ECG ANALYSIS FOR RISK PREDICTION VIA GENERALIZED FUNCTIONAL REGRESSION MODELS

M. Oliani, A. Taverna, P. Zecchin

June 26, 2015

Department of Mathematics, University of Milan

Table of contents

INTRODUCTION

Introduction

- ECGs for a set of patients is given
- ECGs data is composed by 8 different signals for each patient

Goal

Predict the patients' risk for a cardiac pathology

Knowing in advance the risk factors for each different pathology improves response time and quality.

Goal

How we did it: construct a binary classifier for each pathology

- treat each ECG data as functional data
- perform dimensional reduction via *Functional Principal Component Analysis* (FPCA)
- perform classification via regression on a *Generalized Linear Model* (GLM)

Simplifying assumption: we consider each pathology separately from the others, learning to distinguish affected patients from healthy ones.

THEORY

Functional Principal Component Analysis

The purpose of FPCA is to reduce the data dimension, moving from an infinite dimensional space to a finite one.

This method is analogous to the PCA for the multivariate case.
It mainly consists of a selection of the data most variability directions and a consequent axes rotation.

Functional Principal Component Analysis

We consider the covariance function

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

where N is the number of observations, s and t are two temporal instants, $x_i(\cdot)$ the value of the i-th functional data and $\bar{x}(\cdot)$ is the value of the mean functional data.

Functional Principal Component Analysis

We search for eigenvalues μ_l and eigenvectors ξ_l of that matrix such that:

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

Functional Principal Component Analysis

In order to choose how many principal components to keep, we can consider

- the explained variability;
- the scree plot;
- the eigenvalue amount.

Functional Principal Component Analysis

Finally, we have to define some quantities, named *scores*, that will be used as covariates of the model

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

Generalized Linear Models

The Generalized Linear Model (GLM) is a generalization of the linear regression model.

It is characterized by 3 different components:

- random component: $f_Y(y_i, \theta_i) = a(\theta_i)\, b(y_i)\, \exp\left\{y_i Q(\theta_i)\right\}$
- systematic component: $\eta_i = \sum_j \beta_j x_{ij}$
- link function: $\eta_i = g(\mu_i)$

Generalized Linear Models

Let's take

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

Then we obtain the final model

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

called *logistic model*.

This model gives an estimation of π instead of Y, so we put

$$\hat{Y} = \begin{cases} 0, & \text{if } \hat{\pi} < \pi_0 \\ 1, & \text{otherwise} \end{cases}$$

where π_0 is a threshold usually fixed at 0,5.

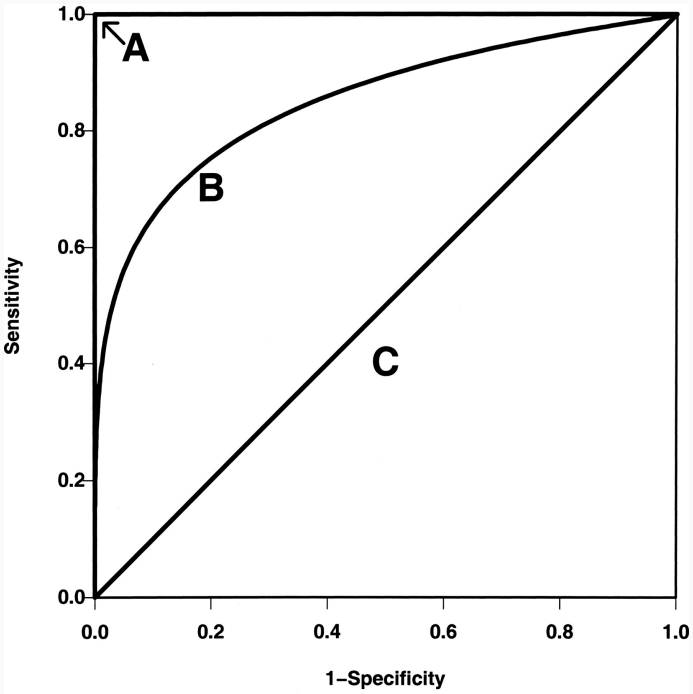
Classification

The threshold π_0 is chosen in order to minimize the misclassification rate

	$\hat{Y} = 1$ (sick)	$\hat{Y} = 0$ (healthy)
$Y = 1$ (sick)	true positive	false negative
$Y = 0$ (healthy)	false positive	true negative

We check two parameters: sensitivity and specificity

Optimal threshold Choice



Point $A = (1-\beta, \alpha) = (0, 1)$ in the ROC plane is the "utopia"

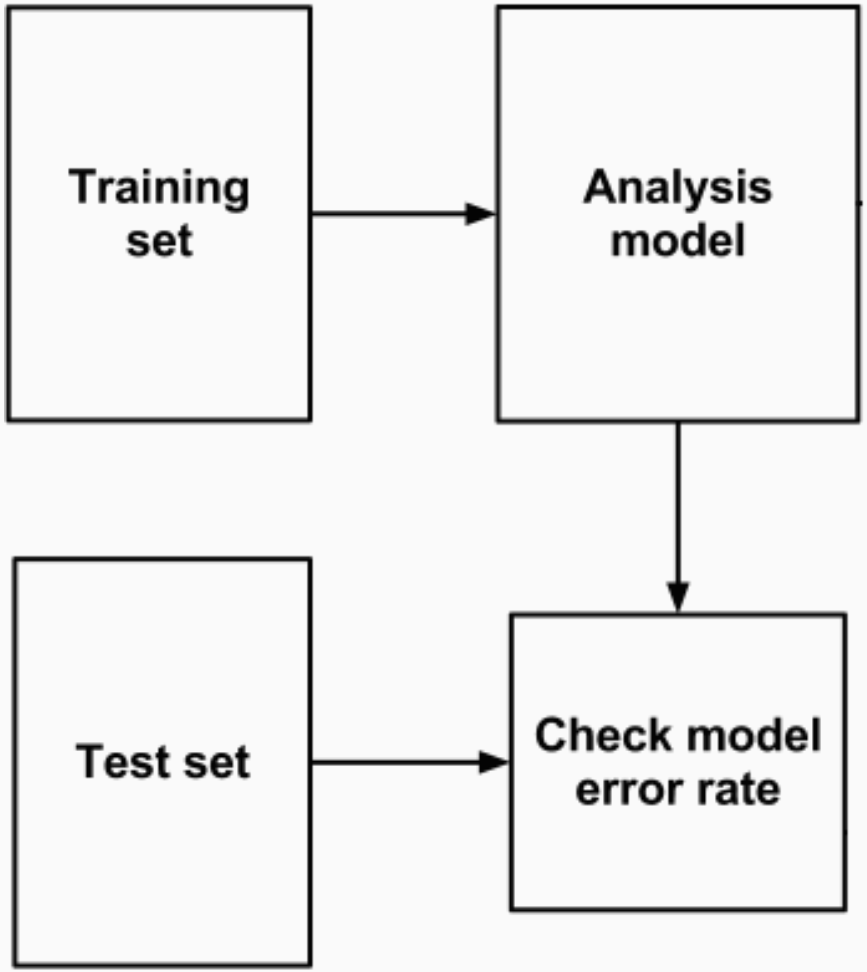
- max. sensitivity and specificity

Chosen threshold is

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

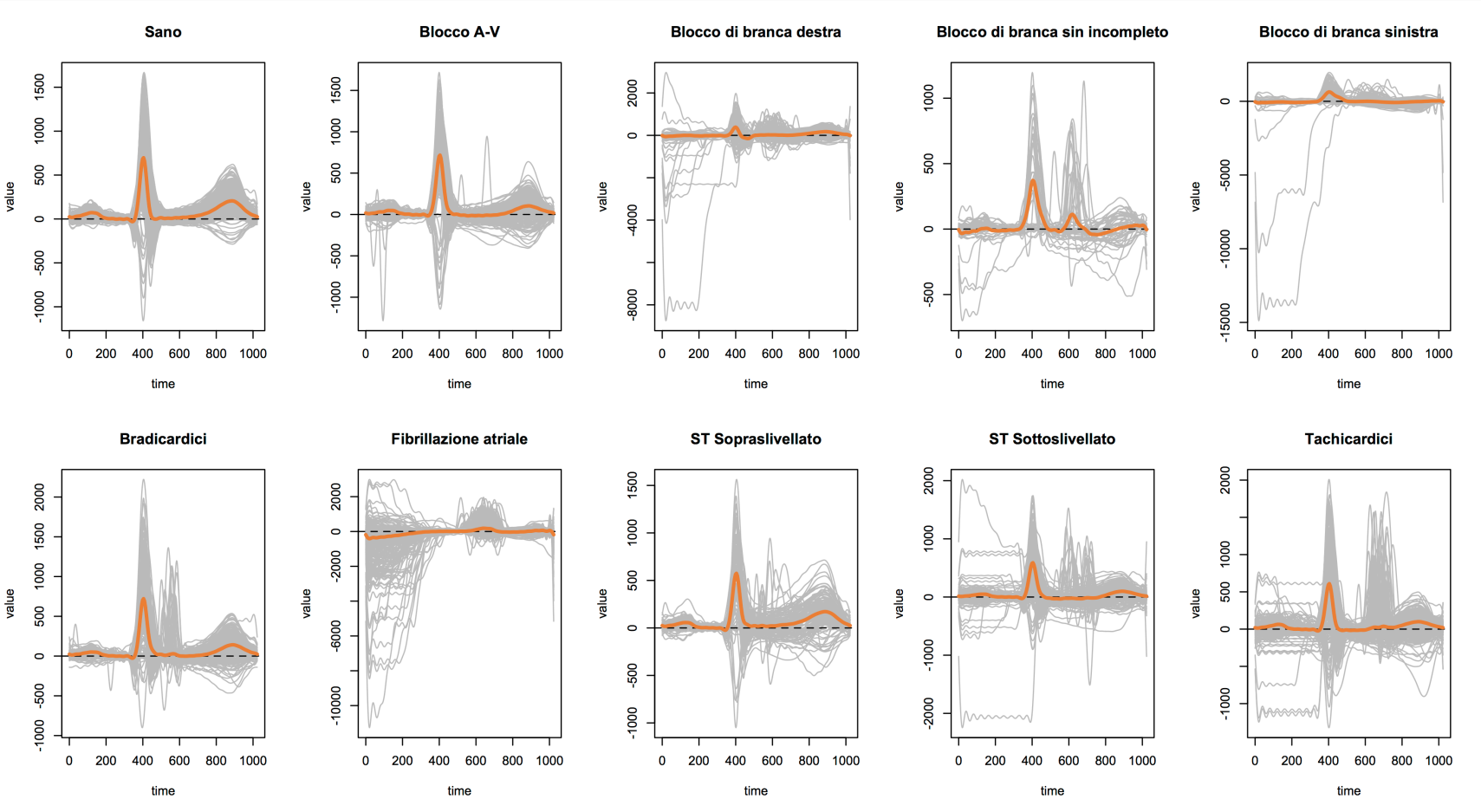
where α_π and β_π are sensitivity and specificity resp. for threshold π

Training and Test



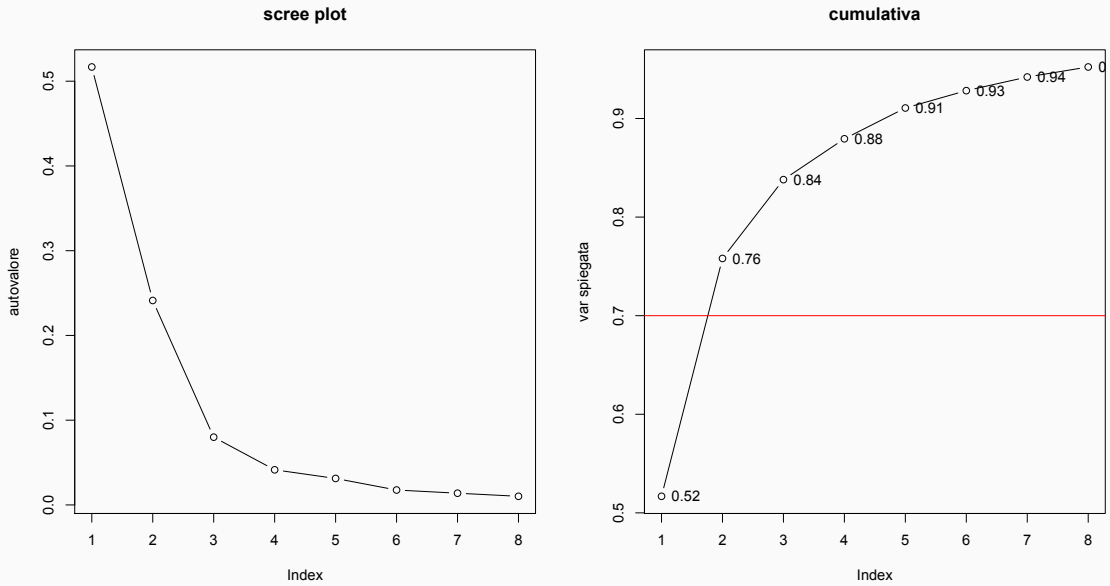
IMPLEMENTATION

First steps


Sano
Blocco A-V
Blocco di branca destra
Blocco di branca sin incompleto
Blocco di branca sinistra
Bradicardici
Fibrillazione atriale
ST Sopraslivellato
ST Sottoslivellato
Tachicardici
value
time

A simplified case

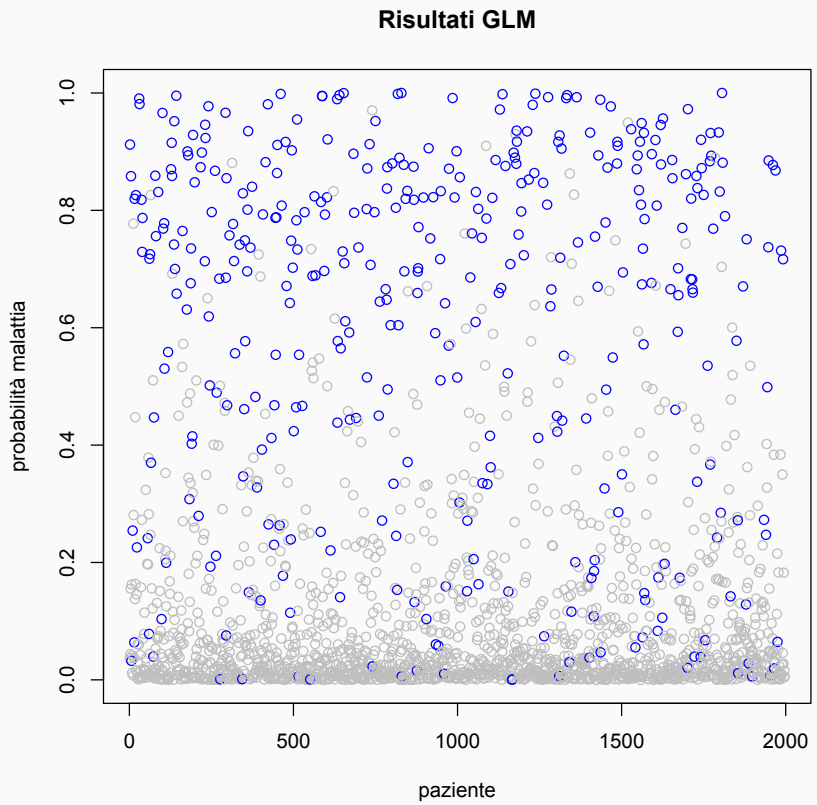
For the sake of simplicity, we study first Right Bundle Brunch Block. After a descriptive analysis, we perform a Principal Components decomposition



obtaining the scores of the subjects

A simplified case

We can now use these values in a GLM



	$\widehat{\text{sick}}$	$\widehat{\text{healthy}}$
sick	264	126
healthy	52	1555

Main Experiment

Test scheme:

1. randomly divide dataset in a training set and a test set
2. compute PCA scores
3. choose the number of PCs to use for each dimension
4. fit a GLM on the training set's scores
5. choose a threshold for the binary classifier, based on the training set
6. verify classifier accuracy on the test set
7. compute confusion matrix

For each pathology we repeat the test N times.

Main Experiment/cont.

Decisions:

- number of tries N for each test for each pathology
- training and test set size
- number of PC K to use for each dimension
- classifier threshold π_0

Main Experiment/cont.

- number of PCs to use
  - same number of PCs for each dimension
  - two criteria:

fixed: choose an a priori fixed number $\bar{K}$

variable: choose the number of PCs that explain at least $\underline{\rho}$ proportion of variance

- classifier threshold
  - two criteria:

fixed: choose an a priori fixed threshold $\bar{\pi}_0$

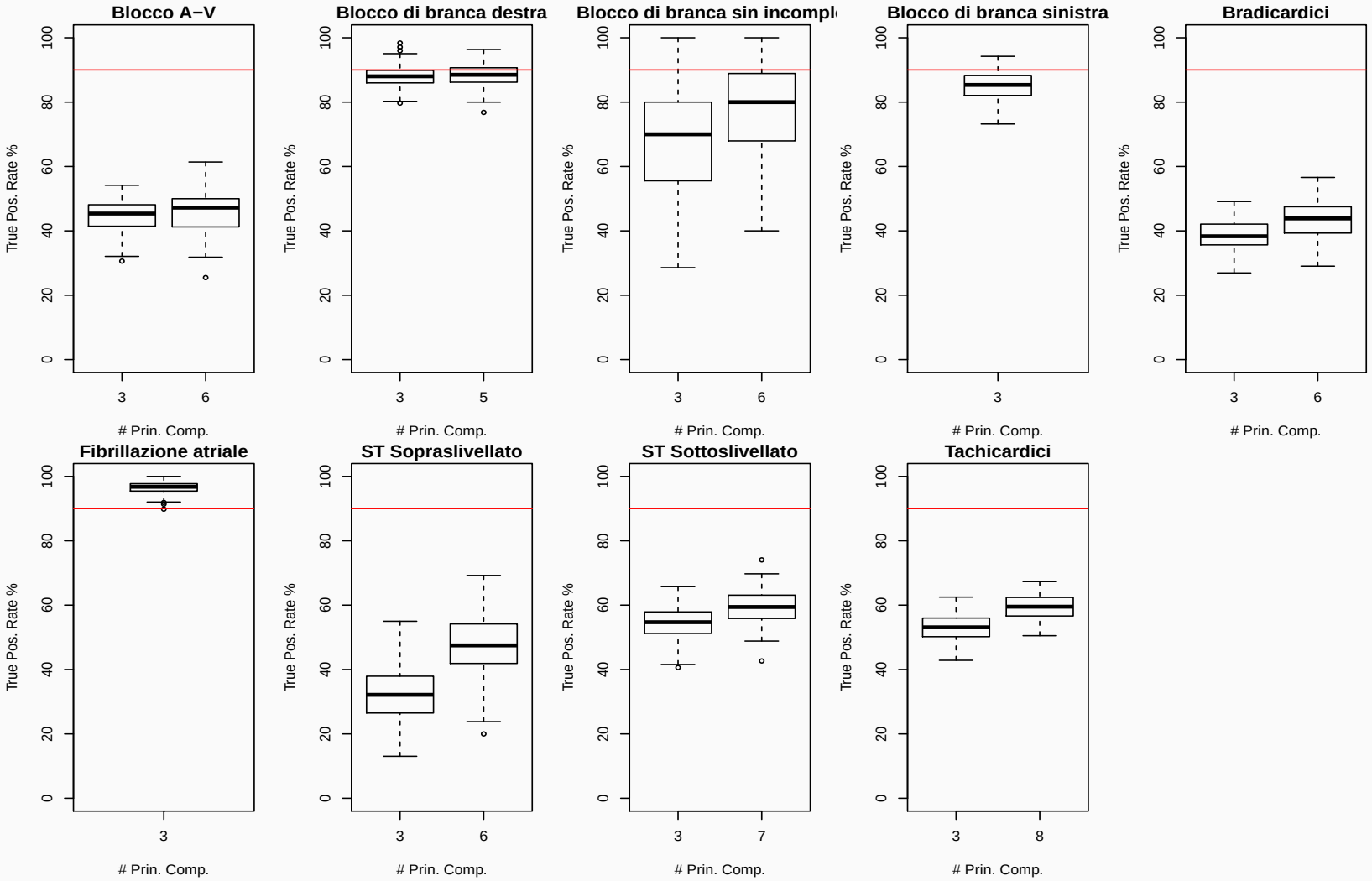
variable: choose the threshold that maximizes the sensitivity of the classifier via the “utopia criterion”

Parameters

N	100	number of tries for each case
$\underline{\rho}$	70%	minimum explained variance required to determine the number of PCs
$\bar{K}$	3	fixed number of PCs
$\bar{\pi}_0$	0.5	fixed classifier threshold
θ	85%	size of training set relative to original dataset

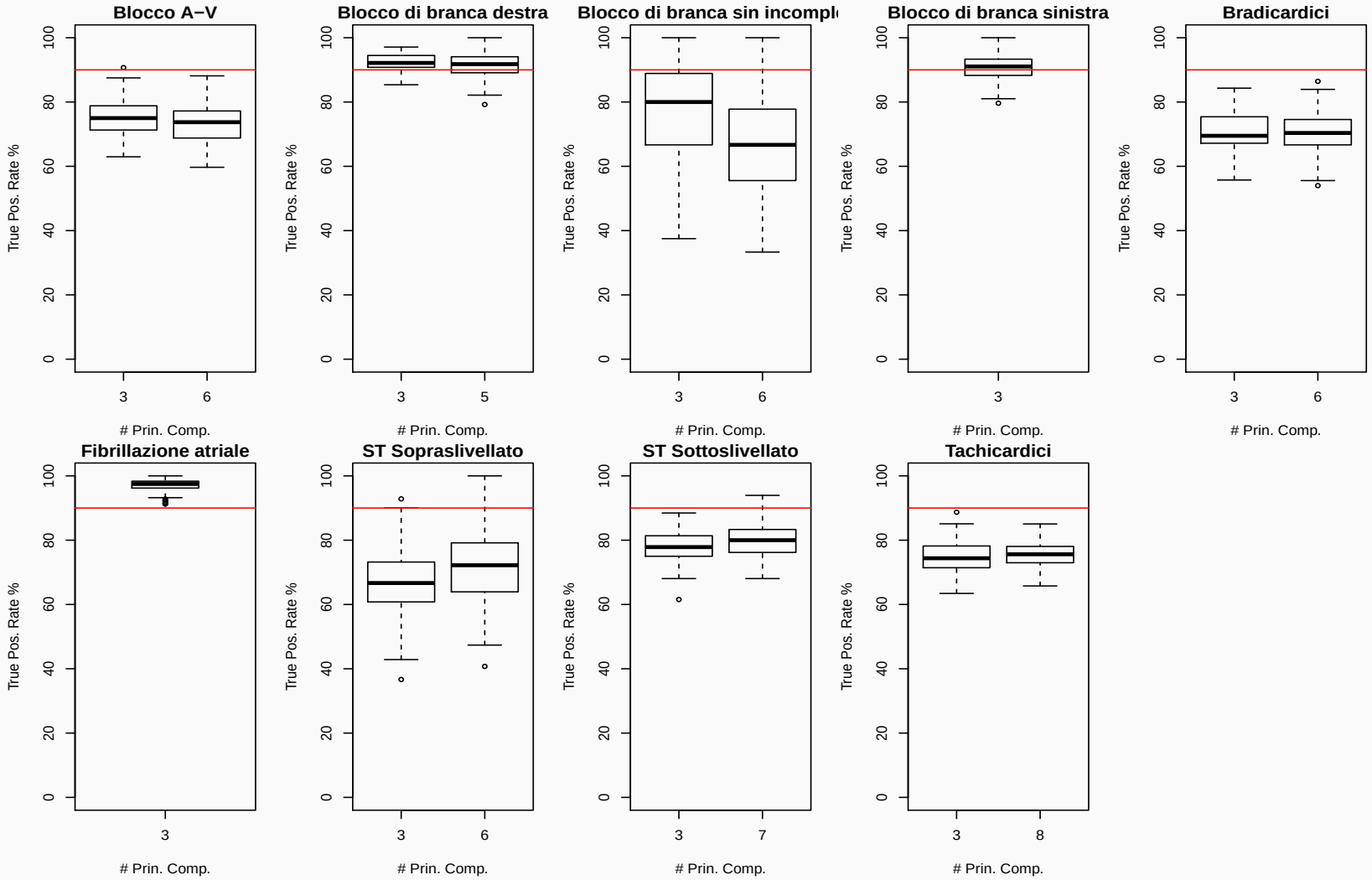
Four different combinations of criteria to choose K and π_0 have to be tried, yielding a total of $4N$ randomized tries for each pathology.

Results - Sensitivity for Fixed Threshold



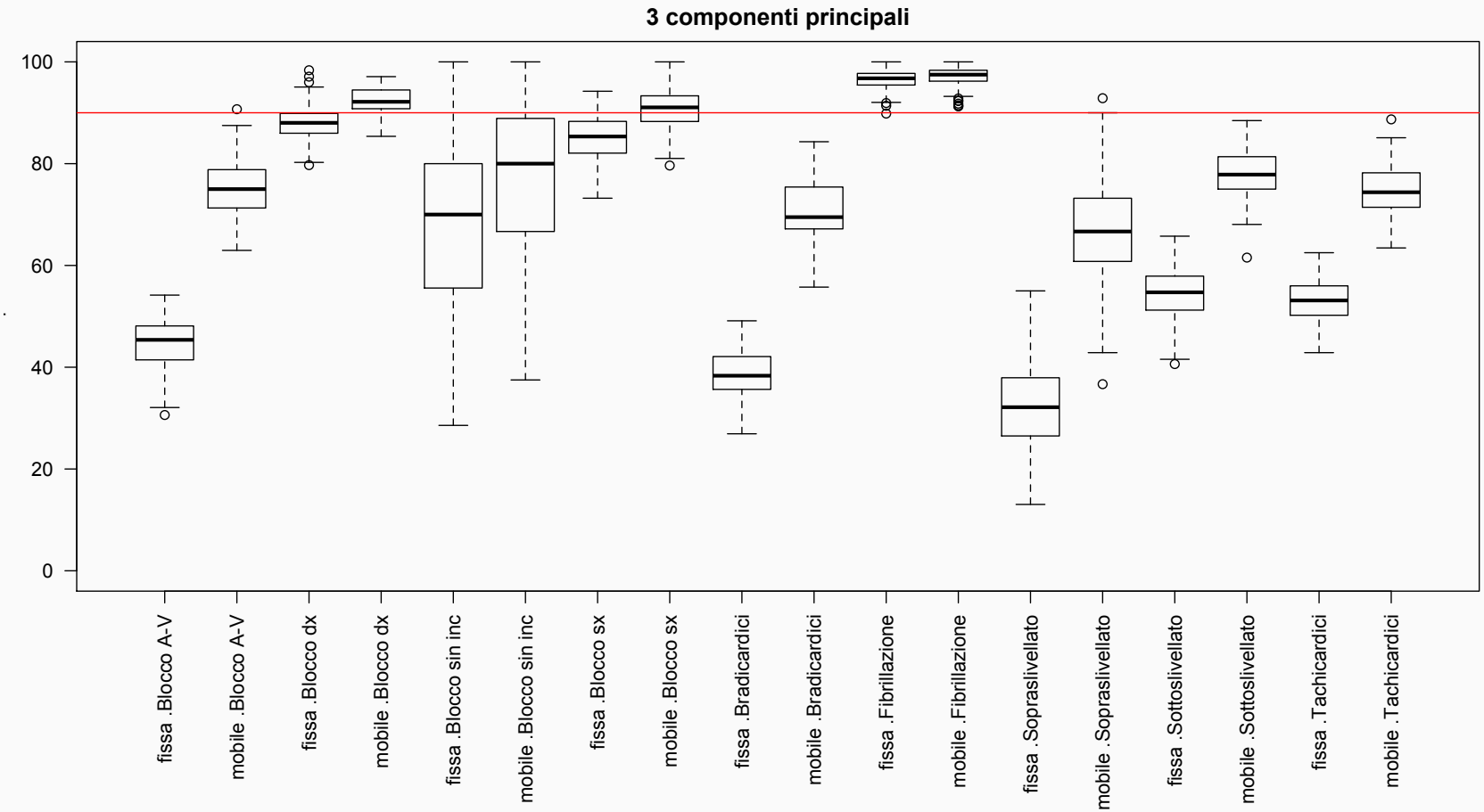
red line = 90% of sensitivity threshold.

Results - Sensitivity for Variable Threshold



red line = 90% of sensitivity threshold.

Results - 3 PCs



Comments

- LBBB, RBBB and Atrial Fibrillation can be effectively separated
- optimizing the threshold improves accuracy, but mostly for poorly classified illnesses
- classifiers perform well even when few PCs are chosen

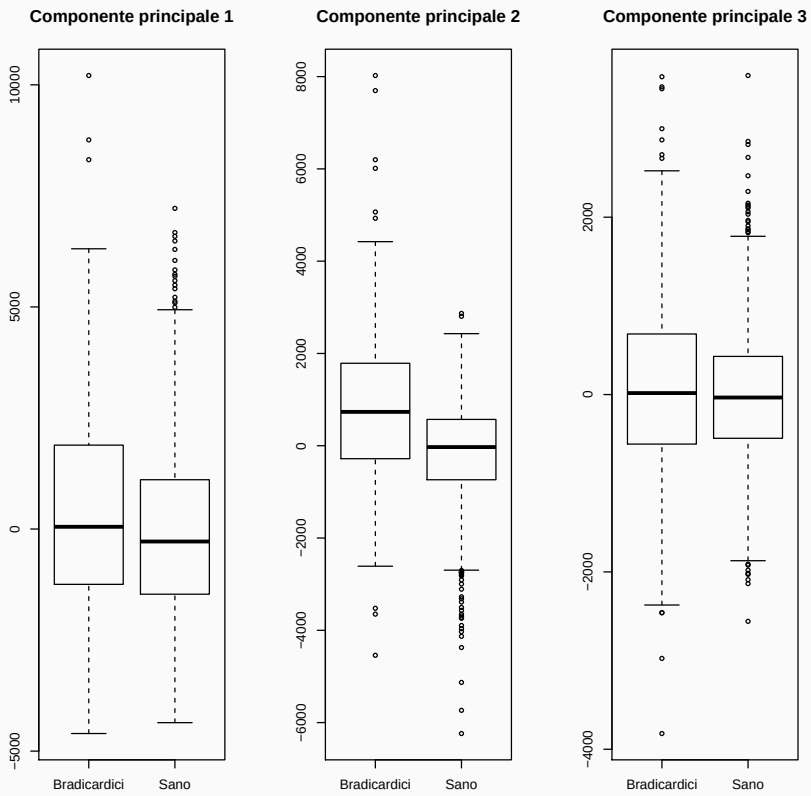
What makes other pathologies harder to separate from healthy subjects?

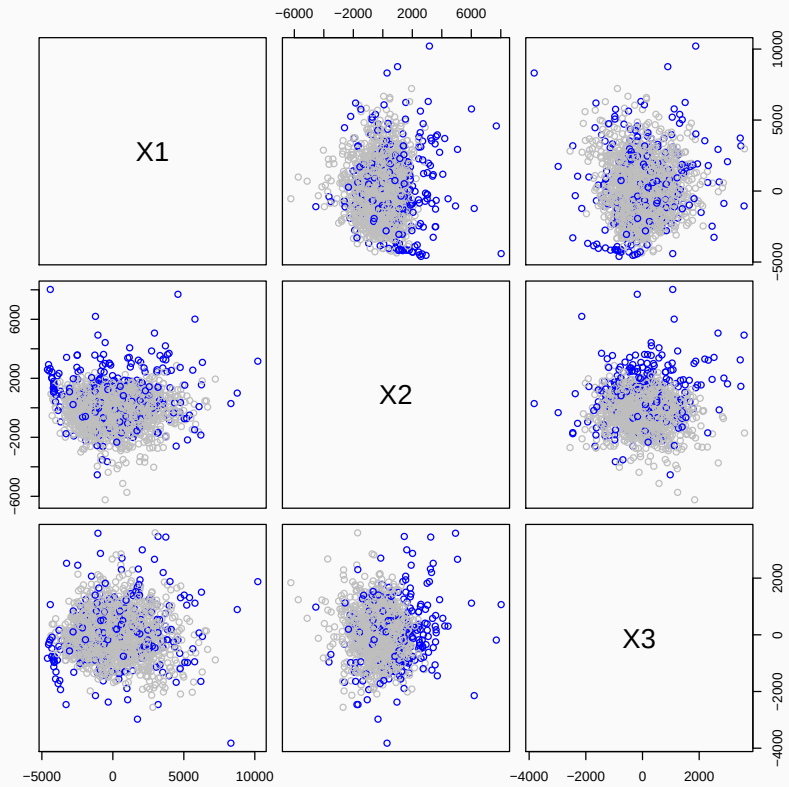
Answer

The corresponding PCA scores are harder to separate.

Separability: easy vs hard cases/1

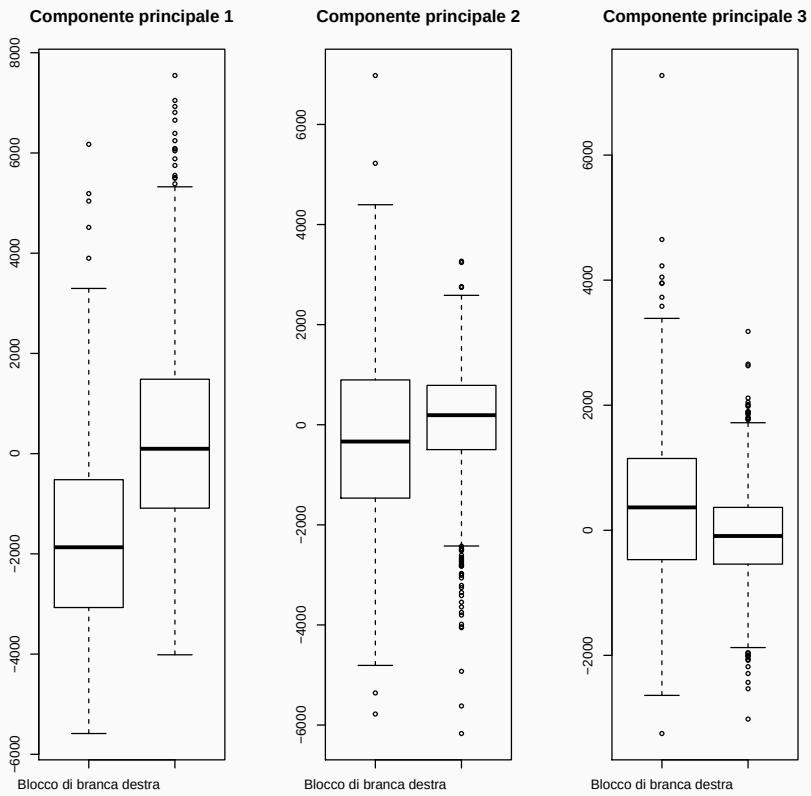
Hard case: Bradycardics

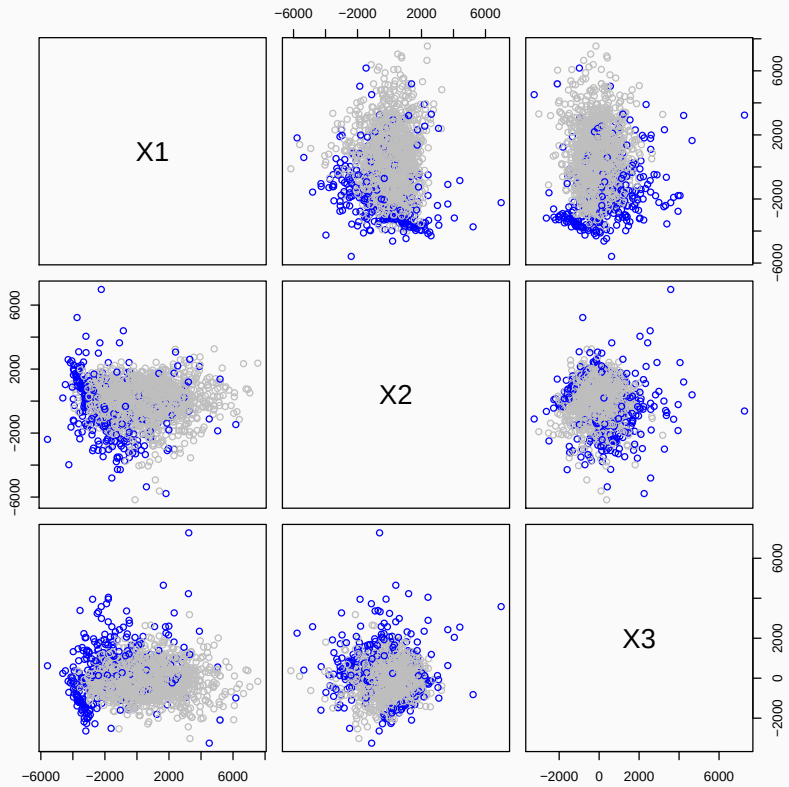




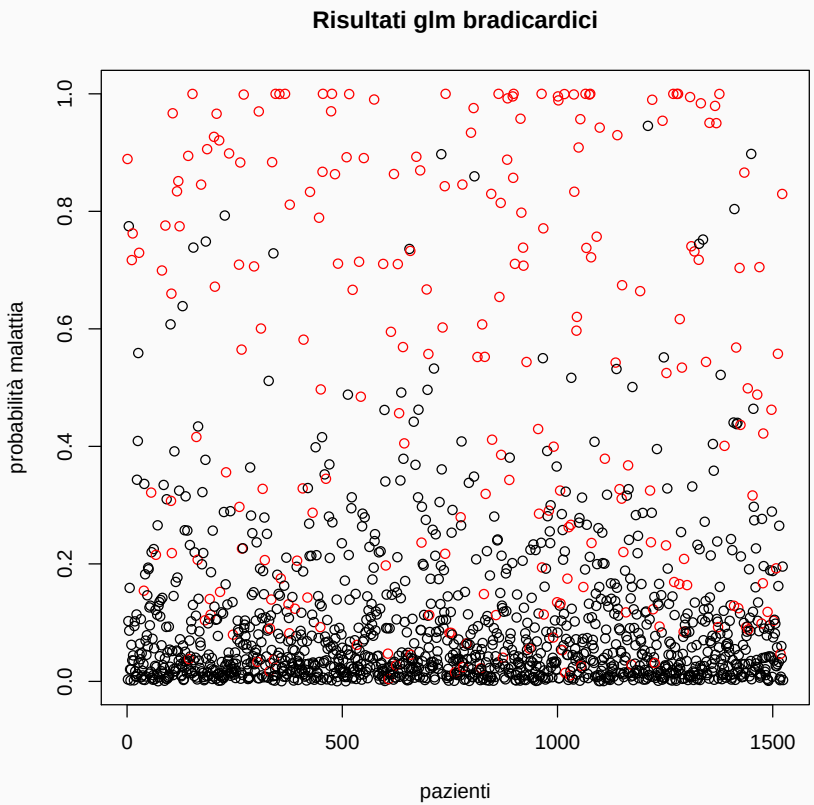
Separability: easy vs hard cases/2

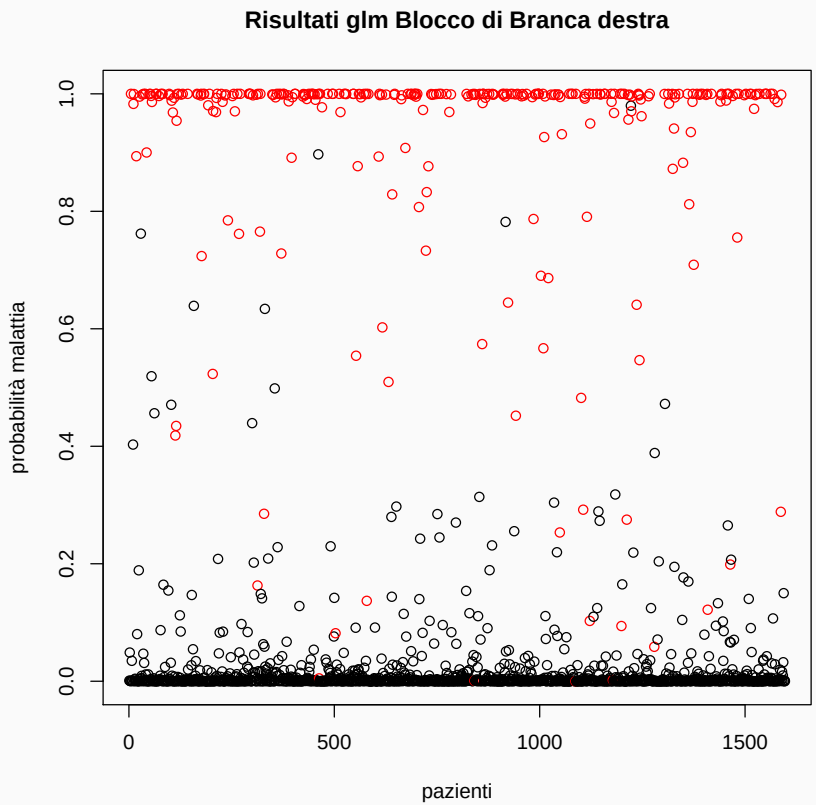
Easy case: Right Bundle Brunch Block



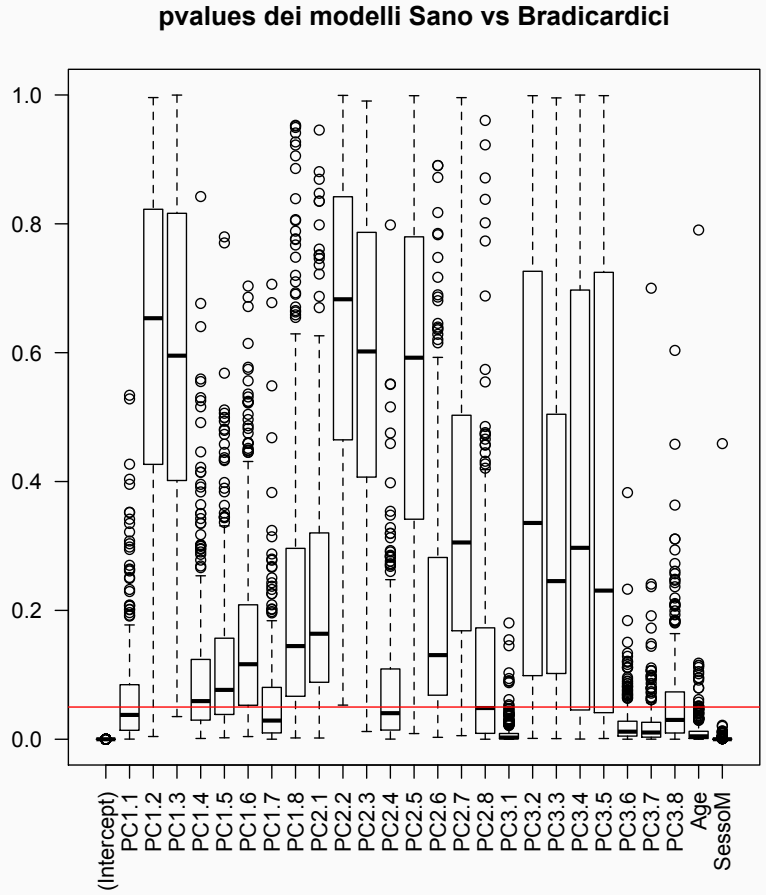


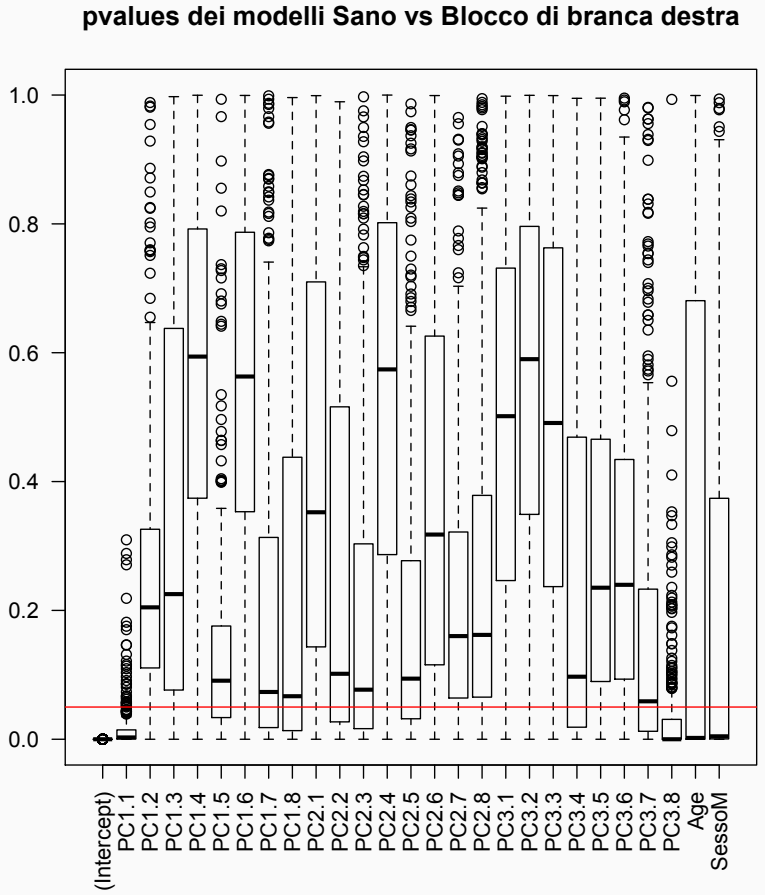
Separability: easy vs hard cases/3





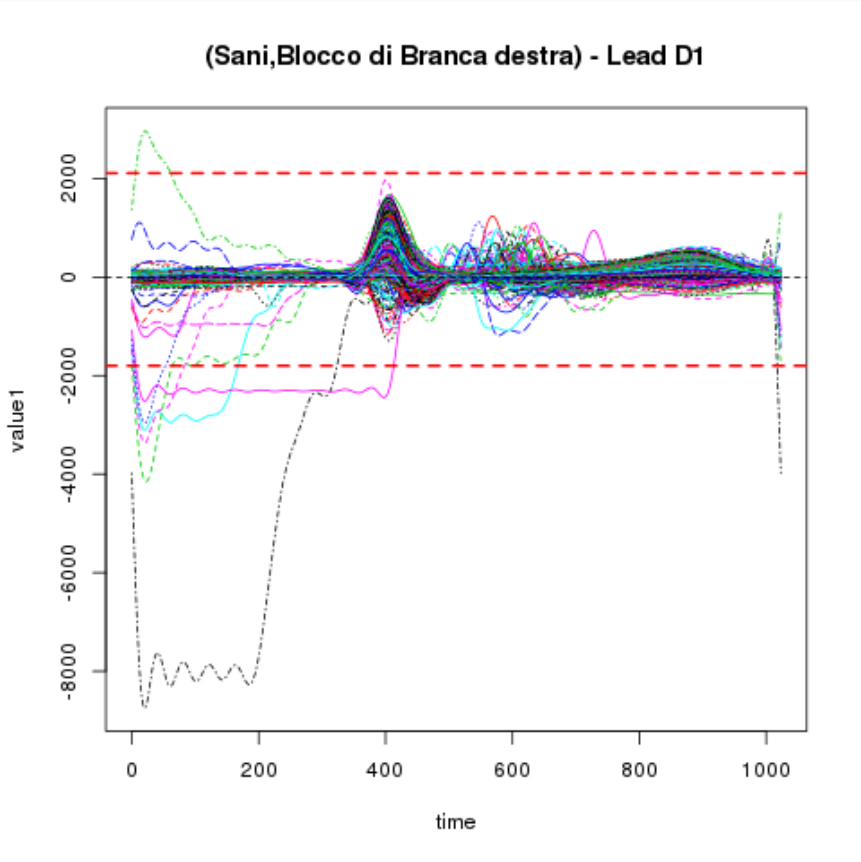
Separability: easy vs hard cases/4





Outliers

Outliers can be easily spotted. Often due to misplacement of sensors.



Outliers/2

- how can we remove outliers?
  - **formal, robust methods:** band depth.
    Open problem. Falls outside our main goal.
  - **heuristic:** filter derivatives with “huge” values at the beginning of the period
    - easy to filter too many patients.
    - need to find a sensible threshold

- should outliers be removed?
  - we successfully removed outliers with the heuristic.
  - classifiers’ accuracy did not improve
    - data hard to separate are not outliers in the scores’ space.

Multiple Classifier

Let $I = 1, \ldots, M$ be a set of trained binary classifiers and let $\mathbf{x}$ be the ECG data for a new patient.

1. for each binary classifier $i \in I$
   1.1 compute scores $\mathbf{c}_i(\mathbf{x})$ for the new patient via regression on eigenfunctions ξ^i
   1.2 apply the classifier on scores $\mathbf{c}_i(\mathbf{x})$, yielding an indicator $\delta^i_{\mathbf{x}}$ and an estimated probability $\pi^i_{\mathbf{x}}$
2. if $\delta^i_{\mathbf{x}} = 0 \; \forall i \in I$ the classifier has determined the patient to be *healthy*
3. otherwise the pathology with index $i^* = \underset{i \in I : \delta^i = 1}{\arg\max} \{\pi^i\}$ is selected.

Multiple Classifier/2

The multiple classifier returns a couple

$$(\Delta, \pi)$$

where

- $\Delta \in 0 \dots |I|$ is the pathology index, with $0 =$ healthy
- $\pi \in [0,1]^{|I|}$ for each $i \in I$ is the probability the patient is actually affected by pathology i, as returned by the i^{th} GLM, if the i^{th} classifier yielded a positive results, or 0 otherwise

CONCLUSIONS

Conclusions

- FPCA synthesizes functional data in a direct and simple way
  - few or no parameters need to be specified
- ECGs scores do not allow for easy separation of all the pathologies. Why? For those cases either
  - ECGs are actually hard to separate or
  - more effective feature extraction methods can be used on ECGs

Questions?