Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

NUOVI METODI DI ANALISI DI DATI EPIGENETICI PER LA PREVISIONE DELL'ETÁ DEL PAZIENTE

Tesi di Laurea in Fisica Matematica Applicata

Relatore:
Chiar.mo Prof.
MARCO LENCI

Presentata da:
RICCARDO PASCUZZO

Sessione III
Anno Accademico 2012/2013

It's tough to make predictions
especially about the future.
Yogi Berra

Introduzione

Questa tesi si inserisce nel progetto di ricerca FARB denominato *'Studio dell'invecchiamento e di malattie associate all'età tramite l'analisi di dati genomici ed epigenomici con metodologie innovative multiscala'*, condotto da un gruppo di docenti e ricercatori del Dipartimento di Fisica, di Matematica e di Biologia dell'Università di Bologna. Scopo del progetto è combinare competenze di ambito fisico, matematico e biologico per lo studio dei meccanismi di invecchiamento e delle patologie associate all'età, che con il progressivo aumento della speranza di vita nell'uomo hanno un sempre maggiore impatto nella società, in termini sia di qualità della vita del singolo sia di politica economica nazionale del Sistema Sanitario. Comprendere meglio la fisiopatologia dell'invecchiamento, identificando singoli elementi e meccanismi biologici maggiormente coinvolti, può portare a trovare nuove strategie e potenziali target terapeutici, che possono essere oggetto di successivi studi specifici.

In questa tesi analizzeremo dati di metilazione di diversi gruppi di pazienti, mettendoli in relazione con le loro età, intesa in senso anagrafico e biologico. Adatteremo metodi che sono già stati usati in altri studi, in particolare di tipo statistico, cercando di migliorarli e proveremo ad applicare a questi dati anche dei metodi nuovi, non solo di tipo statistico. La nostra analisi vuole essere innovativa soprattutto perché, oltre a guardare i dati in maniera locale attraverso lo studio della metilazione di particolari sequenze genetiche più o meno note per essere collegate all'invecchiamento, andremo a considerare i dati anche in maniera globale, analizzando le proprietà della

distribuzione di tutti i valori di metilazione di un paziente.

Nel primo capitolo di questa tesi forniremo un'introduzione generale degli argomenti di genetica che sono stati alla base delle nostre sperimentazioni. Nel secondo capitolo illustreremo invece gli strumenti matematici e statistici dei test d'ipotesi e della regressione multipla con cui abbiamo elaborato i dati a nostra disposizione, i quali verranno descritti nel terzo capitolo. In esso commenteremo principalmente il lavoro di genetica che è stato il punto di partenza della nostra ricerca e da cui abbiamo costruito i nostri modelli di previsione dell'età. Nel quarto capitolo presenteremo dei nuovi metodi di analisi di dati epigenetici (in particolare una nuova funzione di previsione dell'età e l'utilizzo della trasformata di Fourier) e li applicheremo ai dati a nostra disposizione, confrontando i risultati con i precedenti. Infine nel quinto capitolo vedremo come tutti questi metodi (di regressione e di analisi di Fourier) rispondono nel caso di soggetti affetti da sindrome di Down, che è nota avere dirette conseguenze sull'invecchiamento del paziente.

Indice

Introduzione **i**

1 Genetica ed epigenetica **1**

1.1 Elementi essenziali di genetica 1

1.2 Epigenetica e invecchiamento 3

2 Metodi matematico-statistici **7**

2.1 Test di significatività 7

2.2 Esempi di test 12

2.2.1 Test Z 12

2.2.2 Test t 15

2.2.3 Test F 17

2.2.4 Test di Wilcoxon-Mann-Whitney 19

2.2.5 Test di Kolmogorov-Smirnov 22

2.3 Analisi della regressione 27

2.3.1 Regressione lineare multipla 31

2.3.2 Metodo dei minimi quadrati 32

2.3.3 Regressione *ridge* 34

2.3.4 Regressione *lasso* 35

2.3.5 Regressione *elastic-net* 38

3 Analisi dati **39**

3.1 Il *dataset* di riferimento 39

3.2 Costruzione della nostra *signature* 43

3.2.1 Preselezione dei marcatori 43
3.2.2 Applicazione dei metodi statistici 45

4 Nuovi metodi non statistici di previsione dell'età 55
4.1 Nuova funzione di previsione 55
4.2 Analisi spettrale di Fourier 57

5 Caso di studio: pazienti affetti da sindrome di Down 67
5.1 Età anagrafica o età biologica? 67
5.2 Un risultato dall'analisi di Fourier 76

Bibliografia 83

Capitolo 1

Genetica ed epigenetica

In questo primo capitolo illustriamo i principali concetti di genetica che saranno poi ripresi nei capitoli finali, senza però entrare nei dettagli, che esulano dagli scopi di questa tesi. Per approfondimenti si rimanda ai testi [1, 2].

1.1 Elementi essenziali di genetica

Il 1865 è ampiamente riconosciuto come l'anno della nascita della genetica, ad opera del frate agostiniano Gregor Johann Mendel, con i suoi famosi studi sull'ereditarietà dei caratteri. Con la statistica e il calcolo delle probabilità, egli riuscì a elaborare una legge scientifica che spiega come alcune caratteristiche fisiche si trasmettono da una generazione all'altra, pur senza avere ancora la nozione del substrato materiale alla base di questa trasmissione, il DNA. Infatti l'acido deossiribonucleico fu osservato per la prima volta solo qualche anno dopo, nel 1869, da Friedrich Miescher, uno studente di Chimica a Zurigo. Mentre lavorava sul pus (un liquido infiammatorio ricchissimo in globuli bianchi degenerati), per caso aggiunse dell'alcool etilico a questa sostanza e osservò la pronta precipitazione di una massa biancastra, acida e che si trovava nel nucleo delle cellule esaminate: lo chiamò quindi 'acido nucleico', anche se non sapeva di cosa si trattasse in realtà, dato che

gli strumenti dell'epoca non consentivano un'analisi più approfondita. Infatti fu solo con il miglioramento dei microscopi degli anni successivi che si poterono osservare cellule in divisione, cosa che permise nel 1879 a Walter Fleming di scoprire l'esistenza dei 'cromosomi': egli notò la comparsa di alcuni bastoncini microscopici evidenziati dalla colorazione con coloranti basici, visibili nelle cellule durante la divisione cellulare. La parola 'cromosoma' fu attribuita dunque come termine puramente descrittivo, senza una conoscenza di una precisa funzione biologica.

Queste tre osservazioni concentrate temporalmente in una decina d'anni alla fine dell'Ottocento (Leggi di Mendel, scoperta del DNA, descrizione dei cromosomi) sono alla base di tutta la Genetica Moderna, perché è stato in seguito stabilito un nesso fra questi tre elementi. Il quadro interpretativo che abbiamo oggi prevede che il genoma, l'insieme di tutto il patrimonio genetico di un individuo, è organizzato in unità discrete (i cromosomi), che contengono l'informazione genetica sotto forma di una lunga molecola di DNA. I geni sono le regioni del DNA cromosomiale che sono coinvolti nella produzione di proteine e ogni gene risiede in una particolare posizione del cromosoma, chiamata locus genetico.

Guardando alla struttura chimica del DNA, proposta dal biologo James Watson e dal fisico Francis Crick nel 1953, essa consiste in una macromolecola lineare a forma di doppia elica, composta da due catene di nucleotidi, ognuno dei quali è formato di uno scheletro laterale (una gruppo fosfato e uno zucchero, il deossiribosio), che permette il legame con i nucleotidi adiacenti, e da una di quattro differenti basi azotate - adenina (A), guanina (G), citosina (C) e timina (T) - che instaura legami idrogeno con la corrispondente base azotata presente sul filamento opposto. Mentre l'ossatura della catena nucleotidica è una ripetizione della sequenza zucchero-fosfato-zucchero-fosfato, le 'braccia' dei nucleotidi hanno invece una variabilità sequenziale data dalla possibile presenza di una qualunque delle 4 basi azotate. L'unico vincolo è però l'appaiamento delle basi tra i due filamenti: adenina con timina e guanina con citosina. Quando si descrive una sequenza di DNA, per convenzione

si elencano semplicemente le iniziali delle basi azotate che si succedono lungo un filamento, presupponendo con questo che esiste comunque un'ossatura fissa di zuccheri e fosfati, e che la sequenza del filamento non dichiarato si ricavi per complementarità, ossia secondo gli appaiamenti canonici appena descritti.

Infine è da notare che una delle sorprese più grosse che si sono avute nello studio del DNA è stata la scoperta che le sequenze dei geni codificanti per le proteine di solito non sono continue, ma sono invece interrotte da sequenze non codificanti, ossia da sequenze nucleotidiche che non vengono tradotte in proteine. Queste interruzioni non codificanti all'interno di un gene sono dette introni, mentre le sequenze codificanti sono chiamate esoni. L'esoma rappresenta l'insieme di tutte le sequenze esoniche del DNA e, nell'uomo, esso costituisce solo l'1% circa del genoma.

1.2 Epigenetica e invecchiamento

Il 'dogma centrale' della biologia molecolare sostiene che le informazioni ereditarie sono trasmesse attraverso meccanismi genetici tali che l'informazione genetica fluisce dal DNA fino alle proteine, ma non può andare in direzione inversa. In realtà, lungo le generazioni, una cellula scambia con le cellule figlie anche informazioni non contenute nella sequenza di basi del DNA. L'epigenetica studia la trasmissione di caratteri ereditari non attribuibili direttamente alla sequenza di DNA, ovvero le modificazioni del fenotipo (caratteristiche osservabili) che non riguardano alterazioni del genotipo (l'insieme di tutti i geni del DNA), come ad esempio il silenziamento o l'attivazione di un gene. Questa attività risulta di fondamentale importanza dato che ogni tipo di cellula, pur avendo il patrimonio genetico identico a quello di qualunque altra dello stesso organismo, produce soltanto le sue proteine caratteristiche e non quelle di altri tipi di cellule.

Un evento molecolare noto che ostacola l'espressione genica, ossia la 'lettura' delle basi, è la metilazione del DNA, processo in cui un gruppo metilico

(CH_3) si lega ad una base azotata. Nei vertebrati la metilazione avviene tipicamente sulle citosine dei siti CpG[1], parti del DNA dove una citosina appare accanto ad una guanina nella sequenza lineare di basi. Le regioni del genoma che hanno un'alta concentrazione di siti CpG, conosciute come *CpG island*, sono frequentemente localizzate nel tratto iniziale (chiamato promotore) di molti geni. Se tali isole CpG sono metilate allora i geni corrispondenti sono silenziati, altrimenti sono espressi.

Dunque i diversi pattern di metilazione regolano l'accensione o lo spegnimento genico e, poiché rimangono inalterati nel momento della replicazione del DNA, permettono il passaggio alle generazioni cellulari successive anche di un'eredità epigenetica, non direttamente coinvolta nella sequenza nucleotidica. Mentre però le mutazioni nel DNA sono cambiamenti permanenti, i livelli di metilazione sono variabili nel tempo e, da studi recenti, sembrano correlati con l'invecchiamento dell'organismo. Una relazione tra epigenetica ed *aging* fu osservata per la prima volta circa 40 anni fa in uno studio sui salmoni rosa, che mostrava una globale diminuzione della metilazione del DNA genomico all'aumentare dell'età [3]. Questa diminuzione fu osservata in seguito anche in altre specie, incluso l'uomo [4, 5, 6].

L''aging epigenetics' è una disciplina emergente che promette interessanti scoperte nel prossimo futuro, come la identificazione di un metiloma del DNA che possa portare a definire il concetto di cellula 'giovane' o 'vecchia'. Negli ultimi due decenni, un crescente numero di ricerche ha riportato associazioni tra età e lo stato dell'epigenoma, l'insieme delle modificazioni del DNA diverse dai cambiamenti nella sequenza nucleotidica [7]. In particolare, cambiamenti nella metilazione sono stati associati a malattie legate all'età come malattie metaboliche e cancro [8]. Studi hanno anche osservato il fenomeno denominato 'epigenetic drift', in cui la differenza della metilazione del DNA in gemelli identici aumenta in funzione dell'età [9, 10]. Così, l'idea dell'epigenoma come una impronta fissa sta iniziando ad essere sostituita ad un

[1] La notazione 'CpG' specifica che è presente un gruppo fosfato (p) tra una citosina (C) e una guanina (G) e viene usata per distinguere questa sequenza lineare dall'appaiamento di basi complementari CG (citosina e guanina) su due diversi filamenti.

modello di epigenoma come un panorama dinamico che riflette una varietà di cambiamenti cronologici. L'attuale sfida è determinare se queste modificazioni possono essere descritte sistematicamente e modellate per determinare differenti velocità di invecchiamento umano e di legare queste alterazioni a variabili cliniche o ambientali. Nelle scienze forensi, un tale modello permetterebbe di stimare l'età di una persona, basandosi su di un unico campione biologico. Inoltre in campo medico tali analisi sarebbero utili per valutare il rischio di malattie legate all'età attraverso screening di routine e interventi medici basati sull'età biologica invece che su quella cronologica.

Capitolo 2

Metodi matematico-statistici

Uno dei primi obiettivi di un'analisi statistica è quello di effettuare una stima o una previsione riguardo ad una popolazione, basandosi sull'informazione contenuta in un campione casuale. Tra i metodi matematici che permettono queste operazioni, nel primo e secondo paragrafo di questo capitolo analizzeremo alcune tecniche di inferenza statistica, in particolare i test di verifica d'ipotesi: illustreremo dapprima le loro principali caratteristiche e poi presenteremo alcuni esempi di test statistici che sono stati utilizzati nel nostro lavoro sperimentale descritto nei capitoli seguenti. Per approfondimenti su questa parte si vedano [11, 12, 13, 14]. Infine nel terzo e ultimo paragrafo di questo capitolo discuteremo di analisi della regressione, che comprende tecniche per stimare ed analizzare una relazione tra due o più variabili. Per ulteriori dettagli si vedano [15, 16, 17].

2.1 Test di significatività

Un'ipotesi solitamente emerge da riflessioni su un comportamento osservato, un fenomeno naturale o una teoria provata. Se l'ipotesi è espressa in termini di parametri di una popolazione, come media e varianza, allora essa è detta **ipotesi statistica**. I dati da un campione (che potrebbe essere un esperimento) sono usati per testare la validità dell'ipotesi. Una procedu-

ra che permetta di accettare o rifiutare l'ipotesi statistica usando i dati dal campione è chiamata **test di verifica d'ipotesi**, o di significatività.

La verifica delle ipotesi statistiche inizia con la definizione di un insieme di due affermazioni sui parametri in questione. Queste sono di solito in forma di semplici relazioni matematiche che riguardano i parametri. Le due asserzioni sono esclusive ed esaustive: o è vera la prima oppure lo è la seconda, ma non possono esserlo entrambe. La prima è chiamata ipotesi nulla e l'altra ipotesi alternativa. Più formalmente:

Definizione 2.1 (Ipotesi nulla)**.** L'*ipotesi nulla* è un'affermazione riguardo ai valori di uno o più parametri e la si indica usualmente con H_0. Essa rappresenta lo status quo ed è quella che si tenta screditare in favore dell'ipotesi alternativa. Per farlo, i risultati del campione devono indicare in maniera convincente che H_0 sia falsa.

Definizione 2.2 (Ipotesi alternativa)**.** L'ipotesi alternativa H_1 è l'asserzione che contraddice l'ipotesi nulla. Essa è accettata se H_0 viene rifiutata.

Osservazione 2.1. Bisogna sottolineare che con la verifica d'ipotesi non si arriva alla dimostrazione di una delle due ipotesi, ma si ha solo un'indicazione del fatto che l'ipotesi (nulla o alternativa) sia o meno avvalorata dai dati disponibili: in caso non si possa escludere H_0, ciò non vuol dire che essa sia vera, ma solo che il campione non fornisce prove sufficienti a garantirne il rifiuto e dunque a sostenere l'ipotesi alternativa. Il giudizio è dunque sospeso e pertanto saranno necessarie ulteriori osservazioni sul fenomeno studiato.

Dopo aver stabilito le ipotesi, si specifica quali risultati campionari possano portare a non accettare H_0. Intuitivamente, essi dovrebbero essere sufficientemente in contraddizione con l'ipotesi nulla per giustificarne il rifiuto. In altre parole, se il campione statistico ha un range di valori che risultano improbabili se l'ipotesi nulla è vera, allora rifiutiamo H_0 e decidiamo che l'ipotesi alternativa è probabilmente vera. Cosa vogliono dire 'sufficientemente in contraddizione' e 'risultano improbabili', lo si stabilisce attraverso lo studio della distribuzione campionaria di una statistica, detta **statistica test**:

essa è una funzione che fa corrispondere ad ogni campione casuale un valore numerico che può essere classificato come coerente o meno con l'ipotesi specificata dalla H_0. Le statistiche test si dividono in parametriche e non parametriche. Le prime sono quelle in cui è (o si assume che sia) noto il tipo di distribuzione di probabilità della popolazione (distribuzione normale, di Student, ...), mentre nelle seconde non si fa alcun tipo di supposizione sulla distribuzione che ha generato i dati. Vedremo esempi di statistiche test nei paragrafi successivi di questo capitolo.

Utilizzando le proprietà della distribuzione di campionamento della statistica soggetta a test, si può identificare un intervallo di valori di quella statistica che verosimilmente non si presentano se l'ipotesi nulla è vera.

Definizione 2.3. La *regione di rifiuto*, o regione critica, è l'insieme dei valori di un campione statistico che portano al rifiuto dell'ipotesi nulla. I *valori critici* sono i valori della statistica test che sono agli estremi della regione di rifiuto.

A seconda del tipo di regione critica, i test di ipotesi possono essere distinti in **test a una coda** (o unilaterali) e **test a due code** (o bilaterali): se la regione di rifiuto è costituita da un intervallo siamo nel caso di test a una coda, se invece la regione critica è formata da due intervalli disgiunti, quindi abbiamo due code della distribuzione, si parla di test bilaterale. Un semplice modo per stabilire di che tipo è un test, senza dover conoscere la regione di rifiuto, è il seguente: per un test a due code in H_1 compare il segno $\neq$ (o comunque, in generale, la negazione dell'ipotesi nulla), nell'altro caso invece compare uno dei segni $<$ (test ad una coda sinistra) o $>$ (test ad una coda destra).

In un test statistico, è impossibile stabilire la verità di un'ipotesi con il 100% di sicurezza, ma si corre sempre il rischio di giungere ad una conclusione sbagliata. Ci sono due possibili tipi di errore: si può rifiutare H_0 quando invece è vera oppure non rifiutarla quando invece è falsa. Poiché gli errori compaiono come risultato di scelte sbagliate e le decisioni stesse sono basate

su campioni casuali, ne segue che gli errori hanno delle probabilità associate a loro. Diamo pertanto le seguenti definizioni:

Definizione 2.4 (Errori di tipo I e II)**.** Un **errore di tipo I** si commette se H_0 è rifiutata quando invece è vera. La probabilità di questo tipo di errore si denota con α ed è chiamata **livello si significatività**. Formalmente

$$P(\text{rifiutare } H_0 \,|\, H_0 \text{ è vera}) = \alpha$$

Un **errore di tipo II** si commette se H_0 non è rifiutata quando invece è falsa. La probabilità di questo tipo di errore si denota con β. Formalmente

$$P(\text{non rifiutare } H_0 \,|\, H_0 \text{ è falsa}) = \beta$$

Nella tabella 2.1 vediamo uno schema dei possibili risultati di una decisione sulla base di un test statistico.

	H_0 **vera**	H_0 **falsa**
Non rifiuto H_0	Decisione corretta	Errore di tipo II (β)
Rifiuto H_0	Errore di tipo I (α)	Decisione corretta

Tabella 2.1: Risultati di un test statistico

Sarebbe auspicabile che un test avesse $\alpha = \beta = 0$, o almeno che fosse possibile minimizzare entrambi i tipi di errore. Sfortunatamente, per un campione di fissata grandezza, accade che, al diminuire di α, il valore di β tende a crescere e viceversa. Comunque è possibile determinare quale dei due errori sia il più grave: l'errore di tipo II è solo un errore nel senso che si è persa la possibilità di rifiutare correttamente l'ipotesi nulla, mentre non è errore nel senso che siamo arrivati ad una conclusione scorretta, perché nessuna decisione è presa quando l'ipotesi nulla non è rifiutata. Nel caso invece di errore di I tipo, si è presa la decisione di rifiutare H_0 quando invece essa è vera, per questo gli errori di I tipo sono considerati in generale più gravi di quelli del II tipo. L'approccio usuale ai test d'ipotesi è trovare una statistica test che limiti α ad un livello accettabile mentre tiene β il più

basso possibile. Per questo scopo, storicamente e tradizionalmente, si sceglie un valore di α pari a 0.10, 0.05 (l'opzione più frequente) o 0.01, anche se la scelta è del tutto arbitraria.

Capita che però, per lo stesso insieme di dati, il test risponda in un modo con la scelta di un α, in un altro con un α diverso. Molti statistici preferiscono usare un metodo per riportare i risultati di un test di significatività senza dover scegliere un preciso α, ma invece lasciare quel compito a coloro che dovranno decidere cosa fare sulla base delle conclusioni del test. A questo metodo di riportare i risultati ci si riferisce come il metodo di riportare il ***p*-value**, o valore p.

Definizione 2.5 (p-value)**.** Il *valore p* è la probabilità di ottenere un risultato altrettanto estremo o più estremo di quello osservato se la diversità è interamente dovuta alla sola variabilità campionaria, assumendo quindi che l'ipotesi iniziale nulla sia vera, ovvero il valore p indica il minimo livello di significatività per il quale l'ipotesi nulla viene rifiutata.

Nella pratica il p-value si calcola in questo modo, assumendo che il valore della statistica test (ST) del campione casuale considerato sia risultato uguale a V:

$$p\text{-value} = \begin{cases} P(ST < \mathrm{V} \,|\, H_0) & \text{in un test ad una coda sinistra,} \\ P(ST > \mathrm{V} \,|\, H_0) & \text{in un test ad una coda destra,} \\ P(ST > |\mathrm{V}| \,|\, H_0) & \text{in un test a due code.} \end{cases}$$

Nelle situazioni dove l'analista statistico non è la stessa persona che prende le decisioni, l'analista produce il p-value e colui che deve prendere le decisioni determina il livello di significatività basato sul costo di commettere un errore di tipo I. Per queste ragioni, molte riviste scientifiche ora richiedono che i risultati di tali test siano pubblicati in questo modo.

Riassumendo, elenchiamo i passi per costruire un test di verifica di ipotesi:

1. Stabilire l'ipotesi nulla H_0 (che si pensa non sia vera) e l'ipotesi alternativa H_1 (che si ritiene vera).

2. Decidere un livello di significatività α.

3. Scegliere un'appropriata statistica test (ST) e calcolarne il valore corrispondente al campione osservato (V).

4. Determinare la regione di rifiuto (RR) usando la distribuzione campionaria ed α

5. Conclusione: se V cade in RR, rifiutare H_0. Altrimenti, concludere che non ci sono sufficienti prove per rifiutare H_0.

Se si vuole utilizzare l'approccio del p-value, al posto degli ultimi due punti si deve calcolare il p-value e trarre infine le conclusioni confrontando con il livello di significatività α desiderato: se il p-value è più piccolo di α, allora si rifiuta l'ipotesi nulla, in caso contrario invece non è possibile farlo.

2.2 Esempi di test

Vediamo adesso alcuni esempi di test statistici, tra i più utilizzati nelle applicazioni: come test parametrici presentiamo il test Z, il test t e il test F, per i non parametrici invece il test di Wilcoxon-Mann-Whitney e il test di Kolmogorov-Smirnov.

2.2.1 Test Z

Sia $X_1, \ldots, X_n$ un campione preso da una distribuzione normale con media incognita μ e varianza nota σ^2. Supponiamo di voler testare l'ipotesi nulla che la media μ sia uguale ad uno specifico valore μ_0 contro l'alternativa che non lo sia, dunque

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0 .$$

Notiamo che, poiché compare $\neq$ in H_1, stiamo effettuando un test statistico bilaterale, che prende in questo caso il nome di 'test Z a due code'. Poiché lo stimatore corretto della media μ della popolazione è la media campionaria[1]

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

sembra ragionevole rifiutare l'ipotesi nulla quando $\overline{X}$ è lontano da μ_0. Così, la regione critica del test risulta essere della forma

$$C = \{X_1, \ldots, X_n : \quad |\overline{X} - \mu_0| \geq c\}$$

per un adeguato valore c. Fissato un livello di significatività α, c deve essere tale che

$$P\{|\overline{X} - \mu_0| \geq c\} = \alpha\,, \quad \text{quando } \mu = \mu_0. \tag{2.1}$$

Supponendo che l'ipotesi nulla sia vera, $\overline{X}$ ha distribuzione normale con media μ_0 e deviazione standard $\sigma/\sqrt{n}$, così la variabile standardizzata Z, definita da

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0)$$

avrà distribuzione normale standard. Ora, dato che la disuguaglianza

$$|\overline{X} - \mu_0| \geq c$$

è equivalente a

$$\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| \geq \frac{\sqrt{n}}{\sigma}c\,,$$

segue che (2.1) è equivalente a

$$P\{|Z| \geq \frac{\sqrt{n}}{\sigma}c\} = \alpha$$

[1]Per la teoria degli estimatori, si veda [11].

o anche, per la simmetria della normale standard,

$$P\{Z \geq \frac{\sqrt{n}}{\sigma} c\} = \frac{\alpha}{2}\,.$$

Definendo $z_{\alpha/2}$ in maniera tale che

$$P\{Z \geq z_{\alpha/2}\} = \frac{\alpha}{2}\,,$$

ne viene che

$$c = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\,.$$

Dunque, l'ipotesi nulla che la media μ della popolazione sia uguale al valore μ_0 contro l'alternativa che non lo sia è da rifiutare a livello di significatività α se

$$\frac{\sqrt{n}}{\sigma} |\overline{X} - \mu_0| \geq z_{\alpha/2}\,.$$

Il nome 'test Z' è dovuto alla supposizione che la densità della statistica test $\sqrt{n}\,\left|\overline{X} - \mu_0\right| / \sigma$ sia una normale standard quando H_0 è vera. In maniera analoga si mostra che se

1. $H_0 : \mu \leq \mu_0$ e $H_1 : \mu > \mu_0$

2. oppure $H_0 : \mu \geq \mu_0$ e $H_1 : \mu < \mu_0$

la statistica test rimane la stessa del caso bilaterale, mentre la conclusione cambia così: detto z_α il valore di Z per cui l'area a destra di z_α al di sotto della gaussiana standard è uguale ad α, H_0 è da rifiutare nel primo caso se

$$\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) \geq z_\alpha$$

e nel secondo caso se

$$\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) \leq -z_\alpha\,.$$

Questo test può essere condotto in modo diverso usando il p-value: dopo aver calcolato il valore v della statistica test, allora abbiamo che

$$p\text{-value} = \begin{cases} 2P(Z \geq |v|) & \text{nel test a due code} \\ P(Z \geq v) & \text{nel primo test ad una coda} \\ P(Z \leq v) & \text{nel secondo test ad una coda.} \end{cases}$$

L'ipotesi nulla è da rifiutare ad un qualunque livello di significatività maggiore o uguale al p-value trovato.

2.2.2 Test t

Sia $X_1, \ldots, X_n$ un campione preso da una distribuzione normale con media incognita μ e varianza pure incognita σ^2. Supponiamo di voler testare l'ipotesi nulla che la media μ sia uguale ad uno specifico valore μ_0 contro l'alternativa che non lo sia, dunque

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0 .$$

Rispetto al test Z, ora dobbiamo stimare, oltre che la media, anche la varianza: lo stimatore corretto di σ è la deviazione standard campionaria[2]

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2} .$$

In questo caso dunque la statistica test che dobbiamo utilizzare diventa

$$T = \frac{\sqrt{n}}{S} \left(\overline{X} - \mu_0 \right)$$

[2]Si veda [11].

ed è noto[3] che, nel caso sia vera l'ipotesi nulla $\mu = \mu_0$, la variabile aleatoria T appena introdotta ha distribuzione t di Student

$$t_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

con $\nu = n-1$ gradi di libertà. Fissato un livello di significatività α e definito $t_{n-1,\alpha/2}$ tale che

$$P(T \geq t_{n-1,\alpha/2}) = \frac{\alpha}{2}\,,$$

ne segue che se

$$|T| \geq t_{n-1,\alpha/2}$$

allora l'ipotesi nulla è da rifiutare, altrimenti non è possibile farlo. Tale test è chiamato test t a due code. Inoltre è possibile, come per il test Z, costruire un test t ad una coda: se abbiamo i casi

1. $H_0 : \mu \leq \mu_0$ e $H_1 : \mu > \mu_0$
2. oppure $H_0 : \mu \geq \mu_0$ e $H_1 : \mu < \mu_0$

allora H_0 è da rifiutare a livello α nel primo caso se

$$\frac{\sqrt{n}}{S}\left(\overline{X} - \mu_0\right) \geq t_{n-1,\alpha}$$

e nel secondo caso se

$$\frac{\sqrt{n}}{S}\left(\overline{X} - \mu_0\right) \leq -t_{n-1,\alpha}\,.$$

Infine il calcolo del p-value è identico a quello del test Z, effettuando l'opportuno cambio della statistica test.

Test t per due campioni. Vediamo ora come con il test t sia possibile effettuare un'inferenza sulla differenza fra le medie di due popolazioni aventi varianze incognite ma uguali[4]. Supponiamo di avere due campioni

[3] Si veda [11].

[4] Vedremo nel paragrafo successivo come il fatto che le varianze di due popolazioni siano uguali può, a sua volta, essere oggetto di un test statistico, il test F.

indipendenti $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, estratti da due popolazioni aventi distribuzioni normali con medie rispettivamente μ_X e μ_Y e varianze incognite ma uguali. Vogliamo testare l'ipotesi nulla

$$H_0 : \mu_X - \mu_Y = \delta$$

dove δ è una costante specificata. In tal caso, si ricava la stima congiunta S^2 della varianza, attraverso una media pesata delle varianze campionarie S_X^2 e S_Y^2 delle singole popolazioni:

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{(n-1)+(m-1)} .$$

Indicati con $\overline{X}$ e $\overline{Y}$ le due medie campionarie, si dimostra come prima che la statistica test

$$T = \frac{\left(\overline{X} - \overline{Y}\right) - \delta}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

ha distribuzione t di Student con $\nu = n + m - 2$ gradi di libertà. Le conclusioni vengono prese allo stesso modo del test t, confrontando il valore della statistica test con il quantile $t_{n+m-2,\,\alpha/2}$.

2.2.3 Test F

Il test F per il confronto di due varianze è un test di ipotesi basato sulla distribuzione F di Fisher-Snedecor e volto a verificare l'ipotesi che due popolazioni normali abbiano la stessa varianza.

Siano $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ due campioni casuali indipendenti estratti da due popolazioni, aventi distribuzioni normali $N(\mu_X, \sigma_X^2)$ e $N(\mu_Y, \sigma_Y^2)$ rispettivamente, e siano S_X^2 e S_Y^2 le corrispondenti varianze campionarie. Testiamo l'ipotesi nulla

$$H_0 : \sigma_X^2 = \sigma_Y^2$$

contro l'ipotesi alternativa

$$H_1 : \sigma_X^2 \neq \sigma_Y^2 \,.$$

Considerando la variabile

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

si dimostra[5] che essa ha distribuzione di Fisher

$$F_{\nu_1,\nu_2}(x) = \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \frac{x^{\frac{\nu_1}{2}-1}}{(\nu_1 x + \nu_2)^{\frac{\nu_1+\nu_2}{2}}}$$

con $\nu_1 = n-1$ e $\nu_2 = m-1$. Assumendo H_0 vera, abbiamo che la statistica test

$$F = \frac{S_X^2}{S_Y^2}$$

ha distribuzione F_{ν_1,ν_2} e, assegnato un livello di significatività α, l'ipotesi nulla è da rifiutare se

$$F > f_{\nu_1,\,\nu_2,\,\alpha/2} \quad \text{oppure} \quad F < f_{\nu_1,\,\nu_2,\,1-\alpha/2}$$

per la non simmetria della distribuzione di Fisher, dove $f_{\nu_1,\,\nu_2,\,\alpha/2}$ è definito in modo tale che

$$P(F \geq f_{\nu_1,\,\nu_2,\,\alpha/2}) = \frac{\alpha}{2} \,.$$

Da notare che, per questioni di praticità, si ricorre all'identità seguente

$$f_{\nu_1,\,\nu_2,\,1-\alpha/2} = \frac{1}{f_{\nu_2,\,\nu_1,\,\alpha/2}}$$

quando si devono calcolare i quantili di ordine $1-\alpha/2$.

[5] Si veda [11].

2.2.4 Test di Wilcoxon-Mann-Whitney

Il nome *Wilcoxon-Mann-Whitney test* deriva in realtà dall'unione dei nomi di due test distinti ma equivalenti, il *Wilcoxon rank sum test* e il *Mann-Whitney test*, o test U. Descriviamo ora entrambi i test e infine proveremo la loro l'equivalenza.

Test di Wilcoxon. Si tratta di un test non parametrico usato al posto del test t per confrontare i campioni di due popolazioni indipendenti quando esse non hanno distribuzione normale. Siano $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ due campioni casuali indipendenti estratti dalle due popolazioni da confrontare. Vogliamo testare l'ipotesi nulla H_0 che le distribuzioni delle due popolazioni siano uguali contro l'alternativa che non lo siano. Per iniziare, si uniscono i due campioni in uno unico di grandezza $N = n+m$, si ordinano i dati dal più piccolo al più grande e si assegna ad ognuno un grado secondo la posizione che occupa nel nuovo campione[6]: al primo dato si assegna il valore 1, al secondo il 2, ..., all'ultimo il valore N. A questo punto si considera come statistica test la somma dei gradi dei dati provenienti dal primo campione

$$W_1 = \sum_{i=1}^{N} iS_i$$

$$\text{con} \quad S_i = \begin{cases} 1\,, & \text{se l'i-esimo dato appartiene al primo campione} \\ 0\,, & \text{se l'i-esimo dato appartiene al secondo campione} \end{cases}$$

dove si può considerare come primo campione uno qualunque dei due iniziali; per fissare le idee stabiliamo che il primo sia X, quello di dimensione n. Guardando più attentamente alla statistica test, possiamo notare che quando H_0 è vera gli N dati provengono dalla stessa distribuzione e dunque l'insieme dei gradi del primo campione avrà la stessa distribuzione di una selezione casuale uniforme di n valori tra $\{1, 2, \ldots, n+m\}$. Da ciò si dimostra che se l'ipotesi nulla è vera allora il valore atteso e la varianza della statistica test

[6]Supponiamo che non ci sia nessun ex aequo, cosa estremamente improbabile se le X_i e le Y_j hanno distribuzioni continue.

sono

$$\mathrm{E}[W_1] = \frac{n(n+m+1)}{2} \tag{2.2}$$

$$\mathrm{Var}[W_1] = \frac{nm(n+m+1)}{12}\,. \tag{2.3}$$

Ora, supposto che il valore osservato della statistica test sia w, si può rifiutare H_0 con un livello di significatività α se

$$P(W_1 \leq w) \leq \frac{\alpha}{2} \quad \text{oppure} \quad P(W_1 \geq w) \leq \frac{\alpha}{2}$$

dove entrambe le probabilità precedenti sevono essere calcolate sotto l'assunzione che H_0 sia vera. In altre parole, l'ipotesi nulla sarà da rifiutare se la somma dei ranghi del primo campione è troppo grande o troppo piccola per essere spiegata dalla casualità. Di conseguenza, segue che il test di Wilcoxon a livello α porta a rifiutare H_0 se il p-value, dato da

$$p \text{ value} = 2\min\left(P(W_1 \leq w), P(W_1 \geq w)\right)$$

è minore o uguale ad α. Per calcolare le probabilità precedenti, notiamo dapprima che la statistica test W_1 raggiunge il suo minimo quando il primo campione è interamente più piccolo del secondo, così $\min W_1 = n(n+1)/2$. Analogamente $\max W_1 = n(2N - n + 1)/2$. Notiamo pertanto che, detto $k_{n,m}(w)$ il numero di configurazioni di n uni ed m zeri in $S_1, \ldots, S_N$ tali che $W_1 = w$, allora la distribuzione di probabilità

$$P(W_1 = w) = \frac{k_{n,m}(w)}{\binom{N}{n}}$$

$$\text{con} \quad \frac{n(n+1)}{2} \leq w \leq \frac{n(2N-n+1)}{2} \tag{2.4}$$

può essere usata per effettuare un test esatto, anche se le tavole dove consultare i valori della distribuzione di probabilità di Wilcoxon arrivano a considerare campioni di piccola dimensione, nello specifico n ed m minori di 30.

Quando invece n ed m sono entrambi più grandi di tali valori, si può dimostrare che la statistica test W_1 ha distribuzione approssimativamente uguale a quella normale con valore atteso (2.2) e varianza (2.3), dunque è possibile (e conveniente) ricorrere al test Z con l'uso della statistica test

$$Z = \frac{W_1 - \mathrm{E}[W_1]}{\sqrt{\mathrm{Var}[W_1]}}.$$

Test di Mann-Whitney ed equivalenza con Wilcoxon. Come il test di Wilcoxon, il test U, non parametrico, si applica per trovare differenze tra due popolazioni senza l'assunzione che le due distribuzioni siano normali o di altro genere. Costruiamo ora solo la statistica test, senza andare a precisare in che modo si possa rifiutare l'ipotesi nulla: provando l'equivalenza con il test di Wilcoxon, sarà sufficiente basarsi sui risultati precedenti.

Siano $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ come prima e definiamo $D_{ij} = \mathbf{1}_{Y_j<X_i}$ per $i = 1, \ldots, n$ e $j = 1, \ldots, m$, con $\mathbf{1}_{Y_j<X_i}$ funzione indicatrice che vale 1 quando $Y_j < X_i$ e 0 altrimenti. Si sceglie come statistica test la seguente

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} D_{ij}$$

cioè il numero di coppie (i, j) tali che $Y_j < X_i$. Proviamo ora che i test U e di Wilcoxon sono equivalenti. Fissato i, consideriamo

$$\sum_{j=1}^{m} D_{ij} = D_{i1} + D_{i2} + \ldots + D_{im}$$

che corrisponde al numero di indici j per cui $Y_j < X_i$. Questa somma è uguale ad $r(X_i)$, il grado di X_i nel campione di grandezza N, meno k_i, il

numero di valori del primo campione che sono minori o uguali ad X_i. Allora

$$U = \sum_{i=1}^{n} (r(X_i) - k_i) = \sum_{i=1}^{n} r(X_i) - (k_1 + k_2 + \ldots + k_n)$$
$$= \sum_{i=1}^{n} iS_i - (1 + 2 + \ldots + n) = W_1 - \frac{n(n+1)}{2}$$

dunque la statistica U di Mann-Whitney è equivalente a W_1, statistica test di Wilcoxon. Data la formula precedente e gli estremi di W_1 esplicitati nella (2.4), l'intervallo di valori che può assumere U va da 0 ad nm. Inoltre si dimostra[7] che il valore medio e la varianza di U sono

$$\mathrm{E}[U] = \frac{nm}{2} \tag{2.5}$$
$$\mathrm{Var}[U] = \frac{nm(n+m+1)}{12} \tag{2.6}$$

e che, come per W_1 nel caso di n ed m abbastanza grandi, U ha distribuzione approssimativamente normale con valore medio (2.5) e varianza (2.6).

2.2.5 Test di Kolmogorov-Smirnov

Il test di Kolmogorov-Smirnov è un test non parametrico usato per verificare la forma delle distribuzioni campionarie. Esso può essere utilizzato per confrontare un campione di dati con una distribuzione di probabilità di riferimento (test K-S per un campione) oppure due campioni tra loro per verificare che entrambi provengano dalla stessa distribuzione (test K-S per due campioni).

Test K-S per un campione. Supponiamo che $X_1, \ldots, X_n$ sia un campione ordinato di osservazioni reali di dimensione n, indipendenti e identicamente distribuite, con funzione di ripartizione $F(x) = P(X_i \leq x)$ e consideriamo il problema di testare l'ipotesi nulla $H_0 : F = F_0$ contro l'alternativa $H_1 : F \neq F_0$, dove F_0 è la funzione di ripartizione continua di una distribu-

[7] Si veda [11].

zione nota. Detta $\mathbf{1}_{X_i \leq x}$ la funzione indicatrice uguale ad 1 se $X_i \leq x$ e zero altrimenti, definiamo la funzione di ripartizione empirica

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}$$

che indica la frequenza relativa dei valori campionari minori o uguali a x. Risulta evidente che

$$\begin{aligned} \mathrm{E}[I(X_i \leq x)] &= F(x) \\ \mathrm{Var}[I(X_i \leq x)] &= F(x)[1 - F(x)]\,. \end{aligned}$$

e dunque $\forall x \in \mathbb{R}\,,\ \sum_{i=1}^{n} I(X_i \leq x)$ ha distribuzione $\mathrm{Bin}(n, F(x))$ con densità di probabilità pari a

$$P\left(\sum_{i=1}^{n} I(X_i \leq x) = k\right) = \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}\,.$$

Di conseguenza $F_n(x)$ è una variabile aleatoria a valori in $\{0,\ \frac{1}{n},\ \frac{2}{n},\ \ldots,\ \frac{n-1}{n},\ 1\}$ e, per il teorema del limite centrale, si ha che per ogni $x \in \mathbb{R}$, quasi sicuramente e dunque in distribuzione,

$$\sqrt{n}(F_n(x) - F(x)) \approx N(0,\ F(x)[1 - F(x)]) \quad \text{per } n \text{ grande.}$$

Si dimostra che anche

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

converge in distribuzione[8], risultato più forte del precedente che mette in luce l'importanza della funzione di ripartizione empirica come strumento di stima non parametrica di una funzione di ripartizione teorica. Dimostriamo

[8]Per tale dimostrazione si utilizzano teoremi avanzati della teoria dei processi stocastici, in particolare in questo caso si vede che $F_n(x) - F_0(x)$ converge al cosiddetto ponte Browniano o *Brownian bridge*; si veda per i dettagli il paragrafo 14.2 di [14].

inoltre che D_n non dipende dalla particolare distribuzione F_0 scelta.

Teorema 2.2.1. *Se $F_0(x)$ è continua allora la distribuzione di*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|$$

non dipende da F_0 .

Dimostrazione. Proviamo il teorema prima nel caso semplificato che F_0 sia strettamente crescente. In questo caso esiste ed è ben definita la funzione inversa di F_0, anch'essa strettamente crescente. Allora effettuando il cambio di variabile $x = F_0^{-1}(y)$ possiamo scrivere

$$P\left(\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| \leq t\right) = P\left(\sup_{y\in]0,1[} |F_n(F_0^{-1}(y)) - F_0(F_0^{-1}(y))| \leq t\right) .$$

Intanto $F_0(F_0^{-1}(y)) = y$ e, usando la definizione della funzione di ripartizione empirica F_n, abbiamo che

$$F_n(F_0^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq F_0^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(F_0(X_i) \leq y)$$

e quindi

$$P\left(\sup_{y\in]0,1[} |F_n(F_0^{-1}(y)) - y)| \leq t\right) = P\left(\sup_{y\in]0,1[} \left|\frac{1}{n}\sum_{i=1}^{n} I(F_0(X_i) \leq y) - y\right| \leq t\right) .$$

Notiamo che le variabili aleatorie $U_i = F_0(X_i)$ per $i \leq n$ sono indipendenti e hanno distribuzione uniforme in $[0, 1]$. Infatti fissato $u \in [0, 1]$, esiste $x_u \in \mathbb{R}$ tale che $F(x_u) = u$ e poiché

$$F(u) = P(U_i < u) = P(F(X_i) < F(x_u)) = P(X_i < x_u) = F(x_u) = u$$

dunque U_i ha distribuzione uniforme in [0,1]. Così la distribuzione di D_n è la stessa della statistica di Kolmogorov-Smirnov per una distribuzione uniforme in $[0, 1]$, quindi non dipende da F_0. Nel caso generale, F_0^{-1} non esiste

necessariamente, per questo si considera la funzione *pseudoinversa* di F_0

$$F^+(y) = \inf\{t|F_0(t) \geq y\}\,.$$

Utilizzando la proprietà che la condizione $F^+(y) \leq z$ equivale a $F_0(z) \leq y$, si conclude la prova sostituendo F^+ al posto di F_0^{-1} nella dimostrazione del caso precedente. □

D'altra parte, se si è interessati alla stima di $F(\cdot)$ in ogni punto dell'asse reale, ovvero se si vuole valutare la distanza tra $F_n(x)$ e $F(x)$ per ogni x, si ha bisogno di un risultato più forte del precedente, fornito dal seguente

Teorema 2.2.2 (Glivenko-Cantelli)**.** *Siano $X_1, \ldots, X_n$ variabili aleatorie reali i.i.d. con funzione di ripartizione F. Detta F_n la funzione di ripartizione empirica, allora*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0 \quad \textit{quasi sicuramente.}$$

Dimostrazione. Diamo solo una traccia, i dettagli si possono trovare in [18]: si noti che per ogni x fissato, dalla legge forte dei grandi numeri si ha che $F_n(x) \to F(x)$ quasi sicuramente. Per concludere si sfrutti la monotonia della F per avere che la convergenza appena descritta è uniforme in x. □

Il teorema di Glivenko-Cantelli implica che

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| \longrightarrow \sup_{x\in\mathbb{R}} |F(x) - F_0(x)| \quad \text{q. s.}$$

e quindi una convergenza in probabilità, dove il valore del limite è zero se e solo se $F = F_0$, cioè se H_0 è vera. Di conseguenza se n è sufficientemente grande, la statistica D_n tenderà ad essere piccola se H_0 è vera, grande altrimenti. Tutte queste considerazioni portano a scegliere D_n come statistica test di Kolmogorov-Smirnov.

Resta da vedere come si calcola la distribuzione della statistica D_n quando l'ipotesi nulla è vera. Si dimostra[9] che la distribuzione di D_n è tale che

$$P(D_n \leq t) \longrightarrow H(t) = 1 - 2\sum_{i=1}^{+\infty} (-1)^{i-1} e^{-2i^2 t}$$

dove $H(t)$ è la funzione di ripartizione della cosiddetta distribuzione K di Kolmogorov-Smirnov. Quindi per n abbastanza grande possiamo usare i valori critici di K per costruire la regione critica del test. Come tavole di riferimento si prendano, per campioni di numerosità $n \leq 35$, quelle fornite in [20], mentre per $n > 35$ si guardino le tavole in [21], integrate da [22]. Infine possiamo affermare che l'ipotesi nulla H_0 è da rifiutare a livello α se $D_n > k_\alpha$, dove k_α è tale che $H(k_\alpha) = P(K \leq k_\alpha) = 1 - \alpha$.

Test K-S per due campioni. Questo test risulta molto simile al precedente. Siano $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ due campioni casuali indipendenti estratti dalle due popolazioni da confrontare, con funzioni di ripartizione rispettivamente F ed G. Testiamo l'ipotesi nulla

$$H_0 : F = G$$

contro l'alternativa

$$H_1 : F \neq G\,.$$

Dette F_n e G_m le corrispondenti funzioni di ripartizione empiriche, allora la statistica test

$$D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_t |F_n(t) - G_m(t)|$$

è l'analoga della precedente statistica D_n ed il test di Kolmogorov-Smirnov per due campioni si esegue allo stesso modo di quello per un campione.

[9] La prova si può trovare in [19].

2.3 Analisi della regressione

Per stimare la relazione tra diverse variabili si utilizza un processo statistico noto come **analisi della regressione**: da un campione si osservano le misurazioni riguardo a differenti grandezze, dette **regressori** o **variabili indipendenti**, e si esamina la relazione tra queste e una **variabile dipendente**, o **risposta**. Questa relazione è espressa attraverso una funzione f, detta **modello di regressione**, che può servire a predire la risposta sulla base dei regressori, i quali per questo motivo sono detti anche **predittori**. Più in generale, supponiamo di osservare una variabile risposta Y e p regressori differenti $X_1, X_2, \ldots, X_p$. Assumiamo che ci sia una qualche relazione tra Y e $X = (X_1, X_2, \ldots, X_p)$, che può essere scritta nella forma

$$Y = f(X) + \epsilon\,,$$

dove f è una funzione deterministica fissata ma incognita dei regressori ed ϵ è un errore casuale, che supponiamo indipendente da X e con media nulla. Poiché non conosciamo f, vogliamo trovare una stima $\tilde{f}$ di tale funzione che ci permetta di fornire una previsione $\tilde{Y}$ della risposta Y sulla base delle variabili indipendenti X, ovvero

$$\tilde{Y} = \tilde{f}(X). \tag{2.7}$$

Ci sono due aspetti principali da tenere in conto per la stima di f: il primo è l'accuratezza della previsione, mentre il secondo è l'interpretazione.

Accuratezza della previsione. Dal momento che $\tilde{f}$ è una stima del modello di regressione, sarà inevitabile la presenza di un qualche tipo di errore. Consideriamo lo stimatore $\tilde{f}$ e un insieme di predittori x, che supponiamo per il momento fissati entrambi. Per avere un buon stimatore, appare evidente che esso dovrebbe avere in media un basso errore di previsione, misurabile attraverso la quantità $E[(Y - \tilde{Y})^2]$. Vediamo che questo valor medio può

essere scomposto in due parti:

$$\begin{aligned} E\left[(Y-\tilde{Y})^2\right] &= E\left[(f(x)+\epsilon-\tilde{f}(x))^2\right] \\ &= E\left[(f(x)-\tilde{f}(x))^2\right] + E\left[\epsilon^2\right] + 2E\left[\epsilon(f(x)-\tilde{f}(x))\right] \\ &= E\left[(f(x)-\tilde{f}(x))^2\right] + Var(\epsilon) + 2\underbrace{E[\epsilon]}_{=0} E\left[f(x)-\tilde{f}(x)\right] \end{aligned}$$

cioè

$$E\left[(Y-\tilde{Y})^2\right] = \underbrace{E\left[(f(x)-\tilde{f}(x))^2\right]}_{\text{riducibile}} + \underbrace{Var(\epsilon)}_{\text{irriducibile}} . \tag{2.8}$$

Come si vede, l'errore di previsione dipende da due quantità dette **errore riducibile** ed **errore irriducibile**: il primo perché è possibile migliorare l'accuratezza di $\tilde{f}$ usando una più appropriata tecnica statistica per stimare f, mentre il secondo è irriducibile perché Y è anche funzione di ϵ che invece, essendo per definizione indipendente dai regressori, non è possibile prevedere attraverso x. Quindi ha senso concentrarsi sul cercare i metodi per minimizzare l'errore riducibile, così da poter avere una buona accuratezza della previsione anche utilizzando dati diversi da quelli adoperati per costruire il modello. Sull'errore riducibile, detto anche MSE (*mean squared error*) possiamo dire di più: infatti osserviamo che, fissato $f(x)$, esso può ulteriormente

essere scomposto in due parti. Poiché vale che

$$\begin{aligned}
MSE(\tilde{f}) &= E\left[(f(x) - \tilde{f}(x))^2\right] \\
&= E\left[(f(x) - E[\tilde{f}(x)] + E[\tilde{f}(x)] - \tilde{f}(x))^2\right] \\
&= E\left[(f(x) - E[\tilde{f}(x)])^2\right] + E\left[(E[\tilde{f}(x)] - \tilde{f}(x))^2\right] \\
&\quad + 2E\left[(f(x) - E[\tilde{f}(x)])(E[\tilde{f}(x)] - \tilde{f}(x))\right] \\
&= E\left[(f(x) - E[\tilde{f}(x)])^2\right] + E\left[(E[\tilde{f}(x)] - \tilde{f}(x))^2\right] \\
&\quad + 2E\left[f(x)E[\tilde{f}(x)] - f(x)\tilde{f}(x) - E[\tilde{f}(x)]^2 + \tilde{f}(x)E[\tilde{f}(x)]\right] \\
&= \left(f(x) - E[\tilde{f}(x)]\right)^2 + E\left[(E[\tilde{f}(x)] - \tilde{f}(x))^2\right] \\
&\quad + 2E[f(x)]E[\tilde{f}(x)] - 2E[f(x)\tilde{f}(x)] - 2E[\tilde{f}(x)]^2 + 2E[\tilde{f}(x)]E[\tilde{f}(x)]
\end{aligned}$$

allora, dal momento che $E[f(x)] = f(x)$ dato che $f(x)$ è una funzione deterministica, si ha

$$E\left[(f(x) - \tilde{f}(x))^2\right] = \text{Bias}^2(\tilde{f}(x)) + \text{Var}(\tilde{f}(x)) \tag{2.9}$$

dove $\text{Bias}(\tilde{f}(x)) = E[\tilde{f}(x)] - f(x)$ è il *bias*, o distorsione, dello stimatore $\tilde{f}(x)$ di $f(x)$ e invece $\text{Var}(\tilde{f}(x)) = E[(E[\tilde{f}(x)] - \tilde{f}(x))^2]$ è la sua varianza. Intuitivamente, il *bias* della stima si riferisce all'errore che si introduce approssimando un problema reale, che può essere estremamente complicato, con un modello più semplice, mentre la varianza dice quanto cambia $\tilde{f}$ se lo testiamo su dati differenti da quelli adoperati per costruirlo. Tipicamente, all'aumentare della complessità del modello stimato $\tilde{f}$ (*overfitting*), corrisponde una diminuzione del bias (al quadrato) e contemporaneamente una crescita della varianza, mentre l'utilizzo di modelli troppo semplici (*underfitting*) produrrà un'elevata distorsione e una bassa varianza, quindi non esistono metodi che azzerino completamente l'errore riducibile. L'unica possibilità è trovare un compromesso tra bias e varianza in modo da minimizzare MSE. Due modi per farlo sono l'utilizzo di metodi di regolarizzazione, che vedremo nei prossimi paragrafi sulla regressione multilineare, e della *cross-validation*, di cui

parleremo nel prossimo capitolo.

Interpretazione del modello. Vorremmo anche che fosse possibile comprendere come Y cambi non solo in relazione a X, ma ad ogni singolo predittore X_i. Ciò è importante perché, in molte situazioni, solo una parte dei regressori è effettivamente associata ad Y. Identificare quindi i più importanti di essi può essere estremamente utile in molte applicazioni. Inoltre alcuni predittori possono avere una relazione positiva con Y, nel senso che la crescita dell'uno è associata all'aumento dell'altro. Altri predittori possono invece avere la relazione contraria, oppure la relazione tra la risposta e un dato regressore può addirittura dipendere dai valori degli altri, situazione che evidentemente non è sempre deducibile da un modello troppo semplice. Sarebbe auspicabile che la relazione tra Y e ogni predittore sia lineare, per una questione di semplicità e chiarezza del modello, ma d'altro canto la vera relazione è spesso più complicata.

A seconda che il nostro obiettivo principale sia l'accuratezza o l'interpretazione, o una combinazione delle due, può essere appropriato adoperare differenti metodi per stimare f. Per esempio, i modelli lineari danno un'interpretazione abbastanza semplice, ma non sono accurati come altri. Infatti esistono alcuni metodi non lineari che potenzialmente possono raggiungere un grande grado di accuratezza nella previsione, ma a spese di una scarsa interpretazione del modello. In questa tesi affronteremo solo il caso di una funzione di regressione *lineare* nelle variabili indipendenti[10]. Una tale funzione potrà essere pertanto descritta geometricamente da una retta se c'è solo un predittore (regressione lineare semplice) o da un iperpiano se ce ne sono diversi (regressione lineare multipla o multivariata). Esporremo anche metodi di regolarizzazione per la regressione lineare che ci consentiranno di raggiungere buoni livelli sia di previsione che di interpretazione.

[10]In generale un modello di regressione è detto lineare se lo è nei parametri β_i e non è necessario che lo sia anche nelle variabili indipendenti.

2.3.1 Regressione lineare multipla

Siano $x = (x_1, x_2, \ldots, x_p)$ un vettore di p predittori e Y la variabile dipendente. Il modello di regressione lineare multipla è della forma

$$\tilde{Y} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{2.10}$$

dove $\tilde{Y}$ è la risposta predetta della variabile Y e $\beta_0, \beta_1, \ldots, \beta_p$ sono i coefficienti di regressione. Il termine β_0 è chiamato **intercetta** e, nel caso di una regressione lineare semplice in cui $p = 1$, β_1 è detto **pendenza**.

Per ora abbiamo considerato una singola risposta, quindi $\tilde{Y}$ è uno scalare; supponiamo in generale di avere N campioni sperimentali $\mathbf{y} = (y_1, \ldots, y_N)^T$ della variabile Y e i corrispondenti valori osservati $\mathbf{x} = (x_1, \ldots, x_N)^T$ dei regressori, dove $x_i = (1, x_{i1}, \ldots, x_{ip})$ per $i = 1, \ldots, N$; dunque

$$\mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}.$$

è una matrice $N \times (p + 1)$. Si noti che, per comodità, abbiamo inserito la costante 1 nella prima componente di ogni x_i. Con questo insieme di misure $(x_1, y_1) \ldots (x_N, y_N)$, noto come *training data*, andremo a fornire una previsione $\tilde{\mathbf{y}}$ della variabile dipendente $\mathbf{Y}$. Occorre dunque utilizzare qualche metodo per trovare i parametri β_i per cui la funzione di regressione si avvicini il più possibile al nostro training data. Esamineremo adesso quello più popolare, il metodo dei minimi quadrati, e in seguito ne illustreremo altri che meglio si adattano al nostro caso di studio.

2.3.2 Metodo dei minimi quadrati

Vogliamo trovare il vettore dei coefficienti $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ che minimizzi la seguente quantità

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 \\ &= \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \right)^2 \end{aligned}$$

o in forma matriciale

$$RSS(\beta) = (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta)\,.$$

cioè una funzione quadratica in $p+1$ parametri. Differenziando rispetto a β otteniamo

$$\begin{aligned} \frac{\partial RSS}{\partial \beta} &= -2\mathbf{x}^T(\mathbf{y} - \mathbf{x}\beta) \\ \frac{\partial^2 RSS}{\partial \beta^2} &= 2\mathbf{x}^T\mathbf{x}. \end{aligned}$$

Assumendo che $\mathbf{x}$ abbia rango completo, $\mathbf{x}^T\mathbf{x}$ risulta invertibile e definita positiva e imponendo che

$$\frac{\partial RSS}{\partial \beta} = 0\,,$$

cioè

$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}\beta\,, \tag{2.11}$$

otteniamo la soluzione unica

$$\tilde{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}\,. \tag{2.12}$$

e quindi la previsione

$$\tilde{\mathbf{y}} = \mathbf{x}\tilde{\beta}\,. \tag{2.13}$$

In generale però la stima (2.13) del metodo dei minimi quadrati non è spesso soddisfacente.

1. Per quanto riguarda l'accuratezza della previsione, se la vera relazione tra la variabile dipendente e i regressori è approssimativamente lineare, questo metodo produce un *bias* basso. Se $N \gg p$ allora il metodo dei minimi quadrati tende ad avere anche una varianza bassa e quindi rappresenta bene le osservazioni test. Comunque, se N non è molto più grande di p, allora può esserci molta variabilità nell'adattamento ai dati e di conseguenza una scarsa previsione su osservazioni future esterne al *training data*. E se $p > N$, allora la soluzione di (2.11) non è unica dal momento che la matrice $\mathbf{x}^T\mathbf{x}$ di dimensione $(p+1) \times (p+1)$ ha rango massimo N e dunque è singolare e non può essere invertita. Le infinite soluzioni del sistema sotto-determinato hanno tutte un buon adattamento al *training data*, ma scarso per osservazioni diverse.

2. Per l'interpretazione, questo metodo non permette di selezionare i regressori più significativamente correlati con la variabile dipendente. Accade spesso che molti regressori siano in effetti non correlati con la risposta e dunque includere queste variabili irrilevanti porta a complicare inutilmente il modello.

Un modo per migliorare sia l'accuratezza che l'interpretazione è utilizzare i cosiddetti *shrinkage methods*: di questa classe vediamo, nell'ordine, i metodi *ridge*, *lasso* ed *elastic-net*. Con tali tecniche, si vincola la grandezza dei parametri e, producendo un lieve aumento del bias in favore di una diminuzione della varianza dei valori predetti, si migliora la previsione globale, anche per osservazioni non appartenenti al *training data*; allo stesso tempo questi metodi prevedono una selezione dei parametri di regressione rilevanti, riducendo a zero il valore di quelli irrilevanti, e quindi rendono il modello più facilmente interpretabile.

2.3.3 Regressione *ridge*

Questo metodo riduce i coefficienti di regressione imponendo un limite alla loro grandezza. I coefficienti *ridge* devono infatti essere tali che

$$\tilde{\beta}_R = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.14}$$

dove $\lambda \geq 0$ è il parametro che controlla la quantità di riduzione dei coefficienti $\beta_1 \ldots, \beta_p$ (l'intercetta β_0 è esclusa da tale riduzione). Possiamo scriverlo anche nella forma

$$\begin{aligned} \tilde{\beta}_R = \arg\min_{\beta} & \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \\ \text{soggetto a } & \sum_{j=1}^{p} \beta_j^2 \leq t \end{aligned} \tag{2.15}$$

dove t è in corrispondenza biunivoca con λ. Si dimostra, allo stesso modo del metodo dei minimi quadrati, che la soluzione di (2.14) è

$$\tilde{\beta}_R = (\mathbf{x}^T\mathbf{x} + \mathbf{J})^{-1}\mathbf{x}^T\mathbf{y} \tag{2.16}$$

con

$$\mathbf{J} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \lambda\mathbf{I} & \\ 0 & & & \end{bmatrix},$$

dove $\mathbf{I}$ è la matrice identità $p \times p$, $\mathbf{x}$ è la matrice $N \times (p+1)$ delle osservazioni dei p regressori e $\mathbf{y}$ è il vettore N-dimensionale delle misurazioni della variabile dipendente. Notiamo intanto che, aggiungendo una costante positiva alla diagonale di $\mathbf{x}^T\mathbf{x}$, il problema diventa non singolare anche se $\mathbf{x}$ non ha rango pieno. Inoltre per $\lambda = 0$ ritroviamo il metodo dei minimi quadrati, mentre per λ tendente all'infinito, i coefficienti in (2.16) tendono a zero, ma

per valori intermedi di λ, si dimostra[11] che nessuno dei parametri sarà mai esattamente nullo. Dunque la regressione *ridge* non può effettuare una selezione delle variabili (perché non ne esclude nessuna non potendo azzerare nessun coefficiente) e per questo, anche se fornisce una migliore accuratezza della previsione rispetto al metodo dei minimi quadrati, non fa altrettanto per quanto riguarda la chiarezza dell'interpretazione. Ritorneremo su questa questione tra poco, quando parleremo del metodo *lasso*, con il quale sarà possibile raggiungere anche un buon livello di interpretazione del modello di regressione. Osserviamo infine che per ogni valore di λ abbiamo un diverso insieme di parametri $\tilde{\beta}_R$. Per ottenere un metodo migliore dei minimi quadrati occorre dunque trovare un modo per selezionare un buon valore λ: lo vedremo parlando di *cross-validation* nel prossimo capitolo.

2.3.4 Regressione *lasso*

Vogliamo ora trovare i parametri di regressione che soddisfino, invece che (2.14), la seguente condizione

$$\tilde{\beta}_L = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.17}$$

o l'equivalente

$$\begin{aligned} \tilde{\beta}_L = \arg\min_{\beta} & \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \\ \text{soggetto a } & \sum_{j=1}^{p} |\beta_j| \leq t\,. \end{aligned} \tag{2.18}$$

Tale metodo per stimare i coefficienti è chiamato *lasso* (*least absolute shrinkage and selection operator*) e, come *ridge*, esso impone un vincolo alla grandezza dei coefficienti. L'unica differenza è l'uso della norma ℓ^1, al posto della norma ℓ^2, dei parametri $\beta_1, \ldots, \beta_p$. Questo sottile ma importante cam-

[11] Si veda il capitolo 3 di [16].

biamento permette al metodo *lasso* di porre alcuni coefficienti esattamente a zero, cosa che il metodo *ridge* non è in grado di fare. Il motivo è dovuto alla natura del vincolo sui parametri: consideriamo ad esempio il caso di $p = 2$ e vediamo in Figura 2.1 una illustrazione geometrica del perché *lasso* riesce ad azzerare alcuni coefficienti, mentre *ridge* no. La somma dei quadrati

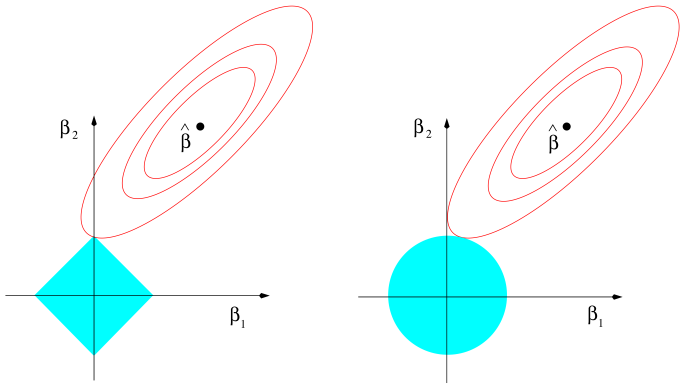


Figura 2.1: Illustrazione geometrica della stima dei metodi *lasso* (a sinistra) e *ridge* (a destra). Si guardi la spiegazione nel testo.

dei residui ha contorni ellittici, centrati attorno alla stima del metodo dei minimi quadrati $\tilde{\beta}$. Il vincolo per la regressione *ridge* corrisponde al disco $\beta_1^2 + \beta_2^2 \leq t$, mentre quello per *lasso* è il quadrato $|\beta_1| + |\beta_2| \leq t$. Per entrambi i metodi, la loro stima dei parametri coincide con l'intersezione dell'ellisse con il bordo della regione corrispondente al proprio vincolo ma, a differenza del disco, il quadrato ha vertici, nei quali uno dei parametri è esattamente zero. Per $p > 2$ si hanno più vertici, spigoli e facce e dunque ci sono più possibilità per i parametri del *lasso* di essere stimati a zero, permettendo di scartare dal modello i regressori non rilevanti e dunque effettuando una selezione delle variabili indipendenti. Quindi, non solo questo metodo riduce la varianza a spese di un lieve aumento del bias come il *ridge*, ma in più seleziona i predittori e permette una più facile interpretazione del modello.

In generale, ci si aspetta che *lasso* si comporti meglio quando abbiamo pochi predittori con coefficienti considerevoli e i rimanenti con coefficienti

piccoli o anche nulli. La regressione *ridge* invece è più accurata quando la risposta è una funzione di molti regressori, tutti con coefficienti di ampiezza abbastanza simile. Comunque, il numero di predittori legato alla variabile dipendente non è mai noto a priori per dati reali. Una tecnica come la *cross-validation*, vedremo, può essere usata per determinare quale metodo sia migliore a seconda dei casi.

Anche il *lasso* però ha delle limitazioni. Consideriamo i seguenti tre scenari.

1. Nel caso $p > N$, il lasso può selezionare al massimo N variabili (si veda [23]) e questo dunque limita le capacità del metodo quando N è di molto inferiore a p.

2. Se c'è un gruppo di variabili tra loro molto correlate, allora *lasso* tende a selezionare una sola delle variabili del gruppo, escludendo le altre (si veda [24]).

3. Per le situazioni in cui $N > p$ invece, se ci sono alte correlazioni tra i predittori, è stato empiricamente osservato che la stima della previsione del *lasso* è dominata da quella della regressione *ridge* (si veda [25]).

Gli scenari 1 e 2 rendono il lasso un metodo di selezione delle variabili inappropriato per alcune situazioni, come nel caso di applicazioni in genetica, in cui si hanno centinaia di migliaia di predittori (ognuno associato ad un gene, per esempio ad un nucleotide del DNA) e poche centinaia di campioni. Per quei predittori che sono fortemente correlati, che possiamo pensare facenti parte come di un gruppo, la selezione ideale dovrebbe essere capace di fare due cose: scartare i predittori non rilevanti e includere automaticamente l'intero gruppo nel modello una volta che un predittore appartenente ad esso viene selezionato. Sarà capace di ciò il metodo che chiude questo capitolo, l'*elastic-net*, che inoltre riesce a produrre una previsione più accurata del *lasso* in una situazione come la 3.

2.3.5 Regressione *elastic-net*

Questo metodo combina i vincoli del *ridge* e del *lasso* in uno unico: la stima dei coefficienti per la regressione *elastic-net* è infatti definita per $\lambda > 0$ e $\alpha \in [0, 1]$ da

$$\tilde{\beta}_{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda P_\alpha(\beta) \right\} \tag{2.19}$$

$$\text{dove } P_\alpha(\beta) = \sum_{j=1}^{p} \left((1-\alpha)\beta_j^2 + \alpha|\beta_j| \right) \tag{2.20}$$

o equivalentemente fissato $t > 0$

$$\tilde{\beta}_{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}$$
$$\text{soggetto a } P_\alpha(\beta) \leq t\,. \tag{2.21}$$

Notiamo subito che per $\alpha = 1$ ritroviamo *lasso*, mentre per $\alpha = 0$ riotteniamo *ridge*. Invece per $\alpha \in]0, 1[$ abbiamo un utile compromesso tra i due metodi: l'*elastic-net* può selezionare le variabili come *lasso* col vantaggio che, nel caso $p > N$, non si ferma ad N predittori selezionati e inoltre, come *ridge*, riduce allo stesso valore i coefficienti dei predittori correlati. In letteratura si è discussa approfonditamente questa capacità di selezionare tutti i predittori correlati tra loro (detta *grouping effect*) dei metodi di regressione: si veda l'articolo [24], in cui viene per la prima volta introdotto il metodo *elastic-net*, per le dimostrazioni di alcuni casi particolari che riguardano tale metodo e anche altri. Sempre dallo stesso articolo sono forniti i risultati empirici e le simulazioni che dimostrano la buona prestazione di questo metodo e la sua superiorità su *lasso* in termini di accuratezza della previsione. Infine, come per i due metodi precedenti, sarà indispensabile utilizzare la *cross-validation* per trovare i parametri α e λ ottimali che rendono minimo l'errore di previsione.

Capitolo 3

Analisi dati

Come abbiamo visto nel primo capitolo, i cambiamenti nella metilazione del DNA sono legati all'invecchiamento dell'organismo, ma non abbiamo ancora chiarito come il metiloma possa essere impiegato per misurare e confrontare i tassi di invecchiamento, in particolare nell'essere umano. Di recente uno studio sui profili di metilazione negli esseri umani [26] ha rivelato in maniera quantitativa come i pattern di metilazione rappresentino un forte e riproducibile indicatore del tasso di invecchiamento. I risultati principali ottenuti da questa ricerca saranno illustrati nel primo paragrafo di questo capitolo, mentre nel secondo discuteremo delle nostre analisi iniziali, basate sui dati messi a disposizione dal suddetto lavoro, e dei nostri primi risultati.

3.1 Il *dataset* di riferimento

L'esperimento ha interessato due differenti gruppi di persone ($N_1 = 482$, $N_2 = 174$), con una finestra di età tra i 19 e i 101 anni. I loro campioni di sangue sono stati trattati con uno specifico *chip a DNA*[1] che misura il livello

[1]Un *microarray di DNA* (comunemente conosciuto come *gene chip* o *chip* a *DNA*) è un insieme di microscopici tratti di DNA (detti sonde o *probe*) attaccati ad una superficie solida come vetro, plastica, o chip di silicio formanti un array (matrice). Tali array permettono di rilevare simultaneamente, in un campione di DNA, la presenza di moltissimi geni, determinati dalle sequenze nucleotidiche complementari a quella del probe.

di metilazione di 485577 marcatori[2] di siti CpG. Di questi, sono stati scartati i siti appartenenti a cromosomi sessuali, riducendo il numero a $p = 473034$ marcatori. La metilazione di un certo marker è stata registrata come una frazione tra zero e uno, rappresentante la proporzione di cellule che hanno quel sito metilato rispetto al totale di cellule sanguigne analizzate. Il risultato è salvato in una matrice che ha per righe i p marcatori e per colonne gli $N = N_1 + N_2 = 656$ pazienti ed è disponibile su GEO (*Gene Expression Omnibus*), uno dei più grandi archivi di dati genomici al mondo.

L'obiettivo dell'analisi di questi dati è riuscire a fornire per ogni paziente una stima dell'età (variabile dipendente) sulla base dei valori di metilazione osservati (variabili indipendenti). Per fornire questa stima è stato utilizzato il modello di regressione dell'*elastic-net*. Poiché siamo in una situazione in cui $p \gg N$ e sarebbe auspicabile restringere il numero di regressori effettivamente associati all'età, tale metodo è sicuramente adatto per quanto esposto nel capitolo precedente. L'algoritmo per la computazione dell'*elastic-net* è implementato nel pacchetto 'glmnet' [27].

Ricordando che il metodo *elastic-net* ha due parametri liberi, λ e α, si è utilizzata la tecnica della *10-fold cross-validation* per trovarne i valori ottimali. In generale, la *K-fold cross-validation* prevede che si svolgano i seguenti passi.

1. Si divide in maniera casuale il *training data* T in K insiemi della stessa grandezza, nel nostro caso $K = 10$. Si ottiene così $T = (T_1, \ldots, T_K)$, con dimensione di ogni T_i, indicata con $|T_i|$, costante $\forall i = 1 \ldots, K$.

2. Fissato λ, per ogni $i = 1 \ldots, K$ si calcola la stima dei coefficienti di regressione (i β nel caso multilineare) del modello sulla base del *training data* ottenuto escludendo T_i da T.

3. Usando i coefficienti trovati nel punto precedente, si calcolano i valori predetti $\tilde{f}_i^{(\lambda)}(x)$ della variabile dipendente per le osservazioni x

[2]Un marcatore genetico (*marker*) è un tratto di DNA con sequenza e localizzazione note, per mezzo del quale è possibile individuare una regione cromosomica.

dell'insieme T_i.

4. Si calcola la stima dell'errore di previsione dovuta alla *cross-validation* per ogni T_i

$$(\text{CV Error})_i^{(\lambda)} = |T_i|^{-1} \sum_{(x,y)\in T_i} \left(y - \tilde{f}_i^{(\lambda)}(x)\right)^2 \tag{3.1}$$

5. Il modello allora avrà, per λ fissato, un errore globale di *cross-validation*

$$(\text{CV Error})^{(\lambda)} = K^{-1} \sum_{i=1}^{K} (\text{CV Error})_i^{(\lambda)}$$

6. Si ripete dal punto 2 al 5 variando il λ e infine si sceglie come λ^* ottimale per il modello quello che ha minimo $(\text{CV Error})^{(\lambda)}$.

Se c'è un altro parametro libero, come nel nostro caso in cui abbiamo anche α, si fa variare il secondo parametro e per ogni valore di esso si effettua una *cross-validation* per determinare il minimo errore e il λ^* ottimale. Si prenderà dunque come α^* ottimale per il modello, assieme al relativo $\lambda^* = \lambda^*(\alpha)$, quello che ha minimo errore $(\text{CV Error})^{(\lambda^*)}$.

Infine è stato adoperato anche il metodo *bootstrap*, che consiste nel ricampionare casualmente B volte (in questo caso è stato scelto $B = 500$) con reimmissione dell'intero *training data*: se abbiamo un numero D di dati a disposizione, si estraggono a caso (in maniera uniforme) D valori da quei dati, con la possibilità di ripescare anche più volte gli stessi, e dunque si producono B nuovi insiemi di dati, ognuno di dimensione D. Per ognuno dei 500 nuovi gruppi si è applicato il metodo dell'*elastic-net* con i parametri forniti dalla *cross-validation* e dunque sono stati selezionati 500 gruppi di regressori. Sono stati inclusi nel modello finale solo i predittori (*marker*) comuni ad almeno 250 gruppi, trovando così 71 marcatori altamente predittivi dell'età. L'accuratezza del modello è alta, con una correlazione tra età anagrafica e

predetta del 96%[3] e un errore di 3.9 anni[4] (Figura 3.1). È stato altresì no-

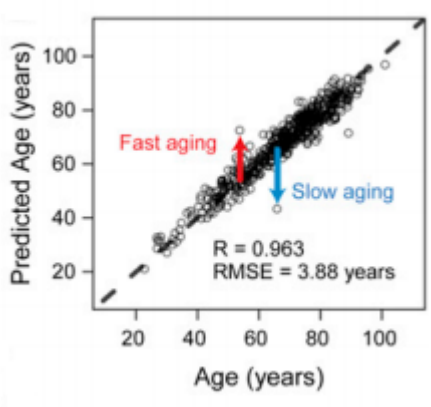


Figura 3.1: Confronto tra età predetta ed età anagrafica per gli individui del gruppo 1.

tato che quasi tutti i marker selezionati dal modello si trovano all'interno o vicino a geni con note funzioni regolatrici delle condizioni di invecchiamento. Oltre a riuscire a prevedere l'età di molti individui con una buona accuratezza, questo modello permette anche di identificare i cosiddetti *outlier*, i valori distanti dalle previsioni: per esempio, in Figura 3.1 sono evidenziati due individui le cui età sono molto sovra- o sottostimate sulla base dei loro dati di metilazione.

Infine il modello è stato testato sull'altro gruppo dei 174 campioni: con i valori dei parametri del modello costruito sui campioni del primo gruppo, sono state predette le età del secondo gruppo, ottenendo una correlazione tra età anagrafica e predetta del 91% e un errore di 4.9 anni (Figura 3.2).

[3]Tale percentuale è espressa in termini del coefficiente di correlazione di Bravais-Pearson, definito come il rapporto tra la covarianza e il prodotto delle deviazioni standard delle due variabili.

[4]L'errore indicato corrisponde alla radice quadrata del MSE (*mean squared error*) introdotto nel capitolo precedente.

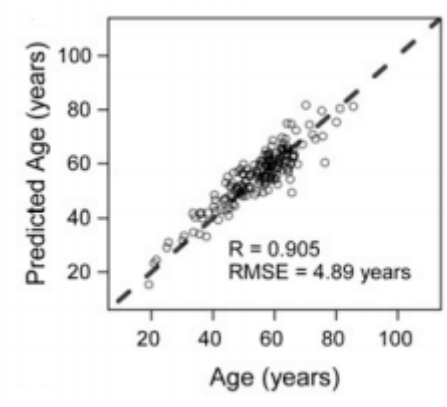


Figura 3.2: Confronto tra età predetta ed età anagrafica per gli individui del gruppo 2.

3.2 Costruzione della nostra *signature*

Uno dei primi risultati di questo lavoro di tesi è il seguente: partendo dai valori di metilazione dei 656 pazienti dello studio appena descritto, abbiamo effettuato una preselezione di $b = 10, 20, 50, 100, 200, 500, 1000$ *marker* (preselezione descritta qui di seguito), ottenendo dunque 7 diversi gruppi, o *signature*, di marcatori. Poi, per ogni gruppo, abbiamo applicato i metodi *ridge*, *lasso* ed *elastic-net* per stimare le età di tutti i pazienti. Con i coefficienti di regressione trovati per ogni metodo, testeremo i nostri modelli (per tutti i b considerati) sui dati del capitolo 5.

3.2.1 Preselezione dei marcatori

La nostra prima analisi è stata osservare la correlazione tra i valori di metilazione di ogni singolo *marker* e l'età di tutti gli $N = 656$ pazienti: utilizzando il metodo dei minimi quadrati, abbiamo calcolato l'intercetta I_j e la pendenza P_j della retta di regressione per ognuno dei $p = 473034$ *marker* sulla base delle età E_i osservate. Il modello quindi appare della seguente forma

$$\tilde{M}_{ij} = P_j\, E_i + I_j \quad \text{per } i = 1, \ldots, N \quad \text{e } j = 1, \ldots, p \tag{3.2}$$

dove $\tilde{M}_{ij}$ è la stima del livello di metilazione del marcatore j-esimo per il paziente i-esimo. In Figura 3.3 vediamo un esempio del grafico ottenuto per uno dei marcatori.

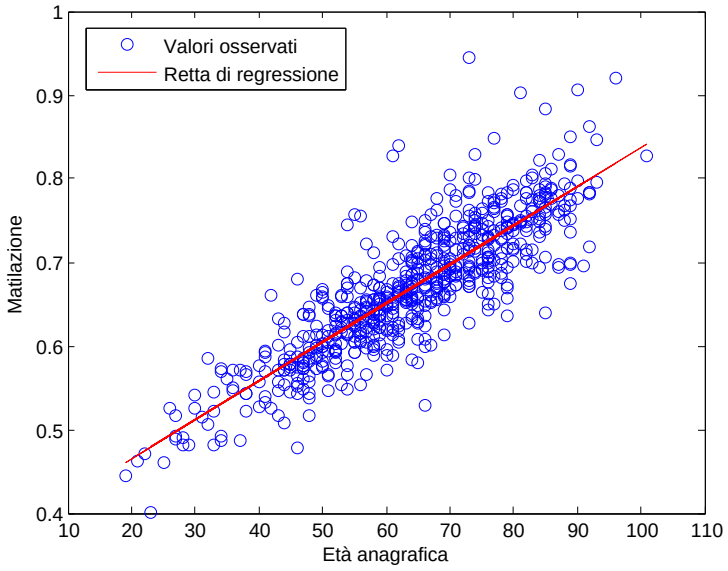


Figura 3.3: Valori di metilazione osservati e predetti per il marcatore $j = 301867$.

In seguito abbiamo calcolato per ogni marcatore la radice dell'errore quadratico medio secondo la formula

$$\text{RMSE}_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_{ij} - \tilde{M}_{ij})^2}$$

dove M_{ij} è il valore di metilazione osservato dell'i-esimo paziente per il j-esimo *marker*. Abbiamo ottenuto così 3 vettori p-dimensionali, con rispettivamente registrate le intercette, le pendenze e l'RMSE per ogni *marker*.

La nostra analisi si è poi concentrata sul decidere cosa volesse dire 'trovare i migliori marcatori'. Abbiamo deciso di graficare l'RMSE in funzione del

valore assoluto della pendenza (VAP) e stilare una classifica dei migliori b *marker* secondo il minimo rapporto tra RMSE e VAP, però solo per quei marcatori con RMSE ≤ 0.15 e VAP ≥ 0.001. Tali restrizioni sono dovute al fatto che non volevamo selezionare *marker* in cui la variazione dei livelli di metilazione al crescere dell'età fosse troppo bassa (ecco perché VAP ≥ 0.001) o che la variabilità all'interno dei dati dello stesso marcatore fosse troppo grande (ecco perché RMSE ≤ 0.15) da poter essere spiegata bene con una correlazione lineare. Vediamo in Figura 3.4 alcuni dei nostri risultati per $b = 10, 50, 200, 1000$ *marker* selezionati. Si noti infine che in tali figure abbiamo anche confrontato le diverse *signature* con i marcatori trovati in [26]: in tabella Tabella 3.1 vediamo il numero di marcatori presenti anche nel gruppo dei 71.

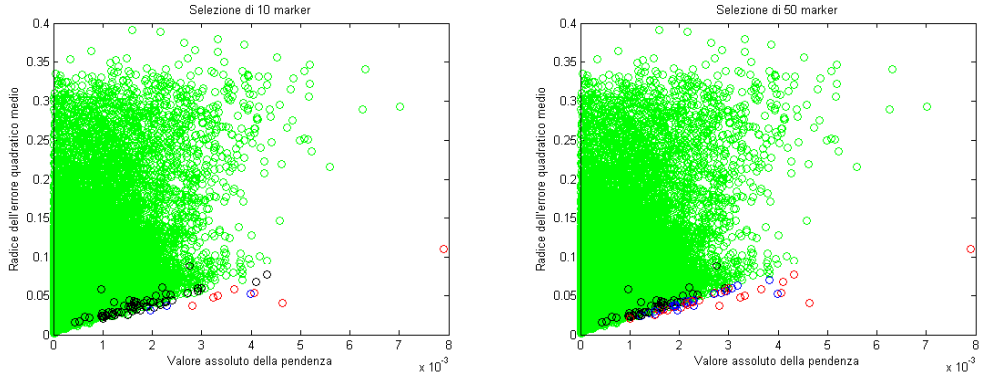
b	Marcatori comuni
10	7
20	15
50	25
100	35
200	48
500	58
1000	60

Tabella 3.1: Numero di *marker* comuni alle nostre *signature* e a quella dei 71 marcatori.

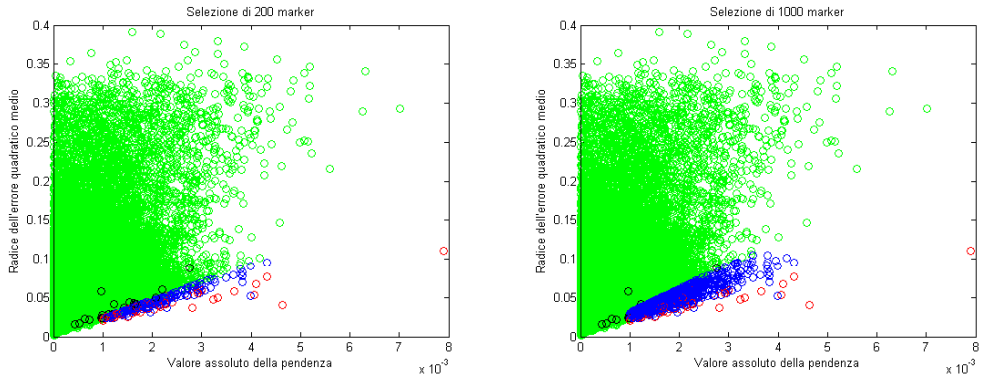
3.2.2 Applicazione dei metodi statistici

Il passo successivo è stato utilizzare i diversi metodi di regressione visti nel paragrafo 2.3 per fornire una previsione delle età per gli N individui in esame e, allo stesso tempo, ottenere per ogni metodo i coefficienti da utilizzare per prevedere le età anche di altri gruppi di pazienti. I nostri risultati sono stati ottenuti lavorando con l'ambiente Matlab.

Minimi quadrati. Attraverso la funzione 'regress.m' di Matlab, abbiamo implementato il metodo dei minimi quadrati per trovare la previsione



(a) $b = 10$ (b) $b = 50$



(c) $b = 200$ (d) $b = 1000$

Figura 3.4: Preselezione dei b marcatori: in rosso vediamo i marker in comune con la *signature* dei 71 del lavoro [26], in blu i restanti selezionati da noi, in nero quelli dei 71 non in comune e in verde tutti gli altri.

delle età degli N pazienti. Vediamo i risultati in Tabella 3.2 e qualche esempio grafico della previsione effettuata con questo metodo in Figura 3.5. Si noti che l'errore di *cross-validation* (CVerr) in questo caso è definito similmente a (3.1), senza però la dipendenza da alcun parametro, ed è utilizzato per stimare l'errore di previsione su campioni diversi da quelli analizzati fino ad ora.

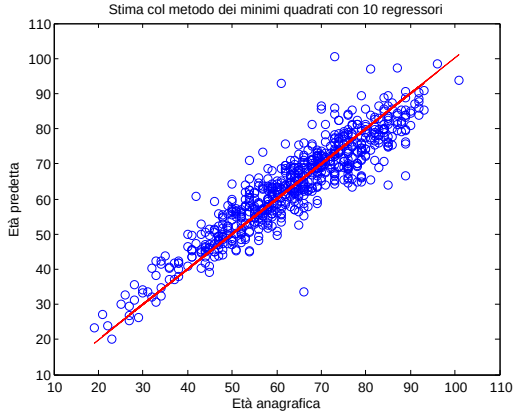
b	RMSE	r	CVerr
10	5.9119	0.9159	36.6650
20	5.5242	0.9270	33.2573
50	4.8523	0.9441	30.1713
100	4.3671	0.9550	28.8038
200	3.7184	0.9676	33.8795
500	2.0228	0.9905	$1.3828 \cdot 10^2$
1000	$2.4986 \cdot 10^{-13}$	1	$1.1169 \cdot 10^3$

Tabella 3.2: Per ogni numero b di marker selezionati vediamo la radice dell'errore quadratico medio della previsione (RMSE), il coefficiente di correlazione (r) di Bravais-Pearson e l'errore di *cross-validation* (CVerr), che indica una stima dell'errore di previsione (o *test error*) su altri campioni.

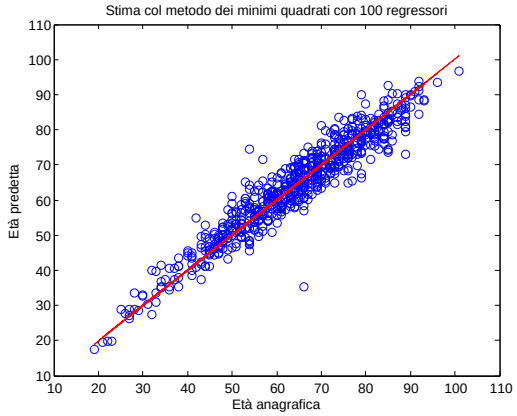
Ridge. Abbiamo applicato il metodo *ridge* con la funzione omonima 'ridge.m' di Matlab e, attraverso la tecnica della *cross-validation* spiegata nel paragrafo precedente, abbiamo trovato i valori ottimali di λ per i $b = 10, 20, 50, 100, 200, 500, 1000$ marker selezionati. In Tabella 3.3 vediamo riassunti i risultati ottenuti e in Figura 3.6 qualche esempio grafico della previsione effettuata con il metodo *ridge*.

Lasso. Abbiamo applicato il metodo *lasso* con la funzione omonima di Matlab, e abbiamo trovato i valori ottimali di λ utilizzando come prima la *cross-validation*. In Tabella 3.4 vediamo riassunti i risultati che abbiamo ottenuto e in Figura 3.7 qualche esempio grafico della previsione effettuata con il metodo *lasso*.

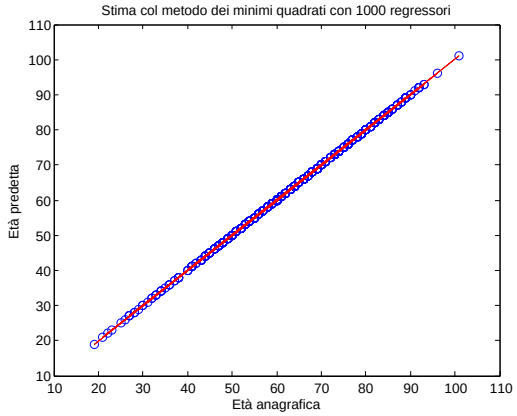
Elastic-net. Abbiamo applicato il metodo *elastic-net* sempre con la funzione 'lasso.m', che permette di selezionare il parametro α grazie all'opzione 'Alpha' $= \alpha$ (di default è uno, cioè implementa *lasso*). I valori ottimali



(a) $b = 10$

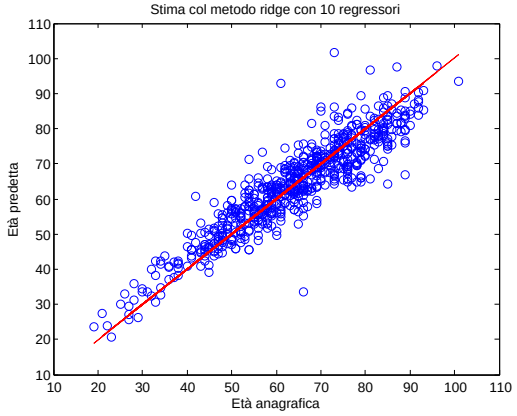


(b) $b = 100$

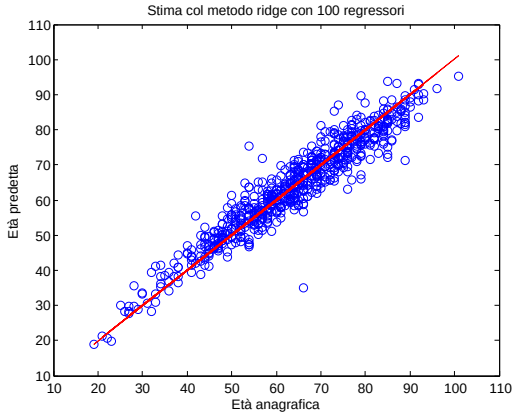


(c) $b = 1000$

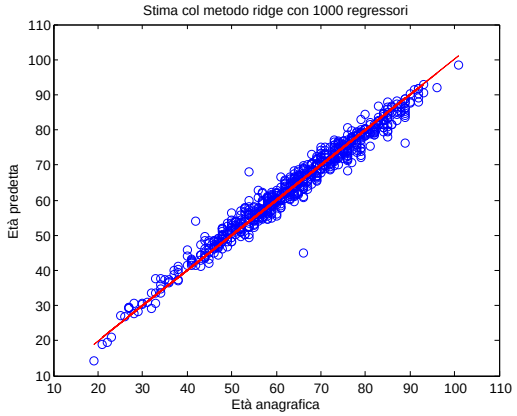
Figura 3.5: Età predette contro età anagrafiche per i 656 pazienti del *dataset* di riferimento, usando il metodo dei minimi quadrati. La bisettrice (in rosso) indica il valore per cui l'età predetta coincide con l'età anagrafica e quindi aiuta visivamente a valutare la differenza tra le due età.



(a) $b = 10$

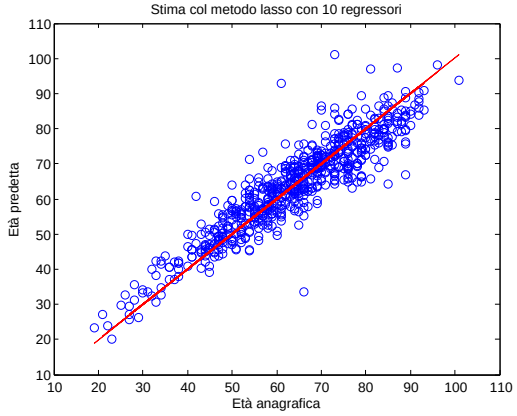


(b) $b = 100$

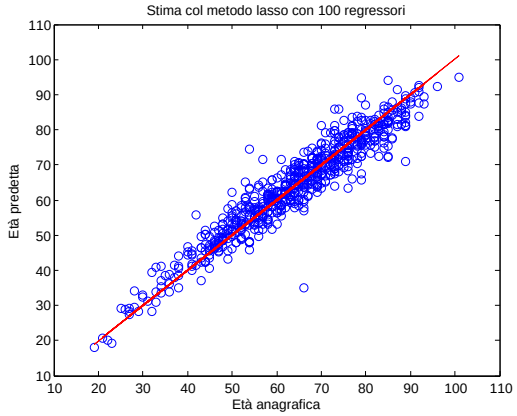


(c) $b = 1000$

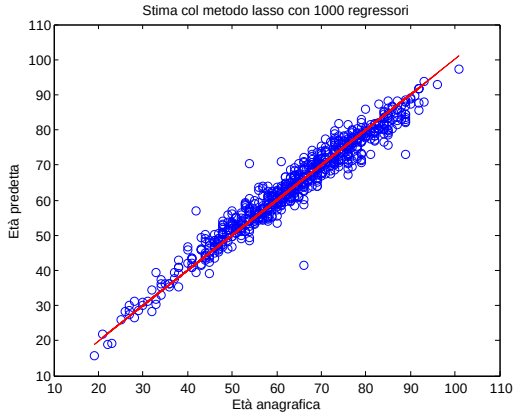
Figura 3.6: Età predette contro età anagrafiche per i 656 pazienti del *dataset* di riferimento, usando il metodo *ridge*. La bisettrice (in rosso) indica il valore per cui l'età predetta coincide con l'età anagrafica e quindi aiuta visivamente a valutare la differenza tra le due età.



(a) $b = 10$



(b) $b = 100$



(c) $b = 1000$

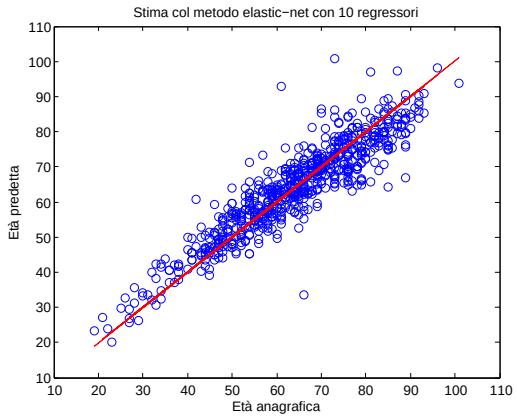
Figura 3.7: Età predette contro età anagrafiche per i 656 pazienti del *dataset* di riferimento, usando il metodo *lasso*. La bisettrice (in rosso) indica il valore per cui l'età predetta coincide con l'età anagrafica e quindi aiuta visivamente a valutare la differenza tra le due età.

b	λ^*	RMSE	r	CVerr
10	15	5.9176	0.9157	36.8899
20	20	5.5330	0.9267	33.0022
50	110	4.9438	0.9421	28.5192
100	100	4.4919	0.9525	26.4784
200	265	4.2222	0.9584	25.7667
500	375	3.4771	0.9722	23.0012
1000	380	2.6081	0.9846	21.4980

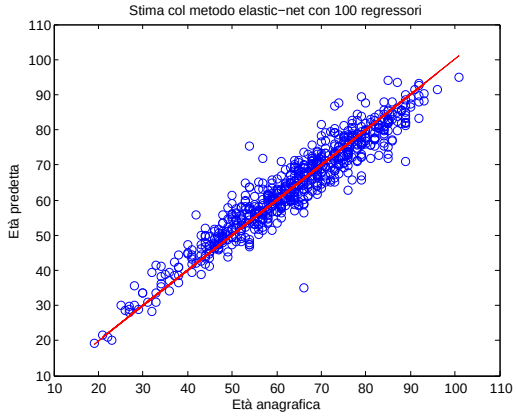
Tabella 3.3: Per ogni numero b di marker selezionati vediamo il valore ottimale λ^* del metodo *ridge* ottenuto via *cross-validation*, la radice dell'errore quadratico medio della previsione (RMSE), il coefficiente di correlazione (r) di Bravais-Pearson e l'errore di *cross-validation* (CVerr), che indica una stima dell'errore di previsione (o *test error*) su altri campioni.

dei parametri α e λ sono stati trovati sempre attraverso la *cross-validation*. In Tabella 3.5 vediamo riassunti i risultati ottenuti e in Figura 3.8 qualche esempio grafico della previsione effettuata con il metodo *elastic-net*.

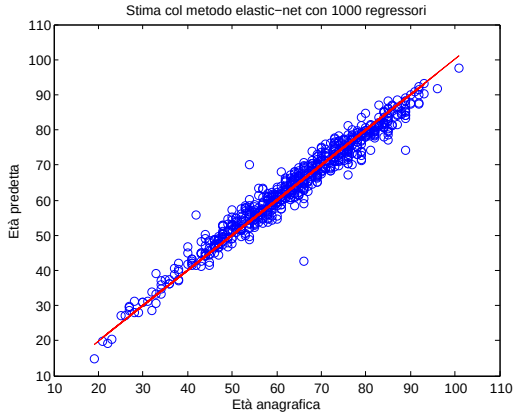
Confronto tra i metodi. Osserviamo subito che il metodo dei minimi quadrati presenta una diminuzione nell'errore di *cross-validation* all'aumentare del numero b di marcatori selezionati solo fino a $b = 200$, mentre per i due valori superiori ($b = 500$ e 1000) tale errore cresce rapidamente. Questo è un esempio del fenomeno dell'*overfitting*, in cui un modello statistico, con troppi parametri liberi rispetto al numero di dati usati per fissarli, si adatta ai dati osservati e dunque peggiora nella previsione delle risposte su dati diversi. Quindi tale metodo è affidabile per il nostro studio solo fino a $b = 200$. Come si nota dal confronto delle rispettive tabelle, i metodi di regressione *ridge*, *lasso* ed *elastic-net* danno risultati molto vicini tra loro e al metodo dei minimi quadrati, il quale per $b \leq 200$ ha sempre minor RMSE rispetto agli altri metodi, ma maggior errore di *cross-validation* (tali differenze si accentuano all'aumentare di b). Invece per $b > 200$ il metodo che presenta minor RMSE ed errore di *cross-validation* è, seppur di poco, il *ridge*. Si noti anche che i metodi *elastic-net* e *lasso* sono pressoché equivalenti fino a $b = 100$, mentre per $b \geq 200$ il primo è significativamente migliore del secondo in termini di



(a) $b = 10$



(b) $b = 100$



(c) $b = 1000$

Figura 3.8: Età predette contro età anagrafiche per i 656 pazienti del *dataset* di riferimento, usando il metodo *elastic-net*. La bisettrice (in rosso) indica il valore per cui l'età predetta coincide con l'età anagrafica e quindi aiuta visivamente a valutare la differenza tra le due età.

b	λ^*	RMSE	r	CVerr
10	0.02	5.9125	0.9159	36.3959
20	0.14	5.5517	0.9263	32.7649
50	0.12	4.9236	0.9425	28.0640
100	0.06	4.4530	0.9532	27.5489
200	0.14	4.3319	0.9560	27.1869
500	0.15	3.8356	0.9658	26.6544
1000	0.15	3.3262	0.9746	24.0421

Tabella 3.4: Per ogni numero b di marker selezionati vediamo il valore ottimale λ^* del metodo *lasso* ottenuto via *cross-validation*, la radice dell'errore quadratico medio della previsione (RMSE), il coefficiente di correlazione (r) di Bravais-Pearson e l'errore di *cross-validation* (CVerr), che indica una stima dell'errore di previsione (o *test error*) su altri campioni.

accuratezza della previsione, situazione ben visibile per $b = 1000$, unico caso in cui abbiamo più predittori che osservazioni.

Oltre alla precisione della stima, siamo interessati anche ad avere una buona interpretazione della relazione esistente tra i valori di metilazione e l'età. Per questo confrontiamo anche il numero di variabili selezionate dai diversi metodi (quelle per cui il coefficiente è diverso da zero): vediamoli in Tabella 3.6. Come già avevamo detto nel capitolo precedente, il metodo *ridge* non effettua selezione di variabili, mentre invece *lasso* ed *elastic-net* riescono a farlo. Si noti che nel metodo dei minimi quadrati nel caso $b = 1000$ la matrice dei livelli di metilazione $\mathbf{x}$ non ha rango pieno. In tal caso la funzione 'regress.m' di Matlab risolve il problema ponendo il massimo numero di coefficienti possibili a zero[5], ottenendo almeno r coefficienti non nulli, con r rango della matrice $\mathbf{x}$.

[5]Si veda la documentazione della funzione 'regress' di Matlab per ulteriori dettagli.

b	α^*	λ^*	RMSE	r	CVerr
10	0.9	0.01	5.9122	0.9159	36.9827
20	0.9	0.13	5.5513	0.9263	33.0272
50	0.8	0.13	4.9233	0.9425	28.1254
100	0.1	0.18	4.5191	0.9519	26.9789
200	0.1	0.20	4.0533	0.9616	26.0965
500	0.1	0.50	3.6651	0.9691	24.2508
1000	0.1	0.50	2.9813	0.9798	21.0943

Tabella 3.5: Per ogni numero b di marker selezionati vediamo i valori ottimali α^* e λ^* del metodo *elastic-net* ottenuti via *cross-validation*, la radice dell'errore quadratico medio della previsione (RMSE), il coefficiente di correlazione (r) di Bravais-Pearson e l'errore di *cross-validation* (CVerr), che indica una stima dell'errore di previsione (o *test error*) su altri campioni.

b	Minimi quadrati	Ridge	Lasso	Elastic-net
10	10	10	9	9
20	20	20	15	15
50	50	50	32	33
100	100	100	77	86
200	200	200	90	183
500	500	500	126	346
1000	656	1000	187	592

Tabella 3.6: Per ogni numero b di marker utilizzati vediamo il numero di coefficienti (e quindi di marcatori) selezionati dai diversi metodi.

Capitolo 4

Nuovi metodi non statistici di previsione dell'età

Oltre agli esperimenti descritti in precedenza, abbiamo guardato in una maniera diversa i dati a nostra disposizione. Nel prossimo paragrafo illustriamo un nostro nuovo metodo per predire l'età, solo in parte statistico, e lo confronteremo con i precedenti. Nel secondo paragrafo invece vedremo quali informazioni può darci la trasformata di Fourier sul profilo di metilazione dei pazienti.

4.1 Nuova funzione di previsione

Nel capitolo precedente si è visto come abbiamo effettuato una preselezione dei b marcatori 'migliori' secondo la più alta correlazione con l'età. Sulla base della classifica ottenuta, in seguito abbiamo associato ad ogni *marker* due tipi di peso, proporzionali alla loro posizione in questa graduatoria. Nel primo caso abbiamo utilizzato per il marcatore j-esimo il peso

$$\frac{\left(\frac{1}{j}\right)^{\gamma}}{S_1} \quad \text{dove} \quad S_1 = \sum_{i=1}^{b} \left(\frac{1}{i}\right)^{\gamma} \tag{4.1}$$

con $j = 1, \ldots, b$ e $\gamma > 0$; nel secondo invece

$$\frac{q^j}{S_2} \quad \text{dove} \quad S_2 = \sum_{i=1}^{b} q^i \tag{4.2}$$

con $q > 0$ e diverso da 1. Scegliendo uno dei due pesi, la nostra funzione di previsione dell'età per ogni paziente è così costruita: invertendo la formula (3.2) della regressione lineare fra età e metilazione di un marker, si calcola una previsione dell'età $\tilde{E}_{ij}$ del paziente i-esimo dovuta al singolo j-esimo marcatore

$$\tilde{E}_{ij} = \frac{M_{ij} - I_j}{P_j} \, .$$

Infine si calcola la media pesata degli $\tilde{E}_{ij}$ su tutti i *marker*, secondo uno dei due pesi scelti in precedenza, ottenendo la previsione $\tilde{E}_i$ dell'età del paziente i-esimo.

Si noti che in tale metodo i parametri I_j e P_j sono fissati dall'inizio[1], i pesi dipendono solo dalla classifica dei marcatori e per questo non ha senso allenare il nostro metodo su qualche *training data*. Abbiamo dunque testato questa funzione con $b = 10, 20, 50, 100, 200, 500, 1000$ marcatori sui 656 pazienti, trovando i valori ottimali dei parametri γ del peso polinomiale (4.1) e q del peso esponenziale (4.2) che minimizzano l'errore RMSE. I risultati per i due tipi di peso si possono vedere in Tabella 4.1.

Notiamo subito che per $b \leq 50$ entrambi i metodi presentano una diminuzione dell'errore RMSE e di *cross-validation* e un aumento del coefficiente di correlazione, mentre per $b \geq 100$ si comportano diversamente: con il primo peso si ha un lieve aumento del RMSE e del CVerror e una sensibile diminuzione del coefficiente di correlazione, mentre con il secondo peso non si ha nessun aumento o diminuzione al crescere del numero di *marker* considerati. Ciò è dovuto alla natura dei pesi (Figura 4.1): entrambi diminuiscono rapidamente per $b \leq 50$ (il primo più velocemente del secondo), invece per

[1] In teoria questi parametri vengono dai dati e quindi il metodo dovrebbe essere 'allenato' per trovare la stima migliore di essi, ma consideriamo per semplicità quei due parametri come molto stabili.

(a) Peso polinomiale ('pol')

b	γ^*	RMSE	r	CVerror
10	0.35	7.2284	0.8977	52.2058
20	0.43	6.6391	0.9116	44.1217
50	0.64	6.2213	0.9212	38.7174
100	0.98	6.3859	0.9174	40.7451
200	1.11	6.4449	0.9161	41.5341
500	1.23	6.5377	0.9140	42.7757
1000	1.29	6.5973	0.9126	43.5355

(b) Peso esponenziale ('exp')

b	q^*	RMSE	r	CVerror
10	0.98	7.2483	0.8972	52.5052
20	0.95	6.6461	0.9115	44.1917
50	0.95	6.2742	0.9200	39.3501
100	0.94	6.3111	0.9191	39.8541
200	0.94	6.3130	0.9191	39.8623
500	0.94	6.3130	0.9191	39.8424
1000	0.94	6.3130	0.9191	39.8654

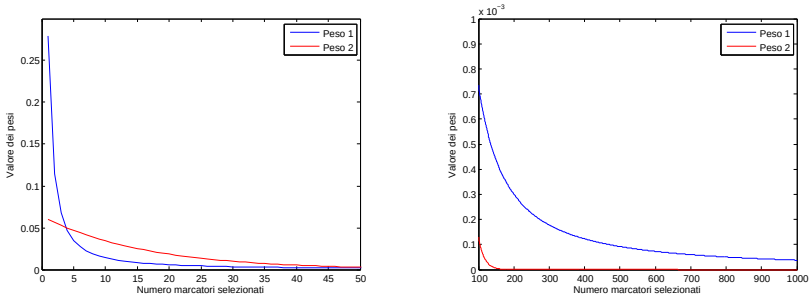
Tabella 4.1: Per ogni numero b di marker selezionati vediamo parametri ottimali γ^* e q^* per i due metodi, la radice dell'errore quadratico medio (RMSE), il coefficiente di correlazione lineare (r) e l'errore di *cross-validation* (CVerror).

$b \geq 100$ il secondo ha una convergenza a zero più veloce del primo e per questo esso produce risultati costanti mentre l'altro cambia seppur di poco.

Dal confronto con i risultati del capitolo precedente possiamo dire che i nostri metodi non si discostano di molto dagli altri fino a $b = 50$, invece per $b \geq 100$ le loro previsioni non sono altrettanto accurate. Vediamo infatti in Figura 4.2 le età predette con i due differenti pesi per alcuni valori di b.

4.2 Analisi spettrale di Fourier

Tutti i metodi descritti fino a qui sono serviti per indagare sul legame tra i valori di metilazione di un numero più o meno grande di *marker* e



(a) Da 1 a 50 (b) Da 100 a 1000

Figura 4.1: Andamento del valore dei due pesi al variare del numero di *marker* considerati.

l'età di un paziente. Oltre a questo approccio, abbiamo voluto analizzare anche globalmente i dati a nostra disposizione, considerando l'intero profilo di metilazione di ognuno dei 656 individui. Per questo i livelli di metilazione sono stati divisi in $C = 500$ classi

$$[0, 0.002[\quad [0.002, 0.004[\quad \dots \quad [0.998, 1]$$

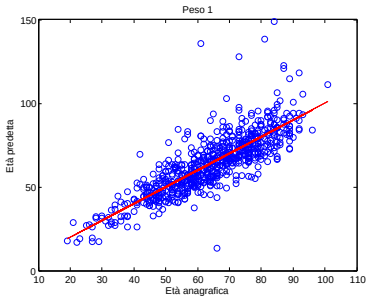
e abbiamo osservato la distribuzione empirica dei valori di metilazione per ogni paziente (Figura 4.3).

A questo punto, per estrarre delle informazioni da queste distribuzioni e cercare eventuali correlazioni significative con l'età, abbiamo deciso di utilizzare la trasformata di Fourier.

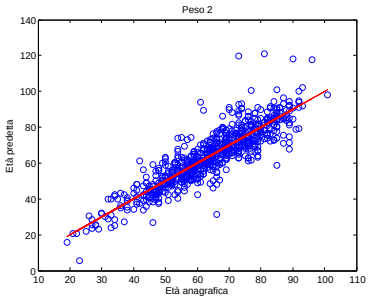
Definizione 4.1 (Coefficienti di Fourier)**.** Consideriamo il caso particolare di f, funzione di densità di probabilità definita in [0,1]. Per $k \in \mathbb{Z}$, si definisce k-esimo coefficiente di Fourier di f la seguente quantità

$$\hat{f}(k) = \int_0^1 e^{-2\pi i k x} f(x) dx. \tag{4.3}$$

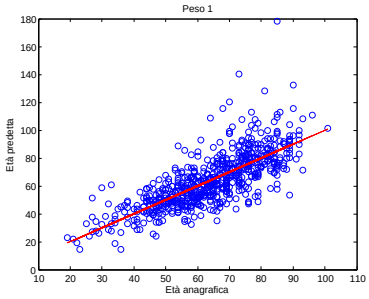
Osservazione 4.1. La successione dei coefficienti di Fourier in questo caso



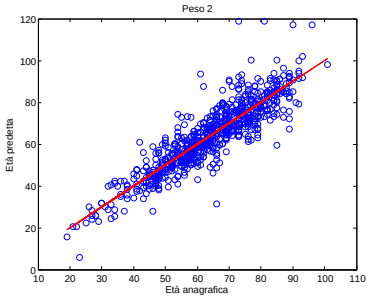
(a) Peso pol, $b = 10$



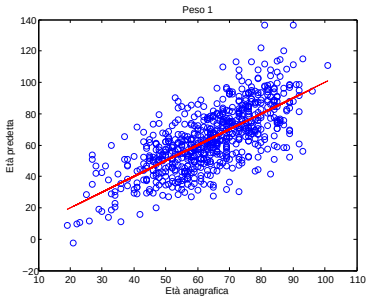
(b) Peso exp, $b = 10$



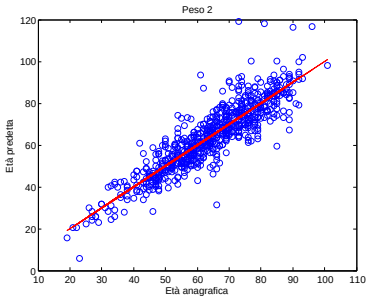
(c) Peso pol, $b = 50$



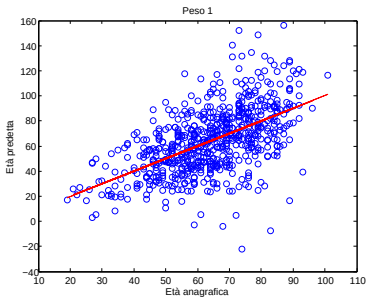
(d) Peso exp, $b = 50$



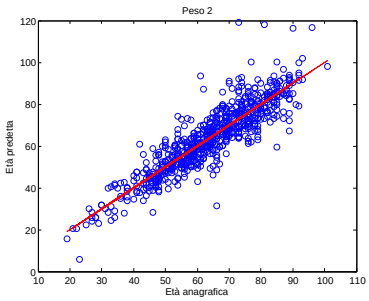
(e) Peso pol, $b = 200$



(f) Peso exp, $b = 200$



(g) Peso pol, $b = 1000$



(h) Peso exp, $b = 1000$

Figura 4.2: Età predette contro età anagrafiche per i 656 pazienti del *dataset* di riferimento, usando la nostra funzione di previsione con i due differenti pesi. La bisettrice (in rosso) indica il valore per cui l'età predetta coincide con l'età anagrafica e quindi aiuta visivamente a valutare la differenza tra le due età.

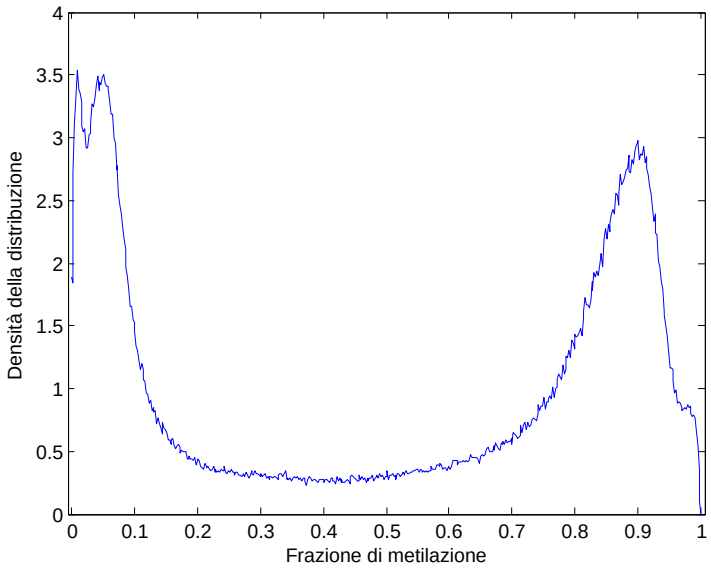


Figura 4.3: Esempio di distribuzione dei valori di metilazione per il primo paziente del *dataset* di riferimento.

coincide[2] con quella dei valori della trasformata di Fourier su $\mathbb{Z}$. Per brevità, nel seguito la (4.3) verrà chiamata semplicemente trasformata di Fourier.

Osservazione 4.2. Fissato $k \in \mathbb{Z}$, dalla definizione (4.3), poiché per la formula di Eulero

$$e^{-2\pi ikx} = \cos(2\pi kx) - i\sin(2\pi kx),$$

si vede che $\hat{f}(k) \in \mathbb{C}$ e dunque possiamo dividere la trasformata di Fourier di f nelle sue parti reale $\mathrm{Re}(\hat{f})$ ed immaginaria $\mathrm{Im}(\hat{f})$. Notiamo anche che $\mathrm{Re}(\hat{f})$ è una funzione pari, mentre $\mathrm{Im}(\hat{f})$ invece è dispari, dunque è sufficiente studiare la due parti della trasformata di Fourier per $k \geq 0$. Inoltre, poiché f è densità di probabilità in [0,1] e $\forall k \in \mathbb{Z}$ si ha $|\hat{f}(k)| \leq 1$, si vede subito dalla definizione (4.3) che per $k = 0$ vale

$$\hat{f}(0) = \int_0^1 f(x)dx = 1 \tag{4.4}$$

[2]Si confronti con [28].

Osservazione 4.3. Nel nostro caso di studio, per ogni pazienti in esame, partiamo da un insieme discreto di valori di metilazione, appartenenti all'intervallo $[0,1]$, e decidiamo di dividerli in $C = 500$ raggruppamenti, o *bin*, di ampiezza $h = 1/C$, che indichiamo con $J_n = [nh, (n+1)h]$, per $n = 0, \ldots, C-1$. Per ogni bin, sia ϕ_n la percentuale di marker la cui metilazione è compresa in J_n, dunque

$$\sum_{n=0}^{C-1} \phi_n = 1\,.$$

Decidiamo di trattare tale distribuzione empirica come una distribuzione su $[0,1]$ fatta a gradini attraverso la seguente

$$f(x) = \sum_{n=0}^{C-1} \frac{1}{h} \phi_n \mathbf{1}_{J_n}(x)\,, \tag{4.5}$$

con $\mathbf{1}_{J_n}$ funzione indicatrice dell'intervallo J_n e $x \in [0,1]$. Si noti che, scegliendo di inserire nella (4.5) la quantità $1/h$, allora la (4.4) è banalmente verificata. A questo punto, riscriviamo la (4.3) con la (4.5): fissato $k \in \mathbb{Z}$ e diverso da zero, abbiamo che

$$\begin{aligned}
\hat{f}(k) &= \int_0^1 e^{-2\pi i k x} f(x) dx \\
&= \sum_{n=0}^{C-1} \frac{1}{h} \phi_n \int_{nh}^{(n+1)h} e^{-2\pi i k x} dx \\
&= \sum_{n=0}^{C-1} \frac{1}{h} \phi_n \left[\frac{e^{-2\pi i k x}}{-2\pi i k} \right]_{nh}^{(n+1)h} \\
&= \sum_{n=0}^{C-1} \frac{1}{h} \phi_n \frac{e^{-2\pi i k (n+1) h} - e^{-2\pi i k n h}}{-2\pi i k}
\end{aligned}$$

e quindi

$$\hat{f}(k) = \sum_{n=0}^{C-1} \phi_n \frac{e^{-2\pi i k h} - 1}{-2\pi i k h}\, e^{-2\pi i k n h}\,. \tag{4.6}$$

Dunque è ben definita la seguente

$$\hat{f}(k) = \begin{cases} \displaystyle\sum_{n=0}^{C-1} \phi_n \frac{e^{-2\pi ikh} - 1}{-2\pi ikh} \, e^{-2\pi iknh} & k \neq 0 \\ 1 & k = 0 \end{cases} \tag{4.7}$$

che utilizzeremo per calcolare *esattamente* la trasformata di Fourier della funzione a gradini (4.5).

Proposizione 4.2.1. *Per ogni $k \in \mathbb{Z}$ vale*

$$\hat{f}(k + C) = \frac{k}{k + C} \hat{f}(k)$$

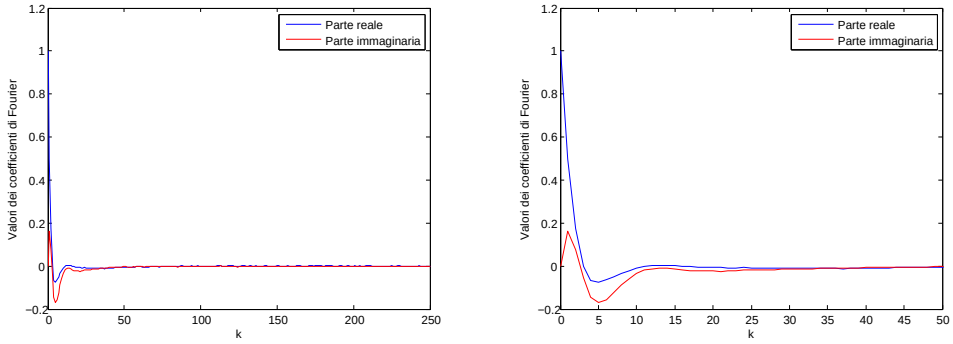
Dimostrazione. Se $k = 0$ allora, poiché $h = 1/C$, si ha

$$\begin{aligned} \hat{f}(C) &= \sum_{n=0}^{C-1} \phi_n \frac{e^{-2\pi iCh} - 1}{-2\pi iCh} \, e^{-2\pi iCnh} \\ &= \sum_{n=0}^{C-1} \phi_n \frac{e^{-2\pi i} - 1}{-2\pi i} \, e^{-2\pi in} \\ &= 0 \end{aligned}$$

dato che $e^{-2\pi i} = 1$. Sia ora $k \neq 0$, allora

$$\begin{aligned} \hat{f}(k + C) &= \sum_{n=0}^{C-1} \phi_n \frac{e^{-2\pi i(k+C)h} - 1}{-2\pi i(k + C)h} \, e^{-2\pi i(k+C)nh} \\ &= \sum_{n=0}^{C-1} \phi_n \frac{e^{-2\pi ikh} e^{-2\pi iCh} - 1}{-2\pi i(k + C)h} \, e^{-2\pi iknh} e^{-2\pi iCnh} \\ &= \sum_{n=0}^{C-1} \phi_n \frac{e^{-2\pi ikh} - 1}{-2\pi i(k + C)h} \, e^{-2\pi iknh} \\ &= \frac{k}{k + C} \hat{f}(k) \, . \end{aligned}$$

□

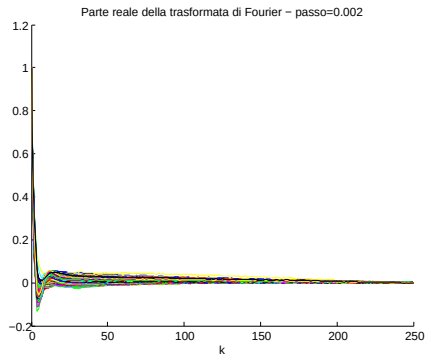


(a) $0 \leq k \leq 250$ (b) $0 \leq k \leq 50$

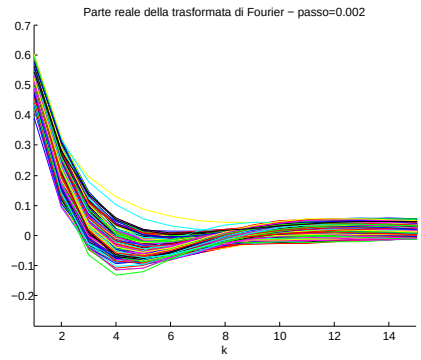
Figura 4.4: Per ogni k, il punto in ordinata dà il valore del corrispondente coefficiente di Fourier (parte reale in blu, parte immaginaria in rosso) di un paziente di 67 anni appartenente al *dataset* di riferimento.

Osservazione 4.4. Per la Proposizione 4.2.1 e le osservazioni precedenti sulla parità della parte reale ed immaginaria di $\hat{f}$, possiamo dunque limitarci a valutare $\hat{f}$ per $k \in [0, C/2]$.

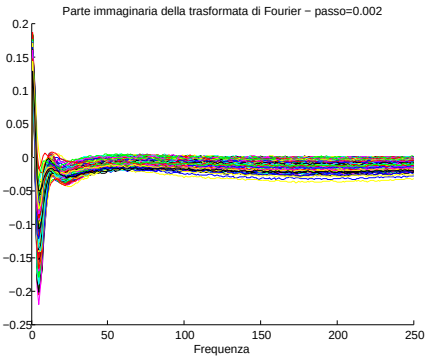
La nostra sperimentazione è dunque continuata calcolando la parte reale ed immaginaria della trasformata di Fourier $\hat{f}$ della distribuzione dei valori di metilazione per ognuno dei 656 individui del *dataset* di riferimento (in Figura 4.4 vediamo un esempio per il primo di questi pazienti). Osservando complessivamente tutte le trasformate reali ed immaginarie (Figura 4.5), si può notare che la variabilità maggiore delle trasformate si ha per $k \in [0, 15]$: ciò si vede chiaramente dalla (Figura 4.6), in cui abbiamo graficato, per ogni k, la differenza tra il massimo e il minimo del corrispondente coefficiente di Fourier su tutti i 656 pazienti. Non abbiamo però osservato correlazioni significative tra età e valori delle trasformate a k fissato. Vedremo nel prossimo capitolo se, analizzando i dati di metilazione di soggetti affetti o meno da specifiche anomalie genetiche, potremo trovare correlazioni sufficientemente rilevanti e che permettano di evidenziare differenze tra tali gruppi di individui.



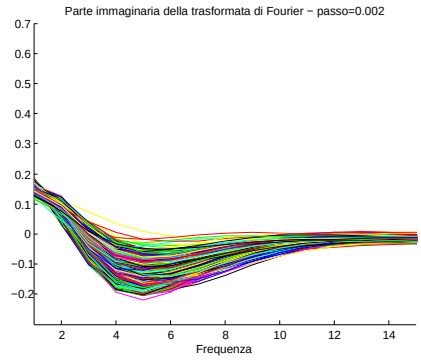
(a) Parte reale, per $0 \leq k \leq 250$



(b) Parte reale, per $1 \leq k \leq 15$

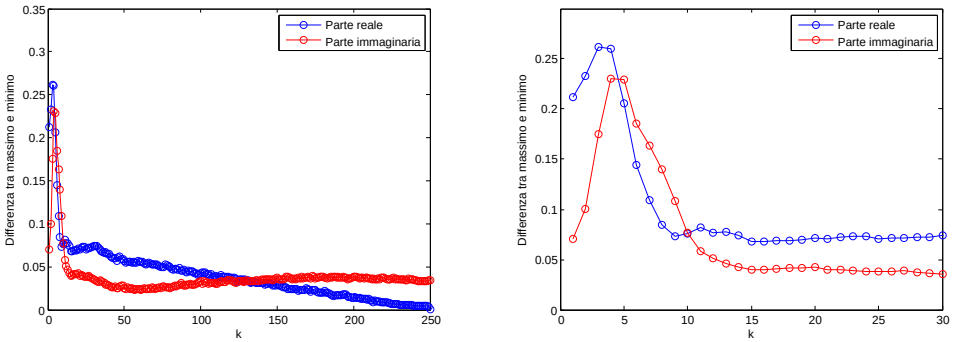


(c) Parte immaginaria, per $0 \leq k \leq 250$



(d) Parte immaginaria, per $1 \leq k \leq 15$

Figura 4.5: Trasformate di Fourier per i 656 pazienti del *dataset* di riferimento.



(a) $1 \leq k \leq 250$ (b) $1 \leq k \leq 30$

Figura 4.6: Per ogni k, il punto in ordinata dà la differenza tra massimo e minimo del corrispondente coefficiente di Fourier su tutti i 656 pazienti del *dataset* di riferimento.

Capitolo 5

Caso di studio: pazienti affetti da sindrome di Down

Concludiamo la nostra esposizione analizzando i dati di metilazione di alcuni pazienti affetti da sindrome di Down[1], assieme ai dati dei loro fratelli e delle loro madri. Nel primo paragrafo utilizzeremo i risultati dei metodi minimi quadrati, *ridge*, *lasso*, *elastic-net* e della nostra funzione di previsione descritti nei paragrafi 3.2.2 e 4.1 per prevedere le età di questo nuovo gruppo. Infine, grazie alla trasformata di Fourier, andremo a considerare globalmente i profili di metilazione dei pazienti.

5.1 Età anagrafica o età biologica?

Nei capitoli precedenti si è visto come i diversi metodi utilizzati stimassero, più o meno accuratamente, l'età anagrafica. Ci chiediamo però se, in presenza di qualche anomalia genetica, la stima fornita dai nostri modelli sia

[1]La sindrome di Down è una malattia genetica causata dalla presenza di un cromosoma 21 (o parte di esso) in più nel genotipo di tutte le cellule (o, più raramente, di una parte, in tal caso si parla di mosaicismo), da qui la definizione di trisomia 21 come sinonimo della sindrome stessa. Le conseguenze mediche provocate dal materiale genetico in soprannumero sono molto variabili e possono influenzare la funzione di qualsiasi organo dell'organismo, portando così il soggetto ad avere generalmente una minore aspettativa di vita.

relativa piuttosto all'età biologica dell'individuo, intesa come quella età che approssimativamente è attribuibile a una persona, indipendentemente dalla sua età anagrafica, sulla base di una globale valutazione dell'efficienza fisica e mentale.

Consideriamo i dati di metilazione di $p = 485577$ marcatori provenienti da $N = 96$ individui[2]: di questi, 32 sono affetti da sindrome di Down (D), mentre i rimanenti pazienti sani sono divisi equamente tra i loro fratelli (F) e le loro madri (M). La matrice a disposizione è dunque di dimensione 485577×96. All'interno di essa ci sono dei valori mancanti, che in lettura abbiamo registrato come *Nan* (*Not a Number*). Le funzioni 'regress.m', 'ridge.m' e 'lasso.m' di Matlab eliminano automaticamente i campioni in cui sono presenti valori mancanti, ma per valori crescenti di b marcatori selezionati, il numero di pazienti da rimuovere dal *dataset* diventa eccessivo (Tabella 5.1) e in tali casi non sarebbe possibile effettuare un'analisi completa dei dati.

b	NP	NaN
10	0	0
20	0	0
50	0	0
100	3	3
200	8	8
500	27	30
1000	45	59

Tabella 5.1: Per ogni numero b di marker selezionati vediamo il numero di pazienti da rimuovere (NP) dal *dataset 'Down e familiari'*, perché hanno almeno un valore mancante, e il numero complessivo di dati assenti (NaN).

Classicamente ci sono diversi metodi per aggirare questo problema: uno dei più adottati è la tecnica della *K Nearest-Neighbors imputation* (KNN), che stima i valori sostitutivi dei dati mancanti da quelli dei K 'più vicini', calcolati attraverso una certa distanza (di solito quella euclidea). La procedura è la seguente:

[2]Questo nuovo *dataset*, a cui ci riferiremo in seguito chiamandolo anche *'Down e familiari'*, proviene dal lavoro [29] in attesa di pubblicazione.

1. Detto x^* il regressore N-dimensionale in cui ci sono uno o più dati assenti da stimare, si calcola la distanza (euclidea) tra x^* e i regressori che non hanno valori mancanti (regressori completi), usando solo le coordinate che non sono mancanti in x^*. Ad esempio, se x^* ha un valore mancante nella componente j-esima, si escluderà tale componente dal calcolo delle distanze tra i regressori completi e x^*.

2. Si individuano i K regressori con distanza minore da x^*.

3. Si sostituisce ai NaN in x^* la media pesata delle corrispondenti coordinate dei primi K regressori più vicini, dove i pesi sono inversamente proporzionali alle distanze.

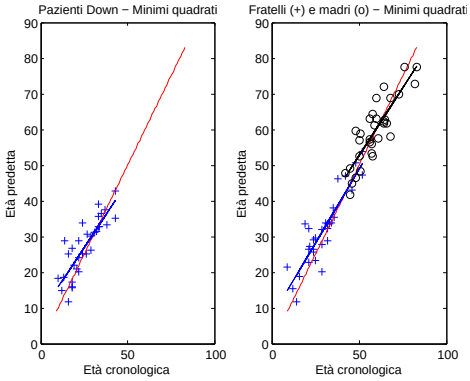
Abbiamo applicato tale tecnica (con $K = 10$) prima di utilizzare i diversi metodi di previsione dell'età: in Tabella 5.2 e Tabella 5.3 vediamo i risultati ottenuti e in Figura 5.1 alcuni esempi grafici delle età predette[3].

Da tali figure si nota che molti dei metodi utilizzati restituiscono un'età predetta per il gruppo D più alta della reale, sia in termini assoluti che se raffrontata ai fratelli (che hanno età nella stessa finestra di valori). Questo è significativo perché è noto che i pazienti affetti da sindrome di Down hanno un'aspettativa di vita più breve della norma[4] e quindi sarebbe interessante se queste funzioni di previsione fornissero effettivamente una stima dell'età biologica di tali individui piuttosto che di quella anagrafica.

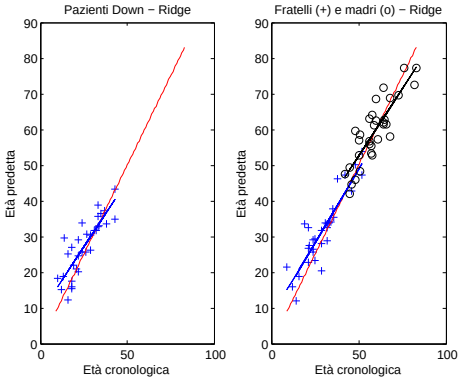
Consideriamo infine le variabili '*differenza delle età predette e anagrafiche per i pazienti affetti da sindrome di Down*' (DD) e l'analoga per i fratelli (DF). Per contraddire, da un punto di vista statistico, l'ipotesi nulla che tali variabili provengono dalla stessa distribuzione, abbiamo effettuato i seguenti

[3]La nostra funzione di previsione è stata adoperata anche non avendo prima applicato la *KNN imputation* alla matrice dei dati di metilazione, ma tali risultati sono stati omessi perché ritenuti ridondanti, in quanto non abbiamo osservato differenze significative dai risultati ottenuti col metodo precedente.

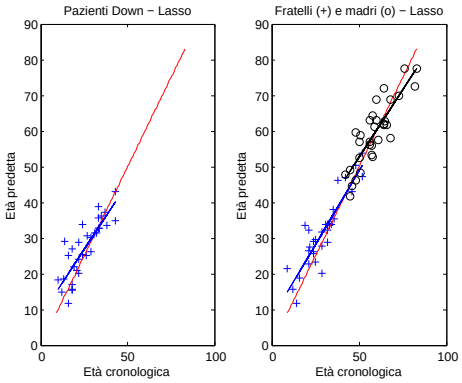
[4]A seguito dei miglioramenti nelle cure mediche, in particolare nei problemi cardiaci, l'aspettativa di vita tra le persone con sindrome di Down è aumentata dai 12 anni nel 1949, ai 60 anni dei nostri giorni, rimanendo comunque al di sotto della speranza di vita media [30].



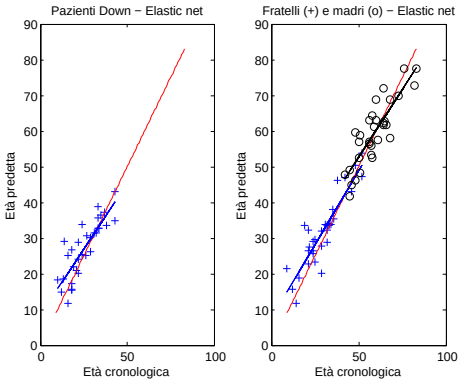
(a) Minimi quadrati, $b = 10$



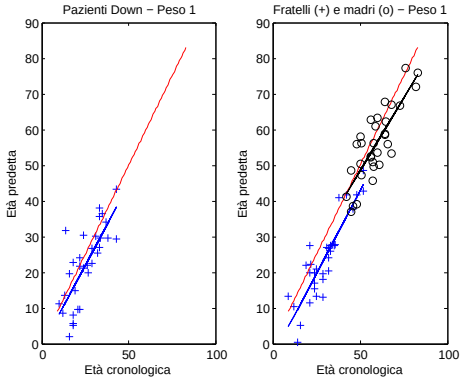
(b) Ridge, $b = 10$



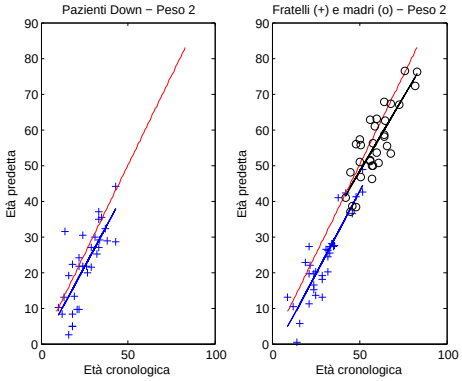
(c) Lasso, $b = 10$



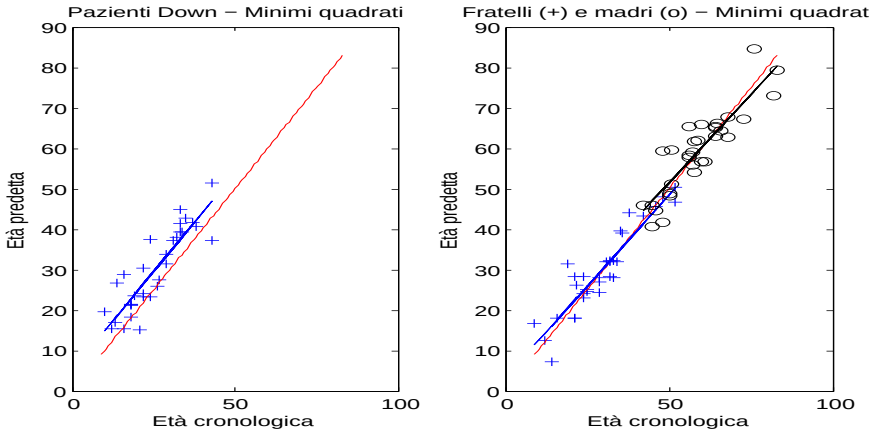
(d) Elastic-net, $b = 10$



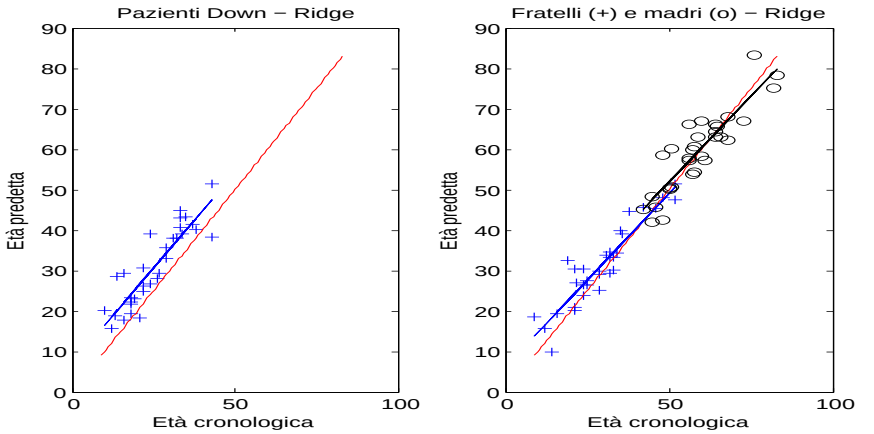
(e) Peso pol, $b = 10$



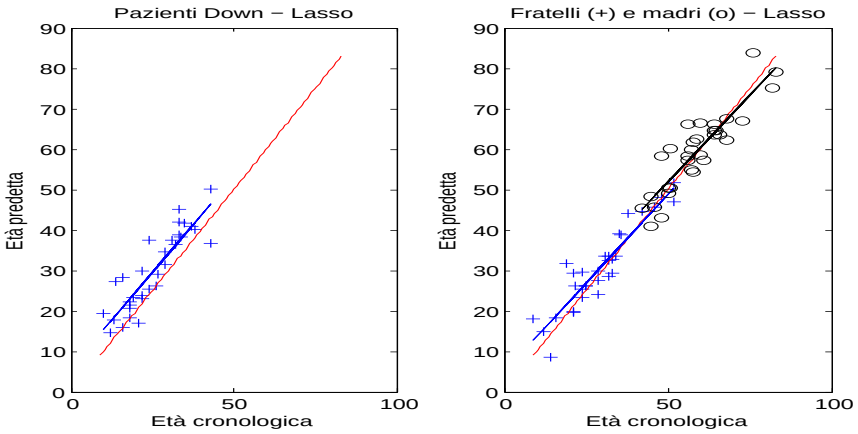
(f) Peso exp, $b = 10$



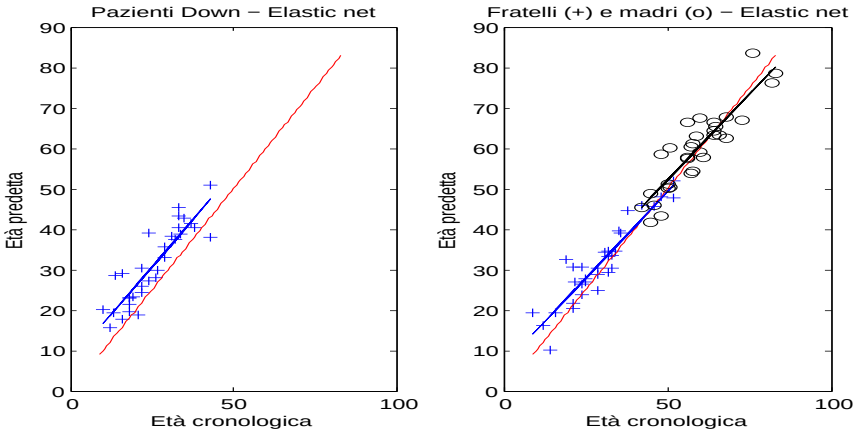
(g) Minimi quadrati, $b = 100$



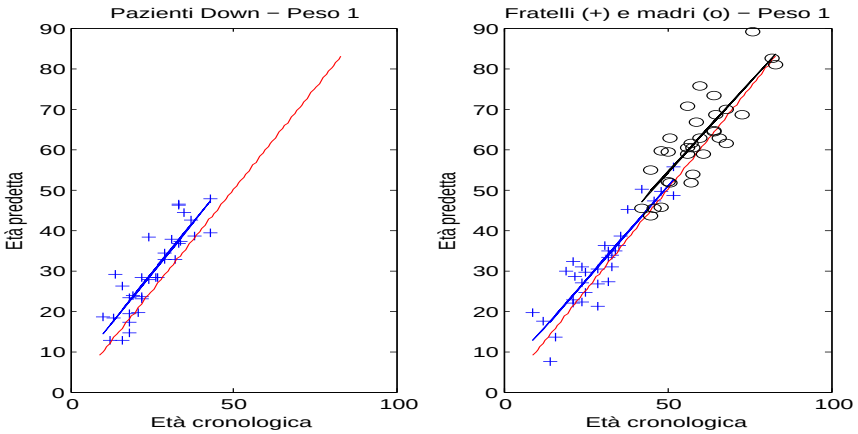
(h) Ridge, $b = 100$



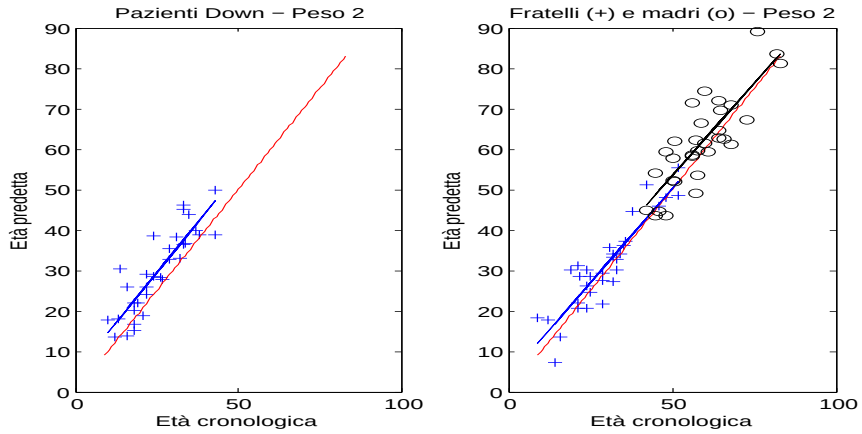
(i) Lasso, $b = 100$



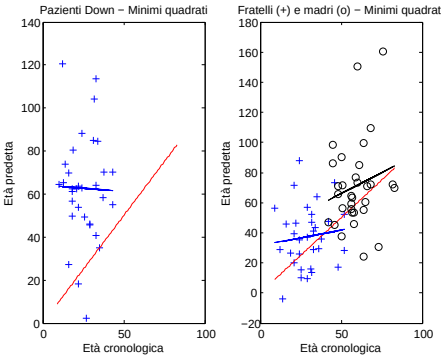
(j) Elastic-net, $b = 100$



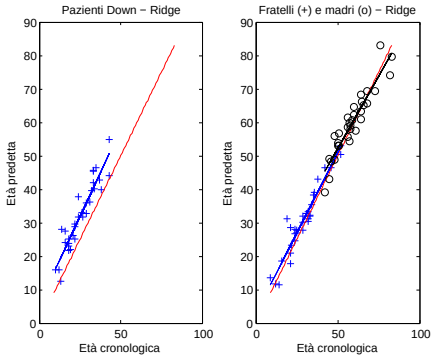
(k) Peso pol, $b = 100$



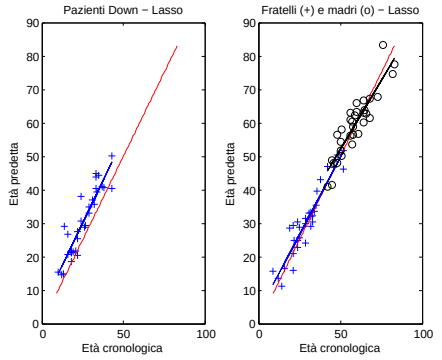
(l) Peso exp, $b = 100$



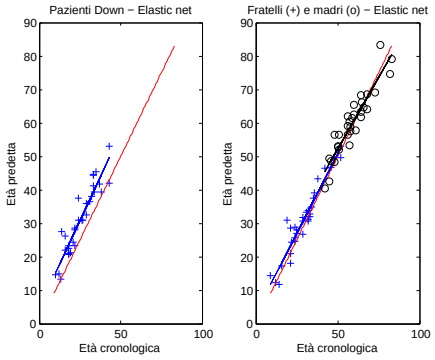
(m) Minimi quadrati, $b = 1000$



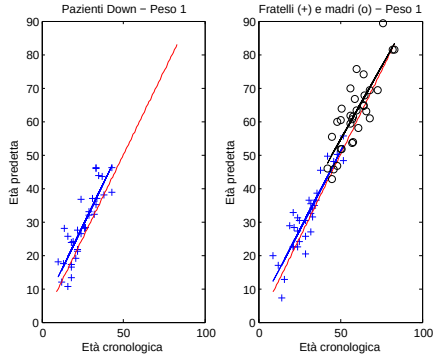
(n) Ridge, $b = 1000$



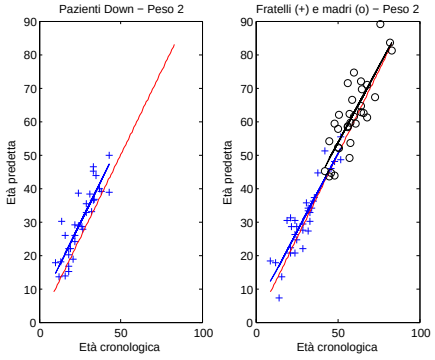
(o) Lasso, $b = 1000$



(p) Elastic-net, $b = 1000$



(q) Peso pol, $b = 1000$



(r) Peso exp, $b = 1000$

Figura 5.1: Età predette contro età anagrafiche dei pazienti del *dataset* '*Down e familiari*', effettuate con diversi metodi di previsione. La bisettrice (in rosso) indica il valore per cui l'età predetta coincide con l'età anagrafica e quindi aiuta visivamente a valutare la differenza tra le due età. Invece i segmenti di retta più corti (in blu per i pazienti Down e i loro fratelli e in nero per le loro madri) rappresentano le rette che meglio si adattano ai valori dell'età predette per ogni gruppo, descrivendone così l'andamento lineare e permettendo il confronto con la bisettrice.

(a) $b = 10$

Metodi	RMSE (D)	r (D)	RMSE (F)	r (F)	RMSE (M)	r (M)
Minimi Q.	5.1156	0.8450	5.2918	0.8969	5.2448	0.8598
Ridge	5.2203	0.8417	5.3039	0.8970	5.2239	0.8602
Lasso	5.1304	0.8432	5.2735	0.8975	5.2513	0.8592
Elnet	5.1298	0.8439	5.2901	0.8971	5.2483	0.8595
Peso pol	7.5983	0.7627	7.7720	0.8852	6.7444	0.8276
Peso exp	7.6969	0.7642	7.9305	0.8824	6.7533	0.8344

(b) $b = 20$

Metodi	RMSE (D)	r (D)	RMSE (F)	r (F)	RMSE (M)	r (M)
Minimi Q.	5.7005	0.8372	4.9657	0.9124	5.4365	0.8544
Ridge	5.9927	0.8417	5.1369	0.9136	5.4602	0.8540
Lasso	5.9368	0.8290	5.0885	0.9111	5.4321	0.8506
El.-net	6.0181	0.8324	5.1349	0.9118	5.4253	0.8518
Peso pol	6.3099	0.8216	4.8577	0.9079	6.1928	0.8204
Peso exp	6.2054	0.8303	4.7902	0.9086	6.1442	0.8254

(c) $b = 50$

Metodi	RMSE (D)	r (D)	RMSE (F)	r (F)	RMSE (M)	r (M)
Minimi Q.	5.0353	0.8759	4.2562	0.9283	4.4204	0.9039
Ridge	6.2329	0.8744	4.9145	0.9257	4.5329	0.8988
Lasso	5.7857	0.8592	4.6555	0.9216	4.5935	0.8950
El.-net	5.9430	0.8610	4.7260	0.9219	4.6166	0.8944
Peso pol	7.1116	0.8597	4.8339	0.9098	6.5106	0.8414
Peso exp	6.9501	0.8647	4.9006	0.9076	6.6482	0.8390

(d) $b = 100$

Metodi	RMSE (D)	r (D)	RMSE (F)	r (F)	RMSE (M)	r (M)
Minimi Q.	6.5626	0.8783	4.1070	0.9254	4.7430	0.8906
Ridge	7.1530	0.8870	4.4781	0.9263	4.6384	0.8943
Lasso	6.3470	0.8846	4.1972	0.9276	4.6142	0.8960
El.-net	7.1312	0.8864	4.6017	0.9259	4.6617	0.8949
Peso pol	6.6297	0.8732	5.1778	0.9053	6.7822	0.8386
Peso exp	6.6442	0.8713	4.9410	0.9070	6.6779	0.8389

Tabella 5.2: Per ogni metodo di previsione utilizzato vediamo, per $b = 10, 20, 50$ e 100 marcatori considerati, i valori della radice dell'errore quadratico medio (RMSE) e il relativo coefficiente di correlazione lineare (r) per i gruppi D, F ed M del *dataset 'Down e familiari'*.

test: dapprima valutiamo l'ipotesi nulla che le varianze delle due variabili siano diverse attraverso il test F del Paragrafo 2.2.3. Come si vede dalla Tabella 5.4, per tutte le *signature* considerate non è possibile rifiutare tale ipotesi. Supponendo allora che DD e DF si distribuiscano normalmente con uguali varianze, applichiamo il test t del Paragrafo 2.2.2 per valutare l'ipotesi nulla che le due variabili abbiano la stessa media: dalla Tabella 5.5 notiamo che per $b = 10, 20, 50$ non possiamo rifiutare tale ipotesi a livello 0.01 (solo con i nostri nuovi metodi per $b = 20, 50$ potremmo rifiutare l'ipotesi nulla a livello 0.05), mentre per $b \geq 100$ le due medie sono statisticamente diverse e dunque possiamo pensare che le differenze riscontrate tra tali variabili siano

(a) $b = 200$

Metodi	RMSE (D)	r (D)	RMSE (F)	r (F)	RMSE (M)	r (M)
Minimi Q.	8.5000	0.9204	4.0944	0.9323	4.5121	0.9058
Ridge	8.2060	0.9178	4.3567	0.9421	4.0897	0.9191
Lasso	5.9873	0.8947	4.2030	0.9311	4.3663	0.9063
El.-net	8.0225	0.9190	3.9154	0.9446	4.0536	0.9198
Peso pol	6.4870	0.8743	5.2683	0.9048	6.8683	0.8387
Peso exp	6.6511	0.8718	4.9535	0.9070	6.6922	0.8389

(b) $b = 500$

Metodi	RMSE (D)	r (D)	RMSE (F)	r (F)	RMSE (M)	r (M)
Minimi Q.	12.0284	0.7548	6.2336	0.8696	6.8290	0.8772
Ridge	7.4932	0.9375	3.5321	0.9558	3.7928	0.9301
Lasso	5.4370	0.9191	3.9419	0.9362	4.6328	0.8956
El.-net	6.8905	0.9297	3.7257	0.9504	3.9002	0.9265
Peso pol	6.2759	0.8746	5.2948	0.9040	6.9411	0.8356
Peso exp	6.6511	0.8719	4.9536	0.9070	6.6923	0.8389

(c) $b = 1000$

Metodi	RMSE (D)	r (D)	RMSE (F)	r (F)	RMSE (M)	r (M)
Minimi Q.	45.6410	-0.0205	23.4047	0.1057	31.1752	0.1934
Ridge	7.7468	0.9323	3.4574	0.9604	3.6486	0.9389
Lasso	6.4722	0.9131	3.7862	0.9425	4.3028	0.9083
El.-net	6.8638	0.9314	3.4672	0.9599	3.7861	0.9334
Peso pol	6.2682	0.8689	5.2668	0.9025	6.9119	0.8330
Peso exp	6.6511	0.8718	4.9536	0.9070	6.6923	0.8389

Tabella 5.3: Per ogni metodo di previsione utilizzato vediamo, per $b = 200$, 500 e 1000 marcatori considerati, i valori della radice dell'errore quadratico medio (RMSE) e il relativo coefficiente di correlazione lineare (r) per i gruppi D, F ed M del *dataset 'Down e familiari'* .

sufficientemente significative (sappiamo già che i nostri metodi non migliorano per questi numeri di marcatori selezionati e infatti i corrispondenti valori p del test aumentano). Anche i test di Wilcoxon del Paragrafo 2.2.4 (Tabella 5.6) e di Kolmogorov-Smirnov del Paragrafo 2.2.5 (Tabella 5.7) supportano tali conclusioni.

b	Minimi Q.	Ridge	Lasso	El.-net	Peso pol	Peso exp
10	0.955656	0.908043	0.917404	0.935328	0.132274	0.164222
20	0.487140	0.512303	0.436692	0.454638	0.205590	0.271202
50	0.523910	0.607745	0.494531	0.507479	0.463017	0.606550
100	0.412324	0.627427	0.538513	0.654051	0.766714	0.742035
200	0.961011	0.633804	0.662297	0.555559	0.772664	0.751123
500	0.128733	0.396625	0.953561	0.514005	0.792443	0.751143
1000	0.327158	0.287694	0.522782	0.281138	0.734174	0.751143

Tabella 5.4: Per ogni numero b di marker selezionati vediamo il valore p del test F adoperato per le variabili DD e DF, per tutti i metodi di previsione dell'età da noi utilizzati.

b	Minimi Q.	Ridge	Lasso	El.-net	Peso pol	Peso exp
10	0.652107	0.731376	0.641623	0.646019	0.104163	0.134605
20	0.704981	0.543831	0.658539	0.593301	0.014221	0.013638
50	0.450163	0.146994	0.264217	0.216816	0.011143	0.013164
100	0.000531	0.001046	0.003884	0.002182	0.073813	0.025044
200	0.000001	0.000006	0.011574	0.000001	0.140441	0.025568
500	0.000001	0.000001	0.005027	0.000024	0.231108	0.025571
1000	0.000012	0.000001	0.000181	0.000010	0.242460	0.025571

Tabella 5.5: Per ogni numero b di marker selezionati vediamo il valore p del test t adoperato per le variabili DD e DF, per tutti i metodi di previsione dell'età da noi utilizzati.

b	Minimi Q.	Ridge	Lasso	El.-net	Peso pol	Peso exp
10	0.506278	0.568234	0.480857	0.489252	0.117756	0.137888
20	0.930452	0.711944	0.845633	0.752353	0.030122	0.027191
50	0.514906	0.134360	0.304337	0.234714	0.011816	0.013238
100	0.000296	0.000403	0.001566	0.000698	0.134360	0.024509
200	0.000001	0.000005	0.004327	0.000001	0.285765	0.025377
500	0.000001	0.000001	0.002154	0.000058	0.343835	0.025377
1000	0.000019	0.000001	0.000166	0.000009	0.386462	0.025377

Tabella 5.6: Per ogni numero b di marker selezionati vediamo il valore p del test di Wilcoxon adoperato per le variabili DD e DF, per tutti i metodi di previsione dell'età da noi utilizzati.

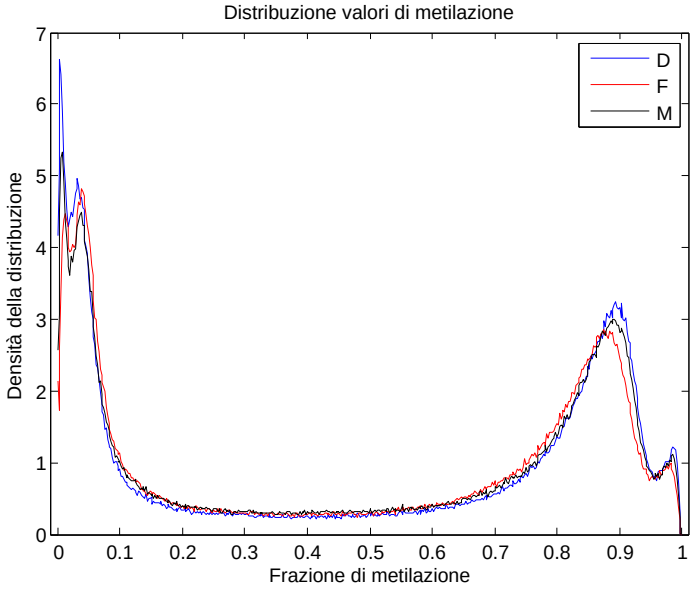
b	Minimi Q.	Ridge	Lasso	El.-net	Peso pol	Peso exp
10	0.382777	0.580885	0.382777	0.382777	0.069487	0.232549
20	0.998157	0.998157	0.950913	0.950913	0.069487	0.069487
50	0.580885	0.131531	0.580885	0.580885	0.015846	0.006841
100	0.000368	0.001042	0.006841	0.002762	0.232549	0.034317
200	0.000001	0.000011	0.006841	0.000001	0.580885	0.034317
500	0.000011	0.000003	0.001042	0.000121	0.580885	0.034317
1000	0.000037	0.000001	0.000368	0.000011	0.950913	0.034317

Tabella 5.7: Per ogni numero b di marker selezionati vediamo il valore p del test di Kolmogorov-Smirnov adoperato per le variabili DD e DF, per tutti i metodi di previsione dell'età da noi utilizzati.

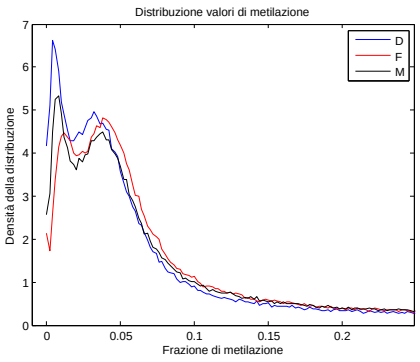
5.2 Un risultato dall'analisi di Fourier

Abbiamo voluto applicare lo stesso metodo di analisi del Paragrafo 4.2 per questo caso di studio. Vediamo innanzitutto in Figura 5.2 il profilo di metilazione di alcuni pazienti del nuovo insieme di dati.

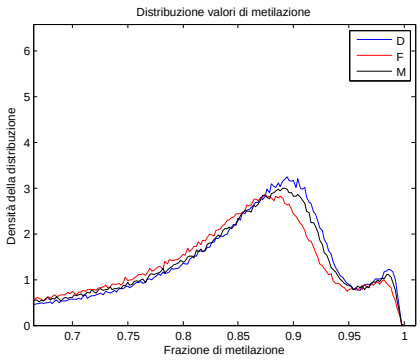
Guardando invece le trasformate reali ed immaginarie (Figura 5.3) dei tre gruppi, esse mostrano comportamenti differenti per alcuni valori di k, in cui la variabilità delle trasformate dei pazienti Down è più alta di quella degli altri due gruppi. Ci siamo poi concentrati sul confronto tra i pazienti Down (gruppo D) e i loro fratelli (gruppo F) poiché, diversamente dalle madri, han-



(a) Profilo globale

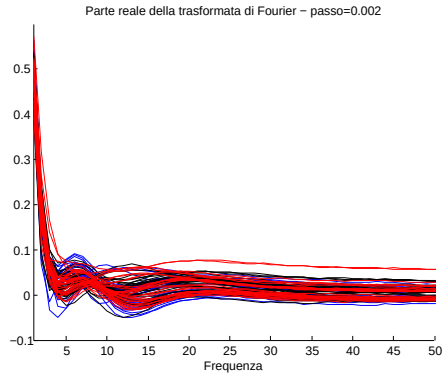


(b) Ingrandimento sul primo picco

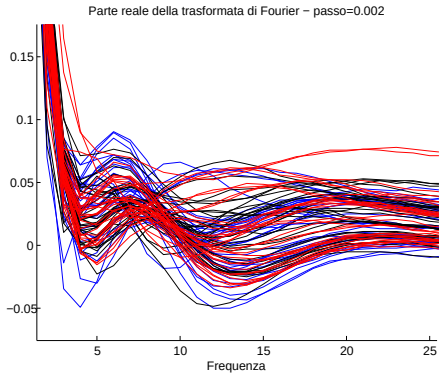


(c) Ingrandimento sul secondo picco

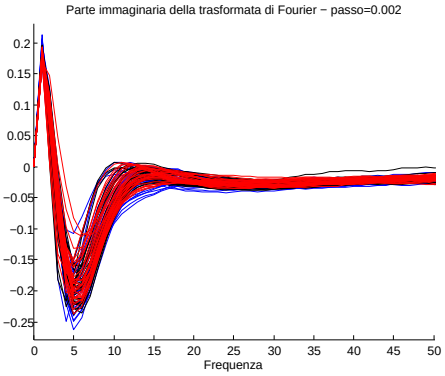
Figura 5.2: Profilo di metilazione di 3 individui: un paziente di 18 anni affetto da sindrome di Down (in blu), il fratello di anni 25 (in rosso) e la madre di anni 58 (in nero).



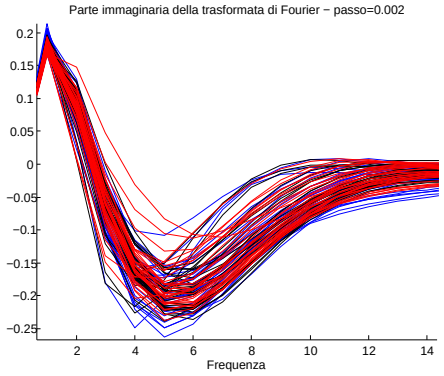
(a) Parte reale, per $0 \leq k \leq 50$



(b) Parte reale, per $0 \leq k \leq 25$

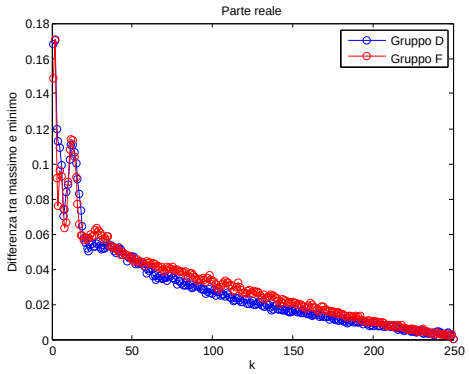


(c) Parte immaginaria, per $0 \leq k \leq 50$

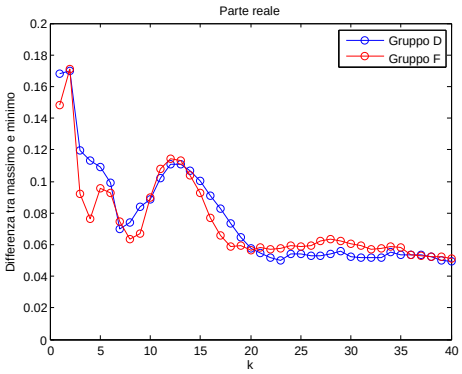


(d) Parte immaginaria, per $1 \leq k \leq 14$

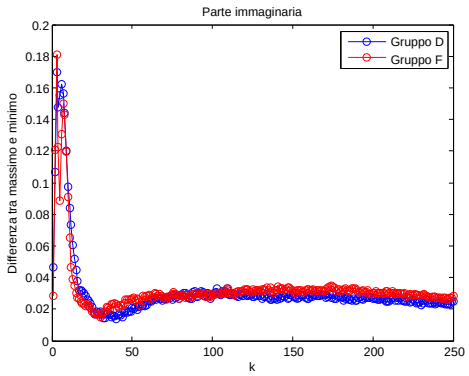
Figura 5.3: Trasformate di Fourier per i soggetti del *dataset 'Down e familiari'*: in blu i pazienti affetti da sindrome di Down, in rosso i fratelli e in nero le madri.



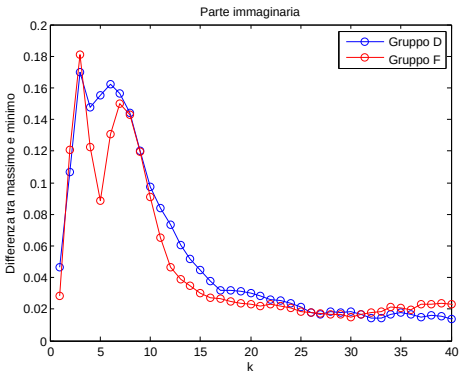
(a) Parte reale, per $1 \leq k \leq 250$



(b) Parte reale, per $1 \leq k \leq 40$



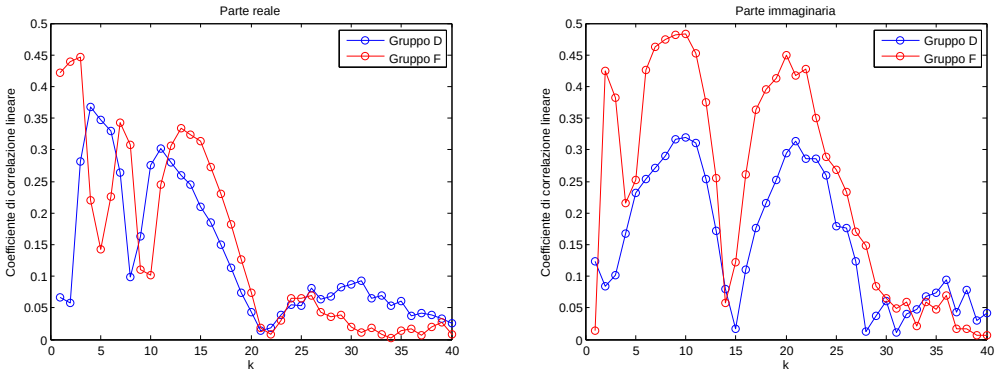
(c) Parte immaginaria, per $1 \leq k \leq 250$



(d) Parte immaginaria, per $1 \leq k \leq 40$

Figura 5.4: Differenza tra il massimo e il minimo dei coefficienti di Fourier per i soggetti del *dataset 'Down e familiari'* al variare di k: in blu i pazienti affetti da sindrome di Down e in rosso i fratelli.

no età comparabili tra loro. Partendo dalla differenza tra massimo e minimo delle trasformate reali ed immaginarie dei due gruppi (Figura 5.4), abbiamo poi analizzato se tali variabilità fossero correlate o meno con l'età: fissato k, calcoliamo il coefficiente di correlazione lineare tra l'età di un paziente (sia del gruppo D, sia del gruppo F) e il suo valore $\mathrm{Re}(\hat{f}(k))$ o $\mathrm{Im}(\hat{f}(k))$ (cioè la parte reale o immaginaria k-esimo coefficiente di Fourier della sua distribuzione di metilazione). Osserviamone l'andamento al variare di k in (Figura 5.5). In maniera analoga, fissato k, valutiamo le pendenze delle rette di regressione, trovate col metodo dei minimi quadrati, che meglio si adatta-



(a) Parte reale, per $1 \leq k \leq 40$

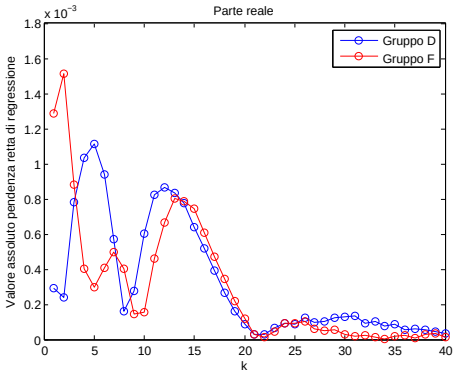
(b) Parte immaginaria, per $1 \leq k \leq 40$

Figura 5.5: Andamento del coefficiente di correlazione lineare tra l'età anagrafica e i valori dei coefficienti di Fourier per i soggetti del *dataset 'Down e familiari'* al variare di k: in blu i pazienti affetti da sindrome di Down e in rosso i fratelli.

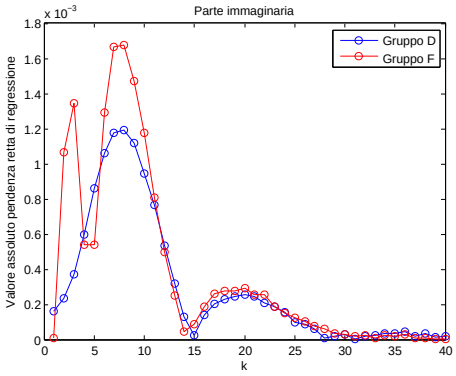
no ai valori $\mathrm{Re}(\hat{f}(k))$ e $\mathrm{Im}(\hat{f}(k))$ contro l'età dei gruppi D ed F (Figura 5.6).

Abbiamo così deciso di stabilire quali fossero i numeri k per cui le differenze tra i due gruppi fossero globalmente più significative. Abbiamo adottato il seguente criterio per ordinare tali valori: per prima cosa, si selezionano i k tali che la pendenza (in valore assoluto) della retta di regressione per il gruppo D sia maggiore di almeno 2 volte quella del gruppo F e infine ordiniamo i k in base al maggior rapporto tra i coefficienti di correlazione con l'età del gruppo D rispetto al gruppo F. Come risultato per la parte immaginaria abbiamo nell'ordine $k = 1, 14$ (Figura 5.7) e per la parte reale $k = 10, 5, 4, 6$ (Figura 5.8)

Notiamo come, all'aumentare dell'età, sia evidente un maggior tasso di crescita dei valori $\mathrm{Re}(\hat{f}(k))$ per $k = 5, 4, 6$ e di quelle immaginarie per $k = 1$, per i pazienti del gruppo D rispetto a quelli del gruppo F, mentre invece abbiamo un maggior tasso di decrescita per $k = 10$ nel caso reale e $k = 14$ nel caso immaginario. Riteniamo infine che le correlazioni di tali quantità con l'età possano essere oggetto di studio in ambito biologico.

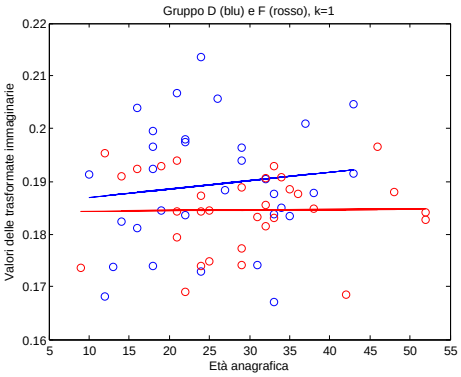


(a) Parte reale, per $1 \leq k \leq 40$

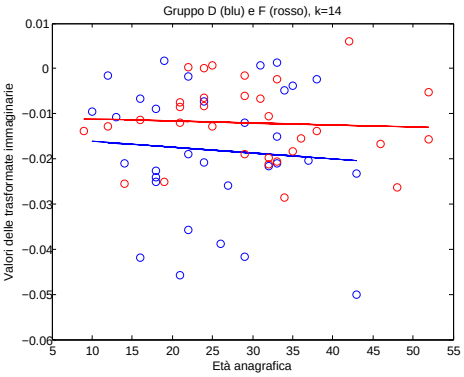


(b) Parte immaginaria, per $1 \leq k \leq 40$

Figura 5.6: Andamento della pendenza della retta di regressione che meglio fitta l'età anagrafica e i valori dei coefficienti di Fourier per i soggetti del *dataset 'Down e familiari'* al variare di k: in blu i pazienti affetti da sindrome di Down e in rosso i fratelli.

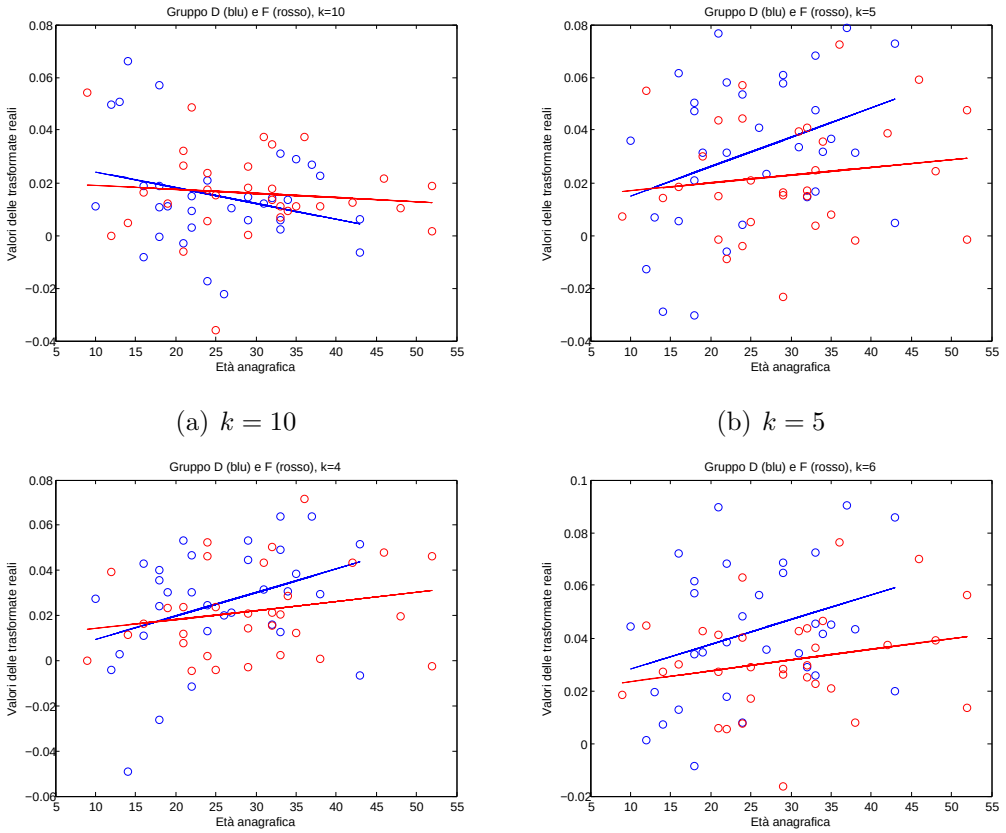


(a) $k = 1$



(b) $k = 14$

Figura 5.7: Risultati, per i k selezionati, della regressione lineare ai minimi quadrati tra età e la parte immaginaria dei coefficienti di Fourier per i soggetti del *dataset 'Down e familiari'*: in blu i pazienti affetti da sindrome di Down e in rosso i fratelli.



(a) $k = 10$ (b) $k = 5$

(c) $k = 4$ (d) $k = 6$

Figura 5.8: Risultati, per i k selezionati, della regressione lineare ai minimi quadrati tra età e la parte reale dei coefficienti di Fourier per i soggetti del *dataset 'Down e familiari'*: in blu i pazienti affetti da sindrome di Down e in rosso i fratelli.

Bibliografia

[1] Lewin, B., *Genes VIII*, Oxford University Press, 2003.

[2] Russell, P.J., *iGenetics: A Molecular Approach*, Pearson-Benjamin Cummings, 2009.

[3] Berdyshev, G.D., Korotaev, G.K., Boiarskikh, G.V., Vaniushin, B.F., *Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning*, Biokhimiia, 1967.

[4] Wilson, V.L., Jones, P.A., *DNA methylation decreases in aging but not in immortal cells*, Science, 1983.

[5] Singhal, R.P., Mays-Hoopes, L.L., Eichhorn, G.L., *DNA methylation in aging of mice*, Mechanisms of Ageing and Development, 1987.

[6] Drinkwater, R.D., Blake, T.J., Morley, A.A., Turner, D.R., *Human lymphocytes aged in vivo have reduced levels of methylation in transcriptionally active and inactive DNA*, Mutation Research, 1989.

[7] Fraga, M.F., Esteller, M., *Epigenetics and aging: the targets and the marks*, Trends in Genetetics, 2007.

[8] Esteller, M., *Epigenetics in cancer*, New England Journal of Medicine, 2008.

[9] Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al.,

Epigenetic differences arise during the lifetime of monozygotic twins, Proceedings of the National Academy of Sciences, 2005.

[10] Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I.E., Kahn, R.S., Ophoff, R.A., *The relationship of DNA methylation with age, gender and genotype in twins and healthy controls*, PLoS ON, 2009.

[11] Rice, J.A., *Mathematical Statistics and data Analysis*, Duxbury Press, 1995.

[12] Ross, S.M., *Introductory Statistics*, Academic Press, 2010.

[13] Freund, R.J., Wilson, W.J., Mohr, D.L., *Statistical Methods*, Academic Press, 2010.

[14] Lehmann, E.L., Romano, J.P., *Testing Statistical Hypotheses*, Springer, 2008.

[15] James, G., Witten, D., Hastie, T., Tibshirani, Ro., *An Introduction to Statistical Learning*, Springer, 2013.

[16] Hastie, T., Tibshirani, Ro., Friedman, J.H., *The Element of Statistical Learning*, Springer, 2009.

[17] Freedman, D.A., *Statistical Models Theory and Practice*, Cambridge University Press, 2005.

[18] Durrett, R., *Probability: Theory and Examples*, Cambridge University Press, 2010.

[19] Feller, W., *On the Kolmogorov-Smirnov limit theorems for empirical distributions*, Annals of Mathematical Statistics, 1948.

[20] Massey, F.J., *The Kolmogorov-Smirnov test for goodness of fit*, Journal of the American Statistical Association, 1951.

[21] Birnbaum, Z.W., *Numerical tabulation of the distribution of Kolmogorov statistic for finite sample size*, Journal of the American Statistical Association, 1952.

[22] Smirnov, N., *Table for estimating the goodness of fit of empirical distributions*, The Annals of Mathematical Statistics, 1948.

[23] Tibshirani, Ry., *The Lasso Problem and Uniqueness*, Electronic Journal of Statistics, Vol. 7, 2013.

[24] Zou, H., Hastie, T., *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society, 2005.

[25] Tibshirani, Ro., *Regression Shrinkage and Selection Via the Lasso*, Journal of the Royal Statistical Society, 1994.

[26] Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., Zhang, K., *Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates*, Molecular Cell, Vol. 49, 2013.

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

[28] Rudin, W., *Analisi reale e complessa*, Boringhieri, 1974.

[29] Bacalini, M.G., Gentilini, D., Boattini, A., Giampieri, E., Pirazzini, C., Giuliani, C., Fontanesi, E., Scurti, M., Remondini, D., Capri, M., Cocchi, G., Ghezzo, S., Collino, S., Del Rio, A., Luiselli, D., Vitale, G., Mari, D., Castellani, G., Fraga, M., Di Blasio, A.M., Salvioli, S., Franceschi, C., Garagnani, P., *Identification of a DNA methylation signature in blood from Down syndrome subjects*, in press.

[30] Urbano, R., *Health Issues Among Persons With Down Syndrome*, Academic Press, 2010.

Ringraziamenti

I preziosi consigli e le indispensabili correzioni del prof. Lenci mi hanno permesso di arrivare e a realizzare e concludere questo lavoro, altrimenti per me impossibile da terminare. Lo ringrazio profondamente per la straordinaria disponibilità e attenzione che ha avuto nei miei riguardi, nonostante i tanti dubbi e inconvenienti che abbiamo incontrato in questa ricerca! Ringrazio anche il dott. Cristadoro e il dott. Remondini per i suggerimenti (e i database!) che mi hanno dato, sono stati davvero preziosi. Desidero inoltre ringraziare la dott.ssa Scardovi, che mi ha fornito commenti e referenze fondamentali per comprendere e spiegare alcuni argomenti di statistica: senza tale aiuto non avrei potuto completare questo lavoro! E infine tutti gli insegnanti di matematica del mio passato scolastico che hanno saputo trasmettermi un po' del loro amore per questa materia, in particolare la prof. Gaetani del liceo, devo veramente ringraziarla di cuore per la passione che metteva nelle sue lezioni!

Oltre alla parte accademica, devono ricevere i miei ringraziamenti un altro po' di persone: innanzitutto mia madre e mio padre, non solo per il finanziamento dei miei studi (che chiaramente però è stato necessario e, spero, un buon investimento!) ma anche per tutto l'affetto e la cura che hanno avuto per me. Ringrazio anche mia sorella Sara, che forse non lo sa ma ho imparato da lei molto più di quanto non sembra in apparenza (di sicuro non la matematica, ma ci sono tante altre cose ugualmente importanti!). Vorrei ringraziare tanti altri: penso che le persone con cui condividi delle esperienze, lunghe o brevi che siano, un po' ti cambino (spero sempre in meglio!); i miei

parenti più stretti, che in quest'ultimo anno ho veramente conosciuto forse più che in tutti gli altri anni passati; i miei amici, dagli storici ai più recenti, sempre lì quando ce n'è bisogno; i compagni vecchi e nuovi; la Beverara; i cestisti (non solo Di Vimini, ma anche del Lame, Trebbo, Santa Viola)... Non posso nominare tutti (e quindi eviterò i nomi) o rischio di dimenticarmi qualcuno! Rimedierò a questa mancanza ringraziandovi uno per uno appena possibile, promesso!