Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN
Scienze Statistiche

XXXI Ciclo

Settore Concorsuale: 13/D1

Settore Scientifico disciplinare: SECS-S/01

ROC curves and the generalization to multiple classes

Presentata da
Elena Nardi

Coordinatore Dottorato
Prof.ssa Alessandra Luati

Supervisore
Prof.ssa Rossella Miglio

Esame finale anno 2019

Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. Rossella Miglio for her steady guidance, support and patience. It was a really great pleasure to work with her during my PhD experience. I would also like to thank Prof. Alessandra Luati for her encouragement to work with rigor and devotion.

I'm grateful to my friends: Vincenza, for being my inspirer, Sara to have kept my spirit high during the three years and Mario for his relaxed point of view during our statistical chats.

I wish to say a special thank to my sister for always believing in me.

Finally, my special thanks to the big loves of my life: my husband, without whose constant motivation and deep love all this would not have been possible and my sons, my greatest supporters. I hope that this experience will teach them that with determination, dignity and self-love great goals can be achieved at every season of the life.

Abstract

The present work focuses on the study and extension of ROC analysis methodology for multiple-class classification problems. In clinical medical research, the need for developing an approach to measure the diagnostic accuracy of biomedical tests in classifying the true status of a patient is a critical point when doing both diagnosis and prognosis. In a two-category classification setting, the ROC analysis is the natural approach and the Area Under the Curve (AUC) is a summary measure of the diagnostic accuracy. However, many real classification problems rely to more than two classes; thus, the ROC manifold generalization of curve and the hypervolume (HUM) generalization of area recently appeared in the literature to address classification problems with more than two classes. Motivated by a real research question arose during a four-class classification study for early detection of colorectal cancer, we review the literature on ROC analysis and on its extension to multiple classes. Then, we develop a new estimator of the accuracy measure of a diagnostic marker. We derive the analytical form of the HUM estimator and the analytical representation of its variance. To assess the performance of the proposed estimator and compare it with the two alternatives existing in the literature, we perform simulation exercises and empirical applications. The first application deals with the topic that initially moved our interest, the early detection of colorectal cancer patients; the second concerns the classification of synovial tissue inflammatory cells, a typical case study in the biostatistics literature. Finally, in the last part of our work, we suggest a statistical method to combine multiple tests for multicategory classification. The novelty of our approach is the use of the classification accuracy (HUM) of the combined marker as the objective function to be maximized. The methodology is evaluated trough a simulation study and two empirical applications.

Contents

Introduction

In clinical studies the accurate diagnosis of a patient condition is crucial for an appropriate treatment, as well as to evaluate the prognosis. Thus, before implementing a new test, it is of primary importance to quantify how well the medical test discriminates among different status. In practical medicine there exist many kinds of diagnostic tests, the simplest we can imagine are, for example, serum ferritin levels in blood to check for anaemia, diagnostic imaging tests such as mammogram to detect any breast abnormalities, or faecal immunochemical test to detect colorectal cancer risk patients. More complicated data, such as the genetic expression profile or a clinical score obtained as result of different simple tests, can be even considered as diagnostic test and could be used, for instance, to establish the severity degree of a particular disease.

Although the interest in the present work is mainly on the clinical research, these problems are part of the general classification issues that could arise in almost all the fields of scientific and social research. The procedures considered in this thesis, in fact, could be generalized to whatever procedure of classification that assigns a subject or an object to a class on the basis of the information observed.

It could happen that the diagnostic test, or more generally, the assignment procedure, might fail and assign the individual to an incorrect class. It becomes, thus, fundamental to measure the qualitative performance of the procedure. Obviously, the stakes in clinical classification are extremely high and a quantification of the risk of an erroneous classification would help to evaluate the accuracy and implementability of the diagnostic strategy. As it is to imagine, the list of situations in which we pursue this objective is practically unlimited; in this thesis we will refer to medical and clinical situations especially in the applications of the proposed methodologies to real data.

In order to depict the quality of a diagnostic marker or a diagnostic test in a supervised classification problem, the Receiver Operating Characteristic (ROC) curve analysis plays a prominent role. This analysis was introduced

in the second half of the last century in a two class classification problem. It consists of a graphical representation of the relationship between *sensitivity* and *specificity* of a test as the cut-off of the test varies. At each value of the decision making threshold, the curve depicts the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity).

The AUC, Area Under the Curve, is perhaps the most frequently used summary measure of the information reported in the ROC curve. It is a global measure of the performance of a diagnostic marker in discriminating the two status. An alternative interpretation refers to the measure of separability between the statistical distributions of the diagnostic test in the two populations.

Nowadays, ROC curve and AUC are still major instruments in the evaluation of a twofold classifier. However, many real situations in diagnostic decisions are not limited to a binary choice; an example could be the case of staging a level of a illness or classify a subject as low risk, moderate risk or high risk for a certain pathology. To address this more complicated m-class classification problem, later in the last century, different contributions in ROC analysis focused on deriving suitable generalizations of the curve. The ROC surface has been introduced to cope with three-class issues while the ROC manifold when more than three classes were considered. Consequently, the notion of the Area Under the Curve has been extended to the Volume Under the Surface (VUS) and, in the more complex situations of more than three classes, to the hypervolume (HUM). From a statistical perspective, (theoretical) inferential studies about generalized ROC analysis appear only at the beginning of 2000. Furthermore, since those first works, only few theoretical and empirical contributions have been developed in the literature, leaving the four-class classification issue almost unexplored. The present work mainly focuses on this specific topic.

The idea of dealing with classifying subjects in a four-class framework was born few years ago while we were investigating the classification ability of biomarkers in detecting colorectal cancer. In that specific case, the population was divided in four groups according to the level of the disease: healthy subjects, positives to the faecal immunochemical test with negative colonoscopy, positives to the faecal immunochemical test with small polyps, faecal immunochemical test with confirmed diagnosis of colorectal cancer. In situations like that, the largely used approach to conduct ROC analysis consists in reducing the dimensionality of the problem by pursuing pairwise two-class ROC curve. Although its simplicity, the standard approach has the undesirable limit of attacking the problem by considering a subset of the entire sample at the time, and, perhaps, ignoring hidden patterns detectable only through a detailed analysis of the whole sample. The aim of our work,

thus, is to shed light on the state of the art on the four-class classification literature and to propose a new methodology to address the problem.

The thesis is organized in four main parts. The first one introduces the problem through a detailed review of the literature; in particular, in Chapter 1 we present the ROC curves theory in the dichotomous case while Chapter 2 summarizes the relevant literature on the generalization of ROC analysis. The second part, Chapter 3, is dedicated to the presentation of the statistical methodology we propose; specifically, we show how we derive an estimator of the volume under the ROC surface in a four-class framework and, moreover, how we derive the analytical form of the variance of the estimator. The third part, instead, is devoted to the evaluation of the performances of our estimator. In Chapter 4 we perform some simulation exercises under different data generating processes, while in chapter 5 we apply our methodology to real data concerning two clinical studies. In both simulated and real data, we compare, our methodology to other two alternatives already existing in the literature. Finally, the last part, Chapter 6, addresses the purpose of combining multiple tests for multiple-category classification. We propose a new statistical approach, a simulation exercise and an application to real data.

Chapter 1

ROC curves and dichotomous outcome

Receiver Operating Characteristic (ROC) analysis was born and developed in statistical decision theory and later applied to signal detection theory. Lusted (1960), for the first time, introduced the ROC curve in the field of diagnostic medicine to distinguish between the criteria that an observer uses to decide whether or not a condition is present and the observer's ability for detecting the condition. Nowadays, the ROC curve has become a standard tool in medicine for evaluating the diagnostic accuracy of a classifier and it is widely treated in many books (Pepe, 2003).

Medical tests, such as biomarkers for cancer, standard biochemical measurements in blood test or subjective probability estimates by a physician who makes diagnostic judgments, can result in binary, ordinal or continuous measures depending on the level of measurement of each variable. Even though, they are often used to make a medical decision in order to classify a subject in a dichotomous way (i.e. diseased/healthy or to be treated/not treated, presence or absence of pain). Therefore, to make a dichotomous decision based on a continuous or ordinal test, a decision rule involving the choice of a threshold, is needed. In most cases the choice of the threshold will depend on the trade-off between failing to detect and falsely identifying the ill. Such a decision often depends on specific circumstances that may change in time, thus is helpful to have a way of displaying and summarizing the performance of the test over a wide range of conditions. The ROC curve is a useful device that simply describes the diagnostic test performance as the choice of the threshold varies. In this chapter we revise the ROC topic before extend it to accommodate problems of multiple-class classification.

1.1 Definition of ROC curve

Although ROC curves are defined for both continuous and ordinal data, without loss of generality, in this work we will focus on continuous test results only. Using the correct diagnostic test to implement a dichotomous decision rule is of fundamental importance to make medical decisions. Let think, for example, to the very common situation where the decision is whether a patient should be treated or not; the decision rule is based on whether or not a test result is greater of a fixed threshold. For example, in evaluating thyroid function and diagnose thyroid disease, the Thyroxine test (commonly indicated by T4) is performed. It consists in measuring the amount of thyroxine (a thyroid hormone) in the blood. Depending on the level of T4 one fixes as pathologic threshold, the subjects will be classified as hypothyroid or euthyroid and consequently treated or not. Generalizing, let X be the continuous diagnostic test result with probability density function $f(x)$, cumulative distribution function $F(x)$ and survival function $S(x) = 1 - F(x)$. Suppose also that larger values of the test are more indicative of disease. The corresponding binary test is defined, according to a particular classification rule, as positive if $X \geq c$ and negative if $X < c$, where c is a fixed threshold. If the diagnostic test is positive, the individual will be classified in the diseased population ($D = 1$), if the test is negative will be classified in the healthy (non diseased) population ($D = 0$). In order to asses the efficacy of the test in classifying subjects, we need to calculate the probability of making an error of misclassification. Such a probability tells us the rate at which a new individual will be misallocated. More generally, at each threshold c, we can define four probabilities associated with the diagnostic test:

$$\begin{aligned}
\text{True Positive Rate:} \quad & \text{TPR}(c) = P[X \geq c | D = 1] \\
\text{False Positive Rate:} \quad & \text{FPR}(c) = P[X \geq c | D = 0] \\
\text{True Negative Rate:} \quad & \text{TNR}(c) = P[X < c | D = 0] \\
\text{False Negative Rate:} \quad & \text{FNR}(c) = P[X < c | D = 1].
\end{aligned}$$

A true positive arises when a subject belonging to class 1 presents a test measure greater than the threshold and is correctly assigned to class 1, a false positive arises when a subject which really belongs to class 0 is assigned to class 1 because his test score falls above the threshold. Evaluating the four quantities above at different values of c will provide full information about the classification performance of the diagnostic test. Given that $TPR+FNR = 1$ and $FPR + TNR = 1$, knowing TPR and FPR is enough to summarize all the information about the classification test. The TPR is also known as

the *sensitivity* of the test while the TNR is the *specificity* of the test. In biomedical research, sensitivity and specificity (1-FPR) are commonly used as descriptive measures of the test performance. The ROC curve is the curve obtained by plotting on orthogonal axis the set of all possible true positive (sensitivity) and false positive (1-specificity) fractions that are attainable dichotomizing X with different thresholds:

$$ROC(\cdot) = \{(FPR(c), TPR(c); \quad c \in (-\infty, \infty))\}$$

where $\lim_{c\to\infty} TPR(c) = 0$ and $\lim_{c\to\infty} FPR(c) = 0$ while $\lim_{c\to-\infty} TPR(c) = 1$ and $\lim_{c\to-\infty} FPR(c) = 1$.

In a continuous domain, we can rewrite the ROC curve as the function that maps p into t, where c is the threshold such that $p = FPR(c)$ and $t = TPR(c)$

$$ROC(\cdot) = \{p, ROC(p); \quad p \in (0, 1)\}$$

or alternatively in a more compact form

$$t = h(p)$$

where t is the true positive rate corresponding to the p false positive rate at threshold c. When there is not possibility to chose a particular threshold for the test result to categorize it as positive, then a ROC curve is very useful in providing a complete description of the possible operating characteristics of the test.

1.1.1 Principal properties of the ROC curve

We now report some important properties of the ROC curve that can be useful further on:

- ROC curve is a monotone increasing function in the positive quadrant, lying between the points $(0, 0)$ and $(1, 1)$.
- ROC curve is invariant to strictly increasing transformations of X.
- ROC curve has a functional relation with the survivor function. Let S_1 and S_0 be the survivor functions of X in the diseased and healthy populations respectively: $S_1(x) = P[X \geq y|D = 1]$ and $S_0(x) = P[X \geq x|D = 0]$, then the ROC curve can be represented as:

$$ROC(p) = S_1(S_0^{-1}(p)) \quad p \in (0, 1). \tag{1.1}$$

  Being the survivor function $S(x) = 1 - F(x)$ where $F(x)$ is the cumulative distribution function of the test random variable X, the *ROC*

may also be interpreted as the curve which summarizes the information on the cumulative distribution functions of the test measures in the two classes.

- The ROC's slope is

$$\frac{\partial ROC(p)}{\partial p} = \frac{f_1(S_0^{-1}(p))}{f_0(S_0^{-1}(p))} = \frac{f_1(x)}{f_0(x)}$$

where f_1 is the probability density function of X in the diseased population and f_0 is the probability density function of X in the healthy population. Interestingly, the slope of the curve can also be interpreted as the likelihood ratio at the threshold x

$$\mathcal{LR}(x) = \frac{f_1}{f_0}$$

where $x = S_0^{-1}(p)$ is the threshold corresponding to the point $(p, ROC(p))$.

From the last property we can conclude that the slope of the ROC curve in a generic point x, tells us how much more probable is a value x of the test to have occurred in the diseased population than in healthy population.

Moreover, a connection with the Neyman-Pearson theory of hypothesis testing can be deduced. Consider the case of testing the null hypothesis H_0 that a subject belongs to the healthy population $D = 0$ against the alternative, H_1, that subject belongs to population $D = 1$. The classification test X is performed to allocate the individuals. Let S_R be the set of values of X for which the subject is allocated in population $D = 1$ (the rejection region), then the Neyman-Pearson lemma states that the most powerful test of size α is the test that has $S_R = \{x : \mathcal{LR}(x) \geq k\}$ where k is determined under the condition of $P(x \in S_R | D = 0) = \alpha$. Thus, in our diagnostic framework, $P(x \in S_R | D = 0) = \alpha$ if and only if $FPR = \alpha$, consequently FPR is the size of the test while TPR is the power of the test.

This connection with the Neyman-Pearson theory assures that a classification rule based on $\mathcal{LR}(x) \geq k$ is an optimal decision rule for classifying subjects as positive for disease. If test result X is such that $\mathcal{LR}(\cdot)$ is monotone increasing, than decision rules based on X exceeding a threshold are as optimal as decision rules based on $\mathcal{LR}(x)$ exceeding a threshold. In other words, for a fixed FPR value, this rule is the one who allows to achieve the highest value of the TPR among all possible criteria based on X. In many practical settings the biological theory behind the diagnostic tests assures that $\mathcal{LR}(x)$ is monotonic increasing in X. Furthermore our assumption that

higher values of X are more indicative of disease is the same as stating that $\mathcal{LR}(x)$ is monotonic increasing.

Due to the above properties, the ROC curve has found large application in different biostatistics frameworks such as guide on the choice of the best threshold, compare between two different medical tests or measure the separation between two different distributions S_1 and S_0.

1.1.2 Area Under the Curve

To summarize information about the ROC curve, numerical indexes are often used. They play an important role in interpreting the curve and, in inferential statistics, to compare different curves and different medical tests. The Area Under the ROC curve (AUC) is the most used summary measure, it is a global measure of the separability between the distributions of test measures in the diseased and healthy populations. It is defined as:

$$AUC = \int_0^1 ROC(p)\mathrm{d}p.$$

The values of AUC are in the range (0,1); a perfect test (in the sense of fully discriminant) has $AUC = 1$ and a completely uninformative test has $AUC = 0.5$, in the latter case the ROC curve is the diagonal line $ROC(p) = p$.
AUC has different interpretations; from the definition we stated above immediately follows the interpretation of AUC as the average true positive rate taken uniformly over all possible false positive rates in the range $(0, 1)$. Some other interpretations are also possible, such as a more probabilistic one according to which AUC can be seen as the probability that two random sampled cases, one from diseased population and one from healthy population, are correctly ordered, i.e. $AUC = P[X_1 > X_0]$.

Sometimes, in the clinical practice, the whole set of TPR is not of interest due to the fact that some values of p are not acceptable. In these cases, partial AUC can be calculated restricting the attention to $p < p_0$:

$$AUC_{par}(p_0) = \int_0^{p_0} ROC(p)\mathrm{d}p.$$

Other summary measures have been formalized and available in the literature although not reported in the present work.

1.1.3 The binormal ROC curve

A special case of the ROC curve is the binormal ROC. It derives from the assumption of normally distributed tests in the two populations of interest.

If we state that Y_0 and Y_1 are two independent Gaussian random variables:

$$X_0 \sim N(\mu_0; \sigma_0^2) \qquad \text{and} \qquad X_1 \sim N(\mu_1; \sigma_1^2)$$

then

$$Z_0 = \frac{X_0 - \mu_0}{\sigma_0^2} \qquad \text{and} \qquad Z_1 = \frac{X_1}{-}\mu_1\sigma_1^2$$

are two standard Normal distributions. The false positive rate is

$$\begin{aligned} S_0(c) &= P(X > c|D = 0) = P\left(Z > [c - \mu_0]/\sigma_0\right) \\ &= P(Z \leq [\mu_0 - c]/\sigma_0) \\ &= \Phi\left(\frac{\mu_0 - c}{\sigma_0}\right) \end{aligned}$$

thus

$$z_p = \Phi^{-1}(S_0(c)) = \frac{\mu_0 - c}{\sigma_0}$$

and

$$c = \mu_0 - \sigma_0 z_p.$$

Hence, the ROC curve at this FPR is

$$\begin{aligned} S_1(c) &= P(X > c|D = 1) = P\left(Z > [c - \mu_1]/\sigma_1\right) \\ &= P(Z \leq [\mu_1 - c]/\sigma_1) \\ &= \Phi\left(\frac{\mu_1 - c}{\sigma_1}\right). \end{aligned}$$

Now, substituting the value of c from above, we obtain

$$S_1(c) = \Phi\left(\frac{\mu_1 - \mu_0 + \sigma_0 z_p}{\sigma_1}\right). \tag{1.2}$$

The ROC curve, thus, takes the form:

$$ROC(p) = \Phi(a + b\Phi^{-1}(p)) \tag{1.3}$$

where

$$a = \frac{\mu_1 - \mu_0}{\sigma_1} \qquad \text{and} \qquad b = \frac{\sigma_0}{\sigma_1} \tag{1.4}$$

and Φ is the Normal standard cumulative distribution function.

It is important to notice that any monotonic transformation of the test variable changes the form of the test result distributions but does not change the ROC, which depends only on the order of test results. Despite its simplicity, this model presents some points of weakness. For instance, the ROC

curve is not concave in the whole domain (0,1). This is a problem when the likelihood ratio function of X is monotone (that, as we have said before, it is intuitively reasonable for most clinical tests) because, as we said before, the slope of the ROC(p) curve can be interpreted as the likelihood ratio function of X at the threshold c so, if the likelihood ratio function is monotone, the optimal ROC curve must be concave. However, if $b \neq 1$ the binormal ROC is not concave.[1]

The AUC of a binormal ROC curve has a simple functional form; in fact, it can be proved that it takes the form

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{1.5}$$

with a and b defined as in eq. (1.4).

1.2 Estimation and Inference

Several approaches have been proposed in the literature to estimate the ROC curve. Some of them are based on parametric techniques and some others on non parametric or semi-parametric ones. In this sections we briefly illustrate some of these estimation approaches.

1.2.1 Empirical estimation

The empirical estimation of the ROC curve is the most popular approach in settings with continuous data due to its simplicity and low computational efforts. In fact, the empirical ROC is obtained simply by plot the empirical estimator of the true positive rate of a test (sensitivity) versus the empirical estimator of the false-positive rate (1-specificity)[2] for all possible cut points c. Assume we have n_1 test results from disease population and n_0 test results from non diseased population, realizations of the X_{1_i} and X_{0_j} independent random variables with $i = 1, \ldots, n_1$, $j = 1, \ldots, n_0$ and population survivor function S_1 and S_0, respectively. The empirical estimators of the TPR and FPR are:

$$\widehat{TPR}(c) = \frac{\sum_{i=1}^{n_1} \mathbb{1}[X_{1_i} \geq c]}{n_1} \quad \text{and} \quad \widehat{FPR}(c) = \frac{\sum_{j=1}^{n_0} \mathbb{1}[Y_0 \geq c]}{n_0}.$$

[1] Other parametric models that constraint the ROC to be concave have been proposed in the literature. For further details see: Dorfman et al. (1997) and Metz and Pan (1999).

[2] We recall that, in a binary classification test, the sensitivity is defined as the proportion of true positives identified by the test while the specificity is defined as the proportion of true negatives.

In terms of survival functions, the empirical ROC is

$$\widehat{ROC} = \hat{S}_1(\hat{S}_0^{-1}(p)).$$

The result is an increasing step function on the unit square, with steps of size $1/n_0$ horizontally and steps of size $1/n_1$ vertically. Although in many representations of the ROC curves the points are interpolated in a smoothed curve, the empirical estimation produces a discrete function since FPR can only assume values in $\left\{0, \frac{1}{n_0}; \frac{2}{n_0}; \ldots; 1\right\}$. To obtain a measure of variability of the empirical ROC curve there are different methods depending on the definition of the sampling variability adopted. The corresponding estimator for the AUC is the empirical AUC:

$$\widehat{AUC} = \int_0^1 \widehat{ROC}(p)\mathrm{d}p.$$

As the ROC curve has been estimated empirically we can exploit the definition of the AUC we have seen before in Section 1.1.2; AUC is equal to the probability that a randomly chosen subject from diseased population yields a value of the diagnostic test larger than that of a randomly chosen individual from healthy population. This definition recalls the two sample rank Mann-Whitney U-statistic, which is an unbiased estimator of AUC (see Mann and Whitney (1947)). The definition of the U statistic is:

$$U = \sum_{j=1}^{n_0}\sum_{i=1}^{n_1}\left\{I[X_{1_i} > X_{0_j}] + \frac{1}{2}I[X_{1_i} = X_{0_j}]\right\}.$$

If we consider all possible pairs of individuals one from each sample, than U is the sum of the proportion of pairs for which the score for an individual from sample $D = 1$ is higher than that for the subject from sample $D = 0$ and half the proportion of ties. In this dissertation, since we deal with continuous statistics, the probability of obtaining ties is negligible. Thus, it follows that:

$$E(U) = p(X_1 > X_0) = AUC.$$

Thus,the U-statistic framework may be used to estimate the AUC. As we will see in the next chapter, this result is crucial in the development of higher dimensional ROC framework.

1.2.2 Estimation based on diagnostic test distributions

The approach we are introducing is based on the idea of modeling the distribution of the classification test in a parametric way, in both populations

(diseased and healthy). This assumption provides induced smooth parametric form for the two distribution functions and hence a smooth estimate of the ROC curve. Assuming a distribution function for X_1 and X_0 and estimating the associated parameters, the induced ROC curve can be consequently derived.

Let

$$S_0(x) = S_{\alpha,0}(x) \quad \text{and} \quad S_1(x) = S_{\beta,1}(x)$$

be the two survivor functions with vector of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively; estimating $\boldsymbol{\alpha}$ from the diseased subjects test results and $\boldsymbol{\beta}$ from the healthy subjects test results, the estimated ROC curve will be

$$\widehat{ROC}_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}}(p) = S_{\hat{\boldsymbol{\beta}},1}(S^{-1}_{\hat{\boldsymbol{\alpha}},0}(p)).$$

An example of this approach is given by the binormal ROC. We can assume two Gaussian distributions for the diagnostic test in the two populations such as $X_1 \sim N(\mu_1; \sigma_1^2)$ and $X_0 \sim N(\mu_0; \sigma_0^2)$ and the vectors of parameters denoted by $\boldsymbol{\alpha} = (\mu_1, \sigma_1^2)$ and $\boldsymbol{\beta} = (\mu_0, \sigma_0^2)$. Estimating the parameters with their sample values leads to obtain the estimates of the distribution functions for X_1 and X_0. As we have seen in eq. (1.3), the ROC curve is estimated by:

$$\widehat{ROC}(p) = \Phi\left(\frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_1} + \left(\frac{\hat{\sigma}_0}{\hat{\sigma}_1}\right)\Phi^{-1}(p)\right).$$

This method is extremely sensitive to the validity of its distributional assumptions and it may introduce unnecessary nuisance since the ROC curve depends on the relationship of the two distributions, not on the distributions themselves. Anyway, the estimated AUC is obtained by substituting the estimated parameters in eq. (1.5), obtaining thus

$$\widehat{AUC} = \Phi\left(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}\right). \tag{1.6}$$

Differently, a semi-parametric approach for estimating the ROC curve requires to assume a location-scale model for the diagnostic test results

$$X_{1_i} = \mu_1 + \sigma_1\epsilon_i; \quad X_{0_j} = \mu_0 + \sigma_0\epsilon_j$$

where the ϵ's are zero-mean unit-variance random variables with survivor function S_0. If we estimate the location-scale parameters with the sampled data and the survivor function by a non parametric method based on the residuals

$$\left\{(X_{1_i} - \hat{\mu}_1)/\hat{\sigma}_1, \quad i = 1, 2, \ldots, n_1; \quad (X_{0_j} - \hat{\mu}_0)/\sigma_0, \quad j = 1, 2, \ldots, n_0\right\},$$

the empirical survivor function becomes

$$\hat{S}_0(x) = \frac{1}{n_1 + n_0} \left\{ \sum_i \mathbb{1} \left[\frac{X_{1_i} - \hat{\mu}_1}{\hat{\sigma}_1} \geq x \right] + \sum_i \mathbb{1} \left[\frac{X_{0_i} - \hat{\mu}_0}{\hat{\sigma}_0} \geq x \right] \right\}$$

. It is a consistent estimator of S_0. Thus, the ROC curve estimator can be written as

$$\widehat{ROC(p)} = \hat{S}_0((\hat{\mu}_0 - \hat{\mu}_1)/\hat{\sigma}_1 + (\hat{\sigma}_0/\hat{\sigma}_1)\hat{S}_0^{-1}(p)).$$

The latter model is defined as semi-parametric because it depends on some parameters whilst the S_0 functional form is not specified.

Another approach to the estimate of the ROC curve has been proposed by Lloyd (1998); the author suggests the use the Kernel smoothing estimator of the test result distributions. Defining the kernel estimators as

$$\hat{S}_1 = 1 - \frac{1}{n_1} \sum_{i=1}^{n_1} \Phi\left(\frac{x - x_i}{h_1} \right) \quad \text{and} \quad \hat{S}_0 = 1 - \frac{1}{n_0} \sum_{j=1}^{n_0} \Phi\left(\frac{x - x_i}{h_0} \right),$$

where Φ is the standard normal distribution function and h is the smooth parameter, the estimate of the ROC curve can be easily derived as

$$\widehat{ROC} = \hat{S}_1(\hat{S}_0^{-1}(p)).$$

The author also shows that the resulting kernel estimate of the AUC can be expressed as:

$$\widehat{AUC} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \Phi\left(\frac{x_{1_j} - x_{0_j}}{\sqrt{h_0^2 + h_1^2}} \right). \tag{1.7}$$

1.2.3 Parametric distribution-free methods

Metz et al. (1998) propose a parametric distribution-free approach to ROC curve estimate. In their contribution the author parametrize the form of the curve without any assumption on the distribution of the two variables Y_D and $Y_{\bar{D}}$. They get inspired by the work of Dorfman and Alf (1968) that was conceived for use with ordered categorical data. Such procedure assumes the existence of a latent random variable that underlies the distribution of the diagnostic test. Assuming a Normal distribution for the latent variable in the intervals corresponding to the classes of the categorical observed variable, it is possible to define a binormal model for each class of the categorical variable. The authors developed an iterative algorithm to maximize the log likelihood of the sample with respect to the parameters of the model.

According to that approach, Metz et al. (1998) note that the maximum likelihood estimation of a ROC curve from continuously distributed data is equivalent to the maximum likelihood estimation from ordinal data if the resulting runs of truth state test values are interpreted as categorical data, conditional on having pooled and arranged the test result in increasing order while maintaining their disease labels. Doing that, any information relevant to ROC curve fitting is retained. Hence, they propose to fit binormal ROC curves to continuous data by means of two algorithms. The first performs a ML estimate and the second, less computationally demanding, is based on quasi-ML estimation.

The ROC-GLM estimator is the result of a different parametric distribution-free method based on the idea of estimating the ROC curve within the generalized linear model binary regression framework proposed by Pepe (2000). The use of GLM framework is suggested by the interpretation of the ROC curve as the set of conditional probabilities that X_1 exceeds X_0 given that X_0 is the $(1-t)$-th quantile of the test result distribution in the healthy population. In fact,

$$\begin{aligned} P[X_1 \geq X_0 | S_0(X_0) = p] &= P[X_1 \geq X_0 | X_0 = S_0^{-1}(p)] \\ &= P[X_1 \geq S_0^{-1}(p)] \\ &= S_1(S_0^{-1}(p)) = ROC(p). \end{aligned}$$

The estimator of the curve is derived starting from a parametric model for the curve such as

$$ROC_s(p) = g\left(\sum_s \alpha_s h_s(p)\right)$$

where g is a link function and $h = \{h_1, h_2, \ldots, h_S\}$ are basis functions. If $g = \Phi$, $h_1(p) = 1$ and $h_2(p) = \Phi^{-1}$ the binormal model arises. To fit the model to data, the author proposes to construct indicator variables $\{U_{ij}, i = 1, \ldots, n_1; j = 1, \ldots, n_0\}$ using all $n_1 \times n_0$ possible pairs of test results. These indicator variables are defined as:

$$U_{ij} = \mathbb{1}[X_{1_i} \geq X_{0_j}].$$

Thus, the expected value of U_{ij} conditional on $S_0(X_{0_j}) = p_j$ becomes

$$E[U_{ij} | S_0(X_{0_j}) = p_j] = ROC_s(p_j).$$

The estimation of the s parameters can be performed trough few steps: first calculate the $\hat{S}_0$ estimator using $\{X_{0_j}; j = 1, \ldots, n_0\}$, second calculate $\hat{p}_j = \hat{S}_0(X_{0_j})$ with $j = 1, \ldots, n_0$ measures; third use the $n_1 \times n_0$ pairs

(X_{1_i}, X_{1_j}) to calculate $\{U_{ij}; i = 1, \ldots, n_1, j = 1, \ldots, n_0\}$ finally fit the generalized linear model, with link function g and predictors $h_s(p)$, to the binary variables U_{ij}. When the ROC curve is estimated by fitting a smooth curve, the corresponding AUC estimate can be obtained by numerical integration.

1.2.4 Lehmann family of ROC curves

A semi-parametric estimator of the ROC curve and its AUC, based on the proportional hazards specification of the test results, was provided for the first time by Gönen and Heller (2010). The authors propose a semi-parametric model for the marker values based, not on the functional form of the marker, but on the relationship between the survivor functions of the marker in the two groups of interest. If we define X the diagnostic marker random variable, $D = \{0, 1\}$ the binary indicator representing the groups ($D = 0$ non diseased, $D = 1$ diseased) and S_0 and S_1 the survival functions of the marker for the two different values of the binary indicator, the semi-parametric relationship suggested by the author is:

$$S_1(x) = S_0(x)^\theta \tag{1.8}$$

where the two survival functions are left unspecified and the only parameter of the model is θ who governs the relationship. More generally, the family of distributions defined by this relation is called the Lehmann family, due to the fact that it was used for the first time by Lehmann (1953) in the study of the power function of statistical tests. If we define p as the false positive rate and t the true positive rate, the relation between p and t, as we have seen in eq. (1.1), defines the ROC curve:

$$t = S_1(S_0^{-1}(p)). \tag{1.9}$$

If we use eq. (1.8) in (1.9), the general form of the Lehmann family ROC curve arises:

$$t = p^\theta \tag{1.10}$$

As we will be much more clear in the next sections, it is worth stressing that the Lehmann relationship can be written in terms of hazard functions. In fact, given the definition of the hazard function of the marker:

$$h(x) = \lim_{x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x}$$

with simple algebra, the Lehmann condition in eq. (1.8) may be rewritten in the following, alternative, specification:

$$\frac{h_1(x)}{h_0(x)} = \theta \tag{1.11}$$

where h_1 and h_0 are the hazard functions in the diseased and non diseased group, respectively. The latter equality shows the relationship of the Lehmann specification with the Cox proportional hazards model (Cox, 1972) and provides the opportunity of using a well developed framework for estimation and inference.[3] The proportional hazards model is a well known survival model who specifies the way that the covariates affect the hazard function; if we use x as the realization of the diagnostic test random variable X and D as the covariate, the Cox proportional hazards model is of the form:

$$h_1(x|d) = h_0(x) \exp\{\beta d\} \tag{1.12}$$

thus $\theta = \exp\{\beta\}$. The β parameter can be estimated using the Cox partial maximum likelihood (Cox, 1975). As the author states, under these conditions, the usual large-sample properties of maximum likelihood estimates and tests apply when partial likelihood is used. Estimation and inference of the ROC curve and AUC can be derived from the proportional hazards framework, therefore:

$$\hat{\theta} = \exp\{\hat{\beta}\}. \tag{1.13}$$

The area under the curve is estimated as:

$$\widehat{AUC} = \int_0^1 p^{\hat{\theta}} dp = (\hat{\theta} + 1)^{-1}$$

and its variance as:

$$V(\widehat{AUC}) = (\hat{\theta} + 1)^{-4} V(\hat{\theta}).$$

[3] See Appendix A for the details of the Cox proportional hazards model and how to estimate the unknown parameters.

Chapter 2

Generalization of the ROC curve to multiple class outcomes

In the last twenty years, several authors have approached the concept of high dimensional ROC in different ways related to the different aspects that each researchers was concerned with and to the different ways they had chosen to deal with the complexity of the problem. In fact, when there are more than two classes, there is some flexibility about which aspects of the classification problem are of interest. Moreover, the generalization of the dichotomous framework is not so trivial as, when moving from the binary ROC analysis to the M-class analysis, the dimension of the problem increases much more than proportionally.

As an example we can think at the three-class problem compared to the two-class one. In the two-class framework, as we have seen before, there are four decision outcomes (false positive, false negative, true positive, and true negative), associated with four diagnostic accuracy fractions (FPR, FNR, TPR, and TNR). Since the relationships $FPR+TNR=1$ and $FNR+TPR=1$ hold, only two diagnostic accuracy fractions are needed to fully describe the two-class classification accuracy. In fact, the ROC curve is the set off all points (FPR; TPR) at different thresholds. In the three-class problem there are nine decision outcomes and nine diagnostic accuracy fractions with three relationships holding $TCR_{11} + FCR_{21} + FCR_{31} = 1$, $FCR_{12} + TCR_{22} + FCR_{32} = 1$ and $FCR_{13} + FCR_{23} + TCR_{33} = 1$ where the generic FCR_{ij} indicates the fraction of individuals from class j classified in class i. Therefore six diagnostic accuracy fractions are needed to fully describe the three-class classification accuracy. As the ROC curve for binary diagnosis represents the trade-off between sensitivity and specificity

for the two categories framework, the ROC surface represents the three-way trade-off among the correct classification probabilities for the three categories. Hence going from two to three classes, the total number of fractions needed to fully describe the classification performance increases from 2 to 6 (He et al. (2006)). In this chapter we review, in a chronological order, some of the foundational works which addressed the generalization of ROC curves. Particular emphasis has been devoted to two contributions (Scurfield (1996) and Mossman (1999)) that, while presenting study designs quite different from the one treated in this thesis, are foundational works in the field of generalization of ROC curve and AUC.

2.1 The Scurfield's approach

Scurfield (1996) generalized the ROC curve introducing the concept of ROC surface in the Theory of Signal Detectability (TSD) framework. In TSD the ROC curves were already employed to distinguish between discriminability of the observer and his decision bias. In this framework, the ROC curve represents the relationship between the probability of detecting the 'signal' event when it occurs and the probability of detecting the 'signal' event when the 'noise' event occurs at all possible levels of the threshold. In this sense, it takes into account the decision bias.

Starting from the assumption that the ROC curve is the most satisfactory among all the measures of discriminability for two events task, the author develops the generalization to three events. A couple of years later, Scurfield wrote a compound to his first work extending the concept of the ROC surface to the case of multiple classes and multiple events, introducing the notion of ROC *manifold* (Scurfield, 1998) and providing the basis for the theoretical development of a higher dimensional ROC framework at the population level. However in his works, no inferential procedures are suggested. In the next section we describe the Scurfield (1998) contribution in detail.

2.1.1 The general model

The theory of signal detectability assumes that an observer is exposed to a realization of one of n possible events. After the observation, the subject is asked to guess which event occurred and to select one decision among a set of n decisions $C = \{1, 2, \ldots, n\}$, where $C = i$ with $i = 1, 2, \ldots, n$ indicates that the observer decided an event of class $D = i$ occurred. In the model

proposed by the author the observer is assumed to represent the stimulus as a numerical value x, thus X is a random variable, and $X_1, X_2, \ldots, X_n$ are conditional random variable associated to each event. However, because of the overlapping of stimulus distributions, the perfect discrimination of events is not possible. In what follows, we will see how in this context the ROC curve is adopted as a discriminability measure.

For simplicity, imagine a three-event forced task,[1] in which the observer has to discriminate among three events $C = \{1, 2, 3\}$. The observer represents the stimulus associated with the event as a numerical value x, which is a realization of the univariate random variable X. The quantity x constitutes the evidence the decision is based on, while X_1, X_2 and X_3 are the conditional random variables with distribution $P(X|D = 1), P(X|D = 2)$ and $P(X|D = 3)$, respectively. Assume that the decision is taken with reference to the values of two criteria, denoted as c_1 and c_2. The decision rule is

$$\begin{aligned} &\text{if } x < c_1 \text{ then } C = 1, \\ &\text{if } c_1 < x < c_2 \text{ then } C = 2, \text{ else} \\ &\text{if } x > c_2 \text{ then } C = 3. \end{aligned} \tag{2.1}$$

Table 2.1: Decision Matrix.

	Decision		
Event	$C = 1$	$C = 2$	$C = 3$
$D = 1$	$P(C = 1\|D = 1)$	$P(C = 2\|D = 1)$	$P(C = 3\|D = 1)$
$D = 2$	$P(C = 1\|D = 2)$	$P(C = 2\|D = 2)$	$P(C = 3\|D = 2)$
$D = 3$	$P(C = 1\|D = 3)$	$P(C = 2\|D = 3)$	$P(C = 3\|D = 3)$

The decision matrix of such scenario is represented in Table 2.1. Each element of the matrix represents the probability that the observer will make a particular decision given that a particular event occurred. The matrix has six degrees of freedom because each row sums up to one. Under the decision rule above and if X is a continuous variable, we have that, for each permutation, the equations below hold:

$$P(C = 1|D = \alpha(1)) = P(X_{\alpha(1)} < c_1) \tag{2.2}$$
$$P(C = 2|D = \alpha(2)) = P(c_1 < X_{\alpha(2)} < c_2) \tag{2.3}$$
$$P(C = 3|D = \alpha(3)) = P(X_{\alpha(3)} > c_2) \tag{2.4}$$

[1] In m-event forced task, m randomly sampled events, one from each of the m classes are produced simultaneously, and the observer has to categorize the events to each of the m classes. The decision is correct if all the m events are correctly classified.

where α indicates a permutation that maps the set $\{1, 2, 3\}$ to the set $\{1, 2, 3\}$, while $\alpha(i)$ is the index obtained after the permutation α is applied. Varying the two criteria over the domain of X, and plotting the probability presented in eq.s (2.2)-(2.4), one can generate the $(123) - \text{ROC}$ surface. In Figure 2.1, we report six ROC surfaces and the related six volumes under the surface, each corresponding to a different permutation of the indexes. The volume under each ROC surfaces is related to the discriminability of the events, and thus to the separation degree of the distributions.

If X_1, X_2, X_3 are identically distributed then, from equations (2.2)-(2.4), we have:

$$P(C = 1|D = 1) + P(C = 2|D = 2) + P(C = 3|D = 3) = 1.$$

In each ROC space, the ROC surface is a triangular plane with vertices $(1, 0, 0), (0, 1, 0), (0, 0, 1)$ and the Volume Under the Surface (VUS) equals $1/6$, that is the minimum value possible. On the other hand, if the three variables are perfectly separated, such that every value of X_1 is less than every value of X_2, that is less than every value of X_3, the 123-ROC surface is determined by the three planes

$$\begin{aligned} P(C = 1|D = 1) &= 1 \\ P(C = 2|D = 2) &= 1 \\ P(C = 3|D = 3) &= 1. \end{aligned}$$

In this case the ROC curves collapse to the axes of their respective spaces. The VUS of 123-ROC will be one, that is its maximum value.

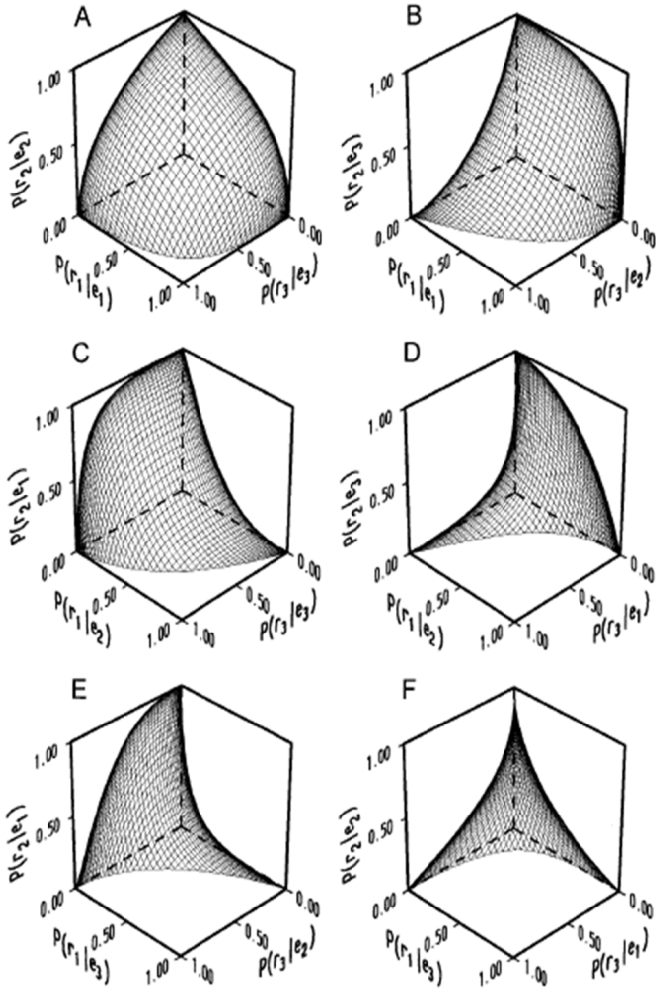
In general if there exist a permutation α such that $P(X_{\alpha(1)} < X_{\alpha(2)} < X_{\alpha(3)}) = 1$, than the observer will be able to perfectly discriminate among the three distributions and one VUS will be equal to one, while the remaining five will be equals to zero.

As a consequence of the above statements, the author states that the volumes under the ROC surfaces are related to the ordering of the variables X_1, X_2, X_3, i.e.

$$VUS_{\alpha(1,2,3)} = P(X_{\alpha(1)} < X_{\alpha(2)} < X_{\alpha(3)}). \tag{2.5}$$

Finally, Scurfield (1998) provides the generalization of the above concepts to the case of n-events and m-evidences, and states the definition of ROC manifold as an extension of the ROC surface and of the Hypervolume Under the Manifold (HUM) as extension of the volume under the surface.

Figure 2.1: Examples of ROC surfaces.



Notes: Six illustrative ROC surfaces generated from three normal distributions with means $\mu_1 = -1$, $\mu_2 = 0$, and $\mu_3 = 1$, and variances $\sigma_1^2 = 1$, $\sigma_2^2 = 1.96$, and $\sigma_3^2 = 1.44$. The ROC surfaces are: (A) the 123-ROC surface, (B) the 132-ROC surface, (C) the 213-ROC surface, (D) the 231-ROC surface, (E) the 312-ROC surface, and (F) the 321-ROC surface. Each contour line represents the variation in one criterion with respect to a fixed value of the other criterion. The contour lines bunch at two outer edges of each surface because the tails of the normal distributions extend to infinity in either direction.

2.2 The Mossman's approach

Approximately in the same period as Scurfield, Mossman (1999) developed high dimensional ROC concepts too, extending the ROC curve analysis to multiple class medical classification settings. Starting from a three-class problem, the author gives a major contribution suggesting a method to estimate the volume under the curve (VUS) in a non-parametric way. Furthermore, bootstrap techniques are proposed to estimate the variance of the estimator. In the next section we deal more diffusely with the Mossman's contribution.

2.2.1 Decision task

The author starts from a practical problem slightly similar as the Scurfield's one. A subject is asked to examine n images of three figures (Circle, Pentagon, Square) presented in a highly degraded form. For each image the subject must estimate a triplet of probabilities. If we denote T_i the triplet for the i-th imagine

$$T_i = (p_{1i}, p_{2i}, p_{3i})$$

where p_{1i}, p_{2i}, p_{3i} represent the subject's confidence about the i-th true status. More in detail, p_{1i} is the subject's probability estimate that image i ia a circle, p_{2i} is the subject's probability estimate that image i ia a pentagon and p_{3i} is the subject's probability estimate that image i ia a square, for each image the estimates should sum to unity so that

$$p_{1i} + p_{2i} + p_{3i} = 1.$$

In addition, the subject is asked to use his estimates to make a decision about the true shape (indicated by $D = 1$, $D = 2$, $D = 3$ for circle, pentagon and square respectively). Let

$$C_{ij} = 1,\ C_{ij} = 2,\ C_{ij} = 3$$

be the three possible subject's decisions about the i-th image, where $C = 1, C = 2, C = 3$ stand for circle, pentagon and square choice respectively, whilst let R_j be the decision rule that guides the subject's choice. Even if there are infinite numbers of possible decision rules that the subject could use, the author suggests three possible rules according to the experimental design that the researcher has to deal with. The rules are reported and discussed in the Section 2.2.3.

2.2.2 Diagnostic Performance

If a series of test results T_i is interpreted under a decision rule R_j, the diagnostic performance can be depicted by a 3×3 contingency table. In this table there are three correct classification rates $TCR_{1j} = P(C_j = 1|D = 1)$, $TCR_{2j} = P(C_j = 2|D = 2)$, $TCR_{3j} = P(C_j = 3|D = 3)$, which are the probabilities that the subject, using rule R_j, makes the correct choice. All off-diagonal values indicate diagnostic errors. As Scurfield pointed out, Mossman also outlines that, in a task with a three-class outcome, there will be six possible diagnostic errors, but, given that the focus is to find a ROC index that summarizes the diagnostic performance, he proposes to ignore the misclassified values (the off diagonal values). Thus, while extending the two-way ROC approach into a third dimension, he suggests to consider only the values lying on the main diagonal of the decision matrix and plot the correct classification rates in a three dimensional space for a set of decision rules. Hence, the points produced can be connected with line segments to obtain a polyhedral ROC surface, the volume under this surface is the VUS. The author proposes an empirical calculation of VUS as average of AUCs at different vertical and horizontal cut-offs of the decision plane. According to the Mossman's interpretation, the ROC curve is obtained starting from the probability triplet of each subject; the VUS is equivalent to the probability that three randomly chosen subjects, one from each of the three classes, will be rated correctly.

2.2.3 Decision rules

We can now report the three decision rules proposed by the author, largely used in subsequent works. If one has to rate three randomly sampled images, each from one of the three classes, the following rules can be used:

R_I: if $p_{1i} \geq \alpha$, treat case i as a circle; if $p_{1i} < \alpha$ and $p_{1i} - p_{3i} \geq \beta$, treat case i as a pentagon; if $p_{1i} < \alpha$ and $p_{2i} - p_{3i} < \beta$, treat case i as a square. In figure 2.2 a graphical representation of this rule is showed. R_I is valid if the outcome doesn't present pre-ordered levels.

R_{II}: given a triplet T_i, treat case i as a circle if p_{1i} is the greatest element in the triplet T_i, treat it as a pentagon if p_{2i} is the greatest element in the triplet T_i, treat it as square if p_{3i} is the greatest element in the triplet T_i. Put differently, R_{II} is a general decision rule to be used when the subject must sort a trio of images drawn randomly from the population. This rule assigns the triplet closest to the circle vertex

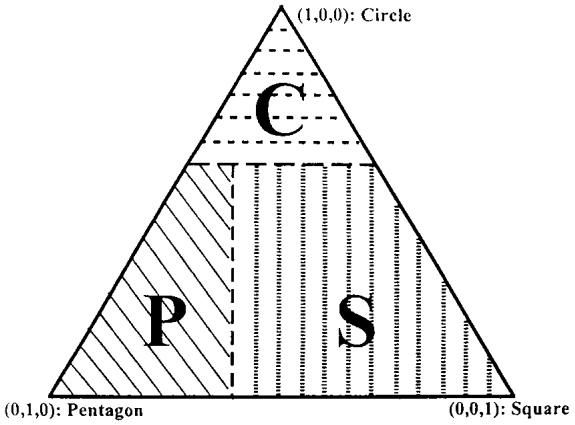


Figure 2.2: Partition of the triangular estimate plane associated with the following type R_I decision rule: If $p_{1_i} \geq 0.6$ treat case as circle, if $p_{1_i} < 0.6$ and $p_{1_i} - p_{3_i} \geq 0.2$ treat case as a pentagon, if $p_{1_i} < 0.6$ and $p_{2_i} - p_{3_i} < 0.2$.

to the circle category, of the two remaining triplets the closest to the pentagon vertex to the pentagon category and the closest to the square vertex to the square category.

R_{III}: if one plotted the trio's three triplets on the triangular estimate plane with circle vertex (1,0,0), pentagon vertex (0,1,0) and square vertex (0,0,1), six different combinations of line segments to link one triplet to one of the three vertex are available. This rule states that one must find the combination of vertex to triplet connections such that the sum of the lengths of the three lines is the shortest, and classify the images associated with each triplet according to the vertex with which each triplet is linked.

2.3 The estimation of the ROC surface and the VUS

In the present section we will revise the main approaches to the estimation of multidimensional ROC surface and volume under the surface (VUS). We will move toward non parametric, semi-parametric and completely parametric estimators of the ROC surface and the ROC volume. Although in the

literature the volume under the ROC surface is commonly named VUS and it is defined as the probability of correctly classifying m subjects, each of which randomly chosen from each of the m classes, the estimation procedure changes depending on whether the true classes are ordered or not.

2.3.1 Non parametric approach for non-ordered classes: Dreiseitl et al. (2000)

Imagine a scenario in which a three-class diagnostic test (that results in a triplet of probabilities) is performed to classify subjects into three classes ($C = \{1, 2, 3\}$) and in which subjects can belong to one of the three-disease classes ($D = \{1, 2, 3\}$). In this case, despite of what happens in the dichotomous case, it is not possible to plot TPR versus FPR since for each TPR there exist two alternative FPR. As we have previously pointed out, in a 3×3 decision matrix there are two off-diagonal values for each row representing the misclassified values. Dreiseitl et al. (2000), according to Mossman (1999), state that the trichotomous version of plotting sensitivity versus specificity is to plot $TCR_1 = P(C = 1|D = 1)$ versus $TCR_2 = P(C = 2|D = 2)$ versus $TCR_3(C = 3|D = 3)$, that are the three diagonal values of the decision matrix. The authors also formalize the empirical estimator of the VUS as a measure of the discriminatory power and propose a non-parametric method to estimate its variance using the Mann-Whitney U statistic.

Starting from Mossman's definition of the VUS as the probability of correctly classifying three subjects, one from each class, and applying his third rule to decide whether the subjects are correctly classified, Dreiseitl et al. (2000) give the expression of the VUS estimator. We now define p_{ij_i} as the vector of probability of subject j_i to belongs to class i, where the first element of the vector is the probability of belonging to class 1, the second element is the probability of belonging to class 2 and the third is the probability of belonging to class 3. Specifically, p_{1j_1} $j_1 = 1, \ldots, n_1$ are the triples for the n_1 subjects of class 1, p_{2j_2} $j_2 = 1, \ldots, n_2$ the triples for the n_2 subjects of class 2 and p_{3j_3} $j_3 = 1, \ldots, n_3$ the triples for the n_3 subjects of class 3. Under the assumption of independence and identical distribution of the triples, an unbiased estimator of $VUS = P[CR(p_1, p_2, p_3) = 1]$, is given by

$$\widehat{VUS} = \hat{\theta}_V = \frac{1}{n_1 n_2 n_3} \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \sum_{j_3=1}^{n_3} CR(p_{1j_1}, p_{2j_2}, p_{3j_3}) \tag{2.6}$$

where

$$CR(p_1, p_2, p_3) = \begin{cases} 1 & \text{if the three triplets } (p_1, p_2, p_3) \text{ are correctly classified} \\ 0 & \text{otherwise} \end{cases}$$

is the function denoting the correctly rated triples. As the VUS is the probability of rating three estimate triples correctly, $\hat{\theta}_V$ gives the fraction of all possible three subject combinations that are rated correctly. For a non discriminatory test the chance of correctly rating the triples p_1, p_2, p_3 is $\frac{1}{3!}$. Finally, the variance of the estimator is

$$\begin{aligned} Var(\hat{\theta}_V) = \frac{1}{n_1 n_2 n_3}[&\theta_V(1-\theta_V) + (n_3-1)(q_{12}-\theta_V^2) + \\ &+ (n_2-1)(q_{13}-\theta_V^2) + (n_1-1)(q_{23}-\theta_V^2) \\ &+ (n_2-1)(n_3-1)(q_1-\theta_V^2) + (n_1-1)(n_3-1)(q_2-\theta_V^2) \\ &+ (n_1-1)(n_2-1)(q_3-\theta_V^2)] \end{aligned} \tag{2.7}$$

where

$$q_{12} = P[CR(P_{1j_1}, P_{2j_2}, P_{3j_3}) = CR(P_{1j_1}, P_{2j_2}, P_{3j_3})]; \quad J_3 \neq j_3$$

is the probability of correctly classifying three subjects of class 1,2,3 and correctly classifying two subjects of class 1,2 and a different subject of class 3 (see Dreiseitl et al., 2000, for a proof of eq. 2.7). In the same way we have:

$$\begin{aligned} q_{13} &= P[CR(P_{1j_1}, P_{2j_2}, P_{3j_3}) = CR(P_{1j_1}, P_{2j_2}, P_{3j_3})]; \quad J_2 \neq j_2 \\ q_{23} &= P[CR(P_{1j_1}, P_{2j_2}, P_{3j_3}) = CR(P_{1j_1}, P_{2j_2}, P_{3j_3})]; \quad J_1 \neq j_1 \\ q_1 &= P[CR(P_{1j_1}, P_{2j_2}, P_{3j_3}) = CR(P_{1j_1}, P_{2j_2}, P_{3j_3})]; \quad J_2 \neq j_2, \quad J_3 \neq j_3 \\ q_2 &= P[CR(P_{1j_1}, P_{2j_2}, P_{3j_3}) = CR(P_{1j_1}, P_{2j_2}, P_{3j_3})]; \quad J_1 \neq j_1, \quad J_3 \neq j_3 \\ q_3 &= P[CR(P_{1j_1}, P_{2j_2}, P_{3j_3}) = CR(P_{1j_1}, P_{2j_2}, P_{3j_3})]; \quad J_1 \neq j_1, \quad J_2 \neq j_2. \end{aligned} \tag{2.8}$$

The quantity above can be estimated by counting the fraction of triples combinations for which the definitions hold, for example:

$$\hat{q}_{12} = \frac{1}{n_1 n_2 n_3} \sum_{j_1} \sum_{j_2} \sum_{j_3} \sum_{j_3 \neq J_3} CR(P_{1j_1}, P_{2j_2}, P_{3j_3}) CR(P_{1j_1}, P_{2j_2}, P_{3j_3}).$$

Thus, an estimator of variance of the VUS can be obtained by substituting in eq. (2.7) the estimators $\hat{q}_{12}, \hat{q}_{13}, \ldots$

$$\begin{aligned} \hat{\sigma}^2 = \frac{1}{n_1 n_2 n_3}[&\hat{\theta}_v(1-\hat{\theta}) + (n_3-1)(\hat{q}_{12}-\hat{\theta}_v^2) + \\ &+ (n_2-1)(\hat{q}_{13}-\hat{\theta}_v^2) + (n_1-1)(\hat{q}_{23}-\hat{\theta}_v^2) \\ &+ (n_2-1)(n_3-1)(\hat{q}_1-\hat{\theta}_v^2) + (n_1-1)(n_3-1)(\hat{q}_2-\hat{\theta}_v^2) \\ &+ (n_1-1)(n_2-1)(\hat{q}_3-\hat{\theta}_v^2)]. \end{aligned} \tag{2.9}$$

According to the Mann-Whitney U-statistic theory, the authors use the asymptotic normality of the estimator to test the hypothesis that two VUS values resulting from two raters who classify the same sample of subjects, are equal. Let $\hat{\theta}_{V1}$ be the VUS estimator obtained from estimates triples of rater 1 and $\hat{\theta}_{V2}$ the VUS estimator obtained from estimates triples of rater 2, their asymptotic normality implies that, for large samples, $\hat{\theta}_{V1} \xrightarrow{d} N(\mu_1; \sigma_1^2)$ and $\hat{\theta}_{V2} \xrightarrow{d} N(\mu_2; \sigma_2^2)$, and therefore, $(\theta_{V1} - \theta_{V2}) \xrightarrow{d} N(\mu_1 - \mu_2; \sigma_1^2 + \sigma_2^2 - 2Cov(\hat{\theta}_{V1}, \hat{\theta}_{V2}))$. Then, the test statistics,

$$z = \frac{\hat{\theta}_{V1} - \hat{\theta}_{V2}}{\sqrt{\hat{\sigma}_1 + \hat{\sigma}_2 - 2\hat{r}\hat{\sigma}_1\hat{\sigma}_2}} \tag{2.10}$$

is the quantity to be compared to the critical values of the normal distribution to determine whether to reject the null hypothesis or not, where $\hat{r}$ is the estimator of the correlation between the two VUS's, that can be calculated similarly as the variance of VUS (see Dreiseitl et al. (2000) for the precise formula).

2.3.2 Non parametric approach for ordered classes : Nakas and Yiannoutsos (2004)

Nakas and Yiannoutsos (2004) unify the approach of Mossman and Scurfield and state the theoretical basis to extend the ROC curve analysis to the multiclass ($M > 3$) ordered classification problems. The authors describe the functional form of the ROC manifold when a continuous diagnostic marker is used to discriminate patients that belong to M ordered classes and discuss a non parametric estimator for Hypervolume Under the Manifold (HUM). Although the condition on the order in the population distributions may seem restrictive, in medical studies is very common to deal with naturally ordered groups. The authors suppose that subjects in class 3 tend to have higher values of the marker than subjects in class 2 and that the latter tend to have values greater than those in class 1 (see Fig. 2.3). For simplicity of exposure we rewrite the ROC curve in eq. (1.1) in terms of distribution function. Let $X_1 \sim F_1(\cdot), X_2 \sim F_2(\cdot), X_3 \sim F_3(\cdot)$ three overlapping distributions corresponding to the distribution functions of the marker in the three populations of patients, and let the rule adopted to classify the subjects, based on the marker values, be the following:

- if marker $X < c_1$ then assign subject to class 1
- if $c_1 < X < c_2$ then assign subject to class 2

- else assign subject to class 3.

Then the functional form of the ROC surface is

$$ROC(p_1, p_3) = F_2(F_3^{-1}(1 - p_3)) - F_2(F_1^{-1}(p_1)),\ 0 \leq p_1, p_3 \leq 1 \qquad (2.11)$$

where $p_1 = TCR_1 = P(C = 1|D = 1) = P(X_1 < c_1)$ and $p_3 = TCR_3 = P(C = 3|D = 3) = P(X_3 > c_2)$ are the probabilities of correct classifications into the first and the third class respectively. The ROC surface is obtained by plotting the points $\{F_1(c_1), F_2(c_2) - F_2(c_1), 1 - F_3(c_2); \quad -\infty < c_1 < c_2 < \infty\}$ in a three-dimensional space at different values of the thresholds c_1 and c_2. This corresponds to the definition provided in equations (2.2)-(2.4) already discussed in Section 2.1.1, but now expressed in terms of distribution functions.
According to this approach the volume under the ROC surface can be seen as an index of the overlap of the three distributions under study and it is obtained by

$$\begin{aligned} VUS = \theta_V &= P[X_1 < X_2 \cap X_2 < X_3] \\ &= \int_0^1 \int_0^1 [F_2(F_3^{-1}(1 - p_3)) - F_2(F_1^{-1}(p_1))]\mathrm{d}p_1\mathrm{d}p_3. \end{aligned}$$

The non parametric estimator of the volume under the ROC surface, as we can see below, has a very similar form as that of Dreiseitl et al. (2000) the only difference is in the rule to determine the correct classifications.

$$\widehat{VUS} = \hat{\theta}_v = \frac{1}{n_1 n_2 n_3} \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \sum_{j_3=1}^{n_3} \mathbb{1}(X_{1j_1}, X_{2j_2}, X_{3j_3})$$

where $\{j_1, j_2, j_3\}$ is a permutation of $\{1, 2, 3\}$ and $\mathbb{1}(X_{1j_1}, X_{2j_2}, X_{3j_3})$ is the indicator function which equals one if X_1, X_2, X_3 are in the correct order, that is if subjects from class 1, 2, 3 are correctly classified in classes 1,2,3 respectively, and zero otherwise.

Nakas and Yiannoutsos (2004) provide the generalization of the estimator in the case of M ($M > 3$) populations using M-1 ordered decision thresholds, $c_i, i = 1, 2, \ldots, M - 1$, together with the same decision rule of the three class framework. Thus, when $X_1, X_2, \ldots, X_M$ are randomly selected from each diagnostic category, the generalized volume under the hypersurface (HUM) is

$$HUM = \theta_H = P[X_1 < X_2 \cap X_2 < X_3 \cap \cdots \cap X_{M-1} < X_M]$$

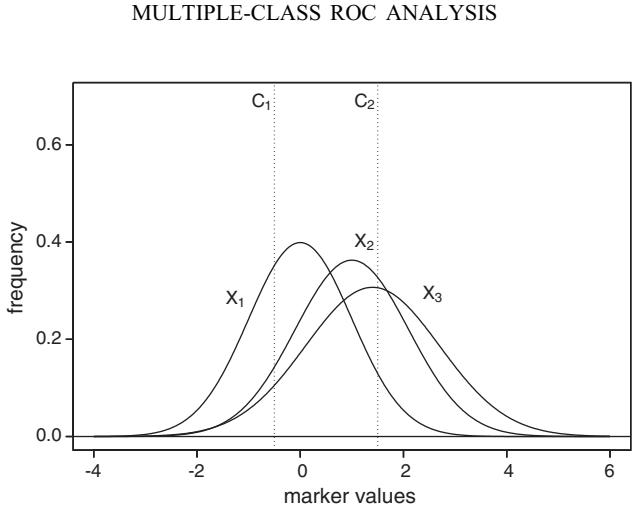


Figure 2.3: Three overlapping continuous distributions based on a diagnostic marker with three different Gaussian distribution. Two ordered decision threshold are fixed to define each point of the ROC surface in the unit cube.

whose value varies from $\frac{1}{M!}$ to 1.
A non parametric estimate of the hypervolume is given by

$$\widehat{HUM} = \hat{\theta}_h = \frac{1}{n_1, n_2, \dot{,} n_M} \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \cdots \sum_{j_M=1}^{n_M} \mathbb{1}(X_{1j_1}, X_{2j_2}, \ldots, X_{Mj_M}).$$

The variance of the hypervolume under the manifold estimator is obtained by generalizing to M classes the variance formula presented in eq. (2.7) and has the following form

$$\begin{aligned} Var(\hat{\theta}_h) = \frac{1}{n_1, n_2, \ldots, n_M} [\theta_H(1-\theta_H) + \sum_i (n_i - 1)(q_i - \theta_H^2) + \\ + \sum_{i_1} \sum_{i_2 \neq i_1} (n_{i_1} - 1)(n_{i_2} - 1)(q_{i_1 i_2} - \theta_H^2) + \cdots + \\ + \sum_{i_1} \cdots \sum_{i_{k-1}} (n_1 - 1) \ldots (n_{k-1} - 1)(q_{12 \ldots k-1} - \theta_H^2)]. \end{aligned} \tag{2.12}$$

It can be estimated in the same way as eq. (2.9) by substituting the estimates of θ_H and q_i in the equation above.

2.3.3 Kernel smoothing based approach: Kang and Tian (2013)

More recently, Kang and Tian (2013) propose an extension of the work of Lloyd (1998) to the case of three ordered classes . The estimator they suggest refers to the situation in which there are three independent random variables X_1, X_2, X_3 that are the result of a continuous diagnostic test, evaluated in three different populations, basically the same scenario as the one we have seen in Section 2.3.2 with a different proposal for the estimator. The authors suggest a new VUS estimator obtained with kernel smoothing techniques. We recall the VUS definition:

$$VUS = \int_0^1 \int_0^{1-F_3(F_1^{-1}(p_1))} F_2(F_3^{-1}(1-p_3)) - F_2(F_1^{-1}(p_1))\mathrm{d}p_1 p_3. \quad (2.13)$$

According to the result that VUS is mathematically equivalent to $P(X_1 < X_2 < X_3)$ and stating the independence of the three random variables, we can rewrite the VUS as:

$$\begin{aligned} VUS &= P(X_1 < X_2 < X_3) = E_{(X_1,X_2,X_3)}[\mathbb{1}(X_1 < X_2 < X_3)] \\ &= E_{(X_2)}E_{(X_1,X_3)}[\mathbb{1}(X_1 < X_2 < X_3)|X_2 = x] \\ &= E_{(X_2)}E_{(X_1,X_3)}[\mathbb{1}(X_1 < X_2) \cap \mathbb{1}(X_2 < X_3)|X_2 = x] \\ &= E_{(X_2)}P(X_1 < x)P(X_3 > x) \\ &= \int_{-\infty}^{+\infty} [F_1(x)(1 - F_3(x)]f_2(x)\mathrm{d}x). \end{aligned} \quad (2.14)$$

Applying the Gaussian kernel estimator, the probability density function f_2, becomes:

$$\hat{f}_2(x) = \frac{1}{n_2}\sum_{j_2=1}^{n_2} \frac{1}{h_2}\phi\left(\frac{x - x_{2j_2}}{h_2}\right) \quad (2.15)$$

if we then apply the same Gaussian kernel estimator to the cumulative distribution functions F_2 and F_3 we obtain the estimates:

$$\hat{F}_1(x) = \frac{1}{n_1}\sum_{j_1=1}^{n_1} \Phi\left(\frac{x - x_{1j_1}}{h_1}\right) \quad (2.16)$$

$$\hat{F}_3(x) = \frac{1}{n_3}\sum_{j_3=1}^{n_3} \Phi\left(\frac{x - x_{3j_3}}{h_3}\right) \quad (2.17)$$

that, substituted in eq. 2.13, gives the VUS estimator:

$$\widehat{VUS} = \hat{\theta}_v = \int_{-\infty}^{+\infty} \hat{F}_1(x)(1 - \hat{F}_3(x)\hat{f}_2(x)\mathrm{d}x). \tag{2.18}$$

The terms h_i in the equations above, are the parameters of bandwidth which control the amount of smoothing of the curve. According to Silverman (1986), the asymptotic value

$$h_i = \left(\frac{4}{3n_i}^{1/5}\right) \min(SD_i, IQR_i/1.349)$$

is suggested since it works well with a wide range of density functions. In the formula above, SD_i and IDR_i are the standard deviation and the interquartile range of the sample distributions respectively.

2.3.4 Non parametric approach for multi-category multiple tests: Li and Fine (2008)

The generalization of the AUC to the multi-category and multiple test problems has been definitely formalized with the contribution of Li and Fine (2008). Starting from the works of Dreiseitl et al. (2000) and Nakas and Yiannoutsos (2004), they generalize the problem to multiple classes and multiple tests. The authors interpret the third rule of Mossman in a mathematical way and, doing that, they develop an inferential method that, some years later, will be even implemented in the statistical software *R*. They argue that the correct classification may not need to be established by the order of the test results, but can be determined by a more complicated geometrical rule derived from the third rule of Mossman. We now illustrate the rule in the simpler case of a forced-choice task with M subjects, each sampled from one of the M different classes. Let p_i, with $i = 1, 2, \ldots, M$, be the probability ratings of subjects coming from class i such that $p_i = (E_{i,1}, E_{i,2}, \ldots, E_{i,M})$ where the generic $E_{i,k}$ stands for the assessed probability of subject i to belong to class k, with $k = 1, 2, \ldots, M$. Let $\nu_1 = (1, 0 \ldots, 0)$, $\nu_2 = (0, 1 \ldots, 0)$, $\ldots$, $\nu_M = (0, 0 \ldots, 1)$, according to R_{III} the authors establish to assign subjects to class $\alpha(1), \alpha(2), \ldots, \alpha(M)$ [2] such that $\|p_1 - \nu_{\alpha(1)}\|^2 + \|p_2 - \nu_{\alpha(2)}\|^2 + \cdots + \|p_M - \nu_{\alpha(M)}\|^2$ is minimized among all possible assignments $\alpha(1) \neq \alpha(2) \neq \ldots \neq \alpha(M)$ where $\|\cdot\|$ is the euclidean distance.

Consider now the case in which there are N subjects sampled from M different populations. Let p_{ij} with $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, n_i$ be

[2] $\alpha(k)$ as we as seen in 2.1.1 indicates the permutation of the index

the probability ratings of subject j coming from class i such that $p_{ij} = (E_{i,j,1}, E_{i,j,2}, \ldots, E_{i,j,M})$, where the generic $E_{i,j,k}$ stands for the probability of subject j from class i to belong to class k, with $k = 1, 2, \ldots, M$. The non parametric HUM estimator proposed by the author becomes:

$$HUM = \theta_h = \frac{1}{\prod_{i=1}^{M} n_i} \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \ldots \sum_{j_M=1}^{n_M} CR\left(p_{1j_1}, p_{2j_2}, \ldots, p_{Mj_M}\right)$$

This formula for the hypervolume seems very similar to that suggested by Nakas and Yiannoutsos (2004), nevertheless it is more general because it applies to the case of unordered variables. Moreover Li and Fine (2008) addressed a very common issue in applied research, that is the case in which the probabilities p_{ij} are unknown and must be estimated from diagnostics tests. Let now see their proposal. Suppose each subject has q test scores, let $\mathbf{T}_{ij} = (T_{ij1}, \ldots, T_{ijq})^T, i = 1, \ldots, M; j = 1, \ldots, n_i$ be the vector of scores, the class probabilities can be modelled with multinomial logistic regression where

$$P_{ij,c}(\beta) = \Pr(\text{subject ij is from class } c|T_{ij}) = \frac{\mathbb{1}(c > 1) \exp\left(\beta_{c-1}^T T_{ij}\right) + \mathbb{1}(c = 1)}{1 + \sum_{k=1}^{M-1} \exp\left(\beta_k^T T_{ij}\right)} \tag{2.19}$$

The $M \times q$ matrix of parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_{M-1}^T)$ can be estimated maximizing the log-likelihood function. Substituting $\hat{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{M-1}^T)^T$ for $\boldsymbol{\beta}$ in eq.(2.19), the estimated probabilities $\hat{p}_{ij} = (\hat{p}_{ij,1}, \ldots, \hat{p}_{ij,M})$ are obtained. These probability multiplets may then be replaced in the HUM estimator to obtain

$$\widehat{HUM} = \hat{\theta}_v = \frac{1}{\prod_{h=1}^{M} n_h} \sum_{j_1}^{n_1} \sum_{j_2}^{n_2} \ldots \sum_{j_M}^{n_M} CR(\hat{p}_{1j_1}, \hat{p}_{1j_2}, \ldots, \hat{p}_{Mj_M}). \tag{2.20}$$

For $M > 2$ and/or $q > 1$ the variance of the HUM estimator is influenced by the estimation of the probability assessment $\hat{\beta}$. As a consequence, in this case it is not possible to use the U-statistic variance formula as the previous authors did. Thus, to make inference about HUM, the bootstrap standard errors are preferable.

Denoting the bootstrap estimator as $\widehat{HUM}_b$, $b = 1, 2, \ldots, B$, where b indicates the bootstrap sample, the bootstrap standard error is

$$\widehat{se}_B\left(HUM\right) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{HUM}_b - \widehat{HUM}\right)^2}.$$

A $100(1-\alpha)\%$ confidence interval for HUM can be calculated using the $\alpha/2$ percentiles of the standard distribution as below:

$$\widehat{HUM} \pm z_{\alpha/2}\hat{se}_B\left(HUM\right).$$

As an alternative, the interval between the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap distribution of $\widehat{HUM}$ may be used.

2.3.5 Parametric approach for ordered classes: Xiong et al. (2006)

The first parametric approach to multiple ROC estimate has been offered by Xiong et al. (2006). The authors propose a parametric framework under two main hypothesis: the Gaussian distribution of the diagnostic test and the ordering of the subjects diagnostic groups. Starting from the same idea of ordered groups as that of Nakas and Yiannoutsos (2004) illustrated in the previous section, the authors provide the functional form of the three-class ROC together with the functional form of the VUS and partial VUS_{par}. In the special case in which rather than general distributions for X_1, X_2 and X_3, three normal distributions are assumed $X_1 \sim N(\mu_1;\sigma_1)$; $X_2 \sim N(\mu_2;\sigma_2)$ $X_3 \sim N(\mu_3;\sigma_3)$, with $\mu_1 < \mu_2 < \mu_3$, the parametric form of the surface becomes:

$$\begin{aligned} ROC(p_1,p_3) &= \left\{\Phi(\beta_1+\beta_2\Phi^{-1}(1-p_3)) - \Phi(\beta_3+\beta_4\Phi^{-1}(p_1))\right\} \\ &\qquad \mathbb{1}_{[\beta_3+\beta_4\Phi^{-1}(p_1)\leq\beta_1+\beta_2\Phi^{-1}(1-p_3)]}(p_1,p_3) \end{aligned}$$

where $\mathbb{1}$ is the indicator function, Φ is the distribution function of the standard normal random variable and $\boldsymbol{\beta} = (\beta_1,\beta_2,\beta_3,\beta_4)$ is the parameters vector of the ROC surface.

Holding these assumptions, the volume under the curve can be written as:

$$VUS = \int_0^1\int_0^1 \Phi(\beta_1+\beta_2\Phi^{-1}(1-p_3)) - \Phi(\beta_3+\beta_4\Phi^{-1}(p_1))\mathrm{d}p_1p_3$$

where

$$\beta_1 = \frac{\mu_3-\mu_1}{\sigma_2};\ \beta_2 = \frac{\sigma_3}{\sigma_2};\ \beta_3 = \frac{\mu_1-\mu_2}{\sigma_2};\ \beta_4 = \frac{\sigma_1}{\sigma_2}.$$

The maximum likelihood estimate of the VUS is obtained by replacing the parameters vector $\boldsymbol{\beta}$ with its maximum likelihood estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4)$. Thus, the estimated ROC surface is

$$\begin{aligned} \widehat{ROC}(p_1,p_3) &= \left\{\Phi(\hat{\beta}_1+\hat{\beta}_2\Phi^{-1}(1-p_3)) - \Phi(\hat{\beta}_3+\hat{\beta}_4\Phi^{-1}(p_1))\right\} \\ &\qquad \mathbb{1}_{[\hat{\beta}_3+\hat{\beta}_4\Phi^{-1}(p_1)\leq\hat{\beta}_1+\hat{\beta}_2\Phi^{-1}(1-p_3)]}(p_1,p_3), \end{aligned}$$

and therefore the VUS estimator is

$$\widehat{VUS} = \int_0^1 \int_0^1 \Phi(\hat{\beta}_1 + \hat{\beta}_2 \Phi^{-1}(1 - p_3)) - \Phi(\hat{\beta}_3 + \hat{\beta}_4 \Phi^{-1}(p_1)) \mathrm{d}p_1 p_3 \quad (2.21)$$

The authors suggest to adopt numerical methods to solve the integral. Due to the asymptotic normality of the ML estimates, the Delta method can be applied to yield the asymptotic variance of the estimated parameters (for more detailed explanation see Xiong et al. (2006)).

Although the above results apply to the particular situation of Normal distributed tests, we would interestingly point out that, due to the invariance of the ROC surface under monotonic transformations, the estimating procedure holds if a known monotonic transformation can be applied to non-normal data to transform it into normal ones.

2.3.6 Lehmann family approach

Recently, Nze Ossima et al. (2015) developed a model extending the results reached by Gönen and Heller (2010) presented in Section 1.2.4 to the three-class problem. They obtain an analytical form for both the ROC surface and the volume under the surface. Furthermore the model they propose allows for the covariate adjustments.

The model

Let X_1, X_2, X_3 be the random variables corresponding to the marker distribution in the three classes, and S_1, S_2, S_3 the corresponding survival functions. The unique assumption they pose is that of the Lehmann alternatives:

$$\begin{aligned} S_2 &= S_1^{\theta_1} \quad 0 < \theta_1 \leq 1 \\ S_3 &= S_2^{\theta_2} \quad 0 < \theta_2 \leq 1 \end{aligned}$$

whereas the survival functions are left unspecified. They define two thresholds c_1 and c_2 with $c_1 < c_2$, and then the correct classification probabilities as:

$$u_1 = 1 - S_1(c_1); \quad u_2 = S_2(c_1) - S_1(c_2); \quad u_3 = S_3(c_2).$$

The ROC surface for the marker is obtained by expressing u_2 as a function of u_1 and u_3:

$$u_2 = S_2[S_1^{-1}(1 - u_1)] - S_2[S_3^{-1}(u_3)]; \quad 0 \leq u_1 \leq 1; 0 \leq u_3 \leq S_3[S_1^{-1}(1 - u_1)].$$

after some substitutions, the semi-parametric form of the surface is obtained:

$$u_2 = (1 - u_1)^{\theta_1} - u_3^{1/\theta_2}.$$

The volume under the surface is the overall accuracy measure they derive by integrating the surface:

$$\int_0^1 \int_0^{(1-u_1)^{\theta_1\theta_2}} u_2(u_1, u_3)\mathrm{d}u_3\mathrm{d}u_1 = \int_0^1 \int_0^{(1-u_1)^{\theta_1\theta_2}} (1-u_1)^{\theta_1} - u_3^{1/\theta_2}\mathrm{d}u_3\mathrm{d}u_1.$$

As a result of the integration the authors get the analytical form of the entire volume under the surface:

$$VUS = \frac{1}{(\theta_2+1)(\theta_1(\theta_2+1)+1)}.$$

The estimation procedure they adopt exploits the proportional hazards framework as in the spirit of Gönen and Heller (2010). More in detail, in this approach the marker plays the role of the dependent variable while the class indicators are the regressors. Clearly, given the generality of the Cox regression model, other possible explanatory variables are also allowed. The estimation of the regression parameters is based on the well known Cox partial maximum likelihood. Instead, the θ's can be easily obtained from these latter exactly as shown in Section 1.2.4, eq.s (1.11)-(1.13).

As a result, the authors derive the maximum likelihood estimate for the VUS, that takes the form:

$$\widehat{VUS} = \frac{1}{(\hat{\theta}_2+1)(\hat{\theta}_1(\hat{\theta}_2+1)+1)}.$$

Furthermore the authors evaluate the asymptotic variance of the parameters using the Delta method and consequently the variance of the VUS estimator is obtained. The variance-covariance matrix for the parameters is:

$$\Sigma_{\hat{\theta}_1,\hat{\theta}_2} = \begin{bmatrix} \sigma^2_{\hat{\theta}_1} & \sigma_{\hat{\theta}_1,\hat{\theta}_2} \\ \sigma_{\hat{\theta}_2,\hat{\theta}_1} & \sigma^2_{\hat{\theta}_2} \end{bmatrix} \tag{2.22}$$

where $\sigma^2_{\hat{\beta}_i} = \exp(2\hat{\beta}_i)\sigma^2_{\hat{\beta}_i}$, $i = 1, 2$ and the covariance $\sigma_{\hat{\theta}_1,\hat{\theta}_2} = \frac{\partial\hat{\theta}_1}{\partial\hat{\beta}_1} \times \frac{\partial\hat{\theta}_2}{\partial\hat{\beta}_2} \times \sigma_{\hat{\beta}_1,\hat{\beta}_2} = \exp(\hat{\beta}_1 + \hat{\beta}_2) \times \sigma_{\hat{\beta}_1,\hat{\beta}_2}$ and the asymptotic variance of $\widehat{VUS}$ is:

$$\begin{aligned} \sigma^2_{\widehat{VUS}} &= \sigma^2_{\hat{\theta}_1}\left(\frac{\partial\hat{VUS}}{\partial\hat{\theta}_1}\right)^2 + \sigma^2_{\hat{\theta}_2}\left(\frac{\partial\hat{VUS}}{\partial\hat{\theta}_2}\right)^2 \\ &+ 2\left(\frac{\partial\hat{VUS}}{\partial\hat{\theta}_1}\frac{\partial\hat{VUS}}{\partial\hat{\theta}_2}\right)\sigma_{\hat{\theta}_1\hat{\theta}_2}. \end{aligned} \tag{2.23}$$

In the next chapter we extend this result to a four-class problem providing an analytical form of the ROC manifold and of its hypervolume.

Chapter 3

Four-class classification models

In a recent contribution Rodia et al. (2018) evaluate a panel of four RNA messengers as putative markers of colorectal cancer detection. Specifically, the authors investigated a sample of patients which can be classified in four distinct groups according to the severity level of the disease. In that paper, the ROC analysis is conducted and based on comparisons between the markers observed in the control group, composed by the healthy subjects, and the values observed for the other groups, taken one at the time. Such a way of proceeding through a sequence of pairwise analysis is quite standard in the empirical literature when the sample is characterized by more than two groups. This clearly represents a limit for this approach, that does not perform the investigations using simultaneously all the data composing the sample of observations.

Thus, motivated by this real research question, in this chapter we develop a new theoretical framework to investigate the ability of a specific marker to distinguish among four groups characterizing the population. In parallel, we explore different approaches to estimate the hypervolume under the manifold (HUM) as a measure of classification accuracy. We propose a new HUM estimator, that we call HUM_{L4}. It is a generalization of the recent approach proposed by Gönen and Heller (2010) for the dichotomous framework and already presented in Section 1.2.4. As we have exposed in Section 2.3.6, Nze Ossima et al. (2015) generalized the idea of Gönen and Heller (2010) to the estimate of the volume under the ROC surface in a three-group classification framework. In the same spirit we derive the analytical formula for the four-category HUM estimator and for the calculation of its variance. As we will discuss in the following sections, this last point represents a novelty in the literature.

Our approach is, also, compared to other estimators existing in the literature and characterized by both different theoretical assumptions and es-

timation procedures. The first, which we call HUM_{EX}, as already seen in Section 2.3.1 and Section 2.3.2, is based on a non parametric approach and relies on the Mann-Whitney U statistic from which it takes the theoretical and inferential framework The second, which we indicate HUM_{LF}, is the estimator presented by Li and Fine (2008) and described in Section 2.3.4. It is based on the multinomial logistic model and, importantly, is not affected by the order of the outcome's categories. . In this chapter we also deal with the issue of computational efforts. In fact, we recall that the HUM is mathematically equivalent to the probability that random variables corresponding to the test results are in a defined order. Hence, a four-class classification problem should require the computation of $4! = 24$ HUMs in order to find the highest one. As we will see later, we propose a technique to avoid the computational burden of calculating 24 HUMs and, consequently, to reduce the time complexity.

The chapter is organized as follows: the next section discusses the Lehmann assumption in a four-class classification framework, that plays a relevant role in our theoretical contribution. Section 3.2 presents the new estimator we propose to classify subjects in a population characterized by four distinct groups, while Section 3.3 presents an analytical formula for the variance of the proposed estimator. Given the key role played by the Lehmann assumption, in Section 3.4 we discuss a testing approach to verify whether such assumption is supported by the data. In Section 3.5 we outline how the proposed methodology can provide information about the separation degree of the classes. Section 3.6 is dedicated to a comparison between our proposal and the alternative methodologies already existing in the literature. Finally, in Section 3.7 we discuss on how to detect the optimal ordering of the groups and introduce the notion of *relative effects*.

3.1 Lehamnn family assumption in a four-class problem

Suppose we employ a diagnostic test with continuous values to distinguish between four classes of disease. Let X_1, X_2, X_3, X_4 be the continuous variables of the test result for subjects from classes 1 to 4. Moreover, let D be an ordinal categorical variable taking on values from 1 to 4, and indicating for each subject, the class he belongs to. Suppose, further, that the test results for class 1, $X_{1,i}$, with $i = 1, 2, \ldots, n_1$ are i.i.d., and the same for all the other classes. Moreover, let S_1, S_2, S_3 and S_4 indicate the corresponding survival functions.

The survival distributions are assumed to have the family of Lehmann alternatives, i.e.:

$$S_2(x) = S_1(x)^{\theta_1}, \quad 0 < \theta_1 \leq 1 \tag{3.1}$$
$$S_3(x) = S_2(x)^{\theta_2}, \quad 0 < \theta_2 \leq 1 \tag{3.2}$$
$$S_4(x) = S_3(x)^{\theta_3}, \quad 0 < \theta_3 \leq 1. \tag{3.3}$$

Using the log transformation allows to rewrite the relationships among the survival functions as

$$\log(S_{i+1}(x)) = \theta_i \log(S_i(x)). \tag{3.4}$$

Moreover, based on the general definition of the hazard function

$$h(x) = \frac{-dS(x)}{dx}\frac{1}{S(x)} = \frac{-d[logS(x)]}{dx}$$

and taking the first derivative in 3.4 with respect to x, we obtain

$$h_{i+i}(x) = h_i\theta_i.$$

Thus, the unknown parameters θ can be modelled through the Cox proportional hazards model assuming the marker value instead of the time index as the argument of the hazard function. The general formula of the Cox model is:

$$h(x|d) = h_1(x)exp\{\boldsymbol{\beta}^{'}\boldsymbol{d}\} \tag{3.5}$$

where x is the marker value, $\boldsymbol{d}$ is the vector of appropriate dummy variables to detect the group, $\boldsymbol{\beta}$ is the vector of parameters with $\theta_i = \exp\{\beta_i\}$ and h_1 is the hazard function of the baseline group.[1] Note that the condition on the parameters θs indicates that subjects from class 4 tend to have higher levels of the diagnostic test than subjects from class 3, and that subjects of class 3 tend to have higher measurements than those from class 2 and so on.

3.1.1 The ROC manifold

Suppose to have three assigned thresholds $c_1 < c_2 < c_3$. The four probabilities of correct classification, in this case, are:

$u_1 = P(X_1 < c_1)$; $u_2 = P(c_1 \leq X_2 < c_2)$; $u_3 = P(c_2 \leq X_3 < c_3)$; $u_4 = P(X_4 \geq c_3)$

[1] See Appendix A for a discussion on the Cox proportional hazards model.

in terms of survival functions we can rewrite:

$$\begin{aligned} u_1 &= 1 - S_1(c_1) && (3.6)\\ u_2 &= S_2(c_1) - S_2(c_2) && (3.7)\\ u_3 &= S_3(c_2) - S_3(c_3) && (3.8)\\ u_4 &= S_4(c_3). && (3.9) \end{aligned}$$

Now, starting from the definition of ROC surface as the function obtained by writing a correct-classification probability for a class as a function of the other classes, the equation for the ROC surface can be obtained by writing the correct-classification probability of, for example, class three u_3 as a function of u_1, u_2 and u_4.

$$\begin{aligned} u_3 &= S_3(c_2) - S_3(c_3) && (3.10)\\ &= S_3[S_2^{-1}(S_2(c_1) - u_2)] - S_3^{-1}(u_4)\\ &= [S_2(c_1) - u_2]^{\theta_2} - u_4^{1/\theta_3}\\ &= [S_1(c_1)^{\theta_1} - u_2]^{\theta_2} - u_4^{1/\theta_3}\\ &= [(1-u_1)^{\theta_1} - u_2]^{\theta_2} - u_4^{1/\theta_3} \end{aligned}$$

The ROC hypersurface is thus a four-dimensional manifold with the following expression:

$$ROC(\mathbf{u}) = ((1-u_1)^{\theta_1} - u_2)^{\theta_2} - u_4^{1/\theta_3} \quad (3.11)$$

where

$$\mathbf{u} = (u_1, u_2, u_4), \text{ with } u_i \in [0,1], \quad i = \{1,2,3\},$$

and

$$0 \leq u_1 \leq 1;$$
$$0 \leq u_2 \leq S_2(c_1)$$
$$0 \leq u_4 < S_4(c_2).$$

Moreover, from eq.s (3.6)-(3.9), we can rewrite

$$0 \leq u_2 \leq (1-u_1)^{\theta_1}$$

and

$$0 \leq u_4 < [(1-u_1)^{\theta_1} - u_2]^{\theta_2\theta_3}.$$

Note that if the four distributions are identical, the discriminating power of the diagnostic test is null and the ROC hypersurface satisfies the equation $u_1 + u_2 + u_3 + u_4 = 1$.

3.1.2 The hypervolume under the manifold

As for the simpler cases of two- and three-classification issues, the accuracy measure of the discriminating function $ROC(\mathbf{u})$ can be given by the hypervolume under the manifold, denoted by HUM. In the next theorem we report an analytical formula for calculating the HUM, that represents the first important theoretical result in this section.

Theorem 1. *Consider a four-class classification problem where the survival functions, under the Lehmann conditions, are given as in eq.s (3.1)-(3.3). Moreover, let the quantities u_1, u_2, u_3 and u_4 be defined as in eq.s (3.6)-(3.9).*

If the discriminating function is given by the $ROC(\boldsymbol{u})$ function as defined in eq. (3.11), then the hypervolume under the manifold HUM is given by

$$HUM(\theta_1, \theta_2, \theta_3) = \frac{1}{(\theta_3 + 1)(\theta_2(\theta_3 + 1) + 1)(\theta_1(\theta_2(\theta_3 + 1) + 1) + 1)} \quad (3.12)$$

for some parameter $0 < \theta_1 \leq 1$, $0 < \theta_2 \leq 1$ and $0 < \theta_3 \leq 1$.

Proof. The hypervolume under the ROC manifold represents the accuracy measure we are interested in and it is obtained integrating the ROC surface defined in eq. (3.11) over its domain:

$$
\begin{aligned}
HUM(\theta_1,\theta_2,\theta_3) &= \int_0^1\int_0^{(1-u_1)^{\theta_1}}\int_0^{((1-u_1)^{\theta_1}-u_2)^{\theta_2\theta_3}} S_3(c_2)-S_3(c_3)\mathrm{d}u_4\mathrm{d}u_2\mathrm{d}u_1 \\
&= \int_0^1\int_0^{(1-u_1)^{\theta_1}}\int_0^{((1-u_1)^{\theta_1}-u_2)^{\theta_2\theta_3}} ((1-u_1)^{\theta_1}-u_2)^{\theta_2}-u_4^{1/\theta_3}\mathrm{d}u_4\mathrm{d}u_2\mathrm{d}u_1 \\
&= \int_0^1\int_0^{(1-u_1)^{\theta_1}} \left| u_4\Big\{((1-u_1)^{\theta_1}-u_2)^{\theta_2}-\frac{\theta_3 u_4^{1/\theta_3}}{\theta_3+1}\Big\}\right|_0^{((1-u_1)^{\theta_1}-u_2)^{\theta_2\theta_3}} \\
&\qquad\qquad \mathrm{d}u_2\mathrm{d}u_1 \\
&= \int_0^1\int_0^{(1-u_1)^{\theta_1}} ((1-u_1)^{\theta_1}-u_2)^{\theta_2\theta_3}\Big\{((1-u_1)^{\theta_1}-u_2)^{\theta_2}+ \\
&\qquad\qquad -\frac{\theta_3((1-u_1)^{\theta_1}-u_2)^{\theta_2}}{\theta_3+1}\Big\}\mathrm{d}u_2\mathrm{d}u_1 \\
&= \int_0^1\int_0^{(1-u_1)^{\theta_1}} \frac{((1-u_1)^{\theta_1}-u_2)^{\theta_2(\theta_3+1)}}{\theta_3+1}\mathrm{d}u_2\mathrm{d}u_1 \\
&= \int_0^1 \left| -\frac{((1-u_1)^{\theta_1}-u_2)^{\theta_2(\theta_3+1)+1}}{(\theta_3+1)(\theta_2(\theta_3+1)+1)}\right|_0^{(1-u_1)^{\theta_1}} \\
&= \int_0^1 \frac{(1-u_1)^{\theta_1(\theta_2(\theta_3+1)+1)}}{(\theta_3+1)(\theta_2(\theta_3+1)+1)} \\
&= \left| -\frac{(1-u_1)^{\theta_1(\theta_2(\theta_3+1)+1)}}{(\theta_3+1)(\theta_2(\theta_3+1)+1)(\theta_1(\theta_2(\theta_3+1)+1)+1)}\right|_0^1 \\
&= \frac{1}{(\theta_3+1)(\theta_2(\theta_3+1)+1)(\theta_1(\theta_2(\theta_3+1)+1)+1)}
\end{aligned}
$$

□

Eq. (3.12) represents the entire hypervolume under the manifold in a four-dimensional classification problem. As it can be seen, the closed form depends only on the parameters of the Lehmann assumption. The following corollary to Theorem 1 provides an interesting result when the ROC function associated to the classification problem is unable to discriminate among the different classes.

Corollary 1. *Consider a four-class classification problem as the one defined in Theorem 1. If the diagnostic test is non-discriminatory, the HUM is equal to*

$$HUM(\theta_1, \theta_2, \theta_3) = 1/24 \tag{3.13}$$

which is its minimum possible value.

Proof. In the case of a non-discriminatory test, we have that $\theta_1 = \theta_2 = \theta_3 = 1$, indicating that there is no difference among the four survival functions. As a consequence, the solution of the integral in eq. (3.12) simply becomes $HUM = 1/24$, which is also equal to $1/4!$, that represents the minimum value as described before. □

Corollary 2. *When θ_3 goes to zero, $HUM = VUS = 1/((\theta_2 + 1)(\theta_1(\theta_2 + 1) + 1))$, as in Nze Ossima et al. (2015). If both θ_3 and θ_2 go to zero, $HUM = VUS = AUC = 1/(\theta_1 + 1)$, i.e. the volume under the surface collapses to the area under the Lehmann family ROC curve as shown in Gönen and Heller (2010).*

Proof. The result can be easily obtained from the proof of Theorem 1, by solving the integral fixing $\theta_3 = 0$ first, and then both $\theta_3 = 0$ and $\theta_2 = 0$. □

The result in Corollary 2 simply states that the analytic formula for the HUM developed in Theorem 1 is a generalization, in the four-class classification framework, of the findings by Nze Ossima et al. (2015) and, originally, by Gönen and Heller (2010) in the three- and two-class classification framework, respectively.

3.2 Estimation

In the previous section we have derived the ROC manifold and the associated summary measure represented by the HUM. In this section, instead, we propose a semi-parametric approach to estimate the hypersurface and the hypervolume.

Under the Lehmann condition, the four survival functions are related one to the others by means of the parameters θ_1, θ_2 and θ_3. As a consequence, both the ROC surface and the HUM are functions of such unknown parameters, that represent the object of our inferential analysis.

Following the intuition by Gönen and Heller (2010) for the two-class classification issue, we propose to estimate the parameters θ_1, θ_2 and θ_3 by means of the proportional hazards regression model already presents in several statistical packages. As we will show here below, the problem, in fact, can be written in terms of regression model. Let x be the generic realization of the diagnostic test X. Under the Lehmann condition we have that:

$$\theta_1 = \frac{h_2(x)}{h_1(x)} \tag{3.14}$$

$$\theta_2 = \frac{h_3(x)}{h_2(x)} \tag{3.15}$$

$$\theta_3 = \frac{h_4(x)}{h_3(x)} \tag{3.16}$$

where $h_i(x)$ are the hazard functions of X in the group i-th, with $i = 1, \ldots, 4$.

In order to estimate the parameters, we define the Cox model with the diagnostic test X in place of the usual "time"variable. To implement the model, the four levels categorical variable D can be replaced by a combination of three *ad hoc* dummy variables D_1,D_2,D_3, defined as follows. If $D = 1$, then $D_1 = 0$, $D_2 = 0$, $D_3 = 0$; if $D = 2$, then $D_1 = 1$, $D_2 = 0$, $D_3 = 0$; if $D = 3$ then $D_1 = 1$, $D_2 = 1$, $D_3 = 0$ and, finally, if $D = 4$ then $D_1 = 1$, $D_2 = 1$, $D_3 = 1$.

The Cox proportional hazards model, thus, can be written as:

$$h(x|d_1, d_2, d_3) = h_1(x) \exp\{\beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3\} \tag{3.17}$$

where d_1,d_2,d_3 are the realizations of the dummies D_1, D_2, D_3, respectively, and $h_1(x)$ is the baseline hazard function. Specifically, the hazard in the first group is:

$$h(x|d_1 = 0, d_2 = 0, d_3 = 0) = h_1(x), \tag{3.18}$$

in group 2 is:

$$h(x|d_1 = 1, d_2 = 0, d_3 = 0) = h_2 = h_1(x) \exp\{\beta_1\} \tag{3.19}$$

in group 3 :

$$h(x|d_1 = 1, d_2 = 1, d_3 = 0) = h_3(x) = h_1(x) \exp\{\beta_1 + \beta_2\} \tag{3.20}$$

and in group 4:

$$h(x|d_1 = 1, d_2 = 1, d_3 = 1) = h_4(x) = h_1(x) \exp\{\beta_1 + \beta_2 + \beta_3\}, \tag{3.21}$$

where the scalars β_1, β_2 and β_3 are the parameters of the regression model we have to estimate. Now, substituting the hazard functions in eq.s (3.18)-(3.21) into the definition of the θ parameters in eq.s (3.14)-(3.16), it is possible to rewrite the latter as a function of the β parameters:

$$\theta_1 = \frac{h_1(x) \exp\{\beta_1\}}{h_1(x)} = \exp\{\beta_1\} \tag{3.22}$$

$$\theta_2 = \frac{h_1(x) \exp\{\beta_1 + \beta_2\}}{h_1(x) \exp\{\beta_1\}} = \exp\{\beta_2\} \tag{3.23}$$

$$\theta_3 = \frac{h_1(x) \exp\{\beta_1 + \beta_2 + \beta_3\}}{h_1(x) \exp\{\beta_1 + \beta_2\}} = \exp\{\beta_3\}. \tag{3.24}$$

Therefore we can estimate the vector of parameters $\boldsymbol{\theta}$ by estimating the vector of parameters $\boldsymbol{\beta}$. Concerning the latter, we can use the well known estimation techniques for the Cox proportional hazards model based on the maximization of the partial likelihood. Moreover, as the properties of the ML estimators hold for the parameters βs and, given that, the θs are obtained by applying a monotonic and continuous transformation, they maintain the same properties. Under the usual regularity conditions, thus, the estimators $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ are consistent and asymptotically normal distributed.

Finally, substituting the $\hat{\theta}$s in eq. (3.12) we obtain the partial maximum likelihood estimate of the hypervolume under the manifold $\widehat{HUM}_{L4}$:

$$\widehat{HUM}_{L4} = \frac{1}{(\hat{\theta}_3 + 1)(\hat{\theta}_2(\hat{\theta}_3 + 1) + 1)(\hat{\theta}_1(\hat{\theta}_2(\hat{\theta}_3 + 1) + 1) + 1)}. \tag{3.25}$$

The procedure described above is extremely simple and has the enormous advantage of being easily implementable with packages already developed in practically all the statistical software, without the need of *ad hoc* codes to be written by researchers interested in the field. Moreover, as will be largely discussed in the following Chapter 4 devoted to the implementation of the procedure using simulated datasets, the simplicity of the procedure goes in parallel with enormous gains in terms of computational time. Finally, another point of interest, is the relatively simple way of obtaining an analytical formula for the standard error of the estimator of the HUM, as presented in the next section.

3.3 An analytical formula for the variance of the estimator

The asymptotic variance for the HUM_{L4} estimator can be obtained, according to Nze Ossima et al. (2015), with the Delta method.

The variance-covariance matrix for the vector of parameters $\hat{\boldsymbol{\theta}}$ can be decomposed as

$$\begin{aligned} \Sigma_{\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3} &= \begin{bmatrix} \sigma^2_{\hat{\theta}_1} & \sigma_{\hat{\theta}_1,\hat{\theta}_2} & \sigma_{\hat{\theta}_1,\hat{\theta}_3} \\ \sigma_{\hat{\theta}_2,\hat{\theta}_1} & \sigma^2_{\hat{\theta}_2} & \sigma_{\hat{\theta}_2,\hat{\theta}_3} \\ \sigma_{\hat{\theta}_3,\hat{\theta}_1} & \sigma_{\hat{\theta}_3,\hat{\theta}_2} & \sigma^2_{\hat{\theta}_3} \end{bmatrix} \\ &= J^T V J \end{aligned} \tag{3.26}$$

where J is the Jacobian of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$:

$$J = \begin{bmatrix} \exp\{\hat{\beta}_1\} & 0 & 0 \\ 0 & \exp\{\hat{\beta}_2\} & 0 \\ 0 & 0 & \exp\{\hat{\beta}_3\} \end{bmatrix} \tag{3.27}$$

and V is the variance-covariance matrix of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$:

$$V = \Sigma_{\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3} = \begin{bmatrix} \sigma^2_{\hat{\beta}_1} & \sigma_{\hat{\beta}_1,\hat{\beta}_2} & \sigma_{\hat{\beta}_1,\hat{\beta}_3} \\ \sigma_{\hat{\beta}_2,\hat{\beta}_1} & \sigma^2_{\hat{\beta}_2} & \sigma_{\hat{\beta}_2,\hat{\beta}_3} \\ \sigma_{\hat{\beta}_3,\hat{\beta}_1} & \sigma_{\hat{\beta}_3,\hat{\beta}_2} & \sigma^2_{\hat{\beta}_3} \end{bmatrix}. \tag{3.28}$$

Substituting (3.27) and (3.28) in (3.26) the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$ becomes:

$$\begin{aligned} \Sigma_{\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3} &= \begin{bmatrix} \exp\{2\hat{\beta}_1\}\sigma^2_{\hat{\beta}_1} & \exp\{\hat{\beta}_1\}\exp\{\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{\hat{\beta}_1\}\exp\{\hat{\beta}_3\}\sigma_{\hat{\beta}_1\hat{\beta}_3} \\ \exp\{\hat{\beta}_1\}\exp\{\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{2\hat{\beta}_2\}\sigma^2_{\hat{\beta}_2} & \exp\{\hat{\beta}_2\}\exp\{\hat{\beta}_3\}\sigma_{\hat{\beta}_2\hat{\beta}_3} \\ \exp\{\hat{\beta}_3\}\exp\{\hat{\beta}_1\}\sigma_{\hat{\beta}_3\hat{\beta}_1} & \exp\{\hat{\beta}_3\}\exp\{\hat{\beta}_2\}\sigma_{\hat{\beta}_3\hat{\beta}_2} & \exp\{2\hat{\beta}_3\}\sigma^2_{\hat{\beta}_3} \end{bmatrix} \\ &= \begin{bmatrix} \exp\{2\hat{\beta}_1\}\sigma^2_{\hat{\beta}_1} & \exp\{\hat{\beta}_1+\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{\hat{\beta}_1+\hat{\beta}_3\}\sigma_{\hat{\beta}_1\hat{\beta}_3} \\ \exp\{\hat{\beta}_1+\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{2\hat{\beta}_2\}\sigma^2_{\hat{\beta}_2} & \exp\{\hat{\beta}_2+\hat{\beta}_3\}\sigma_{\hat{\beta}_2\hat{\beta}_3} \\ \exp\{\hat{\beta}_3+\hat{\beta}_1\}\sigma_{\hat{\beta}_3\hat{\beta}_1} & \exp\{\hat{\beta}_3+\hat{\beta}_2\}\sigma_{\hat{\beta}_3\hat{\beta}_2} & \exp\{2\hat{\beta}_3\}\sigma^2_{\hat{\beta}_3} \end{bmatrix}. \end{aligned}$$

Using the Delta method, thus, the variance for HUM_{L4} is:

$$\begin{aligned}\sigma^2_{\widehat{HUM}_{L4}} = & \sigma^2_{\hat{\theta}_1}\left(\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_1}\right)^2+\sigma^2_{\hat{\theta}_2}\left(\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_2}\right)^2+\sigma^2_{\hat{\theta}_3}\left(\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_3}\right)^2+\\ & +2\left(\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_1}\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_2}\right)\sigma_{\hat{\theta}_1\hat{\theta}_2}+2\left(\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_1}\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_3}\right)\sigma_{\hat{\theta}_1\hat{\theta}_3}+\\ & +2\left(\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_2}\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_3}\right)\sigma_{\hat{\theta}_2\hat{\theta}_3} \qquad (3.29)\end{aligned}$$

where the single partial derivatives are given by:

$$\begin{aligned}\frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_1} &= -\frac{1}{(\hat{\theta}_3+1)[\hat{\theta}_1(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)+1]^2}\\ \frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_2} &= -\frac{2\hat{\theta}_1(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)-1}{(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)^2[\hat{\theta}_1(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)+1]^2}\\ \frac{\partial\widehat{HUM}_{L4}}{\partial\hat{\theta}_3} &= \frac{-\hat{\theta}_2}{(\hat{\theta}_3+1)[\hat{\theta}_2(\hat{\theta}_3+1)+1]^2[\hat{\theta}_1(\hat{\theta}_2(\hat{\theta}_3+1)+1)+1]}+\\ &- \frac{\hat{\theta}_1\hat{\theta}_2}{(\hat{\theta}_3+1)[\hat{\theta}_2(\hat{\theta}_3+1)+1][\hat{\theta}_1[\hat{\theta}_2(\hat{\theta}_3+1)+1]+1]^2}+\\ &- \frac{1}{(\hat{\theta}_3+1)^2[\hat{\theta}_2(\hat{\theta}_3+1)+1][\hat{\theta}_1(\hat{\theta}_2(\hat{\theta}_3+1)+1)+1]}. \qquad (3.30)\end{aligned}$$

This result is extremely interesting as, to the best of our knowledge, it represents the first time a standard error for the HUM related to a four-class classification problem has been derived analytically. Being all the derivatives derived analytically, the empirical calculation is extremely simple and notably faster than any other technique based on simulation. Examples will be provided in the next Chapter 4.

3.4 Validation of the Lehmann assumption

As we have largely discussed in the previous sections, the HUM_{L4} estimator is derived under the Lehmann assumption, i.e. by imposing a proportionality among the distributions of the different groups. Conditional on the validity of such an assumption, we can consider our estimator more or less reliable. Here below, we describe two different approaches to check the validity of the Lehmann assumption: the first is a graphical one and the other is a

statistical test. Before introducing the graphical method, we briefly recall some relationship satisfied by the proportional hazards model (for details see Appendix A).

In general, the cumulative hazard function $H(t)$ relates to the survival function $S(t)$ by $H(t) = -log(S(t))$. Given that under the Cox model $H(t) = H_0(t)\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}$ then it is possible to write $S(t) = S_0^{\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}}$. From this follows that the plot of the $log(-logS(t))$ function versus the $t/log(t)$ should produce approximately parallel curves. Hence, a graphical evidence of crossing curves is indicative of violation of proportional hazards assumption.

Even though graphical methods are easy to implement, they present some limitations. In fact, it is not possible to quantify to what extent the curves are far from the perfect parallel situation and, consequently, the decision to accept the proportionality assumption can be subjective.

Another way to investigate the departure from Lehmann condition is through a test based on the residuals of the model. The test proposed here below is a re-adaptation of the work developed by Grambsch and Therneau (1994). The authors take inspiration from the proportional hazards regression model by Cox (1972). As described in Appendix A, the Cox model describes the hazard rate of a particular event to occur at a certain moment in time as a combination of a baseline hazard function and some covariates. Specifically, it is assumed the effect of the covariates on the hazards to be multiplicative. Moreover, the effect of such covariates is transmitted to the hazards through fixed coefficients, generally indicated by the vector $\boldsymbol{\beta}$.

The general assumption in the Cox regression is that such parameters remain fixed, and do not depend on the time t. The test proposed by Grambsch and Therneau (1994) aims to check about the plausibility of fixed $\boldsymbol{\beta}$ when compared to a generic alternative function $\boldsymbol{\beta}(t)$. This test represents a generalization of many other contributions in the same field, which, instead, focused on specific deviations from the proportional hazards (see, among others, Cox, 1972; Schoenfeld, 1980 and Wei, 1984).

Using the traditional notation, the hazard at time t for the i-th subject with covariates $\boldsymbol{Z}_i$ can be written as

$$h(t) = h_1(t)\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}$$

where $h_1(t)$ is the baseline hazard function and describes the risk for the subject with $\boldsymbol{Z}_i = 0$, while $\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}$ is the relative risk associated with the set of covariates $\boldsymbol{Z}_i$. The null hypothesis is that the vector of parameters $\boldsymbol{\beta}$ remains constant over time. The alternative, instead, is that the generic parameter β_j, $j = 1, \ldots, p$, with p indicating the number of covariates, is

time-varying, and specifically, taking the form

$$\beta_j(t) = \beta_j + \delta_j\, g_j(t)$$

where $g_j(t)$ is a predictable function of t (e.g., a polynomial in t). The null hypothesis, thus, reduces to a joint test on $\delta_j = 0$, $j = 1, \ldots, p$. Grambsch and Therneau (1994) define a test statistics, which depends on the function $g(t)$, in general indicated by $T(G)$, whose asymptotic distribution follows a χ^2 with p degrees of freedom.

If we move from the traditional specification of the Cox model, generally used in survival analysis, to the one introduced in Section 3.2, where the time variable is replaced with the diagnostic test result variable and the covariates are the three *ad hoc* dummy variables D_1, D_2, D_3 indicating the different groups, then the Grambsch and Therneau's test can be interpreted as a test for constant hazards within the groups, and proportional between the groups. In fact, if we write the Cox model as in eq. (3.17)

$$h(x|d_1, d_2, d_3) = h_1 \exp\{\beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3\}$$

where x is the generic realization of the diagnostic test X, d_1,d_2,d_3 are the realizations of the dummies D_1, D_2, D_3, respectively, and $h_1(x)$ is the baseline hazard function, then the test verifies the assumption of constant $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$, against the alternative of the βs to be a function of the values of the markers within each group, i.e.

$$\beta_j(x) = \beta_j + \delta_j\, g_j(x),$$

for any $j = \{1, 2, 3\}$. If the null hypothesis $H_0 : \delta_j = 0$, $j = \{1, 2, 3\}$, cannot be rejected, then β_1, β_2, β_3 are constant scalars and, according to eq.s (3.14)-(3.16) and eq.s (3.22)-(3.24), the hazards functions are proportional between groups. The implementation of the test, however, requires the specification of the deterministic function $g_j(x)$, $j = \{1, 2, 3\}$, that could be any simple polynomial in the marker values x, e.g. $g_j(x) = x$.

A test on the β_j coefficients in the Cox regression framework, hence, becomes a test on the Lehmann assumption. This re-adaptation of the Grambsch and Therneau's approach allows to statistically test whether the main theoretical assumption underlying the HUM_{L4} estimator is likely to be supported by the data.

3.5 Assessing separability among the classes

A very important point in any classification issue is whether all the classes are effectively different, at least according to the information we have and, more

importantly, for the marker we use. In the first stages of the experimental design the researcher proposes a distinction between different classes, based on some *a priori* information he has about the phenomenon. However, it might be the case that for some specific marker we are dealing with, two or more classes do not present substantial differences in their distribution patterns. We do not say that the researcher belief was definitely wrong, but rather that the classification power of this specific marker does not allow to discriminate among the different classes. Verifying this issue through appropriate inferential statistical approach would provide the researcher with important suggestions about the prior belief and/or discriminatory power of the marker at hand and decide to pursue the analysis in this direction or maybe, focus on alternative markers.

In the previous sections of this chapter we have considered an overall indicator about the discriminatory power of the marker, such as the HUM. Specifically, in Section 3.2 we have proposed an estimator of the HUM under the assumption that the distributions of the marker in the classes follow a specific ordering and that the relationships among such distributions can be described parametrically, without however imposing any assumption on the distribution within each class. This assumption, that is indicated as the Lehmann assumption, can be verified through the appropriate testing strategy described in Section 3.4. The obtained HUM indicator, thus, can be seen as a measure about the discriminatory power of the marker in classifying the subjects as a whole.

Although it is not the aim of the present research, a by-product of our methodology is that it can easily allow to verify whether the classes are all sufficiently different according to the marker we are dealing with. If this is effectively the case, then it would be worth continuing with the analysis of four distinguished classes and use the previous methodology to quantify the discriminatory power of the marker through the estimated HUM. If, instead, two classes are not sufficiently different, then we should either revise the empirical investigation by changing the marker, or combining it with one more marker (see Chapter 6, below), or even moving to a three-class classification approach, if it is allowed to aggregate the two similar classes.

Thus, the methodology we proposed in the previous sections presents an important advantage in this direction too. Specifically, the Cox proportional hazards regression model, provides a way to estimate the parameters describing the relationships among the different distributions $\theta_i = \exp\{\beta_i\}$, $i = \{1, 2, 3\}$, as shown in eq.s (3.22)-(3.24). Moreover, based on the well known property of the partial likelihood estimator used for estimating the coefficients of the Cox proportional hazards model, under the usual regularity conditions, the inference on the $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ parameters can be conducted

in the usual way, being asymptotically normally distributed. As the generic θ_i coefficient can be interpreted as the relationship between the distribution of marker in the i-th class with respect to that in the *i-1*-th one, the null hypothesis $H_0 : \theta_i = 1$ indicates that the marker's distribution in the two mentioned classes is the same. Given the transformations in eq.s (3.22)-(3.24), the test can be directly implemented on the parameters of the Cox proportional hazards regression model, i.e. $H_0 : \beta_i = 0$. If the null hypothesis cannot be rejected by the data, from a statistical point of view, the two classes are distributed in the same way and, importantly, the marker is unable to discriminate between subjects belonging to the two classes. As we have said before, this represents a signal that the *a priori* distinction among classes is not supported by the data and, as a consequence, either more information should be used in the empirical analysis, such as considering a new marker, or alternatively, if it makes sense, try to move to a simpler three-class classification framework. In this latter perspective, the test could be also repeated sequentially to further reduce the complexity of the problem to the standard dichotomous setting. Furthermore, this simple and standard test of hypothesis can be extremely useful to explain potential low values of the estimated HUM indicator. Investigating the significance of the coefficients one can detect the classes the marker is unable to discriminate among.

3.6 Comparisons between HUM_{L4} and other estimators for the HUM

The literature on ROC surface analysis when the disease presents more than three stages is rather limited. To the best of our knowledge, there are only two contributions focusing on estimators of the hypervolume as a measure of the discrimination accuracy of a continuous biomarker. In this section we discuss the main characteristics of each estimator and make a comparison with our proposal, with particular attention to settings in which they can find application and to the distributional assumptions they need.

The three methods, our contribution and the other two alternatives, draw on the same idea of disregarding the distribution of the marker; in fact, two of them are non parametric while our proposal can be defined as semi-parametric.

The HUM_{EX} estimator, introduced by Nakas and Yiannoutsos (2004) and largely discussed in Section 2.3.2, is a completely non parametric estimator. This approach assumes an inherited order in the marker measurements between the classes and do not account for multiple marker analysis. Thus,

in order to compute the HUM value, the order of the classes should be determined in advance or, in alternative, all possible 4! HUM values must be calculated. The largest HUM will be the sensible measure of the accuracy of the test. The inferential framework is based on the Mann-Whitney U statistic. Clearly, as the hypervolume under the ROC surface is calculated from the ranks of the test measurement results, it is invariant under monotonically increasing transformations. This is, in applied research, a highly desirable property. The variance of the estimator, as suggested by the authors, must be calculated using bootstrap methodology because, when the sample size increases and the number of classes is greater than three, the U-statistic approach presents computational burden.

From an operational point of view, this approach is highly demanding firstly because, as we pointed out in Section 2.3.2, it requires to evaluate all the permutations of the marker measures and secondly due to fact that 4! hypervolumes need to be calculated in order to identify the largest one. This estimator has been implemented in the "Biocomb"R package (Novoselova et al., 2017).

The HUM_{LF} estimator, as well as the HUM_{EX}, does not require any assumption on the functional form of the distribution of the biomarker in the population. Furthermore, the condition on the ordering of test results with respect to the class of disease is relaxed. This estimator has been derived by Li and Fine (2008) trough an approach based on the estimated class probabilities, thus it allows to overcome the limit to analyse one marker at a time. The multinomial logistic regression model simply provides the estimated class probabilities jointly taking into account several markers, moreover any classification method generating estimated class probabilities could be used. It is worth noting that this approach is the only one that accounts for multiple markers. An analytical form for the variance of the estimator is not provided by the authors, they suggest to adopt bootstrap techniques. This estimator has been implemented in R and, as we will see in the following chapter, it is computationally demanding. In a recent package called "mcca", provided by Gao and Li (2018), different methods for the estimation of the class probability vectors, such as classification tree, support vector machine and linear discrimination analysis, are available.

Our proposal, HUM_{L4}, is a semi-parametric estimator in the sense that it is obtained with an approach that does not require a full parametric specification of the marker distribution for the four populations. As we have seen, it is based on the Lehmann assumption, that postulates the existence of a monotone transformation producing marker values with an extreme value distribution without specifying and estimating the transformation. Thus, the only parameters to be estimated are those governing the relationship among

the survival distributions. This approach, like the one of the HUM_{EX}, needs an assumption on the ordering of the disease categories. Even though it presents the same limit as HUM_{EX} in modelling only one marker, it has the advantage of allowing to control for covariate effects and as we have stated in the previous section, to give information on the separation among the classes through the significance of the coefficients of the Cox proportional hazards regression model. The variance of HUM_{L4} has the analytical form obtained with the Delta method showed in Section 3.3. It is computationally very fast due to the fact that relies on proportional hazards framework, thus inference using standard statistical software is enabled.

3.7 Defining the optimal ordering

In many real world biomedical situations the categories of the outcome we are interested in have an intrinsic natural order. However, quite often, the quantitative test results do not display the corresponding actual ordering. As a consequence, the problem of identifying the relative order of test among groups still remains. This is extremely relevant for estimators as HUM_{L4} or HUM_{EX} that depend on the ordering of the test measure over the classes. In a four-class classification problem, as we stated before, in order to find out the highest HUM value it is necessary to calculate all possible 4! HUM values. This might cause high computational efforts.

3.7.1 Relative effects

A way to avoid the computational burden of calculating $M!$ HUM values is to compute the sample mean for each class. Li et al. (2014), in their work, demonstrate (Theorem 3.1) that, if the test result is Gaussian distributed, the order of the sample means can be used to prescribe the order of the M classes. As in practice the data could present an empirical distribution far from the normal one even after normalization attempts, parametric methods are not always suitable. To overcome this issue Li et al. (2014), suggest a distribution-free non parametric method. The idea is to sort categories using a descriptive statistic called *relative effect*. More in detail, for each test X from each category k, the relative effect e_k can be calculated based on the cumulative distribution of the marker in the k-th group, denoted by $G_k(x)$. Specifically, the relative effect is defined as:

$$e_k = \int G^+(x) \mathrm{d}G_k(x), \qquad k = 1, 2, \ldots, M$$

where

$$G^{+} = \sum_{k=1}^{M} w_k G_k(x), \qquad \text{with} \qquad \left(\sum_{k=1}^{M} w_k = 1 \right)$$

is a weighted average of the k distribution functions G_k and w_k is the relevance of the k-th category. In order to define the correct ordering based upon the sample data, the estimated relative effect $\hat{e_k}$ can be obtained by replacing the unknown distribution functions with their empirical counterpart:

$$\hat{G_k}(x) = \frac{\sum_{i=1}^{n_k} \mathbb{1}(X_{ik} \leq x)}{n_k}.$$

The ordering of the relative effects of each group suggests the optimal ordering of the groups maximizing the HUM. Furthermore, Li et al. (2014) prove the almost surely convergence of the sample relative effect to the true one. The $\hat{e}_k$ can be used to determine the category ordering before applying our hypervolume estimator. As we will see later in the empirical applications, implementing this method allows us to save a huge amount of computational time when estimating the HUM.

Chapter 4

Simulation studies

In this chapter we present a comparison among three different approaches in estimating the HUM value. We report simulations studies with different scenarios in order to evaluate the impact of different data generating processes (DGP) on the estimators performances.

4.1 General setting

We hypothesize a setting in which a diagnostic marker is evaluated in a sample with subjects belonging to four different groups, for example four stages of disease. We consider a set of scenarios according to different sample size and different characteristics of the DGP. As we have deeply discussed in the previous sections, an important assumption characterizing our approach is the Lehmann condition, which is however not necessary for the other estimators already existing in the literature. An important distinction in the following simulation exercises is whether the DGP satisfies such condition or not.

As we have introduced in Section 2.3.6, the Lehmann assumption states that

$$S_2 = S_1^{\theta_1}; \qquad S_3 = S_2^{\theta_2}; \qquad S_4 = S_3^{\theta_3}$$

where S_i, with $i = 1, \ldots, 4$, represents the survival function in each stage of disease, and θ_1, θ_2 and θ_3 are the parameters. The main difficulty in the simulation exercises is that the Cox model, as reported in Section 3.1, is formulated in terms of the hazard functions, while we need to generate the data starting from probability distributions. To overcome the problem we apply the Monte Carlo inversion method. Given that

$$S(x) = \exp\{-H_1(x)\exp(\boldsymbol{\beta}^{'}\boldsymbol{d})\}, \tag{4.1}$$

where $H_1(x) = \int_0^x h_1(u)\mathrm{d}u$ is the cumulative hazard function of the baseline class and $\boldsymbol{d}$ is the vector of appropriate dummy variables to detect the class, it is possible to obtain the distribution function under this model, i.e.

$$F(x) = 1 - \exp\{-H_1(x)\exp(\boldsymbol{\beta}'\boldsymbol{d})\}. \tag{4.2}$$

Given that $F(x)$ is a cumulative distribution function, it can assume values in the interval $(0, 1)$. Denoting with X the marker value in the Cox model and with Y the Uniform distribution, $Y \sim Uni(0, 1)$, then from eq. (4.2) it is possible to write:

$$Y = \exp\left\{-H_1(x)\exp\{\boldsymbol{\beta}'\boldsymbol{d}\}\right\} \sim Uni(0, 1). \tag{4.3}$$

If $h_1(x) > 0$ for all x, then H_1 can be inverted and X can be obtained following the algebra below:

$$\begin{aligned} log(y) &= -H_1(x)\exp\{\boldsymbol{\beta}'\boldsymbol{d}\} \\ -H_1^{-1}[log(y)] &= x\exp\{\boldsymbol{\beta}'\boldsymbol{d}\} \\ x &= H_1^{-1}[-log(y)\exp\{-\boldsymbol{\beta}'\boldsymbol{d}\}]. \end{aligned} \tag{4.4}$$

Generating random values for the marker, thus, reduces to generate random numbers from a $Uni(0, 1)$ distribution and make the transformation in eq. (4.4) as shown in Bender et al. (2005).

One of the probability distribution functions that presents the proportional hazards property is the Weibull distribution. The hazard function of a Weibull random variable X is:

$$h(x) = \lambda\nu x^{\nu-1} \tag{4.5}$$

where $\lambda \in \Re^+$ is the scale parameter and $\nu \in \Re^+$ is the shape parameter. Its cumulative hazard function is:

$$H_1(x) = \lambda x^{\nu} \tag{4.6}$$

and the inverse cumulative hazard function is:

$$H_1^{-1}(x) = \lambda^{-1}x^{1/\nu}. \tag{4.7}$$

Applying the transformation in eq. (4.4), the marker variable can be expressed as:

$$\begin{aligned} x &= \lambda^{-1}[-log(y)\exp\{-\boldsymbol{\beta}'\boldsymbol{d}\}]^{1/\nu} \\ &= -\left(\frac{log(y)}{\lambda\exp\{\boldsymbol{\beta}'\boldsymbol{d}\}}\right)^{1/\nu}. \end{aligned} \tag{4.8}$$

The corresponding hazard function in the Cox model is:

$$h(x|\boldsymbol{d}) = \lambda\nu x^{\nu-1}\exp\{\boldsymbol{\beta}'\boldsymbol{d}\}, \tag{4.9}$$

meaning that the marker we generated follows a Weibull distribution with scale parameter $\lambda(\boldsymbol{d}) = \lambda\exp\{\boldsymbol{\beta}'\boldsymbol{d}\}$ and fixed shape parameter equal to ν. In the first simulation exercise, the Cox regression model with known regression coefficients is simulated.

4.2 Data Generating Process under the Lehmann condition

The first set of simulations are obtained by assuming that the Lehmann condition is met. As described above, we generate the data starting from different Weibull distributions according to the procedure proposed in eq.s (4.4)-(4.9). Specifically, we generate from a Cox proportional hazards model with different vectors of parameters $\boldsymbol{\beta} = (\beta_1; \beta_2; \beta_3)'$ and groups covariate vector d. The hazard functions, thus, assume the form

$$h_i(t|d_1, d_2, d_3) = \begin{cases} h_1(t) & d_{1i} = 0, d_{2i} = 0, d_{3i} = 0 \\ h_1(t)\exp\{\beta_1 d_{1i}\} & d_{1i} = 1, d_{2i} = 0, d_{3i} = 0 \\ h_1(t)\exp\{\beta_1 d_{1i} + \beta_2 d_{2i}\} & d_{1i} = 1, d_{2i} = 1, d_{3i} = 0 \\ h_1(t)\exp\{\beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i}\} & d_{1i} = 1, d_{2i} = 1, d_{3i} = 1. \end{cases}$$

Generating under these conditions corresponds to assume that $X_i \sim Wei(\lambda_i, \nu)$ is the marker's distribution in the i-th class, with $i = 1, \ldots, 4$. In other words, we consider distributions with class-specific scale parameter and constant shape parameter. The three cases presented below simply differ according to different vectors of parameters $\boldsymbol{\beta}$.

Interestingly, for any scenario, it is possible to calculate the exact value of the HUM. The true HUM, thus, represents the reference value for comparing the estimated one and for evaluating the performance of the estimators in terms of the bias.

The first scenario, denote with *Case 1*, is characterized by the vector of parameters $\boldsymbol{\beta} = (-1.4; -0.8; -0.6)'$. In detail, the four distributions characterizing the four classes of disease are:

$$\begin{aligned} X_1 &\sim Wei(4, 2) \\ X_2 &\sim Wei(4 * \exp\{-1.4\}, 2) \\ X_3 &\sim Wei(4 * \exp\{-1.4 - 0.8\}, 2) \\ X_4 &\sim Wei(4 * \exp\{-1.4 - 0.8 - 0.6\}, 2). \end{aligned}$$

As said above, in the second and third scenario we use the same strategy as for *case 1* but with different vectors of parameters $\boldsymbol{\beta}$. Specifically, the second scenario (*case 2*) is characterized by the parameters $\boldsymbol{\beta} = (-2.5; -1.2; -1.7)'$, while the third one (*case 3*) by the parameters $\boldsymbol{\beta}$=$(-4.1; -3.5; -3.8)'$. In Figure 4.1 we show the four probability density functions we use for generating the data in the three cases (*case 1*, *case 2* and *case 3*). As can be easily noted from the figure, from *case 1* to *case 3* the distributions become more and more separate. The true HUMs, thus, are expected to increase from *case 1* to *case 3*. In fact, when solving the integral in eq. (3.13), the true values of the HUM for *case 1*, *case 2* and *case 3* are 0.263, 0.561 and 0.933, respectively.

Finally, for each of the three cases, we generate the data for different sample sizes, N=120, N=200 and N=320, under the assumption of groups of equal dimension (30 individuals, 50 individuals and 80 individuals, respectively). These sample sizes are absolutely in line with clinical studies using real data.

The estimation results are shown in Table 4.1. For each of the three estimators presented in Chapter 3: the HUM_{EX} by Nakas and Yiannoutsos (2004) presented in Section 2.3.2, the HUM_{LF} by Li and Fine (2008) described in Section 2.3.4 and our estimator HUM_{L4} derived in Section 3.1, the first three columns of the table report the characteristics of the DGP, in terms of the parameters β, the true value of the HUM, and the sample size of each simulation. The following columns, instead, are devoted to the simulation results, for each of these estimators we show the estimated HUM (hum), the standard error (se) and the bias, expressed both in absolute value and in percentage, with respect to the true HUM. Concerning the calculation of the standard errors, for our estimator we use the formula in eq. (3.29), while for the other two estimators, as there do not exist analytical results, we apply the bootstrap technique.

Looking at Table 4.1, for *case 1* and *case 2* the HUM_{L4} and HUM_{EX} perform very well and in a very similar way. Both estimators present reduced bias even in small sample, although our estimator has systematically smaller standard errors. For *case 3*, these first two estimators have practically no bias. In all the three cases, the third estimator, HUM_{LF}, performs systematically worse, both in terms of bias and standard errors.

Interestingly, in Table 4.2 we report, for each simulation, the computational times, in seconds, for obtaining the results presented in Table 4.1. The clear message, from this table, is that our estimator is computationally extremely efficient when compared to the other two estimators. Moreover, the computational time for the HUM_{L4} remains practically the same as the sample size increases while it explodes for the other two, in particular for

the HUM_{LF}. This is clearly not a surprise, as the HUM_{LF} estimator needs to compute, for each subject of each class, all the combinations of the four probabilities to belong to a class, to find out the HUM value.

Overall, when the DGP supports the Lehmann condition, our estimator performs extremely well both in small and larger samples, and reveals to be extremely efficient in term of computational time. Among the other two estimators, the HUM_{EX} also performs very well, but the computational time is much longer and increases exponentially with the size of the sample. The HUM_{LF} estimator, under the Lehmann condition, presents the poorest performances in term of bias, precision and computational time.

	$\boldsymbol{\beta}$	true hum	N	est. hum	se	bias (abs value)	bias %
				HUM_{L4}			
			120	0.270	0.042	0.006	2.133
case 1	$(-1.4; -0.8; -0.6)'$	0.264	200	0.267	0.032	0.003	0.996
			320	0.266	0.026	0.002	0.638
			120	0.565	0.053	0.004	0.712
case 2	$(-2.5; -1.2; -1.7)'$	0.561	200	0.563	0.041	0.002	0.396
			320	0.561	0.033	<0.001	0.069
			120	0.933	0.024	<0.001	0.005
case 3	$(-4.1; -3.5; -3.8)'$	0.933	200	0.933	0.018	<0.001	0.034
			320	0.933	0.015	<0.001	0.036
				HUM_{EX}			
			120	0.270	0.047	0.006	2.283
case 1	$(-1.4; -0.8; -0.6)'$	0.264	200	0.266	0.037	0.003	0.952
			320	0.265	0.030	0.001	0.283
			120	0.565	0.061	0.004	0.740
case 2	$(-2.5; -1.2; -1.7)'$	0.561	200	0.563	0.048	0.002	0.360
			320	0.561	0.039	<0.001	0.083
			120	0.934	0.032	0.001	0.130
case 3	$(-4.1; -3.5; -3.8)'$	0.933	200	0.933	0.025	<0.001	0.012
			320	0.933	0.020	<0.001	0.006
				HUM_{LF}			
			120	0.252	0.048	0.012	4.609
case 1	$(-1.4; -0.8; -0.6)'$	0.264	200	0.249	0.038	0.015	5.563
			320	0.248	0.031	0.016	6.127
			120	0.531	0.063	0.030	5.315
case 2	$(-2.5; -1.2; -1.7)'$	0.561	200	0.529	0.049	0.032	5.680
			320	0.526	0.038	0.035	6.218
			120	0.925	0.035	0.008	0.879
case 3	$(-4.1; -3.5; -3.8)'$	0.933	200	0.922	0.027	0.011	1.175
			320	0.922	0.022	0.011	1.205

Table 4.1: Simulation results for the Weibull case under the Lehmann assumption and three different vectors of parameters $\boldsymbol{\beta}$ of the Cox proportional hazards regression model. Bias is expressed in absolute value, while the percentage is calculated with respect to the true HUM value.

		HUM_{L4}	HUM_{EX}	HUM_{LF}
	n_1=n_2=n_3=n_4=30	2.93	78.510	506.79
case 1	n_1=n_2=n_3=n_4=50	3.09	339.220	3052.530
	n_1=n_2=n_3=n_4=80	3.55	1430.370	19739.520
	n_1=n_2=n_3=n_4=30	3.04	76.69	502.89
case 2	n_1=n_2=n_3=n_4=50	3.22	347.85	2782.18
	n_1=n_2=n_3=n_4=80	3.64	1458.08	20393.03
	n_1=n_2=n_3=n_4=30	3.08	76.42	479.56
case 3	n_1=n_2=n_3=n_4=50	3.20	342.68	2751.72
	n_1=n_2=n_3=n_4=80	3.69	1445.08	20036.58

Table 4.2: Computational time in seconds for the Weibull case under the Lehmann assumption and three different vectors of parameters $\boldsymbol{\beta}$ of the Cox proportional hazards regression model.

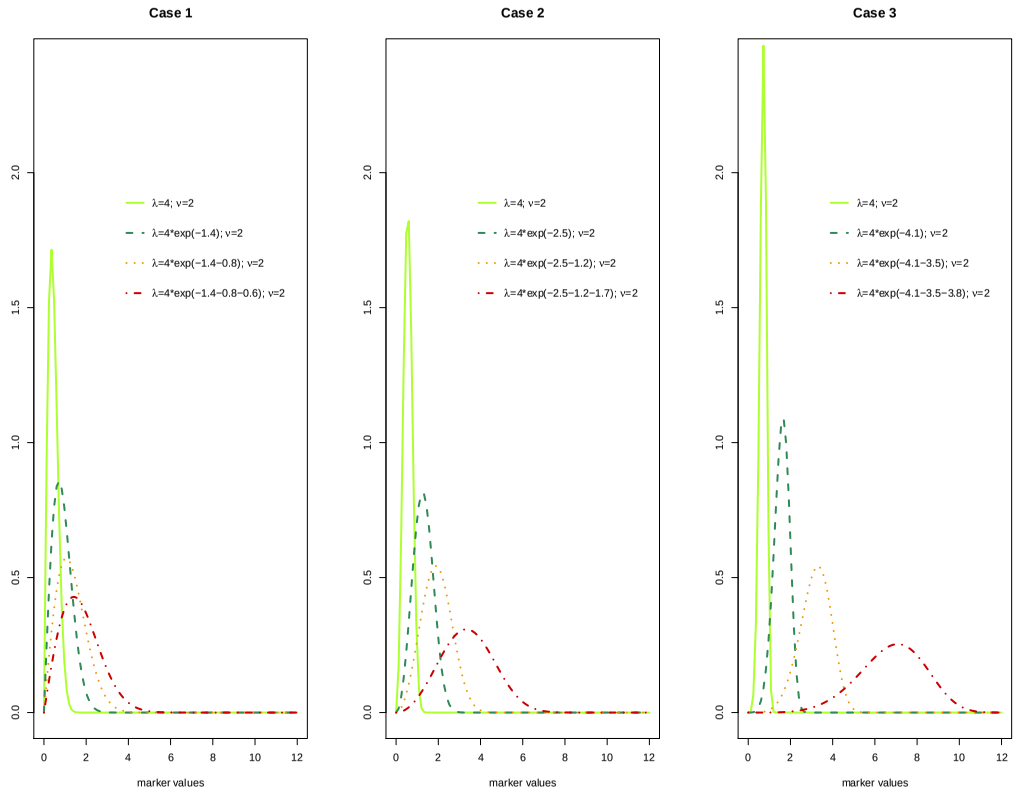


Figure 4.1: Cases: 1,2,3. Representation of the marker distributions under the Lehmann condition and three different Weibull parametrizations.

4.3 Data Generating Process under departures from the Lehmann condition

In this section we evaluate the performance of our estimator when the data are generated under less favourable conditions. Our estimator, the HUM_{L4}, is based on the validity of the Lehmann condition. The DGP featuring the next simulation exercises presents departures from the Lehmann condition in different directions. In particular, we first continue to consider Weibull distributions, although with group-specific shape parameters $\nu(d)$, then we move to Gaussian distributions, for which the Lehmann condition is never satisfied.

4.3.1 DGP from Weibull distributions with group-specific shape parameters

In Section 4.1 we have presented the case of Weibull distributions with the same shape parameter ν, for which the Lehmann condition does hold. When the shape parameter is group-dependent, i.e. $\nu = \nu(d_i)$, the Lehmann condition is not supported anymore. The next three scenario, *case 4* to *case 6*, consider samples generated from Weibull distributions with different scale (λ) and shape (ν) parameters, each corresponding to a different HUM value. In particular, from *case 4* to *case 6*, the distributions are progressively more separated, showing thus increasing values of the true HUM. Figure 4.2 shows the distributions of the three cases described here below.

In *case 4* we generate using the Monte Carlo inversion method we illustrated in Section 4.1 with vector of parameters $\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$ and $\nu_1 = 5$, $\nu_2 = 4$, $\nu_3 = 2$, $\nu_4 = 5$. The real HUM is 0.234. The results on the performance of the three estimators are shown in Table 4.3. The structure of the table is the same as before, where we first reports the parameters of the simulation setting, then the performance of the estimators in terms of estimated HUM, standard errors and bias. In all cases but one, regardless of the size of the sample, all the estimators tend to underestimate the true value of the HUM, with a remarkable bias. The only exception is represented by the HUM_{EX} estimator, that in large samples performs extremely well.

In *case 5* we generate using the Monte Carlo inversion method with vector of parameters $\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$ and $\nu_1 = 2$, $\nu_2 = 2$, $\nu_3 = 3$, $\nu_4 = 4$. The true value of the HUM is 0.553. The performance of the three estimators is practically the same as in the previous *case 4*, with a tendency of underestimating the value of the true HUM, except for the HUM_{EX} estimator in large samples.

In *case 6*, the data are generated with parameters $\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$ and $\nu_1 = 2$, $\nu_2 = 2.5$, $\nu_3 = 4$, $\nu_4 = 7$. As can be seen in Figure 4.2, right panel, the four distributions are rather well separated; this is consistent with the extremely high real HUM, which is 0.836. The results of the simulation exercise are shown in Table 4.3. The bias is quite large for all the estimators, which continue to systematically underestimate the true value of the HUM. Interestingly, for all the estimators, the bias slightly increases with the raise of the sample size.

In general, we can remark that our estimator, the HUM_{L4}, although the failure of the Lehmann condition, continues to perform systematically better than the HUM_{LF}, in terms of both bias and precision (standard error). Regardless the sample size and the magnitude of the hum, the best estimator in terms of bias is the HUM_{EX}.

In Table 4.4 we show the computational time for estimating the HUM for the three estimators. It clearly emerges that our estimator is much faster, regardless of the size of the sample. The computational time, instead, explodes with both HUM_{EX} and HUM_{LF}.

	$\boldsymbol{\beta}$	true hum	N	est. hum	se	bias (abs value)	bias %
				HUM_{L4}			
	$\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$		120	0.160	0.024	0.075	31.972
case 4	$\nu_1 = 5$, $\nu_2 = 4$,	0.234	200	0.159	0.018	0.076	32.332
	$\nu_3 = 2$, $\nu_4 = 5$		320	0.158	0.015	0.076	32.447
	$\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$		120	0.481	0.055	0.073	13.150
case 5	$\nu_1 = 2$, $\nu_2 = 2$,	0.553	200	0.478	0.045	0.076	13.644
	$\nu_3 = 3$, $\nu_4 = 4$		320	0.475	0.035	0.079	14.267
	$\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$		120	0.721	0.049	0.115	13.753
case 6	$\nu_1 = 2$, $\nu_2 = 2.5$,	0.836	200	0.716	0.039	0.120	14.375
	$\nu_3 = 4$, $\nu_4 = 7$		320	0.713	0.032	0.123	14.683
				HUM_{EX}			
	$\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$		120	0.186	0.034	0.048	20.558
case 4	$\nu_1 = 5$, $\nu_2 = 4$,	0.234	200	0.182	0.028	0.053	22.484
	$\nu_3 = 2$, $\nu_4 = 5$		320	0.235	0.028	0.001	0.237
	$\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$		120	0.513	0.059	0.040	7.283
case 5	$\nu_1 = 2$, $\nu_2 = 2$,	0.553	200	0.514	0.048	0.040	7.195
	$\nu_3 = 3$, $\nu_4 = 4$		320	0.554	0.038	0.001	0.094
	$\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$		120	0.773	0.049	0.063	7.508
case 6	$\nu_1 = 2$, $\nu_2 = 2.5$,	0.836	200	0.771	0.038	0.064	7.715
	$\nu_3 = 4$, $\nu_4 = 7$		320	0.771	0.030	0.065	7.818
				HUM_{LF}			
	$\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$		120	0.117	0.042	0.118	50.175
case 4	$\nu_1 = 5$, $\nu_2 = 4$,	0.234	200	0.111	0.033	0.124	52.768
	$\nu_3 = 2$, $\nu_4 = 5$		320	0.104	0.023	0.130	55.513
	$\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$		120	0.473	0.058	0.080	14.540
case 5	$\nu_1 = 2$, $\nu_2 = 2$,	0.553	200	0.473	0.047	0.080	14.487
	$\nu_3 = 3$, $\nu_4 = 4$		320	0.470	0.036	0.083	15.049
	$\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$		120	0.714	0.056	0.122	14.547
case 6	$\nu_1 = 2$, $\nu_2 = 2.5$,	0.836	200	0.711	0.044	0.125	14.939
	$\nu_3 = 4$, $\nu_4 = 7$		320	0.710	0.035	0.126	15.073

Table 4.3: Simulation results for the Weibull case with group-specific shape parameters (Lehmann condition is not satisfied) and three different vectors of parameters $\boldsymbol{\beta}$. Bias is expressed in absolute value, while the percentage is calculated with respect to the true HUM value.

		HUM_{L4}	HUM_{EX}	HUM_{LF}
case 4	n_1=n_2=n_3=n_4=30	2.95	77.39	508.87
	n_1=n_2=n_3=n_4=50	3.14	346.70	2802.82
	n_1=n_2=n_3=n_4=80	3.59	1471.26	20311.33
case 5	n_1=n_2=n_3=n_4=30	3.01	76.66	515.41
	n_1=n_2=n_3=n_4=50	3.20	344.50	2813.94
	n_1=n_2=n_3=n_4=80	3.61	1448.56	20181.73
case 6	n_1=n_2=n_3=n_4=30	3.01	76.39	498.56
	n_1=n_2=n_3=n_4=50	3.16	343.69	2800.65
	n_1=n_2=n_3=n_4=80	3.61	1453.50	20238.43

Table 4.4: Computational time for the Weibull case with group-specific shape parameters (Lehmann condition is not satisfied) and three different vectors of parameters $\boldsymbol{\beta}$ of the Monte Carlo inversion method.

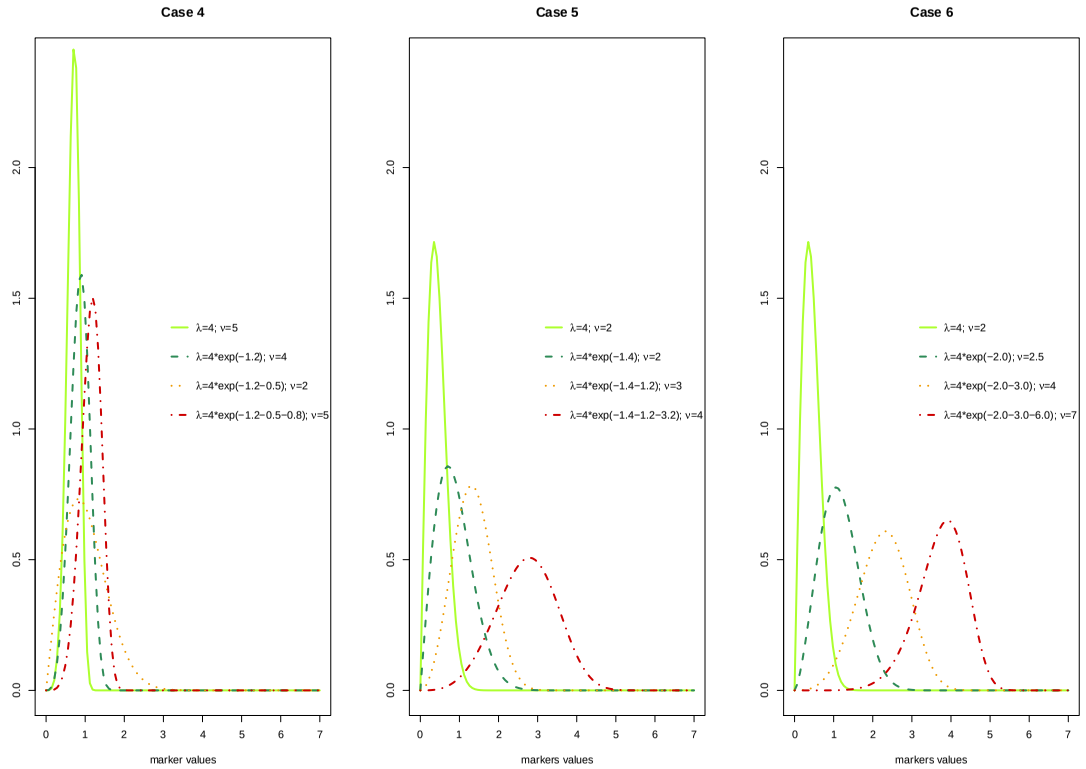


Figure 4.2: Cases: 4, 5 and 6. Representation of the marker distributions using different Weibull distributions with group-specific shape parameters and, thus, featuring departure from the Lehmann condition.

4.3.2 DGP from Normal distributions with equivalent variances

The following set of simulations continues to move away from the Lehmann condition, and focuses on DGPs based on Gaussian distributions. We consider three new scenarios - *case 7*, *case 8* and *case 9* - where the Normal variables have different group-specific expected values but constant variances. As for the previous cases, the parameters are selected in order to have differently separated distributions and, accordingly, featuring different values of the real HUM. The probability density functions of the distributions for each scenario are reported in Figure 4.3. As before, in each scenario we consider three different sample sizes, with groups having the same dimension.

The results of the simulation exercises are shown in Table 4.5, where we report the parameters of the DGP, the true value of the HUM, the sample size N, and the performance (estimated value, standard error and bias) of each of the three estimators: HUM_{L4}, HUM_{EX} and HUM_{LF}.

As can be seen in Figure 4.3, *case 7* is characterized by highly overlapped distributions. The expected values are $\mu_1 = 0.1$, $\mu_2 = 0.3$, $\mu_3 = 0.5$, $\mu_4 = 0.7$, while the standard deviation is common in all groups, $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$. As expected, the true HUM is very small and amounts to 0.077. Focusing on the performances of the estimators, it emerges that, differently with respect to the Weibull distributions, the best estimator is the HUM_{LF}, that presents a very small bias for small and large samples. The HUM_{L4} and HUM_{EX}, instead, show similar bias, in term of magnitude, although the former tends to underestimate, while the latter to overestimate.

The *case 8* is characterized by Normal distributions with expected values $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, $\mu_4 = 4$ and constant standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$. The real value of the HUM is 0.369. For this scenario, regardless of the size of the sample, the HUM_{EX} estimator performs systematically better, in terms of bias, than the other two estimators. Our estimator, instead, presents the highest bias, around the double than those observed for the HUM_{LF}. However, HUM_{L4} continues to be the most efficient one, presenting the lowest values of the standard error for all sample sizes.

The last simulation, *case 9*, considers rather separated Gaussian distributions, with parameters given by $\mu_1 = 1$, $\mu_2 = 3$, $\mu_3 = 5$, $\mu_4 = 7$ and $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$. The true HUM is 0.771. The HUM_{EX} estimator practically presents no bias and a very small standard error. Both the HUM_{L4} and HUM_{LF} substantially overestimate the true HUM, with worse performance for the former. Interestingly, our HUM_{L4} shows slightly increasing bias the larger the sample size. This is due to the fact that the

HUM_{L4} is based on the Lehmann condition that, as said before, is at odds with DGPs coming from Gaussian distributions. The larger is the sample size, the bigger we expect to be the bias.

In Table 4.6 we report the computational time for *case 7*, *case 8* and *case 9*. What emerges, practically, is that our estimator continues to be enormously less time consuming than the other two alternatives, in particular when the sample size does increase.

Overall, when considering departures from the Lehmann conditions originating from Gaussian distributions, the HUM_{L4} estimators show increasing bias the larger the sample size and the more separate variables for the groups. Although such unfavorable conditions, the bias remains relatively contained and in line with the HUM_{LF} estimator, which is, however, much more computationally demanding. The HUM_{EX} estimator, in general, is the one presenting the lowest bias, although it becomes computationally time consuming as the sample size increases.

	parameters β	true hum	N	est. hum	se	bias (abs value)	bias %
			HUM_{L4}				
case 7	$\mu_1 = 0.1$, $\mu_2 = 0.3$	0.077	120	0.069	0.016	0.009	11.218
	$\mu_3 = 0.5$, $\mu_4 = 0.7$		200	0.069	0.012	0.008	10.792
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.069	0.010	0.008	10.633
case 8	$\mu_1 = 1$, $\mu_2 = 2$	0.369	120	0.305	0.050	0.064	17.397
	$\mu_3 = 3$, $\mu_4 = 4$		200	0.300	0.039	0.069	18.633
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.299	0.031	0.071	19.148
case 9	$\mu_1 = 1$, $\mu_2 = 3$	0.771	120	0.672	0.060	0.099	12.818
	$\mu_3 = 5$, $\mu_4 = 7$		200	0.664	0.047	0.107	13.876
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.659	0.039	0.111	14.450
			HUM_{EX}				
case 7	$\mu_1 = 0.1$, $\mu_2 = 0.3$	0.077	120	0.090	0.018	0.012	15.754
	$\mu_3 = 0.5$, $\mu_4 = 0.7$		200	0.084	0.014	0.007	8.552
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.082	0.012	0.004	5.712
case 8	$\mu_1 = 1$, $\mu_2 = 2$	0.369	120	0.365	0.053	0.004	1.206
	$\mu_3 = 3$, $\mu_4 = 4$		200	0.367	0.042	0.003	0.717
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.369	0.034	<0.001	0.037
case 9	$\mu_1 = 1$, $\mu_2 = 3$	0.771	120	0.768	0.049	0.002	0.322
	$\mu_3 = 5$, $\mu_4 = 7$		200	0.769	0.038	0.002	0.248
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.770	0.031	0.001	0.080
			HUM_{LF}				
case 7	$\mu_1 = 0.1$, $\mu_2 = 0.3$	0.077	120	0.077	0.021	0.001	0.844
	$\mu_3 = 0.5$, $\mu_4 = 0.7$		200	0.074	0.016	0.003	4.354
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.073	0.014	0.005	5.928
case 8	$\mu_1 = 1$, $\mu_2 = 2$	0.369	120	0.334	0.052	0.035	9.558
	$\mu_3 = 3$, $\mu_4 = 4$		200	0.337	0.041	0.033	8.875
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.339	0.033	0.030	8.085
case 9	$\mu_1 = 1$, $\mu_2 = 3$	0.771	120	0.703	0.057	0.068	8.768
	$\mu_3 = 5$, $\mu_4 = 7$		200	0.702	0.045	0.068	8.872
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.703	0.036	0.068	8.808

Table 4.5: Simulation results for Normal distributions with group-specific expected values and equal variances (Lehmann condition is not satisfied). Bias is expressed in absolute value, while the percentage is calculated with respect to the true HUM value.

		HUM_{L4}	HUM_{EX}	HUM_{LF}
	n_1=n_2=n_3=n_4=30	2.92	77.55	493.97
case 7	n_1=n_2=n_3=n_4=50	3.15	348.39	2814.59
	n_1=n_2=n_3=n_4=80	3.57	1469.82	20528.60
	n_1=n_2=n_3=n_4=30	2.95	77.39	506.53
case 8	n_1=n_2=n_3=n_4=50	3.14	346.70	2792.03
	n_1=n_2=n_3=n_4=80	3.59	1471.26	19866.11
	n_1=n_2=n_3=n_4=30	2.86	76.33	504.18
case 9	n_1=n_2=n_3=n_4=50	3.12	346.87	2795.04
	n_1=n_2=n_3=n_4=80	3.54	1450.50	20416.80

Table 4.6: Computational time for the Normal case with group-specific expected values and equivalent variances.

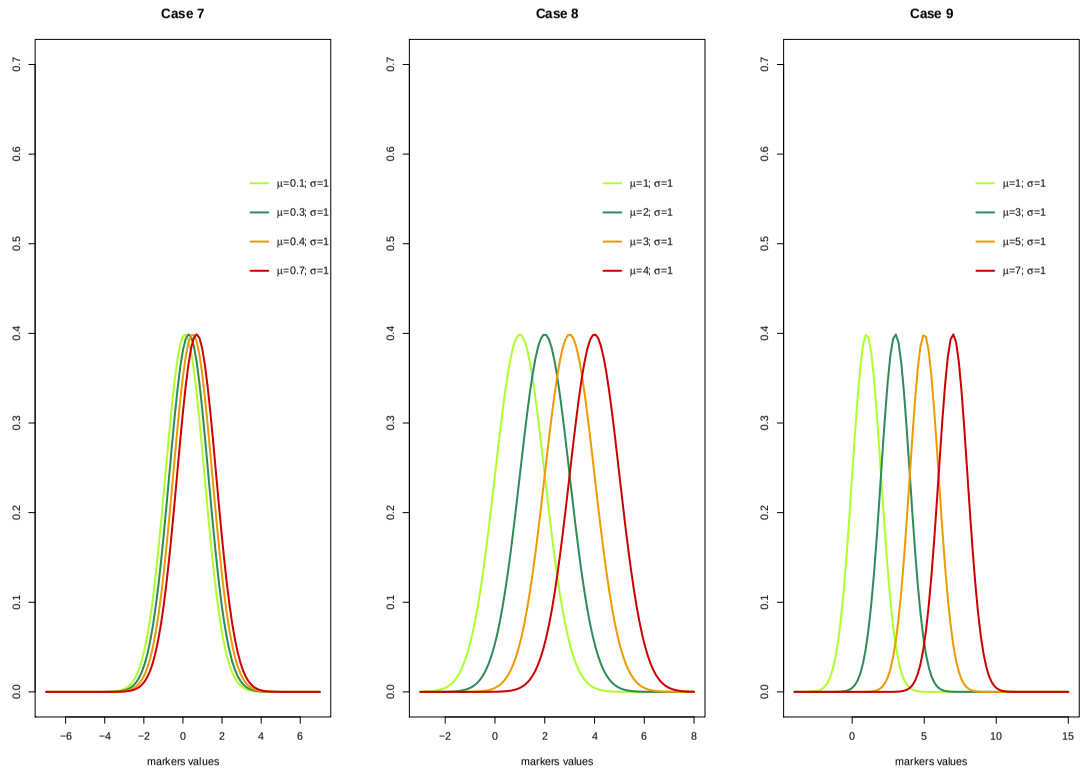


Figure 4.3: Probability density functions of the Normal distributions in *case 7* (left panel), *case 8* (middle panel) and *case 9* (right panel).

Chapter 5

Empirical applications

In this chapter we report two empirical applications of the methodology developed in Chapter 3 on how one can measure the accuracy of a classifier in a four-class classification framework. The first empirical study investigates the classification power of four blood markers in detecting four different levels of colorectal disease, up to the highest level characterizing the full-blown colorectal cancer. The second empirical study, instead, concerns a problem of immunohistological synovial tissue classification presented in Della Beffa et al. (2013) and broadly used in the literature on multicategory diagnostic accuracy (see for example (Li et al., 2017)).

5.1 Blood markers for colorectal cancer

Cancer detection at early stage has been one of the main research topics undertaken by the scientific community over the recent decades. In particular, colorectal cancer (henceforth CRC) has been largely studied as it represents the last stage of genetic and molecular alterations of the originating tumour Such a process is generally quite slow (in some case up to 15 years) and having early detection can represent the most effective strategy for a complete recovery.

As the deterioration from pre malignant lesion to carcinoma and metastasis involves several molecular events, the idea of detecting solid tumours through simple blood tests has received growing interest. At the same time, medical and chemical research has enormously widen the amount of testable components using human blood samples, including cell-free DNA (cfDNA) and RNA (cfRNA), as well as proteins and circulating vesicles, known as exosomes.

With the aim of early detection of CRC, over the recent years many coun-

tries have promoted a massive campaign to resort to faecal immunochemical test (FIT) as a simple, non-invasive and acceptable test. As an example, in most Italian regions, positive patients are considered as those with an FIT> 100 ng/ml. Under these premises, about 5.5% of the screened population (aged 50-69 years) will be FIT positive and only 2.9% will receive a diagnosis of CRC and 20.1% a diagnosis of advanced adenomas (at first round). The rate of false positive FITs, hence, is the incentive for the investigation of new non-invasive and more specific screening tests, including blood markers.

5.1.1 Data and descriptive statistics

The data we analyse come from a recent study on colorectal cancer conducted with the purpose of evaluating a panel of four messenger RNAs (mRNAs) as putative markers of the cancer (Rodia et al., 2018). Specifically, the authors tested four markers: carcinoembryonic antigen-related cell-adhesion molecule 6 (CEACAM6), lectin galactoside binding soluble 4 (LGALS4), tetraspanin 8 (TSPAN8), collagen type I alpha 2 chain (COL1A2), hereafter referred to with the acronym of CELTiC (CEACAM6, LGALS4, TSPAN8 and COL1A2), on subjects positive to the faecal immunochemical test (FIT) and undergoing colonoscopy. The researchers investigated 231 participants that can be classified in four distinct groups: 67 healthy subjects (N), 36 FIT positive with negative colonoscopy (NFIT), 36 low risk that is FIT positive with small polyps (LR), 92 FIT positive with advanced adenomas or a histologically confirmed diagnosis of colorectal cancer (HR/CCR). Before proceeding with the discussion, a clarification about the study design is in order. In fact, as recognized by the authors, the study presents some limitations in the data. In particular, an in-depth analysis of the design would suggest to examine additional healthy subjects as well as FIT negative subjects and to increase the record of FIT positive and CRC subjects. Although we are aware of the preliminary nature of the study, the reason we decided to apply our methodology to the CELTiC dataset is twofold: firstly, the dataset presents the characteristics of classifying the subjects in four categories; secondly, a recent parallel paper investigating the same data offers a way to compare our results to those obtained through the traditional dichotomous-forced approach.

Table 5.1 provides the descriptive statistics of the four biomarkers characterizing each group. In particular, for each biomarker, we report means, standard deviations and relative effects divided by class. As we can see, means and relative effects present the same trend going from Normal to HR/CCR group for all the markers with the exception of COL1A2. It is

important to note that, in line with standard practice in molecular genetics, the marker measures have been transformed such that they are inversely correlated to the amount of gene expression, thus high values indicate low levels of the relative gene.

In Figure 5.2, the relative values for the four groups of healthy control subjects (N), negative colonoscopy (NFIT), low risk lesion (LR), high risk lesion or colorectal cancers (HR/CRC) are reported for each marker CEACAM6, LGALS4, TSPAN8 and COL1A2.

A more detailed descriptive statistical analysis can be found in Rodia et al. (2018). Anyway, the simple descriptive investigation of the relative effects confirms that we should not consider the four categories as completely ordered.

5.1.2 Statistical analysis using the HUM

In the recent paper by Rodia et al. (2018), the authors propose to use a multinomial logistic regression model in order to study the association between outcome and a linear combination of the proposed markers; two-tailed p-values less than 0.05 were considered statistically significant; the reference group is N (healthy subjects). However, the authors force the statistical methodology and use dichotomous ROC curve and AUC analysis to assess the accuracy of the model in discriminating among the four groups of subjects. In this section, instead, we implement the methodology developed in Chapter 3, which is explicitly designed for classification problems characterized by four groups. Specifically, we aimed to compute the ability of the four biomarkers in discriminating subjects among the four groups. The accuracy summary measure we adopt, thus, is given by the HUM.

As proposed in Section 3.2 we estimate HUM_{L4} for each single marker using the relative effect values to establish the correct categories ordering. Furthermore, exploiting the results developed in Section 3.3, we also estimate the analytical asymptotic standard errors of the HUM_{L4}, which provide a measure of the sample uncertainty associated to the accuracy summary indicator.

The HUM_{L4}, moreover, is compared with the other two estimators already existing in the literature: the HUM_{LF} by Li and Fine (2008) described in Section 2.3.4 and the HUM_{EX} by Nakas and Yiannoutsos (2004) presented in Sections 2.3.1. For the latter, however, given that there are no analytical formula for calculating the standard errors, bootstrap techniques have been used. Finally, for the sake of completeness, bootstrap standard errors have been also calculated for the HUM_{L4} estimator. As suggested in Li and Fine (2008), bootstrap estimation of the standard errors for the three estimates

are calculated with $B = 100$ bootstrap re-samples.

Specifically, the HUM_{EX} has been estimated by the R-function *Calculate HUM-EX* included in the R-package *Biocomb* (Novoselova et al., 2017), which internally computes the maximal HUM value between all the possible permutations of class labels. We recall that HUM_{LF}, instead, is not affected by the ordering of the categories.

5.1.3 Results

Before presenting the results, we check for the Lehmann assumption through the graphical method and the statistical test introduced in Section 3.4. As already widely discussed, our estimator HUM_{L4} is based the Lehmann condition, and the result of the test together with analysis of the survival curves can provide a useful evaluation about the reliability of the obtained estimates. The plots of the survival curves for all the markers are reported in Figure 5.1. Even thought it is not easy to interpret the plots in four-class framework, when considering TSPAN8 and COL1A2, it emerges that the curves are overall parallel, although three of them are practically indistinguishable for a wide range of marker values. For the two remaining markers, instead, the parallelism is questionable, especially for LGALS4.

To deeply understand if the Lehmann condition holds, we present in Table 5.2 the p-values of the statistical test, as discussed in Section 3.4. As can be seen in the table, the null hypothesis of the proportional hazards assumption (compatible with the Lehmann condition), cannot be rejected for the TSPAN8 and COL1A2, while only at the 1% critical level for CEACAM6. It has to be rejected, instead, for the LGALS4 marker. Overall, however, the p-values remain relatively low, even when the null hypothesis cannot be rejected by the data.

The results of the estimation procedures are reported in Table 5.3. In particular, for each of the three estimators we show the point estimates and the bootstrap standard errors. Moreover, given the theoretical result shown in Section 3.3, for the HUM_{L4}, in brackets, we also report the asymptotic analytical standard errors. The procedure has been performed for each of the four blood markers, CEACAM6, LGALS4, TSPAN8 and COL1A2, and the results are reported row-by-row.

From Table 5.3, we can deduce five general results: a) for all the estimators and for all the markers, the HUM is rather low; b) the HUM_{EX} estimator always produces the highest values of the HUM; c) the HUM_{L4}, on the contrary, is the one giving the lowest values of the HUM; d) the worst results for the HUM_{L4} are those associated to LGALS4 and CEACAM6; e) the HUM_{L4} is largely the most efficient one, both in terms of analytical and

bootstrapped standard errors. A possible explanation for the poor performance of the HUM_{L4} in terms of the magnitude of the point estimate can be ascribed to the fact that the Lehmann condition is only marginally supported for the first two markers while has to be rejected for the last two, at least at the 5% critical level (see Table 5.2). Moreover, as we have stated in Section 3.5, our approach, through the analysis of the p-values of the Cox regression coefficients, offers detailed information about the discriminatory power of the marker for each single class. In Table 5.4, for each marker, we show the estimated coefficients of the Cox model with the associated p-values. As we can see, the intermediate classes, where the ordering is suggested by the relative effects discussed in Section 3.7.1, are practically indistinguishable for almost all the markers.

On the other side, regardless of the estimator used, the estimated HUM values confirm the LGALS4 biomarker as the most powerful blood marker discriminating among the four groups. Our result reinforces the one obtained in Rodia et al. (2018) with the rough pairwise ROC analysis. This marker is able to correctly classifying four subjects randomly chosen from the four groups with a probability that ranges between 0.129 of the HUM_{L4} estimator and 0.219 of the HUM_{EX} estimator. If we recall that the null value of HUM for a four-category classification problem is $1/4! = 0.042$, we can argue that the accuracy of this marker is sufficiently better than a random guess.

	N			N FIT			LR			HR-CCR		
	mean	sd	rel. effect	mean	sd	rel. effect	mean	sd	rel. effect	mean	sd	rel. effect
TSPAN8	11.330	1.718	0.675	9.997	1.199	0.456	9.924	1.420	0.456	9.558	1.851	0.412
COL1A2	11.449	1.920	0.679	9.674	1.286	0.427	9.674	1.373	0.422	9.608	1.973	0.434
LGALS4	12.893	1.971	0.270	15.662	1.300	0.703	15.284	0.775	0.661	14.693	1.275	0.530
CEACAM6	12.343	1.893	0.364	14.249	1.096	0.699	13.589	1.206	0.561	13.346	1.247	0.503

Table 5.1: Means, standard deviations and relative effects of biomarkers by class.

marker	p-value
TSPAN8	0.185
COL1A2	0.139
LGALS4	< 0.01
CEACAM6	0.011

Table 5.2: Test for proportional hazards assumption.

	HUM_{L4}		HUM_{EX}		HUM_{LF}	
	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se
TSPAN8	0.087	0.012 (0.013)	0.110	0.019	0.108	0.029
COL1A2[a]	0.102	0.015 (0.015)	0.111	0.020	0.111	0.035
LGALS4	0.129	0.021 (0.020)	0.219	0.034	0.151	0.041
CEACAM6	0.096	0.016 (0.015)	0.144	0.024	0.139	0.029

Table 5.3: Estimated HUMs by marker.
[a] The HUM_{EX} estimator suggests a different categories order (HR/CCR, LR, NFIT, N); however, if we impose this order in the HUM_{L4} estimator, the estimated HUM decreases to 0.081.

5.2 Tissue biomarkers of synovitis

In this section we provide an application of the HUM_{L4} estimator to a real dataset obtained from a medical study presented by Della Beffa et al. (2013). As before, we also give a comparison between the HUM_{L4} estimator and the two non parametric estimators we have seen before: the one proposed by Li and Fine (2008) (HUM_{LF}) and the one based on the idea of Nakas and Yiannoutsos (2004) and implemented by Novoselova et al. (2014) (HUM_{EX}).

The dataset we analyse comes from a study in which the number of cells expressing a certain marker is used in immunohistological synovial tissue classification. It is quite common in such a literature to employ absolute densities (e.g. number of cells/mm^2 of subintimal tissue) of specific inflammatory cell types to classify synovial tissue samples for diagnostic or prognostic purposes. Della Beffa et al. (2013) refine this way of proceeding by looking at qualitative features of inflammatory cell populations as a possible further source of information. In this line, the authors consider subjects belonging to six different disease groups and, in each subject, five major inflammatory cell

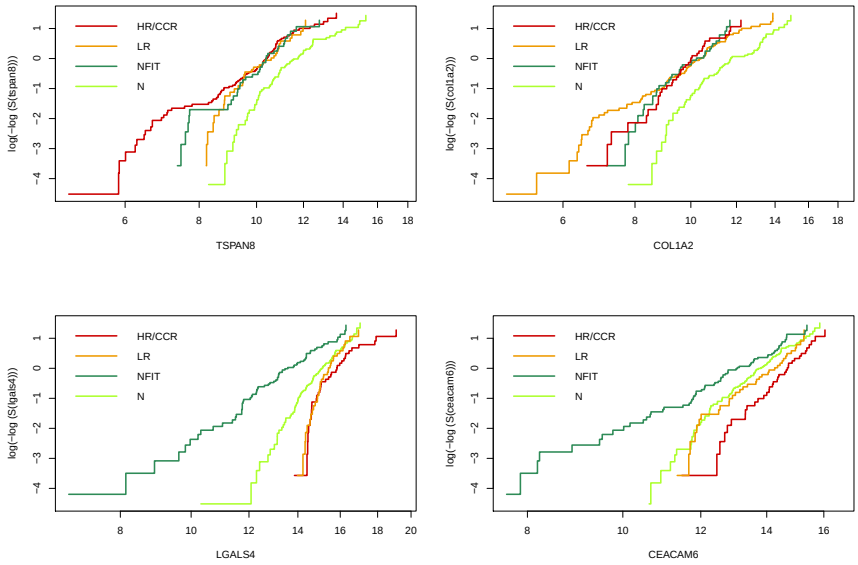


Figure 5.1: Log(-log(survival))curves as function of marker value for TSPAN8, COL1A2, LGALS4 and CEACAM6.

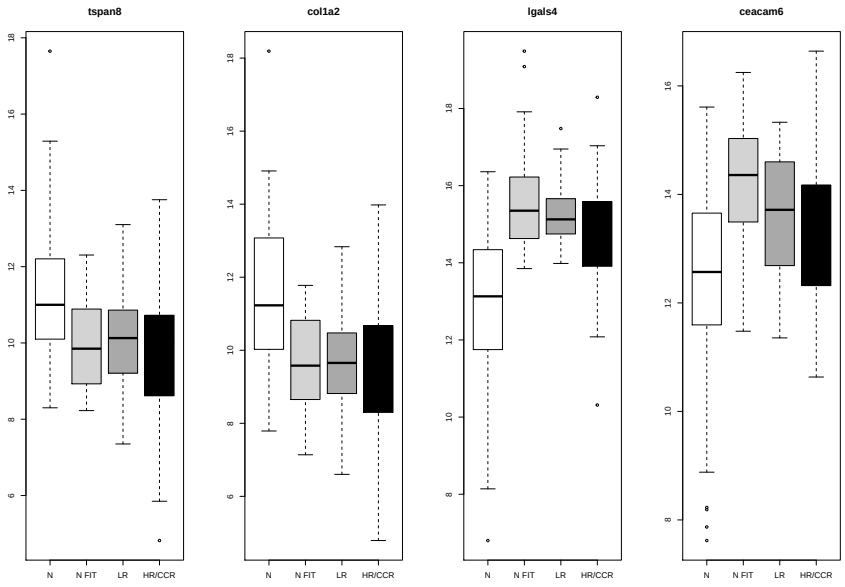


Figure 5.2: Box plot of the four blood markers for colorectal cancer detection in the four groups.

TSPAN8					
	β	$\theta = \exp(\beta)$	se	z	p-value
LR	-0.065	0.937	0.197	-0.330	0.740
N FIT	0.014	1.014	0.236	0.060	0.950
N	-0.839	0.432	0.216	-3.890	< 0.001
COL1A2					
	β	$\theta = \exp(\beta)$	se	z	p-value
N FIT	-0.006	0.994	0.236	-0.030	0.980
HR/CCR	-0.207	0.813	0.199	-1.040	0.300
N	-0.836	0.433	0.168	-4.960	< 0.001
LGALS4					
	β	$\theta = \exp(\beta)$	se	z	p-value
HR/CCR	-0.918	0.399	0.164	-5.590	< 0.001
LR	-0.228	0.796	0.197	-1.160	0.250
N FIT	-0.356	0.700	0.243	-1.460	0.140
CEACAM6					
	β	$\theta = \exp(\beta)$	se	z	p-value
HR/CCR	-0.446	0.641	0.162	-2.750	0.006
LR	-0.099	0.906	0.199	-0.500	0.620
N FIT	-0.465	0.628	0.239	-1.950	0.052

Table 5.4: Estimated coefficients for the Cox proportional hazards regression model. Note that, for each marker, the ordering is the one established by the relative effects and the omitted class is the reference one.

types values (markers) are observed. The dataset utilized is available in the `R` package called `HUM`.

The sample is composed by $N = 92$ patients with chronic septic arthritis (SeA, n=11), rheumatoid arthritis (RA, n=25), early undifferentiated arthritis (EA, n=10), osteoarthritis (OA, n=26), "non-inflammatory" orthopedic arthropathies (OrthArthr, n=6) and healthy volunteers (Normal, n=15). For simplicity, we focus on the analysis of two synovial tissue biomarkers only, *neutrophilic granulocytes* called CD15 and T cells called CD3. The variables are measured as the absolute cell densities expressed as the number of positive staining cells per mm^2. The aim of our analysis is twofold: firstly, we want to assess the diagnostic accuracy of the biomarkers in classifying a subject in one of the six categories using the definition of HUM derived in eq. (3.12) as measure of accuracy; secondly, we aim to analyse the performance of HUM_{L4} estimator making a comparison with the HUM_{EX} and the

HUM_{LF} estimators. The standard errors are computed using a bootstrap procedure. Concerning the HUM_{L4} estimator, standard errors obtained analytically through the formula presented in eq. (3.29) are also provided.

As the focus of this work is on four-class classification issues, we restrict the analysis to four classes considering all the possible sets of four groups containing at least the "Normal" and the "SeA" patients.

Marker CD15

We first present some descriptive statistics and the box plot of the marker's distribution in the four groups. Means and standard deviations for each category, together with the relative effects for each of the six possible sets of four categories, are reported in Table 5.5, while the box plots are shown in Figure 5.2.

	mean	sd	relative effect					
			(1)	(2)	(3)	(4)	(5)	(6)
Normal	0.057	0.112	0.142	0.203	0.160	0.214	0.175	0.133
OrthArthr	0.330	0.380		0.431	0.295	0.405		
OA	0.536	0.594	0.365	0.531			0.453	
Early	4.223	3.336				0.621	0.702	0.495
RA	6.713	7.013	0.702		0.611			0.576
SeA	37.664	21.747	0.910	0.914	0.878	0.874	0.915	0.882

Table 5.5: Descriptive statistics of marker CD15 by the six groups.

In the first two columns of Table 5.5 we report the mean and the standard deviation for each of the six groups. Furthermore, column (1) reports the relative effects of marker CD15 in subset 1 (Normal, OA, RA and SeA), column (2) reports the relative effects of marker CD15 in subset 2 (Normal, OrthArthr, OA and SeA), column (3) reports the relative effects of marker CD15 in subset 3 (Normal, OrthArthr, RA and SeA), column (4) reports the relative effects of marker CD15 in subset 4 (Normal, OrthArthr, Early and SeA), column (5) reports the relative effects of marker CD15 in subset 5 (Normal, OA, Early and SeA) and column (6) reports the relative effects of marker CD15 in subset 6 (Normal, Early, RA and SeA). They are used to determine the ordering of the categories that allows to obtain the maximum value of the estimated hypervolume.

The following step consists in evaluating the Lehmann assumption using the test for the proportional hazards assumption proposed by Grambsch and

Therneau (1994) together with the plots of the log-minus-log transformation of the survival curve. In Table 5.6 we show the significance levels of the test. As the test starts with the null hypothesis of validity of the proportional hazards assumption (i.e. the Lehmann condition does hold), from the table we can see that the test is not significant for all the subsets, indicating that our HUM_{L4} estimator should work well in each subset of patients. The inspection of the curves, see Figure 5.4, strongly confirms the indication given by the statistical test.

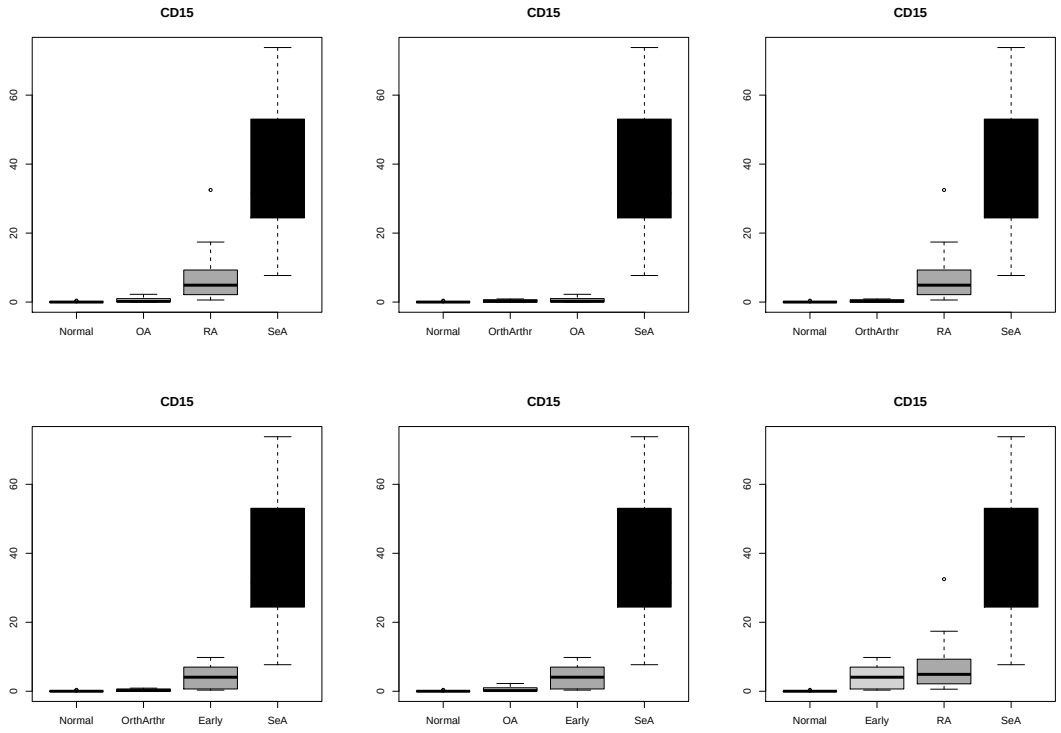


Figure 5.3: Box plots comparing absolute cell densities of marker CD15 in the 6 subsets of patient's groups

Finally, we apply the HUM_{L4} estimator developed in Section 3.2 to the data in order to evaluate the discrimination capability of CD15 marker. The results for the estimated HUMs are summarized in Table 5.7. The performance of the HUM_{L4} is compared with the two other estimators, HUM_{EX} and HUM_{LF}, described before. For each estimator we report the estimated HUM and the associated standard errors obtained with bootstrap techniques. Moreover, for the HUM_{L4} estimator, we show the analytic asymptotic stand-

Group	p-value
Normal, OA, RA, SeA	0.945
Normal, OrthArthr, OA, SeA	0.980
Normal, OrthArthr, RA, SeA	0.949
Normal, OrthArthr, Early, SeA	0.974
Normal, OA, Early, SeA	0.997
Normal, Early,RA, SeA	0.921

Table 5.6: Test for proportional hazards assumption in the six subgroups.

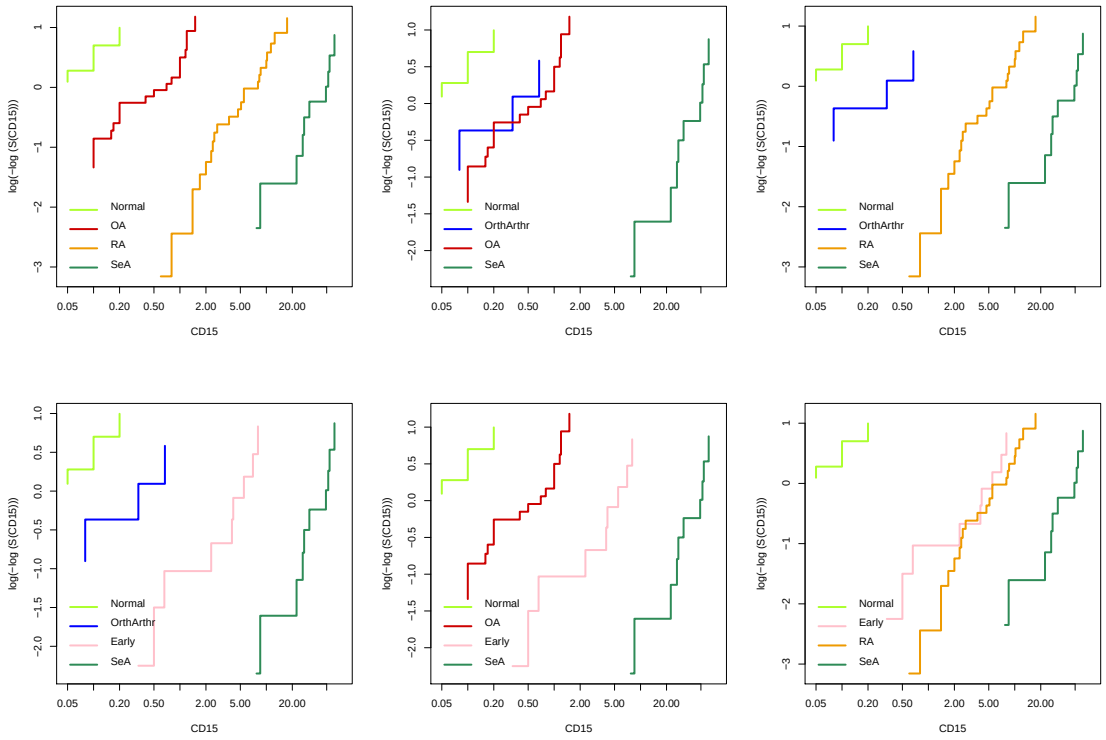


Figure 5.4: Log(-log(survival)) curves as function of CD15 marker value for the six different subsets.

ard errors developed in Section 3.3. Apart for the second combination (Normal, OrthArthr, OA, SeA), the estimated values of the HUM is relatively high. The three different estimators provide very similar values of the HUM. However, our estimator, based on the Lehmann assumption, systematically produces smaller standard errors (both analytical and bootstrapped) compared to the other non parametric estimators. In Table 5.8, we also report the summaries of the estimated Cox proportional hazards models to show how informative our method can be if a researcher wants to have a deeper insight on the differences among the classes in each marker. The p-values reported in the last column of the table confirm that the distribution of CD15 is statistically different among the classes in all the subgroups except in the second, when the classes "OA"and "SeA"are compared with the class "OrthArthr", and marginally in the sixth, when the class "RA"is compared with the class "Early".

Group	HUM_{L4}		HUM_{EX}		HUM_{LF}	
	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se
Normal, OA, RA, SeA	0.657	0.068 (0.069)	0.616	0.082	0.688	0.087
Normal, OrthArthr, OA, SeA	0.388	0.075 (0.068)	0.281	0.104	0.503	0.133
Normal, OrthArthr, RA, SeA	0.650	0.119 (0.097)	0.544	0.162	0.708	0.101
Normal, OrthArthr, Early, SeA	0.621	0.086 (0.097)	0.479	0.150	0.605	0.124
Normal, OA, Early, SeA	0.669	0.076 (0.077)	0.564	0.095	0.605	0.129
Normal, Early,RA, SeA	0.526	0.077 (0.087)	0.528	0.082	0.463	0.120

Table 5.7: Estimated HUMs with marker CD15 by subgroup.

Marker CD3

The same kind of analysis is performed for the marker CD3, too. This marker presents higher means and standard deviations in all groups compared with marker CD15. Descriptive statistics and box plots are shown in Table 5.9 and Figure 5.5, respectively. Moreover, as reported in Table 5.10, the proportional hazards condition strongly holds for all the subsets of subjects (the test presents extremely large p-values, denoting that it is never significant). Further evidence in favor of the Lehmann assumption derive by the graphical inspection of the curves reported in Figure 5.6. Thus, as before, we expect our HUM_{L4} estimator to perform rather well, being the

underlying assumption satisfied.

In Table 5.11 we report the estimated HUMs using the three estimators HUM_{L4}, HUM_{EX} and HUM_{LF}. Overall, the estimated values are lower than those produced by marker CD15, indicating the that marker CD3 is substantially less powerful in discriminating between the four categories of patients. Table 5.12 corroborates those results, showing that, in some subgroups, CD3 has not a significantly different distribution among classes. Such as, for example, in the second subgroup, where the coefficient of class "OA"is not statistically different from zero, indicating that the distribution in class "OA"is substantially the same as the one in class "OrthArthr". Concerning the three different estimators, the results are pretty in line and it is impossible to highlight one estimator systematically performing better than the others. However, HUM_{LF} always presents higher standard errors than the other two estimators.

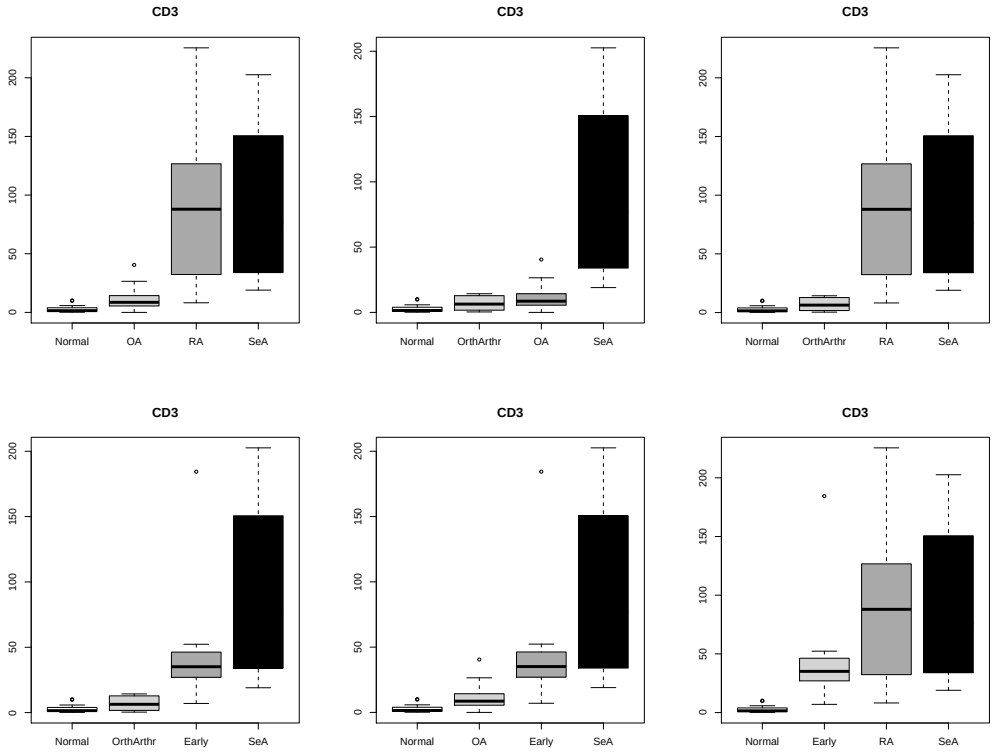


Figure 5.5: Box plots comparing absolute cell densities of marker CD3 in the 6 subsets of patient's groups.

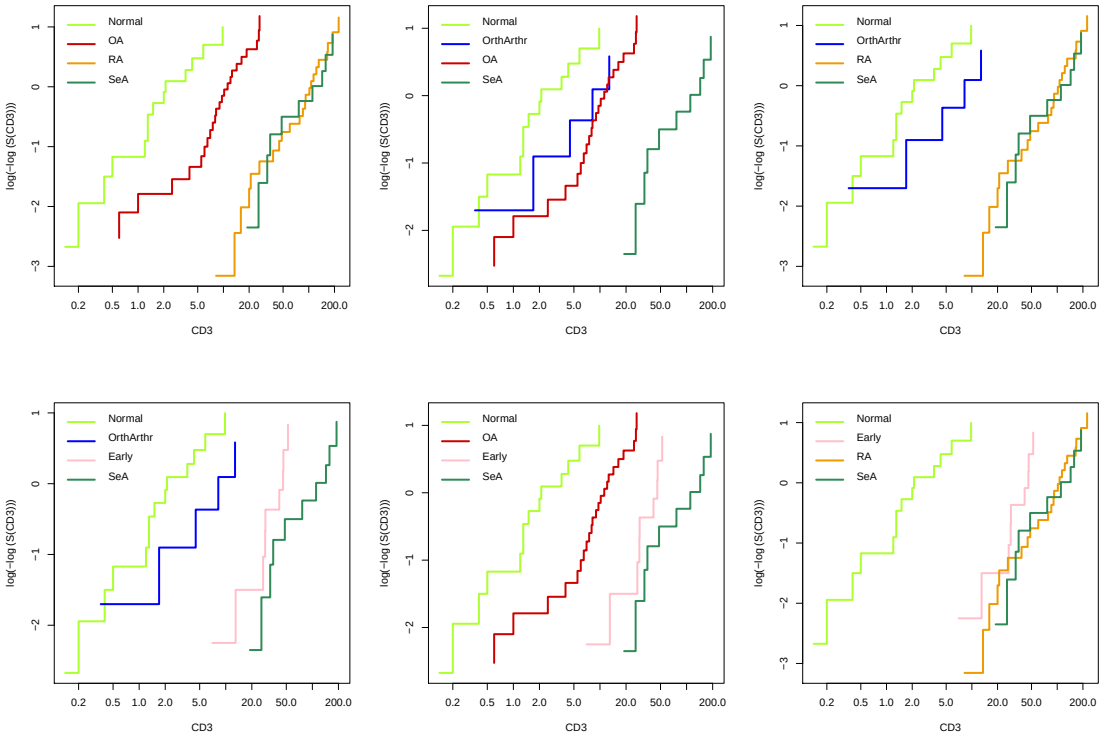


Figure 5.6: Log(-log(survival)) curves as function of CD3 marker value for the six different subsets.

Normal, OA, RA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OA	-1.477	0.228	0.388	-3.810	< 0.001
RA	-2.828	0.059	0.494	-5.730	< 0.001
SeA	-1.921	0.147	0.495	-3.880	< 0.001
Normal, OrthArthr, OA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OrthArthr	-0.877	0.416	0.502	-1.750	0.081
OA	-0.578	0.561	0.473	-1.220	0.222
SeA	-20.400	0.000	4020.000	-0.010	0.996
Normal, OrthArthr, RA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OrthArthr	-1.237	0.290	0.600	-2.060	0.039
RA	-3.572	0.028	0.839	-4.260	< 0.001
SeA	-1.921	0.147	0.495	-3.880	< 0.001
Normal, OrthArthr, Early, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OrthArthr	-1.219	0.296	0.577	-2.110	0.035
Early	-2.032	0.131	0.718	-2.830	0.005
SeA	-2.842	0.058	0.805	-3.530	< 0.001
Normal, OA, Early, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OA	-1.423	0.241	0.382	-3.720	< 0.001
Early	-2.224	0.108	0.638	-3.490	< 0.001
SeA	-2.852	0.058	0.804	-3.550	< 0.001
Normal, Early,RA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
Early	-4.500	0.011	1.092	-4.120	< 0.001
RA	-0.592	0.553	0.396	-1.500	0.130
SeA	-1.942	0.144	0.493	-3.940	< 0.001

Table 5.8: Estimated coefficients for the Cox proportional hazards regression model for marker CD15. Note that the order of the groups is the one established by the relative effects and the missing class is the reference one.

	mean	sd	relative effect					
			1	2	3	4	5	6
Normal	2.956	3.274	0.177	0.259	0.174	0.232	0.217	0.137
OrthArthr	6.977	5.776		0.425	0.265	0.357		
OA	11.096	9.641	0.357	0.506			0.431	
Early	47.460	50.299				0.693	0.756	0.512
RA	91.198	66.185	0.753		0.690			0.662
SeA	94.691	69.092	0.770	0.900	0.703	0.812	0.859	0.674

Table 5.9: Descriptive statistics of marker CD3 by the six possible groups of four categories.

Group	p-value
Normal, OA, RA, SeA	0.918
Normal, OrthArthr, OA, SeA	0.863
Normal, OrthArthr, RA, SeA	0.996
Normal, OrthArthr, Early, SeA	0.997
Normal, OA, Early, SeA	0.922
Normal, Early,RA, SeA	0.984

Table 5.10: Test for the proportional hazards assumption for marker CD3.

Group	HUM_{L4}		HUM_{EX}		HUM_{LF}	
	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se
Normal, OA, RA, SeA[a]	0.335	0.076 (0.074)	0.358	0.063	0.377	0.078
Normal, OrthArthr, OA, SeA	0.334	0.062 (0.064)	0.331	0.074	0.251	0.097
Normal, OrthArthr, RA, SeA[b]	0.347	0.088 (0.083)	0.350	0.085	0.229	0.121
Normal, OrthArthr, Early, SeA	0.463	0.105 (0.097)	0.448	0.105	0.406	0.119
Normal, OA, Early, SeA	0.434	0.091 (0.081)	0.463	0.090	0.423	0.104
Normal, Early,RA, SeA	0.250	0.079 (0.071)	0.268	0.067	0.040	0.120

Table 5.11: Estimated HUMs with marker CD3 by subgroup.
[a] The HUM_{EX} estimator suggests a different categories order (Normal, OA, SeA, RA) if we impose this order in the HUM_{L4} estimator the estimated HUM remains practically the same (0.336).
[b] The HUM_{EX} estimator suggests a different categories ordering (Normal, OrthArthr, SeA, RA), if we impose this last ordering in the HUM_{L4} estimator the estimated HUM slightly decreases to 0.346.

Normal, OA, RA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OA	-1.443	0.236	0.378	-3.820	< 0.001
RA	-2.478	0.084	0.409	-6.050	< 0.001
SeA	0.002	1.002	0.371	0.010	0.996
Normal, OrthArthr, OA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OrthArthr	-0.914	0.401	0.504	-1.810	0.070
OA	-0.510	0.600	0.469	-1.090	0.280
SeA	-2.419	0.089	0.570	-4.250	< 0.001
Normal, OrthArthr, RA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OrthArthr	-1.001	0.367	0.550	-1.820	0.069
RA	-3.831	0.022	0.850	-4.510	< 0.001
SeA	-0.002	0.998	0.371	0.000	0.997
Normal, OrthArthr, Early, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OrthArthr	-1.016	0.362	0.551	-1.840	0.065
Early	-2.935	0.053	0.856	-3.430	0.001
SeA	-0.812	0.444	0.474	-1.710	0.087
Normal, OA, Early, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
OA	-1.416	0.243	0.377	-3.750	< 0.001
Early	-1.881	0.152	0.460	-4.090	< 0.001
SeA	-0.788	0.455	0.475	-1.660	0.097
Normal, Early,RA, SeA					
	β	$\theta = \exp(\beta)$	se	z	p-value
Early	-3.780	0.023	0.817	-4.630	< 0.001
RA	-0.790	0.454	0.395	-2.000	0.045
SeA	0.005	1.005	0.371	0.010	0.990

Table 5.12: Estimated coefficients for the Cox proportional hazards regression model for marker CD3 in the synovitis dataset. Note that, for each marker, the ordering is the one established by the relative effects and the missing class is the reference one.

Chapter 6

Combining multiple markers

6.1 Introduction

Often, in medical research, multiple diagnostic markers are measured on the same individual to assess an optimal result for prognosis as it is well established that, in most cases, one single biomarker is not sufficient to perform a screening for early detection or for a correct prognosis. The statistical analysis, however, continues to focus mainly on one single marker at the time. A natural theoretical development, thus, is to try to combine information from different biomarkers in order to maximize the discrimination of patients belonging to different classes of disease. Recently, especially in medical research, different methods for combining biomarkers in the best fashion have been proposed; starting from a binary diagnostic category setting, hence maximizing the AUC, up to approach aiming to maximize the HUM in a multiple categories outcome setting.

In this section we focus on a new proposal for finding the optimal linear combination of continuous markers. The intuition is based on the idea of searching for the linear combination of markers that achieves the maximum accuracy over all possible linear combinations. Su and Liu (1993) provide a way to estimate the coefficients of the best linear combination of markers that maximizes the AUC. The authors derived the results under the assumption of multivariate Gaussian distribution of markers with both proportional and non proportional covariance matrices. Pepe (2000) proposes two different approaches in finding the linear combination that maximizes the AUC without any restriction on the markers distribution. Moreover, Pepe et al. (2006) provide a comparison between the method based on maximization of empirical AUC and the multinomial logistic approach. As we have seen before in this thesis, the method introduced by Li and Fine (2008) allows to

deal with multiple markers and multi-category outcome through a multinomial logistic model. However, as Zhang and Li (2011) said in their paper "it is not clear if the method yields the best combination to maximize VUS or HUM" because in their approach the objective function to be maximized is not the VUS/HUM function. More recently Zhang and Li (2011) overcame the less general assumption of binary outcome providing an approach that is valid in a medical classification problem with more than two potential outcomes. Unfortunately, we are not able to compare our approach with the one proposed by Zhang and Li (2011) since no code or packages to implement the estimator are available at the moment.

6.2 A new proposal

In this section we propose a new semi-parametric approach based on the Lehmann assumptions to linearly combine multiple biomarkers with the purpose of maximizing the most important diagnostic accuracy index for a four-category outcome, the HUM. Our proposal is to use numerical methods to estimate the coefficients of the optimal linear combination of markers with HUM_{L4} as objective function.

Consider the simplest situation in which we have four classes of patients, as, for example, in a case-control study on carcinoma as presented in Rodia et al. (2018). Let D_1, D_2, D_3 and D_4 denote the four classes, that could be the codification for "Normal", "Very low risk", "Low risk", "High risk or carcinoma". Moreover, suppose that two biomarkers, M_1, M_2, are measured in all subjects. Our purpose is to find the vector of coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$ of the linear combination of markers that maximizes the Hypervolume Under the Manifold defined in eq.(3.13), i.e.

$$HUM_{L4} = \frac{1}{(\theta_3+1)(\theta_2(\theta_3+1)+1)(\theta_1(\theta_2(\theta_3+1)+1)+1)}.$$

More specifically, define a linear combination of the markers as follows

$$\begin{aligned} LC_\alpha &= M_1\alpha_1 + M_2\alpha_2 \\ &= \boldsymbol{M}\ \boldsymbol{\alpha} \end{aligned}$$

where the vector $\boldsymbol{\alpha} = (\alpha_1\,,\, \alpha_2)^T$ contains the loadings of the linear combination, while $\boldsymbol{M} = (M_1\,,\, M_1)$ contains the two markers. However, for sake of identification, we fix one of the two coefficients of the linear combination as equal to one, say $\beta_1 = 1$ and rescale the second coefficient as $\beta = \alpha_2/\alpha_1$.

Thus, the new vector of coefficients can be defined as $\boldsymbol{\beta} = (1\,,\,\beta)^T$. In this way, the linear combination of markers becomes

$$\begin{aligned} LC_\beta &= M_1 + M_2\beta \\ &= \boldsymbol{M}\,\boldsymbol{\beta}. \end{aligned} \tag{6.1}$$

The generalization to more than two markers is straightforward. In fact, in the more general case of q markers $M_1, \ldots, M_q$ providing information on a specific four-class classification issue, the linear combination can be written as

$$\begin{aligned} LC_\beta &= M_1 + M_2\beta_2 + \ldots + M_q\beta_q \\ &= \boldsymbol{M}\,\boldsymbol{\beta}, \end{aligned} \tag{6.2}$$

where, as before, β_1 is normalized to one, and the vector containing the scores of the linear combination is $\boldsymbol{\beta} = (1\,,\,\beta_2\,,\ldots\,,\,\beta_q)^T$, while $\boldsymbol{M} = (M_1\,,\,\ldots\,,\,M_q)$.

The quantity LC_β, that clearly depends on the unknown parameters βs, is our *latent* marker that should have the best diagnostic power in classifying the subjects to the four classes characterizing our population. As largely discussed in previous sections, for a four-class classification problem, the best diagnostic power can be calculated in terms of the HUM. Our optimization problem, thus, can be written as

$$\max_{\boldsymbol{\beta}} \left(HUM(LC_\beta) \right) \tag{6.3}$$

where $HUM(LC_\beta)$ indicates the HUM calculated on the LC_β (latent) marker.

Clearly, any possible estimator for the HUM can be used in the maximization problem. However, given the results of the simulation analysis performed in Chapter 4, the HUM_{L4} estimator developed in Section 3.2, might be a very good candidate. In fact, in order to find the optimal value of $\boldsymbol{\beta}$ maximizing the quantity in eq. (6.3) we propose to use numerical methods based on grid search on a range of reasonable values for $\boldsymbol{\beta}$. In this respect, having the possibility to deal with a well performing and fast estimator for the HUM can be extremely convenient.

Our proposal, thus, is to focus on the HUM_{L4} estimator in the maximization problem in eq. (6.3). An estimator for the optimal linear combination in term of the HUM, hence, becomes

$$\hat{\boldsymbol{\beta}}_{L4} = \max_{\boldsymbol{\beta}} \left(\widehat{HUM}_{L4}(LC_\beta) \right) \tag{6.4}$$

$$= \max_{\boldsymbol{\beta}} \left(\widehat{HUM}_{L4}(\boldsymbol{M}\,\boldsymbol{\beta}) \right) \tag{6.5}$$

where eq. (6.5) explicitly emphasizes the argument of the function to be maximized.

Importantly, as we have discussed in the previous chapters, the HUM_{L4} depends on the ordering of the classes, that, however, are generally unknown in advance. Thus, our implemented method, while searching for the maximum HUM, must take into account this issue. Before proceeding with the calculation of the HUM, our method calculates the relative effects (see Section 3.7) of each marker and sorts the markers according to the order suggested by the relative effects.

In the following sections, the performances of the proposed method are investigated through and extensive simulation study and two applications to real data. Specifically, in Section 6.3 we simulate data from a simple set up for which the optimal linear combination of markers can be known a priori. In Section 6.4 and Section 6.5, instead, we implement our procedure using real datasets: we first investigate the data collected in the colorectal cancer study explained in Section 5.1, then the same approach is applied to the Synovitis data described in Section 5.2.

6.3 Simulation study

In this section we evaluate the performances of the proposed estimator of the linear combination maximizing the HUM. We consider a simulation study in which, for each individual, two markers are observed. Moreover, suppose the population be divided into four classes according to different levels of the disease.

6.3.1 Assumptions on the Data Generating Process

In this simulation exercise, for the reason that will be more clear below, we assume a specific parametric case in which the random vectors of the markers in the four classes follow a Gaussian distribution. Being $\boldsymbol{M}_1 = (M_{11}\,,\, M_{21})$ the markers vector in group D_1, $\boldsymbol{M}_2 = (M_{12}\,,\, M_{22})$ the markers vector in group D_2, $\boldsymbol{M}_3 = (M_{13}\,,\, M_{23})$ the markers vector in group D_3 and $\boldsymbol{M}_4 = (M_{14}\,,\, M_{24})$ the markers vector in group D_4, we thus assume:

$$\boldsymbol{M}_1^T \sim N(\boldsymbol{\mu}_1; \boldsymbol{\Sigma});\;\; \boldsymbol{M}_2^T \sim N(\boldsymbol{\mu}_2; \boldsymbol{\Sigma});\;\; \boldsymbol{M}_3^T \sim N(\boldsymbol{\mu}_3; \boldsymbol{\Sigma});\;\; \boldsymbol{M}_4^T \sim N(\boldsymbol{\mu}_4; \boldsymbol{\Sigma})$$

with:

$$\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_3 = \boldsymbol{\mu}_3 - \boldsymbol{\mu}_4.$$

and

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4 = \mathbf{I}_2.$$

Under these assumptions, the exact order of the markers can be determined in advance and it is equal to the order of the groups mean. Moreover, the differences in the vector of the means for subsequent groups remain constant and equal to the vector $\boldsymbol{\delta}$. Finally, we assume that the variance-covariance matrices are all equivalent and equal to the identity matrix. Within and across groups, thus, the markers are not correlated.

These set of assumptions, although rather restrictive and relatively unlikely in practical experiments, are extremely useful from a theoretical point of view. In fact, according to Su and Liu (1993) and Zhang and Li (2011), under these assumptions, the coefficients of the best linear combination $\boldsymbol{\beta}_0$ have the characteristic that $\boldsymbol{\beta}_0 \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. Moreover, for any given value of $\boldsymbol{\beta}$, the distribution of the linear combination is also known. More precisely, if we define the linear combination in each group as $LC_i = \boldsymbol{M}_i\boldsymbol{\beta}$, with $i = 1, 2, 3, 4$, for any vector of coefficients $\boldsymbol{\beta}$, then $LC_i \sim N(\boldsymbol{\mu}_i^T\boldsymbol{\beta}\,;\, \boldsymbol{\beta}^T\boldsymbol{\Sigma}_i\boldsymbol{\beta})$.

Importantly, as we assume Gaussian distributions of the markers, the Lehmann condition is clearly not satisfied, invalidating the theoretical background of the HUM_{L4} estimator. However, based on the reassuring results of the simulation exercises presented in Section 4.3.2, we expect limited consequences on our estimator of the optimal linear combination coefficients.

6.3.2 Data Generating Process and results

Given the distributional assumptions described in the previous section, we assume the following values for the parameters:

$$\boldsymbol{\mu}_1 = (3; 0)^T\,;\;\; \boldsymbol{\mu}_2 = (5; 1)^T\,;\;\; \boldsymbol{\mu}_3 = (7; 2)^T\,;\;\; \boldsymbol{\mu}_4 = (9; 3)^T\,.$$

The distance between the barycentre of adjacent distributions, as discussed before, is constant and equal to $\boldsymbol{\delta} = (2\,,\,1)^T$. We hypothesize three different sample sizes in line with standard empirical applications: $N = 120$, $N = 200$ and $N = 320$, respectively. Moreover, we impose the groups to have the same dimension.

Based on these simulated data, we apply the method developed in Section 6.2 to obtain the optimal linear combination which maximizes the HUM. For each of the three dimensions, $k = 500$ Monte Carlo samples are generated. Importantly, based on the set of assumptions discussed in the previous section, and normalizing for the first coefficient, the true value of the coefficients of the optimal linear combination is $\boldsymbol{\beta}_0 = (1\,,\,\beta)^T = (1\,,\,0.5)^T$. Furthermore, the real HUM, obtained by numerical integration, is 0.833. For each draw, thus, we calculate the optimal value of $\boldsymbol{\beta}$ through the $\hat{\boldsymbol{\beta}}_{L4}$ estimator defined in eq.s (6.4)-(6.5).

The results of this simulation exercise are reported in Table 6.1. In the table we report the true value of the coefficient β, the average of the estimated coefficients $mean(\hat{\beta})$ and standard deviation of the estimated coefficients $sd(\hat{\beta})$. As we can see from the results, the distribution of the estimated coefficients is centred very close to the real value of the parameter and we can notice that the standard deviations decrease as the sample size increases. We can conclude that, even if the proportional hazards model does not hold, our method works in a very good way.

Finally, in order to evaluate the improvement in the discrimination accuracy due to the combined information of the two markers, we calculate the estimated HUM for one single marker at the time and we make a comparison with the HUM obtained with the linear combination of the two markers. The results are reported in Table 6.2. For each sample size, the sample mean of the estimated HUM are reported. In columns M_1 and M_2 we show the hypervolumes obtained with the single markers while in column $M_1 + \hat{\beta}M_2$ we report the hypervolume obtained with the linear combination of the two markers. The first marker is clearly more informative than the second in discriminating the four classes. In fact, it correctly classifies around the 70% of the subjects while the second only slightly more that 33%, nevertheless combining the two markers we can improve the classification accuracy obtaining a HUM value of almost 0.8 for all the sample dimensions we considered. Although our estimator underestimates the true value of the HUM, which is 0.833, it works pretty well in increasing the classification accuracy with respect to single markers.

sample size	β	mean($\hat{\beta}$)	sd($\hat{\beta}$)
120	0.50	0.509	0.231
200	0.50	0.504	0.182
320	0.50	0.506	0.134

Table 6.1: Simulation results for multivariate normal distributions with constant difference of expected values vector among groups and constant variance-covariance matrix equal to the identity matrix.

6.4 Empirical application on CRC data

In this section we apply our methodology to a set of markers studied in Rodia et al. (2018) and widely analysed in Section 5.1. Data refer to a sample of 231 subjects investigated with the aim of early detecting the presence of colorectal

	HUM_{L4}		
	M_1	M_2	$M_1 + \hat{\beta}M_2$
sample size	$\widehat{\text{hum}}$	$\widehat{\text{hum}}$	$\widehat{\text{hum}}$
120	0.705	0.335	0.786
200	0.704	0.337	0.781
320	0.703	0.338	0.777

Table 6.2: Estimated hypervolumes with single markers and optimal linear combination of makers.

carcinoma. Among the entire sample, 67 individuals are classified as normal subjects (N), 36 present positive faecal immunochemical test (NFIT) with a negative colonoscopy, 36 present positive faecal immunochemical test and small, low risk polyps (LR), while 92 are positive to faecal immunochemical test and have high risk polyps or carcinoma (HR/CCR). We consider, thus, the population of subjects to be divided in four classes. Moreover, for each subject, four markers have been measured in the blood sample. Differently with respect to the simulation exercise, the number of markers, thus, is $q = 4$.

The statistical analysis consists, first, in estimating the HUM associated with each single marker and, second, in estimating the HUM associated with the best linear combination using the $\hat{\boldsymbol{\beta}}_{L4}$ estimator. Concerning the first step, it is the one already performed in Section 5.1, whose results are shown in Table 5.3. Among the four markers, the highest HUM is the one associated with the lectin, galactoside binding soluble 4 marker (lgals4), obtained through the HUM_{EX} estimator (HUM=0.219). For all the other markers, instead, we obtained values of the HUM slightly larger than 0.1. In order to exploit all the variability contained in the four markers and increase their discriminatory ability on future observations we adopt the proposed method based on the linear combination of markers maximizing the HUM.

We perform, thus, the $\hat{\boldsymbol{\beta}}_{L4}$ estimator and search for the optimal linear combination considering 61 values ranging from -3 to 3 for each β coefficient. As for the simulation exercise, for identification purposes, we normalize for the first coefficient. For seek of comparison, we also performed the HUM estimation by means of the HUM_{LF} estimator that, as we highlighted in Section 3.6, allows to compute the HUM taking jointly into account multiple markers, too.

In Table 6.3 the estimated coefficients obtained with the proposed method are reported together with bootstrap standard deviations. Coefficient β_2 refers to the col1a2 marker, β_3 to the lgals4 and β_4 to the ceacam6 marker. As expected, the highest coefficient is the one associated to the lgals4 marker.

Moreover, the hypervolume corresponding to the combined marker is listed in column $\widehat{hum}$. As we can see, using the four markers panel the measure of the discrimination accuracy doubles in comparison to the one obtained with each single marker (see the results reported in Table 5.3). Finally, for seek of comparison, the HUM estimated with HUM_{LF} is reported in the last column. We recall that the HUM_{LF} approach, while estimating the joint HUM, does not search for an optimal combination of markers. Thus, the comparison necessarily remains limited to the HUM values, and not on the coefficients of the linear combination, that cannot be obtained from their approach. However, interestingly, we notice that the magnitude of the hypervolumes obtained with the two approaches is essentially the same.

		HUM_{L4}					HUM_{LF}
$\hat{\beta}_2$	se	$\hat{\beta}_3$	se	$\hat{\beta}_4$	se	$\widehat{\text{hum}}$	$\widehat{\text{hum}}$
-1.818	0.059	2.364	0.063	1.091	0.057	0.266	0.295

Table 6.3: Estimated coefficients of the optimal linear combination together with their bootstrap standard errors ($k = 500$ replications), estimated hypervolume for the combined marker. For identification issues, we fix $\beta_1 = 1$. The last column reports the hypervolume value relative to the four markers, resulting from the Li and Fine (2008) approach.

6.5 Empirical application on Synovitis data

In this second empirical application, we employ our methodology to the dataset presented in Section 5.2, concerning synovial tissue biomarkers classification. In particular, the dataset consists in five inflammatory cells markers observed in a sample of 92 patients. The patients are classified in six different disease groups: chronic septic arthritis (SeA, n=11), rheumatoid arthritis (RA, n=25), early undifferentiated arthritis (EA, n=10), osteoarthritis (OA, n=26), "non-inflammatory" orthopedic arthropathies (OrthArthr, n=6) and healthy volunteers (Normal, n=15). For sake of simplicity, we focus only on two markers: CD15 and CD3, that are both largely used to classify synovial tissues in the literature. Thus, the number of markers, in this empirical example, is $q = 2$.

As the aim of our analysis is to evaluate the classification accuracy in a four-class framework, we generate all possible combinations of four groups

taken from the set of six disease groups, always including the Normal and the SeA groups. The analysis, thus, refers to six different subsets of disease, as shown in Table 6.4.

For each of the investigated subset of groups, reported in the first column of the table, we implement our $\hat{\boldsymbol{\beta}}_{L4}$ estimator. Table 6.4 shows the estimated HUM, as well as the estimated coefficients of the linear combination and their standard errors obtained with the proposed approach. The coefficient associated to the CD15 marker has been fixed to one, while the β_2 coefficient is the one associated to the CD3 marker. Finally, in the last column of the table, the joint HUM calculated with the HUM_{LF} estimator is reported.

The estimated β_2 coefficients are relatively low, although they present very small standard errors. The reduced magnitude of the coefficients denotes a limited contribution of CD3 to the optimal linear combination. This result, however, is not surprising if we refer to Table 5.7 and Table 5.11 in Chapter 5, where both markers were investigated individually. As also reported for convenience in Table 6.4, when taken one at the time, the classification accuracy of CD15 was always almost the double than the one of CD3. Interestingly, when taken together, the estimated HUM of the linear combination, for each of the subsets of disease, is always substantially higher than the one associated to the individual marker. This fundamental result further stresses the importance of combining markers in order to increase the accuracy in classifying subjects to the different classes. Finally, the comparison of the last two columns shows an HUM value completely in line with the one estimated through the HUM_{LF} estimator, reinforcing what we have already concluded in the previous section.

			HUM_{L4}			HUM_{LF}
			CD15	CD3	CD15+$\hat{\beta}_2$CD3	
Subset	$\hat{\beta}_2$	se	$\widehat{\text{hum}}$	$\widehat{\text{hum}}$	$\widehat{\text{hum}}$	$\widehat{\text{hum}}$
Normal, OA, RA, SeA	0.038	0.010	0.657	0.346	0.730	0.699
Normal, OrthArthr, OA, SeA	0.149	0.032	0.388	0.334	0.442	0.450
Normal, OrthArthr, RA, SeA	0.025	0.014	0.650	0.347	0.693	0.719
Normal, OrthArthr, Early, SeA	0.031	0.013	0.621	0.463	0.740	0.757
Normal, OA, Early, SeA	0.101	0.037	0.669	0.463	0.749	0.719
Normal, Early,RA, SeA	0.086	0.063	0.526	0.250	0.587	0.604

Table 6.4: Estimated coefficients and bootstrap standard errors ($k = 500$ replications) of the optimal linear combination, hypervolume values for the combined marker. For identification issues, we fix $\beta_1 = 1$.The last column reports the hypervolume value relative to the two markers, resulting from the Li and Fine (2008) approach.

Concluding remarks and further researches

This work was motivated by a real research question in the field of biostatistics. We were interested in finding out a method to evaluate the accuracy measure of a biomarker in discriminating a sample of patients divided in four classes according to the severity of a disease. The natural way to afford the problem has been to start reviewing the literature over the Receiver Operating Characteristic curve analysis. The ROC curve is a graphical tool used to perform analysis of classification in different scientific fields. In medicine it is often used to evaluate the accuracy of a diagnostic test for discriminating between two classes with respect to a gold standard. A summary measure of the classification accuracy is provided by the Area Under the Curve. The ROC curve analysis is applicable, by construction, only to problems with dichotomous outcomes.

However, it often happens to deal with studies in which the aim is to classify in more than two classes. The ROC manifold and that of Hypervolume Under the Manifold are two theoretical concepts introduced in the literature at the beginning of this century, with the aim of generalizing the idea of the ROC curve to more than two groups. These theoretical contributions, however, have not been accompanied by a large diffusion of implementable methodologies and related empirical applications.

As our problem was related to a four-class sample, we focused on ROC Manifolds and its relative HUM. We proposed a semi-parametric approach to derive the ROC manifold and the hypervolume under the manifold. Our contribution relies on the Lehmann assumptions and constitutes the generalization of the work by Gönen and Heller (2010) and Nze Ossima et al. (2015) to a four-class setting. We provided the analytical representation of the HUM estimator, that we called HUM_{L4}, and of its variance. Furthermore an inferential solution for the estimator and for the variance has been proposed.

To evaluate the performance of the suggested estimator we carried out

extensive simulation studies in which we made a comparison with two other estimators present in the literature. As expected, our estimator presents highly satisfactory performances under the Lehmann conditions in both small and large samples. In addition, we observed that the behaviour of HUM_{L4} does not dependent on the distance among the distributions. In fact, it performs well in estimating both small and large hypervolumes. Moreover, the coverage rate is always large and very close to the nominal level.

Departures from the Lehmann assumptions influence the bias of the estimator that increases as the sample size increases. Overall, the performances of our estimator are never dramatically inferior with respect to those of the two other estimators, even when the data are generated under very unfavourable conditions, i.e. large departures from the Lehmann assumption.

Furthermore, we also evaluated our proposed method in terms of computational time. Using our approach, calculating the HUM takes no more than few seconds, even for large samples. This is a good advantage of our procedure, especially when compared to the two alternative estimators existing in the literature, that, in addition to requiring more than the double in small samples, presents exploding computational times when the sample size becomes relatively large.

Moreover, our method is based on a well developed framework, easy to handle and implementable in all standard statistical packages. In addition, the regression framework at the base of our approach, provides multiple advantages. Firstly, it enables to control for the possible effects of covariates on the accuracy of the diagnostic test. Secondly, the estimates of the coefficients related to the class variables indirectly provides a way to test whether the classes are significantly different one to another. If it were not the case, one could simplify the analysis by grouping classes with the same distributions. These are remarkable points of our methodology.

Finally, with the aim of extending the scope of our methodological contribution, we considered the case of a linear combination of markers. Often, in applied medical research, using a linear combination of distinct markers can produce a different indicator which enhances diagnostic capability. Thus, in the last part of this thesis, we proposed a method to linearly combine multiple markers for four-category classifications, directly based on the optimization of the accuracy of the combined marker under the ROC criteria. We suggest to numerically maximize the HUM_{L4} while searching for the optimal linear combination of markers. The data analyses proved that the resulting models based upon the related linear combinations generate a HUM notably larger than the one obtained with the single markers. Thus, we can conclude that our method can be considered a good candidate when a more informative insight in the capability of a panel of markers to distinguish and classify

patients in different categories is needed.

Some points may deserve future investigation. As we stated diffusely in this thesis, our estimator is a generalization to four classes of the original idea by Gönen and Heller (2010) and Nze Ossima et al. (2015) for two- and three-classification frameworks. Thus, a first line of research would be to search for the existence of a recursive formula for the general m-class estimator. Second, as the scope of this thesis was principally empirical, we have not treated in detail the theoretical properties of the new estimator, that thus remains an open issue for future research. Finally, in medical research the interest in combining the information of multiple markers is still increasing; thus we think that another research point might focus on the optimal combination of biomarkers, such as attempting to solve for a more efficient algorithm in maximizing the HUM and trying to deal with potential non linear combinations.

Appendix A

Some notes on the Cox Regression

In the previous chapters we have introduced the Cox regression as an important tool in the estimation process of the classification accuracy of a particular marker. The Cox regression (sometimes also indicated as proportional hazards regression) has been originally proposed by Cox (1972) as a method for investigating the effect of several variables upon the time a particular event occurs. While it is a reference model in survival analysis, it finds much less applications outside this specific field. In this appendix we recall some main concepts about this methodology.

A.1 The model

Suppose T is a non negative random variable representing a survival time, $f(t)$ is the probability density function, $F(t)$ is the cumulative distribution function and $S(t) = P(T \geq t) = 1 - F(t)$ is the survival function. Furthermore, we also assume that T is a continuous random variable. The variable T is non negative by definition and, in medicine, usually denotes the elapsed time until an event takes to happen. It is commonly characterized by the so called "hazard function". Based on the definition of cumulative density function and survival function, we know that:

$$P(t \leq T < t + \delta t | T \geq t) = \frac{P(t \leq T < t + \delta t)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{S(t)}.$$

Considering infinitesimal variations, we can define the hazard function $h(t)$ as

$$h(t) = \lim_{\delta \to 0} \left\{ \frac{F(t + \delta t)}{\delta t} \right\} \frac{1}{S(t)} = \frac{f(t)}{S(t)}.$$

The hazard function, thus, is the ratio between the probability density function and the survival function. More simply, it can be interpreted as the instantaneous risk that the event of interest happens, within a very narrow time frame.

We also know that:

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}[1 - S(t)],$$

thus, an alternative formulation of the hazard function becomes:

$$h(t) = \frac{-dS(t)}{dt}\frac{1}{S(t)} = \frac{-d[logS(t)]}{dt}$$

Finally, the cumulative hazard function is defined as:

$$H(t) = \int_0^t h(u)du = -logS(t).$$

As an example, if we assume an hazard function constant over time, then, it is possible to show that the survival time will be distributed as an Exponential random variable of parameter λ. In fact, let

$$h(t) = \lambda$$

be the hazard function. Then

$$S(t) = \exp\{-\int_0^t \lambda du\} = e^{-\lambda t}$$

and

$$f(t) = \lambda e^{-\lambda t}.$$

Thus, the survival time will be distributed as an Exponential random variable $T \sim$Exp(λ). If, instead, the hazard function is allowed to depend on time, then the hazard function will be modeled, for example, has a Weibull distribution. In fact, let

$$h(t) = \lambda\nu t^{\nu-1}$$

be the hazard function with parameters λ, $\nu > 0$. Then the survival function will be

$$S(t) = \exp\{-\int_0^t \lambda\nu u^{\nu-1}du\} = \exp(-\lambda t^{\nu})$$

and the probability density function will be

$$f(t) = \lambda\nu t^{\nu-1}\exp(-\lambda t^{\nu}).$$

In this particular case, $T \sim Wei(\lambda, \nu)$, where λ, $\nu > 0$ are the scale and shape parameters respectively.

If the lifetimes of all units of our sample are not governed by the same survival function $S(t)$, the survival model should allow for the presence of a vector of covariates or explanatory variables that may affect survival time. The Proportional hazards model introduced by Cox (1992) is one of the simplest survival models. It focuses directly on modeling the hazard function rather than the survival time. The hazard at time t for the i-th subject with covariates $\boldsymbol{Z}_i$ can be written as

$$h(t) = h_0(t) \exp\{\boldsymbol{z}^{'}\boldsymbol{\beta}\}$$

where $h_0(t)$ is the baseline hazard function and describes the risk for subject with $\boldsymbol{Z}_i = 0$, while $\exp\{\boldsymbol{Z}^{'}\boldsymbol{\beta}\}$ is the relative risk associated with the set of covariates $\boldsymbol{Z}_i$. For example, consider a situation similar to the one showed in Section 5.1.1, in which the population is characterized by four distinct groups. In this case, the covariates are represented by variables indicating the groups and there will be four different hazard functions. More specifically, if we define three dichotomous variables d_1, d_2, d_3 which serve to indicate the groups, then, for the i-th individual, the proportional hazards model is

$$h_i(t|d_1, d_2, d_3) = \begin{cases} h_0(t) & d_{1i} = 0, d_{2i} = 0, d_{3i} = 0 \\ h_0(t) \exp\{\beta_1 d_{1i}\} & d_{1i} = 1, d_{2i} = 0, d_{3i} = 0 \\ h_0(t) \exp\{\beta_1 d_{1i} + \beta_2 d_{2i}\} & d_{1i} = 1, d_{2i} = 1, d_{3i} = 0 \\ h_0(t) \exp\{\beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i}\} & d_{1i} = 1, d_{2i} = 1, d_{3i} = 1. \end{cases}$$

Thus, for example, if we want to compare group 1 with group 0, $h_0(t)$ represents the risk at time t in group 0 (the reference group), and $\theta_1 = \exp\{\beta_1 d_1\}$ is the ratio of the risk in group 1 relative to group 0 at any time t.

Finally, given that all the elements in the model are positive, if we take the logs, the proportional hazards model becomes an additive model for the log of the hazard, with

$$\log\left(h_i(t \mid d_1, d_2, d_3)\right) = \log\left(h_0(t)\right) + \boldsymbol{d}_i^{'}\boldsymbol{\beta}$$

with $\boldsymbol{d}_i^{'}$ collecting information on the group indicators for the i-th subject .

A.2 Estimation of the Cox model

Let the hazard function be a function of a set of p covariates, whose realization for the i-th subject is denoted as $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$. The hazard

function in the Cox specification, thus, can be written as

$$\begin{aligned} h(t\,|\boldsymbol{Z}_i) &= h_0(t)\exp\left(\beta_1 Z_{i1} + \cdots + \beta_p Z_{ip}\right) \\ &= h_0(t)\exp\left(\boldsymbol{Z}_i\,\boldsymbol{\beta}\right). \end{aligned} \tag{A.1}$$

As we have defined before, the probability that at time T_i the event occurs for subject i can be defined as

$$\begin{aligned} L_i(\boldsymbol{\beta}) &= \frac{h(T_i\,|\boldsymbol{Z}_i)}{\sum_{j:\,T_j \geq T_i} h(T_i\,|\boldsymbol{Z}_j)} \\ &= \frac{h_0(T_i)\theta_i}{\sum_{j:\,T_j \geq T_i} h_0(T_i)\theta_j} \\ &= \frac{\theta_i}{\sum_{j:\,T_j \geq T_i} \theta_j} \end{aligned} \tag{A.2}$$

where, for simplicity, we have defined $\theta_j = \exp\left(\boldsymbol{Z}_j\boldsymbol{\beta}\right)$. Importantly, the summation at the denominator is over the set of subjects for which the event has not yet occurred before time Y_i (including subject i itself). Cox named this quantity, that is only a function of the unknown parameters, "partial likelihood". Moreover, he observed that the unknown parameters $\boldsymbol{\beta}$ without the need to model the change of the hazard function over time.

Under the standard assumption of subjects being independent of each other, it is possible to derive the joint probability of all the events as

$$L\left(\boldsymbol{\beta}\right) = \prod_{i:\,C_i=1} L_i(\boldsymbol{\beta}) \tag{A.3}$$

where $C_i = 1$ indicates the occurrence of the event. Moreover, given that $L_i\left(\boldsymbol{\beta}\right) > 0$ by definition, is becomes easier to transform $L\left(\boldsymbol{\beta}\right)$ in log terms:

$$l\left(\boldsymbol{\beta}\right) = \sum_{i:\,C_i=1} \boldsymbol{Z}_i\,\boldsymbol{\beta} \;-\; \log \sum_{j:\,T_j \geq T_i} \theta_j \tag{A.4}$$

that represents the so called log partial likelihood. Maximizing the function over $\boldsymbol{\beta}$ produces the maximum partial likelihood estimator of the unknown parameters of the Cox proportional hazards model. The maximization of the partial likelihood function is generally performed using the Newton-Raphson algorithm, which benefits of the relatively simple way of obtaining analytically the first two derivatives. Specifically, the partial score function (vector of first derivatives) is

$$l'(\boldsymbol{\beta}) = \sum_{i:\,C_i=1} \boldsymbol{Z}_i - \left(\frac{\sum_{j:\,T_j \geq T_i} \theta_j \boldsymbol{Z}_j}{\sum_{j:\,T_j \geq T_i} \theta_j}\right), \tag{A.5}$$

while the Hessian matrix (matrix of second derivatives) is

$$l''(\boldsymbol{\beta}) = -\sum_{i:\, C_i=1} \left(\frac{\sum_{j:\, T_j \geq T_i} \theta_j \boldsymbol{Z}_j \boldsymbol{Z}_j^T}{\sum_{j:\, T_j \geq T_i} \theta_j} - \frac{\left(\sum_{j:\, T_j \geq T_i} \theta_j \boldsymbol{Z}_j\right)\left(\sum_{j:\, T_j \geq T_i} \theta_j \boldsymbol{Z}_j^T\right)}{\left(\sum_{j:\, T_j \geq T_i} \theta_j\right)^2} \right). \tag{A.6}$$

Importantly, the Hessian matrix, evaluated at the estimate of $\boldsymbol{\beta}$, provides information on how precise the estimates are. In fact, as the Fisher Information Matrix in the maximum likelihood estimator, the inverse of the observed Hessian matrix approximates the variance-covariance matrix for the estimates, whose elements on the main diagonal produce standard errors for the regression coefficients.

Finally, from an inferential point of view, it is extremely important to stress that the partial likelihood estimators have the desirable properties of being consistent and asymptotically normally distributed. The inference on the unknown parameters β, thus, is standard and, asymptotically, refers to multivariate normal distribution, where the asymptotic covariance matrix can be estimated through the inverse of the observed Hessian matrix.

Appendix B

Simulations coverage rate

In this appendix we report the results of the empirical coverage probability of the 95% confidence intervals associated with the estimates obtained under the three data generation process exposed in Chapter 4.

The first set of simulations are based on data generated under the Lehmann condition (see Section 4.2). In order to evaluate if the expected coverage probability is achieved, in Table B.1 we present the empirical coverage probability of the 95% confidence intervals associated with the estimates obtained under the approach we propose. The last column reports the percentage of the 1000 computed independent confidence intervals which contains the true HUM value. We can notice that the empirical coverage probabilities are reasonably close to the nominal level.

		HUM_{L4}	se	coverage
	n_1=n_2=n_3=n_4=30	0.270	0.042	94.4
case 1	n_1=n_2=n_3=n_4=50	0.267	0.032	94.8
	n_1=n_2=n_3=n_4=80	0.266	0.026	94.0
	n_1=n_2=n_3=n_4=30	0.565	0.053	95.0
case 2	n_1=n_2=n_3=n_4=50	0.563	0.041	94.9
	n_1=n_2=n_3=n_4=80	0.561	0.033	94.5
	n_1=n_2=n_3=n_4=30	0.933	0.024	92.5
case 3	n_1=n_2=n_3=n_4=50	0.933	0.018	93.6
	n_1=n_2=n_3=n_4=80	0.933	0.015	93.8

Table B.1: Empirical coverage rate of the 95% CI in the Weibull case under the Lehmann assumption and three different vectors of parameters $\boldsymbol{\beta}$ of the Cox proportional hazards regression model.

If the data are generated from Weibull distributions with group specific

shape parameters the Lehmann condition does not hold (see Section 4.3.1). Looking at Table B.2, it emerges that the coverage rates are dramatically low. However, this can be due to the combination of two factors: the bias of our estimator under departure from the Lehmann condition and the low standard errors of the estimates.

		HUM_{L4}	se	coverage
case 4	n_1=n_2=n_3=n_4=30	0.160	0.024	0.313
	n_1=n_2=n_3=n_4=50	0.159	0.018	0.096
	n_1=n_2=n_3=n_4=80	0.158	0.015	0.015
case 5	n_1=n_2=n_3=n_4=30	0.481	0.055	0.681
	n_1=n_2=n_3=n_4=50	0.478	0.045	0.508
	n_1=n_2=n_3=n_4=80	0.475	0.035	0.307
case 6	n_1=n_2=n_3=n_4=30	0.721	0.049	0.299
	n_1=n_2=n_3=n_4=50	0.716	0.039	0.083
	n_1=n_2=n_3=n_4=80	0.713	0.032	0.004

Table B.2: Empirical coverage rate of the 95% CI in the Weibull case with group-specific shape parameters (Lehmann condition is not satisfied) and three different vectors of parameters $\boldsymbol{\beta}$ of the Monte Carlo inversion method.

The last simulations proposed in the thesis focus on data generated from Gaussian distributions (see Section 4.3.2). The coverage probabilities associated with the estimation procedure reflects the bias of the HUM_{L4} estimator, as shown in Table B.3 where the empirical coverage probabilities are listed. Although in the first scenario (*case 7*) the coverage is greater than 80 per cent, in *case 8* and *case 9* it systematically decreases.

		HUM_{L4}	se	coverage
	n_1=n_2=n_3=n_4=30	0.069	0.016	85.0
case 4	n_1=n_2=n_3=n_4=50	0.069	0.012	83.9
	n_1=n_2=n_3=n_4=80	0.069	0.010	81.5
	n_1=n_2=n_3=n_4=30	0.305	0.050	57.9
case 5	n_1=n_2=n_3=n_4=50	0.300	0.039	42.3
	n_1=n_2=n_3=n_4=80	0.299	0.031	26.2
	n_1=n_2=n_3=n_4=30	0.672	0.060	48.0
case 6	n_1=n_2=n_3=n_4=50	0.664	0.047	22.1
	n_1=n_2=n_3=n_4=80	0.659	0.039	5.7

Table B.3: Empirical coverage rate of the 95% CI in the Normal case with group-specific expected values and equivalent variances.

Bibliography

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11): 1713–1723, 2005.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

D. R. Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

C. Della Beffa, E. Slansky, C. Pommerenke, F. Klawonn, J. Li, L. Dai, H. R. Schumacher Jr, and F. Pessler. The relative composition of the inflammatory infiltrate as an additional tool for synovial tissue classification. *PloS one*, 8(8):e72494, 2013.

D. D. Dorfman and E. Alf. Maximum likelihood estimation of parameters of signal detection theory - a direct solution. *Psychometrika*, 33(1):117–124, 1968.

D. D. Dorfman, K. S. Berbaum, C. E. Metz, R. V. Lenth, J. A. Hanley, and H. A. Dagga. Proper receiver operating characteristic analysis: the bigamma model. *Academic Radiology*, 4(2):138–149, 1997.

S. Dreiseitl, L. Ohno-Machado, and M. Binder. Comparing three-class diagnostic tests by three-way roc analysis. *Medical Decision Making*, 20(3): 323–331, 2000.

M. Gao and J. Li. *mcca: Multi-Category Classification Accuracy*, 2018. URL `https://CRAN.R-project.org/package=mcca`. R package version 0.2.0.

M. Gönen and G. Heller. Lehmann family of roc curves. *Medical Decision Making*, 30(4):509–517, 2010.

P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

X. He, C. E. Metz, B. M. Tsui, J. M. Links, and E. C. Frey. Three-class roc analysis-a decision theoretic approach under the ideal observer framework. *IEEE Transactions on Medical Imaging*, 25(5):571–581, 2006.

L. Kang and L. Tian. Estimation of the volume under the roc surface with three ordinal diagnostic categories. *Computational Statistics & Data Analysis*, 62:39–51, 2013.

E. L. Lehmann. The power of rank tests. *The Annals of Mathematical Statistics*, pages 23–43, 1953.

J. Li and J. P. Fine. Roc analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, 9(3): 566–576, 2008.

J. Li, Y. Chow, W. K. Wong, and T. Y. Wong. Sorting multiple classes in multi-dimensional roc analysis: parametric and nonparametric approaches. *Biomarkers*, 19(1):1–8, 2014.

J. Li, J. P. Fine, and M. J. Pencina. Multi-category diagnostic accuracy based on logistic regression. *Statistical Theory and Related Fields*, 1(2): 143–158, 2017.

C. 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.

L. B. Lusted. Logical analysis in roentgen diagnosis: memorial fund lecture. *Radiology*, 74(2):178–193, 1960.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

C. E. Metz and X. Pan. Proper binormal roc curves: theory and maximum-likelihood estimation. *Journal of mathematical psychology*, 43(1):1–33, 1999.

C. E. Metz, B. A. Herman, and J.-H. Shen. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in medicine*, 17(9):1033–1053, 1998.

D. Mossman. Three-way rocs. *Medical Decision Making*, 19(1):78–89, 1999.

C. T. Nakas and C. T. Yiannoutsos. Ordered multiple-class roc analysis with continuous measurements. *Statistics in medicine*, 23(22):3437–3449, 2004.

N. Novoselova, C. Della Beffa, J. Wang, J. Li, F. Pessler, and F. Klawonn. Hum calculator and hum package for r: easy-to-use software tools for multicategory receiver operating characteristic analysis. *Bioinformatics*, 30 (11):1635–1636, 2014.

N. Novoselova, J. Wang, and F. K. Pessler. *Biocomb: Feature Selection and Classification with the Embedded Validation Procedures for Biomedical Data Analysis*, 2017. URL `https://CRAN.R-project.org/package=Biocomb`. R package version 0.3.

A. D. Nze Ossima, J.-P. Daurès, F. Bessaoud, and B. Trétarre. The generalized lehmann roc curves: Lehmann family of roc surfaces. *Journal of Statistical Computation and Simulation*, 85(3):596–607, 2015.

M. S. Pepe. An interpretation for the roc curve and inference using glm procedures. *Biometrics*, 56(2):352–359, 2000.

M. S. Pepe. *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

M. S. Pepe, T. Cai, and G. Longton. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229, 2006.

M. T. Rodia, R. Solmi, F. Pasini, E. Nardi, G. Mattei, G. Ugolini, L. Ricciardiello, P. Strippoli, R. Miglio, and M. Lauriola. Lgals4, ceacam6, tspan8, col1a2: blood markers for colorectal cancer. validation in a cohort of subjects positive to faecal immunochemical test. *Clinical colorectal cancer*, 17(2):217–228, 2018.

D. Schoenfeld. Chi-square goodness of fit tests for the proportional hazards model. *Biometrika*, 67:145–53, 1980.

B. K. Scurfield. Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, 40(3):253–269, 1996.

B. K. Scurfield. Generalization of the theory of signal detectability toneventm-dimensional forced-choice tasks. *Journal of Mathematical Psychology*, 42(1):5–31, 1998.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

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

L. J. Wei. Testing goodness of fit for the proportional hazards model with censored observations. *Journal of the American Statistical Association*, 79: 649–652, 1984.

C. Xiong, G. van Belle, J. P. Miller, and J. C. Morris. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in medicine*, 25(7):1251–1273, 2006.

Y. Zhang and J. Li. Combining multiple markers for multi-category classification: An roc surface approach. *Australian & New Zealand Journal of Statistics*, 53(1):63–78, 2011.