principali definite dalla PCA.

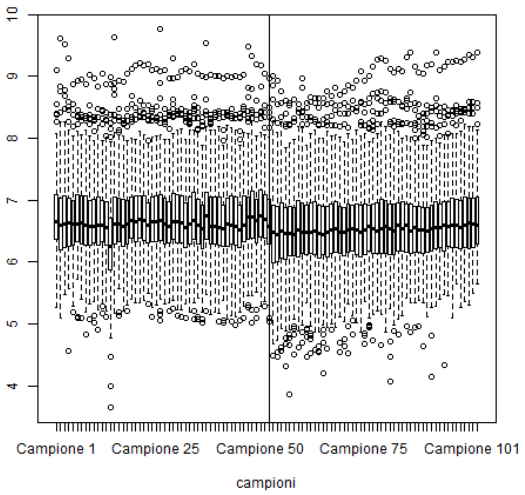


Figura 3.7: Analisi preliminare dell'intero dataset mediante *box-plot*, che riassume le concentrazioni dei vari metaboliti per ogni campione.

3.3 Pre-elaborazione del *dataset*

Come spiegato dettagliatamente nella Sezione 2.4.1, per rimuovere le variazioni non biologiche tra i campioni é opportuno innanzitutto normalizzare i campioni. Abbiamo quindi applicato la normalizzazione *scale*, implementata nel *software* *R* [17] dalla funzione *NormalizeBetweenArray* del pacchetto *limma* [12]. L'effetto di tale normalizzazione é evidente nella Figura 3.8: rispetto allo stesso grafico in Figura 3.7 (creato sui dati originali), notiamo come le mediane dei vari *box-plot* siano ora allineate; purtroppo si identificano ancora due distinti gruppi, confermati ulteriormente nei grafici della PCA (Figura 3.9) e delle concentrazioni dei singoli metaboliti (Figura 3.10).
Seppur gli algoritmi per la rimozione degli effetti di *batch* vengano generalmente usati quando si devono confrontare soggetti provenienti da esperimenti diversi, si é ritenuta necessaria la loro applicazione anche al caso in esame, cosí da ripristinare l'omogeneitá del dataset che risultava alterata dall'analisi strumentale.
Abbiamo applicato dunque ai dati l'algoritmo di rimozione degli effetti di *batch* che segue un approccio empirico bayesiano (paragrafo 2.4.1). Questo metodo, implementato in *R* dalla funzione *ComBat* nel pacchetto *sva* [13], richiede come input il *batch* di appartenenza di ogni campione. Tale informazione non risulta disponibile in quanto non é noto come siano stati definiti i gruppi nell'algoritmo di estrazione; abbiamo basato quindi la suddivisione sulle considerazioni emerse dall'osservazione dei grafici riportati in Figure 3.5 e 3.7: i primi 52 campioni apparterranno al *batch*

1, i restanti 50 al *batch* 2. Rimosso l'effetto *batch* abbiamo proceduto con la normalizzazione.
Osservando le Figure 3.11, 3.12 e 3.13 notiamo come i campioni risultino ora piú sparsi e disordinati (indice di variabilitá prettamente biologica).

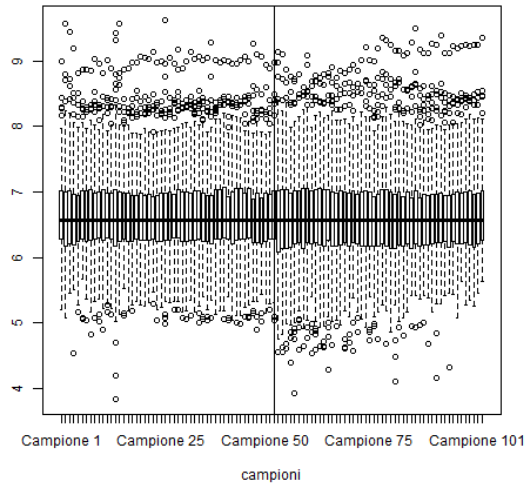


Figura 3.8: Analisi preliminare dell'intero dataset mediante *box-plot*, che riassume le concentrazioni dei vari metaboliti per ogni campione, dopo aver applicato la normalizzazione *scale*.

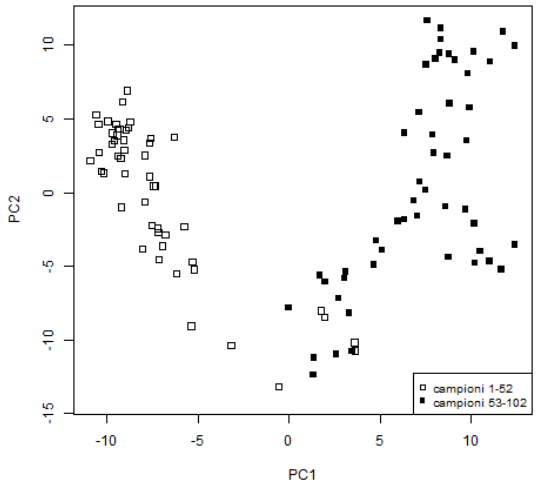
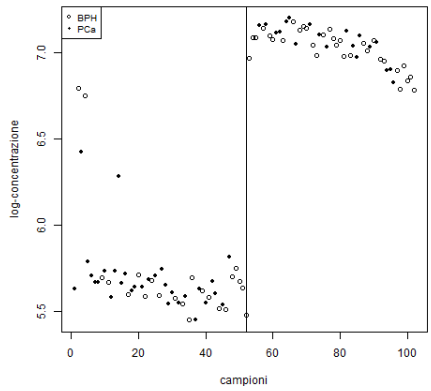
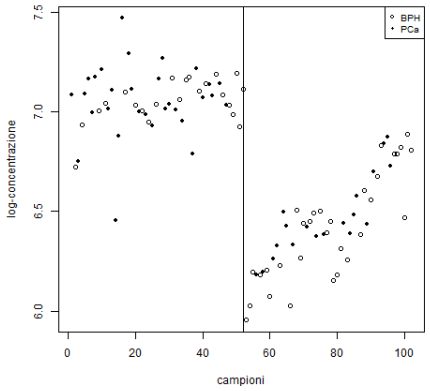


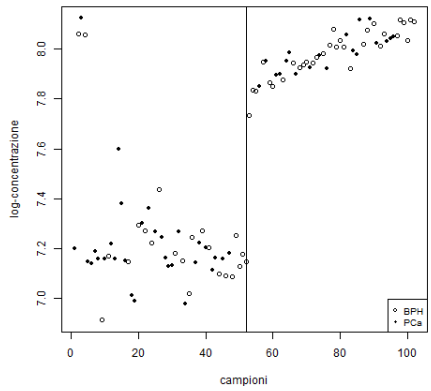
Figura 3.9: Proiezione dei vari campioni sulle prime due componenti principali definite dalla PCA, dopo aver applicato la normalizzazione *scale*.



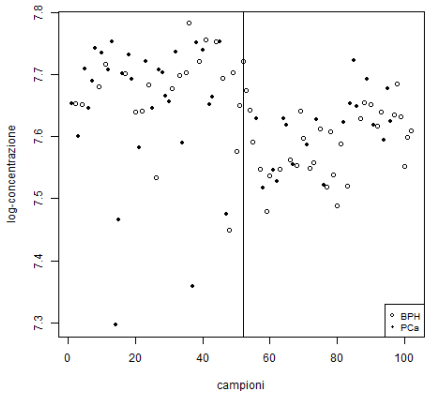
(a) ID11_312_734



(b) ID12_313_1469



(c) ID21_343_1069



(d) ID25_401_1257

Figura 3.10: Concentrazioni di alcuni metaboliti per i vari campioni, dopo aver applicato la normalizzazione *scale*.

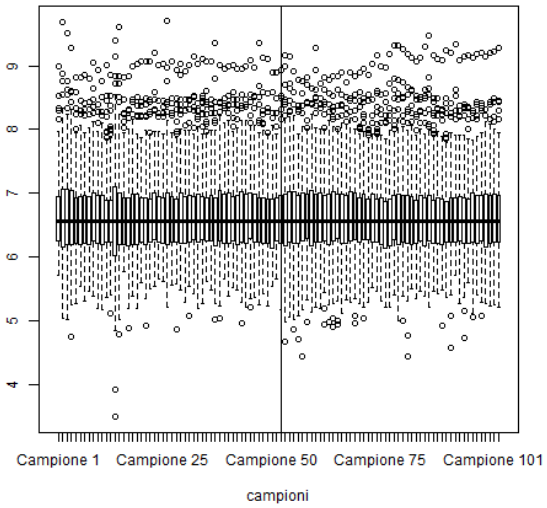


Figura 3.11: Analisi preliminare dell'intero dataset mediante *box-plot*, che riassume le concentrazioni dei vari metaboliti per ogni campione, dopo aver applicato la rimozione degli effetti *batch* (*ComBat*) e la normalizzazione (*scale*).

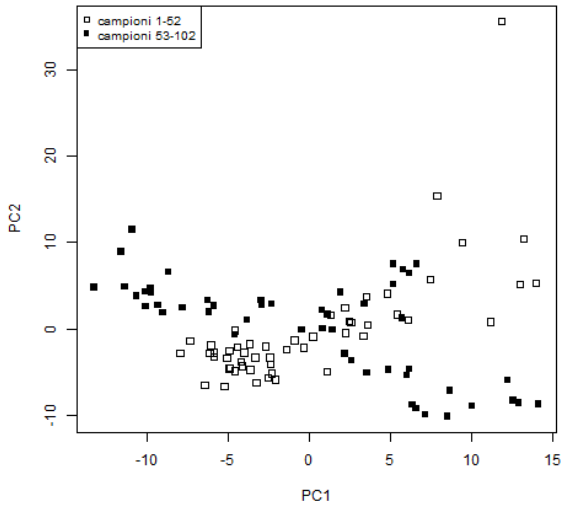
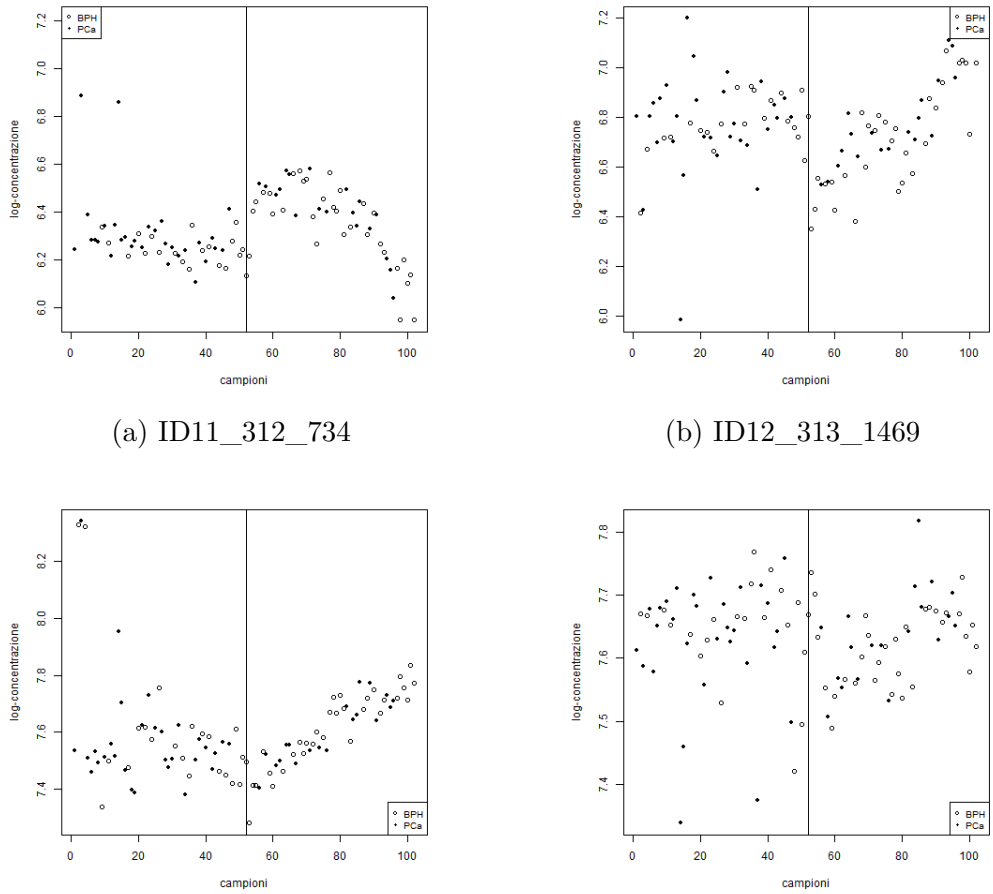


Figura 3.12: Proiezione dei vari campioni sulle prime due componenti principali definite dalla PCA, dopo aver applicato la rimozione degli effetti *batch* (*ComBat*) e la normalizzazione (*scale*).



(a) ID11_312_734

(b) ID12_313_1469

(c) ID21_343_1069

(d) ID25_401_1257

Figura 3.13: Concentrazioni di alcuni metaboliti per i vari campioni, dopo aver applicato la rimozione degli effetti (*ComBat*) e la normalizzazione (*scale*).

A questo punto, abbiamo poi imputato i valori mancanti.
Una prima imputazione col metodo media era giá stata effettuata per poter verificare, mediante la PCA, la presenza di eventuali sottogruppi (*batch*).
Facendo riferimento alla Figura 3.4:

- si era osservato che un metabolita risultava mancante per piú dell' 80% dei campioni e pertanto era stato rimosso dall'analisi del dataset;

- si osserva che un campione presenta 12 valori di concentrazione mancanti e dunque si ipotizza che la causa sia attribuibile a un problema di analisi strumentale (quindi rientrante nella casistica MCAR o MAR): per tale campione verrá applicato il metodo di imputazione *Random Forest* ;

- si individuano ulteriori 5 valori mancati sparsi, relativi a 4 diversi campioni; si ipotizza che la causa sia attribuibile a concentrazioni con valori al di sotto di una soglia, rientranti quindi nella casistica MNAR; verrá dunque applicato il metodo *QRILC*.

Naturalmente, le ipotesi effettuate hanno una validitá limitata in quanto non suffragate da indicazioni specifiche giustificanti i valori mancati. Tuttavia l'impatto di tali ipotesi sulle analisi statistiche risulta trascurabile data la ridotta numerositá di tali valori mancanti (17 valori mancanti in totale, 0.06% del dataset). I metodi di imputazione sono stati applicati in seguito alla normalizzazione e alla rimozione di effetti di *batch* per garantire l'uniformitá della base di dati da cui stimare i valori mancanti.

3.4 Analisi univariata

Terminata la fase di pre-elaborazione, abbiamo eseguito un'analisi univariata al fine di identificare le variabili piú significative per la costruzione del modello di predizione del carcinoma alla prostata.
Abbiamo utilizzato la funzione *univariate* del pacchetto *R muma* [16].
Per ogni variabile, *univariate* valuta la normalitá della distribuzione nei due gruppi attraverso il test di *Shapiro-Wilk*. Per le variabili che presentano una distribuzione normale, la diversificazione tra il gruppo BPH e PCa viene valutata tramite il *t-test* di *Welch's*, in cui le varianze dei due gruppi non vengono assunte uguali. Per le restanti, si utilizza il test non parametrico di *Wilcoxon Mann-Whitney*.
I risultati di tali test sono riportati in Figura 3.14 attraverso il grafico *volcano*. Si osservi come correggendo i *p-value* per confronti multipli nessuna variabile risulti significativa, in quanto tali *p-value* aggiustati presentano valori molto superiori al 5%, adottato come criterio di significativitá.
Se invece si considerano p-value non corretti (Figura 3.15), alcuni metaboliti risultano significativi. Tali valori sono riassunti nella Tabella 3.2a.
Abbiamo poi ripetuto l'analisi univariata utilizzando il pacchetto *limma* applicando la versione "moderata" del t-test [14] per individuare i metaboliti diversamente espressi nei due gruppi. Di nuovo, correggendo i *p-value* per confronti multipli non compaiono variabili significative (Figura 3.16), mentre per *p-value* non corretti il *volcano* in Figura 3.17 ne evidenzia alcune. Le variabili piú significative ($p\text{-value} < 0.1$) sono elencate nella Tabella 3.2b.
Confrontando i risultati in Tabella 3.2, si nota come i primi quattro metaboliti (in ordine di significativitá) siano gli stessi: i risultati del t-test "moderato" di *limma* e della funzione *univariate* di *muma* sono pertanto confrontabili.
Abbiamo infine applicato alcune trasformazioni, in particolare *autoscaling* (quella piú diffusamente impiegata per questo tipo di dati) e *level scaling* come suggerito da [7]. L'analisi é stata condotta solamente applicando il *t-test* "moderato". Nelle

Figure 3.18, 3.19, 3.20 e 3.21 sono riportati i *volcano plot* associati a tali analisi; anche in questo caso la correzione dei *p-value* non permette di identificare variabili significative.
In Tabella 3.3 vengono riportati i *p-value*, non corretti per test multipli, dei metaboliti piú significativi.
Confrontando le Tabelle 3.2 e 3.3, osserviamo che i metaboliti con *p-value* (non aggiustato) inferiore a 0.1 sono pressoché gli stessi; le diverse trasformazioni utilizzate non influiscono dunque sul risultato del t-test "moderato", o comunque hanno un effetto trascurabile.

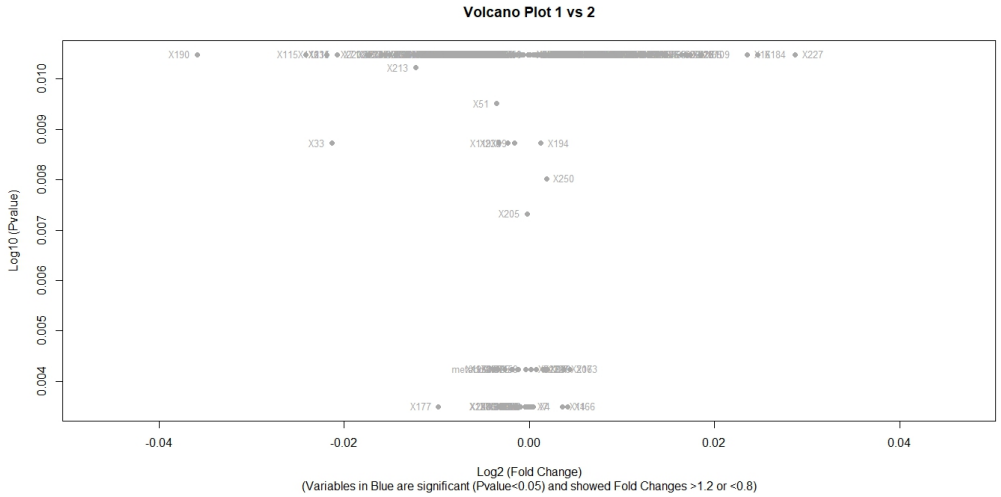


Figura 3.14: *Volcano plot*: p-value corretti per confronti multipli (muma).

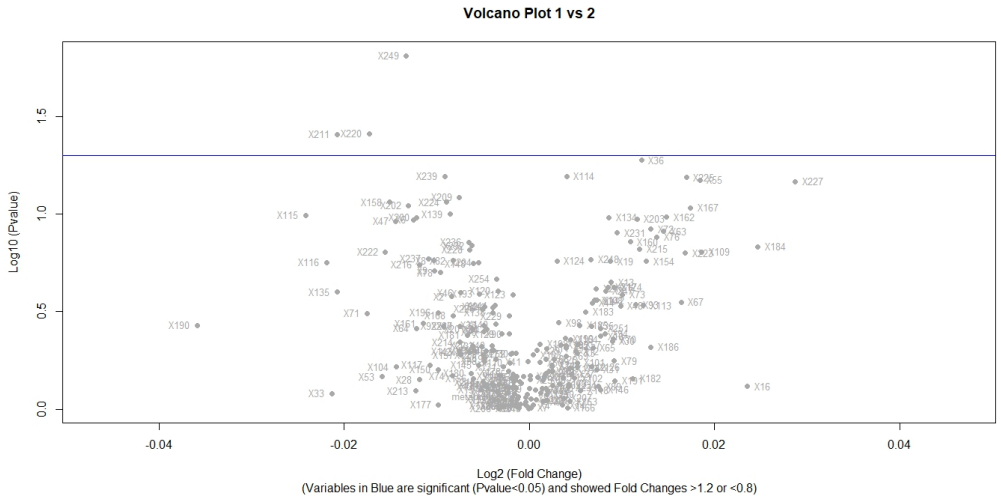


Figura 3.15: *Volcano plot*: p-value non aggiustati (muma).

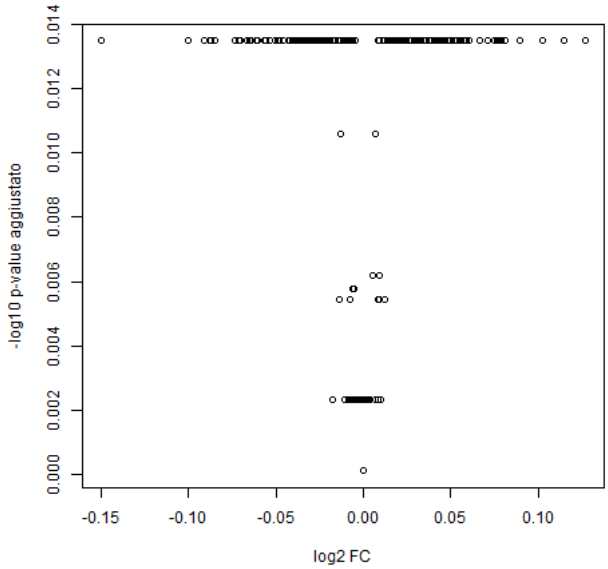


Figura 3.16: *Volcano plot*: *p-value* aggiustati per confronti multipli (limma).

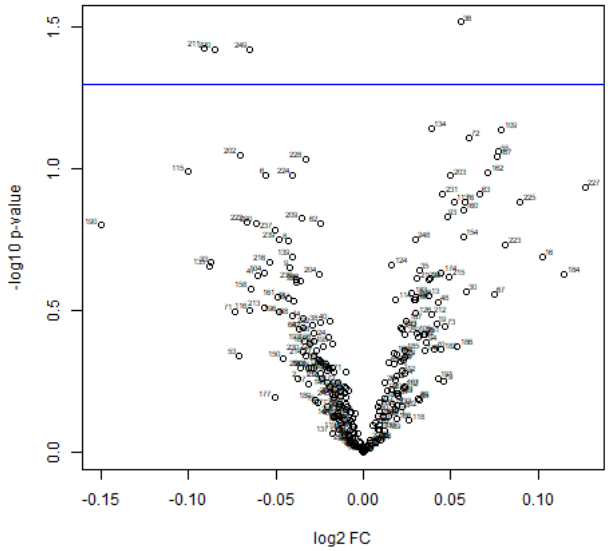


Figura 3.17: *Volcano plot*: *p-value* non aggiustati (limma).

Tabella 3.2: Metaboliti con *p-value* non corretti inferiori a 0.1.

(a) muma

metaboliti	p-value
ID249_329_770	0.0155
ID220_441_1036	0.0386
ID211_485_1041	0.0391
ID36_417_1257	0.0529
ID114_388_1204	0.0640
ID239_354_1023.	0.0640
ID225_326_1520	0.0650
ID55_247_1794	0.0670
ID227_298_1408	0.0680
ID209_329_1145	0.0822
ID224_330_1056.	0.0870
ID158_465_1391	0.0870
ID202_397_1030.	0.0902
ID167_314_1220	0.0931

(b) limma

metaboliti	p-value
ID36_417_1257	0.0302
ID211_485_1041	0.0375
ID249_329_770	0.0377
ID220_441_1036	0.0379
ID134_471_1438	0.0719
ID109_372_1392	0.0724
ID72_433_1169	0.0776
ID55_247_1794	0.0868
ID202_397_1030.	0.0890
ID167_314_1220	0.0901
ID228_343_1201	0.0924

Tabella 3.3: Metaboliti con *p-value* non corretti inferiori a 0.1, dopo aver applicato le trasformazioni *autoscaling* e *level scaling*.

(a) auto scaling

metaboliti	p-value
ID36_417_1257	0.0296
ID249_329_770	0.0384
ID211_485_1041	0.0400
ID220_441_1036	0.0403
ID134_471_1438	0.0661
ID109_372_1392	0.0755
ID72_433_1169	0.0788
ID228_343_1201	0.0827
ID55_247_1794	0.0901
ID202_397_1030.	0.0918
ID167_314_1220	0.0934

(b) level scaling

metaboliti	p-value
ID36_417_1257	0.0301
ID211_485_1041	0.0374
ID249_329_770	0.0380
ID220_441_1036	0.0383
ID134_471_1438	0.0706
ID109_372_1392	0.0722
ID72_433_1169	0.0775
ID55_247_1794	0.0866
ID167_314_1220	0.0901
ID202_397_1030.	0.0903
ID228_343_1201	0.0946

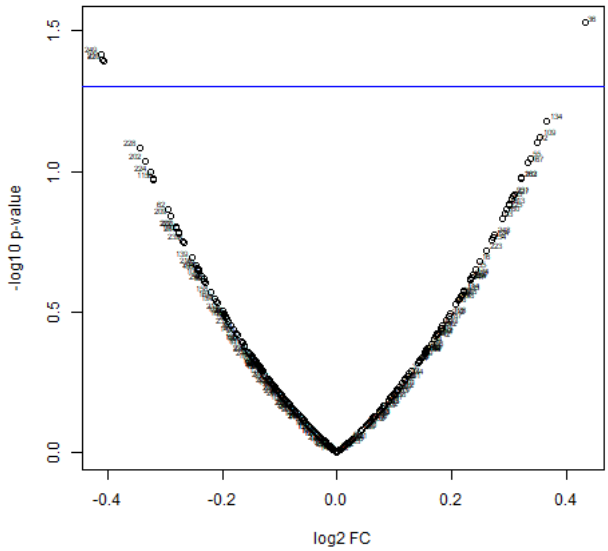


Figura 3.18: *Volcano plot* per dataset trasformato tramite *autoscaling* (p-value non aggiustati).

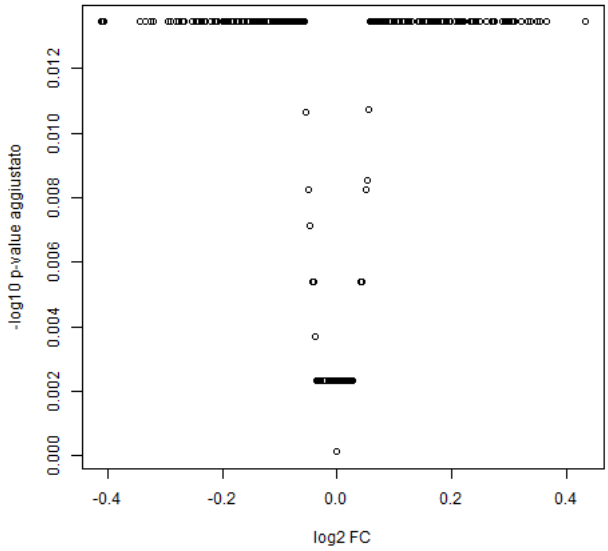


Figura 3.19: *Volcano plot* per dataset trasformato tramite *autoscaling* (p-value aggiustati).

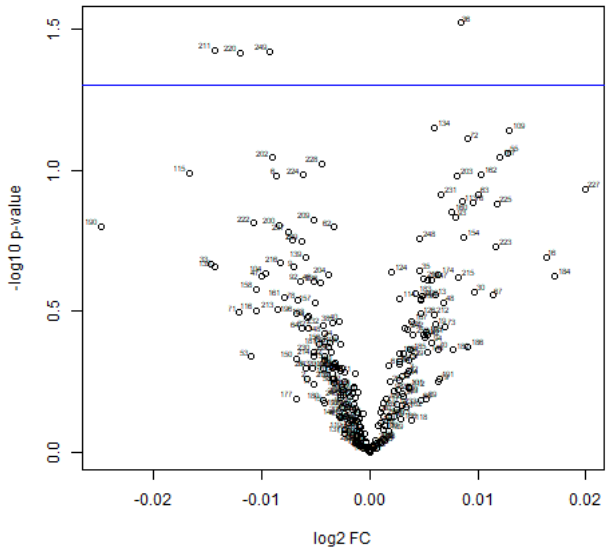


Figura 3.20: *Volcano plot* per dataset trasformato tramite *level scaling* (p-value non aggiustati).

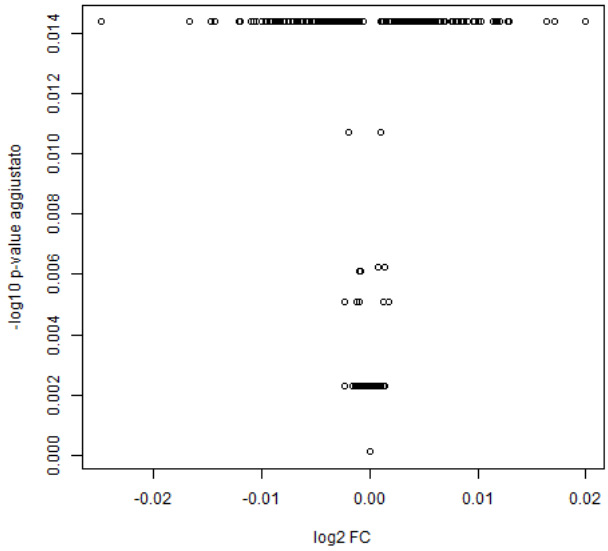


Figura 3.21: *Volcano plot* per dataset trasformato tramite *level scaling* (p-value aggiustati).

3.5 Analisi multivariata

Per quanto riguarda l'analisi multivariata abbiamo applicato all'intero *set* di metaboliti, inclusi l'etá e il PSA, il metodo *Lasso* implementato in *R* dalla funzione *glmnet* dell'omonimo pacchetto [15]. Tale metodo pone a zero i coefficienti dei predittori poco associati alla variabile risposta.
La standardizzazione delle variabili, che risulta indispensabile come descritto nel paragrafo 2.2.2, viene effettuata in automatico dalla funzione precedentemente citata.
La selezione del parametro λ, che penalizza i coefficienti $\boldsymbol{\beta}$ associati alle variabili, é effettuata attraverso la *cross validation* (con $k = 10$) ; si ricerca quel valore di λ per cui l'errore commesso sia minimo, e in particolare, nel caso della regressione logistica l'errore viene associato alla devianza.

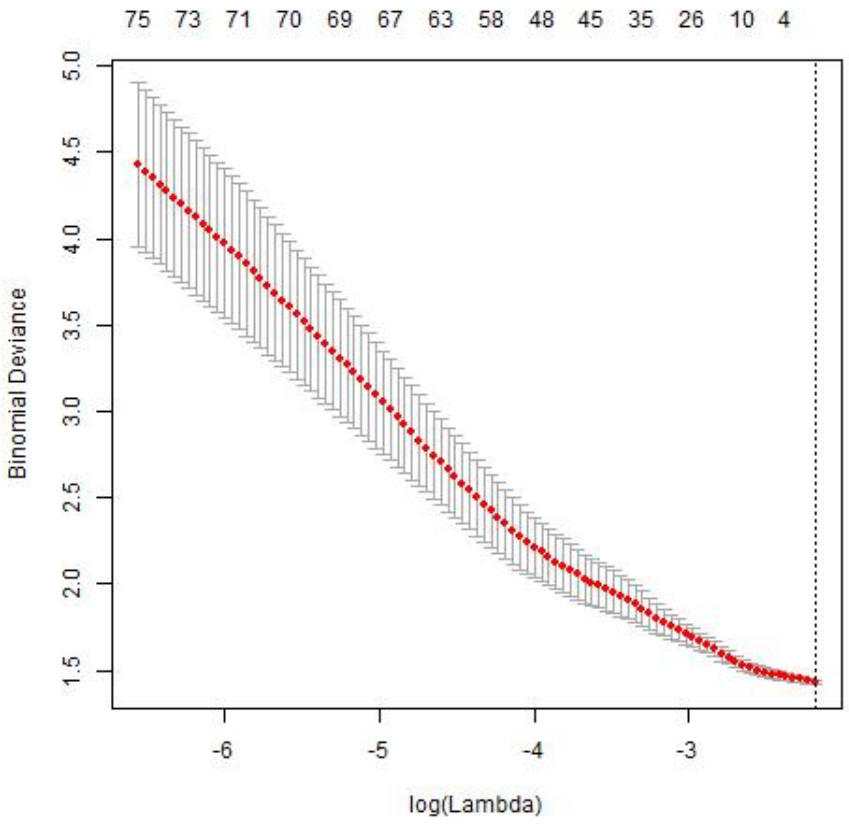


Figura 3.22: Andamento della devianza al variare del parametro di penalizzazione λ (Metodo *Lasso*).

In Figura 3.22 viene riportato l'andamento della devianza al variare del parametro λ; nel caso specifico si puó osservare una funzione monotona decrescente, il cui minimo viene raggiunto annullando tutti i coefficienti.
Nonostante ció, si proverá ad investigare le prestazioni del modello costruito utilizzando le ultime 5 variabili di cui *Lasso* pone a zero i coefficienti, di seguito riportate:

- ID36_417_1257,
- Etá,
- ID249_329_770,
- ID220_441_1036,
- ID167_314_1220.

3.6 Identificazione del modello e valutazione

Prima di analizzare i modelli statistici basati sui metaboliti, facciamo alcune osservazioni sulle prestazioni del PSA nel dataset in esame.
Si era osservato in Figura 3.2 che i valori del PSA per tutti i pazienti erano superiori a 4; tale valore corrisponde al cut-off attualmente adottato nella diagnosi del carcinoma alla prostata. Da questa osservazione possiamo affermare che, nel caso in esame, il test del PSA ha specificitá uguale a zero: nessun paziente viene predetto come "sano". Inoltre la curva ROC derivante da tale biomarcatore, sempre nel dataset in esame, ha un indice $AUC = 0.523$ (valore prossimo al classificatore casuale).
Procediamo ora alla descrizione dei modelli costruiti basandoci sui risultati delle analisi univariata e multivariata.
I risultati delle analisi univariate hanno identificato tutte, indipendentemente dalla trasformazione applicata, le seguenti variabili significative:

- ID36_417_1257,
- ID220_441_1036
- ID211_485_1041
- ID249_329_770.

Si é pertanto deciso di costruire un modello composto dalla variabile Etá (che come osservato in Figura 3.1 risulta abbastanza differente tra i due gruppi) e dai metaboliti sopra elencati, senza ricorrere alle trasformazioni di Classe II (*auto scaling* e *level scaling*) testate nel paragrafo 3.4.

Al modello logistico costruito su tali variabili é stata applicata la *cross validation* imponendo un numero di sottoinsiemi $k = 10$.
La valutazione del modello cosí ottenuto é stata effettuata tramite la costruzione della curva *ROC* e in particolare attraverso la valutazione dell'indice *AUC* (Figura 3.23). Il cut-off ottimale é stato quindi calcolato a partire da tale curva con i criteri *Youden* e della minima distanza dal punto con specificitá e sensibilitá unitarie.
Le prestazioni del modello sono sinteticamente riportate di seguito:

$$\text{AUC} = 0.612$$
$$\textit{Youden} = 0.221$$
$$\textit{min-distance} = 0.571$$

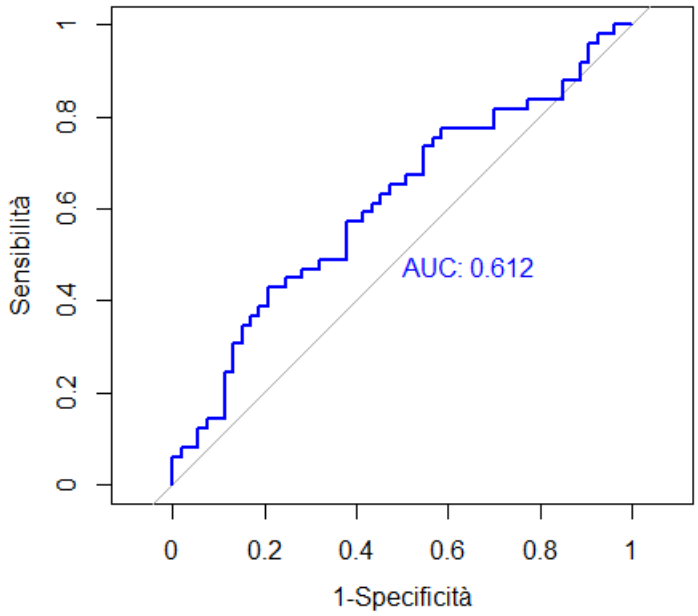


Figura 3.23: Curva ROC per il modello derivante dalle analisi univariata.

La procedura precedentemente descritta é stata ripetuta anche per i modelli che seguono. In particolare, si sono effettuati alcuni tentativi per identificare delle varianti di tale modello che portassero a un miglioramento delle prestazioni; il piú promettente é risultato essere quello definito dalle seguenti variabili:

- ID36_417_1257,
- ID220_441_1036,

- ID211_485_1041,
- Etá.

In Figura 3.24 é riportata la curva ROC ottenuta, mentre le prestazioni del modello sono:

$$\begin{aligned} \text{AUC} &= 0.645 \\ \textit{Youden} &= 0.303 \\ \textit{min-distance} &= 0.513\,. \end{aligned}$$

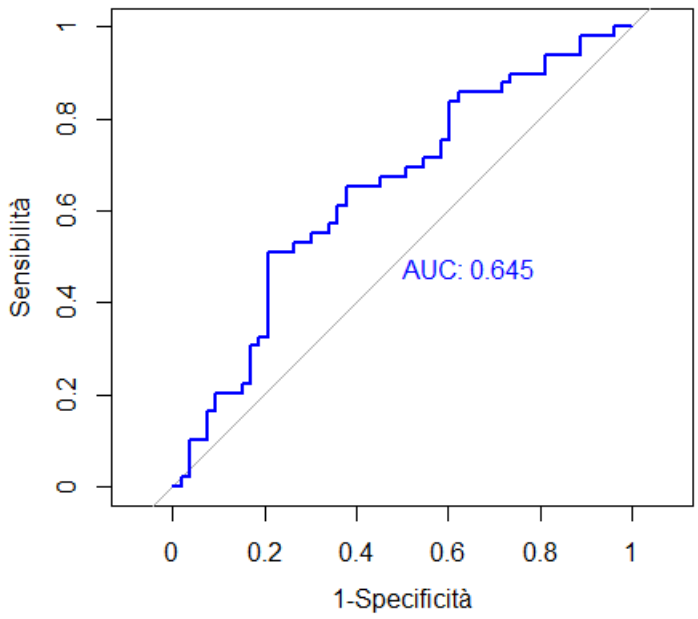


Figura 3.24: Curva ROC per il modello modificato derivante dalle analisi univariata.

Per tale modello, scelto il cut-off definito dall'indice *Youden*, riportiamo la matrice di confusione in Figura 3.25.
Da tale tabella ricaviamo la specificitá e la sensibilitá:

$$\begin{aligned} \text{specificitá} &= 0.793 \\ \text{sensibilitá} &= 0.510\,. \end{aligned}$$

		valore predetto	
		PCa	BPH
valore vero	PCa	25	24
	BPH	11	42

Figura 3.25: Matrice di confusione per il modello modificato derivante dalle analisi univariata.

Lo stesso modello, privato della variabile "Etá" dá invece i seguenti risultati:

$$\text{AUC} = 0.610$$
$$\text{Youden} = 0.254$$
$$\text{min-distance} = 0.528\,.$$

La curva ROC corrispondente é rappresentata in Figura 3.26.
Confrontando tali indici, con quelli del modello precedente, si constata una netta inferioritá di quest'ultimo modello; l'etá risulta dunque una variabile importante ai fini della predizione.
Infine, é stato costruito un modello sulla base dei risultati dell'analisi multivariata (paragrafo 3.5) considerando le seguenti variabili:

- ID36_417_1257,
- Etá,
- ID249_329_770,
- ID220_441_1036,
- ID167_314_1220.

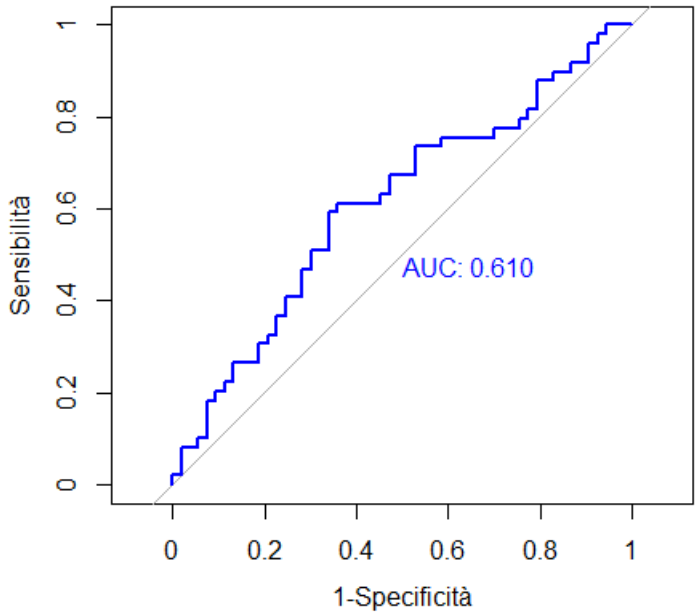


Figura 3.26: Curva ROC per il modello modificato in cui é stata rimossa dal *set* di predittori anche etá.

In Figura 3.27 viene mostrata la curva ROC associata a tale modello. I valori dei principali indicatori sono:

$$\begin{aligned} \text{AUC} &= 0.637 \\ \textit{Youden} &= 0.246 \\ \textit{min-distance} &= 0.547\,. \end{aligned}$$

Mentre la sensibilitá e la specificitá (utilizzando il cut-off associato all'indice *Youden*):

$$\begin{aligned} \text{specificitá} &= 0.490 \\ \text{sensibilitá} &= 0.755 \end{aligned}$$

In accordo con quanto indicato da Xia [3], le prestazione dei modelli analizzati, seppur migliori del PSA, ricadendo negli intervalli $AUC = 0.6 - 0.7$ risultano ancora limitate.
L'analisi dei modelli ha permesso di rafforzare quanto giá inizialmente evidenziato dalle analisi univariata e multivariata; in particolare nelle analisi univariata non

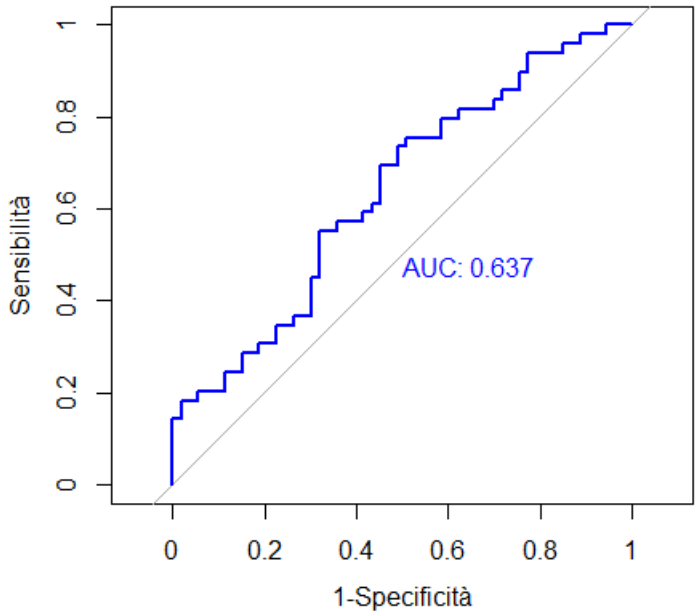


Figura 3.27: Curva ROC per il modello derivante dalle analisi multivariata.

comparivano metaboliti significativi correggendo per test multipli, mentre nell'analisi multivariata il valore minimo della devianza al variare del parametro di penalitá era identificato dal modello con nessun predittore.
Ricordiamo che i vari modelli e gli indicatori ad essi associati sono stati definiti tramite la *cross validation* per evitare il fenomeno di *overfitting*; un diverso partizionamento (casuale) dei campioni in fase di *cross validation* puó quindi portare a risultati leggermente differenti.

Capitolo 4

Conclusioni

Il recente sviluppo della metabolomica é stato favorito alla nascita di tecniche strumentali, quali la risonanza magnetica nucleare e la spettrometria di massa, in grado di misurare, attraverso un unico esperimento, migliaia di metaboliti contemporaneamente (approccio omico).
La disponibilitá di questa vasta mole di informazioni ha permesso di migliorare il processo di identificazioni dei reali fattori associabili ai diversi stati fisiologici.
La necessitá di elaborazione di questa grande varietá di dati ha richiesto e favorito lo sviluppo di numerosi metodi statistici specifici.
Si é osservarvato in letteratura una proliferazione di studi e ricerche che descrivono l'applicazione di tali metodi statistici ai piú svariati campi della biomedicina per l'identificazione di nuovi biomarcatori non invasivi.
A differenza della genomica, la metabolomica presenta una difficoltá maggiore legata all'elevata variabilitá dei metaboliti stessi (essi sono infatti il risultato dell'interazione dell'espressione genica con l'ambiente in cui viviamo).
Il caso analizzato nella presente trattazione riguarda uno dei tumori piú diffusi, quello prostatico. Ad oggi, l'antigene prostatico specifico (PSA) é il biomarcatore piú utilizzato per la diagnosi di tumore alla prostata. Tuttavia, a causa delle elevate percentuali di falsi positivi e falsi negativi, non risulta essere un buon indicatore diagnostico. Il set di dati usato per le analisi é stato fornito dal laboratorio di Farmacogenomica dei Tumori, Fondazione Edo e Elvo Tempia di Biella e consisteva in 102 campioni per i quali sono stati misurati, mediante la spettrometria di massa accoppiata con la cromatografia liquida, le concentrazioni di 254 metaboliti. Di questi 102 pazienti, 53 risultavano affetti da iperplasia prostatica benigna (BPH) mentre i restanti hanno contratto il carcinoma della prostata (PCa).
L'obiettivo dello studio condotto é stato di identificare nuovi possibili biomarcatori tra i metaboliti misurabili nel sangue. Si é costruito un modello logistico, selezionando le variabili sulla base dei risultati di analisi univariate e multivariate. Diversi metodi di normalizzazione e trasformazione sono stati applicati per ridurre le variabilitá non biologiche del dataset. I modelli analizzati hanno permesso di

individuare alcuni metaboliti "promettenti" e di costruire un classificatore la cui capacitá di predizione appare superiore al PSA. Tuttavia i risultati ottenuti sono ancora troppo limitati per poter includere questi metaboliti come nuovi biomarcatori diagnostici del tumore alla prostata.
Una continua ricerca per migliorare le analisi strumentali associata ad una maggior conoscenza dei fenomeni biologici in esame ed alla possibilitá di raccolta di dati relativi a una popolazione piú estesa potranno facilitare in futuro l'identificazione di piú efficaci biomarcatori.

Appendice A

Codice R utilizzato

1. "Analisi preliminare del *dataset* ":

```
# Leggo dataset, trasformo in NA i valori 0, applico log10

dataset=read.table("HighFreq_Represented_mod.txt",...
...sep = "\t",dec=".",header = T)
rownames(dataset)=dataset$Name
dataset$Name=NULL
metaboliti=dataset[,1:254]
metaboliti[metaboliti==0]=NA
dataset[,1:254]=metaboliti
sum(is.na(metaboliti))
metaboliti=log10(metaboliti)

#  Box-plot PSA
png("boxPlotPSA.png")
boxplot(as.numeric(as.character(dataset$PSA))~...
...as.character(dataset$Type),ylab="PSA")
text(y=fivenum(dataset$PSA[which(dataset$Type=="BPH")]),...
... labels =fivenum(dataset$PSA[which(dataset$Type=="BPH")]),...
... x=1.45,cex = 0.8)
text(y=fivenum(dataset$PSA[which(dataset$Type=="PCa")]),...
...labels =fivenum(dataset$PSA[which(dataset$Type=="PCa")]),...
...x=2.45,cex = 0.8)
dev.off()

# Density plot PSA
datasetOrd=dataset[order(rownames(dataset)),]
png("densitypsa.png")
plot(density(as.numeric(as.character(datasetOrd[1:53,"PSA"]))),...
...main="",lty=1)
```

```
lines(density(as.numeric(as.character(datasetOrd[54:102,"PSA"]))),...
...lty=2)
legend("topleft", legend=c("BPH", "PCa"), lty=c(1,2), cex=0.8)
dev.off()

# Box-plot Eta
png("boxPlotEta.png")
boxplot(dataset$Age~dataset$Type)
text(y=fivenum(dataset$Age[which(dataset$Type=="BPH")]),...
...labels =fivenum(dataset$Age[which(dataset$Type=="BPH")]),...
... x=1.45,cex = 0.8)
text(y=fivenum(dataset$Age[which(dataset$Type=="PCa")]),...
... labels =fivenum(dataset$Age[which(dataset$Type=="PCa")]),...
... x=2.45,cex = 0.8)
dev.off()

# Valutazione valori mancanti: shadow matrix
metabolitiOrdinatiperNA=metaboliti[order(-apply(FUN = sum,...
...X = is.na(metaboliti),MARGIN = 1)),order(-apply(FUN = sum,...
...X = is.na(metaboliti),MARGIN = 2))]
install.packages("VIM")
library(VIM)
png("shadowmatrix.png")
matrixplot(metabolitiOrdinatiperNA[1:40,1:30], interactive = F,...
...xlab  = "metaboliti",ylab = "samples")
dev.off()

# Elimino metabolita 176 poiche ha
#una percentuale di valori mancanti >80%
metaboliti=metaboliti[,-176]
dataset=dataset[,-176]
mancanti=is.na(metaboliti)

# Imputo con medie x PCA
medie=apply(X = metaboliti,MARGIN = 2,mean,na.rm=T)
for(j in 1:ncol(metaboliti)){
metaboliti[is.na(metaboliti[,j]),j]=medie[j]
}
dataset[,1:253]=metaboliti[,1:253]

# Per valutare se sia necessario
# normalizzare i dati:
# box-plot campioni, esempi di concentrazioni, PCA

# Box-plot  campioni
png("boxplotcampioni.png")
```

```
boxplot(t(metaboliti),xlab="campioni")
abline(v=52)
dev.off()

# Plot delle concentrazioni di alcuni metaboliti
png("concentrazione11.png")
plot(metaboliti[,11],xlab="campioni",ylab="log-concentrazione",...
...pch=ifelse(dataset$Type=="PCa",20,1))
abline(v=52)
legend("topleft", legend=c("BPH", "PCa"),
pch=c(1,20), cex=0.8)
dev.off()
png("concentrazione12.png")
plot(metaboliti[,12],xlab="campioni",ylab="log-concentrazione",...
...pch=ifelse(dataset$Type=="PCa",20,1))
abline(v=52)
legend("topright", legend=c("BPH", "PCa"),
pch=c(1,20), cex=0.8)
dev.off()
png("concentrazione21.png")
plot(metaboliti[,21],xlab="campioni",ylab="log-concentrazione",...
...pch=ifelse(dataset$Type=="PCa",20,1))
abline(v=52)
legend("bottomright", legend=c("BPH", "PCa"),
pch=c(1,20), cex=0.8)
dev.off()
png("concentrazione25.png")
plot(metaboliti[,25],xlab="campioni",ylab="log-concentrazione",...
...pch=ifelse(dataset$Type=="PCa",20,1))
abline(v=52)
legend("bottomright", legend=c("BPH", "PCa"),
pch=c(1,20), cex=0.8)
dev.off()
png("concentrazione125.png")
plot(metaboliti[,125],xlab="campioni",ylab="log-concentrazione",...
...pch=ifelse(dataset$Type=="PCa",20,1))
abline(v=52)
legend("bottomright", legend=c("BPH", "PCa"),
pch=c(1,20), cex=0.8)
dev.off()
png("concentrazione185.png")
plot(metaboliti[,184],xlab="campioni",ylab="log-concentrazione",...
...pch=ifelse(dataset$Type=="PCa",20,1))
abline(v=52)
legend("topright", legend=c("BPH", "PCa"),
pch=c(1,20), cex=0.8)
```

```
dev.off()

# PCA
pca <- prcomp(metaboliti,
center = TRUE,
scale. = TRUE)

png("PCAimpMedieType.png")
plot(pca$x[,1:2],pch=ifelse(dataset$Type=="PCa",20,1))
legend("topright", legend=c("BPH", "PCa"),
pch=c(1,20), cex=0.8)
dev.off()

png("PCAimpMediePos.png")
plot(pca$x[,1:2],pch=c(rep(0,52),rep(2,50)))
legend("bottomright", legend=c("campioni 1-52",...
... "campioni 53-102"),
pch=c(0,2), cex=0.8)
dev.off()

save(metaboliti,mancanti,dataset,file"FineAnalisiPreliminare.Rda")
```

2. "Pre-elaborazione del *dataset*":

```
load("FineAnalisiPreliminare.Rda")
# Normalizzazione scale
source("https://bioconductor.org/biocLite.R")
biocLite("limma")
library(limma)
metaboliti=as.data.frame(t(normalizeBetweenArrays(t(metaboliti),...
...method = "scale" ) ))

pca <- prcomp(metaboliti,
center = TRUE,
scale. = TRUE)
jpeg("PCAimpMedieNorm.jpg")
plot(pca$x[,1:2],col=c(rep("red",52),rep("blue",50)))
dev.off()
jpeg("boxplotcampioniNorm.jpg")
boxplot(t(metaboliti))
abline(v=52,col="red")
dev.off()
jpeg("concentrazione11Norm.jpg")
plot(metaboliti[,11],xlab="disposizione campioni",...
...ylab="log-concentrazione", col=c(rep("red",52),rep("blue",52)))
dev.off()
```

```
jpeg("concentrazione12Norm.jpg")
plot(metaboliti[,12],xlab="disposizione␣campioni",...
...ylab="log-concentrazione", col=c(rep("red",52),rep("blue",52)))
dev.off()
jpeg("concentrazione21Norm.jpg")
plot(metaboliti[,21],xlab="disposizione␣campioni",...
ylab="log-concentrazione",...
col=c(rep("red",52),rep("blue",52)))
dev.off()
jpeg("concentrazione25Norm.jpg")
plot(metaboliti[,25],xlab="disposizione␣campioni",...
...ylab="log-concentrazione", col=c(rep("red",52),rep("blue",52)))
dev.off()

#La normalizzazione da sola non e' sufficiente
# Rimozione dell'effetto  batch (con ComBat) e
# normalizzazione scale
source("https://bioconductor.org/biocLite.R")
biocLite("sva")
library(sva)
load("FineAnalisiPreliminare.Rda")
metaboliti=as.data.frame(t(ComBat(t(metaboliti),...
...batch =c(rep("A",52),rep("B",50)),par.prior = T)))
metaboliti=as.data.frame(t(normalizeBetweenArrays(t(metaboliti),...
...method = "scale" ) ))
# Controllo per vedere se ComBat funziona
pca <- prcomp(metaboliti, center = TRUE, scale. = TRUE)
jpeg("PCAimpMedieNormRem.jpg")
plot(pca$x[,1:2],col=c(rep("red",52),rep("blue",50)))
dev.off()
jpeg("boxplotcampioniNormRem.jpg")
boxplot(t(metaboliti))
abline(v=52,col="red")
dev.off()

jpeg("concentrazione11NormRem.jpg")
plot(metaboliti[,11],xlab="disposizione␣campioni",...
...ylab="log-concentrazione",col=c(rep("red",52),...
...rep("blue",52)))
dev.off()
jpeg("concentrazione12NormRem.jpg")
plot(metaboliti[,12],xlab="disposizione␣campioni",...
...ylab="log-concentrazione", col=c(rep("red",52),...
...rep("blue",52)))
dev.off()
jpeg("concentrazione21NormRem.jpg")
```

```
plot(metaboliti[,21],xlab="disposizione␣campioni",...
...ylab="log-concentrazione",...
...col=c(rep("red",52),rep("blue",52)))
dev.off()
jpeg("concentrazione25NormRem.jpg")
plot(metaboliti[,25],xlab="disposizione␣campioni",...
...ylab="log-concentrazione",...
...col=c(rep("red",52),rep("blue",52)))
dev.off()

# Imputazione
# Scarico QRLIC
impute.QRILC = function(dataSet.mvs,tune.sigma = 1){
# get the dimension of the data
.....................................................
nFeatures = dim(dataSet.mvs)[1]
nSamples = dim(dataSet.mvs)[2]
# - initialize the matrix of complete abundances
.....................................
dataSet.imputed = dataSet.mvs
# - initialize QR.obj which contains estimates of the
distribution parameters
#   for all samples
QR.obj = list()
for (i in 1:nSamples){
curr.sample = dataSet.mvs[,i]
# - calculate the percentage of missing values
.....................................
pNAs = length(which(is.na(curr.sample)))/length(curr.sample)
# - estimate the mean and standard deviation of the original
#   distribution using quantile regression
upper.q = 0.99
q.normal = qnorm(seq((pNAs+0.001),(upper.q+0.001),...
...(upper.q-pNAs)/(upper.q*100)), mean = 0, sd = 1)
q.curr.sample = quantile(curr.sample,
probs = seq(0.001,(upper.q+0.001),0.01),
na.rm = T)
temp.QR = lm(q.curr.sample ~ q.normal)
QR.obj[[i]] = temp.QR
mean.CDD = temp.QR$coefficients[1]
sd.CDD = as.numeric(temp.QR$coefficients[2])
# generate data from multivariate distributions with MLE parameters
.................
data.to.imp = rtmvnorm(n=nFeatures,
mean = mean.CDD,
sigma = sd.CDD*tune.sigma,
```

```
upper = qnorm((pNAs+0.001),
mean = mean.CDD,
sd = sd.CDD),
algorithm=c("gibbs"))
curr.sample.imputed = curr.sample
curr.sample.imputed[which(is.na(curr.sample))] =
data.to.imp[which(is.na(curr.sample))]
dataSet.imputed[,i] = curr.sample.imputed
}
results = list(dataSet.imputed,QR.obj)
return(results)}

install.packages("tmvtnorm")
library(tmvtnorm)
metaboliti[mancanti]=NA
sum(is.na(metaboliti))

# Imputo gli NA sparsi con "QRILC"
metaboliti=impute.QRILC(t(metaboliti))
metaboliti=as.data.frame(t(metaboliti[[1]]))

# imputo il campione con piu NA con Random Forest
# prima gli reimposto i valori mancanti
reimpostaNulli=which(rownames(metaboliti)=="PCa508")
mancanti[-reimpostaNulli,]=F
sum(mancanti)
metaboliti[mancanti]=NA
sum(is.na(metaboliti))

### imputo con Random Forest
install.packages("missForest")
library(missForest)
metaboliti=missForest(metaboliti)[[1]]
dataset[,1:253]=metaboliti
save(dataset,metaboliti,file = "FinePreelaborazione.rda")
```

3. "Analisi univariata":

```
# carico le trasformazioni di classe II
source("Trasformazioni.R")
load("FinePreelaborazione.rda")

# Uso il paccheto  muma
install.packages("muma")
library(muma)
all.equal(rownames(dataset),rownames(metaboliti))
COMBAT=cbind(metaboliti,metaboliti[,1])
```

```
COMBAT[,1]=ifelse(dataset$Type=="PCa",1,2)
write.csv(x=COMBAT,file = "COMBAT.csv")
univariate(file="COMBAT.csv",normalize = F, multi.test=T,...
...plot.volcano=T)
univariate(file="COMBAT.csv",normalize = F, multi.test=F,...
...plot.volcano=T)

# Moderate t-test (limma) per diversi tipi di trasformazioni
# trasformazione logaritmica
load("FinePreelaborazione.rda")
library(limma)
design.PCa.BPH <- cbind(BPH=c(0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0,...
... 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1,...
... 0, 1, 0,1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1,...
... 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1,...
... 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1,...
... 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1),
PCa=c(1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,...
... 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0,...
1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0,...
.... 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0,...
... 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1,...
1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0))
fit_l2 <- lmFit(t(metaboliti), design.PCa.BPH)
cont.matrix <- makeContrasts (PCavsBPH=PCa-BPH, levels=design.PCa.BPH)
fit2_l2 <- contrasts.fit (fit_l2, cont.matrix)
efit_l2 <- eBayes(fit2_l2)
dim(efit_l2)
results_l2 <- topTable (efit_l2,number=254, adjust="BH")
write.table(results_l2, "PCa_vs_BPH_log2_combat_scale_Limma.txt",...
... sep="\t", col.names=T, row.names=T)
riassuntoModerate=read.table("PCa_vs_BPH_log2_combat_scale_Limma.txt")
head(riassuntoModerate,4)
## pacchetto per mettere testo nei grafi
install.packages("calibrate")
library(calibrate)
# Volcano plot per p-value non aggiustati
jpeg("classcomppvalue.jpg")
plot(x=riassuntoModerate$logFC,y=-log10(riassuntoModerate$P.Value))
textxy(riassuntoModerate$logFC,-log10(riassuntoModerate$P.Value),...
...rownames(riassuntoModerate))
abline(h=-log10(0.05),col="blue")
dev.off()
# Volcano plot per p-value aggiustati
jpeg("classcomppvalueAd.jpg")
plot(x=riassuntoModerate$logFC,y=-log10(riassuntoModerate$adj.P.Val))
```

```
dev.off()

# Trasformazione level scaling
load("FinePreelaborazione.rda")
design.PCa.BPH <- cbind(BPH=c(0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0,...
... 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1,...
... 0, 1, 0,1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1,...
... 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1,...
... 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1,...
... 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1),
PCa=c(1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,...
... 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0,...
1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0,...
.... 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0,...
... 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1,...
1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0))
metaboliti=levelscaling(metaboliti)
fit_l2 <- lmFit(t(metaboliti), design.PCa.BPH)
cont.matrix <- makeContrasts (PCavsBPH=PCa-BPH, levels=design.PCa.BPH)
fit2_l2 <- contrasts.fit (fit_l2, cont.matrix)
efit_l2 <- eBayes(fit2_l2)
dim(efit_l2)
results_l2 <- topTable (efit_l2,number=254, adjust="BH")
write.table(results_l2,"PCa_vs_BPH_log2_combat_scale_Limma_level.txt",
... sep="\t", col.names=T, row.names=T)
riassuntoModerate=read.table("PCa_vs_BPH_log2_combat_scale_␣...
...␣Limma_level.txt")
head(riassuntoModerate,4)
# Volcano plot per p-value non aggiustati
jpeg("classcomppvalue_level.jpg")
plot(x=riassuntoModerate$logFC,y=-log10(riassuntoModerate$P.Value))
textxy(riassuntoModerate$logFC,-log10(riassuntoModerate$P.Value),...
...rownames(riassuntoModerate))
abline(h=-log10(0.05),col="blue")
dev.off()
# Volcano plot per p-value aggiustati
jpeg("classcomppvalueAd_level.jpg")
plot(x=riassuntoModerate$logFC,y=-log10(riassuntoModerate$adj.P.Val))
dev.off()

# Trasformazione autoscaling
load("FinePreelaborazione.rda")
design.PCa.BPH <- cbind(BPH=c(0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0,...
... 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1,...
... 0, 1, 0,1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1,...
... 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1,...
```

```
... 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1,...
... 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1),
PCa=c(1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,...
... 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0,...
1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0,...
.... 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0,...
... 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1,...
1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0))
metaboliti=autoscaling(metaboliti)
fit_l2 <- lmFit(t(metaboliti), design.PCa.BPH)
cont.matrix <- makeContrasts (PCavsBPH=PCa-BPH, levels=design.PCa.BPH)
fit2_l2 <- contrasts.fit (fit_l2, cont.matrix)
efit_l2 <- eBayes(fit2_l2)
dim(efit_l2)
results_l2 <- topTable (efit_l2,number=254, adjust="BH")
write.table(results_l2,"PCa_vs_BPH_log2_combat_scale_Limma_auto.txt",.
... sep="\t", col.names=T, row.names=T)
riassuntoModerate=read.table("PCa_vs_BPH_log2_combat_scale␣...
...␣_Limma_auto.txt")
head(riassuntoModerate,4)
# Volcano plot per p-value non aggiustati
jpeg("classcomppvalue_auto.jpg")
plot(x=riassuntoModerate$logFC,y=-log10(riassuntoModerate$P.Value))
textxy(riassuntoModerate$logFC,-log10(riassuntoModerate$P.Value),...
...(riassuntoModerate))
abline(h=-log10(0.05),col="blue")
dev.off()
# Volcano plot per p-value  aggiustati
jpeg("classcomppvalueAd_auto.jpg")
plot(x=riassuntoModerate$logFC,y=-log10(riassuntoModerate$adj.P.Val))
dev.off()
```

4. "Analisi multivariata":

```
load("FinePreelaborazione.rda")

#Regressione logistica penalizzata
install.packages("glmnet")
library(glmnet)
all.equal(rownames(metaboliti),rownames(dataset))

fit.lasso=glmnet(x = as.matrix(cbind(metaboliti,dataset$Age,...
dataset$PSA)), y = dataset$Type,family = "binomial")
plot(fit.lasso,xvar="lambda")
# cross validation con 10 fold
cv.lasso=cv.glmnet(x = as.matrix(cbind(metaboliti,dataset$Age,...
dataset$PSA)), y = dataset$Type,family = "binomial")
```

```
jpeg("lassoNoTrasf.jpg")
plot(cv.lasso)
dev.off()

# Metaboliti associati al lambda ottimale sono
sa=coef(fit.lasso,s= cv.lasso$lambda.1se)
sa@i

# 5 metaboliti  che lasso "elimina" per ultimi
sa=coef(fit.lasso,s= 0.086746215)
sa@i
```

5. "Identificazione del modello e valutazione":

```
load("FinePreelaborazione.rda")
install.packages("caret")
library(caret)
install.packages("pROC")
library(pROC)
install.packages("e1071")
library(e1071)

# Funzione per definire uno score per i vari campioni:
# modello logistico piu k-fold CV
# INPUT: vettore di variabili (nome) da utilizzare
#        dataset
#       suddivisione del campione in k-fold
#                      (risultato funzione createFolds)
# OUTPUT: dataframe(V1=score, V2=Type)
creaScore=function(dataset,variabili,flds){
dataset$valori=ifelse(dataset$Type=="PCa",1,0)
score=NULL
miniset=dataset[,c(variabili,"valori")]
for(j in 1:length(flds)){
mod=glm(miniset$valori[-flds[[j]]]~., data=miniset[-flds[[j]],],...
...family="binomial")
vettore=as.vector(coef(mod))[-1]
temp=cbind(as.matrix(miniset[flds[[j]],-ncol(miniset)]) %*% ...
... vettore,as.character(dataset$Type[flds[[j]]]))
score=rbind(score,temp)   }
mod=glm(miniset$valori~., data=miniset,family="binomial")
print(summary(mod))
score=as.data.frame(score)
score[,1]=as.numeric(as.character(score[,1]))
return(score)
}
```

```
# INPUT:   elemento classe roc (pROC)
#            risultato funzione creaScore
#OUTPUT    miglior cut off secondo Youden
#          tabella contingenza associata
findBestYouden=function(rr,score){
max=0
threshold=0
for(cutoff in  rr$thresholds[2:102]){
confu=confusionMatrix(ifelse(score$V1>cutoff,"PCa","BPH"),...
... score$V2,positive="PCa")
temp=confu$byClass[1]+confu$byClass[2]
if(temp>max){
finale=confu
max=temp
threshold=cutoff
}
}
names(max)="Youden"
return(list(finale,threshold,max))
}

###INPUT:   elemento classe roc (pROC)
#            risultato funzione creaScore
#OUTPUT    miglior cut off secondo vicinanza a punto (0,1) ,
#                   tabella contingenza associata
findBestNear=function(rr,score){
min=10
threshold=0
for(cutoff in  rr$thresholds[2:102]){
confu=confusionMatrix(ifelse(score$V1>cutoff,"PCa","BPH"),...
... score$V2,positive="PCa")
temp=sqrt((confu$byClass[1]-1)^2+(confu$byClass[2]-1)^2)
if(temp<min){
finale=confu
min=temp
threshold=cutoff
}
}
names(min)="bestNear(0,1)"
return(list(finale,threshold,min))
}

set.seed(152)

# Modello definito da univariata
flds <- createFolds(seq(1:nrow(dataset)), k = 10, list = TRUE,...
```

```
... returnTrain = FALSE)
variabili=c("ID249_329_770","ID36_417_1257", "ID211_485_1041",...
... "ID220_441_1036",   "Age")
score=creaScore(dataset,variabili,flds)
rr1=roc(score$V2,score$V1)
plot.roc(rr1,print.auc = TRUE, col = "blue")
findBestYouden(rr1,score)
findBestNear(rr1,score)
boxplot(score$V1~score$V2)

# Modello definito da univariata senza ID249
flds <- createFolds(seq(1:nrow(dataset)), k = 10, list = TRUE,...
... returnTrain = FALSE)
variabili=c("ID36_417_1257", "ID211_485_1041", "ID220_441_1036", ...
... "Age")
score=creaScore(dataset,variabili,flds)
rr1=roc(score$V2,score$V1)
plot.roc(rr1,print.auc = TRUE, col = "blue")
findBestYouden(rr1,score)
findBestNear(rr1,score)
boxplot(score$V1~score$V2)

# Modello definito da univariata senza ID249 e Eta'
flds <- createFolds(seq(1:nrow(dataset)), k = 10, list = TRUE,...
... returnTrain = FALSE)
variabili=c("ID36_417_1257", "ID211_485_1041", "ID220_441_1036")
score=creaScore(dataset,variabili,flds)
rr1=roc(score$V2,score$V1)
plot.roc(rr1,print.auc = TRUE, col = "blue")
findBestYouden(rr1,score)
findBestNear(rr1,score)
boxplot(score$V1~score$V2)

# Modello definito da multivariata
flds <- createFolds(seq(1:nrow(dataset)), k = 10, list = TRUE,...
... returnTrain = FALSE)
variabili=c("ID36_417_1257","ID249_329_770","ID167_314_1220",...
..."ID220_441_1036","Age")
score=creaScore(dataset,variabili,flds)
rr1=roc(score$V2,score$V1)
plot.roc(rr1,print.auc = TRUE, col = "blue")
findBestYouden(rr1,score)
findBestNear(rr1,score)
boxplot(score$V1~score$V2)
```

Infine riportiamo il codice per le trasformazioni di classe II:

```
centering=function(frame){
medie=apply(X = frame,MARGIN = 2,FUN = mean,na.rm=T)
return(t(t(frame)-medie))
}

autoscaling=function(frame){
frame=centering(frame)
devstand=sqrt(apply(X = frame,MARGIN = 2,FUN = var,na.rm=T))
return(t(t(frame)/devstand))
}

levelscaling=function(frame){
medie=apply(X = frame,MARGIN = 2,FUN = mean,na.rm=T)
frame=centering(frame)
return(t(t(frame)/medie))
}
```

Bibliografia

[1] Zhou B., Xiao J. F., Tuli L., et al, *LC-MS-based metabolomics.* Mol. Biosyst., 2012.

[2] Wishart D. S., Jewison T., Guo A. C., *HMDB 3.0 The Human Metabolome Database in 2013.* Nucleic Acids Res ,2013.

[3] Xia J., Broadhurst D. I., Wilson M., *Translational biomarker discovery in clinical metabolomics: an introductory tutorial.* Metabolomics, 2013.

[4] Hastie T., Tibshirant R., Friedman J., *The Elements of Statistical Learning.* New York: Springer, 2009.

[5] Ross S. M., *Probabilitá e statistica per l'ingegneria e le scienze.* APOGEO, 2008.

[6] Rogantin M. P. *Modelli lineari generali e generalizzati.* 2017.

[7] van den Berg R. A., Hoefsloot H. C. J. , Westerhuis J. A., et al, *Centering, scaling, and transformations: improving the biological information content of metabolomics data.* BMC Genomics, 2006.

[8] Wei R., Wang J., Su M., et al, *Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data.* Scientific Reports, 2018.

[9] Stekhoven D.J., BÃijhlmann P., *MissForest - nonparametric missing value imputation for mixed-type data.* Bioinformatics, 2012.

[10] Koenker R., Bassett Jr. G., *Regression Quantiles.* The Econometric Society, 1978.

[11] Johnson W. E., et al, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007.

[12] Ritchie M. E., Phipson B., Wu D-, Hu Y., et al, *limma powers differential expression analyses for RNA-sequencing and microarray studies* Nucleic Acids Research, 2015.

[13] Leek J. T., Johnson W. E. , Parker H. S., et al *sva: Surrogate Variable Analysis.* 2016.

[14] Smyth G. K., *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.* Stat Appl Genet Mol Biol., 2004.

[15] Friedman J., Hastie T., Tibshirani R. *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010.

[16] Gaude E., Chignola F., Spiliotopoulos D., et al *muma: Metabolomics Univariate and Multivariate Analysis.* 2012.

[17] R Core Team, *R: A Language and Environment for Statistical Computing.* 2016.