POLITECNICO
DI TORINO

UNIVERSITÀ
DEGLI STUDI
DI TORINO

Doctoral Dissertation
Doctoral Program in Pure and Applied Mathematics (32^{nd}cycle)

Statistical methods for biomarker discovery and multivariate classifier evaluation

By

Lidia Sacchetto

******

Supervisor:
Prof. Mauro Gasparini

Doctoral Examination Committee:
Prof. Monica Chiogna, Università di Bologna, Italy (Chair)
Prof. Colin Begg, Memorial Sloan Kettering Cancer Center, New York, US
Prof. Fulvio De Santis, Sapienza Università di Roma, Italy

Università di Torino - Politecnico di Torino
July, 2020

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Lidia Sacchetto
July, 2020

* This dissertation is presented in partial fulfillment of the requirements for the degree of ***Philosophiæ Diploma*** **(PhD degree)** in **Pure and Applied Mathematics**.

Acknowledgements

First of all and foremost I would like to thank my supervisor Prof. Mauro Gasparini for motivating and supporting me during these years, for showing how research in academia works, for introducing me to the fascinating world of statistics in the pharmaceutical industry. Thanks for the patience and the help and for having always found time to discuss problems and to solve my doubts.
Besides my advisor, I would like to thank my thesis reviewers: Prof. Colin Begg (Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, US) and Prof. Monica Chiogna (Department of Statistical Sciences, Università di Bologna, Italy) for the careful reading of the thesis and their insightful comments.
Furthermore, my sincere thanks go to my co-authors: in particular the researchers of the Cancer Genomic Lab (Tempia Foundation, Biella, Italy) who firstly initiated me to biostatistics and bioinformatics, shedding lights on applications of maths to concrete problems; Dr Roberto Zanetti and Dr Stefano Rosso (Piedmont Cancer Registry, Turin, Italy) for giving countless opportunities to internationally show my competences and for always spurring me to do better. They had a fundamental role in my decision to join a PhD program and I really want to thank them for the hard but stimulating work of the last ten years.
In addition, I would like also to thank the Clinical Statistics Europe Oncology group (Bayer Pharmaceutical, Berlin, Germany) for revealing how research in pharmaceutical industry develops and how interesting and challenging it could be: the internship has been a milestone in the choice of starting my working career abroad and I am thankful for this amazing experience.
Certainly, I can not forget to thanks my colleagues PhD students and my office mates: you transformed the long hours spent at the Department in an exciting time; I already miss our coffee breaks, lunches and smart mathematical questions! I enjoyed a lot my second student life.
I am also grateful for the old and new friends of these years: thanks for being as you are, for your support and for finding moments to dedicate to our friendship.
Last, but not least, an heartfelt thank you to my family who always encouraged and sustained me, especially in the tough period of the writing of the thesis; to my parents for their example and love; to my sisters for their patience in listening my worries and for making me more confident in my abilities; to my grandmother for being proud of me (without understanding what I am doing) and for remembering me in her prayers.

Ringraziamenti

Innanzitutto vorrei ringraziare il mio tutor il Prof. Mauro Gasparini per avermi saputo motivare e supportare in questi anni, per avermi mostrato cosa vuol dire fare ricerca in ambito accademico e per avermi introdotto nell'affascinante mondo della statistica in ambito farmaceutico. Grazie per la pazienza e l'aiuto e per aver trovato sempre il tempo per discutere insieme dei problemi e per risolvere i miei dubbi.
Oltre al mio tutor, desidero ringraziare i reviewer della mia tesi: il Prof. Colin Begg (Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, US) e la Prof. Monica Chiogna (Dipartimento di Scienze Statistiche, Università di Bologna, Italia) per l'attenta lettura e gli acuti commenti.
Un sincero grazie poi ai miei co-autori: in particolare le ricercatrici del Cancer Genomic Labs della Fondazione Tempia di Biella che, per prime, mi hanno iniziato alla biostatistica e alla bioinformatica, mostrandomi come la matematica possa essere applicata a problemi reali; il Dr Roberto Zanetti e il Dr Stefano Rosso del Registro Tumori del Piemonte per avermi dato innumerevoli opportunità di esibire le mie competenze in ambito internazionale, spronandomi sempre a far meglio; la mia decisione di intraprendere il dottorato la devo essenzialmente a loro e voglio ringraziarli per il duro ma stimolante lavoro degli ultimi dieci anni.
Inoltre vorrei ringraziare anche il gruppo Clinical Statistics Europe Oncology della Bayer di Berlino per avermi mostrato come funziona la ricerca in ambito farmaceutico, quanto possa essere interessante e ricca di sfide: il tirocinio è stato una pietra miliare nella scelta di proseguire la mia carriera lavorativa all'estero e sono grata di aver avuto questa entusiasmante esperienza.
Certamente non posso dimenticare di dire grazie ai miei colleghi studenti di dottorato e d'ufficio: avete trasformato le lunghe ore trascorse in Dipartimento in un tempo estremamente piacevole; mi mancano le pause caffè, i pranzi insieme e i nostri sagaci quesiti matematici! Ho davvero apprezzato il mio ritorno alla vita da studente.
Sono ovviamente grata a tutti i mie amici, quelli che ci sono da sempre e i nuovi conosciuti negli ultimi anni: grazie per essere come siete, per il vostro supporto e per trovare sempre del tempo da dedicare alla nostra amicizia.
Infine, ma non per importanza, un grazie di cuore alla mia famiglia per avermi sempre incoraggiato e sostenuto, soprattutto nello strenuante periodo della stesura della tesi; ai miei genitori per il loro amore e il loro esempio; alle mie sorelle per la loro pazienza nell'ascoltare le mie preoccupazioni e nel darmi fiducia; a mia nonna per essere fiera di me pur non capendo bene che cosa io faccia e per ricordarmi sempre nelle sue preghiere.

Abstract

The accurate diagnosis of cancer, its prognosis and the correct classification of patients into molecular and/or phenotypical subclasses is crucial. The new biomedical technologies and methodologies available nowadays produce large amount of data and appropriate statistical techniques are needed for rigorous analyses. Biomarkers play a fundamental role and the discovery of new ones to help the detection of disease or its development and to predict the usefulness of treatments constitutes a great scientific result. To critically evaluate the discriminant power, the Receiver Operating Characteristic (ROC) curve is the common used tool. Optimal, always concave, non-decreasing and above the 45-degree line curves should be preferred, but, actually, everybody uses staircase-shape ROC curves which are not so.
This thesis, starting from a real case study which aims to discover new biomarkers to improve the diagnostic route of prostate cancer, proposes an original proper default alternative to the usual empirical implementation of the binary classification curve. The optimality of the likelihood-ratio based approach, a nonparametric extension of naïve Bayes estimation and the strict relationship between the ROC definition and a general notion of concentration of two probability measures are deeply exploited. The new classification procedure can be applied to finite, continuous and even more complex data types, under the only wide assumption that the likelihood-ratio is meaningful. Furthermore, particularly in multivariate settings, as in genomic studies where the analysis of different biomarkers can improve the performance, the likelihood-ratio based rule outperforms commonly used classifiers based on linear combinations of predictors. On the other hand, it is noteworthy to also consider the classification problem by itself in an unsupervised manner, to investigate up to which point, with modern technologies, is possible to adopt basic methods (such as maximum likelihood and hard assignment) to classify subjects. Indeed, at present, large part of the literature is devoted to soft approaches, related to mixture models. Therefore, a new implementation of hard assignment is proposed, exploring parallel computing and comparing results on a small highly cited dataset.
Lastly, some more applied works are presented in the final part of the thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

AAPC	Average Annual Percent Change
AIC	Akaike Information Criterion
APC	Annual Percent Change
ASAP	Atypical Small Acinar Proliferation
ASR(W)	Age Standardized Rate on World population
AUC	Area Under the receiver operating characteristic Curve
BIC	Bayesian Information Criterion
BPH	Benign Prostatic Hyperplasia
CDF	Cumulative Distribution Function
CI	Confidence Interval
CMM	Cutaneous Malignant Melanoma
CRs	Cancer Registries
cTNM	Clinical TNM staging
DAG	Direct Acyclic Graph
DCO	Death Certificate Only
DRE	Digital Rectal Examination
EDTA	Ethylenediaminetetraacetic acid
EM	Expectation-Maximization algorithm

FPF	False Positive Fraction
GS	Gleason Score
HD	Healthy Donor
HGPIN	High-Grade Prostatic Intraepithelial Neoplasia
HN	Head and Neck
KDE	Kernel Density Estimation
LCA	Latent Class Analysis
LCI	Lower Confidence Interval
LDA	Linear Discriminant Analysis
LM	Lentigo Maligna melanoma
LR	Likelihood Ratio
miR/miRNA	microRNA
MISE	Mean Integrated Squared Error
mp-MRI	Multiparametric Magnetic Resonance Imaging
MSE	Mean Squared Error
MV	Microscopic Verification
NM	Nodular Melanoma
NN	Neural Network
OOB	Out-Of-Bag
OS	Overall Survival
PCa	Prostate Cancer
PSA	Prostate-Specific Antigen
QDA	Quadratic Discriminant Analysis
RF	Random Forests

ROC	Receiver Operating Characteristic Curve
RT	Reverse Transcription
RT-qPCR	Quantitative Real-Time Polymerase Chain Reaction
SEER	Surveillance, Epidemiology, and End Results Program
SSM	Superficial Spreading Melanoma
TPF	True Positive Fraction
TRRBPCH	Best Percentage Change from Baseline
TTP	Time-To-Progression
UCI	Upper Confidence Interval
WHO	World Health Organization

Chapter 1

Introduction

1.1 Introduction and motivation

Nowadays, biomarkers play a fundamental role in many different fields, especially in medicine, helping the diagnosis, prognosis and treatments decision for an increasing number of pathologies, from cancers to cardiovascular illness, from infectious diseases and inflammation to food disorders, just to cite some examples. There is not an unique definition and they can be almost every objectively measurable patient characteristic. The great advancement in technologies which characterises the current times has offered the tools to realize what is indicated by someone as a new industrial revolution: at present it is possible to analyse thousands of genes and molecules simultaneously, to obtain relevant information on a subject health status by a simple non-invasive blood or urine exam, to collect and store a quantity of details just unthinkable thirty years ago. On the other hand, the availability of a such large amount of data gives rise to unexpected problems and challenges which feed the newborn Data Science; it is sufficient to do a quick search on the web looking for personalized medicine, big data, machine learning or deep learning to find millions of results: the rigour of mathematics and statistics is the necessary instrument to navigate this universe.
Certainly, in this perspective, it is essential to face problems from a multivariate point of view: for example, in medical diagnostic, various biomarkers can highlight different aspects of the same disease and their diagnostic, prognostic or predictive value can be strengthened considering them all together. For this reason in the thesis the focus is on multivariate methods. In addition to discovering new biomarkers, it is imperative to critically evaluate their discriminant power (i.e. their ability to correctly classify subjects, to predict a recurrence of a disease or the usefulness of a therapy): the receiver operating characteristic (ROC) curve is the most

popular tool with this aim. It had been firstly introduced in the field of signal detection theory in the 1960s, but then it sprung up in many different disciplines from statistics to medical diagnostics, from machine learning to psychology and educational assessment. Actually the ROC curve for binary populations is defined as a parametric two-dimensional locus of points identified by the sensitivity and the specificity of the classifier. From a theoretical point of view, it is noteworthy to restate some well-known properties and to highlight the relationship with a general definition of concentration function. On the other hand, most interesting applications arise when the two underlying and partially overlapping populations are not known and should be estimated, moving towards a data-driven approach. Usually, empirical estimators of the ROC curve are adopted: they are step functions, whereas smooth and always concave curves would be preferred. Therefore a new non-parametric estimation method is proposed: it directly considers the likelihood-ratio as the decision variable, from one side rooting the procedure in the classical statistical theory of decision making and exploiting the Neyman-Person lemma, on the other side winking to machine learning and involving the näive Bayes classifier.

In the ROC curve framework, it is common to assume to know the true membership label of each unit in a training sample (supervised learning). However, it is also of interest to investigate if collected data allow to classify units in different groups without *a priori* knowledge of their true membership (this problem is usually known as clustering or unsupervised learning). The assignment becomes itself a parameter of the model. The focus is primarily on categorical data and, in particular, on multivariate binary data which commonly originate in pools, questionnaires, online automatic interviews, ... Specific methodologies to deal with this kind of information have been formulated since longtime: at present, a soft approach related to mixture models and latent class analysis gains popularity over the more basilar hard assessment, which dates back to 1970s and is a combinatorial clustering problem. Nevertheless, it is stimulating to investigate up to which point, with modern technologies, is possible to use basic methods, such as maximum likelihood and hard assignment, to identify the class membership of given subjects.

Lastly, part of this dissertation is devoted to present more applied work: I believe that in Statistics as well as in many other fields, a strong theoretical background is fundamental, but it should be complemented by concrete examples and case studies to become more understandable and to reach a larger target.

1.2 Organization of the dissertation

This thesis is organized as follow. In Chapter 2 the general framework is introduced, recalling the well known definition of ROC curve, its properties and the important connection with the likelihood ratio, as well as the relationship with the Lorenz curve and the Gini coefficient. In addition, Section 2.3 is devoted to an overview of most common ROC estimation methods.

In Chapter 3 the real problem that stimulated the interest for biomarkers, classification and ROC curve is presented and the published results are reported and discussed. In particular, a new signature for prostate cancer diagnosis is proposed in Section 3.3: it is an easily applicable blood-based classifier that requires testing of two microRNAs and the prostate specific antigene (PSA) and which shows a discriminant power higher than PSA alone and allows to avoid about one-third of unnecessary biopsies. Furthermore, some medical and biological notions and a general definition of biomarker are provided for a better understanding of the project, which has been developed in collaboration with the Cancer Genomics Lab of the Tempia Foundation (Biella, Italy).

In Chapter 4 a novel algorithm to estimate a proper ROC curve in the multivariate setting is proposed, exploiting the flexible Bayes assumption, the Gaussian kernel density estimation and the local independence of the features. Firstly, in Section 4.1 Fisher linear and quadratic discriminant analysis is presented as an example of likelihood-based classifier; then the Näive Bayes classifier and an extension that overcomes the Gaussian assumption (the flexible Bayes) are introduced in Section 4.2. The core of the chapter is constituted by the new algorithm, detailed in all its steps. In addition, some simulations are run to compare the performance of the new classifier with respect to the widely adopted empirical estimate of the ROC curve associated to a logistic score. Lastly, the new estimation method is applied to the real case study previously presented in Chapter 3.

In Chapter 5 the relationship between the likelihood-based ROC curve and the general definition of concentration function is explored from an exquisitely theoretical point of view. After having introduced a likelihood ratio based classification rule which entails a randomization device to deal with atoms (Section 5.1), an alternative formulation of the ROC curve is presented and the connection with the concentration function is proven in Section 5.2. Furthermore, a critical appraisal of a recent published paper is proposed in Section 5.4, whereas some considerations on discrete ROC curve are provided in the last section of this chapter.

In Chapter 6 the focus of the work switches from a supervised learning perspective to an unsupervised one and model-based clustering problems are presented for binary outcomes and binary populations. In particular, it is of interest to investigate up to which point, with modern technologies, it is possible to adopt traditional and basic methods, such as maximum likelihood and hard assignment, to classify units (usually subjects) in different groups. The differences between hard and soft approach are presented in Section 6.1, K-means and EM algorithms are introduced in Section 6.2, while latent class analysis models and their implementation in R are shortly summarized in Section 6.3. The core of the chapter is built around a new implementation of hard assignment in R; computational limits and parallel computing solutions are explored to solve the NP-hard optimization problem (Section 6.4). Finally, the proposed algorithm is applied to a toy example and results are compared to those obtained with the soft approach (Section 6.5).

Lastly, in Chapter 7, some more applied works are collected. The first, described in Section 7.1, was developed during a six months internship at Bayer Pharmaceutical (Berlin, Germany); it was the first attempt to look in depth into all clinical data routinely collected in trials to understand if there was valuable unused information. Multivariate predictive models for treatment efficacy with time-to-event outcomes (overall survival), using only baseline data routinely collected in clinical trials, are presented; machine learning and deep learning methods are compared to more traditional Cox models.
The other two studies are epidemiological analyses of cutaneous malignant melanoma, developed in collaboration with the Piedmont Cancer Registry (Turin, Italy). In particular, I leaded the first and largest observational studies on European incidence trends by thickness's level and on trends of lethal melanomas: published results are reported in Section 7.2, whereas in Section 7.3 a new approach for analysing fatal cases is presented.

More relevant R code, as well as some biological and technical details for the motivating example are presented in Appendices.

Chapter 2

Theoretical fundamentals

In this chapter I present the Receiver Operating Characteristic (ROC) curve, its most important properties and the principal methods to estimate it. In addition, the well known relationship with the likelihood ratio function and the connection with Lorenz curve and Gini coefficient are explored. These theoretical fundamentals will be largely used in the other chapters which contain the original developments of the thesis and some applications.

2.1 The Receiver Operating Characteristic Curve

The Receiver Operating Characteristic curve (ROC curve, also indicated as relative operating characteristic by some authors) originated in the context of signal detection theory in 1960s [9, 10]; since then its use spreads in different fields from statistics to medical diagnostics and radiology, from machine learning to psychophysics, as made evident by the large amount of literature available (as [1, 11, 12] just to cite some recent books).
Essentially the ROC curve is a graphical tool that allows to evaluate the discrimination performance of a classifier, i.e. to asses the accuracy of a statistical model aimed at dividing units into two groups (henceforth indicated as P_+ and P_-). Extensions to multi-class problems and to ROC surface are present in the literature, but will not be detailed in this chapter.
In the medical framework, as the one presented here (motivated by applications), in a typical binary classification problem, units are subjects, the two populations could represent healthy (P_-) and diseased (P_+) people, while the classifier is a screening or diagnostic test.
From a mathematical point of view, the ROC curve is the parametric two-dimensional locus

of the points of coordinates

$$\{(\text{FPF}(t), \text{TPF}(t)),\ t \in (-\infty, \infty)\} \tag{2.1}$$

where the TPF (True Positive Fraction) and the FPF (False Positive Fraction) are the following conditional probabilities:

$$\begin{aligned} \text{TPF} &= \mathbb{P}(S > t | P_+) \\ \text{FPF} &= \mathbb{P}(S > t | P_-), \end{aligned}$$

respectively related to the sensitivity (Se, the probability to assign a subject to P_+, given it comes from P_+) and 1-specificity ($1-Sp$, the probability to wrongly assign a subject from P_- to P_+) of the test. S is the diagnostic variable and, often, it is a score of multiple measurements obtained applying an appropriate function to the vector of predictors. $t \in \mathcal{R}$ is the threshold parameter (the cut-off) of the classification rule. It is common to assume that a greater value of S is more indicative of disease and therefore to assign a subject to P_+ if $S > t$ (as highlighted by Figure 2.1). Hence the ROC curve is described by varying t. If the

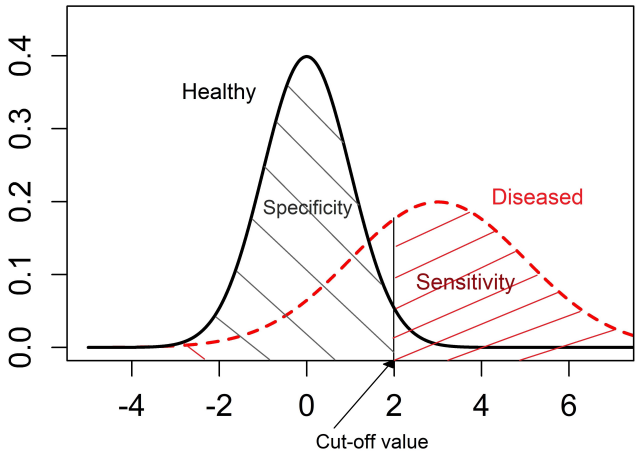


Figure 2.1 Given two partially overlapped populations and a threshold value t, the specificity (shaded in grey) is the proportion of healthy people with a score lower than the cut-off; the sensitivity (in red) is the proportion of diseased people with $S > t$.

two populations completely overlap, the ROC curve reduces to the chance diagonal (the line C in green in Figure 2.2) and the classifier assigns subjects to populations at random; on the other hand, if the two populations are totally separated, there exists a value t which allows to perfectly discriminate subjects: this is the ideal classifier, whose ROC is formed by two segments connecting the point $(0,0) - (0,1)$ and $(0,1) - (1,1)$ (the curve A in red in Figure

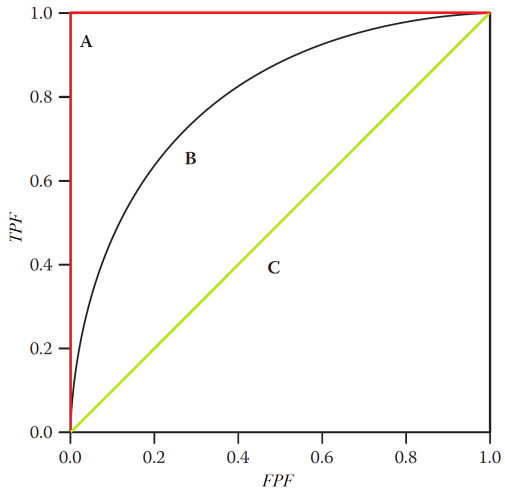


Figure 2.2 The curve A in red represents the ideal classifier; the line C in green describes a random classifier; the curve B in black is an example of a common classifier in between. (Figure adapted from [1]).

2.2). Actually the ROC curve of a diagnostic test lies between these two extremes, with the better the closer to the upper left corner of the square.

2.1.1 ROC curve properties

The ROC curve has some well known properties (see [11, 12] for proofs and details), reported here for their importance in the development of the work.

P1. The ROC curve is invariant to strictly increasing transformations of the diagnostic variable S.

P2. If the decision variable S is a continuous random variable with positive densities f_+ and f_-, a convenient equation for the ROC curve can be derived: let F_+ and F_- be the distribution functions of S in the two populations and let F^{-1} be the inverse of F.

$$\begin{aligned} \text{FPF}(t) &= \mathbb{P}(S > t|P_-) = 1 - F_-(t) \quad \Rightarrow \quad t = F_-^{-1}(1-s) \\ \text{TPF}(t) &= \mathbb{P}(S > t|P_+) = 1 - F_+(t) \end{aligned}$$

Eliminating t:

$$\text{ROC}(s) = 1 - F_+(F_-^{-1}(1-s)) \tag{2.2}$$

P3. Under the hypotheses of the previous property, and assuming that the slope of the ROC curve at threshold value t is well-defined:

$$\frac{\partial \mathrm{ROC}(t)}{\partial t} = \frac{\mathbb{P}(S \leq t | P_+)}{\mathbb{P}(S \leq t | P_-)}.$$

Therefore, the slope corresponds to the likelihood ratio L, evaluated at the threshold value t. More details and implication of this property will be presented in Section 2.2.

2.1.2 AUC and C-index

In addition to the principal properties of the ROC curve just recalled, it is useful to introduce the Area Under the Curve (AUC), a widely used summary measure of the overall accuracy of a diagnostic test, and the strictly related definition of concordance index (usually indicated as $C-$index).
The AUC is defined as

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(z)dz.$$

It varies between 0.5 (uninformative test) and 1 (perfect classifier), and greater AUC values indicate better tests. Clearly, if a ROC curve C_1 dominates another ROC curve C_2 for each $z \in (0,1)$, the same ordered relation is inherited by the AUC, i.e. $\mathrm{AUC}_{C_1} \geq \mathrm{AUC}_{C_2}$; however, a higher AUC does not necessarily imply a dominance of the correspondent ROC (because curves can cross). Finally, an interesting interpretation that precisely states the connection with the Mann-Whitney-Wilcoxon statistical test (as highlighted in [13] and proven in [14]) is:

$$\mathrm{AUC} = \mathbb{P}(S_+ > S_-).$$

In other words, the AUC is equal to the probability that a classifier assigns an higher score to a randomly selected subject from P_+ than to one independently and randomly chosen from P_-.
A generalization of the AUC is the concordance index (C-index), the most popular measure of goodness-of-fit in survival analysis. It is a function of the rank sum statistic and it can be applied to evaluate the model discriminatory power with non-continuous and censored data. Many different estimators have been proposed (as recently reviewed in [15]). Following the formulation given by Harrell *et al.* [16]

$$C = \frac{\sum_{i \neq j} \Delta_i \mathbb{1}_{[X_i < X_j]} \mathbb{1}_{[\hat{\beta} Z_i > \hat{\beta} Z_j]}}{\sum_{i \neq j} \Delta_i \mathbb{1}_{[X_i < X_j]}} \tag{2.3}$$

where $\mathbb{1}_{[\cdot]}$ is the indicator function and, for the ith subject, $X_i = \min(T_i, D_i)$, T_i is the survival time, D_i is the censoring variable, $\Delta_i = 1$ if $X_i = T_i$ (0 otherwise) and $\hat{\beta} Z_i$ is the prognostic risk score (based on baseline evaluation) with Z_i vector of covariates. Therefore, among all pairs of patients for which it is possible to determine the order of survival times, the C-index represents the proportion of concordant pairs, i.e. the fraction of pairs of patients with larger predictive survival who lived longer. Of course, if both patients are alive, the date of death is not known and the pair must be excluded; the same happens for patients who died together and for pairs in which one died and the other is censored at a time which does not allow to determine if he lived longer than the first.
As for the AUC, the C-index varies between 0.5 (random prediction) and 1 (the baseline data always allows to determine the patient with better prognosis with certainty).

2.2 The likelihood ratio

In general, the likelihood ratio function L for a decision (diagnostic) variable S is defined as:

$$L(s) = \frac{\mathbb{P}(S = s|P_+)}{\mathbb{P}(S = s|P_-)}$$

where $\mathbb{P}$ is a probability density function if S is continuous and a mass if S is discrete.
This quantity gives an idea of the extent to which the data support one condition with respect to the other and it allows to map a multivariate vector onto an one-dimensional decision axis (becoming a very useful tool when considering multiple tests).
Indeed, the fundamental role of likelihood ratio in decision making process has been highlighted since the first works on the matter more than forty years ago [9, 10]. However, it is important to re-state the following well-known result (extensively used hereinafter).

Result 2.2.1. The likelihood ratio criterium, i.e. the decision rule which classifies subjects as diseased if

$$L(S) > t \quad \text{for some threshold } t \in \mathcal{R}$$

is optimal, attaining the highest TPF at a given $\mathbb{P}(\mathrm{L}(S) > t|P_-)$, among all possible criteria based on S.

Actually, this result originates from the theory of hypothesis testing and it is essentially the Neyman-Pearson Lemma [17, 18].

Lemma 2.2.2 (The Neyman - Pearson Lemma (from [18]))**.** *Let F_- and F_+ be probability distributions having densities f_- and f_+ respectively, with respect to a measure μ.*

(i) Existence. For testing $H_0 : f_-$ against the alternative $H_1 : f_+$ there exists a test ψ and a constant k such that

$$E_0\psi(S) = \alpha \tag{2.4}$$

and

$$\psi(s) = \begin{cases} 1 & when \;\; f_+(s) > kf_-(s), \\ 0 & when \;\; f_+(s) < kf_-(s). \end{cases} \tag{2.5}$$

(ii) Sufficient condition for a most powerful test. If a test satisfies (2.4) and (2.5) for some k, then it is the most powerful for testing f_- against f_+ at level α.

(iii) Necessary condition for a most powerful test. If ψ is most powerful at level α for testing f_- against f_+, then for some k it satisfies (2.5) a.e.μ. It also satisfies (2.4) unless there exists a test of size less than α and with power 1.

In other words, in the medical diagnostic framework, the belonging to the healthy population (P_-) is the null hypothesis H_0, the decision variable S generates the data on which the test is conducted and the most powerful test of size α is given by $L(s) = \frac{\mathbb{P}(S>t|P_+)}{\mathbb{P}(S>t|P_-)} \geq k$. In addition, the Type I error α is simply the false positive fraction, while the power of the test $(1-\beta)$ and TPF are exactly the same entity.
As a consequence:

- The ROC curve for the $L(S)$ is optimal, in the sense that it is uniformly above all other curves based on S.
- If $L(\cdot)$ is monotone increasing with respect to the decision variable S, then classification rules based on S are optimal, because equivalent to rules based on $L(S)$. The monotonicity of the likelihood ratio is an assumption often made in publications to obtain more efficient results [19–21]; however the optimality of the likelihood ratio based ROC curve persists independently from this assumption (and, unless differently stated, it will not be retained in the following of the work).
- The likelihood ratio based ROC curve is concave everywhere in $[0,1]$: its slope decreases monotonically with the false positives probability (this can be proven using the relation $L(L(S)) = L(S)$, as in [9]).

In particular, the ROC curve based on the likelihood ratio, which satisfies all these nice properties, is usually indicated as a *proper* curve, following the definition given by Egan in 1975 [10]. On the other hand, an *improper* ROC curve (i.e. a curve not based on the

likelihood ratio) can reverse its convexity at certain points, presenting hooks (where the classifier performs worse than chance), as clearly highlighted by Figure 2.3, extracted from [2]. However, the likelihood ratio approach is so powerful only if the joint conditional

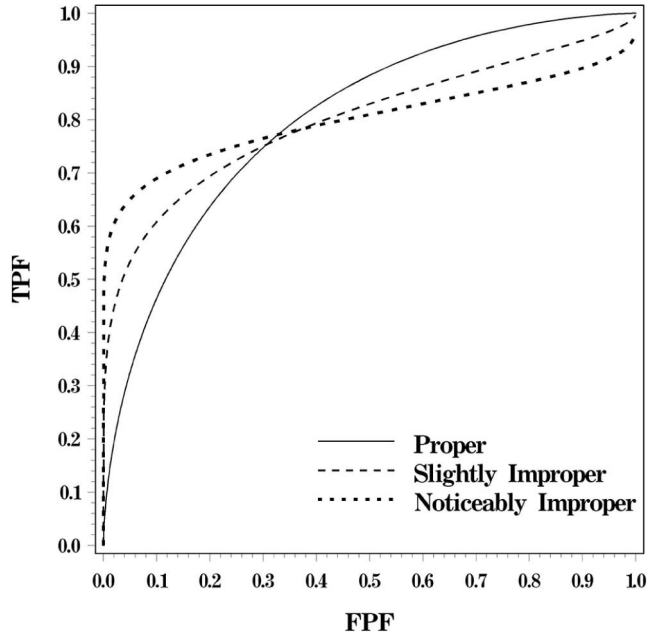


Figure 2.3 ROC curves with different levels of improperness, extracted from [2].

distributions F_+ and F_- are correctly specified. Some authors [11, 22, 23] claimed that using the LR is equivalent to use the (more familiar) risk score, defined as the posterior probability of the positive population corresponding to any given prior probability. However, in the following of the dissertation, I will focus specifically on the LR.
If P_+ and P_- are not known, it is necessarily to estimate them (and any parameter which is a function of them) from the data. The next Section 2.3 is properly devoted to review principal estimates of ROC curves.

2.3 ROC curves estimation

Hitherto, the ROC curve and its most important properties have been presented from a theoretical point of view, assuming to know the quantities which generate the curve; however in many real situations, one only has empirically collected datasets and, therefore, the underlying populations and the associated ROC curve must be learned from the data (i.e. from the realization of given covariates or features). To achieve this, different approaches can be adopted, from a fully non-parametric empirical ROC estimation to the binormal

model, passing through kernel smoothing techniques ([11, 12, 24, 25]), as detailed in the next paragraphs. Other methods have been proposed to estimate the ROC function, to gain flexibility and better fit for atypical data (using Gaussian mixture [26]) or to obtain simpler analytic and closed form expressions [27, 28]. In addition, an attempt to exploit optimality of LR based procedures has been made with a semiparametric approach by Qin and Zhang [29], who proposed to estimate the log likelihood ratio function assuming an exponential link between f_+ and f_-:

$$\log \frac{f_+(x)}{f_-(x)} = \alpha + \beta^T r(x),$$

where $r(\cdot)$ is a smooth function and the parameters α and β can be estimated by maximum semi-parametric likelihood. However, this approach suffers from a certain level of arbitrariness in the choice of $r(x)$.

2.3.1 The empirical ROC

The empirical ROC curve is the simplest and most popular estimate: it is easily obtained computing (and plotting) the sample frequencies of false positive and true positive fractions for different threshold values t, i.e.

$$\widehat{FPR}(t) = \frac{\sum_{i=1}^{N_-} \mathbb{1}_{[S>t|P_-]}}{N_-}$$
$$\widehat{TPF}(t) = \frac{\sum_{i=1}^{N_+} \mathbb{1}_{[S>t|P_+]}}{N_+},$$

where N_- and N_+ are the number of healthy and diseased subjects. It is a discrete set of points linearly joined, and therefore the estimated ROC curve

$$\widehat{\mathrm{ROC}}(t) = \{(\widehat{FPR}(t), \widehat{TPF}(t)), t \in \mathcal{R}\}$$

results in a step function (as in Figure 2.4): if only one between $\widehat{FPR}(t)$ and $\widehat{TPF}(t)$ changes with t, horizontal and vertical jumps are reported on the graph (with larger size in presence of ties), whereas mutual changes of samples proportions produce diagonal segments.
The empirical ROC estimator is free from structural assumptions, non-parametric, depends only on the rank of the data and it is invariant for monotone increasing transformation of the decision variable. In addition, thanks to the law of large numbers, it uniformly converges to the theoretical curve ($\widehat{\mathrm{TPR}}$ converges to TPR and $\widehat{\mathrm{FPR}}$ converges to FPR for every fixed threshold) and, as detailed in [24], it also inherits good asymptotic properties of maximum

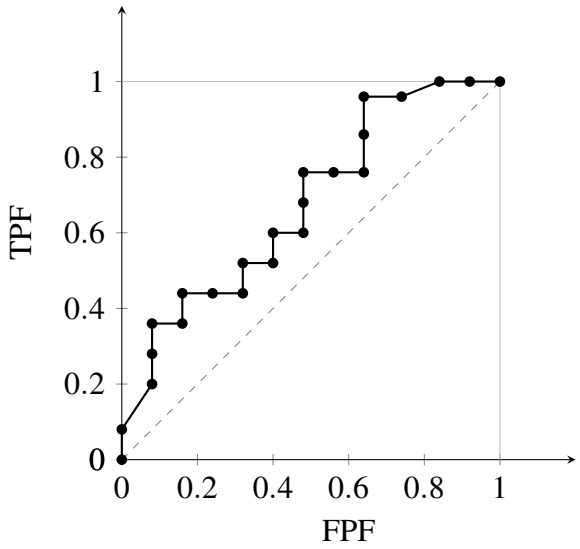


Figure 2.4 Example of an empirical ROC curve on fictitious data

likelihood estimates. However, empirical ROC curves are generally improper: they easily switch from roughly concave to roughly convex and vice versa and, actually, to speak of convexity or concavity it is not even applicable, since they are staircase-shaped functions. To overcome this drawback other smoother estimators are presented in the next paragraphs.

2.3.2 Nonparametric kernel smoothing

To build smooth curves retaining good properties seen before (in particular the distributional free assumption), a widely used approach relies on kernel density estimation (KDE) and it was firstly presented by Zou [25] and Llyod [30]. Indeed, Zou proposed to apply kernel density methods to estimate density functions f_+ and f_- under P_+ and P_- and then to numerically integrate to obtain the correspondent distribution functions estimates which appear in the ROC definition:

$$\widehat{\mathrm{ROC}}(x) = 1 - \widehat{F}_+(\widehat{F}_-^{-1}(1-x)).$$

Specifically, recalling the kernel density estimation theory, as reported in e.g. [31],

$$\hat{f}_s(x) = \frac{1}{n_s \lambda_s} \sum_{i=1}^{n_s} K\left(\frac{x - X_i}{\lambda_s}\right) \quad \text{with } s \in \{+,-\}$$

where λ_s is the bandwidth (i.e. the smoothing parameter) and K is the kernel function (a non-negative function integrating to 1). Intuitively, it is the sum of "bumps" placed at the

observations X_i, where the shape of bumps depends on the kernel. Common choices for the kernel function are:

- Gaussian: $\frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$
- Epanechnikov: $\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right) \quad \text{for} \quad |t| < \sqrt{5}$
- Biweight: $\frac{15}{16}(1-t^2)^2 \quad \text{for} \quad |t| < 1.$

Density estimation is quite robust with respect to the choice of the kernel, while it appears more sensible to bandwidth values; in particular, for $\lambda_s \to 0$ the estimator reduces to a sum of Dirac Delta functions spiked at the observations, whereas for large values of λ_s many details (e.g. multimodality of the data) are concealed by the resulting flattening function. Therefore, kernel functions and especially bandwidth values must be chosen carefully. Zou made some suggestions, however the optimal smoothing parameter for estimating the densities f_- and f_+ does not imply the correspondent optimality for the distributions and for the ROC curve. To overcome this, Lloyd directly estimated F_+ and F_- via KDE (providing optimal values for bandwidth selection of order $O(n^{-1/3})$) and showed that the mean squared error (MSE) for kernel estimators is asymptotically smaller than the empirical one [32], and therefore he recommended their use. In addition, Zhou and Harezlak [33] compared non-parametric kernel smoothing methods to find the best one for ROC estimation and suggested adopting the Altman approach (again of order $O(n^{-1/3})$), while Hall and Hyndman [34] investigated optimal bandwidths when F_+ and F_- are quite different functions. Nevertheless, kernel-based estimators of the ROC curve converge only pointwise to the theoretical true ROC and are not invariant under monotone data transformation; therefore other smoothing methods can be considered, such as local linear smoothing [35], Bayesian bootstrap [36], bandwidth-free smoothing of the empirical CDFs [37], direct smoothing of a distribution function $Z = 1 - F_-$ for which it is possible to express ROC curve as $\text{ROC}(t) = \mathbb{P}(Z \leq t)$ [38].

2.3.3 The binormal model

The binormal model is the reference model for ROC analysis. It assumes normal distributions for the two underlying and partially overlapping populations P_+ and P_-; it is highly robust and, in practice, it can be applied with good fitting results to a large set of different types of data: it is sufficient that data can be considered normal after a monotone transformation. In particular, let μ_+, μ_- and σ_+^2, σ_-^2 be the means and variances of P_+ and P_- populations respectively, with the common assumption that diseased subjects have higher values of the

decision variables; therefore, as reported by every book on ROC curves (see for example [11] or [12]), the ROC curve has the form:

$$\mathrm{ROC}(x) = \Phi(a + b\Phi^{-1}(x))$$

where $a = \frac{(\mu_+ - \mu_-)}{\sigma_+}$, $b = \frac{\sigma_-}{\sigma_+}$, $\Phi(\cdot)$ is the normal standard distribution function and $\Phi^{-1}(x)$ its quantile. If $b = 1$ (homoscedasticity) the binormal ROC curve is concave everywhere, whereas for $b \neq 1$ the curve presents always an hook (and the likelihood ratio of the decision variable is no longer monotone). This can be straightforwardly extended to the multivariate case (i.e considering multivariate normal distributions). In addition, for binormal ROC curve the AUC can be easily written in closed form, recalling that $AUC = \mathbb{P}(S_+ > S_-) = \mathbb{P}(S_+ - S_- > 0)$. Therefore, for independent S_+ and S_- normal distributions, $S_+ - S_- \sim \mathcal{N}(\mu_+ - \mu_-, \sigma_+^2 + \sigma_-^2)$ and

$$\begin{aligned}
\mathrm{AUC} =& \mathbb{P}\left(Z > 0 - \frac{(\mu_+ - \mu_-)}{\sqrt{(\sigma_+^2 + \sigma_-^2)}}\right) \\
=& 1 - \Phi\left(\frac{\mu_- - \mu_+}{\sqrt{\sigma_+^2 + \sigma_-^2}}\right) \\
=& \Phi\left(\frac{\mu_+ - \mu_-}{\sqrt{\sigma_+^2 + \sigma_-^2}}\right) \\
=& \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right)
\end{aligned}$$

2.4 The connection to Lorenz curve and Gini coefficient

It is well known that there is a strict relationship between the ROC and it summary index AUC and the Lorenz curve with the Gini coefficient. Actually, for situations with monotone likelihood ratio, the ROC curve can be obtained from the Lorenz curve applying a linear transformation. In Chapter 5 some interesting consequences due to this connection will be presented, while here definitions and results useful afterwards are introduced.

The Lorenz curve is a tool popular among economists to measure the wealth distribution in populations and to highlight income inequalities; as commonly defined, it is the plot of the cumulative percentage of income against the cumulative percentage of population: the more the wealth is concentrated in a small fraction of the population, the more the curve

is near the point $(1,0)$, whereas if there is a uniform distribution of the income the curve reduced to the $45-$degrees line from the origin ("line of perfect equality"). This idea has been transferred to the medical diagnostic field [3] with a slightly different interpretation, as the curve of the cumulative percentage of diseased subjects against the cumulative percentage of the non-diseased (as it can be seen from Figure 2.5), where the subject are classified in the two groups based on their likelihood-ratio score. To evaluate the "bowedness" of the curve

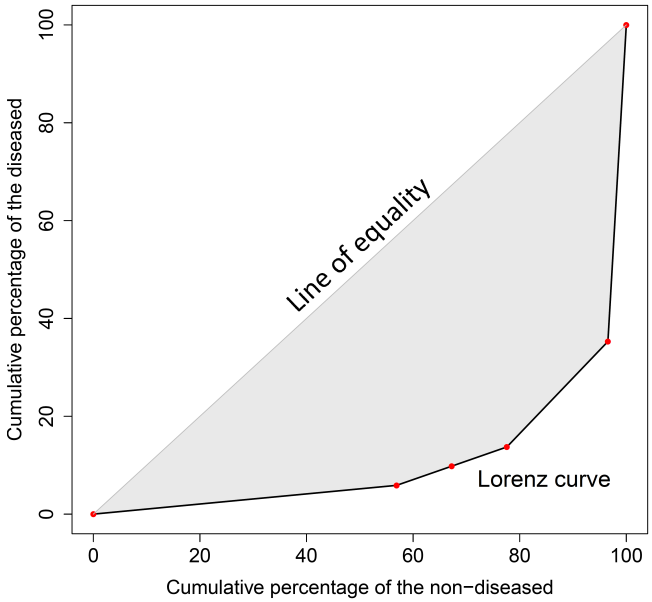


Figure 2.5 The Lorenz curve, adapted from [3].

it is convenient to use the Gini index, a summary measure of the area between the Lorenz curve itself and the diagonal line (the gray shaded area in Figure 2.5) which varies between 0 and 1, as the AUC. Indeed, often in classification and medical diagnostic papers the Gini index is directly defined via AUC, in the following way:

$$G = 2 * AUC - 1$$

even if this relation is precisely correct only if there is a one-to-one correspondence between the summary indices for the curves. In other words, the Lorenz curve should be a mirror image of the ROC curve: this happens when the two underlying populations have normal distributions with the same variance, and therefore the likelihood ratio is a monotone increasing function of the data (as already highlighted in [3, 39, 40]). If this is not the case, the Lorenz curve appears to be better, because it continues to be a concave function, while the (traditional) correspondent ROC curve presents hooks and a "wiggly" shape; it is therefore

necessary to apply a concavifying transformation, reordering the likelihood ratio values. More details will be presented in following chapters 4, 5.

Chapter 3

The motivating example

In this chapter the real problem that stimulated my interest for cancer biomarkers, classification and ROC curves is introduced. It has been raised by the Cancer Genomics Lab of the Edo and Elvo Tempia Foundation in Biella (Italy); I have been involved since the beginning and I took care of all the statistical analyses and evaluations needed for the development of the project.
The project aims to improve the diagnostic route of prostate cancer, finding new non-invasive and easy-to-detect biomarkers, to be used in combination, possibly with the prostate specific antigene (PSA) measurement (the standard non-invasive tool for prostate cancer detection). Beyond the general definition of biomarker and considerations on the relevance of a multivariate approach, some background information on the disease and on biological instruments and techniques are provided for a better understanding of the topic and of the data analysis. Part of the material presented here is extracted from the paper I co-authored "Circulating microRNAs combined with PSA for accurate and non-invasive prostate cancer detection", published on Carcinogenesis in 2018 [41].

3.1 Biomarkers: a general definition

Following the definition given by the US National Institute of Health [42], biomarkers are objective and measurable characteristics used as indicators of normal biological, pathogenic, pharmacologic processes or responses to an exposure or intervention. They are categorized based on their role and application: from susceptibility, diagnostic, prognostic and monitoring biomarkers for the evaluation of the risk, the detection and the progression of a medical condition of interest, to predictive, pharmacodynamic, and safety biomarkers related to the

assessment of the effects of a medical product or an environmental agent as well as the adverse events of treatments (e.g. the toxicity). Groups can partly overlap and the same quantity can belong to different categories; biomarkers can be simply physiologic measurements (blood pressure, body temperature, ...), radiological images or highly complex molecular variables (such as microarrays, miRNAarrays, gene polymorphisms and mutations, proteins, metabolites, ...). Their distinctive features are the non-invasiveness, the easy way for the measurement (for example through a blood sample) and the reproducibility. Furthermore, it is important to highlight that "a biomarker is not an assessment of how an individual feels, functions, or survives" [42] and it is distinct from the clinical outcome of interest (even if a good marker consistently and accurately predicts it).
Actually, it should be clear from the broader definition above that almost everything showing an interaction between a biological system and a potential hazard could be regarded as a biomarker. In particular, in the following, I will consider only diagnostic markers, i.e. biomarkers used to detect or confirm the presence of the condition of interest or to identify subjects with a subtype of the disease [42].
Usually, in medical field and in particular in oncology, it is important to detect diseases as early as possible to allow more effective treatments and better prognosis; diagnostic tests based on biomarkers measurements play an essential role for screening purposes if sensitivity (the ability of correctly identify diseased patients) and specificity (the ability to correctly classified healthy subject) are sufficiently high. Indeed, the perfect test should reach an accuracy of 100% (where the accuracy is a summary index for classification problem). However, often, single biomarkers are not enough accurate to enter the clinical practice and a multivariate approach is to be preferred: different biomarkers can catch different aspects of the disease and their combination results in an higher diagnostic power. Therefore, the best way to combine them acquires increasing interest: it is well known that the optimal classifier relies on the log-likelihood ratio statistics for diseased and healthy populations (for the Neyman-Pearson lemma, as already stated in Chapter 2, and as it will be detailed in Capter 4 with original results). However, many authors insisted on linear combinations [5, 43, 44] and logistic score [11]. The latter approach, more intuitive and easy to interpret also for non-statisticians, has been adopted in the study presented in the next paragraphs.

3.2 Some medical and biological notions

Prostate cancer (PCa): highly incident disease which occurs in men (especially after 50 years of age) and one of the most common causes of cancer-related death [45]. At present its

correct diagnosis relies on invasive tests for histological verification (such as multiple bioptic sampling), because the digital rectal examination (DRE) has low sensitivity [46], whereas the prostate-specific antigene (PSA) measurement (which was the standard screening tool till few years ago) is organ - but not tumor-specific and it leads to too high percentages of false positives and false negatives (with a positive predictive value of only $30-35\%$ [47]). Actually, it is still common to adopt a PSA cut-off of 4 ng/ml to firstly discriminate healthy from disease subjects, but PSA can result high and/or increased for factors different from the presence of cancer (for example due to the benign prostatic hyperplasia (BPH)), as well as aggressive tumours can be found in subjects with low levels of PSA. Other important prognostic factors are gleason score (GS, a value based on the microscopic appearance of prostate cancer; it varies from 2 to 10 and higher values indicate more aggressive tumors) and clinical TNM staging (cT). The latter is the internationally accepted standard for cancer staging, related to the anatomical extent and spread of cancer, where T category describes the primary tumor site, N category describes the regional lymph node involvement and M category describes the presence of distant metastases.

microRNA (miRNAs or miRs): endogenous, single-stranded non-coding small RNA molecules (containing about 22 nucleotides) usually located in cells and found in plants, animals and some viruses. miRNAs can be secreted into the extracellular space and can be detected in different body fluids (blood, saliva, urine ...) Circulating miRNAs have a remarkably stable form in plasma and serum; they control major pathways such as cell growth, proliferation, differentiation and survival, they can act as tumor suppressors or oncogenes and appear dysregulated in the early stage of cancers [48]. Tumor-associated miRs participate in intercellular communication and disseminate through the extracellular fluid to reach and influence the phenotype of remote targets. Therefore, free-cell circulating miRNAs have been proposed as possible biomarkers for several diseases, including PCa [49, 50]. On the other hand, most of the potential miRNAs biomarkers show small differences in their expression levels between healthy and diseased people: attention must be paid to blood sampling methods and a sufficient large sample size should be considered; in addition, often, several miRNAs are evaluated together to improve their diagnostic/prognostic effect.

DNA microarrays: molecular biological technique which allows to analyse the gene expression profiling of thousands of genes simultaneously. Actually, the microarray is a solid surface (the chip) filled by microscopic DNA spots (the features), each containing specific DNA sequences (the probes). The process strongly relies on the hybridization principle and the complementarity of the two DNA strands: fluorescently labelled target sequences bind to

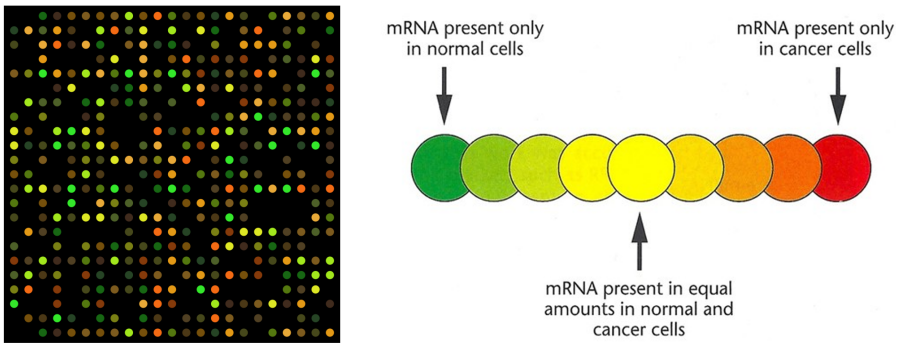


Figure 3.1 Example of a microarray chip with graphical interpretation of the features intensity signal. (Figures extracted from https://www.perkinelmer.com/it/category/cytogenetic-microarrays and https://www.discoveryandinnovation.com/BIOL202/notes/lecture26.html).

probe sequences and generate a signal; the intensity of the signal depends on the quantity of target sample bound to the probe and it is quantified for each feature. It is important to highlight that microarrays provide relative measurements of the intensity of a feature with respect to the intensity of the same feature under a different condition, and therefore, usually, the quantity of interest is the logarithm of the ratio of these intensities.

Real-time quantitative polymerase chain reaction (RT- qPCR): laboratory technique to evaluate gene expression in an accurate and reproducible way; it combines the DNA amplification and detection in a single step to find the exact amount of a target sequence or gene; in particular, during the amplification phase, a fluorescence is generated and measured during each PCR cycle (in this sense it is a real-time (immediate) procedure); the fluorescence increases with the increase of the number of gene copies during the reaction. The process essentially consists in a series of temperature changes.
Two important quantities in RT-qPCR are the threshold line and the C_t (see Figure 3.2):

- the threshold line (the black dashed line in the figure) highlights the point at which the reaction fluorescent intensity exceeds the background level and it is also indicated as the level of detection;
- the C_t value is the PCR cycles number at which the sample's reaction curve intersects the threshold line, i.e. it is the point in time when the target amplification is first detected. It is inversely proportional to the amount of target sequence: the lower the C_t level, the greater the number of target copies in the sample. Often, to be sure that the observed variations are due to real biological changes and not to technical issues,

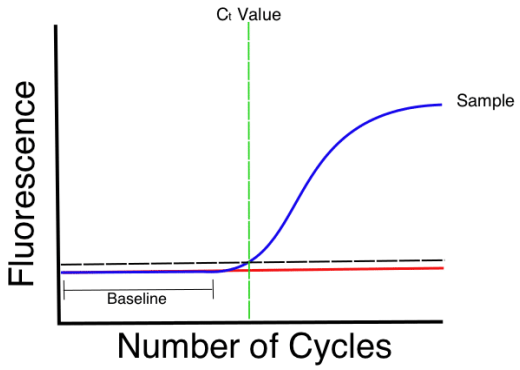


Figure 3.2 Threshold level and C_t value on a RT-qPCR amplification curve. (Figure extracted from https://bitesizebio.com/24581/what-is-a-ct-value/).

C_t levels are normalized by comparison with C_t levels of reference genes. The ΔC_t obtained in this way can be considered similar to the logarithm of the ratio of intensities of gene expression in microarrays analyses.

3.3 A new signature for prostate cancer detection

Studies conducted on normal and (pre)tumoral prostate tissues showed patterns of differentially expressed miRs according to disease status and/or severity [50, 51], with a certain degree of concordance among different laboratories. As already stated, due to their stability in body fluids, and particularly in serum/plasma, miRs could be ideal non-invasive biomarkers for PCa detection or prognosis prediction [49, 50, 52]. However, circulating miRs analysis is affected by variability in sampling procedures, RNA extraction and analysis methods, and high quality study designs are scarce. Only few published studies strictly checked hemolysis, discarded hemolysis-related miRs, optimized RNA extraction and miR quantification, and used independent validation cohorts.

Most of the proposed miRs are able to distinguish HD from PCa or HD from BPH, but perform very badly in distinguishing BPH from PCa. Some perform very well in specific datasets or have been found deregulated in more than one study, but the sign of deregulation is not always concordant [50] and their accuracy might suffer of overfitting. Indeed, validation of the results gained from a screening cohort on a prospective one, using techniques more suitable for diagnostic purposes such as RT-qPCR, remains still a hard task. Validation should be intended as applying the same classification rule, without any change to improve classification results according to the independent dataset. Moreover, several studies do not

include BPH in the dataset, or the size of this class is limited and does not reflect a real representation of BPH incidence. In addition, many prognostic miRs proposed so far derive from studies where only high-risk PCa are included in the analyses, without considering that these markers are aberrantly expressed in low-risk samples as well, questioning their use as prognostic markers.
The project, developed with the Cancer Genomic Labs of the Edo and Elvo Tempia Foundation, aimed to improve the diagnostic route of prostate cancer, proposing a new signature (i.e. a combination of genetic biomarkers) based on microRNAs freely circulating in plasma. It presents several added values, such as the quality of study design, adequate sample size, accurate sample collection and processing, and appropriate classifier validation. In particular, two independent datasets were analysed adopting different biological techniques (microarrays and RT-qPCR):

a) Discovery phase: miR profiling of 120 plasma samples to identify candidate miRs able to detect PCa more accurately than PSA alone;

b) Validation phase: cohort of 242 subjects, analysed by RT-qPCR, to test the classifier (obtained in phase a)).

Starting from more than two thousands miRNAs, using both univariate and multivariate statistical methods, a combination of two miRNAs (miR-103a-3p, let-7a-5p) and log-transformed PSA has been found able to classify PCa samples better than PSA alone and to avoid about one-third of unnecessary biopsies. Details are provided in the next paragraphs.

3.3.1 Materials & Methods

Samples from men with PCa, suspected symptoms of PCa but negative biopsies (BPH), high-grade prostatic intraepithelial neoplasia (HGPIN) or atypical small acinar proliferation (ASAP) were collected at the San Giovanni Battista Hospital of Turin prior to standard 12-core transrectal ultrasonography-guided biopsy. PCa were labelled, according to Gleason Score (GS) values, as $GS6$, $GS7$ or $GS > 7$, whereas, according to PSA, GS and tumor size (cT) within clinical TNM staging as low risk if $GS = 6$, $PSA < 10$, $cT < 2b$; as intermediate risk if $10 \leq PSA \leq 20$ or $GS = 7$ or $cT2b - cT2c$; as high risk if $GS = 8 - 10$ and/or $PSA > 20$ and/or $cT3 - cT4$. Clinically significant tumors comprised intermediate/high-risk PCa. Plasma from healthy donors (HDs) was gathered at Edo and Elvo Tempia Foundation from the same geographic area of the patients. HDs were in the same age range as patients, had negative DRE and $PSA < 4$ ng/ml, were not under any pharmacologic treatment nor had

any previous prostatic pathology.
The study was approved by the Ethics Committee. All subjects provided written informed consent with guarantees of confidentiality.
Plasma collection, processing and storage adhered to good practice rules. Hemolyzed samples or samples belonging to men with other cancer diagnosis were excluded from analyses. Plasma was isolated from ethylenediaminetetraacetic acid (EDTA) or lithium heparin blood samples within one hour from collection, with a standard procedure to prevent hemolysis (technical details on plasma isolation, storage and circulating RNA extraction can be found in Appendix A).
miR profiling was carried out on 138 samples, 120 of which (60 PCa, 51 BPH and 9 HD) were homogeneously collected in heparin tubes and represented the discovery phase dataset (Table 3.1). The remaining 18 HD samples were collected in EDTA tubes and were used for comparison purposes only.
Cases and controls were homogeneous for geographic area, collection times, plasma separation method, storage and hemolysis level. For the discovery phase, samples homogeneous for age, within each disease class, were selected. In the validation phase, samples were consecutively collected and included HGPIN or ASAP lesions as well. This cohort was enriched in older than 65 year-old men and in intermediate-risk PCa.

Statistical Analysis Raw data were processed using the limma R package [53] for microarray analysis. Background correction and inter-array normalization were performed applying normexp (offset= 20) and quantile methods, respectively. (Raw and average normalized $\log_2$Intensities are available in the Gene Expression Omnibus public functional genomics data repository (www.ncbi.nlm.nih.gov/geo/) with the following identifier: GSE113234).
An online tool (https://biostatistics.mdanderson.org/MicroarraySampleSize/) was used to compute sample size for class comparisons: knowing the total number of analyzed miRs (2006) and inputting 1% as acceptable percentage of false positives, 0.8 as desired statistical power, 0.25 as standard deviation and 0.2 as minimum $\log_2$fold-change, the minimum number of samples per class should be 36.
To compare miR profiles between classes, linear model and empirical Bayesian analyses were combined using the limma R package. Top differentially expressed miRs were selected using 0.2 as cutoff for the $\log_2$fold-change in PCa versus non-PCa samples and 0.05 for raw p-value. To compare quantitative real-time polymerase chain reaction (RT-qPCR) data between PCa and non-PCa samples, Student's t-test with Benjamini– Hochberg correction for multiple testing was applied, whereas to compare PSA levels among different disease classes, analysis of variance with Dunnett correction for multiple testing was used. In both

Table 3.1 Study populations for the discovery and the validation phases

	Discovery phase dataset	**Validation phase dataset**
Samples	120	242
PCa	60	68
BPH	51	93
HGPIN/ASAP		8
HD*	9	73
Age		
Median	65	68
≤ 65	61	89
> 65	59	153
PSA	5.96 $(4.42 - 8.40)^{\dagger}$	4.91 $(1.80 - 7.26)^{\dagger}$
PSA ≤ 4	26	105
$4 <$ PSA ≤ 16	79	123
PSA > 16	15	14
Gleason score		
GS6	25	10
GS7	22	40
GS > 7	13	18
Risk class		
PCa low risk	21	9
PCa intermediate risk	24	36
PCa high risk	15	23

* Only heparin-collected HD are reported. Other 18 EDTA-collected HD were analyzed by microarrays but they were not included in the classifier construction.

†Median (first to third quartile).

cases, differences were considered statistically significant if adjusted p-values were < 0.05. Looking at the data, the $\log_2$PSA distribution density of the PCa-BPH cohort of more than 400 consecutively collected samples available at the San Giovanni Battista Hospital in Turin showed an inflection at 4 that corresponds to PSA $= 16$ (Figure 3.3, red curve) and, within PSA > 16 ng/ml samples, there was a strong PCa enrichment; therefore, in the proposed classifier, the latter were directly considered as PCa and only samples with PSA ≤ 16 ng/ml were used to build a score that combines PSA with other variables. The appropriateness of this cutoff was further verified by observing the $\log_2$PSA distribution density of another big cohort of nearly 14000 asymptomatic men over 50 years who tested their PSA levels at Tempia Foundation from 2012 to middle 2018 (Figure 3.3, black curve), within a spontaneous adhesion context. From PSA $= 16$ on, this curve is very close to zero. Using $\log_2$Intensities of differentially expressed miRs in PCa within PSA ≤ 16 ng/ml samples, $\log_2$PSA and $\log_2$Age as input variables, a logistic regression model with Least Absolute Shrinkage and

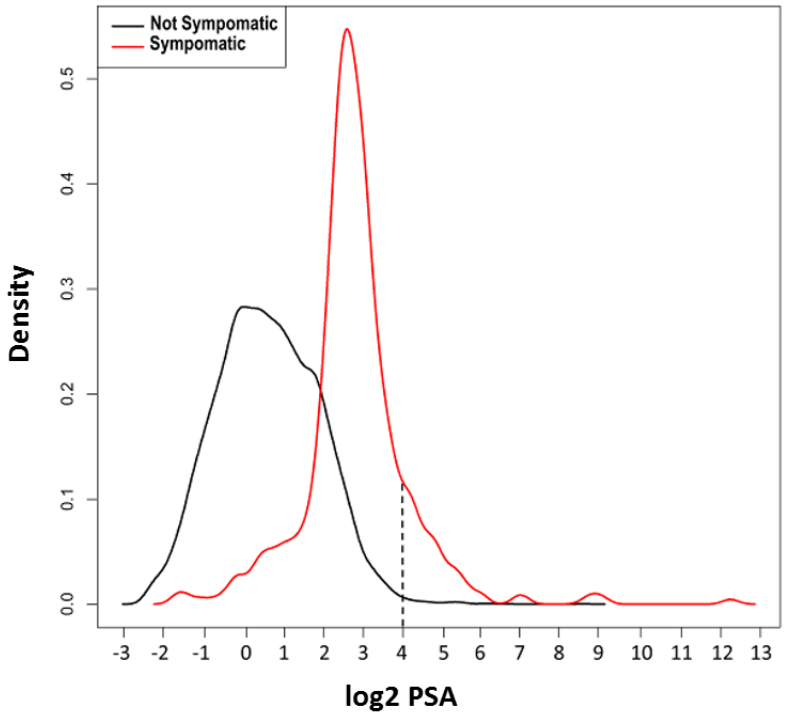


Figure 3.3 $\log_2$PSA distribution density. Red curve: cohort of 430 BPH or PCa patients available at the San Giovanni Battista Hospital of Turin. Black curve: cohort of nearly 14000 not symptomatic men over 50 who tested their PSA levels at our Foundation from 2012 to middle 2018, within a spontaneous adhesion context. The value of 4, corresponding to PSA $= 16$, is highlighted.

Selection Operator (LASSO) penalty [54] was fitted to build a classifier able to discriminate PCa from non-PCa, using the glmnet R package [55]. Five-fold cross-validation was applied to find the best tuning parameter. Following the standards of binary regression, the estimated log odds-ratios of these variables were multiplied by their values and then summed to build a score: if the score was higher than 2.02 (an *ad hoc* threshold based on data to minimize the number of false negatives and false positives), the sample was classified as PCa. This way, the resulting overall classifier turned out to be a combination of the initial PSA check ($>$ 16 ng/ml) and the score. The accuracy of the classifier was measured by the area under the receiver operating characteristic curve (AUC).

Quantitative real-time polymerase chain reaction First, seven miRs were evaluated on the same EDTA-HD samples used for miR profiling to validate the expression changes using RT-qPCR. Subsequently, they were tested on an independent set of EDTA plasma samples (10 HD, 10 BPH, 10 PCa). The two validated miRs were then tested on a validation phase dataset

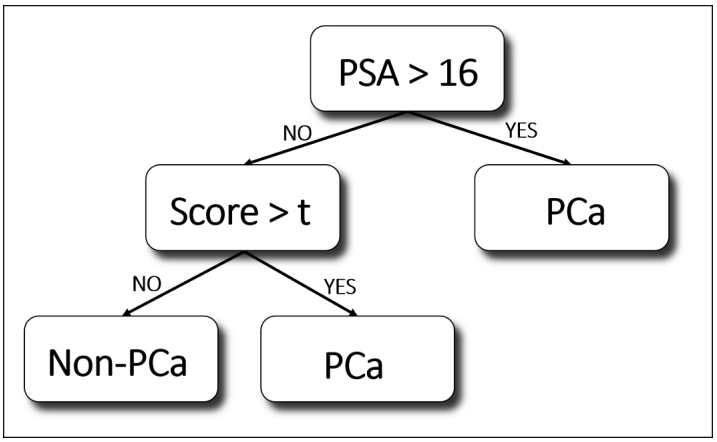


Figure 3.4 Decision tree of the final classifier developed in the discovery set and applied to the validation set.

(Table 3.1), a larger set of 242 consecutively and prospectively collected plasma samples (68 PCa, 93 BPH, 8 HGPIN/ASAP, 73 HD), all collected in EDTA tubes. cel-miR-39-3p was used as exogenous normalizer and each miR as well as negative controls were run in triplicate. Data transformation was done with the ΔC_t method, where C_t^{miR} = average C_t of the three replicates and

$$\Delta C_t^{\text{miR of interest}} = C_t^{\text{miR of interest}} - C_t^{\text{cel-miR}-39-3\text{p}}.$$

Final C_t (C_{tn}) was given by:

$$C_{tn}^{\text{miR of interest}} = -\Delta C_t^{\text{miR of interest}} + K,$$

where $K = 6.2$ is a constant chosen to make $-\Delta C_t$ ranges comparable with microarray $\log_2$ intensity ranges. To calculate K, only BPH and PCa samples were considered (because the two cohorts differed in terms of percentages of HDs and precancerous lesions were not included in the validation set).
The means of microarray $\log_2$Intensities for the two miRs were calculated and then averaged, yielding 6.6. Then K was calculated in order to make the average of mean $C_{tn}^{\text{miR}-103\text{a}-3\text{p}}$ and mean $C_{tn}^{\text{let}-7\text{a}-5\text{p}}$ equal to 6.6.
An independent set was used to test the classifier: the same previously generated coefficients (estimated log odds-ratios) of the two miRs were multiplied by the corresponding C_{tn} and the coefficient of PSA was multiplied by $\log_2$PSA value; the results were summed to build the score for each sample; the same classification rule (Figure 3.4) was applied to assign samples to the PCa or non-PCa groups.

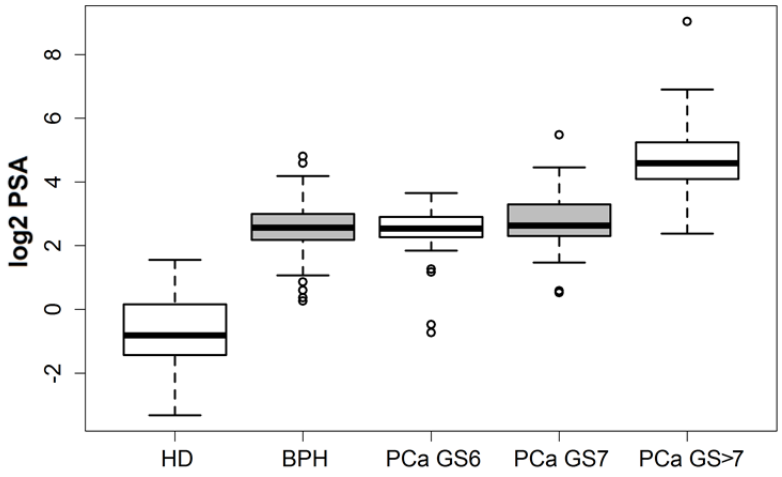


Figure 3.5 Box-plots for $\log_2$PSA values in the 138 plasma samples cohort profiled with microarrays, divided according to risk class.

3.3.2 Results

Discovery phase. One hundred thirty-eight not hemolyzed plasma samples were analyzed by miR profiling; $\log_2$PSA distribution in HD, BPH and PCa groups according to risk class is shown in Figure 3.5. Age was homogeneously distributed among classes. Dunnett-corrected p-values for comparisons between low-risk PCa or intermediate-risk PCa and BPH were equal to 0.99 and 0.91, respectively, and for comparisons between HD or high-risk PCa and BPH were < 0.001. PSA distinguished HD from BPH/PCa, high-risk from low/intermediate risk tumors, but was incapable of discriminating between BPH and low/intermediate-risk tumors.

EDTA-HD versus heparin-HD comparison revealed that the use of heparin-coated tubes did not affect miR profiling (Figure 3.6, panel A and B). Indeed the procedure was not inhibited by heparin, as it would be expected by methods requiring the use of reverse transcriptase or DNA/RNA polymerase. However, the collection method influenced the average intensity of detectable miRs, higher although not statistically significant in EDTA-HD samples; both methods performed similarly in terms of number of detectable miRs, even if each of them could detect a specific set of molecules. One hundred twenty homogeneously collected samples (Table 3.1) were used to highlight numbers and types of miRs detected in each class (Figure 3.6, panel C and D) and compare their levels: circulating miRs could be detected in the plasma of both patients and controls, with the highest average number observed in BPH, followed by GS>6 PCa, GS6 PCa and then HD. Overall, 107 circulating miRs were detected in all classes: 8 were shared by BPH and PCa, 4 by BPH and HD, 1 by HD and PCa, 9 were

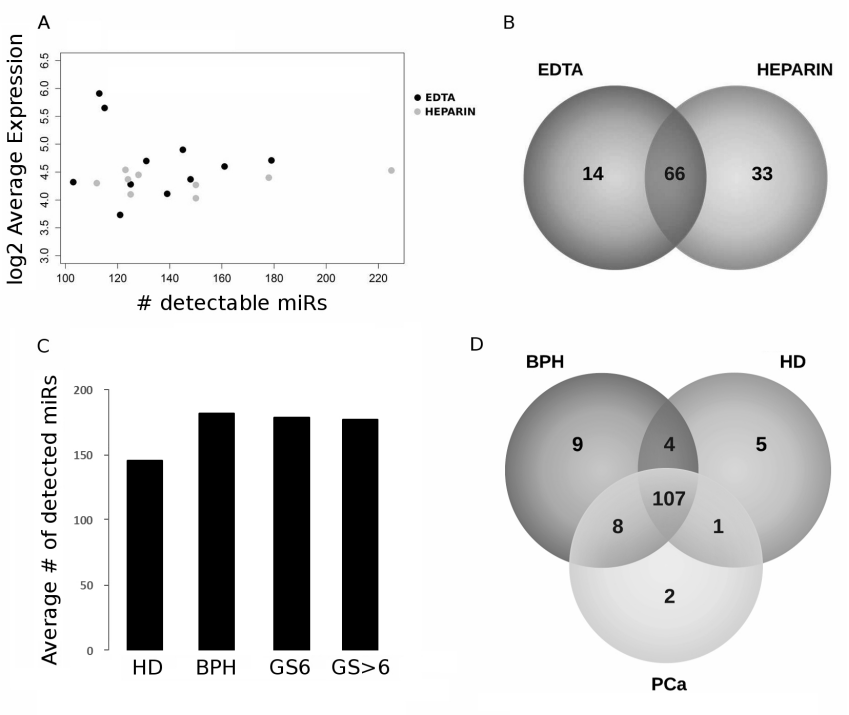


Figure 3.6 Comparison between EDTA-HD versus heparin-HD tube.

exclusively found in BPH, 5 exclusively in HD and 2 in PCa.

PCa (n = 60) versus HD + BPH (n = 60) analysis evidenced one downregulated (miR-4530) and nine upregulated miRs (let-7a-5p, miR-103a-3p, let-7f-5p, miR-17-5p, miR-4454, miR-130a-3p, miR- 15b-5p, miR-24-3p, miR-21-5p) in PCa. Expression changes and p-values are shown in Figure 3.7, panel A, whereas sample and miR clusters are depicted in Figure 3.7, panel B. Unsupervised clustering was not able to clearly separate PCas from BPHs and HDs. On the other hand, the comparison between PCa (n = 60) and BPH (n = 51) resulted in only two upregulated miRs (let-7a-5p and miR-103a-3p).

Restriction to PSA $\leq$ 16 ng/ml samples (48 PCa versus 57 BPH + HD) resulted in 11 upregulated miRs (miR-103a-3p, let-7a-5p, let-7d-5p, miR-17-5p, let-7f-5p, let-7b-5p, miR-24-3p, miR-26a-5p, miR-20a-5p, miR-130a-3p and miR-15b-5p), whereas analysis of $4-16$ ng/ml PSA samples (39 PCa versus 40 BPH) yielded only downregulated miRs (miR-4530, miR-1207-5p, miR-575, miR- 4739, miR-1202, miR-3679-5p, miR-6085, miR-3656, miR-663a, miR-4687-3p, miR-5739).

In order to find a strategy to classify samples, individuals with PSA $>$ 16 were considered independently and directly classified as PCa: 12 were true PCa whereas 3 were false positives. Using only samples with PSA $\leq$ 16, the most recurrent variables (out of the 11 upregulated miRs, $\log_2$Age and $\log_2$PSA) with coefficient $\neq 0$, selected by the penalized

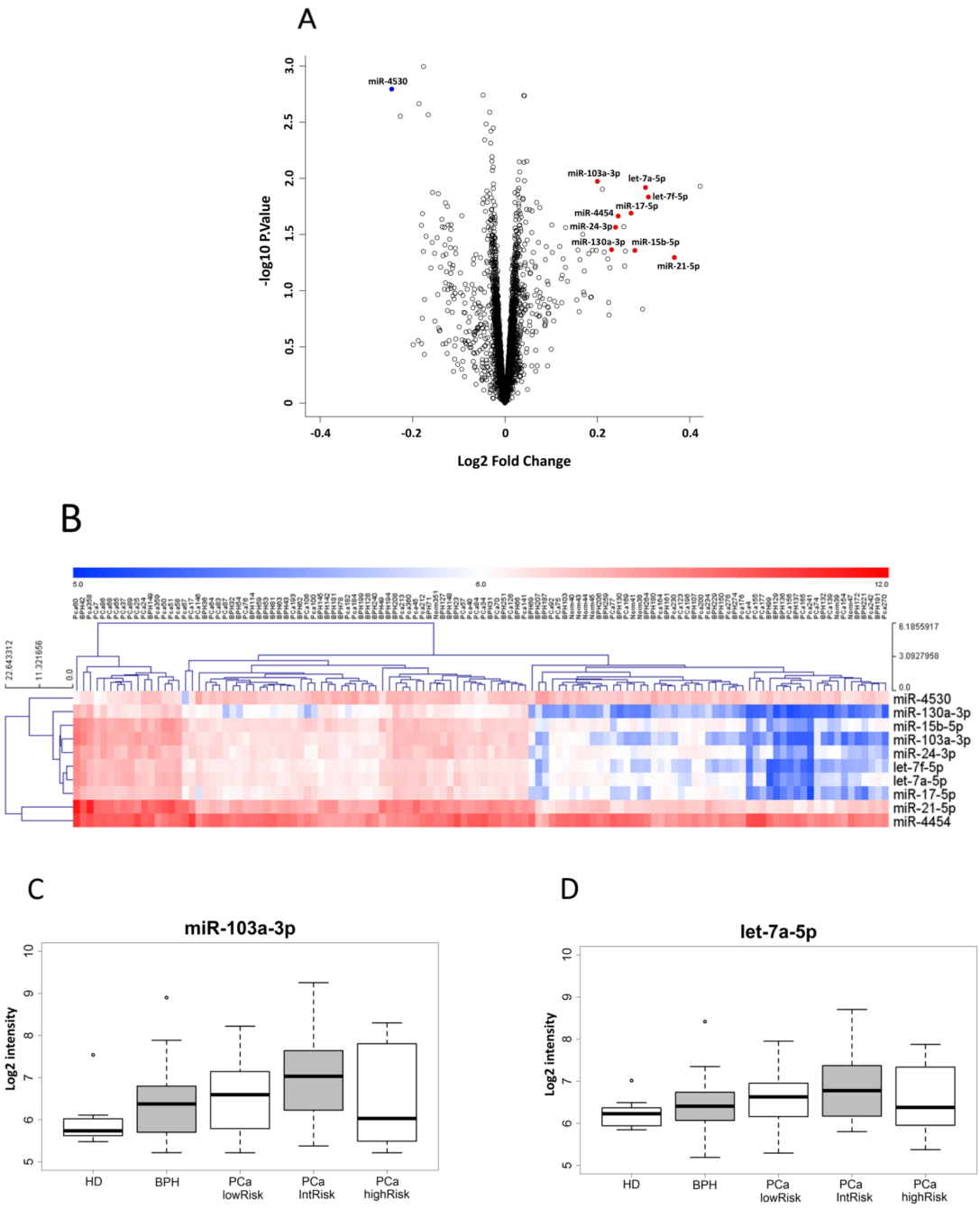


Figure 3.7 (A) Volcano plot showing $\log_2$fold-changes (x axis) and $-\log_{10}$p-values (y axis) of the miR probes analysed, highlighting up-regulated (red circles) and down-regulated (blue circles) miRs in the comparison between PCa (n=60) and BPH+HD (n=60). (B) Unsupervised hierarchical clustering of the expression matrix of 120 plasma samples (columns) and the 10 miRs (rows) that are differentially expressed between PCa and BPH+HD. (C) Box-plots of $\log_2$Intensities for miR-103a-3p in the discovery cohort, according to sample class. (D) Box-plots of $\log_2$Intensities for let7a-5p in discovery cohort according to sample class.

logistic regression model, were miR-103a-3p, let-7a-5p and $\log_2$PSA, all associated with a positive coefficient (0.1994, 0.1294, 0.0385 respectively). Figure 3.7 (panel C and D) depicts $\log_2$Intensity box-plots for the two miRs in each sample class. Dunnett corrected p-values for comparisons of each sample class with HD are all higher than 0.05, except for intermediate-risk PCas (0.0096 and 0.0482 for miR-103a-3p and let-7a-5p, respectively); similarly, Dunnett corrected p-values for comparisons of each sample class with BPH are all higher than 0.05, except for intermediate-risk PCas (0.0197 and 0.0276 for miR-103a-3p and let-7a-5p, respectively).
A score was built by summing the products of the coefficients by the variable values, and samples were ordered by increasing score. Samples with score greater than 2.02 (chosen to optimize accuracy) were classified as PCa. The final classifier was then built combining this score with the initial PSA check (Figure 3.4), obtaining an AUC of 0.68 (95% CI: $0.59-0.78$), whereas the AUC of PSA alone was 0.62 (95% CI: $0.53-0.73$; Figure 3.8). $36/39$ (92%) clinically significant PCa were detected by the final classifier, whereas $34/39$ (87%) by PSA alone. Only three intermediate and five low-risk PCa were misclassified. All high-risk PCa were identified, as well as $7/9$ (77%) PCa with PSA ≤ 4 ng/ml, all in men falling in the $50-69$ age range. Moreover, $9/40$ (22.5%) BPH with PSA > 4 ng/ml and all HD were correctly classified.

Validation of miR expression by an independent technique Seven potentially interesting miRs, selected from the discovery phase, were further evaluated by RT-qPCR (Table 3.2) on the EDTA-HD samples profiled with microarrays and on 30 EDTA-collected independent samples (10 PCa, 10 BPH, 10 HD). Because heparin inhibits RT, RT-qPCR could only be applied on EDTA-HD samples as independent technique to validate expression changes. The correlation between array intensities and RT-qPCR relative expression was higher than 0.8 only for miR-103a-3p and let-7a-5p (the top two upregulated miRs in all previous analyses, and the ones included in the classifier), reinforcing their robustness. Even if miR-21-5p was detectable in all samples, correlation between the two techniques was not satisfactory (correlation coefficient for miR-21-5p = -0.08, versus 0.87 and 0.90 for let-7a-5p and miR-103a-3p, respectively). Because the other miRs did not yield detectable and reproducible C_t in all the samples, their correlation coefficients were not calculated. However, they were further evaluated in the independent group of 30 samples including BPHs and PCas. Still they gave either C_t higher than 40 or non-specific amplification products (as observed by melting curve analysis), or coefficients of variation higher than 0.1. They were then considered as undetectable. In this independent cohort, statistically significant upregulation in PCa versus

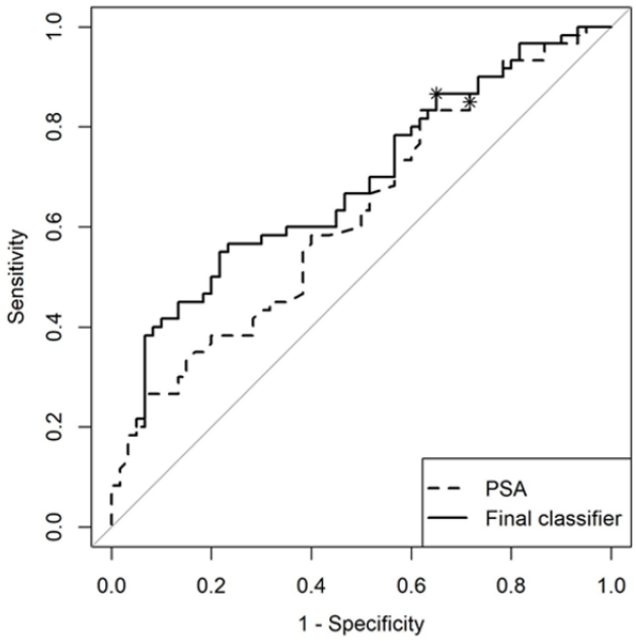


Figure 3.8 ROC curve for PSA (dotted line) and for the final classifier (continuous line) in the discovery cohort; asterisks correspond to the thresholds for PSA (4 ng/ml) and for the final classifier (2.02).

HD + BPH resulted for both miR-103a-3p and let-7a-5p (Figure 3.9) with $\log_2$fold-change > 1 and Benjamini-Hochberg corrected t-test p-value < 0.05.

Validation of the classifier on an independent cohort Additional 242 independent plasma samples (Table 3.1), consecutively collected and checked for hemolysis, were used for the validation phase. Figure 3.10, panel A shows $\log_2$PSA box-plot among different sample classes; Dunnett corrected p-values for comparisons between HGPIN/ASAP or PCa low-risk and BPH are equal to 1, between intermediate-risk PCa and BPH equal to 0.99 and between HD or high-risk PCa and BPH less than 0.0001. Figure 3.10 panel B and C refer to the $\log_2$Intensities of the two miRs; Dunnett corrected p-values for comparisons of each sample class with HD are all less than 0.0001, while comparison with BPH yielded statistically significant p-values only for HD (< 0.0001, both for miR-103a-3p and let-7a-5p). Out of 242, 14 samples had PSA > 16 ng/ml and were directly classified as PCa: 10 were high-risk PCa, whereas 4 were false positives. The remaining 228 samples were classified using the same exact score coefficients and cutoff generated in the discovery step.
The final classifier yielded an AUC of 0.76 (95% CI: $0.70 - 0.82$) (Figure 3.11, panel A), whereas the AUC of PSA alone was 0.74 (95% CI: $0.68 - 0.80$). In particular, the new

Table 3.2 Differentially expressed miRs based on the class comparisons performed and ordered according to their occurrence in the analysis results.

microRNA	**60 PCa vs 60 BPH+HD** (0<PSA<500)	**48 PCa vs 57 BPH+HD** (PSA<16)	**60 PCa vs 51 BPH** (0<PSA<500)	**39 PCa vs 40 BPH** (4<PSA<16)	**RT-qPCR**
miR-103a-3p	↑	↑	↑		↑
let-7a-5p	↑	↑	↑		↑
miR-21-5p	↑				↑
miR-4530	↓			↓	undetectable
let-7d-5p		↑			
let-7f-5p	↑	↑			
miR-17-5p	↑	↑			
miR-26a-5p		↑			
miR-130a-3p	↑	↑			
miR-15b-5p	↑	↑			
miR-24-3p	↑	↑			
miR-4454	↑				
let-7b-5p		↑			
miR-20a-5p		↑			
miR-1202				↓	undetectable
miR-3679-5p				↓	undetectable
miR-6085				↓	undetectable
miR-3656				↓	
miR-1207-5p				↓	
miR-575				↓	
miR-4739				↓	
miR-663a				↓	
miR-4687-3p				↓	
miR-5739				↓	

classifier correctly identified $8/9$ (89%) patients with PCa and PSA ≤ 4 ng/ml, 7 of which harboured clinically significant (3 high-risk and 4 intermediate-risk) tumors. Of note, three of them fell in the $50-69$ age range and had negative DRE. $70/73$ HD (96%) were correctly identified. The AUC of the score, for PSA values lower than 4 ng/ml, was 0.86 (95% CI: $0.77-0.95$) whereas PSA alone had an AUC of 0.79 (95% CI: $0.59-0.98$; Figure 3.11, panel B). In the $4-16$ PSA range, the classifier yielded an AUC of 0.6 (95% CI: $0.43-0.70$) and correctly identified $38/49$ (78%) PCa, 31 of which were clinically significant, and $25/74$ (34%) non-PCa. The AUC of PSA alone, in the same interval, was only 0.47 (95% CI: $0.36-0.57$; Figure 3.11, panel C).

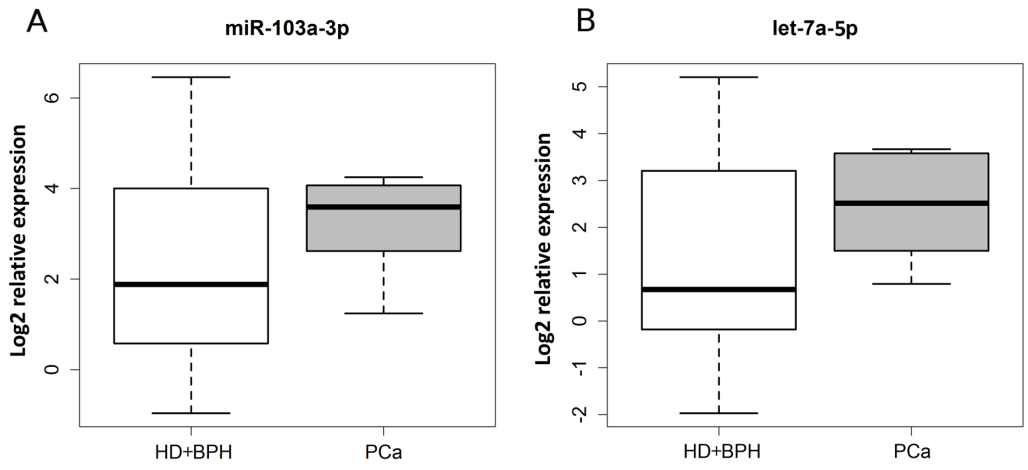


Figure 3.9 Box-plots of ΔC_t values for miR-103a-3p and let7a-5p, showing statistically significant upregulation for both microRNAs.

3.3.3 Some considerations

The two miRs included in the diagnostic score are miR-103a-3p and let-7a-5p. miR-103a-3p was proposed as a plasmatic endogenous normalizer even by the Exiqon protocol, but several recent studies highlighted that its levels in body fluids are not stable and it may work with other miRs as putative diagnostic [56] or prognostic [57] circulating biomarker. It was included into diagnostic/prognostic PCa serum scores by Mihelich et al. [58], and was able to predict biochemical relapse together with PSA after prostatectomy [59]. Another study, where miRs were analyzed in expressed prostatic secretion, found miR-103 associated with prostatitis [60]. Let-7a-5p belongs to the let7 family of tumor suppressors and is usually downregulated in PCa versus normal or BPH tissues [61]. Although it is now well established that cancer cells may release tumor suppressor miRs in the blood stream to get rid of them and prevent their antitumor effect, let-7a-5p levels have been found either up - or downregulated in PCa compared with controls [50], or positively associated with PCa reclassification upon active surveillance [49]. Its plasma levels strongly vary depending on whether extracellular vesicle-incorporated or cell-free miRs are analyzed [62].

This study meant to translate results into clinical practice. Therefore, the samples have not been enriched in extracellular vesicle-incorporated miRs, as this analysis would have inserted other sources of variability and technical challenges, even though this fraction of miRs might be informative for specific PCa diagnosis or prediction of prognosis [63, 64]. Instead, the new proposed classifier is easily applicable as long as blood is carefully collected and plasma quickly isolated.

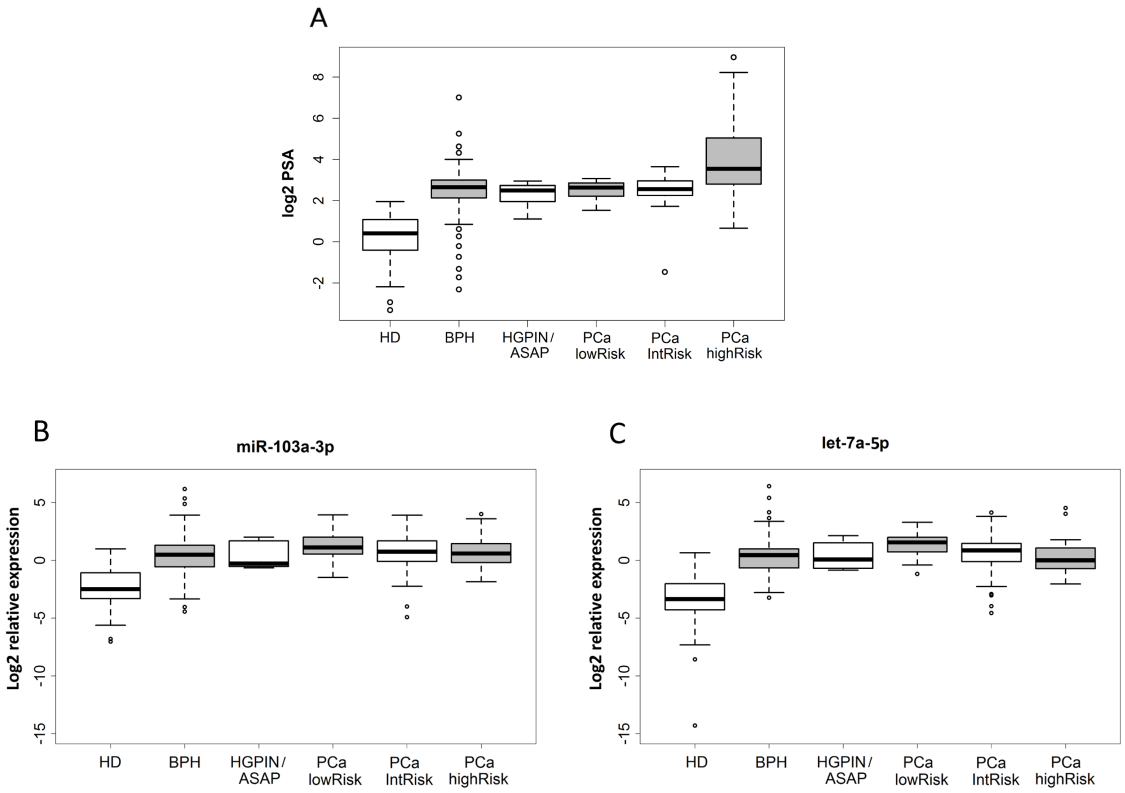


Figure 3.10 (A) $\log_2$PSA in the 242 sample validation cohort, according to sample class. (B) Box-plots of $-\Delta C_t$ values for miR-103a-3p. (C) Box-plots of $-\Delta C_t$ values for let7a-5p.

The proposed classifier combines two miRs with PSA and was able to discriminate PCa from non-PCa and to identify clinically significant PCa, better than PSA alone in the discovery cohort. The same methodology, applied to the validation cohort, allowed for identification of all but one low-PSA tumors: three high-risk, four intermediate-risk and one low-risk PCa. This is an important improvement, given that three of them were found in men with negative DRE and age in the $50-69$ range that represents the group for which screening for PCa could provide benefits [65]. Moreover, for two of them PSA was even < 2.5 ng/ml, limit after which free-PSA is dosed (if free to total PSA ratio is < 0.2, further investigations are recommended). Free to total PSA ratio was not available for the studied cohorts, therefore no comparison between scores has been considered. However, the use of free to total PSA ratio or of other PSA-related markers, which are currently under investigation [66–68], is still questionable. In the $4-16$ ng/ml PSA range, the classifier identified $25/74$ (34%) carriers of non malignant lesions, who may avoid unnecessary and harmful biopsies, and correctly classified 78% of PCa. Although in the discovery phase the final classifier outperformed PSA

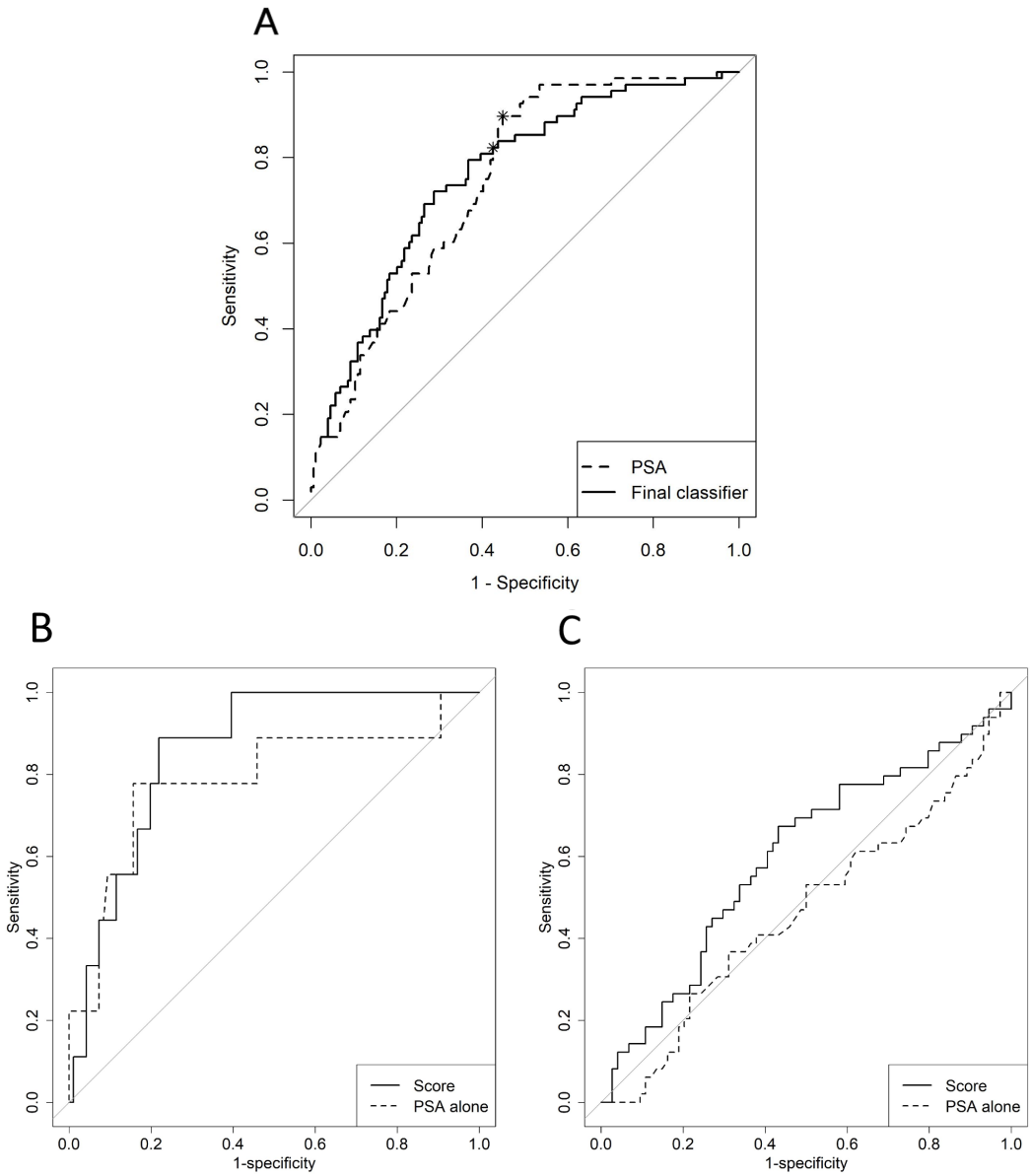


Figure 3.11 (A) ROC curve for PSA (dotted line) of the final classifier (continuous line) in the validation cohort of 242 samples; asterisks correspond to the thresholds for PSA (4 ng/ml) and for the final classifier (2.02). (B) ROC curve for PSA (dotted line) or the score (continuous line) in the validation cohort, considering 105 samples with PSA $\leq$ 4 ng/ml. (C) ROC curve for PSA (dotted line) or the score (continuous line) in the validation cohort, considering 123 samples with 4 $<$ PSA $\leq$ 16 ng/ml.

alone in detecting overall (Figure 3.8) and clinically significant PCa, in the validation cohort its AUC, specificity and positive predictive value were higher than those of PSA alone, at the cost of a lower overall sensitivity (Figure 3.11, panel A). However, it strongly outperformed PSA sensitivity in the $0-4$ PSA interval (Figure 3.11, panel B) and PSA specificity (Figure 3.11, panel C) in the $4-16$ PSA interval. Indeed, all but one missed PCa fall in this critical interval for PSA, where intense research is still underway [69]. It is also important to notice that diagnosis was based on standard rather than mp-MRI-targeted biopsy.
Furthermore, it is noteworthy to highlight that, comparing discovery and validation cohort results, unexpectedly, the AUC is higher in the validation cohort (0.76) than in the discovery cohort (0.68); this unlikely occurrence characterizes not only the new proposed signature, but also the classifier based on PSA alone. It is probably due to the composition of the validation cohort itself, which was enriched in older than 65 year-old men and in intermediate-risk PCa: the classification task resulted therefore easier in this dataset.

To further validate and improve the proposed classifier, a large multicentre prospective study for men in the $50-69$ age range has been planned, in which PSA and DRE will be coupled with circulating miR analysis and mp-MRI, and standard biopsy to mp-MRI targeted biopsy. Clearly, besides microRNAs, other genomic data could be studied as new potential biomarkers; for example metabolites (small molecules related to metabolic processes, with different functions and which represent the interaction between genes and the environment, and therefore characterized by high variability) are quantities of interest. In this perspective a pilot study has been developed on a subset of 102 samples (with $4 < \text{PSA} < 16$) from the validation cohort. Preliminary results shows that a combination of three metabolites, a miRNAs and the age returns a score able to discriminate PCa and BPH with a AUC$= 0.75$. Choosing an *ad hoc* cut-off to avoid false negative in the group of high risk PCa, sensibility raises to 0.92 (with a specificity of 0.42), lowering the percentage of false positive of about 40%. Further analyses are necessary to confirm these first encouraging findings.

Chapter 4

A minimal binary classifier providing a proper ROC curve

As seen in Chapter 2, the ROC curve is a common graphical tool to evaluate the discriminant power of a classifier in binary populations, well-defined and deeply studied from a theoretical point of view. Interesting issues originate in the estimation: it would be desirable to have optimal, always concave, continuous and non-decreasing curve. Thanks to the Neyman-Pearson lemma, the population ROC curve associated with a likelihood-ratio based classification rule satisfies all these nice properties. However, in practice suboptimal ROC curves are often used, because they may be associated to convenient scores (for example, linear combinations of features) which are not necessarily based on the likelihood ratio, and also because, when it comes to estimate them based on sample data, the most popular ROC estimates are the empirical curves (stepwise functions, neither convex nor concave). To the best of my knowledge, the ROC curve of a LR-based classifier has not yet been implemented. In this chapter, after an example of a LR-based classifier for two multivariate normal measures (which actually boils down to Fisher [70]) and a short introduction to näive Bayes classifier (so popular at present in the Machine Learning literature), a novel algorithm to draw a proper ROC curve in a multivariate setting is proposed, exploiting the flexible Bayes assumption, the Gaussian kernel density estimation and the local independence of the features. The procedure allows to overcome other methods' drawbacks: the jagged shape of the empirical ROC and the possible non-concavity of the standard binormal model (which always has hooks in the heteroscedastic case).
Lastly, the new estimation method is applied to a real case study (the validation cohort of the work on circulating miRNAs presented in Chapter 3 and kindly provided by the Genomics Lab of the Tempia Foundation in Biella, Italy) and shows a definitive advantage over the

empirical ROC and the ROC curves associated with common classifiers based on best linear and logistic scores.
In the following a standard binary classification problem is considered: the two population are represented by absolutely continuous probability measures P_+ (the diseased subjects for example) and P_- (the healthy), with respective density f_+ and f_-; $\boldsymbol{x}$ is the p-dimensional random vector of features and the population LR is defined as

$$\mathrm{LR}(\boldsymbol{x}) = \frac{f_+(\boldsymbol{x})}{f_-(\boldsymbol{x})}.$$

The material presented here is partly extracted from the paper: "Sacchetto L, Gasparini M. A minimal binary classifier providing a proper ROC curve" just submitted for publication.

4.1 An example of LR-based classifier: Fisher's LDA and QDA

It is widely recognised [9–12] and already stated in Chapter 2 that the criterion which classifies a subject as positive if

$$\mathrm{LR}(x) > t \quad \text{for some threshold} \quad t \in \mathcal{R} \tag{4.1}$$

is optimal. However, often, it is preferred to use a decision rule based on a one-dimensional score S which is not necessarily the LR, i.e. to classify a subject as positive if

$$S(x) > t \quad \text{for some threshold} \quad t \in \mathcal{R} \tag{4.2}$$

for some random variable S conveniently chosen. Of course, if S happens to be a monotone function of the LR, then decisions (4.1) and (4.2) are equivalent.
As an example, X is a one-dimensional normal variate under both P_+ and P_- with different means $\mu_+ > \mu_-$ and different variances σ_+^2 and σ_-^2, then the LR is a quadratic function of X and decision rule (4.1) is not amenable to (4.2). This is the well-know binormal model, which can be generalized to Fisher [70] linear and quadratic discriminant analyses (LDA, QDA). Even if the optimality of these scores in the normal case should be well known, it is continuously rediscovered [71, 72]; therefore it seems useful to provide an example.
Assume P_- is multivariate normal with mean $\boldsymbol{\mu}_-$ and variance $\boldsymbol{\Sigma}_-$ and P_+ is multivariate normal with mean $\boldsymbol{\mu}_+$ and variance $\boldsymbol{\Sigma}_+$ and both densities exist. By taking the logarithmic

transformation of the LR, it can easily be seen that for the normal case the LR based classification rule declares positive if the quadratic score

$$(\boldsymbol{X}-\boldsymbol{\mu}_-)^{\mathrm{T}}\boldsymbol{\Sigma}_-^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_-)-(\boldsymbol{X}-\boldsymbol{\mu}_+)^{\mathrm{T}}\boldsymbol{\Sigma}_+^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_+) \tag{4.3}$$

is large. Indeed, this is the Fisher's Quadratic Discriminant Analysis (QDA) rule [70], which reduces to linear – hence the corresponding Linear Discriminant Analysis (LDA) – in the case $\boldsymbol{\Sigma}_- = \boldsymbol{\Sigma}_+$ (homoscedasticity). The original work by Fisher did not actually focus on the normality assumption, but QDA and LDA are well established terminology in the literature. Being based on the LR, QDA has a proper ROC curve and it is optimal.
Insisting on a linear classifier leads to suboptimal procedures in the case of heteroscedasticity. The classifier which is optimal within the class of linear ones is proposed by Su and Liu [5] and it declares positive if

$$(\boldsymbol{\mu}_+-\boldsymbol{\mu}_-)^{\mathrm{T}}(\boldsymbol{\Sigma}_-+\boldsymbol{\Sigma}_+)^{-1}X \tag{4.4}$$

is large. It gives an improper ROC curve, which always has a "hook" and which is dominated by the ROC curve of the corresponding quadratic score in expression (4.3).
As an example, consider a bivariate normal vector (X,Y) which in population P_- has a bivariate standard normal distribution, whereas in population P_+ has independent components X distributed normally with mean $\mu_x > 0$ and variance σ_x^2 and Y distributed normally with mean $\mu_y > 0$ and variance $\sigma_y^2 \neq \sigma_x^2$. According to equation (4.3), the QDA classifier declares positive if

$$\left(\frac{X-\mu_x}{\sigma_x}\right)^2+\left(\frac{Y-\mu_y}{\sigma_y}\right)^2-X^2-Y^2<c$$

where c is an arbitrary threshold. By varying c and calculating the appropriate probabilities under P_- and P_+, we can obtain the ROC curve, by simulation or, if greater precision is needed, by using non-central chi-square distributions. The ROC curve for the case $\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 2$, $\sigma_y = 4$ is plotted as a solid line in Figure 4.1.
The best linear classifier according to Expression (4.4) is instead

$$S=\frac{\mu_x}{1+\sigma_x^2}X+\frac{\mu_y}{1+\sigma_y^2}Y.$$

S has normal distributions under P_- and P_+ and by a well-known result its ROC is

$$\mathrm{ROC}(t)=\phi(A+\phi^{-1}(t)B) \tag{4.5}$$

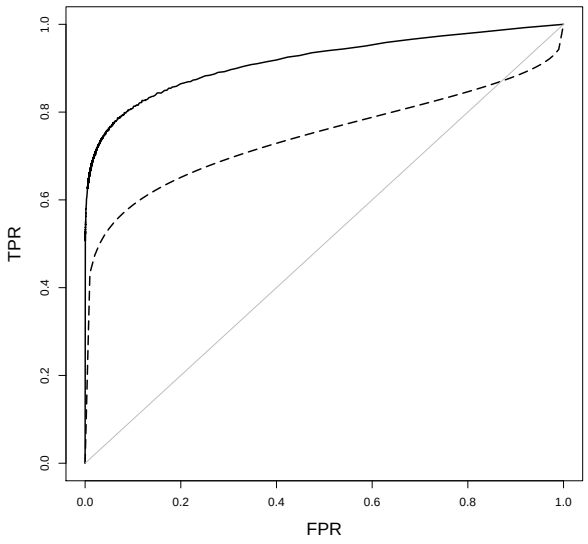


Figure 4.1 QDA (solid) and best linear ROC (dashed) curves for the bi-bivariate normal case, assuming $\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 2$, $\sigma_y = 4$.

where $\phi(\cdot)$ is the standard normal distribution function,

$$A = \frac{\mu_x^2(1+\sigma_y^2) + \mu_y^2(1+\sigma_x^2)}{\sqrt{\mu_x^2\sigma_x^2(1+\sigma_y^2)^2 + \mu_y^2\sigma_y^2(1+\sigma_x^2)^2}}$$

and

$$B = \frac{\sqrt{\mu_x^2(1+\sigma_y^2)^2 + \mu_y^2(1+\sigma_x^2)^2}}{\sqrt{\mu_x^2\sigma_x^2(1+\sigma_y^2)^2 + \mu_y^2\sigma_y^2(1+\sigma_x^2)^2}}.$$

The ROC curve of the best linear classifier for the case $\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 2$, $\sigma_y = 4$ is plotted as a dashed line in Figure 4.1. We can easily see that the QDA ROC curve is concave and dominates the best linear ROC curve.

4.2 Näive and Flexible Bayes classifier

The Näive Bayes, also known as Idiot's Bayes or Simple Bayes with reference to its simplicity, is a largely used classification methodology, which acquired popularity especially in Machine Learning literature, as testified by thousands of citations of the seminal paper published in 1995 by John and Langley [4]. Actually, it provides an easy way to represent, use and learn

probabilistic knowledge in a supervised manner (i.e. knowing the group labels), exploiting the Bayes formula and priors and posteriors probabilities; therefore it is often indicated as a Bayesian classifier, even if subjective probabilities are not formally considered.
The simplicity of this method lies in the strong assumption of local independence of the features, which means that the predictors are conditionally independent given the class; this can be graphically represented by a Bayesian network, in which all the arcs are directed from the node representing the class towards the predictors nodes (Figure 4.2). As a consequence,

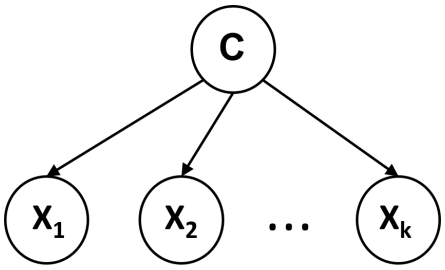


Figure 4.2 A näive Bayes classifier depicted as a Bayesian network with k predictors. Figure extracted from John and Langley [4]

the joint distribution of the features reduces to the product of marginals distributions:

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|C = c) = \mathbb{P}(\cap_i X_i = x_i|C = c) = \prod_{i=1}^{k} \mathbb{P}(X_i = x_i|C = c).$$

and the probability of belonging to class c given measurements $\boldsymbol{x}$ is obtained by

$$\mathbb{P}(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{\mathbb{P}(C = c)\prod_{i=1}^{k} \mathbb{P}(X_i = x_i|C = c)}{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})},$$

where, usually, the denominator is not directly estimated and it is considered a normalizing factor.
It is important to highlight that the local independence assumption is often violated in many real applications; however the näive Bayes classifier performs surprisingly well and better than common alternatives [73]. The robustness of the method is related to the set of parameters to be estimated, smaller than other methodologies, which also limits the estimators variance. On the other hand, biased probability estimates can be more common with the independent model, but this is of low interest in a classification perspective (where only rank order matters). In addition, often, variables selection techniques are applied to reduce the number of predictors and therefore high correlated variables are discarded in a pre-processing step.

Another assumption, not necessary but commonly made when dealing with näive Bayes, is the Gaussian distribution of the underlying populations: in this case

$$\mathbb{P}(X_i = x_i | C = c) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(x_i - \mu_C)^2}{2\sigma_C^2}\right)$$

and it is straightforward to estimate the mean and the standard deviation from sample frequencies.

Of course, apart this basic formulation of the independent model, different more sophisticated variants have been proposed to improve performance and to take into account a certain degree of dependence between predictors. A detailed review is presented by Hand and Yu [73], while here I focus on flexible näive Bayes classifier, due to its relevance in the implementation of the likelihood-ratio based algorithm for the ROC curve estimation discussed in the next paragraphs.

The flexible Bayes algorithm is a non-parametric extension of the näive Bayes, which relies on kernel density estimation of continuous predictors (instead of assuming normality) and which performs well precisely when the Gaussian hypothesis can not be retained. This means that the estimated density is given by kernels function average. Differently from näive Bayes, it requires the estimation of i parameters μ_i (one for each observation x_i) and the evaluation of the kernel function n times; therefore there is an increase in complexity and storage with respect to the independent model. However, the flexible Bayes estimate satisfies some asymptotic properties. In particular it is possible to prove (considering nominal and continuous cases separately, see [4] for details) that the flexible Bayes estimate of $\mathbb{P}(C|\boldsymbol{X})$ with Gaussian kernels is strongly pointwise consistent, i. e. $\mathbb{P}(\lim_{n\to\infty} |\hat{\mathbb{P}}_n(C = c|\boldsymbol{X} = \boldsymbol{x}) - \mathbb{P}(C = c|\boldsymbol{X} = \boldsymbol{x})| < \varepsilon) = 1$ for every ε. This property is important because it guarantees that the Bayes error rate for classification is minimal, if the number of observations is sufficiently large.

4.3 A minimal proposal for a non-parametric ROC curve

The LR is a theoretically sound tool to find optimal decision rules based on the population measures P_+ and P_-, but in practice P_+ and P_- are often not known and it is necessary to estimate them and the associated ROC curves as functions of the data.

In the multivariate case, the problem is that there is no universally accepted optimal estimates of the densities, even if, in the last half century, lots of efforts have been dedicated to find appropriate methods of estimation. Actually, two options are available to find a LR based

estimated binary classifier: either use the ratio of genuine multidimensional density estimates, or adopt some simplifying assumptions. I focused on the latter and the new proposed methodology relies on the crucial flexible Bayes assumption of local independence, i.e. the components of X are assumed to be independent, so that simple one-dimensional density estimates can be used for each of the marginals and, finally, a proper ROC curve can be estimated based on the LR of the joint densities.
In details, consider two population measures P_+ and P_- absolutely continuous with respect to one another and assume two multivariate random samples $\{x_-^{ik}; i = 1,\dots n_-, k = 1,\dots,p\}$ and $\{x_+^{ik}; i = 1,\dots n_+, k = 1,\dots,p\}$ have been observed, where x_-^{ik} (respectively x_+^{ik}) is the value of the k-th feature previously recorded on the i-th object under condition P_- (respectively P_+). A kernel estimate of the k-th marginal density $f_s^k, k = 1,\dots,p, s \in \{-,+\}$ has the well-known form

$$\hat{f}_s^k(x) = \frac{1}{n_s \lambda_s^k} \sum_{i=1}^{n_s} K_\lambda(x, x_s^{ik}) \quad -\infty < x < +\infty \tag{4.6}$$

where, in the Gaussian case (other options exist),

$$K_\lambda(x, x_s^{ik}) = \phi\left(\frac{x - x_s^{ik}}{\lambda_s^k}\right) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - x_s^{ik}}{\lambda_s^k}\right)^2\right\}.$$

λ_s^k is generally called the bandwidth, and it equals the standard deviation in the Gaussian kernel case. Gaussian kernels are widely used in density estimation and dedicated software exists; some default proposals for bandwidth selection have been used.
The LR based Gaussian kernel flexible Bayes classifier is a nonparametric classification rule which assigns a new object $\mathbf{X} = (X_1,\dots,X_p)$ to P_+, given a fixed threshold t, if

$$\hat{L} = \prod_{k=1}^{p} \frac{\hat{f}_+^k(X_k)}{\hat{f}_-^k(X_k)} > t. \tag{4.7}$$

It is important to notice that this is a LR based classification rule which takes as P_- (respectively P_+) the product measure with density $\prod_{k=1}^p \hat{f}_-^k$ (respectively $\prod_{k=1}^p \hat{f}_+^k$).

4.3.1 The simulation algorithm

The ROC curve associated to rule (4.7) can be obtained by estimating TPR and FPR under P_+ and P_- varying the threshold $t \in \mathcal{R}$ and it is proper, being the ROC of a genuine LR based decision rule. It is often too complicated to estimate the distribution of L in closed form, but a viable solution is to calculate the true and false positive rates by simulation. The particular

shape of the kernel estimate (4.6) lends itself to a simple simulation procedure which allows for reliable Monte Carlo calculations; thanks to the law of large numbers, as the number of trials increases, the sample proportions converge point-wise to the theoretical ones. In particular, we want to calculate:

$$\widehat{\mathrm{TPR}} = P_+(\hat{L} > t)$$
$$\widehat{\mathrm{FPR}} = P_-(\hat{L} > t),$$

to obtain the ROC curve as the locus of points $\{\widehat{\mathrm{FPR}}(t), \widehat{\mathrm{TPR}}(t), t \in \mathcal{R}\}$.
Given n subjects from the two populations P_- and P_+ (with $n = \sum_s n_s$ and $s \in \{-,+\}$), $\hat{L}$ is estimated using a Gaussian kernel, centred at the observations $x_s^{ik}, i = 1,\ldots,n_s$, $k = 1,\ldots,p$, $s \in \{-,+\}$. Actually, the estimate in equation (4.6) is a mixture of n_s equally probable Gaussian distributions and it is easy to simulate, selecting one of the component at random and then independently drawing elements from it. For the bandwidth, in this work, the value reported in [31] is adopted:

$$\lambda_s = 1.06 \min(\hat{\sigma}_s, \widehat{\mathrm{IQR}}_s/1.34) n_s^{-0.2},$$

where $\hat{\sigma}_s$ and $\widehat{\mathrm{IQR}}_s$ are the standard deviation and the interquartile range estimated for the two populations.Therefore the following algorithm has been developed; it is stated for the Gaussian kernel Flexible Bayes case, but generalizes easily to other options. The consistency of the Flexible Bayes estimate, reported in the paper by John and Langley [4] and briefly recalled in Section 4.2, could possibly be extended to consistency of the estimated ROC function via a continuous mapping argument.
The true LR-based ROC curve is optimal and dominates all the curves associated to different scores, as already stated. This means that the area under the ROC curve (AUC) assumes its greatest value for the LR-based classifier. However, in practice, the new proposed estimator strongly relies on the local independence of the features as working tool and its optimality can fail when predictors are really dependent. In addition, kernel density estimation methods can weaken the good properties of the curve. In this perspective the LR-based Gaussian kernel Flexible Bayes ROC curve estimate is a minimal proper alternative to the empirical common one.
With respect to the AUC, it is interesting to construct confidence intervals by asymptotic methods or in a purely data-driven way [11, 12]. It is well know that $\mathrm{AUC} = \mathbb{P}(S_+ > S_-)$, i.e. the AUC equals the probability that chosen at random a subject from P_+, its score S_+ is higher than the score S_- of an independent and random subject from P_-. In our case, to

Algorithm 1 Likelihood Based Gaussian Kernel Flexible Bayes ROC

To draw the graph of LR-based ROC curve proceed parametrically in t as follows:

- for t taking values on a finite positive grid
  - for $k = 1, \ldots, p$
    - for $b = 1, \ldots, B$, with large B
      - draw x^*_{-b} uniformly from one of the n_- Gaussian variables with mean $x^{ik}_-, i = 1, \ldots, n_-$ and standard deviation λ^k_-
      - compute $\hat{f}^k_-(x^*_{-b})$ and $\hat{f}^k_+(x^*_{-b})$
      - draw x^*_{+b} uniformly from one of the n_+ Gaussian variables with mean $x^{ik}_+, i = 1, \ldots, n_+$ and standard deviation λ^k_+
      - compute $\hat{f}^k_-(x^*_{+b})$ and $\hat{f}^k_+(x^*_{+b})$
  - compute $\widehat{\mathrm{FPR}}(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\left[\prod_{k=1}^{p} \frac{\hat{f}^k_+(x^*_{-b})}{\hat{f}^k_-(x^*_{-b})} > t\right]}$
  - compute $\widehat{\mathrm{TPR}}(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\left[\prod_{k=1}^{p} \frac{\hat{f}^k_+(x^*_{+b})}{\hat{f}^k_-(x^*_{+b})} > t\right]}$
  - plot $(\widehat{\mathrm{FPR}}(t), \widehat{\mathrm{TPR}}(t))$

where $\mathbb{1}_{[A]}$ is the indicator function of event A, which equals 1 if A is true and 0 otherwise.

calculate AUC and its confidence limits we have to deal with likelihood-ratio based scores and therefore a resampling approach such as the bootstrap seems viable.

4.3.2 Comparison with empirical ROCs

In the previous paragraph an estimated proper ROC curve based on the likelihood ratio has been obtained. It is of interest to compare it with the common empirical ROC curve estimates via simulations. Actually, a stress test for the new proposed method is performed, considering unfavourable situations with increasing degree of correlation between the features.

Given any decision rule, like the one in formula (4.2) for example, it is possible to apply it to the same training data and to obtain a classification which can be compared with the observed true labels. This gives empirical $\widehat{\mathrm{TPR}}$ and $\widehat{\mathrm{FPR}}$ for any given t; by varying t the empirical ROC curve, a step function, is built.

In addition simulations appear also useful to check if the new estimator over-estimates the true performance of the classifier (an overfitting problem, due to the double use of the same data for both fitting a score and calculating the associated ROC curve, can not be completely

avoided). In particular, the true ROC curve can be compared with the ones obtained using the flexible Bayes algorithm and the ones empirically built on the logistic score (i.e. the risk score estimated via logistic regression); the differences are evaluated in term of the mean integrated squared error (MISE), calculated for both the smooth and the empirical curve. In the simulations two underlying bivariate Gaussian distributions are assumed. Both the homoscedastic and the heteroscedastic cases are considered, as well as possible correlation among features. Consider a bivariate vector of predictors $\boldsymbol{x} = (x_1, x_2)$ and two bi-normal populations, i.e. $F_+ \sim \mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma}^+)$ and $F_- \sim \mathcal{N}(\boldsymbol{\mu}^-, \boldsymbol{\Sigma}^-)$.
For the homoscedastic case ($\boldsymbol{\Sigma}^+ = \boldsymbol{\Sigma}^- = \boldsymbol{\Sigma}$), assuming $\text{cov}(x_i, x_j) = 0$ for $i \neq j$, reasoning as in the univariate binormal model (see for example [12]), it is easy to prove that the true ROC curve is the locus of points:

$$\begin{aligned}
\text{FPR}(t) &= \mathbb{P}(\text{LR}^- > t) = \mathbb{P}(\log(\text{LR}^-) > \log(t)) = \\
&= 1 - \Phi\left(\frac{2\sigma_1^2\sigma_2^2\log(t) + A\mu_1^- + B\mu_2^- + C}{\sqrt{(A^2\sigma_1^2 + B^2\sigma_2^2)}}\right) \\
\text{TPR}(t) &= \mathbb{P}(\text{LR}^+ > t) = \mathbb{P}(\log(\text{LR}^+) > \log(t)) = \\
&= 1 - \Phi\left(\frac{2\sigma_1^2\sigma_2^2\log(t) + A\mu_1^+ + B\mu_2^+ + C}{\sqrt{(A^2\sigma_1^2 + B^2\sigma_2^2)}}\right)
\end{aligned}$$

where $\text{LR}^- = \text{LR}(\boldsymbol{x}|-)$ and $\text{LR}^+ = \text{LR}(\boldsymbol{x}|+)$,

$$\begin{aligned}
\text{LR} &= \frac{f_+}{f_-} = \exp\left(-\frac{1}{2}\left(\frac{(x_1-\mu_1^+)^2}{\sigma_1^2} + \frac{(x_2-\mu_2^+)^2}{\sigma_2^2} - \frac{(x_1-\mu_1^-)^2}{\sigma_1^2} - \frac{(x_2-\mu_2^-)^2}{\sigma_2^2}\right)\right) \\
A &= 2\sigma_2^2(\mu_1^- - \mu_1^+), \quad B = 2\sigma_1^2(\mu_2^- - \mu_2^+), \\
C &= \sigma_2^2((\mu_1^+)^2 - (\mu_1^-)^2) + \sigma_1^2((\mu_2^+)^2 - (\mu_2^+)^2).
\end{aligned}$$

As reported in Table 4.1, for 100 simulations, the proposed estimator for the likelihood ratio based ROC curve does not overfit the data and it is slightly better than the empirical one with respect to the mean integrated squared error for uncorrelated and low correlated predictors. This is no longer true when an interaction term is included in the logistic score: in this case, in particular when variables are strongly correlated ($\rho = 0.9$) (and therefore violating the local independence assumption) the MISE of flexible Bayes is about three times higher than MISE for logistic score (Table 4.2); this happens also, in a more relevant way, when the logistic score includes quadratic terms and interaction (Table 4.3).
In the heteroscedastic case, instead, it is not possible to find an analytical closed formulation

for the true ROC curve, neither under the local independence assumption; the LR is a quadratic form (Fisher Quadratic Discriminant analysis) and the evaluation of true and false positive proportions entails linear combinations of non central Chi-squared distributions. Therefore, it is necessary to resort to simulations also to obtain the theoretical ROC.
In this case, as reported in Tables 4.1, for 100 simulations the MISE for flexible Bayes ROC curve is always a bit lower than the one for the ROC associated to logistic score. When an interaction term is included in the logistic score the performances of the two methodologies are totally comparable (Table 4.2); in addition, when also quadratic terms are included in the logistic approach (i.e. when the ideal score is considered) the performances of the Flexible Bayes estimate still appear slightly better (Table 4.3). Nonetheless, ways to decorrelate the variables could be considered as a pre-processing step to improve the performance of the likelihood-based flexible Bayes classifier, but some preliminary trials have not shown a clear winning strategy.

Table 4.1 For 100 simulations, comparison between ROC associated to Flexible Bayes and ROC associated to a logistic score.

$\boldsymbol{\mu}^+$	$\boldsymbol{\mu}^-$	$\boldsymbol{\Sigma}^+$	$\boldsymbol{\Sigma}^-$	ρ	$\mathbf{MISE}_{\text{Flex Bayes}}$	$\mathbf{MISE}_{\text{Logistic}}$
$\begin{pmatrix}1\\1\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\rho=0$	0.00121	0.00217
				$\rho=0.4$	0.00214	0.00224
				$\rho=0.9$	0.00660	0.00239
$\begin{pmatrix}2\\1.5\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1.5 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & 1.8\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\rho=0$	0.00083	0.00100
				$\rho=0.4$	0.00113	0.00134
				$\rho=0.9$	0.00104	0.00141

Table 4.2 For 100 simulations, comparison between ROC associated to Flexible Bayes and ROC associated to a logistic score (with an interaction term).

$\boldsymbol{\mu}^+$	$\boldsymbol{\mu}^-$	$\boldsymbol{\Sigma}^+$	$\boldsymbol{\Sigma}^-$	ρ	$\mathbf{MISE}_{\text{Flex Bayes}}$	$\mathbf{MISE}_{\text{Logistic}}$
$\begin{pmatrix}1\\1\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\rho=0$	0.00159	0.00210
				$\rho=0.4$	0.00251	0.00238
				$\rho=0.9$	0.00737	0.00248
$\begin{pmatrix}2\\1.5\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1.5 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & 1.8\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\rho=0$	0.00086	0.00086
				$\rho=0.4$	0.00123	0.00123
				$\rho=0.9$	0.00116	0.00120

Table 4.3 For 100 simulations, comparison between ROC associated to Flexible Bayes and ROC associated to a logistic score (with quadratic terms and interaction).

$\boldsymbol{\mu}^+$	$\boldsymbol{\mu}^-$	$\boldsymbol{\Sigma}^+$	$\boldsymbol{\Sigma}^-$	ρ	$\mathbf{MISE}_{\text{Flex Bayes}}$	$\mathbf{MISE}_{\text{Logistic}}$
$\begin{pmatrix}1\\1\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\rho = 0$	0.00124	0.00242
				$\rho = 0.4$	0.00328	0.00260
				$\rho = 0.9$	0.00820	0.00289
$\begin{pmatrix}2\\1.5\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1.5 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & 1.8\end{pmatrix}$	$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$	$\rho = 0$	0.00067	0.00069
				$\rho = 0.4$	0.00095	0.00098
				$\rho = 0.9$	0.00098	0.00108

4.3.3 Some considerations

The new ROC curve estimate presented in this section is a non-parametric minimal proposal for a default alternative to the broadly used empirical estimator, allowing to obtain a proper curve without making any assumptions on the underlying populations distributions. Its development relies on three key concepts: the adoption of the likelihood-ratio as the decision variable, the kernel density estimation as the common non-parametric approach in multivariate setting and the local independence of the features as a strong working assumption. Certainly, these ideas are not new by themselves, but, to the best of my knowledge, their joint use has not yet been explored, especially in the medical diagnostic field. In particular, often, the focus is on continuous variable and monotone likelihood ratio: the proposed approach allows to overcome this, dealing with cases never appreciated by the relevant literature.
From a theoretical point of view, it is largely recognized that the likelihood-ratio criterion (namely a classification rule directly based on the likelihood-ratio of two underlying populations) guarantees some optimality properties inherited by the associated true population ROC curve. However, in practice population distributions are not known: they must be estimated from the data, making some simplifying assumptions. In this perspective, the local independence of the features appears a useful working tool and its popularity in Machine Learning literature motivated its adoption for a first approximation also here. Clearly, when features are really dependent (as shown by simulations) the ROC curve estimator is no longer optimal, but it still preserves good properties (the smoothness, the concavity, ...).
All in all, the likelihood-ratio should be considered the rational guideline to combine features in a multivariate setting: it has a higher discriminant power than simply linear combinations. On the other hand, in the univariate framework, there are situations in which a single variable could have good predictive power based on medical knowledge, with a really heteroscedastic behaviour in the two populations; its likelihood-ratio assumes low values for both low and high values of the only predictor and therefore the resulting LR-based classification is in

contrast with the clinician expectation. Nevertheless, also in this case, the likelihood ratio should be regarded as the guiding principle, suggesting that the single variable alone, even if good, is not enough to properly classify subjects.

4.4 Case study: diagnosis of PCa using biomarkers

PCa (prostate cancer) is the most frequent neoplasia diagnosis in men in Europe and one of the most common causes of cancer related death. As said before, a lot of efforts are currently devoted worldwide to finding non-invasive and easy-to-detect diagnostic biomarkers.The diagnosis of prostate cancer disease can be considered a classification problem, and a binary one if it is simplified to PCa versus non-PCa. ROC curves are used to evaluate the performance of the classifiers.

The likelihood-ratio based flexible Bayes classifier and the associated ROC curve estimate, presented in the previous paragraphs, are developed for the validation dataset of the study on circulating miRNAs described in Chapter 3 and in the article by Mello-Grand et al. [41]. In particular, the dataset consists of 58 PCa (the P_+ sample) and 170 non-PCa patients (the P_- sample), including 89 benign hyperplasias, 8 precancerous lesions and 73 healthy controls – but this finer subdivision is not used here. For each patient, two microRNAs and (log-transformed) PSA were combined to build the classifier. The two microRNAs were selected after a cumbersome feature selection procedure which combined statistical and practical aspects, as detailed in Section 3.3.1. Let $\boldsymbol{X} = (X_1, X_2, X_3)$ be the observation vector (microRNA1, microRNA2, log(PSA)). The idea was to combine these biomarkers in such a way to obtain the best classifier, i.e. the one associated to the dominant ROC curve.

The maximum likelihood estimates of means and variance covariance matrices under P_- and P_+ are

$$\hat{\boldsymbol{\mu}}_- = \begin{pmatrix} 4.952 \\ 5.463 \\ 1.403 \end{pmatrix} \quad \hat{\boldsymbol{\Sigma}}_- = \begin{pmatrix} 7.233 & 5.260 & 1.639 \\ 5.259 & 4.927 & 1.165 \\ 1.638 & 1.165 & 2.490 \end{pmatrix}$$

$$\hat{\boldsymbol{\mu}}_+ = \begin{pmatrix} 6.833 \\ 6.939 \\ 2.518 \end{pmatrix} \quad \hat{\boldsymbol{\Sigma}}_+ = \begin{pmatrix} 3.570 & 3.167 & -0.098 \\ 3.167 & 3.086 & -0.150 \\ -0.098 & -0.149 & 0.656 \end{pmatrix}.$$

Assuming $\boldsymbol{X}$ multivariate normal, then $\boldsymbol{X} \sim \text{MVN}(\hat{\boldsymbol{\mu}}_-, \hat{\boldsymbol{\Sigma}}_-)$ under P_- and $\boldsymbol{X} \sim \text{MVN}(\hat{\boldsymbol{\mu}}_+, \hat{\boldsymbol{\Sigma}}_+)$ under P_+. The best parametric classifier is Fisher QDA in equation (4.3), since the covariance matrices differ. The associated ROC curve is displayed in Figure 4.3, left panel, dashed line.

Insisting on linear transformations of $\boldsymbol{X}$, then the best one proposed by Su and Liu [5], given in equation (4.4), is

$$0.099 \times X_1 + 0.043 \times X_2 + 0.292 \times X_3$$

with Gaussian univariate distributions $\mathcal{N}(1.136, 0.461)$ under P_- and $\mathcal{N}(1.711, 0.114)$ under P_+. The associated ROC curve is displayed in Figure 4.3, left panel, solid line.
On the other hand, adopting a less restrictive nonparametric point of view, we could apply Algorithm 1 to get a good approximation of a non-parametric LR based estimated ROC curve for the Flexible Bayes classifier, displayed in Figure 4.3, right panel, dashed line. The solid line is instead the usual stepwise empirical ROC curve (obtained in this case with the R library described by [74]) associated to the following logistic score:

$$0.399 \times X_1 - 0.142 \times X_2 + 0.560 \times X_3.$$

In addition, it seems also noteworthy to compare this score with the best linear combination proposed by Chen [6]:

$$4.349 \times X_1 - 0.640 \times X_2 + 4.723 \times X_3.$$

The associated ROC curve is the dotted line in the right panel of Figure 4.3. The LR-based ROC curve dominates the others and this is confirmed by the AUC values (with nearly non-overlapping confidence intervals), reported in Table 4.4.

Table 4.4 Area under the ROC curve (AUC) and it 95% confidence intervals for different estimators.

Estimation method	**AUC**	**95% Confidence Intervals**
Logistic score	0.760	0.697 - 0.824
Chen best linear combination	0.765	0.702 - 0.828
LR-based Flexible Bayes	0.884	0.813 - 0.924

The ROC of the Flexible Bayes classifier is different from the QDA ROC, since there is a price to pay for the greater generality of the non-parametric approach, but it exhibits a definite advantage over the empirical ROC, which appears to be a less efficient estimate of the underlying true ROC.

The R code [75] to obtain the ROC curve associated to the LR-based Flexible Bayes classifier, following Algorithm 1, and the code to compare the proper and minimal curve with the ROC curves associated to logistic score and to Chen's best linear combination score [6], their AUC

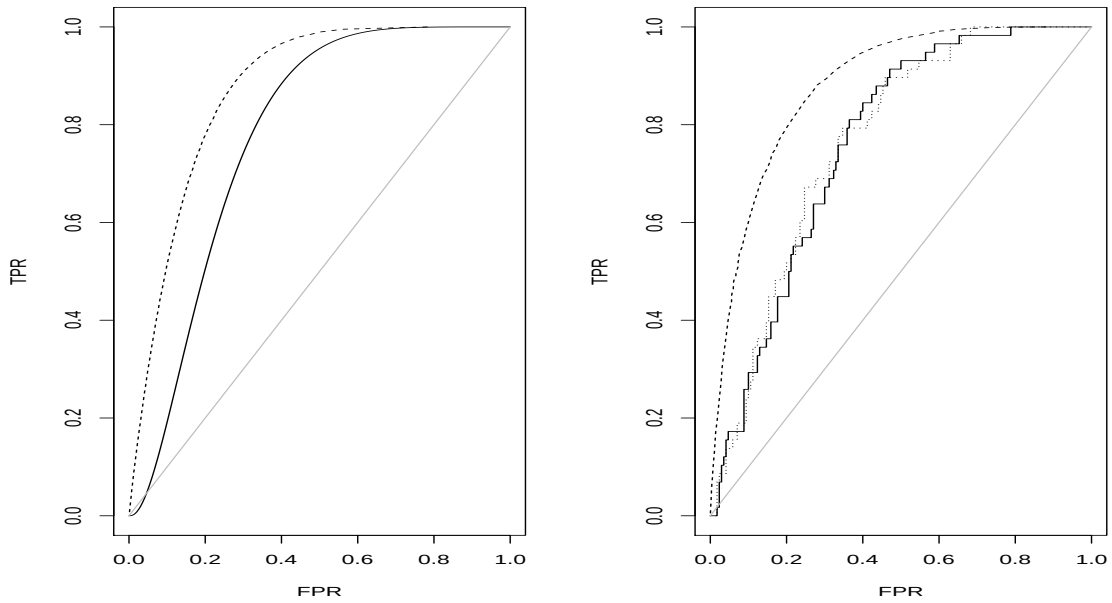


Figure 4.3 Left panel: parametric ROC curves comparison: QDA (dashed line) versus best linear combination as provided in [5](solid line). Right panel: non-parametric ROC curves comparison: empirical ROC associated to the logistic score (solid line), the ROC associated to Chen best linear combination [6](dotted line) and the LR-based Flexible Bayes one (dashed line).

values and confidence intervals in the case study are provided in Appendix B.

Chapter 5

On the meaning and measure of concentration

It is well known that there is a connection between the definition of the LR based classification rule for general data spaces and the general definition of concentration function for two probability measures (as given by Cifarelli and Regazzini in late 1980s [76, 77]) and, in this chapter, it is deeply investigated from an exquisitely theoretical point of view. After having introduced a definition of LR-based classification rule (which entails the use of randomization in case the distribution of the LR contains atoms), an alternative formulation of the ROC curve is presented and the connection with the concentration function is highlighted. As a consequence, the ROC curve parameter appears to be a theoretical quantification of the relationship between two probability measures, and not merely a descriptive tool of the performance of a classifier. To be clearer a couple of examples of special cases are reported; in addition, a critical appraisal of a recent published paper [40] is given and, lastly, some considerations on discrete ROC curves are provided.
Part of the material presented here is extracted from the paper I co-authored "On the definition of a concentration function relevant to the ROC curve" [78].

5.1 LR based ROC curve for general types of data

Assume that P_+ and P_- are absolutely continuous with respect to one another and have densities f_+ and f_-, respectively, with respect to a common dominating measure. Then,

without loss of generality, f_- can be taken to be positive, so that the Likelihood Ratio

$$L = \frac{f_+}{f_-} \tag{5.1}$$

is a well defined non negative random variable. As such, L then has distribution functions under P_- and P_+, which are denoted by H_- and H_+ respectively. More precisely, for each $\ell \in \mathcal{R}$:

$$H_-(\ell) = P_-(L \leq \ell)$$

and

$$H_+(\ell) = P_+(L \leq \ell).$$

Next, define the quantile function associated with H_- in the usual way as follows :

$$q_t = \inf\{y \in \mathcal{R} : H_-(y) \geq t\} \quad 0 < t < 1 \tag{5.2}$$

and recall that, for any real number ℓ,

$$q_t \leq \ell \quad \text{if and only if} \quad H_-(\ell) \geq t.$$

For any given value $t \in (0,1)$, it may or may not happen that $t = H_-(q_t)$, depending on whether t does not correspond or does correspond to a jump of H_-. More specifically, if $t \neq H_-(q_t)$, then $H_-(q_t^-) \leq t < H_-(q_t)$, where the notation $^-$ indicates left limits (nothing to do with P_-), a particularly relevant occurrence for the discussion below.
H_- and H_+ may have jumps, even if P_- and P_+ are absolutely continuous laws on the real line. H_- and H_+ do not have jumps for two Gaussian probability measures, but P_- and P_+ may be absolutely continuous yet L be a finite random variable which takes on a finite set of values, almost surely. This happens, for example, if P_+ and P_- have piecewise constant densities as provided by an example in Section 5.3.
Actually, in this chapter, the following definition of LR is used:

Definition 5.1.1. *Given two alternative probability laws P_- or P_+ mutually absolutely continuous with respective densities f_- and f_+, define the likelihood ratio $L = f_+/f_-$, its respective distribution functions H_- and H_+ and the following classification rule. For each $0 < t < 1$:*

- *if $L > q_t$, declare positive;*
- *if $L < q_t$, declare negative;*

- *if $L = q_t$, then perform an auxiliary independent randomization and declare positive with probability*

$$r(t) = \frac{H_-(q_t) - t}{H_-(q_t) - H_-(q_t^-)}$$

and negative otherwise.

It parallels the definition of a randomized LR test [18], but it is presented here in a classification context. Therefore, the ROC curve can be defined in an alternative way:

Theorem 5.1.1. *The ROC function of the classification rule of Definition 5.1.1 is*

$$\mathrm{ROC}(x) = 1 - H_+(q_{1-x}) + q_{1-x}(H_-(q_{1-x}) - (1-x)), \quad 0 < x < 1. \tag{5.3}$$

As usual, we can complete the result by setting $ROC(0) = 0$ and $ROC(1) = 1$.

Proof. First of all, the FPR and the TPR are computed separately.

$$\begin{aligned}
\mathrm{FPR} &= P_-(\text{declare positive}) = P_-(L > q_t) + P_-(L = q_t) r(t) \\
&= 1 - H_-(q_t) + (H_-(q_t) - H_-(q_t^-)) r(t) \\
&= 1 - H_-(q_t) + H_-(q_t) - t = 1 - t.
\end{aligned}$$

Notice that if $t = H_-(q_t)$ then $H_-(q_t^-) - H_-(q_t) = 0$; in other words the expression simplifies for points which are not H_--atoms.

$$\begin{aligned}
\mathrm{TPR} &= P_+(\text{declare positive}) = P_+(L > q_t) + P_+(L = q_t) r(t) \\
&= 1 - H_+(q_t) + (H_+(q_t) - H_+(q_t^-)) \frac{H_-(q_t) - t}{H_-(q_t) - H_-(q_t^-)} \\
&= 1 - H_+(q_t) + q_t(H_-(q_t) - t)
\end{aligned}$$

since, P_+ and P_- being mutually absolutely continuous, they will both have or not have an atom in q_t and their LR in q_t will be exactly $(H_+(q_t) - H_+(q_t^-))/(H_-(q_t) - H_-(q_t^-))$, i.e. q_t itself. Next, set $\mathrm{FPR} = x$, i.e. $t = 1 - x$, to eliminate the parameter t and obtain the explicit form of the ROC curve:

$$\mathrm{TPR} = 1 - H_+(q_{1-x}) + q_{1-x}(H_-(q_{1-x}) - (1-x)).$$

□

5.2 Relationship with a general concentration function

Expression (5.3) does not come out of nowhere. It corresponds to a definition of concentration function given by Cifarelli [76], and further expanded by Regazzini [77], with the aim of extending the classical definition of concentration given by Gini. Such a definition is naturally based on the LR, and given the strict relationship existing between ROC curves and LRs, the connection comes easily [76] and, for convenience, it is recalled here for the case in which P_+ and P_- are mutually absolutely continuous:

Definition 5.2.1. *Let P_+ and P_- be mutually absolutely continuous probability measures, let f_+ and f_- be their respective derivatives with respect to a common dominating measure μ, let their LR be defined as the real-valued random variable $L = f_+/f_-$, let H_- be its distribution function under P_- and let q_x be its quantile function. Then Cifarelli [76] defines the concentration function of P_+ with respect to P_- as $\varphi(0) = 0$, $\varphi(1) = 1$ and*

$$\varphi(x) = P_+(L < q_x) + q_x(x - H_-(q_x^-)).$$

The connection between this definition and the classification rule of the previous section is established in the next Theorem.

Theorem 5.2.1. *Under the hypotheses described in Definition 5.1.1,*

$$\mathrm{ROC}(x) = 1 - \varphi(1-x) \quad \forall 0 \le x \le 1.$$

where $\varphi(\cdot)$ is the concentration function of P_+ with respect to P_-.

Proof. The equivalent relationship

$$1 - \mathrm{ROC}(1-x) = \varphi(x) \quad \forall 0 \le x \le 1.$$

can be verified directly for $x = 0, 1$ and as follows for $0 < x < 1$:

$$\begin{aligned}
1 - \mathrm{ROC}(1-x) &= H_+(q_x) - q_x(H_-(q_x) - x) \\
&= H_+(q_x) \pm H_+(q_x^-) + q_x(x - H_-(q_x) \pm H_-(q_x^-)) \\
&= H_+(q_x^-) + q_x(x - H_-(q_x^-)) + \\
&\qquad (H_+(q_x) - H_+(q_x^-)) - q_x(H_-(q_x) - H_-(q_x^-)) \\
&= H_+(q_x^-) + q_x(x - H_-(q_x^-)) +
\end{aligned}$$

$$
\begin{aligned}
&(H_-(q_x) - H_-(q_x^-)) \left(\frac{H_+(q_x) - H_+(q_x^-)}{H_-(q_x) - H_-(q_x^-)} - q_x \right) \\
&= P_+(L < q_x) + q_x(x - H_-(q_x^-)) \\
&= \varphi(x).
\end{aligned}
$$

□

Corollary 5.2.2. *Under the hypotheses described in Definition 5.1.1,* $\mathrm{ROC}(\cdot)$ *is a non-decreasing, continuous and concave function on* $[0,1]$. *In particular,* $\mathrm{ROC}(\cdot)$ *is proper.*

Proof. This is a consequence of Theorem 2.3 in Cifarelli [76]. In particular, $\varphi(x)$ is always convex over its domain, i.e. $\forall x_1, x_2$ and $\nu \in [0,1], \varphi(\nu x_1 + (1-\nu)x_2) \leq \nu\varphi(x_1) + (1-\nu)\varphi(x_2)$. By Theorem 5.2.1:

$$1 - \mathrm{ROC}(1 - (\nu x_1 + (1-\nu)x_2)) \leq \nu(1 - \mathrm{ROC}(1-x_1)) + (1-\nu)(1 - \mathrm{ROC}(1-\mathrm{x}_2)).$$

The left hand side of the previous equality becomes:

$$
\begin{aligned}
1 - \mathrm{ROC}(1 - (\nu x_1 + (1-\nu)x_2)) &= 1 - \mathrm{ROC}(\nu + (1-\nu) - \nu x_1 - (1-\nu)x_2) \\
&= 1 - \mathrm{ROC}(\nu(1-x_1) + (1-\nu)(1-x_2)),
\end{aligned}
$$

while the right hand side can be rewritten as:

$$
\begin{aligned}
&\nu(1 - \mathrm{ROC}(1-x_1)) + (1-\nu)(1 - \mathrm{ROC}(1-x_2)) = \\
&\nu - \nu\mathrm{ROC}(1-x_1) + 1 - \nu - (1-\nu)\mathrm{ROC}(1-x_2) = \\
&1 - \nu\mathrm{ROC}(1-x_1) - (1-\nu)\mathrm{ROC}(1-x_2).
\end{aligned}
$$

Therefore:

$$\mathrm{ROC}(\nu t_1 + (1-\nu)t_2) \geq \nu\mathrm{ROC}(t_1) + (1-\nu)\mathrm{ROC}(t_2), \quad \forall t_1, t_2, \nu \in [0,1]$$

where $t_1 = 1 - x_1, t_2 = 1 - x_2$. □

As already stated, it is important to stress that a proper ROC curve is possible under the very general assumption that the LR is meaningful. Instead, in the applied literature, the existence of a proper ROC curve is often believed to be limited to models with a monotone likelihood ratio on a certain score.

5.3 A theoretical example: two absolutely continuous measures with discrete LR

Let P_- be an absolutely continuous probability measure on the real line with density f_- uniform between 0 and 3 and let P_+ have a piecewise constant density f_+ defined as follows:

$$f_+(s) = \frac{1}{18}\mathbb{1}_{[0<s\leq 1]} + \frac{10}{18}\mathbb{1}_{[1<s\leq 2]} + \frac{7}{18}\mathbb{1}_{[2<s\leq 3]} = \begin{cases} \frac{1}{18} & \text{if } 0 < s \leq 1 \\ \frac{10}{18} & \text{if } 1 < s \leq 2 \\ \frac{7}{18} & \text{if } 2 < s \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbb{1}_{[A]}$ is the indicator function for the event A, i.e. the function which equals 1 if A is true and 0 otherwise. Suppose S is a real random variable with density f_- under P_- and f_+ under P_+. It is easy to see that the LR $L = f_+/f_-$ is piecewise constant and not monotone in S, being:

$$L = \begin{cases} \frac{1}{6} & \text{if } 0 < s \leq 1 \\ \frac{10}{6} & \text{if } 1 < s \leq 2 \\ \frac{7}{6} & \text{if } 2 < s \leq 3. \end{cases}$$

A classification rule based only on S gives rise to a ROC curve

$$\mathrm{ROC}_S(x) = \begin{cases} \frac{21}{18}x & \text{if } 0 \leq x < \frac{1}{3} \\ -\frac{3}{18} + \frac{30}{18}x & \text{if } \frac{1}{3} \leq x < \frac{2}{3} \\ \frac{15}{18} + \frac{3}{18}x & \text{if } \frac{2}{3} \leq x < 1 \end{cases}$$

which is not concave, shown as dashed line in Figure 5.1. Using instead the LR based classification rule, the ROC curve is:

$$\mathrm{ROC}_L(x) = \begin{cases} \frac{30}{18}x & \text{if } 0 \leq x < \frac{1}{3} \\ \frac{3}{18} + \frac{21}{18}x & \text{if } \frac{1}{3} \leq x < \frac{2}{3} \\ \frac{15}{18} + \frac{3}{18}x & \text{if } \frac{2}{3} \leq x < 1 \end{cases}$$

which is concave and dominates the previous one as shown in Figure 5.1. This example deals with absolutely continuous densities which, nonetheless, have a finite discrete likelihood ratio: it is an exquisitely theoretical exercise, with few (if any) concrete applications. However, it

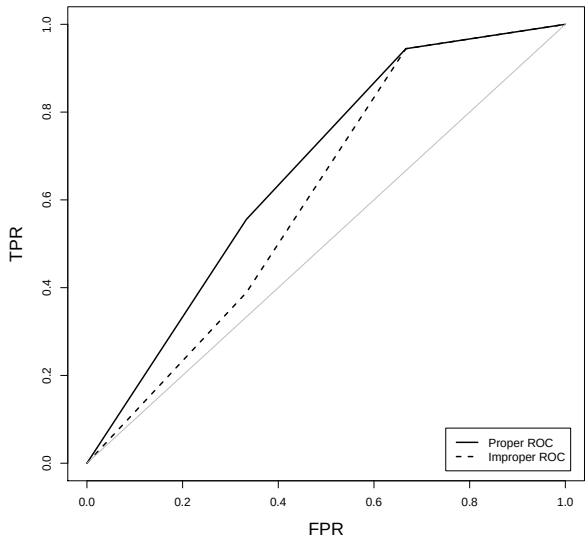


Figure 5.1 Proper ROC curve based on the LR of S (solid line); improper ROC curve based on S (dashed line).

could be of help for a better understanding of proper/improper curves; in addition it allows to address a case particularly difficult for the usual approach to ROC curves (which emphasizes a continuous score is necessary).

5.4 The correct definition of concentration function for diagnostics

In a paper appeared recently on Metron journal Schechtman and Schechtman [40] try to shed some light on the relationship between the Gini Mean Difference (Gini), the Gini Covariance (co-Gini), the Lorenz curve, the Receiver Operating Characteristic (ROC) curve and a particular definition of concentration function. The purpose of the paper is commendable, since there is a lot of confusion regarding the various relationships among these concepts. In particular, it is clearly stated that the ROC curve and its functions (such as the AUC), as well as an appropriate definition of relative concentration of a probability distribution with respect to another, are bivariate objects tying together two different distributions, and can not be reduced to univariate indices such as the Gini. Schechtman and Schechtman [40] build on the wealth of research reviewed in the monograph by Yitzhaki and Schechtman [79], where a whole technology based on the Gini and the co-Gini is proposed as an alternative to

traditional variance and covariance based methods to study variability, correlation, regression and the like.
Of course, studying how jointly distributed random variables interrelate is a very fundamental problem in Statistics and its applications to Economics and the Sciences. However, when turning to the diagnostic (or classification) setup, where ROC curves are naturally used, one observes diagnostic variables (called features in the Machine Learning literature) from two populations and tries to set up a rule that discriminates between them. Some special requirements can then be identified:

1. Two probability distributions should be evaluated as alternative, mutually exclusive explanations of the data, rather than from a joint point of view; for example, a diagnostic marker observed on a patient has either the sick patient distribution or the healthy patient distribution, and in no way the same marker can be observed jointly under both sick and healthy conditions.

2. The definition of the ROC curve and the associated concentration function should be viable also in the multivariate setup; for example, more than one diagnostic marker can be observed on the same patient.

3. The definition of the ROC curve and the associated concentration function must be given both at the population and at the sample level, as widely discussed in the ROC literature (see for example [12]); a clear definition of the ROC curve at the population level is necessary to understand basic ideas and to give appropriate definitions.

It appears that the definition of (absolute and) relative concentration curve contained in [40] is not appropriate for the diagnostic setup since:

a. conditional distributions are used in the Definition 1 of [40], thus contradicting requirement 1);

b. percentiles are used in the same definition, thus contradicting requirement 2);

c. in [40], the discussion on the ROC curve is mantained at the sample level only, making it hard to understand what is, for example, the definition of population ROC curve.

On the other hand, the correct definition of concentration function for the diagnostic setup was the one given by Cifarelli and Regazzini in [76] and previously recalled. In my opinion, that definition is more suitable to the diagnostic setup since it is a one-to-one transformation of the ROC curve of the optimal diagnostic test, i.e. the one based on the likelihood ratio.

Indeed, the likelihood ratio is the fundamental measure of comparison of two distributions and plays a role similar to the role the conditional expectation plays in [40].

5.5 The Lorenz curve and the AUC of the optimal test

As hinted in Chapter 2, ROC curves and the Lorenz-Gini curve are strictly related: an interesting special case arises when X is a positive random variable with finite mean $\mu_x = \int_0^\infty x f_X(x)dx$ and Y is the length-biased version of X, i.e.

$$f_Y(y) = \frac{y f_X(y)}{\mu_X}, \quad y > 0.$$

In economic applications, Y represents wealth; in general, it may be a transferable character, i.e. some characteristic which can in theory be transported from one unit of the population to another. The likelihood ratios in this case simplify to

$$L_X = \frac{f_Y(X)}{f_X(X)} = \frac{X f_X(X)}{\mu_X f_X(X)} = \frac{X}{\mu_X}$$

and

$$L_Y = \frac{f_Y(Y)}{f_X(Y)} = \frac{Y f_X(Y)}{\mu_X f_X(Y)} = \frac{Y}{\mu_X}$$

so that $H_X(\ell) = F_X(\mu_x \ell)$ and $H_Y(\ell) = F_Y(\mu_x \ell)$ and finally

$$\varphi_{Lorenz}(p) = H_Y(H_X^{-1}(p)) = F_Y(F_x^{-1}(p)) = \frac{\int_0^{F_x^{-1}(p)} y f_X(y)dy}{\int_0^\infty x f_X(x)dx},$$

in which the usual forms of the Lorenz curve can be recognized. Indeed, it has been proven:

Corollary 5.5.1. *In the Lorenz-Gini scenario, i.e. when $f_Y(y) = \frac{y f_X(y)}{\mu_X}$, the concentration curve is the usual Lorenz curve.*

An important consequence of Theorem 5.2.1 is about the AUC of the optimal likelihood ratio based test, which can be easily computed as follows.

$$\text{AUC}_{opt} = \int_0^1 ROC_{opt}(q)dq = \int_0^1 (1 - \varphi(1-q))dq = 1 - \int_0^1 \varphi(s)ds. \tag{5.4}$$

Now, in this Lorenz-Gini scenario, the Gini concentration coefficient (Gini) is defined as twice the area between the diagonal and the Lorenz curve:

$$\text{Gini} = 2\int_0^1 (p - \varphi_{Lorenz}(p))dp = 1 - 2\int_0^1 \varphi_{Lorenz}(p)dp$$

Since the concentration curve is a generalization of the Lorenz curve which describes the concentration of one variable with respect to another (and not necessarily its length-biased version), the generalized Gini can be defined as

$$\text{Gini}_{gen} = 2\int_0^1 (p - \varphi(p))dp,$$

similarly to the co-Gini in [40]. Substituting into expression (5.4) the following corollary is obtained.

Corollary 5.5.2. *The AUC of the optimal likelihood ratio based diagnostic test equals*

$$AUC_{opt} = \frac{1}{2}(1 + \text{Gini}_{gen}).$$

The same result can be found in [3] and mentioned by several other authors. It is noteworthy to highlight that the result is true for the likelihood ratio based test and, of course, for models with monotone likelihood ratios (like the example considered in [3]) but not in general for the AUC of any ROC, as also noted by [40]. A few more results provided agree with the results in [3], but they have been presented in a more general form at the population level for continuous variables, for which some examples are presented in the next section.
The definition of concentration function given here is a convenient one for the diagnostic problem, since it compares two alternative probability distributions using a natural bivariate generalization of the Lorenz curve. The discussion on the concentration and the ROC curves at the population level allows for a deeper understanding of the concepts and for the proof of Theorem 5.2.1, which ties together the concentration function and the ROC curve of the optimal likelihood ratio based diagnostic test. Similar results were given by [3] at the sample level. All results mentioned in these sections can conceptually be generalized to higher dimensions, although computations may become very hard. In particular, the likelihood ratio is an efficient dimension reduction technique which reduces the comparison to a one-dimensional problem and allows for Definition 5.2.1 of concentration function without involving higher dimensional conditional expectations or quantiles.

5.6 Some examples

Example 1. Let X be exponential with rate parameter λ_X and Y be exponential with rate parameter λ_Y and assume, as it is customary, that $\lambda_X > \lambda_Y$, so that Y is stochastically greater than X (this corresponds to a situation where the greater a diagnostic marker, the more is indicative of disease). Then it is easy to verify that

$$H_X(\ell) = \mathrm{P}\left(\frac{f_Y(X)}{f_X(X)} \le \ell\right) = \mathrm{P}\left(\frac{\lambda_Y e^{-\lambda_Y X}}{\lambda_X e^{-\lambda_X X}} \le \ell\right) = 1 - (\frac{1}{r\ell})^{r/(r-1)}$$

for $\ell > 1/r$ and 0 otherwise, where $r = \lambda_X/\lambda_Y$. Similarly,

$$H_Y(\ell) = \mathrm{P}\left(\frac{f_Y(Y)}{f_X(Y)} \le \ell\right) = 1 - (\frac{1}{r\ell})^{1/(r-1)}$$

for $\ell > 1/r$ and 0 otherwise. Also,

$$H_X^{-1}(p) = \frac{1}{r}(\frac{1}{1-p})^{(r-1)/r}$$

so that the concentration function is

$$\varphi(p) = H_Y(H_X^{-1}(p)) = 1 - (1-p)^{1/r}, \quad p \in (0,1),$$

the ROC curve of the likelihood ratio based optimal test is

$$\mathrm{ROC}_{\mathrm{opt}}(q) = q^{1/r} \qquad 0 \le q \le 1,$$

and

$$\mathrm{AUC}_{\mathrm{opt}} = \frac{r}{r+1}.$$

Example 2. Let X be exponential with rate parameter λ_X and assume Y is its length-biased version, so that

$$f_Y(y) = \frac{y\lambda_X e^{-\lambda_X y}}{1/\lambda_X} = \lambda_X^2 y e^{-\lambda_X y}, \quad y > 0,$$

i.e. Y is a gamma random variable with parameters 2 and λ_X. This is a Lorenz-Gini scenario, where it is easy to verify that

$$H_X(\ell) = \mathrm{P}\left(\frac{f_Y(X)}{f_X(X)} \le \ell\right) = \mathrm{P}\left(\lambda_X X \le \ell\right) = 1 - e^{-\ell}$$

whereas, after some calculus,

$$H_Y(\ell) \quad = \quad \mathrm{P}\left(\frac{f_Y(Y)}{f_X(Y)} \leq \ell\right) \quad = \quad \mathrm{P}(\lambda_X Y \leq \ell) \quad = \quad 1 \; - \; e^{-\ell} \; - \; \ell e^{-\ell}. \quad (5.5)$$

Since $H_X^{-1}(p) = -\log(1-p)$,

$$\varphi(p) = p + (1-p)\log(1-p), \quad \mathrm{ROC}_{\mathrm{opt}}(q) = q - q\log(q).$$

5.7 Discrete ROC

As clearly stated, the LR-based classification rule allows to obtain a proper ROC curve for finite, continuous and even more complex data types, under the only wide assumption that the likelihood ratio is meaningful. For binary data, the LR takes only a discrete set of values and therefore the associated ROC curve is only a discrete set of points (Figure 5.2, left panel).

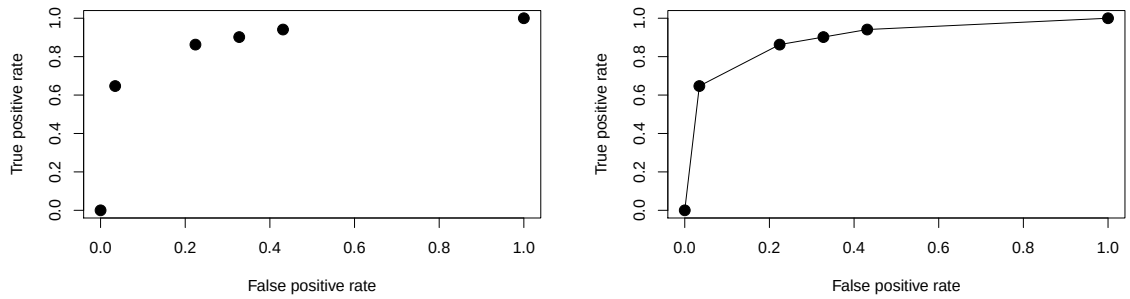


Figure 5.2 The usual definition of a ROC based on discrete data (left) and the ROC on the same data completed via a randomization device (right).

However, thanks to the randomization device introduced in Definition 5.1.1 and as a by-product of the relationship between the ROC curve and the concentration function, the discrete points can be joined to form also in this case a concave and proper ROC curve, as in Figure 5.2, right panel. This could be better understood by a concrete example.

5.7.1 Two finite measures

The following example is taken from the Encyclopedia of Biostatistics [80]. Suppose 109 patients have been classified as diseased (P_+) or not diseased (P_-), based on a gold standard such as biopsy or autopsy. On the basis of radiological exams, they have also been classified over five ordinal levels

$$\begin{aligned} -- &= \text{very mild} \\ - &= \text{mild} \\ +- &= \text{neutral} \\ + &= \text{serious} \\ ++ &= \text{very serious} \end{aligned}$$

Here are the results:

	$--$	$-$	$+-$	$+$	$++$	total
P_-	33	6	6	11	2	58
P_+	3	2	2	11	33	51

In particular, P_+ and P_- are two empirical measures, relative to the diseased and not diseased population respectively, derived from data. There are four possible values for the LR:

$$L = \begin{cases} \frac{58}{561} & \text{if } -- \\ \frac{58}{153} & \text{if } - \text{ or } +- \\ \frac{58}{51} & \text{if } + \\ \frac{319}{17} & \text{if } ++ \end{cases}$$

which give rise to four empirical ROC points $\{$ (25/58, 48/51); (19/58, 46/51); (13/58, 44/51); (2/58, 33/51)$\}$, shown in Figure 5.3. Now, as said, thanks to the randomization device, it is possible to ... connect the dots! This is so since the distribution functions of L under P_- and

P_+ are

$$H_-(\ell) = \begin{cases} 0 & \text{if } 0 \le \ell < \frac{58}{561} \\ \frac{33}{58} & \text{if } \frac{58}{561} \le \ell < \frac{58}{153} \\ \frac{45}{58} & \text{if } \frac{58}{153} \le \ell < \frac{58}{51} \\ \frac{56}{58} & \text{if } \frac{58}{51} \le \ell < \frac{319}{17} \\ 1 & \text{if } \frac{319}{17} \le \ell \end{cases}$$

and

$$H_+(\ell) = \begin{cases} 0 & \text{if } 0 \le \ell < \frac{58}{561} \\ \frac{3}{51} & \text{if } \frac{58}{561} \le \ell < \frac{58}{153} \\ \frac{7}{51} & \text{if } \frac{58}{153} \le \ell < \frac{58}{51} \\ \frac{18}{51} & \text{if } \frac{58}{51} \le \ell < \frac{319}{17} \\ 1 & \text{if } \frac{319}{17} \le \ell. \end{cases}$$

Therefore, the ROC curve can be calculated using equation (5.3):

$$\text{ROC}(x) = \begin{cases} \frac{319}{17}x & \text{if } 0 \le x < \frac{2}{58} \\ \frac{31}{51} + \frac{58}{51}x & \text{if } \frac{2}{58} \le x < \frac{13}{58} \\ \frac{7}{9} + \frac{58}{153}x & \text{if } \frac{13}{58} \le x < \frac{25}{58} \\ \frac{503}{561} + \frac{58}{561}x & \text{if } \frac{25}{58} \le x < 1 \end{cases}$$

The continuous ROC curve interpolates the empirical ROC points, as shown in Figure 5.3.

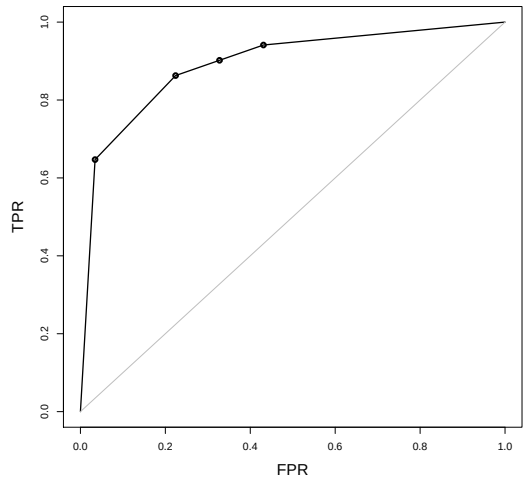


Figure 5.3 The proper ROC curve based on the LR interpolates the empirical ROC points.

Chapter 6

Model-based classification for binary data

Binary classification is a basic problem in Statistics and Machine Learning which could be approached in many different ways. In this chapter classification means clustering, or unsupervised learning. A number r of features, or predictors, is recorded on a sample of n statistical units with the aim of classifying them into two groups, without previously knowing the group memberships (labels) for any of them. Previously, dealing with ROC curve, the true membership class of each subject was assumed to be known (supervised learning).
In the following, I will consider binary features and only two groups (also indicated as clusters, classes or populations), if not differently stated. Extensions to multiple clusters and non binary answers can be obtained straightforwardly. The two clusters are conventionally identified as the positive $(+)$ and the negative $(-)$ group. Multivariate binary data naturally originate from polls, questionnaires, online automated interviews and represent the profile of a statistical unit about which several binary questions have been answered. In this type of problems, usually, data are collected to investigate if there is a separation in classes; therefore there is lower amount of structure and information compared to supervised learning (which, for example, characterizes diagnostic and genomic studies). Nevertheless, some structures will be introduced since binary features will be considered realizations of Bernoulli (i.e. binary) random variables with certain dependence or independence relationships and the focus will be on a model-based approach. In addition, the strong hypothesis of the local independence of the features will be retained also here, given its popularity in Machine Learning literature (see for example Murphy book [81]). Actually, to the best of my knowledge, this working assumption will only affect the clustering rules (namely the way to select units), allowing, on the other hand, possible dependences and correlation among units.

The chapter is organized as follow: after having clarified some basic notions of model-based clustering and having highlighted the well-known difference between hard and soft assignment, the implementation of hard clustering in R is provided, jointly with some considerations on dimensionality problems and parallel computing. Furthermore, latent class models (popular soft clustering methods) are also introduced and a comparison between the two approaches is provided, analysing a small highly cited dataset from educational setting (used as a toy example).
The material presented here is partly extracted from the short paper "ROC curves with binary multivariate data" presented at the 12th meeting CLADAG (Classification and Data Analysis Group) [82]. In addition, some results have been developed jointly with a master student in Mathematical Engineering and have been also included in his master thesis [83].

6.1 Model-based classification, unsupervised learning and clustering

It is important to recall some well known concepts to better frame the work presented in the following.
The focus will be on unsupervised learning approach, in which units (for example subjects) are grouped based on their characteristics, exploiting similarities without prior information or training. In particular units are not labelled and nothing is known on the true class membership. Unsupervised learning is the basic principle for clustering, a way of grouping multivariate data to highlight structures, with a systematic use of numerical methods. Essentially, the idea is to maximize the similarity within groups and the heterogeneity between groups. Often the similarity/dissimilarity among units is measured in terms of distance (assuming a metric space). Certainly, different clustering techniques exist (e.g. combinatorial, partitioning, hierarchical, fuzzy clustering, ... for further details see [84–86]), but I will concentrate here especially on model-based approach, which assumes a probabilistic framework and a formal statistical model for the observations, and therefore adopts standard methods for statistical inference. Indeed, model-based clustering solves in a principled way some difficulties of the analysis, such as the number of clusters or the choice of the method to be used: they reduce to a simpler problem of model selection. In addition, the statistical approach allows to assess uncertainty and to deal with outliers. A good clustering has predictive power, it describes data in a more effective way (allowing lossy compression), and it can highlight units of particular interest. Lastly, it is noteworthy to notice that classification and clustering do not have the exact same meaning, even if they are often used as synonyms:

the word "classification" is preferred in a supervised approach and it aims to identify the membership class of a new object with an algorithm trained on other labelled units.

6.1.1 Hard *versus* soft assignment

In general, in addition to what presented in the previous paragraph, it appears relevant to stress the difference between hard and soft clustering.

- Hard assignment: each unit belongs to one group only, in an exclusive way. The group labels $\gamma_1, \ldots, \gamma_n$ are unknown parameters; the reference clustering algorithm for this approach is k-means, where k indicates the number of clusters to be considered.
- Soft assignment: each unit belongs to different clusters with different probabilities or degrees of membership η_k for $k = 1, \ldots, K$; this approach gives rise to finite mixture models and latent class analysis (LCA) [87], while the Expectation-Maximization (EM) algorithm constitutes the reference.

Actually, in my opinion, hard assignment, which dates back to 1970s [88], is a more principled and straightforward way to assign units to given classes. However, it is a NP-hard optimization problem (i.e. a problem that can be solve in polynomial time by a nondeterministic Turing machine) and therefore it is possible to find the exact solution only for small datasets. Nevertheless, with modern technologies, it is of interest to investigate up to which point such approach can be used towards automatic identification of the two classes and the class labels of the sampled subjects. In particular, in the next paragraph an hard implementation will be proposed in R [75], considering dimensional and memory limits and exploiting parallel computing.
On the other hand, soft clustering acquired lots of popularity at present, thanks to its connection with mixture models, with various R packages and other software routines available for the direct calculus of membership probabilities; furthermore, it appears to be more flexible and it completely avoids dimensionality problems.
The two approaches classify units maximizing the likelihood (different quantities for hard and soft assignment) as detailed below.
Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be independent measurements of r predictors for n different statistical units (i.e. $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ for $i = 1, \ldots, n$), which arise from K possible groups; each unit has probability density function $f(\mathbf{x}; \boldsymbol{\theta}_k)$, $k = 1, \ldots, K$. Define $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ as the set of individual labels: $\gamma_i = k$ implies that $\boldsymbol{x}_i \in C_k$, i.e. C_k is the set of $\mathbf{x}_i$ assigned to the kth

population by $\boldsymbol{\gamma}$. Therefore, in the hard assignment, the likelihood function is:

$$L(\boldsymbol{\gamma}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) = \prod_{\mathbf{x} \in C_1} f(\mathbf{x}; \boldsymbol{\theta}_1) \cdots \prod_{\mathbf{x} \in C_k} f(\mathbf{x}; \boldsymbol{\theta}_k) = \prod_{i:\gamma_i=1} f(\mathbf{x}; \boldsymbol{\theta}_1) \cdots \prod_{i:\gamma_i=k} f(\mathbf{x}; \boldsymbol{\theta}_k), \quad (6.1)$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ have to be jointly estimated from the data. Units are then assigned to a specific group exploiting the maximum likelihood allocation rule [88], which allocates $\boldsymbol{x}$ to the population which maximizes (6.1). Actually, it can be proven that moving a sample point from $\hat{C}_k$ to $\hat{C}_\ell$, where $\hat{C}_k$ and $\hat{C}_\ell$ are the partitions obtained considering the maximum likelihood estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, will reduce the likelihood.
In the soft approach, instead, it is assumed that each $\mathbf{x}_i$ has probability η_k of coming from the kth population, $k = 1, \ldots, K$ and therefore the probability density function of each $\boldsymbol{x}$ is:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \eta_k f(\mathbf{x}, \boldsymbol{\theta}_k).$$

In this case, to correctly allocate $\boldsymbol{x}$ the quantity to be maximized is:

$$\eta_k L_k(\mathbf{x}).$$

In other words, $\mathbf{x}_i$ is assigned to the kth distribution when $\hat{\eta}_\ell f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_\ell) \leq \hat{\eta}_k f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_k)$ for all $\ell \neq k$, where $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\theta}}$ are the estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$. This allocation rule is indicated as the Bayes discriminant rule with respect to $\boldsymbol{\eta}$, even if, often, it is not a full Bayesian procedure, because priors on $\boldsymbol{\theta}$ are not considered.
The hard and the soft approach are equivalent considering $\boldsymbol{\gamma}$ as an (unobservable) random variable whose components are the outcomes of n independent multinomial trials [88, 89].
As already said and as it will be further discussed in the following, to maximize (6.1) with exact procedures is really challenging, because, in general, as detailed in [84], the number of different assignments is:

$$\frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^n.$$

This means that even in the simplest case with only two clusters the exhaustive search requires to evaluate $2^{(n-1)}$ assignments. Therefore, heuristic solutions are often preferred.
Lastly, it is important to notice that, to avoid limiting case with infinite likelihood, the hard approach assumes at least $(r+1)$ observations per group, i.e. $n \geq K(r+1)$. In the proposed implementation this problem is solved adding one success and one failure (binary case) for each predictor (additive Laplace smoothing, a modification of the Agresti and Coull adjustment [90, 91]).

6.2 K-means and EM algorithms

K-means and EM algorithms are the two reference methods for hard and soft clustering respectively. They are both iterative, involving two steps, they can converge to a local optimum instead of to a global one and therefore the starting values have a relevant role. For Gaussian mixture models the two approaches are closely related. In the next paragraphs the peculiar characteristics of the two algorithms are presented.

6.2.1 K-means

K-means clustering is a simple and elegant way for partitioning observations into a pre-specified number K of distinct and non-overlapping C_k groups. As previously stated, a good clustering procedure aims to minimize the variation within-group; therefore the optimization problem to be solved is:

$$\text{minimize} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in C_k} d(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

where $|C_k|$ denotes the size of cluster C_k, $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a measure of dissimilarity, usually the squared Euclidean distance, even if other choices are possible.
At the beginning, K-means randomly assigns numbers from 1 to K to each observation to define the initial cluster (observations with the same number are grouped together). Then, the following two steps are iterated until assignments do not longer change (Figure 6.1):

a. *Assignment step*: the cluster centroid m_k, i.e. the mean of the observations in that cluster, is computed; each observation is assigned to the class identified by the closed centroid.

b. *Update step*: the centroid is recalculated using the observations assigned to the cluster at the previous step.

Centroids are model parameters; at each step the value of $\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, m_k)$ is decreased and the algorithm always converges to a fixed point. However, depending on the initial random assignment, it could find a local solution and therefore it is important to run the algorithm many times and to select the minimal configuration. K-means considers only the distance between the means and the observations: it can not represent the weight or the breadth of each cluster, nor the size or the shape.

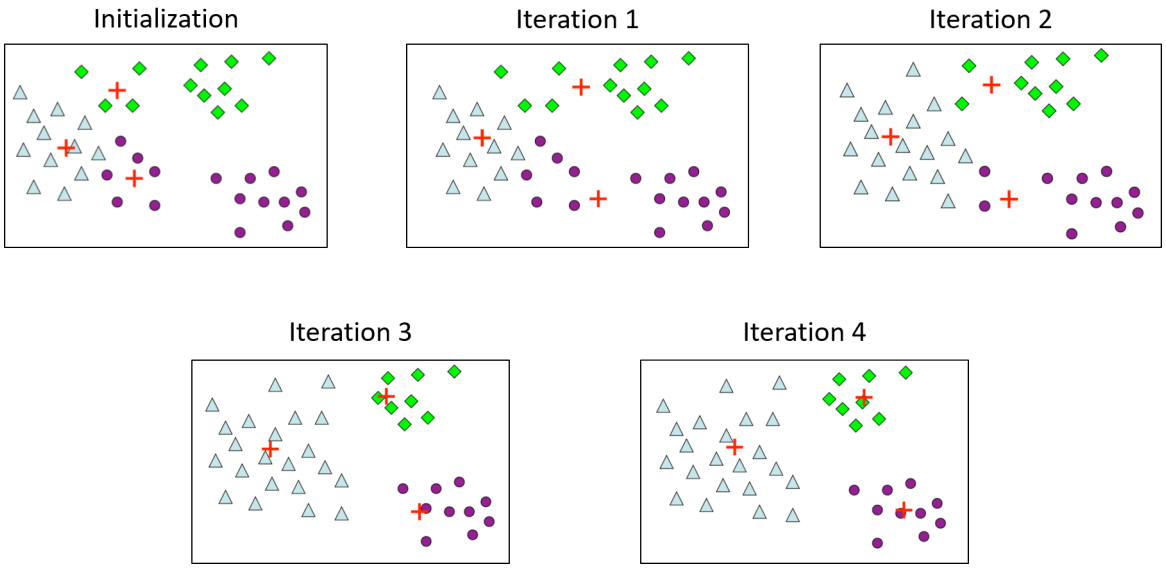


Figure 6.1 Example of how k-means algorithm works

6.2.2 EM algorithm

The Expectation-Maximization (EM) algorithm is a tool for maximum likelihood estimation in the soft assignment approach. It relies on the concept of complete-data likelihood L_C and observed-data likelihood L_O. Let $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$ be independent and identically distributed random variables with probability density function $f(\boldsymbol{t}, \boldsymbol{\theta})$ and $\boldsymbol{t}_i = (\boldsymbol{x}_i, \gamma_i)$ the complete data, where $\boldsymbol{\gamma}$ are latent variables representing the class membership, i.e. $\gamma_{ik} = 1$ if $\boldsymbol{x}_i$ belongs to group k (and 0 otherwise). Therefore:

$$L_C(\boldsymbol{t}_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{t}_i, \boldsymbol{\theta})$$

and integrating the latent variables out: $L_O(\boldsymbol{x}, \boldsymbol{\theta}) = \int L_C(\boldsymbol{t}, \boldsymbol{\theta}) d\boldsymbol{\pi}$, where $\pi_i = \mathbb{P}(\gamma_i = 1)$. The algorithm iterates two steps:

a. The E step computes the conditional expectation of $\mathcal{L}_C$ (the logarithm transformation of L_C), given the observations and the current parameters estimates, i.e.

$$Q(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}) = \mathbb{E}(\mathcal{L}_C(\boldsymbol{\theta}', \boldsymbol{T}) | \boldsymbol{X}, \hat{\boldsymbol{\theta}}), \tag{6.2}$$

where $\boldsymbol{\theta}'$ is a dummy variable.

b. The M step calculates the estimates of parameters $\hat{\boldsymbol{\theta}}$ that maximize the expected log-likelihood from the E step.

To better understand the procedure, the simpler case of a mixture of two Gaussian distributions is presented: $Y = (1-\gamma)Y_1 + \gamma Y_2$, where $\gamma \in \{0,1\}$, with $\mathbb{P}(\gamma = 1) = \pi$. Indicating with $\phi_{\boldsymbol{\theta}}(x)$ the normal probability density with parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, the density of Y is:

$$g_Y(y) = (1-\pi)\phi_{\boldsymbol{\theta}_1}(y) + \pi\phi_{\boldsymbol{\theta}_2}(y).$$

Indeed the model parameters are $\boldsymbol{\theta} = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. The log-likelihood to be maximized is:

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log[(1-\pi)\phi_{\boldsymbol{\theta}_1}(x_i) + \phi_{\boldsymbol{\theta}_2}(x_i)].$$

If the values of γ_i are known, then

$$\mathcal{L}_O(\boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \sum_{i=1}^{n} [(1-\gamma_i)\log\phi_{\boldsymbol{\theta}_1}(x_i) + \gamma_i \log\phi_{\boldsymbol{\theta}_2}(x_i)] + \sum_{i=1}^{n} [(1-\gamma_i)\log(1-\pi) + \gamma_i \log\pi]$$

However, usually the γ_i are not known and, therefore, their expectations are considered. In particular, one can define:

$$\text{responsibility}_i := z_i = \mathbb{E}(\gamma_i | \boldsymbol{\theta}, \boldsymbol{X}) = \mathbb{P}(\gamma_i = 1 | \boldsymbol{\theta}, \boldsymbol{X}).$$

Responsibilities are computed in the E step:

$$\hat{z}_i = \frac{\hat{\pi}\phi_{\hat{\boldsymbol{\theta}}_2}(x_i)}{(1-\hat{\pi})\phi_{\hat{\boldsymbol{\theta}}_1}(x_i) + \hat{\pi}\phi_{\hat{\boldsymbol{\theta}}_2}(x_i)},$$

where $\hat{\pi}$ and $\hat{\boldsymbol{\theta}} = (\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2)$ are initialized respectively to 0.5 and to random observations for $\hat{\mu}_1$ and $\hat{\mu}_2$ and to equal overall sample variance for $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$.
The maximization step evaluates:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n (1-\hat{z}_i)x_i}{\sum_{i=1}^n (1-\hat{z}_i)} \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^n (1-\hat{z}_i)(x_i-\hat{\mu}_1)}{\sum_{i=1}^n (1-\hat{z}_i)}$$
$$\hat{\mu}_2 = \frac{\sum_{i=1}^n \hat{z}_i x_i}{\sum_{i=1}^n \hat{z}_i} \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^n \hat{z}_i (x_i-\hat{\mu}_2)}{\sum_{i=1}^n \hat{z}_i}$$
$$\hat{\pi} = \sum_{i=1}^{n} \hat{z}_i / n$$

The actual maximizer of the likelihood is obtained when spikes of infinite height are considered for each data point ($\sigma_1^2 = 0$), but this is not a useful solution and the additional constraint $\hat{\sigma}_1^2, \hat{\sigma}_2^2 > 0$ is considered.

In general, the EM algorithm works because at each iteration the log-likelihood never decreases, as shown in [84]. However, the convergence to the local maximum of L_O can be slow and problems could arise when some clusters contain only a few observations, when observations are concentrated close to a linear subspace or when, in the multivariate normal case, the covariance matrix of one or more components is singular or nearly singular.

6.3 A short recap on LCA

The latent class analysis (LCA), firstly proposed by Lazarsfeld and Henry [87], is a model-based clustering method, rooted in finite mixture models literature, which classifies multivariate categorical data based on a soft approach. This model aims to describe the heterogeneity and dependence existing in data considering the class membership as a discrete latent (unobservable) variable and exploiting the information provided for each unit by manifest (observed) variables.

In details, consider n units $\boldsymbol{x}_i$ for $i = 1, \ldots, n$ (e.g. subjects replying to a questionnaire), r categorical (for simplicity binary) variables collected for each unit (e.g. replies to r questions in a survey, results of r educational assessment tests) generically indicated with items, and a fixed number K of different latent classes; LCA relies on the strong assumption of local independence, i.e. the manifest variables are independent conditionally on knowing the membership class. In addition, the n individuals constitute a random sample from some population, each has η_k probability to belong to class k for $k = 1, \ldots, K$ (the mixing proportion is the *a priori* probability) and the probability of giving a positive response to a particular item j is the same for all individuals in the same group and it is indicated with $\pi_{jk} = \mathbb{P}(x_j = 1|k)$ for $j = 1, \ldots, r$ and $k = 1, \ldots, K$.

In the binary case, as it is well known, the manifest variables are marginally distributes as Bernoulli, i.e.

$$f(\mathbf{x}_i, \boldsymbol{\theta}_k) = \prod_{j=1}^{r} \pi_{jk}^{x_{ij}} (1 - \pi_{jk})^{1 - x_{ij}}$$

being $\boldsymbol{\theta}_k = (\pi_{1k}, \ldots, \pi_{rk})$ the vector of success probabilities in the k-th population. Therefore, as previously stated

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \eta_k f(\mathbf{x}_i, \boldsymbol{\theta}_k).$$

Maximum likelihood estimates of the parameters of the model are obtained via the EM algorithm, which unfortunately could return a local maximum, instead of a global one. In addition, there could be identifiability problems: to avoid this a necessary but not sufficient

condition is that the number of possible values of the manifest variables is greater than the number of parameters to be estimated [92], i.e

$$2^r - 1 \geq (K-1) + rK.$$

Actually, LCA requires to estimate $(K-1) + K\sum_j (c_j - 1)$ parameters (where c_j is the number of different categories of variable j) and difficulties can arise. More parsimonious variants have been proposed (see for example the recent book by Bouveyron and colleagues [86]).
With respect to clustering, after having fitted a latent class model using the EM algorithm, a classification step should be considered: units are categorised based on the maximum a posteriori principle and assigned to the class which maximize the responsibility z_{ik} of x_i. Actually, for each individual, it is possible to compute the *a posteriori* probability, i.e. the probability that the individual falls into a specific class, given a particular response pattern; fixed the threshold (usually the 50% when $K = 2$), it is easy to decide the membership class. Alternatively, the cluster can be obtained jointly with the estimation of parameters, introducing the classification directly in the algorithm: a C-step is evaluated before the update of the estimates. This "CEM" algorithm could add bias, but the convergence is faster.
Certainly, there are connections with distance-based methods [86] and it is possible to modify latent class models to include covariates. Furthermore, lots of work has been dedicated to find ways to relax the hypothesis of local independence, for example including a new multivariate Bernoulli variable in the parametric mixture model [93], adding random effects and additional (possibly continuous) latent variables to handle dependencies. However these topics will not be detailed here.

6.3.1 The `poLCA` R package

The Polytomous Variable Latent Class Analysis (`poLCA`) is a user-friendly R package for fitting latent class models [94]. It allows to estimate η_k and π_{jk} for $j = 1, \ldots, r$ and $k = 1, \ldots, K$, maximizing the following log-likelihood function via the EM algorithm (the extension to more than two outcomes - i.e. to polytomous variables - is straightforward):

$$\mathcal{L} = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \eta_k \prod_{j=1}^{r} (\pi_{jk})^{x_{ij}} (1 - \pi_{jk})^{1-x_{ij}}.$$

In addition, `poLCA` also provides the posterior probability that each individual belongs to a given class, conditionally on the observed values of the manifest variables:

$$\hat{p}(k_i|\boldsymbol{x}_i) = \frac{\hat{\eta}_k f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_k)}{\sum_{k=1}^{K} \hat{\eta}_k f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_k)}.$$

As already said, the EM algorithm can converge to a local maximum (depending on the initial values and on the complexity of the model): it is always recommended to run the model at least a couple of times and this can be done simply changing the value of `nrep` in the call of the function `poLCA()`. Coding details on LCA models can be found in Appendix B.

```
Model 1: llik = -331.7637 ... best llik = -331.7637
Model 2: llik = -331.7637 ... best llik = -331.7637
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$V1
           Pr(1)  Pr(2)
class 1:  0.7914 0.2086
class 2:  0.2466 0.7534

$V2
           Pr(1)  Pr(2)
class 1:  0.9317 0.0683
class 2:  0.2197 0.7803

$V3
           Pr(1)  Pr(2)
class 1:  0.9821 0.0179
class 2:  0.5684 0.4316

$V4
           Pr(1)  Pr(2)
class 1:  0.9477 0.0523
class 2:  0.2925 0.7075

Estimated class population shares
 0.4134 0.5866

Predicted class memberships (by modal posterior prob.)
 0.4577 0.5423

============================================================
Fit for 2 latent classes:
============================================================
number of observations: 142
number of estimated parameters: 9
residual degrees of freedom: 6
maximum log-likelihood: -331.7637

AIC(2): 681.5273
BIC(2): 708.1298
G^2(2): 8.965682 (Likelihood ratio/deviance statistic)
X^2(2): 9.459244 (Chi-square goodness of fit)
```

Figure 6.2 Example of `poLCA` output

`poLCA` automatically calculates goodness-of-fit values, such as BIC, AIC, Pearson χ^2 and likelihood-ratio chi square G^2 (see [95] for details) to help in deciding the number of clusters (specified in the input by `nclass` command) to be considered; usually one starts from the independence model with $K = 1$ and then increases the number of latent variables iteratively by one until obtaining a satisfying good fit. The aim is to minimize χ^2 and G^2, trying to limit the number of parameters to be estimated, considering that the distributional assumptions for the calculus of these statistics could be violated if groups contains too few observations. The `poLCA` function fits a LCA model on data given in data-frame format (including manifest variables and covariates, if the case), in which categorical items are coded as integers starting from one. Indeed, manifest variables are the responses variables for the model and are included in it in the following "formula" way: `cbind(X1, X2, X3) ~ 1` (with a linear

combination of covariates instead of 1, if the case). Apart from goodness-of-fit values, `poLCA` returns the estimated class-conditional outcome probabilities, the posterior probabilities, the class predictions (the `probs`, `posterior` and `predclass` objects respectively), and, of course, the maximum values of the estimated model log-likelihood (`llik`) (see Figure 6.2 for an example of output). Lastly, it is important to notice that the numerical order of the estimated latent classes returned in the output is totally arbitrary, and is determined solely by the start values of the EM algorithm.

6.4 A new implementation of hard assignment in R

Despite the popularity of LCA models, in my opinion, the hard approach can be considered a more basilar way to assign units to clusters; therefore, considering the technical and technological developments of the recent years, it appears noteworthy to investigate up to which point one can rely on hard assignment for classifying subjects. As previously stated, it is a combinatorial problem and memory allocations limits are explored to find the exact solutions; however, parallel computing on one side and heuristic optimization procedures on the other allow to obtain some interesting results.
In particular, consider n units $\mathbf{X_i} = (X_{i1}, \ldots, X_{ir})$ for $i = 1, \ldots, n$ characterized by unknown group labels $\gamma_1, \ldots, \gamma_n$ (which are therefore parameters from a statistical point of view) and r items. The equality $\gamma_i = +$ (respectively $\gamma_i = -$) would mean that the i-th unit would belong to the positive (respectively negative) group. In the binary case, for a given realization of $\gamma_1, \ldots, \gamma_n$, $\mathbf{X_i} = (X_{i1}, \ldots, X_{ir})$ are independent and each distributed as a MultiBernoulli with independent binary components with parameters $\boldsymbol{\pi}^+ = (\pi_1^+, \ldots, \pi_r^+)$ if $\gamma_i = +$ or $\boldsymbol{\pi}^- = (\pi_1^-, \ldots, \pi_r^-)$ if $\gamma_i = -$ respectively:

$$\mathbf{X_i} = (X_{i1}, \ldots, X_{ir}) \sim \text{MultiBernoulli}(\pi_1^+, \ldots, \pi_r^+) \text{ if } \gamma_i = +,\ i = 1, \ldots, n$$
$$\mathbf{X_i} = (X_{i1}, \ldots, X_{ir}) \sim \text{MultiBernoulli}(\pi_1^-, \ldots, \pi_r^-) \text{ if } \gamma_i = -,\ i = 1, \ldots, n.$$

Also in this approach, features are assumed to be locally independent. Therefore, the likelihood function can be written as:

$$L(\boldsymbol{\gamma}, \boldsymbol{\pi}^+, \boldsymbol{\pi}^-) = \prod_{i:\gamma_i=+} \prod_{k=1}^{r} (\pi_k^+)^{x_{ik}} (1-\pi_k^+)^{1-x_{ik}} \prod_{i:\gamma_i=-} \prod_{k=1}^{r} (\pi_k^-)^{x_{ik}} (1-\pi_k^-)^{1-x_{ik}} \tag{6.3}$$

$$= \prod_{k=1}^{r} (\pi_k^+)^{n_k^+} (1-\pi_k^+)^{\nu - n_k^+} (\pi_k^-)^{n_k^-} (1-\pi_k^-)^{n-\nu-n_k^-} \tag{6.4}$$

where

$$n_k^+ = \sum_{i:\gamma_i=+}^{n} x_{ik}, \quad n_k^- = \sum_{i:\gamma_i=-}^{n} x_{ik} \quad \text{and} \quad \nu = \sum_{i:\gamma_i=+} 1$$

are unknown quantities and should be estimated.
Differently from `poLCA`, here the focus is on profiles m, i.e. on fixed binary strings of length r, which represent possible realizations of each of the random vectors $\mathbf{X_i}$. The following property holds:

Lemma 6.4.1. *Units with identical profiles share the same maximum likelihood group assignment, i.e. if two units i and ℓ have the same profile, then the respective maximum likelihood estimates are equal $\hat{\gamma}_i = \hat{\gamma}_\ell$.*

Proof. Suppose that two units i and ℓ share the same profile and are assigned to different groups, say $\gamma_i = +$ and $\gamma_\ell = -$ for the sake of definiteness. Then the likelihood can be made not smaller by changing the assignment $\gamma_i = +$ to $\gamma_i = -$ if

$$\prod_{k:x_{ik}=1} \pi_k^+ \leq \prod_{k:x_{ik}=1} \pi_k^- \quad \text{(case A)} \tag{6.5}$$

or, vice versa, by changing the assignment $\gamma_\ell = -$ to $\gamma_\ell = +$ if

$$\prod_{k:x_{ik}=1} \pi_k^+ \geq \prod_{k:x_{ik}=1} \pi_k^-, \quad \text{(case B)}. \tag{6.6}$$

To see this, notice that in case A changing $\gamma_i = +$ to $\gamma_i = -$ would imply n_k^- increases by 1 and n_k^+ decreases by 1 for those k such that $x_{ik} = 1$, while ν would change to $\nu - 1$ and no other functions would change. Therefore

$$\prod_{k:x_{ik}=1}^{r} (\pi_k^+)^{n_k^+} (1-\pi_k^+)^{\nu - n_k^+} (\pi_k^-)^{n_k^-} (1-\pi_k^-)^{n-\nu-n_k^-}$$
$$\leq \prod_{k:x_{ik}=1}^{r} (\pi_k^+)^{n_k^+ - 1} (1-\pi_k^+)^{\nu - n_k^+} (\pi_k^-)^{n_k^- + 1} (1-\pi_k^-)^{n-\nu-n_k^-}$$

which is true if and only if $\prod_{k:x_{ik}=1} \pi_k^+ \leq \prod_{k:x_{ik}=1} \pi_k^-$, as stated.
Similarly, in case B, the likelihood would be made no smaller by changing $\gamma_\ell = -$ to $\gamma_\ell = +$. □

The lemma has important consequences stated in the next theorem.

Theorem 6.4.2. *Maximum likelihood estimates* $\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\pi}}^{+}, \hat{\boldsymbol{\pi}}^{-}$ *of* $\boldsymbol{\gamma}, \boldsymbol{\pi}^{+}, \boldsymbol{\pi}^{-}$ *can be obtained with a finite search of at most* 2^r *iterations.*

Proof. By the previous lemma, it is enough to consider the likelihood of each of the 2^r profiles and to maximize it, setting then $\hat{\gamma}_i, i = 1, \ldots, n$ equals to the assignment of the corresponding profile. □

Therefore, the new implementation of hard assignment requires at most 2^r different profiles to be evaluated; this can signify considerable computational savings in the case $r < n$, while in the case $r > n$ dimensional reduction strategies should be implemented.
The building blocks of the new algorithm to implement hard assignment in R (of which a schematic idea is presented in Figure 6.3, whereas the complete code is reported in Appendix B), are:

- the dataset (given in input), containing the units to be classified (by row) and the different (binary) items by column;
- the profile matrix, obtained from the original data, deleting units with the same pattern and adding a variable to take care of the multiplicity of each profile;
- the assignment matrix, showing all possible configurations arisen allocating each profile to different clusters.

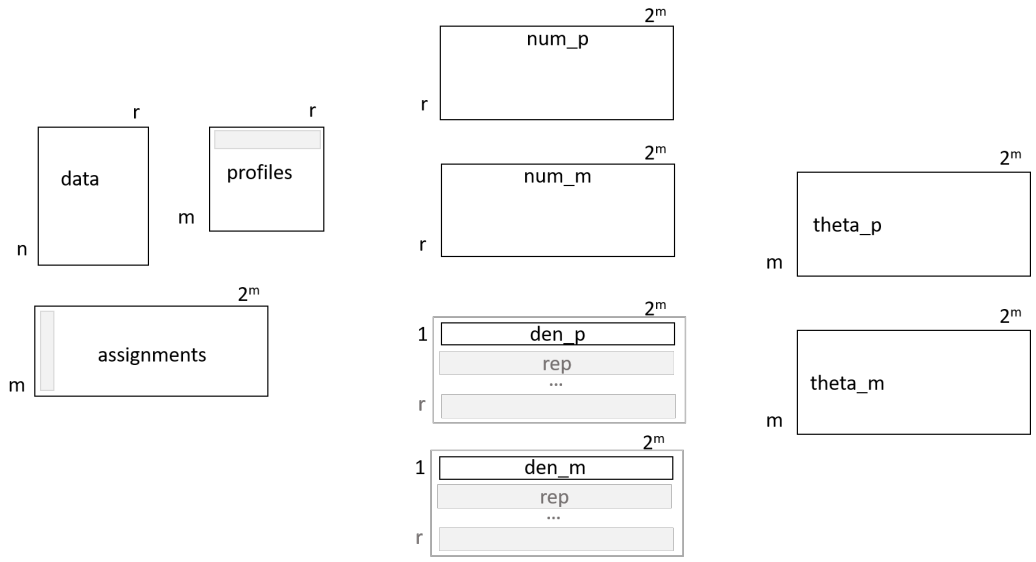


Figure 6.3 Building blocks of the implementation of hard assignment in R

Indeed, dimensionality problems originate from the assignment matrix, whose size grows exponentially with the number of profiles and quickly fills up all the memory available. As it has been showed in [83], the memory space necessary to allocate the matrix is approximately 2^{m-13} Mb and it duplicates just adding one profile. Therefore, it is still impossible to find the maximum of the likelihood with an exhaustive search for more than $m = 30$ profiles. The limit persists even considering parallel computing (with the R package `parallel` for example): in this case the assignment matrix is split in smaller sub-matrices, depending on the number of cores available (here 4), and then the maximum of the likelihood is searched independently in each sub-matrix. Certainly there is a improvement in the time required to obtain the solution (with a speedup of more than 2x).
To concretely implement hard clustering the log-likelihood (the logarithmic transformation of the likelihood function (6.3) is considered only for convenience of calculus) of each configurations (i.e of each column of the assignment matrix) is evaluated. First of all, the probability of profiles (the $\texttt{theta}_p$ and $\texttt{theta}_m$ in Figure 6.3) is computed directly counting the 1s in the profiles matrix corresponding to $+$ (respectively $-$) for a specific configuration `cl` (values stored in $\texttt{num}_p$ and $\texttt{num}_m$ matrices) and dividing by the total number of $+$ (respectively $-$) in the configuration (collected in $\texttt{den}_p$ and $\texttt{den}_m$). Additive Laplace smoothing is applied, adjusting frequencies adding one success to the numerator and one success and one failure to the denominator, to avoid limiting cases. The log-likelihood is then evaluated by `logLikeEval01` function, exploiting the local independence assumption and taking into account the multiplicity of the profiles: in particular the elements *CpL* and *CmL* indicates the components of the log-likelihood with respect to the two clusters ($+$ and $-$). In addition, some vectorized operations have been introduced to improve the code, using the `apply` family functions and the `dyplr` package available in R (in particular, `colLike01` function is the vectorize version of `logLikeEval01`).
Looking at the results obtained on a random dataset and on a small real dataset (described in the next section) it is evident that the worse configuration (i.e. the lowest value of the likelihood) is the one that assigns all the profiles to the same cluster. In addition, symmetrical configurations always return the same value for the log-likelihood, because no weights have been considered and therefore the two clusters $+$ and $-$ are formally equal.
To overcome the limit of about 30 profiles (which means quite small datasets) heuristic methods have been proposed: instead of exhaustive search, only a subset of possible configurations is evaluated for finding the best solution. The simpler idea is to adopt a basic Monte Carlo approach, selecting at random a number of configurations and computing the log-likelihood for them: it is possible that this method does not converge to the true global maximum, but the computation is very fast and therefore good solutions could be found increasing the

number of configurations and iterating the process over time. On the other hand, starting from the algorithm presented by Li et al. [96] to efficiently perform optimization via Monte Carlo method exploiting entropy, a new procedure essentially based on swapping of clusters is proposed. Actually the initialization is completely random and the deterministic assignment

Algorithm 2 Hard clustering via swap

a. Initialization

- Initialize the cluster configuration cl
- Compute initial log–likelihood $L^{(0)}$
- Initialize best log-likelihood $L^{(\text{Best})} = L^{(0)}$
- Initialize best configuration $cl_{\text{Best}} = cl$

b. Iterations from 1 to `nIter`

- Randomly select a profile p in the cluster $+$
- Move p to cluster $-$
- Compute new log-likelihood L
- Compute new cluster configuration cl
- If $L > L^{(\text{Best})}$

  $L^{(\text{Best})} = L$

  $cl_{\text{Best}} = cl$

c. Return L^{Best} and cl_{Best}

of all the profiles to the same cluster could be considered as the starting point. However to improve the convergence a wiser idea is to consider the result of few iterations of the basic Monte Carlo methods as initial guess for cl [83]. Coding details can be found in Appendix B.

6.5 A toy example

To better understand how the new implementation of hard assignment works, it has been tested on a small dataset from educational assessment, commonly used in latent class analysis models, described by [95] and based on data from [8]. It contains answers to $r = 4$ binary questions provided by $n = 142$ subjects. The information was collected to study the learning process in children and, in particular, to classify students in two groups (the P_+ and

Table 6.1 Latent classes analysis (`poLCA` Ass) *versus* hard assignment (Hard Ass) on Macready and Dayton data [8]. Posterior probabilities calculated using `poLCA`.

Profiles	**Frequency**	**Prob.** $\in P_-$	**Prob.** $\in P_+$	`poLCA` **Assign.**	`hard` **Assign.**
1111	15	0.000	1.000	+	+
1101	23	0.002	0.998	+	+
1110	7	0.002	0.998	+	+
0111	4	0.001	0.999	+	+
1011	1	0.003	0.997	+	-
1100	7	0.087	0.913	+	+
1001	6	0.095	0.905	+	-
0101	5	0.025	0.975	+	+
1010	3	0.100	0.900	+	-
0110	2	0.026	0.974	+	+
0011	4	0.029	0.971	+	-
1000	13	0.822	0.178	-	-
0100	6	0.526	0.474	-	+
0001	4	0.550	0.450	-	-
0010	1	0.563	0.437	-	-
0000	41	0.982	0.018	-	-

P_- populations): "masters" students were expected to answer correctly to most questions (majority of 1's in the profile), while mostly 0's were expected from "non-masters". Indeed, students could also answer in the right way by chance, as well as give the wrong answer due to oversight. There are 16 different profiles and it is therefore possible to find the maximum of the likelihood function with an exhaustive search.

```
> exactsol
V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 loglike
1 1 1 1 0 1 0 1 0 1 0 0 1 0 0 0 -240.1591
0 0 0 0 1 0 1 0 1 0 1 1 0 1 1 1 -240.1591
```

Due to the simplicity of the problem also the heuristic procedure quickly converges to the same value.

Table 6.1 reports the comparison of hard and soft assignment approaches (via LCA) and the posterior probabilities calculated by the `poLCA` package: 20 out of 142 subjects are classified differently by the two methods; the reasons for the observed differences are a matter of discussion. In particular, unexpectedly, there are profiles with high posterior probabilities in the latent variable approach, assigned to the other population by the hard algorithm. In

addition, looking at Table 6.1, it seems that the second item has a higher discriminant power in the hard approach: the profile is automatically assigned to $-$ if $\text{Item}_2 = 0$, independently from the correctness of other replies. This observation originates the conjecture summarized below and detailed in [83].
Under the local independence hypothesis, assume that all possible different profiles are present in the dataset, with the same frequency and the same discriminant power (weight) for all the items; in this completely symmetric case the number of optimal configurations (i.e. the number of assignment returning the maximum value of the likelihood) is equal to the number of different items r and the profiles can be divided in two clusters based on answers to a single randomly chosen item q (with $q = 1, \ldots, r$).
Furthermore, in this completely symmetric situation, `poLCA` and the new implementation of hard clustering give the same results.

6.6 Future developments

Results presented in this chapter rely on the strong assumption of local independence of features and are limited to the binary case. Extensions to multivariate (more than binary) populations and to multi-outcomes items (possibly with a different number of categories for different items) follow straightforwardly and constitute future developments of the present work; indeed, in `poLCA` is sufficient to specify the number of clusters in `nclass` and, thanks to goodness-of-fit values automatically reported in the output, it is easy to select the best model.

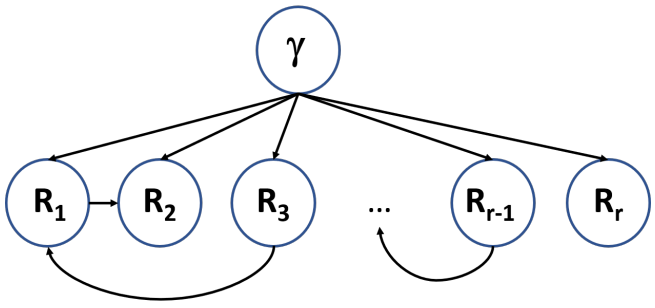


Figure 6.4 Example of proper Bayesian network to model dependencies

In the hard approach, certainly the likelihood function is no longer simply the product of the π_k (for $k = 1, \ldots, r$) and its complementary probability under $+$ and $-$, but it is naturally replaced by the product of probabilities $p_{k,1}, p_{k,2}$ and $p_{k,3}$ for an item k with 3 outcomes for example, still retaining the local independence assumption. Dimensionality problems are

even more relevant and the heuristic methods appear the only feasible.
On the other hand, the DAG (Direct Acyclic Graphs) theory and a proper Bayesian networks approach can be adopted to model possible dependencies among predictors (Figure 6.4). This can be done in R using the package `bnlearn` (as hinted in [83]). New challenges arise: in this framework γ_i becomes a proper parameter and priors should be considered on it; in addition, it is necessary to jointly estimate the structure of the graph (the edges).
A deeper analysis of Bayesian networks and of other ways to relax the hypothesis of local independence represent future developments of this work.

Chapter 7

Applications

This final chapter presents two more applied projects developed in collaboration with Bayer Pharmaceutical (Department of Clinical Statistics Europe Oncology, Berlin) and the Piedmont Cancer Registry (Turin). At this point of the dissertation, it should be quite clear that the "underlying latent variables" around which I have built my PhD work are biomarkers: from the search of new biomarkers for prostate cancer I moved to a deep analysis of ROC curves, likelihood ratio and related properties till the classification/clustering problem by itself. Biomarkers play an important role also here.

- Protein measurements have been included among predictors in the predictive models built on clinical trials data.
- The investigation of genomic characteristics of diseases (in particular cancers) is of great interest at present from diagnostic, prognostic and therapeutic points of view. However, epidemiological analyses are an important preliminary step to provide a broader perspective for pointing out challenging topics, hypotheses and questions on which to focus the biomarkers research. In particular, I have been deeply involved in cutaneous malignant melanoma research as a member of GEM (Genes, Environments and Melanoma) consortia and, in collaboration with the Piedmont Cancer Registry, I leaded the first and largest observational studies on European incidence trends of this malignancy (stratified by thickness's level) and on trends of lethal melanomas.

Part of the material included in this chapter has been extracted from two papers I co-authored "Trends in incidence of thick, thin and in situ melanoma in Europe" [97] and "Skin melanoma deaths within 1 or 3 years from diagnosis in Europe" (just submitted). In addition, a poster

has been presented at the PSI - Statisticians in Pharmaceutical Industry Conference, held in London in June 2019.

7.1 From baseline data to outcomes: are artificial intelligence based models really competitive?

In clinical trials a huge amount of information is routinely collected for each subject, with a large investment of time and resources; however, only a small fraction of these data is commonly used in standard analyses, for stratification and regulatory assessment purposes. At present, the popularity of Big Data and Artificial Intelligence (AI) approaches in different fields paves the way to adopt and adapt these methods for extracting useful information also from clinical data.
The project, done in collaboration with the Department of Clinical Statistics Europe Oncology of Bayer Pharmaceutical, during a six months internship, aims to develop multivariate predictive models for treatment efficacy with time-to-event outcomes (overall survival), using baseline data routinely collected in clinical trials.

7.1.1 A short introduction to (some) machine learning and deep learning methods

Before looking at data and results, it is necessary to shortly introduce the methods used, in particular survival and regression random forests (RF) and feed-forward neural networks (NN); only essential characteristics are recalled here below, while for a detailed review of methods see for example [84]. Actually, the idea was to compare these popular machine learning and deep learning techniques with traditional Cox models. As it is well known, for the machine learning approach, there is not a single algorithm which works best across all possible scenarios, but the performance strongly depends on the specific characteristics of the dataset to be analysed, as well as on applications.

Random Forests They are ensemble non-parametric learning methods that operate by constructing multitudes of decision trees T_b for $b = 1, \ldots, B$ at training time; proposed by Breiman in 2001 [98], they show improvements over bagging procedure evaluating only a subset of predictors selected at random at each split and therefore realising a decorrelation

of trees and a consequent reduction of variance averaging them. Usually classification and regression random forests are considered and predictions are obtained respectively by:

- the majority vote, i.e. for each tree the predicted class is recorded and then the overall prediction is based on the most occurring class;
- $\hat{f}_{\text{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b)$ where Θ_b characterizes the bth random forest tree in terms of split variables, cut-points at each node, and terminal node values.

More recently, random survival forests have been introduced to deal with right-censored data [99]: in this case the quantity to be averaged is the cumulative hazard function of each tree. In general, random forests can handle continuous, binary and categorical variables and, without making any assumptions, allow to model non-linear relations among predictors, as well as possible interactions and high correlation. Two important characteristics are the use of Out-Of-Bag (OOB) samples and the variables importance measure: the latter allows to rank variables based on their relevance in the development of the trees and therefore it is a measure of the prediction strength of each predictor. On the other hand, the OOB samples provide a simply way to estimate the test error, without considering training and validation sets or cross-validation approach: for each observation z_i the random forest predictor is built by averaging only those trees corresponding to bootstrap samples in which z_i did not appear. Commonly OOB observations constitute about the 30% of all data and prevent from overfitted results, even if random forests have the advantages to not overfit the data increasing the number of trees considered (i.e. the depth of the forest). Certainly, there are some important hyper-parameters to be tune using an exhaustive grid search:

- `ntree`: the number of trees to grow the forest;
- `mtry`: the number of variables considered to split a node (default values are $p/3$ for regression RF and $\sqrt{p}$ for classification RF, with $p =$ total number of predictors);
- the node size: the minimum number of subjects in the terminal nodes (default values are 5 for regression RF and 1 for classification RF);
- the sample size: the fraction of observations to sample;
- the split rule: depending on the type of RF, different splitting criteria can be considered in each node to build the forest. In particular, for regression and classification RF they are based on the decrease of the node impurity measured in term of estimates of response variance and Gini index respectively; for survival it relies on log-rank test [99].

The selected random forests model optimizes an appropriate OOB performance measure (the misclassification frequency for classification, the mean squared error for regression and the concordance index C for survival). An interesting extension of random forests (known as multivariate RF) allows to include more than one outcome, i.e. to simultaneously model multiple response data; this is useful when there are dependencies between responses; it is similar to the univariate case and the only difference is in the splitting criteria, which involve predictors related to all the responses; the impurity measure is now based on covariance (instead of variance) [100].

Neural Networks As it is well known and clearly detailed for example in [7, 84, 101], neural networks are parametric models, based on non-linear functions of linear combinations of the features (the input data x_i), as schematically represented in Figure 7.1, panel a, and in the network diagram (panel b); they can process all kind of data coded in numeric form and, depending on the problem at hand and on the type of outcome analysed, similarly to random forests, it is possible to have classification, regression and survival networks. Actually, neural networks, as machine learning and deep learning algorithms in general, aim to transform input data in a meaningful way to better model the target (i.e. the outcome); a layer-wise architecture is considered and in each (hidden) layer an increasingly more abstract representation of the data is stored; in particular, each layer is characterized by: a different number of "neurons" (nodes) fully connected (in case of dense networks) with the previous and the next layer, the associated weights which contain the information learned by the network during the training, the activation function f (not necessary the same for each layer) and the loss function which defines the feedback signal for the learning phase and it is minimized at each iteration by adjusting weights values. Indeed, this adjustment is made by the optimizer, which determines how the learning process proceeds, implementing a back-propagation algorithm applying the chain rule of differentiation to the computation of the gradient values; at each step an error is evaluated (via the loss function), comparing the observed and expected outcomes and an improvement in prediction is realized moving the weights in the direction opposite to gradient descent.
On the other hand, neural networks include a number of hyperparameters to be tuned:

- the number of hidden layers to be considered and the number of neurons to be included in each layer;
- the number of `epochs`, i.e. the number of times the whole training dataset is shown to the network during the training;

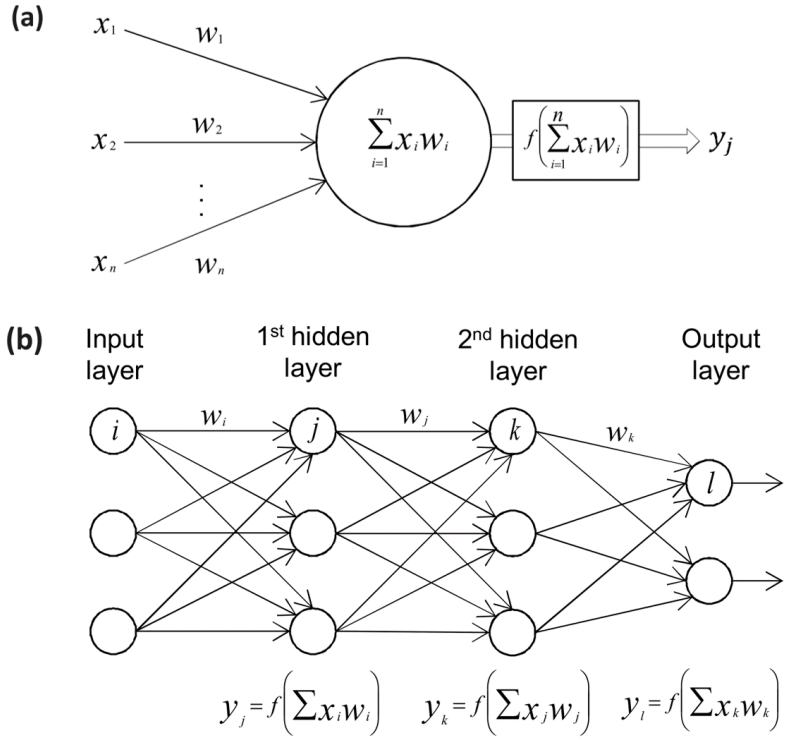


Figure 7.1 Neural network and single neurons (node) scheme (extracted from [7])

- the learning rate which measures how quickly the network updates its parameters and indicates the magnitude of the move of weights at each iteration of the back-propagation algorithm: a too small value could result in a local minimum, whereas a too large one could return a totally random point on the loss function curve;
- the activation functions to model non linearity in the dataset; the most common are: `elu` (exponential linear units, which converges more quickly and it is often preferred with survival outcomes), sigmoid (which takes a real value and returns 0 or 1 and therefore it is useful for classification), `relu` (rectified linear units, defined by $\max(0,x)$, similar to sigmoid but with better performance), tanh (which squeezes real values into $[-1,1]$ interval) and the `softmax` (which returns the probabilities of each target class over all possible target classes); in some specific cases, also a linear function (simply a weighted sum of the neurons) could be useful.

The optimal combination of these hyperparameters can be obtained with an exhaustive grid search, which provides the best combination with respect to some metrics adopted to evaluate the accuracy of the network (for example the mean squared error, the mean absolute

error and the cross-entropy). Cross validation can be exploited to limit the overfitting problem. Actually, in neural networks this could be a problem more relevant than in different frameworks: thanks to the flexibility of the models and the high number of parameters to be set to the optimal value, it is quite easy to find a perfect fit on training data; in addition, due to the information leak phenomenon, it is also possible to have overfitted results on validation data; therefore it is essential to consider also a test set (with data not included neither in the training, nor in the validation) on which to run the final model to obtain a reliable measure of performance. Apart from cross-validation and training-validation-test approaches, specific techniques have been developed to deal with overfittig problems in neural networks; these methods involve adding some penalizing term in the loss function (the $L1$ and $L2$ regularizers) and/or to exclude at random some output features during the training phase (dropout). Certainly, additional hyperparameters should be considered in this case.
Furthermore, it is noteworthy to highlight that it is possible to include more than one output in a neural network to obtain jointly predictions. In particular, multiple outcomes can share all the parameters in the hidden layers, can depend on the same inputs but in a different way or can depend on different inputs; the convergence strategy to calculate the weights and biases in the training process and the adjustment of an output over the others can be irregularly favoured. It would be desirable to have high mutual correlation among outputs (because uncorrelated outcomes increase network complexity) and low correlation among input data. Of course, in these multi-outputs networks the training process is more difficult and there is an increased risk of overfitting.

7.1.2 Materials and methods

Data used in the analyses have been collected in a randomized, double blind, placebo-controlled, multicenter phase III clinical trial. Actually, the following two datasets have been considered:

- Dataset 1: 573 subjects, 184 clinical variables (including only baseline measurements);
- Dataset 2: 499 subjects, 450 variables (about 260 variables on proteins information).

As outcomes variables I focused on overall survival (OS), time-to-progression (TTP) and Best Percentage Change from Baseline (TRRBPCH). As usual, OS is the time from the first visit to death (i.e. without considering censoring information it is the minimum number of days a subject stays alive), TTP is the number of days before the disease progression (evaluated in term of tumour growth), while TRRBPCH is the minimum percentage change

in tumour volume from baseline measurements.
Different models have been fitted on this data: first of all the traditional Cox models, considered as the reference; then survival and regression random forests, also in the multivariate version including two outcomes and, lastly, deep learning architectures, using single output and multi-outputs feed-forward neural networks. Actually, Cox models involved only 8 variables, selected based on physician indications and the medical knowledge of the disease, while random forests and neural networks included all the predictors. For robustness of results all estimations have been repeated 100 times. The different methodologies have been compared using the coefficient of determination R^2, i.e. the percentage of variance explained by the model.
Some technical pre-processing was required to deal with missing values and to transform input data in the more convenient format for neural networks architectures. First of all, a min-max transformation $\tilde{\boldsymbol{x}} = \frac{\boldsymbol{x} - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}$ was applied: all continuous values were rescaled between 0 and 1 to make the predictors comparable. In this way the associated cost function becomes symmetric and the gradient descent algorithm used to find the global minima converges more quickly. Then variables with more than 25% of missing values were deleted and a random forests algorithm was adopted to impute the remaining missing values, assuming that they were missed at random. After having evaluated different possibilities, the twenty-five percent cut-off was selected *ad hoc*, taking into account the number of variables discarded and the additional information that could be provided by variables still containing missings.
Lastly, for neural networks models it has been necessary to convert all categorical variables with more than two levels using the One-Hot-Encoding transformation which creates a number of dummy variables, i.e. variables that take the value 0 or 1 to indicate the absence or presence of some categorical effect.
With respect to softwares, R [75] and Python [102] have been used. In particular, R has been adopted for traditional Cox models (`survival` package), missing imputation (`missForest` package) and random forests (`ranger` and `randomForestSRC` packages), whereas Python with the libraries `Keras`, `TensorFlow`, `scikit-learn` has been exploited for building the neural networks.

7.1.3 Results

Results obtained considering TTP as outcome are totally comparable to those for OS. Therefore only the latter are presented here (unless differently specified). As expected, including all the available variables the performance of models increases a lot: this is already evident

Table 7.1 Comparison of Cox models and regression RFs and NNs in term of % of variance explained R^2 (with its 1st and 3rd quantile)

Method	$R^2(\%)$	$[Q1 - Q3]$
Cox model	15.9	13.5 – 17.9
Regression RFs	28.5	28.3 – 28.7
Neural Networks	33.4	31.8 – 35.5

with random survival forests and even more impressive for the second dataset (the one with protein information) as showed by Figure 7.2, panel A. In addition, boxplots for RF are tighter than Cox one, because traditional models rely more heavily on the partition of data used for the estimation. Switching to a regressive approach and considering neural networks (Figure 7.2, Panel B), the deep learning algorithm still increases the percentage of variance explained (and could be seen as a small improvement also with respect to RF), as shown in Table 7.1. The simultaneous analysis of an additional output (TRRBPCH)

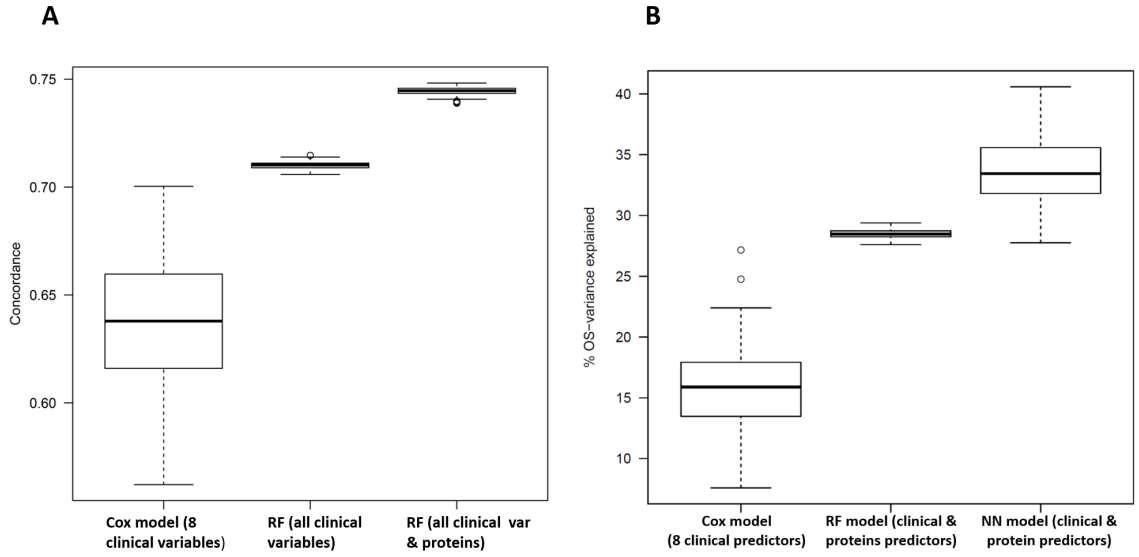


Figure 7.2 Comparison of Cox model, including 8 clinical variables, with machine learning and deep learning methods. Panel A. Cox *versus* RF (with and without proteins information). Model performances evaluated in term of c-index. Panel B. Cox *versus* RFs and NN (including proteins information). Model performance evaluated in term of % of variance explained.

allows explaining an additional 2% of the overall survival related variance in neural networks, whereas multivariate regression RF improves the performance of single output RF of about 8% (Figure 7.3).

Actually, a large part of the work done in this project concerned the tuning of hyperpara-

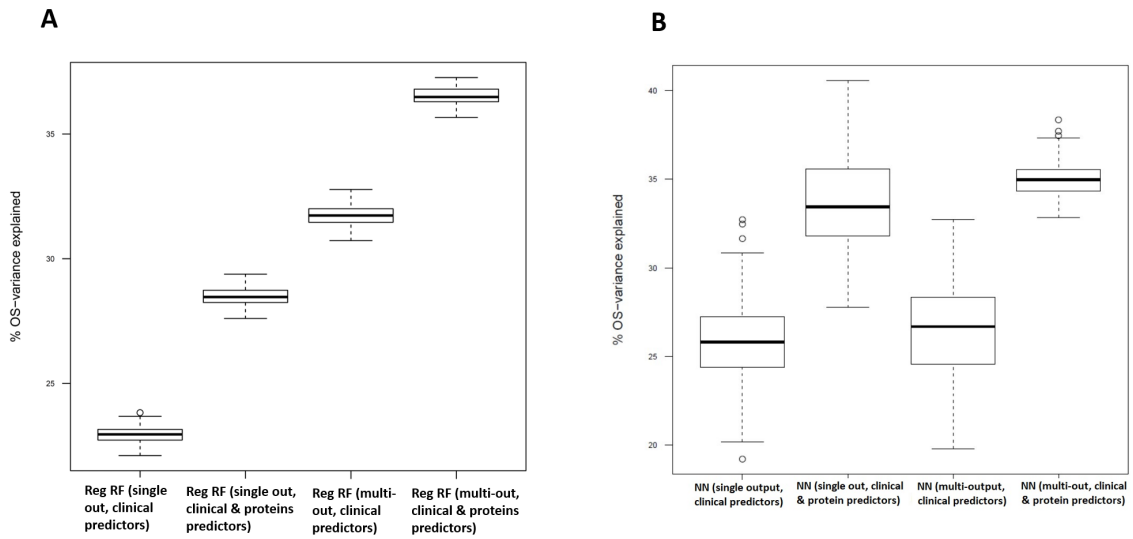


Figure 7.3 Comparison of single and multi-outputs models in term of % of variance explained. Panel A. Regression RFs with and without proteins information. Panel B. Neural Networks, with and without proteins information.

meters for RFs and NNs. Just as an example, the R code to implement the grid search in `ranger` (RANdom forest GEneRator package), as well as the analogue for NNs in Python are reported in Appendix B.

7.1.4 Some considerations

This study was the first attempt to look in depth into all clinical data routinely collected in a trial to understand if there was valuable unused information; different approaches, including traditional methods as well as AI models, have been adopted and relevant results have been compared. Models seem easily applicable to other databases (given in the appropriate format). However, at present, only one clinical database has been analysed; results appears slightly overoptimistic: a residual amount of overfitting could not be excluded and reproducibility should be further discussed. In addition, simpler neural networks including only two hidden layers should be investigated: from preliminary analysis (see Figure 7.4) it seems that the percentage of variance explained by the two models is comparable (and even a little higher for the 2-layers network), whereas the number of trainable parameters to be considered reduces of more than two thirds (with a important reduction of time and resources needed to run the model).
A better tuning of hyperparameters in the multi-outputs neural networks framework, predic-

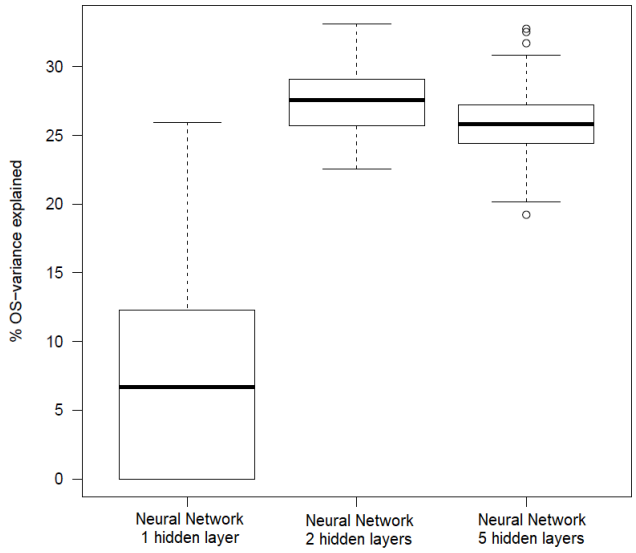


Figure 7.4 Neural Networks models. % variance explained by models with a different number of hidden layers.

tions of the clinical outcome of interest for each subject in a personalized precision medicine perspective and indications of predictors usefulness based on variables importance measures constitute future developments of this work. In addition, in multi-outputs framework, simulations should be evaluated to further investigate how correlation among input and output features affects the network learning.
All in all, this proof of concept study shows that valuable information related to subject clinical outcome is present in baseline data, but not currently used in traditional analyses. It appears evident that AI-based models (random forests and neural networks) allow to better explain OS related variance (using all the available variables); the competitive advantage in term of performance and exploitation of collected data compensates the higher complexity and longer time needed to obtain predictions.

7.2 Trends in incidence of thick, thin and *in situ* melanoma in Europe

The incidence of cutaneous malignant melanoma has increased steeply since the Second World War [103–109] although recent observations show a stabilization of rates in North America and Australia [104]. In these countries prevention campaigns and interventions

aimed at early diagnosis have been implemented, with the aims of limiting solar exposure (especially at younger ages) and of detecting suspicious lesions as early as possible. However, a clear decrease in melanoma mortality in all age groups [110–113] has not yet been observed. Some countries have recently noted evidence of decrease in the incidence of invasive melanoma. However the reasons for the decline are not clear from the data, and may be due to sun prevention campaigns or earlier diagnosis or both. An important element in understanding the trends is the concurrent behaviour of *in situ*, thin and thick lesions. Some studies have shown that the incidence of *in situ* and thin melanomas has been increasing faster than that of thick melanomas [110, 114–121].
At present, information on the incidence of *in situ*, thin and thick lesions is collected by many cancer registries in Europe, but statistics on thickness are not available on a routine basis from most of the public access sites. In addition, published studies often focus on the national burden of the disease, lacking a broader European perspective.
In this work the melanoma incidence and mortality in 13 European countries during the period $1995 - 2012$ have been investigated. To better understand trends, analyses were performed by country, age, sex and Breslow thickness.

7.2.1 Materials and Methods

Individual anonymized incidence data (with corresponding population files) were collected from population based European cancer registries (CRs) through an *ad hoc* call. CRs directors and researchers received the study protocol by e-mail by the end of year 2015. CRs were selected on the basis of a long history of high quality data (as proven by inclusion in the last two editions of Cancer Incidence in Five Continents, low percentages of DCO and high proportions of MV for melanoma); at least ten years of complete registration; a sufficiently large population to avoid large year-to-year variation; and availability of information on thickness. Registries in Cluj, Geneva, Iceland, and Ragusa were also included to ensure representation of a wide range of geographical areas.
To minimize the workload of CRs, data were accepted in any format available and for any period which CRs could provide. This helped ensure a broader perspective and reflected the heterogeneity of European cancer registration activity. However, after data cleaning and the pre-processing of all CRs data files, the analysis focused on cases diagnosed between 1995 and 2012.
A common anonymized database was created containing the following variables for each case: age at diagnosis, sex, year of diagnosis, histological type, tumour location, behaviour (invasive, *in situ*), Breslow thickness, vital status (dead or alive) and cause of death if

melanoma. In the analyses, histological types were categorized as "superficial spreading melanoma" (SSM), "nodular melanoma" (NM), "lentigo maligna melanoma" (LM) and "other"; while cancer body sites were grouped as "head and neck" (HN), "limbs", "trunk" and "other".

Mortality data were retrieved from the World Health Organization (WHO) mortality database [122], to allow a broader comparison among countries.

Annual incidence and mortality rates were age - standardized on the World population (following the Segi standard).

In situ incidence trends were analysed taking care of specific problems highlighted by registries (Cluj CR did not collect this type of lesions and was excluded; Norway CR was included only for the period $1995-2008$, because since 2009 it registered only *in situ* cases that later become malignant).

Invasive lesions were grouped by thickness, as: "thin" for those ≤ 1.00 mm (as defined by TNM since the 6th Edition) and otherwise as "thick". Cases without Breslow thickness information were analysed separately as "unknown". Moreover, in order to minimize bias [123], where a high proportion of cases had missing thickness, incidence trends by Breslow level were also analysed after the imputation of thickness information assuming that it was missed at random. The number of unknown cases n_{unk} was taken as being made up of $n^u_{\text{thin}} + n^u_{\text{thick}}$ (the number of thin and thick lesions). The ratio $\frac{n^u_{\text{thin}}}{n^u_{\text{thick}}}$ was taken to be equal to the known proportion of thin and thick lesions for the same period, sex and age group. Thin/thick ratios were calculated for each age, sex, and period stratum, so the correction was specific to each combination of stratification variables. In particular, the estimated number of thin and thick cases was obtained solving the following system of equations:

$$\begin{cases} \text{Unknown}_{ijk} & = \text{Thin}^U_{ijk} + \text{Thick}^U_{ijk} \\ \frac{\text{Thin}^U_{ijk}}{\text{Thick}^U_{ijk}} & = \frac{\text{Thin}_{ijk}}{\text{Thick}_{ijk}} \end{cases}$$

where i, j, k gave account to the stratification by sex ($i = 1,2$), period ($j = 1,\ldots,18$) and age group ($k = 1,\ldots,18$).

Therefore:

$$\begin{cases} \text{Thick}^U_{ijk} & = \frac{\text{Unknown}_{ijk}}{1+\frac{\text{Thin}_{ijk}}{\text{Thick}_{ijk}}} \\ \text{Thin}^U_{ijk} & = \frac{\text{Thin}_{ijk}}{\text{Thick}_{ijk}} * \text{Thick}^U_{ijk} \end{cases}$$

and

$$\begin{cases} \text{Thin}^F_{ijk} & = \text{Thin}_{ijk} + \text{Thin}^U_{ijk} \\ \text{Thick}^F_{ijk} & = \text{Thick}_{ijk} + \text{Thick}^U_{ijk} \end{cases}$$

where Thin^F_{ijk} nd Thick^F_{ijk} represented the corrected number of thin and thick cases respectively. In this way, a corrected number of thin and thick cases was obtained by adding the calculated components to the original ones, by sex, period, and age group. The correction was done for all cases combined and also on a registry by registry basis: the results were similar (with some more instability due to small numbers, when stratified by registry) and only trends corrected for all cases combined are reported here.
Joinpoint regression models[1] (assuming log transformation of data, homoscedasticity in error variance, and setting other parameters as default) have been used to determine incidence trends and the timing of any changes. From these analyses, the Annual Percent Change (APC) and the Average Annual Percent Change (AAPC) have been reported. Funnel plots have been included to show the precision of single registry estimates.
All analyses were performed using SAS/STAT Version 9.2, R Version 3.3.2 and Joinpoint Version 4.2.02.

7.2.2 Results

The database covered a population of over 117 million inhabitants and included about $415,000$ skin lesions, recorded by 18 European CRs (7 of them with national coverage) in 13 countries, between 1995 and 2012 (Table 7.2). Incidence rates of invasive melanoma in men varied from a low of 5.6 per $100,000$ in Tarragona, Spain to $24/100,000$ in Geneva, Switzerland, where the highest rate of *in situ* lesions was also observed, $23/100,000$ in men in 2012 (Table 7.2). Incidence rates were higher among females in nearly all cancer registries.
The median age at diagnosis was 61 years in men and 56 in women. *In situ* lesions made up about one-quarter of all cases. SSM was the commonest invasive histological subtype (46%) and about 50% of invasive lesions were located on the limbs. However, the body site distribution of lesions was quite different between men and women, with 43% of melanoma on the trunk in men and 57% on the limbs in women. Overall, time trends showed similar patterns for both sexes and are reported together in Figures 7.5 and 7.6. After the exclusion of Austria, Scotland for the period $1995-2005$ and Norway for the period $1995-2008$, due to the lack of thickness information for all cases recorded, thin and thick melanomas were present in our database in similar proportions (38% and 34% respectively); 28% of cases had missing information on thickness, with the highest proportion of missing information in

[1] Joinpoint Regression Program; Statistical Methodology and Applications Branch, Surveillance Research Program, National Cancer Institute. Version 4.2.02

earlier years.

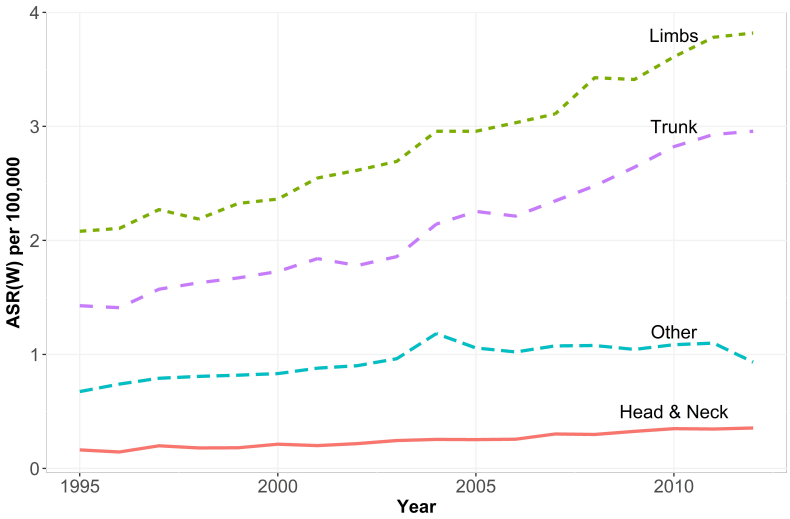


Figure 7.5 Cutaneous malignant melanoma incidence by body site. Both sexes, 1995 - 2012. World age standardized incidence trends.

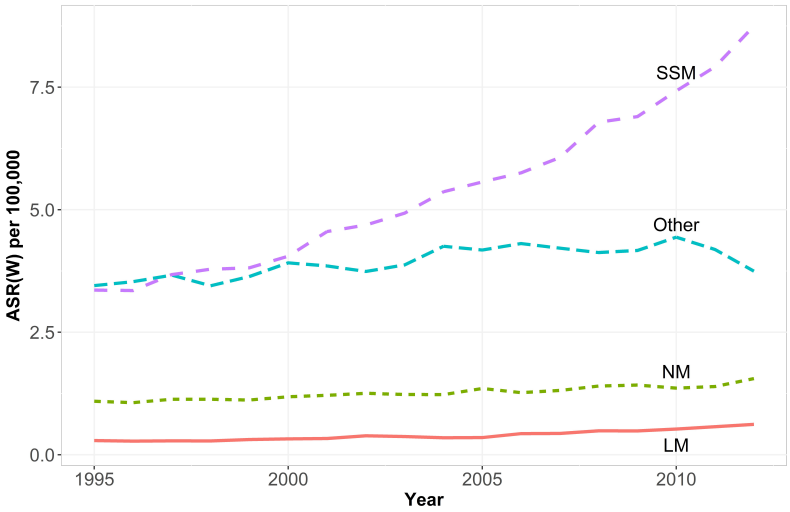


Figure 7.6 Cutaneous malignant melanoma incidence by histological type. Both sexes, 1995 - 2012. World age standardized incidence trends.

Table 7.2 National (N) and regional (R) cancer registries which contributed to the project.

Cancer Registry	**Type (% of coverage)**	**Population** (last year)	**Total number of cases**	**% unknown Breslow** (2008 - 2012)	**CMM Invasive ASRW**(last year)		**CMM *in situ* ASRW**(last year)	
					Men	**Women**	**Men**	**Women**
Austria	N (100)	8388534	25800	-[†]	11.4	11.2	5.6	6.0
Belgium	N (100)	11099554	21664	21%	11.6	18.3	5.1	7.3
Cluj (RO)	R (3.4)	694136	336	20%	6.4	8.4	/	/
England (UK)	R (83.3)	53865817	185259	27%	13.4	14.3	5.2	5.9
Geneva (CH)	R (5.8)	472370	3154	1.5%	23.6	20.0	22.6	20.8
Granada (ES)	R (2.0)	922127	1750	20%	8.0	8.2	2.0	3.6
Iceland	N (100)	327386	1177	3.9%	7.6	13.3	3.8	6.8
Ireland	N (100)	4592900	16092	14%	14.8	15.1	8.6	8.7
Murcia (ES)	R (3.1)	1454250	1984	22%	6.8	9.3	2.2	2.7
Netherlands	N (100)	16865020	74581	3.7%	18.5	21.3	5.6	7.4
Northern Ireland (UK)	R (2.8)	1840498	6282	$< 1\%$	10.8	14.3	7.0	6.3
Norway[‡]	N (100)	5051275	26693	4.2%	20.8	23.0	5.4	7.6
Ragusa (IT)	R (0.5)	307697	568	17%	7.9	6.9	1.3	1.6
Schleswig - Holstein (DE)	R (3.5)	2815955	14289	25%	14.6	13.9	5.7	6.6
Scotland (UK)	R (8.2)	5327700	22548	5.3%	12.8	12.8	6.3	6.5
Slovenia	N (100)	2056262	7662	5.9%	15.8	15.6	6.6	7.7
Terragona (ES)	R (1.7)	806115	1429	34%	5.6	6.4	1.9	1.5
Torino (IT)	R (1.5)	911823	3635	13%	10.3	10.7	8.1	9.2
TOTAL		117799419	414903					

[†]Austria did not provide information on Breslow

[‡]*In situ* collected only till 2008

Incidence trends. The combined data for all participating CRs showed a statistically significant incidence increase in both invasive (AAPC 4.0%, 95% CI: 3.8 – 4.2 in men; AAPC 3.0%, 95% CI: 2.3 – 3.8 in women) and *in situ* (AAPC 7.7%, 95% CI: 7.3 – 8.1 in men; AAPC 6.2%, CI: 5.9 – 6.6 in women) lesions for both men and women, during the 1995 – 2012 period (Table 7.3 and Figure 7.7). The trend in invasive cases seemed mainly

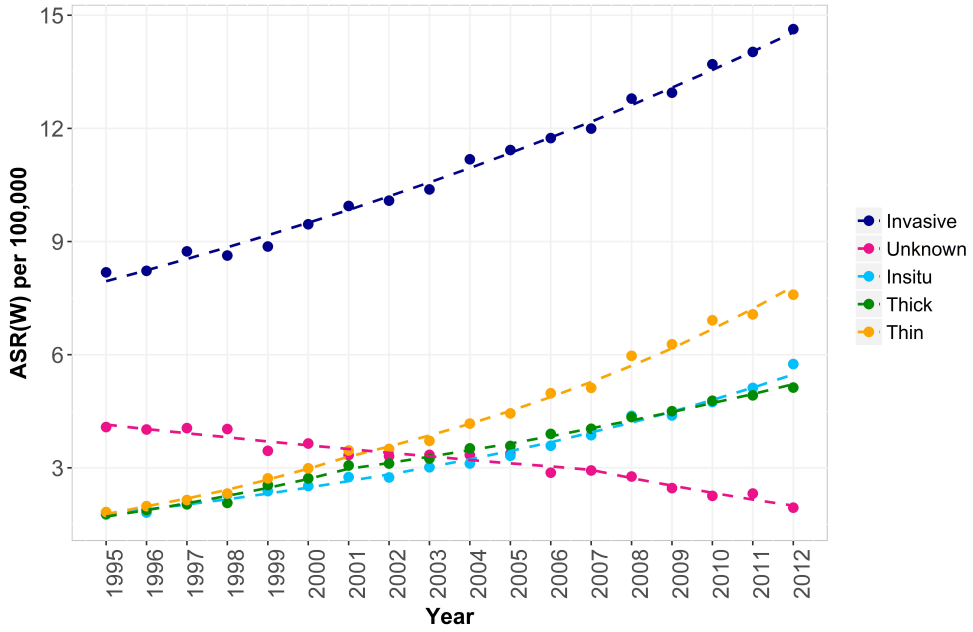


Figure 7.7 Cutaneous malignant melanoma incidence. Europe, 1995 - 2012. Both sexes. World age standardized incidence trends. Dots represent the observed values; dashed lines the Joinpoint models.

due to the increasing incidence of thin lesions, with an average annual increase of 8.3% (95% CI: 7.6 – 9.1) in women and an even greater increase (10%, 95% CI: 9.6 – 11.3) in men. The incidence of thick melanomas also increased; however in recent years the rate of increase was less than for thin melanomas, with a reduction in APC from 12% to 4.3% in women and from 10% to 6.0% in men. A large part of the increase in incidence was for lesions on the limbs and trunk; lesions of head and neck also increased in incidence, although less rapidly (Figure 7.5). SSM, the most prevalent histological sub-type, showed the largest increase, while NM and LM increased more gradually (Figure 7.6).

The steepest increase in incidence occurred in the period 1998 and 2001, in particular in thick melanomas in women. This finding was confirmed after correction for the number of melanomas with unknown thickness (APC 0.6% before 1998, 5.4% between 1998 and 2001, and 1.7% after 2001). The increase in thin melanomas lasted until 2001 (APC 11%) and slowed after that year (APC 8.2%, Table 7.3). However, these changes in trend emerged

Table 7.3 Annual Percent Change (APC) highlighted by the Joinpoint models in melanoma incidence trends in Europe in the period $1995-2012$. Analysis performed before imputing the missing information on Breslow level.

Type	**Sex**	**Joinpoint**	**APC**	**Lower CI** (95%)	**Upper CI** (95%)
Invasive	Men	No JP	4.0 *	3.8	4.2
	Women	2 JP at 1998 and 2001	0.4	-1.7	2.4
			4.9*	0.7	9.3
			3.3*	3.0	3.6
	Both sexes	No JP	3.6*	3.4	3.8
In situ	Men	No JP	7.7 *	7.3	8.1
	Women		6.2 *	5.9	6.6
	Both sexes		6.8*	6.5	7.2
Thin	Men	1 JP at 2000	12.7*	10.1	15.5
			9.5*	8.8	10.2
	Women	1 JP at 2001	10.1*	8.1	12.1
			7.4*	6.6	8.2
	Both sexes	1 JP at 2001	10.8*	9.2	12.4
			8.2*	7.5	8.8
Thick	Men	1 JP at 2001	10.0*	8.2	11.9
			6.0*	5.2	6.7
	Women	2 JP at 1998 and 2001	5.2*	2.5	8.1
			12.5*	6.8	18.6
			4.3*	3.9	4.7
	Both sexes	1 JP at 2001	9.7*	8.0	11.4
			5.2*	4.6	5.9
Unknown	Men	1 JP at 2005	-1.8*	-3.0	-0.5
			-5.7*	-7.7	-3.6
	Women	1 JP at 2007	-3.2*	-4.0	-2.3
			-8.1*	-11.1	-4.9
	Both sexes	1 JP at 2007	-2.8*	-3.6	-1.9
			-7.5*	-10.5	-4.4

*The APC is significantly different from zero at $\alpha = 0.05$

for thin melanomas only before the correction for those of unknown thickness. After the imputation of missing information, thin melanoma increased at a steady (and lower) rate for the entire period (Table 7.4). The completeness of information collected by CRs increased

Table 7.4 Annual Percent Change (APC) highlighted by the Joinpoint models in melanoma incidence trends in Europe in the period $1995-2012$. Analyses performed after the correction for cases with unknown thickness.

Type	**Sex**	**Joinpoint**	**APC**	**Lower CI** (95%)	**Upper CI** (95%)
Thin	Men	No JP	6.3*	6.0	6.6
	Women		4.6*	4.4	4.9
	Both sexes		5.3*	5.0	5.5
Thick	Men	No JP	3.2*	3.0	3.4
	Women	2 JP at 1998 and 2001	0.6	-1.2	2.5
			5.4*	1.6	9.3
			1.7*	1.5	2.0
	Both Sexes	1 JP at 2004	3.3*	2.8	3.7
			2.2*	1.6	2.7

*The APC is significantly different from zero at $\alpha = 0.05$

overtime, most clearly after $2005/2007$ (cases with unknown thickness decreased to -5.7% after 2005 in men and to -8.1% after 2007 in women, Table 7.3).

Correction for lesions of unknown thickness enhanced the differences between thin and thick lesions and flattened their trends: in particular, for thin melanoma, the APC in men was roughly halved (from 13% to 6.3%), while in women the average annual increase was reduced to 4.6% (Table 7.4 and Figure 7.8).

Incidence trends varied considerably among registries (Table 7.5). The average annual percent change (AAPC) for invasive lesions in women varied from -0.5% (Schleswig-Holstein, Germany) to 14% (Cluj, Romania). Incidence trends for *in situ* lesions in men varied from a decrease in AAPC of -2.9% (Ragusa, Italy) to an increase of 24% (Geneva, Switzerland). However, when we compared estimates of AAPC with their precision in a funnel plot we noted that the observed differences across registries were mainly due to random variation (Figure 7.9). Only the AAPC in men in Netherlands was outside the 3σ upper boundary of the funnel plot for invasive lesions. (Figure 7.9, Panel A).

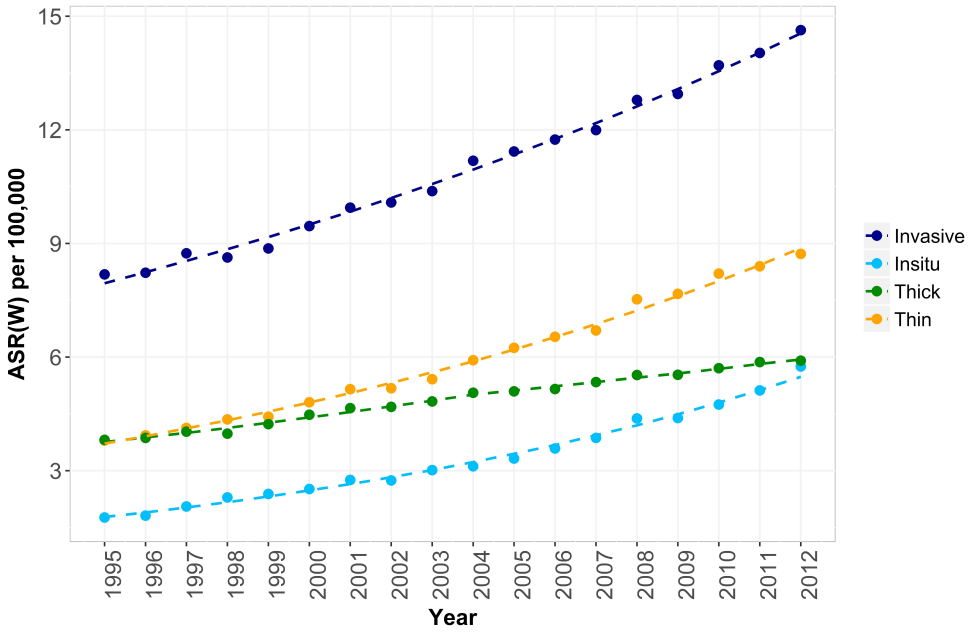


Figure 7.8 Cutaneous malignant melanoma incidence, after correction for unknown. Europe, 1995 - 2012. Both sexes. World age standardized incidence trends. Dots represent the observed values; dashed lines the Joinpoint models.

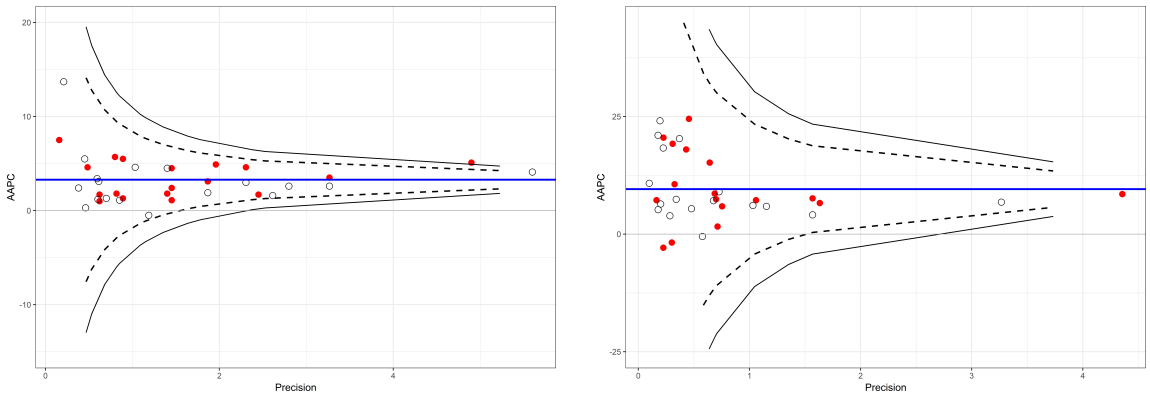
Table 7.5 Average Annual Percent Change (AAPC) for incidence of invasive and *in situ* lesions by single registry, in the period 1995 – 2012.

AAPC	**Men**		**Women**	
Country	**AAPC Invasive**	**AAPC in situ**	**AAPC Invasive**	**AAPC in situ**
Austria	1.7*	20.5*	1.6*	18.3*
Belgium§	5.5*	15.2*	4.5*	21.0*
Cluj (RO)§	7.5	/	13.7*	/
England (UK)	4.5*	8.5*	3.0*	6.8*
Geneva (CH)	1.8*	24.5*	3.1*	20.3*
Granada (ES)	5.7*	7.4*	3.4*	7.1*
Iceland	1.0	-1.8	0.3	6.4
Ireland	4.6*	7.6*	2.6*	4.1*
Murcia (ES)§	1.8	7.2	1.1	3.9
Netherlands	5.1*	6.6*	4.1*	5.9*
Northern Ireland (UK)	3.1*	7.2*	2.6*	6.1*
Norway†	2.4*	1.6	1.9*	-0.5
Ragusa (IT)	4.6*	-2.9	5.5*	5.2
Schleswig - Holstein (DE)§	1.7	5.9*	-0.5	5.4*
Scotland (UK)	3.5*	8.6*	1.3	9.0*
Slovenia	4.9*	18.0*	4.6*	10.8
Tarragona (ES)	1.3	10.6*	2.4	7.4*
Turin (IT)	1.1	19.2*	1.2	24.1*
All	**4.0***	**7.7***	**3.0***	**6.2***

*The AAPC is significantly different from zero at $\alpha = 0.05$

§Shorter period of observation. Belgium 2004 – 2012, Cluj 2006 – 2011, Murcia 1996 – 2009, Schleswig - Holstein 1998 – 2012, Tarragona 1995 – 2011.

†*In situ* collected only till 2008



(a) Invasive (b) *In situ*

Figure 7.9 Average Annual Percent Change (AAPC) by different registries for invasive and *in situ* lesions for both men (solid points) and women (empty circles). The horizontal blue line represent the average AAPC for all the registries together; continuous black lines the $\pm 3\sigma$ bounds; dotted black lines the $\pm 2\sigma$ bounds.

Mortality trends. Melanoma world age - standardized mortality rates ranged from 1 to 4 cases per $100,000$ and the increase in mortality, where present, was much smaller than that in incidence. Rates continue to increase in Northern Europe, in particular in Iceland (AAPC 5.4%), Ireland (AAPC 3.4%) and Norway (AAPC 1.6%). In Western Europe, Netherlands (AAPC 2.1%) and Belgium (AAPC 1.0%) showed a statistically significant increasing trend; in Southern and Eastern Europe only Slovenia (AAPC 2.0%) presented a significant increase in melanoma mortality (Table 7.6 and Figure 7.10). Men exhibited higher rates than women, as survival is poorer among men [124, 125], but, in general, trends were parallel, with the exception of Ireland, UK, Germany and Slovenia, where men showed higher increases than women, in particular in the last period (Table 7.6).

Table 7.6 Average Annual Percent Change (AAPC) for skin cancers mortality in different European countries, in the period 1995 - 2012. Data retrieved from the WHO Mortality database.

Country	**Men**			**Women**			**Both sexes**		
	AAPC	**LCI**(95%)	**UCI**(95%)	**AAPC**	**LCI**(95%)	**UCI**(95%)	**AAPC**	**LCI**(95%)	**UCI**(95%)
Iceland	4.1	-2.2	10.9	6.3	-0.4	13.4	5.4*	0.1	10.9
Ireland	4.2*	2.6	5.7	2.2*	0.4	4.0	3.4*	2.2	4.6
Norway	1.6*	0.9	2.3	1.7*	0.3	3.0	1.6*	1.0	2.2
UK	1.3*	0.9	1.6	-0.1	-0.5	0.3	0.7*	0.4	1.0
Austria	-0.3	-1.5	1.0	-0.9	-2.7	1.0	-0.6	-1.7	0.6
Belgium	1.6*	0.7	2.5	0.4	-0.5	1.3	1.0*	0.3	1.8
Germany	1.0*	0.6	1.5	0.1	-0.6	0.8	0.6	-0.2	1.3
Netherlands	2.2*	1.8	2.6	1.9*	1.2	2.7	2.1*	1.8	2.5
Switzerland	0.0	-0.9	1.0	-0.2	-0.8	0.3	0.0	-0.7	0.6
Italy	0.5	-0.4	1.5	-0.1	-1.8	1.7	0.1	-0.7	0.8
Romania	0.6	-1.0	2.3	-1.1*	-1.7	-0.4	-0.1	-1.4	1.2
Slovenia	2.6*	1.0	4.1	1.5	-0.1	3.1	2.0*	1.0	3.0
Spain	-0.5	-1.5	0.5	-0.9*	-1.4	-0.4	-0.8*	-1.4	-0.1

*The AAPC is significantly different from zero at $\alpha = 0.05$

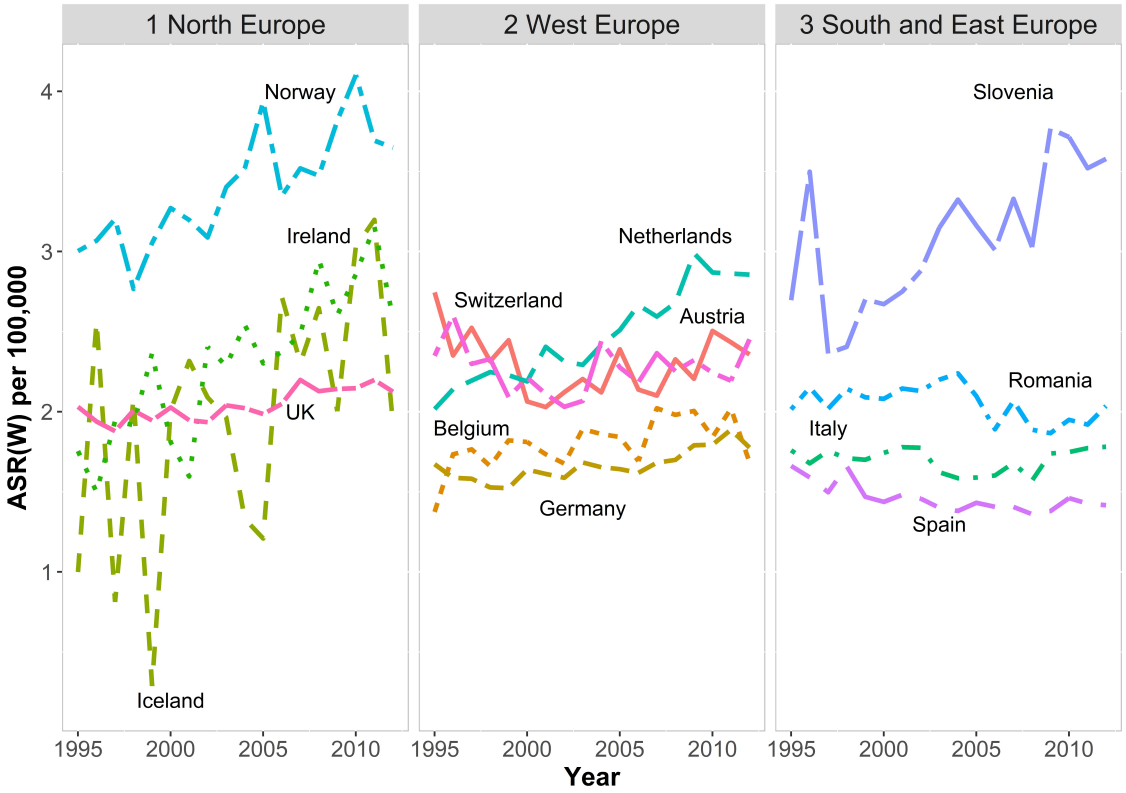


Figure 7.10 Skin cancer mortality, 1995 - 2012. Both sexes. World age standardized mortality trends for melanoma lesions in different European countries. Data retrieved from the WHO Mortality database.

7.2.3 Discussion

Melanoma incidence trends in Europe have been previously analyzed at a national level [105, 110, 111, 113, 116–121, 126–136]. However, the different periods considered and methods used have hampered comparisons between countries. Moreover, differences in incidence and mortality trends in both time and space have made it difficult to interpret mortality-incidence ratios, which are often used as a proxy for survival, and are also used to estimate incidence in areas not covered by registration [137].
In this study, a broad overview of the continent is provided, using data collected by CRs. The analysis has been restricted to a short period ($1995 - 2012$) in order to accommodate the largest number of registries (but there are CRs with data available for a much longer period) and to avoid an overestimation of increase in trends (especially for the *in situ* lesions), simply due to the improvement in the completeness of melanoma cases registered.
A major limitation of this study was that several CRs have registered *in situ* melanomas only in recent years, and have limited information on the thickness of lesions.
Even CRs with a long history of data collection lacked information on Breslow thickness for a large proportion of cases, especially in earlier years. Missing information on thickness, particularly where the proportion missing varied during the period analyzed, could have led to significant bias. The same considerations applied to *in situ* lesions, which were collected and recorded by CRs with greater accuracy and completeness in more recent years. However, the lack of changes in the slope of the increasing incidence trend of *in situ* cases and the attention paid to problems highlighted by different CRs (which led to the exclusion of Cluj CR and (partly) of Norway CR for this analysis) made me confident of the reality of the observed increase.
With respect to invasive cases, the bias has been reduced by re-classifying lesions of unknown thickness as either thin or thick, according to the observed ratio of thin and thick lesions by sex, age and period. Although this method may help to adjust estimates towards the true rate, some bias is unavoidable as the true thin/thick ratio has not been completely observed and it is difficult to know if the estimates have been biased upward or downward. Moreover, it could not be excluded that the correction induced a small bias due to the assumption of randomness in missing information: actually, the distribution of unknown cases with respect to histology, body site and follow up (data non shown) seemed more similar to that of thick lesions.
Within these limitations, there is a clear picture of a continuing upward trend in both invasive and *in situ* melanoma in Europe. As observed, in recent years, the burden of *in situ* lesions has increased in parallel with that of thick lesions. Taking *in situ* and thin melanoma together,

there is an annual burden of 20 new cases per 100,000 inhabitants in Europe, with considerable impact on patients and health care services. The increase in incidence was particularly evident in Norway and Iceland, although the difference between trends in invasive lesions and that of *in situ* melanoma was less striking than in other countries such as Austria, Slovenia, Switzerland (Geneva), Spain (Tarragona) and Italy (Turin).

In the period $1995-2012$, melanoma mortality increased slightly in Norway, Iceland, Ireland, Netherlands, Belgium and Slovenia and overall approached the AAPC increase in incidence of thick lesions. In the most recent years, a steeper increase has been observed among men in Ireland, UK, Germany and Slovenia. However some evidence of decline in mortality has been reported in younger cohorts in some countries [138, 139], and Autier and colleagues proposed a scenario of declining trends in the recent future, due to the aging of younger cohorts supposedly less exposed to UV during childhood [140]. Even if the hypothesis of a decreased UV exposure in younger cohorts will be confirmed, the finding of this work, with an increasing trend of thick lesions in incidence for all ages, are pointing towards a less optimistic scenario for melanoma mortality in the next years.

Previous publications from European countries have shown increasing trends in the incidence of invasive lesions, mainly due to increases in thinner melanomas, as well as a small continuing increase in mortality (with some evidence of stabilization in Switzerland [110, 130] and decreases in mortality in women in Denmark [120] and Northern Ireland [116] which are not statistically significant). Lower incidence and higher mortality rates have been reported in South-Eastern European countries compared to North Western countries [132]. The incidence of *in situ* cases has increased dramatically [117, 118, 120, 128, 131], often at a greater rate than more aggressive lesions.

Outside Europe, melanoma trends have been investigated in US and Australia [106, 107, 114, 115, 141, 142]. Coory and colleagues reported a steep increase in the incidence of *in situ* and thin melanomas in the period $1982-2002$, while Baade et al., looking at more recent years, reported that trends in thinner lesions have now reached a plateau in Australia. On the other hand, mortality trends seem to have been stable for many years, although a non-negligible fraction of thin cases appears to result in death [143]. In the US, SEER data showed increasing trends for invasive (APC = $+3.6\%$) and *in situ* lesions (APC = $+9.5\%$) [141], while mortality scarcely increased (by 0.4% between 1992 and 2004). Analysing SEER data [144] it is possible to notice a stabilization in the incidence of invasive cases in the last years, while in the European database there was a continuing increase, although with rates only half of those in the US. Continuing increases in the incidence of *in situ* lesions are seen in both SEER and Europe trends. Data from the Connecticut Tumor Registry have shown an increase in the proportion of thinner lesions in the most recent period [107].

Although SEER data on Breslow thickness are not completely reliable, due to inconsistencies in the coding of thickness [145], Shaikh and colleagues [142] applied a validated imputation method for missing Breslow information (on about the 13% of cases), and showed that the incidence increased across all thickness groups in the period $1989 - 2009$, most pronounced for thinner lesions.
In conclusion, this study has shown that the incidence of invasive cutaneous malignant melanoma increased sharply in Europe during the period $1995 - 2012$ in both sexes; this rise was mainly attributable to thinner lesions. *In situ* lesions have also increased in incidence, with a steeper trend than that for invasive cases; however, the incidence of thick lesions has also increased to a lesser extent.
Seemingly, earlier diagnosis could not counterbalance a generalized increase in melanoma risk as it has still been observed an increase in thick lesions and also a slight increase of melanoma mortality in six of the 13 European countries included in this study.
For these reasons it is possible to conclude that in Europe we cannot dismiss the need for further efforts in preventive actions for limiting exposure to environmental hazards, especially in childhood as suggest by the European Code Against Cancer [146], as a way for avoiding thick lesions occurrence. Furthermore, advances in research for better targeting earlier detection of aggressive melanoma lesions are much needed.

7.3 Skin melanoma deaths within 1 or 3 years from diagnosis in Europe

The incidence of malignant melanoma of skin continues to rise in Europe in recent years. The increasing trends have been mainly driven by in situ and thin lesions in both sexes; mortality has also increased in some areas, although at a much lower pace and not homogeneously, as highlighted in previous paragraphs and reported in [97]. Although the thickness of the lesion is highly correlated with survival [147–149], the death toll from thinner melanoma is not negligible [142, 143, 150–152]. However, international, data describing trends in lethal melanoma by thickness has rarely been reported by population-based cancer registries (CR) [150] and, to my knowledge, no such study has previously been carried out in Europe.
Considering the substantial (and growing) public health burden of thin melanomas and the difficulty of disentangling the effects of early diagnosis or misclassification from a true increase in risk [153], it appears important to focus attention only on those melanomas which proved to be aggressive and consequently fatal (at one and three years from diagnosis), and to analyse their trends according to principal characteristics of the lesions.

7.3.1 Material and Methods

European CRs initially provided data for a previously published melanoma trends analysis [97]. All these CRs were subsequently requested to give further information on invasive melanoma cases (subject vital status, complete follow-up date and cause of death) to allow more detailed analyses on fatal melanoma cases.

To limit possible difficulties due to the new privacy legislation which entered into force in Europe in May 2018 [154], CRs were allowed to send anonymous data aggregated by the following design variables: sex, year of diagnosis, five-year age group, tumour thickness (≤ 1 mm: thin; > 1 mm: thick), death within one year, death within three years, death after more than three years.

Unlike the classic definition, here a fatal case is defined as a patient with an incident melanoma diagnosed between 1995 and 2012 and who died within 3 years after diagnosis even if the cause of death was different from melanoma. The time span is limited to reduce the bias due to death from other causes (information on causes of death was available for too few registries and could not be used). The analysis of fatal cases offers an opportunity to avoid the censorship bias common in traditional survival analysis. In addition, subjects older than 74 years of age have been excluded, in order to reduce potential differential bias, as the number of deaths not due to melanoma becomes more common in elderly people.

Descriptive statistics and proportions of fatal cases by different variables (period of diagnosis ($1995-2000$, $2001-2006$ and $2007-2012$), sex, tumour thickness, histological type of lesion (superficial spreading melanoma, nodular melanoma, lentigo maligna), tumour site (head and neck, limbs, trunk) and cancer registry) have been calculated to provide some general information on collected data. Trends in fatal melanoma have been analysed using a multivariate generalized linear mixed effects model, in which the frequency of fatal cases within one or three years is assumed to have a Poisson distribution with the corresponding population as offset. Variation between registries has been taken into account including random intercepts and random slopes in the model. The change in fatal cases incidence across the defined periods of diagnosis is the main variable studied (with the first period as reference), while other variables have been treated as confounding/modifying factors. In addition, higher order interactions between variables have been explored and introduced in the final model, if relevant.

The Belgian Cancer Registry started registration activity in 2004, and therefore could not provide data for the first period of analysis. Its figures were then kept separate in the global model.

Slovenia sent additional data for the 5-year period $2013-2017$ (with follow up at the end of

Table 7.7 Populations, number of invasive melanoma, and number and proportion (%) of fatal cases at 1 and 3 years since diagnosis by participating registry and period (all ages included)

Cancer Registry (Abbreviations)	**Period of diagnosis**	**Population***	**Number of invasive cases†**	**Fatal 1 year**		**Fatal 3 year**	
				N	%	N	%
Austria	1995 - 2011	7675506	15683	1566	10	2565	16
Belgium	2004 - 2012	9875432	13640	417	3	1162	9
Geneva	1995 - 2012	424131	1860	24	1	84	5
Iceland	1995 - 2012	299595	697	14	2	43	6
Ireland	1995 - 2012	4297337	8237	332	4	747	9
Norway	1995 - 2012	4476586	16608	538	3	1620	10
Netherlands (Nether.)	1995 - 2012	15425267	49769	1215	2	4035	8
Ragusa	1995 - 2012	281184	364	17	5	51	14
Schleswig-Holstein (Schl.Hol.)	1998 - 2012	2568498	8017	538	7	854	11
Slovenia	1995 - 2012	1887840	5246	336	6	833	16
Tarragona (Tarraco)	1995 - 2011	728671	998	60	6	112	11
Turin	1995 - 2012	800358	2241	76	3	224	10
TOTAL	1995 - 2012	48740404	123360	5133	4	12330	10

*Yearly average population $0-99$ years for the period $2007-2012$
†$0-74$ years of age

year 2018). This new data have been analysed separately with a generalized linear Poisson model (including the same fixed effects as for the previous model) to check the persistence of risk observed in 1995 - 2012.
All analyses have been performed with R Version 3.6.0 [75].

7.3.2 Results

Twelve European CRs provided valid data for the analysis, with follow-up information on life status at 1 and 3 years after diagnosis (Table 7.7). Overall, 123360 invasive cases were collected between $1995-2012$, in the population from 0 to 74 years of age. The largest number of cases was from the Netherlands, with almost 50000 cases diagnosed between 1995 and 2012, while the smallest was Ragusa (Italy) with 364 cases diagnosed during the same period. Approximately four percent of invasive cases died within one year, while 10% had the same fate within 3 years from diagnosis.
The number of fatal cases showed an overall crude decreasing trend (Table 7.8) at both 1

Table 7.8 Number and proportion (%) of fatal cases at 1 and 3 years since diagnosis by sex, age, period, tumour thickness, histological type and body site

		N	**Fatal 1 year**		**Fatal 3 years**	
Sex	Male	56035	3290	5.9%	7860	14.0%
	Female	67325	1843	2.7%	4470	6.6%
Age group	0-44 yrs	37374	694	1.9%	1830	4.9%
	45-64 yrs	57200	2343	4.1%	5610	9.8%
	65-74 yrs	28786	2096	7.3%	4890	17.0%
Period	1995 - 2000	27121	1447	5.3%	3295	12.1%
	2001 - 2006	40030	1781	4.4%	4225	10.6%
	2007 - 2012	56209	1905	3.4%	4810	8.6%
Thickness	≤ 1 mm	23663	289	1.2%	759	3.2%
	> 1 mm	65231	2147	3.3%	6439	9.9%
	Unknown	34466	2697	7.8%	5132	14.9%
Histological Type	Superficial Spreading	65027	812	1.2%	3103	4.8%
	Nodular	14087	857	6.1%	2807	19.9%
	Lentigo maligna	3726	53	1.4%	177	4.8%
	Unspecified	40520	3411	8.4%	6243	15.4%
Body Site	Head & Neck	13861	560	4.0%	1599	11.5%
	Limbs	56833	1192	2.1%	3833	6.7%
	Trunk	7532	1926	25.6%	2593	34.4%
	Unspecified	45134	1455	3.2%	4305	9.5%

and 3 years. As expected, men and older patients showed a greater early lethality. Thick lesions (> 1 mm), nodular melanoma and melanoma on the trunk had a higher percentage of fatal cases, at both 1 and 3 years from diagnosis. More poorly documented cases, lacking information on tumour thickness showed a larger proportion of fatal cases (Table 7.8).

When fatal cases by period were analysed in a multivariate Poisson generalized linear model with mixed effects (random intercept for registries and random slope for period), including sex, age, tumour thickness, histological type and body site, an overall decreasing trend with an estimated 16% decrease in risk of fatal case at 1 year and 8% at 3 years between the first ($1995-2000$) and the last ($2007-2012$) period were found. However, the mixed-effect model showed a remarkable variability within European countries; not only the base rate, but also the direction of changes, was different.

Differences in lethality between periods and among registries can be seen in Figure 7.11 and Figure 7.12, where each estimate is shown with its 95% limits and the black dot represents the reference period ($1995-2000$; $2004-2006$ for Belgium).

The majority of registries showed a decreasing trends in fatal cases at both one and three years, especially in the most recent periods. This was seen in particular in Iceland, Turin (Italy),

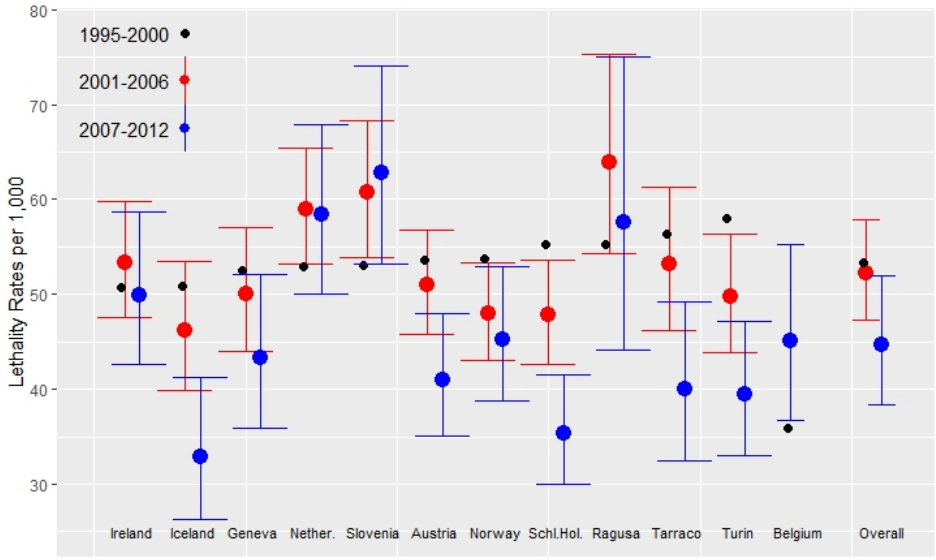


Figure 7.11 Fatal invasive skin melanomas at 1 year. Lethality rates per 1000 cases and 95% confidence limits for period 2001 – 2006 (red) and 2007 – 2012 (blue) vs 1995 – 2000 (reference period represented by a black dot; 2004 – 2006 for Belgium), by cancer registry. Rates estimated from a mixed-effect Poisson model controlling for sex, age, thickness of melanoma lesions, histological type and site.

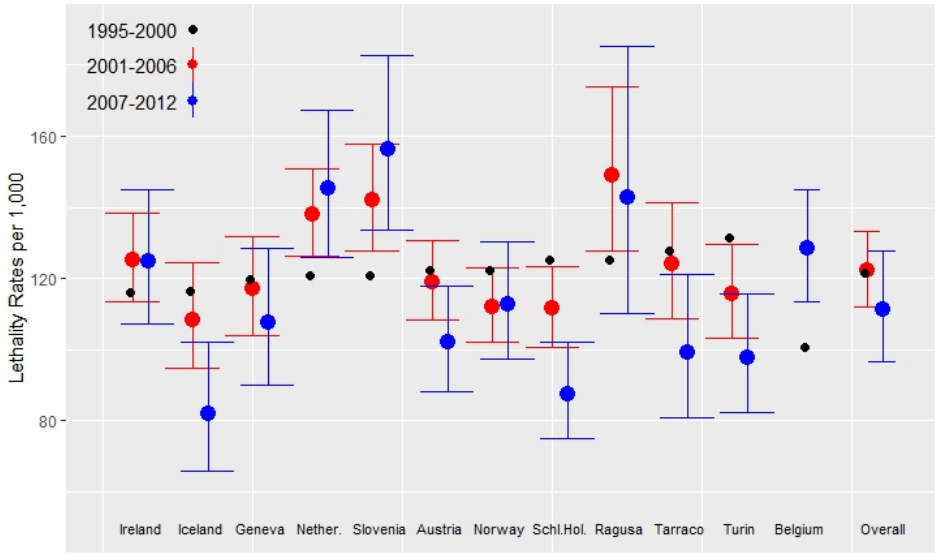


Figure 7.12 Fatal invasive skin melanomas at 3 year. Lethality rates per 1000 cases and 95% confidence limits for period 2001 – 2006 (red) and 2007 – 2012 (blue) vs 1995 – 2000 (reference period represented by a black dot; 2004 – 2006 for Belgium), by cancer registry. Rates estimated from a mixed-effect Poisson model controlling for sex, age, thickness of melanoma lesions, histological type and site.

Schleswig-Holstein (Germany) and Tarragona (Spain). The last three had high lethality in the reference years. A few registries (the Netherlands, Slovenia and Ragusa (Italy)) showed opposite patterns. In the Netherlands and Ragusa a non-significant decrease was seen in the third period after an increase in the second period. Slovenia had increasing lethality rates over time. However, when the analysis included $2013-2017$, it exhibited a plateau in risk, although not a reversal of the trends (data not shown). Finally, Belgium showed a significant increase for the last period compared to $2004-2006$, as data for $1995-2000$ was lacking. Analysis of fatal cases at 3 years confirmed the general pattern shown for 1 year follow-up.

7.3.3 Discussion

An epidemic of melanoma began several decades ago, most likely caused by a global change in outdoor leisure activities that affected most of the affluent white population of Europe, North America and Australia. As a response to this rise in risk and subsequent incidence, increasing awareness among professionals and general public lead to an increased surveillance and early diagnosis, leading in turn to a further increase in diagnosed cases [155]. In particular, the increase in awareness led to the diagnosis of lesions that would probably not have progressed to a lethal disease [156]. However, the lack of a clear-cut marker linking morphological characteristics to prognosis makes it practically impossible to separate the influence of aetiological from diagnostic factors on incidence trends. Statistics on mortality trends can not help in solving the conundrum, as mortality is linked equally to incidence and survival.

To help determine if there has been a true decrease in lethal melanoma, cases with the poorest prognosis, i.e. those that died in a short interval since diagnosis, have been identified. This data has a number of useful characteristics: early deaths should not be influenced, in general, by other concurrent pathologies; their occurrence is not inflated by non-invasive cases or "precautionary" upstaged diagnoses; and finally, these cases are exactly those on which the preventative efforts should be exerted. As information on factors that are linked to prognosis is available only from cancer registries and cannot be retrieved in death certificates, I collected data from those European cancer registries that had information available for a sufficient period of observation. In the analysis, variables such as sex, age, tumour thickness, histological type, and body site have been confirmed as prognostic factors [149]. In particular, as provided by our models, cases of thick melanoma ($>$ 1 mm) had 8 times the risk of dying within one year and 8.5 times the risk of dying within 3 years. However, 289 (1.2%) cases with thin melanoma ($\leq$ 1mm) died at one year and 759 (3.2%) at 3 years. This proportion is comparable to those observed in studies from the USA (3%) [152] and Australia (3%)

[143] for comparable periods. The percentage of fatal thin melanoma in Norway (2.4%) and Sweden (1.9%) was similar to those observed in this analysis.
The main histological type was nodular melanoma (with 6.1% and 19.9% lethality at 1 and 3 years respectively). This was also reported in the Norwegian Malignant Melanoma Registry data for the period $2008-2012$ [157] and from Swedish data [158]. This histological type is a major risk factor for melanoma death, as recently highlighted by Lattanzi and colleagues [159].
In addition, the larger proportion of fatal melanomas found in less documented cases confirmed that a lack of documentation in the registry's records is often an indicator of poorer prognosis [123].
Results showed an overall decrease in the proportion of fatal cases; this has been previously reported in Sweden [160], where the proportion of fatal thin cases decreased from 4.1% in $1989-1993$ to 3.0% in $1999-2003$. However, it has been highlighted that the decrease was not homogeneous across registries, with some of them showing an increasing pattern. Unlike Criscione et al. [150] the proportion of fatal melanomas at 1 and 3 years with unknown thickness decreased slightly over time. On the other hand, as in the US data, the proportion of fatal thick melanomas significantly increased. Some differences between registries might be explained by campaigns devoted to reducing exposure and to recommending change in personal behaviours, or by increasing diagnostic "pressure", as in Germany (Schleswig-Holstein), in Northern Italy (Turin), Iceland [121] and more recently Austria. Some differences might be clarified by contrasting trends according to age, as in the Nordic countries where mortality decreased among individuals younger than 50 years and increased among older individuals [161]. For Ragusa (Italy) findings on fatal cases are consistent with the results of survival analysis of incidence cases overlapping the first two periods of the study ($1995-2000$ and $2001-2006$) [162, 163].
It is worrying that in some countries, despite a nationwide attention to melanoma prevention, rates of fatal cases did not decrease, or even increased. This might indicate an increased diffusion of risk factors in the general population during more recent years. As ionizing radiation is the only proven risk factor for melanoma, increased rates of fatal cases suggest a parallel increase in personal behaviours leading to higher sun exposure or even artificial UV (i.e. sunlamps) exposure. In addition, the countries where an overall downward trend in lethality rates was not observed (in particular Ireland, Netherlands and Slovenia) are those which still have an increasing overall incidence of melanoma looking at the long-term trends [97, 164]; other registries instead showed more stable incidence rates.
In the present study, information on treatment or tumour characteristics, such as for example BRAF markers, was not included. On the other hand, targeted therapies for melanoma have

been approved in Europe only in late 2011 [165], with different availability and adoption across countries; only a small proportion of analysed cases could have benefited from such treatments. In my opinion, the decrease in fatal cases observed in the last period, although not homogeneous, cannot be completely and only attributed to the new promising drugs.
As far as I know, this was the first study that has examined "fatal" melanoma with a definition of observation time since diagnosis. Criscione simply analysed fatal cases overall, and so compared cases with different survival, potentially from 1 to 18 years. In addition, statistical methods used in the present study allowed for a better control of the random effect of choosing only a sample of cancer registries in Europe.
Trends of fatal cases of melanoma highlight real changes in risk, i.e. not related to overdiagnosis. They show a decrease in fatal melanoma cases in most countries, with a few exceptions. Stronger efforts in earlier detection of both fatal and non-fatal cases could lead to a more efficient treatment of melanoma in general.

Appendix A

The motivating example

A.1 Some further biological and technical details

Plasma isolation and storage Blood was centrifuged at 2500 r.p.m (1250g) at 4°C for 10 min. The supernatant was transferred into new tubes and subjected to a second centrifugation step at 2500 r.p.m (1250g) at 4°C for 10 min to remove cell debris and fragments. Plasma was stored in 4.5 ml cryovials at -80°C until transfer to the Cancer Genomics Lab. To calculate hemolysis score [166], $10\mu l$ of plasma was centrifuged at 1000g for 5 min at room temperature and the absorbance at 385 and 414 nm was measured by a NanoDrop spectrophotometer (Thermo Fisher) using the UV-VIS program. Finally, $220\mu l$ aliquots were created for each sample and stored into 1.5 ml tubes at -80°C. Samples with hemolysis score < 0.057 and/or 414 nm/385 nm absorbance ratio below 2 were kept for further processing.

Circulating RNA extraction Before extraction, one 220μl aliquot per sample was centrifuged for 5 min at 1000g at 4°C. Total RNA was extracted with miRNeasy serum/plasma kit (Qiagen) using Exiqon protocol, with the bacteriophage MS2-RNA carrier (Roche Diagnostics) to promote RNA precipitation and purification on membranes. The Caenorhabditis elegans cel-miR-39-3p miR mimic spike-in (Qiagen) was added. RNA samples were eluted in 30μl of nuclease-free water and stored at -80°C.

Appendix B

R programs: details and coding

B.1 A minimal binary classifier providing a proper ROC curve

The R [75] code used to obtain ROC curves reported in Figure 4.3 for the case study (Section 4.4) on circulating miRNAs is presented here.
In details, the program in Section B.1.1 allows to estimate the non-parametric ROC curve associated to a LR-based Flexible Bayes classifier, following the Algorithm 1 in Section 4.3.3. The code in Section B.1.2 enables to implement the ROC curve associated to the best linear combination of three biomarkers, as defined in Su & Liu [5] seminal paper. It is important to remember that Su and Liu combination relies on the normality assumption, i.e. biomarkers present a multivariate normal distribution under diseased (P_+) and healthy (P_-) populations. The function `el.anal.dat` in Section B.1.3 is used to calculate the best linear combination of biomarkers regardless of the data distribution, as presented in [6] (and exploiting the code provided by the authors, with only minor changes). Lastly, in Section B.1.4, the code to plot different ROC curves (including also the one associated to the standard logistic score) and to compare them via the AUC (area under the curve) index is provided.

B.1.1 ROC curve associated to LR based flexible Bayes classifier

```
##############################################################################
#                                                                            #
# ROC_LR_FlexBayes.R                                                         #
# Code to implement the likelihood ratio based flexible Bayes ROC            #
#                                                                            #
##############################################################################

# Data: datapos are diseased subjects
#       dataneg are non-diseased subjects
# t: threshold

myROC <- function(t, dataneg, datapos, B=100000){

  if (is.matrix(dataneg)==F) return("dataneg not a matrix")
  if (is.matrix(datapos)==F) return("datapos not a matrix")

  if (dim(dataneg)[2] != dim(datapos)[2]) return("datapos and dataneg not same number
       of columns!")
  p <- dim(dataneg)[2]

  nneg <- dim(dataneg)[1]
  npos <- dim(datapos)[1]

  xneg <- xpos <- fnegneg <- fposneg <- fnegpos <- fpospos <- matrix(0,B,p)

  for (k in 1:p){
    lambdaneg <- thumb(dataneg[,k])
    lambdapos <- thumb(datapos[,k])
    xneg[,k] <- rnorm(B, mean=dataneg[sample(1:nneg,B,rep=T),k], sd=lambdaneg )
    xpos[,k] <- rnorm(B, mean=datapos[sample(1:npos,B,rep=T),k], sd=lambdapos )
    fnegneg[,k] <- sapply(xneg[,k],mydens,data=dataneg[,k])
    fposneg[,k] <- sapply(xneg[,k],mydens,data=datapos[,k])
    fnegpos[,k] <- sapply(xpos[,k],mydens,data=dataneg[,k])
    fpospos[,k] <- sapply(xpos[,k],mydens,data=datapos[,k])
  }

  FPR <- mean( apply(fposneg,1,prod) > t * apply(fnegneg,1,prod) )
  TPR <- mean( apply(fpospos,1,prod) > t * apply(fnegpos,1,prod))

  c(FPR,TPR)
}

# Gaussian kernel density estimator
mydens <- function(x,data) sum(dnorm(x-data,sd=thumb(data)))/length(data)

# Rule of thumb for bandwidth selection by Silverman (1986)
thumb <- function(x) 1.06*min(sd(x),IQR(x)/1.34)*length(x)^(-.2)

tFLEXgrid_nonpar <- c(0.001,seq(0.01,0.1,by=0.01),seq(0.1,5,by=0.1),seq(5.5,14,by
    =0.5),20,100)
rocFLEXgrid_nonpar <- t(sapply(tFLEXgrid_nonpar, myROC, dataneg=dataneg, datapos=
    datapos))
```

B.1.2 ROC curve associated to Su & Liu best linear combination

```
#############################################################################
#                                                                           #
# ROC_BLC_SuLiu1993.R                                                       #
# Code based on Su & Liu (1993) to obtain the ROC curve associated to       #
# the best linear combination (BLC) of different biomarkers, under          #
# the normality assumption.                                                 #
#                                                                           #
#############################################################################

# Su & Liu best linear combination for 3 biomarkers
# Normality assumption
# Data: datapos are diseased subjects
#       dataneg are non-diseased subjects

linroc_SL <- function(t, mux_H = 0,
                      mux_D = 1,
                      muy_H = 0,
                      muy_D = 2,
                      muz_H = 0,
                      muz_D = 2,
                      sigmax_H = 1,
                      sigmax_D = 2,
                      sigmay_H = 1,
                      sigmay_D = 4,
                      sigmaz_H = 1,
                      sigmaz_D = 4,
                      rhoxy_H = 0,
                      rhoxy_D = 0,
                      rhoxz_H = 0,
                      rhoxz_D = 0,
                      rhoyz_H = 0,
                      rhoyz_D = 0){
  mu_H = c(mux_H, muy_H, muz_H)
  sigma_H = matrix(c(sigmax_H^2, sigmax_H*sigmay_H*rhoxy_H, sigmax_H*sigmaz_H*rhoxz_H
      , sigmax_H*sigmay_H*rhoxy_H, sigmay_H^2, sigmay_H*sigmaz_H*rhoyz_H, sigmax_H*
      sigmaz_H*rhoxz_H, sigmay_H*sigmaz_H*rhoyz_H, sigmaz_H^2 ),3)

  mu_D = c(mux_D, muy_D, muz_D)
  sigma_D = matrix(c(sigmax_D^2, sigmax_D*sigmay_D*rhoxy_D, sigmax_D*sigmaz_D*rhoxz_D
      , sigmax_D*sigmay_D*rhoxy_D, sigmay_D^2, sigmay_D*sigmaz_D*rhoyz_D, sigmax_D*
      sigmaz_D*rhoxz_D, sigmay_D*sigmaz_D*rhoyz_D, sigmaz_D^2),3)

  mu <- mu_D-mu_H

  a <- (solve(sigma_H+sigma_D))%*%mu
  a1 <- a[1,1]
  a2 <- a[2,1]
  a3 <- a[3,1]

  pnorm( ((a1*mux_D+a2*muy_D+a3*muz_D)-(a1*mux_H+a2*muy_H+a3*muz_H))/sqrt(a1^2*sigmax
      _D^2+a2^2*sigmay_D^2+a3^2*sigmaz_D^2+2*a1*a2*rhoxy_D*sigmax_D*sigmay_D+2*a1*a3*
      rhoxz_D*sigmax_D*sigmaz_D+2*a2*a3*rhoyz_D*sigmaz_D*sigmay_D) +
```

```
            qnorm(t)*sqrt(a1^2*sigmax_H^2+a2^2*sigmay_H^2+a3^2*sigmaz_H^2+2*a1*a2*
                rhoxy_H*sigmax_H*sigmay_H+2*a1*a3*rhoxz_H*sigmax_H*sigmaz_H+2*a2*a3*
                rhoyz_H*sigmaz_H*sigmay_H)/sqrt(a1^2*sigmax_D^2+a2^2*sigmay_D^2+a3^2*
                sigmaz_D^2+2*a1*a2*rhoxy_D*sigmax_D*sigmay_D+2*a1*a3*rhoxz_D*sigmax_D*
                sigmaz_D+2*a2*a3*rhoyz_D*sigmaz_D*sigmay_D))

}

mux_H <- mean(dataneg[,2])
muy_H <- mean(dataneg[,3])
muz_H <- mean(dataneg[,1])
sigmax_H <- sd(dataneg[,2])
sigmay_H <- sd(dataneg[,3])
sigmaz_H <- sd(dataneg[,1])
rhoxy_H <- cor(dataneg[,2], dataneg[,3])
rhoxz_H <- cor(dataneg[,2], dataneg[,1])
rhoyz_H <- cor(dataneg[,3], dataneg[,1])

mux_D <- mean(datapos[,2])
muy_D <- mean(datapos[,3])
muz_D <- mean(datapos[,1])
sigmax_D <- sd(datapos[,2])
sigmay_D <- sd(datapos[,3])
sigmaz_D <- sd(datapos[,1])
rhoxy_D <- cor(datapos[,2], datapos[,3])
rhoxz_D <- cor(datapos[,2], datapos[,1])
rhoyz_D <- cor(datapos[,3], datapos[,1])

tSLgrid_param <- seq(0,1,by=0.001)
rocSLgrid_param <- seq(0,1,by=0.001)

for (i in 1:length(rocSLgrid_param))
  rocSLgrid_param[i] <- linroc_SL(tSLgrid_param[i],
                                  mux_H, mux_D, muy_H, muy_D, muz_H, muz_D,
                                  sigmax_H, sigmax_D, sigmay_H, sigmay_D, sigmaz_H,
                                      sigmaz_D,
                                  rhoxy_H, rhoxy_D, rhoxz_H, rhoxz_D, rhoyz_H, rhoyz_
                                      D)
```

B.1.3 ROC curve associated to Chen best linear combination

```
##############################################################################
#                                                                            #
# ROC_BLC_Chen2015.R                                                         #
# Code provided by Chen et al. (2015) to calculate the empirical             #
# likelihood ratio confidence interval estiomates of the best linear         #
# combination (BLC) of biomarkers.                                           #
#                                                                            #
# Minor changes made in function CI.anal (lines 66 - 73); the version        #
# reported in Supplementary material of Chen et al. seems not working;       #
# addition of the score as output of the program.                            #
#                                                                            #
##############################################################################

# installing package 'emplik' for the first time:
if (!require('emplik')) install.packages('emplik')
require(emplik)

##############################################################################
# data.x and data.y are the data sets for the case group and                 #
# the control group, respectively, where rows correspond                     #
# to observations and columns correspond to biomarkers;                      #
# delta is the parameter used to define the bandwidth where                  #
# the bandwidth h=(m+n)^(-delta);                                            #
# initial is a vector of the length of the number of                         #
# biomarkers, the starting point to find the BLC;                            #
# alpha is the nominal siginificance level, e.g., 0.05.                      #
##############################################################################

el.anal.dat <- function(data.x,data.y,delta,initial,alpha){
  n <- nrow(data.x)
  m <- nrow(data.y)
  h <- (m+n)^(-delta) # define bandwidth
  combine<-function(data,lambda) data%*%as.vector(lambda)
  # kernel functions where the normal kernel is considered
  # and can be replace by other kernels
  K.h <- function(x) pnorm (x/h,0,1)
  k.h <- function(x) dnorm (x/h,0,1)/h
  # the objective function to be optimized
  find.lam.kernel <- function(lambda){
    x.c <- combine(data.x,lambda)
    y.c <- combine(data.y,lambda)
    -mean(mapply(function(x) K.h(x-y.c),x.c) )
  }
  # the gradient function
  grn <- function(lambda){
    x.c <- combine(data.x,lambda)
    y.c <- combine(data.y,lambda)
    xy.diff.k <- mapply(function(x) k.h(x-y.c),x.c)
    -sapply(1:ncol(data.x),function(i) {
      mean(xy.diff.k*mapply(function(x) (x-data.y[,i]),data.x[,i]) )
    })
  }
  # obtain BLC coefficients
```

```
  opt.kernel <- optim(initial,find.lam.kernel,method="BFGS",gr=grn,
                      control=list(maxit=30000, ndeps=.1))
  A.kernel <- -opt.kernel$value
  a.kernel <- opt.kernel$par
  # obtain estimates used in the r(A0)
  x.c <- combine(data.x,a.kernel)
  y.c <- combine(data.y,a.kernel)
  xy.diff <- mapply(function(x) K.h(x-y.c),x.c)
  vi <- apply(xy.diff,2,mean)
  wj <- apply(xy.diff,1,mean)
  S10 <- sum( (wj-mean(wj))^2)/(m-1)
  S01 <- sum( (vi-mean(vi))^2)/(n-1)
  S2 <- (n*S01+m*S10)/(m+n)
  crit.val<-qchisq(1-alpha,df=1) # critical value

  # Minor changes made in this function: now it returns a list
  CI.anal <- function(AUC){
    ELR.A0 <- el.test(vi, AUC)#$'-2LLR'
    sigma2.hat <- mean((vi-AUC)^2)
    r <- m/(m+n)*sigma2.hat/S2
    r_ELRA0 <-lapply(ELR.A0, '*', r)
    #return(r*ELR.A0)
    return(r_ELRA0)
  }

  score <- rbind(y.c,x.c)

  tmp=findUL(step=0.1, initStep=0, fun=CI.anal, MLE=A.kernel, level=crit.val)
  upper <- tmp$Up
  lower <- tmp$Low

  out <- list(CI = c(lower,upper),A.est = A.kernel,
              a.est = a.kernel/a.kernel[1]*sign(a.kernel[1]),
              score = score)

  return(out)
}
```

B.1.4 ROC curves comparison

```
#########################################################################
#                                                                       #
# ROC_comparison.R                                                      #
# To compare ROC curve associated with the LR-based Flexible Bayes      #
# classifier and ROC curves associated with common classifier based     #
# on best linear (Chen, 2015) and logistic score.                       #
# AUC and confidence interval calculated for each classifier.           #
#                                                                       #
# Depends on:                                                           #
#              - ROC_LR_FlexBayes.R (for LR Flexible Bayey ROC          #
#                                    (Sacchetto & Gasparini (2020)))    #
#              - ROC_BLC_Chen_2015_CSDA.R (for Chen (2015) ROC)         #
#                                                                       #
#########################################################################

# Data:
#       - datapos are diseased subjects
#       - dataneg are non-diseased subjects
#       - 3 biomarkers measurement for each subject

dataneg=read.table("dataneg.txt")
datapos=read.table("datapos.txt")
dataneg <- as.matrix(dataneg)
datapos <- as.matrix(datapos)

# Logistic score
dataset <- as.data.frame(rbind(dataneg, datapos))
status <- c(rep(0, times=nrow(dataneg)), rep(1, times=nrow(datapos)))
model.logistic<-glm(status~ dataset[,1] + dataset[,2] + dataset[,3],
                    family=binomial, data=dataset)
coef.out <- model.logistic$coefficients
mat.coef <- t(coef.out)
score.logistic <- as.matrix(dataset)%*%mat.coef[2:4]

# Calling the library ROCR (one among many R packages)
library(ROCR)

# ROC
pred.logistic <- prediction(score.logistic, status)
perf.logistic <- performance(pred.logistic,"tpr","fpr")

# AUC of logistic score ROC
auc.logistic <- performance(pred.logistic, measure = "auc")
auc.logistic <- auc.logistic@y.values[[1]]

# Library pROC allows to directly calculate confidence intervals
library(pROC)
roc.graph.logistic=roc(status, score.logistic[,1], smooth = FALSE,
                       auc = TRUE, ci = TRUE, plot = TRUE,
                       identity=TRUE, legacy.axes=TRUE,
                       xlab="1-specificity", ylab = "Sensibility")
logistic_AUC <- roc.graph.logistic$auc
logistic_CI <- roc.graph.logistic$ci
```

```
# ROC associated to Chen (2015) best linear combination of biomarkers
# The BLC of biomarkers provides a score for each subject
# The ROC is calculated for this score
# From ROC_BLC_Chen_2015_CSDA.R
est.Chen <-el.anal.dat(datapos,dataneg,delta=0.08,
                       initial=rep(1,ncol(datapos)),alpha=0.05)

pred.BLC.Chen <- prediction(est.Chen$score, status)
perf.BLC.Chen <- performance(pred.BLC.Chen,"tpr","fpr")

# AUC of Chen best linear combination ROC
# (alternatively to est.Chen$A.est)
auc.Chen <- performance(pred.BLC.Chen, measure = "auc")
auc.Chen <- auc.Chen@y.values[[1]]

# Confidence Interval for BLC Chen 2015
CI_Chen <- est.Chen$CI

# Alternatively, ROC curve, AUC and CI via pROC
roc.graph.Chen=roc(status, as.numeric(est.Chen$score),
                   smooth = FALSE, auc = TRUE, ci = TRUE, plot = TRUE,
                   identity=TRUE, legacy.axes=TRUE, xlab="1-specificity",
                   ylab = "Sensibility")
Chen_AUC <- roc.graph.Chen$auc
Chen_CI <- roc.graph.Chen$ci

# Plot of the curves
pdf("ROC_logistic_BLC_Chen_FlexBayes.pdf", width=8.5, height=8.5)

plot(perf.logistic, xlab="FPR", ylab="TPR")
lines(perf.BLC.Chen@x.values[[1]], perf.BLC.Chen@y.values[[1]],lty=3, lwd=1.5)

# ROC associated to LR-based Flexible Bayes classifier
# From ROC_LR_FlexBayes.R
points(rocFLEXgrid_nonpar, type='l',lty=5)

lines(c(0,1),c(0,1), col = 'gray')

legend(0.75, 0.15, c('Logistic score', 'BLC Chen', 'Flex Bayes'),
       col = c(1,1,1), lty = c(1,3,5), lwd=c(1,1.5,1), cex =1)

dev.off()

# AUC associated to LR-based Flexible Bayes ROC
f <- approxfun(c(1,rocFLEXgrid_nonpar[,1]), c(1, rocFLEXgrid_nonpar[,2]))
AUC.LRroc <- integrate(f,  min(rocFLEXgrid_nonpar[,1]), 1)$value

# To calculate confidence interval for the AUC of the LR-based Flexible
# Bayes ROC curve a bootstrap approach is preferred.
#
# myAUC_i fuction calculate AUC via simulations, applying the definition
# AUC = P(LR+ > LR-)
# Input:
#       data = rbind(datapos, dataneg), where
```

```
#                       datapos: diseased subjects
#                       dataneg: non-diseased subjects
#       p: number of predictors
#       B: number of simulations
#       indices: required for the random selection of rows
#       npos: number of rows of datapos
#       nneg: number of rows of dataneg

myAUC_i <- function(B, p, data=rbind(datapos, dataneg), indices, npos, nneg){
  ineg <- data[indices[indices>npos],]
  ipos <- data[indices[indices<=npos],]
  nneg <- dim(ineg)[1]
  npos <- dim(ipos)[1]

  xneg <- xpos <- LRpos <- LRneg <-matrix(0,B,p)
  for (k in 1:p){
    lambdaneg <- thumb(ineg[,k])
    lambdapos <- thumb(ipos[,k])
    xneg[,k] <- rnorm(B, mean=ineg[sample(1:nneg,B,rep=T),k], sd=lambdaneg )
    xpos[,k] <- rnorm(B, mean=ipos[sample(1:npos,B,rep=T),k], sd=lambdapos )
    LRpos[,k] <- sapply(xpos[,k], myLR, datapos=ipos[,k], dataneg=ineg[,k])
    LRneg[,k] <- sapply(xneg[,k], myLR, datapos=ipos[,k], dataneg=ineg[,k])
  }

  mean( apply(LRpos,1,prod) > apply(LRneg, 1, prod) )
}

# Likelihood ratio
myLR <- function(x, datapos, dataneg) mydens(x, datapos)/mydens(x, dataneg)

# Set seed for reproducibility
set.seed(1)
data <-rbind(datapos, dataneg)
AUC_calc <- myAUC_i(B=10000, p=3, data, indices = c(1:nrow(data)), npos=nrow(datapos)
    , nneg=nrow(dataneg))

# AUC CI, via boostrap
library(boot)
AUC_boot <- boot(data=data, statistic=myAUC_i, R=100, B=10000, p=3, npos=nrow(datapos
    ), nneg=nrow(dataneg))
AUC_boot$t  # Different AUC values for each bootstrap replicate
AUC_boot$t0 # AUC value (via bootstrap)
# Confidence interval
AUC_ci <- boot.ci(AUC_boot, conf = 0.95, type = c( "basic"))

AUC.comparison <-rbind(auc.logistic, auc.Chen=est.Chen$A.est, AUC.LRroc)
```

B.2 A new implementation of hard assignment

The code for the new implementation of hard assignment in R [75] is presented in the following. Apart from the basic function, a parallel version is proposed, as well as an heuristic approach to obtain the maximum of the likelihood. Lastly, the code is applied to the toy example dataset.

```
##############################################################################
#                                                                            #
# Binary populations, binary outcomes                                        #
# Functions to implement hard assignment                                     #
#                                                                            #
##############################################################################

library(dplyr)
library(pryr)

# To convert individual data to profiles
dataToProfiles <- function(data){
  # obtaining profiles
  data %>% # pipe symbol (dplyr package)
  unique(margin = 1) -> profiles

  # cbinding the frequencies
  profiles <- cbind(profiles, profiles %>%  # obtaining frequencies
    apply(1, function(y) apply(data,1,function(x) all(x==y))) %>%
    colSums() -> frequencies)
  return(profiles)
}

# Function likelihoodEval01: to evaluate the log-likelihood
# function, with Laplace smoothing adjustment, for a certain
# cluster configuration cl
# Input
#       cl: possible configuration to be considered
#       profiles: matrix of profiles
# Output: total log-likelihood for the given configuration
likelihoodEval01 <- function(cl, profiles){
  # number of questions (columns - 1) in profiles
  r = ncol(profiles) - 1

  # cl is a logical vector
  # 1 if the element is in the + cluster
  # 0 if the element is in the - cluster

  # To find the number of elements in the + cluster
  den_p = sum(profiles[cl, r + 1])

  # To find the number of elements in the - cluster
  den_m = sum(profiles[!cl, r + 1])
```

```
  # Special case: the + cluster is composed by one element
  if ((length(cl[cl == TRUE]) > 1) || (length(cl[cl == TRUE]) == 0)) {
    num_p = colSums(profiles[cl, 1:r]*profiles[cl, r+1])
  } else {
    num_p = profiles[cl, 1:r]*profiles[cl, r+1]
  }

  # Special case: the - cluster is composed by one element
  if ((length(cl[cl == FALSE]) > 1) || (length(cl[cl == FALSE]) == 0)) {
    num_m = colSums(profiles[!cl, 1:r]*profiles[!cl, r+1])
  } else {
    num_m = profiles[!cl, 1:r]*profiles[!cl, r+1]
  }

  # To evaluate probabilities in + and - using Laplace Smoothing
  p_p = (num_p + 1)/(den_p + 2)
  p_m = (num_m + 1)/(den_m + 2)

  # Total log-likelihood for cluster +
  # Logarithm transformation adopted for convenience
  CpL = sum((profiles[cl,1:r]*profiles[cl,r+1])%*%log(p_p) +
              ((1 - profiles[cl,1:r])*profiles[cl,r+1])%*%log(1-p_p))

  # Total log-likelihood for cluster -
  # Logarithm transformation adopted for convenience
  CmL = sum((profiles[!cl,1:r]*profiles[!cl,r+1])%*%log(p_m) +
              ((1 - profiles[!cl,1:r])*profiles[!cl,r+1])%*%log(1-p_m))

  # Total log-likelihood
  L = CpL + CmL
  return(L)
}

# Function likelihoodEval01Laplace
# Modification of function likelihoodEval01, including adjustment
# smoothing parameters as input (in Laplace case a=1, b = 2)
likelihoodEval01Laplace <- function(cl, profiles, a = 1, b = 2){
  # number of questions (columns - 1) in profiles
  r = ncol(profiles) - 1

  # cl is a logical vector
  # 1 if the element is in the + cluster
  # 0 if the element is in the - cluster

  # To find the number of elements in the + cluster
  den_p = sum(profiles[cl, r + 1])

  # To find the number of elements in the - cluster
  den_m = sum(profiles[!cl, r + 1])

  # Special case: the + cluster is composed by one element
  if ((length(cl[cl == TRUE]) > 1) || (length(cl[cl == TRUE]) == 0)) {
    num_p = colSums(profiles[cl, 1:r]*profiles[cl, r+1])
  } else {
    num_p = profiles[cl, 1:r]*profiles[cl, r+1]
```

```
  }

  # Special case: the - cluster is composed by one element
  if ((length(cl[cl == FALSE]) > 1) || (length(cl[cl == FALSE]) == 0)) {
    num_m = colSums(profiles[!cl, 1:r]*profiles[!cl, r+1])
  } else {
    num_m = profiles[!cl, 1:r]*profiles[!cl, r+1]
  }

  # To evaluate the probabilities in + and - using Laplace Smoothing
  # with parameter a = 1 and b = 2
  p_p = (num_p + a)/(den_p + b)
  p_m = (num_m + a)/(den_m + b)

  # Total log-likelihood for cluster +
  CpL = sum((profiles[cl,1:r]*profiles[cl,r+1])%*%log(p_p) +
              ((1 - profiles[cl,1:r])*profiles[cl,r+1])%*%log(1-p_p))

  # Total log-likelihood for cluster -
  CmL = sum((profiles[!cl,1:r]*profiles[!cl,r+1])%*%log(p_m) +
              ((1 - profiles[!cl,1:r])*profiles[!cl,r+1])%*%log(1-p_m))

  # Total log-likelihood
  L = CpL + CmL
  return(L)
}

# ------------------------------------------------------------------ #
# Example
# Random data generation
r <- 4 # Number of different items
data <- matrix(rbinom(1000, 1, .6), 250, r, byrow = T)
head(data)

# From observation to profiles
profiles <- dataToProfiles(data)
t(profiles)
m <- nrow(profiles)

# m x 2^m assignment matrix
tmp <- split.data.frame(cbind(rep(0, m), rep(1, m)),rep(1:m))
assignments <- matrix(t(expand.grid(tmp)), nrow=m, ncol=2^m)
# Dimensions of assignments (memory allocation)
format(object.size(assignments), units = "Mb")

# Log-likelihood evaluation for a certain cl
cl <- as.logical(assignments[,100])
as.numeric(cl)
likelihoodEval01(cl, profiles)

# Maximum log-likelihood applying function loglikeEval01
results <- apply(assignments, 2,
                 function(cl) likelihoodEval01(as.logical(cl), profiles))
```

```
#############################################################################
#                                                                           #
# Binary populations, binary outcomes                                       #
# Functions to implement hard assignment                                    #
# Parallel version                                                          #
#                                                                           #
#############################################################################

library(parallel)
library(dplyr)
library(pryr)

# Function splitMat
# Input: matrix mat to be splitted
# Output: List of N submatrices of mat
splitMat <- function(mat, N){

  # The list of sub-matrices is inizialized with the first element
  submats = list(mat[,1:round(dim(mat)[2]/N, 0)])

  # To append the remaining matrices
  for (k in 1:(N - 2)) {
    submats = append(submats,
                     list(mat[,(k*round(dim(mat)[2]/N, 0)):((k + 1)*round(dim(mat)[2]
                         /N, 0))]))
  }

  # To append the last one
  submats = append(submats,
                   list(mat[,((N - 1)*round(dim(mat)[2]/N, 0)):dim(mat)[2]]))

  return(submats)
}

# Function colLike01
# To evaluate the log-likelihood of the columns of a sub-assignments matrix
# in a vectorized way (using apply function)
# Output: best configurations (cls) and the relative log-likelihood
colLike01 <- function(submat) {
  loglike <- apply(submat, 2, function(cl) likelihoodEval01(as.logical(cl), df))
  return(unique(as.data.frame(cbind(t(submat), loglike)) %>%
                  filter(loglike == max(loglike))))
}

# ---------------------------------------------------------------------------- #
# Example
set.seed(7) # Set seed for reproducibility

# Number of questions
r <- 5
# Random dataset
data <- matrix(rbinom(1000, 1, .6), 200, r, byrow = T)
head(data)

# From observations to profiles
```

```
profiles <- dataToProfiles(data)
profiles <- profiles[1:17,]

# Total number of profiles
m <- nrow(profiles)

# Assignment matrix generation
df <- profiles
tmp <- split.data.frame(cbind(rep(0, m), rep(1, m)),rep(1:m))
assignments <- matrix(t(expand.grid(tmp)), nrow=m, ncol=2^m)

# Subdivision of the assignment matrix into n (number of cores) submatrices
# to create clusters for parallel computing
numCores <- detectCores()
numCores
clusters <- makeCluster(numCores)

# List containing numCores sub-matrices of assignments
submats <- splitMat(assignments, numCores)

# Dimensions of each sub-matrix
dim(submats[[1]])

# To run the program in parallel with "parallel" package it is necessary
# to export the libraries, data and functions
clusterEvalQ(clusters, {library(dplyr); library(magrittr)})
clusterExport(clusters, c("df", "colLike01", "likelihoodEval01"))

# To run the code in parallel
parallelOutput <- parLapply(clusters, submats, colLike01)

# Same code not run in parallel
output <- colLike01(assignments)
```

```
##############################################################################
#                                                                            #
# Binary populations, binary outcomes                                        #
# Functions to implement hard assignment                                     #
# Heuristic procedures                                                       #
#                                                                            #
##############################################################################

# Heuristic 1.
# Function MC to implement basic Montecarlo method
# Input:  df: dataframe
#         nIter: number of iterations
#         nSim: number of random configurations considered at each
#               iteration
# Output: maximum of log-likelihood and the associated configuration
MC <- function(df, nIter = 100, nSim = 1000){

  # To retrieve number of profiles and answers
  m <- nrow(df)
  r <- ncol(df) - 1

  # To inizialize solution vector (best solution until t step)
  solution <- rep(-Inf, nIter)

  # To inizialize best solution (max log-likelihood found)
  best_sol <- -Inf

  # To inizialize best configuration
  best_conf = rep(0, m)

  # Iterations
  for (t in 1:nIter){
    SIM <- matrix(sample(0:1, nSim*m, replace = TRUE), nrow = m, ncol = nSim)

    # current solution dataframe at iteration t
    current <- (colLike01(SIM) %>% filter(loglike == max(loglike)) %>% unique())

    # current max log-likelihood found at iteration t
    current_sol <- current$loglike %>% unlist() %>% unique()

    # current configuration found at iteration t
    current_conf <- current %>% select(-loglike) %>% head(1) %>% as.numeric()

    # If a new maximum is found
    if (current_sol > best_sol){
      best_sol <- current_sol
      best_conf <- current_conf
    }

    solution[t] <- best_sol
  }

  # Results
  output <- list()
  output[[1]] <- best_conf
```

```
  output[[2]] <- best_sol
  return(output)
}

# Heuristic 2.
# Function swapMC to implement simple iterative heuristic
# with MC initialization (Algorithm 2 in Chapter 6)
# Input:  df: dataframe
#         nIter: number of iterations
#         MCnSim: number of run of Monte Carlo basic heuristic
#         MCnIter: number of iterations of Monte Carlo basic heuristic
# Output: maximum of the log-likelihood and the associated configurations
swapMC <- function(df, nIter = 100, MCnSim = 100, MCnIter = 100){
  m <- nrow(df)
  r <- ncol(df) - 1

  # Inizialization using Monte Carlo basic method
  cat("Initializing the first configuration using MC ...\n")
  init <- MC(df, MCnIter, MCnSim)
  cl <- init[[1]]
  L0 <- init[[2]]

  # Initialization of the best configuration
  cl_best <- cl

  # Initialization of the best log-Likelihood
  L_best <- L0

  # Iterations
  for (it in 1:nIter){
    # Swap
    cl <- cl_best
    rnd <- sample(1:m,1)
    cl[rnd] <- 1 - cl[rnd]

    # If better, mantain the swap
    if (likelihoodEval01(as.logical(cl), df) > L_best) {
      L_best <- likelihoodEval01(as.logical(cl), df)
      cl_best <- cl
    }
  }

  # Results
  output <- list()
  output[[1]] <- cl_best
  output[[2]] <- L_best
  return(output)
}

# ------------------------------------------------------------------------- #
# Examples
# Example 1. MC basic

library(dplyr)
library(pryr)
```

```
# Generating SIM matrix
nSim <- 10
set.seed(5)
SIM <- matrix(sample(0:1, nSim*m, replace = TRUE), nrow = m, ncol = nSim)

# Best configuration
colLike01(SIM)

# Set parameters
set.seed(5)
nIter <- 100
nSim <- 1000

# Initialization of solution vector (best solution until t step)
solution <- rep(-Inf, nIter)

# Initialization of best solution (max log-likelihood found)
best_sol <- -Inf

# Initialization of best configuration
best_conf = rep(0, m)

# Iterations
for (t in 1:nIter){
  SIM <- matrix(sample(0:1, nSim*m, replace = TRUE), nrow = m, ncol = nSim)

  # current solution dataframe at iteration t
  current <- (colLike01(SIM) %>% filter(loglike == max(loglike)) %>% unique())

  # current max log-likelihood found at iteration t
  current_sol <- current$loglike %>% unlist() %>% unique()

  # current configuration found at iteration t
  current_conf <- current %>% select(-loglike) %>% head(1) %>% as.numeric()

  # if a new maximum is found
  if (current_sol > best_sol){
    best_sol <- current_sol
    best_conf <- current_conf
  }

  solution[t] <- best_sol
}

# Results
best_conf
max(solution)

# Example 2. swapMC heuristic
library(parallel)
set.seed(7)# for reproducibility

# Number of questions
r <- 5
```

```
# Random dataset
data <- matrix(rbinom(1000, 1, .6), 200, r, byrow = T)
head(data)

# From observations to profiles
profiles <- dataToProfiles(data)
profiles <- profiles[1:20,]
m <- nrow(profiles)

# Assignment matrix generation
df <- profiles
tmp <- split.data.frame(cbind(rep(0, m), rep(1, m)),rep(1:m))
assignments <- matrix(t(expand.grid(tmp)), nrow=m, ncol=2^m)

# Subdivision of the assignment matrix into n (number of cores) submatrices
# creating clusters for parallel computing
numCores <- detectCores()
clusters <- makeCluster(numCores)

# List containing numCores sub-matrices of assignments
submats <- splitMat(assignments, numCores)

# Exporting the libraries, data and functions
clusterEvalQ(clusters, {library(dplyr); library(magrittr)})
clusterExport(clusters, c("df", "colLike01", "likelihoodEval01"))

# Parallel version - exact solution
start_time = Sys.time()
parallelOutput <- parLapply(clusters, submats, colLike01)
end_time = Sys.time()
end_time - start_time

# Best solution obtained by the exact (exhaustive search) method
Reduce("rbind", parallelOutput) %>%
  filter(loglike == max(loglike))-> exact_sol
exact_sol %>% select(loglike) %>% unique()

# Swap heuristic
swapMC(df)

# How many times maximum is reached
maxima <- rep(0, 100)
for (i in 1:100) {
  cat(paste0(i, "\n"))
  set.seed(i)
  maxima[i] <- swapMC(df)[[2]]
}

# Correct solutions
sum(maxima == max(maxima))
```

```
##############################################################################
#                                                                            #
# Binary populations, binary outcomes                                        #
# Functions to implement hard assignment                                     #
#                                                                            #
# Toy example - Bartholomew et al. dataset                                   #
#                                                                            #
##############################################################################

library(dplyr)
library(pryr)
library(parallel)

data_Bart <- read.table('Bartholomew_dataset.txt', sep ='\t', header=T)
data_Bart <- data.frame(frequencies = data_Bart$Observed_freq,
                        q1 = data_Bart$V1,
                        q2 = data_Bart$V2,
                        q3 = data_Bart$V3,
                        q4 = data_Bart$V4)
t(data_Bart)
m <- dim(data_Bart)[1]

# m x 2^m assignment matrix evaluation
tmp <- split.data.frame(cbind(rep(0, m), rep(1, m)),rep(1:m))
assignments <- matrix(t(expand.grid(tmp)), nrow=m, ncol=2^m)

# Profiles evaluation
profiles <- data.frame(data_Bart$q1, data_Bart$q2, data_Bart$q3, data_Bart$q4,
                       data_Bart$frequencies) %>%
            as.matrix()
df <- profiles

# Exact method
numCores <- detectCores()
clusters <- makeCluster(numCores)
submats <- splitMat(assignments, numCores)
clusterEvalQ(clusters, {library(dplyr); library(magrittr)})
clusterExport(clusters, c("df", "colLike01", "likelihoodEval01"))
start_time = Sys.time()
parallelOutput <- parLapply(clusters, submats, colLike01)
end_time = Sys.time()
end_time - start_time
Reduce("rbind", parallelOutput) %>%
  filter(loglike == max(loglike))-> exact_sol
exact_sol %>% dplyr::select(loglike) %>% unique()
exact_sol

# Comparison with soft assignment (EM algorithm)
# Different solutions found
hard_sol <- as.numeric(exact_sol[1,] %>% dplyr::select(-loglike))
poLCA_sol <- c(1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0)

# Log-likelihoods
likelihoodEval01(as.logical(hard_sol), df)
likelihoodEval01(as.logical(poLCA_sol), df)
```

```
# Cbinding the dataframe with the solutions found
complete_df <- as.data.frame(cbind(df, hard_sol, poLCA_sol))

# Number of individuals
individuals <- complete_df %>%
    dplyr::select(data_Bart.frequencies) %>%
    sum()

# Number of individuals in cluster + for hard assignment
complete_df %>%
  filter(hard_sol == 1) %>%
  dplyr::select(data_Bart.frequencies) %>%
  sum() -> np_hard

# Number of individuals in cluster - for hard assignment
nm_hard <-  individuals - np_hard

# Number of individuals in cluster + for soft assignment
complete_df %>%
  filter(poLCA_sol == 1) %>%
  dplyr::select(data_Bart.frequencies) %>%
  sum() -> np_soft

# Number of individuals in cluster - for soft assignment
nm_soft <-  individuals - np_soft

c(np_hard, np_soft)
c(nm_hard, nm_soft)

# Correct answers for question 1 in cluster + hard assignment
(((complete_df %>%
  filter(hard_sol == 1) %>%
  dplyr::select(data_Bart.q1))*
  (complete_df %>%
     filter(hard_sol == 1) %>%
     dplyr::select(data_Bart.frequencies))) %>%
  sum())/np_hard -> h11

# Correct answers for question 2 in cluster + hard assignment
(((complete_df %>%
     filter(hard_sol == 1) %>%
     dplyr::select(data_Bart.q2))*
    (complete_df %>%
       filter(hard_sol == 1) %>%
       dplyr::select(data_Bart.frequencies))) %>%
    sum())/np_hard -> h12

# Correct answers for question 3 in cluster + hard assignment
(((complete_df %>%
     filter(hard_sol == 1) %>%
     dplyr::select(data_Bart.q3))*
    (complete_df %>%
       filter(hard_sol == 1) %>%
       dplyr::select(data_Bart.frequencies))) %>%
```

```
    sum())/np_hard -> h13

# Correct answers for question 4 in cluster + hard assignment
(((complete_df %>%
     filter(hard_sol == 1) %>%
     dplyr::select(data_Bart.q4))*
    (complete_df %>%
       filter(hard_sol == 1) %>%
       dplyr::select(data_Bart.frequencies))) %>%
    sum())/np_hard -> h14

# Correct answers for question 1 in cluster - hard assignment
(((complete_df %>%
     filter(hard_sol == 0) %>%
     dplyr::select(data_Bart.q1))*
    (complete_df %>%
       filter(hard_sol == 0) %>%
       dplyr::select(data_Bart.frequencies))) %>%
    sum())/np_hard -> h01

# Correct answers for question 2 in cluster - hard assignment
(((complete_df %>%
     filter(hard_sol == 0) %>%
     dplyr::select(data_Bart.q2))*
    (complete_df %>%
       filter(hard_sol == 0) %>%
       dplyr::select(data_Bart.frequencies))) %>%
    sum())/np_hard -> h02

# Correct answers for question 3 in cluster - hard assignment
(((complete_df %>%
     filter(hard_sol == 0) %>%
     dplyr::select(data_Bart.q3))*
    (complete_df %>%
       filter(hard_sol == 0) %>%
       dplyr::select(data_Bart.frequencies))) %>%
    sum())/np_hard -> h03

# Correct answers for question 4 in cluster - hard assignment
(((complete_df %>%
     filter(hard_sol == 0) %>%
     dplyr::select(data_Bart.q4))*
    (complete_df %>%
       filter(hard_sol == 0) %>%
       dplyr::select(data_Bart.frequencies))) %>%
    sum())/np_hard -> h04

# Correct answers in hard approach for each question
hp <- c(h11, h12, h13, h14)
hm <- c(h01, h02, h03, h04)

# Heursitc procedure: swap
set.seed(7)
start_time = Sys.time()
results_swap <- swapMC(df)
```

```
end_time = Sys.time()
end_time - start_time
results_swap[[1]]
results_swap[[2]]

# Exact method
results_exact <- apply(assignments, 2, function(cl) likelihoodEval(as.logical(cl),
    profiles))

dfResults = as.data.frame(cbind(t(assignments), results_exact))
dfResults %>% head()

# Best configuration
dfResults %>% filter(results_exact == max(results_exact))

# Worst configuration
dfResults %>% filter(results_exact == min(results_exact))
```

B.3 Applications

B.3.1 Examples of hyperparameters tuning

Examples of R [75] and Python [102] codes to tune the hyperparameters of random forests and neural networks models are presented.

```
##############################################################################
#                                                                            #
# Example to tune RFs (via ranger package)                                   #
#                                                                            #
##############################################################################

library('ranger')
library('survival')

attach(db_NA25_imputed_ximp_logOS)

# Hyperparameter grid search for RF
hyper_grid2 <- expand.grid(
  mtry       = seq(2, 40, by = 3),
  node_size  = seq(5, 40, by = 5),
  sampe_size = c(.632, .70, .80),
  split_rule = c("variance", "extratrees", "maxstat"),
  OOB_MSE    = 0
)
nrow(hyper_grid2)

for(i in 1:nrow(hyper_grid2)) {
  # Train model
  model <- ranger(
```

```
    formula          = logOS ~ .,
    data             = db_NA25_imputed_ximp_logOS,
    num.trees        = 500,
    mtry             = hyper_grid2$mtry[i],
    min.node.size    = hyper_grid2$node_size[i],
    sample.fraction  = hyper_grid2$sampe_size[i],
    splitrule        = hyper_grid2$split_rule[i],
    seed             = 123
  )
  # Add OOB error to grid
  hyper_grid2$OOB_OS_MSE[i] <- model$prediction.error
}


################################################################################
#
# Example to tune NNs in Python
#
################################################################################

# To implement grid search in Python for NNs
# Use scikit-learn to grid search
import numpy
from sklearn.model_selection import GridSearchCV
from keras.models import Sequential
from keras.layers import Dense
from keras.wrappers.scikit_learn import KerasRegressor

# Function to create model, required for KerasRegressor
def create_model(optimizer='RMSprop', neurons1=1, neurons2=1, activation1 = 'elu',
activation2 = 'relu'):
    # Create model
    model = Sequential()
    model.add(Dense(266, input_dim=266, activation=activation1))
    model.add(Dense(neurons1, activation=activation1))
    model.add(Dense(neurons1, activation=activation1))
    model.add(Dense(neurons1, activation=activation2))
    model.add(Dense(neurons2, activation=activation2))
    model.add(Dense(neurons2, activation=activation2))
    model.add(Dense(1,  activation='elu'))
    # Compile model
    model.compile(loss='mean_squared_error', optimizer=optimizer) #, metrics=['mse'])
    return model

# Fix random seed for reproducibility
seed = 7
numpy.random.seed(seed)

# Load data
X = db_OHE_25NAimputed_OS.iloc[:, 0:266]
Y = db_OHE_25NAimputed_OS.iloc[:, 266]

# Create model
model = KerasRegressor(build_fn=create_model, verbose=0)
```

```
# Define the grid search parameters
epochs = [500, 1000]
optimizer = ['SGD', 'RMSprop', 'Adagrad'] #, 'Adadelta', 'Adam', 'Adamax', 'Nadam']
neurons1 = [100, 200, 266]
neurons2 = [50, 100, 200]
activation1 = ['elu', 'tanh']#['relu', 'elu', 'tanh']
activation2 = ['relu', 'elu']#, 'linear']
param_grid = dict(epochs=epochs, optimizer = optimizer, neurons1 = neurons1,
neurons2 = neurons2, activation1 = activation1,
activation2 = activation2)

# The GridSearchCV process will then construct and evaluate one model for each combination
# of parameters. Cross validation is used to evaluate each individual model and the default
# of 3-fold cross validation. All scores objects follow the convention that higher return
# values are better than lower return values. Thus metrics which measure the distance between
# the model and the data, like metrics.mean_squared_error, are available as neg_mean_squared_error
# which return the negated value of the metric.
grid = GridSearchCV(estimator=model, param_grid=param_grid, n_jobs=-1,
scoring = 'neg_mean_squared_error')
grid_result = grid.fit(X, Y)
```

Bibliography

[1] Kelly H Zou, Aiyi Liu, Andriy I Bandos, Lucila Ohno-Machado, and Howard E Rockette. *Statistical evaluation of diagnostic performance: topics in ROC analysis*. CRC Press, 2011.

[2] Andriy I Bandos, Ben Guo, and David Gur. Estimating the area under ROC curve when the fitted binormal curves demonstrate improper shape. *Academic Radiology*, 24(2):209–219, 2017.

[3] Wen-Chung Lee. Probabilistic analysis of global performances of diagnostic tests: interpreting the lorenz curve-based summary measures. *Statistics in Medicine*, 18(4):455–471, 1999.

[4] George H John and Pat Langley. Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI:338–345, 1995.

[5] John Q Su and Jun S Liu. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355, 1993.

[6] Xiwei Chen, Albert Vexler, and Marianthi Markatou. Empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers. *Computational Statistics & Data Analysis*, 82:186–198, 2015.

[7] Sandra Vieira, Walter HL Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75, 2017.

[8] George B Macready and C Mitchell Dayton. The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2(2):99–120, 1977.

[9] David M Green and John A Swets. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[10] James P Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, London, 1975.

[11] Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.

[12] Wojtek J Krzanowski and David J Hand. *ROC curves for continuous data*. Chapman and Hall/CRC, 2009.

[13] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[14] Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415, 1975.

[15] Adam R Brentnall and Jack Cuzick. Use of the concordance index for predictors of censored survival data. *Statistical Methods in Medical Research*, 27(8):2359–2373, 2018.

[16] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

[17] Jerzy Neyman and Egon S Pearson. On the problem of the most efficient tests of statistical inference. *Biometrika A*, 20:175–240, 1933.

[18] Erich Leo Lehmann. *Testing statistical hypotheses*. Springer, 1986.

[19] Chris J Lloyd. Estimation of a convex ROC curve. *Statistics & Probability Letters*, 59(1):99–111, 2002.

[20] Baojiang Chen, Pengfei Li, Jing Qin, and Tao Yu. Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association*, 111(514):861–874, 2016.

[21] Tao Yu, Pengfei Li, and Jing Qin. Density estimation in the two-sample problem with likelihood ratio ordering. *Biometrika*, 104(1):141–152, 2017.

[22] Martin W McIntosh and Margaret Sullivan Pepe. Combining several screening tests: optimality of the risk score. *Biometrics*, 58(3):657–664, 2002.

[23] Shinto Eguchi and John Copas. A class of logistic-type discriminant functions. *Biometrika*, 89(1):1–22, 2002.

[24] Fushing Hsieh and Bruce W Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24(1):25–40, 1996.

[25] Kelly H Zou, WJ Hall, and David E Shapiro. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16(19):2143–2156, 1997.

[26] Amay SM Cheam and Paul D McNicholas. Modelling receiver operating characteristic curves using Gaussian mixtures. *Computational Statistics & Data Analysis*, 93:192–208, 2016.

[27] Mithat Gönen and Glenn Heller. Lehmann family of ROC curves. *Medical Decision Making*, 30(4):509–517, 2010.

[28] Alicja Jokiel-Rokita and Rafał Topolnicki. Estimation of the ROC curve from the Lehmann family. *Computational Statistics & Data Analysis*, page 106820, 2019.

[29] Jing Qin and Biao Zhang. Best combination of multiple diagnostic tests for screening purposes. *Statistics in Medicine*, 29(28):2905–2919, 2010.

[30] Chris J Lloyd. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, 93(444):1356–1364, 1998.

[31] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC Press, 1986.

[32] Chris J Lloyd and Zhou Yong. Kernel estimators of the ROC curve are better than empirical. *Statistics & Probability Letters*, 44(3):221–228, 1999.

[33] Xiao-Hua Zhou and Jaroslaw Harezlak. Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine*, 21(14):2045–2055, 2002.

[34] Peter G Hall and Rob J Hyndman. Improved methods for bandwidth selection when estimating ROC curves. *Statistics & Probability Letters*, 64(2):181–189, 2003.

[35] Liang Peng and Xiao-Hua Zhou. Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference*, 118(1-2):129–143, 2004.

[36] Jiezhun Gu, Subhashis Ghosal, and Anindya Roy. Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine*, 27(26):5407–5420, 2008.

[37] Alicja Jokiel-Rokita and Michał Pulit. Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions. *Statistics and Computing*, 23(6):703–712, 2013.

[38] Michał Pulit. A new method of kernel-smoothing estimation of the ROC curve. *Metrika*, 79(5):603–634, 2016.

[39] R John Irwin and Michael J Hautus. Lognormal Lorenz and normal receiver operating characteristic curves as mirror images. *Royal Society Open Science*, 2(2):140280, 2015.

[40] Edna Schechtman and Gideon Schechtman. The relationship between Gini terminology and the ROC curve. *Metron*, 77(3):171–178, 2019.

[41] Maurizia Mello-Grand, Ilaria Gregnanin, Lidia Sacchetto, Paola Ostano, Andrea Zitella, et al. Circulating microRNAs combined with PSA for accurate and non-invasive prostate cancer detection. *Carcinogenesis*, 40(2):246–253, 2018.

[42] FDA-NIH Biomarker Working Group et al. BEST (Biomarkers, EndpointS, and other Tools) resource. 2016.

[43] Margaret Sullivan Pepe and Mary Lou Thompson. Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140, 2000.

[44] Aiyi Liu, Enrique F Schisterman, and Yan Zhu. On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 24(1):37–47, 2005.

[45] Melina Arnold, Henrike E Karim-Kos, Jan Willem Coebergh, Graham Byrnes, Ahti Antilla, et al. Recent trends in incidence of five common cancers in 26 European countries since 1988: Analysis of the European Cancer Observatory. *European Journal of Cancer*, 51(9):1164–1187, 2015.

[46] Marc A Dall'Era, Peter C Albertsen, Christopher Bangma, Peter R Carroll, H Ballentine Carter, et al. Active surveillance for prostate cancer: a systematic review of the literature. *European Urology*, 62(6):976–983, 2012.

[47] Shinkan Tokudome, Ryosuke Ando, and Yoshiro Koda. Discoveries and application of prostate-specific antigen, and some proposals to optimize prostate cancer screening. *Cancer Management and Research*, 8:45, 2016.

[48] Hao Wang, Ran Peng, Junjie Wang, Zelian Qin, and Lixiang Xue. Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. *Clinical Epigenetics*, 10(1):1–10, 2018.

[49] Richard SC Liu, Ekaterina Olkhov-Mitsel, Renu Jeyapala, Fang Zhao, Kristina Commisso, et al. Assessment of serum microrna biomarkers to predict reclassification of prostate cancer in patients on active surveillance. *The Journal of Urology*, 199(6):1475–1481, 2018.

[50] Annika Fendler, Carsten Stephan, George M Yousef, Glen Kristiansen, and Klaus Jung. The translational potential of microRNAs as biofluid markers of urological tumours. *Nature Reviews Urology*, 13(12):734, 2016.

[51] Paolo Gontero, Giancarlo Marra, Francesco Soria, Marco Oderda, Andrea Zitella, et al. A randomized double-blind placebo controlled phase I–II study on clinical and molecular effects of dietary supplements in men with precancerous prostatic lesions. Chemoprevention or "chemopromotion"? *The Prostate*, 75(11):1177–1186, 2015.

[52] Marilesia Ferreira de Souza, Hellen Kuasne, Mateus de Camargo Barros-Filho, Heloísa Lizotti Cilião, Fabio A Marchi, et al. Circulating mRNAs and miRNAs as candidate markers for the diagnosis and prognosis of prostate cancer. *PloS One*, 12(9):e0184094, 2017.

[53] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[54] Jia-Xing Zhang, Wu Song, Zhen-Hua Chen, Jin-Huan Wei, Yi-Ji Liao, et al. Prognostic and predictive value of a microRNA signature in stage ii colon cancer: a microRNA expression analysis. *The Lancet Oncology*, 14(13):1295–1306, 2013.

[55] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[56] Tommaso Cavalleri, Laura Angelici, Chiara Favero, Laura Dioni, Carolina Mensi, et al. Plasmatic extracellular vesicle microRNAs in malignant pleural mesothelioma and asbestos-exposed subjects suggest a 2-miRNA signature as potential biomarker of disease. *PLoS One*, 12(5):e0176680, 2017.

[57] Xiumei Jiang, Lutao Du, Weili Duan, Rui Wang, Keqiang Yan, et al. Serum microRNA expression signatures as novel noninvasive biomarkers for prediction and prognosis of muscle-invasive bladder cancer. *Oncotarget*, 7(24):36733, 2016.

[58] Brittany L Mihelich, Joseph C Maranville, Rosalie Nolley, Donna M Peehl, and Larisa Nonn. Elevated serum microRNA levels associate with absence of high-grade prostate cancer in a retrospective cohort. *PloS One*, 10(4):e0124245, 2015.

[59] Prashant K Singh, Leah Preus, Qiang Hu, Li Yan, Mark D Long, et al. Serum microRNA expression patterns that predict early treatment failure in prostate cancer patients. *Oncotarget*, 5(3):824, 2014.

[60] Ye Chen, SuNing Chen, Jian Zhang, YangMin Wang, Zhengping Jia, et al. Expression profile of microRNAs in expressed prostatic secretion of healthy men and patients with IIIA chronic prostatitis/chronic pelvic pain syndrome. *Oncotarget*, 9(15):12186, 2018.

[61] Dejuan Kong, Elisabeth Heath, Wei Chen, Michael L Cher, Isaac Powell, et al. Loss of let-7 up-regulates EZH2 in prostate cancer consistent with the acquisition of cancer stem cell signatures that are attenuated by BR-DIM. *PloS One*, 7(3):e33729, 2012.

[62] Edgars Endzelins, Andreas Berger, Vita Melne, Cristina Bajo-Santos, Kristine Sobolevska, et al. Detection of circulating miRNAs: comparative analysis of extracellular vesicle-incorporated miRNAs and cell-free miRNAs in whole plasma of prostate cancer patients. *BMC Cancer*, 17(1):730, 2017.

[63] Diederick Duijvesz, Theo Luider, Chris H Bangma, and Guido Jenster. Exosomes as biomarker treasure chests for prostate cancer. *European Urology*, 59(5):823–831, 2011.

[64] Xiaoyi Huang, Tiezheng Yuan, Meihua Liang, Meijun Du, Shu Xia, et al. Exosomal miR-1290 and miR-375 as prognostic markers in castration-resistant prostate cancer. *European Urology*, 67(1):33–41, 2015.

[65] David C Grossman, Susan J Curry, Douglas K Owens, Kirsten Bibbins-Domingo, Aaron B Caughey, et al. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *JAMA*, 319(18):1901–1913, 2018.

[66] Nicolas Mottet, Joaquim Bellmunt, Michel Bolla, Erik Briers, Marcus G Cumberbatch, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent. *European Urology*, 71(4):618–629, 2017.

[67] Jeffrey J Tosoian, Sasha C Druskin, Darian Andreas, Patrick Mullane, Meera Chappidi, et al. Prostate Health Index density improves detection of clinically significant prostate cancer. *BJU International*, 120(6):793–798, 2017.

[68] Pär Stattin, Andrew J Vickers, Daniel D Sjoberg, Robert Johansson, Torvald Granfors, et al. Improving the specificity of screening for lethal prostate cancer using prostate-specific antigen and a panel of kallikrein markers: a nested case–control study. *European Urology*, 68(2):207–213, 2015.

[69] Alicia C McDonald, Manish Vira, Jing Shen, Martin Sanda, Jay D Raman, et al. Circulating microRNAs in plasma as potential biomarkers for the early detection of prostate cancer. *The Prostate*, 78(6):411–418, 2018.

[70] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[71] Charles E Metz and Xiaochuan Pan. "Proper" binormal ROC curves: theory and maximum-likelihood estimation. *Journal of Mathematical Psychology*, 43(1):1–33, 1999.

[72] Stephen L Hillis. Equivalence of binormal likelihood-ratio and bi-chi-squared ROC curve models. *Statistics in Medicine*, 35(12):2031–2057, 2016.

[73] David J Hand and Keming Yu. Idiot's Bayes - Not so stupid after all? *International Statistical Review*, 69(3):385–398, 2001.

[74] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.

[75] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[76] Donato M Cifarelli and Eugenio Regazzini. On a general definition of concentration function. *Sankhya: The Indian Journal of Statistics, Series B*, 49(3):307–319, 1987.

[77] Eugenio Regazzini. Concentration comparisons between probability measures. *Sankhya: The Indian Journal of Statistics, Series B*, 54(2):129–149, 1992.

[78] Mauro Gasparini and Lidia Sacchetto. On the definition of a concentration function relevant to the ROC curve. *arXiv preprint arXiv:2001.00944*, 2020.

[79] Shlomo Yitzhaki and Edna Schechtman. *The Gini methodology: A primer on a statistical methodology*, volume 272. Springer Science & Business Media, 2012.

[80] Theodore Colton and Peter Armitage. *Encyclopedia of biostatistics*. John Wiley & Sons, 2005.

[81] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[82] Lidia Sacchetto and Mauro Gasparini. ROC curves with binary multivariate data. In Simona Balzano Giovanni Porzio, Francesca Greselin, editor, *Book of Short Papers - CLADAG*, pages 420 – 423. Edizioni Universitá di Cassino, 2019.

[83] Alessandro Barilli. Maximum likelihood based clustering via parallel computing. Master's thesis, Alta Scuola Politecnica, Politecnico di Torino, Turin, 2019.

[84] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer Series in Statistics, New York, 2001.

[85] Alan J Izenman. *Modern Multivariate Statistical Techniques - Regression, Classification, and Manifold Learning*. Springer Science and Business Media, New York, 2008.

[86] Charles Bouveyron, Gilles Celeux, T Brendan Murphy, and Adrian E Raftery. *Model-Based Clustering and Classification for Data Science: With Applications in R*, volume 50. Cambridge University Press, 2019.

[87] Paul F Lazarsfeld and Neil W Henry. *Latent structure analysis*. Houghton Mifflin Co., 1968.

[88] Kanti V Mardia, John T Kent, and John M Bibby. Multivariate analysis. *Probability and Mathematical Statistics. Academic Press Inc*, 1979.

[89] Allen J Scott and Michael J Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397, 1971.

[90] Alan Agresti and Brent A Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

[91] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[92] Leo A Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

[93] Ajit C Tamhane, Dingxi Qiu, and Bruce E Ankenman. A parametric mixture model for clustering multivariate binary data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(1):3–19, 2010.

[94] Drew A Linzer and Jeffrey B Lewis. poLCA: an R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10):1–29, 2011.

[95] David J Bartholomew, Fiona Steele, and Irini Moustaki. *Analysis of multivariate social science data*. Chapman and Hall/CRC, 2008.

[96] Tao Li, Sheng Ma, and Mitsunori Ogihara. Entropy-based criterion in categorical clustering. In *Proceedings of the twenty-first International Conference on Machine Learning*, page 68. ACM, 2004.

[97] Lidia Sacchetto, Roberto Zanetti, Harry Comber, Christine Bouchardy, David Brewster, et al. Trends in incidence of thick, thin and in situ melanoma in Europe. *European Journal of Cancer*, 92:108–118, 2018.

[98] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[99] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[100] Mark Segal and Yuanyuan Xiao. Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):80–87, 2011.

[101] Francois Chollet and JJ Allaire. *Deep Learning with R*. Manning Publications Co., USA, 1st edition, 2018.

[102] Python Software Foundation. Python Language Reference, version 2.7. http://www.python.org.

[103] Max D Parkin, Jacques Ferlay, Maria-Paula Curado, Freddie Bray, Brenda Edwards, et al. Fifty years of cancer incidence: CI5 I–IX. *International Journal of Cancer*, 127(12):2918–2927, 2010.

[104] Friederike Erdmann, Joannie Lortet-Tieulent, Joachim Schüz, Hajo Zeeb, Rüdiger Greinert, et al. International trends in the incidence of malignant melanoma 1953–2008—are recent generations at higher or lower risk? *International Journal of Cancer*, 132(2):385–400, 2013.

[105] Cynthia Holterhues, Loes M Hollestein, Tamar Nijsten, Elsje R Koomen, Wilma Nusselder, et al. Burden of disease due to cutaneous melanoma has increased in the Netherlands since 1991. *British Journal of Dermatology*, 169(2):389–397, 2013.

[106] Eleni Linos, Susan M Swetter, Myles G Cockburn, Graham A Colditz, and Christina A Clarke. Increasing burden of melanoma in the United States. *Journal of Investigative Dermatology*, 129(7):1666–1674, 2009.

[107] Alan C Geller, Richard W Clapp, Arthur J Sober, Lou Gonsalves, Lloyd Mueller, et al. Melanoma epidemic: an analysis of six decades of data from the Connecticut Tumor Registry. *Journal of Clinical Oncology*, 31(33):4172, 2013.

[108] Stefano Rosso, Lidia Sacchetto, Adriano Giacomin, Roberto Foschi, Roberta De Angelis, et al. Estimates of cancer burden in Piedmont and Aosta Valley. *Tumori Journal*, 99(3):269–276, 2013.

[109] Rafael Marcos-Gragera, Neus Vilar-Coromina, Jaume Galceran, Joan Borràs, Ramòn Clèries, et al. Rising trends in incidence of cutaneous malignant melanoma and their future projections in Catalonia, Spain: increasing impact or future epidemic? *Journal of the European Academy of Dermatology and Venereology*, 24(9):1083–1088, 2010.

[110] Andrea Bordoni, Sandra Leoni-Parvex, Simona Peverelli, Paola Mazzola, Luca Mazzucchelli, et al. Opportunistic screening strategy for cutaneous melanoma does not change the incidence of nodular and thick lesions nor reduce mortality: a population-based descriptive study in the European region with the highest incidence. *Melanoma Research*, 23(5):402–407, 2013.

[111] Alexander Armstrong, Chris Powell, Roy Powell, Nicky Hallam, James Taylor, et al. Are we seeing the effects of public awareness campaigns? A 10-year analysis of Breslow thickness at presentation of malignant melanoma in the South West of England. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 67(3):324–330, 2014.

[112] Anne Kricker, Bruce K Armstrong, Chris Goumas, Nancy E Thomas, Lynn From, et al. Survival for patients with single and multiple primary melanomas: the genes, environment, and melanoma study. *JAMA Dermatology*, 149(8):921–927, 2013.

[113] Melina Arnold, Cynthia Holterhues, Loes M Hollestein, Jan Willem Coebergh, Tamar Nijsten, , et al. Trends in incidence and predictions of cutaneous melanoma across Europe up to 2015. *Journal of the European Academy of Dermatology and Venereology*, 28(9):1170–1178, 2014.

[114] Michael Coory, Peter Baade, Joanne Aitken, Mark Smithers, G Roderick C McLeod, et al. Trends for in situ and invasive melanoma in Queensland, Australia, 1982–2002. *Cancer Causes & Control*, 17(1):21–27, 2006.

[115] Peter Baade, Xingqiong Meng, Danny Youlden, Joanne Aitken, and Philippa Youl. Time trends and latitudinal differences in melanoma thickness distribution in Australia, 1990–2006. *International Journal of Cancer*, 130(1):170–178, 2012.

[116] Anthony Montella, Anna Gavin, Richard Middleton, Philippe Autier, and Mathieu Boniol. Cutaneous melanoma mortality starting to change: a study of trends in Northern Ireland. *European Journal of Cancer*, 45(13):2360–2366, 2009.

[117] Emanuele Crocetti, Adele Caldarella, Alessandra Chiarugi, Paolo Nardini, and Marco Zappa. The thickness of melanomas has decreased in central Italy, but only for thin melanomas, while thick melanomas are as thick as in the past. *Melanoma Research*, 20(5):422–426, 2010.

[118] Andrea Ambrosini-Spaltro, Tomas Dal Cappello, Jenny Deluca, Cinzia Carriere, Guido Mazzoleni, et al. Melanoma incidence and breslow tumour thickness development in the central Alpine region of South Tyrol from 1998 to 2012: a population-based study. *Journal of the European Academy of Dermatology and Venereology*, 29(2):243–248, 2015.

[119] Helen L Hunter, Olivia Dolan, Elizabeth McMullen, David Donnelly, and Anna Gavin. Incidence and survival in patients with cutaneous malignant melanoma: experience in a UK population, 1984–2009. *British Journal of Dermatology*, 168(3):676–678, 2013.

[120] Neel M Helvind, Lisbet Rosenkrantz Hölmich, Sigrun Smith, Martin Glud, Klaus K Andersen, et al. Incidence of in situ and invasive melanoma in Denmark from 1985 through 2012: a national database study of 24059 melanoma cases. *JAMA Dermatology*, 151(10):1087–1095, 2015.

[121] Hreinn Stefansson, Laufey Tryggvadottir, Elinborg J Olafsdottir, Ellen Mooney, Jon H Olafsson, et al. Cutaneous melanoma in Iceland: changing Breslow's tumour thickness. *Journal of the European Academy of Dermatology and Venereology*, 29(2):346–352, 2015.

[122] WHO Mortality Database, World Health Organization. https://www.who.int/healthinfo/mortality_data/en/. Accessed: 8 June 2017.

[123] Waqas R Shaikh, Martin A Weinstock, Allan C Halpern, Susan A Oliveria, Alan C Geller, et al. The characterization and potential impact of melanoma cases with unknown thickness in the United States' Surveillance, Epidemiology, and End Results Program, 1989–2008. *Cancer Epidemiology*, 37(1):64–70, 2013.

[124] Arjen Joosse, Sandra Collette, Stefan Suciu, Tamar Nijsten, Poulam M Patel, et al. Sex is an independent prognostic indicator for survival and relapse/progression-free survival in metastasized stage III to IV melanoma: a pooled analysis of five European organisation for research and treatment of cancer randomized controlled trials. *Journal of Clinical Oncology*, 31(18):2337–46, 2013.

[125] Arjen Joosse, Augustinus PT van der Ploeg, Lauren E Haydu, Tamar Nijsten, Esther de Vries, et al. Sex differences in melanoma survival are not related to mitotic rate of the primary tumor. *Annals of Surgical Oncology*, 22(5):1598–1603, 2015.

[126] Loes M Hollestein, Sanne A Van den Akker, Tamar Nijsten, Henrike E Karim-Kos, Jan Willem Coebergh, et al. Trends of cutaneous melanoma in The Netherlands: increasing incidence rates among all breslow thickness categories and rising mortality rates since 1989. *Annals of Oncology*, 23(2):524–30, Feb 2012.

[127] Johan Lyth, Hanna Eriksson, Johan Hansson, Christian Ingvar, Malin Jansson, et al. Trends in cutaneous malignant melanoma in Sweden 1997-2011: thinner tumours and improved survival among men. *British Journal of Dermatology*, 172(3):700–6, Mar 2015.

[128] Anita Toender, Susanne K Kjær, and Allan Jensen. Increased incidence of melanoma in situ in Denmark from 1997 to 2011: results from a nationwide population-based study. *Melanoma Research*, 24(5):488–95, Oct 2014.

[129] Christiane Bay, Anne Mette Tranberg Kejs, Hans H Storm, and Gerda Engholm. Incidence and survival in patients with cutaneous melanoma by morphology, anatomical site and TNM stage: a Danish population-based register study 1989-2011. *Cancer Epidemiology*, 39(1):1–7, Feb 2015.

[130] Remo Minini, Sabine Rohrmann, Ralph Braun, Dimitri Korol, and Silvia Dehler. Incidence trends and clinical-pathological characteristics of invasive cutaneous melanoma from 1980 to 2010 in the canton of Zurich, Switzerland. *Melanoma Research*, 27(2):145–151, Apr 2017.

[131] Robert JT van der Leest, Judith Zoutendijk, Tamar Nijsten, Wolter J Mooi, Jasper I van der Rhee, et al. Increasing time trends of thin melanomas in The Netherlands: What are the explanations of recent accelerations? *European Journal of Cancer*, 51(18):2833–41, Dec 2015.

[132] Jelena Barbaric, Mario Sekerija, Dominic Agius, Daniela Coza, Nadya Dimitrova, et al. Disparities in melanoma incidence and mortality in South-Eastern Europe: Increasing incidence and divergent mortality patterns. Is progress around the corner? *European Journal of Cancer*, 55:47–55, Mar 2016.

[133] Vesna Zadnik, Maja P Zakelj, Katarina Lokar, Katja Jarm, Urska Ivanus, et al. Cancer burden in slovenia with the time trends analysis. *Radiology and Oncology*, 51(1):47–55, 2017.

[134] Ofelia Suteu, Mihaiela L Blaga, Florian Nicula, Patricia Suteu, Ovidiu Coza, et al. Incidence trends and survival of skin melanoma and squamous cell carcinoma in Cluj County, Romania. *European Journal of Cancer Prevention*, 26:S176–S182, 2017.

[135] Alexander Katalinic, Nora Eisemann, and Annika Waldmann. Skin cancer screening in Germany: documenting melanoma incidence and mortality from 2008 to 2013. *Deutsches Ärzteblatt International*, 112(38):629, 2015.

[136] Trude E Robsahm, Gjøril Bergva, Unn E Hestvik, and Bjørn Møller. Sex differences in rising trends of cutaneous malignant melanoma in Norway, 1954–2008. *Melanoma Research*, 23(1):70–78, 2013.

[137] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5):E359–E386, 2015.

[138] Laufey Tryggvadóttir, Mette Gislum, Timo Hakulinen, Åsa Klint, Gerda Engholm, et al. Trends in the survival of patients diagnosed with malignant melanoma of the skin in the Nordic Countries 1964–2003 followed up to the end of 2006. *Acta Oncologica*, 49(5):665–672, 2010.

[139] Mary J Sneyd and Brian Cox. A comparison of trends in melanoma mortality in New Zealand and Australia: the two countries with the highest melanoma incidence and mortality in the world. *BMC Cancer*, 13(1):372, 2013.

[140] Philippe Autier, Alice Koechlin, and Mathieu Boniol. The forthcoming inexorable decline of cutaneous melanoma mortality in light-skinned populations. *European Journal of Cancer*, 51(7):869–878, 2015.

[141] Simone Mocellin and Donato Nitti. Cutaneous melanoma in situ: translational evidence from a large population-based study. *The Oncologist*, 16(6):896–903, 2011.

[142] Waqas R Shaikh, Stephen W Dusza, Martin A Weinstock, Susan A Oliveria, Alan C Geller, et al. Melanoma thickness and survival trends in the United States, 1989–2009. *JNCI: Journal of the National Cancer Institute*, 108(1), 2016.

[143] David C Whiteman, Peter D Baade, and Catherine M Olsen. More people die from thin melanomas ([= or<, slanted] 1 mm) than from thick melanomas (> 4 mm) in queensland, australia. *Journal of Investigative Dermatology*, 135(4):1190, 2015.

[144] Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1973 - 2014), National Cancer Institute, DCCPS, Surveillance Research Program. www.seer.cancer.gov.

[145] Phyllis A Gimotty, Ronald Shore, Nancy L Lozon, Jeanne Whitlock, Sidan He, et al. Miscoding of melanoma thickness in SEER: research and clinical implications. *Journal of Investigative Dermatology*, 136(11):2168–2172, 2016.

[146] Rüdiger Greinert, Esther De Vries, Friederike Erdmann, Carolina Espina, Anssi Auvinen, et al. European Code against Cancer 4th edition: Ultraviolet radiation and cancer. *Cancer Epidemiology*, 39:S75–S83, 2015.

[147] Charles M Balch, Jeffrey E Gershenwald, Seng-jaw Soong, John F Thompson, Michael B Atkins, et al. Final version of 2009 AJCC melanoma staging and classification. *Journal of Clinical Oncology*, 27(36):6199, 2009.

[148] Adèle C Green, Peter Baade, Michael Coory, Joanne F Aitken, and Mark Smithers. Population-based 20-year survival among people diagnosed with thin melanomas in Queensland, Australia. *Journal of Clinical Oncology*, 30(13):1462–1467, 2012.

[149] Alicia Brunssen, Lina Jansen, Nora Eisemann, Annika Waldmann, Janick Weberpals, et al. A population-based registry study on relative survival from melanoma in Germany stratified by tumor thickness for each histologic subtype. *Journal of the American Academy of Dermatology*, 80(4):938–946, 2019.

[150] Vincent D Criscione and Martin A Weinstock. Melanoma thickness trends in the United States, 1988–2006. *Journal of Investigative Dermatology*, 130(3):793–797, 2010.

[151] Ahmedin Jemal, Mona Saraiya, Pragna Patel, Sai S Cherala, Jill Barnholtz-Sloan, et al. Recent trends in cutaneous melanoma incidence and death rates in the United States, 1992-2006. *Journal of the American Academy of Dermatology*, 65(5):S17–e1, 2011.

[152] Shoshana M Landow, Annie Gjelsvik, and Martin A Weinstock. Mortality burden and prognosis of thin melanomas overall and by subcategory of thickness, SEER registry data, 1992-2013. *Journal of the American Academy of Dermatology*, 76(2):258–263, 2017.

[153] Laura J Gardner, Jennifer L Strunck, Yelena P Wu, and Douglas Grossman. Current controversies in early-stage melanoma: Questions on incidence, screening, and histologic regression. *Journal of the American Academy of Dermatology*, 80(1):1–12, 2019.

[154] Philipp GH Metnitz, Paul Zajic, and Andrew Rhodes. The General Data Protection Regulation and its effect on epidemiological and observational research. *The Lancet Respiratory Medicine*, 8(1):23–24, 2020.

[155] Wolfgang Weyers. The 'epidemic' of melanoma between under- and overdiagnosis. *Journal of Cutaneous Pathology*, 39(1):9–16, 2011.

[156] Gerardo Ferrara and Iris Zalaudek. Is histopathological overdiagnosis of melanoma a good insurance for the future? *Melanoma Management*, 2:21–25, 2015.

[157] Trude E Robsahm, Per Helsing, Yngvar Nilssen, Linda Vos, Syed MH Rizvi, et al. High mortality due to cutaneous melanoma in Norway: a study of prognostic factors in a nationwide cancer registry. *Clinical Epidemiology*, 10:537, 2018.

[158] Christer Lindholm, Ronny Andersson, Monika Dufmats, Johan Hansson, Christian Ingvar, et al. Invasive cutaneous malignant melanoma in Sweden, 1990–1999: A prospective, population-based study of survival and prognostic factors. *Cancer*, 101(9):2067–2078, 2004.

[159] Michael Lattanzi, Yesung Lee, Danny Simpson, Una Moran, Farbod Darvishian, et al. Primary melanoma histologic subtype: impact on survival and response to therapy. *JNCI: Journal of the National Cancer Institute*, 111(2):180–188, 2019.

[160] Hanna Eriksson, Margareta Frohm-Nilsson, Jacob Järås, Lena Kanter-Lewensohn, Petra Kjellman, et al. Prognostic factors in localized invasive primary cutaneous malignant melanoma: results of a large population-based study. *British Journal of Dermatology*, 172(1):175–186, 2015.

[161] Laufey Tryggvadóttir, Mette Gislum, Timo Hakulinen, Åsa Klint, Gerda Engholm, et al. Trends in the survival of patients diagnosed with malignant melanoma of the skin in the Nordic countries 1964–2003 followed up to the end of 2006. *Acta Oncologica*, 49(5):665–672, 2010.

[162] AIRTUM Working Group et al. Italian cancer figures, report 2007: survival. *Epidemiologia e Prevenzione*, 31(Suppl 1), 2007.

[163] AIRTUM Working Group et al. Italian cancer figures, report 2011: Survival of cancer patients in Italy. *Epidemiologia e Prevenzione*, 35(5-6 Suppl 3):1, 2011.

[164] Vesna Zadnik, Maja P Zakelj, Katarina Lokar, Katja Jarm, Urska Ivanus, et al. Cancer burden in Slovenia with the time trends analysis. *Radiology and Oncology*, 51(1):47–55, 2017.

[165] Zahra Hanaizi, Barbara van Zwieten-Boot, Gonzalo Calvo, Arantxa Sancho Lopez, Maaike van Dartel, et al. The European Medicines Agency review of ipilimumab (Yervoy) for the treatment of advanced (unresectable or metastatic) melanoma in adults who have received prior therapy: summary of the scientific assessment of the Committee for Medicinal Products for Human Use. *European Journal of Cancer*, 48(2):237–242, 2012.

[166] Valentina Appierto, Maurizio Callari, Elena Cavadini, Daniele Morelli, Maria Grazia Daidone, et al. A lipemia-independent NanoDrop®-based score to identify hemolysis in plasma and serum samples. *Bioanalysis*, 6(9):1215–1226, 2014.