Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica e Gestione delle Imprese

RELAZIONE FINALE

METODI PER LA COMBINAZIONE OTTIMALE DI BIOMARCATORI DIAGNOSTICI BASATI SULL'INDICE DI YOUDEN

Relatore Prof. Gianfranco Adimari
Dipartimento di Scienze Statistiche

Laureanda: Ilaria Angeli
Matricola N. 1008557

Anno Accademico 2013/2014

Alla mia famiglia

Ringraziamenti

Ringrazio il Professor Gianfranco Adimari per avermi assistita nella redazione di questo lavoro. Ringrazio inoltre Stefano Mussi per la gentile collaborazione.

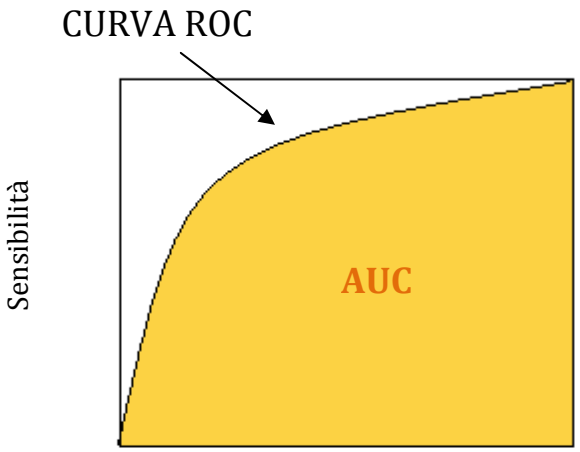
INDICE

INTRODUZIONE

In campo medico vengono impiegati numerosi e diversi biomarcatori per individuare eventuali irregolarità o anomalie fisiologiche nei pazienti.

In particolare, un biomarcatore diagnostico permette di distinguere una popolazione sana da una malata attraverso la definizione di un valore di soglia ottimale, noto come valore di "cut-off"o "cut-point".

La "bontà" di un test diagnostico, cioè la sua accuratezza nella discriminazione tra due popolazioni, è misurata da diversi strumenti, tra i quali il più noto è quello rappresentato dalla curva ROC (Receiver Operating Characteristic). Essa illustra graficamente la capacità diagnostica di un marcatore su base scalare e continua tracciando, per tutti i valori di soglia in un quadrato unitario, la proporzione di veri positivi (sensibilità) rispetto alla proporzione di falsi positivi (1 – specificità).



L'area delimitata al di sotto della curva ROC (AUC) è il più utilizzato indice di accuratezza diagnostica globale, in quanto, a valori maggiori associati a quest'area, corrisponde una migliore abilità selettiva del biomarcatore su tutti i valori di soglia.

Solitamente, più biomarcatori vengono misurati in uno stesso soggetto con lo scopo di incrementare l'accuratezza nella diagnosi di una malattia. Infatti, molti ricercatori sono spinti a ottenere dei risultati basati sulla combinazione ottimale di più biomarcatori, in modo da avere un quadro informativo più completo rispetto a quello che si avrebbe studiandoli in modo separato.

Tale combinazione può essere ottenuta utilizzando diversi metodi, anche se il più facile e diffuso consiste nel combinare i biomarcatori linearmente.

Su e Liu[1], per esempio, hanno ricavato (derivato) le combinazioni lineari che massimizzano l'area sotto le curve ROC. Considerando il fatto che la precisione della diagnosi è spesso verificata attraverso il tasso di classificazione globale corretto (cioè somma tra specificità e sensibilità) relativo alla soglia, il risultato combinato basato proprio sulla massimizzazione dell'AUC si rivela non essere così adeguato per giustificare la capacità discriminatoria tra pazienti sani e malati nel punto di "threshold" ottimale.

L'AUC, infatti, è una misura globale e sintetica utilizzata per valutare la bontà diagnostica per tutte le possibili soglie, non soltanto di quella migliore.

Pertanto, nella ricerca di un ideale marcatore lineare che permetta di raggiungere il più elevato tasso di classificazione corretta globale, proprio quest'ultimo dovrebbe essere impiegato come funzione obiettivo.

Pochi articoli di ricerca si sono occupati di studiare la combinazione di biomarcatori usando altre statistiche riassuntive legate alla curva ROC come funzioni obiettivo, oltre all'AUC. Interessante si è rivelata la proposta di Jiingjing Yin e Lili Tian[2] che hanno provato ad adottare l'indice di Youden. Questo indice è molto utilizzato in ambiti scientifici per la scelta di un valore di soglia per un marcatore in quanto il concetto sul quale si basa è strettamente legato al contesto dell'analisi mediante curva ROC.

A differenza dell'AUC, l'indice di Youden, definito come di seguito

$$J = \max_c \{\text{Sensibilità}(c) + \text{Specificità}(c) - 1\}$$

è una funzione della sensibilità e della specificità massimizzata rispetto al punto "c", che corrisponde proprio al cut-poin ottimale, e fornisce anche una misura immediata del tasso di classificazione corretta globale massimo per un dato marcatore.

Alla luce di quanto appena detto, nei capitoli successivi una prima parte sarà dedicata all'analisi di alcuni metodi per la ricerca della combinazione lineare di biomarcatori basata sulla massimizzazione dell'indice di Youden.

Nella seconda parte si confronteranno i risultati derivanti da alcune simulazioni sui metodi proposti e dimostrerà il vantaggio nell'utilizzo dell'indice di Youden come funzione obiettivo per combinare i marcatori rispetto all'AUC. Nella terza parte si proporrà un esempio di applicazione del metodo non parametrico del lisciamento del nucleo per la ricerca della combinazione lineare ottimale di biomarcatori, utili all'identificazione di portatori della distrofia muscolare di Duchenne.

CAPITOLO 1.

1.1 Test diagnostici e impiego dell'indice dì Youden

Come detto nell'introduzione, l'indice di Youden (J) è una misura sintetica dell'accuratezza di un test diagnostico. Per capire in che modo questa tecnica riesca a stabilire l'efficacia del test, occorre prima specificare cosa esso sia e come venga impiegato nella medicina moderna.

In numerosi campi della scienza, infatti, sono continuamente introdotte procedure di diversa natura e complessità, ma sempre giustamente codificate, con lo scopo di analizzare e verificare determinate ipotesi. Tali metodi prendono il nome di "test".

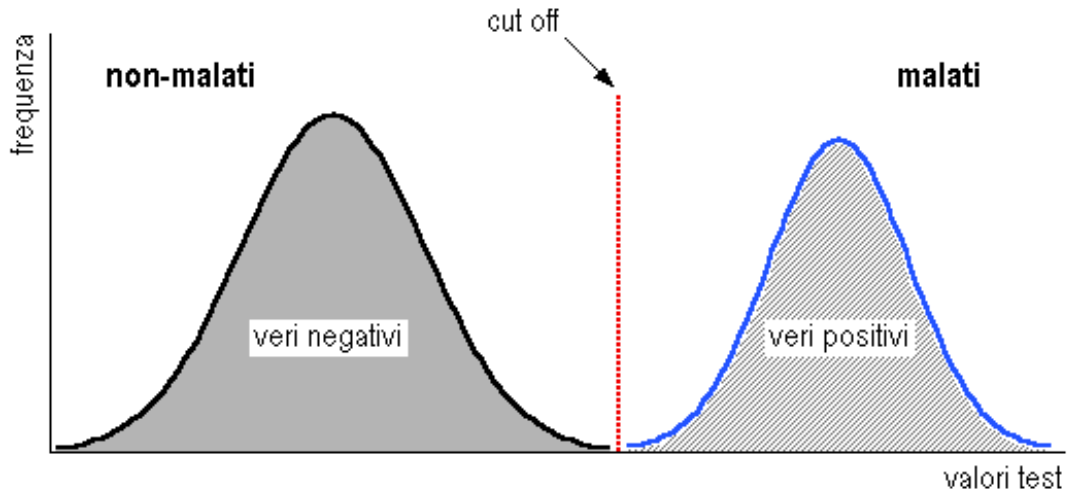
In particolare, in epidemiologia (branca della scienza medica che tratta dello studio dell'incidenza, della distribuzione e del controllo di una malattia in una popolazione[3]) i test sono lo strumento principale delle operazioni di screening su popolazioni considerate presuntivamente sane, attraverso le quali si vuole identificare precauzionalmente la presenza di infezioni o malattie subcliniche.

È soprattutto nell'attività diagnostica di routine che i test rappresentano quei fattori fondamentali e determinanti per la conferma o l'esclusione della presenza di una determinata malattia, già sospettata precedentemente in base ai dati clinici.

Se l'esito di un test (diagnostico) fornisce un output dicotomico ("vero/falso", "positivo/negativo", "presente/assente") esso viene chiamato test "qualitativo", mentre un test è "quantitativo" quando produce risultati espressi attraverso variabili numeriche, siano esse "discrete" , cioè che assumono valori finiti e numerabili derivati da operazioni di conteggio (es. numero di pazienti fumatori), o siano esse "continue", quindi valori non numerabili ma che possono assumere qualsiasi valore all'interno di un intervallo (es. il peso di un paziente).

A differenza dei test qualitativi, nel caso di test quantitativi (discreti o continui) è indispensabile la presenza di un valore di soglia, noto come "cut-off" o "cut-point" in grado di individuare una netta separazione tra risultati "positivi" e "negativi", ossia una distinzione tra soggetti malati e non malati (Figura 1.) .

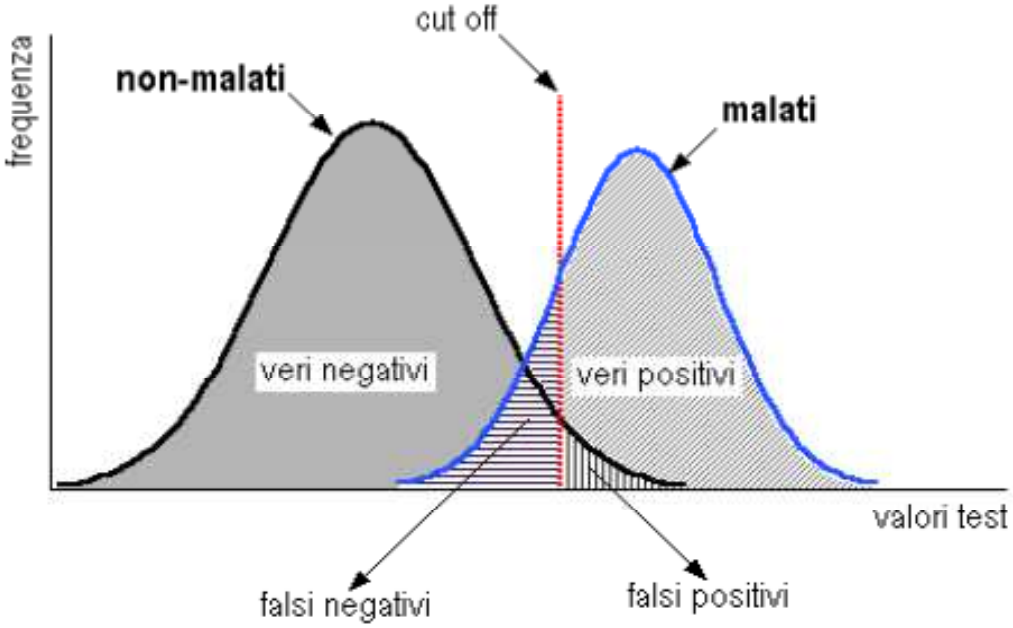
Figura 1.



A tal proposito va però chiarito che, nelle circostanze reali, è praticamente impossibile ritrovarsi in una situazione di assoluta scissione tra la popolazione "sana" e "malata". Infatti, come si vede dalla figura 2. , le distribuzioni degli esiti di un ipotetico test nelle due categorie di individui sono sovrapposte, e ciò dimostra che si ha un certo numero di persone sane che risulta positiva al test ("falsi positivi", "FP"), e un certo numero di persone malate che erroneamente vengono classificate come sane ("falsi negativi", "FN").

Figura 2.

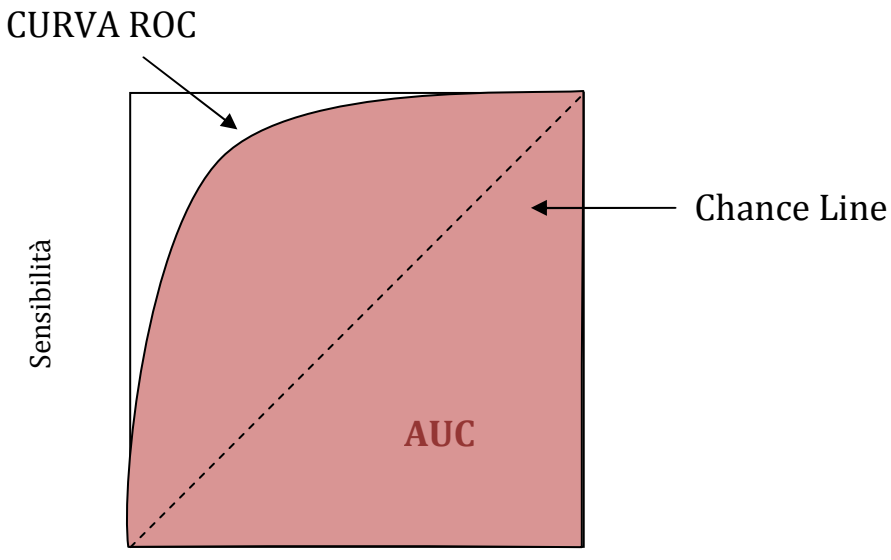
La



performance di un test ad un determinato valore di soglia, detto anche “threshold”, viene valutata perciò calcolando anche due importanti indici sintetici:

- la sensibilità (q(c)), ovvero la probabilità che un soggetto realmente malato risulti positivo nel test (“veri positivi”, “TP”) ed è misurata dal rapporto tra veri positivi e il totale degli individui effettivamente malati;
- la specificità (p(c)), ovvero la probabilità che un soggetto realmente sano risulti negativo nel test (“vero negativo”, “TN”) ed è misurata dal rapporto tra veri negativi e il totale degli individui effettivamente sani.

L’analisi ROC (Receiver Operating Characteristic) è uno tecnica impiegata per studiare la funzione che lega la sensibilità (q(c)) al complemento a uno della specificità (1 - (p(c)). L’obiettivo è quindi quello di analizzare il rapporto tra “veri positivi” e “falsi positivi”. Se i risultati sono registrati su scala continua, la relazione tra questi parametri viene mostrata graficamente attraverso una curva, detta appunto “curva ROC” che si ottiene tracciando, in un sistema di riferimento cartesiano, i punti che in ascissa corrispondono alla proporzione di “falsi positivi” e in ordinata alla proporzione di “veri positivi”, per ogni valore possibile di cut –off.



Qualsiasi modello per la ripartizione ideale propone la massimizzazione di sensibilità e specificità in quanto si ipotizza una situazione in cui non vi siano falsi né negativi, né positivi. Nella realtà ciò non avviene, anzi si osserva una sorta di trade-off tra i due parametri.

E' facile, infatti, verificare che sensibilità e specificità sono fra loro inversamente correlate in rapporto alla scelta del valore di cut off.[4] Infatti, modificando quest'ultimo si può ottenere:

- un aumento della sensibilità e diminuzione della specificità;
- diminuzione della sensibilità ed aumento della specificità.

L'area sottesa alla cura ROC, nota come AUC (Area Under Curve), è uno dei metodi di sintesi che fornisce l'informazione utile per verificare la capacità discriminatoria del test considerato: maggiore è il valore dell'area, maggiore risulta l'accuratezza del test.

Per valutare un test diagnostico, questo metodo assegna lo stesso peso a sensibilità e specificità, quando nella realtà conviene attribuire loro pesi diversi in base alle differenti tipologie di malattia.

Un altro approccio comunemente utilizzato per testare l'efficacia diagnostica dei biomarcatori è dato dall'indice di Youden (J).

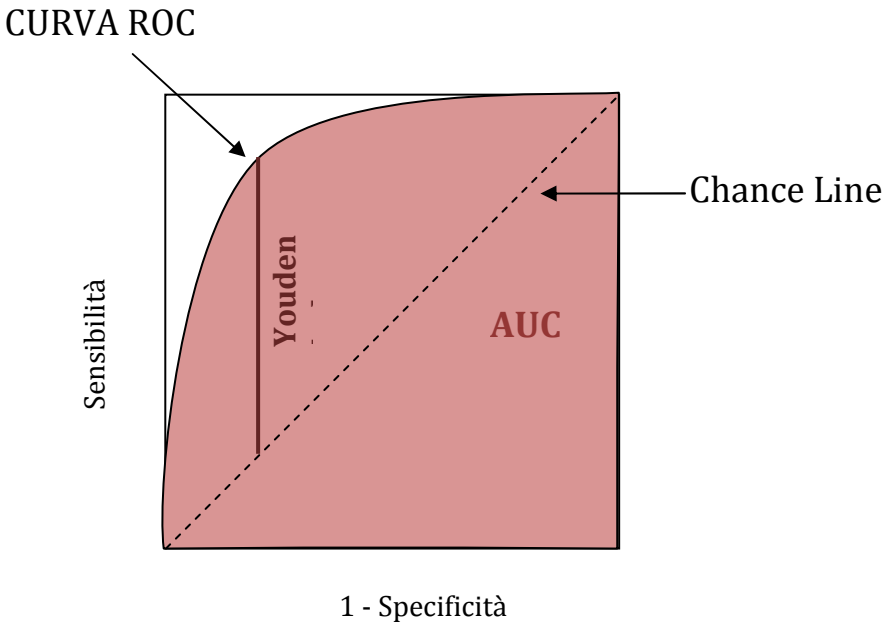
Introdotto nella letteratura medica da Youden nel 1950, esso rappresenta la massimizzazione di una funzione della sensibilità e della specificità rispetto ad un valore di soglia "c", che di conseguenza è quel "cut-point" che identifica più correttamente la distinzione tra sani e malati.

$$\begin{aligned} J &= \max_c \{\text{sensibilità}(c) + \text{specificità}(c) - 1\} \\ &= \max_c \{ q(c) + p(c) - 1\} \\ &= \max_c \{q(c) - (1 - p(c))\} \end{aligned}$$

J è quindi la massima abilità di differenziazione di un biomarcatore quando viene dato lo stesso peso a sensibilità e specificità. Assume valori compresi nell'intervallo chiuso [0 , 1], dove a valore 0 corrisponde un test completamente inefficace, mentre a valore 1 corrisponde un test perfettamente efficace[5].

Questo indice è strettamente legato all'analisi precedente. Convenzionalmente, J si calcola empiricamente sommando specificità e sensibilità meno uno per tutti i possibili cut-points e scegliendo il risultato con il valore massimo. Graficamente, J è la massima distanza verticale tra la curva ROC e la diagonale o "chance line".

A differenza dell'AUC, l'indice di Youden permette anche di calcolare direttamente il valore di cut-off.



1.2 Biomarcatori e loro combinazioni lineari

L'impiego dei biomarcatori nella ricerca di base e clinica, come pure nella pratica clinica, è diventato talmente comune che la loro presenza come endpoint primari, cioè come variabili che forniscono l'informazione più rilevante per l'obiettivo primario di uno studio, è ormai quasi del tutto consolidata. Nel caso specifico di biomarcatori, che sono stati ben caratterizzati e più volte verificati per predire correttamente importanti risultati clinici, il loro uso è completamente giustificato e appropriato.

In molti altri casi, invece, si presume che la "validità" dei biomarcatori debba essere continuamente valutata e rivalutata, soprattutto quando si combinano tra loro.

Il termine "biomarcatore", infatti, si riferisce ad una sottocategoria molto ampia dei cosiddetti segni medici, ossia indicazioni oggettive di stato di salute osservate dall' esterno del paziente e non percepite direttamente da quest'ultimo, che possono essere misurati accuratamente e riprodotti.

Tra le numerose definizioni adottate per definire un marcatore biologico, l'International Programme on Chemical Safety lo considera come una " qualsiasi sostanza, struttura, o processo che può essere misurato nel corpo, o i propri prodotti, e che può influenzare o predire l'incidenza di un risultato o di una malattia".[6]

L'utilità di un marcatore viene accertata basandosi generalmente sulla sensibilità e sulla specificità, definite nel paragrafo precedente.

In molti studi diagnostici, nei quali sono presenti più marcatori contemporaneamente, si pensa che un risultato ottenuto dalla loro combinazione possa notevolmente aumentare l'accuratezza della diagnosi, rispetto a un loro semplice impiego individuale. Perciò, nelle circostanze in cui sono coinvolti due o più biomarcatori, alcuni ricercatori hanno proposto diversi metodi per la valutazione e il confronto delle performance. Già negli anni 1982-83 Hanley e McNeil, e negli anni successivi McClish (1987) ed E. R. Delong, D. M. Delong e D. L. Clarke-Pearson (1988) avevano presentato raffronti di marcatori basandosi sulla differenza delle aree sottese alle curve ROC, mentre Greenhouse e Mantel (1950) e Linnet (1987) avevano comparato due valori di sensibilità ad un unico livello fissato di specificità.

Dal momento in cui, nelle situazioni reali, i diversi marcatori sono sensibili ad altrettanti diversi aspetti della malattia, è importante usare simultaneamente due o più buoni biomarcatori in modo che si riesca ad ottenerne uno nuovo con una sensibilità maggiore. Su e Liu[1] furono interessati a combinare linearmente due o più marcatori analizzando la teoria dell'analisi discriminante lineare; in particolare, dimostrarono che sotto la distribuzione normale multivariata le migliori combinazioni lineari sono uguali alla funzione lineare discriminante che Fisher aveva considerato nel 1936, e scoprirono che le combinazioni lineari ottimali dei biomarcatori sono quelle che massimizzano l'area sotto le curve ROC.

I metodi di combinazione lineare che utilizzano l'AUC come funzione obiettivo sono stati ampiamente studiati. Quando si confronta la precisione diagnostica di parecchi biomarkers, è spesso di grande interesse per i ricercatori porre l'attenzione soltanto sul tratto della curva ROC che presenta maggiore specificità e quindi prediligere i markers che hanno miglior prestazioni ad alte specificità. Benchè le combinazioni lineari di Su e Liu siano ottime in quanto massimizzano l'area sotto la ROC CURVE, esse potrebbero avere performance insoddisfacenti per gli scopi reali, in quanto negli intervalli in cui la specificità è elevata, al contrario, la sensibilità è notevolmente bassa[7].

1.3 Indice di Youden come funzione obiettivo per la combinazione lineare ottimale

Nella ricerca di una combinazione lineare ottimale di marcatori biologici occorre scegliere in modo accurato la funzione obiettivo, ossia quella funzione che minimizzata o massimizzata fornisce le soluzioni più adeguate per lo scopo prefissato.

Finora molti ricercatori hanno affrontato il problema di trovare la combinazione ottimale e lineare basandosi sulla massimizzazione dell'area sottesa alla curva ROC.

Attualmente, si ritiene che questo risultato combinato non sfrutti pienamente le caratteristiche adatte a stabilire la soglia migliore. L'attenzione viene posta quindi sulla necessità di trovare quella funzione obiettivo che sia in grado di dare un valore di cut-off il più accurato possibile. L'accuratezza è la capacità di un test diagnostico di generare valori corrispondenti a quelli reali e viene calcolata frequentemente dal tasso di classificazione globale corretto, dato dalla somma dei valori di sensibilità e di specificità relativi alla soglia. Sarebbe opportuno quindi disporre di una funzione obiettivo che, massimizzata, possa fornire contemporaneamente un valore di cut-point ottimale e una misura diretta del tasso di classificazione globale corretto massimo.

A tal proposito Jingjing Yin and Lili Tian[2] hanno proposto alcuni metodi per la combinazione lineare ottimale di più biomarcatori basandosi sull'indice di Youden.

La combinazione di marcatori ottenuta ponendo l'indice di Youden come funzione obiettivo porterà a un risultato combinato con il massimo tasso di classificazione globale corretto, relativo alla soglia diagnostica, tra tutte le possibili combinazioni lineari. Quindi, come biomarcatore combinato, ha la migliore accuratezza per fare diagnosi.

Prima di procedere con la descrizione dei metodi analizzati da Yin e Tian, è necessario delineare il contesto di partenza e fare alcune precisazioni.

Si assumano n_1 e n_2 soggetti per gruppi di malati e non malati, rispettivamente, e p marcatori su scala continua vengono misurati su ciascun soggetto e, ovviamente, patologia.

Sia Y_1 il vettore p-dimensionale delle misurazioni dei biomarcatori nel gruppo dei malati e sia $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ un campione casuale di numerosità n_1, con

$$Y_{1i} = \left(Y_{1i1}, Y_{1i2}, \dots, Y_{1ip}\right)^T , \;\; i = 1, 2, \dots, n_1,$$

dove Y_{1i} $(k = 1, 2, \dots , p)$ denota la misurazione del k-esimo biomarcatore nell'individuo i-esimo nel gruppo dei malati.

Analogamente, sia Y_2 il vettore p-dimensionale delle misurazioni dei biomarcatori nel gruppo dei non malati e sia Y_{21}, Y_{22}, ..., Y_{2n_2} un campione casuale di numerosità n_2, con

$$Y_{2j} = \left(Y_{2j1}, Y_{2j2}, \dots, Y_{2jp}\right)^T , \;\; j = 1, 2, \dots, n_2,$$

dove Y_{2jk} $(k = 1, 2, \dots , p)$ denota la misurazione del k-esimo biomarcatore nell'individuo j-esimo nel gruppo dei non malati.

Sia $w = (w_1, w_2, ..., w_p)^T$ il vettore dei coefficienti di combinazione. Il risultato (test) combinato è definito da

$$Z_g = w^T Y_g \;\; \text{per } g = 1, 2 \qquad (1)$$

per gruppi di malati e sani, rispettivamente.

Per il k-esimo biomarcatore, $F_{Y_{1k}}(.)$ e $F_{Y_{2k}}(.)$ denotano le funzioni di ripartizione (CDF) delle misurazioni per le due popolazioni, nell'ordine definito poc'anzi. Dato un valore di soglia c, sensibilità (P_{1k}) e specificità (P_{2k}) per il generico k-esimo marcatore sono

$$P_{1k}(c) = 1 - F_{Y_{1k}}(c), \quad P_{2k}(c) = F_{Y_{2k}}(c)$$

e le corrispettive stime sono

$$\hat{P}_{1k}(c) = 1 - \hat{F}_{Y_{1k}}(c), \quad \hat{P}_{2k}(c) = \hat{F}_{Y_{2k}}(c)$$

dove $\hat{F}_{Y_{1k}}(.)$ e $\hat{F}_{Y_{2k}}(.)$ indicano le funzioni di ripartizione stimate per i campioni di malati e non malati.

Nelle applicazioni pratiche, il valore di cut-off c è spesso sconosciuto e viene scelto ottimizzando una funzione obiettivo, come l'indice di Youden.

L'indice di Youden per il k-esimo marcatore (J_k) è definito come $J_k = \max_c\{P_{1k}(c) + P_{2k}(c) - 1\}$. Il valore della soglia ottimale per il k-esimo marcatore determinato dall'indice di Youden è indicato con c_k, e si ottiene da

$$\begin{aligned} c_k &= \arg max_c \left(P_{1k}(c) + P_{2k}(c) - 1\right) ; \\ &= \arg max_c \left(F_{y_{2k}}(c) - F_{Y_{1k}}(c)\right). \end{aligned} \tag{2}$$

La sensibilità e la specificità al cut-point ottimale mediante l'indice di Youden sono

$$P_{1k}(c_k) = 1 - F_{Y_{1k}}(c_k), \quad P_{2k}(c) = F_{Y_{2k}}(c_k)$$

L'indice di Youden stimato per il k-esimo marcatore è

$$\hat{J}_k = \hat{P}_{1k}(\hat{c}_k) + \hat{P}_{2k}(\hat{c}_k) - 1 = \hat{F}_{Y_{2k}}(\hat{c}_k) - \hat{F}_{Y_{1k}}(\hat{c}_k) \tag{3}$$

dove

$$\hat{c}_k = \arg max_c\left(\hat{F}_{Y_{2k}}(c) - \hat{F}_{Y_{1k}}(c)\right) \tag{4}$$

Per il risultato combinato Z_g in (1), si denotano le corrispondenti CDF come $F_{Z_g}(.)$ per $g = 1, 2$, e il cut-point ottimale associato con l'indice di Youden come c_w.
L'indice di Youden stimato per il risultato combinato è

$$\hat{J}_w = \hat{F}_{Z_2}(\hat{c}_w) - \hat{F}_{Z_1}(\hat{c}_w) \tag{5}$$

dove

$$\hat{c}_w = \arg max_c\left(\hat{F}_{Z_2}(c) - \hat{F}_{Z_1}(c)\right). \tag{6}$$

1.4 Due metodi empirici di ricerca non vincolati alla scelta della distribuzione (non parametrici)

In questo capitolo vengono proposti due metodi di ricerca non parametrici: il metodo min-max e l'approccio graduale (stepwise).

Utilizzando le stime empiriche per le funzioni di ripartizione delle popolazioni, malata e sana, nella (3) e nella (5), la stima empirica dell'indice di Youden per il k-esimo marcatore è

$$\hat{J}_k = \hat{F}_{Y_{2k}}(\hat{c}_k) - \hat{F}_{Y_{1k}}(\hat{c}_k) = \frac{\sum_{j=1}^{n_2} I(y_{2jk} \le \hat{c}_k)}{n_2} - \frac{\sum_{i=1}^{n_1} I(y_{1ik} \le \hat{c}_k)}{n_1} \quad (7)$$

dove $\hat{c}_k$ è il valore di cut-off empirico per il k-esimo biomarcatore definito nella (4) per $k = 1, 2, \dots, p$. Allo stesso modo, per il risultato combinato Z_g $(g = 1, 2)$, la stima empirica dell'Indice di Youden (J_w) è

$$\hat{J}_w = \hat{F}_{Z_2}(\hat{c}_w) - \hat{F}_{Z_1}(\hat{c}_w) = \frac{\sum_{j=1}^{n_2} I(w^T y_{2j} \le \hat{c}_w)}{n_2} - \frac{\sum_{i=1}^{n_1} I(w^T y_{1i} \le \hat{c}_w)}{n_1} \quad (8)$$

dove $\hat{c}_w$ è il valore di cut-off empirico definito nella (6).

Lo scopo è quello di ottenere il vettore dei coefficienti di combinazione linerare w che massimizza $\hat{J}_w$, ovvero l'indice di Youden del marcatore combinato. Poiché $\hat{J}_w$ coinvolge la funzione indicatrice $I(.)$, esso non è una funzione lisciata.

Pertanto, piuttosto che un approccio analitico, è necessaria, per tale problema di ottimizzazione, un'ampia griglia di ricerca nello spazio p-dimensionale.

Un simile metodo di ricerca con griglia è stato proposto da Pepe e Thompson[8] con lo scopo di trovare la combinazione lineare ottimale che massimizza l'AUC.

Queste procedure di analisi sono, di solito, computazionalmente complesse quando $p > 3$. Quindi sono necessari alcuni metodi alternativi che alleggeriscano le difficoltà di calcolo, proposti nel seguito.

L'approccio del min-max

C. Liu, A. Liu e S. Halabi[9] hanno proposto un approccio "distribution-free", ossia non dipendente da distribuzioni di probabilità, di tipo min-max, che combina linearmente soltanto i valori di minimo e massimo delle misurazioni dei p biomarcatori, per massimizzare l'AUC. Esso si basa sui ragionamenti di seguito esposti.

Per tutti i marcatori $1 \leq k \leq p$, alla soglia c, la sensibilità soddisfa

$$Pr\,(min\,(Y_{1i}) > c) \leq Pr\,(Y_{1ik} > c) \leq Pr\,(max\,(Y_{1i}) > c)$$

per ogni $i = 1, 2, \dots, n_1$, e la specificità soddisfa

$$Pr\,(max\,(Y_{2j}) \leq c) \leq Pr\,(Y_{2jk} \leq c) \leq Pr\,(min\,(Y_{2j}) \leq c)$$

per ogni $j = 1, 2, \dots, n_2$. Questi risultati dimostrano che un compromesso tra massimo e minimo, come ad esempio una corretta combinazione lineare dei due, potrebbe aumentare i valori di sensibilità e specificità. Liu et al.[9] sperano che, ai fini di massimizzare l'AUC, questa procedura di min-max proposta porti buoni risultati, in quanto, essendo di natura non parametrica, è robusta contro distribuzioni non correttamente specificate.

I precedenti ragionamenti citati si applicano direttamente al criterio di ottimizzazione basato sull'indice di Youden, semplice funzione appunto di sensibilità e specificità.

Seguendo gli studi di Liu et al.[9], Yin e Tian[2] hanno esteso l'approccio del min-max usando l'indice di Youden come funzione obiettivo (al posto dell'AUC) come segue:

$$\hat{J}_w = \frac{\sum_{j=1}^{n_2} I(y_{2j,max} + w y_{2j,min} \leq \hat{c}_w)}{n_2} - \frac{\sum_{i=1}^{n_1} I(y_{1i,max} + w y_{1i,min} \leq \hat{c}_w)}{n_1} \qquad (9)$$

dove $y_{1i,max} = \max(y_{1i})$ e $y_{1i,min} = \min(y_{1i})$ per il gruppo dei malati, $y_{2j,max} = \max(y_{2j})$ e $y_{2j,min} = \min(y_{2j})$ per il gruppo dei non malati, e $\hat{c}_w$ è il valore di cut-off empirico ottimale per tale risultato combinato. Pertanto, un altro vantaggio di questo metodo è dato dal fatto che il problema di ottimizzazione si riduce soltanto alla semplice ricerca di un unico parametro, ossia il coefficiente w.

Considerando 201 valori equispaziati nell'intervallo [-1, 1], il valore di w all'interno di [-1, 1], che dà il più elevato indice di Youden nella (9), è ottenuto attraverso una ricerca empirica sui 201 valori. Per ciascun valori di w dato:

- si ricerca in modo empirico il valore di soglia ottimale $\hat{c}_w$ massimizzando $\hat{J}_w$;
- si calcola il valore di $\hat{J}_w$ al cut-point ottimale selezionato per ciascun valore di w;
- si seleziona il valore ottimale di w che corrisponde al maggior valore di $\hat{J}_w$.

Riconoscendo il fatto che, moltiplicando per qualsiasi costante, l'indice di Youden del risultato combinato non viene influenzato in quanto la curva ROC è invariante rispetto a qualsiasi trasformazione monotona, si può dividere la combinazione lineare per w, cioè $\frac{1}{w} \times (max + min) = r \times (max + min)$. Ponendo $r = \left(\frac{1}{w}\right)$ pari a 201 valori equispaziati in [-1,1], i valori di w al di fuori dell'intervallo vengono altrettanto esplorati. Pertanto viene eseguita un'estesa ricerca empirica di tutti i possibili valori di w che vanno da $-\infty$ a $+\infty$. Infine, dopo aver ottenuto i valori di w all'interno e al di fuori di [-1, 1], rispettivamente, si confrontano i corrispondenti valori di $\hat{J}_w$ per decidere per quale valore di w si ottiene la combinazione ottimale dei coefficienti appartenenti alla retta dei numeri reali.

L'approccio graduale o stepwise

Il primo inconveniente dell'approccio di tipo min-max è causato dal fatto che questo metodo utilizza soltanto il minimo e il massimo dei valori del marcatore, quindi potrebbe non essere il modo migliore per ottenere una combinazione lineare desiderabile. Yin e Tian[2] hanno di conseguenza provato ad adottare un altro metodo studiato da Kang et al.[10] che prevede la combinazione dei marcatori in modo graduale. In particolare, L. Kang, C. Xiong, P. Crane e L. Tian[10] avevano analizzato un approccio graduale utilizzando il volume sotteso alla superficie della curva ROC (VUS) come funzione obiettivo per biomarcatori con tre categorie ordinali di malattia: ad esempio, per alcune malattie come l'Alzheimer si possono definire le categorie "normale", "deterioramento cognitivo lieve" e "malato di Alzheimer" al posto dell'abituale classificazione binaria "sano" o "malato".

Potendo scegliere tra una procedura graduale verso l'alto e verso il basso, Kang et al.[10] hanno verificato attraverso alcuni studi di simulazione il miglior rendimento della seconda che è stata, di conseguenza, scelta da Yin e Tian[2] per cercare la combinazione lineare ottimale usando l'indice di Youden come funzione obiettivo.

I passaggi da sviluppare sono i seguenti:

a) Calcolare la stima empirica dell'indice di Youden per ciascuno dei p biomarcatori seguendo la (7);

b) Ordinare questi p biomarcatori in base ai valori delle stime dell'indice di Youden empirico, dal più grande al più piccolo. Si denotino i corrispondenti valori per il gruppo dei malati con $y_{1i} = \left(y_{1i,(p)},\ y_{1i,(p-1)}, \dots, y_{1i,(1)}\right)^T$ per $i = 1, 2, \dots, n_1$ e per i non malati con $y_{2j} = \left(y_{2j,(p)},\ y_{2j,(p-1)}, \dots, y_{2j,(1)}\right)^T$ per $j = 1, 2, \dots, n_2$.

c) Combinare i primi due biomarcatori empiricamente usando la funzione obiettivo

$$\hat{J}_{w_{p-1}} = \frac{\sum_{j=1}^{n_2} I\left(y_{2j,(p)} + w_{p-1}\, y_{2j,(p-1)} \le \hat{c}_{w_{p-1}}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(y_{1i,(p)} + w_{p-1}\, y_{1i,(p-1)} \le \hat{c}_{w_{p-1}}\right)}{n_1} \tag{10}$$

che è valutata su 201 valori di w_{p-1} equispaziati in [-1, 1]. Per ciascun valore di w_{p-1} dato, si cerca empiricamente il valore della soglia ottimale $\hat{c}_{w_{p-1}}$ attraverso la massimizzazione di $\hat{J}_{w_{p-1}}$ e poi si calcola il valore di $\hat{J}_{w_{p-1}}$ al cut-point ottimale per ciascun w_{p-1}.

d) In modo analogo al passo (c), combinare empiricamente i primi due biomarcatori con

$$\hat{J}_{r_{p-1}} = \frac{\sum_{j=1}^{n_2} I\left(r_{p-1}\, y_{2j,(p)} + y_{2j,(p-1)} \le \hat{c}_{r_{p-1}}\right)}{n_2} - \frac{\sum_{i=1}^{n_1} I\left(r_{p-1}\, y_{1i,(p)} + y_{1i,(p-1)} \le \hat{c}_{r_{p-1}}\right)}{n_1} \tag{11}$$

e $\hat{J}_{r_{p-1}}$ è valutato su 201 valori di r_{p-1} equispaziati in [-1, 1].

e) Indicare i coefficienti $\left(w_{p-1}\ o\ r_{p-1}\right)$, che si riferiscono all'indice di Youden con valore maggiore $\left(\hat{J}_{w_{p-1}} o\ \hat{J}_{r_{p-1}}\right)$, come la combinazione lineare dei coefficienti

per i primi due marcatori al di fuori dei 201 valori di $\hat{J}_{w_{p-1}}$ e degli altri 201 valori di $\hat{J}_{r_{p-1}}$.

f) Avendo derivato il risultato combinato univariato dei primi due marcatori biologici nel punto (e), occorre combinare quest'ultimo con il terzo marcatore, che corrisponde al *(p-2)* esimo marcatore ordinato, utilizzando la stessa procedura dal punto (c) all' (e). Procedere quindi allo stesso modo finchè tutti i biomarcatori vengono inclusi nella combinazione lineare.

1.5 Due metodi di ricerca di derivazione numerica

In questo paragrafo vengono proposti due metodi di derivazione numerica, anch'essi con lo scopo di ottenere una combinazione lineare ottimale basata sull'impiego dell'indice di Youden come funzione obiettivo.

Un primo metodo è un approccio parametrico che si basa sulla distribuzione normale multivariata; il secondo invece è un approccio non parametrico che sfrutta la tecnica del lisciamento basata sul metodo del nucleo.

L'approccio parametrico sotto normale multivariata

Si assuma la distribuzione normale multivariata per Y_1 e Y_2, quindi $Y_g \sim N_p(\eta_g, \Sigma_g)$, $g = 1{,}2$ per i gruppi di, nell'ordine, malati e sani. Dato il vettore dei pesi di combinazione w, il risultato combinato $Z_g = w^T Y_g$ segue la distribuzione normale univariata, ossia $Z_g \sim N(\eta_g, \sigma_g^2)$, dove

$$\mu_g = w^T \eta_g \tag{12}$$

e

$$\sigma_g^2 = w^T \Sigma_g w \tag{13}$$

Schisterman and Perkins[11] avevano presentato dei risultati, per l'indice di Youden e per il punto di soglia ottimale, per un singolo marcatore sotto distribuzione normale. Basandosi su questi, Yin e Tian[2] hanno derivato l'indice di Youden e il corrispondente valore c del marcatore combinato.

Dato il vettore w, quando il marcatore combinato ha varianze diverse tra il gruppo dei malati e non malati, ovvero $\sigma_1^2 \neq \sigma_2^2$, (questa situazione avviene quando $\Sigma_1 \neq \Sigma_2$), il cut-point ottimale per il risultato combinato è

$$c_w = \frac{\mu_2(b^2-1) - a + b\sqrt{a^2 + (b^2-1)\,\sigma_2^2 \ln(b^2)}}{b^2-1} \tag{14}$$

e l'indice di Youden

$$J_w = \Phi\left(\frac{\mu_1 - c_w}{\sigma_1}\right) + \Phi\left(\frac{c_w - \mu_2}{\sigma_2}\right) - 1 \tag{15}$$

dove

$$a = \mu_1 - \mu_2 = w^T\,(\eta_1 - \eta_2), \qquad b = \frac{\sigma_1}{\sigma_2} = \frac{\sqrt{w^T \Sigma_1 w}}{\sqrt{w^T \Sigma_2 w}}$$

e $\Phi(.)$ denota la funzione di ripartizione della normale standard.

Quando $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (i.e., $\Sigma_1 = \Sigma_2 = \Sigma$)

$$c_w = \frac{\mu_1 + \mu_2}{2}$$

e

$$J_w = 2\Phi\left(\frac{\mu_1 - \mu_2}{2\sqrt{\sigma^2}}\right) - 1$$

Una stima dell'indice di Youden $\hat{J}_w$ può essere ottenuta inserendo le stime di $\hat{\eta}_g$ e di $\hat{\Sigma}_g$ nella (14) e nella (15). Ovviamente, $\hat{J}_w$ è una funzione continua esplicita del vettore combinato w sotto normale multivariata ed è differenziabile rispetto a esso.

$\hat{J}_w$ si può ottimizzare numericamente rispetto al vettore w mediante algoritmi quasi-Newton, così da ottenere il vettore combinato w ottimale.

L'approccio non parametrico basato sul lisciamento del nucleo (Kernel Smoothing)

Per i contesti nei quali non vi sono assunzioni di normalità multivariata, si può comunque sviluppare un metodo di derivazione numerica per stimare il vettore w.

Lo stimatore non parametrico dell'indice di Youden per il risultato combinato, che è $\hat{J}_w$ nella (8), contiene alcune funzioni indicatrici *I(.)*, e perciò non è direttamente differenziabile.

Il metodo del lisciamento del nucleo, più noto come "Kernel Smoothing", è un semplice modo per trovare una struttura, all'interno di data sets, senza imporre alcun modello parametrico. La formula dello stimatore della densità basata sul metodo del nucleo è

$$\hat{f}(x; h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}$$

dove K è detto nucleo (kernel) ed è una funzione che deve soddisfare $\int K(x)dx = 1$, mentre h è un numero positivo spesso chiamato "bandwidth" ossia l'ampiezza della banda. Spesso K è scelto in modo da essere una funzione di densità di probabilità unimodale e simmetrica rispetto lo zero[12].

Al fine di ottenere una versione di $\hat{J}_w$ lisciata e differenziabile rispetto a w, Yin e Tian[2] applicano il nucleo Gaussiano per lisciare le funzioni di distribuzione nelle popolazioni di malati e sani come segue

$$\hat{F}_{Z_g}^{KS}(t) = \frac{1}{n_g} \sum_{i=1}^{n_g} \Phi\left(\frac{t - Z_{gi}}{h_{Z_g}}\right), \; g = 1, 2 \tag{16}$$

dove l'ampiezza di banda è

$$h_{Z_g} = 0.9 min\{SD(Z_g), IQR(Z_g)/1.34\}\, n_g^{-0.2}$$

e $SD(Z_g)$ e $IQR(Z_g)$ sono la deviazione standard e lo scarto interquartile del risultato combinato per le due popolazioni. La formula per il bandwidth è stata proposta da Silverman[13].

Dal lisciamento del nucleo quindi si ottiene la seguente stima dell'Indice di Youden

$$\hat{J}_w^{KS} = \hat{F}_{Z_2}^{KS}(\hat{c}_w^{KS}) - \hat{F}_{Z_1}^{KS}(\hat{c}_w^{KS}) \tag{17}$$

dove $\hat{c}_w^{KS}$ denota il cut-point ottimale ottenuto attraverso l'Indice di Youden per il risultato combinato con la tecnica del kernel smoothing.

A differenza del precedente metodo parametrico, dove $\hat{c}_w$ è una funzione esplicita del vettore dei coefficienti w, $\hat{c}_w^{KS}$ è valutato contemporaneamente con w attraverso alcuni algoritmi di derivazione. Il vettore dei parametri sconosciuto da ottimizzare è

$$w^* = (c_w^{KS}, w^T)^T$$

Dopo il lisciamento del nucleo, $\hat{J}_w^{KS}$ è una funzione continua esplicita di w^*, quindi differenziabile rispetto w e c_w^{KS}. Inoltre, in modo analogo al metodo parametrico precedente, si può ottimizzare numericamente la funzione $\hat{J}_w^{KS}$ rispetto a w^* usando algoritmi quasi-Newton.

CAPITOLO 2.

2.1 Confronto tra i quattro metodi proposti attraverso uno studio di simulazione

Yin e Tian[2] hanno eseguito degli studi di simulazione per confrontare i quattro metodi proposti basati sull'indice di Youden che sono, ricordandoli brevemente:

- l'approccio parametrico sotto normale multivariata (MVN);
- l'approccio non parametrico basato sul metodo del nucleo (KS);
- l'approccio non parametrico del min-max (MMX);
- l'approccio non parametrico graduale verso il basso (SWD).

Per far ciò, sono stati presi in considerazione ampi intervalli di distribuzioni congiunte di quattro marcatori biologici diagnostici:

(i) distribuzioni normali multivariate con uguale varianza;
(ii) distribuzioni normali multivariate con varianza diversa;
(iii) distribuzioni esponenziali multivariate;
(iv) una distribuzione congiunta con diverse distribuzioni marginali (chi-quadro, normale, gamma, esponenziale) per ciascun biomarcatore generato usando una copula normale con correlazione intercambiabile.

Per ciascuna distribuzione congiunta sono stati scelti due set di parametri e, per ogni configurazione di simulazione, sono stati generati 10.000 campioni casuali Monte Carlo dalla distribuzione congiunta sottostante al campione di dimensione

$$(n_1, n_2) = (10, 20), (30, 30), (50, 30), (50,50)$$

per il gruppo di malati e per quello dei non malati.

Senza analizzare nel dettaglio come sono stati generati i dati con le diverse distribuzioni congiunte, si possono fare delle considerazioni in base ai risultati esposti di seguito nelle tabelle, ciascuna delle quali si riferisce alla relativa distribuzione adottata.

Table I. Mean Youden index and probability of obtaining largest Youden index for different combination methods under multivariate normal distributions with equal variance.

Normal means $(\eta_{11}, \eta_{12}, \eta_{13}, \eta_{14})$	Sample sizes (n_1, n_2)	Mean Youden index				Probability of largest Youden index			
		MVN	*KS*	*MMX*	*SWD*	*MVN*	*KS*	*MMX*	*SWD*
$\Sigma_1 = \Sigma_2 = 0.7I_{4\times4} + 0.3J_{4\times4}$ (low correlation)									
(0.2, 0.5, 1.0, 0.7)	(10, 20)	0.6350	0.6815	0.5718	0.6885	0.1108	0.3850	0.0950	0.4092
	(30, 30)	0.5505	0.5915	0.4848	0.5932	0.0858	0.4425	0.0542	0.4175
	(50, 30)	0.5284	0.5735	0.4650	0.5662	0.0583	0.5458	0.0600	0.3358
	(50, 50)	0.5013	0.5348	0.4374	0.5330	0.0642	0.4883	0.0442	0.4033
(0.4, 1.0, 1.5, 1.2)	(10, 20)	0.6382	0.6920	0.5460	0.6655	0.1450	0.4475	0.0800	0.3275
	(30, 30)	0.5402	0.5862	0.4630	0.5663	0.0858	0.5342	0.0383	0.3417
	(50, 30)	0.5381	0.5738	0.4426	0.5608	0.0875	0.5025	0.0200	0.3900
	(50, 50)	0.5093	0.5407	0.4209	0.5293	0.1025	0.4575	0.0150	0.4250
$\Sigma_1 = \Sigma_2 = 0.5I_{4\times4} + 0.5J_{4\times4}$ (medium correlation)									
(0.2, 0.5, 1.0, 0.7)	(10, 20)	0.6742	0.7188	0.5365	0.6518	0.2129	0.5471	0.0488	0.1912
	(30, 30)	0.5785	0.6195	0.4502	0.5630	0.2175	0.6475	0.0242	0.1108
	(50, 30)	0.5748	0.6058	0.4379	0.5591	0.2117	0.6217	0.0000	0.1667
	(50, 50)	0.5461	0.5816	0.4206	0.5301	0.1650	0.7150	0.0050	0.1150
(0.4, 1.0, 1.5, 1.2)	(10, 20)	0.7835	0.8158	0.7200	0.8075	0.1567	0.4167	0.1017	0.3250
	(30, 30)	0.7060	0.7393	0.6415	0.7270	0.0933	0.5558	0.0600	0.2908
	(50, 30)	0.7064	0.7401	0.6364	0.7287	0.0683	0.6458	0.0125	0.2733
	(50, 50)	0.6878	0.7139	0.6208	0.7070	0.0767	0.5875	0.0233	0.3125
$\Sigma_1 = \Sigma_2 = 0.3I_{4\times4} + 0.7J_{4\times4}$ (large correlation)									
(0.2, 0.5, 1.0, 0.7)	(10, 20)	0.7888	0.8068	0.6970	0.7958	0.2150	0.3600	0.0717	0.3533
	(30, 30)	0.7087	0.7287	0.6170	0.7202	0.1588	0.4479	0.0362	0.3571
	(50, 30)	0.6947	0.7186	0.6071	0.7155	0.0967	0.4517	0.0200	0.4317
	(50, 50)	0.6752	0.6923	0.5864	0.6882	0.1446	0.4396	0.0179	0.3979
(0.4, 1.0, 1.5, 1.2)	(10, 20)	0.8110	0.8338	0.6908	0.7790	0.2221	0.4871	0.0729	0.2179
	(30, 30)	0.7400	0.7632	0.6213	0.7117	0.2217	0.5833	0.0167	0.1783
	(50, 30)	0.7321	0.7506	0.6074	0.7227	0.1950	0.5625	0.0050	0.2375
	(50, 50)	0.7120	0.7293	0.5742	0.6826	0.1904	0.6504	0.0188	0.1404

MVN, derivation-based multivariate normal approach; *KS*, derivation-based kernel-smoothing non-parametric method; *MMX*, min–max non-parametric approach; *SWD*, stepwise non-parametric approach with downward direction.

Table II. Simulation results for mean Youden index and probability of obtaining largest Youden index for different combination methods under multivariate normal distributions ($\Sigma_1 = 0.3I_{4\times4} + 0.7J_{4\times4}$ and $\Sigma_2 = 0.7I_{4\times4} + 0.3J_{4\times4}$).

Normal means $(\eta_{11}, \eta_{12}, \eta_{13}, \eta_{14})$	Sample sizes (n_1, n_2)	Mean Youden index				Probability of largest Youden index			
		MVN	*KS*	*MMX*	*SWD*	*MVN*	*KS*	*MMX*	*SWD*
(0.2, 0.5, 1.0, 0.7)	(10, 20)	0.6561	0.7043	0.5920	0.6726	0.1728	0.4285	0.1472	0.2515
	(30, 30)	0.5533	0.5945	0.5095	0.5735	0.1762	0.4075	0.1115	0.3048
	(50, 30)	0.5333	0.5685	0.4879	0.5575	0.1620	0.3893	0.0843	0.3643
	(50, 50)	0.5140	0.5469	0.4721	0.5325	0.1422	0.4778	0.0915	0.2885
(0.4, 1.0, 1.5, 1.2)	(10, 20)	0.7906	0.8184	0.6899	0.8088	0.1948	0.3792	0.0755	0.3505
	(30, 30)	0.7041	0.7289	0.6024	0.7170	0.1557	0.4517	0.0427	0.3500
	(50, 30)	0.6920	0.7150	0.5927	0.7107	0.1037	0.4467	0.0330	0.4167
	(50, 50)	0.6752	0.6972	0.5717	0.6897	0.1085	0.4555	0.0215	0.4145

MVN, derivation-based multivariate normal approach; *KS*, derivation-based kernel-smoothing non-parametric method; *MMX*, min–max non-parametric approach; *SWD*, stepwise non-parametric approach with downward direction.

Table III. Mean Youden index and probability of obtaining largest Youden index for different combination methods under multivariate exponential distributions.

Exponential means	Sample sizes	Mean Youden index				Probability of largest Youden index			
$(\eta_{11}, \eta_{12}, \eta_{13}, \eta_{14})$	(n_1, n_2)	*MVN*	*KS*	*MMX*	*SWD*	*MVN*	*KS*	*MMX*	*SWD*
(0.30, 0.35, 0.4, 0.30)	(10, 20)	0.6923	0.7578	0.6290	0.7682	0.1048	0.3422	0.0805	0.4724
	(30, 30)	0.6099	0.6712	0.5505	0.6911	0.0351	0.2816	0.0519	0.6314
	(50, 30)	0.6025	0.6535	0.5362	0.6799	0.0201	0.1730	0.0316	0.7753
	(50, 50)	0.5900	0.6437	0.5164	0.6661	0.0181	0.1974	0.0172	0.7673
(0.35, 0.45, 0.45, 0.35)	(10, 20)	0.8138	0.8490	0.6992	0.8639	0.1678	0.3060	0.0368	0.4894
	(30, 30)	0.7399	0.7829	0.6243	0.8011	0.0577	0.2752	0.0110	0.6561
	(50, 30)	0.7324	0.7682	0.6129	0.7896	0.0323	0.1865	0.0079	0.7732
	(50, 50)	0.7232	0.7617	0.5952	0.7802	0.0267	0.2033	0.0018	0.7683

MVN, derivation-based multivariate normal approach; *KS*, derivation-based kernel-smoothing non-parametric method; *MMX*, min–max non-parametric approach; *SWD*, stepwise non-parametric approach with downward direction.

Table IV. Mean Youden index and probability of obtaining largest Youden index for different combination methods under multivariate chi-squared/normal/gamma/exponential distributions with normal copula.

Parameter	Sample sizes	Mean Youden index				Probability of largest Youden index			
	(n_1, n_2)	*MVN*	*KS*	*MMX*	*SWD*	*MVN*	*KS*	*MMX*	*SWD*
$N(0.3, 1)$	(10, 20)	0.5590	0.6873	0.4311	0.7197	0.0641	0.3478	0.0191	0.5690
$\Gamma(0.4, 1)$	(30, 30)	0.4704	0.6156	0.3203	0.6553	0.0229	0.2797	0.0018	0.6955
	(50, 30)	0.4651	0.6095	0.2985	0.6519	0.0125	0.2056	0.0000	0.7819
	(50, 50)	0.4533	0.5979	0.2704	0.6295	0.0152	0.2553	0.0000	0.7295
$N(0.6, 1)$	(10, 20)	0.7710	0.8515	0.5019	0.8450	0.1366	0.4435	0.0041	0.4159
$\Gamma(0.8, 1)$	(30, 30)	0.7295	0.8071	0.4046	0.8069	0.0760	0.4559	0.0000	0.4681
	(50, 30)	0.7290	0.8013	0.3890	0.8081	0.0519	0.3920	0.0000	0.5561
	(50, 50)	0.7196	0.7911	0.3653	0.7928	0.0547	0.4390	0.0000	0.5062

MVN: derivation-based multivariate normal approach;
KS: derivation-based kernel-smoothing non-parametric method;
MMX: min–max non-parametric approach;
SWD: stepwise non-parametric approach with downward direction.

Osservando i valori ottenuti in ciascuna tabella, si nota che l'approccio non parametrico graduale verso il basso (SWD) è superiore rispetto agli altri tre metodi per campioni che sono distribuiti molto diversamente da distribuzioni normali multivariate e che il metodo non parametrico basato sul lisciamento del nucleo (KS) funziona meglio per dati normali multivariati. Pertanto, per distribuzioni normali multivariate è preferibile utilizzare il metodo "Kernel Smoothing" mentre in assenza di assunzioni di normalità è raccomandato l'impiego del metodo "stepwise-down" (SWD).

Il fatto che l'approccio parametrico sotto normale multivariata sia inferiore rispetto alla tecnica del lisciamento del nucleo potrebbe essere dovuto a un errore di ri-sostituzione delle stime. Un altro possibile motivo è che, mentre per metodi parametrici il cut-point ottimale è una funzione di combinazione di coefficienti, per i

metodi non parametrici, come il Kernel Smoothing, il valore di soglia viene cercato numericamente attraverso la combinazione dei coefficienti, godendo quindi di maggior flessibilità. Inoltre, la ragione per cui il metodo del nucleo per dati non distribuiti normalmente è meno soddisfacente del metodo graduale verso il basso è dovuta all'utilizzo del nucleo Gaussiano, il più diffuso per questo tipo di procedura. Non si esclude quindi che, utilizzando altre funzioni del nucleo, la situazione non possa cambiare e addirittura capovolgersi.

Per quanto riguarda il metodo del min-max, i risultati di simulazione ne hanno dimostrato la minor capacità rispetto agli altri metodi che utilizzano tutte le informazioni di tutti i marcatori, e non solo il minimo e il massimo.

2.2 Preferenza dell'indice di Youden rispetto l'AUC

L'AUC è stata largamente utilizzata come funzione obiettivo per combinare marcatori. L'idea di massimizzare l'indice di Youden è stata proposta con lo scopo di ottenere un marcatore combinato con migliore accuratezza diagnostica.

Perciò è di grande interesse confrontare i metodi di combinazione proposti basati sull'indice di Youden con quelli basati sull'AUC.

Di seguito viene presentata una tabella che fornisce le medie dei tassi di classificazione globale corretta, la sensibilità e la specificità per i risultati (test) combinati basati sull'indice di Youden C_J e sull'AUC C_A. I risultati sono ottenuti mediante 10.000 simulazioni di dati generati da normale multivariata in cui vengono considerati diversi valori per medie e varianze. In particolare, il vettore delle medie per il gruppo dei sani è fissato pari a $\eta_2 = (0,0,0,0)^T$. Per il gruppo dei malati ci sono due configurazioni per la media: (i) $\eta_1 = (0.2, 0.2, 0.5, 0.5)^T$ e (ii) $\eta_1 = (0.1, 0.3, 0.4, 0.6)^T$. Le matrici di varianza relative alle popolazioni, malata e sana, sono $\Sigma_1 = \gamma\, I_{4\times4} + (1-\gamma)\, J_{4\times4}$ e $\Sigma_2 = (1-\gamma) I_{4\times4} + \gamma\, J_{4\times4}$ con $\gamma = 0.1, 0.5, 0.9$

Per quanto riguarda il metodo di combinazione basato sull'indice di Youden, viene considerato quello che utilizza il Kernel Smoothing, che come visto in precedenza, fornisce il risultato combinato migliore sotto normale multivariata.

Indicando con w_J il vettore dei coefficienti ottenuti massimizzando l'indice di Youden e con w_A quello ottenuto massimizzando l'AUC, per ogni simulazione sono stati calcolati empiricamente sensibilità, specificità e tasso di classificazione globale corretta al valore di soglia ottimale per il risultato combinato Z_g $(g = 1, 2)$ basati su w_J e su w_A.

Table V. Comparison of optimal linear combination methods based on Youden index and AUC combination criteria.

		$\eta_1 = (0.2, 0.2, 0.5, 0.5)^T$						$\eta_1 = (0.1, 0.3, 0.4, 0.6)^T$					
		Total rate		Specificity		Sensitivity		Total rate		Specificity		Sensitivity	
Parameter	Sample size	C_J	C_A	C_J	C_A	C_J	C_A	C_J	C_A	C_J	C_A	C_J	C_A
$\gamma = 0.1$	(10, 10)	1.6905	1.6006	0.7588	0.7139	0.9317	0.8867	1.7053	1.6192	0.7689	0.7223	0.9364	0.8969
	(10, 20)	1.6634	1.5701	0.7616	0.7275	0.9018	0.8426	1.6801	1.5920	0.7685	0.7434	0.9116	0.8486
	(30, 30)	1.5312	1.4332	0.5955	0.6187	0.9357	0.8144	1.5427	1.4507	0.6097	0.6220	0.9330	0.8287
	(50, 30)	1.5006	1.4009	0.5688	0.6130	0.9318	0.7879	1.5183	1.4294	0.5818	0.6261	0.9366	0.8033
	(50, 50)	1.4807	1.3822	0.5430	0.5944	0.9377	0.7879	1.5014	1.4047	0.5560	0.5933	0.9455	0.8114
$\gamma = 0.5$	(10, 10)	1.6516	1.5788	0.8182	0.7471	0.8334	0.8317	1.6605	1.5941	0.8161	0.7579	0.8444	0.8362
	(10, 20)	1.5876	1.5220	0.8294	0.7793	0.7582	0.7427	1.5954	1.5291	0.8321	0.7746	0.7633	0.7545
	(30, 30)	1.4601	1.4043	0.7187	0.678	0.7414	0.7263	1.4713	1.4162	0.7229	0.6765	0.7484	0.7397
	(50, 30)	1.4349	1.3819	0.6989	0.6777	0.7359	0.7042	1.4453	1.3940	0.7052	0.6739	0.7401	0.7201
	(50, 50)	1.3936	1.3503	0.6814	0.6522	0.7122	0.6981	1.4150	1.3730	0.6992	0.6655	0.7158	0.7075
$\gamma = 0.9$	(10, 10)	1.6894	1.6087	0.9102	0.8193	0.7792	0.7894	1.6985	1.6127	0.9162	0.8336	0.7823	0.7791
	(10, 20)	1.6185	1.5337	0.9278	0.8296	0.6907	0.7042	1.6245	1.5449	0.9342	0.8493	0.6904	0.6956
	(30, 30)	1.5269	1.4370	0.8999	0.7699	0.6270	0.6671	1.5411	1.4525	0.9051	0.7835	0.6360	0.6690
	(50, 30)	1.5127	1.4187	0.9103	0.7775	0.6024	0.6412	1.5342	1.4449	0.9203	0.8060	0.6140	0.6389
	(50, 50)	1.4811	1.3832	0.9081	0.7507	0.5730	0.6325	1.4976	1.4055	0.9228	0.7800	0.5749	0.6256

C_J, combination method based on Youden index (J); C_A, combination method based on AUC.
Total rate: the maximum total correct classification rate an linear combination can achieve, obtained by the sum of sensitivity and specificity at the optimal diagnostic threshold associated with Youden index.

La tabella mostra che il risultato combinato basato sull'indice di Youden ha un maggiore tasso di corretta classificazione alla soglia ottimale, rispetto a quello basato sull'AUC, in tutti i casi analizzati. Quando $\gamma = 0.1$, il risultato di combinazione con l'indice di Youden è migliore rispetto a quello con l'AUC in termini di tasso di classificazione globale corretta, e in particolare supera il valore di sensibilità ottenuto con l'AUC. Per esempio, quando i campioni casuali sono di ampiezza $(50, 50)$ e η_1 ha la forma $(0.2, 0.2, 0.5, 0.5)^T$, la sensibilità è 0.94 per la combinazione con l'indice di Youden e 0.79 per quella con l'AUC. Anche quando $\gamma = 0.9$ la performance del metodo di combinazione dell'indice di Youden è migliore, sia per quanto riguarda il tasso di classificazione sia per quanto riguarda la specificità. Per $\gamma = 0.5$, la combinazione dell'indice di Youden è leggermente preferibile rispetto l'AUC in tutti i parametri calcolati.

Per concludere quindi, in base alle osservazioni appena riportate, si può sostenere che il metodo di combinazione lineare che utilizza l'indice di Youden come funzione obiettivo produce un marcatore combinato con un'accuratezza diagnostica migliore, se comparato con un altro che si basa invece sull'AUC.

CAPITOLO 3.

Applicazione: Distrofia Muscolare di Duchenne

La distrofia muscolare di Duchenne è una malattia X-linked recessiva (codificata da un gene mutante localizzato sul cromosoma X) del muscolo causata da una mancanza della proteina distrofina. I ragazzi affetti cominciano a manifestare segni di malattia in età precoce, cessano di camminare all'inizio della seconda decade, e di solito muoiono dall'età di 20 anni. Fino a quando il trattamento del difetto genetico di base non sarà disponibile, approcci medici, chirurgici, e riabilitativi possono essere usati per mantenere la funzione motoria del paziente. I corticosteroidi, tra cui prednisone e un composto correlato, deflazacort, hanno recentemente dimostrato di ritardare notevolmente la perdita di forza e funzione muscolare.

In assenza di un test per l'identificazione univoca della distrofia muscolare di Duchenne (DMD), sono necessari metodi per combinazione dei risultati delle prove di singoli test, il più efficaci e razionali possibile. Percy et al.[14] hanno utilizzato la discriminazione logistica per valutare l'efficacia delle misure di siero chinasi della creatina (CK), emopexina (H), piruvato chinasi (PK), e lattato deidrogenasi (LD) e, in varie combinazioni, per identificare i portatori di DMD, analizzando un campione di 127 donne sane e un altro di 63 donne portatrici[15].

Utilizzando lo stesso dataset ho applicato il metodo del lisciamento del nucleo per trovare una possibile combinazione lineare dei quattro biomarcatori utilizzando il programma R. In particolare, è stata creata la seguente funzione, che rappresenta la (17)

```
f<-function(vett,n1,n2,h1,h2,y1,y2)
{a1<-mean(pnorm((vett[1]-vett[2]*y1$CK-vett[3]*y1$H-vett[4]*y1$PK-
vett[5]*y1$LD)/h1))
a2<-mean(pnorm((vett[1]-vett[2]*y2$CK-vett[3]*y2$H-vett[4]*y2$PK-
vett[5]*y2$LD)/h2))
a<- -(a2-a1)
a}
```

dove precedentemente le ampiezze (bandwidth) erano state definite come

```
h1<-(k*min(sd(y1$CK),sd(y1$H),sd(y1$PK),sd(y1$LD))*(n1^(-0.2)))
h2<-(k*min(sd(y2$CK),sd(y2$H),sd(y2$PK),sd(y2$LD))*(n2^(-0.2)))
```

con una variazione rispetto a quelle utilizzate da Yin e Tian[2] che semplifica notevolmente la struttura della funzione obiettivo.
Mediante la procedura optim() la funzione è stata quindi massimizzata e ha fornito i valori dei coefficienti di combinazione e il valore stesso della funzione, ovvero dell'indice di Youden. La tabella di seguito mostra le espressioni delle diverse combinazioni lineari e il relativo valore dell'indice di Youden al variare di k nelle bandwidth.

Valori di K	Combinazione lineare ottimale	Indice di Youden
0.1	– 0.173 CK – 0.357 H – 0.216 PK – 0.027 LD	0.8153
0.3	– 0.453 CK + 0.105 H – 0.334 PK – 0.167 LD	0.6747
0.5	– 0.161 CK – 0.330 H – 0.282 PK – 0.038 LD	0.7893
0.7	– 0.189 CK – 0.398 H – 0.360 PK – 0.049 LD	0.7844
0.9	– 0.209 CK – 0.440 H – 0.395 PK – 0.056 LD	0.7783
1.1	– 0.181 CK – 0.383 H – 0.387 PK – 0.057 LD	0.7621

Dato il valore più elevato dell'indice di Youden (evidenziato in giallo) pari a 0.8153 con k pari a 0.1, si può concludere che l'espressione

$$-0.173\,CK - 0.357\,H - 0.216\,PK - 0.027\,LD$$

rappresenta la combinazione lineare ottimale dei marcatori biologici analizzati per distinguere le madri portatrici da quelle sane, con soglia ottimale pari a -50. Il risultato si può ritenere accettabile in quanto il valore dell'indice di Youden ottenuto nell'applicazione di Yin e Tian[2], che impiegava le bandwidth con espressione più complessa, è pari a 0.8164 quindi simile a quello ottenuto con questa analisi.

BIBLIOGRAFIA

1. Su JL. Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association 1993;88(424):1350-5.

2. Yin JY,Lili Tian. Optimal linear combinations of multiple diagnostic biomarkers based on youden index. Stat Med 2013;33(8):1426-40.

3. Definizione di "epidemiologia" [Internet]. Available from: http://glossario.paginemediche.it/.

4. Bottarelli E, Parodi S. Un approccio per la valutazione della validità dei test diagnostici: Le curve ROC (receiver operating characteristic). Ann.Fac.Medic.Vet.Di Parma 2003;23:49-68.

5. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. Epidemiology 2005;16(1):73-81.

6. Definizione di "marcatore biologico" [Internet]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078627/.

7. Liu, Aiyi Liu, Enrique Schisterman,Yan Zhu. On linear combinations of biomarkers to improve diagnostic accuracy. Stat Med 2005;24(1):37-47.

8. Pepe M, Thompson M. Combining diagnostic test results to increase accuracy. Biostatistics 2000;1:123-40.

9. Liu C, Liu A, Halabi S. A min-max combination of biomarkers to improve diagnostic accuracy. Stat Med 2011;30(16):2005-14.

10. Kang, Le Kang, Chengjie Xiong, Paul Crane,Lili Tian. Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. Stat Med 2012;32(4):631-43.

11. Schisterman EF, Perkins N. Confidence intervals for the youden index and corresponding optimal cut point. 2007;36(3):549-63.

12. Wand MP, Jones MC. The univariate kernel density estimator. In: Kernel smoothing. ; 1995. .

13. Silverman BW. Density estimation for statistics and data analysis. ; 1986. .

14. Percy ME, Andrews DF, Thompson MW. Duchenne muscular dystrophy carrier detection using logistic discrimination: Serum creatine kinase, hemopexin, pyruvate kinase, and lactate dehydrogenase in combination. Am J Med Genet 1982;13(1):27-38.

15. Duchenne Muscular Dystrophy [Internet]. Available from: http://www.jaaos.org/content/10/2/138.short.