TESE DE DOUTORAMENTO

Nonparametric Methods for the Comparison of ROC Curves with Application to Biomedicine

Arís Fanjul Hevia

ESCOLA DE DOUTORAMENTO INTERNACIONAL
DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO
EN ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

SANTIAGO DE COMPOSTELA
2020

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DA AUTORA DA TESE

Nonparametric Methods for the Comparison of ROC Curves with Application to Biomedicine

Arís Fanjul Hevia

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1. A tese abarca os resultados da elaboración do meu traballo.
2. De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.
3. A tese é a versión definitiva presentada para a súa defensa e coincide coa versión enviada en formato electrónico.
4. Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.

En Oviedo, a 8 de decembro de 2020

Asdo.: Arís Fanjul Hevia

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DOS DIRECTORES/TITOR DA TESE

D. Wenceslao González Manteiga
D. Juan Carlos Pardo Fernández

En condición de titor e directores da Tese de Doutoramento titulada

Nonparametric Methods for the Comparison of ROC Curves with Application to Biomedicine

INFORMAN

Que a presente tese se corresponde co traballo realizado por Dna. Arís Fanjul Hevia, baixo a nosa dirección/titorización, e autorizamos a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como directores/titor desta non incorre nas causas de abstención establecidas na Lei 40/2015.

En Santiago de Compostela,
a 8 de decembro de 2020

En Vigo, a 8 de decembro de 2020

Asdo.: Wenceslao González Manteiga

Asdo.: Juan Carlos Pardo Fernández

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimientos

Un proyecto tan largo como una tesis nunca es producto de una sola persona. Estas palabras de agradecimiento van dirigidas a todas aquellas que, desde la cercanía o la distancia, me apoyaron durante estos últimos años.

En primer lugar quiero mostrar mi profundo agradecimiento a mis dos directores, Wences y Juan Carlos. Wences, recuerdo perfectamente el día en que me llamaste a tu despacho y me ofreciste hacer la tesis contigo. Este es un camino que difícilmente habría tomado si no hubiese sido por ti. Gracias por animarme a meterme en este mundo y por guiarme durante todos estos años. Juan Carlos, una vez me dijiste que "o que non está escrito non existe" y, desde luego, esta tesis no estaría escrita ni sería lo que es si no fuera por ti. Gracias por tu dedicación, por tu atención al detalle y por estar siempre disponible a un "Skype" de distancia.

Gracias a los miembros del Departamento de Estadística, Análisis Matemático y Optimización de la USC, a los profesores y muy especialmente a los compañeros doctorandos. Gracias por los consejos, por resolver y aguantar todas mis dudas, por hacer de todos los congresos experiencias memorables. Gracias Jose, María, Andrea, Mercedes, Maribel, Ángel, Álex, David, Laura F., Alejandra, Laura D., Brais y María A. Extiendo los agradecimientos al resto de compañeros de la facultad, por ese ambiente tan bueno que va más allá de las paredes del edificio. Gracias a los compañeros de la Sala π, que hicisteis más fácil y agradable el pasar tantas horas en el lugar de trabajo. Por esas tardes en las que probamos que el núcleo importa, por esos momentos con cámaras secretas, con gatos y con soluciones apareciendo en los sitios más inverosímiles. Gracias a los que ya he mencionado y a Juan Carlos, Chiara, Marta, Mohsen, Ruth y Alfredo. No puedo dejar de nombrar también a aquellos compañeros con los que compartí la experiencia de organizar el Seminario de Iniciación a la Investigación y, por supuesto, a todos los del grupo del café (tomasen o no café). Es difícil mencionaros a todos, y a algunos tendría que daros las gracias más de una vez. Gracias Suso, Lucía, Víctor, Luis, Xabi, Ignacio, Gonzalo, Saray, Aida, Jorge, Bea, Franco, Pilar, David...

I would also like to express my gratitude to Prof. Ingrid Van Keilegom for giving me the opportunity to work in a different environment. Thanks for inviting me twice to Leuven, for your kindness and your thoughtfulness. And of course, thanks to all the colleagues that I met over there, like Motti, Andrea, Edith, Min or Leonard, as well to Ester, for contributing to such a wonderful experience. Gracias de nuevo a Xabi, a Álex, a Jose y a Mercedes por esos momentos en los que coincidimos estando de estancia.

Durante todos estos años en Santiago también conté con el apoyo de muy buena gente fuera de la facultad, como el de Rita y Chus, o como el de mis diversos compañeros de piso. Gracias a las Cristinas, a María, a Laura y a Diego por vuestra compañía y vuestro cariño, por darme un espacio donde poder desconectar.

Dirigiendo mi atención ahora a mi tierra, me gustaría agradecer también a los miembros del Departamento de Estadística, Investigación Operativa y Didáctica de la Matemática de la Universidad de Oviedo, en especial a Sonia, Javi, Irene, Quique y Sara. Gracias por acogerme tan cálidamente en esta "vuelta a casa" y por vuestra ayuda en estos últimos meses tan complicados.

Una parte muy especial de estos agradecimientos va dirigida a mis compañeros de promoción. Pese a la distancia y a solo vernos de cena en cena lleváis siendo un apoyo muy importante para mí desde esas primeras clases en las que nos introdujimos juntos en el mundo de las matemáticas. Gracias Isa, Laura, Víctor, Andrea, Carlos, Alberto, Ángela y Belén.

Gracias a todos los amigos de Celorio por todos esos veranos de desconexión, de playa, Sella, sidra, fiestas de prao y estrellas fugaces. Y por todos los que vengan.

Gracias a Alba y a Paula. Gracias, capicúas, por ser una constante durante más años de los que me atrevo a contar. Por estar ahí para todo y en todo momento.

Y, como no podía ser de otra forma, un agradecimiento muy grande a toda mi familia, por ser esa fuente de energía, cariño y emoción que lo mueve todo. Gracias a la familia de Noreña (esté o no esté en Noreña), a todos mis tíos y a mis numerosos primos paternos, por vuestro interés y ánimo constantes. Un agradecimiento especial a Nacho, por todas las veces que me hiciste de "embajador" en Galicia. Gracias también a mis tías maternas, Paz y Cova, por vuestro cariño y apoyo, y a mis primos Fernando y Cova (y Enric) por todos los consejos, los ánimos y los ratos que pasamos juntos. Y gracias, Nel, por llenarnos a todos de sonrisas.

Gracias Víctor. No sabes la de veces que has hecho mi vida más fácil, abriendo camino para mí. Seguir tus huellas siempre ha hecho que mis pasos lleguen más lejos de lo que podrían hacerlo sin tu guía, incluso cuando caminamos en distintas direcciones. Y gracias, Mer, por tu cariño y tu amabilidad, por estar siempre dispuesta a ayudar.

Gracias papá y mamá. Por aseguraros de que tuviera siempre todas las oportunidades posibles, y algunas más. Por apoyarme en cada decisión y en cada etapa. Todo lo bueno lo he sacado de vosotros

Finalmente, gracias abuela por tu "cuarta parte de genes", por ser un ejemplo de vida a seguir para mí día tras día. Y gracias, abuelo: pese a dejarnos justo cuando me embarcaba en esta aventura, tu recuerdo sigue muy presente en mi memoria, siendo siempre una fuente de inspiración.

Gracias, en definitiva, a todos los que hicisteis que esta etapa del doctorado fuera una experiencia que va más allá de lo que queda recogido en estas páginas.

Arís Fanjul Hevia
Oviedo, diciembre de 2020

Funding

This work has been supported by the Spanish Ministerio de Educación, Cultura y Deporte (fellowship FPU14/05316), as well as by the Spanish Ministerio de Economía, Industria y Competitividad, through grant numbers MTM2013-41383P and MTM2016-76969-P, which includes support from the European Regional Development Fund (ERDF) and the IAP network P7/06 StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Sistems in natural, biomedical and social science) of the Belgian Government (Belgian Science Policy). Part of the research done in Chapter 4 and Chapter 5 was carried out during two visits to KU Leuven, supported by Mobility Grants EST17/00442 and EST18/00673 (from the Spanish Ministerio de Educación y Formación Profesional). The Supercomputing Center of Galicia (CESGA) is acknowledged for providing the computational resources that allowed to run most of the simulations. Finally, Dr. F. Gude (Unidade de Epidemioloxía Clínica, Hospital Clínico Universitario de Santiago) is thanked for providing the data sets employed in this dissertation.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 1

Introduction

The *Receiver Operating Characteristic curve* (ROC curve) is a statistical tool that analyses the accuracy of a certain method of classification. This means that, given a binary classifier, the study of its corresponding ROC curve can tell us how well that classifier is capable of discriminating between two different groups.

The idea of the ROC curve manifests during World War II, where radar receiver operators (hence, the name) were interested in differentiating the signal of a potential enemy aircraft from simple noise. It was developed in the fields of radar signal detection and psychophysics, being Green and Swets (1966) one of the earliest references. Since then it has been an active area of research.

Later on, its potential became evident in medical studies, where the correct diagnosis of patients plays a decisive role. The ROC curve is nowadays a commonly accepted way of analysing the discriminatory capability of a diagnostic method, and it has been involved in the solution of different problems. See the books of Pepe (2003) or Krzanowski and Hand (2009) for an overview on this topic.

The goal of this dissertation is to study and design new nonparametric methods for comparing two or more of these ROC curves, taking into account information of other covariates that may o may not influence the result of the study. This first chapter is devoted to introducing the concept of the ROC curve and specify the objectives that will be pursued (Section 1.1) and clarify the structure of this document (Section 1.2).

1.1 Background on ROC curves

In this first section we show how an ROC curve is built, along with some of its main properties, summary measures and related parametric models. Lastly, we discuss some of its main applications.

Table 1.1: Successes and errors that can be obtained from the different combinations of diagnosis and true status of a subject.

		Status	
		Diseased (D)	Healthy (H)
Diagnosis	Diseased (+)	*True Positive* (TP)	*False Positive* (FP)
	Healthy (−)	*False Negative* (FN)	*True Negative* (TN)

1.1.1 A problem of classification

The starting point of any ROC curve analysis is a problem of classification: there is a population divided in two different categories[1] and, given a subject of such population, we want to determine the category to which it belongs to.

Since the applications of the ROC curve that we will be considering throughout these chapters are set in a biomedical environment, we will consider this population to be a set of patients suspected of having a certain condition or disease. Thus, the binary classifier will be a method of diagnosis whose aim is to tell apart the healthy and the diseased populations (identified as H and D, respectively). Nevertheless, all the methods that will be discussed can be applied in other areas (see Chapter 10 in Krzanowski and Hand, 2009, for other examples).

A diagnostic method, on the basis of available information, will classify each subject as diseased (+) or healthy (−). In order to select an appropriate diagnostic method we have to take into account that there are two types of error measurements involved in the decision process. Those errors depend on the combination of the diagnosis (+ or −) and the true status (D or H) of the subject, as summarized in Table 1.1. The misclassifications happen either when a healthy patient (H) is diagnosed as diseased (+), called a *False Positive* (FP), or a diseased patient (D) is diagnosed as a healthy one (−), called a *False Negative* (FN). Correct diagnoses happen when the healthy (H) are diagnosed as healthy (+) (which is called a *True Positive*, TP) or when the diseased (D) are diagnosed as diseased (−) (called a *True Negative*, TN). When we consider the probabilities associated to these terms, we obtain the concepts of *sensitivity* and *specifitity*:

- *Sensitivity* $= P($ diagnosis $+ \mid$ status $D)$, the probability of correctly detecting the condition of interest.

- *Specificity* $= P($ diagnosis $- \mid$ status $H)$, the probability of diagnosing as healthy a subject that does not have the condition.

[1] The traditional ROC curve always considers a binary classifier. There are some works that consider three different classes (by defining the so-called ROC surface), but they will not be discussed here. For a review on this topic, check Nakas (2014).

It is important to consider the FP and the FN as two different ways of making an error of classification, as those errors may have different consequences depending on the situation at hand. For instance, it is essential for cancer patients to be correctly diagnosed as early as possible (and thus, it is important to have few FN), but the consequences of a FP (a healthy patient receiving a potentially dangerous treatment) could be equally worrisome.

Now, suppose that the diagnostic method depends on a continuous[2] variable Y, and that, for certain value c (called *threshold* or *cutoff*) it classifies as diseased all the subjects with $Y > c$, and as healthy all the subjects with $Y \leqslant c$. We will be referring to this variable as the *diagnostic variable* or *diagnostic marker*. This means that the notions of sensitivity and specificity are going to vary with each threshold c, and that they will depend on the distribution of the diagnostic variable Y on the healthy and on the diseased populations. To be more precise, let the random variables Y^F (with distribution F) and Y^G (with distribution G) represent the diagnostic variables in the diseased and in the healthy population, respectively. In this context the sensitivity is also called the *True Positive Fraction* (TPF), and the probability of misclassifying a diseased subject (which is the specificity complementary) is called the *False Positive Fraction* (FPF). Given a certain threshold c:

- $TPF(c) = P(Y^F > c) = 1 - F(c) = sensitivity(c)$,
- $FPF(c) = P(Y^G > c) = 1 - G(c) = 1 - specificity(c)$.

In Figure 1.1 there is a representation of these concepts. The densities of the diagnostic variables Y^F and Y^G are represented there, along with the threshold $c = 1$ and the corresponding $TPF(c)$ and $FPF(c)$. If we had a perfect diagnostic method, in which the densities of the corresponding diagnostic variables were completely separated, it would be easy to choose a threshold with both values of sensitivity and specificity equal to one. However, that is not often the case, as the densities usually overlap (as in Figure 1.1). In those cases it is not so easy to decide a criterion to choose the optimal threshold. The ideal case would be to obtain a green area ($sensitivity(c)$) as big as possible and a blue area ($1 - specificity(c)$) as reduced as possible, but both areas increase or decrease when we take greater or smaller values of c, respectively. It is not possible to increase the first and, at the same time, reduce the latter.

This is where the ROC curve comes into play. Instead of taking into account only one possible threshold value, the ROC curve considers all of them: for each value c it represents its corresponding TPF against its corresponding FPF. In other words, it represents the sensitivity against the complementary of the specificity for all the possible threshold values:

$$\{(FPF(c), TPF(c)), c \in \mathbb{R}\} = \{1 - specificity(c), sensitivity(c)), c \in \mathbb{R}\}.$$

[2]The ROC curve can be defined for discrete variables as well. However, in this dissertation we will consider that Y is continuous.

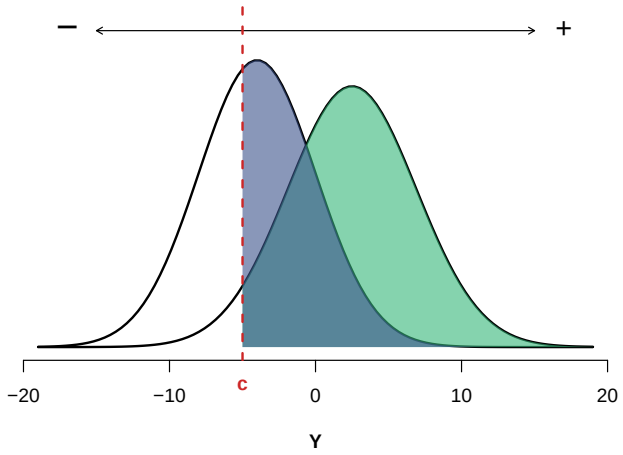


Figure 1.1: Densities of Y^F (to the right) and Y^G (to the left), the diagnostic variables on the diseased and the healthy populations, respectively. For $c = -5$, the obtained sensitivity, $TPF(c)$, is coloured in green, and the $1-$specificity, $FPF(c)$, is coloured in blue.

Taking into account that those values can be expressed using the cumulative distribution functions of the diagnostic variables, the curve can also be expressed as

$$\{(1 - G(c), 1 - F(c)), c \in \mathbb{R}\}.$$

Finally, we can rewrite it in the form of a function, using the most usual way of defining the ROC curve:

$$ROC(p) = 1 - F(G^{-1}(1 - p)),\ p \in (0, 1), \tag{1.1}$$

where G^{-1} is the quantile function associated to the distribution G.

In Figure 1.2 an ROC curve is depicted. The densities of the diagnostic variables are also represented (to the left), showing different combinations of sensitivities and specificities by taking different thresholds c_1, c_2 and c_3. Note that the ROC curve is a monotone increasing continuous function (as long as Y^F and Y^G are continuous variables) and that it takes values above the diagonal on the unit square[3]. Depending on the separation existing between the distributions of the diagnostic variables, the ROC curve will be closer to the diagonal –that represents a method based on random allocation– or will get close to the point of maximum sensitivity and specificity, the point $(0, 1)$.

This can be observed in the three scenarios represented in Figure 1.3. In the first one, the two densities overlap almost completely and, thus, the corresponding ROC curve almost coincides with the diagonal. In the third scenario, the reverse situation happens: the two densities are easily told apart, which yields an ROC curve that comes close to the

[3]Strictly speaking, the ROC curve lies above the diagonal if an only if $F(c) \leqslant G(c)$ for all c, that is, when Y^F is stochastically greater than Y^G. In practical applications the variables Y^F and Y^G may not be fully ordered, and thus some points may lie below the diagonal.

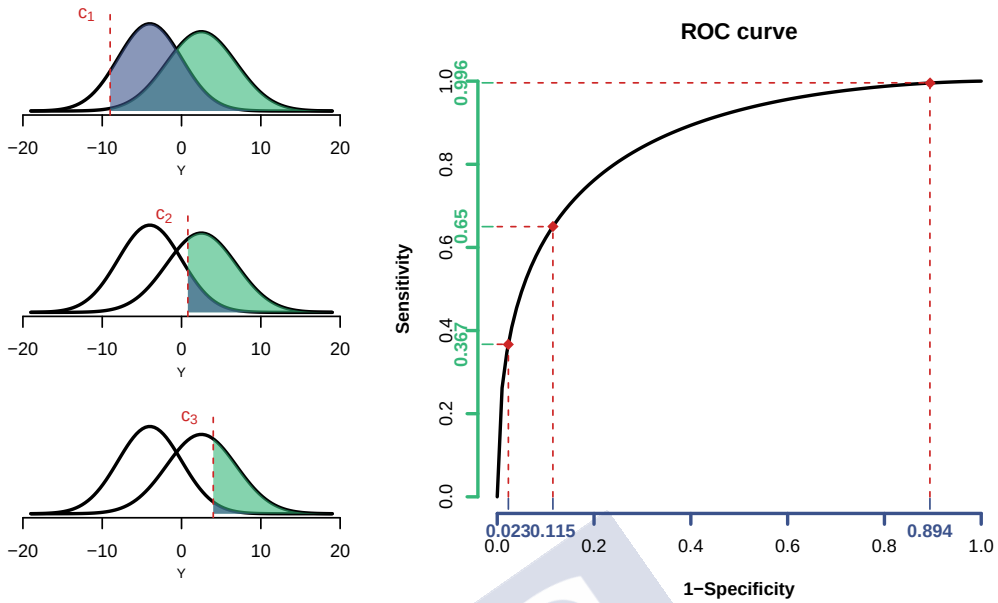


Figure 1.2: The densities of the diagnostic variables are represented to the left, with three different threshold values (c_1, c_2 and c_3) marking three different pairs of sensitivities and specificities. To the right, the ROC curve is drawn, with the pairs of $(FPF(c), TPF(c))$ highlighted in red for $c \in \{c_1, c_2, c_3\}$.

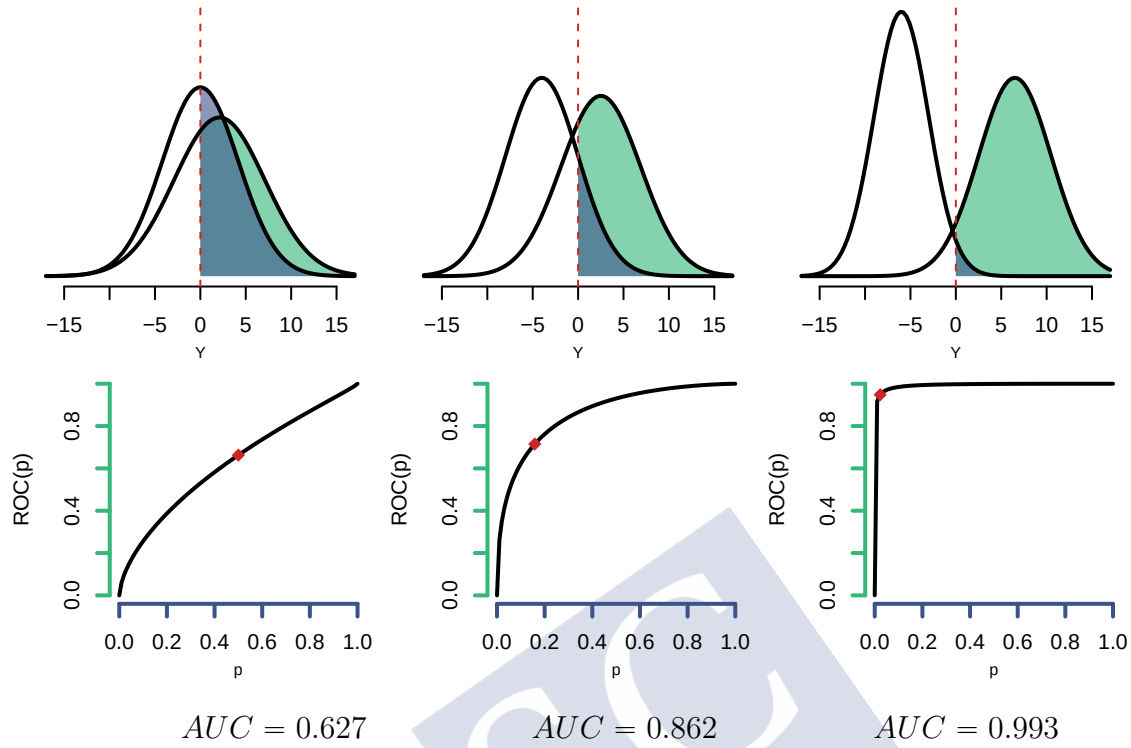
point $(0, 1)$. The second pair of densities represents a situation between those extremes.

Taking all of this into account, the ROC curve offers a general perspective –in the sense that it does not rely on the selection of the threshold for the classification criterion– to visually distinguish the diagnostic methods that have a discriminatory power from those that do not represent an improvement with respect the tossing of a coin.

One property of the ROC curve worth mentioning is that it remains invariant with respect to monotone increasing transformations of the diagnostic markers. This is due to the fact that it measures the degree of separation between two variables regardless of the scale of those variables. A monotone transformation such as a translation will move the values of the variable on its support, but will do it for both the diseased and the healthy populations equally, maintaining the same separation between them.

Furthermore, we have to bear in mind that in all the previously discussed situations we have assumed that the diagnostic marker adopts higher values in the diseased population. If we consider a diagnostic variable with the reverse situation, the resulting ROC curve would fall below the diagonal of the unit interval. In this case, it suffices to take $-Y^F$ and $-Y^G$ to obtain an ROC curve above the diagonal. For other more complex scenarios, in which the diseased subjects could yield both the highest and the lowest values, we would have to consider making some adjustments, either by contemplating a transformation of the diagnostic variable or by using the *generalized ROC curve* (Martínez-Camblor et al., 2014), a generalization of the ROC curve to these kind of situations.

We will keep making the assumption that the diagnostic variable yields higher values



$AUC = 0.627$ $AUC = 0.862$ $AUC = 0.993$

Figure 1.3: Three different scenarios are considered here. The densities of each pair of the diagnostic variables are drawn, along with their corresponding ROC curve. The threshold of value $c = 0$ is highlighted in red, with the areas of $TPF(c)$ coloured in green and the areas of $FPF(c)$ coloured in blue. A summary measure, the Area Under the Curve (AUC), is shown for each situation.

for the diseased population throughout the rest of this dissertation.

1.1.2 Summary indices of the ROC curve

Several indices that summarize the information given by an ROC curve on a single scalar can be found in the literature. We consider some of them here:

- The *Area Under the Curve* (AUC):

  It is probably the most widely used summary index for the ROC curve and, as its own name indicates, represents the area under the ROC curve:

  $$AUC = \int_0^1 ROC(p)dp. \tag{1.2}$$

  Seeing that the ROC curve (provided that the diagnostic variables are well defined) will take values above the diagonal of the unit square, the AUC will thus take values between 0.5 and 1, where 0.5 represents random allocation and 1 the perfect classification. This can be observed in Figure 1.3, where the AUCs are displayed for each one of the situations.

- The *partial Area Under the Curve* (pAUC):

  Sometimes the interest lies only in some range of FPFs contained in $(0, 1)$. In those cases, the pAUC is computed the same way as the AUC, but on that specific range of FPFs, say (δ_1, δ_2):

  $$pAUC(\delta_1, \delta_2) = \int_{\delta_1}^{\delta_2} ROC(p)dp, \qquad \text{with } 0 \leqslant \delta_1 < \delta_2 \leqslant 1.$$

  It has a similar interpretation as the AUC.

- The *Youden index*:

  It represents the maximum difference between the TPF and the FPF:

  $$YI = \max_c\{TPF(c) - FPF(c)\} = \max_c\{sensitivity(c) + specificity(c) - 1\}.$$

  The threshold which leads to the point on the ROC curve corresponding to the Youden index is often taken to be the optimal classification threshold.

1.1.3 Parametric models

Given that the ROC curve is built from cumulative distribution functions, by assuming certain parametric models for those distribution functions we can obtain fully specified ROC curve models. This way we can obtain binormal ROC curves, bi-Weibull ROC curves, bi-exponential ROC curves, bi-gamma ROC curves, bi-logistic ROC curves... For a review on the parametric methods adopted for fitting the ROC curves, see Pundir and Amala (2014).

By far, the most used parametric model that appears in the literature is the binormal model. It can be obtained by assuming that the diagnostic variables Y^F and Y^G follow normal distributions $N(\mu^F, \sigma^F)$ and $N(\mu^G, \sigma^G)$, with $\mu^F > \mu^G$ to ensure that the diseased population is the one with higher values. The three ROC curves displayed in Figure 1.3 are binormal ROC curves, with $\mu^F = 2.1$, $\mu^G = 0$, $\sigma^F = 5$ and $\sigma^G = 4.2$ for the first one, $\mu^F = 2.5$, $\mu^G = -4$, $\sigma^F = 4.4$ and $\sigma^G = 4$ for the second one and $\mu^F = 6.5$, $\mu^G = -6$, $\sigma^F = 4$ and $\sigma^G = 3$ for the last one. The formula for the binormal ROC curve model is displayed in Table 1.2, along with its corresponding AUC.

Apart from the binormal model, Table 1.2 also gathers the details for the bi-exponential and the bi-Weibull ROC curve models, which will be used in further chapters of this dissertation. Note that the bi-exponential model (in which we assume that $\lambda^F < \lambda^G$) is a particular case of the bi-Weibull, when the scale parameters $\alpha^F = \alpha^G = 1$ and considering the shape parameters $\lambda^F = 1/\beta^F$ and $\lambda^G = 1/\beta^G$. In the case of the bi-Weibull ROC curve, the AUC does not have a closed formula.

Table 1.2: Parametric models for the ROC curve.

	Binormal	Bi-exponential	Bi-Weibull
F and G	$Y^F \sim N(\mu^F, \sigma^F)$ $Y^G \sim N(\mu^G, \sigma^G)$	$Y^F \sim Exp(\lambda^F)$ $Y^G \sim Exp(\lambda^G)$	$Y^F \sim Weibull(\alpha^F, \beta^F)$ $Y^G \sim Weibull(\alpha^G, \beta^G)$
$ROC(p)$	$\Phi\left(\frac{\mu^F-\mu^G}{\sigma^F} + \frac{\sigma^G}{\sigma^F}\Phi^{-1}(p)\right)$	$p^{\frac{\lambda^F}{\lambda^G}}$	$\exp\left\{-\left(\frac{\beta^G}{\beta^F}(-\log(p))^{\frac{1}{\alpha^G}}\right)^{\alpha^F}\right\}$
AUC	$\Phi\left(\frac{\mu^F-\mu^G}{\sqrt{(\sigma^F)^2+(\sigma^G)^2}}\right)$	$\frac{\lambda^G}{\lambda^F+\lambda^G}$	–

Note: Φ and Φ^{-1} denote the cumulative distribution and quantile functions of the standard normal.

Goodness-of-fit tests

There are not many articles in the literature devoted to designing goodness-of-fit tests for parametric models for ROC curves, and most of them are dedicated to determining whether the ROC curve follows a binormal model or not.

Given that the binormal ROC curve is obtained from normal distribution functions, one could be tempted to perform the goodness-of-fit tests on both of the diagnostic variables, using the well-known procedures to decide if they follow normal distributions or not. However, we have to take into account that having an ROC curve with a binormal model does not necessary mean that the corresponding diagnostic variables follow normal distributions. For example, if the diagnostic variables followed log-normal distribution functions, such as $Y^F \sim LN(\mu^F, \sigma^F)$ and $Y^G \sim LN(\mu^G, \sigma^G)$, the resulting ROC curve would be the same binormal model detailed in the first row of Table 1.2. This is because the ROC curves have the property of remaining invariant to monotone increasing transformations (such as the logarithm).

Goodness-of-fit tests have been proposed for categorical data (Zhou, 1995; Walsh, 1999) and for continuous data (Zou et al., 2003, 2005). In the latter a test statistic based on the AUC is developed, comparing a nonparametric estimate of the AUC with an efficient estimate of the AUC under some parametric assumption.

1.1.4 Main goals of the ROC curve studies

It has already been established that the ROC curve is used for evaluating the discriminatory capability of a diagnostic marker. However, the analysis of this curve can be used for reaching further goals.

On the one hand, given that it offers the sensitivity and specificity for all the possible thresholds, it is a widely accepted way for selecting an optimal cutoff point that best discriminates between patients with and without the disease. Thus, it can be used for designing new methods of diagnosis that minimize the classification errors.

On the other hand, when there is more than one diagnostic method for a certain disease, the comparison of the ROC curves of the corresponding methods can serve to

compare their accuracy of diagnosis, or to compare the behaviour of one diagnostic method across different groups (i.e., different hospitals, different gender, different age groups...).

Moreover, apart from the diagnostic variables that are used for the classification method, in practice it is usual to have other covariates that provide more information to the study. In those cases it is important to analyse the effect that these covariates may have on the discriminatory capability of the diagnostic method. This can also be studied by using ROC curves, adapting them so they are able to take into account this extra information.

In this dissertation we will focus our attention on the combination of the last two goals: the comparison of ROC curves with the incorporation of covariates to the analysis.

1.2 General objectives of this dissertation

The main goal of this thesis is to propose and study new tests for comparing ROC curves, either with no covariates, with a unidimensional covariate or with a multidimensional one. Seeing that the inclusion of those covariates in the ROC curve analysis could influence the conclusions drawn from those studies, a parallel line of research will be to determine a strategy to assess the significance of the covariate effect. The last objective that will be pursued here is the application of all the new methodologies in real biomedical datasets.

1.2.1 Distribution of the manuscript

Those objectives are developed throughout this document as follows:

Chapter 2: ROC curves in the presence of covariates

The first step will be to see how to include the covariate information into the ROC curve analysis and how to estimate the ROC curve and the AUC in the scenarios with and without covariates. Both aspects are detailed in Chapter 2, along with a discussion for determining when those covariates affect the discriminatory capability of the diagnostic method. It includes the design of a new test related to the significance of the covariate effect and an application to a real dataset.

The contents of this chapter will be collected in a future paper, still under preparation.

Chapter 3: Comparison without covariates

Chapter 3 contains a review of the existing techniques in the literature for comparing two or more ROC curves without covariates. It is focussed on the nonparametric tests that compare independent ROC curves, although some of the methodologies have adaptations for the case of dependent ROC curves. Special attention is given to the philosophy behind the construction of each test statistic (meaning, whether it compares the whole curve or just some summary measure like the AUC) and to the methods that are used to obtain the

distribution of such statistics. It contains a simulation study that compares the behaviour of several methodologies in different scenarios designed to highlight their strengths and weaknesses.

This chapter is mainly based on Fanjul-Hevia and González-Manteiga (2018).

Chapter 4: Comparison with unidimensional covariates

Chapter 4 is devoted to the design of a new test for comparing ROC curves conditional to a unidimensional covariate. It combines existing methods for estimating the conditional ROC curve (seen in Chapter 2) and for comparing ROC curves without covariates (seen in Chapter 3). It includes the asymptotic distribution of the proposed test statistic, as well as a bootstrap mechanism for approximating that distribution. A simulation study is carried out, although no other methods are compared to the new proposal, as we believe it to be the first of the sort for this kind of test with covariates. The proposed methodology is illustrated with an application to real data.

This chapter is mainly based on Fanjul-Hevia et al. (2020a).

Chapter 5: Comparison with multidimensional covariates

In Chapter 5 we extend the methodology proposed for the comparison of ROC curves conditioned to a unidimensional covariate to the case with a multidimensional covariate. The use of random projections provides a way of transforming the problem into a simpler one similar to the problem studied in Chapter 4. A bootstrap algorithm is designed to approximate the distribution of the test statistic proposed, and its behaviour is analysed through a simulation study. An application of the test to a dataset is also provided.

This chapter is mainly based on Fanjul-Hevia et al. (2020b).

Chapter 6: Conclusions and discussion

Finally, Chapter 6 includes some final comments and the discussion of the problems (old or new) that remain open for further research.

Appendices A and B

Appendix A contains the technical proofs for Chapters 4 and 5. Furthermore, Appendix B collects extra simulation results for the studies carried out in Chapters 4 and 5.

Additionally, a summary of this dissertation in both Spanish and Galician languages is provided at the end of the document.

1.2.2 Real datasets

The application to biomedicine of the newly designed methodologies is one of the main goals of this dissertation. An introduction of the two different datasets, *Diabetes* and

Table 1.3: A sample of the Diabetes dataset.

	prediabetes	*GA*	*a1c1*	*GP22*	*age*
1	0	12.21	4.80	5.21	46
2	0	12.21	5.00	9.14	26
3	0	12.77	5.30	5.91	54
4	1	14.36	5.80	7.24	52
5	0	9.64	5.70	4.88	76
6	1	15.18	5.30	7.16	66
⋮	⋮	⋮	⋮	⋮	⋮

Pleural Effusion –that will be employed to illustrate the different methodologies– is given below. Both datasets were provided by Dr. F. Gude Sampedro (Unidade de Epidemioloxía Clínica, Hospital Clínico Universitario de Santiago).

Diabetes

In this first dataset, the disease that will be under study is prediabetes. Here, a patient is considered as prediabetic when it presents a diagnostic of diabetes mellitus or blood glucose levels above 100 mg/dl. The dataset contains information from patients that are suspected of having this prediabetic condition. After removing a few subjects with some missing values, the remaining dataset contains a total of 1496 patients, 405 (27.1%) of whom are considered to have the disease.

Apart from the binary output that indicates if a patient has diabetes or not, and the variable *glucose* (both variables used to obtain *prediabetes*, a binary indicator of the prediabetic condition), there are other variables in the dataset, such as such as blood levels of glycated albumin (*GA*) and haemoglobin (*a1c1*), glycan peaks (*GP22*), *age*, *gender*... Some of these variables (*GA*, *a1c1* or *GP22*) will be used as diagnostic markers for the prediabetes. Others, like *gender*, will be used as a covariate that can affect the ROC curves obtained for each one of those diagnostic markers.

A sample of the datasets with the variables that will be used is on Table 1.3. This dataset is employed in Chapter 2 to illustrate the discussion about how to assess the effect of the covariates on the ROC curve analysis.

Pleural Effusion

Pleural Effussion (PE) is the build-up of excess fluid between the layers of the pleura outside the lungs. Its appearance can be due to numerous factors, but we are interested in the case in which the PE has a malignant origin (MPE), defined by the presence of malignant cells on cytology or pleural biopsy. Thus, in this case the subjects under study will be patients with PE, and the objective will be to analyse the accuracy of the diagnosis methods that are able to differentiate the MPE from the PE whose origins are not cancer-related.

Table 1.4: A sample of the Pleural Effusion dataset.

	gender	*age*	*CA125*	*CA153*	*nse*	*cyfra*	*neo*
1	1	63	81.00	5.00	1.00	11.80	0
2	0	32	1334.00	4.00	4.70	46.70	1
3	0	83	655.10	17.02	0.22	7.65	0
4	0	74	1779.00	12.00	5.40	162.30	1
5	1	22	1147.00	28.16	6.33	98.01	0
6	0	42	159.00	36.00	12.30	26.60	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The PE database is composed by the information of 480 patients on 22 different variables. The one called *neo* is the dichotomous variable that indicates if the PE is malignant or not, having a total of 211 (43.9%) patients with MPE. The other variables contain information about several tumour markers such as the carbohydrate antigen 15-3 (*ca153*, in U/ml), the cancer antigen 125 (*ca125*, in U/ml), the cytokeratin fragment 21-1 (*cyfra*, in ng/ml) or the neuron-specific enolase (*nse*, in ng/ml). It also includes clinical variables such as *gender* or *age* of the patients.

This dataset is used in Chapters 4 and 5 to illustrate the methodology for comparing ROC curves in the presence of unidimensional and multidimensional covariates. It has been used previously in Valdés et al. (2013) to assess the performance of several parameters (tumour markers and clinical-radiological criteria) in the diagnosis of MPE.

A sample of the datasets with the variables that will be employed in the studies is displayed on Table 1.4. Note that not all the variables will be used in both analyses and that, given the existence of some missing values and some pre-processing of the data to eliminate outliers, different sample sizes can be considered for the two studies.

Chapter 2

ROC curves in the presence of covariates

The ROC curve is a statistical tool that analyses the accuracy of a diagnostic test in which a variable is used to decide whether an individual is healthy or not. Along with that diagnostic variable it is usual to have information of some other covariates. In some situations it is advisable to incorporate that information into the study, as the performance of the ROC curves can be affected by them. Using the covariate-adjusted, the covariate-specific or the pooled ROC curves, in this chapter we discuss how to decide if we can exclude the covariates from our study or not, and the implications this may have in further analyses of the ROC curve. A real database is analysed to illustrate the problem.

2.1 Motivation of the study

In the previous chapter we introduced the concept of the ROC curve as a statistical tool that analyses the accuracy of a certain diagnostic test. This diagnostic test was based on the measurement of a certain diagnostic marker on the healthy and diseased population (i.e., the two groups that we hope to be able to differentiate as well as possible). We called those variables Y^F and Y^G. However, in a practical situation it is usual to have some other covariates, either continuous (such as blood pressure, age or body mass index of the patients) or discrete (such as gender, medical history, hospital where the treatment is given,...). This situation raises the question of whether this extra information should or could be included in the ROC curve analysis.

In order to answer that question let us begin by discussing two examples in which the covariate considered affects the study in two different ways. In the first one, a new diagnostic method is being used in three different hospitals (the hospital being the covariate considered). As all of them have different hospital policies, the diagnostic variables are being measured in different scales. In spite of this, all the hospitals are still obtaining the same success when discriminating the healthy and diseased patients. The corresponding

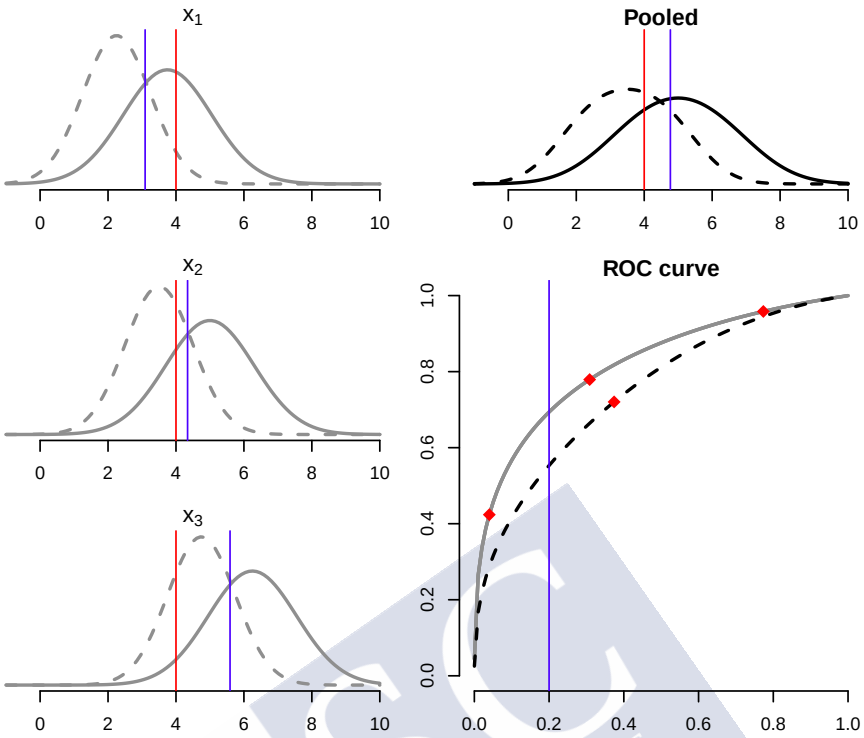


Figure 2.1: Example of a situation where the covariate affects the behaviour of the diagnostic markers, but not their discriminatory capability. In the densities, the dotted lines represent the healthy population, and the continuous lines represent the diseased population. x_1, x_2 and x_3 represent the different hospitals. The threshold $c = 4$ is highlighted in red, and the value at which a specificity of 0.8 is achieved is highlighted in blue for all the densities and ROC curves considered.

densities of the diagnostic variables considered in each hospital are depicted in Figure 2.1 (to the left), along with their corresponding ROC curve (which is the same for the three cases). This is because of one of the properties of the ROC curve mentioned in the previous chapter: the curve remains unchanged when its classification variables undergo a monotone increasing transformation (such as the translation that occurs on this example, due to the use of different scales).

However, if we dismiss the fact that the measurements of the diagnostic markers are coming from different hospitals (i.e., omitting the covariate information and taking the pooled data sample) the resulting ROC curve does change. Note that the corresponding densities of the diagnostic variables of the pooled data (also in Figure 2.1, on the top right) have different shapes than the others.

It is worth mentioning that even though the ROC curves in this example are the same for all the values of the covariate, the threshold that gives a certain pair of values of sensitivity and specificity could not be necessarily the same for every hospital. In Figure 2.1 we have highlighted in red a certain threshold of the diagnostic variable (in particular,

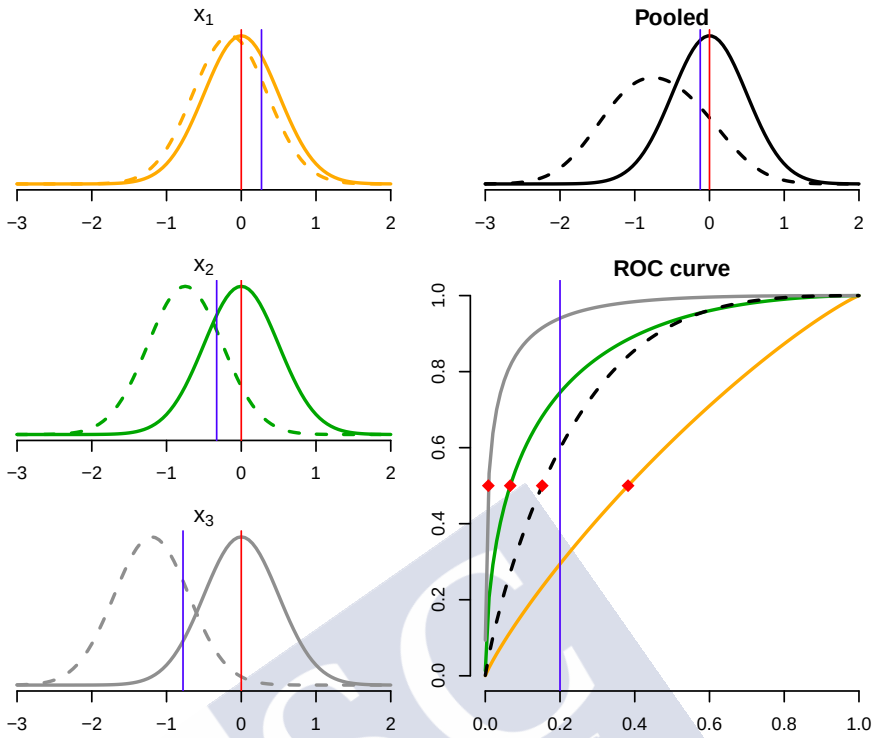


Figure 2.2: Example of a situation in which both the diagnostic variables and the ROC curve are affected by a covariate, obtaining also a different ROC curve when taking the pooled data. The threshold $c = 0$ is highlighted in red, and the value at which a specificity of 0.8 is achieved is highlighted in blue for all the densities and ROC curves considered.

the point $c = 4$) and the corresponding pair of $(1 - specificity(c), sensitivity(c))$ on their ROC curve. The pairs are different for each case. On the other hand, we have also highlighted in blue a certain value of specificity (in particular, 0.8, which means that $1 - specificity(c)$ is 0.2) and the thresholds associated with it (also indicated in blue) are different in each case.

For the second example we consider another disease for which we use a certain variable of diagnosis. This time, however, the covariate at hand is the age of the patients, aggregated in three groups. This example is represented in Figure 2.2, where the densities of all the diagnostic variables and their corresponding ROC curves are different. The diagnostic marker in the diseased population remains unchanged by the covariate, but the densities associated with the healthy populations change, increasing or decreasing the distance to the densities related to the diseased population. Not only that, but when dismissing the knowledge of that covariate and taking the pooled data (densities shown on the top right of Figure 2.2), we obtain yet another different ROC curve.

We have once again highlighted in red a certain threshold ($c = 0$), and its translation to each ROC curve. Note that, as the diagnostic variable is not affected by the covariate, the sensitivity associated to that threshold is the same in all the cases. The specificity of

0.8 is also marked in blue.

We may further complicate the scenarios if we do not assume that the covariate has the same distribution in the healthy and in the diseased population (as shown in the previous examples). For some other examples, check Pardo-Fernández et al. (2014).

In the light of the last examples, it is clear that the presence of a covariate can affect the study of the discriminatory capability of a certain diagnostic method, either by having an effect on the diagnostic variables or on the ROC curve itself. This means that all the possible applications that the study of ROC curves may have (from searching for optimal cutoff points to comparing different diagnostic markers) are as well sensitive to the presence of covariates.

Apart from the strategy of ignoring any covariate information that we may have (which means using the pooled data to conform the *pooled ROC curve*), there are two ways of modelling the effect of a covariate in an ROC curve: using the *covariate-specific* or *conditional ROC curve* or using the *covariate-adjusted ROC curve*.

In Section 2.2 of this chapter we will define and study these curves, commenting some of their properties and seeing how to estimate them. Then, in Section 2.3 we will discuss how to determine whether we should incorporate a covariate to our ROC analysis by studying the relationship between these three curves, including the design of a new methodology for testing the equality of the covariate-adjusted ROC curve and the pooled ROC curve. This is followed by a real-data application for illustration purposes in which we use one of the datasets described in Section 1.2.2. We finish the chapter by discussing, in Section 2.4, how this test for covariate effect can be adapted when considering more than one ROC curve (which brings us back to one of the main objectives of this dissertation: the comparison of ROC curves).

2.2 The three curves

Let Y^F and Y^G be the continuous diagnostic markers in the diseased and healthy population, respectively. Let X be a continuous unidimensional covariate, although the definitions introduced in this section hold for a d-dimensional covariate $\boldsymbol{X}$. X^F will represent the covariate in the diseased population, and X^G in the healthy population. R_X will denote the intersection of R_{X^F} and R_{X^G} (the supports of X^F and X^G, respectively), and is assumed to be non-empty. Furthermore, let $F(y) = P(Y^F \leqslant y)$, $F(y|x) = P(Y^F \leqslant y|X^F = x)$, $G(y) = P(Y^G \leqslant y)$, $G(y|x) = P(Y^G \leqslant y|X^G = x)$ and $F^X(x) = P(X^F \leqslant x)$.

In this section we will discuss the three curves that can be used when dealing with a diagnostic problem with covariate information: the *pooled*, the *conditional* and the *covariate-adjusted* ROC curves. Later we will show how these curves can be viewed as cumulative distribution functions. Next we will introduce some of the summary indices that can be drawn from them and finally we will discuss some methods for their estimation.

Although the definition of these curves hold for the case in which the considered covari-

ate is discrete (as it was in the examples mentioned in the previous section), throughout the rest of this dissertation we will consider the covariate to be continuous.

2.2.1 The pooled ROC curve

The pooled ROC curve is the same curve that was defined in (1.1) in the previous chapter. We added the term *pooled* in this context with covariates to emphasize the fact that, by using all the pooled data to build this kind of curve, we are disregarding the effect that this covariate may have.

One interesting property of this curve is the fact that it can be viewed as a cumulative distribution function of some *placement value*:

$$\begin{aligned} ROC(p) &= 1 - F(G^{-1}(1-p)) = P(Y^F > G^{-1}(1-p)) = P(G(Y^F) > 1-p) \\ &= P(1 - G(Y^F) < p),\ p \in (0,1). \end{aligned} \tag{2.1}$$

Thus, $R = 1 - G(Y^F)$ is a random variable that has the ROC curve as its cumulative distribution function. This can be useful for further analysis, as we can take advantages of the existing techniques for dealing with cumulative distribution functions, adjusting them for the situation at hand.

The summary measure of the ROC curve that we will be using the most throughout this dissertation is the already mentioned AUC (1.2). A useful probabilistic interpretation of this measure can be given when Y^F and Y^G are independent: the AUC can be seen as the probability that tests results from a randomly selected pair of diseased and healthy subjects are correctly ordered (meaning $Y^F > Y^G$):

$$\begin{aligned} AUC &= \int_0^1 ROC(p)dp = \int_0^1 (1 - F(G^{-1}(1-p))dp = \int_{-\infty}^{\infty} (1 - F(t))g(t)dt \\ &= \int_{-\infty}^{\infty} \left(\int_t^{\infty} f(t')dt' \right) g(t)dt = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(t' > t) f(t')g(t)dt'dt \\ &= P(Y^F > Y^G), \end{aligned} \tag{2.2}$$

where $f(\cdot)$ and $g(\cdot)$ are the density functions of Y^F and Y^G, respectively.

Estimation

In practice, the ROC curve must be estimated, as F and G are unknown. In general, we will have two samples, $\{Y_1^F, ..., Y_{n^F}^F\}$ and $\{Y_1^G, ..., Y_{n^G}^G\}$ from the diseased and healthy populations, respectively. The estimation of the ROC curve has been intensively discussed in the literature in the last few years. A review of the existing methodologies can be seen in Gonçalves et al. (2014). This review contains frequentist and Bayesian methods. In this dissertation we will focuss on the frequentist approach.

We summarize here some of the most commonly used estimators:

- *The parametric estimator*:

  If we assume that F and G belong to given parametric families, we can employ parametric estimators to obtain the ROC curve estimator. In particular, if we assume that the ROC curve fits a binormal model (meaning that the diagnostic variables follow a normal distribution, or can be converted to normal by a Box-Cox transformation such as the logarithm) as we saw in Section 1.1.3, then

  $$\widehat{ROC}_p(p) = \Phi(\hat{a} + \hat{b}\Phi^{-1}(p)),\ p \in (0,1),$$

  with $\hat{a}$ and $\hat{b}$ being the Maximum Likelihood Estimators of the parameters $a = (\mu^F - \mu^G)/\sigma^F$ and $b = \sigma^G/\sigma^F$.

  The corresponding parametric estimation of the AUC will be, thus

  $$\widehat{AUC}_p = \Phi\left(\hat{a}(1+\hat{b}^2)^{-1/2}\right).$$

  Of course, it is more usual for the distribution of the diagnostic variables to be unknown, so we must relay on nonparametric estimators.

- *The empirical estimator*:

  It is the simplest nonparametric estimator: it consists on plugging the empirical estimates of F and G on (1.1), obtaining

  $$\widehat{ROC}(p) = 1 - \hat{F}(\hat{G}^{-1}(1-p)),\ p \in (0,1), \tag{2.3}$$

  where $\hat{F}(t) = (n^F)^{-1}\sum_{i=1}^{n^F} I(Y_i^F \leqslant t)$, $\hat{G}(t) = (n^G)^{-1}\sum_{i=1}^{n^G} I(Y_i^G \leqslant t)$ are the empirical distribution functions and $\hat{G}^{-1}(p) = \inf\{t : \hat{G}(t) \geqslant p\}$ is the empirical quantile distribution function.

  This estimator has many good properties: it is, under some basic assumptions for F and G, uniformly convergent to the theoretical curve (Hsieh and Turnbull, 1996) and it is invariant under an increasing monotone transformation of the data. Its main disadvantage is that it is not a continuous estimator.

  The corresponding AUC of the empirical ROC curve, based on the interpretation showed in (2.2), is similar to the Mann-Whitney statistic:

  $$\widehat{AUC} = \frac{1}{n^F n^G}\sum_{j=1}^{n^G}\sum_{i=1}^{n^F}\left(I(Y_i^F > Y_j^G) + \frac{1}{2}I(Y_i^F = Y_j^G)\right). \tag{2.4}$$

- *The smoothed estimator*:

  To solve the lack of smoothness of the empirical estimator, Zou et al. (1997) proposed the use of kernel estimation methods (Silverman, 1986)[1] to estimate the distributions underneath the ROC curve, obtaining

$$\widetilde{ROC}_{s_1}(p) = 1 - \tilde{F}(\tilde{G}^{-1}(1-p)),\ p \in (0,1), \tag{2.5}$$

  where $\tilde{F}(t)$ and $\tilde{G}(t)$ are estimated through smooth versions of the empirical distribution function. As it happens with kernel estimators, the kernel functions employed are relatively unimportant, but the selection of the bandwidth parameters (one for each distribution) is far from straightforward (Lloyd, 1998; Lloyd and Yong, 1999; Jokiel-Rokita and Pulit, 2013). Also, this estimator is not invariant under a monotone transformation of the data, and can be unreliable at boundaries of the ROC curve. Peng and Zhou (2004) gave an alternative smooth estimator that solves both problems, applying the smoothing directly on the ROC curve and using the empirical cumulative distribution functions of F and G

$$\widetilde{ROC}_{s_2}(p) = 1 - \int \hat{F}(\hat{G}^{-1}(1-p+hu))\kappa(u)du,\ p \in (0,1). \tag{2.6}$$

  More recently, Pulit (2016) proposed a smoothed estimator for the ROC curve which is also invariant under non-decreasing data transformations, using the unobserved samples of the variable R discussed in (2.1) to estimate a kernel distribution function.

  The AUC can be also estimated using these kernel-based estimators. For example, Lloyd (1998) proposed, using a Gaussian kernel,

$$\widetilde{AUC}_s = \sum_{i=1}^{n^F}\sum_{j=1}^{n^G} \Phi\left(\frac{Y_j^G - Y_i^F}{\sqrt{h_F^2 + h_G^2}}\right),$$

  where h_F and h_G are the bandwidth parameters of each kernel-estimated cumulative distribution function.

Among these estimators, in this dissertation we will be using the empirical one.

[1] Let $X_1, ..., X_n$ be a sample from a variable X, with density function f and cumulative distribution function F. The kernel estimators of f and F are

$$\hat{f}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n}\kappa\left(\frac{x-X_i}{h}\right), \quad \tilde{F}_{n,h}(x) = \int_{-\infty}^{x}\hat{f}_{n,h}(t)dt = \frac{1}{n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) = \int_{-\infty}^{\infty}\hat{F}(x+hu)\kappa(u)du,$$

respectively, where h is a bandwidth parameter and κ is the kernel (typically, a density), and $K(u) = \int_{-\infty}^{u}\kappa(t)dt$.

2.2.2 The conditional ROC curve

The most usual way of introducing the effect of the covariate on the ROC curve analysis is by using the *covariate-specific* or *conditional ROC curve*. For a fixed value of the covariate $x \in R_X$, the conditional ROC curve is defined as

$$ROC^x(p) = 1 - F(G^{-1}(1-p|x)|x),\ p \in (0,1), \tag{2.7}$$

where $F(\cdot|x)$ and $G(\cdot|x)$ are the cumulative distribution functions of the diagnostic variable in the diseased and healthy population, respectively, conditioned to the value x. Note that its structure is very similar to the pooled ROC curve, except that the distribution functions are now conditioned to the value of the covariate x.

As we did in (2.1), given $x \in R_X$, the conditional ROC curve in that point can be viewed as a cumulative distribution function:

$$\begin{aligned} ROC^x(p) &= 1 - F(G^{-1}(1-p|x)|x) = P(Y^F > G^{-1}(1-p|x)|X^F = x) \\ &= P(G(Y^F|x) > 1-p|X^F = x) = P(1 - G(Y^F|x) \leqslant p|X^F = x),\ p \in (0,1). \end{aligned}$$

The conditional version of the AUC, called the *covariate-specific* or *conditional AUC* (AUC^x), for a fixed $x \in R_X$ is defined as:

$$AUC^x = \int_0^1 ROC^x(p)dp. \tag{2.8}$$

Following similar steps to the ones in (2.2) (and assuming that Y^F and Y^G are independent conditional to $X^F = X^G = x$), for a fixed $x \in R_X$, $AUC^x = P(Y^F > Y^G|X^F = x, X^G = x)$. It also takes values between 0.5 and 1, with the same interpretation as before.

Estimation

There are different approaches that can be found in the literature regarding the estimation of the conditional ROC curve. Some of the methods estimate directly the conditional distribution functions (López-de Ullibarri et al., 2008; Inácio de Carvalho et al., 2013), while others introduce the covariate effect through some regression models. The latter are estimators based either on direct (Alonzo and Pepe, 2002; Rodríguez-Álvarez et al., 2011a) or on induced (González-Manteiga et al., 2011; Rodríguez-Álvarez et al., 2011b) regression methodologies. These and other methods of estimation of conditional ROC curves are reviewed in Pardo-Fernández et al. (2014).

Moreover, there are other authors like Brumback et al. (2006) or Yao et al. (2010) that estimate directly the conditional AUC (2.8).

We focus now on the induced-regression methodology to estimate a conditional ROC curve, particularly on the estimator proposed in González-Manteiga et al. (2011), as it is the approach that we will be using later on. This approach is based on the nonparametric

location-scale regression models

$$\begin{aligned} Y^F &= \mu^F(X^F) + \sigma^F(X^F)\varepsilon^F, \\ Y^G &= \mu^G(X^G) + \sigma^G(X^G)\varepsilon^G, \end{aligned} \tag{2.9}$$

where, for $D \in \{F, G\}$, X^D is the covariate associated with Y^D, $\mu^D(\cdot) = \mathbb{E}(Y^D|X^D = \cdot)$ and $(\sigma^D)^2(\cdot) = \mathbb{V}\text{ar}(Y^D|X^D = \cdot)$ are the conditional mean and the conditional variance functions (both of them unknown smooth functions), and the error ε^D is independent of X^D and has cumulative distribution function H^D. As the healthy and the diseased populations are assumed to be independent, ε^F and ε^G are assumed to be independent as well.

The conditional ROC curve defined in (2.7) can be expressed in terms of those error cumulative distribution functions, H^F and H^G. Using the fact that $F(t|x) = H^F(\{t - \mu^F(x)\}/\sigma^G(x))$ and $G^{-1}(p|x) = \mu^G(x) + \sigma^G(x)(H^G)^{-1}(p)$, it is easy to see that

$$\begin{aligned} ROC^x(p) &= 1 - H^F\left(\frac{(H^G)^{-1}(1-p)\sigma^G(x) + \mu^G(x) - \mu^F(x)}{\sigma^F(x)}\right) \\ &= 1 - H^F\left((H^G)^{-1}(1-p)b(x) - a(x)\right), \ p \in (0,1), \end{aligned}$$

where $a(x) = (\mu^F(x) - \mu^G(x))/\sigma^F(x)$ and $b(x) = \sigma^G(x)/\sigma^F(x)$. The advantage of this alternative way of defining the conditional ROC curve is that now we have an expression that does not depend on conditional probabilities.

Now, let $\{(X_i^F, Y_i^F)\}_{i=1}^{n^F}$ be an i.i.d. sample from the distribution of (X^F, Y^F) and $\{(X_i^G, Y_i^G)\}_{i=1}^{n^G}$ an i.i.d. sample from the distribution of (X^G, Y^G). The following estimator of the conditional ROC curve is proposed:

$$\widehat{ROC}^x(p) = 1 - \hat{H}^F\left((\hat{H}^G)^{-1}(1-p)\hat{b}(x) - \hat{a}(x)\right), \ p \in (0,1), \tag{2.10}$$

where, for $D \in \{F, G\}$,

- $\hat{H}^D(y) = (n^D)^{-1}\sum_{i=1}^{n^D} I(\hat{\varepsilon}_i^D \leqslant y)$,
- $\hat{\varepsilon}_i^D = \frac{Y_i^D - \hat{\mu}^D(X_i^D)}{\hat{\sigma}^D(X_i^D)}$ for $i \in \{1, \cdots, n^D\}$,
- $\hat{\mu}^D(x) = \sum_{i=1}^{n_D} W_i^D(x, g^D)Y_i^D$ is a nonparametric estimator of $\mu^D(x)$ based on local weights $W_i^D(x, g^D)$ depending on a bandwidth parameter g^D,
- $(\hat{\sigma}^D)^2(x) = \sum_{i=1}^{n^D} W_i^D(x, g^D)[Y_i^D - \hat{\mu}^D(X_i^D)]^2$ is a nonparametric estimator of $(\sigma^D)^2(x)$. For simplicity we take the same bandwidth parameter g^D that is used for the estimation of the regression function $\hat{\mu}^D(x)$,
- $W_i^D(x, g^D) = \frac{\kappa_{g^D}(x - X_i^D)}{\sum_{l=1}^{n^D}\kappa_{g^D}(x - X_l^D)}$, for $i \in \{1, \cdots, n^D\}$, are Nadaraya-Watson-type weights, where $\kappa_{g^D}(\cdot) = \kappa(\cdot/g^D)/g^D$ and κ is the kernel (typically, a probability density function).

The estimator in (2.10) is of an empirical type. The continuous nature of the ROC curve though motivates the construction of a continuous estimator. Imitating the smoothing done in (2.6) by Peng and Zhou (2004) for the unconditional case, the following continuous estimator for the conditional ROC curve is proposed:

$$\widehat{ROC}^x(p) = 1 - \int \hat{H}^F \left((\hat{H}^G)^{-1}(1-p+hu)\hat{b}(x) - \hat{a}(x) \right) \kappa(u)du, \ p \in (0,1), \quad (2.11)$$

where h is a bandwidth parameter. The study carried out in González-Manteiga et al. (2011) showed that its effect is not very important, although the introduction of a small amount of smoothing produces better behaviour in terms of Mean Squared Error with respect the empirical estimator (2.10). This estimator can also be written as:

$$\widehat{ROC}^x(p) = \frac{1}{n^F} \sum_{i=1}^{n^F} K_h \left(\hat{H}^G \left(\{\hat{\varepsilon}_i^F + \hat{a}(x)\}/\hat{b}(x) \right) - 1 + p \right), \ p \in (0,1),$$

where $K_h(\cdot) = K(\cdot/h)$ and K is the distribution function corresponding to κ.

Moreover, an estimator for the conditional AUC (2.8) would be:

$$\widehat{AUC}^x = \int_0^1 \widehat{ROC}^x(p)dp. \quad (2.12)$$

In González-Manteiga et al. (2011) (and in Rodríguez-Álvarez et al., 2011b) they also propose a pointwise bootstrap confidence intervals for the conditional AUC based on the percentile method. For a fixed $x \in R_X$, and for $b \in \{1, \cdots, B\}$ (with B large):

1. For $D \in \{F, G\}$, let $\{\varepsilon_i^{D,b*}\}_{i=1}^{n^D}$ be an i.i.d. sample from $\hat{H}^D$.
2. Generate, for $D \in \{F, G\}$, the bootstrap samples $\{(X_i^D, Y_i^{D,b*})\}_{i=1}^{n^D}$, with $Y_i^{D,b*} = \hat{\mu}^D(X_i^D) + \hat{\sigma}^D(X_i^D)\varepsilon_i^{D,b*}$.
3. From $\{(X_i^D, Y_i^{D,b*})\}_{i=1}^{n^D}$, with $D \in \{F, G\}$, repeat the estimation process to obtain $\widehat{ROC}^{x,b*}(p)$ for $p \in (0,1)$. Then, obtain $\widehat{AUC}^{x,b*}$ using (2.12).

Let $\widehat{AUC}^{x,(b)*}$ be the order statistics of the values $\widehat{AUC}^{x,1*}, \cdots, \widehat{AUC}^{x,B*}$ obtained at the last step. The bootstrap confidence interval, for a confidence level of $1-\alpha$, is given by $\widehat{AUC}^{x,([B\alpha/2])*}, \widehat{AUC}^{x,([B(1-\alpha/2)])*}$ (where $\lfloor\cdot\rfloor$ denotes the integer part).

Note that there are also methods for computing confidence intervals for the AUC obtained from the pooled data. However, we only go into detail with the confidence interval of AUC^x because we will be using it in the application with real data.

2.2.3 The covariate-adjusted ROC curve

The *covariate-adjusted ROC curve* (AROC curve) was first introduced by Janes and Pepe (2009). They defined the AROC curve as the ROC curve that is obtained when the

thresholds used for the classification are covariate-specific. This means that the thresholds are chosen to ensure that the covariate-specific (or conditional) FPF is common across all the values of the covariate. Mathematically, it is defined as

$$AROC(p) = P(Y^F > G^{-1}(1-p|X^F)) = 1 - F(G^{-1}(1-p|X^F)),\ p \in (0,1). \quad (2.13)$$

Note that the threshold equal to the quantile $G^{-1}(1-p|X^F)$ yields a FPF of p in the covariate-specific population.

The AROC curve can also be viewed as a weighted average of conditional ROC curves (the weight depending on the distribution of X^F):

$$AROC(p) = \int ROC^x(p) dF^X(x),\ p \in (0,1).$$

This means that it could be particularly useful for giving a summary of the performance of the conditional ROC curve when sample sizes are not large enough to the conditional ROC curve to be estimated accurately.

Equivalently, $AROC(p) = \mathbb{E}(ROC^{X^F}(p))$ (where the expectation is taken with respect to X^F). This means that when the covariate affects the diagnostic variables but not their discriminatory accuracy (i.e., when the performance of the diagnostic marker is the same across populations with different values of the covariate), the AROC curve coincides with the conditional ROC curve (we will see more on this matter in the next section).

On the other hand, this function can also be expressed as a cumulative distribution function:

$$\begin{aligned} AROC(p) &= P(Y^F > G^{-1}(1-p|X^F)) = P(G(Y^F|X^F) > 1-p) \\ &= P(1 - G(Y^F|X^F) \leqslant p) \quad \forall p \in (0,1). \end{aligned} \quad (2.14)$$

Moreover, if we take into account the location-scale regression model considered previously (2.9) to model the conditional ROC curve we could write the AROC curve without conditional distributions:

$$AROC(p) = 1 - F\left(\mu^G(X^F) + \sigma^G(X^F)(H^G)^{-1}(1-p)\right),\ p \in (0,1).$$

The *covariate-adjusted AUC* (AAUC) is defined as a summary index of the AROC curve:

$$AAUC = \int_0^1 AROC(p) dp.$$

We can also consider the covariate-adjusted AUC as a weighted average of the conditional AUC:

$$AAUC = \int_0^1 AROC(p)dp = \int_0^1 \int ROC^x(p)dF^X(x)dp = \int \int_0^1 ROC^x(p)dpdF^X(x)$$
$$= \int AUC^x dF^X(x).$$

Estimation

The samples that we would have in this case are the same ones needed for the estimation of the conditional ROC curve. Janes et al. (2009) and Janes and Pepe (2009) use the definition of the AROC given in (2.13) to propose an estimator of the form

$$\widehat{AROC}(p) = \frac{1}{n^F} \sum_{i=1}^{n^F} I\left(Y_i^F > \hat{G}^{-1}(1-p|X_i^F)\right),$$

where $\hat{G}^{-1}(1-p|X_i^F)$ needs to be estimated. They propose both semiparametic and nonparametric alternatives, but we are more interested in the one proposed by Rodríguez-Álvarez et al. (2011b), who also presented a location-scale regression model (2.9) for the accommodation of the covariate in the ROC curve analysis. They expressed the conditional quantile in terms of the error distribution ($G^{-1}(p|x) = \mu^G(x) + \sigma^G(x)(H^G)^{-1}(p)$, with $p \in (0,1)$), obtaining the estimator:

$$\widehat{AROC}(p) = \frac{1}{n^F} \sum_{i=1}^{n^F} I\left(\frac{Y_i^F - \hat{\mu}^G(X_i^F)}{\hat{\sigma}^G(X_i^F)} > \hat{G}^{-1}(1-p)\right), \ p \in (0,1), \tag{2.15}$$

where $\hat{\mu}^G(\cdot)$ and $\hat{\sigma}^G(\cdot)$ are estimated using local polynomial kernel smoothers. In the application with the real data-set that appears in Section 2.3.3, the AROC curve was estimated using this formula, but with the kernel-type regression estimators for the mean and for the standard deviation detailed in the previous section. More recently, Inácio de Carvalho and Rodríguez-Álvarez (2018) presented a nonparametric Bayesian approach.

As an estimator of the AAUC we will be using

$$\widehat{AAUC} = \int_0^1 \widehat{AROC}(p)dp.$$

2.3 Significance of the covariate effect

Now that we have seen the three curves that can be used in this context we are going to study the relationships that can be established between them, and what can be learned from that. We have to determine whether the covariate affects the behaviour of the discriminatory capability and then, even when it does not, we have to decide if this

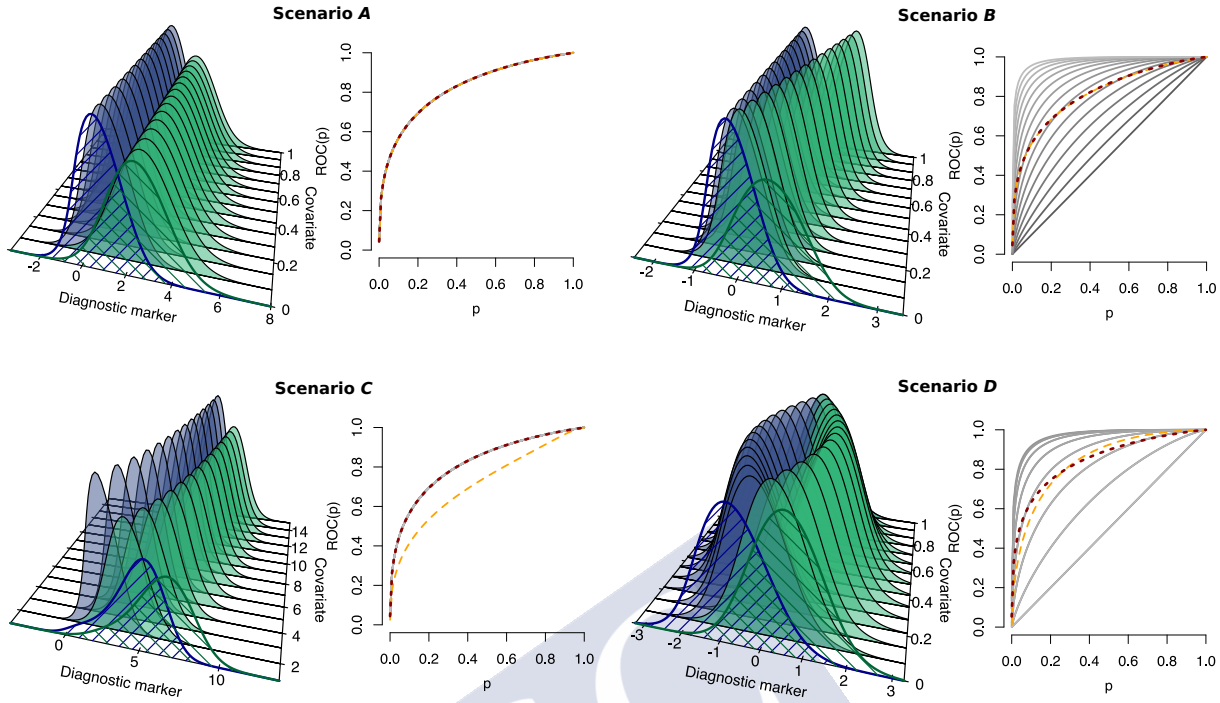


Figure 2.3: Four scenarios with different relationships between the pooled, conditional and covariate-adjusted ROC curves. For each one of them, the conditional densities of the diagnostic variables (blue for the healthy population and green for the diseased population) are depicted for several fixed values of the covariate. The striped densities in the front row represent the densities for the marginal diagnostic markers. The corresponding ROC and AROC curves are drawn for each case in a discontinuous orange and brown line, respectively. The conditional ROC curves are represented for every value of the covariate whose conditional densities are also represented.

covariate has an effect on the diagnostic variables or if it can be directly disregarded.

In order to better illustrate our aim here, let us introduce four scenarios (A, B, C and D), represented in Figure 2.3. Each one of these scenarios is affected differently by a covariate, producing different relationships between the pooled (in orange), conditional (in gray) and covariate-adjusted ROC curves (in brown). Along with those curves, in Figure 2.3 we have drawn the densities of the diagnostic variables for the diseased (in green) and the healthy (in blue) populations, in a similar way to what we did in the examples on Figures 2.1 and 2.2. Apart from representing the densities of the pooled diagnostic samples (the ones with the striped areas) we selected several values of the covariate and draw their corresponding conditional densities (those same values are the same selected for the representation of the conditional ROC curves). The location-scale models assumed for the construction of these scenarios will be specified later on in Section 2.3.2 in Table 2.1.

It is precisely the representation of those conditional densities what indicates when the covariate has an effect on the diagnostic variables. Scenario A is the only one in which those conditional densities remain the same regardless the value of the covariate (and

coincide with the marginal densities as well). In this case, it is obvious that the pooled, conditional and covariate-adjusted ROC curves have to be equal.

In scenario B, the situation changes: although the distribution of the diagnostic variable in the healthy population remains unchanged by the covariate, the same does not apply to the diseased population. Thus, for lower values of the covariate, the conditional densities overlap with each other almost completely, whereas this overlapping is reduced when the covariate increases, separating the conditional densities. This translates into conditional ROC curves that are very close to the diagonal for the lower values and very close to the point of maximum specificity and sensitivity for the highest values. The pooled ROC and the AROC curves coincide in this situation.

Scenario C shows a more curious situation: we have diagnostic variables that are affected by the covariate, but this effect is such that the discriminatory capability remains constant throughout all the values of the covariate. This means that the conditional ROC curves are equal, and that they match the AROC curve. In fact, if we calculate (following the models on Table 2.1 on page 31) the expression of this conditional ROC curve, for a certain $x \in R_X$, we obtain $ROC_C^x(p) = 1 - \Phi\left(\frac{10}{13}\left(\Phi^{-1}(1-p) - \frac{3}{2}\right)\right)$ for $p \in (0,1)$, which is independent of the value of x. However, the effect of the covariate is noted when representing the pooled ROC curve, as it is attenuated with respect to the other two curves. In a practical situation this means that if we disregard the effect of this particular covariate, the performance of the diagnostic method would be compromised.

The last scenario, D, shows a situation in which the three curves are different. This time, both the lower and the higher values of the covariates produce conditional ROC curves close to the diagonal, whereas the medium values procure a wider separation of the corresponding conditional densities.

In these examples (and particularly in scenarios B and D) we can observe the interpretation of the AROC curve as a vertical average of the conditional ROC curves at each FPF p. Note, however, that this average has to take into consideration the distribution of the covariate, which is not reflected in any way in Figure 2.3 (the covariate values chosen to condition the densities and the ROC curves where selected uniformly on the support of the covariate).

Now that we have seen examples for all the different configurations of the three curves, what strategy must be followed in order to decide which one should be employed in the ROC curve analysis?

2.3.1 Tests for assessing the covariate effect

Three different situations can arise when dealing with ROC curves with covariates: in the first one the performance of the ROC curve changes with the value of the covariates (and with it, its discriminatory capability); in the second, the covariates affect the distribution of the diagnostic markers, but not their discriminatory capability; in the last one, the covariates do not affect the ROC curve in any way. Deciding the situation we have at hand is a two-step problem.

The first step should be to test whether the conditional ROC curve is constant for each value of the covariate $x \in R_X$, meaning $ROC^z(p) = ROC^x(p)\ \forall x \in R_X$, with $p \in (0,1)$, for a certain fixed value $z \in R_X$. In this case, this ROC^z would coincide with the covariate-adjusted ROC curve, given that it is a weighted average of the conditional ROC curves. Thus, we would be interested in testing:

$$H_0^1 : ROC^x(p) = AROC(p),\ p \in (0,1)\ \forall x \in R_X, \tag{2.16}$$

versus the general alternative $H_1^1 : H_0^1$ is not true.

If this null hypothesis were to be rejected, we would be in a scenario where the discriminatory capability of the diagnostic marker is affected by the covariate. In this case we should use the conditional ROC curve for further analysis. Going back to the scenarios represented in Figure 2.3, this would be the case for the two that are in the second column, B and D. The scenarios in the first column (A and C), though, satisfy H_0^1.

Otherwise, one could think about eliminating the covariate from the analysis, but accepting that $ROC^x = AROC$ for any x does not necessarily mean that the pooled ROC curve is going to coincide with the AROC curve. Thus, one should make this test:

$$H_0^2 : AROC(p) = ROC(p),\ p \in (0,1), \tag{2.17}$$

versus the general alternative $H_1^2 : H_0^2$ is not true.

If this hypothesis were rejected, the AROC curve should be considered. This is the case for the scenarios represented in the second row of Figure 2.3, C and D, whereas scenarios A and B do satisfy H_0^2. Of course, in the case of B and D we should not be applying this test, as H_0^1 would have been already rejected. Looking closer at scenario B, and taking into account the models used for its construction (see Table 2.1), note that the diagnostic variable Y^G is independent of the covariate. This implies that $G(y|x)=G(y)$ and, thus, $G(Y^F|X^F) \overset{d}{=} G(Y^F)$. Considering that equivalence of distribution functions and the alternative ways of writing both the pooled and the covariate-adjusted curves mentioned at (2.1) and (2.14), we can conclude that, in those conditions, H_0^2 holds. Thus, having the diagnostic variable of the healthy population being independent of the covariate at hand is a sufficient condition to satisfy H_0^2.

Only if both of the above mentioned hypotheses, H_0^1 and H_0^2, hold can we consider removing the covariates from the analysis (using, thus, the pooled ROC curve).

In Figure 2.4 we summarize the strategy that can be followed to decide whether the covariate effect is significant in an ROC curve study with covariates. Following that scheme in the examples already discussed, scenarios B and D would not satisfy H_0^1, and thus, in their case we should employ conditional ROC curves for further analyses (regardless H_0^2 holds for them or not). Scenario C, though, satisfies H_0^1, but not H_0^2, so the AROC curve should be used in this case. Only scenario A satisfies both hypotheses and thus, is the only one in which the use of covariates could be disregarded from the study.

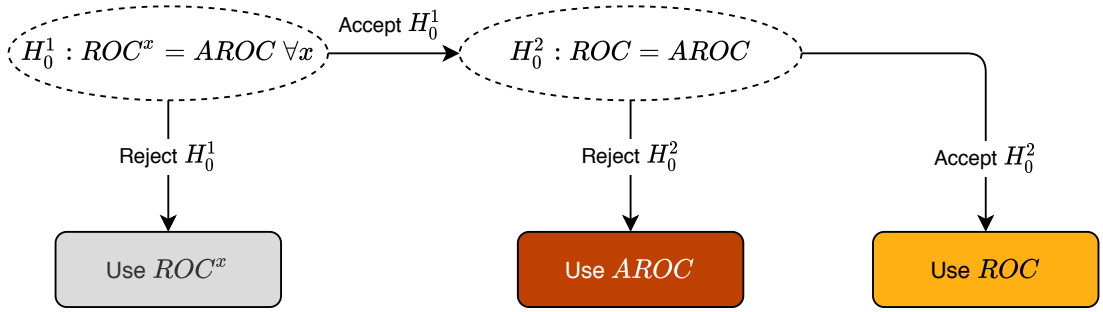


Figure 2.4: Scheme that summarizes the strategy that should be employed to model the ROC curve when there are covariates involved in the study.

One could wonder if we could test directly whether the pooled and the conditional ROC curves are equal or not for all possible values of the covariate. However, in a test like that the null hypothesis would only be satisfied with scenarios like A, whereas for the rest of the examples it would be rejected. Nevertheless, this means that scenarios similar to C would also be rejected, and it could seem that the covariate has some effect on the discriminatory capability of the test. Thus, this kind of test would push us to use the conditional ROC curve when it is not really necessary. The bottom line of this other possible test is that if you want to determine the relationships between the three curves you would still be needing to perform further tests.

There are not many references regarding these problems in the literature. Rodríguez-Álvarez et al. (2011b) propose a test for dealing with the first test (2.16) in a case with a unidimensional covariate and Rodríguez-Álvarez et al. (2018) propose an inferential procedure for testing the effect of covariates over the conditional ROC curve employing generalized additive models, using two different bootstrap-based tests to check the possible effect of the continuous covariates on the ROC curve and the presence of factor-by-curve interaction terms. In the next section we propose a new methodology to test the second hypothesis (2.17), focusing in the case with only one covariate.

2.3.2 Test ROC vs AROC

The objective in this section is to propose a nonparametric test to decide whether the covariate at hand has an effect on the performance of the diagnostic variable or not. In other words, the aim is to test whether the AROC curve coincides with the ROC curve built with the pooled data, i.e.:

$$H_0^2 : AROC(p) = ROC(p), \quad \text{for all} \quad p \in (0,1),$$

versus

$$H_1^2 : AROC(p) \neq ROC(p), \quad \text{for some } p \in (0,1).$$

Several approaches could be followed to handle this problem. We could compare (as it is commonly done in the literature) some summary measures of those curves, such as the AUC and the AAUC. The problem with that approach is that, even if we were able to prove that the summary measures were equal, this does not necessarily mean that their corresponding curves are equal (although the converse implication is true).

For our approach we are going to be comparing the whole curves, taking advantage of the estimators of the ROC and AROC curves discussed in Section 2.2 in (2.3) and (2.15), respectively. We consider the statistic

$$S_\psi = \psi\left(\widehat{ROC}(p) - \widehat{MROC}(p)\right) + \psi\left(\widehat{AROC}(p) - \widehat{MROC}(p)\right), \tag{2.18}$$

where $\widehat{MROC}(p) = \frac{\widehat{ROC}(p)+\widehat{AROC}(p)}{2}$ is a sort of average between the estimated ROC and the estimated AROC curves, and ψ is a continuous function chosen to measure the distance among those curves. In particular we take three different distance functions, ψ_{L_1}, ψ_{L_2} and ψ_{KS}, two based on the L_1 and the L_2 measures and the other on the Kolmogorov-Smirnov criterion based on the supreme. This leaves us with

- $S_{\psi_{L_1}} = \int |\widehat{ROC}(p) - \widehat{MROC}(p)|dp + \int |\widehat{AROC}(p) - \widehat{MROC}(p)|dp,$
- $S_{\psi_{L_2}} = \int (\widehat{ROC}(p) - \widehat{MROC}(p))^2 dp + \int (\widehat{AROC}(p) - \widehat{MROC}(p))^2 dp,$
- $S_{\psi_{KS}} = \sup_{p\in(0,1)} \left|\widehat{ROC}(p) - \widehat{MROC}(p)\right| + \sup_{p\in(0,1)} \left|\widehat{AROC}(p) - \widehat{MROC}(p)\right|.$

These test statistics will take values close to zero when under the null hypothesis, and large positive values when under the alternative hypothesis. In order to approximate their distributions we are going to use a bootstrap algorithm.

At this point we have to be cautious about two things: first, the estimators of the ROC and the AROC curves are not independent, as they are obtained from the same samples; second, replicating the null hypothesis in a study with ROC curves is not a straightforward matter, even in the simpler case of the pooled ROC curve. Take into account that, although we can consider the ROC curve as a cumulative distribution function of a certain random variable ($R = 1 - G(Y^F)$, as proved in (2.1)), this variable is not observable. What we do observe are the samples of the two diagnostic variables, Y^F and Y^G, whose distributions (F and G, respectively) are estimated and used to compute the ROC curve. Thus, instead of having the unobserved samples $\{R_i\}_{i=1}^{n^F}$, we have the samples of the form $\{\hat{R}_i\}_{i=1}^{n^F} = \{1 - \hat{G}(Y_i^F)\}_{i=1}^{n^F}$. Because of that, the usual bootstrap methods are not directly applicable.

This is why the bootstrap algorithm that we will be using is based on the general bootstrap algorithm proposed in Martínez-Camblor and Corral (2012), aimed for problems

with a complex data structure. First, we consider the expression

$$T_\psi = \psi\left(\left(\widehat{ROC}(p) - ROC(p)\right) - \left(\widehat{MROC}(p) - MROC(p)\right)\right) + \psi\left(\left(\widehat{AROC}(p) - AROC(p)\right) - \left(\widehat{MROC}(p) - MROC(p)\right)\right),$$

where ROC and $AROC$ represent the theoretical ROC and AROC curves, respectively, and $MROC(p) = \frac{ROC(p)+AROC(p)}{2}$. Note that, under the null hypothesis, $ROC(p) = AROC(p) = MROC(p)$ for $p \in (0,1)$, and therefore $S_\psi = T_\psi$ (the same applies for ψ_{L_1}, ψ_{L_2}, ψ_{KS} or any other distance considered). The idea behind this methodology is to use T_ψ instead of S_ψ to compute the bootstrap statistic in the algorithm (note that, whereas T_ψ cannot be calculated in practice, it can be calculated in a bootstrap environment). This way, we do not need to assume any hypothesis when generating the bootstrap samples: the null hypothesis is being used when we exchange S_ψ^* by T_ψ^*.

In Martínez-Camblor and Corral (2012) the general bootstrap algorithm was designed for the comparison of a certain parameter or function in different populations, and here the ROC and the AROC curves that we want to compare come from the same place. To avoid the dependency problems that arise from estimating both curves using the same data, the original sample, conformed by $\{(X_i^F, Y_i^F)\}_{i=1}^{n^F}$ and $\{(X_i^G, Y_i^G)\}_{i=1}^{n^G}$, was divided (randomly and evenly) in two sets. One of those, $\{Y_{R,i}^F\}_{i=1}^{n_R^F}$ and $\{Y_{R,i}^G\}_{i=1}^{n_R^G}$ was used for the estimation of the ROC curve (note that the samples of covariates $\{X_{R,i}^F\}_{i=1}^{n_R^F}$ and $\{X_{R,i}^G\}_{i=1}^{n_R^G}$ are not needed to compute the ROC curve), and the other, $\{(X_{A,i}^F, Y_{A,i}^F)\}_{i=1}^{n_A^F}$ and $\{(X_{A,i}^G, Y_{A,i}^G)\}_{i=1}^{n_{AG}}$, for the estimation of the AROC curve, with $n^F = n_R^F + n_A^F$ and $n^G = n_R^G + n_A^G$.

The proposed bootstrap algorithm to approximate the distribution of (2.18) goes as follows:

1. From the original sample, compute the statistic value s_ψ, using $\{Y_{R,i}^F\}_{i=1}^{n_R^F}$ and $\{Y_{R,i}^G\}_{i=1}^{n_R^G}$ for the estimation of the ROC curve and $\{(X_{A,i}^F, Y_{A,i}^F)\}_{i=1}^{n_A^F}$ and $\{(X_{A,i}^G, Y_{A,i}^G)\}_{i=1}^{n_A^G}$ for the estimation of the AROC curve.

2. Generate B random samples (with B large) for the two sets of data. For $b \in \{1, \cdots, B\}$:

   (i) For $D \in \{F, G\}$, let $\{Y_{R,i}^{D,b*}\}_{i=1}^{n_R^D}$ be an i.i.d. sample from the empirical distribution function obtained from the first set of data.

   (ii) For $D \in \{F, G\}$, let $\{\varepsilon_{A,i}^{D,b*}\}_{i=1}^{n_A^D}$ be an i.i.d. sample from the empirical distribution function of the residuals computed using the second set of data, as in (2.10). Build the bootstrap sample $\{(X_{A,i}^D, Y_{A,i}^{D,b*})\}_{i=1}^{n_A^D}$, where $Y_{A,i}^{D,b*} = \hat{\mu}^D(X_{A,i}^D) + \hat{\sigma}^D(X_{A,i}^D)\varepsilon_{A,i}^{D,b*}$.

3. For $b \in \{1, \cdots, B\}$, obtain $\widehat{ROC}^{b*}(p)$ for $p \in (0,1)$ from $\{Y_{R,i}^{D,b*}\}_{i=1}^{n_R^D}$ (with $D \in \{F, G\}$) and $\widehat{AROC}^{b*}(p)$ for $p \in (0,1)$ from $\{(X_{A,i}^D, Y_{A,i}^{D,b*})\}_{i=1}^{n_A^D}$ (with $D \in \{F, G\}$).

4. Using T_ψ instead of S_ψ, compute the statistic bootstrap values $t_\Psi^{b,*}$, replacing $\widehat{ROC}$ by $\widehat{ROC}^{b*}$, ROC by $\widehat{ROC}$, $\widehat{AROC}$ by $\widehat{AROC}^{b*}$, and $AROC$ by $\widehat{AROC}$ for $b \in \{1, \ldots, B\}$.

Table 2.1: Conditional mean and conditional standard deviation functions considered for the construction of Scenarios A, B, C and D.

Scenario	Regression functions	Conditional standard deviation functions	Relationship among the curves
A	$\mu_A^F(x) = 2.5$ $\mu_A^G(x) = 1$	$\sigma_1^F(x) = 1.3$ $\sigma_1^G(x) = 1$	$ROC^x = AROC\ \forall x \in R_X$ $ROC = AROC$
B	$\mu_B^F(x) = 1.5x$ $\mu_B^G(x) = 0$	$\sigma_B^F(x) = 0.5$ $\sigma_B^G(x) = 0.5$	$ROC^x \neq AROC\ x \in R_X$ $ROC = AROC$
C	$\mu_C^F(x) = 2.5 + 2\log(x)$ $\mu_C^G(x) = 1 + 2\log(x)$	$\sigma_C^F(x) = 1.3$ $\sigma_C^G(x) = 1$	$ROC^x = AROC\ \forall x \in R_X$ $ROC \neq AROC$
D	$\mu_D^F(x) = -\sin(\pi(x+1))$ $\mu_D^G(x) = \sin(\pi(x+1))$	$\sigma_D^F(x) = 0.75$ $\sigma_D^G(x) = 0.75$	$ROC^x \neq AROC\ x \in R_X$ $ROC \neq AROC$

5. Use, as a p-value approximation, $p - value = \frac{1}{B}\sum_{b=1}^{B} I(s_\psi \leqslant t_\psi^{b,*})$.

Of course, the splitting of the sample to compute the estimators of the AROC and ROC curves implies a loss of power, as the sample size is divided by half. There is some room for improvement, and the search for an alternative procedure that overcomes this disadvantage is deferred for future studies. Still, to the best of our knowledge, this is the first methodology designed for this kind of test, with the upside of being a nonparametric approach.

Simulation study

In this section we carry out a finite sample study to analyse the performance of this new test in terms of level approximation and power. We consider four different scenarios, the same scenarios (A, B, C and D) that have been discussed before in this chapter. The location-scale regression models assumed for their construction, similar to the one presented in (2.9), are specified in Table 2.1. In that table we also indicate the relationships between the conditional, the covariate-adjusted and the pooled ROC curve.

We would not be following our own advice by testing H_0^2 (2.17) on scenarios like B or C, since we have already established that their conditional ROC curve changes with the value of the covariate. However, we have kept them in our simulation study to show that this test does not need any assumption regarding the behaviour of the conditional ROC curve: it can be conducted regardless of the result of the test H_0^1 (2.16).

The regression errors ε^F and ε^G were considered to follow normal standard distributions. The covariate followed a uniform distribution on the unit interval for both the diseased and the healthy population in Scenarios A, B and D. For Scenario C, the covariate follows a uniform distribution on the interval $[1, 15]$. Three different sample sizes were considered for the study, with $(n^F, n^G) = (100, 100), (250, 350), (500, 500)$. Note that the second sample size is unbalanced. 1000 datasets were simulated to compute the pro-

portion of rejection for each case. The number of bootstrap iterations (B) considered was 200.

Moreover, we used the three different distance functions previously mentioned (ψ_{L_1}, ψ_{L_2} and ψ_{KS}) for the construction of the test statistic, so we would discuss the results for the three of them. We will be denoting them as the L_1 (the one based on the L_1 measure), the L_2 (the one based on the L_2 measure) and the KS (the one based on Kolmogorov-Smirnov criterion) statistics.

Scenarios A and B were the ones selected to calibrate the level of the test (as they have equal ROC and AROC curves, they meet the null hypothesis). We show the results for three different nominal levels: $\alpha \in \{0.025, 0.05, 0.1\}$. Scenarios C and D were used to analyse the power. Note that the separation between the two curves is wider in scenario C, so we expect to obtain a higher power there with respect to scenario D. For those last scenarios we only show the results for $\alpha = 0.05$.

In Figure 2.5 we have four graphs containing the results of the simulation study for each one of the scenarios considered. The proportion of rejections are displayed there for each sample size and each test statistic. In order to obtain more detailed information about the practical performance of the tests under the null hypothesis we have included intervals constructed around the estimated proportion of rejection to verify whether the level is correctly approximated. More specifically, for a given estimated proportion of rejections, $\hat{p}$, the shown interval is $\left[\hat{p} \pm 1.96\sqrt{\frac{\alpha(1-\alpha)}{n_s}}\right]$, where n_s is the number of simulated samples used to obtain the estimated proportion. As long as those intervals contain the nominal level we can say that the test is well calibrated, as this is equivalent to perform a test to check if the actual level of the test equals the nominal level α. Note that the sample n_s is, in this case, 1000 for all intervals considered, and thus their length is not influenced by the sample sizes of the ROC curves of the study. This formula will be the one employed for the computation of the confidence intervals for the proportion that will appear in the rest of the chapters of this document (strictly speaking they are not confidence intervals, but we will use this notation for the sake of simplicity).

In the case of the L_1 and L_2 statistics, the nominal levels are well approximated by the estimated proportions, although in Scenario B the result for the unbalanced sample size is a bit overestimated. The KS statistic, however, seems to be more conservative (which is in line with the conservativeness of the Kolmogorov-Smirnov test).

As for the power, it shows the consistency of the test: it grows when we increase the sample size for all statistics. It is higher on Scenario C, as the difference between the ROC and the AROC curves is greater. In fact, in Scenario D the proportion of rejection is barely over the level of 0.05 (in the graphics, the dotted gray line). The L_1 and the L_2 statistics show a very similar behaviour, but the power obtained with KS is always below (in the case of Scenario C, considerably so).

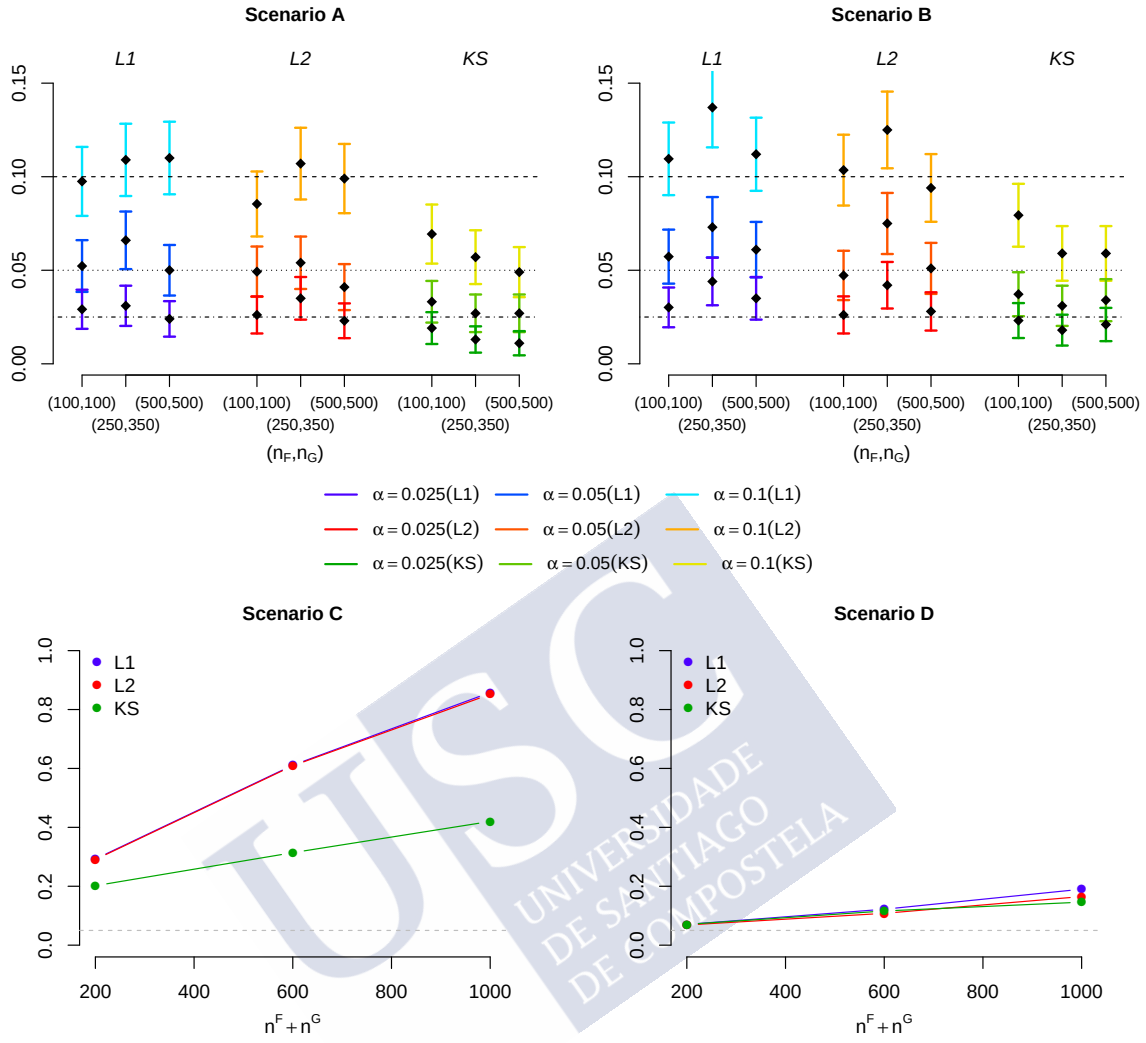


Figure 2.5: Results of the simulations study. The two graphs on the first row (one for each scenario under the null hypothesis) show the estimated proportions of rejection and their corresponding confidence intervals for all the sample sizes and the three test statistics considered. The graphs on the second row show the same but for scenarios under the alternative hypothesis, without the confidence intervals.

2.3.3 Application to real data

In order to illustrate the discussion and the test developed in this chapter we analyse a data set concerning patients suspected of prediabetes. The database at hand is the first one that was introduced in Chapter 1, in Section 1.2.2. Of the 1496 patients contained in this data set, 405 were considered as diseased (prediabetic) and 1091 as healthy. Note that this means that the data is unbalanced.

Out of the ten variables at our disposal we are going to focus our attention on five of them. First of all, we select three variables that we are going to consider as three different diagnostic markers: *GA* (which represents the glycated albumin), *a1c1* (haemoglobin)

Table 2.2: Summary of the variables contained in the Diabetes dataset for the prediabetic (D) and the non-prediabetic (H) subjects.

	GA		*a1c1*		$-$*GP22*		*age*	
	D	H	D	H	D	H	D	H
Minimum	8.50	7.88	4.80	3.10	−9.25	−11.29	24.0	18.00
1st quartile	13.46	12.46	5.60	5.20	−6.11	−7.24	55.0	35.00
Median	15.11	13.55	5.90	5.30	−5.43	−6.27	65.0	47.00
Mean	15.96	13.54	6.31	5.35	−5.53	−6.39	63.6	48.56
3rd quartile	17.63	14.58	6.70	5.50	−4.81	−5.46	73.0	62.00
Maximum	33.96	20.29	12.80	6.90	−3.23	−1.03	90.0	91.00

and $-$*GP22* (glycan peaks). We aim to assess the capability of correctly diagnosing prediabetes to the patients. Then, we are also going to take into account one covariate, the *age*. And finally, the last variable is a binary output that indicates if the subject has the disease (is prediabetic) or not. In Table 2.2 a summary of the continuous variables that are being used is shown.

We begin our analysis by representing the conditional densities of the three diagnostic markers at certain ages, along with their corresponding conditional ROC curves. The resulting graphics are collected in Figure 2.6. Note that the third diagnostic variable appears now under the tab $-$*GP22*. This is because, in this particular case, higher values of the diagnostic variables are more common in the healthy population, whereas the diseased subjects tend to have lower values, which goes against the assumptions made for the construction of a ROC curve. By taking the opposite values of this variable we ensure that the roles are exchanged.

At first sight it could appear that the conditional ROC curves remain constant through all those values, although we can appreciate a sort of hill for the medium age in the *GA* marker, and the $-$*GP22* seems to have better discriminatory power for the youngest patients, as the conditional ROC curves at those lower ages are closer to the point of maximum sensitivity and specificity.

However, there are two different issues that must be taken into consideration. First, the conditional ROC curve is estimated locally, which means that the estimations computed on the extreme values of the covariate are not as reliable, because they have fewer data around (and this condition exaggerates when the covariate is not uniformly distributed). Secondly, on those representations there is no insight on how the covariate is distributed in the healthy and in the diseased populations.

Next, we estimated the pooled and the covariate-adjusted ROC curves for each one of the diagnostic variables. We represented them in Figure 2.7. The conditional ROC curve was also estimated for certain values of the covariate, as well as their respective conditional AUC with a pointwise 0.95 confidence interval. The summary measures AUC and AAUC were also estimated (they are represented as horizontal lines, as they do not depend on fixed values of the covariate).

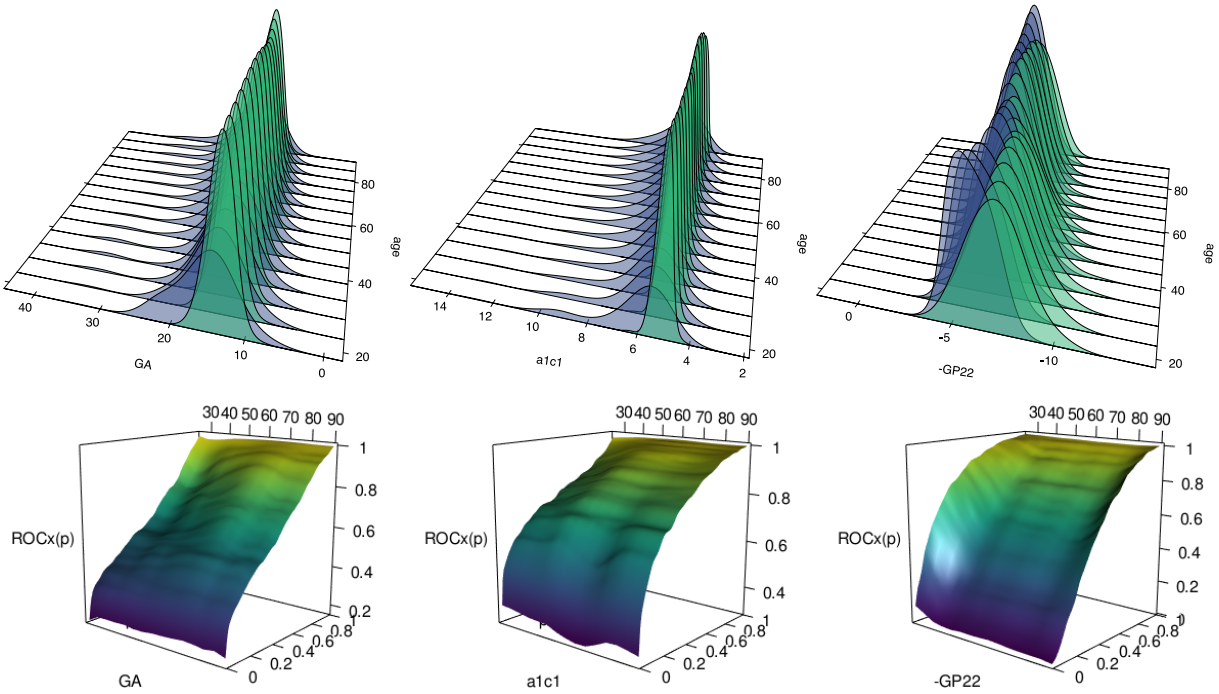


Figure 2.6: Estimated conditional densities of the three diagnostic variables, taking age as the continuous covariate and their corresponding estimated conditional ROC curves.

Setting our attention on those summary measures and the pointwise confidence interval we can have a first insight of the relationship between the curves. For the considered confidence level, the AAUC falls inside the confidence interval for all the values of the covariate, for all the diagnostic markers. Of course, we have to take into account that it is not a confidence band, so the level should be adjusted, but in any case it seems that there may not be differences between those indices. The AUC and the AAUC, despite being presented without confidence intervals, seem to be very similar in the first two variables. The ROC and the AROC curves of $-GP22$, however, are more separated (as their corresponding summary measures are).

Then, we follow the scheme depicted in Figure 2.4 and perform the two-step study for each one of those diagnostic markers (in the test for comparing the ROC and the AROC curves we used the test statistic L_2 and 500 bootstrap iterations). The obtained p-values, with their interpretation when we take a significance level of $\alpha = 0.05$, are summarized in Table 2.3. The conclusions that are drawn from that study match our previous suspicions: the covariate *age* does not seem to have a significant impact on the performance of each diagnostic marker. However, in the case of the $-GP22$ marker we find differences between the ROC and the AROC curve, and thus the latter should be employed for further analysis of this diagnostic variable.

For the sake of the argument (although this is not something that should be done in a practical case) we have reviewed the obtained results, this time for a significance level of $\alpha = 0.1$. The results are summarized in Table 2.4. The conclusions drawn this time

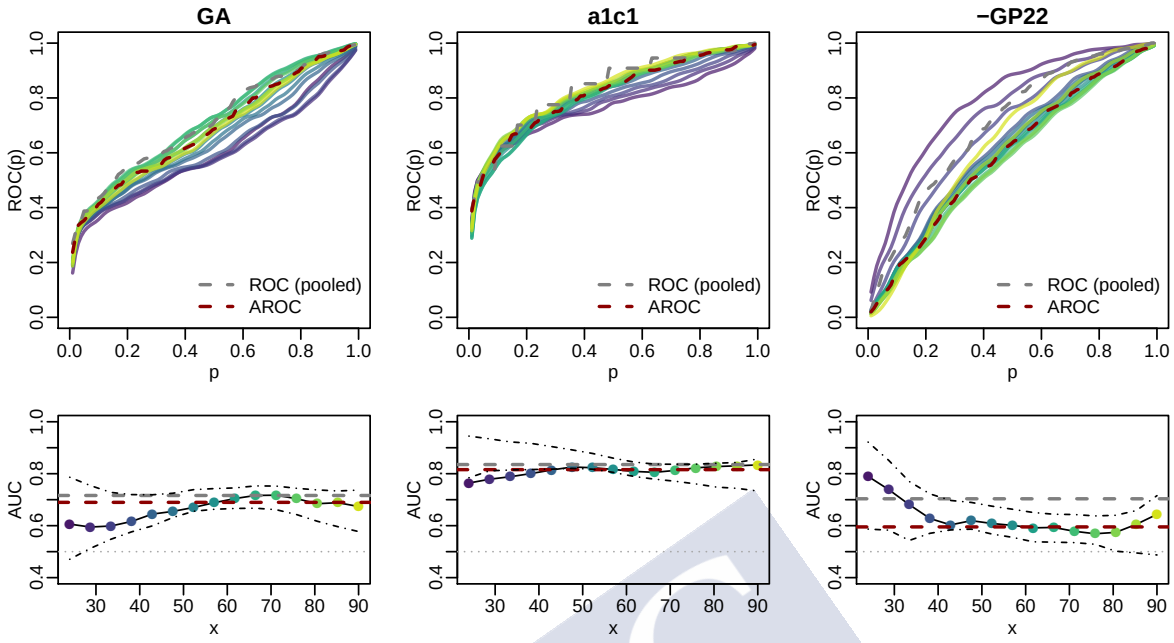


Figure 2.7: Estimated pooled (in discontinuous orange lines), covariate-adjusted (in discontinuous brown lines) and conditional (in continuous lines, one colour for each conditioned value) ROC curves of the three diagnostic, along with their corresponding summary measures, AUC (in discontinuous orange lines), AAUC (in discontinuous brown lines) and AUC^x, with its pointwise confidence interval. The gray horizontal line represents an AUC of 0.5, the hazard.

Table 2.3: Summarized p-values for tests (2.16) and (2.17) for the three diagnostic markers, taking $\alpha = 0.05$.

	GA	*a1c1*	*−GP22*
$\boldsymbol{H_0^1}$	$p-value = 0.08$	$p-value = 0.853$	$p-value = 0.523$
$\boldsymbol{H_0^2}$	$p-value = 0.782$	$p-value = 0.91$	$p-value < 0.001$
	$\Downarrow$	$\Downarrow$	$\Downarrow$
$\alpha = 0.05$	Use ***ROC***	Use ***ROC***	Use ***AROC***

Table 2.4: Summarized p-values for tests (2.16) and (2.17) for the three diagnostic markers, taking $\alpha = 0.1$.

	GA	*a1c1*	*−GP22*
$\boldsymbol{H_0^1}$	$p-value = 0.08$	$p-value = 0.853$	$p-value = 0.523$
$\boldsymbol{H_0^2}$	$p-value = 0.782$	$p-value = 0.91$	$p-value < 0.001$
	$\Downarrow$	$\Downarrow$	$\Downarrow$
$\alpha = 0.1$	Use $\boldsymbol{ROC^x}$	Use ***ROC***	Use ***AROC***

are very similar for the diagnostic markers of *a1c1* and *−GP22*, but for the *GA* variable the first test rejects the null hypothesis, indicating that its performance as a diagnostic

method can change depending on the values of *age*.

Another aspect that should be considered is that we are performing sequential comparisons without taking into account the problems that can arise from multitesting, but we do not elaborate further in this topic, as it is not the aim of this study. However, in a practical situation the level α should be controlled.

Finally, note that, despite the fact that in this section we have been dealing with a unidimensional covariate, the discussion of how to asses the significance of its effect in an ROC curve study is still valid for a multidimensional covariate. The limitation of the new test proposed in Section 2.3.3 (as well as the one proposed by Rodríguez-Álvarez et al., 2011b) comes mostly from the considered estimators of the conditional and covariate-adjusted ROC curves, which are valid only for unidimensional covariates.

2.4 Further analysis with more than one marker

In this chapter we have proven that a covariate can affect the performance of an ROC curve study in several ways. We have introduced and discussed the different curves that can be used to incorporate that covariate effect in the analysis, seeing different estimators for each one of the curves. By studying the relationships between the three curves we have designed a strategy for deciding which one of them (conditional, covariate-adjusted or pooled) is the better fitted for the situation at hand. At this point, we are in a position to apply that knowledge to the further analysis that we may be interested in.

In the first section of this chapter we have already mentioned that one of the motivations for correctly acknowledging the covariate effect in this context is the fact that the thresholds of the diagnostic variables can have different sensitivities and specificities for different covariate values. Thus, we could now look for optimal thresholds for the pooled, conditional or covariate-adjusted ROC curves.

On the other hand, one of the main objectives of this dissertation is the comparison of diagnostic methods throughout the comparison of the corresponding ROC curves. Depending on how the considered covariate affects each marker, the type of ROC curve that should be used for the comparison of the markers can change. A similar discussion to the one held in Section 2.3.2 can be developed here. Supposing that we wish to compare just two different methods of diagnosis (or that we want to compare the behaviour of one method in two separated groups), in Figure 2.8 we have summarized the strategy that could be followed to decide what ROC curves should be used to make the comparison in every possible scenario.

Taking into account all the possible combinations showed in that scheme, in the end we may find the following situations:

(i) The covariate at hand can be disregarded, and the pooled ROC curves can be used in each one of the groups. In this case, we can safely test $H_0^6 : ROC_1 = ROC_2$. There are multiple papers in the literature concerning this problem, and they will be the topic of discussion in Chapter 3.

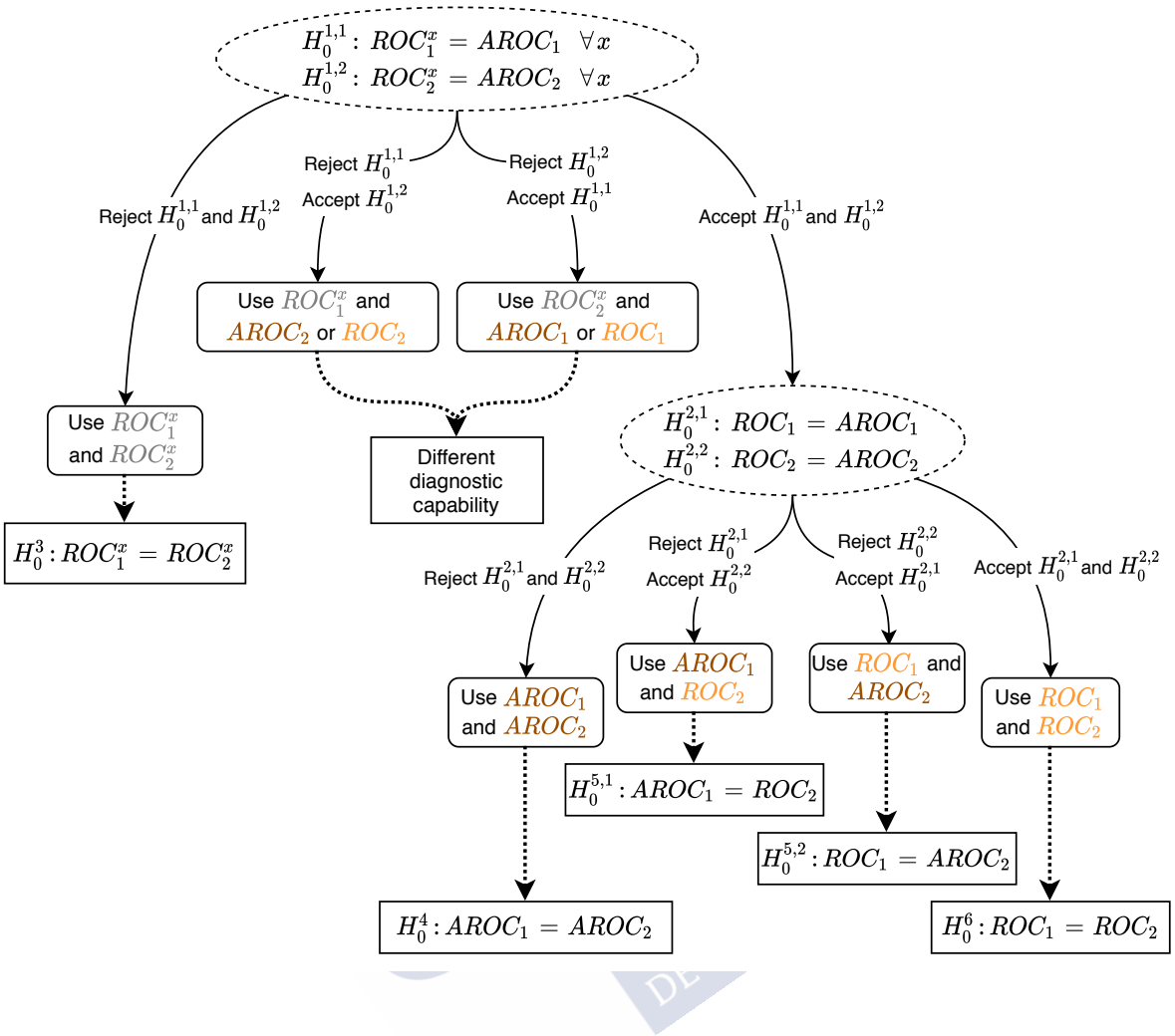


Figure 2.8: Scheme summarizing a strategy that could be employed for comparing two ROC curves when covariate information is available. ROC_1, $AROC_1$ and ROC_1^x are the pooled, covariate-adjusted and conditional ROC curves, respectively, related to one diagnostic method, and ROC_2, $AROC_2$ and ROC_2^x are the same for the other diagnostic method.

(ii) The covariate has no effect on one of the diagnostic methods, but it does influence the diagnostic variables of the other (without implications in its discriminatory capability). This would mean to perform the test $H_0^{5,1} : AROC_1 = ROC_2$ or the test $H_0^{5,2} : ROC_1 = AROC_2$. To the best of our knowledge, nothing has been done in the literature to address this problem, although a possible course of action could be to adapt the methodology proposed in Section 2.3.2 for the comparison of ROC and AROC curves.

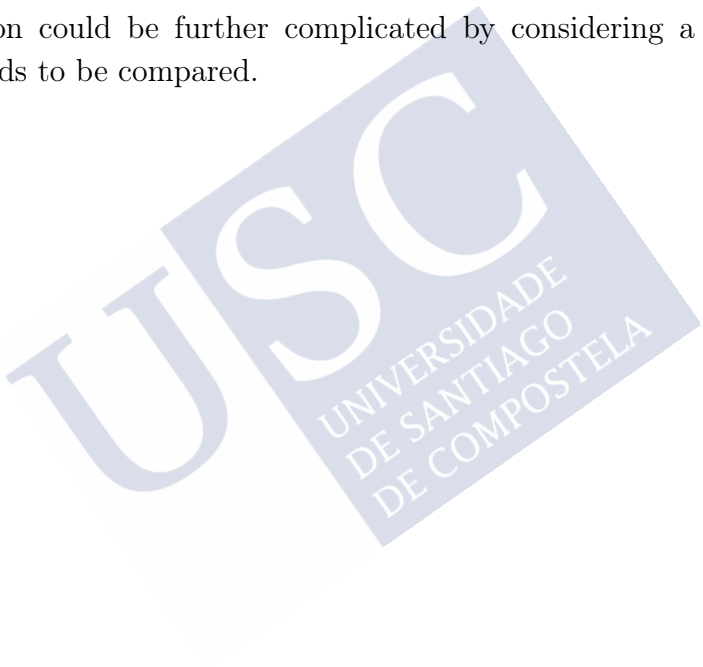
(iii) When the covariate affects the diagnostic markers but not their discriminatory capability in both groups, we should perform test $H_0^4 : AROC_1 = AROC_2$. In Janes et al. (2009) a test for comparing two AROC curves (based on a Wald test statistic) is presented, although the proposed test does not compare the whole curve but some

summary measures (like the AAUC).

(iv) The performance of the ROC curve is affected by the covariate in only one of the diagnostic methods. In this case, it does not make sense to perform further analysis, as we would be comparing a conditional ROC curve that has proven to change with the value of the covariate with another that does not.

(v) In the last situation both methods of diagnosis are affected by the covariate, and thus, their corresponding conditional ROC curves should be used. Then for the comparison of those methods we should test $H_0^3 : ROC_1^x = ROC_2^x,\ x \in R_X$. In general, the comparison of ROC curves in presence of covariates has not been dealt with in depth in the literature. We would dedicate Chapters 4 and 5 to this problem.

This discussion could be further complicated by considering a number $K > 2$ of diagnostic methods to be compared.



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

Comparison of ROC curves without covariates

The problem of comparing the accuracy of diagnostic tests is usually carried out through the comparison of the corresponding ROC curves. This matter has been approached from different perspectives. Usually, ROC curves are compared through their respective areas under the curve (AUCs), but in cases where there is no uniform dominance between the involved curves other procedures are preferred. Although the asymptotic distributions of the statistics behind these methods are, in general, known, resampling plans are also considered. With the purpose of comparing the performance of different approaches, with different ways of calibrating the distribution of the tests, a simulation study is carried out in this chapter to investigate the statistical power and the nominal level of each methodology.

The contents of this chapter appear on Fanjul-Hevia and González-Manteiga (2018).

3.1 Motivation

One of the main applications of the ROC curves is the comparison between different diagnostic markers. In a medical environment, a situation may arise where more than one method is available to diagnose a certain disease. In those cases, we would like to determine if one of those methods of diagnosis makes a better distinction between the healthy and the diseased populations or not. The ROC curve analysis is an appropriate tool to make that comparison.

This problem has been studied widely over the last decades. The first proposals that appeared in the literature address the issue through the comparison of the area under the ROC curves (AUCs). After that, other methods were developed to test the equality among ROC curves through some equivalent expression. One of the most recent approaches use the fact that the ROC curve can be viewed as a cumulative distribution function to make the comparison.

Some of these proposals are parametric approaches, others are nonparametric. Some

are prepared to deal with paired data (which means that the samples from the diagnostic variables from the different categories that we wish to compare come from the same individuals), some others do not. Moreover, while some methods are limited to the comparison of only two ROC curves, others are suited to compare more than two curves. As can be noted, there is quite an amount of different methodologies that approach the problem from different perspectives, and several ways to determine or approximate the distribution behind those methods. One may wonder which one of those would be the most suited to use in each given practical situation.

We can define the ROC curve the same way we have done in the previous chapters, taking into account that now we will be dealing with two (or, in general, $K \geqslant 2$) ROC curves. Let Y_k^F and Y_k^G be the continuous random diagnostic variables for each one of the different groups or categories that we wish to compare, with $k \in \{1, \ldots, K\}$. The corresponding k-th ROC curve to that pair of diagnostic variables would be

$$ROC_k(p) = 1 - F_k({G_k}^{-1}(1-p)), \quad p \in (0,1),$$

where $F_k(y) = P(Y_k^F \leqslant y)$, and $G_k(y) = P(Y_k^G \leqslant y)$ are the cumulative distribution functions of the k-th diagnostic markers and G_k^{-1} is the quantile function associated to the distribution G_k. In practice we will have two samples for each $k \in \{1, \ldots, K\}$ curve: $\{Y_{k,1}^F, \ldots, Y_{k,n_k^F}^F\}$ and $\{Y_{k,1}^G, \ldots, Y_{k,n_k^G}^G\}$ from the diseased and healthy populations, respectively. Note that, in the case of dependent ROC curves, the sample sizes are the same throughout all the different markers, with $n^F = n_1^F = \cdots = n_K^F$ and $n^G = n_1^G = \cdots = n_K^G$. In those situations it would be more appropriate to denote the samples as $\{(Y_{1,i}^F, \ldots, Y_{K,i}^F)\}_{i=1}^{n^F}$ and $\{(Y_{1,i}^G, \ldots, Y_{K,i}^G)\}_{i=1}^{n^G}$. Of course, for each ROC curve we can compute its corresponding summary measures, such as the AUC (that will be denoted as AUC_k, for $k \in \{1, \ldots, K\}$) or the Youden index.

Mathematically, our objective is to test the equality among two (or more) ROC curves:

$$H_0: \ ROC_1(p) = ROC_2(p) \quad \text{or} \quad H_0: \ ROC_1(p) = \cdots = ROC_K(p), \ p \in (0,1). \tag{3.1}$$

Note that the problem here goes beyond the mere comparison of cumulative distribution functions: we are not interested in comparing F_1 with F_2 and G_1 with G_2, but in comparing the degree of separation between F_1 and G_1 and between F_2 and G_2.

In practice, those ROC curves must be estimated, as F_k and G_k are unknown. In this chapter, the empirical estimator $\widehat{ROC}(p)$ (2.3) and the smoothed estimator $\widetilde{ROC}(p)$ (2.5) seen in Chapter 2 are employed to estimate each k ROC curves. F_k and G_k are thus estimated using the empirical distribution function or the kernel estimator (which requires the selection of bandwidth parameters).

In this chapter we make a comparative study of several ROC curve comparison procedures that have been proposed in the literature over the last few decades. First, we introduce the methods that are included in the study in Section 3.2. Then, in Section 3.3 we analyse the results of a simulation study in which we compared different ROC

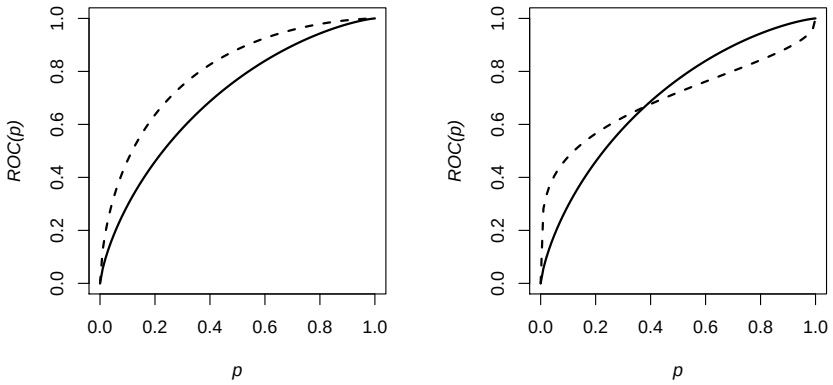


Figure 3.1: Examples of ROC curves for two different scenarios. In the one to the left, one of the ROC curves outperforms the other. In the other case, the curves cross each other and have similar AUC.

curves using those methodologies, to check the nominal levels and power of each test. The scenarios considered in that simulation study were selected to expose some of the weaknesses and limitations of those methods. Finally, some final remarks are discussed (Section 3.4).

3.2 ROC curve comparison methods in the literature

As it was mentioned before, in the last few decades several authors have faced the matter of comparing diagnostic markers. We have focused our attention in the nonparametric approaches, putting aside the ones that assume binormal ROC curves, i.e. curves constructed under the assumption that F and G follow normal distributions (as indicated in Section 1.1.3 of Chapter 1).

The first test to be developed was based on the AUC (Hanley and McNeil, 1983). Instead of checking the equality between the ROC curves, the test was made through the comparison of the corresponding AUCs. DeLong et al. (1988) were the first to propose a fully nonparametrical test of this sort, and despite the existence of further developments (as Wieand et al., 1989; Bandos et al., 2005; Martínez-Camblor, 2007; Braun and Alonzo, 2008), their method is nowadays the most frequently used to compare AUCs.

However, the equality among AUCs is not enough to determine the equivalence among different classification rules. This can be easily seen in Figure 3.1, where two different scenarios are displayed. In the first case, one of the ROC curves dominates the other over all the points, so the difference among them can be detected through the comparison of AUCs. In the second scenario, though, the two curves cross each other and, despite having the same AUCs, they certainly are different.

In order to solve that problem, some alternative approaches were proposed over the

years (Venkatraman and Begg, 1996; Venkatraman, 2000; Antoch et al., 2010; Martínez-Camblor et al., 2011, 2013).

Note that in every approach discussed here the problem has two sides. The first one consists on the selection of a criterion for measuring the possible difference among ROC curves, i.e., an appropriate statistic to perform the test. The second is to obtain the distribution of the designed test statistic under the null hypothesis. The latter is not always possible, so some resampling plans are proposed as an alternative way to approximate that distribution. Depending on the philosophy behind the construction of the test statistics (i.e., the comparison of the ROC curves through some summary measure, or through the whole curve) and the approach followed to obtain the distribution of the statistic, each methodology has its own strengths and weaknesses.

A simulation study was carried out in Section 3.3 to compare some of these methods. In particular, the methods that were included in the study are summarized below.

3.2.1 Comparison methods in the simulation study

DeLong *et al.* (1988)

The best known nonparametric procedure based on the comparison of the AUCs is the one proposed by DeLong et al. (1988). Despite the existence of several developments of their approach it is still one of the most commonly used methods for ROC curve comparison.

The area under an empirical ROC curve, when calculated by the trapezoidal rule, has been shown to be equal to the Mann-Whitney U-statistic for comparing distributions of values from two samples (Bamber, 1975), as shown in (2.4) on Chapter 2.

DeLong et al. (1988) make use of this analogy and exploit the theory for U-statistics for their method. For any contrast $\boldsymbol{L\hat{\theta}'}$, where $\boldsymbol{L}$ is a row vector of coefficients and $\hat{\boldsymbol{\theta}} = (\widehat{AUC}_1, \ldots, \widehat{AUC}_K)$ is a vector representing the estimated areas under the ROC curves, they propose the following test:

$$\frac{\boldsymbol{L\hat{\theta}'} - \boldsymbol{L\theta'}}{\boldsymbol{LSL'}} \sim N(0,1),$$

where $\boldsymbol{S}$ is the estimated covariance matrix for the vector of parameter estimates $\hat{\boldsymbol{\theta}}$, calculated using the jackknife estimator.

They also include a generalization of these results to apply any set of linear contrasts to a vector of AUCs. Despite having an asymptotic distribution of the final statistic, bootstrap procedures are also available, specially for the cases in which the size of the data is small. Although at first this method was designed to compare correlated ROC curves, it can also be applied when the data are independent. All these varieties of the test can be computed in R Core Team (2020) with the package *pROC* provided by Robin et al. (2011).

Venkatraman and Begg (1996) and Venkatraman (2000)

Venkatraman and Begg (1996) and Venkatraman (2000) were the first ones to propose a way of comparing diagnostic tests that was not based on the comparison of the AUCs for both the paired and non-paired situation. Instead, they defined a pair of mixture distributions such as:

- $M_{\kappa 1}(y_1) = \kappa F_1(y_1) + (1-\kappa)G_1(y_1)$
- $M_{\kappa 2}(y_2) = \kappa F_2(y_2) + (1-\kappa)G_2(y_2)$

where $0 < \kappa < 1$, the mixing proportion, is the prevalence of disease in the study population. They showed that, if the ROC curves were equal, then, for every $y_1 \in I_1$, it existed a corresponding $y_2 \in I_2$ such that $M_{\kappa 1}(y_1) = M_{\kappa 2}(y_2) = p$, with I_i the support of the diagnostic variable Y_i, for $i = 1, 2$ and $p \in (0, 1)$.

Let $y_1^p = M_{\kappa 1}^{-1}(p)$ and $y_2^p = M_{\kappa 2}^{-1}(p)$. Then, the misclassification probabilities for each diagnostic marker are given by

- $e_{y1}(p) = \kappa F_1(y_1^p) + (1-\kappa)(1 - G_1(y_1^p))$
- $e_{y2}(p) = \kappa F_2(y_2^p) + (1-\kappa)(1 - G_2(y_2^p))$

They observed that the difference between those probabilities, $e_{y1}(p) - e_{y2}(p)$, is identically zero for all p and for any $0 < \kappa < 1$ if and only if the ROC curves are equal. Thus, testing the null hypothesis (3.1) can be reduced to testing the hypothesis that the parameter $\eta = \int |e_{y1}(p) - e_{y2}(p)|dp$ is zero.

The misclassification rates e_{y1} and e_{y1} can be estimated by substituting the cumulative distributions F_i and G_i, for $i = 1, 2$, for their empirical distribution functions, $\hat{F}_i$ and $\hat{G}_i$. Not being able to develop a sufficiently accurate asymptotic approximation for the distribution of $\hat{\eta}$, they suggested a permutation method instead. This test can also be computed in R Core Team (2020) with the package *pROC*.

Antoch *et al.* (2010)

Taking into account the definition of ROC curves, Antoch et al. (2010) show that the null hypothesis (3.1) can be formulated as the equality of some increasing transformation functions τ_G and τ_F:

$$ROC_1(p) = ROC_2(p)\ \forall p \in (0,1) \Leftrightarrow \tau_G(t) = \tau_F(t)\ \forall t \in I_1,$$

where $\tau_G(t) = {G_2}^{-1}(G_1(t))$ and $\tau_F(t) = {F_2}^{-1}(F_1(t))$, and I_1 is the support of the variable Y_1. The test statistic suggested for this equivalent hypothesis is

$$T_n = n \int_{I_{1*}} (\hat{\tau}_G(t) - \hat{\tau}_F(t))^2 dt,$$

where $\hat{\tau}_F$ is the empirical estimator of τ_F, $\hat{\tau}_F = \hat{F}_2^{-1}(\hat{F}_1(t))$, $\hat{\tau}_G$ is the empirical estimator of τ_G and the integral is on a close interval $I_{1*} \subset I_1$ such that the densities related to the second ROC curve ($g_2(s)$ and $f_2(s)$) are positive and finite for all s in the images of $\tau_G(t)$ and $\tau_F(t)$, $t \in I_{1*}$. T_n is expected to take low values when the two ROC curves are equivalent. However, we detected an important lack of symmetry when running the simulation study for this statistic: the power obtained by testing ROC_1 vs. ROC_2 was significantly different from the one obtained when testing ROC_2 vs. ROC_1.

We proposed an alternative statistic, based on the same idea:

$$W_n = T_n + T_n^{-1},$$

where $T_n^{-1} = n\int_{I_{2*}}(\hat{\tau}_G^{-1}(t) - \hat{\tau}_F^{-1}(t))^2 dt$ is constructed as before but with the order of ROC_1 and ROC_2 permuted.

While Antoch et al. (2010) include the asymptotic distribution of T_n, in this chapter a permutation method was used to estimate the distribution of W_n.

Martínez-Camblor et al. (2011 and 2013)

The ROC curve of a continuous diagnostic marker can be viewed as a cumulative distribution function, as seen in Chapter 2 (2.1):

$$ROC(p) = 1 - F(G^{-1}(1-p)) = P(Y^F > G^{-1}(1-p)) = P(1 - G(Y^F) \leqslant p) = H(p),$$

where H is the cumulative distribution function of the variable $1 - G(Y^F)$. Following this analogy, Martínez-Camblor et al. (2011) use the traditional K-sample criterion for comparing cumulative distribution functions in the ROC curve context. For that reason, they chose to use a statistic of the form:

$$S_n = \sum_{k=1}^{K} d_n(\sqrt{n^F}\{\widehat{ROC}_k(p) - \widehat{ROC}(p)\}), \tag{3.2}$$

where $\widehat{ROC}_k(p)$, with $k \in \{1, \ldots, K\}$, is the empirical estimation of the k-th ROC curve $\widehat{ROC}(p) = K^{-1}\sum_{i=k}^{K}\widehat{ROC}_k(p)$, and $\{d_n\}_{n\in\mathbb{N}}$ is a sequence of real functions such that $d_n \to d$ (a.s.).

In Martínez-Camblor et al. (2011), they considered the test statistics based on the L_1-measure, the L_2-measure and on the classical Cramér-von Mises test:

- $L_1 = \sum_{k=1}^{K}\sqrt{n^F}\int|\widehat{ROC}_k(p) - \widehat{ROC}(p)|dp$,
- $L_2 = \sum_{k=1}^{K} n^F\int(\widehat{ROC}_k(p) - \widehat{ROC}(p))^2 dp$,
- $C_R = \sum_{k=1}^{K} n^F\int(\widehat{ROC}_k(p) - \widehat{ROC}(p))^2 d\widehat{ROC}(p)$.

In that paper, they use a permutation method to calibrate the statistics proposed. They also provided the code in R language (R Core Team, 2020) to carry out these kinds

of tests. Further on in the simulation study performed in Section 3.3 we will see the behaviour on L_1 and L_2 when calibrated that way.

Martínez-Camblor et al. (2013) use again this kind of approach. This time, they consider two d_n functions to construct the statistics:

- $d_n(g(t)) = ||g(t)||_\infty = sup_{t\in\mathbb{R}}|g(t)|$, based on the Kolmogorov-Smirnov criterion,
- $d_n(g(t)) = \int g(t)^2 dt$, based on the L_2-measure.

The second distance function results in the same statistic that was proposed on the previous paper. The innovation of the last paper is the proposal of a different resampling plan. There they introduce a general bootstrap algorithm to approximate the statistic distribution under the null hypothesis. These sort of tests can be computed in Pérez-Fernández (2017) with the library *nsROC*.

3.2.2 Resampling plans

As has already been mentioned, once we have decided on a criterion to compare the curves we need to find its distribution under the null hypothesis. Very often such distribution of the statistic is unknown or difficult to calculate. In those situations, some resampling plans are necessary to calibrate its distribution. Here we include two different procedures that are applied in the previously explained methodologies.

Permutation methods

Permutation methods (Good, 2005) have become more popular in the last few years due to the increase in computational power. They are, in general, robust and almost assumption-free statistical tools.

They consist in generating different samples by permuting the observed data. Then, the value of the test statistic for the original data is compared with all the values obtained for the statistic when the data are exchanged. The application of this test requires that the observations are exchangeable, which they are if the probability of any particular result is the same independently of the order of the observations.

In general, the steps to perform a permutation test are the following:

1. Computation of the test statistic for the observed data.
2. Generation of a certain number of samplings by randomly permuting the data and evaluation of the test statistic for each one of the permutations.
3. Computation of the approximation of the p-value.

This methodology was employed in Martínez-Camblor et al. (2011) to calibrate the statistics considered therein, and in our simulation study it is also used to evaluate the distribution of the statistic based on the one proposed by Antoch *et al.* (2010). However,

a warning must be given: the null hypothesis that is considered in these procedures is the equality among the cumulative distribution functions of the healthy and the diseased population of the different ROC curves ($F_1 = F_2$ and $G_1 = G_2$), which is not exactly the same as the equality among ROC curves. In fact, it is a stronger hypothesis than the one considered in (3.1). Note that, as stated previously, the ROC curve is invariant to monotone increasing transformations on the diagnostic markers. If we take Y_1^F and Y_1^G as a pair of diagnostic markers, and then we take $Y_2^F = \gamma(Y_1^F)$ and $Y_2^G = \gamma(Y_1^G)$, with $\gamma(\cdot)$ a monotone increasing function, both pairs of diagnostic markers will yield the same ROC curve, without necessarily having the same distribution functions. This will have consequences when checking the nominal level of these tests.

Furthermore, this is not the case for Venkatraman's method, since the permutation here is not employed directly: the observed data are first transformed using the mixture distributions $M_{\kappa 1}$ and $M_{\kappa 2}$ so the desired null hypothesis is kept.

The general bootstrap algorithm for hypothesis testing (GB)

When the null does not imply necessarily the equality among the involved cumulative distribution functions (which is what happens here) the usual bootstrap resampling plan may not be a good method for estimating the statistic variability. This is why Martínez-Camblor and Corral (2012) proposed a general bootstrap algorithm which deals with this problem, preserving the data structure within the different studied groups. The key of this algorithm is that the null hypothesis is considered in order to compute the statistic (bootstrap) values instead of at the resampling moment (as usual). They use the fact that the statistic in (3.2), under the null (and only under the null) is equal to:

$$S_n = \sum_{k=1}^{K} d_n \left(\sum_{j=1}^{K} \alpha_{kj} \sqrt{n^F} \{ \widehat{ROC_j}(p) - ROC_j(p) \} \right), \tag{3.3}$$

where $\alpha_{kj} = I\{k = j\} - 1/n^F$.

The algorithm they propose is the following:

1. From the original sample, compute the real statistic value, s_n (3.2).

2. From the smoothed multivariate cumulative empirical distribution functions, $\tilde{F}$ and $\tilde{G}$, generate B K-dimensional random samples for the healthy and the diseased populations.

3. Compute the statistic bootstrap values, $s_n^{*,b}$ using the expression (3.3) and replacing, for $b \in \{1, \ldots, B\}$, $\widehat{ROC_j}(p)$ by $\widehat{ROC_j}^{*,b}(p)$ and $ROC_j(p)$ by $\widetilde{ROC_j}(p)$, where $\widehat{ROC}(p)$ and $\widetilde{ROC_j}(p)$ are the estimators mentioned in Section 3.1.

$$s_n^{*,b} = \sum_{i=1}^{K} d_n \left(\sum_{j=1}^{K} \alpha_{kj} \sqrt{n} \{ \widehat{ROC_j}^{*,b}(p) - \widetilde{ROC_j}(p) \} \right).$$

4. The final p-value approximation will be

$$p - value = \frac{1}{B}\sum_{i=1}^{B} I(s_n < s_n^{*,b}).$$

Note that in Step 2 the term *smooth* was employed. Thus, there is going to be some bandwidth parameters involved (h^F and h^G) which are not always easy to choose. In Step 3, in order to avoid the computation of $\widetilde{ROC}_j(t)$ (and thus, avoiding the selection of another bandwidth parameter) they propose to replace that expression with $\overline{\widehat{ROC}_j}^* = (1/B)\sum_{b=1}^{B} \widehat{ROC}_j^{*,b}$.

3.3 Simulations

In this section we show the results from a Monte Carlo simulation study that compares some of the procedures mentioned before. In particular, it includes the DeLong's AUC comparison method (D_L), the Venkatraman method for independent samples (V_k), a statistic based on the one proposed by Antoch et al. (2010) (W_n), two statistics proposed by Martínez-Camblor et al. (2011) based on the L_1 and the L_2-measures and calibrated through a permutation method (L_1^P and L_2^P) and two statistics proposed by Martínez-Camblor et al. (2013) based on the L_2-measure and the Kolmogorov-Smirnov criterion and calibrated through a general bootstrap algorithm introduced by Martínez-Camblor and Corral (2012) (KS^{GB} and L_2^{GB}). A total of seven methodologies were compared. We recall that each procedure is a combination of a certain statistic and a way of calculating its distribution under the null hypothesis. For example, L_2^P and L_2^{GB} use the same statistic to perform the test, but they are studied separately due to the different resampling plans employed.

In order to investigate the nominal level and the statistical power of these methodologies, nineteen different models were considered, all of them regarding the case in which only two independent ROC curves were compared.

The simulations are based on 1000 Monte Carlo replications, and the p-values where computed using 200 permutations or 500 bootstrap samples, depending on the calibration method adopted. Different sample sizes were considered for each situation, with n_k^F and n_k^G indicating the sample sizes of the k-th diagnostic marker for the diseased and healthy population respectively, with $k \in \{1, 2\}$.

3.3.1 Level of the tests

Twelve scenarios were considered in order to investigate the nominal level of the different statistics. The pairs of ROC curves that were compared were generated from normal ($N(\mu, \sigma)$, with μ and σ being the parameters of mean and standard deviation), Weibull ($Weibull(\alpha, \beta)$, with α and β the scale and shape parameters, respectively), exponen-

Table 3.1: Distribution functions of the diagnostic markers that generate the ROC curves under the null hypothesis.

	Y_G^1	Y_F^1	Y_G^2	Y_F^2
A	$N(0,1)$	$N(0.36,1)$	$N(0,1)$	$N(0.36,1)$
B	$N(0,1)$	$N(1.19,1)$	$N(0,1)$	$N(1.19,1)$
C	$N(0,1)$	$N(2.2,1)$	$N(0,1)$	$N(2.2,1)$
D	$N(0,1)$	$N(2.5,2.5)$	$N(0,1)$	$N(2.5,2.5)$
A_W	$Weibull(0.4,0.25)$	$Weibull(0.5,0.5)$	$Weibull(0.4,0.25)$	$Weibull(0.5,0.5)$
B_W	$Weibull(2.5,1)$	$Weibull(3.5,1.5)$	$Weibull(2.5,1)$	$Weibull(0.5,0.5)$
A_{Exp}	$Exp(2)$	$Exp(1.5)$	$Exp(2)$	$Exp(1.5)$
B_{Exp}	$Exp(1.5)$	$Exp(0.5)$	$Exp(1.5)$	$Exp(0.5)$
$A.1$	$N(0,1)$	$N(0.36,1)$	$LN(0,1)$	$LN(0.36,1)$
$B.1$	$N(0,1)$	$N(1.19,1)$	$LN(0,1)$	$LN(1.19,1)$
$C.1$	$N(0,1)$	$N(2.2,1)$	$LN(0,1)$	$LN(2.2,1)$
$B.2$	$N(0,1)$	$N(1.19,1)$	$N(3,1)$	$N(4.19,1)$

tial ($Exp(\lambda)$), and lognormal ($LN(\mu,\sigma)$) distribution functions detailed in Table 3.1. Those pairs of ROC curves are also represented in Figures 3.2 - 3.6 along with the results of the simulation study. The considered sample sizes were $(n_1^F, n_1^G) = (n_2^F, n_2^G) = (50,50), (100,100), (150,150)$.

The first four models, A, B, C and D are binormal models where the null hypothesis was reached by taking the same cumulative distribution functions for the diseased and the healthy population for the construction of both ROC curves (meaning $F_1 = F_2$ and $G_1 = G_2$). This is also the case for the models A_W, B_W, A_{Exp} and B_{Exp}, though these scenarios do not come from normal distributions. A_W and A_{Exp} have approximately the same shape as the A model, but are constructed using the Weibull and exponential distributions respectively. The same applies for B_W and B_{Exp} regarding scenario B.

The last models ($A.1$, $B.1$, $C.1$ and $B.2$) are slightly different: the pairs of ROC curves that we compare there, despite being equal, have $F_1 \neq F_2$ and $G_1 \neq G_2$. This particularity was achieved due to the fact that the ROC curve is invariant under non-decreasing transformations. These kinds of scenarios are hardly ever considered in other articles to calibrate the nominal level of the statistic that compares ROC curves, although it is a situation that can arise in real-life problems. Note that models A and $A.1$ lead to the same ROC curves, as they do B, $B.1$ and $B.2$, and C and $C.1$.

We set aside these special models for later, and we focus our attention on the other eight. The proportion of rejections observed by the different methodologies is displayed in Figures 3.2 (for binormal scenarios), 3.3 (for non-binormal scenarios) and 3.4. Figures 3.2 and 3.3 contain the approaches that do not depend on any bandwidth parameter. Each row reflects the results obtained by each statistic, and each column represents the model tested. For each scenario and statistic, the proportion of rejections was estimated for three different sample sizes. Moreover, the confidence interval for each proportion was

calculated so that, if the nominal level that we are looking for is not contained in that interval, we can conclude that the observed rejection proportion is significantly different than the one expected. In such cases, we can conclude that the statistic considered is not well calibrated.

In general, looking at the confidence intervals obtained in Figures 3.2 and 3.3, we could say that the nominal level expected was reached. On the other hand, in Figure 3.4 we see the results for KS^{GB} and L_2^{GB}. In these cases, a bandwidth parameter must be chosen. The authors take $h^F = h\hat{\sigma}^F(n^F)^{-1/3}$ to run the smooth bootstrap samples from the diseased population and $h^G = h\hat{\sigma}^G(n^G)^{-1/3}$ for the healthy population, without suggesting any criterion to determine h. Here we include the results for $h = \{1/2, 1, 2\}$. We also propose the use of the Sheather and Jones (1991) criterion for optimal bandwidth selection for smooth density estimation.

The first thing we notice when testing these statistics is the importance of choosing an appropriate bandwidth parameter. For L_2^{GB} the bandwidth selection has small effect on the final result, although its behaviour gets worse as the ROC curve approaches the point of maximum sensitivity and specificity. This effect is greater for KS^{GB}: the appropriate bandwidth changes with the shape of the ROC curve. For example, in model A, $h = 1, 2$ give a good calibration for the statistic, while $h = 1/2$ seems a little anticonservative. In model B, however, $h = 2$ does not perform so well whereas $h = 1/2$ reaches the expected nominal level. As for models C and D all of the values yield conservative results.

The bandwidth based on the Sheather and Jones criterion proposed here seems like a suitable alternative: despite the fact that it still has problems to calibrate the statistic when the ROC curve approaches the point (0,1), it seems to perform at least as good as any of the bandwidths selected manually.

Similar conclusions can be obtained when focusing on the non-binormal models. In particular, model A_W shows the usefulness of having some sort of criterion for choosing the bandwidth, as the fixed values $h = \{1/2, 1, 2\}$ lead to a rejection proportion that falls well below the expected 0.05. It is worth noticing that similar shapes of the ROC curves (as it happens in models A and A_W) does not guarantee that the same bandwidth parameter is suitable for both situations.

Moreover, if we focus our attention in the special scenarios $A.1$, $B.1$, $C.1$ and $B.2$ mentioned before, we find that some of the statistics that worked rather well for the previous models do not perform correctly in these cases. In Figure 3.5 and 3.6 we can see the results for these models for all the methodologies.

W_n, L_1^P and L_2^P underestimate the proportion of rejection in all four scenarios considered. This is because, in the permutation procedure, the null hypothesis that is being replicated is the one that assumes that $F_1 = F_2$ and $G_1 = G_2$. This assumption, despite implying the equality among ROC curves, is not equivalent to the null hypothesis that is being tested.

D_L and V_k still calibrate the nominal level as expected in model $B.2$, although they also seem a little anticonservative when testing $A.1$ and $B.1$. For scenario $C.1$ none of the statistics considered reached the level desired. On the other hand, the good performance of

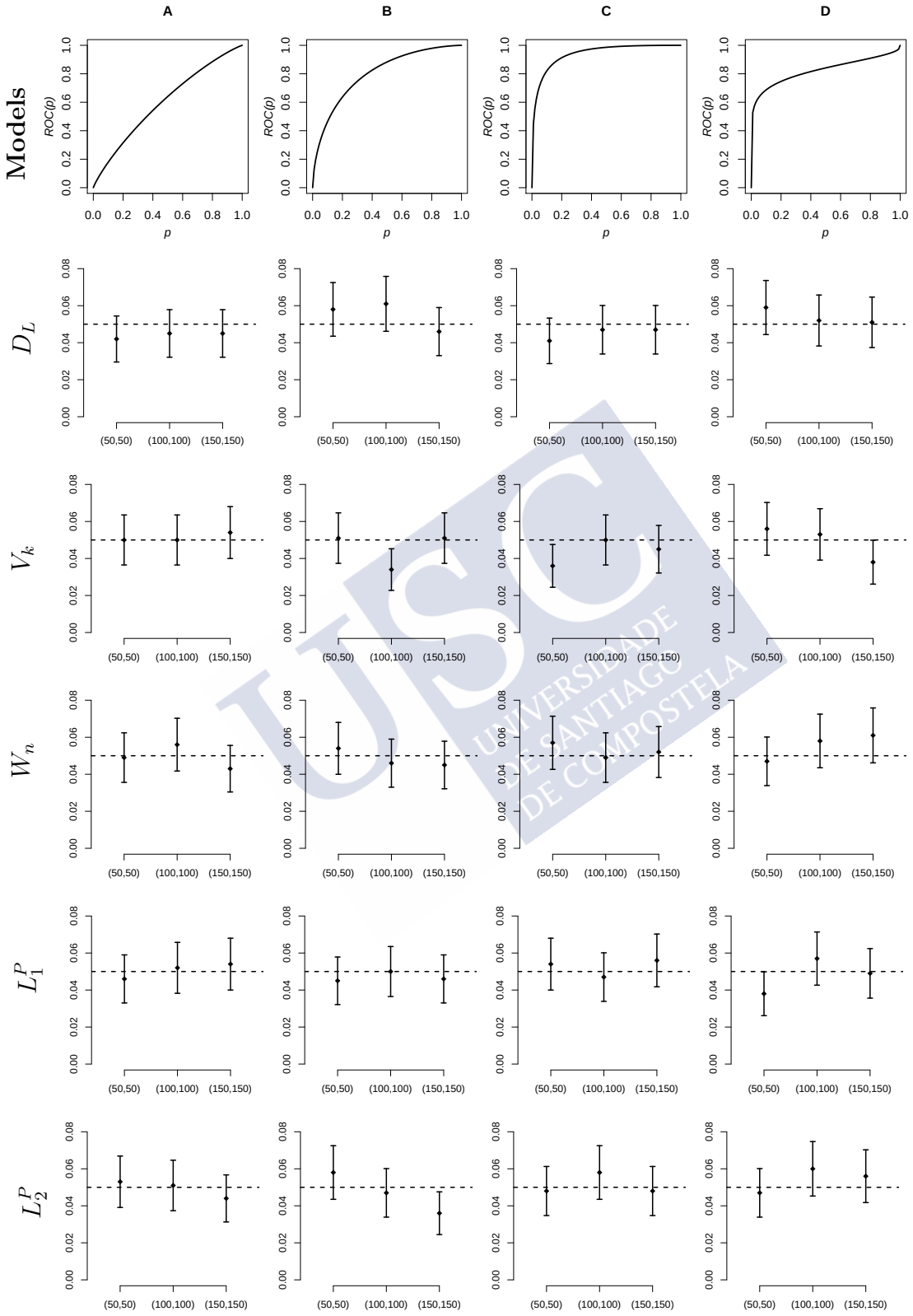


Figure 3.2: Observed rejection proportions in 1000 Monte Carlo simulations under the null hypothesis and their confidence intervals for different sample sizes. The horizontal dashed line represents the theoretical nominal level $\alpha = 0.05$.

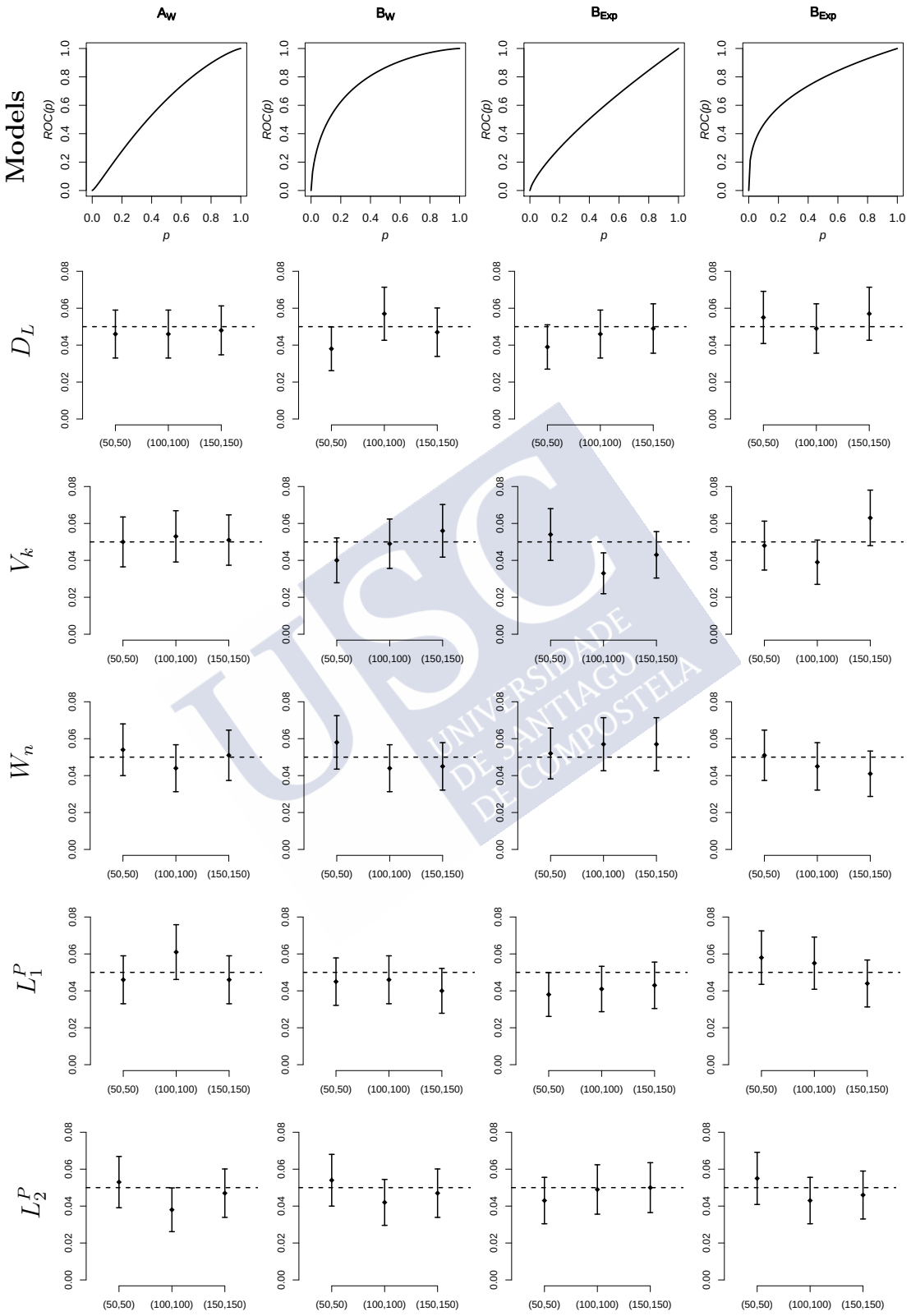


Figure 3.3: Observed rejection proportions in 1000 Monte Carlo non-binormal simulations under the null hypothesis and their confidence intervals for different sample sizes. The horizontal dashed line represents the theoretical nominal level $\alpha = 0.05$.

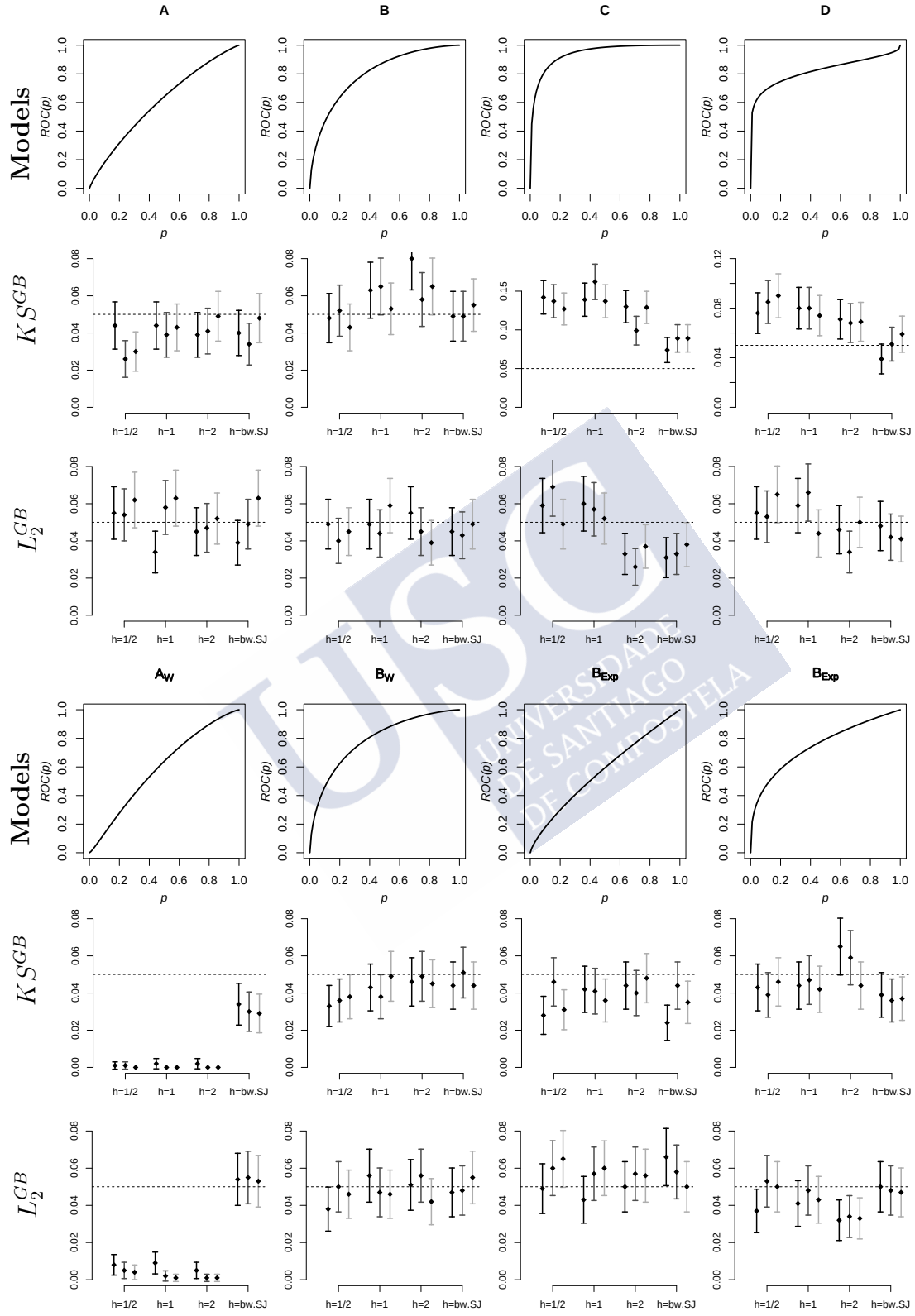


Figure 3.4: Observed rejection proportions in 1000 Monte Carlo simulations under the null hypothesis and their confidence intervals for different sample sizes (from the darkest to the lightest, $(n, m) \in \{(50, 50), (100, 100), (150, 150)\}$) and for different bandwidth choices.

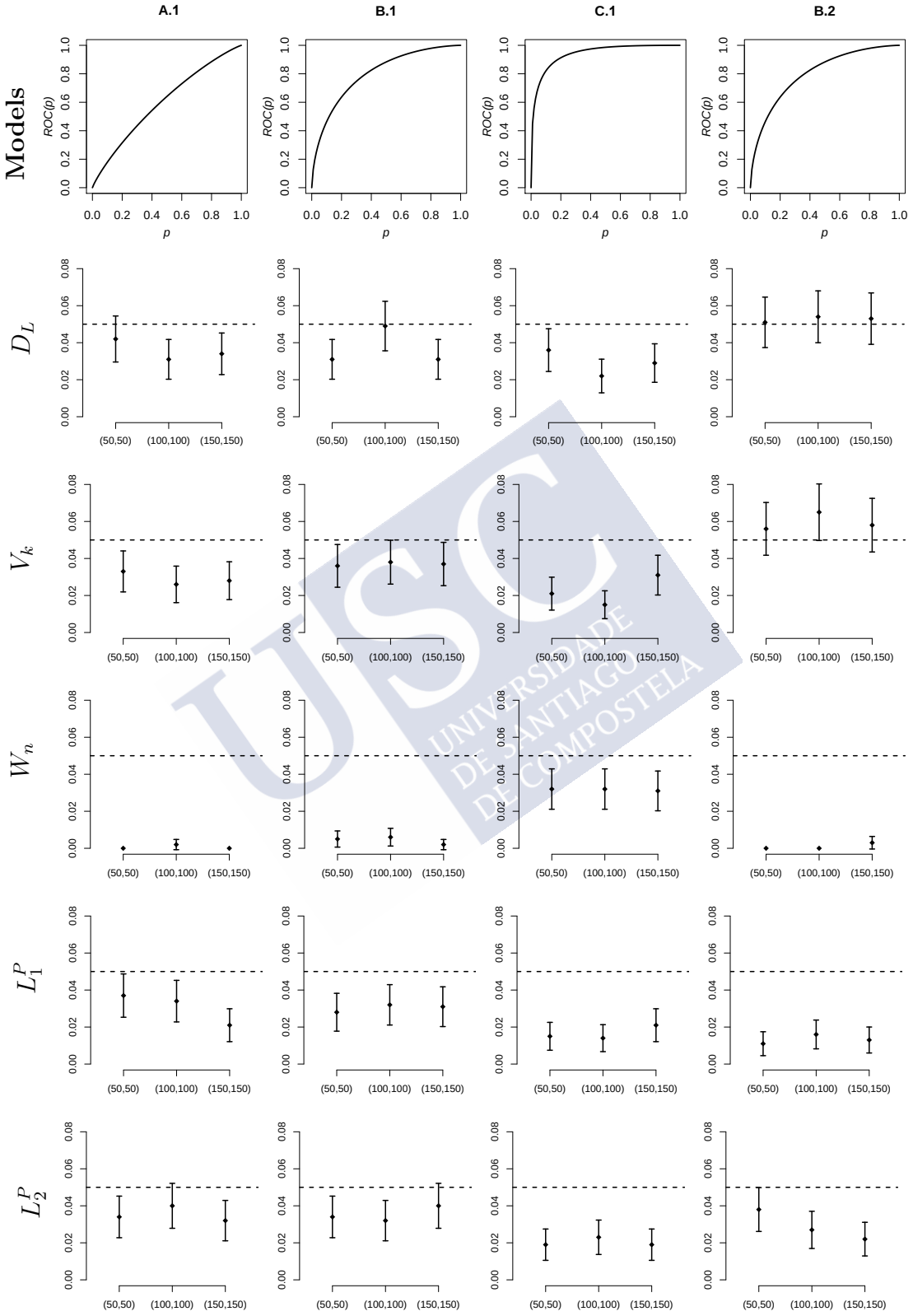


Figure 3.5: Observed rejection proportions in 1000 Monte Carlo simulations and their confidence intervals in one scenario in which, although the null hypothesis holds, $F^1 \neq F^2$ and $G^1 \neq G^2$.

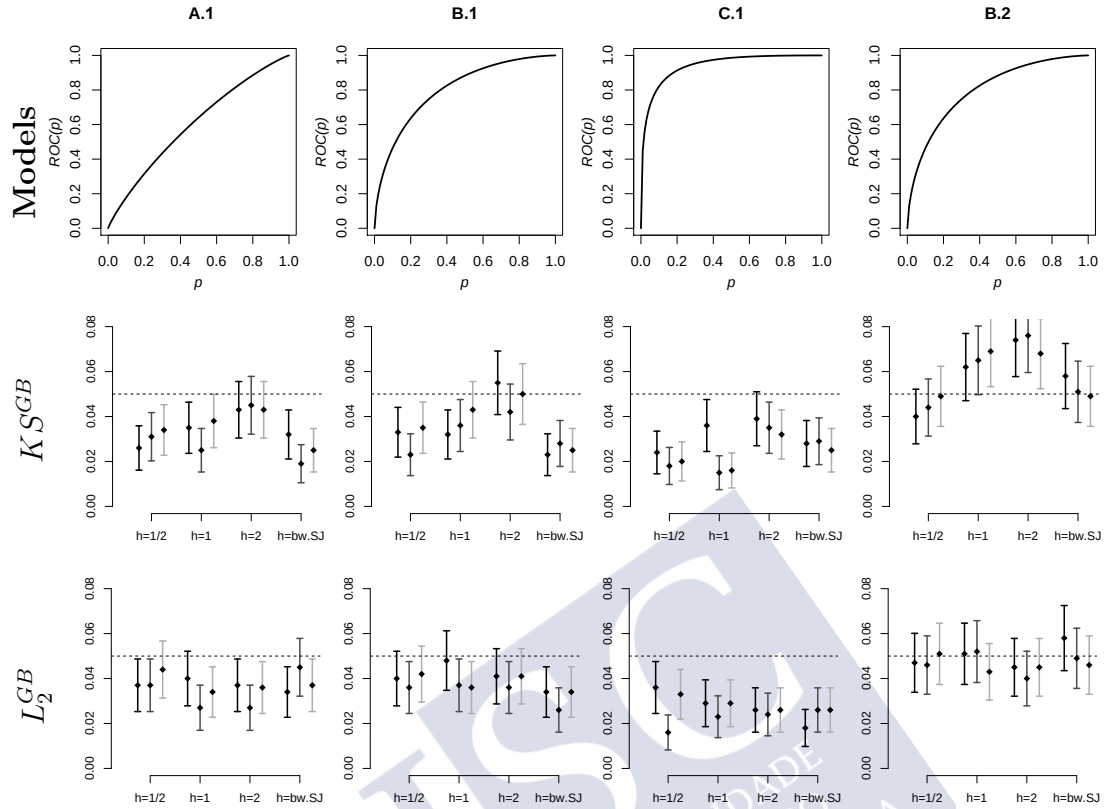


Figure 3.6: Observed rejection proportions in 1000 Monte Carlo simulations and their confidence intervals in scenarios in which, although the null hypothesis holds, $F_1 \neq F_2$ and $G_1 \neq G_2$. For each scenario several sample sizes and several bandwidth choices were considered.

KS^{GB} and L_2^{GB} depends, once again, on the choice of the bandwidth parameter, specially in the case of KS^{GB}. Even so, the general bootstrap general algorithm seems like a suitable alternative to the permutation method.

3.3.2 Power of the tests

In order to investigate the statistical power for the previous statistics, seven different scenarios were chosen to perform a Monte Carlo simulation study. The distribution functions that constitute the pairs of different ROC curves compared in each scenario are detailed in Table 3.2.

Once again, all cases were generated from normal, Weibull or exponential distributions. The resulting ROC curves are plotted in Figure 3.7. Of course, this time the ROC curves that are compared are not equal. In spite of this, some of the pairs of curves can still have similar AUC.

The considered sample sizes were $(n_1^F, n_1^G) = (n_2^F, n_2^G) = (25, 25)$, $(50, 50)$, $(100, 100)$, $(150, 150)$, $(200, 200)$. The bandwidth parameters chosen for the general bootstrap algorithm were the same as before, but, as the Sheather and Jones criterion seemed to

Table 3.2: Distribution functions of the diagnostic markers that generate the ROC curves under the alternative hypothesis.

	Y_G^1	Y_F^1	Y_G^2	Y_F^2
E	$N(0,1)$	$N(0.74,1)$	$N(0,1)$	$N(1.19,1)$
F	$N(0,1)$	$N(0.74,1)$	$N(0,1)$	$N(1.17,2)$
G	$N(0,1)$	$N(1.81,1)$	$N(0,1)$	$N(2.87,2)$
H	$N(0,1)$	$N(0.74,1)$	$N(0,1)$	$N(1.88,2)$
I	$N(0,1)$	$N(1.00,2)$	$N(0,1)$	$N(2.00,4)$
J	$Weibull(2.5,1)$	$Weibull(3.1,1.5)$	$Weibull(1,1)$	$Weibull(2.5,2.5)$
K	$Exp(2)$	$Exp(1)$	$Exp(1.5)$	$Exp(0.5)$

yield higher power for the statistics than the others, only the results for that bandwidth parameter are displayed here.

Figure 3.8 shows the observed statistical power for the seven scenarios considered here. At first sight, none of the proposed methods seem to have higher power than the rest of them, at least not for all situations.

DeLong's method, as expected, is probably the best test for the scenarios in which one ROC curve dominates the other in all points, as occurs in model (E), but fails to discriminate ROC curves that, despite having different shapes, have the same AUC. Venkatraman's method behaves as well as the DeLong's in the case of dominance among ROC curves, but although it outperforms that method in other situations, its power seems lower than the rest.

The statistics that were calibrated through a permutation method also behave differently depending on the scenario. W_n seemed to be one of the most powerful statistics, but we have to take into account that it did not always reach the nominal level desired, so it is not appropriate to compare its power. The same applies for L_1^P and L_2^P. It is worth noting that L_2^P has similar power when applying the general bootstrap algorithm, L_2^{GB}.

As for KS^{GB}, it outperforms L_2^{GB} in most scenarios (excluding models E and K, where there is dominance among curves, and model J, that is obtained from Weibull distributions). Still, once again we have to recall the correct behaviour of this statistic depends highly on the bandwidth choice.

In general, all we can say is that the results are not conclusive: the most powerful methodology depends on the shape of the ROC curves that are being compared.

3.4 Discussion

The comparison of ROC curves has been a widely accepted methodology for comparing the accuracy of different diagnostic markers. For the last few decades, several different methods have been developed to handle that sort of study.

In this chapter we have presented and compared the principal existing methods for ROC curve comparison. In particular, we have focused our attention in the nonparametric

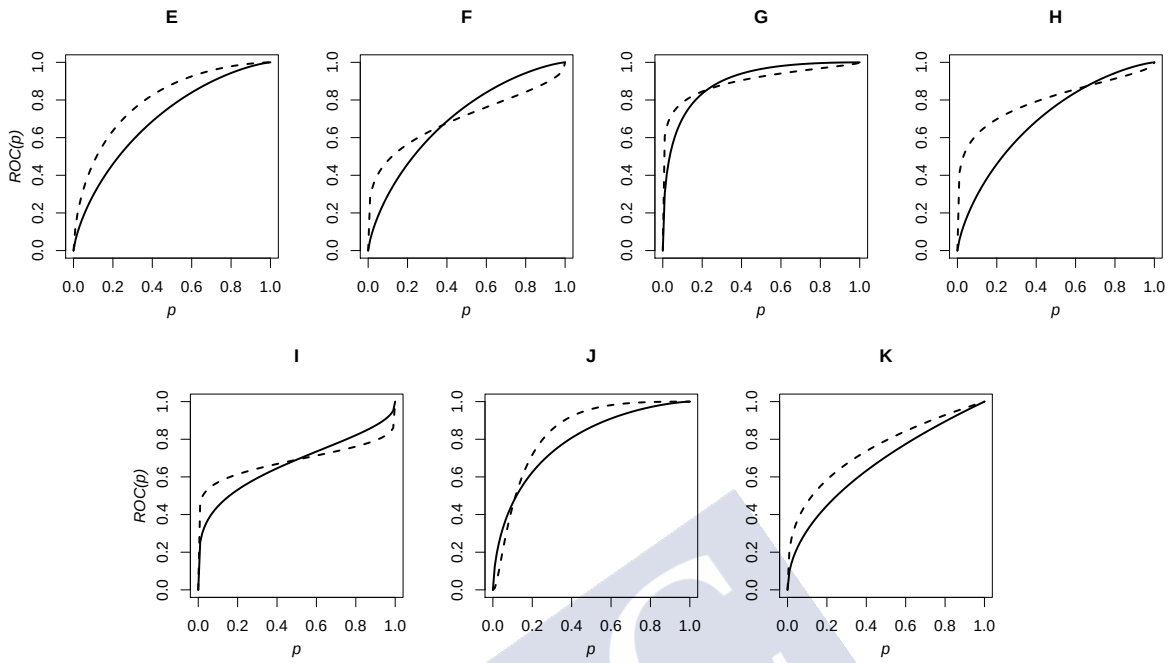


Figure 3.7: Scenarios under the alternative hypothesis that are used to calculate the statistical power of the statistics.

comparison of two independent ROC curves, although some of the methods studied can be applied to the case with paired data (DeLong et al., 1988; Venkatraman, 2000; Martínez-Camblor et al., 2013), or to the case in which more than two ROC curves are to be compared (DeLong et al., 1988; Martínez-Camblor et al., 2011, 2013).

The simulation study detailed in Section 3.3 shows that most of the tests considered reached an appropriate nominal level. However, some of the methods that used the permutation procedure did not always behave properly when comparing equal ROC curves with different cumulative distribution functions defining each curve. Likewise, the desired nominal level was not obtained for certain bandwidth choices, specially with the Kolmogorov-Smirnov type of statistic. The problem of choosing an appropriate smoothing parameter for the general bootstrap algorithm is still open. The proposal discussed here (taking the Sheather and Jones criterion for optimal bandwidth selection for smooth density estimation) behaves at least as well as some fixed parameter h.

As for the power of the tests, none of the methods proposed here outperforms the others in all the scenarios considered. In the cases where one of the curves dominates the other, the DeLong's method yields the highest power, but when the ROC curves cross each other it looses its discriminatory capacity. For the other scenarios it seems that the method based on Antoch's proposal has higher power but, as it did not calibrate the nominal level desired under the null hypothesis, it is not appropriate to compare its power with the other methodologies. The same applies to the L_1 and L_2-measure statistics calibrated though permutation, or for the Kolmogorov-Smirnov sort of statistic,

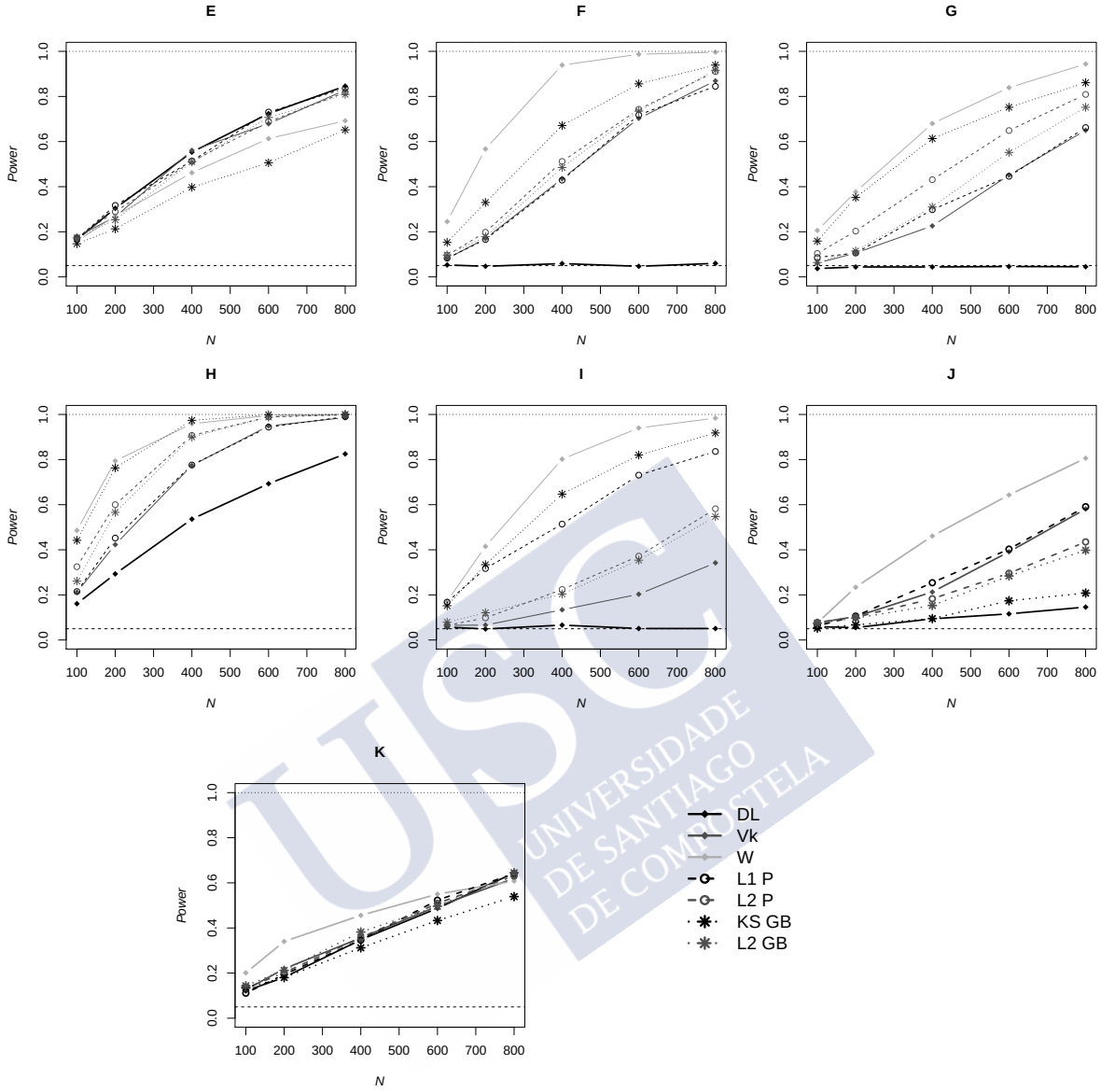


Figure 3.8: Observed statistical power for the different models considered, with $N = n_1 + m_1 + n_2 + m_2$.

whose good behaviour depends on the smoothing parameter. The methods whose power is appropriate to compare (the ones that reach the same nominal level) are, ultimately, DeLong's, Venkatraman's and the L_2-measure kind of statistic proposed by Martínez-Camblor and calibrated through the bootstrap general algorithm. Of these, the one who seems to obtain higher power is the Martínez-Camblor's method.

It is also worth mentioning that the results here obtained are consistent with the simulation studies that are carried out in the papers that are being analysed here, although those studies are limited to the case in which the ROC curves are binormal and they do not consider scenarios under the null hypothesis that are built with different cumulative distribution functions.

We would like to stress that this revision of procedures cannot be viewed as a guideline to choose the method that yields the lower p-value in a comparison study, for that can only result in a higher Type I error. The type of test employed should be selected beforehand depending on the nature of the data.

Furthermore, throughout this chapter we have limited our study to the comparison of diagnostic variables without any additional information. However, in a medical environment, along with the diagnostic markers, it is usual to have some other measurements, some covariates. In those cases it would be advisable to incorporate that information to the study, as the performance of the ROC curves can be affected by them.

Nevertheless, to our knowledge little has been done to combine the comparison between ROC curves and the presence of covariates. This study was motivated by the idea of finding an appropriate ROC curve comparison procedure to extend to the case with covariates. After looking at the results, the methodology developed in Martínez-Camblor et al. (2013) may be a good starting point to extend the ROC curve comparison to a scenario with covariates.

Chapter 4

Comparison of ROC curves with unidimensional covariates

Comparing the accuracy and the behaviour of different diagnostic procedures is one of the main objectives of the ROC curve analysis. As seen in the previous chapter, several procedures may be found in the literature concerning the comparison of two or more ROC curves. Along with the diagnostic variables it is usual to observe other covariates, but that extra information has been hardly ever considered for the comparison of this kind of curves. This information can be taken into account when constructing the ROC curves, usually by means of the conditional ROC curve, introduced in Chapter 2. It is known that the discriminatory capability of these curves may be influenced by this extra information, so it should be included in the study. With this idea, a new nonparametric test is proposed in this chapter for the comparison of conditional ROC curves. A bootstrap mechanism is used to calibrate the test and simulations are run to analyse the practical performance of the test in terms of level approximation and power. An application to real data is also presented to illustrate the procedure.

The main contents of this chapter are collected in Fanjul-Hevia et al. (2020a).

4.1 Introduction

In a biomedical environment, when there is more than one procedure for diagnosing a certain disease, one may wonder if one of the methods is more accurate than the others, or if it is equivalent to an alternative procedure (that could be less expensive or could be of lower risk for the patient). In the same line, it could be of interest to determine whether the diagnostic variable performs equally among different groups (e.g. different hospitals, different age range, different gender). These questions can be solved by comparing the ROC curves of the respective markers. In this chapter we move a step forward to include covariate information by testing the equality of conditional ROC curves.

When the hypothesis of equality of ROC curves is rejected it means that, even if their AUCs (or their pAUCs) were equal (and thus, some could argue that the different

procedures have the same global diagnostic ability) one should treat them differently: in those cases the sensitivity and specificity of each value of the diagnostic marker will not be the same for all the ROC curves considered. Detecting these situations is specially relevant when the objective of the study is to obtain the optimal cut-off value of each ROC curve. Furthermore, the comparison of whole ROC curves can also be of interest for designing new diagnostic variables obtained by the optimization of linear combinations of multiple markers (Su and Liu, 1993; Kim et al., 2015).

The problem of comparing ROC curves has been approached in many ways in the literature: some of these proposals are parametric, some others fully nonparametric. Some compare paired samples, others independent data, and others are prepared to compare more than two ROC curves. Some of these methodologies have been reviewed in the previous chapter, where a simulation study was run in order to compare their advantages and disadvantages. Behind all the methodologies there are two issues to handle: the type of test statistic that is used, and the calibration of its distribution under the null hypothesis (or at least, the procedures that are used to approximate that distribution). The most common procedure is based on the comparison of the AUCs (DeLong et al., 1988; Wieand et al., 1989; Molodianovitch et al., 2006; Martínez-Camblor, 2007; Yin et al., 2017) but, despite being one of the most powerful methods to compare ROC curves when one dominates the others at all points, it fails to reject the cases in which the curves cross each other. Other methods compare the difference among the ROC curves themselves (Venkatraman and Begg, 1996; Venkatraman, 2000; Antoch et al., 2010; Martínez-Camblor et al., 2011, 2013; Braga et al., 2013), approximating the distribution of the statistic by means of permutation or bootstrap algorithms.

Furthermore, along with the diagnosis variable it is usual to have some other covariates. As we discussed in Chapter 2, it is important to take this information into account, because the diagnostic capability of a marker can change with the value of a covariate. By considering the conditional ROC curve (2.7) it is possible to analyse how the performance of the maker is affected by this extra information.

There are different approaches in the literature for the estimation of the conditional ROC curve. Some of the methods estimate directly the conditional distribution functions (López-de Ullibarri et al., 2008; Inácio de Carvalho et al., 2013), while others introduce the covariate effect through some regression models. The latter are estimators based either on direct (Alonzo and Pepe, 2002; Rodríguez-Álvarez et al., 2011a) or on induced (González-Manteiga et al., 2011; Rodríguez-Álvarez et al., 2011b) regression methodologies.

On the other hand, there are also proposals based on Bayesian methodologies (like Inácio de Carvalho et al., 2013; Inácio de Carvalho et al., 2017; de Carvalho et al., 2020).These and other methods of estimation of conditional ROC curves are reviewed in Pardo-Fernández et al. (2014).

Bearing this in mind, the possible effect of the covariates should also be incorporated in the analysis when comparing different methods of diagnosis. ROC curves that seemed equivalent could be different when conditioned to the value of a certain covariate, or the other way around. To our better knowledge, little has been done in the literature

to address this problem. Some authors have dealt with the related issue of determining whether the considered covariate has a significant effect on the ROC curve (Rodríguez-Álvarez et al., 2011b, 2018), but nothing has been done for the comparison of two or more conditional ROC curves.

The main objective of this chapter is to propose a nonparametric methodology to compare two or more ROC curves conditioned to the value of one continuous covariate in the context of independent populations. The proposal of González-Manteiga et al. (2011) detailed in (2.11) in Chapter 2 is used here to estimate the conditional ROC curves, although now we will have K curves to estimate instead of one. The new methodology is presented in Section 4.2, which includes a bootstrap algorithm for approximating the distribution of the proposed statistic. The results from a simulation study are shown in Section 4.3. The database of Pleural Effusion introduced in Chapter 1 is analysed in Section 4.4 to illustrate the procedure, and it is followed by some conclusions (Section 4.5).

4.2 Testing methodology

In this section we explain the new methodology developed in order to compare K conditional ROC curves. We extend the idea behind Martínez-Camblor et al. (2011, 2013), based on the fact that the ROC curves can be viewed as cumulative distribution functions. The advantages of using this method for comparing ROC curves as a starting point are the following: (1) it uses the whole curve to make the comparison (and not just a functional of the curve), (2) it can be adapted to the cases in which the ROC curves are either paired or independent, and (3) it is able to compare more than two ROC curves at once.

First we will present the test statistic, then we show the asymptotic properties of the statistic and finally we propose a bootstrap algorithm to approximate the p-values.

4.2.1 The test statistic

Let Y_k^F and Y_k^G be the continuous diagnostic variables in the diseased and the healthy populations, respectively, of the k-th curve, for $k \in \{1, \ldots, K\}$. Let X_k^F and X_k^G be the covariates associated with those populations, and R_X their common support (supposed to be non-empty). For each $k \in \{1, \ldots, K\}$, given $x \in R_X$ we have the k-th conditional ROC curve:

$$ROC_k^x(p) = 1 - F_k(G_k^{-1}(1-p|x)|x), \quad 0 < p < 1,$$

where $F_k(y|x) = P(Y_k^F \leqslant y | X_k^F = x)$, $G_k(y|x) = P(Y_k^G \leqslant y | X_k^G = x)$ and $G_k^{-1}(\cdot|x)$ is the conditional quantile function of the healthy population.

Given $x \in R_X$, our objective is to test

$$H_0 : ROC_1^x(p) = \cdots = ROC_K^x(p) \quad \text{for all } p \in (0,1). \tag{4.1}$$

against the general alternative $H_1 : H_0$ is not true.

For each $k = 1, \ldots, K$, we will have independent samples:

- $\{(X_{k,i}^F, Y_{k,i}^F)\}_{i=1}^{n_k^F}$ an i.i.d. sample from the distribution of (X_k^F, Y_k^F),
- $\{(X_{k,i}^G, Y_{k,i}^G)\}_{i=1}^{n_k^G}$ an i.i.d. sample from the distribution of (X_k^G, Y_k^G),

with n_k^F and n_k^G the sample size of the k-th diseased and healthy populations, respectively. Define, for $k \in \{1, \ldots, K\}$, $N_k = n_k^F + n_k^G$.

The following test statistic is proposed:

$$S_N^x = \sum_{k=1}^{K} \psi\left(\sqrt{g_k N_k}\{\widehat{ROC_k}^x(p) - \widehat{ROC}_\bullet^x(p)\}\right), \tag{4.2}$$

where:

- for $k \in \{1, \ldots, K\}$, $\widehat{ROC_k}^x$ is the estimated k-th conditional ROC curve given in Chapter 2 (2.11) following the methodology in González-Manteiga et al. (2011),
- $g_k = \frac{n_k^F g_k^F + n_k^G g_k^G}{N_k}$, where g_k^F and g_k^G are the bandwidth parameters related to the estimation of the conditional means and variances of the location-scale regression models of the diseased and the healthy populations involved in the estimation of the k-th conditional ROC curve. Likewise, for each one of the estimated ROC curves there will be another bandwidth parameter, h_k, responsible for the smoothness of the estimator,
- $\widehat{ROC}_\bullet^x(p) = \left(\sum_{k=1}^K g_k N_k\right)^{-1} \sum_{k=1}^K g_k N_k \widehat{ROC_k}^x(p)$ is a sort of weighted average for all the K conditional ROC curves,
- ψ is a non-negative real-valued function that will measure the difference from one estimated ROC curve to the weighted average of all of them. This function may be similar to the ones used for the comparison of cumulative distribution functions (after all, a ROC curve can be viewed as a cumulative distribution function). For example, consider:
  - $S_{N,L2}^x = \sum_{k=1}^K g_k N_k \int \left(\widehat{ROC_k}^x(p) - \widehat{ROC}_\bullet^x(p)\right)^2 dp$, based on the L_2-measure,
  - $S_{N,KS}^x = \sum_{k=1}^K \sqrt{g_k N_k} \sup_p \left|\widehat{ROC_k}^x(p) - \widehat{ROC}_\bullet^x(p)\right|$, based on the Kolmogorov-Smirnov criterion.

The null hypothesis will be rejected for large values of S_N^x.

4.2.2 Theoretical properties of the statistic

Before exploring the asymptotic distribution of the proposed statistic, let us introduce an expression that coincides with the statistic S_N^x as long as the null hypothesis holds:

$$T_N^x = \sum_{k=1}^{K} \psi\left(\sqrt{g_k N_k}\{(\widehat{ROC}_k^x(p) - \widehat{ROC}_\bullet^x(p)) - (ROC_k^x(p) - ROC_\bullet^x(p))\}\right),$$

where $ROC_\bullet^x(p) = \left(\sum_{k=1}^{K} g_k N_k\right)^{-1} \sum_{k=1}^{K} g_k N_k ROC_k^x(p)$ for $0 < p < 1$. This expression can be rewritten as

$$T_N^x = \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} \sqrt{g_j N_j}\alpha_{kj}(N)\{\widehat{ROC}_j^x(p) - ROC_j^x(p)\}\right), \tag{4.3}$$

where $\alpha_{kj}(N) = I(k = j) - \sqrt{g_k N_k}\sqrt{g_j N_j}\left(\sum_{i=1}^{K} g_i N_i\right)^{-1}$. For each $k, j \in \{1, \dots K\}$, $\alpha_{kj}(N)$ converges to a certain $c_{kj} \in \mathbb{R}$ under some conditions given in Appendix A.1. Note that, in general, T_N^x cannot be computed, as it depends on the unknown theoretical conditional ROC curves.

The equivalence between T_N^x and S_N^x under the null hypothesis is of great utility for obtaining the asymptotic properties of the statistic. It is also used in the bootstrap algorithm proposed in the next section for the approximation of the critical values of the test.

Our main theoretical result is presented below. The assumptions required and the proofs are given in detail in Appendix A.1.

Theorem 4.1. *Assume that (A1)-(A4) hold and that the K groups are independent. Then, under the null hypothesis (4.1),*

$$S_N^x \xrightarrow{\mathcal{L}} \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} c_{kj} W_j^x(p)\right)$$

where $c_{kj} \in \mathbb{R}$ is the limit of $\alpha_{kj}(N)$, and $W_j^x(p)$ is the Gaussian process at which converges the process $(N_j g_j)^{1/2}\left\{\widehat{ROC}_j^x(p) - ROC_j^x(p)\right\}$, for $j \in \{1, \dots, K\}$, as stated in González-Manteiga et al. (2011).

4.2.3 The bootstrap algorithm

In the previous section we have obtained the asymptotic distribution of the test statistic under the null hypothesis. However, it includes the true distribution of the populations involved in each of the compared ROC curves, something that in practice will be unknown and difficult to estimate. Therefore, the distribution is not useful in practice in order to approximate critical values or, equivalently, the p-values. Alternatively, a bootstrap algorithm is suggested here to approximate a p-value for the proposed test:

1. From the original sample $(\{(X^F_{k,i}, Y^F_{k,i})\}_{i=1}^{n^F_k}$ and $\{(X^G_{k,i}, Y^G_{k,i})\}_{i=1}^{n^G_k}$ for $k = 1, \ldots, K)$, compute the test statistic value (4.2), that we will denote by s^x_N.

2. For $b = 1, \ldots, B$ and $k = 1, \ldots, K$,

   (i) For each $D \in \{F, G\}$, let $\left\{\varepsilon^{D,b*}_{k,i}\right\}_{i=1}^{n^D_k}$ be an i.i.d. sample from the empirical cumulative distribution function of $\left\{\hat{\varepsilon}^D_{k,i}\right\}_{i=1}^{n^D_k}$.

   (ii) Reconstruct bootstrap samples $\{(X^D_{k,i}, Y^{D,b*}_{k,i})\}_{i=1}^{n^D_k}$, for each $D \in \{F, G\}$, where $Y^{D,b*}_{k,i} = \hat{\mu}^D_k(X^D_{k,i}) + \hat{\sigma}^D_k(X^D_{k,i})\varepsilon^{D,b*}_{k,i}$[1].

3. Compute the test statistic in the bootstrap samples, $t^{x,b*}_N$, for $b = 1, \ldots, B$ using (4.3):

$$t^{x,b*}_N = \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} \sqrt{g_j N_j}\alpha_{kj}(N)\{\widehat{ROC}^{x,b*}_j(p) - \widehat{ROC}^{x}_j(p)\}\right),$$

   where $\widehat{ROC}^{x,b*}_j$ is the estimated k-th conditional ROC curve of the b-th bootstrap sample.

4. The distribution of S^x_N under the null hypothesis is approximated by $\{t^{x,1*}_N, \ldots, t^{x,B*}_N\}$, and the p-value is approximated by

$$p - value = \frac{1}{B}\sum_{b=1}^{B} I(s^x_N \leqslant t^{x,b*}_N).$$

As it happened with the bootstrap algorithms previously discussed in Chapters 2 and 3, this bootstrap algorithm has a particularity that the usual bootstrap in testing setups does not have: the null hypothesis (4.1) is not employed in the generation of the bootstrap samples (Step 2), where each one of them is generated independently from the corresponding empirical distribution and regression functions. Instead, it is used in the construction of the bootstrap statistic (Step 3), by exchanging the role of S^x_N for T^x_N (note that, under the null hypothesis, both expressions are equivalent). This is because, in the field of ROC curves, we may have information about the distribution of the healthy and the diseased populations separately, but not about the ROC curve itself. Thus, it is difficult to generate samples under the hypothesis of equality of K ROC curves. It is true that if we assume that the cumulative distribution functions of the K diseased and the K healthy populations are the same, we ensure obtaining the same K ROC curves, but by doing that we are not considering all the possible scenarios of the null hypothesis, as

[1]Note that the conditional mean and the conditional variance functions, as well as the residuals, come from a location-scale regression model similar to the one introduced in Chapter 2 (2.9), with the difference that here we have K pairs of regression models.

$F_1 = \cdots = F_K$ and $G_1 = \cdots = G_K$ is a sufficient but not a necessary condition to obtain K equal ROC curves. In the simulation study that is described in the following section, Scenario C is an example of this kind of null hypothesis in which the equal ROC curves that are being compared are constructed differently.

This algorithm is based on the general bootstrap algorithm proposed in Martínez-Camblor and Corral (2012) for hypothesis testing, method designed to preserve the data structure. This idea is used in Martínez-Camblor et al. (2013) for the comparison of ROC curves in the case without covariates.

Remark 4.1. The previous methodology was thought for a scenario in which the compared ROC curves are independent. However, in practice it is usual the situation in which there is some dependence structure between the diagnostic variables that are being compared, since those variables are sometimes measured on the same individuals.

In such cases, the samples at our disposal would be of the form $\{(X_i^F, Y_{1,i}^F, \ldots, Y_{K,i}^F)\}_{i=1}^{n^F}$ and $\{(X_i^G, Y_{1,i}^G, \ldots, Y_{K,i}^G)\}_{i=1}^{n^G}$. Note that here $n^F = n_k^F$ and $n^G = n_k^G$ for all $k \in \{1, \ldots, K\}$, and that the residuals $\left(\varepsilon_1^F, \ldots, \varepsilon_K^F\right)$ and $\left(\varepsilon_1^G, \ldots, \varepsilon_K^G\right)$ would now follow a K-dimensional distribution with zero mean and a covariance matrix with ones in the diagonal.

In this case we propose a modification on the previous bootstrap algorithm, in particular a change in Step 2:

2. For $b = 1, \ldots, B$, generate the bootstrap samples $(\{(X_i^F, Y_{1,i}^{F,b*}, \ldots, Y_{K,i}^{F,b*})\}_{i=1}^{n^F}$ and $\{(X_i^G, Y_{1,i}^{G,b*}, \ldots, Y_{K,i}^{G,b*})\}_{i=1}^{n^G})$ as follows:

   (i) For each $D \in \{F, G\}$, let $\left\{\left(\varepsilon_{1,i}^{D,b*}, \ldots, \varepsilon_{K,i}^{D,b*}\right)\right\}_{i=1}^{n^D}$ be an i.i.d. sample from the empirical joint cumulative distribution function of the original residuals.

   (ii) Reconstruct the bootstrap samples $\{(X_i^D, Y_{1,i}^{D,b*}, \ldots, Y_{K,i}^{D,b*})\}_{i=1}^{n^D}$ for each $D \in \{F, G\}$, where $Y_{k,i}^{D,b*} = \hat{\mu}_k^D(X_i^D) + \hat{\sigma}_k^D(X_i^D)\varepsilon_{k,i}^{D,b*}$.

4.3 Simulations

In order to analyse the performance of this new methodology, simulations were run for the comparison of several independent conditional ROC curves. Those curves were all drawn from location-scale regression models like the ones presented in Chapter 2 (2.9). In Table 4.1 the different conditional mean and conditional standard deviation functions employed to construct all the ROC curve used throughout this section are displayed. Note that several of the curves considered contain non-linear expressions.

The regression errors ε^F and ε^G were considered to have standard normal distribution for all the ROC curves except for ROC_5^x, for which ε^F is constructed as a standardization of a exponential variable of mean 2 and ε^G as a standardization of a exponential variable of mean $1/2$. In all the scenarios the covariates X^F and X^G are uniformly distributed on $[0, 1]$. Thus, the value of the covariate x at which the conditional ROC curves will be compared will be in $(0, 1)$. Particularly, the comparisons are made for $x \in \{0.25, 0.5, 0.75\}$.

Table 4.1: Conditional mean and conditional standard deviation functions of the conditional ROC curves considered in the simulation study.

	Regression functions	**Conditional standard deviation functions**
ROC_1^x $\vdots$ ROC_4^x	$\mu^F(x) = cx + \sin(0.5\pi x),\ \mu^G(x) = 0.5x^2$ where: $c = 0 \Rightarrow ROC_1^x \qquad c = 1 \Rightarrow ROC_3^x$ $c = 0.5 \Rightarrow ROC_2^x \qquad c = 1.5 \Rightarrow ROC_4^x$	$\sigma^F(x) = 0.5 + 0.5x,$ $\sigma^G(x) = 0.5 + 0.5x$
ROC_5^x	$\mu^F(x) = 1 + 3x,$ $\mu^G(x) = 0.25 + x + 0.5x^2$	$\sigma^F(x) = 1 + 2x,$ $\sigma^G(x) = 0.25 + 0.5x$
ROC_6^x	$\mu^F(x) = 1 + 2x,$ $\mu^G(x) = 2x^2$	$\sigma^F(x) = 0.5 + x,$ $\sigma^G(x) = 0.5 + 0.5x$
ROC_7^x	$\mu^F(x) = 2 + 2x,$ $\mu^G(x) = 1 + 2x^2$	$\sigma^F(x) = 0.5 + x,$ $\sigma^G(x) = 0.5 + 0.5x$
ROC_8^x	$\mu^F(x) = \frac{2+4x+4x^3}{1+2x},$ $\mu^G(x) = 1 + 2x^2$	$\sigma^F(x) = 0.5,$ $\sigma^G(x) = \frac{1+x}{2+4x}$
ROC_9^x	$\mu^F(x) = x,$ $\mu^G(x) = 0.5x^2$	$\sigma^F(x) = 0.5x,$ $\sigma^G(x) = x$
ROC_{10}^x	$\mu^F(x) = x - 0.5x^2 + 2\sin(\pi x),$ $\mu^G(x) = 2\sin(\pi x)$	$\sigma^F(x) = \frac{\sqrt{19}}{4}x,$ $\sigma^G(x) = 0.25x$
ROC_{11}^x	$\mu^F(x) = 3x + 2.5x^2,$ $\mu^G(x) = 2x + 3x^2$	$\sigma^F(x) = \frac{\sqrt{11}}{4}x,$ $\sigma^G(x) = 0.75x$

The first four ROC curves are similar except for a constant c. Depending on its value, there will be scenarios either under the null or under the alternative hypothesis. On the other hand, note that ROC_6^x, ROC_7^x and ROC_8^x are equal despite being constructed differently (for example, the regression functions of ROC_7^x are a translation of the regression functions of ROC_6^x, but, being the ROC curve invariant to any monotone transformation, they share the same final ROC curve formula). Finally, ROC_9^x, ROC_{10}^x and ROC_{11}^x are different ROC curves that have the same conditional AUC.

The reason for taking into consideration so many curves is to explore different scenarios that could be of interest in order to analyse both the level and the power of the test proposed in this chapter. In these simulations we include comparisons among two and three conditional ROC curves (i.e., $K = 2$ and $K = 3$) and different sample sizes. For simplicity, the same sample sizes were considered for each one of the curves (i.e. $(n^F, n^G) = (n_k^F, n_k^G)$ for each $k \in \{1, \ldots, K\}$ and thus $N_1 = \cdots = N_K$). Similar results were obtained for unbalanced sample sizes (check Appendix B.1.3 for some examples of simulations with different sample sizes).

Moreover, two different functions ψ were considered for the construction of the statistic S_N^x: one based on the L_2-measure and the other one based on the Kolmogorov-Smirnov criterion (from now on denoted by L_2 and KS respectively). The simulations were run for 200 and 500 bootstrap iterations, yielding similar results. Here we show the outcome of

Table 4.2: Computational cost (in seconds) of the test for different sample sizes for $K = 2$ and for $K = 3$ groups. B represents the number of bootstrap iterations considered.

		$K = 2$					$K = 3$				
B	N_k:	200	400	600	800	1000	200	400	600	800	1000
200		1.41	3.43	6.69	13.59	17.08	2.08	5.17	10.01	20.05	24.64
500		3.14	7.62	14.74	31.11	38.17	4.64	11.52	22.84	45.77	56.78

the test for 500 bootstrap samples (to see the results for analogous simulations with 200 bootstrap samples, go to Appendix B.1.2). Furthermore, 1000 datasets were simulated in each scenario.

As for the bandwidth parameters that are needed for the estimation of the ROC curves, they were selected as follows:

- h_k was taken as $1/\sqrt{N_k}$ for each $k \in \{1, \ldots, K\}$. Its value does not seem to have a significant effect neither on the conditional ROC curve estimation, nor on the tests. Other values were also considered for this parameter, including h_k=0, which represents the non-continuous estimator of the conditional ROC curve, described in Chapter 2 (2.10). Check Appendix B.1.1 to see a comparison of the obtained results of running simulations using different ways to determine h_k.

- g_k^F and g_k^G were selected by least-squares cross-validation, using a grid of 25 points ranging from 0.1 to 0.5. Check Appendix B.1.1 for comments and other simulation studies on this matter.

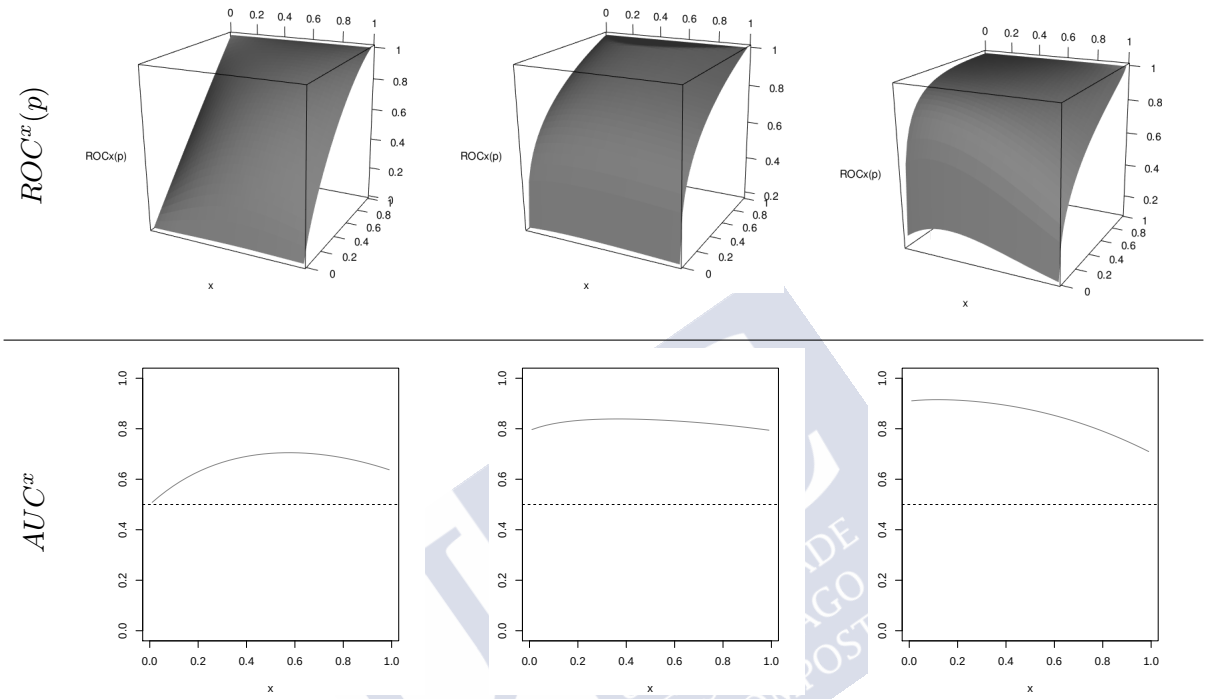
It is worth noticing that, for each bootstrap iteration, different conditional ROC curves are to be estimated, and thus, new bandwidth parameters should be selected. However, h_k remains the same, as the sample sizes do not change, and for computational issues the g_k^F and g_k^G parameters calculated for the original samples were also the ones employed for the bootstrap estimations. This prevents the simulations to become unfeasible due to the computational cost of cross-validation methods. For the data application on Section 4.4 we use this bandwidth selection, although different bandwidths could be used for each iteration of the bootstrap algorithm in practical applications. In Table 4.2 there is a summary of the computational cost of this test for the comparison of 2 and 3 conditional ROC curves for different sample sizes. The time was measured in a computer with Intel(R) Core(TM) i5-4590 CPU, 3.30GHz and 8 GB of RAM. Each measurement represents the time it takes to run a single test.

4.3.1 Level of the test

In order to check if the test is well calibrated, three different scenarios are considered under the null hypothesis (4.1). The number of conditional ROC curves that are compared are

Table 4.3: Scenarios under the null hypothesis considered for calibrating the level of the test.

K	**Scenario A**	**Scenario B**	**Scenario C**
2	ROC_1^x vs. ROC_1^x	ROC_5^x vs. ROC_5^x	ROC_6^x vs. ROC_7^x
3	ROC_1^x vs. ROC_1^x vs. ROC_1^x	ROC_5^x vs. ROC_5^x vs. ROC_5^x	ROC_6^x vs. ROC_7^x vs. ROC_8^x



$K = 2$ and $K = 3$ in each scenario. All of them are summarized and represented in Table 4.3, along with their corresponding AUC^x.

In scenario A the null hypothesis holds, as each conditional ROC curve is constructed with equal regression and conditional variance functions in their corresponding healthy and diseased populations. The same applies for scenario B, with the distinction that, in this case, the errors considered do not follow normal distributions, but exponential ones. The third scenario is a very particular case that is hardly ever considered when comparing ROC curves, despite being a feasible scenario in practice: the regression and conditional variance functions of these curves are different (as it can be seen in Table 4.1) but, nevertheless, the final conditional ROC curves coincide.

The null hypothesis is tested for different values of x: 0.25, 0.5 and 0.75. Keeping in mind that the covariates X^F and X^G are considered as uniformly distributed on $[0, 1]$, it is expected that the test will show a better behaviour at points that are not close to the border of that interval (i.e., for $x = 0.5$). The values closer to 0 or 1 will have less data around, and thus it is reasonable to expect that the corresponding estimation of the conditional ROC curves on those points will be not as good.

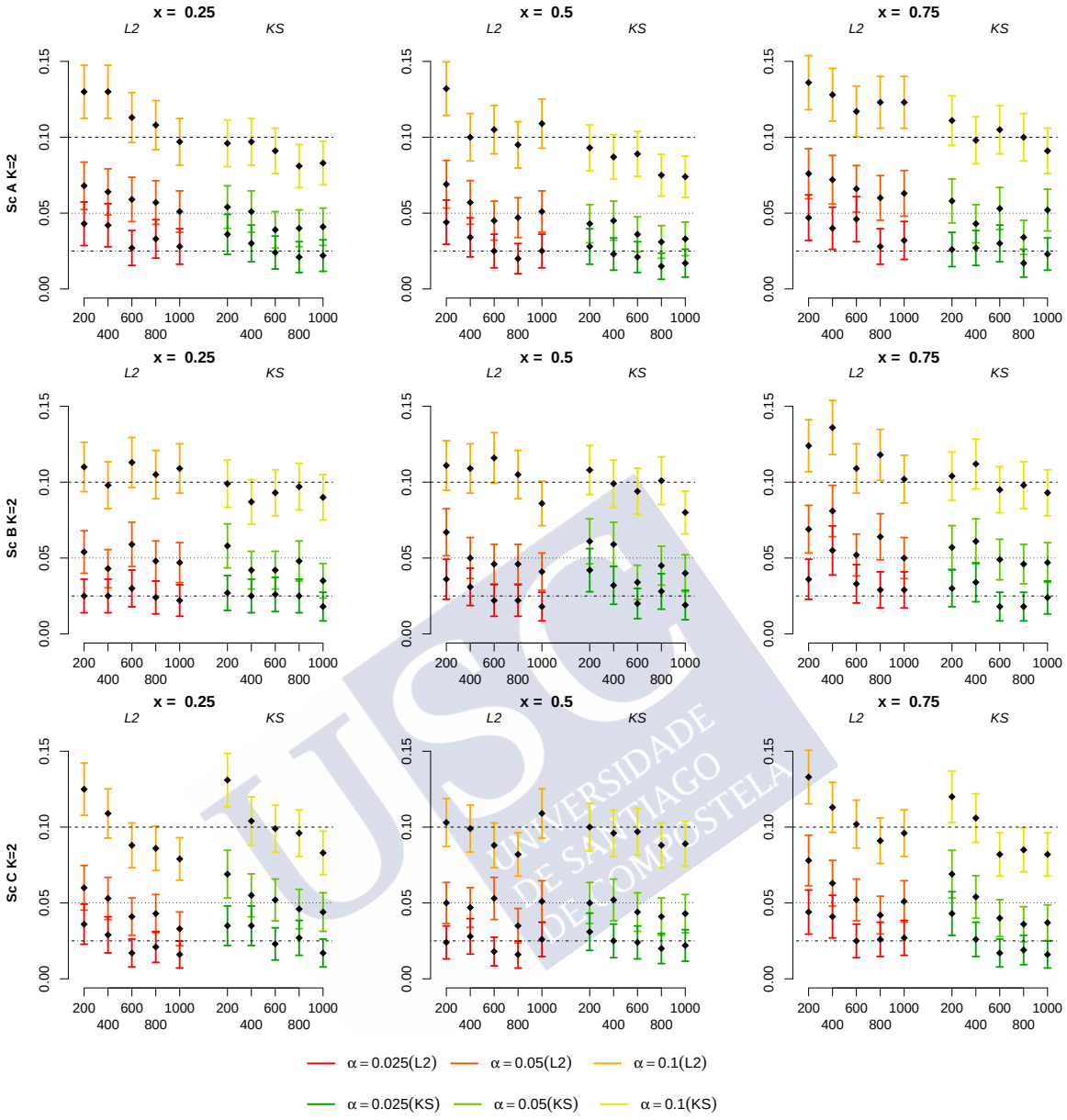


Figure 4.1: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis for Scenarios A, B and C with $K = 2$ for different values of the covariate. Each diagram contains the results for the statistic based on L_2 and KS and for different sample sizes N_k, where $N_k = n_k^F + n_k^G$.

The results of the simulations are summarized in Figures 4.1 (for $K = 2$) and 4.2 (for $K = 3$). Each subfigure represents the test of one scenario for a particular value of the covariate. The nominal levels considered are $\alpha \in \{0.025, 0.05, 0.1\}$. The estimated proportion of rejections over 1000 replications of the datasets is represented along with its confidence interval for each nominal level. The sample sizes considered are $(n^F, n^G) \in \{(100, 100), (150, 250), (300, 300), (550, 250), (500, 500)\}$.

In general it can be said that the expected nominal level is reached, as most of the confidence intervals of the estimated proportions contain the corresponding nominal level,

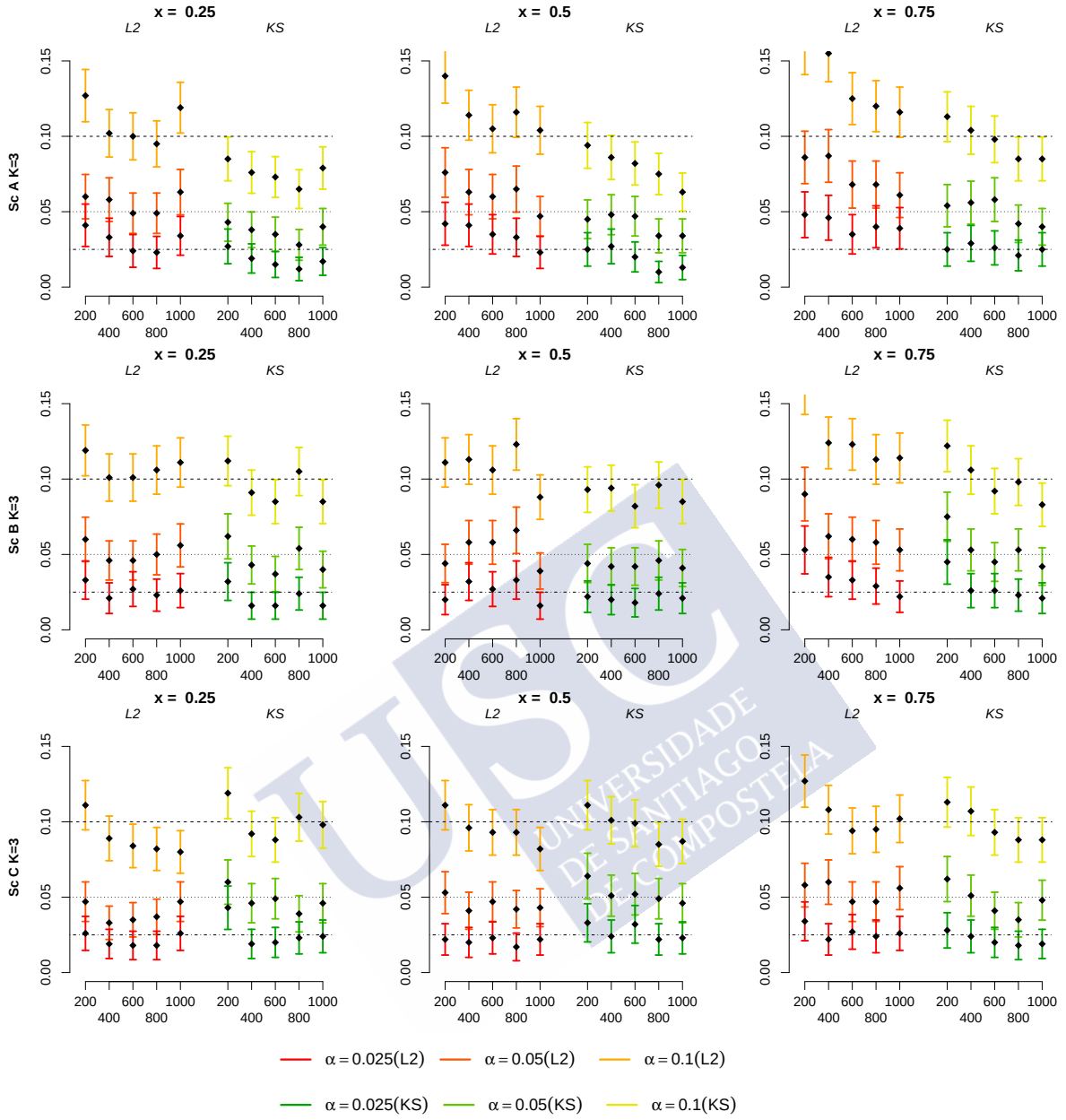


Figure 4.2: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis for Scenarios A, B and C with $K = 3$ for different values of the covariate. Each diagram contains the results for the statistic based on L_2 and KS and for different sample sizes N_k, where $N_k = n_k^F + n_k^G$.

specially when we make the comparison for $x = 0.5$.

The two kinds of statistics considered (L_2 and KS) behave similarly, KS being a little more conservative. It seems that for smaller sample sizes KS is a better option, while L_2 is better for greater sample sizes. Difference of sample sizes on the healthy and diseased populations (i.e., $n^F \neq n^G$) does not seem to have an effect here, except perhaps for the fourth instance, $(550, 250)$. Note that in this case the samples are heavily unbalanced, and thus an inaccurate behaviour in the error rate approximation can be expected. Moreover,

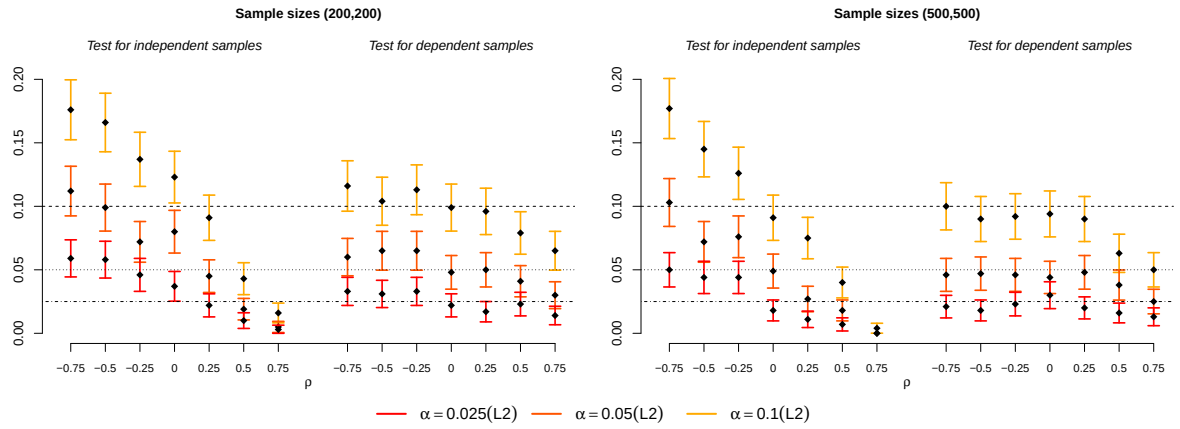


Figure 4.3: Estimated proportion of rejection under the null hypothesis for two different sample sizes $(n^F, n^G) = \{(200, 200), (500, 500)\}$ *and different levels of dependence between the compared ROC curves,* $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$*. Each diagram depicts the results of two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the* L_2 *statistic.*

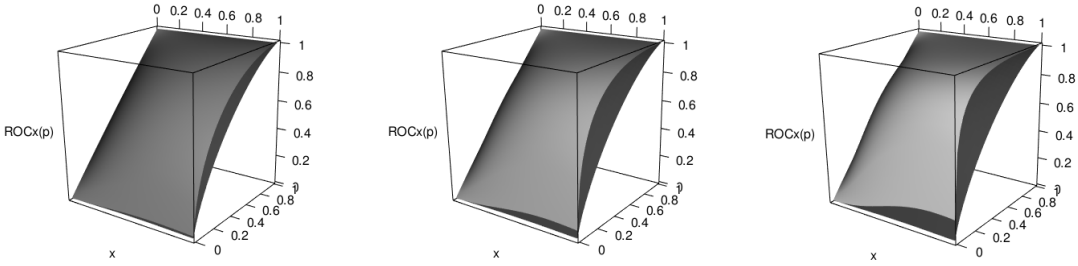
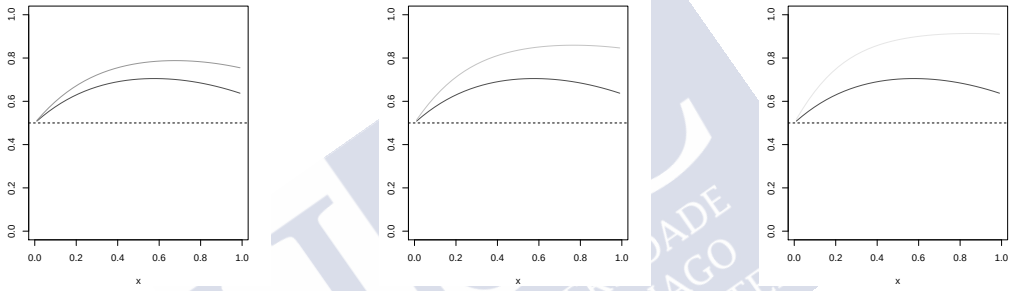
the results obtained for the comparison of two curves ($K = 2$) are similar to the ones obtained for the comparison of three curves ($K = 3$).

Remark 4.2. As pointed out in Remark 4.1, sometimes the diagnostic variables used for the construction of ROC curves have a dependence structure. Here we show what happens with the level of the test when that dependence is not taken into account (meaning, when we use the methodology we have proposed for testing independent ROC curves). We compare it with the level of the test that uses the modification of the bootstrap algorithm proposed in Remark 4.1 to this end.

The scenario considered here was similar to the one used in Scenario A, comparing ROC_2^x vs. ROC_2^x , with the exception that the regression errors in the two ROC curves now follow a binormal distribution with means zero, variances one and correlation ρ (this happens for the healthy and the diseased regression errors). We try the test for different values of ρ, which represent the degree of correlation between the curves that are being compared. In Figure 4.3 we show the results of this brief simulation study performing the test for $x = 0.5$ for several sample sizes. In this case we have only used the L_2 statistic and 200 bootstrap samples. A more extensive simulation study concerning this particular matter can be find at Appendix B.1.4.

Looking at the results it is clear that, when using the test for independent samples in this scenario, the level is highly overestimated for the negative correlations and heavily underestimated for the positive ones. This effect of the correlation is substantially corrected when using the test with the modified version of the bootstrap that takes into account the dependence structure, although for high values of ρ this test is still conservative.

Table 4.4: Scenarios under the alternative hypothesis, with $K = 2$, considered for analysing the power of the test.

K	**Scenario A.1**	**Scenario A.2**	**Scenario A.3**
2	ROC_1^x vs. ROC_2^x	ROC_1^x vs. ROC_3^x	ROC_1^x vs. ROC_4^x
$ROC^x(p)$			
AUC^x			

4.3.2 Power of the test

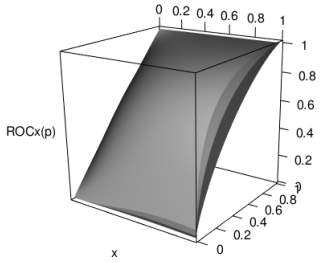
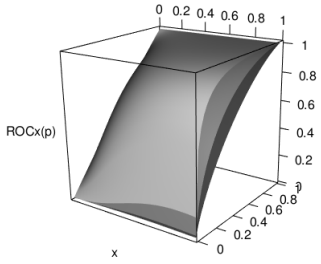
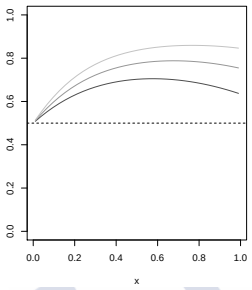
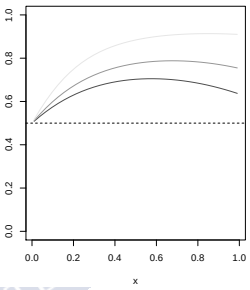
We now propose different scenarios under the alternative hypothesis. Thus, here we compare K conditional ROC curves that are not equal, for $K = 2$ and for $K = 3$.

First, we consider the ROC curve that we had in Scenario A (ROC_1^x), and we compare it with ROC_2^x, ROC_3^x or ROC_4^x (separately). Those ROC curves differ from ROC_1^x in a constant c of the regression function in such way that, the higher the value of c gets, the more different is going to be the curve with regard to ROC_1^x. Different values for c are considered for the comparison of 2 and 3 curves. These scenarios are called $A.1$–$A.5$ and are represented in Table 4.4 and Table 4.5.

Furthermore, we would like to highlight that the analysis could be simplified if we just compared the conditional AUCs. However, in this case we could be failing to differentiate ROC curves like ROC_6^x, ROC_7^x or ROC_8^x: conditional curves that, despite having the same AUC, are quite different from each other. The scenarios constructed with these ROC curves, $D.1$ and $D.2$ can be seen in detail in Table 4.6.

The simulations were run for the same statistics, the same covariate values and the same sample sizes ($(n^F, n^G) \in \{(100, 100), (150, 250), (300, 300), (550, 250), (500, 500)\}$) as in the previous section. However, in this occasion we only show results for one significance level: $\alpha = 0.05$. The results of these simulations are collected in Figure 4.4.

Table 4.5: Scenarios under the alternative hypothesis, with $K = 3$ considered for analysing the power of the test.

K	**Scenario A.4**	**Scenario A.5**
3	ROC_1^x vs. ROC_2^x vs. ROC_3^x	ROC_1^x vs. ROC_2^x vs. ROC_4^x
$ROC^x(p)$		
AUC^x		

We observe that the test is consistent, given that its power grows with the sample sizes, as it does when the difference between the compared ROC curves increases. The different proportions of rejection that were obtained for the different values of the covariate x are due to the fact that the curves that are being compared are more similar when x is closer to zero (as it can be easily seen looking at the conditional AUCs). However, if the distance between the curves was to remain constant for different values of x we would observe higher power for the points farther away from the limits of the interval (0,1) (i.e, for $x = 0.5$).

It can also be observed that the power obtained for the statistic that uses the L_2 measure is higher than the one obtained for the KS, something that goes along with the fact of KS being more conservative. This is not so clear for Scenarios $D.1$ and $D.2$, where it can be seen that this method is indeed able to differentiate conditional ROC curves with similar AUCs.

Furthermore, as happened before with the level of the test, uneven sample sizes for the diseased and healthy populations do not seem to affect its power significantly.

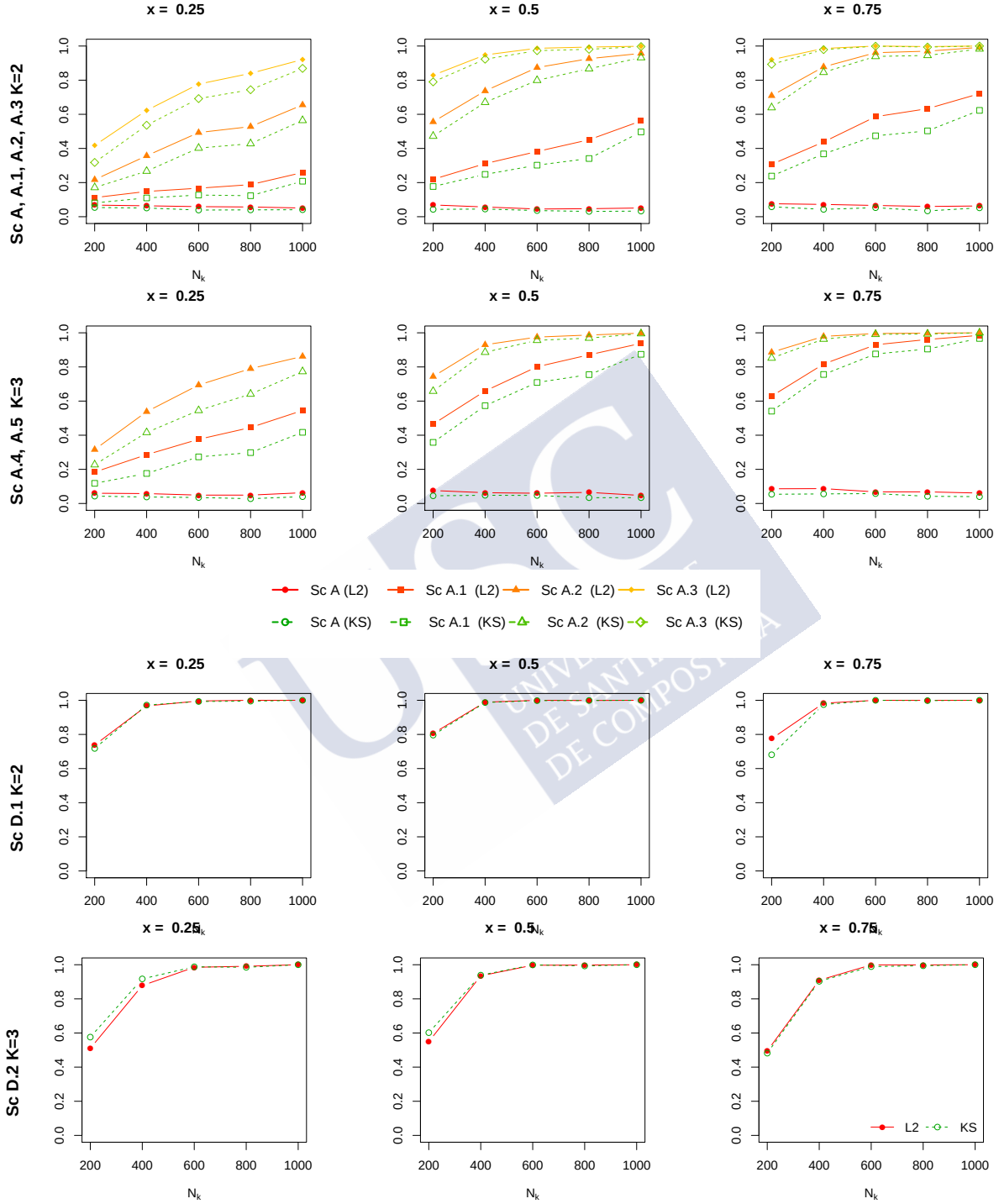
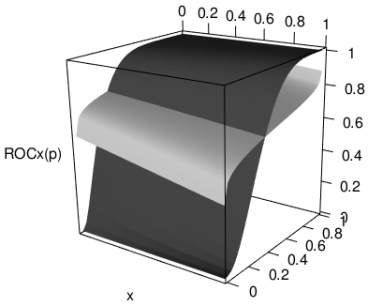
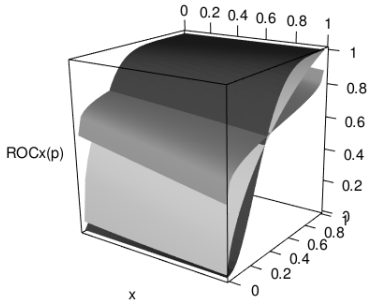
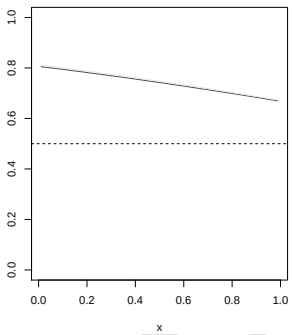
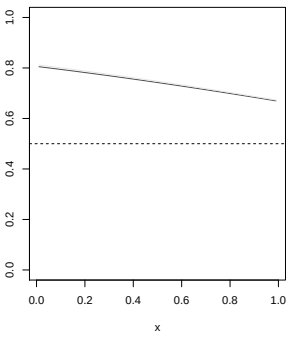


Figure 4.4: Estimated proportion of rejection under the alternative hypothesis for different sample sizes and different scenarios ($\alpha = 0.05$). Scenario A represents the situation under the null hypothesis.

Table 4.6: Scenarios under the alternative hypothesis, but with similar AUC, considered for analysing the power of the test.

K	**Scenario D.1**	K	**Scenario D.2**
2	ROC_9^x vs. ROC_{10}^x	3	ROC_9^x vs. ROC_{10}^x vs. ROC_{11}^x
$ROC^x(p)$		$ROC^x(p)$	
AUC^x		AUC^x	

4.4 Application to real data

An illustration of the proposed test is displayed in this section through the analysis of a dataset concerning patients with Pleural Effusion. These data have been introduced in Section 1.2.2, on Chapter 1.

The objective of analysing this database is to assess the performance of a tumour marker, the carbohydrate antigen (*CA153*), for cancer diagnosis. Along with this diagnostic marker and the variable that indicates if the pleural effusion is malignant or not we have two other covariates: the *age* and the *gender* of the patients.

One of the first questions we could ask ourselves is whether the *gender* of the patients has an effect on the discriminatory capability of the tumour marker. For illustrative purposes, 11 subjects were removed from the original dataset from the healthy population due to their atypically high levels of *CA153*. Thus the remaining dataset resulted in 480 individuals, 304 men (133 with cancer) and 176 women (78 with cancer). A summary of this data is shown in Table 4.7. The ROC curves corresponding to each one of those groups (and to the pooled data) are drawn in Figure 4.5.

Although the ROC curve of women seems to dominate the one of men (meaning that

Table 4.7: A summary of the variables from the Pleural Effusion dataset used in the study, by gender, for the MPE (D) and the non-MPE (H) subjects.

	age				*CA153*			
	Men		Women		Men		Women	
	D	H	D	H	D	H	D	H
Minimum	34.00	17.00	32.00	15.00	4.00	1.00	1.00	1.00
1st quartile	60.00	45.50	59.00	45.50	10.55	9.00	12.00	10.00
Median	73.00	63.00	72.00	67.50	18.00	13.17	32.27	14.46
Mean	69.35	60.89	68.62	62.01	53.95	14.87	111.78	15.18
3rd quartile	78.00	78.00	78.00	80.00	39.95	19.00	143.28	19.24
Maximum	95.00	93.00	90.00	94.00	493.00	36.28	836.70	37.34

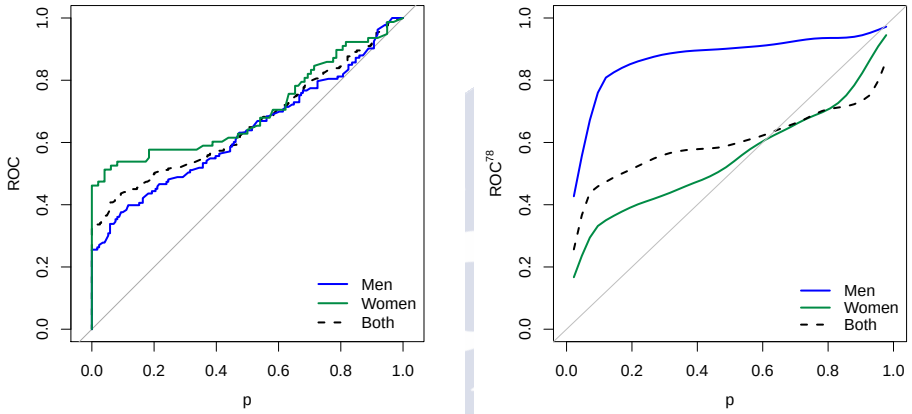


Figure 4.5: Estimated ROC curves for the men, for the women and for both groups, for all ages (to the left) and conditioned to the age 78 (to the right).

the discriminatory capability of the tumour marker seems higher for the women), the difference is not clear. When testing their equality with one of the methods for comparing ROC curves without covariates (e.g. DeLong's method for the comparison of AUCs), a p-value of 0.228 was obtained. Thus, we found no evidence that the tumour marker *CA153* works differently when diagnosing men or women.

However, if we suspect that the *age* of the subjects can affect the diagnosis, we should take it into account to make the comparison. In Figure 4.6 we may find a scatter plot representing the relation between age and *CA153* antigen for both genders, each of them classified as diseased (malignant pleural effusion) or healthy.

Barely any patient under 40 years of age had the disease. Both *genders* presented high values of *CA153* only from a certain *age*. The diseased population seems to behave similarly for both genders, while for the diseased population women present in general higher and more scattered *CA153* levels.

We could then consider the ROC curves conditioned to the value of the continuous covariate *age*. In Figure 4.7 we have the representation of the estimated conditional ROC curve for both groups, along with its conditional AUC with a 95 per cent pointwise

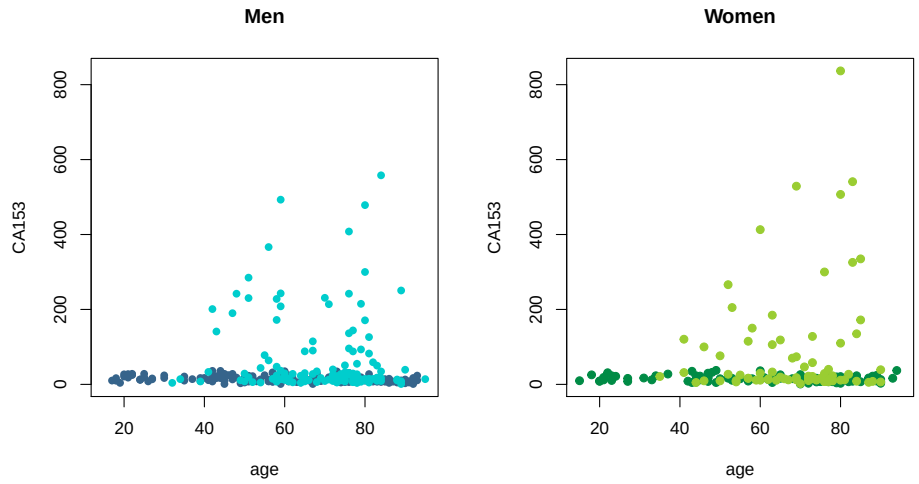


Figure 4.6: CA153 of Men and Women represented with respect the age of the subjects, for the healthy (darker colours) and the diseased (brighter colours) populations.

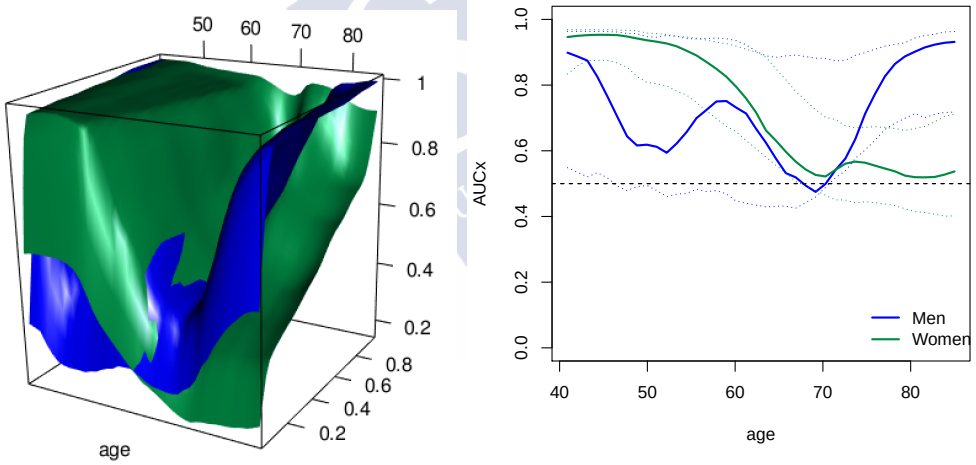


Figure 4.7: Conditional ROC curves and AUC (with their 95 per cent pointwise bootstrap confidence interval in dotted lines) for men and for women.

bootstrap confidence interval. We can observe that the way the ROC curve changes with the *age* seems quite different for men and for women. However, we must not forget that the *age* covariate is not uniformly distributed: the difference we observe for small and larger ages can be a consequence of having few individuals with those extreme cases.

When testing the ROC curves conditioned to the value of the three quartiles of the covariate *age* (53, 69 and 78) we obtain no significant difference for the first two (p-values of 0.130 for L_2 and 0.137 for KS at the age of 53, and p-values of 0.608 for L_2 and 0.452 for KS at median age of 69). However, we do reject the hypothesis of equality of the ROC curves with different *gender* at age 78 (p-values 0.002 for both L_2 and KS). Thus, we obtain different results when testing the equivalence of ROC curves with and without

covariates. The estimated conditional ROC curves are depicted in Figure 4.5. Note that, contrary to what we could expect, the curve estimated form the pooled data (conditioned or not) does not have to lie necessarily in between the curves estimated for the different genders of the population. Moreover, it is also worth noticing that, when conditioned to the *age* of 78, it is the ROC curve corresponding to men dominates that belonging to women, contrary to what happened in the case without covariates.

There are some goodness-of-fit tests for validating the location-scale regression models assumed in (2.9) (Einmahl and Van Keilegom, 2008). Despite not being the purpose of this chapter, we have checked the models for both the healthy and diseased populations in the case of males and females. For a significance level of $\alpha = 0.05$ none of the models were rejected.

4.5 Discussion

In this chapter we presented a new methodology for comparing two or more ROC curves conditioned to the value of one continuous covariate, a problem that had not been addressed thoroughly in the literature. Given that it only depends on fully nonparametric techniques, it is suited for a wide range of scenarios, as it has been shown in the simulation study. On the one hand, it is able to detect difference among curves regardless they have the same conditional AUC or not. On the other hand, it also accommodates correctly to the situations where conditional ROC curves are equal even when the underlying cumulative distribution functions differ.

The asymptotic behaviour of the statistic considered has been presented, although the selection of the bandwidth parameters involved in the statistic is still an open issue, as in other testing problems in nonparametric settings. In our simulation study we have explored the behaviour of our test based on cross-validation bandwidths.

Two different functions were proposed for the construction of the statistic, the L_2 and the KS, the second one being a little more conservative. Different sample sizes have been considered, including uneven ones without any appreciable effect on the test performance.

Furthermore, as it has been pointed out before, the method works better when the conditional value x that is being considered is far from the extremes of the support of the covariate. In the simulation study, the values tested were the three quartiles of the covariates X^F and X^G (which followed a uniform distribution on [0,1]). Special caution must be taken when the selected x falls outside the interquartile range of the covariate at hand.

The methodology was illustrated by means of an application to a dataset, which showed that comparing ROC curves with or without taking covariates into consideration may lead to different conclusions.

Although the test proposed in this chapter is designed for the case in which there is independence among the diagnostic markers that are being compared, an extension to the dependent case, in which the diagnostic markers could be correlated, is possible

(as discussed in Remarks 4.1 and 4.2). Furthermore, the proposed methodology can only handle one covariate, which may limiting in practical studies where more than one covariate may be available. An extension for a setup with multiple covariates will be presented in the next chapter, where we will be using random projections (an idea applied, for example, by Escanciano, 2006, in a regression context) to reduce the dimension of the problem (and, thus, avoiding the need for more complex or demanding assumptions and models, such as generalized additive models).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 5

Comparison of ROC curves with multidimensional covariates

We have already seen in the previous chapters that the comparison of ROC curves is frequently used in the literature to compare the discriminatory capability of different classification procedures based on diagnostic variables. The performance of these variables can be sometimes influenced by the presence of other covariates, and thus they should be taken into account when making the comparison. A new nonparametric test is proposed here for testing the equality of two or more dependent ROC curves conditioned to the value of a multidimensional covariate. Projections are used for transforming the problem into a one-dimensional approach easier to handle. Some simulations of the new methodology and an illustration on a real dataset are presented.

The main contents of this chapter are collected in Fanjul-Hevia et al. (2020b).

5.1 Introduction

Throughout the previous chapters we have seen the usefulness of the comparison of ROC curves. In Chapter 3 we have discussed several methodologies that can be found in the literature for making that sort of comparisons (without covariates). In Chapter 4 we have seen a way of introducing the effect of the covariates into the study by using the conditional ROC curve. Given that in practice it is usual to have more than one covariate along with the diagnostic variables, in this chapter our goal is to propose a test to compare ROC curves that includes the presence of a multidimensional covariate in the analysis.

Let us consider Y^F and Y^G as the continuous diagnostic markers in the diseased and healthy populations, respectively, $\boldsymbol{X}^F = (X_1^F, \ldots, X_d^F)'$ as the continuous d-dimensional covariate of the diseased population and $\boldsymbol{X}^G = (X_1^G, \ldots, X_d^G)'$ as the continuous d-dimensional covariate of the healthy population, then, given a fixed value $\boldsymbol{x} = (x_1, \ldots, x_d)'$ of $\boldsymbol{R_X}$ (where $\boldsymbol{R_X}$ is the intersection of $\boldsymbol{R_{X^F}}$ and $\boldsymbol{R_{X^G}}$, the supports of $\boldsymbol{X}^F$ and $\boldsymbol{X}^G$, and is assumed to be non-empty), the conditional ROC curve (in the case of being conditioned

by a multidimensional covariate) is defined as

$$ROC^{\boldsymbol{x}}(p) = 1 - F(G^{-1}(1-p|\boldsymbol{x})|\boldsymbol{x}),\ p \in (0,1), \tag{5.1}$$

where $F(y|\boldsymbol{x}) = P(Y^F \leqslant y|\boldsymbol{X}^F = \boldsymbol{x})$, and $G(y|\boldsymbol{x}) = P(Y^G \leqslant y|\boldsymbol{X}^G = \boldsymbol{x})$.

By comparing these conditional ROC curves instead of the standard ROC curves it is possible to incorporate the potential effect of the covariates in the analysis of the equivalence of two or more methods of diagnosis. A test for performing this comparison is proposed in the previous chapter for the case of a continuous one-dimensional covariate. The objective here is to extend that methodology to the case in which we have a multidimensional covariate. Thus, the aim is to test, given a certain $\boldsymbol{x} \in \boldsymbol{R_X}$,

$$H_0 : ROC_1^{\boldsymbol{x}}(p) = \cdots = ROC_K^{\boldsymbol{x}}(p) \ \text{ for all } p \in (0,1), \tag{5.2}$$

where K is the number of diagnostic markers (and thus, ROC curves) that are being compared. In this context we would have K diagnostic variables and one d-dimensional covariate in the healthy population, $(\boldsymbol{X}^F, Y_1^F, \cdots, Y_K^F)$, and similar variables in the diseased population, $(\boldsymbol{X}^G, Y_1^G, \cdots, Y_K^G)$. We will assume that there is a dependence among the ROC curves (meaning the covariate is common for all the K curves considered).

In order to be able to make this comparison, we are going to rely on the estimation of the corresponding conditional ROC curves. As already explained in Chapter 3, there is a wide range of estimation methods in the literature: some of them estimate the conditional distribution functions involved in the definition of the conditional ROC curve, others use regression functions to include the effect of the covariates (following direct or indirect approaches).

In Chapter 4 the estimation of the conditional ROC curve that is used is based on the indirect (or induced) regression methodology. This approach incorporates the covariate information through regression models by considering the effect of those covariates in the diagnostic marker in each population of healthy or diseased separately. However, this method was originally designed for one single covariate. One could think of extending that methodology by changing the estimator of the conditional ROC curve for another capable of handling multidimensional covariates. Nevertheless, there are not many methods in the literature capable of considering more than one covariate when estimating the conditional ROC curve, and most of them have some parametric assumptions that we would like to avoid making. See Inácio de Carvalho et al. (2013) as an example of a nonparametric Bayesian model to estimate the conditional distribution functions involved in the ROC curves, Rodríguez-Álvarez et al. (2011a) as an example of a direct ROC curve regression model or Rodríguez and Martínez (2014) as an example of induced methodology (framed in a Bayesian setting).

The goodness-of-fit tests related to multidimensional data tend to become less powerful when the dimension of the problem increases. This is why, in this chapter, the problem of comparing conditional ROC curves is first transformed using projections in such a way

that the multidimensional problem becomes a unidimensional problem easier to handle. This idea has been applied several times in the literature for reducing the dimension in goodness-of-fit problems (see, for example, Escanciano, 2006; García-Portugués et al., 2014; Patilea et al., 2016). In the last few years random projections are increasingly being used as a way to overcome the curse of dimensionality. The characterization of the multidimensional distribution of the original data by the distribution of the randomly projected unidimensional data is what allows for the reduction of the dimension.

To that end, in Section 5.2 we show how the problem in (5.2) can be transformed in a test with one-dimensional covariates by using projections. Then, a methodology is proposed for testing that equivalent hypothesis. In Section 5.3 the results from a simulation study show the practical performance of the test in terms of level approximation and power. An application to real data is also presented to illustrate the procedure in Section 5.4.

5.2 Methodology

This section is divided in three parts. In the first one, 5.2.1, we present a result that allows us to transform the problem discussed in (5.2) into an equivalent one, easier to handle, by using projections to reduce the multidimensional role of the covariate to a unidimensional one.

In Section 5.2.2 we show a methodology to test the equality of conditional ROC curves on a unidimensional problem, very similar to the one proposed in the previous chapter. Finally, in Section 5.2.3, we combine that methodology with the result obtained in Section 5.2.1 to solve our original problem with multidimensional covariates. Both Sections 5.2.2 and 5.2.3 include the statistic proposed to perform the test and a bootstrap algorithm to approximate its distribution.

5.2.1 An equivalent problem

In order to present the transformation of the problem, first we need to introduce the definition of *the ROC curve conditioned to a pair* $(x^F, x^G) \in R_{X^F} \times R_{X^G}$:

$$ROC^{x^F,x^G}(p) = 1 - F(G^{-1}(1-p|x^G)|x^F),\ p \in (0,1). \tag{5.3}$$

This concept is very similar to the conditional ROC curve (5.1): the only difference is that this new definition allows us to condition on different values for the diseased and healthy populations. In this case x^F and x^G are unidimensional, but the definition could be applied on a multidimensional case. Even if the interpretability of this new ROC curve is not very clear in practice, theoretically it does not present any problems (as it will not do its estimation), as the population of healthy and diseased are always considered to be independent.

The following result is the base for developing the test for comparing ROC curves with multidimensional covariates. It is based on the idea used in Escanciano (2006) of using projections for reducing the number of dimensions of the covariate in a regression context. Given $\boldsymbol{\beta}, \boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{\beta}'\boldsymbol{x}$ denotes the scalar product of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{x}$. For now on, all the vectors representing the projections will be considered to be contained in the d-dimensional unit sphere $\mathbb{S}^{d-1} = \{\boldsymbol{\beta} \in \mathbb{R}^d : ||\boldsymbol{\beta}|| = 1\}$. This way we ensure that all possible directions are equally important.

Lemma 5.1. *Assume $\mathbb{E}|Y_k^F| < \infty$ and $\mathbb{E}|Y_k^G| < \infty$ for every $k \in \{1, \ldots, K\}$. Then, given a certain $\boldsymbol{x} \in \boldsymbol{R_X}$, and assuming dependence among the ROC curves (meaning the covariate is common for all the K curves considered), then*

$$ROC_1^{\boldsymbol{x}}(p) = \cdots = ROC_K^{\boldsymbol{x}}(p) \quad \text{for all } p \in (0,1) \text{ a.s.}$$

if and only if

$$ROC_1^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) = \cdots = ROC_K^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) \quad \text{for all } p \in (0,1) \text{ a.s. for any } \boldsymbol{\beta}^F, \boldsymbol{\beta}^G,$$

where $\boldsymbol{\beta}^F$ and $\boldsymbol{\beta}^G$ are d-dimensional coordinates in $\mathbb{S}^{d-1}$ that represent the directions of the projections.

The proof of this Lemma can be found in Appendix A.2. Note that $(\boldsymbol{\beta}^F)'\boldsymbol{x}$ and $(\boldsymbol{\beta}^G)'\boldsymbol{x}$ are one-dimensional values. By using these ROC curves conditioned to a pair of projected covariates (as defined in (5.3)), the problem is reduced to a one-dimensional covariate conditional ROC curve comparison test for each possible direction $\boldsymbol{\beta}^F$ and $\boldsymbol{\beta}^G$.

Thus, given the result in Lemma 5.1, instead of testing for the null hypothesis (5.2), we may use this equivalent formulation to develop a test that, given a certain $\boldsymbol{x} \in R_{\boldsymbol{X}}$, tests

$$H_0 : ROC_1^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) = \cdots = ROC_K^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) \quad \text{for all } p \in (0,1) \ \forall \boldsymbol{\beta}^F, \boldsymbol{\beta}^G \quad (5.4)$$

against the general alternative $H_1 : H_0$ is not true. The notation $\forall$ is used instead of 'for any' to shorten the expression (this applies mainly in the proofs found in Appendix A.2).

In a first step, a statistic for testing the equivalence of these ROC curves is presented for a certain pair of fixed projections, and then that statistic is adapted to include all possible directions.

5.2.2 Test for a unidimensional covariate

The objective in this section is to develop a test for the equivalent problem presented in Lemma 5.1 for a fixed pair of projections $\boldsymbol{\beta}^F$ and $\boldsymbol{\beta}^G$. Here a test is presented for comparing two or more dependent ROC curves conditioned to two unidimensional values.

Given the pair $(x^F, x^G) \in R_{X^F} \times R_{X^G}$, the aim is then to test

$$H_0 : ROC_1^{x^F,x^G}(p) = \cdots = ROC_K^{x^F,x^G}(p) \text{ for all } p \in (0,1) \tag{5.5}$$

against the general alternative $H_1 : H_0$ is not true[1].

The samples available in this context are:

- $\{(X_i^F, Y_{1,i}^F, \ldots, Y_{K,i}^F)\}_{i=1}^{n^F}$ an i.i.d. sample from the distribution of $(X^F, Y_1^F, \ldots, Y_K^F)$,
- $\{(X_i^G, Y_{1,i}^G, \ldots, Y_{K,i}^G)\}_{i=1}^{n^G}$ an i.i.d. sample from the distribution of $(X^G, Y_1^G, \ldots, Y_K^G)$,

with n^F and n^G the sample sizes of the diseased and healthy populations, respectively. Define, for $k \in \{1, \ldots, K\}$, $N = n^F + n^G$. Unlike what happened in the previous chapter (where we compared independent ROC curves), the data of the diagnostic markers under comparison is drawn from the same sets of individuals. Therefore, the sample size used for the construction of each curve is always the same and thus we can avoid the use of the subindex k when denoting sample sizes. Note that both X^F and X^G are here one-dimensional covariates.

The method used for the estimation of the conditional ROC curves is based on the one proposed in González-Manteiga et al. (2011), which relies on nonparametric location-scale regression models. It is, in fact, very similar to the method used in Chapter 4 for comparing unidimensional covariates, but, for the sake of clarity, we include it below. To be more precise, for each $k = 1, \ldots, K$, assume that

$$Y_k^F = \mu_k^F(X^F) + \sigma_k^F(X^F)\varepsilon_k^F \tag{5.6}$$
$$Y_k^G = \mu_k^G(X^G) + \sigma_k^G(X^G)\varepsilon_k^G \tag{5.7}$$

where, for $D \in \{F, G\}$, $\mu_k^D(\cdot) = \mathbb{E}(Y_k^D | X^D = \cdot)$ and $(\sigma_k^D)^2(\cdot) = \mathbb{V}\text{ar}(Y_k^D | X^D = \cdot)$ are the conditional mean and the conditional variance functions (both of them unknown smooth functions), and the error ε_k^D is independent of X^D. The dependence structure between the K diagnostic variables is modelled by introducing a dependence structure between the errors: $(\varepsilon_1^D, \ldots, \varepsilon_K^D)$ will follow a multivariate distribution function with zero mean and a covariance matrix with ones in the diagonal.

Given this location-scale regression model structure for the diagnostic variables, the k-th ROC curve conditioned to a pair of values $(x^F, x^G) \in R_{X^F} \times R_{X^G}$ can be expressed in terms of the marginal cumulative distribution functions of the errors, H_k^F and H_k^G:

$$ROC_k^{x^F,x^G}(p) = 1 - H_k^F\left(\left(H_k^G\right)^{-1}(1-p)b_k(x^F, x^G) - a_k(x^F, x^G)\right), \tag{5.8}$$

[1]This test can be even more general: the values at which we are conditioning could also be different in each conditional ROC curve, meaning that we could be conditioning at the values $(x_1^F, x_1^G), \ldots, (x_K^F, x_K^G) \in R_{X^F} \times R_{X^G}$. In such case, the aim would be to test $H_0 : ROC_1^{x_1^F,x_1^G} = \cdots = ROC_K^{x_K^F,x_K^G}$. The generalization of the methodology is almost immediate. For further detail, check Appendix B.2.1.

where

$$a_k(x^F, x^G) = \frac{\mu_k^F(x^F) - \mu_k^G(x^G)}{\sigma_k^F(x^F)} \text{ and } b_k(x^F, x^G) = \frac{\sigma_k^G(x^G)}{\sigma_k^F(x^F)}.$$

Thus, this $k-th$ conditional ROC curve can be estimated by

$$\widehat{ROC}_k^{x^F,x^G}(p) = 1 - \int \hat{H}_k^F \left(\left(\hat{H}_k^G \right)^{-1} (1 - p + h_k u) \hat{b}_k(x^F, x^G) - \hat{a}_k(x^F, x^G) \right) \kappa(u) du, \quad (5.9)$$

where, for $D \in \{F, G\}$,

- $\hat{H}_k^D(y) = (n^D)^{-1} \sum_{i=1}^{n^D} I(\hat{\varepsilon}_{k,i}^D \leqslant y)$,
- $\hat{\varepsilon}_{k,i}^D = \dfrac{Y_{k,i}^D - \hat{\mu}_k^D(X_i^D)}{\hat{\sigma}_k^D(X_i^D)}$ for $i \in \{1, \cdots, n^D\}$,
- $\hat{\mu}_k^D(x) = \sum_{i=1}^{n^D} W_{k,i}^D(x, g_k^D) Y_{k,i}^D$ is a nonparametric estimator of $\mu_k^D(x)$ based on local weights $W_{k,i}^D(x, g_k^D)$ depending on a bandwidth parameter g_k^D,
- $(\hat{\sigma}_k^D)^2(x) = \sum_{i=1}^{n^D} W_{k,i}^D(x, g_k^D)[Y_{k,i}^D - \hat{\mu}_k^D(X_i^D)]^2$ is a nonparametric estimator of $(\sigma_k^D)^2(x)$. For simplicity we take the same bandwidth parameter g_k^D that is used for the estimation of the regression function $\hat{\mu}_k^D(x)$,
- $W_{k,i}^D(x, g_k^D) = \dfrac{\kappa_{g_k^D}(x - X_i^D)}{\sum_{l=1}^{n^D} \kappa_{g_k^D}(x - X_l^D)}$, for $i \in \{1, \cdots, n^D\}$, are Nadaraya-Watson type weights, where $\kappa_{g_k^D}(\cdot) = \kappa(\cdot/g_k^D)/g_k^D$ and κ is a probability density function symmetric around zero.
- $\hat{a}_k(x^F, x^G) = \left(\hat{\mu}_k^F(x^F) - \hat{\mu}_k^G(x^G)\right)/\hat{\sigma}_k^F(x^F)$ and $\hat{b}_k(x^F, x^G) = \hat{\sigma}_k^G(x^G)/\hat{\sigma}_k^F(x^F)$.
- h_k is a bandwidth parameter responsible for the smoothness of the estimator. Its value does not seem to have a significant effect on the conditional ROC curve estimation.

This way of estimating the conditional ROC curve is similar to the one proposed in González-Manteiga et al. (2011), with the difference that they condition the ROC curve on a single value x and here we have a pair of values x^F and x^G, each one of them related to the diseased and the healthy population, respectively. As both populations are independent, the adaptation of the methodology of González-Manteiga et al. (2011) to this case is straightforward.

Once we know how to estimate this doubly conditional ROC curve we can propose a test statistic for the test (5.5):

$$S^x = \sum_{k=1}^{K} \psi \left(\sqrt{g_k N} \{ \widehat{ROC}_k^{x^F,x^G}(p) - \widehat{ROC}_\bullet^{x^F,x^G}(p) \} \right), \quad (5.10)$$

where:

- for $k \in \{1, \ldots, K\}$, $g_k = \frac{n^F g_k^F + n^G g_k^G}{N}$, where g_k^F and g_k^G are bandwidth parameters involved in the estimation of the k-th conditional ROC curve.

- for $k \in \{1, \ldots, K\}$, $\widehat{ROC}_k^{x^F, x^G}(p)$ is the estimated conditional ROC curve given (x^F, x^G), as seen in (5.9),

- $\widehat{ROC}_\bullet^{x^F, x^G}(p) = \left(\sum_{k=1}^K g_k\right)^{-1} \sum_{k=1}^K g_k \widehat{ROC}_k^{x^F, x^G}(p)$ is a sort of weighted average of the K conditional ROC curves.

- ψ is a real-valued function that measures the difference between each estimated conditional ROC curve and the weighted average of all of them. For example, if one considers the L_2-measure, then the resulting test statistic is

$$S_{L2}^x = \sum_{k=1}^K g_k N \int \left(\widehat{ROC}_k^{x^F, x^G}(p) - \widehat{ROC}_\bullet^{x^F, x^G}(p)\right)^2 dp.$$

  On the other hand, when using the Kolmogorov-Smirnov criterion the resulting test statistic is

$$S_{KS}^x = \sum_{k=1}^K \sqrt{g_k N} \sup_p \left|\widehat{ROC}_k^{x^F, x^G}(p) - \widehat{ROC}_\bullet^{x^F, x^G}(p)\right|.$$

The null hypothesis will be rejected for large values of S^x. In order to obtain the distribution of this statistic, a bootstrap algorithm is proposed. This bootstrap algorithm is adapted from the procedure proposed in Martínez-Camblor and Corral (2012) and has been already used by Martínez-Camblor et al. (2013) in the context of ROC curves. In fact, it is a similar bootstrap that was used in Chapters 2, 3 and 4. The key of this algorithm is that

$$T^x = \sum_{k=1}^K \psi\left(\sqrt{g_k N}\left\{\left(\widehat{ROC}_k^{x^F, x^G}(p) - \widehat{ROC}_\bullet^{x^F, x^G}(p)\right) - \left(ROC_k^{x^F, x^G}(p) - ROC_\bullet^{x^F, x^G}(p)\right)\right\}\right),$$

coincides with the statistic S^x as long as the null hypothesis holds, where

$$ROC_\bullet^{x^F, x^G}(p) = \left(\sum_{k=1}^K g_k\right)^{-1} \sum_{k=1}^K g_k ROC_k^{x^F, x^G}(p),\ p \in (0, 1).$$

The quantity T^x can be rewritten as

$$T^x = \sum_{k=1}^K \psi\left(\sum_{j=1}^K \sqrt{g_j N}\alpha_{kj}\{\widehat{ROC}_j^{x^F, x^G}(p) - ROC_j^{x^F, x^G}(p)\}\right), \tag{5.11}$$

where $\alpha_{kj} = I(k = j) - \sqrt{g_k}\sqrt{g_j}\left(\sum_{i=1}^K g_i\right)^{-1}$. Note that, in general, T^x cannot be computed from the data, as it depends on the unknown theoretical conditional ROC curves, but it is useful when applying the bootstrap algorithm.

The bootstrap algorithm suggested to approximate a p-value for this test is the following:

A.1 From the original samples, $\{(X_i^F, Y_{1,i}^F, \ldots, Y_{K,i}^F)\}_{i=1}^{n^F}$ and $\{(X_i^G, Y_{1,i}^G, \ldots, Y_{K,i}^G)\}_{i=1}^{n^G}$, compute the test statistic value (5.10), that we will denote by s^x.

A.2 For $b = 1, \ldots, B$, generate the bootstrap samples $\{(X_i^F, Y_{1,i}^{F,b*}, \ldots, Y_{K,i}^{F,b*})\}_{i=1}^{n^F}$ and $\{(X_i^G, Y_{1,i}^{G,b*}, \ldots, Y_{K,i}^{G,b*})\}_{i=1}^{n^G}$ as follows:

(i) For each $D \in \{F, G\}$, let $\left\{\left(\varepsilon_{1,i}^{D,b*}, \ldots, \varepsilon_{K,i}^{D,b*}\right)\right\}_{i=1}^{n^D}$ be an i.i.d. sample from the empirical cumulative multivariate distribution function of the original residuals.

(ii) Reconstruct the bootstrap samples $\{(X_i^D, Y_{1,i}^{D,b*}, \ldots, Y_{K,i}^{D,b*})\}_{i=1}^{n^D}$ for each $D \in \{F, G\}$, where $Y_{k,i}^{D,b*} = \hat{\mu}_k^D(X_i^D) + \hat{\sigma}_k^D(X_i^D)\varepsilon_{k,i}^{D,b*}$.

A.3 Compute the test statistic based on the bootstrap samples, for $b = 1, \ldots, B$ using (5.11) as

$$t^{x,b*} = \sum_{k=1}^{K} \psi \left(\sum_{j=1}^{K} \sqrt{g_j N} \alpha_{kj} \{\widehat{ROC}_j^{x^F, x^G, b*}(p) - \widehat{ROC}_j^{x^F, x^G}(p)\} \right),$$

where $\widehat{ROC}_j^{x^F, x^G, b*}$ is the estimated j-th conditional ROC curve of the b-th bootstrap sample.

A.4 The distribution of S^x under the null hypothesis (and thus, the distribution of T^x) is approximated by the empirical distribution of the values $\{t^{x,1*}, \ldots, t^{x,B*}\}$ and the p-value is approximated by

$$p - value = \frac{1}{B} \sum_{b=1}^{B} I(s^x \leqslant t^{x,b*}).$$

In contrast with the usual bootstrap algorithms in testing setups, in this case the null hypothesis is not employed when generating of the bootstrap samples (Step A.2), because replicating the null hypothesis of equal ROC curves is not a straightforward problem. Instead, it is used in the computation of the bootstrap statistic (Step A.3) by using T^x instead of S^x, that are equal under the null hypothesis. This particularity also appears in the bootstrap algorithm of the next section.

There are two kinds of bandwidth parameters that appear in the estimation of the k-th conditional ROC curve (5.9), with $k \in \{1, \ldots K\}$. The first one, h_k, is taken as $1/\sqrt{N}$, and the second ones, g_k^F and g_k^G, are selected by least-squares cross-validation. Note that, for each bootstrap iteration, the bandwidth parameters could change, as their selection depends on the sample. However, h_k remains constant, as we are choosing it in terms of the sample size, and that is the same for each bootstrap iteration. As for g_k^F and g_k^G, for computational issues we have decided to compute them on step A.1 using the original sample, and then apply the same bandwidths for all the bootstrap estimations. The cross-validation method can be very time-consuming, and this simplification prevents the simulations to become infeasible.

5.2.3 Test for a multidimensional covariate

Once having seen a strategy for testing (5.4) for only one pair of fixed directions, the idea now is to modify the previous procedure so the new statistic takes into account all the possible directions that $\boldsymbol{\beta}^F$ and $\boldsymbol{\beta}^G$ can take. For that purpose, consider the test statistic

$$D_S^{\boldsymbol{x}} = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} d\boldsymbol{\beta}^F d\boldsymbol{\beta}^G, \tag{5.12}$$

where $d\boldsymbol{\beta}^F$ and $d\boldsymbol{\beta}^G$ represent the uniform density on the sphere of dimension d, $\mathbb{S}^{d-1}$. This ensures that all directions are equally important.

The expression $S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$ is equal to the statistic used in (5.10) for testing the equality of K ROC curves when conditioned to the value of the pair $\left((\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}\right)$, that is,

$$S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} = \sum_{k=1}^{K} \psi\left(\sqrt{g_k N}\{\widehat{ROC}_k^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) - \widehat{ROC}_\bullet^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p)\}\right).$$

Note that, in this context with multidimensional covariates, the samples at our disposal are $\{(\boldsymbol{X}_i^F, Y_{1,i}^F, \ldots, Y_{K,i}^F)\}_{i=1}^{n^F}$ and $\{(\boldsymbol{X}_i^G, Y_{1,i}^G, \ldots, Y_{K,i}^G)\}_{i=1}^{n^G}$, with $\boldsymbol{X}_i^F = (X_{1,i}^F, \cdots, X_{d,i}^F)'$ and $\boldsymbol{X}_i^G = (X_{1,i}^G, \cdots, X_{d,i}^G)'$.

In practice, as it is done in Colling and Van Keilegom (2017), to compute the test statistic $D_S^{\boldsymbol{x}}$ random directions $\boldsymbol{\beta}_1^F, \ldots, \boldsymbol{\beta}_{n_\beta}^F$ and $\boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_{n_\beta}^G$ are drawn uniformly from $\mathbb{S}^{d-1}$, where $n_{\boldsymbol{\beta}}$ is the number of random directions considered (the same number of directions is taken for $\boldsymbol{\beta}^F$ and for $\boldsymbol{\beta}^G$). With them, the approximated statistic is

$$\tilde{D}_S^{\boldsymbol{x}} = \frac{1}{n_{\boldsymbol{\beta}}^2} \sum_{r=1}^{n_{\boldsymbol{\beta}}} \sum_{l=1}^{n_{\boldsymbol{\beta}}} S^{(\boldsymbol{\beta}_r^F)'\boldsymbol{x},(\boldsymbol{\beta}_l^G)'\boldsymbol{x}}. \tag{5.13}$$

In order to obtain the distribution of the statistic, a bootstrap algorithm (similar to the one described in the previous section) is proposed. To do so, the following expression is introduced:

$$D_T^{\boldsymbol{x}} = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} T^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} d\boldsymbol{\beta}^F d\boldsymbol{\beta}^G, \tag{5.14}$$

where $T^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$ is the same as in (5.11), but for the conditioning values of $\left((\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}\right)$:

$$T^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} = \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} \sqrt{g_j N}\alpha_{kj}\{\widehat{ROC}_j^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) - ROC_j^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p)\}\right).$$

As it happened in (5.11), $T^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$ cannot be computed without knowing the true distribution of the diagnostic markers. However, it can be computed in the bootstrap algorithm below, and there $D_T^{\boldsymbol{x}}$ is approximated by

$$\tilde{D}_T^{\boldsymbol{x}} = \frac{1}{n_{\boldsymbol{\beta}}^2} \sum_{r=1}^{n_{\boldsymbol{\beta}}} \sum_{l=1}^{n_{\boldsymbol{\beta}}} T^{(\boldsymbol{\beta}_r^F)'\boldsymbol{x},(\boldsymbol{\beta}_l^G)'\boldsymbol{x}}. \tag{5.15}$$

As happened before, for two given projections $\boldsymbol{\beta}^F$ and $\boldsymbol{\beta}^G$, $S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$ and $T^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$

coincide as long as the null hypothesis holds, and thus the same happens with $D_S^{\boldsymbol{x}}$ and $D_T^{\boldsymbol{x}}$.

Taking into account these approximations, the resulting bootstrap algorithm goes as follows:

B.1 Draw $n_{\boldsymbol{\beta}}$ random directions $\boldsymbol{\beta}_1^F, \ldots, \boldsymbol{\beta}_{n_\beta}^F$ and $\boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_{n_\beta}^G$ uniformly from $\mathbb{S}^{d-1}$.

B.2 For each random directions $\boldsymbol{\beta}_r^F$ and $\boldsymbol{\beta}_l^G$ (with $r, l \in \{1, \ldots, n_{\boldsymbol{\beta}}\}$) , consider the sample $\left\{\left((\boldsymbol{\beta}_r^F)'\boldsymbol{X}_i^F, Y_{1,i}^F, \ldots, Y_{K,i}^F\right)\right\}_{i=1}^{n^F}$ and $\left\{\left((\boldsymbol{\beta}_l^G)'\boldsymbol{X}_i^G, Y_{1,i}^G, \ldots, Y_{K,i}^G\right)\right\}_{i=1}^{n^G}$ and the conditioning values $\left((\boldsymbol{\beta}_r^F)'\boldsymbol{x}, (\boldsymbol{\beta}_l^G)'\boldsymbol{x}\right)$. With them, following steps A.1–A.3 of the bootstrap algorithm of the previous subsection, compute the value of $s^{(\boldsymbol{\beta}_r^F)'\boldsymbol{x},(\boldsymbol{\beta}_l^G)'\boldsymbol{x}}$ and the B corresponding $t^{(\boldsymbol{\beta}_r^F)'\boldsymbol{x},(\boldsymbol{\beta}_l^G)'\boldsymbol{x},b*}$.

B.3 Compute $\tilde{d}_S^{\boldsymbol{x}} = \frac{1}{n_{\boldsymbol{\beta}}^2}\sum_{r=1}^{n_\beta}\sum_{l=1}^{n_\beta} s^{(\boldsymbol{\beta}_r^F)'\boldsymbol{x},(\boldsymbol{\beta}_l^G)'\boldsymbol{x}}$ and $\tilde{d}_T^{\boldsymbol{x},b*} = \frac{1}{n_{\boldsymbol{\beta}}^2}\sum_{r=1}^{n_{\boldsymbol{\beta}}}\sum_{l=1}^{n_{\boldsymbol{\beta}}} t^{(\boldsymbol{\beta}_r^F)'\boldsymbol{x},(\boldsymbol{\beta}_l^G)'\boldsymbol{x},b*}$ as in (5.13) and (5.15).

B.4 Approximate the p-value of the test by:

$$p - value = \frac{1}{B}\sum_{b=1}^{B} I(\tilde{d}_S^{\boldsymbol{x}} \leqslant \tilde{d}_T^{\boldsymbol{x},b*}).$$

Remark 5.1. Note that $n_{\boldsymbol{\beta}}$ represents the number of random directions drawn from $\mathbb{S}^{d-1}$ considered for the approximation of (5.13) and (5.15), but that, in fact, we are using $n_{\boldsymbol{\beta}}^2$ different combination of pairs $(\boldsymbol{\beta}^F, \boldsymbol{\beta}^G) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ to make that approximation. This could become an issue from the computational point of view, as the complexity of the problem increases very fast when increasing the value of $n_{\boldsymbol{\beta}}$.

As an alternative, we could consider using

$$D_S^{\boldsymbol{x}} = \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} d\boldsymbol{\beta}^F\boldsymbol{\beta}^G,$$

instead of statistic (5.12), where $d\boldsymbol{\beta}^F\boldsymbol{\beta}^G$ represents the uniform density on the torus of dimension d, $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. This ensures, as before, that all pairs of directions are equally important. Thus, in practice, instead of using the approximation (5.13) we could consider

$$\hat{D}_S^{\boldsymbol{x}} = \frac{1}{m_{\boldsymbol{\beta}}}\sum_{r=1}^{m_{\boldsymbol{\beta}}} S^{(\boldsymbol{\beta}_r^F)'\boldsymbol{x},(\boldsymbol{\beta}_r^G)'\boldsymbol{x}},$$

where $(\boldsymbol{\beta}_1^F, \boldsymbol{\beta}_1^G), \ldots, (\boldsymbol{\beta}_{m_\beta}^F, \boldsymbol{\beta}_{m_\beta}^G)$ are pairs of random directions drawn uniformly from $\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}$, and where $m_{\boldsymbol{\beta}}$ would represent here the same as $n_{\boldsymbol{\beta}}^2$ before, with the advantage that it allows for more flexibility because it can assume non-squared values. A similar adaptation could be applied for the approximation of $D_T^{\boldsymbol{x}}$ in (5.14).

Remark 5.2. In the literature there are articles such as Cuesta-Albertos et al. (2007) or Cuesta-Albertos et al. (2019) that use only one random projection. The main idea is to perform the test at hand for a randomly selected projection instead of for all possible projections. The use of projections results in a dimension reduction (as desired), and the use of one single projection results in a reduction of the computational cost.

Table 5.1: Conditional mean and conditional standard deviation functions of the conditional ROC curves considered in the simulation study.

$\boldsymbol{x}$	ROC curves	Regression functions	Conditional standard deviation functions
$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$	$ROC_1^{\boldsymbol{x}}$	$\mu_1^F(\boldsymbol{x}) = \sin(0.5\pi x_1) + 0.1x_2$ $\mu_1^G(\boldsymbol{x}) = 0.5x_1x_2$	$\sigma_1^F(\boldsymbol{x}) = 0.5 + 0.5x_1$ $\sigma_1^G(\boldsymbol{x}) = 0.5 + 0.5x_1$
	$ROC_2^{\boldsymbol{x}}$	$\mu_2^F(\boldsymbol{x}) = 0.3 + \sin(0.5\pi x_1) + 0.1x_2$ $\mu_2^G(\boldsymbol{x}) = 0.5x_1x_2$	$\sigma_2^F(\boldsymbol{x}) = 0.5 + 0.5x_1$ $\sigma_2^G(\boldsymbol{x}) = 0.5 + 0.5x_1$
	$ROC_3^{\boldsymbol{x}}$	$\mu_3^F(\boldsymbol{x}) = \sin(0.5\pi x_1) + 0.1x_2$ $\mu_3^G(\boldsymbol{x}) = -0.3 + 0.4x_2 + 0.5x_1x_2$	$\sigma_3^F(\boldsymbol{x}) = 0.5 + 0.5x_1$ $\sigma_3^G(\boldsymbol{x}) = 0.5 + 0.5x_1$
$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$	$ROC_4^{\boldsymbol{x}}$	$\mu_4^F(\boldsymbol{x}) = \sin(0.5\pi x_1) + 0.1x_2 + 0.5x_3$ $\mu_4^G(\boldsymbol{x}) = 0.5x_1x_2 + x_3$	$\sigma_4^F(\boldsymbol{x}) = 0.5 + 0.1x_3$ $\sigma_4^G(\boldsymbol{x}) = 0.5 + 0.1x_3$
	$ROC_5^{\boldsymbol{x}}$	$\mu_5^F(\boldsymbol{x}) = \sin(0.5\pi x_1) + 0.1x_2 + 0.5x_3$ $\mu_5^G(\boldsymbol{x}) = x_1x_2 + x_3$	$\sigma_5^F(\boldsymbol{x}) = 0.5 + 0.1x_3$ $\sigma_5^G(\boldsymbol{x}) = 0.5 + 0.1x_3$
	$ROC_6^{\boldsymbol{x}}$	$\mu_6^F(\boldsymbol{x}) = \sin(0.5\pi x_1) + 0.1x_2 + 0.5x_3$ $\mu_6^G(\boldsymbol{x}) = -0.3 + 0.5x_1x_2 + x_3$	$\sigma_6^F(\boldsymbol{x}) = 0.5 + 0.2x_2 + 0.3x_3$ $\sigma_6^G(\boldsymbol{x}) = 0.5 + 0.1x_3$

Following that idea, instead of testing the equality of covariate-projected ROC curves for all possible projections, we could test the equality of covariate-projected ROC curves for some random pair of projections given a certain $\boldsymbol{x} \in R_{\boldsymbol{X}}$, meaning:

$$H_0 : ROC_1^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} = \cdots = ROC_K^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} \text{ for some } \boldsymbol{\beta}^F, \boldsymbol{\beta}^G. \tag{5.16}$$

The equivalence between this hypothesis and the one of interest in this chapter given in (5.2) still needs theoretical justification. However, it is a possibility worth studying, if only for computational reasons. A way of perform this approach could be to consider the proposed methodology for $n_{\boldsymbol{\beta}} = 1$.

5.3 Simulations

In order to analyse the performance of the proposed methodology, simulations were run for the comparison of several dependent conditional ROC curves. On a first stage, these simulations were focused on analysing the behaviour of the unidimensional test described in Section 5.2.2, but we do not include them here, as they are very similar to the ones displayed in the previous chapter (see Appendix B.2.1 for more details on the subject). Instead, we show the results for several scenarios (first under the null hypothesis and then under the alternative) in which we compare K ROC curves (with $K \in \{2, 3\}$) conditioned to a d-dimensional covariate (with $d \in \{2, 3\}$).

All the curves used in the simulation study were drawn from location-scale regression models similar to the ones presented in (5.6) and (5.7), only that, in this case, the regression and the conditional standard deviation functions are for d-dimensional covariates. The construction of those curves is summarized in Table 5.1, where all the different conditional mean and conditional standard deviation functions are displayed.

The regression errors were considered to follow multivariate normal distributions with zero

mean, variance one and correlation ρ for all the models, that is, the correlation matrices for $K = 2$ and $K = 3$ are

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

respectively.

In all scenarios the covariates X_1^F, X_1^G, X_2^F, X_2^G, X_3^F and X_3^G are uniformly distributed in the unit interval. Thus, the value of the multidimensional covariate $\boldsymbol{x}$ at which the conditional ROC curves should be compared is contained in $[0,1]^d$. Particularly, the comparisons are made for $\boldsymbol{x} = (0.5, 0.6)'$ and for $\boldsymbol{x} = (0.5, 0.6, 0.5)'$, for $d = 2$ and $d = 3$, respectively.

The study contains simulations for different sample sizes $(n^F, n^G) \in \{(100,100), (250,150), (250,350)\}$ and different values of ρ that represent different possible degrees of correlation between the diagnostic variables under comparison ($\rho \in \{-0.5, 0, 0.5\}$). For the scenarios in which three curves are being compared, the variance-covariance matrix

$$\Sigma = \begin{pmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}$$

was also considered. Note that the different covariances considered for the regression errors represent different dependencies that may exist between the diagnostic variables that are being compared.

Moreover, two different functions ψ were considered for the construction of $S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$: one based on the L_2-measure and the other one based on the Kolmogorov-Smirnov criterion (from now on denoted by L_2 and KS respectively). The number of iterations used in the bootstrap algorithm was 200, and 500 datasets were simulated to compute the proportion of rejection in each scenario.

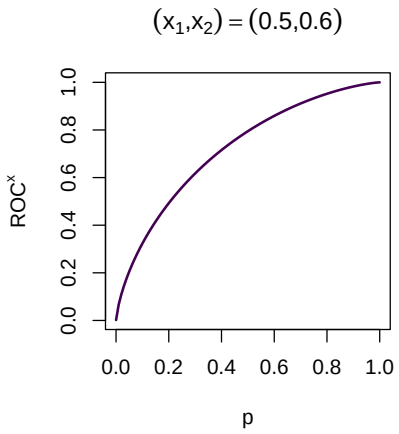
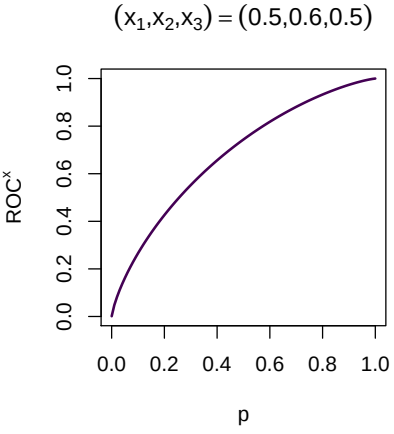
Furthermore, the number of directions that was used for approximating the test statistic $D_S^{\boldsymbol{x}}$ was taken as $n_{\boldsymbol{\beta}} = 5$ (as mentioned in Remark 5.1, notice that this means that $n_{\boldsymbol{\beta}}^2$ =25 different pairs of directions were considered). As it is explained in Section 5.2.3, $D_S^{\boldsymbol{x}}$ is approximated with $\tilde{D}_S^{\boldsymbol{x}}$ by drawing random directions uniformly from $\mathbb{S}^{d-1}$. However, for the case $d = 2$, in which $\mathbb{S}^1$ represents the unit circumference, we also tried to approximate $D_S^{\boldsymbol{x}}$ by generating $n_{\boldsymbol{\beta}}$ evenly spaced directions (something not always possible for $d > 2$ for any number of $n_{\boldsymbol{\beta}}$). The simulations obtained using this approximation are similar to the ones obtained for random directions, and can be observed in Appendix B.2.2. Note that this is just another way of approximating numerically the integrals of the test statistic at hand, and that the philosophy behind the methodology remains the same.

In Table 5.2 there is a summary of the computational cost of this test (with 200 bootstrap iterations and considering $n_{\boldsymbol{\beta}} = 5$) for the comparison of 2 and 3 conditional ROC curves with a covariate with dimension 2 and 3 for different sample sizes. The time was measured in a computer with Intel(R) Core(TM) i5-4590 CPU, 3.30GHz and 8 GB of RAM. Each measurement represents the time it takes to run a single test.

Table 5.2: Computational cost (in seconds) of the test for different sample sizes for $K = 2$ and for $K = 3$ groups and dimensions of the covariate $d = 2$ and $d = 3$. 200 bootstrap iterations were considered for each case.

		$K = 2$			$K = 3$		
d	N:	200	400	600	200	400	600
2		33.13	166.92	279.46	34.10	157.37	313.94
3		34.83	172.38	288.61	34.58	162.52	314.21

Table 5.3: Scenarios under the null hypothesis considered for calibrating the level of the test.

	2-dimensional covariate	**3-dimensional covariate**
$K = 2, 3$	ROC_1^x	ROC_4^x
		

5.3.1 Level of the test

The scenarios that were considered for calibrating the level of the test (by comparing the same conditional ROC curves) are represented in Table 5.3.

The results of the simulations obtained for $n_{\boldsymbol{\beta}} = 5$ are summarized in Figures 5.1 (for $d = 2$) and 5.2 (for $d = 3$). Each subfigure represents the test of one scenario for a particular sample size. The nominal levels considered are $\alpha \in \{0.025, 0.05, 0.1\}$. The estimated proportion of rejections over 500 replications of the datasets is represented along with its confidence interval for each nominal level.

In general it can be said that the expected nominal level is reached, as most of the estimated proportions are close to the corresponding nominal level. The L_2 statistic seems to overestimate the level in a few scenarios, but its behaviour improves when increasing the sample size. The KS statistic is a little more conservative.

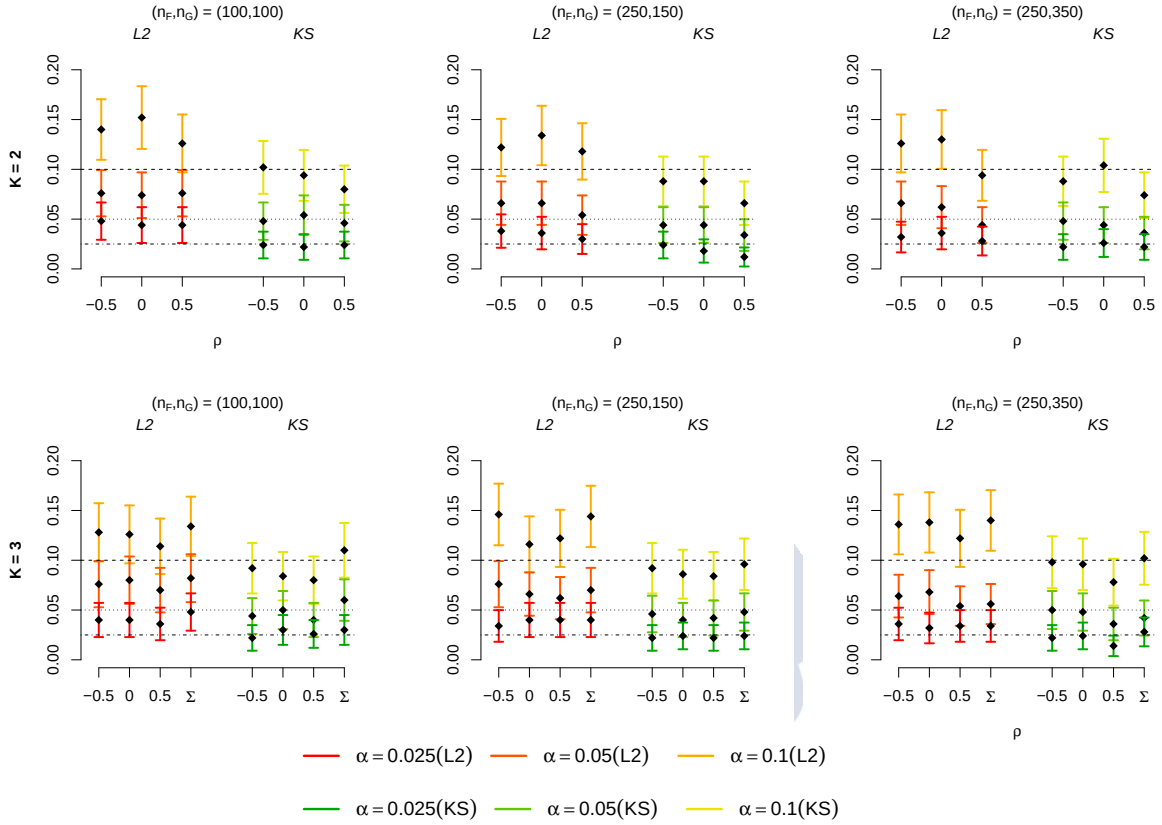


Figure 5.1: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis with $d = 2$ and $n_{\boldsymbol{\beta}} = 5$ for different sample sizes and different values of ρ and for the correlation matrix Σ.

5.3.2 Power of the test

On the other hand, the scenarios that were considered for studying the power of the test (by comparing different conditional ROC curves) are represented in Table 5.4.

The results of the simulations are summarized in Figures 5.3 (for $n_{\boldsymbol{\beta}} = 5$). In those figures the first and second row represent the simulation results for the scenarios with $K = 2$ and $K = 3$, respectively, and the first and the second column represent the simulation results for $d = 2$ and for $d = 3$, respectively. In this case, only $\alpha = 0.05$ was considered.

It can be seen that the power of the test grows with the considered sample sizes. The L_2 statistic yields higher power than the KS statistic, which is consistent with KS being more conservative. Moreover, the difference between the conditional ROC curves considered for the case of $d = 2$ is bigger than the difference between the ROC curves in the scenarios with $d = 3$, which translates in higher power for the cases in which $d = 2$.

We can also observe that for each scenario, the highest power is always obtained for the cases in which the correlation of the diagnostic variables is $\rho = 0.5$, and the lowest for $\rho = -0.5$.

Note that for the scenario with $d = 3$ and $\rho = -0.5$ the power of the test does not increase significantly from the first sample size to the second (in fact, for $K = 3$ it even decreases a little). This can be due to the fact that the case with smaller sample sizes presents balanced data, (n^F, n^G) being $(100, 100)$, whereas for the second set, the considered sample sizes are $(250, 150)$.

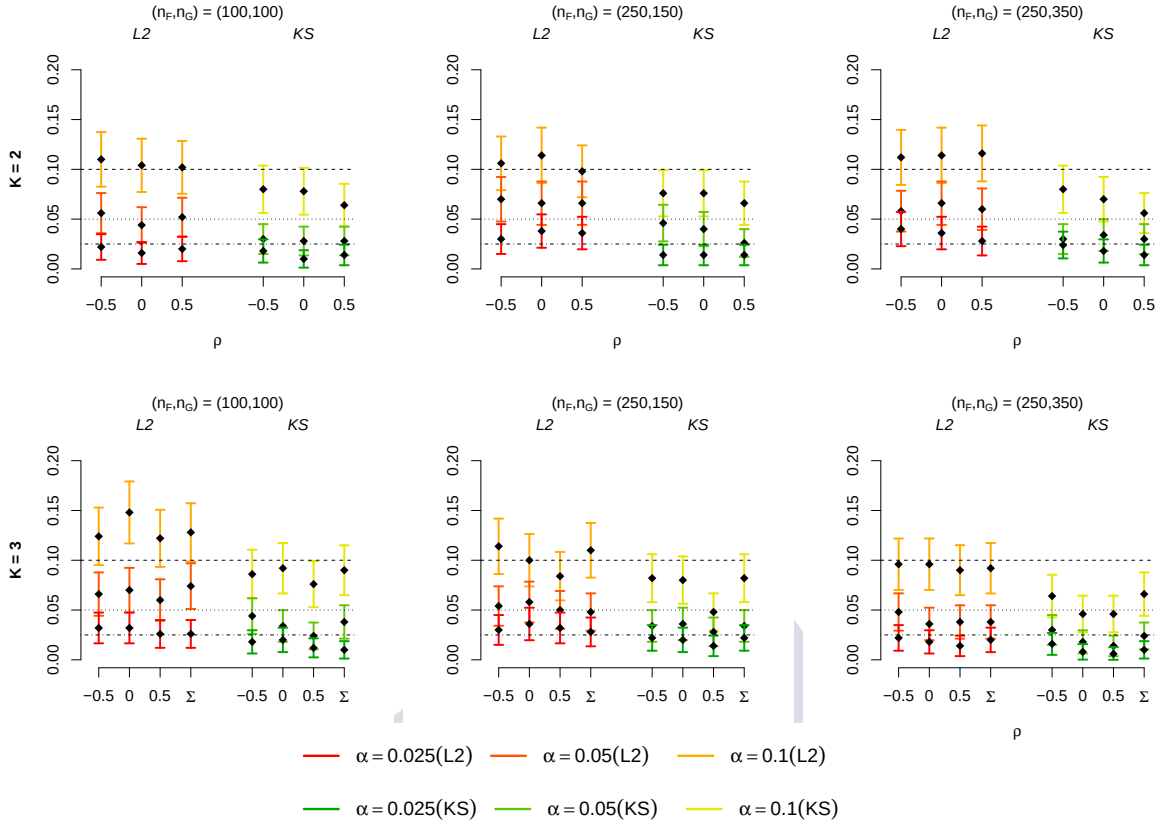


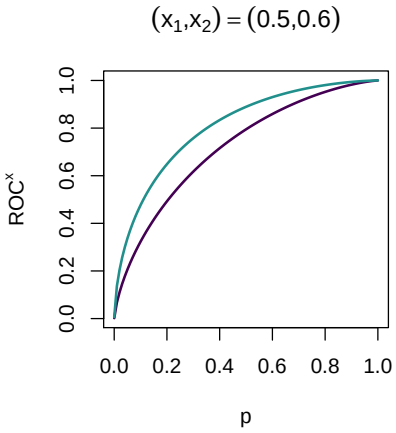
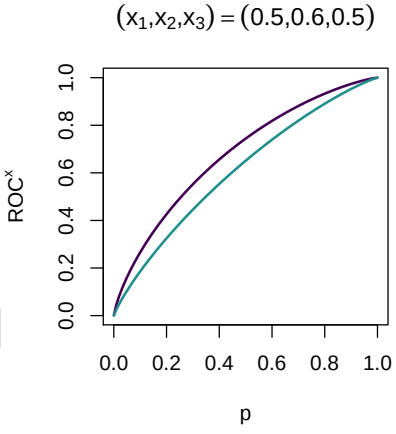
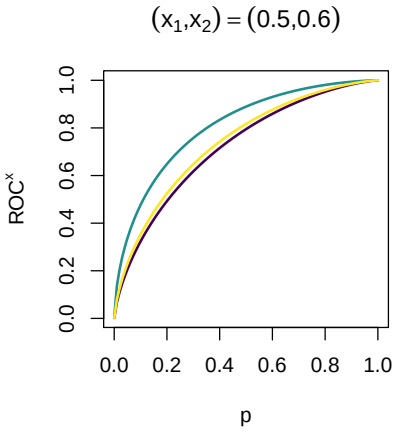
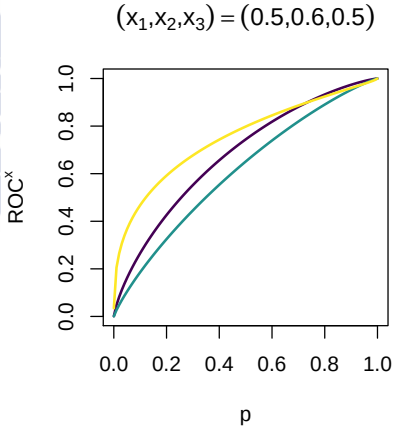
Figure 5.2: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis with $d = 3$ and $n_{\boldsymbol{\beta}} = 5$ for different sample sizes and different values of ρ and for the correlation matrix Σ.

The case with the largest sample sizes is also unbalanced, but not so much.

Remark 5.3. In order to evaluate the modification of the method proposed in Remark 5.1 and 5.2 we have run simulations for the same scenarios previously described. We show here the results for the scenarios with $K = 2$ and $d = 2$ under the null and the alternative hypotheses for assessing the level and the power of the test, respectively. Similar conclusions were obtained with the rest of the scenarios (go to Appendix B.2.3 to see the results of the simulation study for those cases). The parameters that are used here are the same as before, with the exception that now 1000 datasets were simulated instead of 500.

Figure 5.4 shows the results of the simulations when considering the modification of Remark 5.1 for $m_{\boldsymbol{\beta}} = 50$ (first row) and $m_{\boldsymbol{\beta}} = 25$ (second row), and the results for considering only one random projection (Remark 5.2), i.e., $m_{\boldsymbol{\beta}} = n_{\boldsymbol{\beta}} = 1$ (third row). Note that taking $m_{\boldsymbol{\beta}} = 25$ is comparable with $n_{\boldsymbol{\beta}} = 5$ used in the previous simulations (see first row of Figure 5.1), and that the results are very similar: the estimated proportion of rejections is a little overestimated for the L_2 statistic for the smaller sample size and otherwise close to the nominal level, and the KS statistic is always more conservative. Increasing $m_{\boldsymbol{\beta}}$ from 25 to 50 does not seem to affect the results significantly, and neither does reducing it to a single random projection ($m_{\boldsymbol{\beta}} = 1$). In Figure 5.5 we can observe the results for the simulations under the alternative hypothesis, once again for $m_{\boldsymbol{\beta}} = 50$, $m_{\boldsymbol{\beta}} = 25$ and $m_{\boldsymbol{\beta}} = 1$. The first two graphics are very similar to the one

Table 5.4: Scenarios under the alternative hypothesis considered for calibrating the power of the test. ROC_1^x and ROC_4^x are represented in purple, ROC_2^x and ROC_5^x in green, and ROC_3^x and ROC_6^x in yellow.

	2-dimensional covariate	**3-dimensional covariate**
$K = 2$	ROC_1^x vs. ROC_2^x	ROC_4^x vs. ROC_5^x
		
$K = 3$	ROC_1^x vs. ROC_2^x vs. ROC_3^x	ROC_4^x vs. ROC_5^x vs. ROC_6^x
		

obtained for $n_{\boldsymbol{\beta}} = 5$ (see the first graphic of Figure 5.3), but from the last graphic it is obvious that by using only one random projection the power of the test decreases considerably (as it was expected).

In the light of these results it seems that the alternative methodology proposed in Remark 5.1 yields similar conclusions than the first proposal, with no noticeable gain when increasing the number $m_{\boldsymbol{\beta}}$ used to approximate the value of the statistic from 25 to 50. It remains an open problem to determine an optimal value for that parameter.

As for the idea mentioned in Remark 5.2, using only one random projection seems to produce a well calibrated test, despite having considerably lower power.

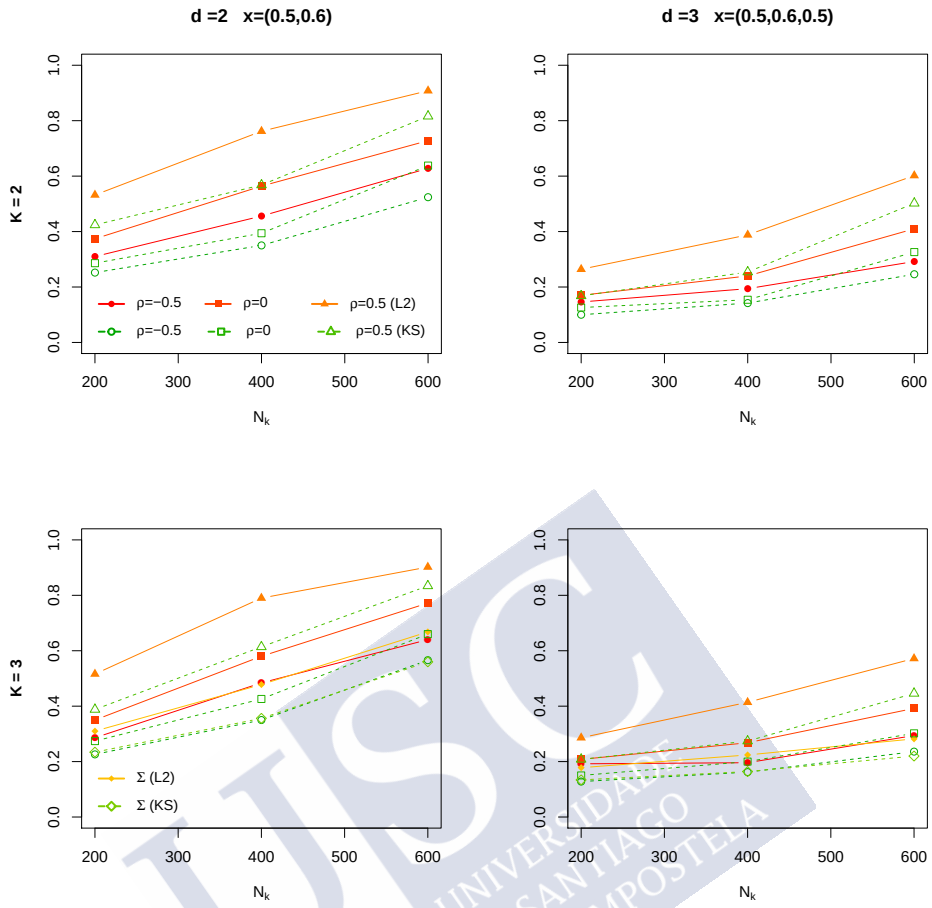


Figure 5.3: Estimated proportion of rejection under the alternative hypothesis for different sample sizes and different values of ρ and for the correlation matrix Σ, for $n_{\boldsymbol{\beta}} = 5$ ($\alpha = 0.05$).

5.4 Application to real data

An illustration of the proposed test is displayed in this section through the analysis of a dataset concerning 463 patients with Pleural Effusion (PE). The database at hand is the second one that was introduced in Chapter 1, in Section 1.2.2.

From a medical perspective, the goal is to find a way to discriminate the patients in which the PE has a malignant origin (MPE) from those in which the PE is due to other non-cancer-related causes. 200 individuals form the sample had MPE (the diseased population in this context), against 263 who did not (healthy population). For that matter, two diagnostic markers were considered, the carbohydrate antigen 152 (*CA125*) and the cytokeratin fragment 21-1 (*cyfra*). Moreover, the information of two different covariates is also available: the *age* and the neuron-specific enolase (*nse*). Due to the characteristics of the data (positive values, most of them close to zero, with some extreme high values), logarithms of those variables –excluding the variable *age*– were considered for the study. Being the logarithm a monotone transformation, its use does not have an effect on the estimation of the pooled ROC curve. However, it does affect the estimation of the conditional ROC curves, as it reduces the effect of the more extreme values of

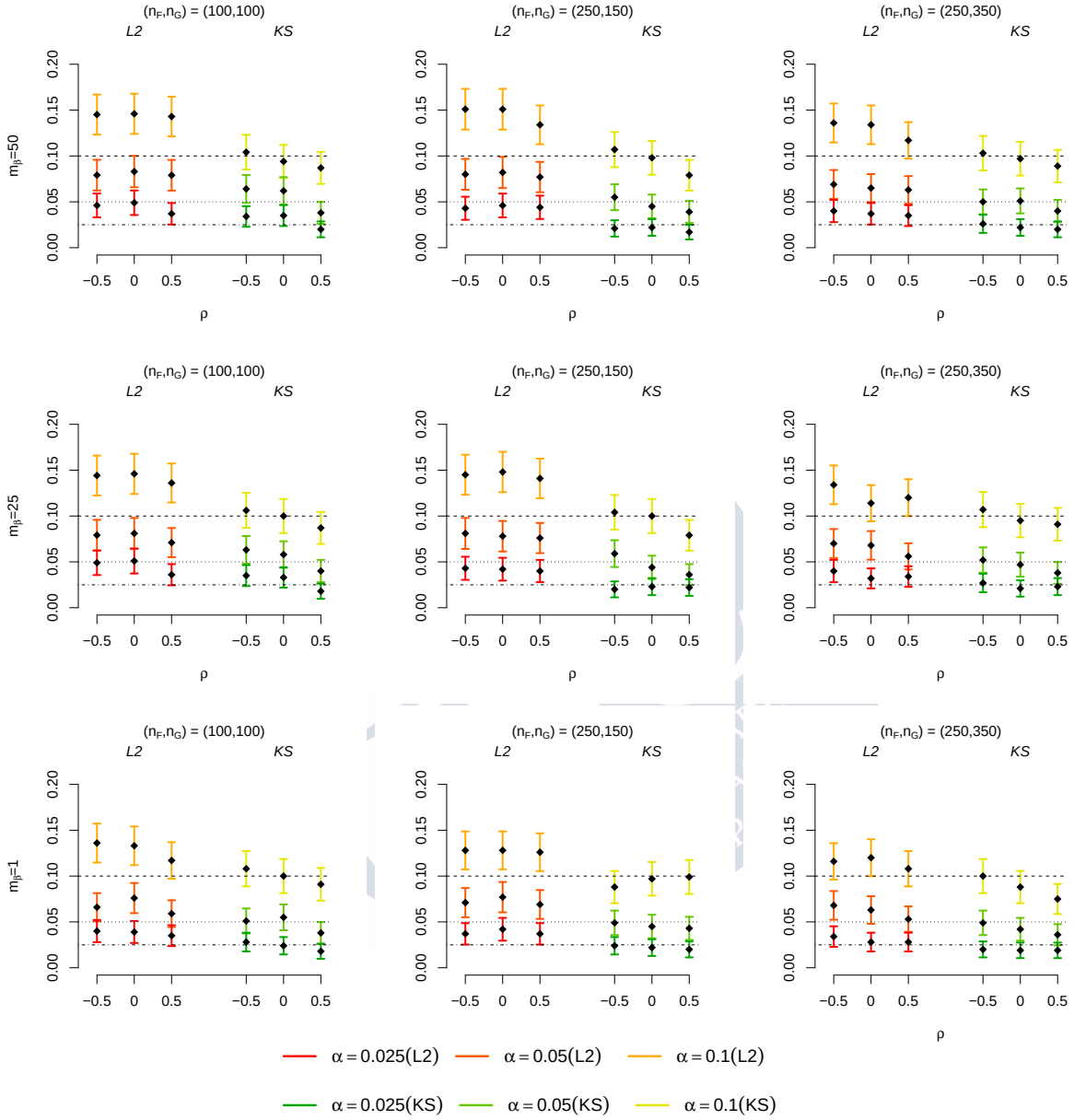


Figure 5.4: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis with $K = 2$, $d = 2$ and $m_{\boldsymbol{\beta}} = 50, 25, 1$ for different sample sizes and different values of ρ.

the variables. The variables used in this study are summarised in Table 5.5. A representation of the relationship of each one of those biomarkers with the two covariates is depicted in Figure 5.6, for both MPE (green) and the non-MPE (blue) patients. It can be observed that the shape of the point clouds of the two populations changes with the values of the covariates, specially in the case of the diseased population.

In order to evaluate whether the discriminatory capability of those markers (Y_1^F and Y_1^G as the variables containing the information of $\log(CA125)$, and Y_2^F and Y_2^G as the variables containing the information of $\log(cyfra)$) is the same when the covariates *age* and $\log(nse)$ are taken into account, the methodology explained in previous sections is applied, comparing

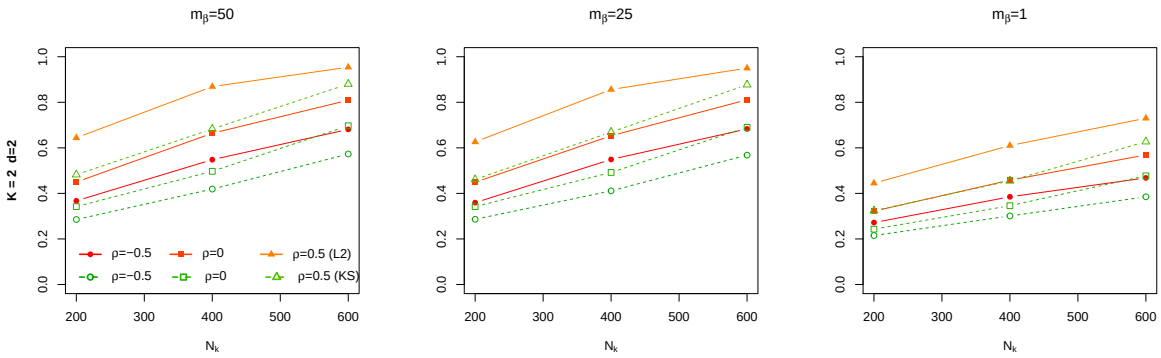


Figure 5.5: Estimated proportion of rejection under the alternative hypothesis for different sample sizes and different values of ρ, for $n_{\boldsymbol{\beta}} = 50, 25, 1$ and for the scenarios with $K = 2$ and $d = 2$ ($\alpha = 0.05$).

Table 5.5: A summary of the variables from the Pleural Effusion dataset used in the study for the MPE (D) and the non-MPE (H) subjects.

	age		log(*nse*)		log(*CA125*)		log(*cyfra*)	
	D	H	D	H	D	H	D	H
Minimum	32.00	15.0	−3.00	−3.00	1.61	1.16	0.43	−0.11
1st quartile	60.75	47.0	0.34	0.46	6.11	5.45	2.55	2.18
Median	73.00	65.0	1.34	1.46	6.83	6.27	3.62	3.00
Mean	69.50	61.3	1.29	1.29	6.55	5.90	3.81	3.02
3rd quartile	78.00	78.5	2.33	2.42	7.38	6.77	4.67	3.97
Maximum	95.00	94.0	5.32	5.31	8.87	8.05	8.07	6.65

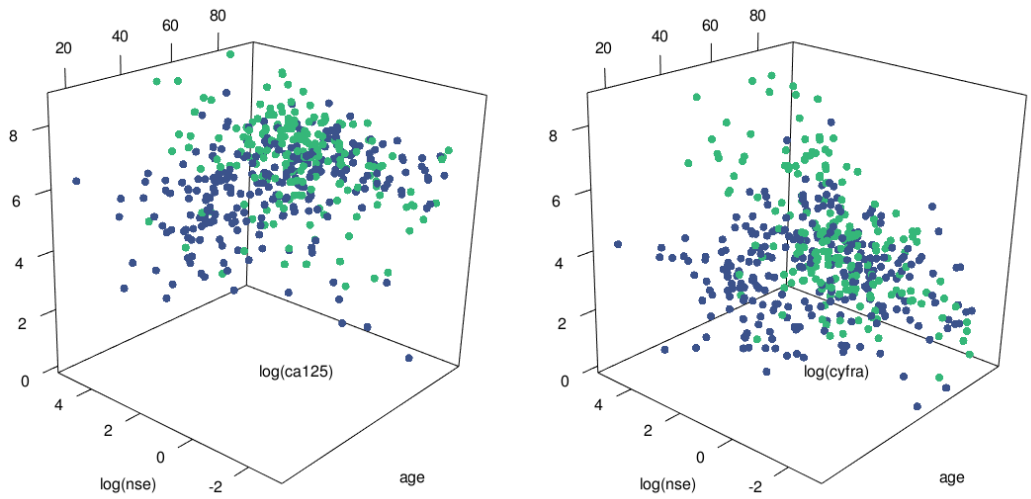


Figure 5.6: Scatterplot of the three different diagnostic biomarkers in function of the two covariates considered: age and log(*nse*)*. The healthy subjects are represented in blue and the diseased ones in green.*

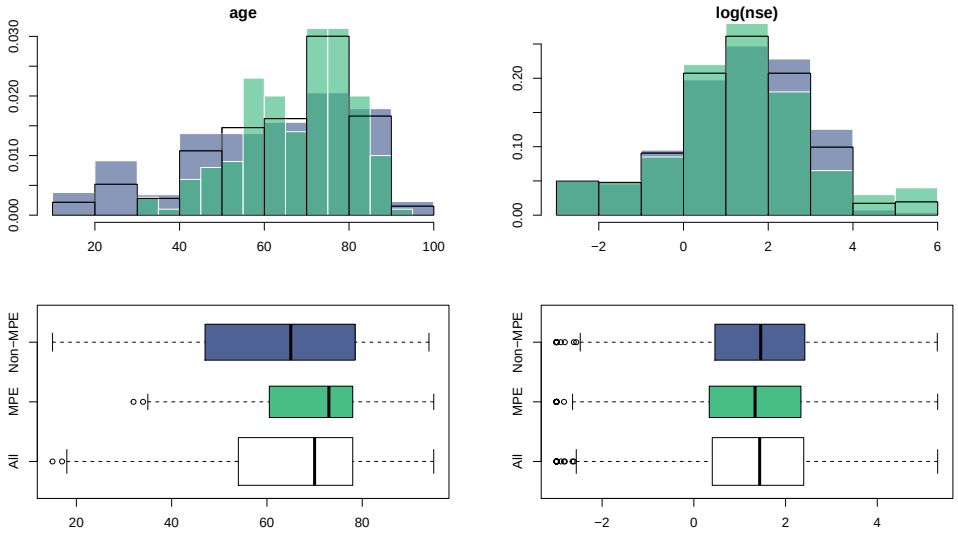


Figure 5.7: Histograms and boxplots of the two covariates considered (age and nse). The healthy subjects are represented in blue and the diseased ones in green. The black histogram lines and the white boxplot correspond to the two populations of the healthy and the diseased patients combined.

their respective ROC curves conditioned to different values of the bidimensional covariate $\boldsymbol{X} = (X_1, X_2)$ with $X_1 = age$ and $X_2 = \log(nse)$. In order to explore the advantages of using this method over the ones that do not consider multidimensional covariates, we also test the equivalence of the ROC curves of those diagnostic markers for the case in which no covariates are taken into account and for the case in which only one of the covariates is included in the analysis.

Note that, despite using the same dataset than in the previous chapter, apart from the fact that here we consider multidimensional covariates, there is another difference between this study and the one carried out in Section 4.4: here we are comparing dependent ROC curves (as the diagnostic variables are being measured over the same subjects), whereas in the previous chapter one diagnostic variable was compared in two groups that were independent. The correlation between Y_1^F and Y_2^F is 0.13, whereas the correlation between Y_1^G and Y_2^G is 0.37.

Figure 5.7 shows how those two covariates are distributed in the diseased and healthy populations. Note that the covariates have different magnitudes: the values that the variable *age* takes are always going to be greater than the values of $\log(nse)$. Thus, if we were to use the procedure directly over these variables, when projecting the multidimensional covariate $\boldsymbol{X}$ on any direction, the effect of the second component would be overshadowed by that of the first component. To prevent this from happening we decided to standardize the variables X_1 and X_2. This also affects the value $\boldsymbol{x}$ at which the conditional ROC curves are being compared. Note that when the covariate is modified by a one-to-one transformation, the ROC curve conditioned to a certain value x of the original covariate coincides with the ROC curve conditioned to the transformed value of the transformed covariate.

To be more precise, given a non-degenerate multidimensional covariate $\boldsymbol{X}$ the standardiza-

tion[2] proposed here is to consider the multidimensional covariate $\boldsymbol{X}_s = \boldsymbol{B}^{-1}(\boldsymbol{X} - \boldsymbol{a})$, with $\boldsymbol{B}$ a diagonal matrix with $(\sqrt{\mathbb{V}\text{ar}(X_1)}, \ldots, \sqrt{\mathbb{V}\text{ar}(X_d)})$ in the diagonal and $\boldsymbol{a} = (\mathbb{E}(X_1), \cdots, \mathbb{E}(X_d))'$. Then, for a given variable Y, a given $y \in \mathbb{R}$ and a certain value of the covariate $\boldsymbol{x}$,

$$P(Y \leqslant y|\boldsymbol{X} = \boldsymbol{x}) = P(Y \leqslant y|\boldsymbol{B}^{-1}(\boldsymbol{X} - \boldsymbol{a}) = \boldsymbol{B}^{-1}(\boldsymbol{x} - \boldsymbol{a})) = P(Y \leqslant y|\boldsymbol{X}_s = \boldsymbol{x}_s)),$$

with $\boldsymbol{x}_s = \boldsymbol{B}^{-1}(\boldsymbol{x} - \boldsymbol{a})$ and, thus,

$$ROC^{\boldsymbol{x}}(p) = 1 - F(G^{-1}(1-p|\boldsymbol{x})|\boldsymbol{x}) = 1 - F(G^{-1}(1-p|\boldsymbol{x}_s)|\boldsymbol{x}_s) = ROC^{\boldsymbol{x}_s}(p).$$

Note that the standardization that takes place here does not care for the covariance between the covariates that conform $\boldsymbol{X}$, as we are only interested on obtaining covariates with similar magnitudes. Also, in practice the standardization is made considering the sample mean and the sample standard deviation of the covariates at hand.

We start the analysis of the performance of the two diagnostic markers by comparing their respective ROC curves without taking into account any covariate information. For that matter we use the method proposed by DeLong et al. (1988). The estimated ROC curves for both markers are depicted in Figure 5.8. The p-value obtained for that comparison was 0.138. Similar results were obtained when using other ways of comparing ROC curves without covariates (like Martínez-Camblor et al. 2013 or Venkatraman and Begg 1996). Thus, we do not find significant differences between the two diagnostic variables in terms of diagnostic accuracy.

Next, we compare the two diagnostic markers taking into account a unidimensional covariate using the test proposed in Chapter 4 for dependent diagnostic markers. We consider the covariates *age* and $\log(nse)$, each one at a time. We test the equality of the ROC curves conditioned to the values of $\{51, 67, 83\}$ in the case of *age* and the values of $\{-0.92, 1.14, 3.27\}$ in the case of $\log(nse)$. The corresponding ROC curve for every case is estimated in Figure 5.8. For each considered covariate and each value of the covariate we obtain a p-value of the test, summarized in Table 5.6. The test is made considering two types of statistics, one based on the L_2-measure and the other in the Kolmogorov-Smirnov criterion, although both of them yield similar results. When comparing the ROC curves conditioned on different values of the *age*, the results are in line with the obtained for the previous case, in which no covariates where taken into account: the equality of the two curves is not rejected. However, when considering the covariate $\log(nse)$, we see that for the value 1.14 the null hypothesis is rejected (for a significance level of 5%). This matches the representation of the conditional ROC curves depicted in Figure 5.8.

Finally, we compare the performance of the two diagnostic variables considering the effect of both the *age* and the $\log(nse)$ at the same time. This is where we use the methodology proposed in this chapter. We test the equality of their respective ROC curves conditioned to nine pairs of values of the two covariates: the ones obtained by making all the possible combinations of $\{51, 67, 83\}$ and $\{-0.92, 1.14, 3.27\}$. As before, two different type of statistics were considered: L_2

[2]We could also consider $\boldsymbol{X}^F$ and $\boldsymbol{X}^G$ as two separate variables (the healthy and diseased populations are always taken as independent) and standardize each variable with their corresponding means and standard deviations. With this approach, the value $\boldsymbol{x}$ has to be standardized differently when is conditioning a distribution related to the diseased population and when the distribution is related to the healthy population instead. This is plausible, but it would imply testing the equality of ROC curves conditioned to a pair $(\boldsymbol{x}^F, \boldsymbol{x}^G)$, as defined in (5.3), this time with multidimensional values. This requires some adjustments in the methodology with little repercussion in the final outcome.

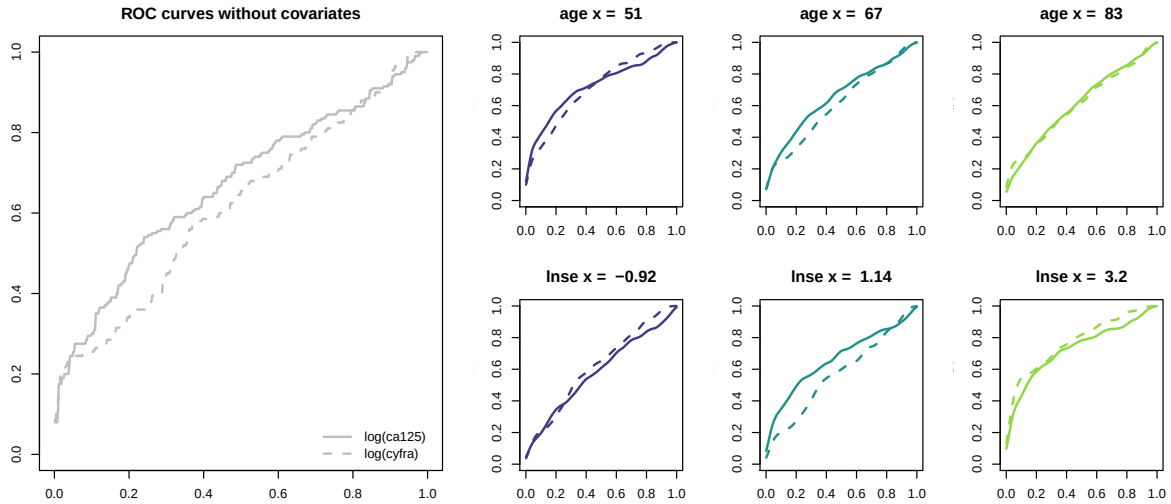


Figure 5.8: ROC curve estimation for both diagnostic variables ($\log(CA125)$ *and* $\log(cyfra)$*, represented by the solid and the dashed line, respectively) without covariates and conditioned to different values of the covariates age and* $\log(nse)$*.*

Table 5.6: Results for the comparison of the ROC curves of the diagnostic markers $\log(CA125)$ *and* $\log(cyfra)$ *when considering a unidimensional covariate, that covariate being the age or the* $\log(nse)$*.*

age	**51**	**67**	**83**
p-values (L_2)	0.454	0.218	0.936
$\log(nse)$	**-0.92**	**1.14**	**3.20**
p-values (L_2)	0.844	0.012	0.470

age	**51**	**67**	**83**
p-values (KS)	0.512	0.202	0.762
$\log(nse)$	**-0.92**	**1.14**	**3.20**
p-values (KS)	0.900	0.008	0.412

Table 5.7: Results for the comparison of the ROC curves of the diagnostic markers $\log(CA125)$ *and* $\log(cyfra)$ *when considering the multidimensional covariate (*age,$\log(nse)$*).*

$\log(nse)$ \ *age*	**51**	**67**	**83**
-0.92	0.000	0.030	0.258
1.14	0.152	0.070	0.004
3.20	0.026	0.056	0.010

$\log(nse)$ \ *age*	**51**	**67**	**83**
-0.92	0.004	0.048	0.424
1.14	0.212	0.050	0.016
3.20	0.066	0.196	0.032

and KS (and once again, the results are similar in both cases), with $n_{\boldsymbol{\beta}} = 10$ and 500 bootstrap iterations (similar p-values where obtained for $n_{\boldsymbol{\beta}} = 20$). The results obtained are summarized in Table 5.7. Note that in this case we did not represent the estimated ROC curves conditioned to the bidimensional covariate $(age, log(nse))$. This is to stress the fact that, with this methodology, $\widehat{ROC}^{\boldsymbol{x}}$ (with $\boldsymbol{x}$ bidimensional) does not need to be computed at all.

Table 5.8: Results for the comparison of the ROC curves of the diagnostic markers $\log(CA125)$ *and* $\log(cyfra)$ *when considering the multidimensional covariate* $(age, \log(nse))$ *using the alternative approximation of the test statistic proposed in Remark 5.1.*

$\log(nse)$ \ *age*	**51**	**67**	**83**
-0.92	0.006	0.016	0.170
1.14	0.180	0.052	0.016
3.20	0.014	0.002	0.002

$\log(nse)$ \ *age*	**51**	**67**	**83**
-0.92	0.000	0.028	0.208
1.14	0.124	0.044	0.010
3.20	0.002	0.002	0.014

The obtained p-values show that, depending on the pair of values of the considered covariate, we can find significative differences between the ROC curves of the $\log(CA125)$ and the $\log(cyfra)$ markers, including pairs of values that when considered separately in the previous test did not reject the null hypothesis. Likewise, finding differences between the ROC curves conditioned to marginal covariates at certain values does not mean that those differences will be significant when considering the multidimensional covariates (for example, when we conditioned the ROC curves marginally to the value of 1.14 of $\log(nse)$ we find differences, but when considering both covariates this difference between the ROC curves only remains significant for the *age* of 83).

Remark 5.4. If we were to use the alternative way of approximating the test statistic for the comparison of the ROC curves with multidimensional covariates proposed in Remark 5.1, with $m_{\boldsymbol{\beta}} = 100$ (note that $m_{\boldsymbol{\beta}} = 100$ is comparable with the use of $n_{\boldsymbol{\beta}} = 10$) and 500 bootstrap iterations (similar p-values where obtained for $m_{\boldsymbol{\beta}} = 50$), we would obtain the results shown in Table 5.8. As it can be observed, the p-values obtained were very similar to the ones obtained in Table 5.7, with the exception of the conditioned pair $(age, \log(nse)) = (67, 3.20)$.

5.5 Discussion

In this chapter a new nonparametric methodology has been presented for comparing two or more dependent ROC curves conditioned to the value of a continuous multidimensional covariate. This method combines existing techniques for reducing the dimension in goodness-of-fit tests and for estimating and comparing ROC curves conditioned to a one-dimensional covariate.

A simulation study was carried out in order to analyse the practical performance of the test. Two different functions were proposed for the construction of the statistic, the L_2 and the KS, the second one being a little more conservative. Different correlations between the diagnostic variables and different sample sizes have been considered, including uneven ones without any appreciable effect on the test performance.

Finally, the methodology was illustrated by means of an application to a dataset. With this new test it was possible to detect differences on the discriminatory ability of two diagnostic variables conditioned to two different covariates without the need of an estimator of an ROC curve conditioned to a multidimensional covariate. With this application it becomes clear the importance of being able to include the effect of multidimensional covariates to the ROC curves

analysis, as different conclusions could be drawn of the comparison of those curves when considering a multidimensional covariate, when considering unidimensional covariates or when excluding the covariates from the study.

Chapter 6

Conclusions and discussion

This dissertation was dedicated to the design and study of different methods for comparing ROC curves, with and without covariates. In this last chapter we summarize the obtained results on each chapter, pointing out the problems or extensions that could be addressed in further studies.

Chapter 2: ROC curves in the presence of covariates

The first step towards that objective was to determine when and how the covariates should be taken into account when dealing with ROC curves. Chapter 2 was dedicated to designing a strategy to decide in which situations it is more appropriate to use the pooled ROC curve, the conditional ROC curve or the covariate-adjusted ROC curve. It included a new methodology to test the equivalence between the pooled ROC curve and the AROC curve. Due to the dependence between those curves, the sample had to be divided to estimate both curves separately. However, this is not a very efficient way to deal with that problem, and it needs further research. An uneven division of the sample in favour of the estimation of the AROC curve could be considered.

Moreover, all the discussion and the tests considered in that chapter are limited to the case in which the covariate is unidimensional. Given that in practical situations this may not be realistic, an extension for multidimensional covariates could also be the topic for further research.

Chapter 2 finished with a discussion about how the strategy of deciding what kind of ROC curve should be used can be extended to the case where multiple ROC curves are being compared. Some of the settings described, such as the comparison among ROC curves without covariates or the comparison of conditional ROC curves have been discussed in this dissertation, but others remain an open problem. For the comparison of the pooled ROC curve and the covariate-adjusted ROC curve of different diagnostic markers (or independent samples), a similar test to the one proposed in Section 2.3.2 could be used.

Furthermore, another approach for incorporating the covariate effect in the ROC curve analysis is to design a new diagnostic maker that includes those covariates. Linear combination of multiple diagnostic variables have been considered in the literature (Su and Liu, 1993; Schisterman et al., 2004), using the corresponding AUC to choose the optimal configuration that ensures the highest discriminatory capability. Regardless of the limitations of such procedures (the parametric assumptions, the inflexibility of the linear model, the limitations inherent to the use of the AUC,...) this opens the discussion of which variables should be considered as diagnostic variables and which variables should be considered as covariates. On the one hand,

the study of the significance of the covariate effect proposed in Chapter 2 could be used to detect the variables that should be incorporated to the diagnostic methodology, and, moreover, this could be done without assuming any linear model. On the other hand, the diagnostic variables are usually given beforehand, with the specific objective of studying the illness at hand, whereas the covariates are all the extra information that can be collected from the subjects.

Chapter 3: Comparison of ROC curves without covariates

The next step in the study of the methods for comparing ROC curves was to analyse the problem without covariates. This was done in Chapter 3, where we performed a review of several techniques that exist in the literature to address this problem. This review included a simulation study that presented several scenarios designed to show the advantages and disadvantages of the various methodologies. They included situations in which two (or more) ROC curves were different despite having the same AUC (which showed the limitations of the methods based on the comparison of summary measures of the ROC curves), and situations in which the null hypothesis was reached despite the ROC curves being constructed by using different distribution functions. This last kind of scenario was hardly ever considered in the literature to show the calibration of these tests, and shows the problems than can arise from the use of resampling plans that do not replicate correctly the null hypothesis.

Most of the methods discussed in this chapter to compare ROC curves have a modified version that allows them to compare either independent or dependent ROC curves. However, the simulation study did not include scenarios with any dependence structure. Thus, the comparison of those methods using scenarios with dependent ROC curves (and the inclusion of other methods designed specifically for the dependent case that were not considered here) could be a subject for further research.

Chapter 4: Comparison of ROC curves with unidimensional covariates

The comparative study carried out in Chapter 3 laid the foundations for the design of a test for comparing ROC curves with covariates. Following the same philosophy of the methodology proposed in Martínez-Camblor et al. (2013) for the comparison of ROC curves without covariates (that was one of the methods studied in Chapter 3 that better behave in all scenarios) and using the estimator of the conditional ROC curve introduced in González-Manteiga et al. (2011), in Chapter 4 we designed a new methodology for comparing ROC curves with a unidimensional covariate.

The simulation study performed to analyse the size and power of the test did not include any comparison with other similar methodologies because, to the best of our knowledge, this is the first work that addresses this problem.

As for the consistency of the proposed bootstrap algorithm, despite not being addressed in this dissertation, it could be proved provided that the consistency of the bootstrap estimator of the conditional ROC curve is obtained. This is deferred to future research.

Another problem that remains open is the selection of the bandwidth parameters that are involved in the estimation of the conditional ROC curves. Also, there are several considerations

that should be taken into account when using this test, like the selection of the covariate values at which the ROC curves are conditioned. Enough data from the covariate around the selected values is required for both the healthy and diseased populations.

On the other hand, whereas this methodology is thought for the comparison of ROC curves at a fixed x value, it would be of great interest to be able to compare the curves at a finite number of points, or even through all the values of the covariate. This is a problem related to the significance of the covariate effect on the conditional ROC curves, something studied in Rodríguez-Álvarez et al. (2011b, 2018) and in Chapter 2 of this dissertation.

Moreover, the extension of these methodologies to the case in which the covariates at hand were longitudinal or functional could also be of great interest for future research. Functional data have already been included in ROC curve studies in Inácio et al. (2012), in Inácio de Carvalho et al. (2016) or in Estévez-Pérez and Vieu (2020). Following the idea behind our piece of research, the combination of existing techniques for comparing ROC curves without covariates and the methodologies for estimating the conditional ROC curves could help to develop further procedures in these more challenging settings.

In fact, given a longitudinal covariate X_t, with t denoting an instant on time, the interest of comparing ROC curves while taking into account this kind of covariate could go further: we could even consider the comparison of the ROC curve that is modelled using X_t with the ROC curve obtained when all the information of that longitudinal covariate from X_1 to X_{t-1} is taken into account (i.e., we could compare ROC_{X_t} with $ROC_{X_1,\ldots,X_{t-1}}$). This could determine whether the knowledge of the evolution of the covariate affects the discriminatory capability of the diagnostic method or not.

Chapter 5: Comparison of ROC curves with multidimensional covariates

Chapter 5 was dedicated to the extension of the methodology proposed in Chapter 4 to the case of multidimensional covariates. This extension was carried out by using random projections to transform a multidimensional problem into a unidimensional one.

This means that the problems that needed further research on the previous chapter are inherited here, like the problem of bandwidth selection. However, other issues require further study, like the optimal number of projections $n_{\boldsymbol{\beta}}$ (or $m_{\boldsymbol{\beta}}$, depending on the approach considered) in the approximation of the test statistic.

Another possible course of study could be to reproduce the methodology seen in Chapters 4 and 5 but with a different estimator for the conditional ROC curve, like the ones based on a direct methodology. Note that this would require some adjustments on the bootstrap algorithm.

On a final note, bear in mind that, when discussing the objectives of this dissertation in Chapter 1, the application of the developed techniques to real biomedical problems was emphasized. Two different datasets were analysed throughout this document to that end. Nonetheless, the methodologies designed for the comparison of conditional ROC curves could be applied with other goals in mind, focussing on personalized medicine. For example, take a subject that is suspected to have a certain disease. If he/she could choose among several hospitals to be diagnosed (as an example of independent ROC curve comparison problem) or if several diagnostic procedures were at his/her disposal (as an example of dependent ROC curve comparison prob-

lem), we could decide which option is the best by comparing the corresponding ROC curves while conditioning to the covariate values that this specific subject holds, such as age, body mass index, blood pressure, etc. This would mean a step towards a more personalized diagnostic methodology.

Appendix A

Some theoretical results

This appendix is devoted to prove the theoretical results obtained in Chapters 4 and 5. Section A.1 contains the assumptions and proofs needed for the asymptotic distribution of the test statistic proposed in Chapter 4. In Section A.2 we present all the proofs needed for the result in Chapter 5 that establishes the equivalence of comparing ROC curves conditioned to multidimensional covariates and comparing ROC curves conditioned to those same covariates projected in any direction.

A.1 Assumptions and proofs for Chapter 4

Here we present the assumptions and proofs needed for the theoretical results presented in Chapter 4. In the following lines, $F_{X_k^D}$ and $f_{X_k^D}$ are the cumulative distribution function and the density function of the covariate X_k^D, and h_k^D is the density function of the error ε_k^D, with $D \in \{F, G\}$ and $k \in \{1, \ldots, K\}$. In addition, the superscript (j) for $j \in \{1, 2, 3\}$ will represent the first, second or third derivative of a function.

Assumptions

(A1) (i) $n_k^D/N_k \longrightarrow \lambda_k^D$ for some $0 < \lambda_k^D < 1$ ($D \in \{F, G\}$ and $k \in \{1, \ldots, K\}$). Moreover, $N_k g_k^5 \longrightarrow C_k$ for some $0 \leqslant C_k < \infty$, $N_k g_k^{3+\alpha_k} (\log g_k^{-1})^{-1} \longrightarrow \infty$ for some $\alpha_k < 0$ and $N_k h_k^4 g_k \longrightarrow 0$ for each $k \in \{1, \ldots, K\}$.

(ii) $N_i/N_j \longrightarrow \Lambda_{ij}$ for some $0 < \Lambda_{ij} < \infty$ and $g_i^D/g_j^{D'} \longrightarrow \gamma_{ij}^{DD'}$ for some $0 < \gamma_{ij}^{DD'} < \infty$, with $i, j \in \{1, \ldots, K\}$ and $D, D' \in \{F, G\}$.

(iii) the kernel κ has compact support, $\int u\kappa(u)du = 0$ and κ is twice continuously differentiable.

(A2) (i) R_{X^F} and R_{X^G} are bounded intervals in $\mathbb{R}$.

(ii) $F_{X_k^D}$ is thrice continuously differentiable and $\inf_{x \in R_{X_k^D}} f_{X_k^D}(x) > 0$ ($D \in \{F, G\}$, $k \in \{1, \ldots, K\}$).

(iii) μ_k^D and σ_k^D are twice continuously differentiable and $\inf_{x \in R_{X_k^D}} \sigma_k^D(x) > 0$ ($D \in \{F, G\}$, $k \in \{1, \ldots, K\}$).

(A3) H_k^D is thrice continuously differentiable and $\sup_y |y^2 (H_k^D)^{(j)}(y)| < \infty$ for $j \in \{1,2,3\}$, $D \in \{F,G\}$, $k \in \{1,\ldots,K\}$. Moreover, for any $\delta > 0$, $\inf_{\delta<p<1-\delta} h_k^G((H_k^G)^{-1}(p)) > 0$.

(A4) $\psi(\cdot)$ is continuous and defines a norm over the space of functions.

Note that the assumptions provided in (A1.i), (A1.iii), (A2) and (A3) are the same that are needed in González-Manteiga et al. (2011) to show the convergence of the process (for only one ROC curve).

Proof of Theorem 4.1. It is easy to see that, under the null hypothesis, $S_N^x = T_N^x$. Thus, it is enough to proof that

$$T_N^x \xrightarrow{\mathcal{L}} \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} c_{kj} W_j^x(p)\right),$$

where $W_j^x(p)$ is the Gaussian process at which converges the process

$$(N_j g_j)^{1/2} \left\{\widehat{ROC}_j^x(p) - ROC_j^x(p)\right\},$$

with $j \in \{1,\ldots,K\}$, given a fixed $x \in R_X$ and $\delta < p < 1-\delta$ with δ small (González-Manteiga et al., 2011).

Let us define

$$\hat{R}^x(p) = \begin{pmatrix} (N_1 g_1)^{1/2} \left\{\widehat{ROC}_1^x(p) - ROC_1^x(p)\right\} \\ \vdots \\ (N_K g_K)^{1/2} \left\{\widehat{ROC}_K^x(p) - ROC_K^x(p)\right\} \end{pmatrix} \quad \text{and} \quad \mathcal{W}^x(p) = \begin{pmatrix} W_1^x(p) \\ \vdots \\ W_K^x(p) \end{pmatrix},$$

where $\mathcal{W}(p)$ is a multivariate Gaussian process with vector of means 0 in the case of under-smoothing (with $C_k = 0$ for all $k \in \{1,\ldots,K\}$) and a diagonal covariance matrix.

Then, due to the independence of the groups and the weak convergence of the conditional ROC processes, we obtain that

$$\hat{R}^x(p) \xrightarrow{\mathcal{L}} \mathcal{W}^x(p). \tag{A.1}$$

Now we define the function $\Psi : \mathbb{R}^K \longrightarrow \mathbb{R}$ as $\Psi(\mathbf{V}) = \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} c_{kj} V_j\right)$, with $\mathbf{V} = (V_1,\ldots,V_K)^t$, $\mathbf{V} \in \mathbb{R}^K$ and $c_{kj} \in \mathbb{R}$ constants.

We also define the function $\Psi_N : \mathbb{R}^K \longrightarrow \mathbb{R}$ as $\Psi_N(\mathbf{V}) = \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} \alpha_{kj}(N) V_j\right)$, where $\alpha_{kj}(N) = I(k=j) - \sqrt{g_k N_k}\sqrt{g_j N_j}\left(\sum_{i=1}^{K} g_i N_i\right)^{-1}$. Note that Ψ is continuous as long as ψ is continuous.

To obtain the result that we seek, it suffices to demonstrate that

$$\Psi_N\left(\hat{R}^x(p)\right) \xrightarrow{\mathcal{L}} \Psi\left(\mathcal{W}^x(p)\right). \tag{A.2}$$

If we prove that, for every sequence $v_N \in \mathbb{R}^K$ such that $v_{N'} \longrightarrow v$ along a subsequence and $v \in \mathbb{R}^K$ we obtain that $\Psi_{N'}(v_{N'}) \longrightarrow \Psi(v)$, we have the result (A.2) desired thanks to the Continuous Mapping Theorem (van der Vaart, 1998, Theorem 18.11) and to (A.1).

Let $v_N = (v_{N,1}, \ldots, v_{N,K}) \in \mathbb{R}^K$ be a sequence such that $v_{N'} \longrightarrow v_0$ along a subsequence and $v_0 = (v_{0,1}, \ldots, v_{0,K}) \in \mathbb{R}^K$. Then, since ψ is continuous and using Lemma A.1 (see below),

$$\Psi_{N'}(v_{N'}) = \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} \alpha_{kj}(N') v_{N',j}\right) \longrightarrow \sum_{k=1}^{K} \psi\left(\sum_{j=1}^{K} c_{kj} v_{0,j}\right) = \Psi(v_0),$$

and thus, the convergence in (A.2) holds.

Lemma A.1. *Assuming* $n_k^D/N_k \longrightarrow \lambda_k^D$ *for some* $0 < \lambda_k^D < 1$, $N_i/N_j \longrightarrow \Lambda_{ij}$ *for some* $0 < \Lambda_{ij} < \infty$ *and* $g_i^D/g_j^{D'} \longrightarrow \gamma_{ij}^{DD'}$ *for some* $0 < \gamma_{ij}^{DD'} < \infty$, *for* $D, D' \in \{F, G\}$, *and defining* $g_i/g_j \longrightarrow \gamma_{ij}$ *for some* $0 < \gamma_{ij} < \infty$, *with* $i, j \in \{1, \ldots, K\}$, *we have the convergence*

$$\alpha_{ij}(N) \longrightarrow c_{ij},$$

where $c_{ij} = I(i = j) - \frac{1}{\sum_{k=1}^{K} (\gamma_{ki}\gamma_{kj}\Lambda_{ki}\Lambda_{kj})^{1/2}} \in \mathbb{R}$ *for* $i, j \in \{1, \ldots, K\}$.

Proof. First, given $i, j \in \{1, \ldots, K\}$, assuming that $n_k^D/N_k \longrightarrow \lambda_k^D$ for some $0 < \lambda_k^D < 1$ ($D \in \{F, G\}$) and $N_i/N_j \longrightarrow \Lambda_{ij}$ for some $0 < \Lambda_{ij} < \infty$ we obtain

$$\frac{n_i^D}{n_j^{D'}} = \frac{\frac{n_i^D}{N_i}}{\frac{n_j^D}{N_j}} \frac{N_i}{N_j} \longrightarrow \frac{\lambda_i^D}{\lambda_j^{D'}} \Lambda_{ij}$$

Then, knowing that $g_i^D/g_j^{D'} \longrightarrow \gamma_{ij}^{DD'}$ for some $0 < \gamma_{ij}^{DD'} < \infty$, for $D, D' \in \{F, G\}$, we can see that $g_i/g_j \longrightarrow \gamma_{ij}$ for some $0 < \gamma_{ij} < \infty$:

$$\begin{aligned}
\frac{g_i}{g_j} &= \frac{(n_i^F g_i^F + n_i^G g_i^G)}{(n_j^F g_j^F + n_j^G g_j^G)} \frac{N_j}{N_i} = \left(\frac{n_i^F g_i^F}{n_j^F g_j^F + n_j^G g_j^G} + \frac{n_i^G g_i^G}{n_j^F g_j^F + n_j^G g_j^G}\right) \frac{N_j}{N_i} \\
&= \left(\frac{1}{\frac{n_j^F}{n_i^F}\frac{g_j^F}{g_i^F} + \frac{n_j^G}{n_i^F}\frac{g_j^G}{g_i^F}} + \frac{1}{\frac{n_j^F}{n_i^G}\frac{g_j^F}{g_i^G} + \frac{n_j^G}{n_i^G}\frac{g_j^G}{g_i^G}}\right) \frac{N_j}{N_i} \\
&\longrightarrow \left(\frac{1}{\frac{\lambda_j^F}{\lambda_i^F}\Lambda_{ji}\gamma_{ji}^{FF} + \frac{\lambda_j^G}{\lambda_i^F}\Lambda_{ji}\gamma_{ji}^{GF}} + \frac{1}{\frac{\lambda_j^F}{\lambda_i^G}\Lambda_{ji}\gamma_{ji}^{FG} + \frac{\lambda_j^G}{\lambda_i^G}\Lambda_{ji}\gamma_{ji}^{GG}}\right)\Lambda_{ji} \\
&= \left(\frac{1}{\frac{\lambda_j^F}{\lambda_i^F}\gamma_{ji}^{FF} + \frac{\lambda_j^G}{\lambda_i^F}\gamma_{ji}^{GF}} + \frac{1}{\frac{\lambda_j^F}{\lambda_i^G}\gamma_{ji}^{FG} + \frac{\lambda_j^G}{\lambda_i^G}\gamma_{ji}^{GG}}\right) =: \gamma_{ij}
\end{aligned}$$

Then, given $i, j \in \{1, \ldots, K\}$,

$$\begin{aligned}
\alpha_{ij}(N) &= I(i = j) - \frac{(g_i N_i)^{1/2}(g_j N_j)^{1/2}}{\sum_{k=1}^{K} g_k N_k} \\
&= I(i = j) - \frac{1}{\sum_{k=1}^{K} \frac{g_k}{g_i^{1/2} g_j^{1/2}} \frac{N_k}{N_i^{1/2} N_j^{1/2}}} \longrightarrow I(i = j) - \frac{1}{\sum_{k=1}^{K} \gamma_{ki}^{1/2} \gamma_{kj}^{1/2} \Lambda_{ki}^{1/2} \Lambda_{kj}^{1/2}} =: c_{ij} \in \mathbb{R}.
\end{aligned}$$

A.2 Proofs for Chapter 5

Here we present the proofs needed for the Lemma 5.1 presented in Chapter 5.

Lemma A.2. *Escanciano (2006) or Cuesta-Albertos et al. (2019): Given a random variable Y such that $\mathbb{E}|Y| < \infty$,*

$$\mathbb{E}[Y|\boldsymbol{X}] = 0 \ a.s. \Leftrightarrow \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}] = 0 \ a.s. \ for \ any \ vector \ \boldsymbol{\beta} \in \mathbb{S}^{d-1}. \tag{A.3}$$

From now on it will be assumed that all projections $\boldsymbol{\beta}$ considered satisfy $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$.

Lemma A.3. *Let $Y_1, \cdots, Y_K$ be K dependent random variables with cumulative distribution functions $F_1, \ldots, F_K$, respectively, such that $\mathbb{E}|Y_k| < \infty$ for every $k \in \{1, \ldots, K\}$. Let $\boldsymbol{X}$ be a multidimensional covariate. Then, given $c_1, \ldots, c_K$,*

$$F_1(c_1|\boldsymbol{X}) = \cdots = F_K(c_K|\boldsymbol{X}) \ a.s. \Leftrightarrow F_1(c_1|\boldsymbol{\beta}'\boldsymbol{X}) = \cdots = F_K(c_K|\boldsymbol{\beta}'\boldsymbol{X}) \ a.s. \ \forall \boldsymbol{\beta}, \tag{A.4}$$

with $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$.

Proof. It is proven for $K = 2$:

$$\begin{aligned} F_1(c_1|\boldsymbol{X}) = F_2(c_2|\boldsymbol{X}) \ a.s. &\Leftrightarrow \mathbb{E}[I(Y_1 \leqslant c_1)|\boldsymbol{X}] = \mathbb{E}[I(Y_2 \leqslant c_2)|\boldsymbol{X}] \ a.s. \\ &\overset{(*)}{\Leftrightarrow} \mathbb{E}[I(Y_1 \leqslant c_1) - I(Y_2 \leqslant c_2)|\boldsymbol{X}] = 0 \ a.s. \\ &\overset{(A.3)}{\Leftrightarrow} \mathbb{E}[I(Y_1 \leqslant c_1) - I(Y_2 \leqslant c_2)|\boldsymbol{\beta}'\boldsymbol{X}] = 0 \ a.s. \ \forall \boldsymbol{\beta} \\ &\Leftrightarrow \mathbb{E}[I(Y_1 \leqslant c_1)|\boldsymbol{\beta}'\boldsymbol{X}] = \mathbb{E}[I(Y_2 \leqslant c_2)|\boldsymbol{\beta}'\boldsymbol{X}] \ a.s. \ \forall \boldsymbol{\beta} \\ &\Leftrightarrow F_1^{\boldsymbol{\beta}}(c_1|\boldsymbol{\beta}'\boldsymbol{X}) = F_2^{\boldsymbol{\beta}}(c_2|\boldsymbol{\beta}'\boldsymbol{X}) \ a.s. \ \forall \boldsymbol{\beta}, \end{aligned}$$

where $F_i^{\boldsymbol{\beta}}(c_i|\boldsymbol{\beta}'\boldsymbol{X}) = P(Y_i \leqslant c_i|\boldsymbol{\beta}'\boldsymbol{X} = \boldsymbol{\beta}'\boldsymbol{X})$ for $i = 1, 2$.

Note that in (*) the fact that the random variables are dependent is being used in the sense that they are conditioned to the same covariate $\boldsymbol{X}$ (i.e., there is no X_1 and X_2 as there would be in the independent case). □

Definition A.1. The *inverted conditional ROC curve (IROC)* is defined as:

$$IROC(p) = 1 - G(F^{-1}(1-q)), \ q \in (0,1).$$

Related to the previous definition, the *inverted conditional ROC curve* ($IROC^{x^G,x^F}$), given the pair $(x^F, x^G) \in R_{X^F} \times R_{X^G}$, can also be defined as:

$$IROC^{x^G,x^F}(q) = 1 - G(F^{-1}(1-q|x^F)|x^G), \ q \in (0,1).$$

Lemma A.4. *The equality of ROC curves is equivalent to the equality of the inverted ROC curves, i.e.,*

$$ROC_1(p) = \cdots = ROC_K(p) \ \forall p \in (0,1) \ \Leftrightarrow \ IROC_1(q) = \cdots = IROC_K(q) \ \forall q \in (0,1).$$

Moreover, the same property holds when talking about conditional ROC curves. Given the pair $(x^F, x^G) \in R_{X^F} \times R_{X^G}$,

$$\begin{aligned} ROC_1^{x^F,x^G}(p) = \cdots = ROC_K^{x^F,x^G}(p)\ \forall p \in (0,1) \\ \Leftrightarrow\ IROC_1^{x^G,x^F}(q) = \cdots = IROC_K^{x^G,x^F}(q)\ \forall q \in (0,1). \end{aligned} \tag{A.5}$$

Proof. It is proven for the unconditional case, and for $K = 2$. The conditional case is similar.

$$ROC_1(p) = ROC_2(p)\ \forall p \in (0,1) \Leftrightarrow 1 - F_1(G_1^{-1}(1-p)) = 1 - F_2(G_2^{-1}(1-p))\ \forall p \in (0,1)$$

Take $q = 1 - F_2(G_2^{-1}(1-p))$ (and hence, $q = ROC_2(p)$). q will take all the values in $(0,1)$, and thus, $p = 1 - G_2(F_2^{-1}(1-q)) = IROC_2(q)$.

Then,

$$\begin{aligned} ROC_1(p) = ROC_2(p)\ \forall p \in (0,1) &\Leftrightarrow 1 - F_1(G_1^{-1}(1-(1-G_2(F_2^{-1}(1-q))))) = q\ \forall q \in (0,1) \\ &\Leftrightarrow 1 - G_2(F_2^{-1}(1-q) = 1 - G_1(F_1^{-1}(1-q))\ \forall q \in (0,1) \\ &\Leftrightarrow IROC_2(q) = IROC_1(q)\ \forall q \in (0,1). \end{aligned}$$

□

Proof of Lemma 5.1

Proof. It is proven for $K = 2$. For $p \in (0,1)$,

$$\begin{aligned} & ROC_1^{\boldsymbol{x}}(p) = ROC_2^{\boldsymbol{x}}(p)\ a.s. \Leftrightarrow \\ \Leftrightarrow\ & 1 - F_1(G_1^{-1}(1-p|\boldsymbol{x})|\boldsymbol{x}) = 1 - F_2(G_2^{-1}(1-p|\boldsymbol{x})|\boldsymbol{x})\ a.s. \\ \Leftrightarrow\ & F_1(G_1^{-1}(1-p|\boldsymbol{x})|\boldsymbol{x}) = F_2(G_2^{-1}(1-p|\boldsymbol{x})|\boldsymbol{x})\ a.s. \\ \overset{(A.4)}{\Leftrightarrow}\ & F_1^{\boldsymbol{\beta}^F}\left(G_1^{-1}(1-p|\boldsymbol{x})|(\boldsymbol{\beta}^F)'\boldsymbol{x}\right) = F_2^{\boldsymbol{\beta}^F}\left(G_2^{-1}(1-p|\boldsymbol{x})|(\boldsymbol{\beta}^F)'\boldsymbol{x}\right)\ a.s.\ \forall \boldsymbol{\beta}^F \\ \Leftrightarrow\ & ROC_1^{(\boldsymbol{\beta}^F)'\boldsymbol{x},\boldsymbol{x}}(p) = ROC_2^{(\boldsymbol{\beta}^F)'\boldsymbol{x},\boldsymbol{x}}(p)\ a.s.\ \forall \boldsymbol{\beta}^F \\ \overset{(A.5)}{\Leftrightarrow}\ & IROC_1^{\boldsymbol{x},(\boldsymbol{\beta}^F)'\boldsymbol{x}}(q) = IROC_2^{\boldsymbol{x},(\boldsymbol{\beta}^F)'\boldsymbol{x}}(q)\ a.s.\ \forall \boldsymbol{\beta}^F\ \text{ for } q \in (0,1) \\ \Leftrightarrow\ & G_1((F_1^{\boldsymbol{\beta}^F})^{-1}(1-q|(\boldsymbol{\beta}^F)'\boldsymbol{x})|\boldsymbol{x}) = G_2((F_2^{\beta^F})^{-1}(1-q|(\boldsymbol{\beta}^F)'\boldsymbol{x})|\boldsymbol{x})\ a.s.\ \forall \boldsymbol{\beta}^F \\ \overset{(A.4)}{\Leftrightarrow}\ & G_1^{\boldsymbol{\beta}^G}((F_1^{\boldsymbol{\beta}^F})^{-1}(1-q|(\boldsymbol{\beta}^F)'\boldsymbol{x})|(\boldsymbol{\beta}^G)'\boldsymbol{x}) = G_2^{\boldsymbol{\beta}^G}((F_2^{\boldsymbol{\beta}^F})^{-1}(1-q|(\boldsymbol{\beta}^F)'\boldsymbol{x})|(\boldsymbol{\beta}^G)'\boldsymbol{x})\ a.s.\ \forall \boldsymbol{\beta}^F, \boldsymbol{\beta}^G \\ \Leftrightarrow\ & IROC_1^{(\boldsymbol{\beta}^G)'\boldsymbol{x},(\boldsymbol{\beta}^F)'\boldsymbol{x}}(q) = IROC_2^{(\boldsymbol{\beta}^G)'\boldsymbol{x},(\boldsymbol{\beta}^F)'\boldsymbol{x}}(q)\ a.s.\ \forall \boldsymbol{\beta}^F, \boldsymbol{\beta}^G \\ \overset{(A.5)}{\Leftrightarrow}\ & ROC_1^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(\tilde{p}) = ROC_2^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(\tilde{p})\ a.s.\ \forall \boldsymbol{\beta}^F, \boldsymbol{\beta}^G \text{ for } \tilde{p} \in (0,1), \end{aligned}$$

where

$$\begin{aligned} F_1^{\boldsymbol{\beta}^F}\left(c|(\boldsymbol{\beta}^F)'\boldsymbol{x}\right) &= P\left(Y_1^F \leqslant c|(\boldsymbol{\beta}^F)'\boldsymbol{X}^F = (\boldsymbol{\beta}^F)'\boldsymbol{x}\right), \\ F_2^{\boldsymbol{\beta}^F}\left(c|(\boldsymbol{\beta}^F)'\boldsymbol{x}\right) &= P\left(Y_2^F \leqslant c|(\boldsymbol{\beta}^F)'\boldsymbol{X}^F = (\boldsymbol{\beta}^F)'\boldsymbol{x}\right), \\ G_1^{\boldsymbol{\beta}^G}\left(c|(\boldsymbol{\beta}^G)'\boldsymbol{x}\right) &= P\left(Y_1^G \leqslant c|(\boldsymbol{\beta}^G)'\boldsymbol{X}^G = (\boldsymbol{\beta}^G)'\boldsymbol{x}\right), \\ G_2^{\boldsymbol{\beta}^G}\left(c|(\boldsymbol{\beta}^G)'\boldsymbol{x}\right) &= P\left(Y_2^G \leqslant c|(\boldsymbol{\beta}^G)'\boldsymbol{X}^G = (\boldsymbol{\beta}^G)'\boldsymbol{x}\right). \end{aligned}$$

□

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix B

Extra simulations

The methodologies presented in Chapters 4 and 5 depend on several parameters. These parameters are sometimes related to the approximation of the statistic, other times related with the bandwidth parameter selection, or simply related with the properties of the data at hand. This results in a quite extensive collection of finite sample studies. For the sake of readability and simplicity of this document, only the main simulation studies were included in the main body of the text. In this appendix we included the rest of them, which allow us to study each case in greater depth.

B.1 Supplementary material for Chapter 4

In this first part of the appendix we collected the simulations that were left out of Chapter 4. In Section B.1.1 we see how the different bandwidth parameters involved in the estimation of the conditional ROC curves can affect the results of the test. In Section B.1.2 we see analogous simulations to the ones shown in Chapter 4, but using fewer bootstrap iterations. Section B.1.3 has some results for scenarios with unbalanced data in the different ROC curves that are being compared, and Section B.1.4 contains a simulation study for the comparison of dependent ROC curves.

B.1.1 Bandwidth parameters involved in the estimation of the conditional ROC curve

As it has been pointed out repeatedly throughout this dissertation, the selection of bandwidth parameters is still an open problem. There are no optimal bandwidths designed for the estimation of the conditional ROC curve presented by González-Manteiga et al. (2011) and, even if there were, there is no guarantee that they would be optimal for a comparison test such as the ones presented here.

In this subsection we have tried different configurations for these bandwidth parameters (for both h_k, that determines the smoothness of the estimator, and g_k^F and g_k^G, used for the estimation of the regression functions, with $k \in \{1, \ldots, K\}$). We run simulations using similar scenarios as the ones considered in Section 4.3. We focused only in the approximation of the level of the test, as our concern in this case is mainly with its calibration and not so much with its power.

Table B.1: Estimated proportion of rejections for scenario C (for $K = 2$) for $\alpha = 0.05$ for different values of the bandwidth parameter h_k.

			L_2				KS			
x	(n_k, m_k)	h_k:	$h_0 = 0$	$h_{1,k}$	$h_{2,k}$	$h_{3,k}$	h_0	$h_{1,k}$	$h_{2,k}$	$h_{3,k}$
0.25	(100, 100)		0.062	0.060	0.061	0.064	0.060	0.069	0.069	0.064
	(150, 250)		0.050	0.053	0.055	0.056	0.043	0.055	0.057	0.061
	(300, 300)		0.037	0.041	0.039	0.037	0.037	0.052	0.049	0.049
	(550, 250)		0.038	0.043	0.042	0.039	0.038	0.046	0.048	0.050
	(500, 500)		0.031	0.033	0.032	0.033	0.029	0.044	0.041	0.042
0.5	(100, 100)		0.053	0.050	0.049	0.051	0.046	0.050	0.053	0.049
	(150, 250)		0.047	0.047	0.047	0.043	0.036	0.052	0.047	0.049
	(300, 300)		0.050	0.052	0.052	0.049	0.030	0.044	0.044	0.047
	(550, 250)		0.036	0.035	0.035	0.038	0.033	0.041	0.039	0.043
	(500, 500)		0.052	0.051	0.052	0.057	0.033	0.043	0.042	0.050
0.75	(100, 100)		0.079	0.078	0.078	0.078	0.053	0.069	0.070	0.081
	(150, 250)		0.062	0.063	0.067	0.068	0.035	0.054	0.056	0.059
	(300, 300)		0.049	0.052	0.053	0.054	0.029	0.040	0.039	0.042
	(550, 250)		0.043	0.042	0.043	0.043	0.026	0.036	0.040	0.043
	(500, 500)		0.050	0.051	0.051	0.050	0.022	0.037	0.041	0.047

Bandwidth h_k: smoothness of the ROC curve estimation

We studied the effect of the bandwidth parameter h_k on the scenario C, with $K = 2$, described in Section 4.3.1. The bandwidths considered were $h_0 = 0$ (representing the empirical estimation of the conditional ROC curve, introduced in (2.10)), $h_{1,k} = \frac{1}{\sqrt{N_k}}$, $h_{2,k} = \frac{1}{\sqrt{N_k/2}}$ and $h_{3,k} = \frac{2}{\sqrt{N_k/2}}$. The rest of the parameters (from the covariate distribution to the sample sizes) are similar to the ones considered previously. The results (the proportion of rejections under the null hypothesis) are collected in Table B.1. Note that $h_{1,k}$ was the one used for the simulation study on Chapter 4.

In general, this bandwidth parameter does not seem to affect the results. The proportions of rejections obtained for each conditional x value, each sample size and each type of statistic are very similar for the four bandwidths, although the empirical version of the estimator of the conditional ROC curve (i.e., the results for h_0) yield more conservative results in the case of the KS statistic.

Bandwidths g_k^F and g_k^G: estimating the regression functions

The bandwidths used for the estimation of the regression functions (g_k^F and g_k^G for $k \in \{1, \cdots, K\}$) needed for the estimation of the conditional ROC curves were selected by a cross-validation method. We made several trials fixing different values for these bandwidths, and it became obvious that its selection affects the calibration of the test: one fixed bandwidth could yield overestimated proportions of rejections under the null hypothesis for some scenarios and for some sample sizes, and at the same time yield underestimated proportions for other scenarios and other sample sizes. The selection of the grid from which the cross-validation was made can

also influence the behaviour of the test.

We have repeated the same simulation study described in 4.3.1 changing that grid, considering one slightly thinner and wider. Here we show the results when the grid is made out of 100 points (instead of the 25 used previously) and its range goes from 0.075 to 0.7, instead of going from 0.1 to 0.5. Figures B.1 and B.2 contain the results of the same simulation study carried out in section 4.3.1 using that alternative grid. They are very similar to Figures 4.1 and 4.2 of Chapter 4, but they seem to obtain better approximation of the level of the test for the larger sample sizes (particularly when conditioning to $x = 0.5$).

B.1.2 Number of bootstrap iterations

In this subsection we present the results of a similar study (with the same scenarios and the same sample sizes) as the one performed in Section 4.3. The only difference is that, in this case, the number of bootstrap iterations considered was 200 instead of 500. This does not seem to affect significantly the result of the simulations.

On the one hand, Figures B.3 and B.4 hold the results for the simulations run under the null hypothesis (and are similar to Figures 4.1 and 4.2 of Chapter 4).

On the other hand, the results for the study of the power (showed in Chapter 4 at Figure 4.4 for $B = 200$) can be found in Figure B.5.

B.1.3 Unbalanced data

Given that in practical situations the sample sizes do not have to be the same for each diagnostic marker considered (or for the healthy and the diseased populations involved), in this subsection we run some simulations using unbalanced data. We considered the scenarios described in Section 4.3 that had $K = 2$, using the same parameters, with the exception of the number of bootstrap iterations (here we considered 200 instead of 500, although for what we saw in Subsection B.1.2 it does not affect the results of the test), and the sample sizes. Here we studied the results for four different sets of sample sizes, which we denote by M_1, M_2, M_3 and M_4:

$$
\begin{aligned}
&M_1 : (n_1^F, n_1^G) = (n_2^F, n_2^G) = (250, 150),\ N_1 = N_2 = 400,\\
&M_2 : (n_1^F, n_1^G) = (n_2^F, n_2^G) = (150, 250),\ N_1 = N_2 = 400\\
&M_3 : (n_1^F, n_1^G) = (150, 300),\ (n_2^F, n_2^G) = (250, 100),\ N_1 = 450,\ N_2 = 350\\
&M_4 : (n_1^F, n_1^G) = (300, 150),\ (n_2^F, n_2^G) = (250, 100),\ N_1 = 450,\ N_2 = 350.
\end{aligned}
$$

Note that M_1 was already considered in Section 4.3. There, N_1 and N_2 show the sample sizes involved in each compared ROC curve. If we computed the sample sizes of the number diseased (N^F) and healthy individuals (N^G), we would obtain the pairs $(N^F, N^G) = (500, 300), (300, 500), (400, 400), (550, 250)$ for M_1, M_2, M_3 and M_4, respectively. This means that, of this sets of sample sizes, the most unbalanced is M_4. Note that the total sample size in each case is 800.

Figure B.6 displays the results for the scenarios under the null hypothesis. In general, the confidence intervals of the estimated proportions contain the nominal level, specially for $x = 0.5$ (which is to be expected, as it is the covariate value with most data around).

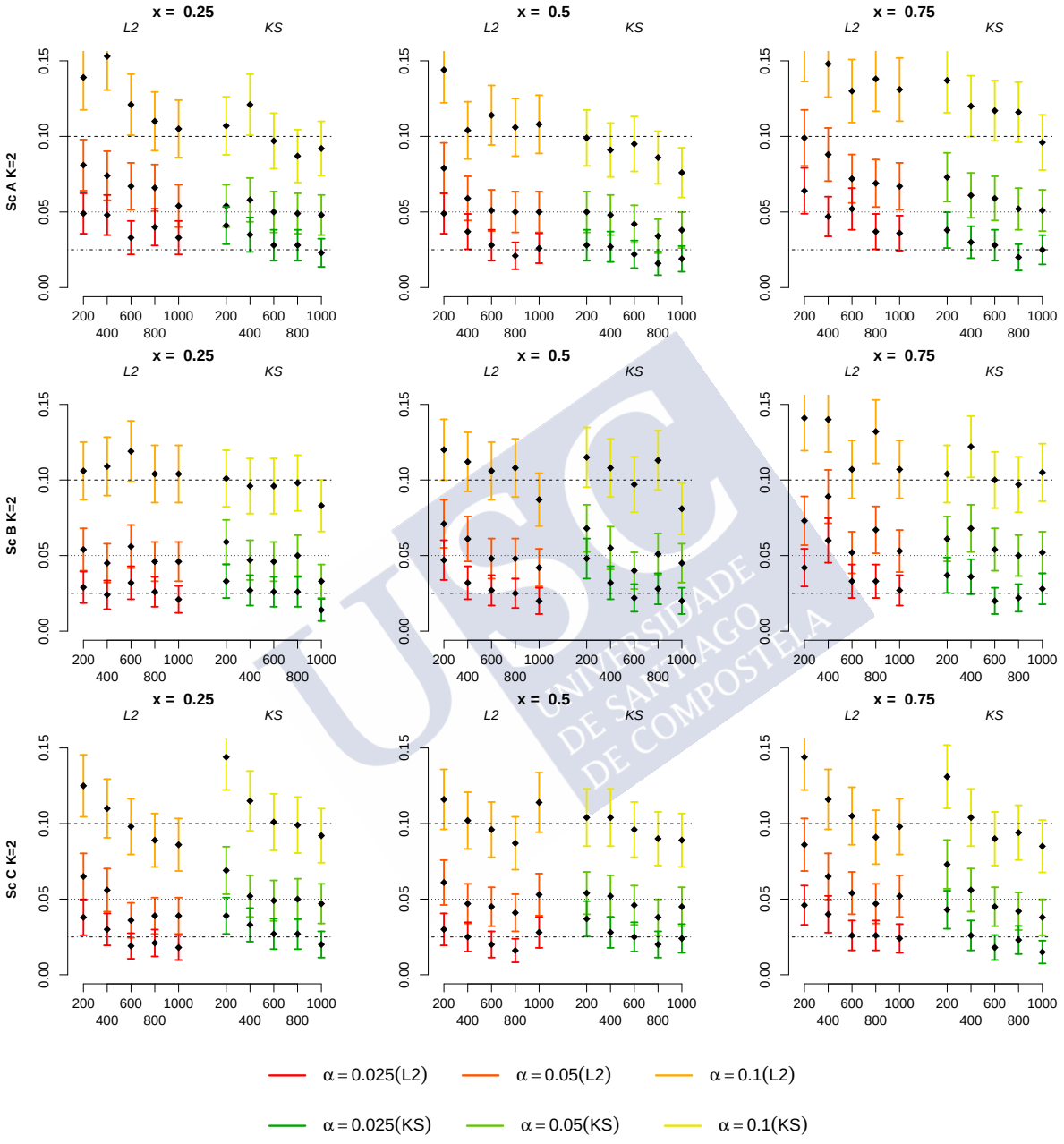


Figure B.1: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis for Scenarios A, B and C with $K = 2$ for different values of the covariate. The grid used for the cross-validation selection of the bandwidths had 100 points ranging from 0.075 to 0.7. Each graph contains the results for the statistic based on L_2 and KS and for different sample sizes N_k, where $N_k = n_k^F + n_k^G$.

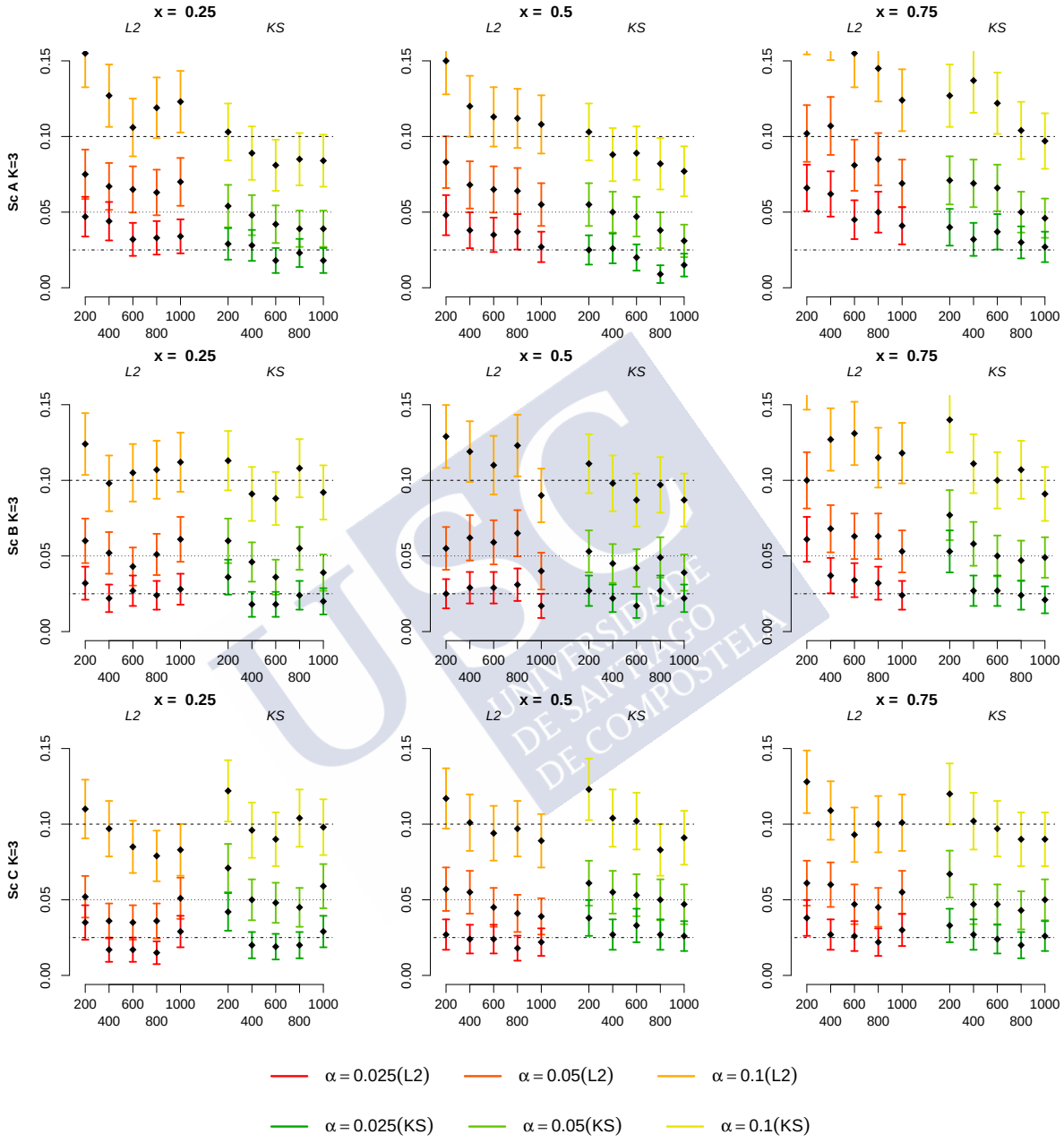


Figure B.2: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis for Scenarios A, B and C with $K = 3$ for different values of the covariate. The grid used for the cross-validation selection of the bandwidths had 100 points ranging from 0.075 to 0.7. Each graph contains the results for the statistic based on L_2 and KS and for different sample sizes N_k, where $N_k = n_k^F + n_k^G$.

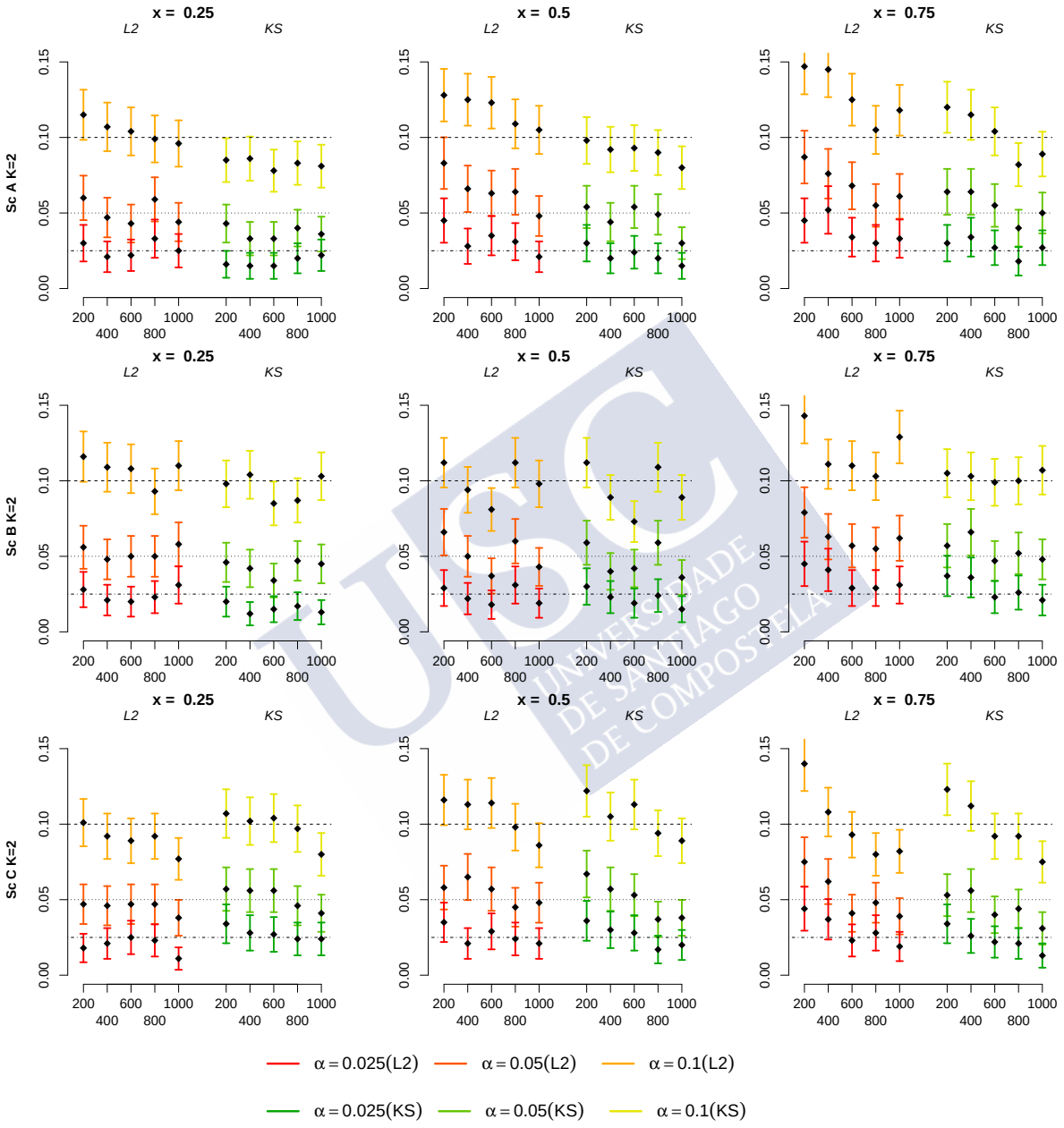


Figure B.3: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis for Scenarios A, B and C with $K = 2$ for different values of the covariate, using 200 bootstrap iterations. Each graph contains the results for the statistic based on L_2 and KS and for different sample sizes N_k, where $N_k = n_k^F + n_k^G$.

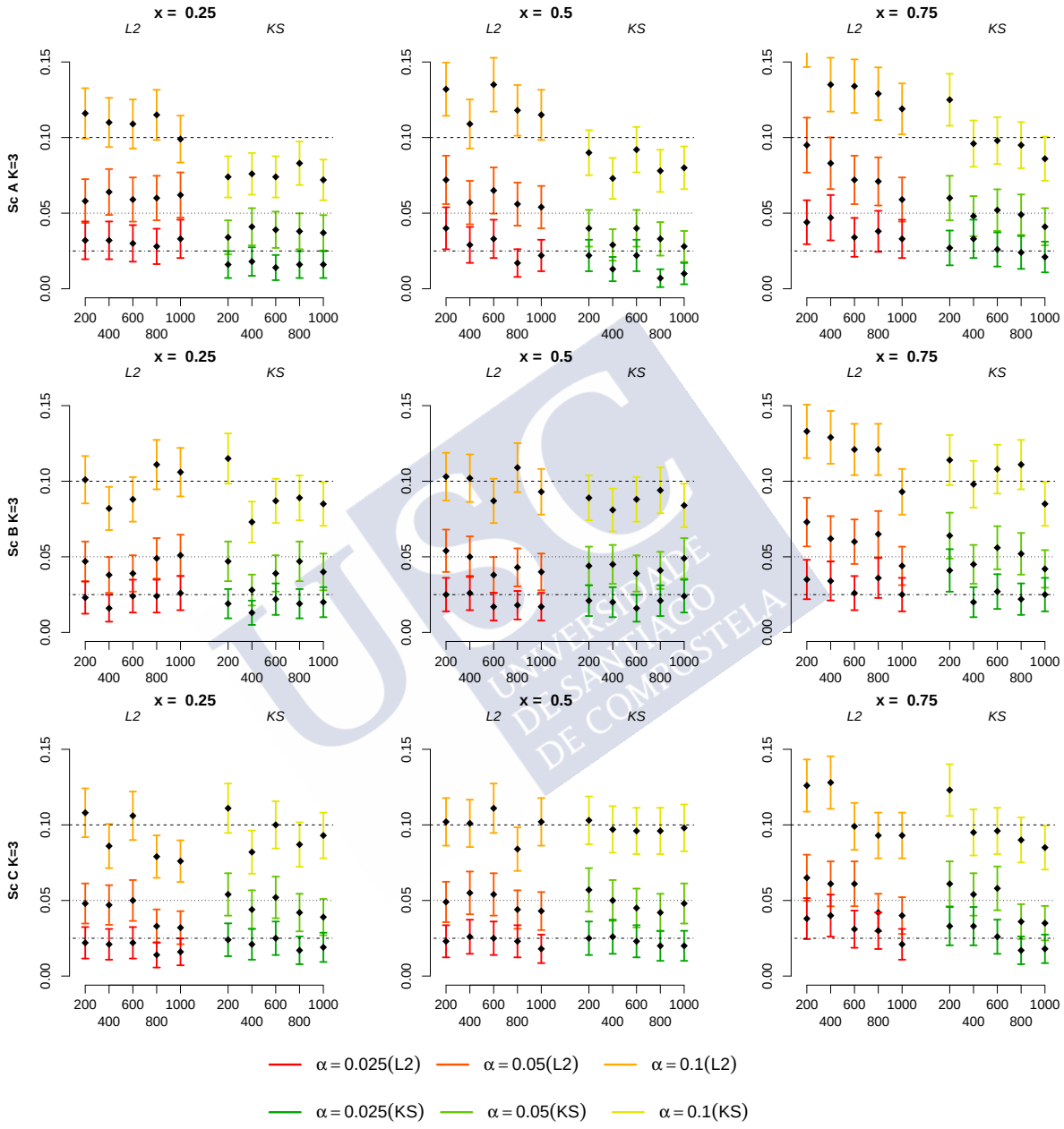


Figure B.4: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis for Scenarios A, B and C with $K = 3$ for different values of the covariate, using 200 bootstrap iterations. Each graph contains the results for the statistic based on L_2 and KS and for different sample sizes N_k, where $N_k = n_k^F + n_k^G$.

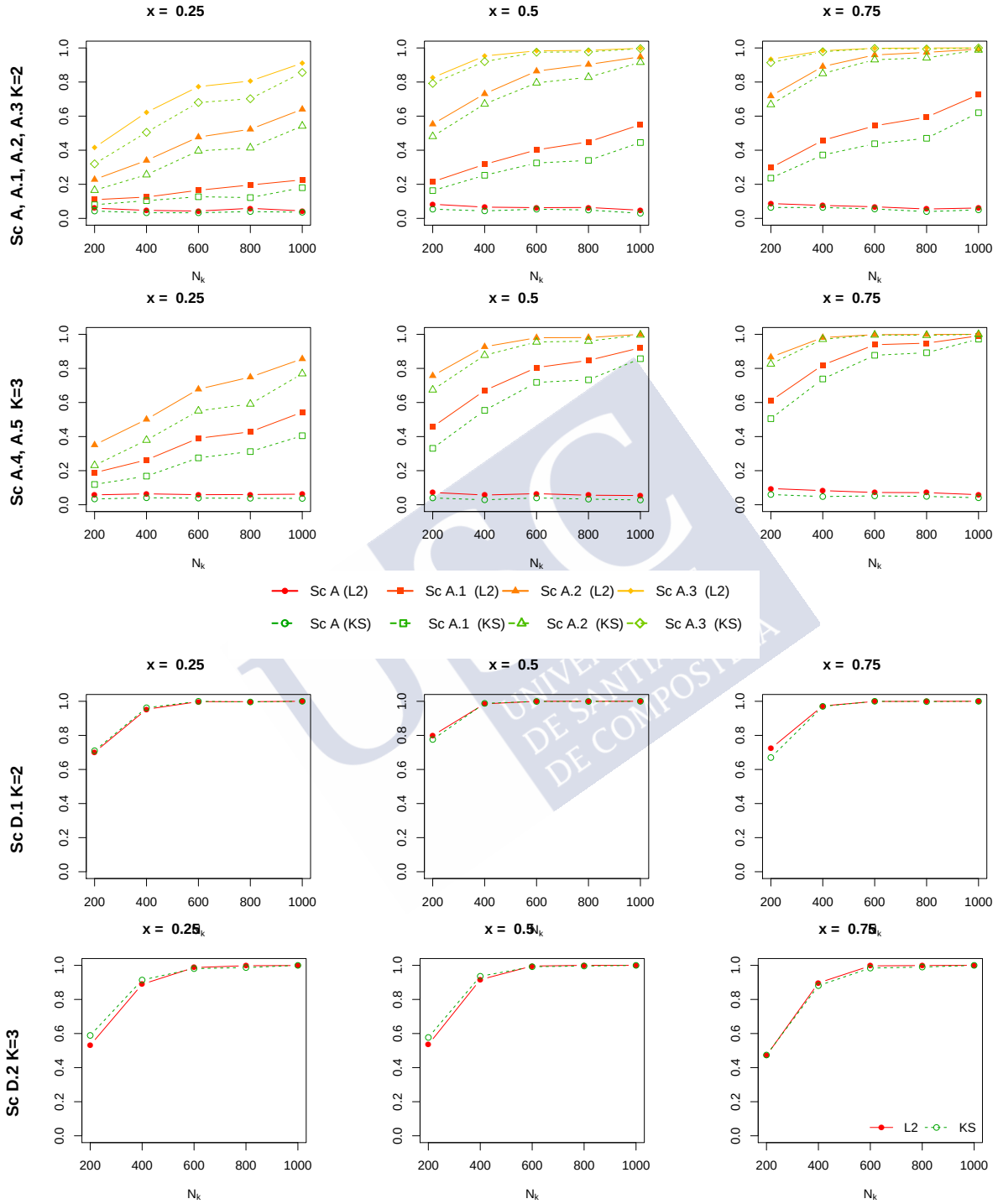


Figure B.5: Estimated proportion of rejection under the alternative hypothesis for different sample sizes and different scenarios ($\alpha = 0.05$) using 200 bootstrap iterations. Scenario A represents the situation under the null hypothesis.

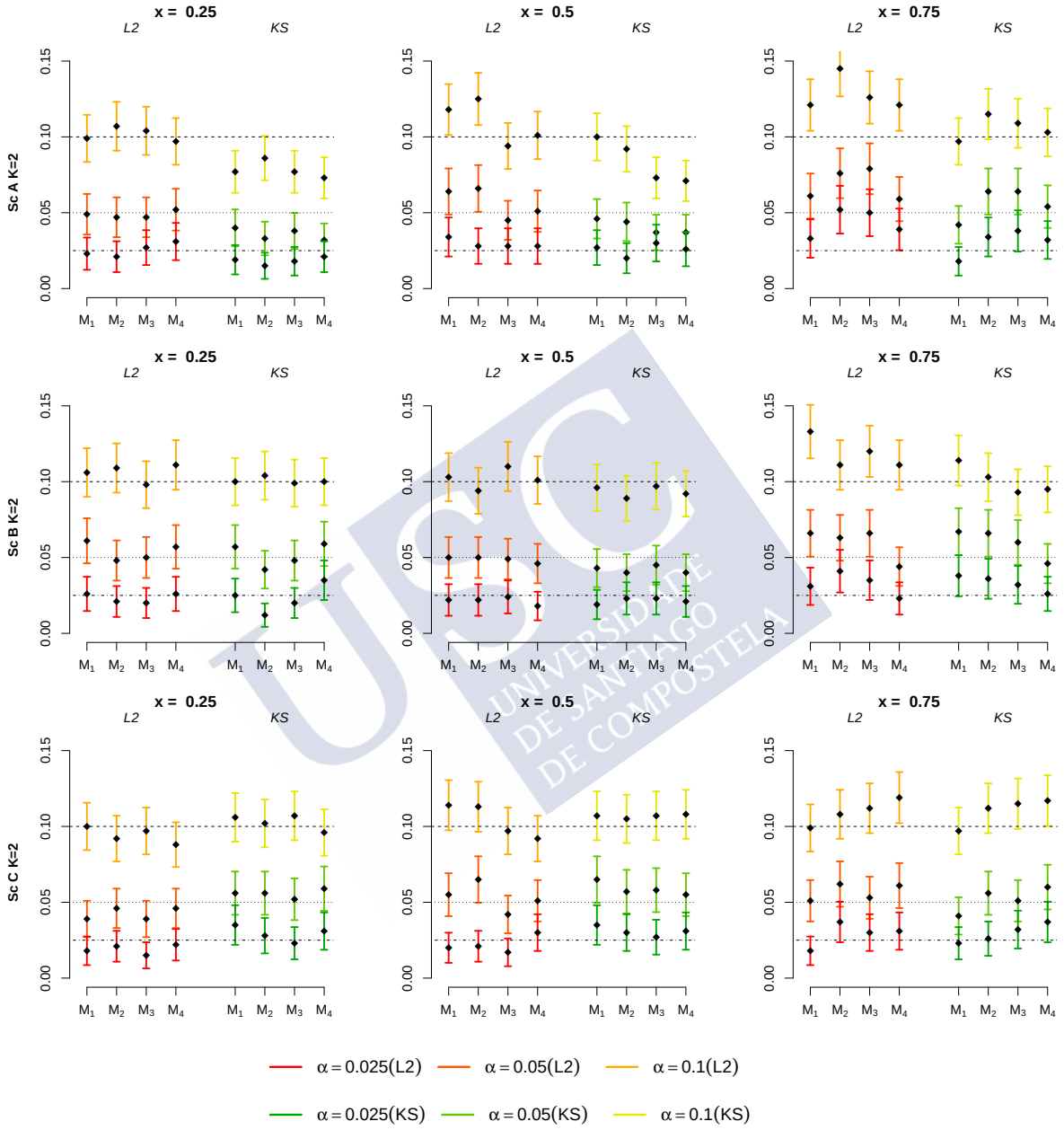


Figure B.6: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis for Scenarios A, B and C with $K = 2$ for different values of the covariate, using 200 bootstrap iterations. Each graph contains the results for the statistic based on L_2 and KS and for several unbalanced sample sizes (all of them adding up to a total of 800).

Table B.2: Estimated proportion of rejection under the alternative hypothesis for Scenarios A.1, A.2, A.3 and D.1, all of them with $K = 2$ (in columns), for different values of the covariate, using 200 bootstrap iterations. Each graph contains the results for the statistic based on L_2 and KS and for several unbalanced sample sizes (all of them adding up to a total of 800).

	Sample	L_2				KS			
x	size	$A.1$	$A.2$	$A.3$	$D.1$	$A.1$	$A.2$	$A.3$	$D.1$
0.25	M_1	0.131	0.349	0.599	0.949	0.092	0.259	0.493	0.943
	M_2	0.125	0.340	0.622	0.953	0.104	0.256	0.505	0.962
	M_3	0.115	0.303	0.560	0.971	0.070	0.208	0.426	0.972
	M_4	0.110	0.326	0.585	0.933	0.079	0.226	0.456	0.917
0.5	M_1	0.307	0.731	0.944	0.984	0.243	0.636	0.906	0.965
	M_2	0.317	0.731	0.953	0.985	0.253	0.671	0.920	0.988
	M_3	0.305	0.717	0.935	0.996	0.220	0.632	0.889	0.996
	M_4	0.282	0.726	0.928	0.984	0.209	0.608	0.877	0.963
0.75	M_1	0.445	0.895	0.992	0.983	0.332	0.831	0.979	0.946
	M_2	0.458	0.891	0.984	0.973	0.372	0.850	0.978	0.970
	M_3	0.463	0.878	0.985	0.997	0.355	0.803	0.972	0.995
	M_4	0.434	0.875	0.984	0.979	0.322	0.784	0.961	0.950

On the other hand, we collected the results for the scenarios under the alternative hypotheses in Table B.2. Given that all sets of sample size have the same total sample size (800), the power obtained in each scenario should be (and are) similar for all M_1, M_2, M_3 and M_4.

B.1.4 Comparison of dependent ROC curves

Remark 4.1 showed an adjustment that can be done to the bootstrap algorithm proposed in Section 4.2.3 to take into account the dependency that exists among ROC curves that are estimated using the same samples (i.e., with a common covariate). In Remark 4.1 we included a brief simulation study for that modification, and here we extend that analysis.

We considered three different scenarios (whose regression and conditional standard deviation functions are described in Table B.3) with similar parameters than the ones used for the independent simulation study, with some exceptions: the number of bootstrap iterations (here we use 200 instead of 500, although, as seen in Subsection B.1.2 this should not affect the results of the test), the sample sizes (now we took $(n^F, n^G) \in \{(100, 100), (200, 200), (500, 500)\}$) and the correlation that exists between the regression errors of the different conditional ROC curves ($\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$).

The simulations were run for scenarios A, B and C, only for the case with $K = 2$. The null hypothesis was tested for three values of the covariate x: 0.25, 0.5 and 0.75. The results for the scenario A are collected in Figures B.7, B.8 and B.9, each figure containing the simulations for a certain conditional value x. Likewise, the results for scenario B are collected in Figures B.10, B.11 and B.12, and the results for scenario C in Figures B.13, B.14 and B.15. As we did in Remark 4.1, we included the results for both the original bootstrap algorithm, that does not

Table B.3: Scenarios considered for the study of the comparison of dependent ROC curves.

Scenario	Regression functions	Conditional standard deviation functions
A	$\mu_A^F(x) = 0.5, \quad \mu_A^G(x) = 0$	$\sigma_1^F(x) = \sigma_1^G(x) = 0.5$
B	$\mu_B^F(x) = x, \quad \mu_B^G(x) = 0$	$\sigma_B^F(x) = \sigma_B^G(x) = 0.5$
C	$\mu_C^F(x) = 0.5x + \sin(0.5\pi x), \quad \mu_C^G(x) = 0.5x^2$	$\sigma_C^F(x) = \sigma_C^G(x) = 0.5 + 0.5x$

take the dependence structure into account, to show how it should not be used in this kind of scenarios.

In general, the modification of the proposed bootstrap algorithm seems to be able to capture the dependence structure, at least for the case of the L_2 test statistic. The results for the KS statistic are even more conservative than in the independence case studied in Section 4.3.1. Both statistics underestimate the proportion of rejections when the correlation between the ROC curves is high, but nevertheless it supposes a considerable improvement regarding the behaviour of the test for the independence case.

B.2 Supplementary material for Chapter 5

In this second part of this appendix we gather some extra simulation studies that were left out of Chapter 5. In Section B.2.1 we show a generalization of the test for comparing dependent ROC curves conditioned to unidimensional covariates. In Section B.2.2 we study an alternative way of approximating the statistic (5.12) proposed in Section 5.2.3. Finally, in Section B.2.3 we complete the simulation study briefly mentioned in Remark 5.4.

B.2.1 Generalization of the test for unidimensional covariates

In Section 5.2.2 of Chapter 5 we saw a test for comparing dependent ROC curves conditioned to a pair of values of a unidimensional covariate. This test is very similar to the one seen Chapter 4 (in that case, it was a test for independent ROC curves, and conditioning in only one value of the covariate). Here we provide a generalization of the test of Section 5.2.2. Find below the new proposal, along with a simulation study.

In this case, we present a test for comparing two or more dependent ROC curves conditioned to K pairs of one-dimensional values. Given the values $(x_1^F, x_1^G), \ldots, (x_K^F, x_K^G) \in R_{X^F} \times R_{X^G}$, the aim is then to test

$$H_0 : ROC_1^{x_1^F, x_1^G}(p) = \cdots = ROC_K^{x_K^F, x_K^G}(p) \quad \text{for all } p \in (0,1) \tag{B.1}$$

against the general alternative $H_1 : H_0$ is not true. In fact, in Chapter 5 what we are doing is using this test with $x_1^F = \cdots = x_K^F = (\boldsymbol{\beta}^F)'\boldsymbol{x}$ and $x_1^G = \cdots = x_K^G = (\boldsymbol{\beta}^G)'\boldsymbol{x}$.

The samples available are the same as for the test in Section 5.2.2, and the same nonparametric location-scale regression models are assumed to accommodate the diagnostic variables (this is because the changes we are doing for this generalization only affect the values of the covariate at which we are conditioning, not the variables themselves).

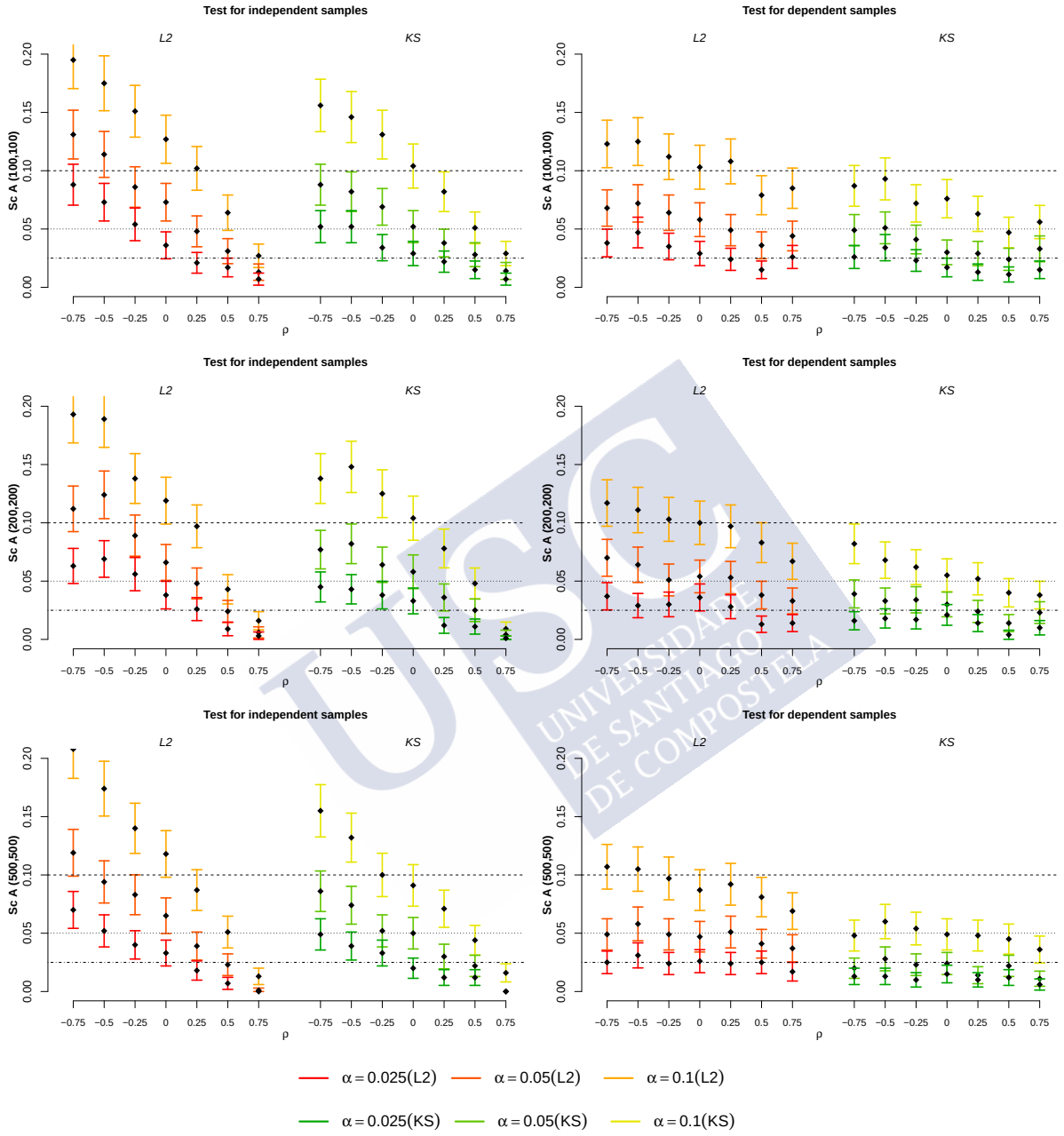


Figure B.7: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario A (with $K = 2$) at $x = 0.25$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

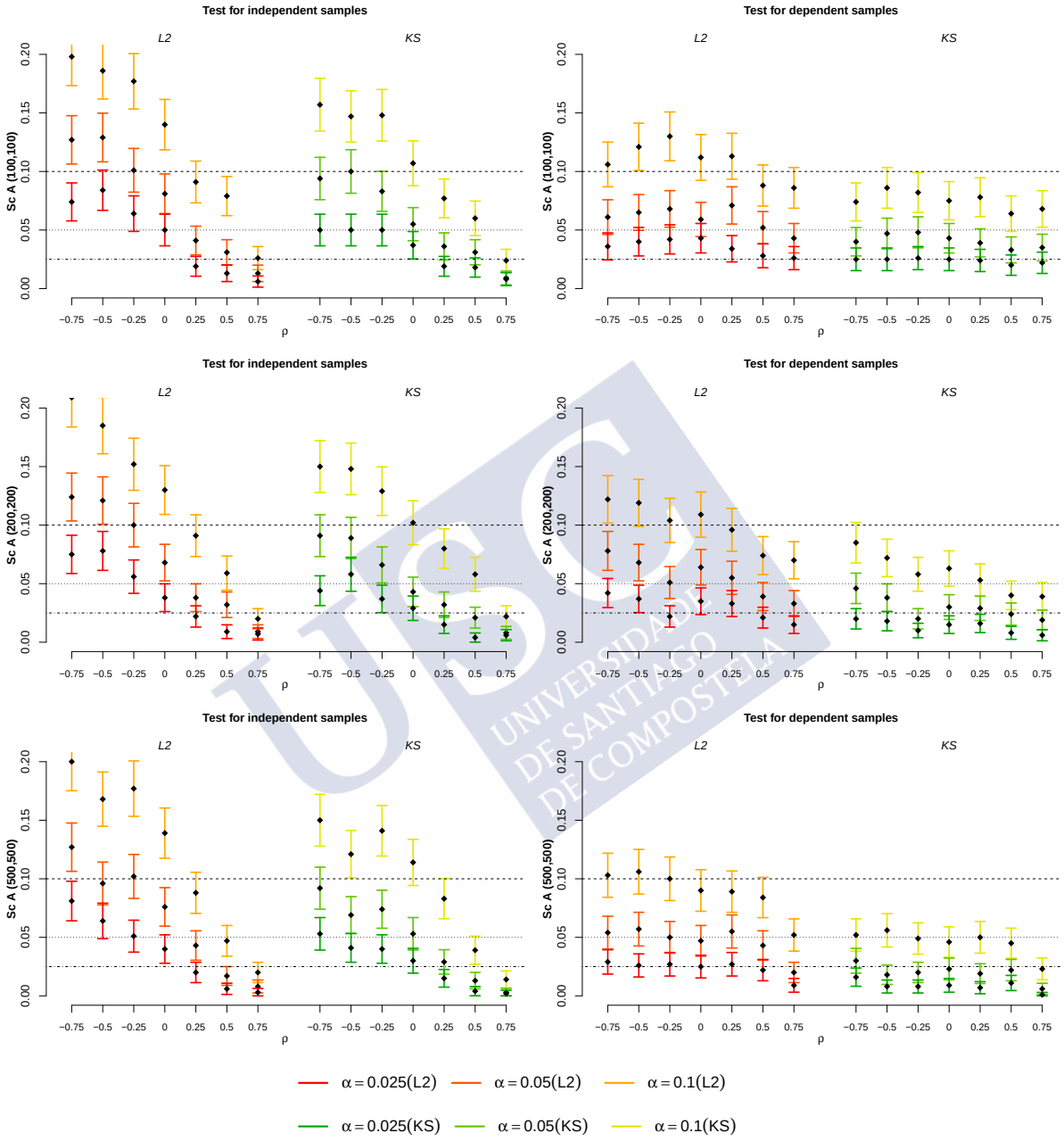


Figure B.8: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario A (with $K = 2$) at $x = 0.5$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

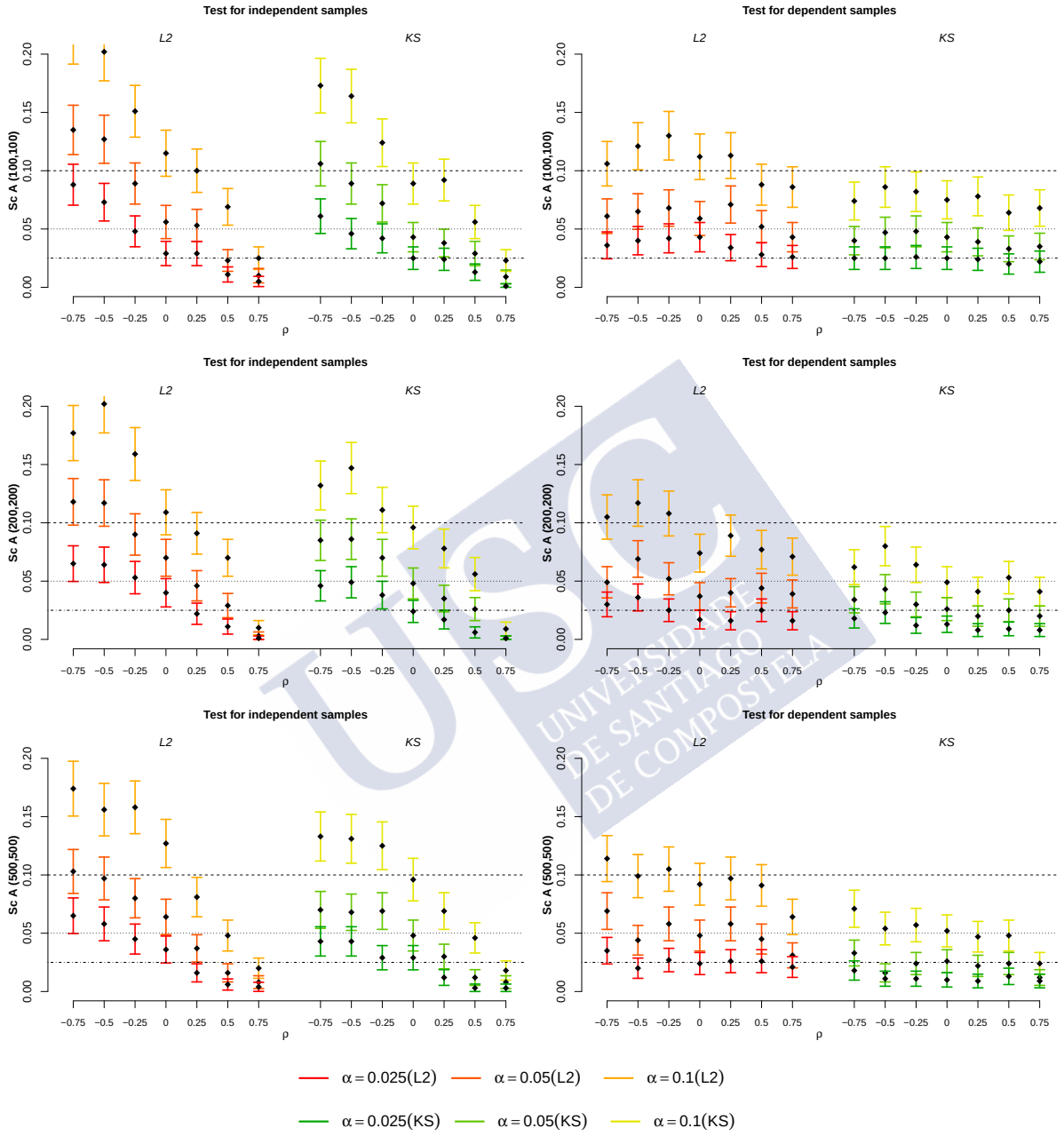


Figure B.9: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario A (with $K = 2$) at $x = 0.75$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

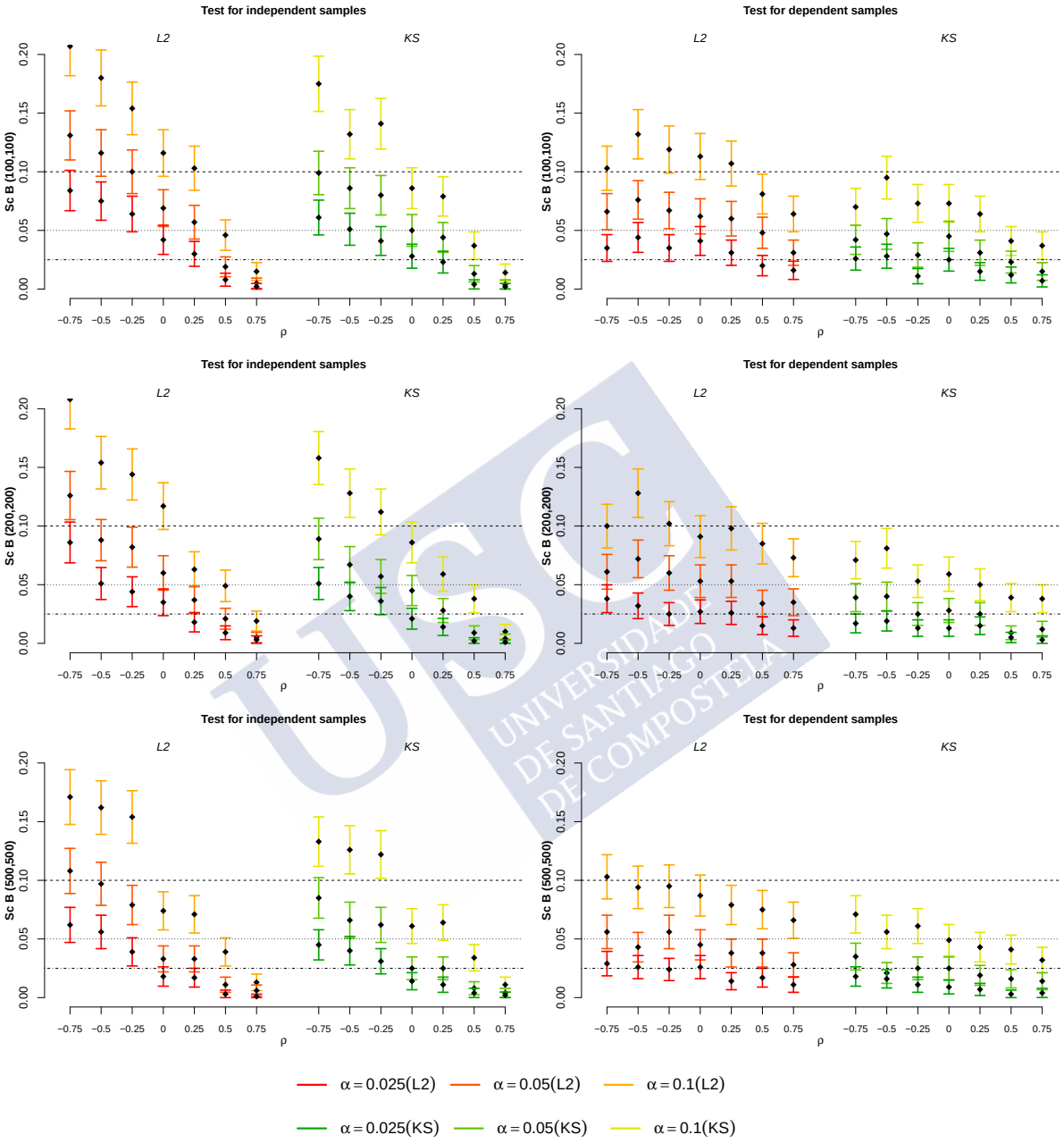


Figure B.10: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario B (with $K = 2$) at $x = 0.25$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

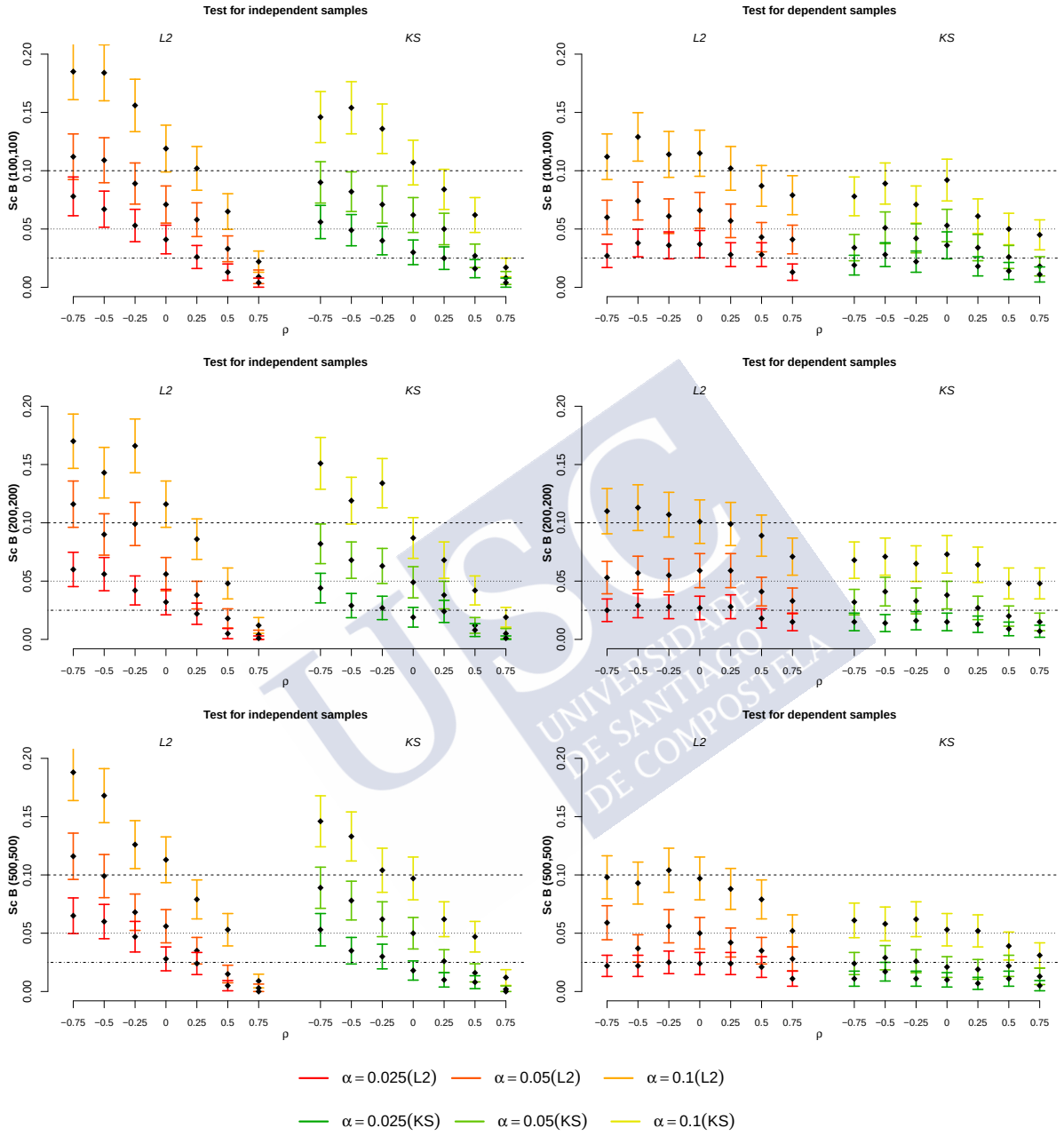


Figure B.11: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario B (with $K = 2$) at $x = 0.5$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

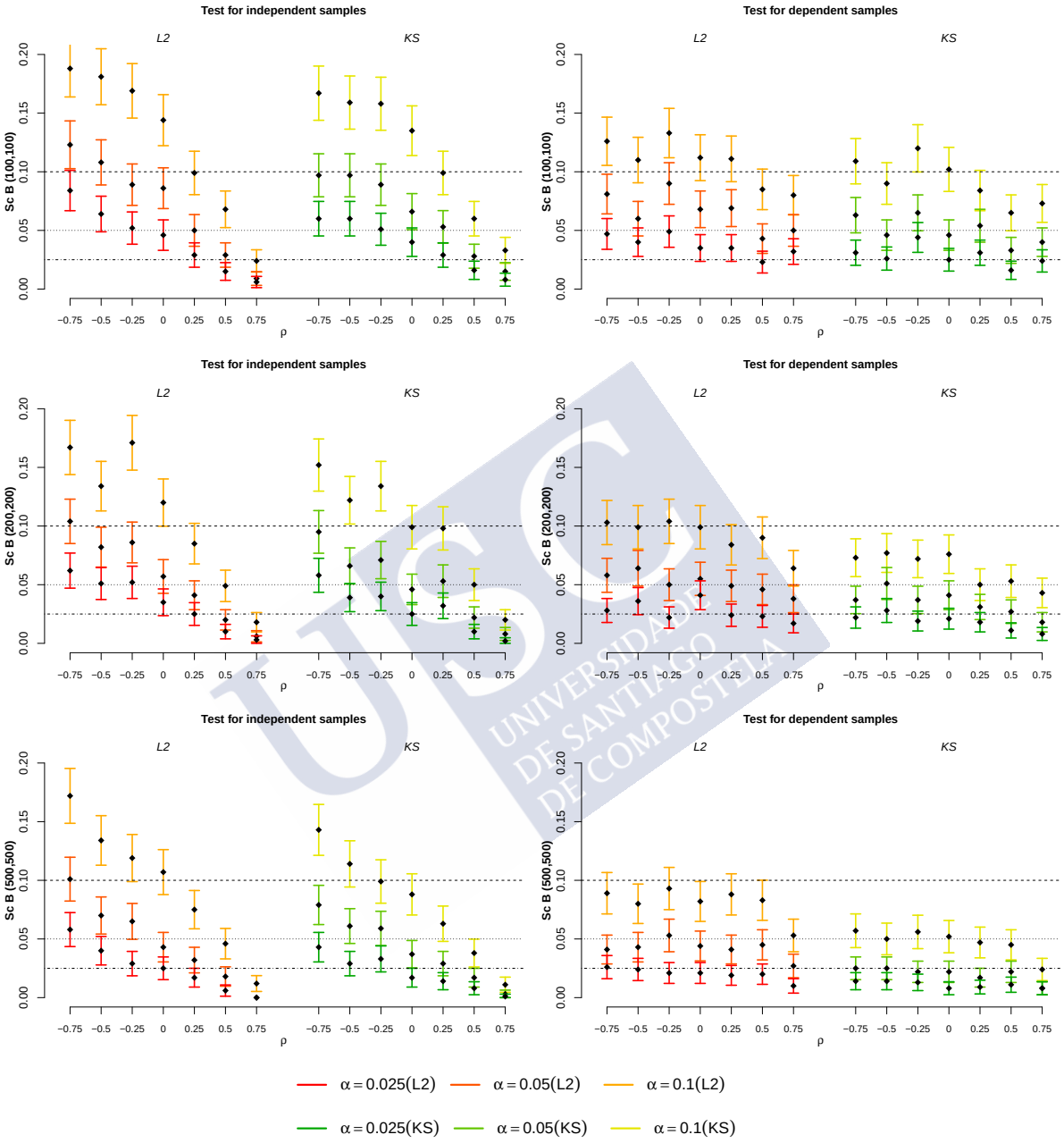


Figure B.12: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) *and different levels of dependence between the compared ROC curves,* $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$*, of the scenario B (with* $K = 2$*) at* $x = 0.75$*. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the* L_2 *and* KS *test statistics.*

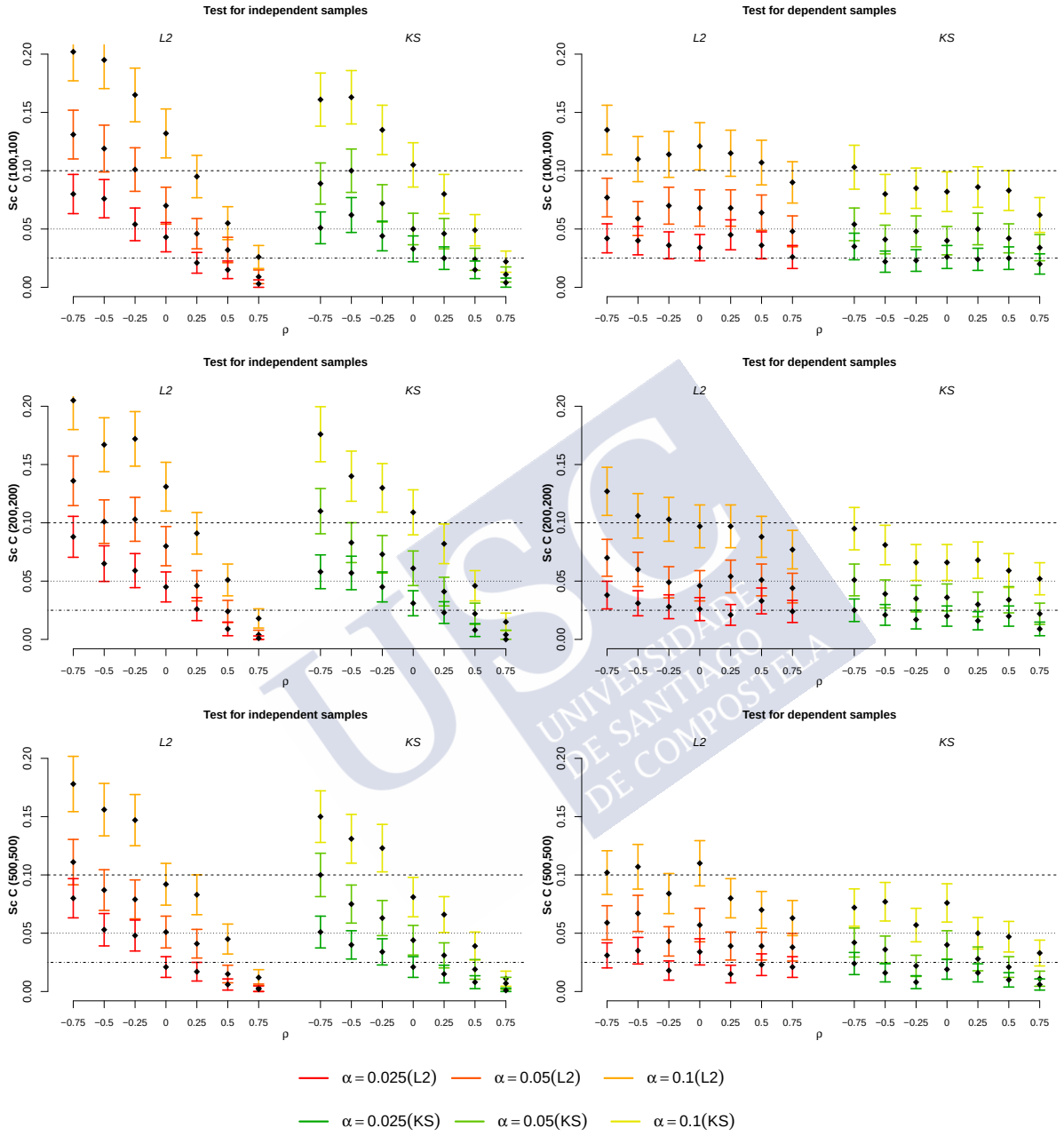


Figure B.13: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario C (with $K = 2$) at $x = 0.25$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

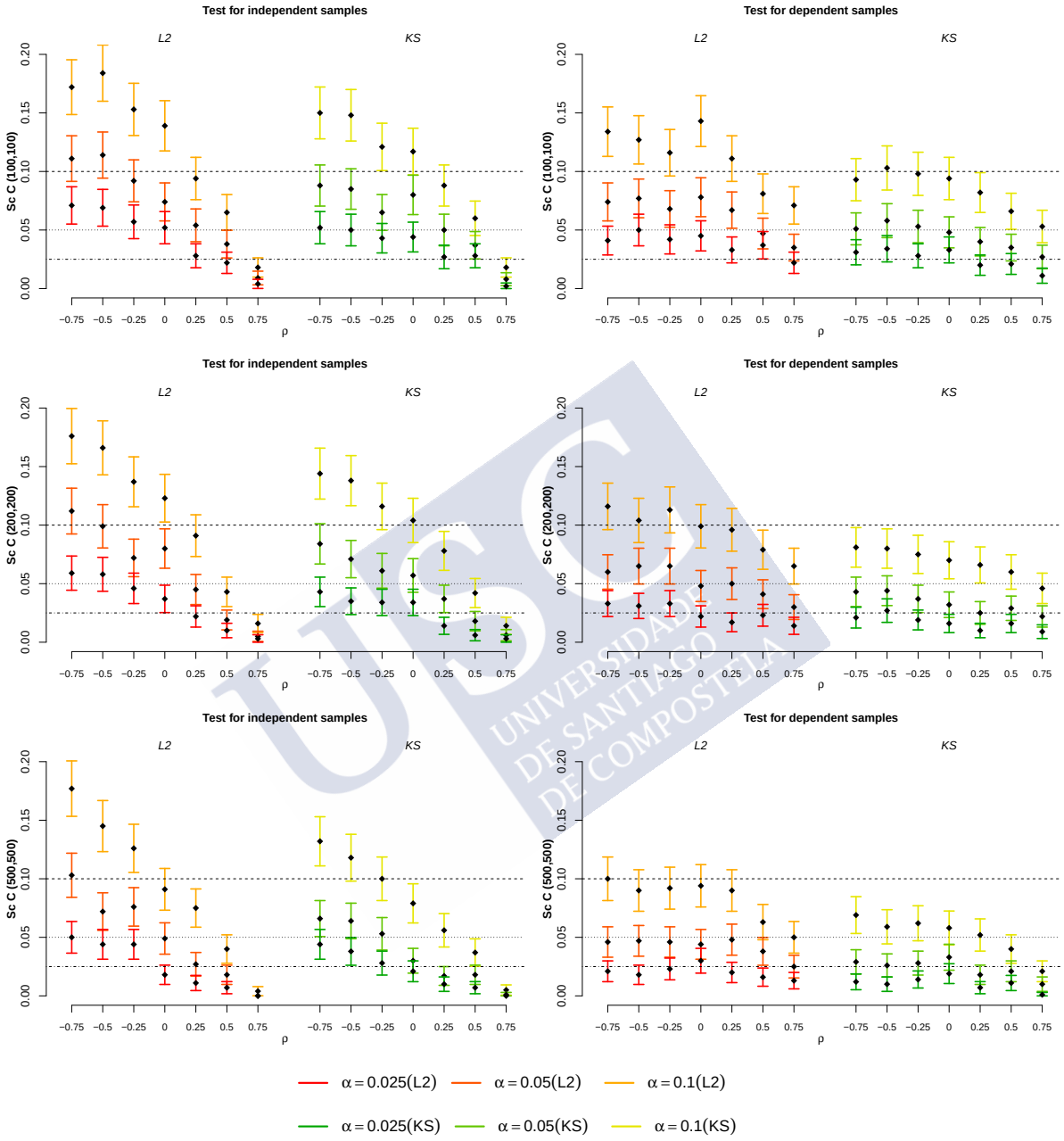


Figure B.14: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario C (with $K = 2$) at $x = 0.5$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

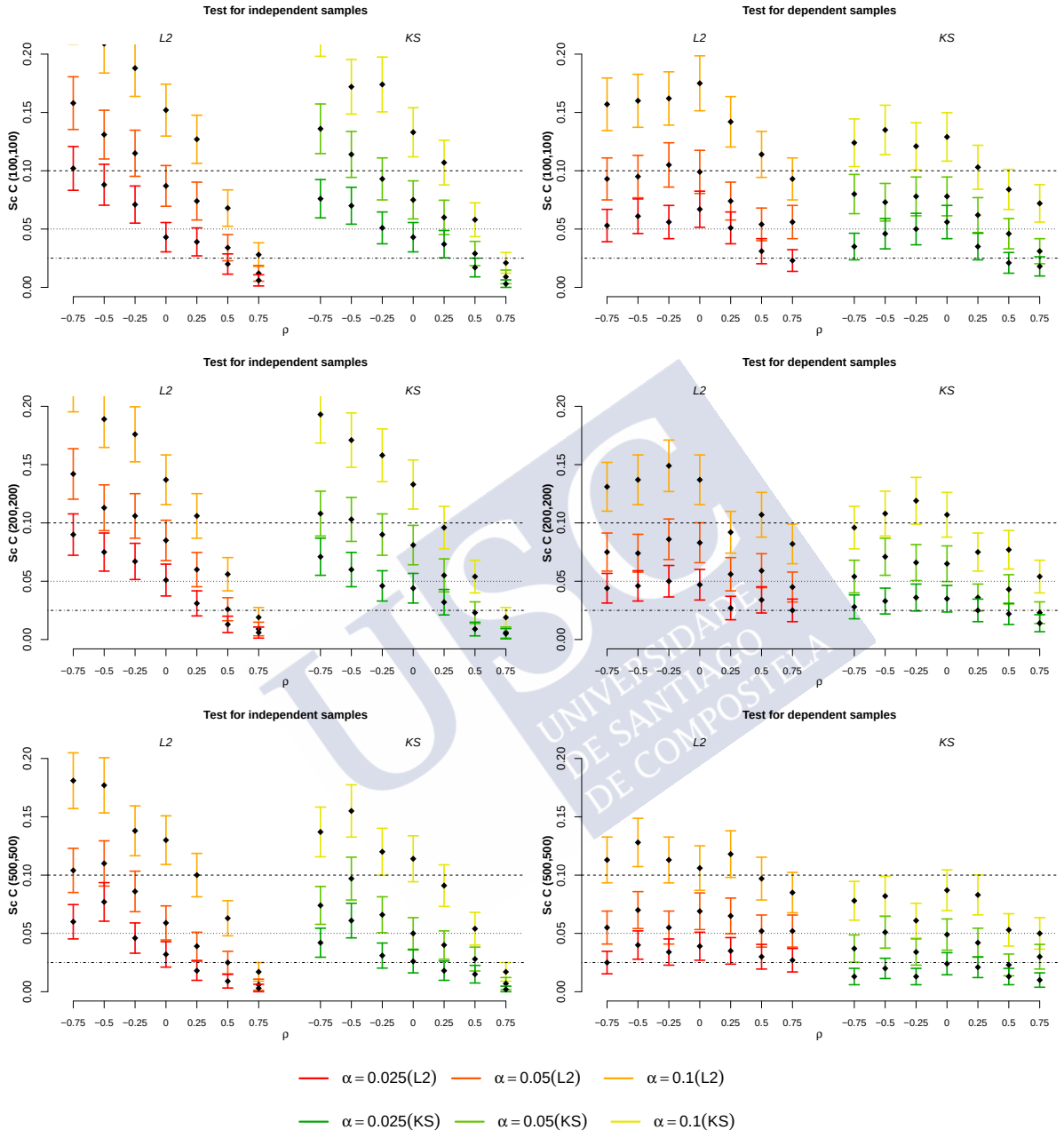


Figure B.15: Estimated proportion of rejections under the null hypothesis for different sample sizes (n^F, n^G) and different levels of dependence between the compared ROC curves, $\rho \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, of the scenario C (with $K = 2$) at $x = 0.75$. Each row contains the results for two tests: one that ignores the dependence structure (left) and one that acknowledges that dependence (right), both using the L_2 and KS test statistics.

In order to test (B.1), the following test statistic is proposed:

$$S^x = \sum_{k=1}^{K} \psi\left(\sqrt{g_k N}\{\widehat{ROC}_k^{x_k^F, x_k^G}(p) - \widehat{ROC}_\bullet^{x^F, x^G}(p)\}\right),$$

where:

- for $k \in \{1, \ldots, K\}$, given a pair (x_k^F, x_k^G),

$$\widehat{ROC}_k^{x^F, x^G}(p) = 1 - \int \hat{H}_k^F\left((\hat{H}_k{}^G)^{-1}(1-p+h_k u)\hat{b}_k(x_k^F, x_k^G) - \hat{a}_k(x_k^F, x_k^G)\right)\kappa(u)du$$

  is the estimated conditional ROC curve given (x_k^F, x_k^G), where, for $D \in \{F, G\}$,

  - $\hat{H}_k^D(y) = (n^D)^{-1}\sum_{i=1}^{n^D} I(\hat{\varepsilon}_{k,i}^D \leqslant y)$,
  - $\hat{\varepsilon}_{k,i}^D = \dfrac{Y_{k,i}^D - \hat{\mu}_k^D(X_i^D)}{\hat{\sigma}_k^D(X_i^D)}$, for $i \in \{1, \cdots, n^D\}$,
  - $\hat{\mu}_k^D(x) = \sum_{i=1}^{n^D} W_{k,i}^D(x, g_k^D)Y_{k,i}^D$ is a nonparametric estimator of $\mu_k^D(x)$ based on local weights $W_{k,i}^D(x, g_k^D)$ depending on a bandwidth parameter g_k^D,
  - $(\hat{\sigma}_k^D)^2(x) = \sum_{i=1}^{n^D} W_{k,i}^D(x, g_k^D)[Y_{k,i}^D - \hat{\mu}_k^D(X_i^D)]^2$ is a estimator of $(\sigma_k^D)^2(x)$. For simplicity we take the same bandwidth parameter g_k^D that is used for the estimation of the regression function $\hat{\mu}_k^D(x)$,
  - $W_{k,i}^D(x, g_k^D) = \dfrac{\kappa_{g_k^D}(x - X_i^D)}{\sum_{l=1}^{n^D}\kappa_{g_k^D}(x - X_l^D)}$, for $i \in \{1, \cdots, n^D\}$, are Nadaraya-Watson-type weights, where $\kappa_{g_k^D}(\cdot) = \kappa(\cdot/g_k^D)/g_k^D$ and κ is a probability density function symmetric around zero.
  - $\hat{a}_k(x_k^F, x_k^G) = \frac{\hat{\mu}_k^F(x_k^F) - \hat{\mu}_k^G(x_k^G)}{\hat{\sigma}_k^F(x_k^F)}$ and $\hat{b}_k(x_k^F, x_k^G) = \frac{\hat{\sigma}_k^G(x_k^G)}{\hat{\sigma}_k^F(x^F)}$.

- $g_k = \frac{n^F g_k^F + n^G g_k^G}{N}$, where g_k^F and g_k^G are the bandwidth parameters related to the estimation of the conditional means and variances of the location-scale regression models of the diseased and the healthy populations involved in the estimation of the k-th conditional ROC curve. Likewise, for each one of the estimated ROC curves there will be another bandwidth parameter, h_k, responsible for the smoothness of the estimator.
- $\widehat{ROC}_\bullet^{x^F, x^G}(p) = \left(\sum_{k=1}^K g_k N\right)^{-1}\sum_{k=1}^K g_k N \widehat{ROC}_k^{x_k^F, x_k^G}(p)$ is a weighted average of the K conditional ROC curves.
- ψ is a real-valued function that will measure the difference from one estimated conditional ROC curve to the weighted average of all of them.

The null hypothesis will be rejected for large values of S^x. In order to obtain the distribution of this statistic, we propose a bootstrap algorithm analogous to the one use previously in Section 5.2.2. The key of this algorithm is, once again, that

$$T^x = \sum_{k=1}^{K} \psi\left(\sqrt{g_k N}\{(\widehat{ROC}_k^{x_k^F, x_k^G}(p) - \widehat{ROC}_\bullet^{x^F, x^G}) - (ROC_k^{x_k^F, x_k^G}(p) - ROC_\bullet^{x^F, x^G}(p))\}\right),$$

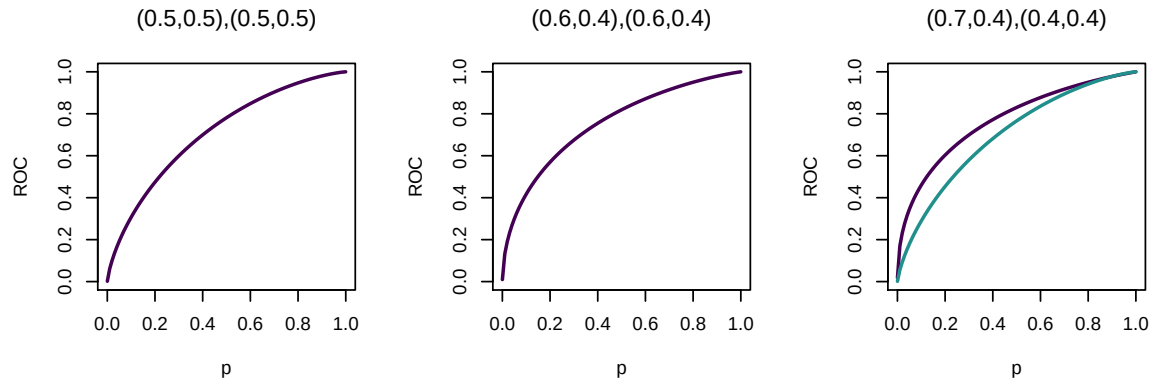


Figure B.16: Conditional ROC curves considered for the simulation study ($K = 2$).

coincides with the statistic S^x as long as the null hypothesis holds, where

$$ROC_{\bullet}^{x^F,x^G}(p) = \left(\sum_{k=1}^{K} g_k N\right)^{-1} \sum_{k=1}^{K} g_k N ROC_k^{x_k^F,x_k^G}(p),\ 0 < p < 1.$$

This bootstrap algorithm, used to approximate the distribution of this test statistic, follows the same steps than the one explained in Section 5.2.2 of Chapter 5. Note that the main change of this methodology with the one described in Section 5.2.2 is limited to the estimation of the conditioned ROC curve, and the only differences in both estimations is the estimation of functions $a_k(\cdot,\cdot)$ and $b_k(\cdot,\cdot)$.

Simulation study

We run a simulation study to analyse the practical performance of the test in terms of level approximation and power. We considered two different scenarios, one with $K = 2$ and the other with $K = 3$.

In all these scenarios only one conditional ROC curve (ROC^{x^F,x^G}) was considered: the one based on the location-scale regression models with $\mu^F(x^F) = \sin(0.5\pi x^F)$, $\mu^F(x^G) = 0.5(x^G)^2$, $\sigma^F(x^F) = 0.5 + 0.5x^F$ and $\sigma^G(x^G) = 0.5 + 0.5x^G$, with normal regression errors. The covariates X^F and X^G were considered to follow a uniform distribution on the unit interval.

In the first scenario we compare two ROC curves ($K = 2$), both of them obtained using the same functions previously described. In this case we made the comparison at three different points, that we will call A, B and C:

$$\begin{aligned} A &: (x_1^F, x_1^G) = (0.5, 0.5), (x_2^F, x_2^G) = (0.5, 0.5), \\ B &: (x_1^F, x_1^G) = (0.6, 0.4), (x_2^F, x_2^G) = (0.6, 0.4), \\ C &: (x_1^F, x_1^G) = (0.7, 0.4), (x_2^F, x_2^G) = (0.4, 0.4). \end{aligned}$$

The ROC curve conditioned at those points are depicted in Figure B.16. The first and second set of points considered result in equal ROC curves (and thus, uphold the null hypothesis (B.1)) but the third ones result in two different ROC curves (and hence, under the alternative hypothesis).

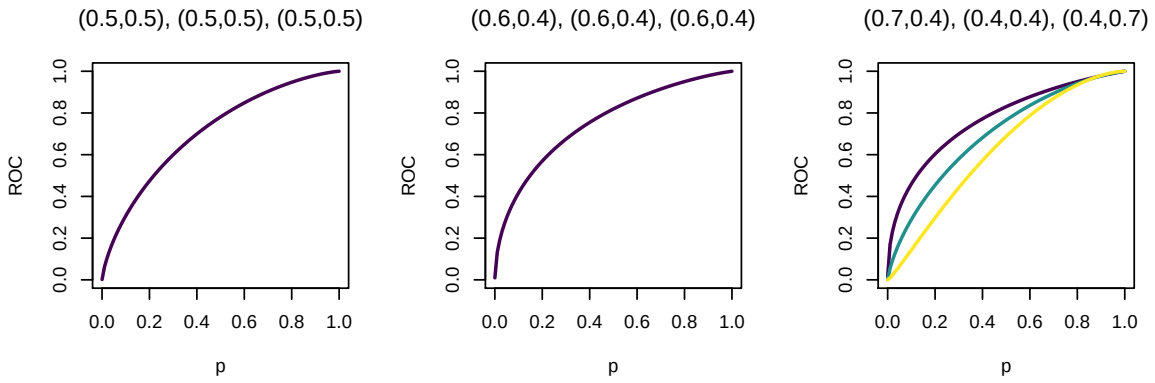


Figure B.17: Conditional ROC curves considered for the simulation study ($K = 3$).

For the second scenario, three curves were compared ($K = 3$). We use the same one as before, but this time the sets of points at which we condition the curves are:

$$\begin{aligned} D &: (x_1^F, x_1^G) = (0.5, 0.5), (x_2^F, x_2^G) = (0.5, 0.5), (x_3^F, x_3^G) = (0.5, 0.5), \\ E &: (x_1^F, x_1^G) = (0.6, 0.4), (x_2^F, x_2^G) = (0.6, 0.4), (x_3^F, x_3^G) = (0.6, 0.4), \\ F &: (x_1^F, x_1^G) = (0.7, 0.4), (x_2^F, x_2^G) = (0.4, 0.4), (x_3^F, x_3^G) = (0.4, 0.7). \end{aligned}$$

The ROC curve conditioned at those points are depicted in Figure B.17. As it happened for the previous scenario, the first and second sets of points at which we made the comparison result in equal ROC curves (under the null hypothesis), whereas the third one results in different curves (under the alternative hypothesis).

The four sample sizes considered for the study were $(n, m) \in \{(100, 100), (250, 150), (250, 350), (400, 400)\}$. 1000 replications were used to estimate the proportion of rejection in each case. As we did in Chapter 5, different correlations were assumed between the diagnostic markers, with $\rho \in \{-0.5, 0, 0.5\}$, and, in the case of $K = 3$, we also considered

$$\Sigma = \begin{pmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}$$

as the variance-covariance matrix of the regression errors. Likewise, both the L_2 and the KS kind of statistic were considered.

The results of the simulation study for the first scenario (with $K = 2$) are collected in Figure B.18, and the results for the second scenario (with $K = 3$) are in Figure B.19. Note that, for the first two columns of both figures, we are calibrating the test under the null hypothesis, whereas in the third column we are studying the power of the test.

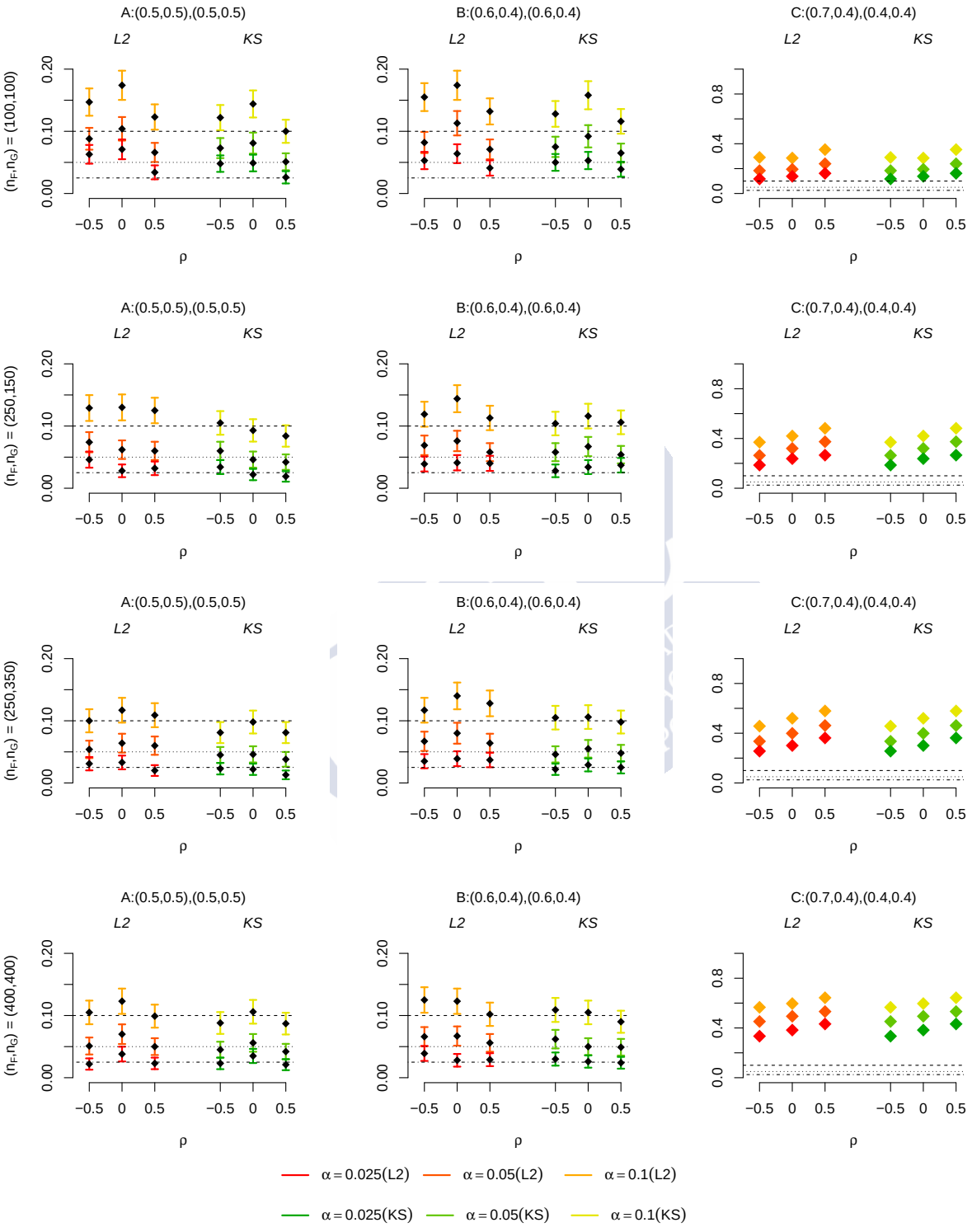


Figure B.18: Estimated proportion of rejections (for different levels $\alpha \in \{0.025, 0.05, 0.1\}$ in different colours) of the scenario with $K = 2$ for different pairs of values of the covariate (columns) and different sample sizes (rows). The confidence intervals were added in the first two columns (situations under the null hypothesis). Note that the axes of the third column are in a different scale.

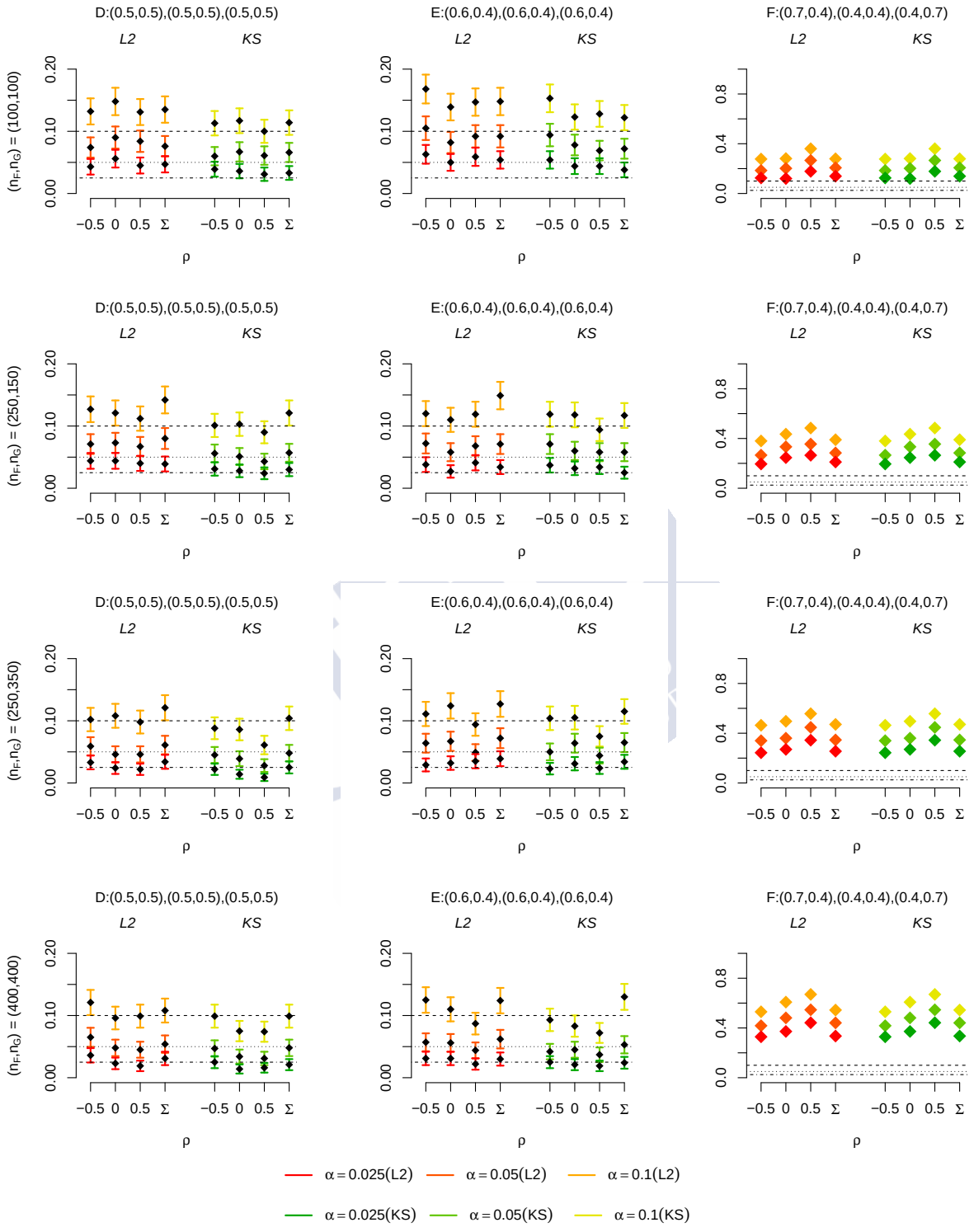


Figure B.19: Estimated proportion of rejections (for different levels $\alpha \in \{0.025, 0.05, 0.1\}$ in different colours) of the scenario with $K = 3$ for different pairs of values of the covariate (columns) and different sample sizes (rows). The confidence intervals were added in the first two columns (situations under the null hypothesis). Note that the axes of the third column are in a different scale.

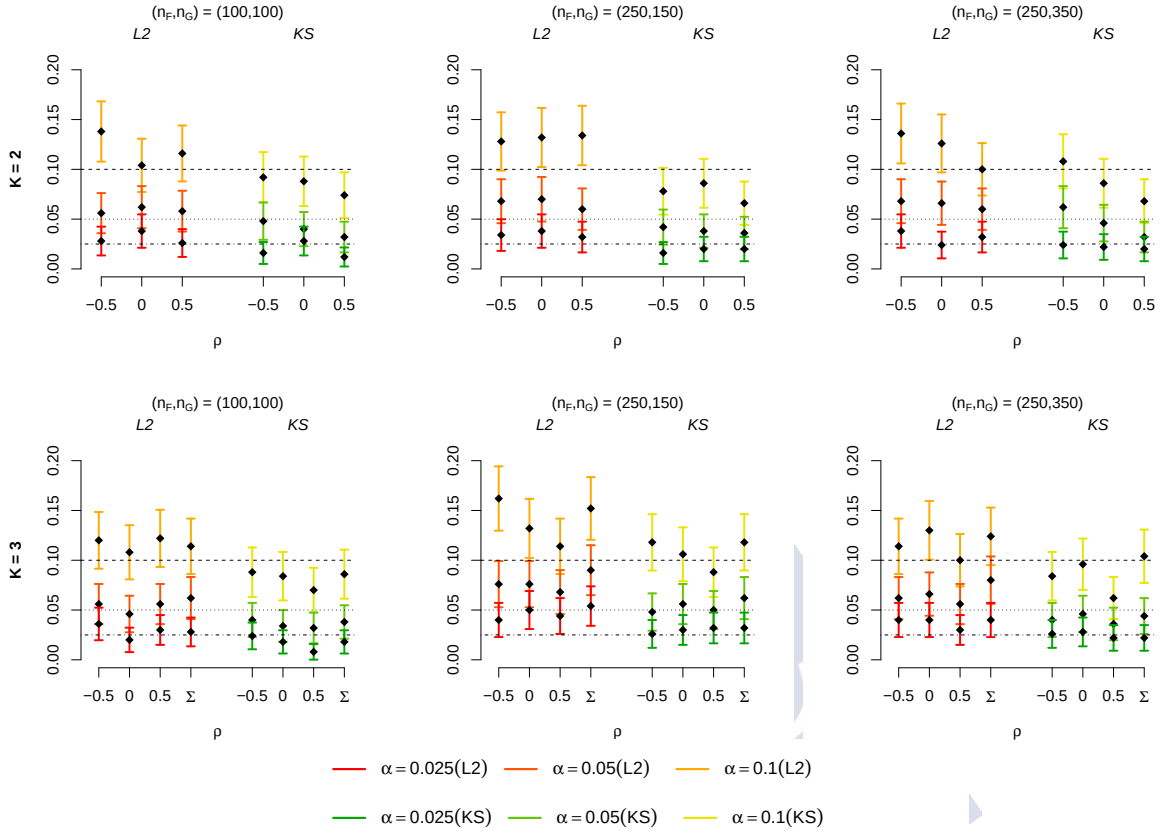


Figure B.20: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis with $d = 2$ and $n_{\boldsymbol{\beta}} = 5$ for different sample sizes and different ρ, using evenly spaced directions for the approximation of the integral of the test statistic.

B.2.2 Evenly spaced projections

In Section 5.2.3 we discussed a way of approximating the integral of the statistic (5.12) by taking random directions $\boldsymbol{\beta}_1^F, \ldots, \boldsymbol{\beta}_{n_\beta}^F$ and $\boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_{n_\beta}^G$ from $\mathbb{S}^{d-1}$. Here, instead of taking random directions, we approximate the integral by taking evenly spaced directions in the unit circumference (for the particular case of $d = 2$).

We repeat the same simulation study carried out in Section 5.3, although only for $d = 2$. For $d > 2$ the d-dimensional sphere $\mathbb{S}^{d-1}$ is not easily covered with evenly distributed points (and it is not even possible to do so for any number $n_{\boldsymbol{\beta}}$).

Figure B.20 is analogous to Figure 5.1 of Chapter 5, but considering equidistant projections instead random projections.

Likewise, Figure B.21 is analogous to Figure 5.3 of Chapter 5: it shows the results of the simulations concerning the power of the test but considering equidistant projections instead random projections.

In general, the proportions of rejection are close to their corresponding nominal levels, and the power of the test increases when the sample size increases. The results are very similar to the ones obtained in Chapter 5, although the power obtained with this evenly spaced configuration is slightly higher.

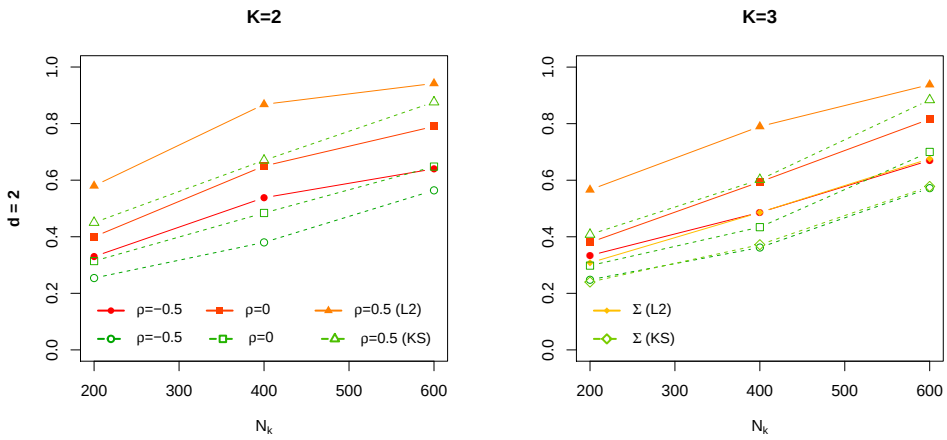


Figure B.21: Estimated proportion of rejection under the alternative hypothesis for different sample sizes and different ρ, with $d = 2$ and $n_{\boldsymbol{\beta}} = 5$ ($\alpha = 0.05$), using evenly spaced directions for the approximation of the integral of the test statistic.

B.2.3 Alternative test statistic approximation

In Remarks 5.1 and 5.2, an alternative way of approximating the test statistic (5.12) was proposed. The behaviour of the statistic with this approximation was briefly studied in the simulation section of Chapter 5, in Remark 5.3. Here we complete that simulation study.

First we have the results concerning the study of the level of the test: we considered the scenarios under the null hypothesis with $d = 2$ and $K = 3$ (Figure B.23), with $d = 3$ and $K = 2$ (Figure B.22) and with $d = 3$ and $K = 3$ (Figure B.24). Note that the corresponding simulations for the scenario with $d = 2$ and $K = 2$ is already displayed in Chapter 5 (Figure 5.4). For all of those cases we observed what happened for $m_{\boldsymbol{\beta}} = 50$, 25 and 1.

Secondly, we study the power of the test, as we did in Figure 5.5 of Chapter 5, this time with the scenarios under the alternative hypothesis with $K = 2, 3$ and $d = 2, 3$ (note that the configuration $K = 2$ and $d = 2$ was the one set as an example in Chapter 5). The results can be observed in Figure B.25.

The results are very similar to the ones obtained for the case with $d = 2$ and $K = 2$, for both the study of the nominal level and the power of the test.

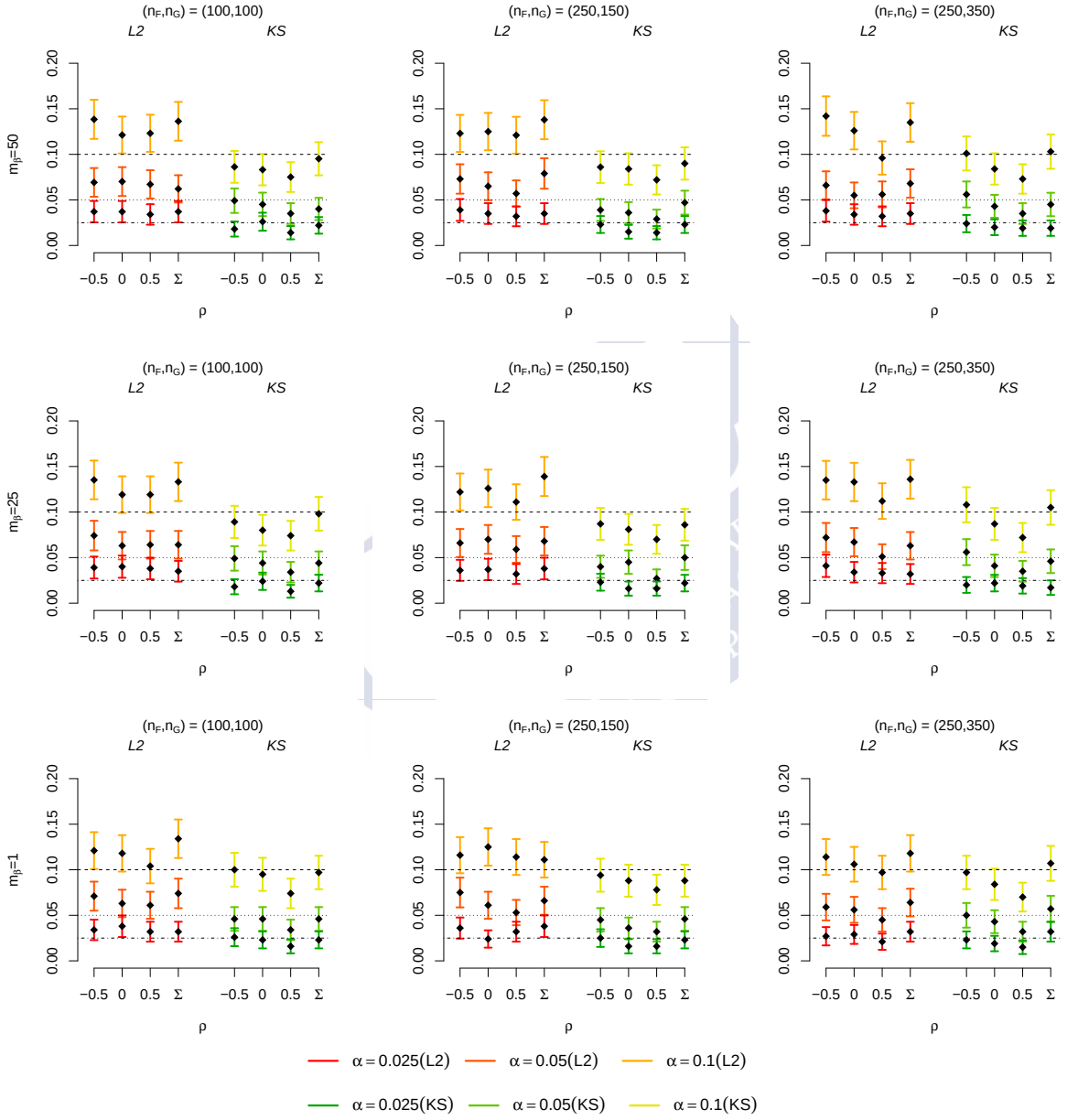


Figure B.22: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis with $K = 3$, $d = 2$ and $m_{\boldsymbol{\beta}} = 50, 25, 1$ for different sample sizes and different ρ.

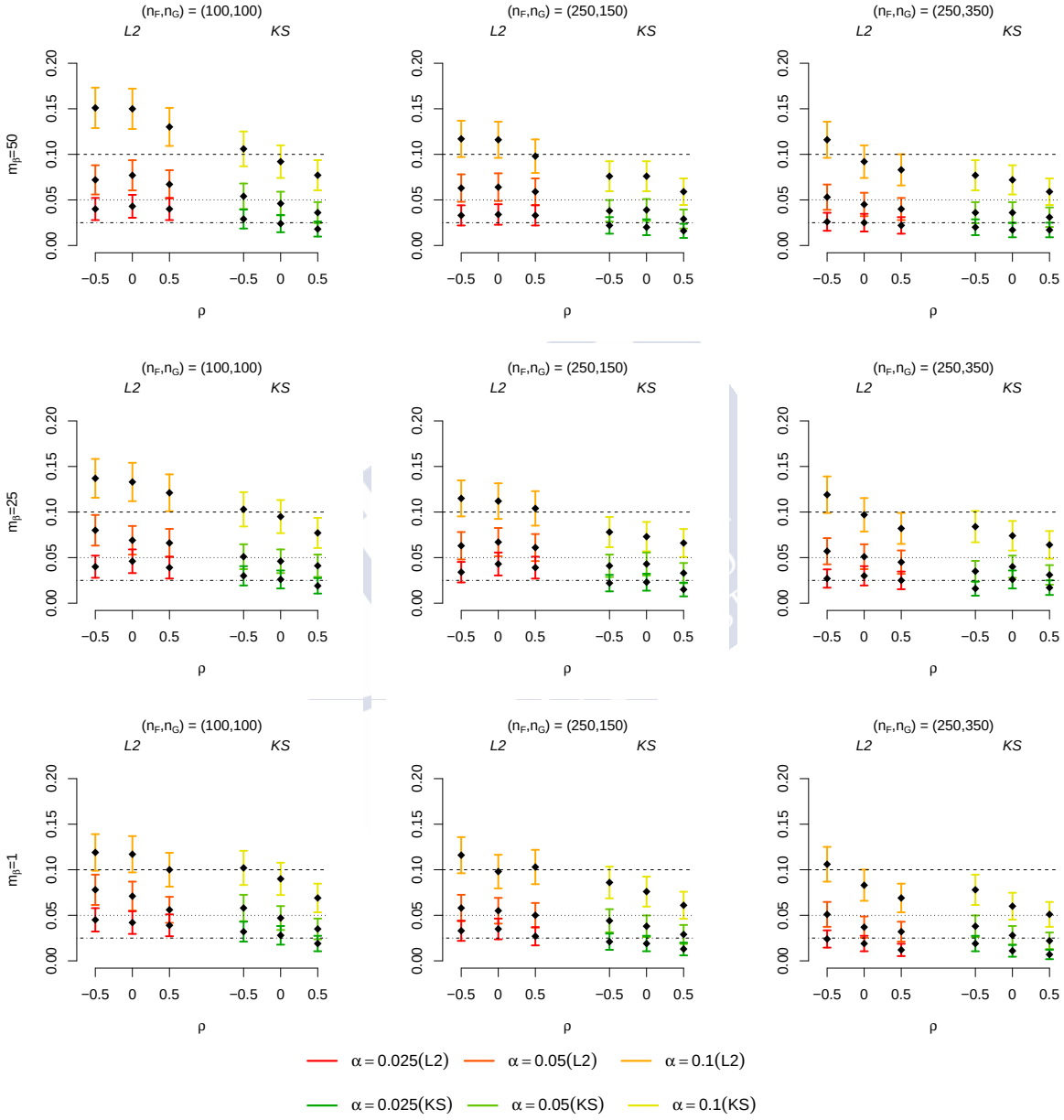


Figure B.23: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis with $K = 2$, $d = 3$ and $m_\beta = 50, 25, 1$ for different sample sizes and different ρ.

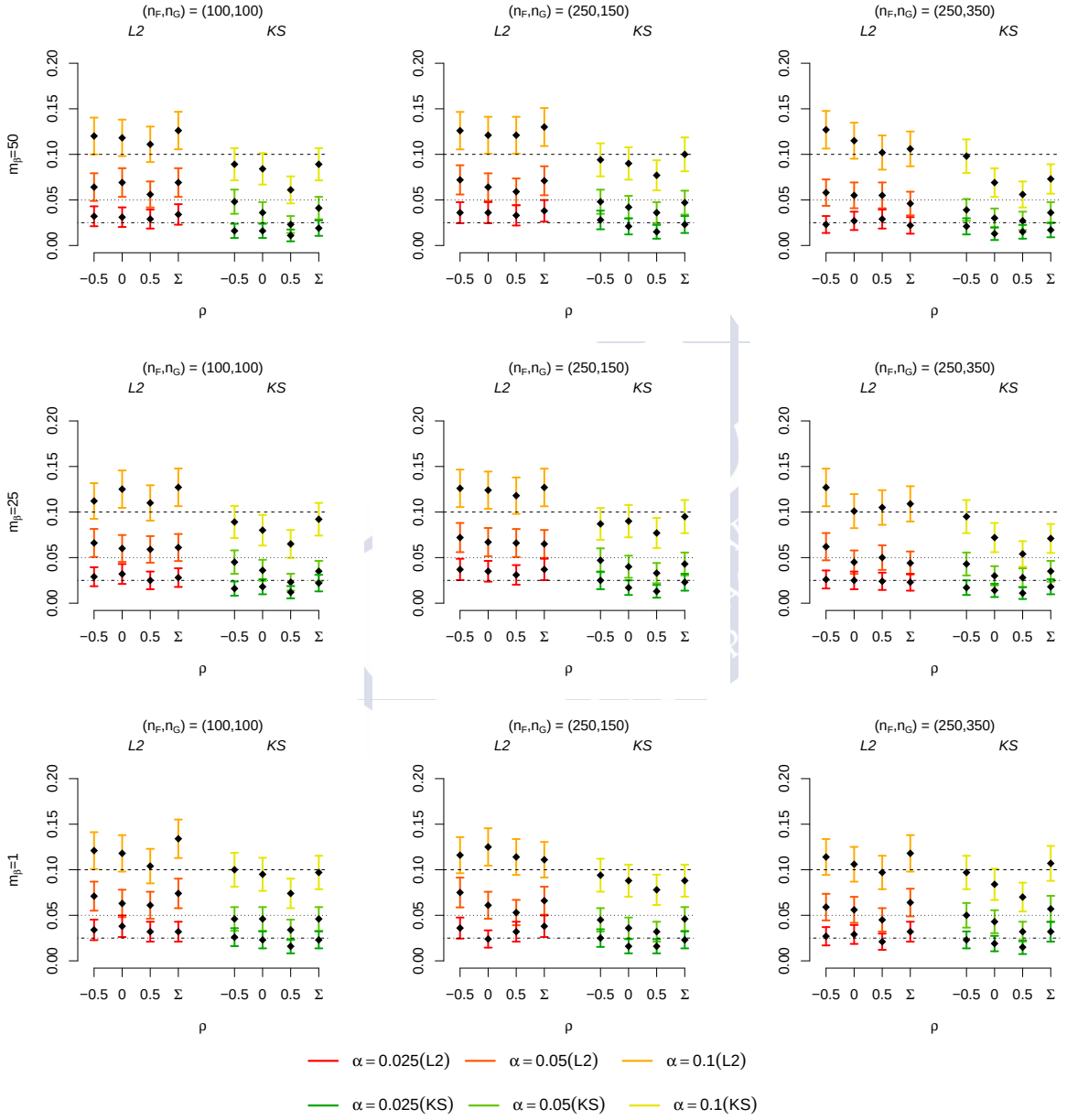


Figure B.24: Estimated proportion of rejection and the corresponding confidence intervals under the null hypothesis with $K = 3$, $d = 3$ and $m_{\boldsymbol{\beta}} = 50, 25, 1$ for different sample sizes and different ρ.

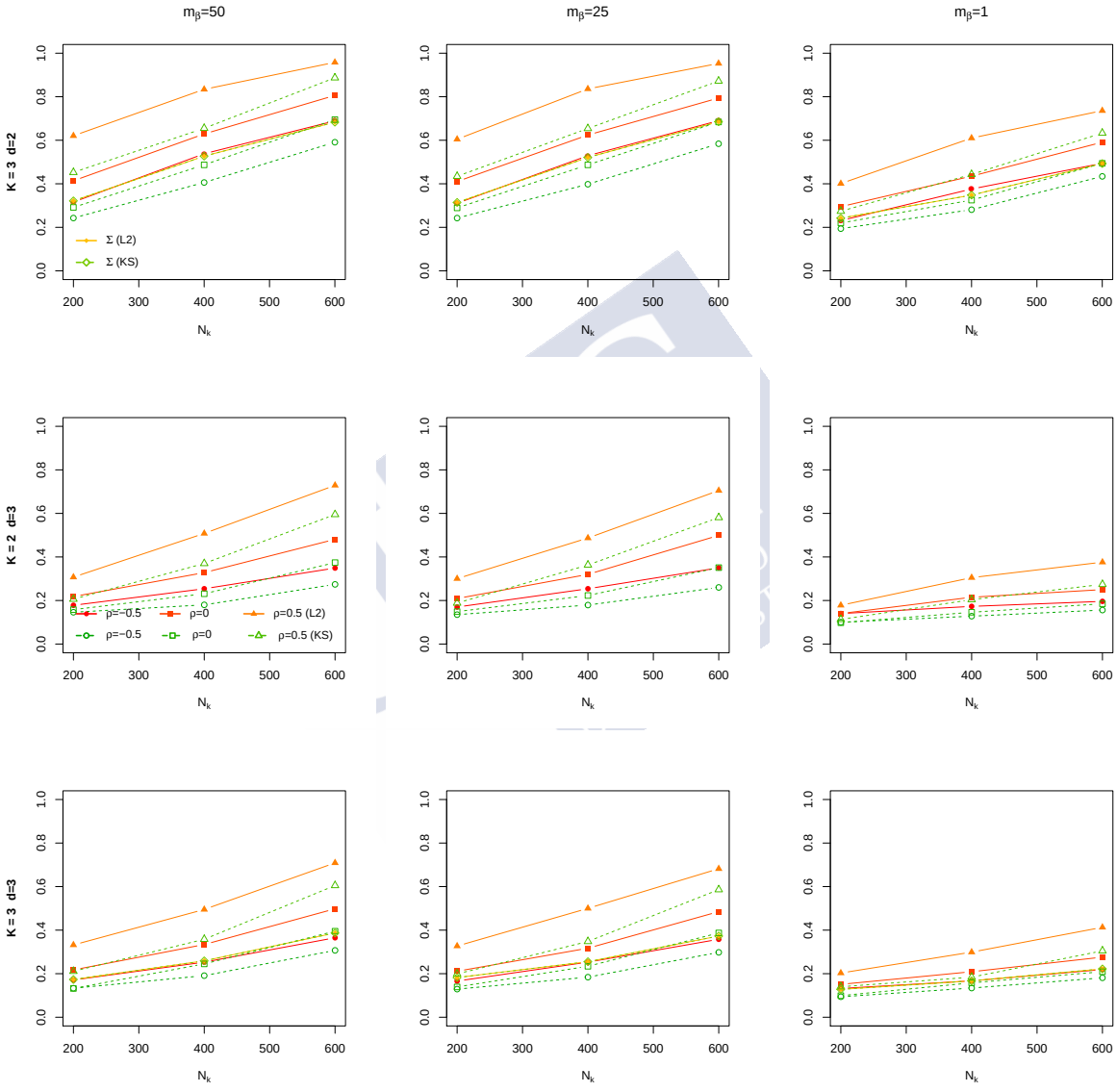


Figure B.25: Estimated proportion of rejection under the alternative hypothesis for different sample sizes and different ρ, for $n_{\boldsymbol{\beta}} = 50, 25, 1$ and for the scenarios with different combinations of K and d ($\alpha = 0.05$).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumen en castellano

La curva ROC (del inglés, Receiver Operating Characteristic curve) es una herramienta estadística utilizada para evaluar la capacidad discriminativa de un sistema de clasificación. Dado un método de clasificación binario, su correspondiente curva ROC indica en qué medida se consiguen diferenciar las dos poblaciones que se pretenden clasificar.

Una de las principales utilidades de este tipo de curvas aparece a la hora de comparar la capacidad discriminativa de dos o más métodos de clasificación. Mediante la comparación de las correspondientes curvas ROC se puede determinar si dichos métodos son igual de eficaces o no.

El principal objetivo de esta tesis es precisamente el estudio y desarrollo de contrates para comparar curvas ROC. Se estudiará la comparación de curvas en tres situaciones: en la primera la comparación se realizará sin tener en cuenta información extra que puedan aportar otras covariables, en la segunda se contemplará la comparación de curvas ROC con covariables unidimensionales y en la tercera se extenderá esa metodología al caso multidimensional. Dado que la inclusión de covariables puede alterar las conclusiones que se obtienen en cada análisis, paralelamente se estudiará una estrategia para determinar si el efecto de dichas covariables es significativo o no en el contexto de curvas ROC. El último objetivo perseguido en esta tesis será la ilustración de las distintas metodologías presentadas con datos reales.

A continuación se presenta un resumen de los capítulos desarrollados a lo largo de esta tesis, indicando las principales aportaciones que se realizan de acuerdo con los objetivos de la misma.

Capítulo 1: Introducción

En el primer capítulo se comienza por presentar la curva ROC y algunas de sus propiedades. El punto de partida de cualquier análisis relacionado con estas curvas es un método de clasificación: se tiene una población dividida en dos categorías y, dado un elemento de esa población, interesa determinar a qué categoría pertenece.

Como las curvas ROC se aplican en un ámbito mayormente biosanitario, consideraremos que esa población consiste en un conjunto de pacientes susceptibles de tener una cierta enfermedad. El criterio de clasificación que se usará estará basado en una variable Y, a la que llamaremos variable diagnóstica o variable marcador: dado un determinado punto de corte c, un individuo será diagnosticado como enfermo para $Y > c$ y como sano para $Y \leqslant c$.

A la hora de evaluar un método de diagnosis hay que tener en cuenta que existen dos tipos de errores que se pueden cometer al clasificar a los individuos. Asociados a estos errores surgen los conceptos de sensibilidad (la probabilidad de clasificar como enfermo a un individuo enfermo) y especificidad (la probabilidad de clasificar correctamente a un individuo sano).

Para que un método de clasificación fuera lo mejor posible lo ideal sería que se eligiera un punto de corte c tal que se maximizara tanto la sensibilidad como la especificidad. Sin embargo,

esto no es posible, ya que al mover el punto c para aumentar una de ellas, la otra disminuirá inevitablemente. En este contexto surge la curva ROC, que no se limita a la selección de un único punto de corte, sino que representa los valores de la sensibilidad en función del complementario de la especificidad para cada posible valor del punto de corte. La definición más usual de la curva ROC se obtiene al expresar esta relación a través de las funciones de distribución de las variables diagnósticas en la población de los sanos y de los enfermos:

$$ROC(p) = 1 - F(G^{-1}(1-p)),\ p \in (0,1),$$

donde F y G son las funciones de distribución de las variables de diagnosis en las poblaciones de los enfermos y de los sanos, respectivamente, y G^{-1} es la función cuantil asociada a la distribución G.

La curva ROC, que es continua siempre y cuando las variables de diagnosis lo sean, toma valores en el cuadrado unidad. Tiene la propiedad de permanecer invariante ante transformaciones monótonas crecientes de las variables diagnósticas. Una curva ROC cercana a la diagonal representa un método de diagnosis donde las densidades de las variables diagnósticas están muy solapadas, mientras que una curva ROC que se acerca al punto $(0,1)$, el de máxima sensibilidad y especificidad, representa un método con una gran capacidad discriminativa.

En la literatura existen diversos resúmenes numéricos de las curvas ROC, de entre los que destaca el AUC (del inglés Area Under the Curve), que representa el área bajo la curva ROC. En este capítulo se describen este y otros indicadores, además de recoger los modelos paramétricos más utilizados para modelizar las curvas ROC.

Finalmente, se describen los dos conjuntos de datos que se utilizan a lo largo del manuscrito para ilustrar la aplicación de las distintas metodologías. El primero de ellos consta de datos de sujetos sospechosos de tener prediabetes. Se dispone de varias variables que pueden ser utilizadas como variables diagnósticas y otras covariables (como la edad) que pueden influir en su capacidad discriminativa. El segundo conjunto de datos contiene información de pacientes con derrame pleural. En este caso interesa diferenciar cuándo ese derrame es por causas cancerígenas de cuándo no, y para ello se dispone de diversas variables marcadores y de otras covariables que pueden influir en el estudio.

Capítulo 2: Curvas ROC en presencia de covariables

Es habitual que, junto con las variables de diagnosis, en un estudio con curvas ROC se disponga de más información en forma de covariables. Estas covariables pueden afectar de distintas formas el comportamiento de las curvas ROC, y por ello es conveniente estudiar si ese efecto es significativo o no.

En este capítulo se comienza exponiendo dos ejemplos para motivar el estudio. En el primero se describe el caso en el que la covariable, pese a dar lugar a las mismas curvas ROC para todos sus valores, genera una curva ROC distinta de la que se obtendría al construir la curva con todos los datos agregados. En el segundo, la covariable afecta tanto a las variables de diagnosis como a su capacidad discriminativa. En ambos casos se ve que no es suficiente con estudiar la curva ROC de los datos agregados.

En este contexto se pueden utilizar tres curvas para incorporar una covariable al estudio de curvas ROC: la curva ROC agregada, que es equivalente a desestimar el efecto de la covariable;

la curva ROC condicionada, que, dado un valor de la covariable x, se define como

$$ROC^x(p) = 1 - F(G^{-1}(1-p|x)|x),\ p \in (0,1);$$

y la curva ROC ajustada o curva AROC, que es una media ponderada de las curvas ROC condicionadas. En este capítulo se da la definición y alguna de las propiedades de cada una de esas tres curvas. En concreto se muestran algunos de los principales estimadores que se pueden encontrar en la literatura para cada una de ellas, destacando la propuesta de González-Manteiga et al. (2011) para la estimación de la curva ROC condicionada basada en la regresión inducida.

A continuación se propone una estrategia, basada en las relaciones que se pueden establecer entre las tres curvas, para determinar en qué medida afecta una covariable al estudio de las curvas ROC. Esta estrategia consta de dos pasos. En el primero se contrasta si la curva ROC condicionada es constante para cada valor de la covariable o no (es decir, si es igual a la curva AROC o no). En caso negativo, se debe utilizar la curva ROC condicionada para posteriores análisis. En caso afirmativo se continúa al segundo paso, en el que se compara la igualdad de las curvas ROC y AROC. Solo en el caso de que se aceptara también esta segunda hipótesis se podría desestimar por completo el efecto de la covariable (usando la curva ROC agregada). En caso contrario, se debe emplear la curva AROC.

En Rodríguez-Álvarez et al. (2011b) proponen un método para realizar el contraste relativo al primer paso de esta estrategia. Para el segundo paso se desarrolla una nueva metodología en este capítulo. Junto con el estadístico de contraste (basado en la suma de las distancias entre las curvas ROC y AROC con su curva promedio) se propone un algoritmo bootstrap para aproximar su distribución. La descripción de este contraste va seguida de un estudio de simulación para analizar su calibrado y su potencia.

A continuación se ilustra esta estrategia de dos pasos aplicándola a la base de datos de prediabetes, estudiando, para tres variables de diagnosis distintas, qué tipo de curva hay que utilizar cuando se dispone de información sobre la covariable edad.

El capítulo finaliza con una reflexión sobre cómo se podría extender esta estrategia al caso en el que se quisiera comparar dos o más curvas y al mismo tiempo tener en cuenta la información de las covariables.

Capítulo 3: Comparación de curvas ROC sin covariables

Este capítulo está dedicado a la comparación de curvas ROC sin covariables. En él se hace una revisión de las distintas metodologías que existen en la literatura para realizar tal comparación, enfocándonos al ámbito no paramétrico y en aquellas que comparan curvas ROC independientes. En concreto, el objetivo es, dados K métodos de diagnosis, contrastar

$$H_0:\ ROC_1(p) = \cdots = ROC_K(p),\quad \text{para todo } p \in (0,1).$$

En las distintas metodologías existentes destacan dos aspectos a tener en cuenta. Por un lado, la construcción del estadístico, ya que algunas de las propuestas realizan la comparación de las curvas a través la comparación de algún resumen numérico como el AUC. Estas técnicas tienen la limitación de no ser capaces de detectar diferencias entre curvas ROC distintas que al cruzarse tienen el mismo AUC. Por otro lado, es importante el método que se utiliza para obtener la distribución del estadístico. En este caso la complicación viene al tratar de replicar

la hipótesis nula en los métodos de remuestreo, pues la igualdad de curvas ROC no implica que las funciones de distribución de sus correspondientes variables diagnósticas sean necesariamente iguales.

En este capítulo se describen las metodologías propuestas en DeLong et al. (1988), Venkatraman (2000), Antoch et al. (2010), Martínez-Camblor et al. (2011) y Martínez-Camblor et al. (2013) relativas a este problema. A continuación se realiza un estudio de simulación para comparar su comportamiento. Los escenarios que se utilizan para realizar el estudio de simulación están pensados para sacar a relucir las ventajas o inconvenientes que pueda tener cada metodología tanto a nivel de construcción del estadístico como a nivel de aproximación de su distribución.

En concreto, se observa que las metodologías que utilizan métodos de remuestreo basándose en la igualdad de funciones de distribución y no en la igualdad de curvas no aproximan bien el nivel en algunos escenarios. Por otro lado, se aprecia la falta de potencia en las metodologías basadas en la comparación de AUCs en los casos donde las curvas se cruzan.

Capítulo 4: Comparación de curvas ROC con covariables unidimensionales

Una vez se han visto las distintas técnicas existentes en la literatura para comprar curvas ROC, en este capítulo se diseña y se estudia una nueva metodología para realizar esa comparación, esta vez condicionada al valor de una covariable unidimensional. Es decir, dado un valor x de la covariable, se pretende realizar el contraste

$$H_0 : ROC_1^x(p) = \cdots = ROC_K^x(p), \quad \text{para todo } p \in (0,1).$$

Para ello se parte de una de las metodologías estudiadas en el capítulo anterior para el caso sin covariables, la de Martínez-Camblor et al. (2011), que está basada en la comparación de toda la curva ROC, es capaz de comparar más de dos curvas, se puede adaptar al caso en el que las curvas ROC sean dependientes y que utiliza un algoritmo bootstrap que conserva la estructura de los datos. Dicha metodología se combina con la propuesta de González-Manteiga et al. (2011) para la estimación de las curvas ROC condicionadas.

El estadístico de contraste se construye sumando las distancias entre cada una de las curvas condicionadas que se pretenden comparar con respecto a una curva promedio. Se obtiene la distribución asintótica del estadístico bajo la hipótesis nula pero, dado que dicha distribución depende de funciones que en la práctica se desconocen, también se propone el uso de un algoritmo bootstrap.

Dicho algoritmo bootstrap, al igual que el resto de algoritmos bootstrap desarrollados en esta tesis, está basado en el algoritmo bootstrap generalizado propuesto en Martínez-Camblor y Corral (2012). La principal diferencia con respecto al bootstrap usual es que, a la hora de generar las muestras bootstrap, no se hace bajo la hipótesis nula de igualdad de curvas ROC. En su lugar, la hipótesis nula se aplica al calcular el estadístico bootstrap, pues en vez de utilizar el estadístico de contraste propuesto originalmente se pone en su lugar una expresión que es igual a este cuando la hipótesis nula es cierta. De esta forma se preserva la estructura de los datos que se tienen que generar, además de que nos ahorra el problema de cómo remuestrear bajo la hipótesis nula de igualdad de curvas ROC.

Para comprobar el buen funcionamiento de la metodología propuesta se presentan los resultados de un estudio de simulación. Se consideran escenarios en los que se comparan dos o tres curvas ROC condicionadas a distintos valores de la covariable y para distintos tamaños muestrales. En general se obtiene una buena aproximación para el nivel del test y una potencia que aumenta con el tamaño muestral y al aumentar la diferencia entre las curvas ROC que se están comparando.

La aplicación de esta metodología se ilustra utilizando los datos de pacientes con derrame pleural. En particular se compara el comportamiento de una variable diagnosis basada en un antígeno cuando el sujeto en estudio es un hombre o una mujer a la vez que se tiene en cuenta la covariable edad. Los resultados muestran que se puede llegar a distintas conclusiones al incluir o no la edad en el estudio.

Cabe destacar que, inicialmente, el test desarrollado en este capítulo estaba enfocado al contraste de curvas ROC independientes. Sin embargo, es posible realizar una adaptación para el caso de curvas ROC dependientes simplemente alterando la manera en la que se generan las muestras bootstrap en el algoritmo de remuestreo. Esta adaptación se incluye en este capítulo, así como un estudio de simulación en el que se consideran variables de diagnosis con distintos grados de correlación entre ellas. En dicho estudio se puede observar que el no tener en cuenta la dependencia existente entre las curvas ROC puede llevar a un mal calibrado del test para correlaciones muy extremas, algo que se consigue corregir al aplicar la metodología adaptada.

Capítulo 5: Comparación de curvas ROC con covariables multidimensionales

En este capítulo el objetivo es extender la metodología vista previamente para el caso de una covariable unidimensional al caso de una covariable multidimensional. Es decir, dado cierto valor de una covariable multidimensional $\boldsymbol{x}$, el objetivo es realizar el contraste

$$H_0 : ROC_1^{\boldsymbol{x}}(p) = \cdots = ROC_K^{\boldsymbol{x}}(p), \quad \text{para todo } p \in (0,1).$$

En este caso, se asumirá que las curvas ROC comparadas son dependientes en el sentido de que se ven condicionadas por la misma covariable multidimensional.

Teniendo en cuenta cómo se construyó el estadístico de contraste en el capítulo anterior, aquí se podría considerar la misma idea utilizando esta vez una metodología para estimar la curva ROC condicionada que pudiera incorporar una covariable multidimensional. Sin embargo, en vez de eso se optó por utilizar proyecciones para reducir la dimensión del problema y convertirlo en un problema más fácil de manejar, técnica ya utilizada en Escanciano (2006) en el contexto de la regresión.

Esta idea da lugar a que, en vez de realizar el contraste en el que se condiciona la curva ROC a covariables multidimensionales, en su lugar se realiza el siguiente contraste equivalente

$$H_0 : ROC_1^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) = \cdots = ROC_K^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p), \text{ para todo } p \in (0,1) \text{ y cualquier } \boldsymbol{\beta}^F, \boldsymbol{\beta}^G,$$

donde $\boldsymbol{\beta}^F$ y $\boldsymbol{\beta}^G$ son coordenadas d-dimensionales que representan las direcciones de las proyecciones en una esfera unidad d-dimensional.

La ventaja de realizar este contraste alternativo es que las variables con las que en ese caso se está condicionando a las curvas ROC son unidimensionales, que es precisamente el problema que se ha tratado en el capítulo anterior. Sin embargo, hay que hacer dos ajustes con respecto

a la metodología empleada anteriormente. En primer lugar, como se tienen que emplear dos proyecciones distintas para cada valor de la covariable (que viene del hecho de que cada curva ROC se construye a partir de dos funciones de distribución), la curva ROC condicionada que se está comparando está ahora condicionada a dos valores distintos, $(\boldsymbol{\beta}^F)'\boldsymbol{x}$ y $(\boldsymbol{\beta}^G)'\boldsymbol{x}$. Esto requiere la adaptación del estadístico que se tenía para un solo valor de la covariable mediante la modifición del método de estimación de cada curva ROC. En segundo lugar, este contraste alternativo requiere comprobar la igualdad de estas curvas doblemente condicionadas para cualquier par de direcciones $\boldsymbol{\beta}^F$ y $\boldsymbol{\beta}^G$. Para tener en cuenta esto, el estadístico de contraste que se propone es

$$D_S^{\boldsymbol{x}} = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} d\boldsymbol{\beta}^F d\boldsymbol{\beta}^G,$$

donde $S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$ es el estadístico de contraste que se utilizaría para realizar el test en el caso unidimensional para un par de direcciones $\boldsymbol{\beta}^F$ y $\boldsymbol{\beta}^G$ fijas.

En la práctica esa cantidad es difícil de calcular, así que en su lugar se utiliza una aproximación numérica de esa integral doble por medio de las combinaciones de $n_{\boldsymbol{\beta}}$ direcciones $\boldsymbol{\beta}^F$ y $n_{\boldsymbol{\beta}}$ direcciones $\boldsymbol{\beta}^G$ extraídas aleatoriamente de la esfera unidad d-dimensional.

Para aproximar la distribución de este estadístico se recurre una vez más a un algoritmo bootstrap que sigue la misma filosofía que el utilizado en el capítulo anterior y emplea la hipótesis nula en el cómputo del estadístico bootstrap en vez de en la generación de las muestras bootstrap.

La explicación de la construcción de este estadístico y del correspondiente algoritmo bootstrap va seguida de la presentación de los resultados de un estudio de simulación en el que se consideran escenarios que comparan dos o tres curvas ROC condicionadas a covariables de dimensión dos o tres. En general se obtienen buenos resultados tanto para el nivel del test como para la potencia.

Finalmente, se analizan de nuevo los datos reales de los individuos con derrame pleural para comparar el comportamiento de dos variables de diagnosis. Se realiza la comparación en las tres situaciones que se han visto a lo largo de la tesis: en primer lugar sin tener en cuenta la información de ninguna covariable, en segundo lugar teniendo en cuenta covariables unidimensionales, y en tercer lugar teniendo en cuenta covariables multidimensionales. Las conclusiones que se pueden extraer en cada una de dichas situaciones es distinta, con lo que queda de manifiesto la importancia de tener una herramienta capaz de tratar con covariables multidimensionales para realizar este tipo de comparaciones.

Capítulo 6: Conclusiones

En este último capítulo se comentan los resultados obtenidos a lo largo de la tesis, destacando aquellos aspectos de las diferentes metodologías que quedan aún por estudiar. También se proponen nuevas vías de investigación, como pueden ser la adaptación de la metodología propuesta en el Capítulo 2 para el caso de tener una covariable multidimensional o la extensión de las técnicas de comparación de curvas para el caso de que las covariables fueran funcionales o longitudinales.

Apéndices A y B

En el Apéndice A vienen recogidas las demostraciones relativas a los resultados teóricos de los Capítulos 4 y 5. Por otra parte, el Apéndice B contiene resultados adicionales de los estudios de simulación llevados a cabo en los Capítulos 4 y 5. Estas simulaciones incluyen estudios relaciona-

dos con la selección de los diferentes parámetros de suavizado involucrados en las estimaciones no paramétricas, resultados para distintas configuraciones del número de muestras bootstrap o del número $n_{\boldsymbol{\beta}}$, resultados en los que se consideran distintos tamaños muestrales y resultados para adaptaciones de las metodologías descritas, como para el caso de datos apareados.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumo en galego

A curva ROC (do inglés, Receiver Operating Characteristic curve) é unha ferramenta estatística utilizada para avaliar a capacidade discriminativa dun sistema de clasificación. Dado un método de clasificación binario, a súa correspondente curva ROC indica en que medida se conseguen diferenciar as dúas poboacións que se pretenden clasificar.

Unha das principais utilidades deste tipo de curvas aparece á hora de comparar a capacidade discriminativa de dous ou máis métodos de clasificación. Mediante a comparación das correspondentes curvas ROC pódese determinar se devanditos métodos son igual de eficaces ou non.

O principal obxectivo desta tese é precisamente o estudo e desenvolvemento de contrates para comparar curvas ROC. Estudarase a comparación de curvas en tres situacións: na primeira a comparación realizarase sen ter en conta información extra que poidan achegar outras covariables, na segunda contemplarase a comparación de curvas ROC con covariables unidimensionais e na terceira estenderase esa metodoloxía ao caso multidimensional. Dado que a inclusión de covariables pode alterar as conclusións que se obteñen en cada análise, paralelamente estudarase unha estratexia para determinar se o efecto de ditas covariables é significativo ou non no contexto de curvas ROC. O último obxectivo perseguido nesta tese será a ilustración das distintas metodoloxías presentadas con datos reais.

A continuación preséntase un resumo dos capítulos desenvolvidos ao longo desta tese, indicando as principais achegas que se realizan de acordo cos obxectivos da mesma.

Capítulo 1: Introdución

No primeiro capítulo comézase por presentar a curva ROC e algunhas das súas propiedades. O punto de partida de calquera análise relacionada con estas curvas é un método de clasificación: tense unha poboación dividida en dúas categorías e, dado un elemento desa poboación, interesa determinar a que categoría pertence.

Como as curvas ROC se aplican nun ámbito maiormente biosanitario, consideraremes que esa poboación consiste nun conxunto de pacientes susceptibles de ter unha certa enfermidade. O criterio de clasificación que se usará estará baseado nunha variable Y, á que chamaremos variable diagnóstica ou variable marcador: dado un determinado punto de corte c, un individuo será diagnosticado como enfermo para $Y > c$ e como san para $Y \leqslant c$.

Á hora de avaliar un método de diagnose hai que ter en conta que existen dous tipos de erros que se poden cometer ao clasificar aos individuos. Asociados a estes erros xorden os conceptos de sensibilidade (a probabilidade de clasificar como enfermo a un individuo enfermo) e especificidade (a probabilidade de clasificar correctamente a un individuo san).

Para que un método de clasificación fose o mellor posible o ideal sería que se elixise un punto de corte c tal que se maximizase tanto a sensibilidade como a especificidade. Con todo, isto non é posible, xa que ao mover o punto c para aumentar unha delas, a outra diminuirá inevitablemente. Neste contexto xorde a curva ROC, que non se limita á selección dun único punto de corte, senón que representa os valores da sensibilidade en función do complementario da especificidade para cada posible valor do punto de corte. A definición máis usual da curva ROC obtense ao expresar esta relación a través das funcións de distribución das variables diagnósticas na poboación dos sans e dos enfermos:

$$ROC(p) = 1 - F(G^{-1}(1-p)),\ p \in (0,1),$$

onde F e G son as funcións de distribución das variables de diagnose nas poboacións dos enfermos e dos sans, respectivamente, e G^{-1} é a función cuantil asociada á distribución G.

A curva ROC, que é continua a condición de que as variables de diagnose o sexan, toma valores na cadrado unidade. Ten a propiedade de permanecer invariante ante transformacións monótonas crecentes das variables diagnósticas. Unha curva ROC próxima á diagonal representa un método de diagnose onde as densidades das variables diagnósticas están moi solapadas, mentres que unha curva ROC que se achega ao punto $(0,1)$, o de máxima sensibilidade e especificidade, representa un método cunha gran capacidade discriminativa.

Na literatura existen diversos resumos numéricos das curvas ROC, entre os que destaca a AUC (do inglés Area Under the Curve), que representa a área baixo a curva ROC. Neste capítulo descríbense este e outros indicadores, ademais de recoller os modelos paramétricos máis utilizados para modelizar as curvas ROC.

Finalmente, descríbense os dous conxuntos de datos que se utilizan ao longo do manuscrito para ilustrar a aplicación das distintas metodoloxías. O primeiro deles consta de datos de suxeitos sospeitosos de ter prediabetes. Disponse de varias variables que poden ser utilizadas como variables diagnósticas e outras covariables (como a idade) que poden influír na súa capacidade discriminativa. O segundo conxunto de datos contén información de pacientes con derrame pleural. Neste caso interesa diferenciar cando ese derrame é por causas cancerígenas de cando non, e para iso disponse de diversas variables marcadores e doutras covariables que poden influír no estudo.

Capítulo 2: Curvas ROC en presenza de covariables

É habitual que, xunto coas variables de diagnose, nun estudo con curvas ROC se dispoña de máis información en forma de covariables. Estas covariables poden afectar de distintas formas o comportamento das curvas ROC, e por iso é conveniente estudar se ese efecto é significativo ou non.

Neste capítulo comézase expoñendo dous exemplos para motivar o estudo. No primeiro descríbese o caso no que a covariable, a pesar de dar lugar ás mesmas curvas ROC para todos os seus valores, xera unha curva ROC distinta da que se obtería ao construír a curva con todos os datos agregados. No segundo, a covariable afecta tanto ás variables de diagnose como á súa capacidade discriminativa. En ambos os casos vese que non é suficiente con estudar a curva ROC dos datos agregados.

Neste contexto pódense utilizar tres curvas para incorporar unha covariable ao estudo de

curvas ROC: a curva ROC agregada, que é equivalente a desestimar o efecto da covariable; a curva ROC condicionada, que, dado un valor da covariable x, se define como

$$ROC^x(p) = 1 - F(G^{-1}(1-p|x)|x),\ p \in (0,1);$$

e a a curva ROC axustada ou curva AROC, que é unha media ponderada das curvas ROC condicionadas. Neste capítulo dáse a definición e algunha das propiedades de cada unha desas tres curvas. En concreto móstranse algúns dos principais estimadores que se poden atopar na literatura para cada unha delas, destacando a proposta de González-Manteiga et al. (2011) para a estimación da curva ROC condicionada baseada na regresión inducida.

A continuación proponse unha estratexia, baseada nas relacións que se poden establecer entre as tres curvas, para determinar en que medida afecta unha covariable ao estudo das curvas ROC. Esta estratexia consta de dous pasos. No primeiro contrástase se a curva ROC condicionada é constante para cada valor da covariable ou non (é dicir, se é igual á curva AROC ou non). En caso negativo, débese utilizar a curva ROC condicionada para posteriores análises. En caso afirmativo continúase ao segundo paso, no que se compara a igualdade das curvas ROC e AROC. Só no caso de que se aceptase tamén esta segunda hipótese poderíase desestimar por completo o efecto da covariable (usando a curva ROC agregada). En caso contrario, débese empregar a curva AROC.

En Rodríguez-Álvarez et al. (2011b) propoñen un método para realizar o contraste relativo ao primeiro paso desta estratexia. Para o segundo paso desenvólvese unha nova metodoloxía neste capítulo. Xunto co estatístico de contraste (baseado na suma das distancias entre as curvas ROC e AROC coa súa curva media) proponse un algoritmo bootstrap para aproximar a súa distribución. A descrición deste contraste vai seguida dun estudo de simulación para analizar o seu calibrado e a súa potencia.

A continuación ilústrase esta estratexia de dous pasos aplicándoa á base de datos de prediabetes, estudando, para tres variables de diagnose distintas, que tipo de curva hai que utilizar cando se dispón de información sobre a covariable idade.

O capítulo finaliza cunha reflexión sobre como se podería estender esta estratexia ao caso no que se quixese comparar dúas ou máis curvas e ao mesmo tempo ter en conta a información das covariables.

Capítulo 3: Comparación de curvas ROC sen covariables

Este capítulo está dedicado á comparación de curvas ROC sen covariables. Nel faise unha revisión das distintas metodoloxías que existen na literatura para realizar tal comparación, enfocándonos ao ámbito non paramétrico e naquelas que comparan curvas ROC independentes. En concreto, o obxectivo é, dados K métodos de diagnose, contrastar

$$H_0:\ ROC_1(p) = \cdots = ROC_K(p),\quad \text{para todo } p \in (0,1).$$

Nas distintas metodoloxías existentes destacan dous aspectos a ter en conta. Por unha banda, a construción do estatístico, xa que algunhas das propostas realizan a comparación das curvas a través a comparación dalgún resumo numérico como a AUC. Estas técnicas teñen a limitación de non ser capaces de detectar diferenzas entre curvas ROC distintas que ao cruzarse teñen a mesma AUC. Doutra banda, é importante o método que se utiliza para obter a distribución do

estatístico. Neste caso a complicación vén ao tratar de replicar a hipótese nula nos métodos de remostraxe, pois a igualdade de curvas ROC non implica que as funcións de distribución das súas correspondentes variables diagnósticas sexan necesariamente iguais.

Neste capítulo descríbense as metodoloxías propostas en DeLong et al. (1988), Venkatraman (2000), Antoch et al. (2010), Martínez-Camblor et al. (2011) e Martínez-Camblor et al. (2013) relativas a este problema. A continuación realízase un estudo de simulación para comparar o seu comportamento. Os escenarios que se utilizan para realizar o estudo de simulación están pensados para sacar a relucir as vantaxes ou inconvenientes que poida ter cada metodoloxía tanto a nivel de construción do estatístico como a nivel de aproximación da súa distribución.

En concreto, obsérvase que as metodoloxías que utilizan métodos de remostraxe baseándose na igualdade de funcións de distribución e non na igualdade de curvas non aproximan ben o nivel nalgúns escenarios. Doutra banda, apréciase a falta de potencia nas metodoloxías baseadas na comparación de AUCs nos casos onde as curvas se cruzan.

Capítulo 4: Comparación de curvas ROC con covariables unidimensionais

Unha vez que se viron as distintas técnicas existentes na literatura para comprar curvas ROC, neste capítulo deséñase e estúdase unha nova metodoloxía para realizar esa comparación, esta vez condicionada ao valor dunha covariable unidimensional. É dicir, dado un valor x da covariable, preténdese realizar o contraste

$$H_0 : ROC_1^x(p) = \cdots = ROC_K^x(p) \quad \text{para todo } p \in (0,1).$$

Para iso pártese dunha das metodoloxías estudadas no capítulo anterior para o caso sen covariables, a de Martínez-Camblor et al. (2011), que está baseada na comparación de toda a curva ROC, é capaz de comparar máis de dúas curvas, pódese adaptar ao caso no que as curvas ROC sexan dependentes e que utiliza un algoritmo bootstrap que conserva a estrutura dos datos. Dita metodoloxía combínase coa proposta de González-Manteiga et al. (2011) para a estimación das curvas ROC condicionadas.

O estatístico de contraste constrúese sumando as distancias entre cada unha das curvas condicionadas que se pretenden comparar con respecto a unha curva media. Obtense a distribución asintótica do estatístico baixo a hipótese nula pero, dado que dita distribución depende de funcións que na práctica se descoñecen, tamén se propón o uso dun algoritmo bootstrap.

Este algoritmo bootstrap, do mesmo xeito que o resto de algoritmos bootstrap desenvolvidos nesta tese, está baseado no algoritmo bootstrap xeralizado proposto en Martínez-Camblor e Corral (2012). A principal diferenza con respecto ao bootstrap usual é que, á hora de xerar as mostras bootstrap, non se fai baixo a hipótese nula de igualdade de curvas ROC. No seu lugar, a hipótese nula aplícase ao calcular o estatístico bootstrap, pois no canto de utilizar o estatístico de contraste proposto orixinalmente ponse no seu lugar unha expresión que é igual e este cando a hipótese nula é certa. Desta forma presérvase a estrutura dos datos que se teñen que xerar, ademais de que nos aforra o problema de como remostrear baixo a hipótese nula de igualdade de curvas ROC.

Para comprobar o bo funcionamento da metodoloxía proposta preséntanse os resultados dun estudo de simulación. Considéranse escenarios nos que se comparan dúas ou tres curvas ROC condicionadas a distintos valores da covariable e para distintos tamaños mostrais. En xeral

obtense unha boa aproximación para o nivel do test e unha potencia que aumenta co tamaño mostral e ao aumentar a diferenza entre as curvas ROC que se están comparando.

A aplicación desta metodoloxía ilústrase utilizando os datos de pacientes con derrame pleural. En particular compárase o comportamento dunha variable diagnose baseada nun antíxeno cando o suxeito en estudo é un home ou unha muller á vez que se ten en conta a covariable idade. Os resultados mostran que se pode chegar a distintas conclusións ao incluír ou non a idade no estudo.

Cabe destacar que, inicialmente, o test desenvolvido neste capítulo estaba enfocado ao contraste de curvas ROC independentes. Con todo, é posible realizar unha adaptación para o caso de curvas ROC dependentes simplemente alterando a maneira na que se xeran as mostras bootstrap no algoritmo de remostraxe. Esta adaptación inclúese neste capítulo, así como un estudo de simulación no que se consideran variables de diagnose con distintos graos de correlación entre elas. Neste estudo pódese observar que o feito de non ter en conta a dependencia existente entre as curvas ROC pode levar a un mal calibrado do test para correlacións moi extremas, algo que se consegue corrixir ao aplicar a metodoloxía adaptada.

Capítulo 5: Comparación de curvas ROC con covariables multidimensionais

Neste capítulo o obxectivo é estender a metodoloxía vista previamente para o caso dunha covariable unidimensional ao caso dunha covariable multidimensional. É dicir, dado certo valor dunha covariable multidimensional $\boldsymbol{x}$, o obxectivo é realizar o contraste

$$H_0 : ROC_1^{\boldsymbol{x}}(p) = \cdots = ROC_K^{\boldsymbol{x}}(p) \text{ para todo } p \in (0,1).$$

Neste caso, asumirase que as curvas ROC comparadas son dependentes no sentido de que están condicionadas pola mesma covariable multidimensional.

Tendo en conta como se construíu o estatístico de contraste no capítulo anterior, aquí poderíase considerar a mesma idea utilizando esta vez unha metodoloxía para estimar a curva ROC condicionada que puidese incorporar unha covariable multidimensional. Con todo, no canto diso optouse por utilizar proxeccións para reducir a dimensión do problema e convertelo nun problema máis fácil de manexar, técnica xa utilizada en Escanciano (2006) no contexto da regresión.

Esta idea dá lugar a que, en vez de realizar o contraste no que se condiciona a curva ROC a covariables multidimensionais, no seu lugar realízase o seguinte contraste equivalente

$$H_0 : ROC_1^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p) = \cdots = ROC_K^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}(p), \text{ para todo } p \in (0,1) \text{ e calquera } \boldsymbol{\beta}^F, \boldsymbol{\beta}^G,$$

onde $\boldsymbol{\beta}^F$ e $\boldsymbol{\beta}^G$ son coordenadas d-dimensionaies que representan as direccións das proxeccións nunha esfera unidade d-dimensional.

A vantaxe de realizar este contraste alternativo é que as variables coas que nese caso se están a condicionar ás curvas ROC son unidimensionais, que é precisamente o problema que se tratou no capítulo anterior. Aínda así, hai que facer dous axustes con respecto á metodoloxía empregada anteriormente. En primeiro lugar, como se teñen que empregar dúas proxeccións distintas para cada valor da covariable (que vén do feito de que cada curva ROC se constrúe a partir de dúas funcións de distribución), a curva ROC condicionada que se está comparando está agora condicionada a dous valores distintos, $(\boldsymbol{\beta}^F)'\boldsymbol{x}$ e $(\boldsymbol{\beta}^G)'\boldsymbol{x}$. Isto require a adaptación

do estatístico que se tiña para un só valor da covariable mediante a modificación do método de estimación de cada curva ROC. En segundo lugar, este contraste alternativo require comprobar a igualdade destas curvas dobremente condicionadas para calquera par de direccións $\boldsymbol{\beta}^F$ e $\boldsymbol{\beta}^G$. Para ter en conta isto, o estatístico de contraste que se propón é

$$D_S^{\boldsymbol{x}} = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}} d\boldsymbol{\beta}^F d\boldsymbol{\beta}^G,$$

onde $S^{(\boldsymbol{\beta}^F)'\boldsymbol{x},(\boldsymbol{\beta}^G)'\boldsymbol{x}}$ é o estatístico de contraste que se utilizaría para realizar o test no caso unidimensional para un par de direccións $\boldsymbol{\beta}^F$ e $\boldsymbol{\beta}^G$ fixas.

Na práctica esa cantidade é difícil de calcular, así que no seu lugar utilízase unha aproximación numérica desa integral dobre por medio das combinacións de $n_{\boldsymbol{\beta}}$ direccións $\boldsymbol{\beta}^F$ e $n_{\boldsymbol{\beta}}$ direccións $\boldsymbol{\beta}^G$ extraídas aleatoriamente da esfera unidade d-dimensional.

Para aproximar a distribución deste estatístico recórrese unha vez máis a un algoritmo bootstrap que segue a mesma filosofía que o utilizado no capítulo anterior e emprega a hipótese nula no cómputo do estatístico bootstrap en vez de na xeración das mostras bootstrap.

A explicación da construción deste estatístico e do correspondente algoritmo bootstrap vai seguida da presentación dos resultados dun estudo de simulación no que se consideran escenarios que comparan dúas ou tres curvas ROC condicionadas a covariables de dimensión dous ou tres. En xeral obtéñense bos resultados tanto para o nivel do test como para a potencia.

Finalmente, analízanse de novo os datos reais dos individuos con derrame pleural para comparar o comportamento de dúas variables de diagnose. Realízase a comparación nas tres situacións que se viron ao longo da tese: en primeiro lugar sen ter en conta a información de ningunha covariable, en segundo lugar tendo en conta covariables unidimensionais, e en terceiro lugar tendo en conta covariables multidimensionais. As conclusións que se poden extraer en cada unha das devanditas situacións é distinta, co que queda de manifesto a importancia de ter unha ferramenta capaz de tratar con covariables multidimensionais para realizar este tipo de comparacións.

Capítulo 6: Conclusións

Neste último capítulo coméntanse os resultados obtidos ao longo da tese, destacando aqueles aspectos das diferentes metodoloxías que quedan aínda por estudar. Tamén se propoñen novas vías de investigación, como poden ser a adaptación da metodoloxía proposta no Capítulo 2 para o caso de ter unha covariable multidimensional ou a estensión das técnicas de comparación de curvas para o caso de que as covariables fosen funcionais ou lonxitudinais.

Apéndices A e B

No Apéndice A veñen recollidas as demostracións relativas aos resultados teóricos dos Capítulos 4 e 5. Por outra banda, o Apéndice B contén resultados adicionais dos estudos de simulación levados a cabo nos Capítulos 4 e 5. Estas simulacións inclúen estudos relacionados coa selección dos diferentes parámetros de suavizado involucrados nas estimacións non paramétricas, resultados para distintas configuracións do número de mostras bootstrap ou do número $n_{\boldsymbol{\beta}}$, resultados nos que se consideran distintos tamaños mostrais e resultados para adaptacións das metodoloxías descritas, como para o caso de datos emparellados.

References

Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3(3):421–432.

Antoch, J., Prchal, L., and Sarda, P. (2010). Nonparametric comparison of ROC curves: testing equivalence. In Antoch, J., Hušková, M., and Sen, P., editors, *Nonparametrics and robustness in modern statistical inference and time series analysis: a Festschrift in honor of Professor Jana Jurečková*, volume 7 of *Collections*, pages 12–24. Institute of Mathematical Statistics, Beachwood.

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

Bandos, A. I., Rockette, H. E., and Gur, D. (2005). A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statistics in Medicine*, 24(18):2873–2893.

Braga, A. C., Costa, L., and Oliveira, P. (2013). An alternative method for global and partial comparison of two diagnostic systems based on ROC curves. *Journal of Statistical Computation and Simulation*, 83(2):307–325.

Braun, T. M. and Alonzo, T. A. (2008). A modified sign test for comparing paired ROC curves. *Biostatistics*, 9(2):364–372.

Brumback, L. C., Pepe, M. S., and Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25(4):575–590.

Colling, B. and Van Keilegom, I. (2017). Goodness-of-fit tests in semiparametric transformation models using the integrated regression function. *Journal of Multivariate Analysis*, 160:10–30.

Cuesta-Albertos, J. A., del Barrio, E., Fraiman, R., and Matrán, C. (2007). The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, 51(10):4814–4831.

Cuesta-Albertos, J. A., García-Portugués, E., Febrero-Bande, M., and González-Manteiga, W. (2019). Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *Annals of Statistics*, 47(1):439–467.

de Carvalho, M., Barney, B. J., and Page, G. L. (2020). Affinity-based measures of biomarker performance evaluation. *Statistical Methods in Medical Research*, 29(3):837–853.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845.

Einmahl, J. and Van Keilegom, I. (2008). Specification tests in nonparametric regression. *Journal of Econometrics*, 143(1):88–102.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051.

Estévez-Pérez, G. and Vieu, P. (2020). A new way for ranking functional data with applications in diagnostic test. *Computational Statistics*.

Fanjul-Hevia, A. and González-Manteiga, W. (2018). A comparative study of methods for testing the equality of two or more ROC curves. *Computational Statistics*, 33:357–377.

Fanjul-Hevia, A., González-Manteiga, W., and Pardo-Fernández, J. C. (2020a). A non-parametric test for comparing conditional ROC curves. *Computational Statistics & Data Analysis*. 07146, https://doi.org/10.1016/j.csda.2020.107146.

Fanjul-Hevia, A., Pardo-Fernández, J. C., Van Keilegom, I., and González-Manteiga, W. (2020b). A test for comparing conditional ROC curves with multidimensional covariates. *Manuscript under preparation*.

García-Portugués, E., González-Manteiga, W., and Febrero-Bande, M. (2014). A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics*, 23(3):761–778.

Gonçalves, L., Subtil, A., Oliveira, M. R., and Bermudez, P. (2014). ROC curve estimation: An overview. *REVSTAT–Statistical Journal*, 12(1):1–20.

González-Manteiga, W., Pardo-Fernández, J. C., and Van Keilegom, I. (2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics*, 38(1):169–184.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer-Verlag, New York, third edition.

Green, D. and Swets, J. (1966). *Signal Detection Theory and Psychophysics*. John Wiley, New York.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, 24(1):25–40.

Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F., Alonzo, T. A., and Cadarso-Suárez, C. (2012). Extending induced ROC methodology to the functional context. *Biostatistics*, 13(4):594–608.

Inácio de Carvalho, V., de Carvalho, M., Alonzo, T. A., and González-Manteiga, W. (2016). Functional covariate-adjusted partial area under the specificity-ROC curve with an application to metabolic syndrome diagnosis. *Annals of Applied Statistics*, 10(3):1472–1495.

Inácio de Carvalho, V., de Carvalho, M., and Branscum, A. J. (2017). Nonparametric bayesian covariate-adjusted estimation of the Youden index. *Biometrics*, 73(4):1279–1288.

Inácio de Carvalho, V., Jara, A., Hanson, T. E., and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling. *Bayesian Analysis*, 8(3):623–646.

Inácio de Carvalho, V. and Rodríguez-Álvarez, M. X. (2018). Bayesian nonparametric inference for the covariate-adjusted ROC curve. arXiv.

Janes, H., Longton, G., and Pepe, M. S. (2009). Accommodating covariates in receiver operating characteristic analysis. *Stata Journal*, 9:17–39.

Janes, H. and Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96(2):371–382.

Jokiel-Rokita, A. and Pulit, M. (2013). Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions. *Statistics and Computing*, 23(6):703–712.

Kim, E., Zeng, D., and Zhou, X.-H. (2015). Semiparametric transformation models for multiple continuous biomarkers in ROC analysis. *Biometrical Journal*, 57(5):808–833.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. Chapman & Hall/CRC, Boca Ratón.

Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, 93(444):1356–1364.

Lloyd, C. J. and Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statistics & Probability Letters*, 44(3):221–228.

López-de Ullibarri, I., Cao, R., Cadarso-Suárez, C., and Lado, M. J. (2008). Nonparametric estimation of conditional ROC curves: Application to discrimination tasks in computerized detection of early breast cancer. *Computational Statistics & Data Analysis*, 52(5):2623–2631.

Martínez-Camblor, P. (2007). Comparación de pruebas diagnósticas desde la curva ROC. *Revista Colombiana de Estadística*, 30(2):163–176.

Martínez-Camblor, P., Carleos, C., and Corral, N. (2011). Powerful nonparametric statistics to compare k independent ROC curves. *Journal of Applied Statistics*, 38(7):1317–1332.

Martínez-Camblor, P., Carleos, C., and Corral, N. (2013). General nonparametric ROC curve comparison. *Journal of the Korean Statistical Society*, 42(1):71–81.

Martínez-Camblor, P. and Corral, N. (2012). A general bootstrap algorithm for hypothesis testing. *Journal of Statistical Planning and Inference*, 142(2):589–600.

Martínez-Camblor, P., Corral, N., Rey, C., Pascual, J., and Cernuda-Morollón, E. (2014). Receiver operating characteristic curve generalization for non-monotone relationships. *Statistical Methods in Medical Research*, 26(1):113–123.

Molodianovitch, K., Faraggi, D., and Reiser, B. (2006). Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biometrical Journal*, 48(5):745–757.

Nakas, C. T. (2014). Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *REVSTAT – Statistical Journal*, 12:43–65.

Pardo-Fernández, J. C., Rodríguez-Álvarez, M. X., and Van Keilegom, I. (2014). A review on ROC curves in the presence of covariates. *REVSTAT–Statistical Journal*, 12(1):21–41.

Patilea, V., Sánchez-Sellero, C., and Saumard, M. (2016). Testing the predictor effect on a functional response. *Journal of the American Statistical Association*, 111(516):1684–1695.

Peng, L. and Zhou, X.-H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference*, 118(1):129–143.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Pérez-Fernández, S. (2017). *nsROC: Non-Standard ROC Curve Analysis*. R package version 1.0.

Pulit, M. (2016). A new method of kernel-smoothing estimation of the ROC curve. *Metrika*, 79(5):603–634.

Pundir, S. and Amala, R. (2014). Parametric receiver operating characteristic modeling for continuous data: A glance. *Model Assisted Statistics and Applications*, 9:121–135.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

Rodríguez, A. and Martínez, J. C. (2014). Bayesian semiparametric estimation of covariate-dependent ROC curves. *Bioestatistics*, 15(2):353–369.

Rodríguez-Álvarez, M. X., Roca-Pardiñas, J., and Cadarso-Suárez, C. (2011a). A new flexible direct ROC regression model: Application to the detection of cardiovascular risk factors by anthropometric measures. *Computational Statistics & Data Analysis*, 55(12):3257–3270.

Rodríguez-Álvarez, M. X., Roca-Pardiñas, J., and Cadarso-Suárez, C. (2011b). ROC curve and covariates: extending induced methodology to the non-parametric framework. *Statistics and Computing*, 21(4):483–499.

Rodríguez-Álvarez, M. X., Roca-Pardiñas, J., Cadarso-Suárez, C., and Tahoces, P. G. (2018). Bootstrap-based procedures for inference in nonparametric receiver-operating characteristic curve regression analysis. *Statistical Methods in Medical Research*, 27(3):740–764.

Schisterman, E. F., Faraggi, D., and Reiser, B. (2004). Adjusting the generalized ROC curve for covariates. *Statistics in Medicine*, 23(21):3319–3331.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

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

Valdés, L., San-José, E., Ferreiro, L., González-Barcala, F.-J., Golpe, A., Álvarez-Dobaño, J. M., Toubes, M. E., Rodríguez-Núñez, N., Rábade, C., Lama, A., and Gude, F. (2013). Combining clinical and analytical parameters improves prediction of malignant pleural effusion. *Lung*, 191(6):633–643.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics*, 56(4):1134–1138.

Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83(4):835–848.

Walsh, S. J. (1999). Goodness-of-fit issues in ROC curve estimation. *Medical Decision Making*, 19(2):193–201.

Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76(3):585–592.

Yao, F., Craiu, R. V., and Reiser, B. (2010). Nonparametric covariate adjustment for receiver operating characteristic curves. *The Canadian Journal of Statistics*, 38(1):27–46.

Yin, L., Diao, G., and Liu, A. (2017). A semiparametric method for comparing the discriminatory ability of biomarkers subject to limit of detection. *Statistics in Medicine*, 36(26):4141–4152.

Zhou, X. H. (1995). Testing an underlying assumption on a ROC curve based on rating data. *Medical Decision Making*, 15(3):276–282.

Zou, K. H., Gastwirth, J. L., and McNeil, B. J. (2003). A goodness-of-fit test for a receiver operating characteristic curve from continuous diagnostic test data. In Kolassa, J. E. and Oakes, D., editors, *Crossing boundaries: statistical essays in honor of Jack Hall*, volume 43 of *Lecture Notes–Monograph Series*, pages 59–68. Institute of Mathematical Statistics, Beachwood.

Zou, K. H., Hall, W. J., and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16(19):2143–2156.

Zou, K. H., Resnic, F. S., Talos, I.-F., Goldberg-Zimring, D., Bhagwat, J. G., Haker, S. J., Kikinis, R., Jolesz, F. A., and Ohno-Machado, L. (2005). A global goodness-of-fit test for receiver operating characteristic curve analysis via the bootstrap method. *Journal of Biomedical Informatics*, 38(5):395–403.