UNIVERSITÀ DEGLI STUDI DI MILANO

Relazione conclusiva "ECG analysis for risk prediction via generalized functional regression models"

Marta Oliani Andrea Taverna Patrick Zecchin

Corso: Laboratorio di Modellistica Matematica

Docente: Francesca Ieva

19 giugno 2015

Sommario

L'obiettivo di questo progetto è la costruzione di un classificatore per l'individuazione di patologie cardiache a partire dai segnali elettrocardiografici di un paziente.

Gli ECG sono rappresentati da dati funzionali multivariati. La rappresentazione funzionale consente l'uso di tecniche di *Functional Data Analysis* (FDA), in particolare la *Functional Principal Component Analysis* (FPCA), in grado di sintetizzare i dati attraverso un numero finito di dimensioni.

Nel nostro progetto usiamo gli scores della FPCA come covariate per un modello di regressione logistica con cui miriamo a distinguere i pazienti sani da quelli affetti da una particolare patologia. Per facilitare la separazione del dataset tra i due gruppi di soggetti consideriamo ciascuna patologia separatamente. I nostri esperimenti, ripetuti e ottimizzati con le metodologie consigliate in letteratura, consentono di valutare la separabilità dei dati per diverse patologie.

Indice

1 Aspetti teorici

1.1 FPCA

La Functional Principal Component Analysis (FPCA) è una procedura per dati funzionali analoga alla PCA per dati finito-dimensionali. Essa consente di analizzare la struttura di covarianza dei dati, permettendo di evidenziare features e patterns rilevanti. Permette inoltre di operare una riduzione dimensionale dei dati, selezionando solo un numero limitato di componenti, e in particolare di passare da uno spazio funzionale, caratterizzato da un numero infinito di dimensioni, ad uno di dimensioni finite.
La tecnica che prende il nome di FPCA consiste nel decomporre la matrice di varianza-covarianza tramite i suo autovalori, al fine di trovare le direzioni lungo le quali i dati assumono maggiore variabilità e poter poi effettuare una rotazione degli assi basata su essa.

Consideriamo pertanto la funzione di covarianza

$$\nu(s,t) = \frac{1}{N-1} \sum_i \big[x_i(s) - \bar{x}(s)\big] \cdot \big[x_i(t) - \bar{x}(t)\big]$$

dove con N indichiamo il numero di osservazioni, con s e t due istanti "temporali" in cui valutiamo il nostro dato funzionale, con $x_i(\cdot)$ il valore assunto dall'i-esimo dato funzionale e con $\bar{x}(\cdot)$ il valore assunto dal dato funzionale medio. Di questa matrice andiamo a calcolare gli autovalori μ_l e le relative autofunzioni (o armoniche) ξ_l, che devono rispettare le seguenti condizioni:

- $\int \xi_l(t)^2 dt = 1$;
- $\int \xi_j(t)\xi_l(t)dt = 0 \qquad j = 1, ..., l-1$;
- $\int \nu(s,t)\xi_j(t)dt = \mu_j \xi_j(s)$

Occorre quindi trovare un criterio per stabilire il numero di autofunzioni da trattenere per formare la base ortonormale da utilizzare. Ad esempio, se abbiamo r autovalori non nulli, scegliamo di tenere i primi l se vale la relazione

$$\frac{\sum_{j=1}^{l} \mu_j}{\sum_{j=1}^{r} \mu_j} \geq S$$

cioè se la varianza spiegata è maggiore di S, che rappresenta una soglia percentuale arbitrariamente fissata (solitamente posta intorno al 70-80%). Alternativamente, si può scegliere di trattenere soltanto gli autovalori che corrispondono ad un valore maggiore o uguale a 1, oppure effettuare uno *scree plot*, cioè un grafico i cui punti (i, μ_i) indicano il valore dell'i-esimo autovalore, e trattenere le prime l componenti se il grafico è fortemente decrescente fino all'l-esimo autovalore e successivamente assume un andamento costante o debolmente decrescente.
L'obiettivo finale di questa trasformazione è quello di poter utilizzare i dati all'interno di un modello di regressione logistica. Per questo motivo, è necessario introdurre dei coefficienti c_{ij}, chiamati *scores*, che andranno inseriti come covariate all'interno del modello, definiti come

$$c_{ij} = \int \xi_j(t)[x_i(t) - \bar{x}(t)]dt$$

e che verranno poi associati a ciascuna x_i. Questi valori c_{ij} altro non sono che la proiezione del dato funzionale i sul nuovo asse dato dall'autofunzione j.

1.2 GLM

In presenza di dati categorici come una diagnosi, che dà esito positivo o negativo, utilizziamo una generalizzazione del modello di regressione lineare detta GLM (*Generalized Linear Model*), costituita da tre componenti:

1. **componente casuale**: è data da una variabile Y di cui si osservano N realizzazioni $\{y_1, ..., y_N\}$, che provengono da una distribuzione appartenente ad una famiglia esponenziale, ovvero

   $$f_Y(y_i, \theta_i) = a(\theta_i)\, b(y_i)\, \exp\left\{y_i Q(\theta_i)\right\}$$

   con $Q(\theta_i)$ parametro naturale;

2. **componente sistematica**: è costituita da un vettore $\boldsymbol{\eta} = (\eta_1, ...\eta_N)$ le cui componenti si ottengono dalla combinazione lineare dei valori x_{ij} del predittore j del soggetto i, ossia

   $$\eta_i = \sum_j \beta_j x_{ij}$$

   Ciascun η_i è chiamato predittore lineare;

3. **link function**: serve a connettere la componente casuale con quella sistematica. Indicata con μ_i la media di Y_i il modello connette le due componenti tramite la link function g, monotona e differenziabile, nel modo seguente:

   $$\eta_i = g(\mu_i)$$

In particolare, consideriamo un modello di regressione logistica in cui prendiamo

$$Y \sim Be\big(\pi(x)\big) = \begin{cases} 0, & 1 - \pi(x) \\ 1, & \pi(x) \end{cases}$$

e creiamo la seguente relazione

$$\pi(x) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}$$

che restituisce un numero compreso tra 0 e 1. Otteniamo quindi

$$\frac{\pi(x)}{1 - \pi(x)} = \exp\{\alpha + \beta x\}$$

da cui si ricava infine

$$\log\Big(\frac{\pi(x)}{1 - \pi(x)}\Big) = \alpha + \beta x$$

Abbiamo così costruito un modello in cui β è strettamente collegato alla variazione del rapporto tra la probabilità di successo e quella di insuccesso corrispondente ad una variazione unitaria di x.

Possiamo però notare come in questo modo non otteniamo una stima di Y bensì una funzione del suo valore atteso, in particolare una stima della probabilità π. Per questo motivo è necessario fissare una soglia π_0 e porre $\hat{Y} = 1$ se $\hat{\pi} \geq \pi_0$ e $\hat{Y} = 0$ altrimenti.
La scelta della soglia π_0 viene fatta per provare a controbilanciare il problema del numero di classificati 0 (nel caso in esame, il numero di sani) e di classificati 1 (i malati) da cui il modello "impara", nonché soprattutto per provare a minimizzare gli errori a cui il modello è inevitabilmente soggetto. Infatti, il modello può classificare un dato in modo corretto o errato, andando a creare quattro diversi casi: veri positivi, veri negativi, falsi positivi e falsi negativi. Nel nostro caso specifico di ambito medico avremo in particolare la seguente configurazione:

	predetto malato	predetto sano
realmente malato	veri positivi	falsi negativi
realmente sano	falsi positivi	veri negativi

Naturalmente l'obiettivo è quello di minimizzare l'errore, in particolare è importante ridurre il più possibile il numero di falsi negativi. Per testare la bontà del modello si può effettuare un grafico che prende in considerazione due parametri:

sensitività α**:** data da $P(\hat{Y} = 1|Y = 1)$, rappresenta la probabilità di ottenere un vero positivo;

specificità β**:** data da $P(\hat{Y} = 0|Y = 0)$, rappresenta invece la probabilità di ottenere un vero negativo.

Più precisamente il grafico, che prende il nome di curva ROC (*Receiver Operating Characteristic*), pone lungo l'asse delle ordinate la sensitività e su quello delle ascisse 1-specificità al variare della soglia. Il caso migliore si ottiene per sensitività e specificità massime, ovvero per un punto della curva ROC con ordinata e ascissa molto vicine rispettivamente a 1 e a 0 (il punto A in Figura 1).
In generale non è possibile determinare il punto migliore in modo univoco, ma è possibile scegliere tra diversi compromessi. Riportiamo qui due criteri per la selezione della soglia ottima. Sia $(1 - \beta_\pi, \alpha_\pi)$ un punto della curva ROC.

- soglia di sensitività: si sceglie a priori un valore minimo di sensitività $\underline{\alpha}$ e si seleziona il punto della curva ROC che massimizza la specificità

$$\pi^* = \arg\max_{\pi \in [0,1]} \{\beta_\pi : \alpha_\pi \geq \underline{\alpha}\}$$

- distanza minima dall'utopia: si sceglie il punto della curva ROC che è più vicino al punto di utopia $(0, 1)$, corrispondente a sensitività e specificità massime, secondo la norma euclidea

$$\pi^* = \arg\min_{\pi \in [0,1]} \{(1 - \alpha_\pi)^2 + (1 - \beta_\pi)^2\}$$

Infine, per testare la bontà del modello e dunque la sua capacità predittiva, si provvede a dividere in modo casuale ciascuna tipologia di pazienti in due gruppi

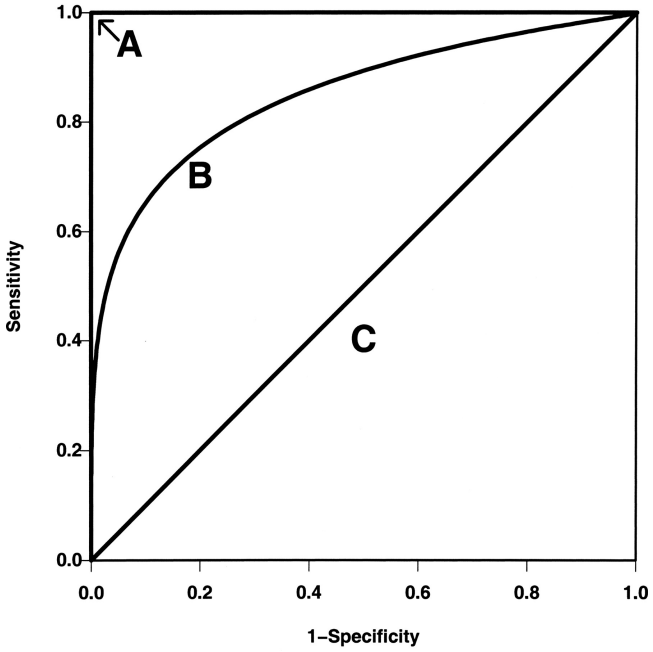

Figura 1: Esempio di curva ROC

training set: contenente la maggior parte della popolazione per la diagnosi considerata, è l'insieme dei dati su cui viene impostato il modello

test set: contenente la rimanente parte delle osservazioni, corrisponde ai pazienti inediti per il modello appena creato sui cui dati si prova dunque a fare delle previsioni.

Questa suddivisione è resa necessaria per un corretto studio della percentuale di misclassificati: valutare l'accuratezza di un modello sugli stessi dati su cui tale modello è stato costruito porterebbe infatti inevitabilmente ad una sottostima dell'errore, cosa che vogliamo evitare. Inoltre, per evitare che la divisione effettuata possa essere in qualche modo "sfortunata", si ripete la procedura un numero di volte sufficiente ad eliminare ogni possibile distorsione.

2 Il progetto sviluppato

2.1 Analisi del problema

Il problema che ci è stato chiesto di affrontare riguarda lo studio di una popolazione di persone che possono risultare affette da malattie cardiovascolari. Il Centro di Emergenza del 118 di Milano, tramite la società "Mortara Rangoni", ha fornito un dataset composto da 8 files, ciascuno dei quali contiene una delle derivazioni principali su cui viene solitamente registrato l'ECG dei pazienti (indicate con $D1$, $D2$, $V1$, $V2$, $V3$, $V4$, $V5$ e $V6$).
Tra i dati a disposizione per ciascun paziente erano compresi età, sesso, diagnosi e *landmarks*, ossia i punti caratteristici del tracciato elettrocardiografico.

Poichè i dati erano già lisciati, l'istante temporale iniziale ($POnset$) è stato considerato pari a 0, mentre quello finale ($TOffset$) è stato posto a 565 per tutti, indistintamente.
Il nostro obiettivo era quello di creare un modello, attraverso il metodo del GLM, che, una volta immesse tutte le informazioni disponibili relative ad un paziente specifico, restituisse una diagnosi rappresentata in forma binaria. Nel nostro caso, se la risposta è uguale a 0 significa che il paziente in questione è considerato Sano, se invece la risposta è 1 allora il paziente è etichettato come Malato.
Più precisamente, le diagnosi possibili erano 11 (il gruppo "Sano" più le 10 differenti patologie), perciò è stato necessario costruire 10 modelli per poter effettuare un confronto tra la possibilità di essere sano e quella di essere affetti dalla malattia i-esima. Tuttavia, poichè non è sufficiente sapere se il paziente è malato ma è necessario conoscere anche da quale patologia è affetto, abbiamo provato, come ultimo passaggio, a creare un modello che confronti le probabilità di essere affetto da ciascuna malattia e assegni infine al paziente quella che risulta avere la probabilità maggiore o, alternativamente, lo classifichi definitivamente come sano.

2.2 Applicazione ad un caso semplificato

Analizziamo ora brevemente il procedere della nostra analisi, volta a trovare un modello predittivo riguardante la diagnosi di malattie cardiovascolari.
Inizialmente abbiamo scelto per semplicità di studiare una sola derivazione principale, in particolare quella relativa a $D1$, e una sola malattia, nello specifico "Blocco di branca destra", in modo da avere una prima idea del comportamento del nostro sistema e procedere successivamente a una sua generalizzazione sull'intero dataset per tutte le diagnosi.

campo	valore
Age	68
Sesso	F
Ponset	0
POffset	73.60915
QQRSOnset	174.0757
Rpeak	214.8592
QRSOffset	267.5792
Tpeak	491.3908
TOffset	565
PAxis	-34
QRSAxis	9
Taxis	8
Diagnosi	”Sano”
Fibrillazione.atriale	N
V1	18.705
V2	15.672
...	...
V1024	23.255

Tabella 1: Prima riga del dataset considerato

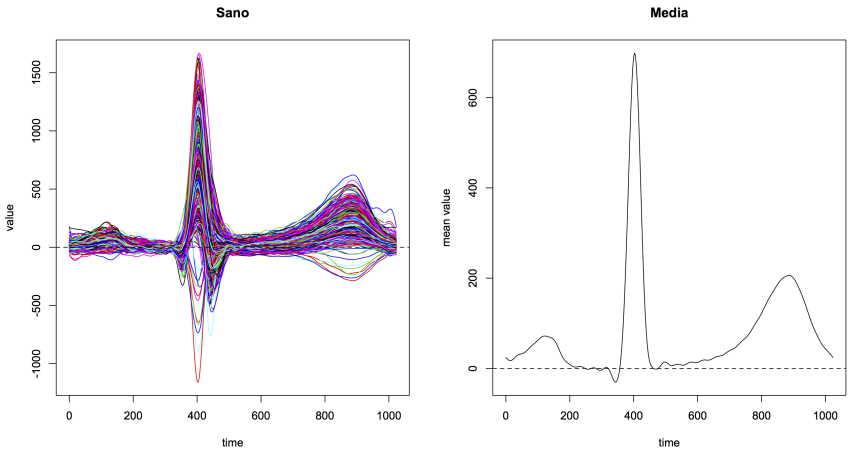

Figura 2: Tracciati del segnale ECG e loro media, pazienti sani, derivazione $D1$

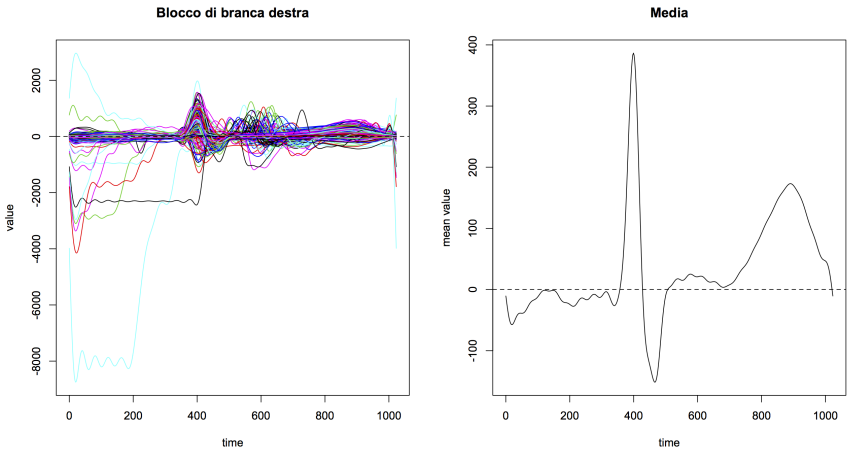

Figura 3: Tracciati del segnale ECG e loro media, pazienti affetti da Blocco di branca destra, derivazione $D1$

Dopo una prima analisi descrittiva dei dati a nostra disposizione, abbiamo iniziato ad eseguire la scomposizione in componenti principali della variabile funzionale costituita dal segnale ECG. Per questa operazione, per quanto non fosse strettamente necessario, poichè i dati fornitici erano già stati lisciati e standardizzati, ci siamo serviti della libreria `FDA`, rilasciata per `R` da *J. O. Ramsay et al.* (vedi [3]). Tramite tale libreria abbiamo ricreato i nostri dati funzionali, sfruttando, in particolare, un sistema di basi di Fourier, a causa della periodicità dei segnali.
Come prima scelta di parametri abbiamo deciso di mantenere 8 componenti principali di cui abbiamo effettuato un grafico relativo ai loro comportamenti (Figura 2.2) e scree plot (Figura 2.2), in modo da poterne osservare la variabilità spiegata.

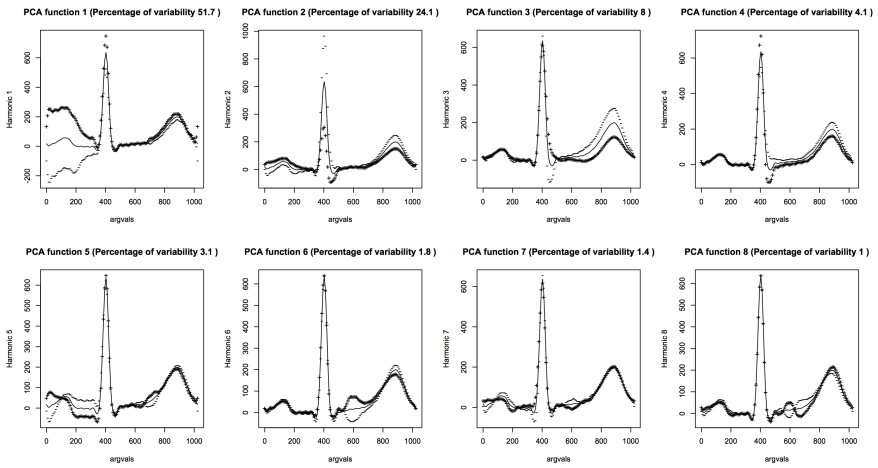

Figura 4: Prime componenti principali per "Sani vs Blocco di branca destra", derivazione $D1$

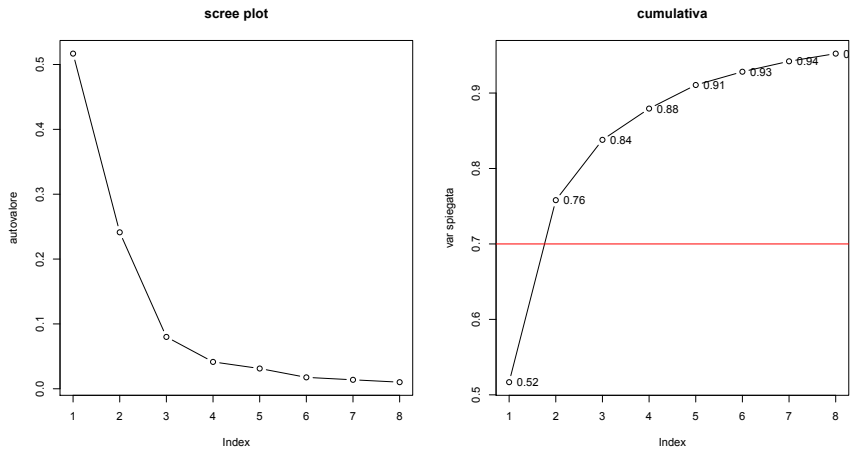

Figura 5: Varianza spiegata delle rime componenti principali per "Sani vs Blocco di branca destra", derivazione $D1$

Già a questo primo livello di analisi possiamo notare (Figura 2.2) che due componenti principali sono sufficienti per spiegare più del 70% della variabilità dell'insieme di dati considerato, mentre lo scree plot ci indicherebbe di porre a tre il numero di componenti da tenere.
Mantenendo per il momento quest'ultima scelta di numero di componenti principali, otteniamo la matrice degli scores, ossia il valore della proiezione di ciascun dato funzionale (il tracciato ECG del singolo paziente) sui tre nuovi assi

	score 1	score 2	score 3
Paziente 1	160.9850	-1107.7695	303.77154
Paziente 2	-655.0137	2875.5500	-578.73629
Paziente 3	259.6181	-545.4188	2065.13823
Paziente 4	215.0350	-1105.3881	963.42275
...	...	...	...

Possiamo quindi utilizzare i dati così ricavati come regressori in un modello lineare generalizzato, assieme all'età e al sesso del paziente. Per un corretto funzionamento del modello abbiamo escluso i soggetti per i quali i dati anagrafici non erano disponibili.

```
glm(formula = factor(record$Diagnosi2) ~
            ris$scores[, 1] + ris$scores[, 2] + ris$scores[, 3] +
            record$Age + record$Sesso,
            family = binomial(link = "logit"),
            data = factor(record$Diagnosi2))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6519  -0.4030  -0.1902  -0.0544   4.0123

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.273e+00  4.994e-01 -18.569  < 2e-16 ***
ris$scores[, 1] -2.696e-04  6.853e-05  -3.934 8.37e-05 ***
ris$scores[, 2]  6.094e-04  5.102e-05  11.945  < 2e-16 ***
ris$scores[, 3] -7.034e-04  6.521e-05 -10.786  < 2e-16 ***
record$Age       1.022e-01  6.365e-03  16.056  < 2e-16 ***
record$SessoM    9.973e-01  1.705e-01   5.851 4.89e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1972.3  on 1996  degrees of freedom
Residual deviance: 1044.5  on 1991  degrees of freedom
AIC: 1056.5

Number of Fisher Scoring iterations: 6
```

Dall'analisi effettuata risulta evidente che, nel caso in esame, tutte le componenti del modello hanno un'elevata significatività.
Inseriamo ora gli scores nel regressore ottenuto, ricavando così la probabilità di ciascun paziente di essere malato, che viene ovviamente rappresentata con un numero compreso tra 0 e 1. Per una corretta interpretazione occorre tenere presente che abbiamo posto a 0 la condizione "Sano" e a 1 quella di "Malato" e, per semplicità, abbiamo posto la soglia a $\pi_0 = 0,5$.

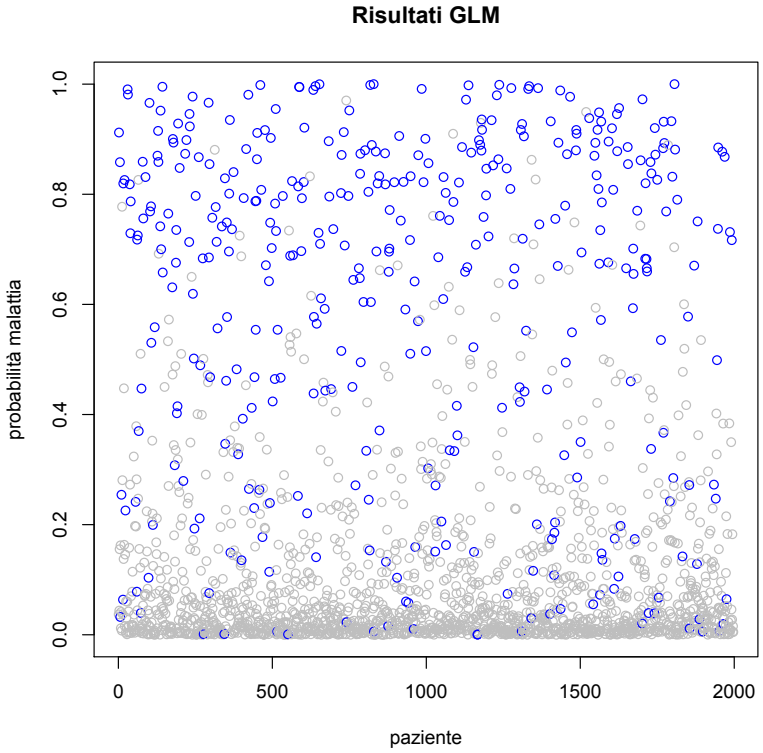

Figura 6: Classificazione dei pazienti ottenuta da GLM (in grigio i soggetti realmente Sani, in blu i Malati)

	predetto malato	predetto sano
realmente malato	264	126
realmente sano	52	1555

Tabella 2: Risultati di una prima regressione logistica per "Sani vs Blocco di branca destra", derivazione $V1$

Dai risultati in tabella si può vedere che, in generale, la percentuale di individui classificati correttamente è molto alta (91%). Tuttavia se osserviamo soltanto i pazienti malati, la percentuale dei ben classificati scende al 68%, che equivale a considerare erroneamente come sane circa un terzo delle persone affette dalla patologia.

Alla luce dei risultati ottenuti possiamo affermare che questo modello sembra non funzionare come vorremmo, in quanto il nostro scopo è quello di minimizzare il numero dei malati ai quali non riesco ad effettuare una diagnosi corretta. Possiamo ipotizzare che ciò sia dovuto anche all'elevato numero di sani sul totale dei pazienti in esame (circa 80%). Questo può facilmente condizionare le

capacità predittive del modello, che apprende molto bene come classificare i sani ma dispone di poche informazioni che permettano una buona classificazione dei malati. Un'altra causa del non corretto funzionamento del modello potrebbe essere dovuta alla presenza di outliers, rappresentati da segnali con picco negativo o con valori iniziali e/o finali eccessivamente distanti dalla media (questi ultimi dovuti al rumore), come si può già vedere dal grafico in Figura 2.2. Non è detto però che essi contribuiscano negativamente, in quanto, se appartenenti alla popolazione malata, potrebbero contribuire a caratterizzarla e se esclusi ridurrebbero i dati a disposizione, penalizzando ulteriormente l'apprendimento delle proprietà specifiche della popolazione malata.
Per le considerazioni effettuate alla fine del caso semplice, una volta passati al caso generale (in cui facciamo uso di tutte le derivazioni) abbiamo deciso di provare ad apportare le seguenti modifiche:

- confrontare la bontà del modello con un numero prefissato di componenti principali con quella ottenuta scegliendo un numero di componenti sufficiente a spiegare una prefissata quantità della variabilità del sistema;
- utilizzare una soglia di classificazione π_0 variabile, che tenga conto della differente proporzione di sani e malati all'interno del campione;
- suddividere il campione in training set e test set, per poter stimare correttamente l'errore di classificazione;
- rimuovere gli outliers identificati con uno dei criteri sopra citati.

Giungiamo in questo modo alla versione definitiva del nostro possibile classificatore, che andiamo a discutere nella sezione successiva.

3 Applicazione al caso generico

La costruzione e verifica dei classificatori si articola nelle seguenti fasi:

1. divisione casuale del dataset in un insieme di training ed uno di test
2. calcolo degli scores della PCA sul dataset completo
3. scelta del numero di componenti principali da utilizzare per ciascuna dimensione
4. generazione del modello di regressione logistica sui dati di training
5. scelta della soglia del classificatore sui dati di training
6. verifica del classificatore sui dati di test con conseguente calcolo della matrice di confusione

I passi 3 e 5 richiedono la scelta di opportuni parametri. Per determinare i valori ottimali di questi parametri e verificare la robustezza dei classificatori ripetiamo lo schema di cui sopra N volte per ciascuna combinazione dei parametri.
I parametri da scegliere sono:

- numero di componenti principali K: come suggerito in [5] conviene considerare gli scores delle prime componenti di ciascuna dimensione. Per semplicità supponiamo di scegliere uno stesso numero K di componenti principali per ciascuna dimensione. Per determinare K consideriamo due criteri:
  - fisso: consideriamo un numero ridotto di componenti principali $\bar{K}$ che riteniamo sufficiente a sintetizzare il dataset
  - variabile: troviamo il valore di K che permetta di spiegare una percentuale di varianza superiore ad una soglia $\underline{\rho}$.
- soglia del classificatore π_0: determina oltre quale valore della probabilità restituita dal modello di regressione il classificatore restituisce esito positivo. Utilizziamo due criteri:
  - fisso: scegliamo a priori una soglia
  - variabile: calcoliamo una soglia che ottimizzi la sensibilità e la specificità del classificatore, in particolare scegliendo π_0 secondo il criterio di "utopia".

Per ogni classificatore abbiamo quindi 4 combinazioni di criteri, ognuno dei quali viene provato N volte con N dataset diversi costruiti per campionamento.

3.1 Risultati delle analisi

Di seguito riportiamo i risultati dei test per tutte le patologie considerate. Abbiamo eseguito i test con la seguente configurazione:

Parametro	Valore	Descrizione
N	100	numero di prove per ciascuna combinazione di parametri
$\underline{\rho}$	70%	percentuale di varianza spiegata per la determinazione del numero di componenti principali nel caso variabile
$\bar{K}$	3	numero PC nel caso fisso
$\bar{\pi}_0$	0.5	soglia classificatore nel caso fisso
θ	85%	frazione del dataset originale utilizzata per costruire il dataset di training.

Per ogni combinazione di criteri per i parametri K e π_0 riportiamo la distribuzione, come boxplot, della sensitività dei classificatori nei test. Nei grafici indichiamo con una linea rossa la soglia della sensitività al 90%, sotto la quale individuiamo le prove "fallite", ovvero i classificatori che non sono risultati abbastanza sensibili.
I risultati mostrano che il nostro modello riesce a classificare efficacemente le diagnosi di Blocco di branca destra, Blocco di branca sinistra e Fibrillazione atriale. Per le altre patologie determinare un buon classificatore risulta molto più complicato poiché i dati sono più difficili da separare nello spazio delle scores. Si vedano ad esempio la Figura 12 e la Figura 10, nonché le successive considerazioni.

L'uso della soglia variabile consente di migliorare la sensibilità dei classificatori, anche se il miglioramento è più significativo per le malattie non ben classificate. Il numero di componenti principali K utilizzato nel modello di regressione non risulta influire significativamente sui risultati dei classificatori. In particolare, ciò significa che il nostro modello permette di ottenere buoni risultati anche usando un numero limitato $\bar{K} = 3$ di componenti principali.

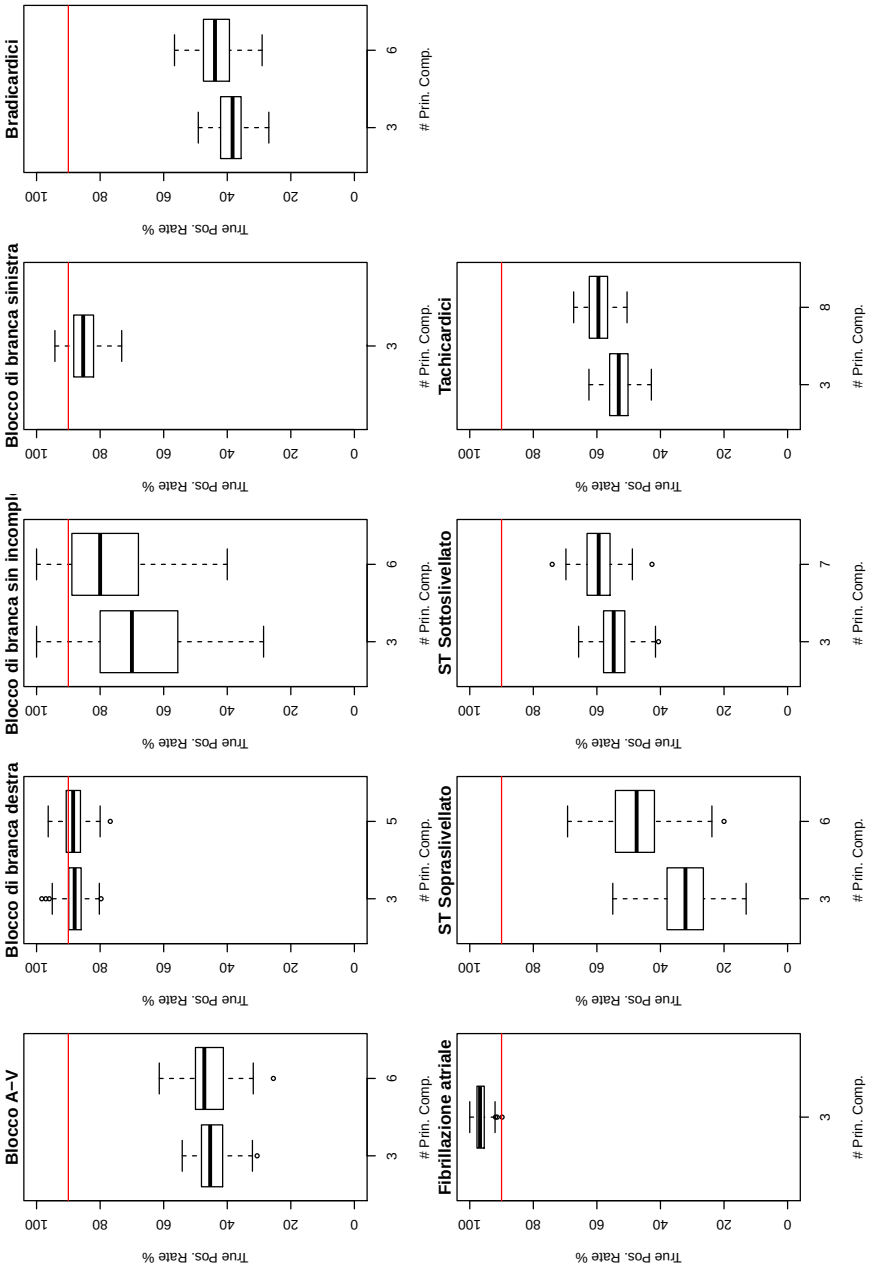

Figura 7: Sensitività dei classificatori per soglia fissa $\pi_0 = \bar{\pi_0}$ al variare di K

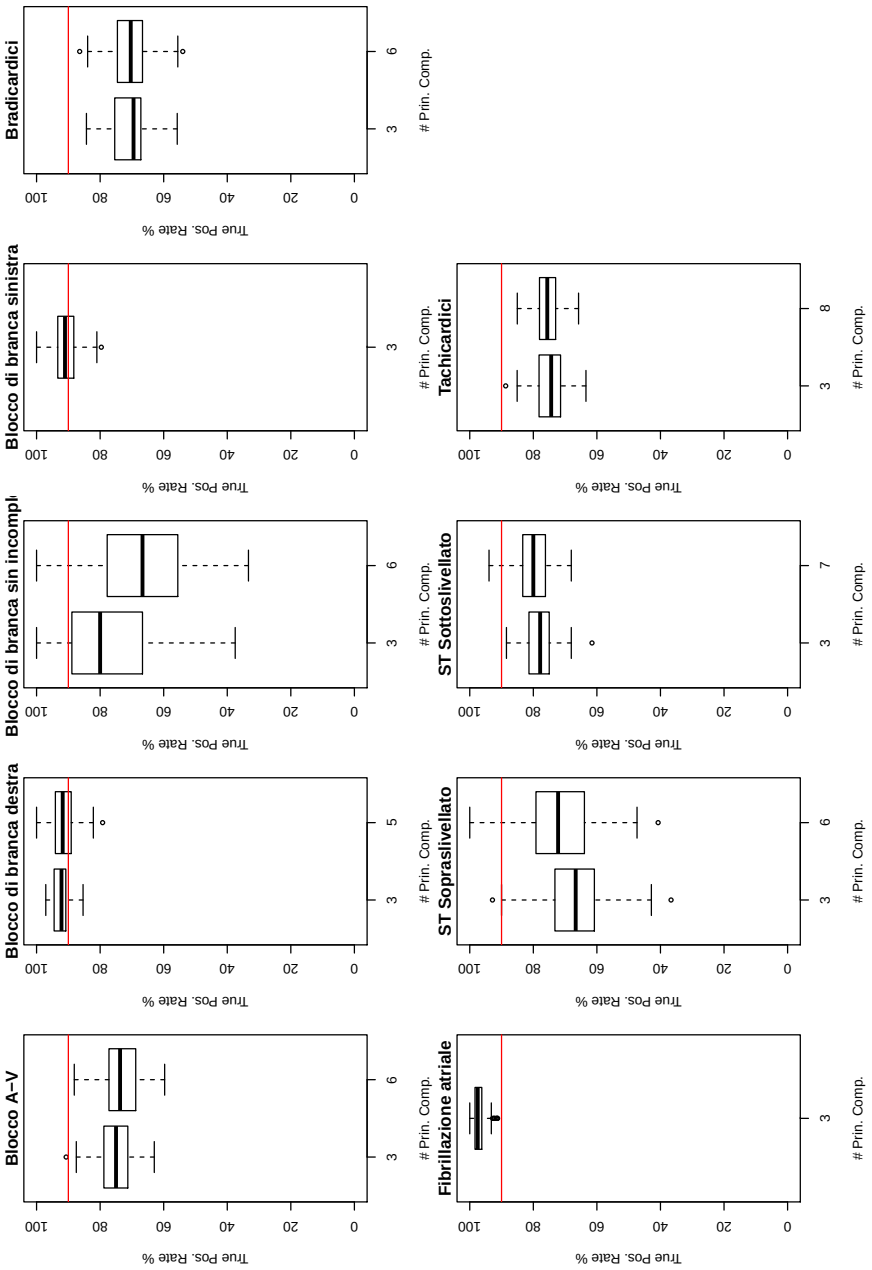

Figura 8: Sensitività dei classificatori per soglia variabile al variare del numero di K

3.2 Analisi di due casi contrapposti

Analizziamo ora, nello specifico, due patologie che consentono di illustrare il comportamento di alcuni parametri del modello sia nel caso in cui la classifi-

cazione sia particolarmente efficiente, che in presenza di un'alta percentuale di misclassificazione.
Un esempio di popolazione che fornisce pessimi risultati è quello dei bradicardici, di cui riportiamo, di seguito, due grafici significativi (Figura 3.2 e 10).

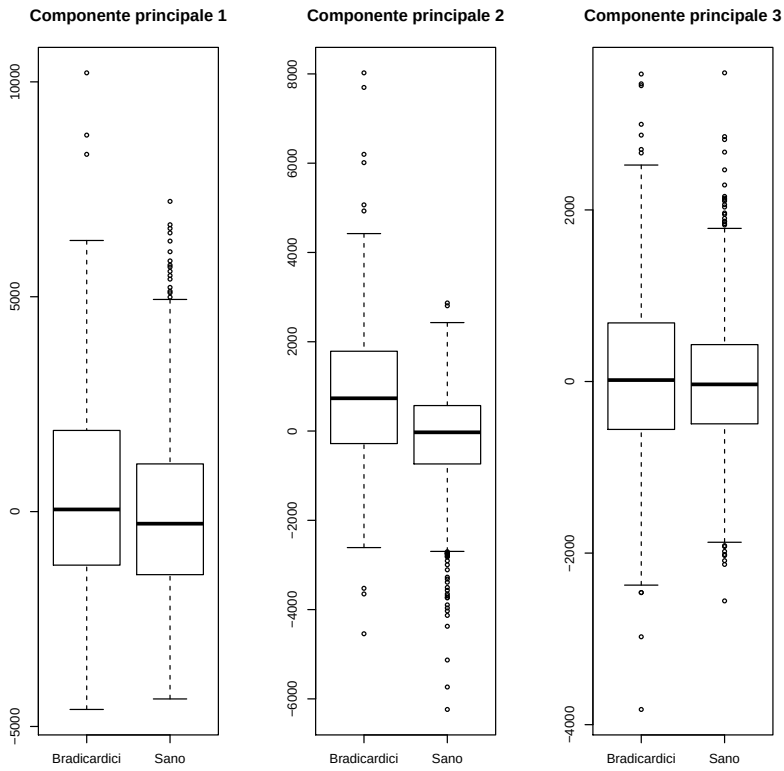

Figura 9: Boxplot delle prime 3 componenti principali per "Sani vs Bradicardici", derivazione $D1$

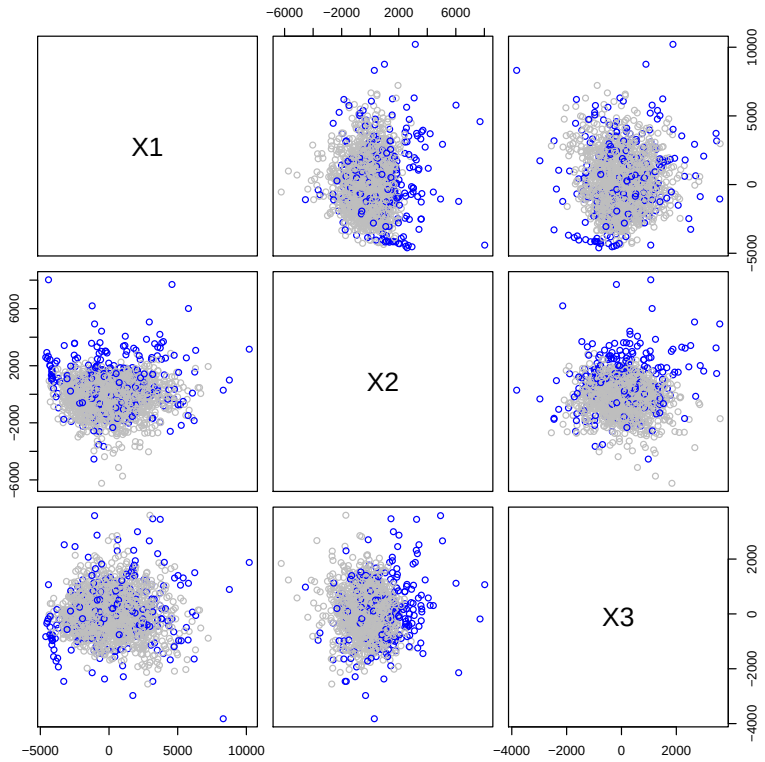

Figura 10: Relazione tra gli scores delle prime 3 componenti principali per "Sani vs Bradicardici", derivazione $D1$ (in blu i Malati, in grigio i Sani)

Dalla sovrapposizione dei boxplot delle componenti principali della popolazione sana e di quella malata possiamo intuire la difficoltà del modello nell'effettuare una corretta classificazione dei pazienti. Questa supposizione è avvalorata anche dalla mancanza di una netta separazione dei dati osservabile nel grafico degli scores (Figura 10).
Prendiamo poi come esempio di modello efficiente quello relativo agli affetti da Blocco di branca destra (Figura 3.2 e 12).

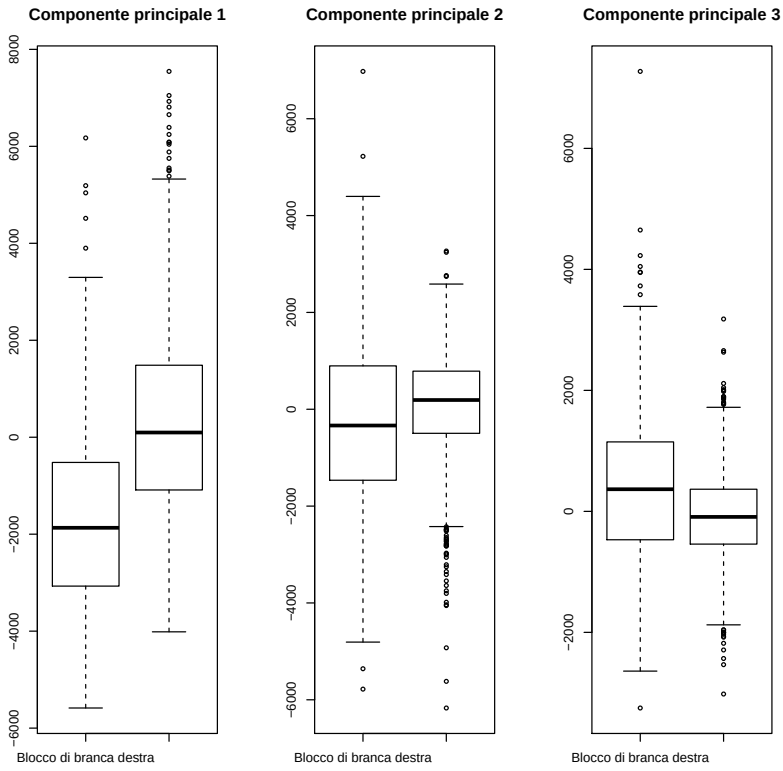

Figura 11: Boxplot delle prime 3 componenti principali per “Sani vs Blocco di branca destra”, derivazione $D1$

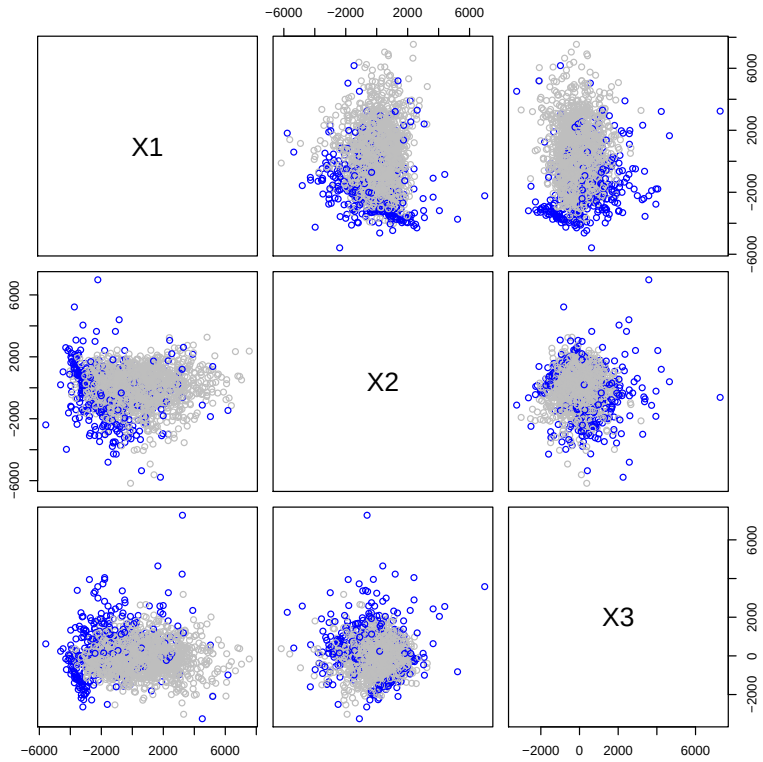

Figura 12: Relazione tra gli scores delle prime 3 componenti principali per "Sani vs Blocco di branca destra", derivazione $D1$ (in blu i Malati, in grigio i Sani)

Si può osservare che, anche in questo caso, i boxplot tendono a sovrapporsi, ad eccezione della prima componente. Tuttavia, poichè essa è ritenuta tra le più significative dal GLM, possiamo aspettarci ugualmente una buona classificazione dei soggetti. Ciò viene inoltre supportato dal grafico degli scores, dove si può vedere la presenza di alcune zone in cui la densità dei malati è decisamente elevata mentre i sani tendono a concentrarsi in altre zone (Figura 12).

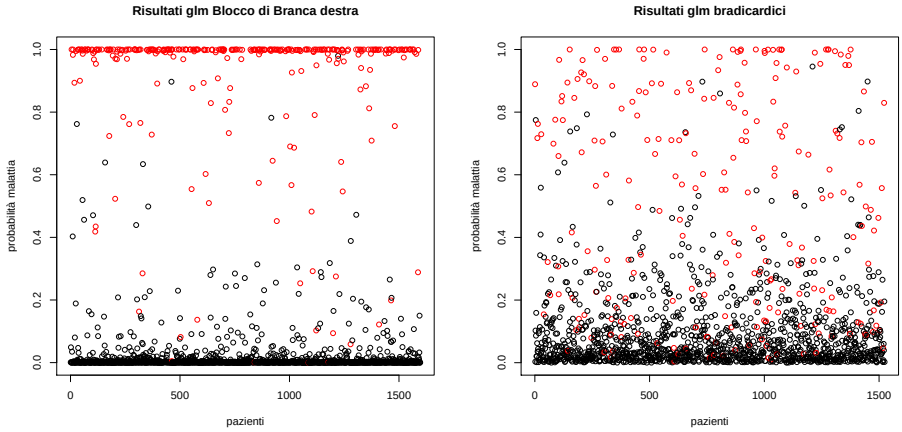

Figura 13: Confronto risultati GLM tra esempio classificato bene (Blocco di branca destra) ed esempio classificato male (Bradicardici), in rosso i pazienti effettivamente malati

Abbiamo provato infine a studiare il comportamento dei p-values delle componenti del modello di regressione per entrambe le patologie precedentemente esaminate. Basandoci anche su quanto scritto in [5], ci aspettiamo che i regressori più significativi siano quelli relativi alle prime componenti di ciascuna derivazione. Notiamo invece che questo non corrisponde a quanto si verifica nelle nostre analisi, ma ciò può essere dovuto ad una peculiarità del nostro sistema.

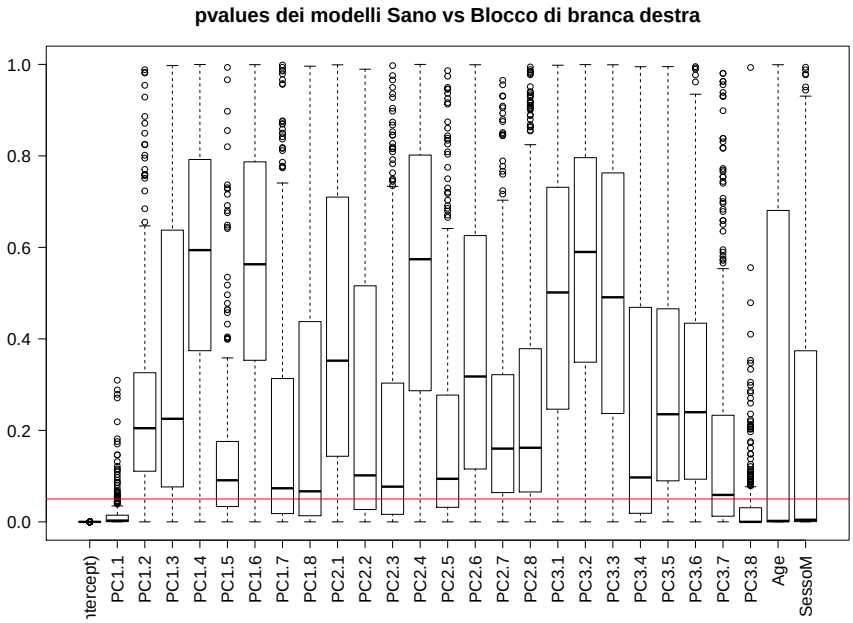

Figura 14: esempio classificati bene, valore pvalues delle pc da glm

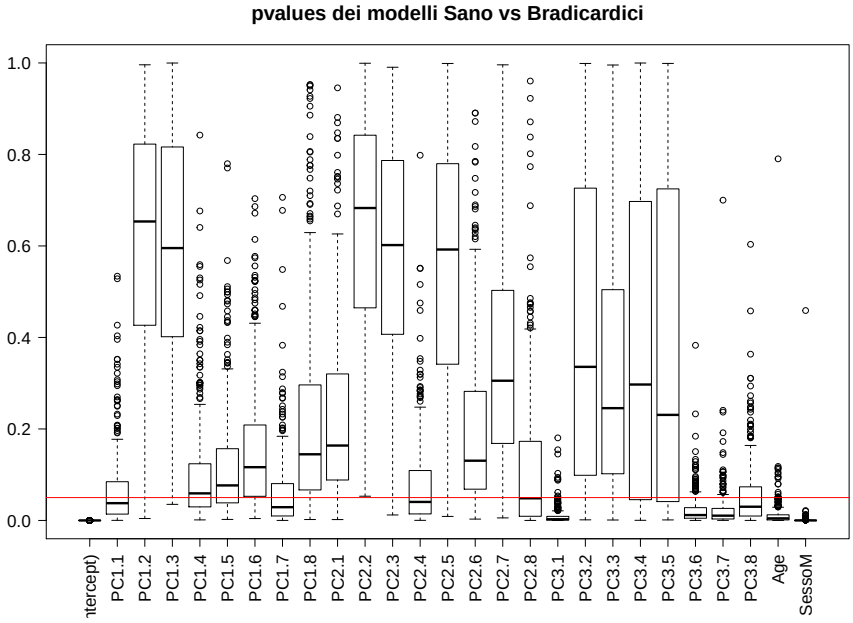

Figura 15: esempio classificati male, valore pvalues delle pc da glm

3.3 Gli outliers

Abbiamo riscontrato la presenza di potenziali outliers nel dataset, si veda ad esempio la Figura 16. Gli outliers nella realtà sono spesso dovuti ad un erroneo posizionamento dei sensori dell'ECG.

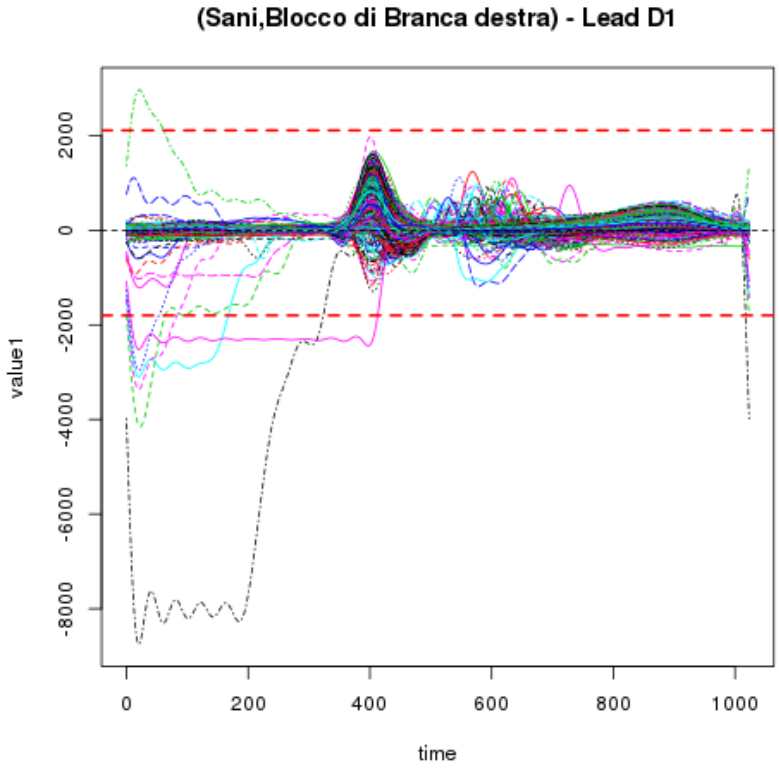

Figura 16: Outliers nella derivativa D1 per il dataset contenente pazienti Sani e con Blocco di Branca destra. Le linee tratteggiate separano i tratti di curva più lontani, appartenenti a potenziali outliers, da quelli più vicini alla media.

La presenza di outliers in un campione può comprometterne l'analisi e può essere quindi vantaggioso separarli dal resto del campione.
L'individuazione di outliers per dati funzionali è un problema aperto, oggetto di ricerca e l'uso di strumenti robusti per il filtraggio degli outliers va oltre lo scopo del nostro progetto. Abbiamo perciò utilizzato un metodo euristico per l'individuazione degli outliers, basato sull'osservazione diretta dei dati. Dato un campione abbiamo rimosso i pazienti le cui derivative assumevano valori in modulo molto grandi nei primi istanti dell'ECG, come suggerirebbe la figura 16. Abbiamo notato tuttavia che questa euristica può portare all'eliminazione di molti pazienti malati se le soglie utilizzate risultano troppo piccole. Occorrerebbe quindi un'analisi più accurata degli outliers per determinare un criterio efficace di filtraggio. D'altra parte, anche disponendo di soglie efficienti, che

filtrino gli outliers eliminando un numero limitato di pazienti, non abbiamo riscontrato miglioramenti significativi nell'accuratezza dei classificatori.

3.4 Proposta per un classificatore multiplo

Possiamo infine combinare più classificatori binari per costruire un classificatore multiplo con il seguente algoritmo. Rappresentiamo ciascun classificatore binario come una funzione

$$\phi(\mathbf{x}|\mathbf{Y}, \pi_{\mathbf{0}}) = (\delta, \pi)$$

dove

$$\delta = \begin{cases} 1 & \pi \geq \pi_0 \\ 0 & \pi < \pi_0 \end{cases}$$

e $\pi = \pi(\mathbf{x})$ è la probabilità che il modello di regressione associa all'unità $\mathbf{x}$ di essere positiva. Ogni classificatore è inoltre asssociato ad un insieme di autofunzioni $\xi = (\xi^l)_{1 \leq l \leq K}$ ottenuto dalla FPCA applicata al dataset di training. Dati n classificatori binari $\{(\phi_i, \xi^i)\}_{i \in I}$, con $I = \{1, \ldots, n\}$, ciascuno utilizzato per separare i pazienti sani da quelli affetti da una particolare patologia, ed una nuova unità statistica $\mathbf{x}'$, il classificatore multiplo opera come segue:

1. per ogni classificatore $i \in I$
   (a) calcola gli scores $\mathbf{c}(\mathbf{x})$ del nuovo paziente tramite regressione sulle autofunzioni ξ^i
   (b) applica il classificatore agli scores $\mathbf{c}(\mathbf{x})$, ottenendo un indicatore $\delta^i_{\mathbf{x}}$ e una probabilità $\pi^i_{\mathbf{x}}$
2. se $\delta^i_{\mathbf{x}} = 0 \ \forall i \in I$ il classificatore determina che il paziente è sano
3. altrimenti il classificatore associa al paziente la patologia corrispondente all'indice $i^* = \underset{i \in I : \delta^i = 1}{\arg\max} \{\pi^i\}$. In caso di parità tra malattie diverse viene scelta quella con indice $i \in I$ minore.

Il classificatore, analogamente al caso binario, ritorna una coppia (Δ, π) dove Δ è l'indice della malattia i prevista, o 0 se il paziente è sano, e $\pi \in [0,1]^n$ è il vettore di probabilità che associa alla patologia i-esima la probabilità che il paziente sia affetto da essa, se il corrispondente classificatore ϕ_i ha dato esito positivo, altrimenti associa una probabilità nulla.

4 Conclusioni

Con questo lavoro abbiamo provato a sviluppare un modello matematico in grado di identificare l'eventuale presenza di una malattia cardiovascolare, basandosi semplicemente sull'osservazione del segnale dell'ECG del paziente. A tal scopo abbiamo fatto uso della FPCA e del GLM, due tecniche statistiche che si sono rivelate semplici da un punto di vista sia implementativo che computazionale, nonchè di facile interpretazione.
Purtroppo il loro utilizzo non ha portato ai risultati sperati, probabilmente a causa della scarsità di informazioni contenute nel segnale ECG che porta ad

una sovrapposizione degli scores generati dalla decomposizione in componenti principali. Siamo infatti riusciti a classificare efficacemente soltanto 3 patologie (Blocco di branca destra, Blocco di branca sinistra e Fibrillazione atriale), mentre le rimanenti diagnosi presentano una percentuale di errore superiore al 20%.
Si tratta quindi di un metodo potenzialmente valido, che ci permetterebbe anche di costruire un classificatore unico per tutte le patologie, ma che risulta invece poco preciso a causa dei dati di partenza.

Riferimenti bibliografici

[1] F. Ieva, A.M. Paganoni (2013), *Risk prediction for myocardial infarction via generalized functional regression models*, Statistical Methods in Medical Research.

[2] J.O. Ramsay, G. Hooker, S. Graves (2009), *Functional Data Analysis with R and MATLAB*, Springer.

[3] J.O. Ramsey, H. Wickham, S. Graves, G. Hooker (2015), *Package 'fda'*, Version 2.4.4, `http://cran.r-project.org/web/packages/fda/fda.pdf`.

[4] A. Agresti (2002), *Categorical Data Analysis*, Second Edition, John Wiley & Sons.

[5] L. M. Sangalli, P. Secchi, S. Vantini, A. Veneziani (2009), *A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery*, Journal of the American Statistical Association March 2009, Vol. 104, No. 485.