UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Estimación de curvas ROC condicionales en presencia de covariables composicionales

Gonzalo Carabajal

Directora: Dra. Graciela Boente

Octubre de 2024

Estimación de curvas ROC condicionales en presencia de covariables composicionales

La curva ROC es una herramienta ampliamente utilizada para evaluar la capacidad discriminatoria de una variable continua a la hora de distinguir individuos de dos clases o poblaciones. En casos en los que, en forma adicional, se miden covariables que brindan información sobre la condición de interés, puede resultar conveniente utilizar la curva ROC condicional en su lugar. Ésta es una función que depende de la distribución condicional de los datos, por lo que puede estimarse de diversas formas. Entre ellas, la metodología inducida, que consideraremos en este trabajo, asume, en cada población, un modelo de regresión que vincula a la variable clasificadora con las covariables. A partir de estos modelos es posible encontrar una forma explícita de la curva ROC condicional en términos de las funciones de regresión y de varianza, así como de la distribución de los errores. Por otro lado, en los últimos años, ha crecido el interés en el análisis de datos composicionales (CoDA), que surge cuando las observaciones son vectores no negativos cuyas componentes suman una constante. En este trabajo, adaptamos la metodología inducida utilizada para la estimación de curvas ROC condicionales al caso en que las covariables son de naturaleza composicional, obteniendo un estimador basado en funciones de distribución empírica que no hace suposiciones sobre la distribución de los datos y una versión suavizada de éste. Bajo condiciones de regularidad, obtenemos resultados de consistencia para ambos estimadores. Por otra parte, el comportamiento para muestras finitas se analiza mediante un estudio de simulación. Finalmente, aplicamos las herramientas desarrolladas a un conjunto de datos reales relacionados con el diagnóstico de diabetes a partir de mediciones de glucosa en sangre.

Palabras Clave: Curva ROC condicional, Biomarcador, Covariables, Datos composicionales.

Agradecimientos

A Graciela por haber aceptado dirigirme y haber estado siempre disponible y dispuesta a ayudarme. Junto a ella aprendí muchísimo de lo que es hacer investigación en Estadística.

A mi familia, por su apoyo incondicional. A mi mamá, por ser quien me introdujo al mundo de la matemática, y a mi papá, por alegrarse por cada meta cumplida. A mis hermanos, por ser mis mejores amigos.

A Ana y Marina, por tomarse el tiempo de leer este trabajo.

A mis compañeros y amigos de los últimos años de la carrera, por la compañía y por hacer más llevaderas las cursadas.

A mis alumnos del Washington School por las mañanas compartidas y a los directivos, por confiar en mí.

A la Comisión Fulbright, por haberme dado la posibilidad de vivir un viaje inolvidable con la beca Friends of Fulbright.

A todos y cada uno de los docentes de la carrera que participaron directa o indirectamente en mi formación.

¡Gracias!

Índice general

Capítulo 1
Introducción

Las Curvas ROC (*Receiver Operating Characteristic*), por sus siglas en inglés, son una herramienta estadística utilizada para evaluar la capacidad discriminatoria de una variable continua a la hora de distinguir entre dos grupos. Si bien surgieron durante la Segunda Guerra Mundial para asistir en la detección por medio de radares de objetos enemigos, hoy en día se utilizan ampliamente en diversos campos como la medicina, economía y psicología. En el contexto médico, la curvas ROC se utilizan, por ejemplo, para evaluar la precisión de una prueba diagnóstica basada en la medición de una variable continua, usualmente denominada *biomarcador*, que intenta distinguir entre enfermos y sanos. Esta será la terminología que adoptaremos en este trabajo a modo de ejemplificaión. Más en general, la curva ROC puede utilizarse para evaluar el desempeño de cualquier clasificador que divida observaciones en dos clases y esté basado en una variable continua.

Cuando se quiere clasificar individuos en dos grupos, por ejemplo, enfermos y sanos, a través de la medición de una variable continua $Y \in \mathbb{R}$ en cada uno de ellos, es de esperar que se cometan errores de clasificación, dando lugar, entre otros, a *falsos positivos*, individuos sanos clasificados como enfermos, y *verdaderos positivos*, individuos correctamente clasificados como enfermos. En resumidas palabras, la curva ROC describe la relación entre la proporción de *falsos positivos* y *verdaderos positivos* y puede definirse a partir de las distribuciones del biomarcador en las poblaciones sana y enferma, las cuales dan lugar a una expresión funcional que resulta monótona creciente y cuyo gráfico queda contenido en el cuadrado $[0,1] \times [0,1]$ pasando por los puntos $(0,0)$ y $(1,1)$. Para evaluar la capacidad discriminatoria del biomarcador, usualmente se utiliza el *área bajo la curva* como medida resumen, que mide el área bajo la curva ROC. Cuánto más cercana a uno sea esta área, mejor será la capacidad del biomarcador de distinguir entre enfermos y sanos. Dado que la curva ROC y, por consiguiente, el área bajo la curva, son objetos que se definen a partir de funciones de distribución, en un contexto práctico, el objetivo será estimar la curva utilizando como observaciones mediciones del biomarcador, tanto en personas enfermas como sanas.

En algunos escenarios, conocer información adicional de los individuos en forma de covariables puede mejorar la evaluación del desempeño de un biomarcador, en el sentido de que distintos valores de las covariables pueden afectar el resultado del biomarcador. Por ejemplo, en estudios clínicos que se llevan a cabo en distintos centros médicos, las mediciones del biomarcador podrían verse afectadas por el centro en el cual fueron tomadas o el profesional que realizó la medición. Otra covariable que puede afectar la capacidad discriminatoria del biomarcador es el estadío de la enfermedad, ya que casos menos severos pueden ser más difíciles de detectar. Una alternativa para estimar la curva ROC es la metodología inducida, que supone un modelo de regresión entre el biomarcador y las covariables. Esta metodología es muy general, con lo cual existen propuestas tanto paramétricas como no paramétricas

para estimar la curva ROC condicional, que es la curva que resulta de condicionar a un valor específico del vector de covariables de interés.

Algunos trabajos, como por ejemplo el de Inácio et al. (2012), han explorado la metodología inducida en el contexto funcional, es decir, cuando las covariables son funcionales. Sin embargo, en la literatura existente no hay precedentes de trabajos que investiguen el desempeño de los procedimientos de estimación de la curva ROC condicional en escenarios en los que las covariables sean de otro tipo. Es por eso que en este trabajo nos proponemos considerar covariables de naturaleza *composicional*, es decir, cuando las covariables son vectores aleatorios de coordenadas no negativas que suman 1. Dado que la aplicación de técnicas estadísticas estándar a datos composicionales puede dar lugar a conclusiones erróneas, es necesario adoptar una estrategia que tenga en cuenta la estructura de estos datos, cuyo espacio muestral se conoce como *simplex*. Este tipo de datos se caracteriza por contener información relativa y no absoluta ya que lo que importa es cuánto contribuye cada variable a un todo.

El objetivo principal de esta tesis, por lo tanto, será adaptar la metodología inducida al caso en que las covariables son de tipo composicional a través del estudio de dos estimadores semiparamétricos distintos de la curva ROC condicional. Uno de ellos está basado en las funciones de distribución empíricas de los errores de los modelos de regresión en los que se apoya la metodología inducida. El otro es una versión suavizada de éste y está inspirado en las ideas presentadas en Pulit (2016) para estimar la curva ROC no condicional de forma suave.

Una situación práctica en la que podría surgir la necesidad de estimar una curva ROC condicional con covariables composicionales es el caso del diagnóstico de una enfermedad en la que la composición de glóbulos blancos de un paciente es tenido en cuenta como covariable para evaluar la capacidad discriminatoria de cierto biomarcador. En este caso, dado que los niveles absolutos de monocitos, granulocitos y linfocitos pueden variar mucho de persona a persona, puede ser conveniente analizar dicha composición desde un enfoque composicional, considerando sólo la información relativa contenida en dicha composición.

A continuación, describimos brevemente la estructura de esta tesis.

En el **Capítulo 2** damos una introducción al llamado Análisis de Datos Composicionales (CoDA). En el mismo, describimos la geometría de Aitchison, que le da al simplex estructura de espacio vectorial, junto con una serie de transformaciones que vinculan al simplex con un espacio vectorial real y hacen posible el análisis estadístico de este tipo de datos.

En el **Capítulo 3**, introducimos primero la curva ROC en general y luego la curva ROC condicional y describimos la metodologí inducida como método de estimación de esta última. Luego, proponemos los dos estimadores semiparamétricos mencionados anteriormente en el escenario en el que las covariables son composicionales. Resultados de la consistencia de estos estimadores se presentan en el **Capítulo 4** junto con sus demostraciones.

En el **Capítulo 5**, presentamos los resultados de un estudio de simulación llevado a cabo con el fin de evaluar el desempeño de los estimadores propuestos bajo distintos escenarios. Además, consideramos dos casos distintos: el *balanceado*, en el que la cantidad de enfermos y sanos de la muestra utilizada para la estimación es la misma, y el *no balanceado*, en el que una de las poblaciones tiene un tamaño de muestra mucho mayor.

Por último, en el **Capítulo 6**, analizamos un conjunto de datos reales utilizando las técnicas desarrolladas a lo largo del trabajo.

Capítulo 2
Análisis de Datos Composicionales

2.1. ¿Qué es un dato composicional?

Muchas veces, en la práctica, los datos con los que trabajamos vienen dados de forma tal que expresan las partes de un todo, es decir, se tienen variables, todas medidas en la misma unidad, que suman cierta constante. Este es el caso cuando los datos corresponden a proporciones. Esto se puede deber a dos razones: ya sea porque sólo se cuenta con ese tipo de información, o bien porque que no son las mediciones en términos absolutos de las variables las que son relevantes para el análisis, sino las proporciones relativas entre ellas. A este tipo de datos se los conoce como *datos composicionales*.

Tomemos como ejemplo el caso de los gastos mensuales de un hogar. Analizar cómo están compuestos estos gastos en cierta población podría proporcionar información valiosa acerca de la relación entre los tipos de gastos con la ventaja de que hogares con poderes adquisitivos muy diferentes se vuelven comparables, lo cual no sería posible si los datos se analizaran en términos absolutos. Por otro lado, esto último sí sería conveniente si uno estuviera interesado en estudiar, por ejemplo, si un mayor poder adquisitivo lleva a un mayor gasto en alimentos.

Otros ejemplos de datos composicionales incluyen la composición geoquímica de rocas, la composición de los glóbulos blancos en sangre de acuerdo al tipo y la composición de la dieta de una persona.

La Tabla 2.1 muestra un ejemplo ficticio tomado de Filzmoser et al. (2018) donde se observa la composición de los gastos mensuales (en euros) de tres hogares, tanto en términos absolutos (como datos multivariados) como en términos relativos (como datos composicionales). Se incluyen sólo gastos específicos y se excluyen otros, como pueden ser gastos en salud o en ocio.

	Hogar	Vivienda	Alimentos	Transporte	Comunicaciones	Suma
Información	1	1710	950	570	570	3800
absoluta	2	540	300	180	180	1200
en euros	3	900	500	300	300	2000
Información	1	45	25	15	15	100
expresada	2	45	25	15	15	100
en %	3	45	25	15	15	100

Tabla 2.1: Ejemplo artificial de los gastos mensuales de tres hogares, expresados en euros y en porcentajes.

Como podemos ver, si bien en términos absolutos los gastos son muy diferentes, el dato composicional correspondiente es el mismo para los tres hogares. Esto nos lleva a definir a las composiciones y al operador de clausura, que transforma un dato multivariado de $\mathbb{R}_+^m$ en la composisición correspondiente.

2.1.1. El simplex como espacio muestral

Definición 2.1. *Un vector* $\mathbf{x} = (x_1, x_2, \ldots, x_m)^{\mathrm{T}}$ *es una* ***composición*** *de* $m-$*partes si sus componentes son todas estrictamente positivas y su suma es igual a una constante* κ*.*

Definición 2.2. *La* ***clausura*** *de un vector con coordenadas estrictamente positivas* $\mathbf{z} \in \mathbb{R}_+^m$ *se define como*

$$\mathcal{C}(\mathbf{z}) = \left(\frac{\kappa \cdot z_1}{\sum_{i=1}^m z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^m z_i}, \cdots, \frac{\kappa \cdot z_m}{\sum_{i=1}^m z_i} \right)^{\mathrm{T}}$$

En base a estas definiciones, es claro que las composiciones son en realidad clases de equivalencia dentro de $\mathbb{R}_+^m$ en donde dos vectores son equivalentes si son proporcionales entre sí. Además, la clausura resulta un representante conveniente de la clase.

El primer paso en cualquier análisis estadístico es determinar un espacio muestral para los datos. En este caso, el espacio muestral natural para los datos composicionales es el *simplex*, el cual se define a continuación.

Definición 2.3. *El* ***simplex*** *se define por*

$$\mathcal{S}^m = \left\{ \mathbf{x} = (x_1, \ldots, x_m)^{\mathrm{T}} \in \mathbb{R}^m \ : \ x_i > 0, \ 1 \leq i \leq m \ ; \ \sum_{i=1}^m x_i = \kappa \right\}.$$

Notemos que la dimensión de $\mathcal{S}^m$ es $m-1$. Además, sin pérdida de generalidad, podemos asumir que la constante κ de la definición es igual a 1, lo cual haremos de aquí en más.

Los diagramas de la Figura 2.1 representan a los espacios $\mathcal{S}^2$ y $\mathcal{S}^3$.

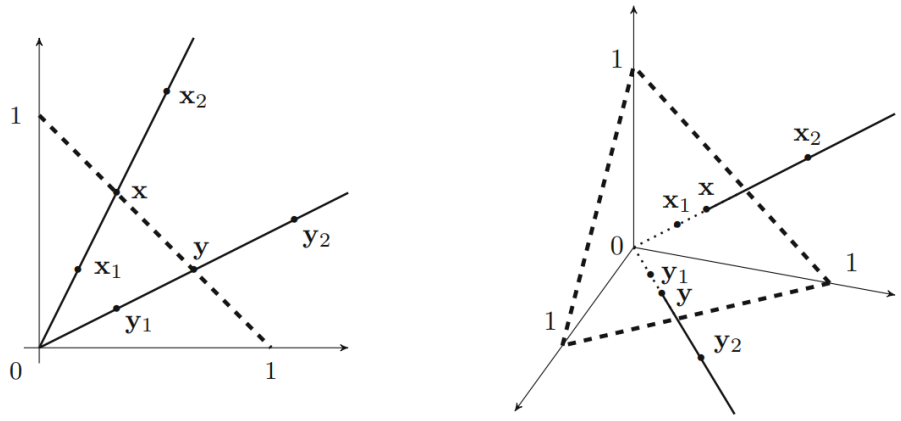
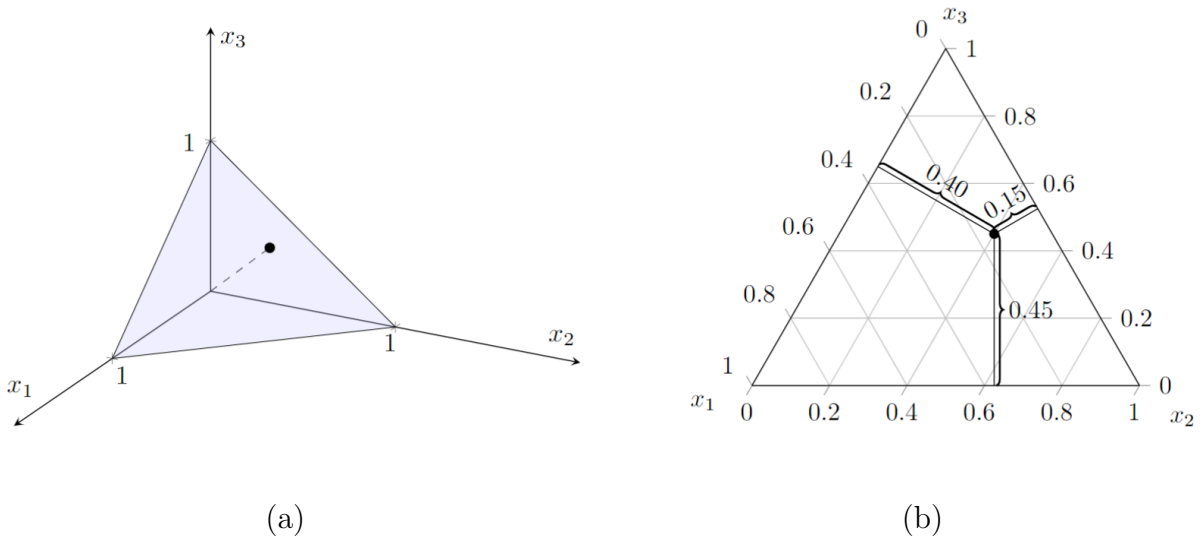


Figura 2.1: Representación de los simplex $\mathcal{S}^2$ (panel izquierdo) y $\mathcal{S}^3$ (panel derecho). Los vectores $\mathbf{x}, \mathbf{x}_1$ y $\mathbf{x}_2$ así como los vectores $\mathbf{y}, \mathbf{y}_1$ y $\mathbf{y}_2$ son equivalentes en el sentido composicional. Las proyecciones sobre el simplex en cada caso resultan en las composiciones $\mathbf{x}$ y $\mathbf{y}$. Figura extraída de Filzmoser et al. (2018).

2.1.2. Diagramas ternarios

Cuando $m = 3$, si bien el espacio muestral está incluido en $\mathbb{R}^3$, existe una forma de representar los datos en el plano. Esto es posible a través de los llamados *diagramas ternarios*. En la Figura 2.2 se representa a la composición $(0.15, 0.40, 0.45)^{\mathrm{T}}$ tanto sobre el simplex en $\mathbb{R}^3$ como sobre dicho diagrama.



(a) (b)

Figura 2.2: (a) Simplex como subconjunto de $\mathbb{R}^3$. (b) Diagrama ternario.

Un *diagrama ternario* es un triángulo equilátero en el que el dato $\mathbf{p} = (p_1, p_2, p_3)^{\mathrm{T}}$ se sitúa en su interior a una distancia p_1 desde el lado opuesto al vértice x_1, a una distancia p_2 desde el lado opuesto al vértice x_2 y a una distancia p_3 desde el lado opuesto al vértice x_3. La longitud del lado del triángulo puede elegirse libremente, por lo que a efectos prácticos la tomaremos como 1. Las coordenadas en el plano (u, v) se obtienen a partir de una combinación lineal convexa de los vértices del triángulo equilátero (prefijados por ejemplo en $A = (0, 0)^{\mathrm{T}}$, $B = (1, 0)^{\mathrm{T}}$ y $C = (0.5, \sqrt{3}/2)^{\mathrm{T}}$ para los vértices x_1, x_2 y x_3 respectivamente) utilizando las coordenadas del punto $\mathbf{p}$ en $\mathbb{R}^3$ como coeficientes:

$$(u, v)_{\mathbf{p}} = p_1\ A + p_2\ B + p_3\ C = p_2\ (1, 0)^{\mathrm{T}} + p_3 \left(0.5, \frac{\sqrt{3}}{2}\right)^{\mathrm{T}} = \left(p_2 + 0.5p_3\ ,\ \frac{\sqrt{3}}{2} p_3\right)^{\mathrm{T}}.$$

Si los vértices se fijan en otros puntos del plano, las nuevas coordenadas se obtienen de forma análoga. Por otro lado, la grilla dentro del triángulo se utiliza para visualizar regiones de puntos con cierto valor de cada parte (coordenada) de la composición. En este caso, el lado izquierdo del triángulo indica los distintos niveles de la coordenada x_1, el inferior, los de la coordenada x_2, y el derecho, indica los distintos niveles de la coordenada x_3. Dado que el orden en que se ubican las coordenadas en los vértices puede variar, es usual que se incluya sobre cada lado del triángulo una etiqueta con la coordenada correspondiente.

2.1.3. Problemas al aplicar métodos tradicionales

En esta sección describiremos brevemente algunos de los problemas que surgen al aplicar las nociones usuales de vectores aleatorios en $\mathbb{R}^m$ a vectores composicionales.

El problema de la correlación espúrea

Es importante destacar que aplicar métodos tradicionales a datos composicionales puede traer problemas, el más importante de los cuales fue descripto por Pearson (1897), quien demostró que existe una correlación espúrea en este tipo de datos, la cual puede llevar a interpretaciones erróneas. Más precisamente, debido a la estructura de estos datos, sucede que necesariamente existirá una correlación negativa entre las variables.

En efecto, como muestra Aitchison (1986), dada una composición $\mathbf{x}$ de m partes, sujeta a la restricción $x_1 + x_2 + \cdots + x_m = 1$, si x_i es alguna de sus componentes, se tiene que

$$\textsc{Cov}(x_i, x_1 + x_2 + \cdots + x_m) = \textsc{Cov}(x_i, 1) = 0\,.$$

Luego,

$$\textsc{Cov}(x_i, x_1)+\textsc{Cov}(x_i, x_2)+\cdots+\textsc{Cov}(x_i, x_{i-1})+\textsc{Cov}(x_i, x_{i+1})+\cdots+\textsc{Cov}(x_i, x_m) = -\textsc{Var}(x_i)\,,$$

con lo cual, a menos que x_i sea constante, el lado derecho será negativo y, por lo tanto, también deberá serlo alguna de las covarianzas del lado izquierdo. Además, como esto vale para $1 \leq i \leq m$, al menos m covarianzas en total serán necesariamente negativas.

El promedio como medida de centralidad

Cuando se tienen datos composicionales en el simplex, la media aritmética usual no representará correctamente el centro de los datos. En su lugar, la clausura de la media geométrica, es decir $g(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathcal{C}\left(\left(\prod_{i=1}^n \mathbf{x}_{i1}\right)^{1/n}, \ldots, \left(\prod_{i=1}^n \mathbf{x}_{im}\right)^{1/n}\right)$, será mejor opción como medida de centralidad. La Figura 2.3 muestra un ejemplo de esta situación.

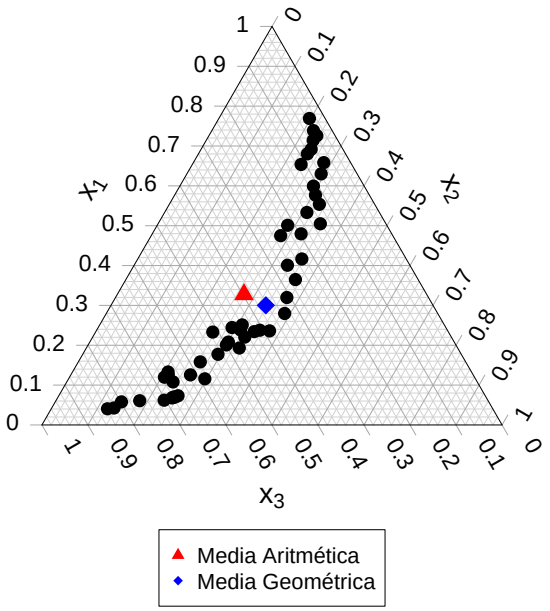


Figura 2.3: Media aritmética (en rojo) y geométrica (en azul) de un conjunto de datos composicionales.

Distancia entre composiciones

Consideremos los siguientes pares de composiciones y calculemos la distancia euclídea entre cada uno de ellos.

- $\mathbf{u}_1 = (0.05, 0.65, 0.30)^{\mathrm{T}} \quad \mathbf{u}_2 = (0.10, 0.60, 0.30)^{\mathrm{T}}$
- $\mathbf{w}_1 = (0.50, 0.20, 0.30)^{\mathrm{T}} \quad \mathbf{w}_2 = (0.55, 0.15, 0.30)^{\mathrm{T}}$

Si computamos el vector diferencia entre cada par, obtenemos que

$$\mathbf{u}_2 - \mathbf{u}_1 = (0.5, -0.5, 0)^{\mathrm{T}} \quad \text{y} \quad \mathbf{w}_2 - \mathbf{w}_1 = (0.5, -0.5, 0)^{\mathrm{T}},$$

de lo cual se deduce que la distancia euclídea entre ambos pares de vectores es la misma. Sin embargo, en términos relativos, la primera componente de $\mathbf{u}_2$ es el doble de la de $\mathbf{u}_1$, mientras que las primeras coordenadas de $\mathbf{w}_1$ y $\mathbf{w}_2$ difieren sólo en un 10 % y este hecho debería tenerse en cuenta al medir la cercanía entre los pares de puntos.

Los apartados anteriores muestran que la geometría Euclídea no es capaz de detectar correctamente similitud entre composiciones, por lo que será necesario el uso de otra geometría que dé cuenta de la naturaleza de estos datos. La pregunta, entonces, es ¿cómo analizamos este tipo de datos?

2.2. Geometría de Aitchison

2.2.1. Principios básicos

De acuerdo a la literatura existente, véase Filzmoser et al. (2018) ó Pawlowsky-Glahn et al. (2015), hay tres principios que deben ser respetados por cualquier método estadístico que se proponga analizar datos composicionales:

Invariancia de escala. Cualquier información relevante contenida en una composición debe ser invariante bajo cambios de escala, ya que las composiciones son en realidad clases de equivalencia formada por vectores proporcionales. Este principio apoya el hecho de tomar $\kappa = 1$ en la definición del espacio muestral y trabajar con composiciones cuyas componentes suman 1.

Invariancia por permutaciones. Una permutación de las partes de una composición no debe alterar la información contenida en ella, al igual que en la estadística multivariada estándar.

Coherencia subcomposicional. La información contenida en una composición de m partes no debería contradecirse con aquella contenida en una subcomposición (es decir, la composición que resulta de considerar sólo algunas de las partes de la composición original, y que por lo tanto, pertencece a un simplex de menor dimensión) de m_0 partes, con $m_0 < m$. En particular, la distancia entre dos composiciones debe ser mayor o igual a cualquiera de las distancias posibles entre las subcomposiciones de cada composición. Además, los cocientes entre dos partes de una subcomposición deben ser los mismos que en la composición original.

Veamos que la distancia euclídea usual no satisface la coherencia subcomposicional, lo cual, una vez más, la vuelve inadecuada para el tratamiento de datos composicionales. Si consideramos las composiciones $\mathbf{x} = (0.55, 0.40, 0.05)^{\mathrm{T}}$ e $\mathbf{y} = (0.10, 0.80, 0.10)^{\mathrm{T}}$, la distancia euclídea entre $\mathbf{x}$ e $\mathbf{y}$ es $d(\mathbf{x}, \mathbf{y}) = 0.604$, mientras que si tomamos las subcomposiciones que resultan de considerar las primeras dos partes de cada composición, es decir,

$$\mathbf{x}_s = \left(\frac{0.55}{0.55+0.40}, \frac{0.40}{0.55+0.40}\right)^{\mathrm{T}} \quad \text{e} \quad \mathbf{y}_s = \left(\frac{0.10}{0.10+0.80}, \frac{0.80}{0.10+0.80}\right)^{\mathrm{T}},$$

se tiene que $0.661 = d(\mathbf{x}_s, \mathbf{y}_s) > d(\mathbf{x}, \mathbf{y}) = 0.604$, lo cual viola el principio de coherencia subcomposicional.

2.2.2. Breve historia del análisis de datos composicionales

Tal como describen Pawlowsky-Glahn et al. (2015), el enfoque adoptado para analizar datos composicionales fue cambiando a lo largo del tiempo. A pesar de las advertencias de Pearson y, más tarde, del geólogo Felix Chayes, hasta los años 80, el análisis de datos composicionales se basó en la implementación de métodos estadísticos estándar.

Fue recién en la década de los 80 que John Aitchison descubrió que la información relativa que contienen las composiciones podía ser expresada en términos de *ratios* o cocientes entre sus componentes e introdujo una serie de transformaciones biyectivas basadas en *logratios* (logaritmos de cocientes entre componentes) que mapeaban los datos composicionales en el espacio real conocido. Si bien este enfoque fue ampliamente aceptado por la comunidad estadística, asumía implícitamente una geometría y medida euclídea en el espacio muestral composicional.

Finalmente, a partir del 2000, se comenzó a desarrollar la teoría que hoy en día sigue vigente y que describe formalmente la estructura geométrico-algebraica del simplex. En una propuesta pionera, Aitchison (1982) estableció las bases de una geometría para el simplex, ver también Aitchison (2003) o su primera edición Aitchison (1986). Esta geometría, conocida como *geometría de Aitchison*, resuelve los problemas de tener una suma constante y provee una estructura de espacio vectorial euclídeo al simplex, adaptanto las nociones de traslación, escalado, producto interno y ortogonalidad que conocemos en el espacio $\mathbb{R}^m$. A partir de esta geometría, se desarrollaron múltples procedimientos de análisis estadístico, adecuando al contexto de los datos composicionales las técnicas usuales de regresión, análisis de conglomerados o análisis discriminante, entre otras.

2.2.3. Estructura de espacio vectorial

La idea detrás de la geometría de Aitchison es dotar al simplex de estructura de espacio vectorial y realizar allí todas las operaciones involucradas en el análisis estadístico, sin recurrir a transformaciones de los datos.

Las operaciones análogas a la suma (o traslación) y al producto por escalar de $\mathbb{R}^m$ son la perturbación y la potenciación, definidas a continuación.

Definición 2.4. *La* ***perturbación*** *de* $\mathbf{x} \in \mathcal{S}^m$ *por* $\mathbf{y} \in \mathcal{S}^m$ *se define como*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_m y_m) = \left(\frac{x_1 y_1}{\sum_{j=1}^m x_j y_j}, \ldots, \frac{x_m y_m}{\sum_{j=1}^m x_j y_j}\right)^{\mathrm{T}} \in \mathcal{S}^m.$$

Definición 2.5. *La* ***potenciación*** *de* $\mathbf{x} \in \mathcal{S}^m$ *por una constante* $\alpha \in \mathbb{R}$ *se define por*

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_m^\alpha) = \left(\frac{x_1^\alpha}{\sum_{j=1}^m x_j^\alpha}, \ldots, \frac{x_m^\alpha}{\sum_{j=1}^m x_j^\alpha} \right)^{\mathrm{T}} \in \mathcal{S}^m.$$

Consideramos como ejemplo ilustrativo el conjunto de datos `WhiteCells` del paquete `ggtern` de `R`, analizado en Aitchison (1986), donde cada observación corresponde a la composición de los glóbulos blancos de 30 muestras de sangre, separándolos en tres tipos: granulocitos (G), linfocitos (L) y monocitos (M) calculada por medio de una técnica de análisis de imágenes. En la Figura 2.4 puede observarse el efecto que tienen la perturbación y la potenciación con distintos parámetros sobre estos datos.

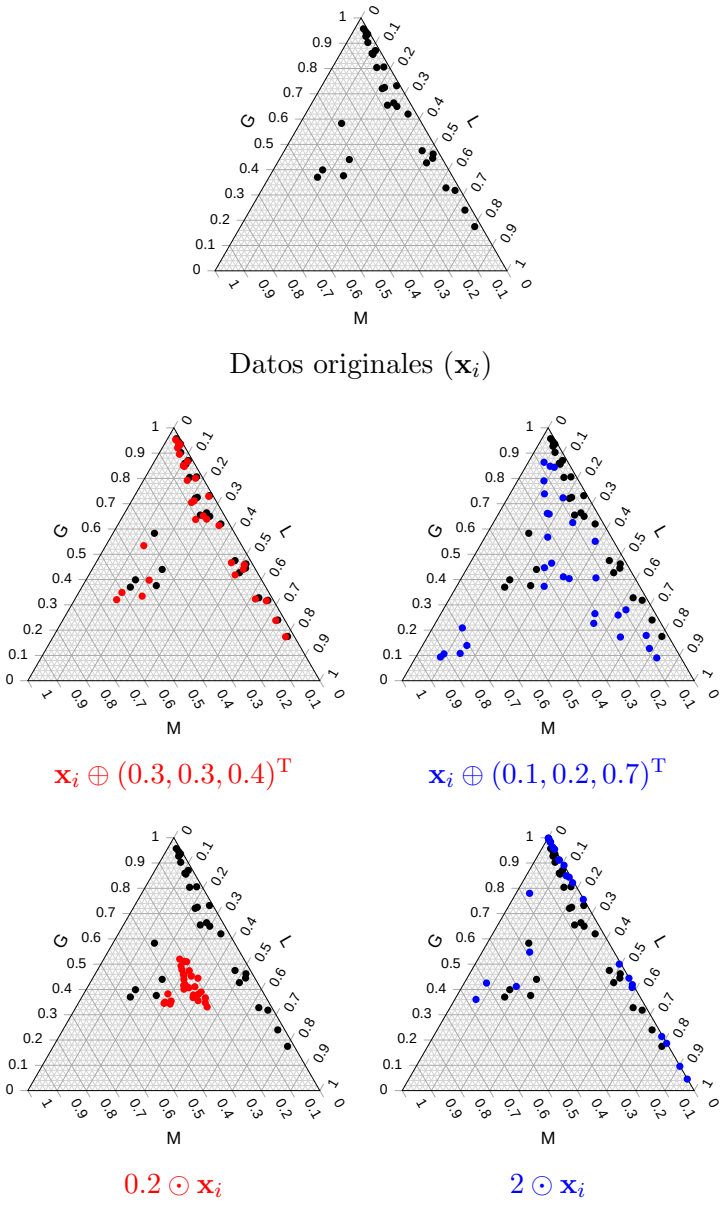


Figura 2.4: Efecto de la perturbación y potenciación sobre los datos de `WhiteCells`.

Con las operaciones de perturbación y potenciación, el simplex resulta un espacio vectorial real, lo cual se detalla en la siguiente proposición, cuya demostración omitimos.

Proposición 2.1. *El simplex, $(\mathcal{S}^m, \oplus, \odot)$, con las operaciones de perturbación y potenciación es un espacio vectorial. Más precisamente,*

- *Dados $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^m$, la perturbación satisface las siguientes propiedades:*

  A1. Conmutatividad: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$.

  A2. Asociatividad: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$.

  A3. Existencia de elemento neutro único

  $$\mathbf{n} = \mathcal{C}(1, \ldots, 1) = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right).$$

  A4. Existencia de elemento inverso: $\mathbf{x}^{-1} = \mathcal{C}(x_1^{-1}, x_2^{-1}, \ldots, x_m^{-1})$, *es decir, se cumple que* $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$.

- *Dados $\mathbf{x}, \mathbf{y} \in \mathcal{S}^m$, $\alpha, \beta \in \mathbb{R}$, la potenciación satisface las siguientes propiedades:*

  B1. Asociatividad: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$.

  B2. Propiedad distributiva 1: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus \mathbf{y}$.

  B3. Propiedad distributiva 2: $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$.

  B4. Existencia de elemento neutro: $1 \odot \mathbf{x} = \mathbf{x}$ *y el elemento neutro es único.*

De forma análoga a como se hace en $\mathbb{R}^m$, notaremos por $\mathbf{x} \ominus \mathbf{y}$ a la diferencia de la perturbación, dada por

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1} = \mathcal{C}(x_1 y_1^{-1}, x_2 y_2^{-1}, \ldots, x_m y_m^{-1}).$$

Notemos que, como es de esperar, la diferencia de la perturbación de una composición con si misma resulta en el elemento neutro:

$$\mathbf{x} \ominus \mathbf{x} = \mathcal{C}(x_1/x_1, x_2/x_2, \ldots, x_m/x_m) = \mathcal{C}(1, \ldots, 1) = \mathbf{n}.$$

Para obtener un espacio vectorial euclídeo, es decir, un espacio vectorial con producto interno de dimensión finita, consideramos el producto interno de Aitchison, con la norma y distancia asociadas, las cuales cumplen las propiedades análogas a las que valen en $\mathbb{R}^m$. Notaremos por $\log(\,\cdot\,)$ al logaritmo natural.

Definición 2.6.

- *El* ***producto interno de Aitchison*** *se define para $\mathbf{x}, \mathbf{y} \in \mathcal{S}^m$ como*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right) = \frac{1}{m} \sum_{i<j} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right).$$

- *La* ***norma de Aitchison*** *se define para $\mathbf{x} \in \mathcal{S}^m$ como*

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\{\log\left(\frac{x_i}{x_j}\right)\right\}^2} = \sqrt{\frac{1}{m} \sum_{i<j} \left\{\log\left(\frac{x_i}{x_j}\right)\right\}^2}.$$

- *La* ***distancia de Aitchison*** *entre* $\mathbf{x}$ *e* $\mathbf{y} \in \mathcal{S}^m$ *se define como*

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\{ \log\left(\frac{x_i}{x_j}\right) - \log\left(\frac{y_i}{y_j}\right) \right\}^2}$$
$$= \sqrt{\frac{1}{m} \sum_{i<j} \left\{ \log\left(\frac{x_i}{x_j}\right) - \log\left(\frac{y_i}{y_j}\right) \right\}^2}.$$

La distancia de Aitchinson tiene las siguientes dos propiedades, en relación a la perturbación y potenciación que son análogas a las que se verifican en $\mathbb{R}^m$

- *Relación entre la distancia y la perturbación*: Dados $\mathbf{x}$, $\mathbf{y}$, $\mathbf{p} \in \mathcal{S}^m$,

$$d_a(\mathbf{x} \oplus \mathbf{p}, \mathbf{y} \oplus \mathbf{p}) = d_a(\mathbf{x}, \mathbf{y}).$$

- *Relación entre la distancia y la potenciación*: Dados $\mathbf{x}$ e $\mathbf{y} \in \mathcal{S}^m$, $\alpha \in \mathbb{R}$,

$$d_a(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y}) = |\alpha| d_a(\mathbf{x}, \mathbf{y}).$$

2.3. Representación por medio de coordenadas

Una forma de analizar datos composicionales que surgió de las ideas dadas en Aitchison (1982) es a través de transformaciones que mapeen el simplex en un espacio real en el que se puedan utilizar técnicas multivariadas estándar. Veremos que aplicar alguna de estas transformaciones podrá ser visto como expresar a las composiciones en un sistema de coordenadas ortogonal respecto de la geometría de Aitchison.

Antes de seguir, veamos una forma de obtener una base de $\mathcal{S}^m$. El primer paso es construir un sistema generador y una forma de hacerlo es considerar la base canónica usual de $\mathbb{R}^m$, $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\}$ ya que $\mathcal{S}^m \subset \mathbb{R}^m$. Esto tiene el claro problema de que este conjunto de vectores no es ni un sistema de generadores ni una base de $\mathcal{S}^m$. De hecho, no toda combinación lineal de estos vectores estará en el simplex. Sin embargo, a partir de esa base, definiendo

$$\mathbf{w}_i = \mathcal{C}(\exp(\mathbf{e}_i)) = \mathcal{C}(1, 1, \ldots, e, \ldots, 1) \quad \text{para} \quad 1 \leq i \leq m, \tag{2.1}$$

obtenemos un sistema de generadores de $\mathcal{S}^m$ como espacio vectorial, ya que

$$\mathbf{x} = \bigoplus_{i=1}^{m} \log(x_i) \odot \mathbf{w}_i,$$

donde x_i es la coordenada $i-$ésima de $\mathbf{x}$ en la base $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\}$. Luego, se puede obtener una base de $\mathcal{S}^m$ descartando alguno de los $\mathbf{w}_i$. Más aún, si se quiere una base ortonormal, se puede aplicar el procedimiento de Gram-Schmidt, ver Egozcue et al. (2003).

2.3.1. Transformaciones *alr* y *clr*

La primera transformación considerada por Aitchison (1982) fue la *additive logratio* (*alr*), definida a continuación.

Definición 2.7. *Sea* $\mathbf{x} = (x_1, x_2, \ldots, x_m)^{\mathrm{T}}$ *una composición perteneciente al simplex* $\mathcal{S}^m$ *y consideremos la coordenada* $m-$*ésima,* x_m*, como parte referencial. La transformación alr definida de* $\mathcal{S}^m$ *en* $\mathbb{R}^{m-1}$ *viene dada por:*

$$alr(\mathbf{x}) = \left(\log\left(\frac{x_1}{x_m}\right), \ldots, \log\left(\frac{x_{m-1}}{x_m}\right)\right)^{\mathrm{T}}.$$

La transformación *alr* es un isomorfismo entre $\mathcal{S}^m$ y $\mathbb{R}^{m-1}$, con inversa $alr^{-1} : \mathbb{R}^{m-1} \to \mathcal{S}^m$ dada por

$$alr^{-1}(\mathbf{y}) = \mathcal{C}\left(\exp(y_1), \ldots, \exp(y_{m-1)}, 1\right).$$

Si bien los datos bajo la transformación *alr* pueden ser analizados con técnicas multivariadas estándar, ésta no es una isometría, es decir, no preserva distancias ni ángulos, lo cual la vuelve poco práctica. Además, no es una transformación simétrica en las componentes en el sentido de que se debe elegir la parte referencial que ocupa el lugar del denominador en los log-cocientes involucrados.

Es por estas desventajas que se introdujo la transformación *centered logratio* (*clr*), que efectivamente resulta una isometría y trata a todas las componentes por igual, reemplazando los denominadores de la transformación *alr* por la *media geométrica*, g_m, de las componentes de $\mathbf{x}$,

$$g_m(\mathbf{x}) = \left(\prod_{i=1}^{m} x_i\right)^{1/m}.$$

Definición 2.8. *Dada una composición* $\mathbf{x} \in \mathcal{S}^m$*, se define la transformación clr, como*

$$clr(\mathbf{x}) = \left(\log\left(\frac{x_1}{g_m(\mathbf{x})}\right), \ldots, \log\left(\frac{x_m}{g_m(\mathbf{x})}\right)\right)^{\mathrm{T}}.$$

La inversa de esta transformación viene dada por

$$clr^{-1}(\mathbf{y}) = \mathcal{C}(\exp(\mathbf{y})),$$

donde $\exp(\cdot)$ se aplica componente a componente.

Notemos que, a diferencia de la transformación *alr*, la transformación *clr* tiene como espacio de llegada a $\mathbb{R}^m$, y no $\mathbb{R}^{m-1}$. Sin embargo, las componentes de $clr(\mathbf{x})$ suman 0, pues

$$\begin{aligned}\sum_{k=1}^{m} y_k &= \sum_{k=1}^{m} \log\left[\frac{x_k}{\left(\prod_{j=1}^{m} x_j\right)^{1/m}}\right] = \sum_{k=1}^{m} \log\left[\frac{x_k}{\exp\left(\frac{1}{m}\sum_{l=1}^{m}\log x_l\right)}\right] \\ &= \sum_{k=1}^{m}\left(\log x_k - \frac{1}{m}\sum_{l=1}^{m}\log x_l\right) = \sum_{k=1}^{m}\log x_k - m\frac{1}{m}\sum_{l=1}^{m}\log x_l = 0\,,\end{aligned}$$

es decir, $clr(\mathbf{x})$ pertenece al hiperplano $\mathbb{H} = \{\mathbf{z} \in \mathbb{R}^m : \sum_{j=1}^m z_j = 0\}$ lo que significa que $clr(\mathcal{S}^m) = \mathbb{H}$, con $\dim(\mathbb{H}) = m - 1$. Esto da lugar a una limitación práctica de la transformación *clr* que es que la matriz de covarianzas de la transformación *clr* de cualquier vector aleatorio composicional resultará singular, ya que su rango columna no es completo. Observemos además que para cualquier $\mathbf{x} \in \mathcal{S}^m$, tenemos que $clr(\mathbf{x})$ es ortogonal al vector $\mathbf{1}_m$ que tiene todas sus coordenadas iguales a uno.

Notemos que la transformación *clr* aplicada a una composición simplemente devuelve las coordenadas de la misma respecto al sistema de generadores de (2.1), ya que se puede verificar que:

$$\mathbf{x} = \bigoplus_{i=1}^{m} \log\left(\frac{x_i}{g_m(\mathbf{x})}\right) \odot \mathbf{w}_i\,,$$

donde debemos recordar que la escritura de $\mathbf{x}$ respecto de $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$ no es única ya que no se trata de una base.

Si bien $clr(\mathbf{x})$ no corresponde a las coordenadas de $\mathbf{x}$ respecto de una base, satisface algunas propiedades importantes como las siguientes:

- $clr(\mathbf{n}) = \mathbf{0}_m$,
- $clr(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha\, clr(\mathbf{x}) + \beta\, clr(\mathbf{y})$,
- $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle clr(\mathbf{x}), clr(\mathbf{y}) \rangle$,
- $\|\mathbf{x}\|_a = \|clr(\mathbf{x})\| \implies d_a(\mathbf{x}, \mathbf{y}) = d(clr(\mathbf{x}), clr(\mathbf{y})) = \|clr(\mathbf{x}) - clr(\mathbf{y})\|$.

Estas propiedades nos muestran que otra diferencia con respecto a la transformación *alr* es que esta transformación preserva distancias. Más precisamente, si bien la *alr* es un isomorfismo entre $\mathcal{S}^m$ y $\mathbb{R}^{m-1}$ no es una isometría, en cambio, *clr* es una isometría (y por lo tanto un isomorfismo) pero entre $\mathcal{S}^m$ y el subespacio $\mathbb{H}$ de $\mathbb{R}^m$ de dimensión $m - 1$. Por este comportamiento y por el hecho de que estas transformaciones no pueden ser asociadas a un sistema de coordenadas ortogonal en el simplex, Egozcue et al. (2003) propusieron una nueva transformación que llamaron la transformación *isometric logratio* (*ilr*) de $m-$partes o composiciones. Dicha transformación da coordenadas en $\mathbb{R}^{m-1}$, llamadas $ilr-$coordenadas y tiene ciertas ventajas respecto de las definidas previamente que describiremos en la próxima sección.

2.3.2. Transformación *ilr*

Dado que con la geometría de Aitchison el simplex resulta un espacio vectorial euclídeo, y existe un único tal espacio de dimensión $m - 1$, salvo isometrías, sabemos que existe un isomorfismo isométrico entre $\mathcal{S}^m$ y $\mathbb{R}^{m-1}$, al cual nos referiremos como la transformación *ilr*.

Debido a que la transformación *ilr* es un isomorfismo y además una isometría, valen las siguientes propiedades, donde $\langle \cdot\,, \cdot \rangle$ y $\|\cdot\|$ denotan al producto interno y norma asociada usuales de $\mathbb{R}^{m-1}$.

- $ilr(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot ilr(\mathbf{x}) + \beta \cdot ilr(\mathbf{y})$.
- $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle ilr(\mathbf{x}), ilr(\mathbf{y}) \rangle$
- $d_a(\mathbf{x}, \mathbf{y}) = d(ilr(\mathbf{x}), ilr(\mathbf{y})) = \|ilr(\mathbf{x}) - ilr(\mathbf{y})\|$

Dada una composición $\mathbf{x} \in \mathcal{S}^m$, dicha transformación se construye a partir de una base ortonormal de $\mathcal{S}^m$. La idea es expresar a $\mathbf{x}$ en dicha base, con lo cual las $m-1$ componentes $ilr(\mathbf{x})$ corresponderán a las coordenadas de $\mathbf{x}$ en esa base. Esto quiere decir que habrá tantas transformaciones *ilr* como bases se puedan definir en $\mathbb{H}$, o sea, infinitas. Es por eso que estudiaremos sus propiedades en general. Para ello, necesitamos la siguiente definición.

Definición 2.9. *Sea* $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{m-1}\}$ *una base ortonormal de* $\mathcal{S}^m$*. Llamamos matriz de contraste a la matriz de* $m \times (m-1)$ *cuyas columnas están dadas por* $clr(\mathbf{w}_i)$ *con* $i = 1, \ldots, m-1$*.*

La siguiente proposición nos da una forma de expresar a la transformación *ilr* asociada a una base ortonormal del simplex como un simple producto matricial, explotando el hecho de que la transformación *clr* es una isometría.

Proposición 2.2. *Sea* $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{m-1}\}$ *una base ortonormal de* $\mathcal{S}^m$ *y* $\mathbf{U}$ *la matriz de contraste asociada. La transformación* $ilr : \mathcal{S}^m \to \mathbb{R}^{m-1}$ *asociada a* $\mathbf{U}$ *se puede expresar como*

$$\mathbf{x}^* = ilr(\mathbf{x}) = \mathbf{U}^{\mathrm{T}} clr(\mathbf{x}) = \mathbf{U}^{\mathrm{T}} \log(\mathbf{x}) \,, \tag{2.2}$$

y su inversa, $ilr^{-1} : \mathbb{R}^{m-1} \to \mathcal{S}^m$*, como*

$$ilr^{-1}(\mathbf{z}) = \mathcal{C}\left(\exp\{\mathbf{U}\mathbf{z}\}\right) . \tag{2.3}$$

Demostración. Comencemos notando que como $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{m-1}\}$ es una base de $\mathcal{S}^m$, entonces

$$\langle clr(\mathbf{w}_i), clr(\mathbf{w}_j) \rangle = \langle \mathbf{w}_i, \mathbf{w}_j \rangle_a = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} .$$

Luego, la matriz de contraste cumple que $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{m-1}$, con $\mathbf{I}_{m-1} \in \mathbb{R}^{(m-1)\times(m-1)}$ la matriz identidad. Sea $x_i^\star$ la $i-$ésima coordenada de $\mathbf{x}$ en la base $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{m-1}\}$, es decir,

$$\mathbf{x} = \bigoplus_{i=1}^{m-1} x_i^\star \odot \mathbf{w}_i \,,$$

y llamemos $\mathbf{x}^\star = (x_1^\star, \ldots, x_{m-1}^\star)^{\mathrm{T}}$. Como la base $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{m-1}\}$ es ortonormal respecto del producto interno de Aitchison y como la transformación *clr* es una isometría, tenemos que

$$x_i^\star = \langle \mathbf{x}, \mathbf{w}_i \rangle_a = \langle clr(\mathbf{x}), clr(\mathbf{w}_i) \rangle \,.$$

Por lo tanto, $\mathbf{x}^\star = \mathbf{U}^{\mathrm{T}} clr(\mathbf{x})$ como queríamos probar.

La última igualdad de (2.2), es decir, $\mathbf{x}^* = ilr(\mathbf{x}) = \mathbf{U}^{\mathrm{T}} \log(\mathbf{x})$, se sigue del hecho que $\mathbf{1}_m^{\mathrm{T}}\ clr(\mathbf{w}_j) = 0$, para $j = 1, \ldots, m-1$, donde $\mathbf{1}_m$ denota al vector m-dimensional de unos de donde $\mathbf{U}^{\mathrm{T}}\mathbf{1}_m = \mathbf{0}_{m-1}$. Efectivamente,

$$\mathbf{U}^{\mathrm{T}} clr(\mathbf{x}) = \mathbf{U}^{\mathrm{T}} \{\log(\mathbf{x}) - \mathbf{1}_m \log(g_m(\mathbf{x}))\} = \mathbf{U}^{\mathrm{T}} \log(\mathbf{x}) - \mathbf{U}^{\mathrm{T}}\mathbf{1}_m \log(g_m(\mathbf{x})) = \mathbf{U}^{\mathrm{T}} \log(\mathbf{x}) \,,$$

lo que concluye la demostración de (2.2).

Veamos ahora que vale (2.3). Sea $\mathbf{z} \in \mathbb{R}^{m-1}$ y veamos que al aplicar la transformación *ilr* a $\mathcal{C}(\exp\{\mathbf{Uz}\})$ recuperamos el vector $\mathbf{z}$. Por simplicidad de notación, para $\mathbf{u} \in \mathbb{R}^m_+$ indiquemos por $\|\mathbf{u}\|_1 = \sum_{j=1}^m u_j$, entonces

$$\log\left(\mathcal{C}\left(\mathbf{u}\right)\right) = \log(\mathbf{u}) - \mathbf{1}_m \log\left(\|\mathbf{u}\|_1\right) ,$$

de donde deducimos que

$$\begin{aligned} ilr\left(\mathcal{C}\left(\exp\{\mathbf{Uz}\}\right)\right) &= \mathbf{U}^{\mathrm{T}} \log\left(\mathcal{C}\left(\exp\{\mathbf{Uz}\}\right)\right) = \mathbf{U}^{\mathrm{T}}\left\{\log\left(\exp\{\mathbf{Uz}\}\right) - \mathbf{1}_m \log\left(\|\mathbf{Uz}\|_1\right)\right\} \\ &= \mathbf{U}^{\mathrm{T}} \log\left(\exp\{\mathbf{Uz}\}\right) - \mathbf{U}^{\mathrm{T}}\mathbf{1}_m \log\left(\|\mathbf{Uz}\|_1\right) = \mathbf{U}^{\mathrm{T}}\mathbf{Uz} \\ &= \mathbf{z} , \end{aligned}$$

lo que concluye la demostración de (2.3). □

En el trabajo de Egozcue et al. (2003), se construye una base ortonormal en el simplex aplicando el método de Gram-Schmidt a una base particular de $\mathcal{S}^m$ y obteniendo las coordenadas *ilr* de $\mathbf{x}$ dadas por

$$x_i^\star = \sqrt{\frac{i}{i+1}} \log\left(\frac{g_i(x_1, \ldots, x_i)}{x_{i+1}}\right) ,$$

donde $g_i : \mathbb{R}^i \to \mathbb{R}$ denota la media geométrica, es decir,

$$g_i(x_1, \ldots, x_i) = \left(\prod_{j=1}^{i} x_j\right)^{1/i} .$$

En el caso $m = 3$, las coordenadas *ilr* de $\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}}$ vienen dadas por

$$x_1^\star = \sqrt{\frac{1}{2}} \log\left(\frac{x_1}{x_2}\right) \quad \text{y} \quad x_2^\star = \sqrt{\frac{2}{3}} \log\left(\frac{\sqrt{x_1 x_2}}{x_3}\right) . \tag{2.4}$$

Esta será la transformación utilizada en el Capítulo 5 donde se presenta el estudio de simulación sobre la propuesta de curvas ROC condicionales que daremos en la Sección 3.4 del Capítulo 3. Esta transformación es la que se emplea por defecto en los paquetes de `R` más difundidos para el análisis de datos composicionales.

Una vez que definimos la transformación *ilr*, que es un isomorfismo isométrico entre el simplex, $\mathcal{S}^m$, y el espacio real $\mathbb{R}^{m-1}$, podemos pasar a trabajar directamente con las coordenadas *ilr* de los datos composicionales que querramos analizar, ya que éstas se mueven libremente en $\mathbb{R}^{m-1}$ y podemos aplicar allí técnicas de análisis estadístico estándar. Este procedimiento será aceptable siempre y cuando el método de análisis sea invariante por rotaciones, lo cual nos asegurará que se obtendrán los mismos resultados sea cual sea la base ortonormal elegida en el simplex para construir la transformación *ilr*.

Existen varias formas de construir la transformación *ilr*, eligiendo la base ortonormal de acuerdo al problema que se quiere analizar. En Pawlowsky-Glahn et al. (2015) pueden consultarse la definición de las *coordenadas pivote* y los *balances*, que son dos alternativas ampliamente utilizadas en las que la base se construye dando una mayor interpretación a las coordenadas.

También debemos notar que en casos en los que haya composiciones en los que alguna de las coordenadas es nula, las transformaciones *alr*, *clr* y *ilr* no podrán ser aplicadas, ya que en ellas hay logaritmos o cocientes involucrados. Este problema se conoce en la literatura como *presencia de ceros* y existen diversas estrategias para lidiar con él, las cuales no abordaremos en este trabajo. Puede consultarse Pawlowsky-Glahn et al. (2015) para profundizar en el tema.

2.4. Distribuciones en el simplex

Con el objetivo de llevar a cabo el análisis estadístico de datos composicionales, resultará esencial definir modelos de probabilidad sobre el simplex. En particular, nos serán de utilidad más adelante a la hora de realizar estudios de simulación.

Recordemos que para modelar una variable aleatoria es necesario definir un espacio de probabilidad, dado por tres elementos:

- un espacio muestral Ω formado por todos los valores posibles de la variable,
- una $\sigma-$álgebra $\mathcal{F}$ definida sobre Ω, y
- una función de probabilidad $\mathbb{P} : \mathcal{F} \to [0, 1]$.

Para variables aletorias reales, el espacio de probabilidad usual viene dado por $(\mathbb{R}^{m-1}, \mathcal{B}^*, \mathbb{P})$, donde $\mathcal{B}$ es la σ-álgebra generada por los conjuntos borelianos de $\mathbb{R}^{m-1}$ y $\mathbb{P}$ denota la función de probabilidad definida a partir de la medida de Lebesgue para el caso de variables aleatorias continuas.

Definición 2.10. *Decimos que un vector aleatorio $\mathbf{X} \in \mathbb{R}^m$ es una* ***composición aleatoria*** *si todos sus valores posibles pertenecen al simplex, $\mathcal{S}^m$, el cual se considera el espacio muestral y está dotado de la geometría de Aitchison, con las operaciones de perturbación y potenciación junto con el producto interno, norma y distancia de Aitchison.*

Sea $\mathbf{X} = (X_1, X_2, \ldots, X_m)^{\mathrm{T}}$ una composición aleatoria y consideremos sus coordenadas *ilr*, definidas a partir de cierta base ortonormal de $\mathcal{S}^m$ y la matriz de contraste $\mathbf{U}$ asociada:

$$\mathbf{X}^* = ilr(\mathbf{X}) = (X_1^*, X_2^*, \ldots, X_m^*)^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}} \log(\mathbf{X}).$$

Así, $\mathbf{X}^*$ es una variable aleatoria definida en $\mathbb{R}^{m-1}$, en particular, en el espacio de probabilidad usual dado por $(\mathbb{R}^{m-1}, \mathcal{B}^\star, \mathbb{P}^\star)$. Esto nos lleva a definir a las composiciones aleatorias continuas:

Definición 2.11. *Una composición aleatoria $\mathbf{X} \in \mathcal{S}^m$ es continua si existe una función real no negativa $f^* : \mathbb{R}^{m-1} \to \mathbb{R}$ definida en casi todo punto tal que para todo boreliano $B^\star$ de $\mathbb{R}^{m-1}$,*

$$\mathbb{P}(ilr(\mathbf{X}) \in B^\star) = \int_{B^\star} f^*(\mathbf{x}^*)\, d\mathbf{x}^*.$$

El espacio de probabilidad con el que trabajaremos será $(\mathcal{S}^m, \mathcal{B}, \mathbb{P})$, donde $\mathcal{B} = ilr^{-1}(\mathcal{B}^\star)$, y $\mathbb{P}(\mathbf{X} \in B) = \mathbb{P}^\star(ilr(\mathbf{X}) \in B^\star)$ con $B = ilr^{-1}(B)$.

La función f^*, de existir, es la función de densidad de las coordenadas *ilr* de $\mathbf{X}$, mientras que f, definida por $f(\mathbf{x}) = f^*(\mathbf{x}^*)$, es la función de densidad de $\mathbf{X}$ en el simplex. Si bien la expresión de f^* depende de la base ortonormal elegida en $\mathcal{S}^m$, no sucede lo mismo con la distribución en el simplex, ya que ambas describen la misma ley de probabilidad. Además, la medida de referencia en el caso de la distribución de las coordenadas *ilr* de $\mathbf{X}$ es la medida de Lebesgue, mientras que si trabajamos con la distribución de $\mathbf{X}$ en el simplex, ésta estará definida respecto de la medida de Aitchison. La medida de Aitchison se construye a partir de la medida de Lebesgue: si $\lambda(B^*)$ denota la medida de Lebsgue del boreliano $B^* \in \mathbb{R}^{m-1}$, la medida de Aitchison de un conjunto boreliano $B \in \mathcal{S}^m$, con $B = ilr^{-1}(B^*)$, viene dada por $\lambda_a(B^*) = \lambda(B)$.

2.4.1. Distribución Normal Logística

La distribución Normal Logística fue introducida por Aitchison & Shen (1980), quienes asumían que las coordenadas *alr* de una composición aleatoria seguían una distribución normal multivariada en $\mathbb{R}^{m-1}$, expresando la densidad respecto de la medida de Lebesgue en $\mathcal{S}^m$. Resulta que si en lugar de usar las coordenadas *alr* se utilizan las coordenadas *ilr* y se asume que éstas siguen una distribución normal, se obtiene la misma distribución, con la ventaja de que las coordenadas *ilr* corresponden a una base ortonormal del símplex (Pawlowsky-Glahn et al., 2015).

Definición 2.12. *Decimos que una composición aleatoria $\mathbf{X} \in \mathcal{S}^m$ sigue una distribución normal logística o una distribución normal en el simplex si su vector de coordenadas ilr, $\mathbf{X}^* = ilr(\mathbf{X})$, sigue una distribución normal multivariada.*

Esto quiere decir que si $\mathbf{X}$ sigue una distribución normal en el simplex, entonces $\mathbf{X}^*$ sigue una distribución normal en $\mathbb{R}^{m-1}$, o sea, $\mathbf{X}^* \sim \mathcal{N}_{m-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con función de densidad respecto de la medida de Lebesgue dada por:

$$f^*(\mathbf{x}^*) = \frac{1}{(2\pi)^{(m-1)/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu})\right\}, \tag{2.5}$$

donde $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son el vector de medias y la matriz de covarianzas de las coordenadas *ilr* de $\mathbf{X}$, respectivamente. Asumiremos que $\boldsymbol{\Sigma}$ es definida positiva y tomaremos a estos parámetros como los parámetros de la distribución de $\mathbf{X}$ en el simplex, la cual notaremos $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La expresión de la función de densidad de $\mathbf{X}$ en el simplex con respecto a la medida de Aitchison tendrá la forma

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{(m-1)/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(ilr(\mathbf{x}) - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(ilr(\mathbf{x}) - \boldsymbol{\mu})\right\}, \tag{2.6}$$

la cual dependerá de la base ortonormal elegida para construir la transformación *ilr*. Notemos que si bien el valor de los parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ también dependerá de esta elección, éstos cambiarán de forma tal que la distribución se preserve. Éste es el contenido del siguiente resultado.

Proposición 2.3. *Consideremos dos bases ortonormales de $\mathcal{S}^m$ con matrices de contraste asociadas $\mathbf{U}_1$ y $\mathbf{U}_2$, tal que las coordenadas ilr asociadas vienen dadas por $ilr_i(\mathbf{x}) = U_i^{\mathrm{T}} \log(\mathbf{x})$, $i = 1, 2$. Sea $\mathbf{X}$ una composición aleatoria con distribución normal en el simplex, o sea, $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ donde los parámetros corresponden a las densidades de las ecuaciones (2.5) y (2.6) con la transformación ilr_1.*

Luego, las coordenadas de $\mathbf{X}$ obtenidas utilizando la transformación ilr_2 siguen una distribución normal en $\mathbb{R}^{m-1}$, $ilr_2(\mathbf{X}) \sim \mathcal{N}_{m-1}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, donde $\boldsymbol{\mu}_2 = \mathbf{U}_2\mathbf{U}_1^{\mathrm{T}}\boldsymbol{\mu}_1$ y $\boldsymbol{\Sigma}_2 = \mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\boldsymbol{\Sigma}_1\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1$, es decir, $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ cuando las coordenadas de referencia son $ilr_2(\mathbf{x})$.

Demostración. Primero probemos que cualquier matriz de contraste $\mathbf{U}$ cumple que

$$\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\,. \tag{2.7}$$

Para eso, notemos que de la descomposición en valores singulares, $\mathbf{U}^{\mathrm{T}}\mathbf{U}$ y $\mathbf{U}\mathbf{U}^{\mathrm{T}}$ tienen los mismos autovalores excepto por un autovalor nulo adicional en $\mathbf{U}\mathbf{U}^{\mathrm{T}}$. Como $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{m-1}$, todos los autovalores no nulos son iguales a 1 (multiplicidad $m-1$) y los autovectores asociados son los de la base canónica de $\mathbb{R}^{m-1}$. Sin pérdida de generalidad, podemos asumir que $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}_m + \mathbf{M}$ donde $\mathbf{M} = k\,\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}$, es decir, es una matriz cuyas entradas son todas iguales a k. Premultiplicando por $\mathbf{U}^{\mathrm{T}}$ obtenemos que $\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}} + \mathbf{U}^{\mathrm{T}}\mathbf{M} = \mathbf{U}^{\mathrm{T}}$, pues las columnas de $\mathbf{U}$ suman 0. Análogamente, $\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{U}$ ya que $\mathbf{M}\mathbf{U}$ es una matriz nula pues $\mathbf{U}^{\mathrm{T}}\mathbf{1}_m = \mathbf{0}_{m-1}$. Luego, $\mathbf{U}^{\mathrm{T}}$ es una pseudoinversa de $\mathbf{U}$. Como $tr(\mathbf{U}\mathbf{U}^{\mathrm{T}}) = tr(\mathbf{U}^{\mathrm{T}}\mathbf{U}) = tr(\mathbf{I}_{m-1}) = m-1$ y $tr(\mathbf{I}_m + \mathbf{M}) = m + km$, obtenemos que $k = -1/m$.

Habiendo probado (2.7), si tomamos las coordenadas ilr_1 e ilr_2 de $\mathbf{x}$, tenemos que

$$\begin{cases} ilr_1(\mathbf{x}) = \mathbf{U}_1^{\mathrm{T}}\log(\mathbf{x}) \\ ilr_2(\mathbf{x}) = \mathbf{U}_2^{\mathrm{T}}\log(\mathbf{x}) \end{cases} \implies \begin{cases} \mathbf{U}_1 ilr_1(\mathbf{x}) = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}} \\ \mathbf{U}_2 ilr_2(\mathbf{x}) = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}} \end{cases} \implies \mathbf{U}_1 ilr_1(\mathbf{x}) = \mathbf{U}_2\, ilr_2(\mathbf{x}),$$

con lo cual, multiplicando a izquierda por $\mathbf{U}_1^{\mathrm{T}}$, obtenemos que $ilr_1(\mathbf{x}) = \mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\, ilr_2(\mathbf{x})$. Ahora, como por hipótesis, $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, tenemos que la densidad de $\mathbf{X}$ respecto de la medida de Aitchison viene dada por

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{(m-1)/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(ilr_1(\mathbf{x}) - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(ilr_1(\mathbf{x}) - \boldsymbol{\mu}_1)\right\}. \tag{2.8}$$

Definamos $\boldsymbol{\mu}_2 = \mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_2^{-1} = \mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\boldsymbol{\Sigma}_1^{-1}\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1$. Notemos que $\boldsymbol{\mu}_2 = \mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\boldsymbol{\mu}_1$, luego

$$\mathbf{U}_2\,\boldsymbol{\mu}_2 = \left(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\,\mathbf{1}_m^{\mathrm{T}}\right)\mathbf{U}_1\,\boldsymbol{\mu}_1 = \mathbf{U}_1\boldsymbol{\mu}_1 - \frac{1}{m}\mathbf{1}_m\,\mathbf{1}_m^{\mathrm{T}}\mathbf{U}_1\,\boldsymbol{\mu}_1\,,$$

por lo tanto, premultiplicando por $\mathbf{U}_1$ obtenemos

$$\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\,\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 - \frac{1}{m}\mathbf{U}_1^{\mathrm{T}}\mathbf{1}_m\,\mathbf{1}_m^{\mathrm{T}}\mathbf{U}_1\,\boldsymbol{\mu}_1 = \boldsymbol{\mu}_1\,,$$

donde en el último paso usamos que, como las columnas de $\mathbf{U}_1$ suman 0 por estar definida en base a la transformacón *clr*, o sea, $\mathbf{1}_m^{\mathrm{T}}\mathbf{U}_1 = \mathbf{0}_{m-1}$.

Reemplazando $ilr_1(\mathbf{x})$ en (2.8) por $\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\, ilr_2(\mathbf{x})$, obtenemos lo siguiente:

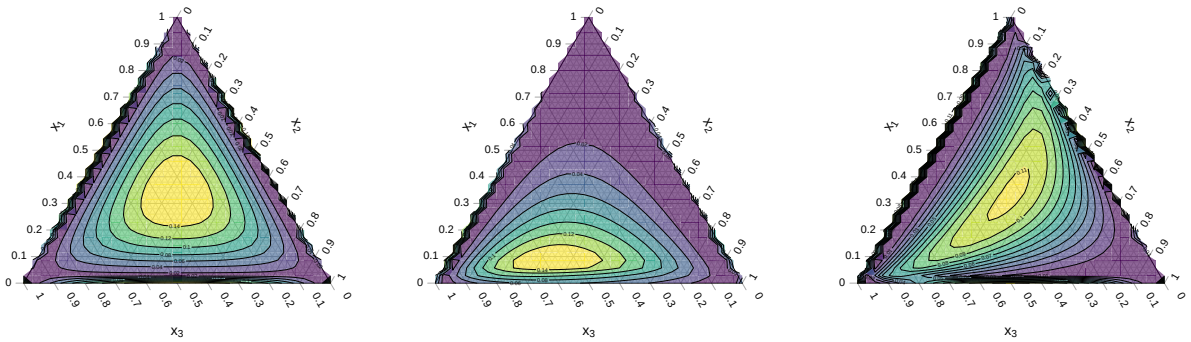
$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{(2\pi)^{(m-1)/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(ilr_1(\mathbf{x}) - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(ilr_1(\mathbf{x}) - \boldsymbol{\mu}_1)\right\} \\ &= \frac{1}{(2\pi)^{(m-1)/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(ilr_2(\mathbf{x}) - \boldsymbol{\mu}_2)^T\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\boldsymbol{\Sigma}_1^{-1}\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1(ilr_2(\mathbf{x}) - \boldsymbol{\mu}_2)\right\}. \end{aligned}$$

Para probar que $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, faltaría ver que $\boldsymbol{\Sigma}_2 = \mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\boldsymbol{\Sigma}_1\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1$. Para ello, alcanza con probar que $(\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1)^{-1} = \mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2$:

$$\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2 = \mathbf{U}_2^{\mathrm{T}}\left(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\right)\mathbf{U}_2 = \mathbf{U}_2^{\mathrm{T}}\mathbf{U}_2 - \frac{1}{m}\mathbf{U}_2^{\mathrm{T}}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\mathbf{U}_2 = \mathbf{I}_{m-1},$$

pues $\mathbf{1}_m^{\mathrm{T}}\mathbf{U}_2 = \mathbf{0}_{m-1}$. Análogamente, $\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1 = \mathbf{I}_{m-1}$ y $(\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_2)^{-1} = \mathbf{U}_2^{\mathrm{T}}\mathbf{U}_1$. □

La Figura 2.5 muestra las curvas de nivel para la densidad de la ecuación (2.6) en el simplex.



$\boldsymbol{\mu} = (0,0)$ $\quad$ $\boldsymbol{\mu} = (1,1)$ $\quad$ $\boldsymbol{\mu} = (0,0)$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 2 \\ 1 & 4 \end{pmatrix}$$

Figura 2.5: Curvas de nivel de la densidad de una composición con distribución normal para tres casos distintos y respecto de la medida de Aitchison.

Es importante destacar que valen propiedades análogas a las que se verifican en $\mathbb{R}^{m-1}$, como la normalidad de toda combinación lineal de composiciones normales e incluso el Teorema Central del Límite. Estas propiedades, junto con sus demostraciones pueden consultarse en Pawlowsky-Glahn et al. (2015).

2.4.2. Distribución de Dirichlet

La otra distribución propuesta en la literatura para trabajar con datos composicionales es la distribución de Dirichlet, que tiene como soporte exactamente el simplex $\mathcal{S}^m$ y resulta de la clausura de m variables positivas independientes con distribución Gamma con mismo parámetro de escala.

Definición 2.13. *Una composición aleatoria continua $\mathbf{X} \in \mathcal{S}^m$ sigue una distribución de Dirichlet de parámetro $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)^{\mathrm{T}}$ con $\alpha_i > 0$, para $i = 1, \ldots, m$, si su función de densidad respecto de la medida de Lebesgue se escribe como*

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_m)}{\prod_{j=1}^m \Gamma(\alpha_j)} \prod_{j=1}^m x_j^{\alpha_j - 1}\, \mathbb{I}_{\{\mathbf{x} \in \mathcal{S}^m\}}\,, \tag{2.9}$$

donde $\Gamma(\,\cdot\,)$ es la función Gamma de Euler.

Notemos que como $\mathbf{x} \in \mathcal{S}^m$, una de las coordenadas de la composición puede expresarse en términos de las otras, por ejemplo, $x_m = 1 - \sum_{i=1}^{m-1} x_i$, lo cual es cómun en la literatura. En este caso, la densidad de (2.9) está expresada en términos de la medida de Lebesgue, si queremos expresarla respecto de la medida de Aitchison, obtenemos la siguiente expresión:

$$f_a(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\sqrt{m}\,\Gamma(\alpha_+)}{\prod_{j=1}^m \Gamma(\alpha_j)} \prod_{j=1}^m x_j^{\alpha_j}\, \mathbb{I}_{\{\mathbf{x}\in\mathcal{S}^m\}}, \quad \text{donde} \quad \alpha_+ = \sum_{i=1}^m \alpha_i.$$

La Figura 2.6 muestra las curvas de nivel de dos distribuciones Dirichlet distintas, respecto de ambas medidas.

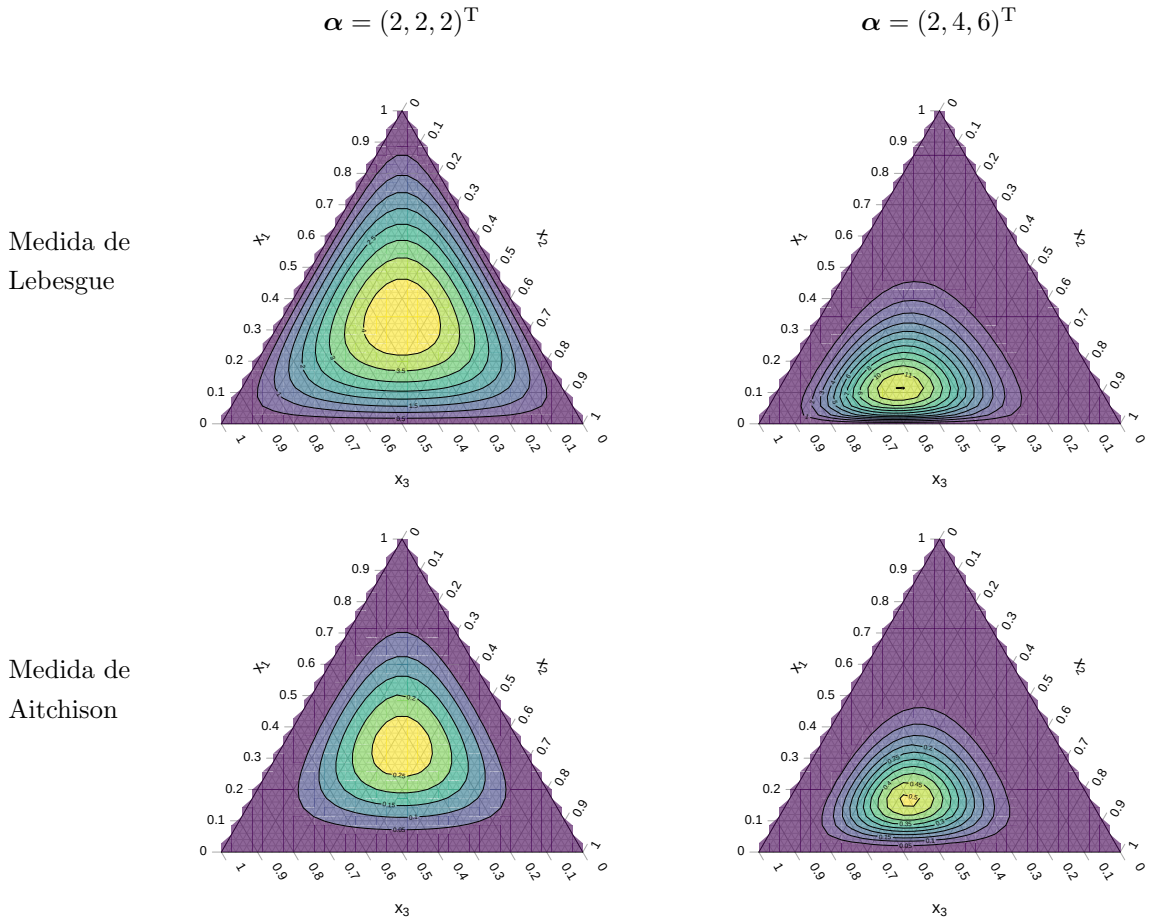


Figura 2.6: Curvas de nivel de la densidad de una composición con distribución Dirichlet para dos parámetros distintos. En el panel superior, se presentan las curvas de nivel de las densidades respecto de la medida de Lebesgue, mientras que en el inferior las de las densidades respecto de la medida de Aitchison.

La distribución de Dirichlet ha sido ampliamente utilizada en el campo de la estadística bayesiana, ya que es la distribución conjugada a priori de la distribución multinomial. Además, tiene la ventaja de tener soporte exactamente igual a $\mathcal{S}^m$, por lo que no es necesario recurrir a ninguna transformación. Es por estas razones que la distribución de Dirichlet cobró popularidad.

2.4.3. Otras distribuciones

Resulta importante mencionar que, utilizando las coordenadas *ilr* de una composición, es posible definir distribuciones en el simplex a partir de distribuciones multivariadas conocidas. Por ejemplo, en el Capítulo 5 que presenta los resultados de un estudio de simulación, contemplaremos el caso en que las coordenadas *ilr* de una composición siguen una distribución uniforme en una región acotada de $\mathbb{R}^{m-1}$.

2.5. Modelo de regresión lineal simplicial-real

Dado que en el siguiente capítulo consideraremos modelos lineales que involucren covariables composicionales, en esta sección describimos la extensión del modelo de regresión lineal al caso en el que las covariables son de naturaleza composicional.

El modelo lineal relaciona a la variable respuesta real Y con las covariables composicionales $\mathbf{X}$ a través de la siguiente relación:

$$\mathbb{E}(Y|\mathbf{X}=\mathbf{x}) = \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x} \rangle_a, \tag{2.10}$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^{\mathrm{T}} \in \mathcal{S}^m$ y $\beta_0 \in \mathbb{R}$.

Dado que la respuesta es real, si se cuenta con observaciones $(y_i, \mathbf{x}_i)$, $1 \leq i \leq n$, independientes que cumplen el modelo (2.10), la forma de ajustar el modelo es la misma que con covariables reales, es decir, a través de los estimadores de mínimos cuadrados de los parámetros. Esto es, el estimador viene definido por

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathcal{S}^m}{\operatorname{argmin}} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_a\right)^2 . \tag{2.11}$$

Para dar una expresión explícita de $\widehat{\boldsymbol{\beta}}$, recordemos que el producto interno de Aitchison satisface que $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle ilr(\mathbf{x}), ilr(\mathbf{y}) \rangle = ilr(\mathbf{x})^{\mathrm{T}} ilr(\mathbf{y})$, con lo cual si $\boldsymbol{\beta}^*$ representa a las coordenadas *ilr* del vector de parámetros $\boldsymbol{\beta}$, es decir, $\boldsymbol{\beta}^* = ilr(\boldsymbol{\beta})$ tenemos que

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_a\right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - ilr(\mathbf{x}_i)^{\mathrm{T}} \boldsymbol{\beta}^*\right)^2 .$$

Por lo tanto, el problema de minimización dado en (2.11) es equivalente a hallar estimadores de β_0 y $\boldsymbol{\beta}^*$ utilizando las coordenadas *ilr* de las covariables, es decir,

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}^*}) = \underset{(\beta_0, \boldsymbol{\beta}^*) \in \mathbb{R} \times \mathbb{R}^{m-1}}{\operatorname{argmin}} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - ilr(\mathbf{x}_i)^{\mathrm{T}} \boldsymbol{\beta}^*\right)^2 .$$

Una vez hallado $\widehat{\boldsymbol{\beta}^*}$, se aplica la inversa de la transformación *ilr* para obtener un estimador de $\boldsymbol{\beta}$, es decir, se define $\widehat{\boldsymbol{\beta}} = ilr^{-1}(\widehat{\boldsymbol{\beta}^*})$. En conclusión, la estimación de los parámetros de la regresión en el caso variable respuesta real y covariables composicionales se reduce a la estimación clásica con covariables en $\mathbb{R}^{m-1}$.

Capítulo 3
Curvas ROC con Covariables

3.1. Curva ROC

3.1.1. Conceptos básicos

Consideremos una variable continua Y que se desea emplear para clasificar individuos en dos poblaciones. A efectos prácticos, supondremos que la variable representa una prueba diagnóstica, o *biomarcador*, que se quiere utilizar para distinguir entre individuos *enfermos* y *sanos* de cierta patología de acuerdo a un valor umbral c, clasificando al individuo en cuestión como enfermo si $Y \geq c$ y como sano si $Y < c$.

Sea F_D la distribución del biomarcador Y en la población enferma y F_H la distribución del biomarcador Y en la población sana. Sean además Y_D e Y_H las variables aleatorias correspondientes al biomarcador en la población enferma y sana, respectivamente, que a lo largo de este capítulo supondremos independientes. Es decir, $Y_D \sim F_D$ e $Y_H \sim F_H$.

Como es de esperar, cuando se mide el biomarcador de un individuo, pueden cometerse dos tipos de errores de clasificación:

- Se clasifica como enferma a una persona sana, lo que se conoce como un *falso positivo*.
- Se clasifica como sana a una persona enferma, lo que se conoce como un *falso negativo*.

Para dar cuenta de estos errores de clasificación, veamos algunos términos clave. Para ello, notemos que podemos presentar los resultados de una prueba diagnóstica a través de una tabla de contingencia, llamada *matriz de confusión*, como la presentada en la Tabla 3.1, donde se representan la cantidad de verdaderos positivos (VP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (VN).

	Estado del individuo	
Resultado de la prueba	Enfermos ($D = 1$)	Sanos ($D = 0$)
Positiva ($Y \geq c$)	VP	FP
Negativa ($Y < c$)	FN	VN

Tabla 3.1: Resultados de una prueba diagnóstica sobre un cierto número de individuos.

En base a esta información, definimos los siguientes términos:

- Sensibilidad o Fracción de Verdaderos Positivos (FVP):
  $\mathrm{Se}(c) = \mathrm{FVP}(c) = \mathbb{P}(Y \geq c \mid D = 1) = \dfrac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$
- Especificidad (Es) o Fracción de Verdaderos Negativos (FVN):
  $\mathrm{Es}(c) = \mathrm{FVN}(c) = \mathbb{P}(Y < c \mid D = 0) = \dfrac{\mathrm{VN}}{\mathrm{FP} + \mathrm{VN}}$
- Fracción de Falsos Positivos (FFP):
  $\mathrm{FFP}(c) = \mathbb{P}(Y \geq c \mid D = 0) = \dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{VN}} = 1 - \mathrm{Es}(c)$
- Fracción de Falsos Negativos (FFN):
  $\mathrm{FFN}(c) = \mathbb{P}(Y < c \mid D = 1) = \dfrac{\mathrm{FN}}{\mathrm{VP} + \mathrm{FN}} = 1 - \mathrm{Se}(c)$

Notemos que estos valores dependen del valor umbral c elegido. Además, una característica deseada de la prueba diagnóstica es que tenga valores de sensibilidad y especificidad cercanos a uno, ya que esto significaría que la proporción de individuos mal clasificados es chica. Para obtener una representación visual de estas dos medidas, observemos la Figura 3.1, donde vemos graficadas densidades hipotéticas de Y_D e Y_H y el valor umbral $c = 18$. Cabe destacar que asumiremos que los valores del biomarcador en la población enferma son, en promedio, mayores que en la población sana. En otras palabras, que la media de la distribución F_D es mayor que la media de F_H.

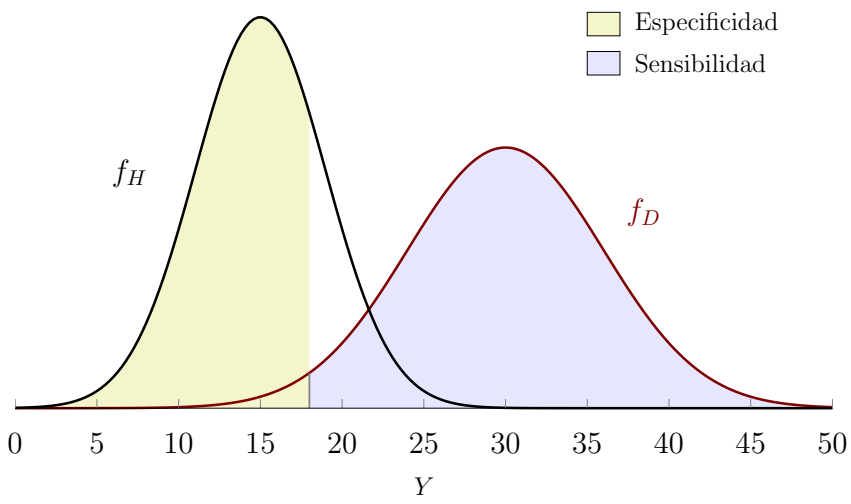


Figura 3.1: Densidades hipotéticas de Y_D e Y_H junto con la especificidad y sensibilidad para el punto de corte $c = 18$

Observando esta representación gráfica, podemos deducir que cuanto mayor sea la separación de las densidades del biomarcador en enfermos y sanos, mejor será la capacidad discriminatoria de la prueba, ya que tanto la sensibilidad como la especificidad tenderán a acercarse a 1 para algún valor de c. Esto se verá reflejado en la curva ROC, la cual definimos a continuación.

3.1.2. Definición de la curva ROC

Dada una prueba diagnóstica que clasifica entre enfermos y sanos en base a un biomarcador Y, se define la curva ROC de la siguiente manera:

Definición 3.1. *La curva ROC se define como el gráfico de* $Se(c)$ *versus* $1-Es(c)$ *para* $-\infty < c < \infty$*, es decir,*

$$ROC = \{(FFP(c)\ ,\ FVP(c))\ : -\infty < c < \infty\}\,.$$

En esta definición, la variable c parametriza a la curva. Podemos distinguir tres casos notables de curvas ROC, representados en la Figura 3.2. La curva de la derecha corresponde a una prueba diagnóstica perfecta que distingue completamente entre enfermos y sanos, es decir, para el cual existe un valor umbral c tal que $\mathrm{FVP}(c) = 1$ y $\mathrm{FFP}(c) = 0$, haciendo que la curva pase por la esquina superior izquierda. Esto ocurre si las densidades del biomarcador en cada población no tienen soporte común. El otro caso extremo es el de la curva de la izquierda, que corresponde a un test no informativo, ya que $\mathrm{FVP}(c) = \mathrm{FFP}(c)$ para todo c. En este caso, la densidad de Y es la misma en ambas poblaciones. Por último, la curva del medio representa una curva ROC típica. A medida que la precisión de la pueba mejora, la curva se desplaza hacia la curva que pasa por el $(0, 1)$. Esto quiere decir que, visualmente, dadas dos curvas ROC, la que más cerca del extremo superior izquierdo esté, mejor capacidad discriminatoria tendrá. Claramente, en casos en los que las dos curvas se crucen, podría no ser tan claro decidir cuál está más cerca del punto $(0, 1)$, lo cual nos llevará más adelante a definir una medida que permita comparar la capacidad discriminatoria de dos biomarcadores distintos.

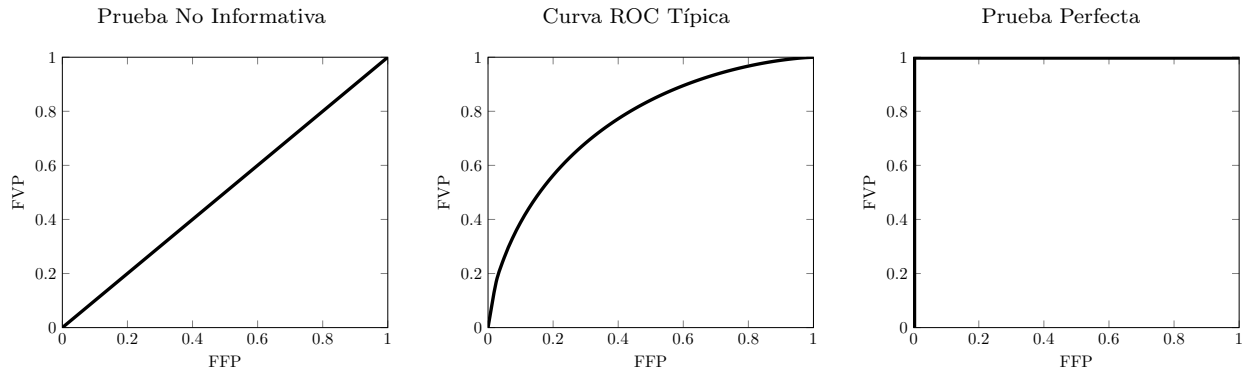


Figura 3.2: Casos notables de curvas ROC.

Veamos una definición equivalente de la curva ROC:

Proposición 3.1. *La curva ROC viene dada por el gráfico de*

$$ROC(p) = 1 - F_D\left(F_H^{-1}(1-p)\right)\ ,$$

para $p \in [0, 1]$*, donde* $F_H^{-1}(p) = \text{ínf}\{x \in \mathbb{R} : F_H(x) \geq p)\}$ *es la función cuantil de* Y_H*.*

Demostración. A partir de la definición de especificidad, tenemos que

$$\mathrm{Es}(c) = \mathbb{P}(Y < c \mid D = 0) = F_H(c) \implies 1 - \mathrm{Es}(c) = 1 - F_H(c).$$

Análogamente, tenemos que $\text{Se}(c) = 1 - F_D(c)$. Ahora bien, para cada c, la curva ROC viene dada por el par

$$(1 - \text{Es}(c)\ ,\ \text{Se(c)}) = (1 - F_H(c), 1 - F_D(c))\ .$$

Sea $p = 1 - F_H(c)$, luego $c = F_H^{-1}(1-p)$, de donde $\text{ROC}(p) = 1 - F_D\left(F_H^{-1}(1-p)\right)$. □

En base a esta definición equivalente y al hecho de que F_D y F_H son funciones de distribución, es claro que la curva ROC resulta monótona no decreciente. Efectivamente, si $p_1 > p_2$, entonces $1 - p_1 < 1 - p_2$. Por lo tanto, como F_H^{-1} es no-decreciente, deducimos que $F_H^{-1}(1-p_1) \leq F_H^{-1}(1-p_2)$, por lo tanto

$$1 - F_D(F_H^{-1}(1-p_1)) \geq 1 - F_D(F_H^{-1}(1-p_2))\,,$$

es decir, $\text{ROC}(p_1) \geq \text{ROC}(p_2)$.

Un resultado interesante es que cualquier transformación estrictamente creciente del biomarcador preserva la curva ROC, como se establece a continuación.

Proposición 3.2. *La curva ROC es invariante respecto de cualquier transformación estrictamente creciente de Y.*

Demostración. Sea h una transformación monótona creciente y consideremos la variable $W = h(Y)$. Indiquemos por $W_D = h(Y_D)$ y $W_H = h(Y_H)$.

Veamos que cualquier punto $(\text{FFP}(c)\ ,\ \text{FVP}(c))$ de la curva ROC de Y también pertenece a la curva ROC de W.

Consideremos $d = h(c)$. Entonces,

$$\begin{aligned}\mathbb{P}(W_D \geq d) &= \mathbb{P}(h(Y_D) \geq h(c)) = \mathbb{P}(Y_D \geq c)\,, \quad \text{y}\\ \mathbb{P}(W_H \geq d) &= \mathbb{P}(h(Y_H) \geq h(c)) = \mathbb{P}(Y_H \geq c)\,.\end{aligned}$$

Luego, si llamamos $\widetilde{\text{FFP}}(d)$ y $\widetilde{\text{FVP}}(d)$ las probabilidades de falsos positivos y verdaderos positivos asociadas a W, respectivamente, tenemos que

$$\begin{aligned}\left(\widetilde{\text{FFP}}(d)\ ,\ \widetilde{\text{FVP}}(d)\right) &= (\mathbb{P}(W_H \geq d)\ ,\ \mathbb{P}(W_D \geq d)) = (\mathbb{P}(Y_H \geq c)\ ,\ \mathbb{P}(Y_D \geq c))\\ &= (\text{FFP}(c)\ ,\ \text{FVP}(c))\ ,\end{aligned}$$

lo que concluye la demostración. □

3.1.3. Área bajo la curva

Entre todas las medidas utilizadas para resumir y comparar dos curvas ROC distintas, la más popular es el *área bajo la curva* , AUC por sus siglas en inglés, definida como

$$\text{AUC} = \int_0^1 \text{ROC}(t)\ dt.$$

Como mencionamos anteriormente, esta medida nos permitirá comparar dos curvas ROC distintas, ya que cuánto más cercana a 1 sea la AUC, mejor será la capacidad discriminatoria del biomarcador. El siguiente resultado da una definición equivalente de la AUC.

Proposición 3.3. *El área bajo la curva de una curva ROC coincide con la probabilidad de que el biomarcador arroje un resultado mayor para un individuo enfermo que para uno sano, es decir,* $AUC = \mathbb{P}(Y_D > Y_H)$.

Demostración. Tenemos que

$$\text{AUC} = \int_0^1 \text{ROC}(t)dt = \int_0^1 \left(1 - F_D(F_H^{-1}(1-t))\right) dt.$$

Sea f_H la función de densidad de Y_H. Aplicamos el Teorema de Cambio de Variables con $w = F_H^{-1}(1-t)$ y observando que $F_H(w) = 1 - t$ implica que $f_H(w)\ dw = -dt$, o sea, $dt = -f_H(w)\ dw$, obtenemos

$$\text{AUC} = \int_{+\infty}^{-\infty} (1 - F_D(w))\,(-f_H(w))\,dw = \int_{-\infty}^{+\infty} (1 - F_D(w))\,f_H(w)\ dw\,.$$

Por lo tanto, deducimos que

$$\begin{aligned}\text{AUC} &= \int_{-\infty}^{+\infty} \mathbb{P}\left(Y_D > w\right) f_H(w)\ dw = \int_{-\infty}^{+\infty} \left(\int_w^{+\infty} f_D(v)\ dv\right) f_H(w)\ dw \\ &= \int_{-\infty}^{+\infty} \int_w^{+\infty} f_D(v)\ f_H(w)\ dv\ dw = \iint_{v>w} f_{DH}(v,w)\ dv\ dw \\ &= \mathbb{P}(Y_D > Y_H)\,,\end{aligned}$$

lo que concluye la demostración. □

Dado que esta medida puede interpretarse como la probabilidad de que los resultados del biomarcador en un individuo sano y otro enfermo, elegidos al azar, resulten correctamente ordenados, es razonable que cuanto más cercana a 1 sea el AUC, mejor será la capacidad de diagnóstico del biomarcador.

3.1.4. El modelo binormal

El modelo más simple surge de asumir que tanto F_D como F_H son normales. En ese caso, podemos deducir la forma funcional de la curva ROC, de acuerdo a la siguiente proposición.

Proposición 3.4. *Si* $Y_D \sim \mathcal{N}(\mu_D, \sigma_D^2)$ *e* $Y_H \sim \mathcal{N}(\mu_H, \sigma_H^2)$*, entonces la curva ROC viene dada por la siguiente expresión*

$$ROC(p) = \Phi(a + b\ \Phi^{-1}(p)) \quad \textit{para } 0 \le p \le 1,$$

donde Φ *denota la fución de distribución acumulada de una variable aleatoria normal estándar,*

$$a = \frac{\mu_D - \mu_H}{\sigma_D} \qquad y \qquad b = \frac{\sigma_H}{\sigma_D}\,.$$

Demostración. Dado un valor umbral c, se tiene que

$$\text{FFP}(c) = \mathbb{P}(Y_H \ge c) = 1 - \Phi\left(\frac{c - \mu_H}{\sigma_H}\right) = \Phi\left(\frac{\mu_H - c}{\sigma_H}\right)\,,$$

mientras que

$$\text{FVP}(c) = \mathbb{P}(Y_D \geq c) = 1 - \Phi\left(\frac{c - \mu_D}{\sigma_D}\right) = \Phi\left(\frac{\mu_D - c}{\sigma_D}\right).$$

Luego, de la primera expresión, obtenemos que $c = \mu_H - \sigma_H \ \Phi^{-1}(\text{FFP}(c))$, con lo cual, llamando $t = \text{FFP}(c)$, decucimos

$$\text{ROC}(p) = \text{FVP}(c) = \Phi\left(\frac{\mu_D - \mu_H}{\sigma_D} + \frac{\sigma_H}{\sigma_D}\ \Phi^{-1}(p)\right),$$

concluyendo la prueba. □

Al mismo tiempo, podemos obtener una fórmula cerrada para el AUC de la curva ROC, como muestra el siguiente resultado.

Proposición 3.5. *Si $Y_D \sim \mathcal{N}(\mu_D, \sigma_D^2)$ e $Y_H \sim \mathcal{N}(\mu_H, \sigma_H^2)$, independientes entre sí,entonces*

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

con a y b definidos como en la Proposición 3.4.

Demostración. Reordemos que por la Proposición 3.3, $\text{AUC} = \mathbb{P}(Y_D > Y_H) = \mathbb{P}(Y_D - Y_H > 0)$. Llamando $W = Y_D - Y_H$, obtenemos que $W \sim \mathcal{N}(\mu_D - \mu_H, \sigma_D^2 + \sigma_H^2)$. Por lo tanto,

$$\begin{aligned}\text{AUC} = \mathbb{P}(W > 0) &= 1 - \Phi\left(\frac{-\mu_D + \mu_H}{\sqrt{\sigma_D^2 + \sigma_H^2}}\right) = \Phi\left(\frac{\mu_D - \mu_H}{\sqrt{\sigma_D^2 + \sigma_H^2}}\right) \\ &= \Phi\left(\frac{\frac{\mu_D - \mu_H}{\sigma_D}}{\sqrt{1 + \frac{\sigma_H^2}{\sigma_D^2}}}\right) = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right),\end{aligned}$$

lo que muestra la igualdad deseada. □

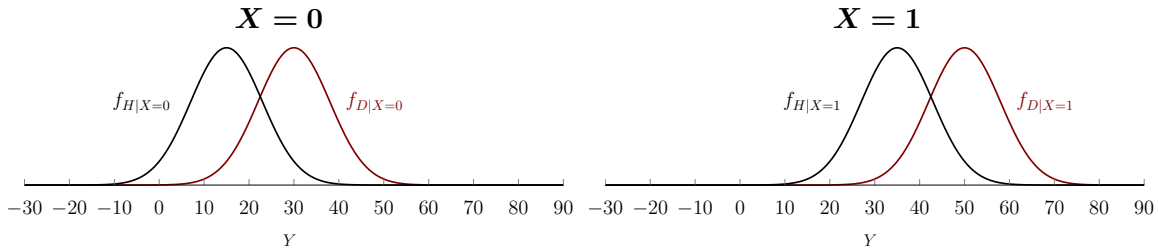
Recordemos que la curva ROC es invariante bajo transformaciones monótonas. Es por eso que el modelo binormal es tan ampliamente utilizado, ya que si el biomarcador Y en cada población no es normal y se asume que existe una transformación monótona de Y tal que el biomarcador bajo dicha transformación sigue una distribución normal, en cada población, la curva ROC del marcador original es la dada por el modelo binormal.

Más propiedades de la curva ROC así como criterios para elegir de forma óptima el punto de corte pueden encontrarse en Pepe (2003).

3.2. Curva ROC con covariables

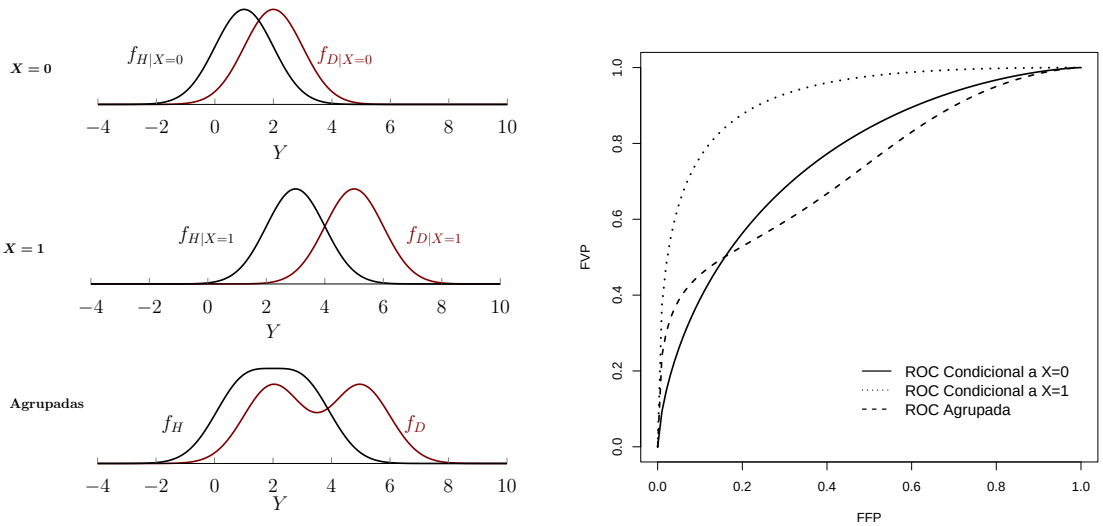
En determinadas situaciones, cuando además del biomarcador Y se cuenta con covariables que proporcionan información adicional acerca de los individuos de la muestra, puede resultar conveniente incorporar dichas covariables a la curva ROC para incrementar su capacidad discriminatoria. Tal como se explica en Pardo-Fernández et al. (2014), hay dos casos distintos en cuanto al efecto de una covariable sobre la curva ROC:

- Cuando afecta sólo el resultado de la prueba, pero no la capacidad discriminatoria del marcador. Esto sucede cuando las distribuciones condicionales están igual de apartadas entre sí que las no condicionales (sin tener en cuenta las covariables). En este caso, las curvas ROC condicionadas a cada valor de la covariable son iguales independientemente del valor condicionante. La siguiente figura ilustra esta situación, donde se toma como ejemplo una covariable binaria X, que podría pensarse, por ejemplo, como el sexo de un individuo:



Bajo estas condiciones, podemos tener distintos casos, dependiendo de la prevalencia de la enfermedad para cada valor de la covariable. En Pardo-Fernández et al. (2014), pueden consultarse algunos ejemplos ilustrativos en los que se puede obervar que si se ignoran las covariables y se construye la curva ROC correspondiente, ésta podría resultar por encima o por debajo de la curva ROC condicional, que será común, en este caso, para ambos valores de la covariable, dando una idea engañosa de la verdadera capacidad discriminatoria del biomarcador.

- Otra situación distinta surge cuando tanto el apartamiento de las distribuciones condicionales como la capacidad de distinguir entre un grupo y otro varían según el valor que tome la covariable. La Figura 3.3 muestra esta situación.



(a) (b)

Figura 3.3: (a) Densidades del biomarcador Y en ambas poblaciones condicionales a cada valor de la covariable (primer y segundo panel), y densidad del biomarcador sin condicionar a la covariable (tercer panel). (b) Curvas ROC condicionales a cada valor de la covariable X y curva ROC *agrupada*, es decir, aquella correspondiente al biomarcador sin condicionar.

En la Figura 3.3 hemos supuesto que el resultado y capacidad discriminatoria de Y depende de la covariable, pero que ésta es independiente del verdadero estado de un

individuo, esto es, que $\mathbb{P}(D = 1|X = 0) = \mathbb{P}(D = 1|X = 1) = 0.5$. En este caso, también se tomó $\mathbb{P}(D = 1) = 0.5$ y $\mathbb{P}(X = 1) = 0.5$. Como podemos observar, las cuvas ROC condicionales son muy distintas a la curva ROC que se obtiene de agrupar la información ignorando la covariable. En particular, el marcador Y distingue mejor entre enfermos y sanos cuando $X = 1$ que cuando $X = 0$, aunque en ambos casos la capacidad discriminatoria es buena. Por el contrario, cuando se agrupan todos los individuos, sin importar el valor de la covariable, la capacidad discriminatoria disminuye notablemente, especialmente para valores de FFP entre 0.2 y 0.8. Es por esto que ignorar la información aportada por las covariables puede llevar a conclusiones erróneas sobre la calidad de la prueba diagnóstica. Más aún, considerar las covariables puede ayudar no sólo a distinguir en qué grupos de individuos la prueba diagnóstica de interés tiene un mejor desempeño, sino también a elegir puntos de corte acordes. En general, cada valor de la covariable, sea ésta continua o discreta, tendrá asociada una curva ROC, la cual definimos a continuación.

3.2.1. Curva ROC condicional ($\mathbf{ROC_x}$)

De ahora en más, supondremos que se cuenta con un vector de covariables y que es el mismo para ambas poblaciones, enfermos y sanos, aunque esto podría no ser el caso, ya que podría ser de interés estudiar covariables específicas de cada población, por ejemplo, estadío de la enfermedad. Como extensión natural de la curva ROC sin covariables, podemos definir la **curva ROC condicional** de la siguiente manera.

Definición 3.2. *Dadas las variables Y_D e Y_H, correspondientes al biomarcador en cada población, y vectores de covariables $\mathbf{X}_D$ y $\mathbf{X}_H$ para enfermos y sanos, respectivamente, la curva ROC condicional a un valor $\mathbf{x}$ de la covariable perteneciente a la intersección de los soportes de $\mathbf{X}_D$ y $\mathbf{X}_H$ viene dada por*

$$ROC_{\mathbf{x}}(p) = 1 - F_D\left(F_H^{-1}(1 - p \mid \mathbf{x}) \mid \mathbf{x}\right), \qquad para\ 0 \leq p \leq 1,$$

donde

$$F_D(\, t \mid \mathbf{x}) = \mathbb{P}\left(Y_D \leq t \mid \mathbf{X}_D = \mathbf{x}\right) \quad y$$
$$F_H(\, t \mid \mathbf{x}) = \mathbb{P}\left(Y_H \leq t \mid \mathbf{X}_H = \mathbf{x}.\right),$$

corresponden a las distribuciones condicionales de Y a cada población.

Al igual que antes, tenemos que el área bajo la curva ROC condicional provee una medida de la capacidad discriminatoria del biomarcador para un valor específico del vector de covariables y se define como

$$\mathrm{AUC}_{\mathbf{x}} = \int_0^1 \mathrm{ROC}_{\mathbf{x}}(p)\, dp\,.$$

3.2.2. Curva ROC ajustada (AROC)

Más allá de la utilidad de las curvas ROC condicionales, resulta de interés tener una medida global que tenga en cuenta la precisión del biomarcador para todos los valores posibles de las covariables. Para ello, se define la **curva ROC ajustada**.

Definición 3.3. *La curva ROC ajustada se define como un promedio de todas las curvas ROC condicionales pesado de acuerdo a la distribución de las covariables en la población enferma*

$$AROC(p) = \int ROC_{\mathbf{x}}(p)\, dH_D(\mathbf{x}),$$

donde $H_D(\mathbf{x}) = \mathbb{P}(\mathbf{X}_D \leq \mathbf{x})$ *es la función de distribución del vector* $\mathbf{X}_D$.

3.3. Estimación mediante la metodología inducida

En esta sección abordaremos el método de regresión inducida, utilizado para estimar la curva ROC condicional, aunque también puede emplearse para construir estimadores de la curva ajustada. Este método consiste en modelar el efecto de las covariables a través de modelos de regresión que vinculen la variable clasificadora (en nuestro caso el biomarcador) con las covariables, en cada población (sanos y enfermos) por separado. Es decir, se asume un modelo de posición y escala para la relación entre el biomarcador y las covariables:

$$\begin{cases} Y_D &= \mu_D(\mathbf{X}_D) + \sigma_D(\mathbf{X}_D)\, \varepsilon_D \\ Y_H &= \mu_H(\mathbf{X}_H) + \sigma_H(\mathbf{X}_H)\, \varepsilon_H\,, \end{cases} \tag{3.1}$$

donde para $j = D, H$, $\mu_j(\mathbf{x}) = \mathbb{E}(Y_j \mid \mathbf{X}_j = \mathbf{x})$, $\sigma_j^2(\mathbf{x}) = \textsc{Var}(Y_j \mid \mathbf{X}_j = \mathbf{x})$ y el error ε_j es independiente de X_j y tiene distribución G_j. Además, para garantizar la identificabilidad de las funciones de regresión y de σ_j, se asume que $\mathbb{E}(\varepsilon_j) = 0$ y $\textsc{Var}(\varepsilon_j) = 1$, respectivamente.

Asumiendo la validez de los modelos dados en (3.1) y utilizando la independencia entre las covariables y los errores, podemos escribir la función de distribución condicional de la población enferma como:

$$\begin{aligned} F_D(y \mid \mathbf{x}) &= \mathbb{P}\left(Y_D \leq y \mid \mathbf{X}_D = \mathbf{x}\right) = \mathbb{P}\left(\mu_D(\mathbf{X}_D) + \sigma_D(\mathbf{X}_D)\, \varepsilon_D \leq y \mid \mathbf{X}_D = \mathbf{x}\right) \\ &= \mathbb{P}\left(\mu_D(\mathbf{x}) + \sigma_D(\mathbf{x})\, \varepsilon_D \leq y \mid \mathbf{X}_D = \mathbf{x}\right) = \mathbb{P}\left(\varepsilon_D \leq \frac{y - \mu_D(\mathbf{x})}{\sigma_D(\mathbf{x})}\right) \\ &= G_D\left(\frac{y - \mu_D(\mathbf{x})}{\sigma_D(\mathbf{x})}\right). \end{aligned}$$

Por otro lado, la función cuantil para la distribución condicional de Y_H puede vincularse con la distribución de los errores mediante:

$$\begin{aligned} F_H^{-1}(p \mid \mathbf{x}) &= \text{ínf}\left\{y \in \mathbb{R} : F_H(y|\mathbf{x}) \geq p\right\} \\ &= \text{ínf}\left\{y \in \mathbb{R} : G_H\left(\frac{y - \mu_H(\mathbf{x})}{\sigma_H(\mathbf{x})}\right) \geq p\right\}. \end{aligned}$$

Por lo tanto, $(F_H^{-1}(p \mid \mathbf{x}) - \mu_H(\mathbf{x}))/\sigma_H(\mathbf{x}) = G_H^{-1}(p)$ de donde se obtiene que $F_H^{-1}(p \mid \mathbf{x}) = \mu_H(\mathbf{x}) + \sigma_H(\mathbf{x})G_H^{-1}(p)$. Luego, para $p \in [0,1]$, y para un valor $\mathbf{x}$ fijo de las covariables, la curva ROC condicional puede expresarse de la siguiente forma:

$$\begin{aligned} \text{ROC}_{\mathbf{x}}(p) &= 1 - F_D\left(F_H^{-1}(1 - p \mid \mathbf{x}) \mid \mathbf{x}\right) \\ &= 1 - F_D\left(\mu_H(\mathbf{x}) + \sigma_H(\mathbf{x})G_H^{-1}(1-p) \mid \mathbf{x}\right) \\ &= 1 - G_D\left(\frac{\mu_H(\mathbf{x}) + \sigma_H(\mathbf{x})G_H^{-1}(1-p) - \mu_D(\mathbf{x})}{\sigma_D(\mathbf{x})}\right) \\ &= 1 - G_D\left(G_H^{-1}(1-p)b(\mathbf{x}) - a(\mathbf{x})\right), \end{aligned} \tag{3.2}$$

donde

$$a(\mathbf{x}) = \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_D(\mathbf{x})} \quad \text{y} \quad b(\mathbf{x}) = \frac{\sigma_H(\mathbf{x})}{\sigma_D(\mathbf{x})}.$$

Esta formulación nos permite expresar la curva ROC condicional en términos de la función de distribución y función cuantil de los errores, que son no condicionales. Esto quiere decir que, en lugar de estimar las distribuciones condicionales de Y_D y Y_H, basta con estimar las distribuciones de los errores en cada población. Esto resulta una gran ventaja, ya que, por lo general, es imposible estimar las distribuciones condicionales a cada valor de las covariables, pues puede haber valores para los que no se cuente con observaciones. Una alternativa para sortear este problema y estimar la función $\mathrm{ROC}_{\mathbf{x}}$ sería estimar las distribuciones condicionales de forma suave mediante núcleos, utilizando observaciones cuyos valores de las covariables sean cercanos a $\mathbf{x}$. Sin embargo, este enfoque también tiene desventajas, como la maldición de la dimensión cuando el número de covariables crece, así como el hecho de que se deben elegir ventanas de suavizado distintas para cada población y se debe contar con suficientes observaciones para que los entornos alrededor del punto $\mathbf{x}$ no resulten vacíos.

La metodología inducida para estimar la curva ROC condicional es muy general, y existen diversos enfoques para estimar las funciones de distribución de los errores en cada población, así como para estimar las medias y varianzas condicionales presentes en el modelo de regresión propuesto. Sin embargo, todas ellas pueden resumirse mediante el siguiente procedimiento.

Dadas muestras independientes de la población sana y enferma, o sea, dados $(y_{j,i}, \mathbf{x}_{j,i})$, $1 \leq i \leq n_j$, independientes con la misma distribución que $(Y_j, \mathbf{X}_j)$, para $j = D, H$, la metodología inducida para estimar la curva ROC condicional puede describirse como sigue:

a) Para $j = D, H$, calcule estimadores $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ de las funciones de posición y escala, respectivamente,

b) Calcule para cada muestra los residuos estandarizados

$$\widehat{\varepsilon}_{j,i} = \frac{y_{j,i} - \widehat{\mu}_j(\mathbf{x}_{j,i})}{\widehat{\sigma}_j(\mathbf{x}_{j,i})} \qquad 1 \leq i \leq n_j \quad j = D, H\,,$$

y un estimador $\widehat{G}_j$ de la distribución de los errores G_j, como por ejemplo, la distribución empírica asociada a $\{\widehat{\varepsilon}_{j,i}\}_{i=1}^{n_j}$, es decir,

$$\widehat{G}_j(s) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{I}_{\{\widehat{\varepsilon}_{j,i} \leq s\}}\,.$$

c) Haga un plug–in de los estimadores calculados en a) y b) en la expresión (3.2) para obtener

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(p) = 1 - \widehat{G}_D\left(\frac{\widehat{\sigma}_H(\mathbf{x})}{\widehat{\sigma}_D(\mathbf{x})}\,\widehat{G}_H^{-1}(1-p) \;-\; \frac{\widehat{\mu}_D(\mathbf{x}) - \widehat{\mu}_H(\mathbf{x})}{\widehat{\sigma}_D(\mathbf{x})}\right).$$

Cabe mencionar que si suponemos el modelo binormal, es decir, si suponemos que $G_j = \Phi$, en b) no hace falta estimar la función de distribución y tomaremos $\widehat{G}_j = \Phi$, es decir, el estimador *plug–in* resulta ser igual a

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(p) = 1 - \Phi\left(\frac{\widehat{\sigma}_H(\mathbf{x})}{\widehat{\sigma}_D(\mathbf{x})}\,\Phi^{-1}(1-p) \;-\; \frac{\widehat{\mu}_D(\mathbf{x}) - \widehat{\mu}_H(\mathbf{x})}{\widehat{\sigma}_D(\mathbf{x})}\right).$$

A lo largo de este trabajo, nos restringiremos al caso en que las covariables son continuas. Además, en la Sección 3.4 adoptaremos un enfoque en particular y consideraremos covariables de naturaleza composicional, la cual ha sido descripta en el Capítulo 2.

3.4. Curva ROC con covariables composicionales

En esta sección, nos concentraremos en la situación en la que las covariables son composicionales, más aún, supondremos que las covariables de interés son las mismas en ambas poblaciones y que $\mathbf{X}_D$ y $\mathbf{X}_H$ tienen soporte en el simplex $\mathcal{S}^m$. De acuerdo a lo descripto en la Sección 3.3 la metodología inducida utiliza un modelo de regresión en cada población como el dado en (3.1) de modo a obtener una expresión para la curva ROC condicional. Más precisamente, para cada $\mathbf{x}$ en el soporte común $\mathcal{S}$ de $\mathbf{X}_D$ y $\mathbf{X}_H$, la curva ROC condicional a $\mathbf{x}$ se obtiene a partir de la expresión dada en (3.2). Por simplicidad, supondremos de ahora en más que el modelo (3.1) es homoscedástico y por lo tanto, obtenemos la expresión

$$\mathrm{ROC}_{\mathbf{x}}(p) = 1 - G_D\left(G_H^{-1}(1-p)\frac{\sigma_H}{\sigma_D} - \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_D}\right), \tag{3.3}$$

donde para $j = D, H$, $\mu_j(\mathbf{x}) = \mathbb{E}(Y_j \mid \mathbf{X}_j = \mathbf{x})$ y el error ε_j es independiente de X_j, con $\mathbb{E}(\varepsilon_j) = 0$, $\textsc{Var}(\varepsilon_j) = 1$ y $\varepsilon_j \sim G_j$.

Para definir el estimador de $\mathrm{ROC}_{\mathbf{x}}$, supondremos que contamos con observaciones independientes tanto del biomarcador como de las covariables en ambas poblaciones: $(y_{D,i}, \mathbf{x}_{D,i}) \in \mathbb{R} \times \mathcal{S}^m$, $1 \leq i \leq n_D$, y $(y_{H,i}, \mathbf{x}_{H,i}) \in \mathbb{R} \times \mathcal{S}^m$, $1 \leq i \leq n_H$. Definiremos dos estimadores, el primero de ellos descripto en la Sección 3.4.1, usa la empíricas de los residuos para estimar G_j y el segundo, definido en la Sección 3.4.2, provee un estimador suave que adapta las ideas dadas en Pulit (2016) al contexto de la ROC condicional. En ambos casos, supondremos que la relación entre el biomarcador y las covariables está dada por un modelo de regresión lineal homoscedástico, es decir, supondremos que

$$Y_D = \beta_{0,D} + \langle \boldsymbol{\beta}_D, \mathbf{X}_D \rangle_a + \sigma_D\, \varepsilon_D \tag{3.4}$$
$$Y_H = \beta_{0,H} + \langle \boldsymbol{\beta}_H, \mathbf{X}_H \rangle_a + \sigma_H\, \varepsilon_H\,, \tag{3.5}$$

donde, como mencionamos anteriormente, ε_j es independiente de X_j, $\mathbb{E}(\varepsilon_j) = 0$, $\textsc{Var}(\varepsilon_j) = 1$ y $\varepsilon_j \sim G_j$. Por lo tanto, $\mu_j(x) = \beta_{0,j} + \langle \boldsymbol{\beta}_j, \mathbf{x} \rangle_a$ y la curva ROC dada en (3.3) se escribe como

$$\begin{aligned}\mathrm{ROC}_{\mathbf{x}}(p) &= 1 - G_D\left(G_H^{-1}(1-p)\frac{\sigma_H}{\sigma_D} - \frac{\beta_{0,D} - \beta_{0,H} + \langle \boldsymbol{\beta}_D, \mathbf{x} \rangle_a - \langle \boldsymbol{\beta}_H, \mathbf{x} \rangle_a}{\sigma_D}\right) \\ &= 1 - G_D\left(G_H^{-1}(1-p)\frac{\sigma_H}{\sigma_D} - \frac{\beta_{0,D} - \beta_{0,H} + \langle \boldsymbol{\beta}_D \ominus \boldsymbol{\beta}_H, \mathbf{x} \rangle_a}{\sigma_D}\right).\end{aligned} \tag{3.6}$$

3.4.1. Estimador semiparamétrico basado en empíricas

El estimador más simple de la curva ROC condicional es aquel que utiliza las distribuciones empíricas de los residuos estandarizados para obtener un estimador de G_j. Para ello, consideremos observaciones independientes $(y_{D,i}, \mathbf{x}_{D,i}) \in \mathbb{R} \times \mathcal{S}^m$, $1 \leq i \leq n_D$, y $(y_{H,i}, \mathbf{x}_{H,i}) \in \mathbb{R} \times \mathcal{S}^m$, $1 \leq i \leq n_H$ que cumplen los modelos (3.4) y (3.5), respectivamente. Los siguientes pasos describen el procedimiento para obtener el estimador semiparamétrico $\widehat{\mathrm{ROC}}_{\mathbf{x}}$.

Paso 1 Como se mencionó en la Sección 3.3, el primer paso consiste en calcular, para $j = D, H$, estimadores $\widehat{\beta}_{0,j}$ y $\widehat{\boldsymbol{\beta}}_j$ de $\beta_{0,j}$ y $\boldsymbol{\beta}_j$. Estos estimadores pueden calcularse utilizando el procedimiento de mínimos cuadrados descripto en la Sección 2.5, es decir, $\widehat{\beta}_{0,j}$ y $\widehat{\boldsymbol{\beta}}_j$ se definen como

$$(\widehat{\beta}_{0,j}, \widehat{\boldsymbol{\beta}}_j) = \underset{(b_0,\mathbf{b})\in\mathbb{R}\times\mathcal{S}^m}{\operatorname{argmin}} \quad \sum_{i=1}^{n_j} \left(y_{j,i} - b_0 - \langle \mathbf{b}, \mathbf{x}_{j,i}\rangle_a\right)^2 .$$

A partir de ellos, se pueden obtener estimadores $\widehat{\sigma}_j$ de los desvíos estándar σ_j, como

$$\widehat{\sigma}_j = \sqrt{\frac{1}{n_j - m}\sum_{i=1}^{n_j}\left(y_{j,i} - \widehat{\beta}_{0,j} - \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x}_{j,i}\rangle_a\right)^2},$$

que resulta insesgado.

Paso 2 Para cada muestra, calcular los residuos estandarizados

$$\widehat{\varepsilon}_{j,i} = \frac{y_{j,i} - \widehat{\beta}_{0,j} - \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x}_{j,i}\rangle_a}{\widehat{\sigma}_j} \qquad 1 \le i \le n_j \quad j = D, H\,,$$

y a partir de ellos obtenga la distribución empírica $\widehat{G}_D$ y la función cuantil empírica mediante $\widehat{G}_H^{-1}$

$$\widehat{G}_D(s) = \frac{1}{n_D}\sum_{i=1}^{n_D} \mathbb{I}_{\{\widehat{\varepsilon}_{D,i}\le s\}}$$
$$\widehat{G}_H^{-1}(p) = \text{ínf}\{s \in \mathbb{R} : \quad \widehat{G}_H(s) \ge p)\} \qquad \text{para} \quad p \in [0,1]$$

Paso 3 Finalmente, dado un valor fijo $\mathbf{x}$ de las covariables podemos hacer un *plug–in* en la expresión (3.6) obteniendo el estimador de $\text{ROC}_{\mathbf{x}}$:

$$\widehat{\text{ROC}}_{\mathbf{x}}(p) = 1 - \widehat{G}_D\left(\widehat{G}_H^{-1}(1-p)\frac{\widehat{\sigma}_H}{\widehat{\sigma}_D} - \frac{\widehat{\beta}_{0,D} - \widehat{\beta}_{0,H} + \langle \widehat{\boldsymbol{\beta}}_D \ominus \widehat{\boldsymbol{\beta}}_H, \mathbf{x}\rangle_a}{\widehat{\sigma}_D}\right). \tag{3.7}$$

Como es de esperar, dado que este estimador se construye principalmente a través de funciones de distribución empíricas, la curva ROC condicional estimada no resulta continua, sino que se trata de una función escalonada. Es por eso que en la siguiente sección proponemos una versión suavizada de este estimador.

3.4.2. Estimador semiparamétrico suavizado: $\widehat{\text{ROC}}_{\mathbf{x},h}$

En su trabajo, Pulit (2016) propone estimar la curva ROC no condicional definiendo la pseudovariable aleatoria $Z = 1 - F_H(Y_D)$, y notando que

$$\mathbb{P}(Z \le p) = \mathbb{P}\left(Y_D > F_H^{-1}(1-p)\right) = 1 - F_D\left(F_H^{-1}(1-p)\right) = \text{ROC}(p)\,,$$

Esto sugiere que se puede estimar la curva $\text{ROC}(p)$ mediante las función de distribución empírica de la pseudovariable Z. La propuesta dada en Pulit (2016) consiste en utilizar

estimadores basados en núcleos para la función distribución de Z. De esta forma, este procedimiento da lugar a una estimación suavizada de la curva ROC. Notemos que es necesario contar con realizaciones de la pseudovariable Z. Para ello, basándonos en observaciones independientes de cada población $y_{j,i}$, $1 \leq i \leq n_j$, puede obtenerse primero la distribución empírica de las observaciones de la población sana, $\widehat{F}_H$, y luego definir $Z_i = 1 - \widehat{F}_H(y_{D,i})$, lo que permitiría definir los estimadores basados en núcleos.

Inspirándonos en el método propuesto por Pulit (2016), intentaremos obtener una estimación suave de $\mathrm{ROC}_{\mathbf{x}}$ definiendo variables aleatorias adecuadas. Para ello, la Proposición 3.6 nos da una forma de expresar a la curva ROC condicional como la función de distribución de cierta variable aleatoria.

Proposición 3.6. *Sea $\mathbf{x} \in \mathcal{S} \subset \mathcal{S}^m$ el soporte común de $\mathbf{X}_j$, que está fijo, y sean $(Y_j, \mathbf{X}_j) \in \mathbb{R} \times \mathcal{S}^m$, $j = D, H$, vectores aleatorios que cumplen el modelo de regresión (3.1) homoscedástico, es decir, $\sigma_j(\mathbf{x}) \equiv \sigma_j$. Entonces, se cumple que*

$$ROC_{\mathbf{x}}(p) = \mathbb{P}(W_{\mathbf{x}} \leq p)\,,$$

donde la variable aleatoria $W_{\mathbf{x}}$ se define como

$$\begin{aligned} W_{\mathbf{x}} &= 1 - G_H\left(\frac{Y_D - \mu_D(\mathbf{X}_D)}{\sigma_H} + \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_H}\right) \\ &= 1 - G_H\left(\varepsilon_D \frac{\sigma_D}{\sigma_H} + \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_H}\right). \end{aligned}$$

Demostración. El resultado se sigue simplemente reescribiendo adecuadamente $\mathbb{P}(W_{\mathbf{x}} \leq p)$. Efectivamente,

$$\begin{aligned} \mathbb{P}(W_{\mathbf{x}} \leq p) &= \mathbb{P}\left(1 - G_H\left(\epsilon_D \frac{\sigma_D}{\sigma_H} + \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_H}\right) \leq p\right) \\ &= \mathbb{P}\left(G_H\left(\epsilon_D \frac{\sigma_D}{\sigma_H} + \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_H}\right) \geq 1 - p\right) \\ &= \mathbb{P}\left(G_H^{-1}(1-p) \leq \epsilon_D \frac{\sigma_D}{\sigma_H} + \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_H}\right) \\ &= \mathbb{P}\left(\varepsilon_D \geq G_H^{-1}(1-p)\frac{\sigma_H}{\sigma_D} - \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_D}\right) \\ &= 1 - G_D\left(G_H^{-1}(1-p)\frac{\sigma_H}{\sigma_D} - \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_D}\right) \\ &= \mathrm{ROC}_{\mathbf{x}}(p)\,, \end{aligned}$$

lo que concluye la demostración. □

La idea será estimar la función de distribución de la variable $W_{\mathbf{x}}$ utilizando un estimador basado en núcleos de la función de distribución, obteniendo así una curva ROC condicional suave. Consideremos entonces observaciones independientes $(y_{D,i}, \mathbf{x}_{D,i}) \in \mathbb{R} \times \mathcal{S}^m$, $1 \leq i \leq n_D$, y $(y_{H,i}, \mathbf{x}_{H,i}) \in \mathbb{R} \times \mathcal{S}^m$, $1 \leq i \leq n_H$ que cumplen los modelos generales homocedásticos, es decir, los dados por (3.1) con $\sigma_j(\mathbf{x}) \equiv \sigma_j$. El estimador se obtiene mediante los siguientes pasos:

Paso S1 Como en la Sección 3.3, el primer paso consiste en calcular, para $j = D, H$, estimadores de las funciones de regresión $\mu_j(\mathbf{x})$ y definir estimadores del desvío estándar σ_j como

$$\widehat{\sigma}_j = \sqrt{\frac{1}{n_j}\sum_{i=1}^{n_j}\left(y_{j,i} - \widehat{\mu}_j(\mathbf{x}_{j,i})\right)^2}\,.$$

Cuando las observaciones cumplen los modelos (3.4) y (3.5), $\mu_j(\mathbf{x}) = \beta_{0,j} + \langle \boldsymbol{\beta}_j, \mathbf{x}\rangle_a$, por lo tanto, podemos considerar los estimadores $\widehat{\beta}_{0,j}$, $\widehat{\boldsymbol{\beta}}_j$ y $\widehat{\sigma}_j$ de $\beta_{0,j}$, $\boldsymbol{\beta}_j$ y σ_j, respectivamente, de mínimos cuadrados descriptos en el **Paso 1** y tomar $\widehat{\mu}_j(\mathbf{x}) = \widehat{\beta}_{0,j} + \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x}\rangle_a$.

Paso S2 Calcular los residuos estandarizados de las observaciones de la muestra de los individuos sanos

$$\widehat{\varepsilon}_{H,i} = \frac{y_{H,i} - \widehat{\mu}_H(\mathbf{x}_{H,i})}{\widehat{\sigma}_H} \qquad 1 \le i \le n_H\,,$$

y a partir de ellos obtener la distribución empírica $\widehat{G}_H$ como

$$\widehat{G}_H(s) = \frac{1}{n_H}\sum_{i=1}^{n_H} \mathbb{I}_{\{\widehat{\varepsilon}_{H,i} \le s\}}\,.$$

Paso S3 Generar n_D pseudo–observaciones de la variable $W_{\mathbf{x}}$ a partir de la distribución empírica de ε_H y la muestra de las covariables en la población enferma, como

$$\begin{aligned}\widehat{W}_{\mathbf{x},i} &= 1 - \widehat{G}_H\left(\frac{y_{D,i} - \widehat{\mu}_D(\mathbf{x}_{D,i})}{\widehat{\sigma}_H} + \frac{\widehat{\mu}_D(\mathbf{x}) - \widehat{\mu}_H(\mathbf{x})}{\widehat{\sigma}_H}\right)\\ &= 1 - G_H\left(\widehat{\varepsilon}_{D,i}\frac{\widehat{\sigma}_D}{\widehat{\sigma}_H} + \frac{\widehat{\mu}_D(\mathbf{x}) - \widehat{\mu}_H(\mathbf{x})}{\widehat{\sigma}_H}\right)\,, \quad 1 \le i \le n_D\,,\end{aligned}$$

donde, para $1 \le i \le n_D$, $\widehat{\varepsilon}_{D,i}$ indica el $i-$ésimo residuo estandarizado de la muestra de los individuos enfermos, o sea,

$$\widehat{\varepsilon}_{D,i} = \frac{y_{D,i} - \widehat{\mu}_D(\mathbf{x}_{D,i})}{\widehat{\sigma}_D}\,.$$

Cuando las observaciones cumplen los modelos (3.4) y (3.5), tenemos que

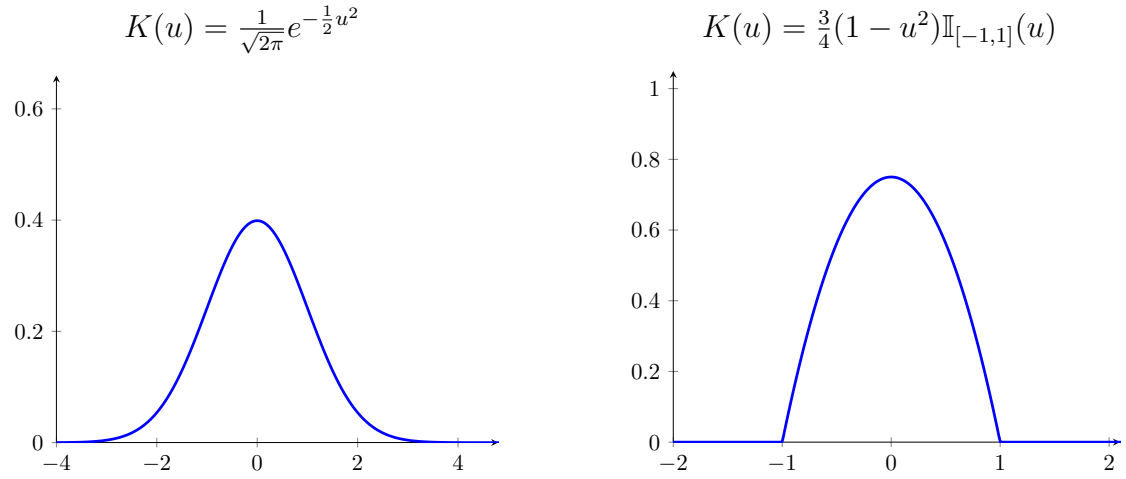
$$\widehat{W}_{\mathbf{x},i} = 1 - \widehat{G}_H\left(\frac{y_{D,i} - \langle \widehat{\boldsymbol{\beta}}_D, \mathbf{x}_{D,i}\rangle_a + \langle \widehat{\boldsymbol{\beta}}_D \ominus \widehat{\boldsymbol{\beta}}_H, \mathbf{x}\rangle_a - \widehat{\beta}_{0,H}}{\widehat{\sigma}_H}\right)\,, \quad 1 \le i \le n_D\,.$$

Paso S4 Sea $K : \mathbb{R} \to \mathbb{R}_+$ una función de densidad par y continua. Indiquemos por $\mathcal{K}(u) = \int_{-\infty}^{u} K(t)\,dt$. El estimador semiparamétrico suavizado viene dado por:

$$\widehat{\mathrm{ROC}}_{\mathbf{x},h}(p) = \frac{1}{n_D}\sum_{i=1}^{n_D}\mathcal{K}\left(\frac{p - \widehat{W}_{\mathbf{x},i}}{h_{n_D}}\right)\,,$$

donde $\{h_{n_D}\}_{n_D\in\mathbb{N}}$ una sucesión decreciente de números positivos, usualmente denominada *ventana*, tal que $\lim_{n_D\to\infty} h_{n_D} = 0$. Observemos que para evitar notación engorrosa hemos denotado $\widehat{\mathrm{ROC}}_{\mathbf{x},h}$ en lugar de $\widehat{\mathrm{ROC}}_{\mathbf{x},h_{n_D}}$.

La función K se denomina *núcleo* y como suponemos que es una densidad, $\mathcal{K}$ es una función de distribución. Algunas opciones frecuentemente utilizadas son el núcleo gaussiano y el de Epanechnikov, cuyos gráficos se muestran en la Figura 3.4 junto con sus expresiones.



(a) Núcleo Gaussiano.

(b) Núcleo de Epanechnikov.

Figura 3.4: Gráfico de dos funciones de núcleos K: Núcleo Gaussiano (panel izquierdo) y núcleo de Epanechnikov (panel derecho).

Vale la pena mencionar que el estimador $\widehat{\mathrm{ROC}}_{\mathbf{x},h}(p)$ coincide con el estudiado en González-Manteiga et al. (2011) para el caso de covariables reales y cuando las funciones de regresión se estiman utilizando medias locales, es decir, cuando se supone un modelo de regresión no-paramétrico.

En el **Paso S4** de la construcción del estimador es necesario elegir una ventana. Para el caso en que no hay covariables, Pulit (2016) propuso la siguiente ventana

$$h^{\star}_{n_D}(p) = c_{n_D} \frac{\sqrt{5p(1-p)}}{\sqrt{2n_D}} \quad \text{donde} \quad c_{n_D} = 1 + 1.8 n_D^{-1/5}. \tag{3.8}$$

que resulta óptima respecto del criterio allí definido. Notemos que $h^{\star}_{n_D}(p)$ no depende de las distribuciones F_D y F_H del biomarcador, a diferencia de otras propuestas previas como las dadas en Peng & Zhou (2004).

Capítulo 4
Consistencia

En este capítulo, estudiaremos las propiedades de consistencia de los estimadores definidos en las Secciones 3.4.1 y 3.4.2, cuando las covariables pertencen al simplex. Consideraremos, por lo tanto, observaciones $(y_{j,i}, \mathbf{x}_{j,i})$, $1 \leq i \leq n_j$, $j = D, H$, independientes tales que $y_{j,i} \in \mathbb{R}$, $\mathbf{x}_{j,i} \in \mathcal{S}^m$, y que cumplen el modelo de regresión lineal homoscedástico en el simplex dado por

$$y_{D,i} = \beta_{0,D} + \langle \boldsymbol{\beta}_D, \mathbf{x}_{D,i} \rangle_a + \sigma_D \, \varepsilon_{D,i} \,, \tag{4.1}$$
$$y_{H,i} = \beta_{0,H} + \langle \boldsymbol{\beta}_H, \mathbf{x}_{H,i} \rangle_a + \sigma_H \, \varepsilon_{H,i} \,, \tag{4.2}$$

donde $\varepsilon_{j,i}$ es independiente de $\mathbf{x}_{H,i}$, $\mathbb{E}(\varepsilon_{j,i}) = 0$, $\textsc{Var}(\varepsilon_{j,i}) = 1$, $\varepsilon_{j,i} \sim G_j$ y donde indicamos por $\langle \cdot, \cdot \rangle_a$ el producto interno en la geometría de Aitchison como se definió en el Capítulo 2.

4.1. Resultados previos e hipótesis

El siguiente resultado que corresponde al Lema S.1 en Bianco et al. (2022), es útil para probar la convergencia uniforme de algunas sucesiones de funciones.

Lema 4.1. *Sea $\mathcal{I} = (u_0, u_1)$ donde u_0 puede ser $-\infty$ y u_1 puede ser $+\infty$. Sean $F_n : \mathcal{I} \to [0,1]$ una sucesión de funciones aleatorias y $F : \mathcal{I} \to [0,1]$ tales que F es continua, $\lim_{t \to u_1} F(t) = 1$ y $\lim_{t \to u_0} F(t) = 0$ y F_n y F son funciones no decrecientes. Luego, si $F_n(t) \xrightarrow{c.s.} F(t)$, para cada $t \in \mathbb{R}$, se tiene que $\|F_n - F\|_\infty \xrightarrow{c.s.} 0$.*

En particular, este lema prueba que las funciones de distribución empíricas convergen casi seguramente uniformemente sobre $\mathbb{R}$.

El siguiente Lema da una versión débil del anterior.

Lema 4.2. *Sea $\mathcal{I} = (u_0, u_1)$ donde u_0 puede ser $-\infty$ y u_1 puede ser $+\infty$. Sean $F_n : \mathcal{I} \to [0,1]$ una sucesión de funciones aleatorias y $F : \mathcal{I} \to [0,1]$ tales que F es continua, $\lim_{t \to u_1} F(t) = 1$ y $\lim_{t \to u_0} F(t) = 0$ y F_n y F son funciones no decrecientes. Luego, si $F_n(t) \xrightarrow{p} F(t)$, para cada $t \in \mathbb{R}$, se tiene que $\|F_n - F\|_\infty \xrightarrow{p} 0$.*

Demostración. Sea $\eta > 0$, queremos ver que $\lim_{n \to \infty} \mathbb{P}\left(\|F_n - F\|_\infty > \eta\right) = 0$. Como en el Lema S.1 de Bianco et al. (2022), sean $u_0 < a$ y $b < u_1$ tales que $F(a) < \eta/2$ y $F(b) > 1 - \eta/2$. Como F es uniformemente continua en $[a, b]$, existe $\delta > 0$ tal que

$$|t - s| < \delta, t, s \in [a, b] \Rightarrow |F(t) - F(s)| < \eta/2$$

Sean $a = a_0 < a_2 < \ldots < a_k = b$, una grilla de puntos tales que $a_j - a_{j-1} < \delta$, $1 \leq j \leq k$.

Entonces, para cualquier $t < a$, $F_n(t) - F(t) \leq F_n(a) \leq F_n(a) - F(a) + F(a) \leq |F_n(a) - F(a)| + \eta/2$, mientras que $F(t) - F_n(t) \leq F(a) < \eta/2$, de donde

$$\sup_{t<a} |F_n(t) - F(t)| \leq |F_n(a) - F(a)| + \frac{\eta}{2}\,. \tag{4.3}$$

Análogamente,

$$\sup_{t>b} |F_n(t) - F(t)| \leq |F_n(b) - F(b)| + \frac{\eta}{2}\,. \tag{4.4}$$

Finalmente, dado $t \in [a, b]$, existe $1 \leq j \leq k$ tal que $t \in [a_{j-1}, a_j]$, con lo cual

$$\begin{aligned} F_n(t) - F(t) &\leq F_n(a_j) - F(a_{j-1}) \leq F_n(a_j) - F(a_j) + F(a_j) - F(a_{j-1}) \\ &\leq \frac{\eta}{2} + \max_{1\leq j\leq k} |F_n(a_j) - F(a_j)|\,, \end{aligned}$$

mientras que

$$\begin{aligned} F(t) - F_n(t) &\leq F(a_j) - F_n(a_{j-1}) \leq F(a_j) - F(a_{j-1}) + F(a_{j-1}) - F_n(a_{j-1}) \\ &\leq \frac{\eta}{2} + \max_{1\leq j\leq k} |F_n(a_j) - F(a_j)|\,. \end{aligned} \tag{4.5}$$

Por lo tanto, de (4.3), (4.4) y (4.5) concluímos que

$$\|F_n - F\|_\infty < \frac{\eta}{2} + \max_{0\leq j\leq k} |F_n(a_j) - F(a_j)|\,. \tag{4.6}$$

Como para cada $0 \leq j \leq k$, $F_n(a_j) - F(a_j) \xrightarrow{p} 0$, tenemos que $\max_{0\leq j\leq k} |F_n(a_j) - F(a_j)| \xrightarrow{p} 0$, por lo tanto,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{0\leq j\leq k} |F_n(a_j) - F(a_j)| > \frac{\eta}{2}\right) = 0\,.$$

Sean $\nu > 0$ y n_0 tales que para todo $n \geq n_0$,

$$\mathbb{P}\left(\max_{0\leq j\leq k} |F_n(a_j) - F(a_j)| > \frac{\eta}{2}\right) < \nu\,,$$

o equivalentemente

$$\mathbb{P}\left(\max_{0\leq j\leq k} |F_n(a_j) - F(a_j)| \leq \frac{\eta}{2}\right) > 1 - \nu\,,$$

Luego, usando (4.6), obtenemos que para todo $n \geq n_0$,

$$\mathbb{P}\left(\|F_n - F\|_\infty < \eta\right) > \mathbb{P}\left(\max_{0\leq j\leq k} |F_n(a_j) - F(a_j)| \leq \frac{\eta}{2}\right) > 1 - \nu\,,$$

lo que implica que $\|F_n - F\|_\infty \xrightarrow{p} 0$. □

Para probar la consistencia de los estimadores propuestos necesitaremos las siguientes hipótesis sobre las funciones de distribución de los errores, así como sobre el núcleo y la ventana elegidos.

H1 $G_H : \mathbb{R} \to (0,1)$ tiene asociada una densidad acotada g_H tal que $g_H(y) > 0,\ \forall\, y \in \mathbb{R}$.

H2 (a) $G_D : \mathbb{R} \to (0,1)$ es continua.

(b) G_D tiene asociada una densidad acotada g_D.

H3 $\widehat{\beta}_{0,j} \xrightarrow{c.s.} \beta_{0,j}$, $\|\widehat{\boldsymbol{\beta}}_j \ominus \boldsymbol{\beta}_j\|_a \xrightarrow{c.s.} 0$, $\widehat{\sigma}_j \xrightarrow{c.s.} \sigma_j$, para $j = D, H$.

H4 K es una función no–negativa, par, acotada, continuamente diferenciable con derivada acotada y $\int K(t)\, dt = 1$.

H5 $n_D/(n_D + n_H) \longrightarrow \tau$ con $0 < \tau < 1$.

H6 $h_{n_D} \longrightarrow 0$ y $n_D\, h_{n_D}^2 \longrightarrow \infty$.

H7 (a) $n_j{}^{1/2}\,(\widehat{\beta}_{0,j} - \beta_{0,j}) = O_{\mathbb{P}}(1)$, $n_j{}^{1/2}\,\|\widehat{\boldsymbol{\beta}}_j \ominus \boldsymbol{\beta}_j\|_a = O_{\mathbb{P}}(1)$, $n_j{}^{1/2}\,(\widehat{\sigma}_j - \sigma_j) = O_{\mathbb{P}}(1)$.

(b) $\mathbb{E}\|\mathbf{x}_{H,1}\|_a < \infty$.

(c) $r_H(u) = u\, g_H(u)$ es acotada.

Vale la pena mencionar que bajo **H1**, la condición **H7**(c) se cumple si el $\lim_{u\to\pm\infty} r_H(u)$ existe. Por otra parte, si el límite existe es 0, ya que $\mathbb{E}(\varepsilon_{H,1}) = 0$ implica que $\int |r_H(u)|du < \infty$.

4.2. Consistencia fuerte uniforme de $\widehat{\mathrm{ROC}}_{\mathbf{x}}$

El Teorema 4.1 muestra que el estimador semipararamétrico definido en el **Paso 3** dado por

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(p) = 1 - \widehat{G}_D\left(\widehat{G}_H^{-1}(1-p)\frac{\widehat{\sigma}_H}{\widehat{\sigma}_D} + \frac{\widehat{\beta}_{0,H} + \langle\widehat{\boldsymbol{\beta}}_H, \mathbf{x}\rangle_a - \beta_{0,D} - \langle\widehat{\boldsymbol{\beta}}_D, \mathbf{x}\rangle_a}{\widehat{\sigma}_D}\right).$$

es uniformemente consistente en p.

Teorema 4.1. *Sean $(y_{j,i}, \mathbf{x}_{j,i})$, $1 \le i \le n_j$, $j = D, H$, observaciones independientes que satisfacen los modelos de regresión dados en (4.1) y (4.2) y sean $\widehat{\beta}_{0,j}$, $\widehat{\boldsymbol{\beta}}_j$ y $\widehat{\sigma}_j$ estimadores de $\beta_{0,j}$, $\boldsymbol{\beta}_j$ y σ_j, respectivamente. Supongamos que valen las hipótesis **H1**, **H2**(a) y **H3**. Entonces, para cada $\mathbf{x} \in \mathcal{S}^m$,*

$$\sup_{0<p<1}\left|\widehat{ROC}_{\mathbf{x}}(p) - ROC_{\mathbf{x}}(p)\right| \xrightarrow{c.s.} 0\,.$$

Para probar el teorema, necesitaremos el siguiente resultado, que prueba la convergencia uniforme casi segura de las funciones de distribución empíricas de los residuos consideradas en la construcción de los estimadores presentadose en las Secciones 3.4.1 y 3.4.2 .

Lema 4.3. *Sean $(y_{j,i}, \mathbf{x}_{j,i})$, $1 \le i \le n_j$, $j = D, H$, observaciones independientes que satisfacen los modelos de regresión dados en (4.1) y (4.2). Consideremos estimadores $\widehat{\beta}_{0,j}$, $\widehat{\boldsymbol{\beta}}_j$ y $\widehat{\sigma}_j$ de $\beta_{0,j}$, $\boldsymbol{\beta}_j$ y σ_j, respectivamente que cumplen **H3**. Para $j = D, H$, definamos los residuos estandarizados*

$$\widehat{\varepsilon}_{j,i} = \frac{y_{j,i} - \widehat{\beta}_{0,j} - \langle\widehat{\boldsymbol{\beta}}_j, \mathbf{x}_{j,i}\rangle_a}{\widehat{\sigma}_j} \qquad 1 \le i \le n_j\,.$$

Entonces, si G_j es continua, para $j = D, H$, tenemos que $\|\widehat{G}_j - G_j\|_\infty \xrightarrow{c.s.} 0$ donde

$$\widehat{G}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{I}_{\{\widehat{\varepsilon}_{j,i} \leq t\}} \, .$$

Demostración. Por el Lema 4.1, basta ver que $\widehat{G}_j(t) \xrightarrow{c.s.} G_j(t)$ para cada $t \in \mathbb{R}$. En primer lugar, observemos que dados $\mathbf{x}, \boldsymbol{\beta} \in \mathcal{S}^m$, $\langle \mathbf{x}, \boldsymbol{\beta} \rangle_a = ilr(\boldsymbol{\beta})^{\mathrm{T}} ilr(\mathbf{x})$ y además $ilr(\mathbf{x} \oplus \boldsymbol{\beta}) = ilr(\mathbf{x}) + ilr(\boldsymbol{\beta})$, mientras que $ilr(\alpha \odot \mathbf{x}) = \alpha \, ilr(\mathbf{x})$. Por lo tanto, usando que $y_{j,i} = \beta_{0,j} + \langle \boldsymbol{\beta}_j, \mathbf{x}_{j,i} \rangle_a + u_{j,i}$, donde $u_{j,i} = \sigma_j \, \varepsilon_{j,i}$, obtenemos que

$$\begin{aligned} \widehat{\varepsilon}_{j,i} &= \frac{y_{j,i} - \widehat{\beta}_{0,j} - \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x}_{j,i} \rangle_a}{\widehat{\sigma}_j} = \frac{u_{j,i} + \beta_{0,j} + \langle \boldsymbol{\beta}_j, \mathbf{x}_{j,i} \rangle_a - \widehat{\beta}_{0,j} - \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x}_{j,i} \rangle_a}{\widehat{\sigma}_j} \\ &= \frac{u_{j,i} + \beta_{0,j} - \widehat{\beta}_{0,j} + ilr(\boldsymbol{\beta}_j \ominus \widehat{\boldsymbol{\beta}}_j)^{\mathrm{T}} ilr(\mathbf{x}_{j,i})}{\widehat{\sigma}_j} \, . \end{aligned}$$

Por lo tanto, si llamamos $\widehat{\boldsymbol{\theta}} = ilr(\widehat{\boldsymbol{\beta}}_j \ominus \boldsymbol{\beta}_j)$ y $\widehat{\alpha}_0 = \widehat{\beta}_{0,j} - \beta_{0,j}$ (notar que obviamos los subíndices para simplificar la notación) tenemos que

$$\widehat{G}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{I}_{\left\{u_{j,i} - (\widehat{\beta}_{0,j} - \beta_{0,j}) - ilr(\widehat{\boldsymbol{\beta}}_j \ominus \boldsymbol{\beta}_j)^{\mathrm{T}} ilr(\mathbf{x}_{j,i}) \leq \widehat{\sigma}_j \, t\right\}} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{I}_{\left\{u_{j,i} - \widehat{\alpha}_{0,j} - \widehat{\boldsymbol{\theta}}_j^{\mathrm{T}} ilr(\mathbf{x}_{j,i}) \leq \widehat{\sigma}_j \, t\right\}} \, .$$

Sea $t_0 \in \mathbb{R}$ fijo. Como en el Lema 1 de Bianco et al. (2022), consideremos la familia de funciones:

$$\mathcal{F} = \left\{ f_{\kappa,\alpha_0,\boldsymbol{\theta}}(u, \mathbf{x}) = \mathbb{I}_{\left\{u - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) \leq \kappa \, t_0\right\}} \, : \, (\kappa, \alpha_0, \boldsymbol{\theta}) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R}^{m-1} \right\} .$$

Observemos que la clase $\mathcal{F}$ tiene como envolvente a $F \equiv 1$ y que $f_{\kappa,\alpha_0,\boldsymbol{\theta}}$ puede escribirse como

$$f_{\kappa,\alpha_0,\boldsymbol{\theta}}(u, \mathbf{x}) = \mathbb{I}_{\left\{u - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) \leq \kappa \, t_0\right\}} = \mathbb{I}_{\left\{\kappa^{-1} u - \kappa^{-1} \alpha_0 - \kappa^{-1} \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) - t_0 \leq 0\right\}} = \mathbb{I}_{A_{\kappa^{-1}, \alpha_0 \, \kappa^{-1}, \boldsymbol{\theta} \, \kappa^{-1}}} \, ,$$

donde $A_{s,\alpha_0,\boldsymbol{\theta}} = \left\{(u, \mathbf{x}) \in \mathbb{R} \times \mathcal{S}^m \, : \, s \, u - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) - t_0 \leq 0 \right\}$. Definamos la clase de conjuntos:

$$\mathcal{A} = \left\{ A_{s,\alpha_0,\boldsymbol{\theta}} : (s, \alpha_0, \boldsymbol{\theta}) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R}^{m-1} \right\} .$$

Dado que la familia de funciones

$$\left\{ g(u, \mathbf{x}) = su - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) - t_0 \leq 0 \ \text{ para } (s, \alpha_0, \boldsymbol{\theta}) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R}^{m-1} \right\}$$

es una familia finito-dimensional de dimensión $m + 1$, aplicando los Lemas 9.6, 9.8 y 9.9 de Kosorok (2008), obtenemos que $\mathcal{A}$ es una clase VC de conjuntos de índice a lo sumo $m + 3$. Usando de nuevo el Lema 9.8 de Kosorok (2008), tenemos que la familia $\mathcal{F}$ es una clase VC de índice $V(\mathcal{F})$ a lo sumo $m + 3$.

Luego, por el Teorema 2.6.7 de van der Vaart & Wellner (1996), existe una constante universal K tal que para toda medida de probabilidad $\mathbb{Q}$, se tiene que para todo $0 < \varepsilon < 1$

$$N\left(\varepsilon, \mathcal{F}, L_1(\mathbb{Q})\right) \leq K \, V(\mathcal{F}) \, (16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{2(V(\mathcal{F})-1)} \, .$$

Como $V(\mathcal{F}) \leq m+3$, la desigualdad anterior implica que

$$\sup_{\mathbb{Q}} N\left(\varepsilon, \mathcal{F}, L_1(\mathbb{Q})\right) < \infty\,,$$

donde el supremo se toma sobre todas las medidas de probabilidad discretas. Aplicando ahora el Teorema 2.4.3 de van der Vaart & Wellner (1996) o el Teorema 2.4 de de Kosorok (2008), obtenemos que $\mathcal{F}$ es una familia de funciones Glivenko–Cantelli, es decir, satisface

$$\sup_{f\in\mathcal{F}} \left|\mathbb{P}_{n_j} f - \mathbb{P}_j f\right| \xrightarrow{c.s.} 0\,, \tag{4.7}$$

donde utilizamos la notación estándar de procesos empíricos, es decir,

$$\mathbb{P}_{n_j} f = \frac{1}{n_j}\sum_{i=1}^{n_j} f(u_{j,i}, \mathbf{x}_{j,i}) \quad \text{y} \quad \mathbb{P}_j f = \mathbb{E} f(u_{j,1}, \mathbf{x}_{j,1})\,.$$

Definamos la función $\mathbb{L}^j_{t_0}(\alpha_0, \boldsymbol{\theta}, \sigma) = \mathbb{P}(u_{j,1} - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{j,1}) \leq \sigma\, t_0)$. De (4.7) se sigue que

$$\sup_{(s,\alpha_0,\boldsymbol{\theta})\in\mathbb{R}_{>0}\times\mathbb{R}\times\mathbb{R}^{m+1}} \left|\frac{1}{n_j}\sum_{i=1}^{n_j} \mathbb{I}_{\left\{u_{j,i}-\alpha_0-\boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{j,i})\leq s\, t_0\right\}} - \mathbb{L}_{t_0}(\alpha_0, \boldsymbol{\theta}, \sigma)\right| \xrightarrow{c.s.} 0\,.$$

Por lo tanto, deducimos que

$$\left|\frac{1}{n_j}\sum_{i=1}^{n_j} \mathbb{I}_{\left\{u_{j,i}-\widehat{\alpha}_0-\widehat{\boldsymbol{\theta}}^{\mathrm{T}} ilr(\mathbf{x}_{j,i})\leq \widehat{\sigma}_j\, t_0\right\}} - \mathbb{L}_{t_0}(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_j)\right| \xrightarrow{c.s.} 0\,,$$

o equivalentemente $\left|\widehat{G}_j(t_0) - \mathbb{L}_{t_0}(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}_j, \widehat{\sigma}_j)\right| \xrightarrow{c.s.} 0$. Por lo tanto, para probar que $\widehat{G}_j(t_0) - G_j(t_0) \xrightarrow{c.s.} 0$, faltaría mostrar que

$$\mathbb{L}_{t_0}(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}_j, \widehat{\sigma}_j) - G_j(t_0) \xrightarrow{c.s.} 0\,. \tag{4.8}$$

Para ello observemos que

$$\begin{aligned}
\mathbb{L}^j_{t_0}(\alpha_0, \boldsymbol{\theta}, \sigma) &= \mathbb{P}(u_{j,1} - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{j,1}) \leq \sigma\, t_0) \\
&= \mathbb{E}\left\{\mathbb{P}(\sigma_j\, \varepsilon_{j,1} - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{j,1}) \leq \sigma\, t_0 | \mathbf{x}_{j,1})\right\} \\
&= \mathbb{E}\left\{G_j\left(\frac{\alpha_0 + \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{j,1}) + \sigma\, t_0}{\sigma_j}\right)\right\}\,.
\end{aligned}$$

La consistencia dada en **H3** implica que, $\widehat{\alpha}_0 \xrightarrow{c.s.} 0$ y $\widehat{\boldsymbol{\theta}} \xrightarrow{c.s.} \mathbf{0}_{m-1}$, luego para todo $\mathbf{x} \in \mathcal{S}^m$

$$\frac{\widehat{\alpha}_0 + \widehat{\boldsymbol{\theta}}^{\mathrm{T}} ilr(\mathbf{x}_{j,1}) + \widehat{\sigma}_j\, t_0}{\sigma_j} \xrightarrow{c.s.} t_0\,,$$

de donde usando la continuidad de G_j concluímos que

$$G_j\left(\frac{\widehat{\alpha}_0 + \widehat{\boldsymbol{\theta}}^{\mathrm{T}} ilr(\mathbf{x}_{j,1}) + \widehat{\sigma}_j\, t_0}{\sigma_j}\right) \xrightarrow{c.s.} G_j\left(t_0\right)\,.$$

El Teorema de Convergencia Mayorada permite entonces deducir (4.8), lo que concluye la demostración. □

Demostración del Teorema 4.1. Notemos que por **H1**, el funcional cuantil es continuo, por lo tanto, del Lema 4.3 y del Lema 3 de Jokiel-Rokita & Pulit (2013), deducimos que $\widehat{G}_H^{-1}(p) \xrightarrow{c.s.} G_H^{-1}(p)$ para cada $0 < p < 1$.

Para simplificar la notación, de ahora en más indicaremos por $\mu_j(\mathbf{x}) = \beta_{0,j} + \langle \boldsymbol{\beta}_j, \mathbf{x} \rangle_a$ y $\widehat{\mu}_j(\mathbf{x}) = \widehat{\beta}_{0,j} + \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x} \rangle_a$, para $j = D, H$, y

$$\widehat{\Delta}(\mathbf{x}, p) = \widehat{G}_H^{-1}(1-p)\frac{\widehat{\sigma}_H}{\widehat{\sigma}_D} + \frac{\widehat{\mu}_H(\mathbf{x}) - \widehat{\mu}_D(\mathbf{x})}{\widehat{\sigma}_D}\,,$$
$$\Delta(\mathbf{x}, p) = G_H^{-1}(1-p)\frac{\sigma_H}{\sigma_D} + \frac{\mu_H(\mathbf{x}) - \mu_D(\mathbf{x})}{\sigma_D}\,.$$

Luego, la consistencia dada en **H3** y el hecho de que $\widehat{G}_H^{-1}(p) \xrightarrow{c.s.} G_H^{-1}(p)$, implican que $\widehat{\Delta}(\mathbf{x}, p) \xrightarrow{c.s.} \Delta(\mathbf{x}, p)$, para cada $\mathbf{x} \in \mathcal{S}^m$ y $0 < p < 1$.

Fijemos $\mathbf{x} \in \mathcal{S}^m$ y $0 < p < 1$, tenemos las desigualdades

$$\begin{aligned}
|\widehat{\mathrm{ROC}}_{\mathbf{x}}(p) - \mathrm{ROC}_{\mathbf{x}}(p)| &= \left|\widehat{G}_D\left(\widehat{\Delta}(\mathbf{x}, p)\right) - G_D\left(\Delta(\mathbf{x}, p)\right)\right| \\
&\leq \left|\widehat{G}_D\left(\widehat{\Delta}(\mathbf{x}, p)\right) - G_D\left(\widehat{\Delta}(\mathbf{x}, p)\right)\right| + \left|G_D\left(\widehat{\Delta}(\mathbf{x}, p)\right) - G_D\left(\Delta(\mathbf{x}, p)\right)\right| \\
&\leq \left\|\widehat{G}_D - G_D\right\|_\infty + \left|G_D\left(\widehat{\Delta}(\mathbf{x}, p)\right) - G_D\left(\Delta(\mathbf{x}, p)\right)\right|\,.
\end{aligned}$$

Por un lado, el primer sumando en el término de la derecha de la desigualdad tiende a cero c.s., por el Lema 4.3. Por otro lado, el hecho que $\widehat{\Delta}(\mathbf{x}, p) \xrightarrow{c.s.} \Delta(\mathbf{x}, p)$ y la continuidad de G_D, implican que el segundo sumando tiende a cero. Por lo tanto, $|\widehat{\mathrm{ROC}}_{\mathbf{x}}(p) - \mathrm{ROC}_{\mathbf{x}}(p)| \xrightarrow{c.s.} 0$ para todo $\mathbf{x} \in \mathcal{S}^m$ y $0 < p < 1$.

Aplicando ahora el Lema 4.1 a las funciones $F_n = \widehat{\mathrm{ROC}}_{\mathbf{x}}$ y $F = \mathrm{ROC}_{\mathbf{x}}$, se deduce que $\sup_{0<p<1} |\widehat{\mathrm{ROC}}_{\mathbf{x}}(p) - \mathrm{ROC}_{\mathbf{x}}(p)| \xrightarrow{c.s.} 0$. □

4.3. Consistencia débil uniforme de $\widehat{\mathrm{ROC}}_{\mathbf{x},h}$

Recordemos que hemos definido

$$a(\mathbf{x}) = \frac{\mu_D(\mathbf{x}) - \mu_H(\mathbf{x})}{\sigma_D}\,, \qquad b = \frac{\sigma_H}{\sigma_D} \qquad \text{y} \qquad W_{\mathbf{x},i} = 1 - G_H\left(\frac{\varepsilon_{D,i} + a(\mathbf{x})}{b}\right)\,, \tag{4.9}$$
$$\widehat{a}(\mathbf{x}) = \frac{\widehat{\mu}_D(\mathbf{x}) - \widehat{\mu}_H(\mathbf{x})}{\widehat{\sigma}_D}\,, \qquad \widehat{b} = \frac{\widehat{\sigma}_H}{\widehat{\sigma}_D} \qquad \text{y} \qquad \widehat{W}_{\mathbf{x},i} = 1 - \widehat{G}_H\left(\frac{\widehat{\varepsilon}_{D,i} + \widehat{a}(\mathbf{x})}{\widehat{b}}\right)\,,$$

donde, $\mu_j(\mathbf{x}) = \beta_{0,j} + \langle \boldsymbol{\beta}_j, \mathbf{x} \rangle_a$ y $\widehat{\mu}_j(\mathbf{x}) = \widehat{\beta}_{0,j} + \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x} \rangle_a$, para $j = D, H$, y para $1 \leq i \leq n_D$,

$$\widehat{\varepsilon}_{D,i} = \frac{y_{D,i} - \widehat{\mu}_D(\mathbf{x}_{D,i})}{\widehat{\sigma}_D} = \frac{\mu_D(\mathbf{x}_{D,i}) - \widehat{\mu}_D(\mathbf{x}_{D,i})}{\widehat{\sigma}_D} + \frac{\sigma_D}{\widehat{\sigma}_D}\,\varepsilon_{D,i}\,.$$

De esta forma, como se mostró en la Proposición 3.6, si $W_{\mathbf{x}} = 1 - G_H\left((\varepsilon_D + a(\mathbf{x}))/b\right) \sim W_{\mathbf{x},1}$ la curva ROC condicional se expresa como

$$\mathrm{ROC}_{\mathbf{x}}(p) = \mathbb{P}(W_{\mathbf{x}} \leq p) = F_{W_{\mathbf{x}}}(p)\,,$$

y el estimador semiparamétrico suavizado dado en el **Paso S4** se define como

$$\widehat{\mathrm{ROC}}_{\mathbf{x},h}(p) = \frac{1}{n_D}\sum_{i=1}^{n_D} \mathcal{K}\left(\frac{p - \widehat{W}_{\mathbf{x},i}}{h_{n_D}}\right),$$

donde $\mathcal{K}(u) = \int_{-\infty}^{u} K(t)\,dt$ y hemos indicado $\widehat{\mathrm{ROC}}_{\mathbf{x},h}$ en lugar de $\widehat{\mathrm{ROC}}_{\mathbf{x},h_{n_D}}$ por simplicidad de notación. Para este estimador el Teorema 4.2 da condiciones para la consistencia débil, para cada $\mathbf{x} \in \mathcal{S}^m$, uniformemente en $0 < p < 1$.

Observemos que por **H1** y **H2**(b), $\mathrm{ROC}_{\mathbf{x}}(p)$ es derivable respecto de p y por lo tanto, $F_{W_{\mathbf{x}}}$ tiene densidad $f_{W_{\mathbf{x}}}(p) = \mathrm{ROC}'_{\mathbf{x}}(p)$.

Teorema 4.2. *Sean $(y_{j,i}, \mathbf{x}_{j,i})$, $1 \le i \le n_j$, $j = D, H$, observaciones independientes que satisfacen los modelos de regresión en (4.1) y (4.2). Sean $\widehat{\beta}_{0,j}$, $\widehat{\boldsymbol{\beta}}_j$ y $\widehat{\sigma}_j$ estimadores de $\beta_{0,j}$, $\boldsymbol{\beta}_j$ y σ_j, respectivamente. Supongamos que valen las hipótesis **H1**, **H2**(b), **H4** a **H7**. Entonces, para cada $\mathbf{x} \in \mathcal{S}^m$, $\sup_{0<p<1}\left|\widehat{ROC}_{\mathbf{x},h}(p) - ROC_{\mathbf{x}}(p)\right| \xrightarrow{p} 0$.*

Para probar el Teorema 4.2, necesitaremos algunos resultados auxiliares.

Lema 4.4. *Sean $(y_{H,i}, \mathbf{x}_{H,i}^{\mathrm{T}})$, $1 \le i \le n_H$, $y_{H,i} \in \mathbb{R}$, $\mathbf{x}_{H,i} \in \mathcal{S}^m$, observaciones independientes que cumplen el modelo (4.2). Sean $\widehat{\beta}_{0,H}, \widehat{\boldsymbol{\beta}}_H$ y $\widehat{\sigma}_H$ estimadores de $\beta_{0,H}, \boldsymbol{\beta}_H$ y σ_H, respectivamente, que cumplen **H7**. Entonces, bajo **H1**, tenemos que*

$$\sqrt{n_H}\|\widehat{G}_H - G_H\|_\infty = O_{\mathbb{P}}(1).$$

Demostración. Utilizaremos argumentos similares a los empleados en la demostración del Lema 4.3. Allí, vimos que, como $y_{H,i} = \beta_{0,H} + \langle \boldsymbol{\beta}_H, \mathbf{x}_{H,i}\rangle_a + u_{H,i}$, donde $u_{H,i} = \sigma_H\,\varepsilon_{H,i}$, se tenía que

$$\begin{aligned}\widehat{\varepsilon}_{H,i} &= \frac{y_{H,i} - \widehat{\beta}_{0,H} - \langle \widehat{\boldsymbol{\beta}}_H, \mathbf{x}_{H,i}\rangle_a}{\widehat{\sigma}_H} = \frac{u_{H,i} + \beta_{0,H} + \langle \boldsymbol{\beta}_H, \mathbf{x}_{H,i}\rangle_a - \widehat{\beta}_{0,H} - \langle \widehat{\boldsymbol{\beta}}_H, \mathbf{x}_{H,i}\rangle_a}{\widehat{\sigma}_H} \\ &= \frac{u_{H,i} + \beta_{0,H} - \widehat{\beta}_{0,H} + ilr(\boldsymbol{\beta}_H \ominus \widehat{\boldsymbol{\beta}}_H)^{\mathrm{T}} ilr(\mathbf{x}_{H,i})}{\widehat{\sigma}_H}.\end{aligned}$$

Por lo tanto, si llamamos $\widehat{\boldsymbol{\theta}}_H = ilr(\widehat{\boldsymbol{\beta}}_H \ominus \boldsymbol{\beta}_H)$ y $\widehat{\alpha}_{0,H} = \widehat{\beta}_{0,H} - \beta_{0,H}$ tenemos que

$$\widehat{G}_H(t) = \frac{1}{n_H}\sum_{i=1}^{n_H} \mathbb{I}_{\left\{u_{H,i} - (\widehat{\beta}_{0,H} - \beta_{0,H}) - ilr(\widehat{\boldsymbol{\beta}}_H \ominus \boldsymbol{\beta}_H)^{\mathrm{T}} ilr(\mathbf{x}_{H,i}) \le \widehat{\sigma}_H\, t\right\}} = \frac{1}{n_H}\sum_{i=1}^{n_H} \mathbb{I}_{\left\{u_{H,i} - \widehat{\alpha}_{0,H} - \widehat{\boldsymbol{\theta}}_H^{\mathrm{T}} ilr(\mathbf{x}_{H,i}) \le \widehat{\sigma}_H\, t\right\}}.$$

Como en el Lema 1 de Bianco et al. (2022), consideremos la familia de funciones:

$$\mathcal{F} = \left\{ f_{\kappa,\alpha_0,\boldsymbol{\theta}}(u, \mathbf{x}) = \mathbb{I}_{\left\{u - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) \le \kappa\, t\right\}} \,:\, (\kappa, t, \alpha_0, \boldsymbol{\theta}) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{m-1} \right\}.$$

Observemos que la clase $\mathcal{F}$ tiene como envolvente a $F \equiv 1$ y que $f_{\kappa,\alpha_0,\boldsymbol{\theta}}$ puede escribirse como

$$f_{\kappa,\alpha_0,\boldsymbol{\theta}}(u, \mathbf{x}) = \mathbb{I}_{\left\{u - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) \le \kappa\, t\right\}} = \mathbb{I}_{\left\{\kappa^{-1}u - \kappa^{-1}\alpha_0 - \kappa^{-1}\boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) - t \le 0\right\}}.$$

Definamos el conjunto $A_{s,\alpha_0,\boldsymbol{\theta},t} = \{(u, \mathbf{x}) \in \mathbb{R} \times \mathcal{S}^m \,:\, s\,u - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) - t \le 0\}$ y la clase de conjuntos:

$$\mathcal{A} = \left\{ A_{s,\alpha_0,\boldsymbol{\beta}_1,t} : (s, t, \alpha_0, \boldsymbol{\theta}) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{m-1} \right\}.$$

Dado que la familia de funciones

$$\left\{g(u,\mathbf{x}) = su - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) - t \leq 0 \text{ para } (s,t,\alpha_0,\boldsymbol{\theta}) \in \mathbb{R}_{>0}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}^{m-1}\right\}$$

es una familia finito-dimensional de dimensión $m+2$, aplicando los Lemas 9.6, 9.8 y 9.9 de Kosorok (2008), obtenemos que $\mathcal{A}$ es una clase VC de conjuntos de índice a lo sumo $m+4$. Usando de nuevo el Lema 9.8 de Kosorok (2008), tenemos que la familia $\mathcal{F}$ es una clase VC de índice $V(\mathcal{F})$ a lo sumo $m+4$. Luego, por el Teorema 2.6.7 de van der Vaart & Wellner (1996), existe una constante universal K tal que para toda medida de probabilidad $\mathbb{Q}$, se tiene que para todo $0<\varepsilon<1$

$$N\left(\varepsilon, \mathcal{F}, L_2(\mathbb{Q})\right) \leq K\, V(\mathcal{F})\, (16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{2(V(\mathcal{F})-1)}.$$

Como $V(\mathcal{F}) \leq m+4$, la desigualdad anterior implica que

$$\int_0^\infty \sup_{\mathbb{Q}} \sqrt{\log N\left(\varepsilon, \mathcal{F}, L_2(\mathbb{Q})\right)}\, d\varepsilon \; < \infty\,,$$

donde el supremo se toma sobre todas las medidas de probabilidad discretas. El Teorema 2.5.2 de van der Vaart & Wellner (1996) permite deducir que $\mathcal{F}$ resulta una clase Donsker, es decir, el proceso indexado por $\mathcal{F}$, $\{(\mathbb{P}_{n_H} - \mathbb{P}_H)\, f \,:\, f \in \mathcal{F}\}$, satisface que

$$\mathbb{G}_{n_H} = \sqrt{n_H}\,(\mathbb{P}_{n_H} - \mathbb{P}_H) \rightsquigarrow \mathbb{G}\,, \tag{4.10}$$

donde $\mathbb{G}$ es un proceso gaussiano de media cero y una vez más utilizamos la notación estándar de procesos empíricos, es decir,

$$\mathbb{P}_{n_H} f = \frac{1}{n_H}\sum_{i=1}^{n_H} f(u_{H,i}, \mathbf{x}_{H,i}) \quad \text{y} \quad \mathbb{P}_H f = \mathbb{E} f(u_{H,1}, \mathbf{x}_{H,1})\,.$$

Como en la demostración del Lema 4.3, definamos $\mathbb{L}_t^H(\alpha_0, \boldsymbol{\theta}, \sigma) = \mathbb{P}(u_{H,1} - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{H,1}) \leq \sigma\, t)$. De (4.10) se sigue que

$$\sqrt{n_H} \sup_{(s,t,\alpha_0,\boldsymbol{\theta})\in\mathbb{R}_{>0}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}^{m+1}} \left| \frac{1}{n_H}\sum_{i=1}^{n_H} \mathbb{I}_{\left\{u_{H,i} - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{H,i}) \leq s\, t\right\}} - \mathbb{L}_t^H(\alpha, \boldsymbol{\theta}, \sigma) \right| = O_{\mathbb{P}(1)}\,.$$

Por lo tanto, obtenemos que

$$\sqrt{n_H} \sup_{t\in\mathbb{R}} \left| \frac{1}{n_H}\sum_{i=1}^{n_H} \mathbb{I}_{\left\{u_{H,i} - \widehat{\alpha}_0 - \widehat{\boldsymbol{\theta}}^{\mathrm{T}} ilr(\mathbf{x}_{H,i}) \leq \widehat{\sigma}_H\, t\right\}} - \mathbb{L}_t^H(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_H) \right| = O_{\mathbb{P}(1)}\,,$$

o equivalentemente

$$\sqrt{n_H} \sup_{t\in\mathbb{R}} \left| \widehat{G}_H(t) - \mathbb{L}_t^H(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_H) \right| = O_{\mathbb{P}(1)}\,.$$

Para probar que $\sqrt{n_H}\|\widehat{G}_H - G_H\|_\infty$ está acotado en probabilidad, faltaría mostrar que

$$\sqrt{n_H} \sup_{t\in\mathbb{R}} \left| \mathbb{L}_t^H(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_H) - G_H(t) \right| = \sqrt{n_H} \sup_{t\in\mathbb{R}} \left| \mathbb{L}_t^H(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_H) - \mathbb{L}_t^H(0, \mathbf{0}_{m-1}, \sigma_H) \right| = O_{\mathbb{P}}(1)\,. \tag{4.11}$$

Para ello observemos que, como en la demostración del Lema 4.3,

$$\begin{aligned}\mathbb{L}_t^H(\alpha_{,0}\,\boldsymbol{\theta},\sigma) - G_H(t) &= \mathbb{P}(u_{H,1} - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{H,1}) \le \sigma\, t) - G_H(t)\\ &= \mathbb{E}\left\{\mathbb{P}\left(u_{H,1} - \alpha_0 - \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{H,1}) \le \sigma\, t \middle| \mathbf{x}_{H,1}\right)\right\} - G_H(t)\\ &= \mathbb{E}\left\{G_H\left(\frac{\alpha_0 + \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{H,1}) + \sigma\, t}{\sigma_H}\right)\right\} - G_H(t)\\ &= \mathbb{E}\left\{\left[G_H\left(\frac{\alpha_0 + \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}_{H,1}) + \sigma\, t}{\sigma_H}\right) - G_H\left(\frac{\sigma}{\sigma_H} t\right)\right]\right\}\\ &\quad + G_H\left(\frac{\sigma}{\sigma_H}\, t\right) - G_H(t)\,. \end{aligned} \tag{4.12}$$

Por **H1**, G_H tiene densidad acotada, con lo cual para cualquier $\mathbf{x} \in \mathcal{S}^m$ se cumple

$$\left|G_H\left(\frac{\alpha_0 + \boldsymbol{\theta}^{\mathrm{T}} ilr(\mathbf{x}) + \sigma t}{\sigma_H}\right) - G_H\left(\frac{\sigma t}{\sigma_H}\right)\right| \le \frac{\|g_H\|_\infty}{\sigma_H}\left\{|\alpha_0| + \|\boldsymbol{\theta}\|\, \|ilr(\mathbf{x})\|\right\}. \tag{4.13}$$

Luego, utilizando (4.12) y (4.13), concluímos que para todo $\sigma > 0$, $\alpha \in \mathbb{R}$ y $\boldsymbol{\theta} \in \mathbb{R}^{m+1}$ se tiene

$$\left|\mathbb{L}_t^H(\alpha, \boldsymbol{\theta}, \sigma) - G_H(t)\right| \le \frac{\|g_H\|_\infty}{\sigma_H}\left\{|\alpha_0| + \|\boldsymbol{\theta}\|\, \mathbb{E}\|ilr(\mathbf{x}_{H,1})\|\right\} + \left|G_H\left(\frac{\sigma}{\sigma_H}\, t\right) - G_H(t)\right|$$

Como $\|ilr(\mathbf{x}_{H,1})\| = \|\mathbf{x}_{H,1}\|_a$, $\widehat{\alpha}_{0,H} = \widehat{\beta}_{0,H} - \beta_{0,H}$, $\widehat{\boldsymbol{\theta}} = ilr(\widehat{\boldsymbol{\beta}}_H \ominus \boldsymbol{\beta}_H)$, $\|\widehat{\boldsymbol{\theta}}_H\| = \|\widehat{\boldsymbol{\beta}}_H \ominus \boldsymbol{\beta}_H\|_a$, de la desigualdad anterior deducimos que

$$\begin{aligned}\sqrt{n_H}\sup_{t\in\mathbb{R}}\left|\mathbb{L}_t^H(\widehat{\alpha}_0, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_H) - G_H(t)\right| &\le \frac{\|g_H\|_\infty}{\sigma_H}\sqrt{n_H}\left\{|\widehat{\beta}_{0,H} - \beta_{0,H}| + \|\widehat{\boldsymbol{\beta}}_H \ominus \boldsymbol{\beta}_H\|_a\, \mathbb{E}\|\mathbf{x}_{H,1}\|_a\right\}\\ &\quad + \sqrt{n_H}\sup_{t\in\mathbb{R}}\left|G_H\left(\frac{\widehat{\sigma}_H}{\sigma_H}\, t\right) - G_H(t)\right|. \end{aligned} \tag{4.14}$$

Observemos que

$$\left|G_H\left(\frac{\widehat{\sigma}_H}{\sigma_H}\, t\right) - G_H(t)\right| \le \left|\frac{\widehat{\sigma}_H}{\sigma_H} - 1\right| g_H(\widehat{\xi}\, t)t\,,$$

donde $\widehat{\xi}$ es un punto intermedio entre 1 y $\widehat{\sigma}_H/\sigma_H$. Por lo tanto, usando que por **H7** $r_H(u) = g_H(u)u$ es acotada y que $\lim_{n_H\to\infty}\mathbb{P}(\widehat{\xi} > 1/2) = 1$ pues $\widehat{\sigma}_H/\sigma_H \xrightarrow{p} 1$, obtenemos que en el conjunto $\{\widehat{\xi} > 1/2\}$

$$\sqrt{n_H}\sup_{t\in\mathbb{R}}\left|G_H\left(\frac{\widehat{\sigma}_H}{\sigma_H}\, t\right) - G_H(t)\right| \le \sqrt{n_H}\frac{|\widehat{\sigma}_H - \sigma_H|}{\sigma_H}\|r_H\|_\infty \frac{1}{\widehat{\xi}} \le 2\sqrt{n_H}\frac{|\widehat{\sigma}_H - \sigma_H|}{\sigma_H}\|r_H\|_\infty\,.$$

Utilizando (4.14) deducimos que en el conjunto $\{\widehat{\xi} > 1/2\}$

$$\begin{aligned}\sqrt{n_H}\sup_{t\in\mathbb{R}}\left|\mathbb{L}_t^H(\widehat{\alpha}_{0,H}, \widehat{\boldsymbol{\theta}}_H, \widehat{\sigma}_H) - G_H(t)\right| &\le \frac{\|g_H\|_\infty}{\sigma_H}\sqrt{n_H}\left\{|\widehat{\beta}_{0,H} - \beta_{0,H}| + \|\widehat{\boldsymbol{\beta}}_H \ominus \boldsymbol{\beta}_H\|_a\, \mathbb{E}\|\mathbf{x}_{H,1}\|_a\right\}\\ &\quad + 2\sqrt{n_H}\frac{|\widehat{\sigma}_H - \sigma_H|}{\sigma_H}\|r_H\|_\infty \end{aligned}$$

y (4.11) es consecuencia de $\lim_{n_H\to\infty}\mathbb{P}(\widehat{\xi} > 1/2) = 1$ y de la hipótesis **H7**. □

Definamos

$$\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p) = \frac{1}{n_D}\sum_{i=1}^{n_D} \mathcal{K}\left(\frac{p - W_{\mathbf{x},i}}{h_{n_D}}\right),$$

donde para cada $1 \leq i \leq n_D$, $W_{\mathbf{x},i}$ está definido en (4.9). Observemos que, como $\mathcal{K}$ es no–decreciente, $\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)$ es no–decreciente. Por otra parte, $W_{\mathbf{x},i}$ no son observables y el estimador $\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}$ corresponde al estimador suavizado de la función de distribución de $W_{\mathbf{x}}$ cuya consistencia fuerte uniforme daremos en el Lema 4.6. Para probarlo necesitaremos el siguiente resultado cuya demostración puede consultarse en el Apéndice B de Pollard (1984).

Lema 4.5 (Desigualdad de Hoeffding)**.** *Sean $Z_1 \dots Z_n$, variables aleatorias independientes tales que $\mathbb{E}(Z_i) = 0$ y $a_i \leq Z_i \leq b_i$. Entonces para todo $\eta > 0$, se tiene*

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} Z_i\right| \geq \eta\right) \leq 2\,\exp\left\{-\frac{2n^2\eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

Lema 4.6. *Sean $(y_{j,i}, \mathbf{x}_{j,i})$, $1 \leq i \leq n_j$, $j = D, H$, observaciones independientes que satisfacen los modelos de regresión en (4.1) y (4.2). Supongamos que valen las hipótesis **H1**, **H2**(a) y que $h_{n_D} \to 0$, entonces para cada $\mathbf{x} \in \mathcal{S}^m$, se tiene que $\sup_{0<p<1}\left|\widehat{ROC}^{\star}_{\mathbf{x},h}(p) - ROC_{\mathbf{x}}(p)\right| \xrightarrow{c.s.} 0$.*

Demostración. Por el Lema 4.1, bastará probar que $\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p) \xrightarrow{c.s.} \mathrm{ROC}_{\mathbf{x}}(p)$, para todo $0 < p < 1$. Notemos que

$$\left|\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p) - \mathrm{ROC}_{\mathbf{x}}(p)\right| \leq \left|\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p) - \mathbb{E}\left\{\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)\right\}\right| + \left|\mathbb{E}\left\{\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)\right\} - \mathrm{ROC}_{\mathbf{x}}(p)\right|.$$

Probaremos que

$$\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p) - \mathbb{E}\left\{\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)\right\} \xrightarrow{c.s.} 0\,, \tag{4.15}$$

$$\mathbb{E}\left\{\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)\right\} \longrightarrow \mathrm{ROC}_{\mathbf{x}}(p)\,. \tag{4.16}$$

Para probar (4.15) usaremos la desigualdad de Hoeffding tomando

$$Z_i = \mathcal{K}\left(\frac{p - W_{\mathbf{x},i}}{h_{n_D}}\right) - \mathbb{E}\left\{\mathcal{K}\left(\frac{p - W_{\mathbf{x},1}}{h_{n_D}}\right)\right\}.$$

Como $\mathcal{K}$ es una función de distribución, tenemos que $-1 \leq Z_i \leq 1$ por lo tanto, el Lema 4.5 implica que para todo $\eta \geq 0$

$$\mathbb{P}\left(\left|\frac{1}{n_D}\sum_{i=1}^{n_D} \mathcal{K}\left(\frac{p - W_{\mathbf{x},i}}{h_{n_D}}\right) - \mathbb{E}\left\{\mathcal{K}\left(\frac{p - W_{\mathbf{x},1}}{h_{n_D}}\right)\right\}\right| \geq \eta\right) \leq 2\,\exp\left\{-\frac{n\,\eta^2}{2}\right\}.$$

Luego,

$$\sum_{n_D=1}^{\infty} \mathbb{P}\left(\left|\frac{1}{n_D}\sum_{i=1}^{n_D} \mathcal{K}\left(\frac{p - W_{\mathbf{x},i}}{h_{n_D}}\right) - \mathbb{E}\left\{\mathcal{K}\left(\frac{p - W_{\mathbf{x},1}}{h_{n_D}}\right)\right\}\right| > \eta\right) < \infty.$$

Aplicando el Lema de Borel-Cantelli, obtenemos

$$\mathbb{P}\left(\bigcap_{N=1}^{\infty}\bigcup_{n_D=N}^{\infty}\left\{\left|\frac{1}{n_D}\sum_{i=1}^{n_D}\mathcal{K}\left(\frac{p-W_{\mathbf{x},i}}{h_{n_D}}\right)-\mathbb{E}\left\{\mathcal{K}\left(\frac{p-W_{\mathbf{x},1}}{h_{n_D}}\right)\right\}\right|>\eta\right\}\right)=0\,,$$

o equivalentemente,

$$\mathbb{P}\left(\bigcup_{N=1}^{\infty}\bigcap_{n_D=N}^{\infty}\left\{\left|\frac{1}{n_D}\sum_{i=1}^{n_D}\mathcal{K}\left(\frac{p-W_{\mathbf{x},i}}{h_{n_D}}\right)-\mathbb{E}\left\{\mathcal{K}\left(\frac{p-W_{\mathbf{x},1}}{h_{n_D}}\right)\right\}\right|<\eta\right\}\right)=1\,,$$

es decir, dado $\eta>0$,

$$\mathbb{P}\left(\exists N\geq 1\,:\,\forall n_D\geq N,\,\left|\frac{1}{n_D}\sum_{i=1}^{n_D}\mathcal{K}\left(\frac{p-W_{\mathbf{x},i}}{h_{n_D}}\right)-\mathbb{E}\left\{\mathcal{K}\left(\frac{p-W_{\mathbf{x},1}}{h_{n_D}}\right)\right\}\right|<\eta\right)=1\,,$$

lo cual prueba (4.15).

Veamos ahora que (4.16) vale. Observemos que

$$\begin{aligned}\mathbb{E}\left\{\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)\right\}&=\mathbb{E}\left\{\frac{1}{n_D}\sum_{i=1}^{n_D}\mathcal{K}\left(\frac{p-W_{\mathbf{x},i}}{h_{n_D}}\right)\right\}=\mathbb{E}\left\{\mathcal{K}\left(\frac{p-W_{\mathbf{x},1}}{h_{n_D}}\right)\right\}\\&=\int_{-\infty}^{+\infty}\mathcal{K}\left(\frac{p-w}{h_{n_D}}\right)dF_{W_{\mathbf{x}}}(w)=\lim_{\substack{a\to-\infty\\b\to+\infty}}\int_a^b\mathcal{K}\left(\frac{p-w}{h_{n_D}}\right)dF_{W_{\mathbf{x}}}(w)\,.\end{aligned}\tag{4.17}$$

Llamemos $\alpha(w)=\mathcal{K}\left((p-w)/h\right)$. Tenemos que $\lim_{a\to-\infty}\alpha(a)=1$, mientras que $\lim_{b\to+\infty}\alpha(b)=0$. Por otra parte, $\alpha(w)$ es derivable y $\alpha'(w)=-K_h(p-w)$, donde $K_h(t)=(1/h)K(t/h)$. De esta forma, si usamos la fórmula de integración por partes tenemos que

$$\int_a^b\alpha(w)dF_{W_{\mathbf{x}}}(w)=\alpha(b)F_{W_{\mathbf{x}}}(b)-\alpha(a)F_{W_{\mathbf{x}}}(a)-\int_a^b F_{W_{\mathbf{x}}}(w)d\alpha(w)\,.\tag{4.18}$$

Como $\lim_{a\to-\infty}\alpha(a)F_{W_{\mathbf{x}}}(a)=0$ y $\lim_{b\to+\infty}\alpha(b)F_{W_{\mathbf{x}}}(b)=0$, de (4.17) y (4.18) deducimos que

$$\mathbb{E}\left\{\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)\right\}=\frac{1}{h_{n_D}}\int_{-\infty}^{+\infty}F_{W_{\mathbf{x}}}(w)K\left(\frac{p-w}{h_{n_D}}\right)dw\,.$$

Por lo tanto, obtenemos que

$$\mathbb{E}\left\{\widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)\right\}=\int_{-\infty}^{+\infty}F_{W_{\mathbf{x}}}(p-h\,u)K\left(u\right)du\,.$$

La continuidad de $\mathrm{ROC}_{\mathbf{x}}(p)$, el hecho que $F_{W_{\mathbf{x}}}(p)=\mathrm{ROC}_{\mathbf{x}}(p)$ junto con la integrabilidad de K y el Teorema de Convergencia Mayorada permiten deducir que $\lim_{h\to0}\int_{-\infty}^{+\infty}F_{W_{\mathbf{x}}}(p-h\,u)K\left(u\right)du=F_{W_{\mathbf{x}}}(p)$, lo que concluye la demostración de (4.16). □

Demostración del Teorema 4.2. Por el Lema 4.2, y usando que, como $\mathcal{K}$ es no–decreciente, $\widehat{\mathrm{ROC}}_{\mathbf{x},h}(p)$ es no–decreciente, para obtener el resultado uniforme bastará probar que, para cada $0<p<1$, $\widehat{\mathrm{ROC}}_{\mathbf{x},h}(p)-\mathrm{ROC}_{\mathbf{x}}(p)\xrightarrow{p}0$.

Fijemos entonces $0 < p < 1$. Podems escribir $\widehat{\mathrm{ROC}}_{\mathbf{x},h}(p) - \mathrm{ROC}_{\mathbf{x}}(p) = \widehat{A}_1(p) + \widehat{A}_2(p)$ donde $\widehat{A}_1(p) = \widehat{\mathrm{ROC}}_{\mathbf{x},h}(p) - \widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p)$ y $\widehat{A}_2(p) = \widehat{\mathrm{ROC}}^{\star}_{\mathbf{x},h}(p) - \mathrm{ROC}_{\mathbf{x}}(p)$. El Lema 4.6 implica que $\widehat{A}_2(p) \xrightarrow{c.s.} 0$, por lo tanto, bastará ver que $\widehat{A}_1(p) \xrightarrow{p} 0$. De ahora en más sea $n = n_D + n_H$ y $n_{\mathrm{mín}} = \mathrm{mín}(n_D, n_H)$.

Para ello tenemos que probar que dado $\eta > 0$, $\mathbb{P}\left(\widehat{A}_1(p) > \eta\right) \to 0$, es decir, que para todo $\nu > 0$, existe n_0 tal que, para todo $n_{\mathrm{mín}} \geq n_0$,

$$\mathbb{P}\left(\left|\widehat{A}_1(p)\right| > \eta\right) < \nu\,. \tag{4.19}$$

Sea $C > 1$ tal que $\mathbb{P}(|\varepsilon_D| + \|\mathbf{x}_D\|_a \leq C) \geq 1 - \eta/4$. Por la ley de los grandes números,

$$\frac{1}{n_D}\sum_{i=1}^{n_D} \mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a > C\}} \xrightarrow{c.s.} \mathbb{P}(|\varepsilon_D| + \|\mathbf{x}_D\|_a > C)\,.$$

Sea $\mathcal{N}_C$ tal que $\mathbb{P}(\mathcal{N}_C) = 0$ y para todo $\omega \notin \mathcal{N}_C$,

$$\frac{1}{n_D}\sum_{i=1}^{n_D} \mathbb{I}_{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a > C} \to \mathbb{P}(|\varepsilon_D| + \|\mathbf{x}_D\|_a > C)\,.$$

Luego, si $\omega \notin \mathcal{N}_C$, existe n_1 tal que, para todo $n_D \geq n_1$,

$$\frac{1}{n_D}\sum_{i=1}^{n_D} \mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a > C\}} < \frac{\eta}{2}\,.$$

Observemos que

$$\widehat{A}_1(p) = \frac{1}{n_D}\sum_{i=1}^{n_D}\left\{\mathcal{K}\left(\frac{p - \widehat{W}_{\mathbf{x},i}}{h_{n_D}}\right) - \mathcal{K}\left(\frac{p - W_{\mathbf{x},i}}{h_{n_D}}\right)\right\}.$$

Por lo tanto, podemos tenemos $\widehat{A}_1(p) = \widehat{A}_{1,1}(p) + \widehat{A}_{1,2}(p)$ donde

$$\widehat{A}_{1,1}(p) = \frac{1}{n_D}\sum_{i=1}^{n_D}\left\{\mathcal{K}\left(\frac{p - \widehat{W}_{\mathbf{x},i}}{h_{n_D}}\right) - \mathcal{K}\left(\frac{p - W_{\mathbf{x},i}}{h_{n_D}}\right)\right\}\mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a > C\}}\,,$$
$$\widehat{A}_{1,2}(p) = \frac{1}{n_D}\sum_{i=1}^{n_D}\left\{\mathcal{K}\left(\frac{p - \widehat{W}_{\mathbf{x},i}}{h_{n_D}}\right) - \mathcal{K}\left(\frac{p - W_{\mathbf{x},i}}{h_{n_D}}\right)\right\}\mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a \leq C\}}\,.$$

Usando que $\mathcal{K}$ es una función de distribución y, por lo tanto, $0 \leq \mathcal{K} \leq 1$, deducimos que $\left|\widehat{A}_{1,1}(p)\right| \leq \eta/2$, para $\omega \notin \mathcal{N}_C$ si $n_{\mathrm{mín}} \geq n_1$. De donde, como $\mathbb{P}(\mathcal{N}_C) = 0$, obtenemos que si $n_{\mathrm{mín}} \geq n_1$

$$\mathbb{P}\left(\left|\widehat{A}_1(p)\right| > \eta\right) \leq \mathbb{P}\left(\left|\widehat{A}_{1,2}(p)\right| > \frac{\eta}{2}\right).$$

Por lo tanto, para probar (4.19), bastará probar que existe $n_0 \geq n_1$ tal que para todo $n_{\mathrm{mín}} \geq n_0$

$$\mathbb{P}\left(\left|\widehat{A}_{1,2}(p)\right| > \frac{\eta}{2}\right) < \nu\,. \tag{4.20}$$

Usando desarrollo de Taylor de orden 2 de la función $f(t) = \mathcal{K}((p-t)/h_{n_D})$ alrededor del punto $W_{\mathbf{x},i}$ con ξ_i obtenemos $\widehat{A}_{1,2}(p) = \widehat{B}_1(p) + \widehat{B}_2(p)$ donde

$$\widehat{B}_1(p) = \frac{1}{n_D\, h_{n_D}} \sum_{i=1}^{n_D} K\left(\frac{W_{\mathbf{x},i} - p}{h_{n_D}}\right) \left(\widehat{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right) \mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a \le C\}}\,,$$

$$\widehat{B}_2(p) = \frac{1}{2\, n_D\, h_{n_D}^2} \sum_{i=1}^{n_D} K'\left(\frac{\xi_i - p}{h_{n_D}}\right) \left(\widehat{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right)^2 \mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a \le C\}}\,,$$

donde ξ_i es un punto intermedio entre $\widehat{W}_{\mathbf{x},i}$ y $W_{\mathbf{x},i}$. Para probar (4.20), mostraremos que existe $n_0 \ge n_1$ tal que, para todo $n_{\text{mín}} \ge n_0$,

$$\mathbb{P}\left(\left|\widehat{B}_1(p)\right| > \frac{\eta}{4}\right) < \frac{\nu}{2}\,, \tag{4.21}$$

$$\mathbb{P}\left(\left|\widehat{B}_2(p)\right| > \frac{\eta}{4}\right) < \frac{\nu}{2}\,. \tag{4.22}$$

Para ello, definamos

$$\widetilde{W}_{\mathbf{x},i} = 1 - G_H\left(\frac{\widehat{\varepsilon}_{D,i} + \widehat{a}(\mathbf{x})}{\widehat{b}}\right)\,, \qquad U_{\mathbf{x},i} = \frac{\widehat{\varepsilon}_{D,i} + \widehat{a}(\mathbf{x})}{\widehat{b}} - \frac{\varepsilon_{D,i} + a(\mathbf{x})}{b}\,,$$

$$\widehat{c} = \left|\frac{\sigma_D}{\widehat{\sigma}_D}\,\frac{1}{\widehat{b}} - \frac{1}{b}\right| \xrightarrow{p} 0\,, \qquad \widehat{d}(\mathbf{x}) = \left|\frac{\widehat{a}(\mathbf{x})}{\widehat{b}} - \frac{a(\mathbf{x})}{b}\right| \xrightarrow{p} 0\,,$$

$$\widehat{\Delta}_0 = |\beta_{0,D} - \widehat{\beta}_{0,D}|/(\widehat{b}\,\widehat{\sigma}_D) \xrightarrow{p} 0\,, \qquad \widehat{\Delta}_1 = \|\boldsymbol{\beta}_D \ominus \widehat{\boldsymbol{\beta}}_D\|_a/(\widehat{b}\,\widehat{\sigma}_D) \xrightarrow{p} 0\,.$$

Observemos que

$$\begin{aligned} |U_{\mathbf{x},i}| &\le \left|\varepsilon_{D,i}\left(\frac{\sigma_D}{\widehat{\sigma}_D}\,\frac{1}{\widehat{b}} - \frac{1}{b}\right) + \frac{\mu_D(\mathbf{x}_{D,i}) - \widehat{\mu}_D(\mathbf{x}_{D,i})}{\widehat{b}\,\widehat{\sigma}_D} + \frac{\widehat{a}(\mathbf{x})}{\widehat{b}} - \frac{a(\mathbf{x})}{b}\right| \\ &\le |\varepsilon_{D,i}|\,\widehat{c} + \left|\frac{\mu_D(\mathbf{x}_{D,i}) - \widehat{\mu}_D(\mathbf{x}_{D,i})}{\widehat{b}\,\widehat{\sigma}_D}\right| + \widehat{d}(\mathbf{x})\,. \end{aligned}$$

Por otra parte, usando que $\mu_D(\mathbf{x}) = \beta_{0,D} + \langle \boldsymbol{\beta}_D, \mathbf{x}\rangle_a$ y $\widehat{\mu}_D(\mathbf{x}) = \widehat{\beta}_{0,D} + \langle \widehat{\boldsymbol{\beta}}_D, \mathbf{x}\rangle_a$, obtenemos que $\mu_D(\mathbf{x}_{D,i}) - \widehat{\mu}_D(\mathbf{x}_{D,i}) = \beta_{0,D} - \widehat{\beta}_{0,D} + \langle \boldsymbol{\beta}_D \ominus \widehat{\boldsymbol{\beta}}_D, \mathbf{x}_{D,i}\rangle_a$, de donde,

$$\left|\frac{\mu_D(\mathbf{x}_{D,i}) - \widehat{\mu}_D(\mathbf{x}_{D,i})}{\widehat{b}\,\widehat{\sigma}_D}\right| \le \widehat{\Delta}_0 + \widehat{\Delta}_1\,\|\mathbf{x}_{D,i}\|_a\,.$$

Por lo tanto, como g_H es acotada por **H1**, concluímos que

$$\left|\widetilde{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right| \le \|g_H\|_\infty\,|U_{\mathbf{x},i}| \le \|g_H\|_\infty \left\{|\varepsilon_{D,i}|\,\widehat{c} + \widehat{\Delta}_0 + \widehat{\Delta}_1\,\|\mathbf{x}_{D,i}\|_a + \widehat{d}(\mathbf{x})\right\}\,. \tag{4.23}$$

Por otra parte, tenemos que

$$\left|\widehat{W}_{\mathbf{x},i} - \widetilde{W}_{\mathbf{x},i}\right| = \left|\widehat{G}_H\left(\frac{\widehat{\varepsilon}_{D,i} + \widehat{a}(\mathbf{x})}{\widehat{b}}\right) - G_H\left(\frac{\widehat{\varepsilon}_{D,i} + \widehat{a}(\mathbf{x})}{\widehat{b}}\right)\right| \le \left\|\widehat{G}_H - G_H\right\|_\infty\,. \tag{4.24}$$

Usando (4.23) y (4.24) deducimos que, si $|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a \le C$,

$$\begin{aligned} \left|\widehat{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right| &\le \left|\widehat{W}_{\mathbf{x},i} - \widetilde{W}_{\mathbf{x},i}\right| + \left|\widetilde{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right| \\ &\le \left\|\widehat{G}_H - G_H\right\|_\infty + \|g_H\|_\infty \left\{|\varepsilon_{D,i}|\,\widehat{c} + \widehat{\Delta}_0 + \widehat{\Delta}_1\,\|\mathbf{x}_{D,i}\|_a + \widehat{d}(\mathbf{x})\right\} \\ &\le \left\|\widehat{G}_H - G_H\right\|_\infty + C\,\|g_H\|_\infty \widehat{d}_1(\mathbf{x})\,, \end{aligned} \tag{4.25}$$

donde $\widehat{d}_1(\mathbf{x}) = \widehat{c} + \widehat{\Delta}_0 + \widehat{\Delta}_1 + \widehat{d}(\mathbf{x})$ y $\widehat{d}_1(\mathbf{x}) \xrightarrow{p} 0$.

Comencemos probando (4.21). Para ello, como por **H6** $h_{n_D} \to 0$ y $n_D\, h_{n_D} \to 0$, usando **H2**(b) deducimos que

$$\widehat{f}_{n_D}(p) = \frac{1}{n_D\, h_{n_D}} \sum_{i=1}^{n_D} K\left(\frac{W_{\mathbf{x},i} - p}{h_{n_D}}\right) \xrightarrow{p} f_{W_x}(p) = \mathrm{ROC}'_{\mathbf{x}}(p)\,,$$

Luego, existe $n_2 \geq n_1$ tal que si $n_D \geq n_2$,

$$\mathbb{P}\left(\widehat{f}_{n_D}(p) > M\right) < \frac{\nu}{4}\,, \tag{4.26}$$

donde $M = 1$ si $\mathrm{ROC}'_{\mathbf{x}}(p) = 0$ y $M = 2\,\mathrm{ROC}'_{\mathbf{x}}(p)$ si $\mathrm{ROC}'_{\mathbf{x}}(p) \neq 0$.

Usando (4.25) obtenemos que

$$\begin{aligned}\left|\widehat{B}_1(p)\right| &\leq \frac{1}{n_D\, h_{n_D}} \sum_{i=1}^{n_D} K\left(\frac{W_{\mathbf{x},i} - p}{h_{n_D}}\right) \left|\widehat{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right| \mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a \leq C\}} \\ &\leq \widehat{f}_{n_D}(p) \left\{ \left\| \widehat{G}_H - G_H \right\|_\infty + C\, \|g_H\|_\infty\, \widehat{d}_1(\mathbf{x}) \right\}.\end{aligned}$$

Teniendo en cuenta que $\widehat{d}_1(\mathbf{x}) \xrightarrow{p} 0$, $\left\| \widehat{G}_H - G_H \right\|_\infty \xrightarrow{p} 0$, tenemos que existe $n_3 \geq n_1$ tal que si $n_{\mín} \geq n_3$,

$$\mathbb{P}\left(\left\| \widehat{G}_H - G_H \right\|_\infty + C\, \|g_H\|_\infty\, \widehat{d}_1(\mathbf{x}) > \frac{\eta}{4\,M} \right) < \frac{\nu}{4}\,.$$

Por lo tanto, usando (4.26), deducimos que si $n_{\mín} \geq \máx(n_2, n_3)$,

$$\mathbb{P}\left(\left|\widehat{B}_1(p)\right| > \frac{\eta}{4} \right) < \frac{\nu}{2}\,,$$

lo que concluye la demostración de (4.21).

Probaremos ahora (4.22). Observemos que

$$\begin{aligned}\left|\widehat{B}_2(p)\right| &\leq \frac{1}{2\, n_D\, h_{n_D}^2} \sum_{i=1}^{n_D} \left| K'\left(\frac{\xi_i - p}{h_{n_D}}\right) \right| \left(\widehat{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right)^2 \mathbb{I}_{\{|\varepsilon_{D,i}| + \|\mathbf{x}_{D,i}\|_a \leq C\}} \\ &\leq \|K'\|_\infty \frac{1}{n_D\, h_{n_D}^2} \left\{ n_D \left\| \widehat{G}_H - G_H \right\|_\infty^2 + C^2\, \|g_H\|_\infty^2 n_D\, \widehat{d}_1^2(\mathbf{x}) \right\},\end{aligned}$$

donde usamos la siguiente acotación

$$\left(\widehat{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right)^2 = \left(\widehat{W}_{\mathbf{x},i} - \widetilde{W}_{\mathbf{x},i} + \widetilde{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right)^2 \leq 2 \left\{ \left(\widehat{W}_{\mathbf{x},i} - \widetilde{W}_{\mathbf{x},i}\right)^2 + \left(\widetilde{W}_{\mathbf{x},i} - W_{\mathbf{x},i}\right)^2 \right\}.$$

El resultado se obtiene ahora del hecho que $n_D/n \to \tau$, $n_D\, h_{n_D}^2 \to \infty$, $\sqrt{n_H} \left\| \widehat{G}_H - G_H \right\|_\infty = O_{\mathbb{P}}(1)$ por el Lema 4.4 y que, por **H5** y **H7**, $n^{1/2} \widehat{d}_1(\mathbf{x}) = O_{\mathbb{P}}(1)$. Efectivamente, como $\|K'\|_\infty \left\{ n_D \left\| \widehat{G}_H - G_H \right\|_\infty^2 + C^2\, \|g_H\|_\infty^2 n_D\, \widehat{d}_1^2(\mathbf{x}) \right\}$ está acotado en probabilidad y $n_D\, h_{n_D}^2 \to \infty$, se tiene que existe n_4 tal que si $n_{\mín} \geq n_4$,

$$\mathbb{P}\left(\left|\widehat{B}_2(p)\right| > \frac{\eta}{4} \right) < \frac{\nu}{2},$$

probando (4.22). □

Capítulo 5
Estudios de Simulación

En este capítulo resumimos los resultados de un experimento numérico llevado a cabo para evaluar el comportamiento de los estimadores de la curva ROC condicional con covariables composicionales introducidos en la Sección 3.4 en distintos escenarios. Los modelos y estimadores considerados se presentan en las Secciones 5.1 y 5.2, respectivamente, mientras que las medidas de resumen utilizadas se describen en la Sección 5.3. Finalmente, los resultados obtenidos se presentan en las Secciones 5.4 y 5.5 para la situación en que las muestras tienen igual tamaño y para el caso desbalanceado, respectivamente.

5.1. Modelo y distribuciones consideradas

Para comparar los dos estimadores definidos en la Sección 3.4, en cada replicación y para los diferentes posibles valores del tamaño muestral de cada población n_D y n_H, se generaron observaciones independientes $(y_{D,i}, \mathbf{x}_{D,i}) \sim (Y_D, \mathbf{X}_D)$, $1 \leq i \leq n_D$, e $(y_{H,i}, \mathbf{x}_{H,i}) \sim (Y_D, \mathbf{X}_D)$, $1 \leq i \leq n_H$, donde $\mathbf{x}_{H,i} \in \mathcal{S}^3$, $\mathbf{x}_{D,i} \in \mathcal{S}^3$. Las covariables y las respuestas se relacionan mediante el siguiente modelo lineal:

$$Y_D = \beta_{0,D} + \langle \boldsymbol{\beta}_D, \mathbf{X}_D \rangle_a + \sigma_D \, \varepsilon_D \,, \tag{5.1}$$

$$Y_H = \beta_{0,H} + \langle \boldsymbol{\beta}_H, \mathbf{X}_H \rangle_a + \sigma_H \, \varepsilon_H \,, \tag{5.2}$$

donde $\beta_{0,D} = 4$, $\beta_{0,H} = 1$, $\boldsymbol{\beta}_D = (0.1, 0.3, 0.6)^{\mathrm{T}}$, $\boldsymbol{\beta}_H = (0.05, 0.55, 0.4)^{\mathrm{T}}$ y $\sigma_D = \sigma_H = 1$. Los errores se generaron independientes de las covariables con distintas distribuciones pero, en todos los casos, $\mathbb{E}(\varepsilon_j) = 0$ y $\textsc{Var}(\varepsilon_j) = 1$, para $j = D, H$.

Consideramos dos modelos para la distribución de las covariables $\mathbf{X}_j$, $j = D, H$ en el simplex y supusimos que las covariables tienen la misma distribución en las dos poblaciones. Los modelos considerados fueron

MX.1 Las covariables siguen una distribución Dirichlet: $\mathbf{X}_D, \mathbf{X}_H \sim \mathcal{D}(3, 7, 10)$,

MX.2 Las coordenadas *ilr* de las covariables siguen una distribución uniforme en el rectángulo $[-2, 2] \times [-2, 2]$, es decir, $\mathbf{X}_D^*, \mathbf{X}_H^* \sim \mathcal{U}([-2, 2] \times [-2, 2])$.

Las dos posibles distribuciones de las covariables se combinaron con cuatro distribuciones distintas para los errores del biomarcador, donde los parámetros de las distribuciones de los errores fueron tomados para satisfacer $\mathbb{E}(\varepsilon_j) = 0$ y $\textsc{Var}(\varepsilon_j) = 1$ para $j = D, H$. En todas las situaciones, elegimos la distribución G_H de los errores en la población sana igual a la distribución G_D de los errores en la población de los enfermos. Los modelos considerados para $G = G_D = G_H$ fueron

G.1 $G = \mathcal{N}(0, 1)$.

G.2 $G = \mathcal{L}og(0, \sigma = \sqrt{3}/\pi)$, es decir,

$$G_j(x) = \frac{1}{1 + \exp\left\{\frac{x}{\sigma}\right\}}\,.$$

En este caso, los errores tienen distribución logística escalada de modo a tener esperanza 0 y varianza 1.

G.3 $\varepsilon_j \sim \sigma\, t_4$, $j = D, H$, donde $\sigma = 1/\sqrt{2}$ para asegurar que $\text{VAR}(\varepsilon_j) = 1$. Por otra parte, se tomaron 4 grados de libertad para garantizar la existencia de primer y segundo momento.

G.4 $G = \mathcal{U}(-\sqrt{3}, \sqrt{3})$ de esta forma los errores ε_j cumplen que $\mathbb{E}(\varepsilon_j) = 0$ y $\text{VAR}(\varepsilon_j) = 1$.

En todos los casos, se tomó el número de replicaciones $N_R = 1000$.

Exploramos el comportamiento para distintos tamaños de muestra. Para ello, dividimos el estudio en dos casos:

- el balanceado, en el que el tamaño de muestra en ambas poblaciones es el mismo, considerando muestras de tamaños $n_D = n_H = 50, 100, 300$, cuyos resultados se presentan en la Sección 5.4 y
- el caso no balanceado, en el que tomamos $n_D = 50, n_H = 100$ y $n_D = 100, n_H = 50$ y que se resumen en la Sección 5.5.

5.2. Los estimadores

Para cada modelo descripto en la Sección 5.1, calculamos, en grillas de puntos, estimadores de la curva ROC condicional, así como del área bajo la curva condicional denotada $\widehat{\text{AUC}}_{\mathbf{x}}$ utilizando los estimadores propuestos en las Secciones 3.4.1 y 3.4.2 y estudiados en la Sección 4.

El estimador semiparamétrico $\widehat{\text{ROC}}_{\mathbf{x}}$ corresponde al descripto en 3.4.1, que se obtiene utilizando la función de distribución empírica $\widehat{G}_D$ y la función cuantil empírica $\widehat{G}_H^{-1}$ obtenidas a partir de los residuos estandarizados y dadas por

$$\widehat{G}_D(s) = \frac{1}{n_D}\sum_{i=1}^{n_D} \mathbb{I}_{\{\widehat{\epsilon}_{D,i} \le s\}}$$
$$\widehat{G}_H^{-1}(p) = \text{ínf}\{s \in \mathbb{R} : \quad \widehat{G}_H(s) \ge p)\}\,.$$

El estimador suavizado $\widehat{\text{ROC}}_{\mathbf{x},h}$ se calcuó utilizando como núcleo K el núcleo Epanechnikov y como ventana h_{n_D} la mencionada en (3.8), es decir,

$$h^{\star}_{n_D}(p) = c_{n_D}\frac{\sqrt{5p(1-p)}}{\sqrt{2n_D}}$$

donde $c_{n_D} = 1 + 1.8\, n_D^{-1/5}$.

Por otra parte, además de estos dos estimadores, calculamos el estimador que llamaremos *plug-in* y denotaremos $\widehat{\mathrm{ROC}}_{\mathbf{x},\mathrm{PI}}$, el cual supone un modelo binormal, o sea, el estimador definido por

$$\widehat{\mathrm{ROC}}_{\mathbf{x},\mathrm{PI}}(p) = 1 - \Phi\left(\Phi^{-1}(1-p)\frac{\widehat{\sigma}_H}{\widehat{\sigma}_D} - \frac{\widehat{\beta}_{0,D} - \widehat{\beta}_{0,H} + \langle \widehat{\boldsymbol{\beta}}_D \ominus \widehat{\boldsymbol{\beta}}_H, \mathbf{x}\rangle_a}{\widehat{\sigma}_D}\right).$$

El objetivo fue evaluar la estabilidad de este estimador ante desviaciones de la distribución subyacente de los errores.

A partir de estos estimadores, se calculó el estimador del área bajo la curva condicional. Más precisamente, una vez obtenido el estimador $\widehat{\mathrm{ROC}}_{\mathbf{x}}(p)$ en una grilla de puntos en $[0,1]$, es decir para $p \in \mathcal{G}_p = \{p_j\}_{j=1}^{N_p}$, el estimador $\widehat{\mathrm{AUC}}_{\mathbf{x}}$ se definió como

$$\widehat{\mathrm{AUC}}_{\mathbf{x}} = \frac{1}{N_p}\sum_{j=1}^{N_p} \widehat{\mathrm{ROC}}_{\mathbf{x}}(p_j)\,.$$

De igual forma, y aunque $\mathrm{AUC}_{\mathbf{x}} = \int_0^1 \mathrm{ROC}_{\mathbf{x}}(p)\,dp$, se calculó $\mathrm{AUC}_{\mathbf{x}}$ como

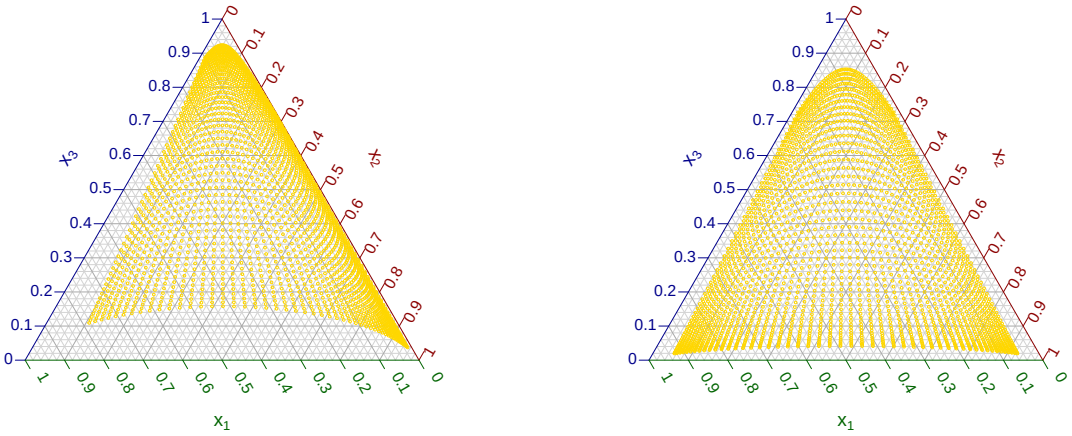
$$\mathrm{AUC}_{\mathbf{x}} = \frac{1}{N_p}\sum_{j=1}^{N_p} \mathrm{ROC}_{\mathbf{x}}(p_j)\,,$$

o sea, se evaluó la verdadera área bajo la curva como el promedio de la curva sobre la grilla $\mathcal{G}_p$ de puntos, aunque, por ejemplo, en el caso binormal como vimos, la $\mathrm{AUC}_{\mathbf{x}}$ tiene una expresión cerrada. Esta decisión se tomó para que no se confundiera el error numérico de aproximar la integral por el promedio en el cálculo del estimador con el error estadístico de la estimación.

5.3. Las medidas resumen

Para resumir la discrepancia entre el estimador y la verdadera superficie ROC, consideramos dos grillas de puntos para evaluar los estimadores de $\mathrm{ROC}_{\mathbf{x}}(p)$, a saber, la grilla $\mathcal{G}_p = \{p_j\}_{j=1}^{N_p}$ que corresponde a una grilla equiespaciada para los valores $p \in [0,1]$ y una grilla de posibles valores para $\mathbf{x}$, $\mathcal{G}_{\mathbf{x}} = \{\mathbf{x}_\ell\}_{\ell=1}^{N_{\mathbf{x}}}$. La grilla $\mathcal{G}_p$ toma valores entre 0 y 1 con paso 0.01, con lo cual $N_p = 101$.

Para definir $\mathcal{G}_{\mathbf{x}}$ para cada posible distribución de las covariables, **MX.1** y **MX.2**, calculamos la región de mayor densidad de la misma sobre la cual calcular la curva ROC condicional. Para ello, en cada caso generamos una grilla rectangular en el espacio *ilr* de $N_{\mathbf{x}} = 50 \times 50$ puntos en la región $\mathcal{R} = [-1.4, 3.2] \times [-0.8, 2.6]$ para el modelo **MX.1** y $\mathcal{R} = [-2,2] \times [-2,2]$ para el modelo **MX.2**. Luego, utilizando la inversa de la transformación *ilr* construimos la grilla $\mathcal{G}_{\mathbf{x}}$ en el símplex. Las grillas resultantes se presentan en la Figura 5.1.



(a) Modelo **MX.1** (b) Modelo **MX.2**

Figura 5.1: Grillas de puntos en el simplex $\mathcal{S}^3$ sobre las cuales se estimó la curva ROC condicional.

Para dar una medida global de la discrepancia entre la estimación, que denotaremos de forma genérica $\widehat{\mathrm{ROC}}_{\mathbf{x}}(p)$, y la verdadera curva ROC condicional, $\mathrm{ROC}_{\mathbf{x}}(p)$, calculamos la media sobre replicaciones de las siguientes medidas resumen:

- El error cuadrático medio, dado por

$$MSE = \frac{1}{N_{\mathbf{x}} N_p} \sum_{\ell=1}^{N_{\mathbf{x}}} \sum_{j=1}^{N_p} \left(\widehat{\mathrm{ROC}}_{\mathbf{x}_\ell}(p_j) - \mathrm{ROC}_{\mathbf{x}_\ell}(p_j) \right)^2 ,$$

- La distancia de Kolmogorov-Smirnov, dada por

$$KS = \sup_{1 \le \ell \le N_{\mathbf{x}}} \sup_{1 \le j \le N_p} \left| \widehat{\mathrm{ROC}}_{\mathbf{x}_\ell}(p_j) - \mathrm{ROC}_{\mathbf{x}_\ell}(p_j) \right| .$$

Por otra parte, para visualizar el comportamiento de ambas medidas resumen para los tres estimadores considerados y para los distintos modelos de distribución, calculamos los boxplots ajustados definidos en Hubert & Vandervieren (2008). En ese caso, se utilizaron los subíndices EMP, SM y PI para identificar a los estimadores $\widehat{\mathrm{ROC}}_{\mathbf{x}}$, $\widehat{\mathrm{ROC}}_{\mathbf{x},h}$ y $\widehat{\mathrm{ROC}}_{\mathbf{x},\mathrm{PI}}$, respectivamente.

Asimismo, para tener una mejor visualización de los estimadores del área bajo la curva, se calcularon los boxplots de superficie de las estimaciones obtenidas como fueron definidos en Genton et al. (2014), sobre las $N_R = 1000$ replicaciones para los dos estimadores considerados. Para estos gráficos, la noción de profundidad de volumen es utilizada para establecer un orden entre superficies. La superficie mediana (o más profunda) se representa en verde oscuro, la región central que contiene el 50 % de las superficies estimadas más profundas se ilustra en celeste, mientras que las superficies amarillas corresponen a los *bigotes*, más allá de cuyos límites una superficie se declara como atípica. Por claridad de visualización, en estos gráficos no se muestran las superficies atípicas. La verdadera superficie sobre el simplex $\mathrm{AUC}_{\mathbf{x}}$ estará representada en color verde lima. En el diagrama ternario inferior se presentan en gris los puntos donde se calcularon los estimadores y la verdadera superficie, tal como se indicó en la Figura 5.1.

5.4. Resultados para el caso balanceado

Las Tablas 5.1 y 5.2 presentan las medias sobre replicaciones de las medidas resumen, con sus respectivos desvíos estándar en gris, para el caso balanceado, es decir, cuando $n_D = n_H \in \{50, 100, 300\}$ para los tres estimadores considerados.

		G.1					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0196 (0.0169)	0.6311 (0.1906)	0.0193 (0.0168)	0.5978 (0.1821)	0.0185 (0.0168)	0.5543 (0.2131)
100	100	0.0110 (0.0102)	0.4836 (0.1713)	0.0109 (0.0101)	0.4664 (0.1633)	0.0104 (0.0102)	0.4190 (0.1877)
300	300	0.0036 (0.0038)	0.2853 (0.1137)	0.0036 (0.0038)	0.2771 (0.1139)	0.0034 (0.0038)	0.2443 (0.1216)
		G.2					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0216 (0.0192)	0.6788 (0.1778)	0.0212 (0.0191)	0.6321 (0.1801)	0.0203 (0.0192)	0.5845 (0.2019)
100	100	0.0116 (0.0115)	0.529 (0.1717)	0.0115 (0.0115)	0.4994 (0.1692)	0.0108 (0.0114)	0.4412 (0.1807)
300	300	0.0041 (0.0043)	0.3337 (0.1218)	0.0041 (0.0043)	0.3193 (0.1242)	0.0039 (0.0042)	0.2834 (0.1135)
		G.3					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0229 (0.0218)	0.7369 (0.1596)	0.0225 (0.0218)	0.6649 (0.1776)	0.0221 (0.0219)	0.6296 (0.1876)
100	100	0.0130 (0.0127)	0.6121 (0.1637)	0.0129 (0.0127)	0.5555 (0.1679)	0.0125 (0.0125)	0.5054 (0.1582)
300	300	0.0047 (0.0049)	0.4105 (0.1381)	0.0047 (0.0049)	0.3812 (0.1378)	0.0053 (0.0054)	0.3557 (0.1061)
		G.4					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0196 (0.0171)	0.5628 (0.2145)	0.0198 (0.0169)	0.5788 (0.1841)	0.0200 (0.0168)	0.5867 (0.1972)
100	100	0.0102 (0.0098)	0.4008 (0.1708)	0.0104 (0.0097)	0.4246 (0.1470)	0.0113 (0.0097)	0.4615 (0.1549)
300	300	0.0034 (0.0036)	0.2280 (0.0988)	0.0035 (0.0036)	0.2269 (0.0975)	0.0048 (0.0037)	0.3428 (0.0843)

Tabla 5.1: Media y desvío estándar (entre paréntesis en gris) sobre replicaciones de las medidas MSE y KS cuando las covariables tiene distribución **MX.1** y los errores tienen distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**).

Estas tablas muestran que los estimadores semiparamétricos $\widehat{\text{ROC}}_{\mathbf{x}}$ y $\widehat{\text{ROC}}_{\mathbf{x},h}$ dan resultados semejantes entre sí y estables a lo largo de las distintas distribuciones para los errores consideradas, lo cual es positivo ya que $\widehat{\text{ROC}}_{\mathbf{x},h}$ tiene la ventaja de ser un estimador suave. Tanto la media sobre replicaciones de los errores cuadráticos medios como de la distancia de Kolmogorov-Smirnov son menores para el caso en que las covariables tienen distribución **MX.2** que cuando tienen distribución Dirichlet. Sorprendentemente, el estimador basado en el supuesto de binormalidad es bastante estable cuando los errores tienen distribución Logística (Modelo **G.2**) o t_4 normalizada (Modelo **G.3**), dando valores medios levemente mayores de MSE y KS cuando las covariables son uniformes y los errores tienen distribución t_4 normalizada y uniforme. Salvo por ese caso, el estimador *plug-in* es muy competente a lo largo de los tamaños de muestra y distribuciones consideradas.

		G.1					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0029 (0.0023)	0.3891 (0.1134)	0.0031 (0.0024)	0.3718 (0.0717)	0.0023 (0.0024)	0.2731 (0.1168)
100	100	0.0015 (0.0013)	0.2923 (0.0889)	0.0017 (0.0013)	0.2880 (0.0614)	0.0012 (0.0013)	0.1929 (0.0871)
300	300	0.0005 (0.0004)	0.1715 (0.0556)	0.0005 (0.0004)	0.1608 (0.0540)	0.0004 (0.0004)	0.1112 (0.049)
		G.2					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0032 (0.0027)	0.4616 (0.1320)	0.0035 (0.0029)	0.4029 (0.0966)	0.0028 (0.0031)	0.3315 (0.127)
100	100	0.0018 (0.0015)	0.3616 (0.1121)	0.0019 (0.0015)	0.3231 (0.0885)	0.0016 (0.0017)	0.2586 (0.0903)
300	300	0.0006 (0.0004)	0.2177 (0.0737)	0.0006 (0.0005)	0.1981 (0.0720)	0.0006 (0.0005)	0.1763 (0.0546)
		G.3					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0037 (0.0032)	0.5598 (0.1517)	0.0040 (0.0037)	0.4474 (0.1225)	0.0043 (0.0061)	0.4262 (0.1365)
100	100	0.0020 (0.0017)	0.4520 (0.1432)	0.0022 (0.0019)	0.3763 (0.1158)	0.0027 (0.0037)	0.3513 (0.1076)
300	300	0.0007 (0.0006)	0.2930 (0.1133)	0.0007 (0.0006)	0.2599 (0.0992)	0.0016 (0.0022)	0.2675 (0.0692)
		G.4					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	50	0.0029 (0.0026)	0.2937 (0.0948)	0.0036 (0.0026)	0.3844 (0.0593)	0.0033 (0.0023)	0.3564 (0.0925)
100	100	0.0015 (0.0013)	0.2116 (0.0649)	0.0018 (0.0013)	0.2906 (0.0436)	0.0023 (0.0012)	0.3138 (0.0702)
300	300	0.0005 (0.0004)	0.1199 (0.0374)	0.0005 (0.0004)	0.1254 (0.0358)	0.0015 (0.0005)	0.2746 (0.0466)

Tabla 5.2: Media y desvío estándar (entre paréntesis en gris) sobre replicaciones de las medidas MSE y KS cuando las covariables tiene distribución **MX.2** y los errores tienen distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**).

Para tener una mejor comprensión del comportantamiento de las medidas resumen, las Figuras 5.2 a 5.5 presentan los boxplots ajustados del MSE y KS bajo los dos modelos para la distribución de las covariables y las cuatro posibles distribuciones de los errores. Las figuras mantienen la escala del eje vertical para mejorar la comparación visual entre las distintas distribuciones. En cada Figura se indica por MSE_{EMP}, MSE_{SM} y MSE_{PI}, los boxplots del MSE de $\widehat{\text{ROC}}_{\mathbf{x}}$, $\widehat{\text{ROC}}_{\mathbf{x},h}$ y $\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$, respectivamente que se muestran en verde, celeste y beige, respectivamente. Una notación análoga se utilizó al considerar los boxplots de la distancia de Kolmogorov–Smirnov, en la que los boxplots de las medidas asociadas a $\widehat{\text{ROC}}_{\mathbf{x}}$, $\widehat{\text{ROC}}_{\mathbf{x},h}$ y $\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$ se muestran en carmesí, verde agua y rosa, respectivamente.

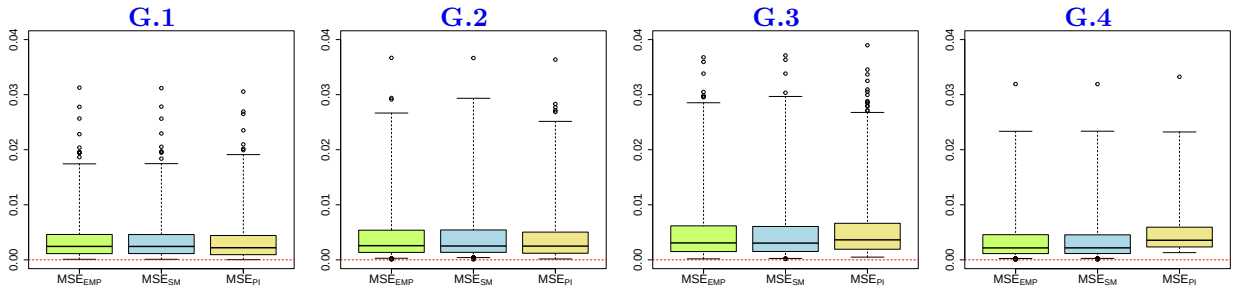


Figura 5.2: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.1** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 300$.

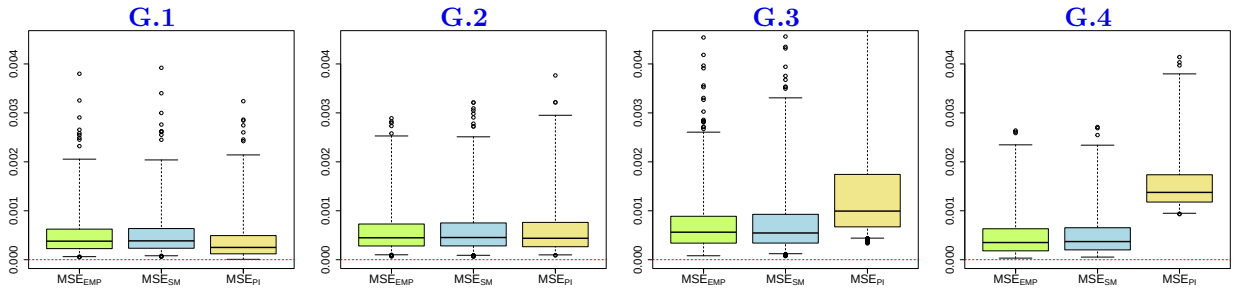


Figura 5.3: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 300$.

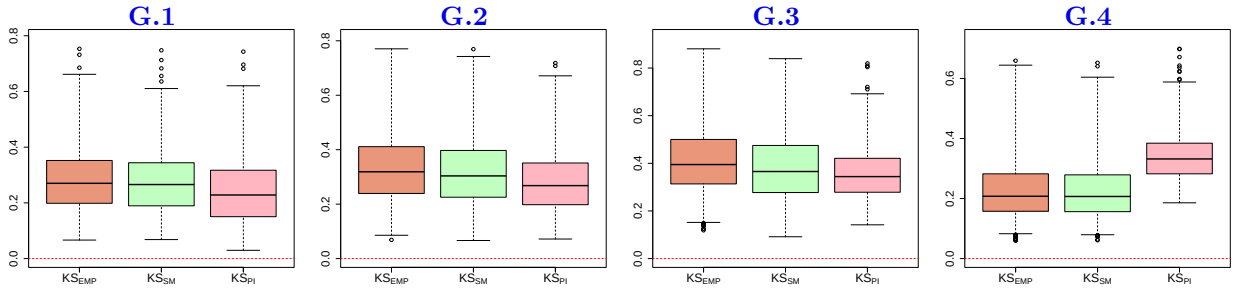


Figura 5.4: Boxplots ajustados de la distancia de Kolmogorov-Smirnov (KS) para los tres estimadores con covariables con distribución **MX.1** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 300$.

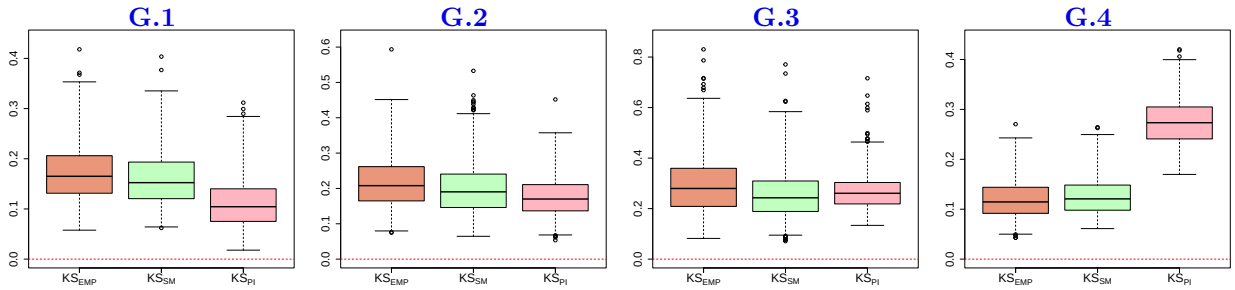


Figura 5.5: Boxplots ajustados de la distancia de Kolmogorov-Smirnov (KS) para los tres estimadores con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 300$.

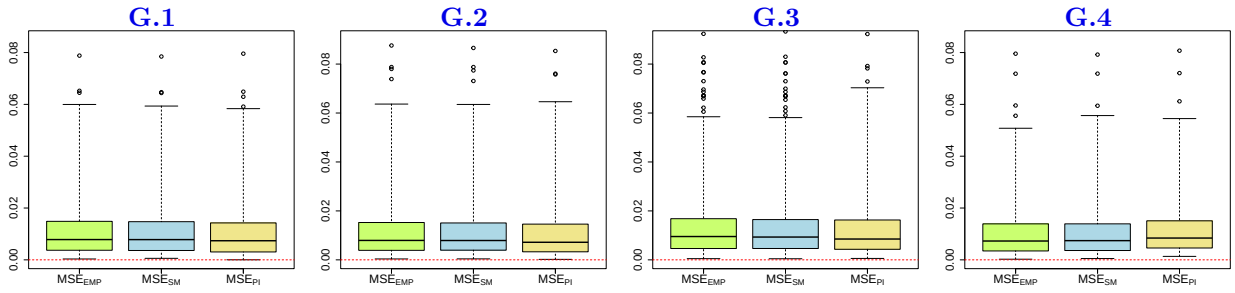


Figura 5.6: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.1** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 100$.

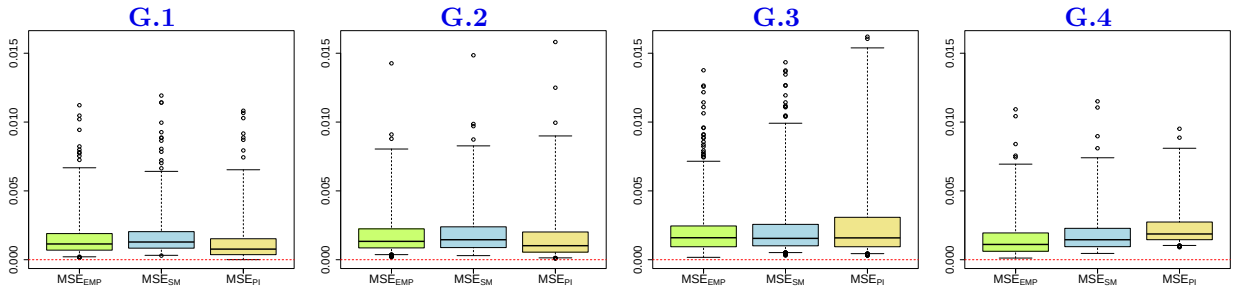


Figura 5.7: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 100$.

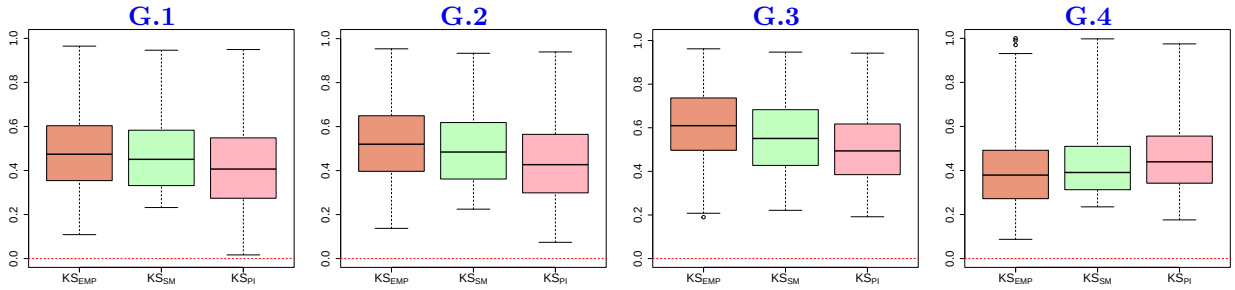


Figura 5.8: Boxplots ajustados de la distancia de Kolmogorov-Smirnov (KS) para los tres estimadores con covariables con distribución **MX.1** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 100$.

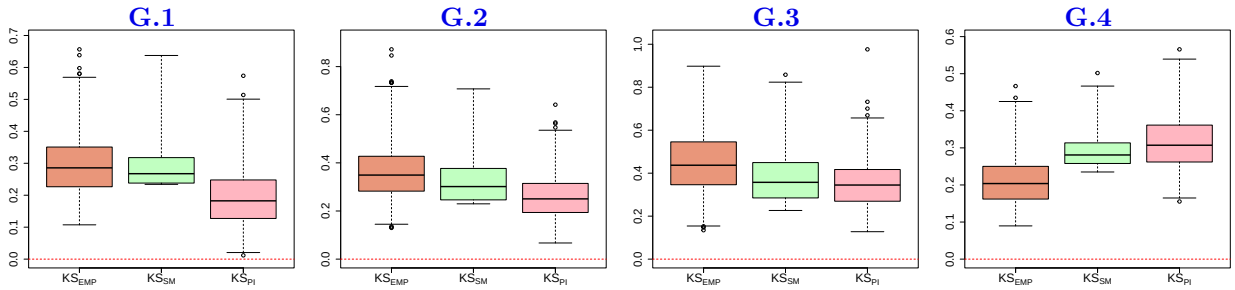


Figura 5.9: Boxplots ajustados de la distancia de Kolmogorov-Smirnov (KS) para los tres estimadores con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 300$.

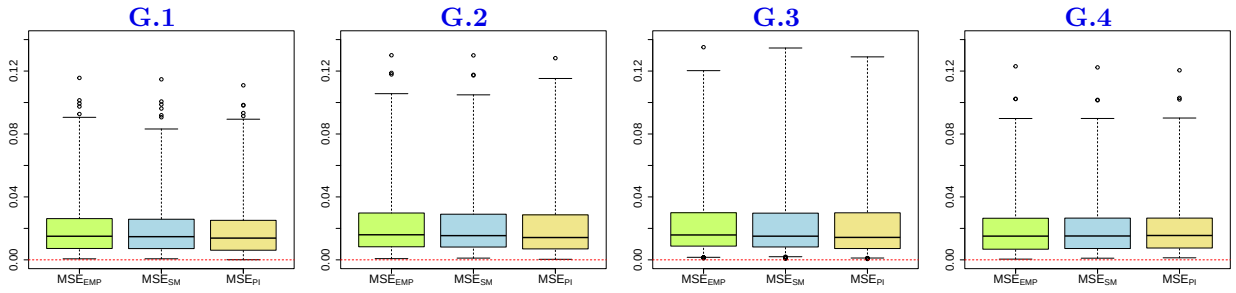


Figura 5.10: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.1** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 50$.

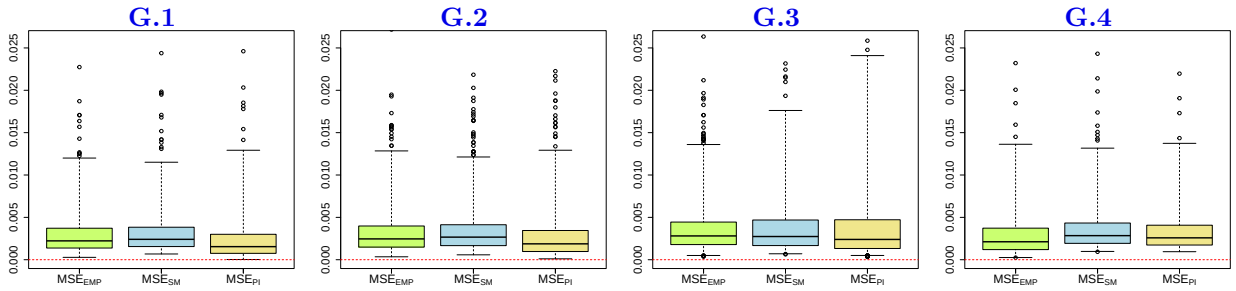


Figura 5.11: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 50$.

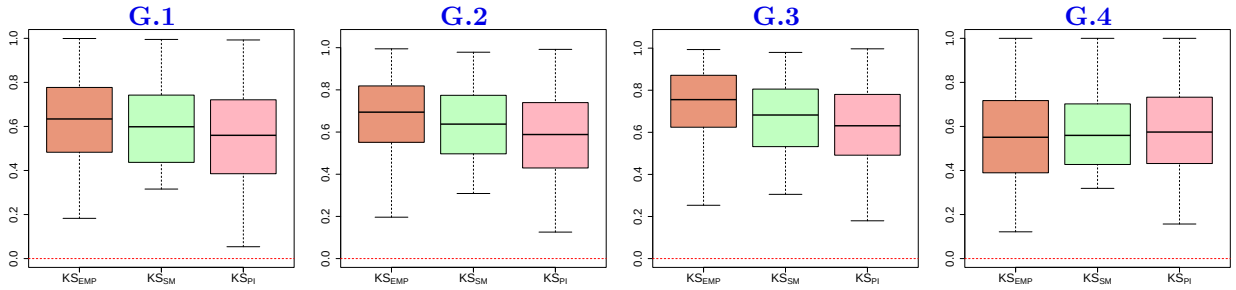


Figura 5.12: Boxplots ajustados de la distancia de Kolmogorov-Smirnov (KS) para los tres estimadores con covariables con distribución **MX.1** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 50$.

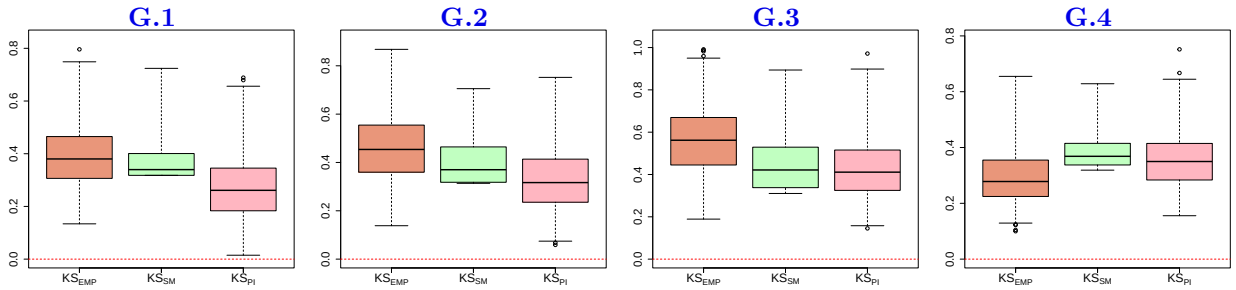


Figura 5.13: Boxplots ajustados de la distancia de Kolmogorov-Smirnov (KS) para los tres estimadores con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**), Logística (**G.2**), t_4 normalizada (**G.3**) y uniformes (**G.4**). En todos los casos, $n_D = n_H = 50$.

Estas figuras permiten ver que, en general, cuando los errores son normales se obtienen los valores más chicos de MSE y KS. Además, cuando $n_H = n_D = 300$, el estimador $\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$ tiene un comportamiento similar o aún mejor en algunos casos (ver Modelo **MX.2**) no sólo en el caso de errores normales, como era de esperar, sino también en el caso logístico. Más aún, cuando ε_j tiene distribución t_4 normalizada (**G.3**), el boxplot de la distancia de Kolmogorov muestra menor dispersón y medianas semejantes o menores a los de $\widehat{\text{ROC}}_{\mathbf{x}}$ y $\widehat{\text{ROC}}_{\mathbf{x},h}$. Este comportamiento no se observa al considerar los boxplots de MSE especialmente en el caso en que las covariables tienen distribución **MX.2**, situación en la que los errores cuadráticos del estimador basado en el modelo binormal son mucho mayores. Estas observaciones muestran que ambas medidas son necesarias para describir globalmente el comportamiento de los estimadores de la ROC condicional. Estas diferencias se atenúan cuando el tamaño de muestra es más chico, pues cuando $n_D = n_H = 50, 100$, el estimador *plug-in* no presenta diferencias tan extremas. En conclusión, como se había observado en las Tablas 5.1 y 5.2, cuando $\varepsilon_j \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$, para $j = D, H$, el estimador $\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$ es mucho peor que sus competidores semiparamétricos, lo cual era de esperar pues la distribución uniforme tiene soporte compacto, aunque esto sólo ocurre de forma notable cuando el tamaño de muestra es relativamente grande. En cuanto a la diferencia entre los estimadores semiparamétricos, es interesante destacar que el estimador suavizado exhibe valores en general menores de KS, así como una menor dispersión de la misma, mientras que el MSE no presenta diferencias notables.

Como hemos mencionado, el área bajo la curva da un indicador de la capacidad discrimatoria de un procedimiento y un resumen global de la curva ROC condicional. Las Figuras 5.14 a 5.19 presentan los boxplots de superficie como fueron definidos en Genton et al. (2014), sobre las 1000 replicaciones para los tres estimadores considerados. Sólo se muestran los resultados para el caso en que los errores tienen distribución normal y uniforme. Como se indicó previamente, la verdadera superficie sobre el simplex $\text{AUC}_{\mathbf{x}}$ está representada en color verde lima, mientras que en el diagrama ternario se presentan en gris los puntos donde se calcularon los estimadores y la verdadera superficie.

Los boxplots de superficies muestran que, como era de esperar, a medida que el tamaño de muestra crece, la variabilidad de los estimadores de la AUC condicional disminuye, pues tanto las bandas que contienen el 50 % de las superficies como los *bigotes* se acercan a la superficie mediana. También es evidente que los tres estimadores se comportan de manera muy similar cuando los errores son normales, pues los boxplots casi no muestran diferencias y para cada valor del tamaño de muestra, la variabilidad de los estimadores para ambas distribuciones de los errores consideradas es comparable. Una característica que sobresale de estos gráficos es que cuando las covariables tienen distribución **MX.2**, la verdadera AUC parecería no estar completamente contenida en la región delimitada por las superficies análogas a los *bigotes* en el caso de los estimadores $\widehat{\text{AUC}}_{\mathbf{x},h}$ y $\widehat{\text{AUC}}_{\mathbf{x},\text{PI}}$. Esto, junto con los resultados anteriores, parecería sugerir que bajo dicho diseño de las covariables las estimaciones son menos estables.

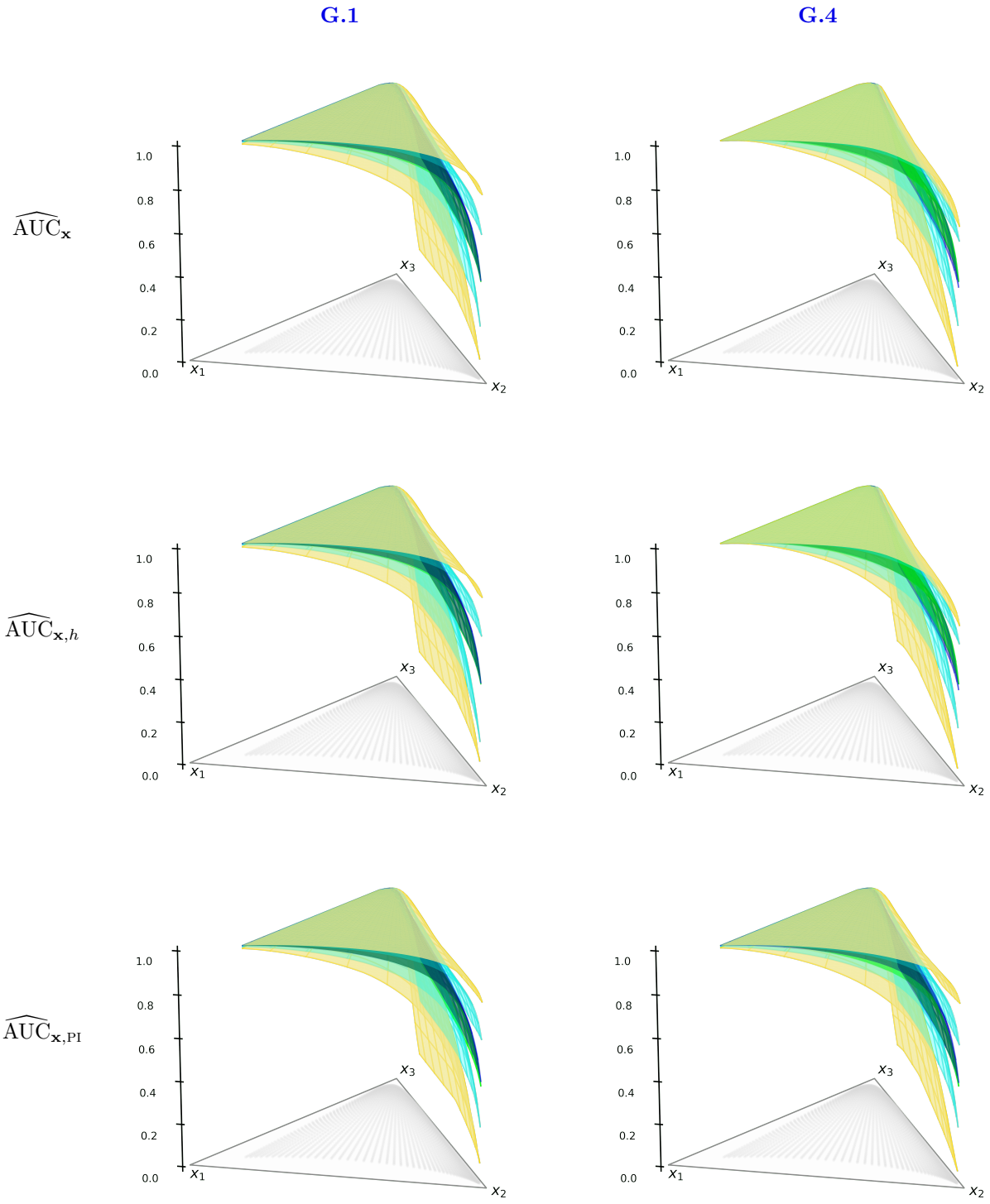


Figura 5.14: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución Dirichlet (**MX.1**) y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 300$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

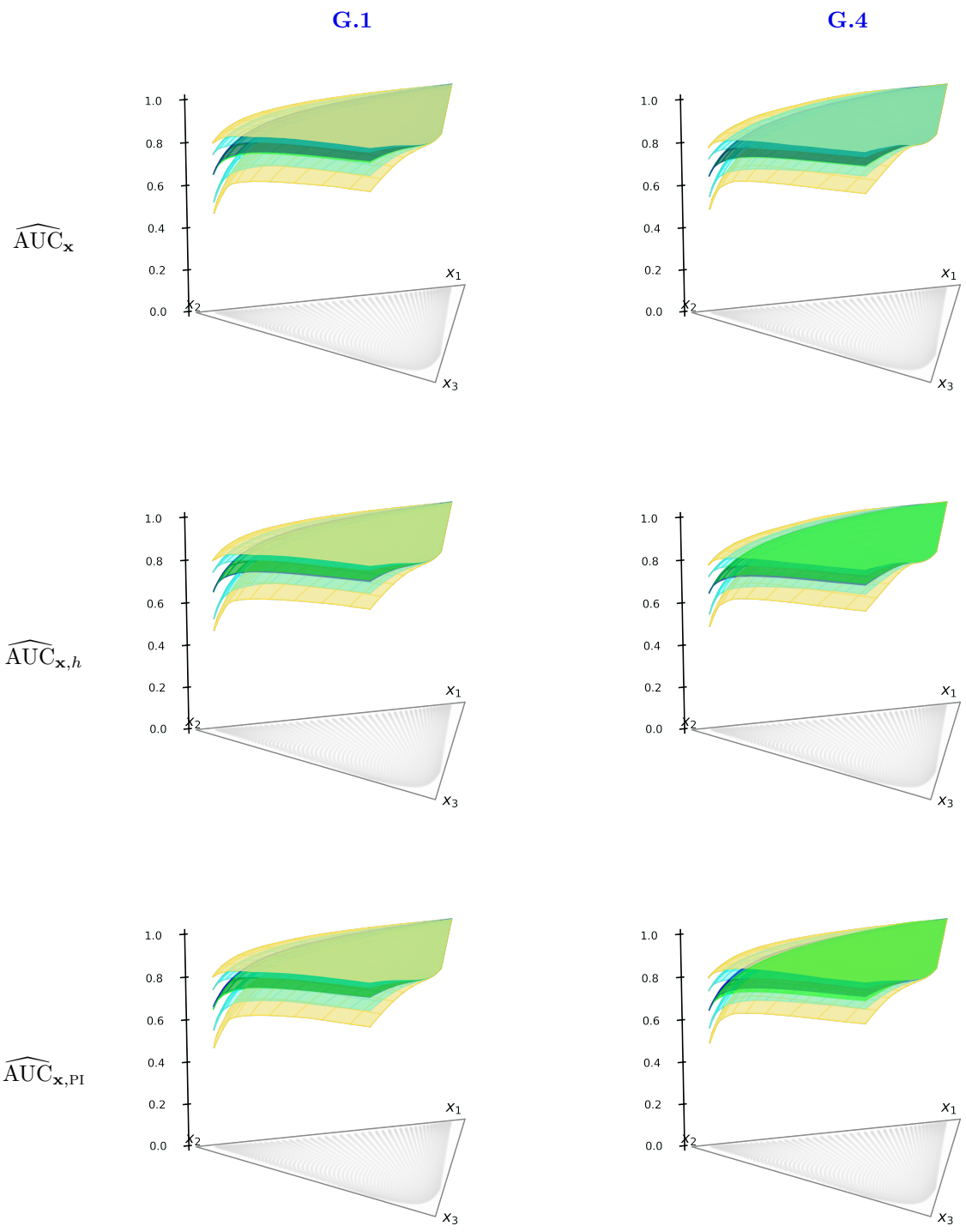


Figura 5.15: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 300$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

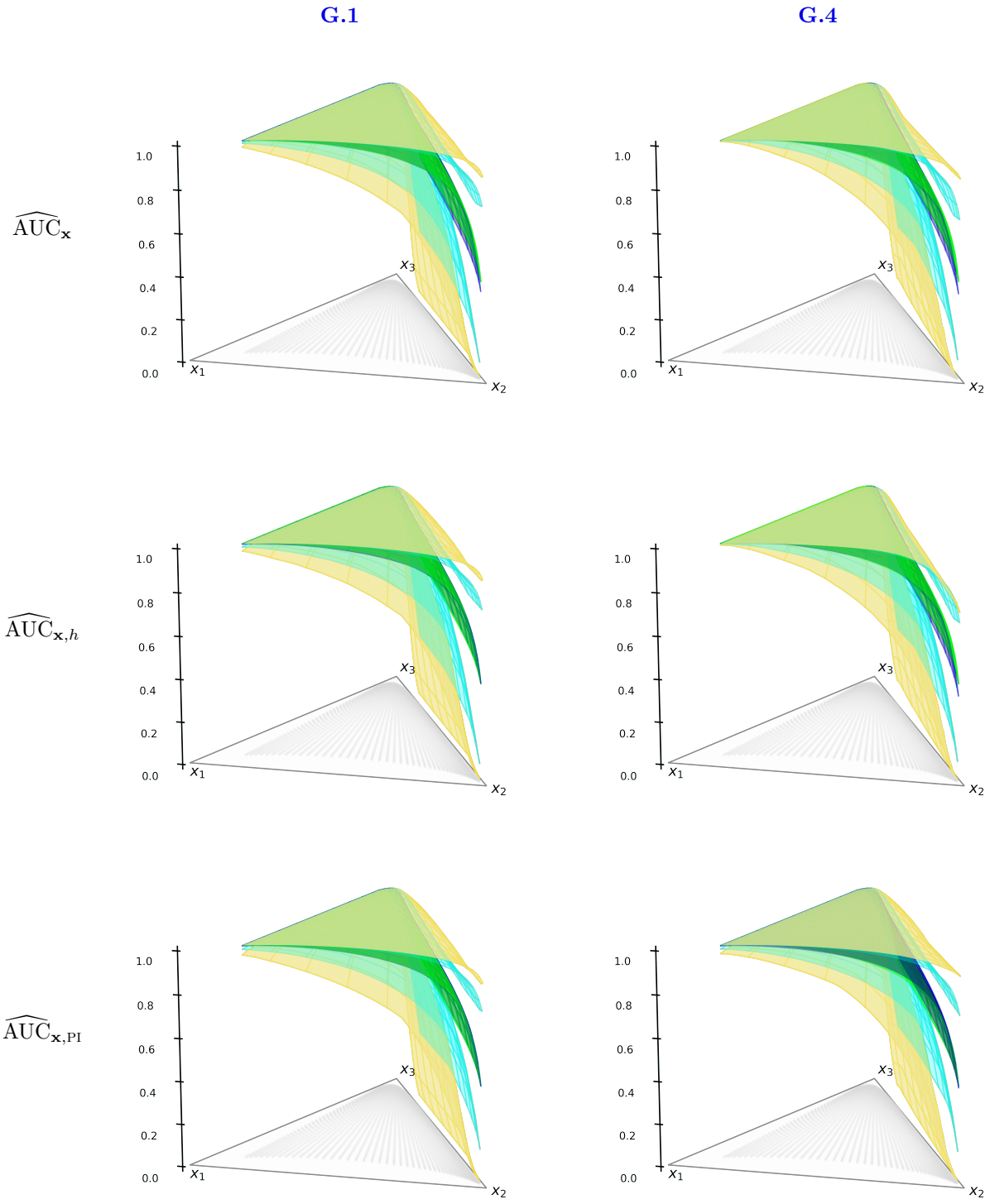


Figura 5.16: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución Dirichlet (**MX.1**) y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 100$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

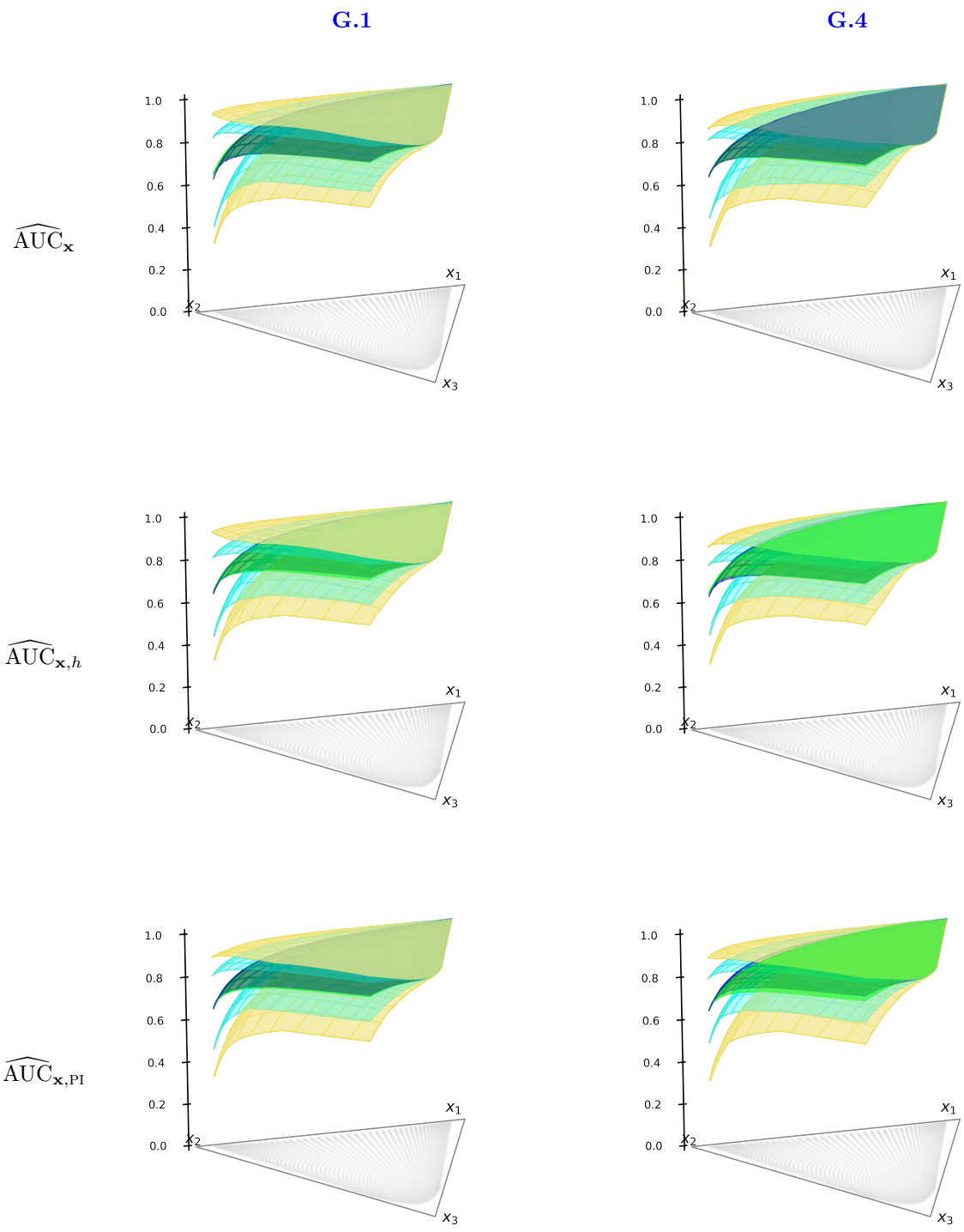


Figura 5.17: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 100$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

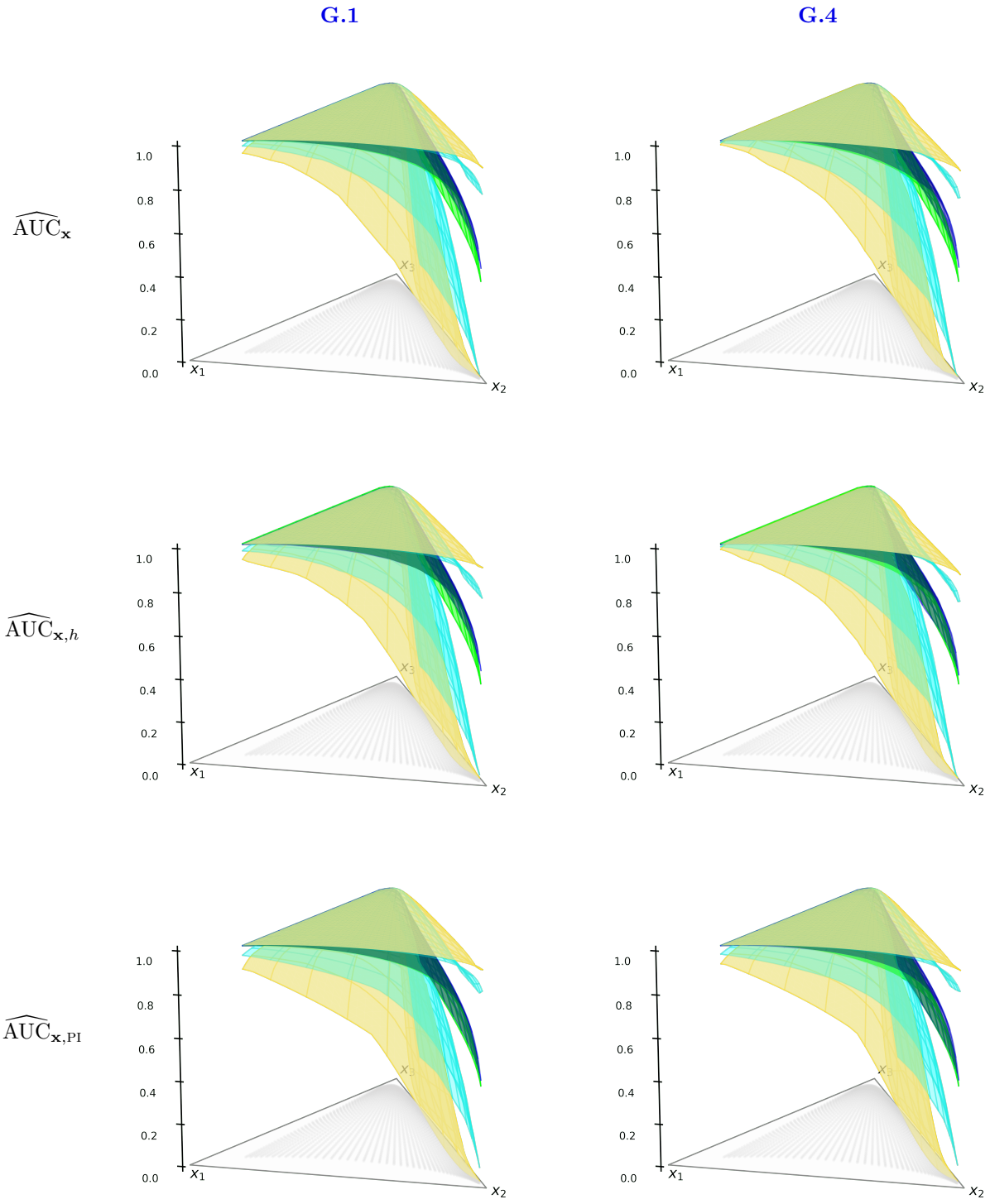


Figura 5.18: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución Dirichlet (**MX.1**) y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 50$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

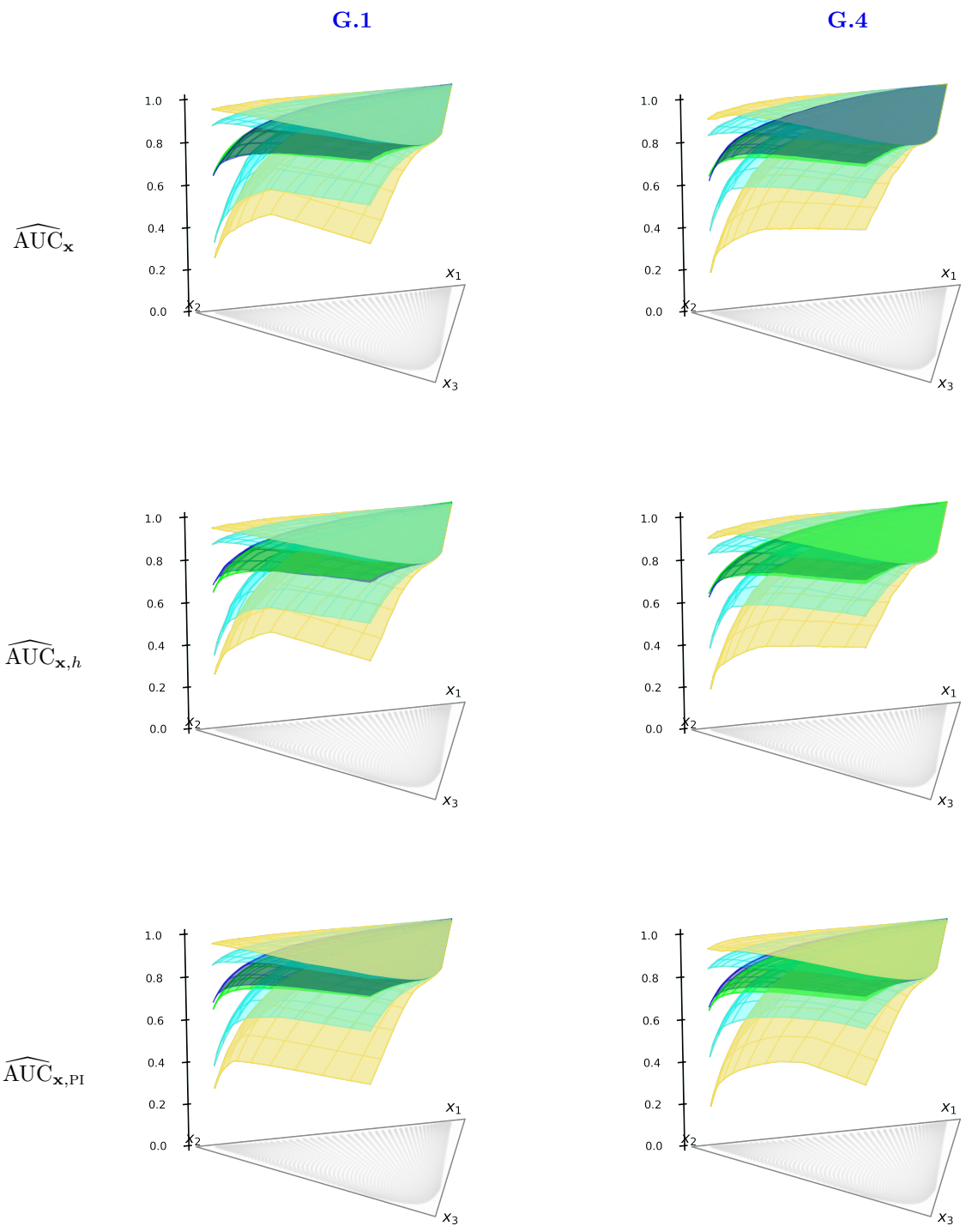
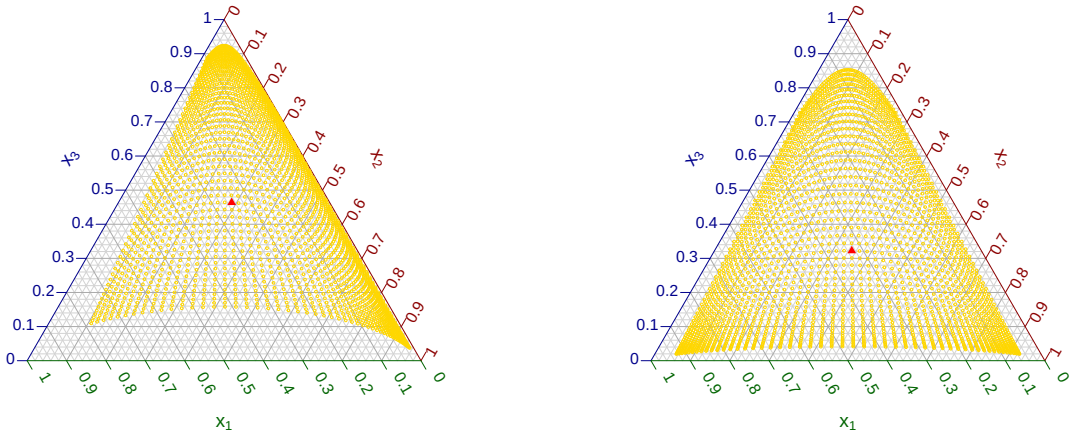


Figura 5.19: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 50$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

Es importante destacar que teniendo en cuenta que la AUC promedia los valores sobre la grilla $\mathcal{G}_p = \{p_j\}_{j=1}^{N_p}$, la ventaja de suavidad del estimador $\widehat{\text{AUC}}_{\mathbf{x},h}$ se pierde en esta visualización. Además, dado que estamos trabajando con covariables de dimensión mayor a 1, no resulta posible visualizar gráficamente la estimación de la curva ROC condicional, ya que cada punto de la grilla en el simplex tiene asociada una curva ROC de dos dimensiones. Por esta razón, calculamos el área bajo la curva estimada y graficamos los boxplots de superficie de esta medida.

Una desventaja a tener en cuenta es que dos curvas ROC podrían reportar el mismo valor de AUC pero ser muy distintas, por lo que si bien el AUC es ampliamente usada, por sí sola no da una visión completa del comportamiento de los estimadores de la ROC condicional. Por esta razón, decidimos graficar las 1000 estimaciones de la curva ROC condicional en cada caso, junto con la verdadera curva para dos puntos particulares de la grilla, graficados en la Figura 5.20. Los gáficos de las Figuras 5.21 a 5.26 muestran que en algunos casos, cuando los errores son uniformes, la verdadera curva se sale de la región de curvas estimadas. Claramente, la verdadera curva $\text{ROC}_{\mathbf{x}}$ presenta puntos de no suavidad, lo cual podría ser la causa de este fenómeno. En otras palabras, el estimador suavizado no captura del todo bien la forma de la curva cuando los errores son uniformes. Podemos ver también cómo el estimador $\widehat{\text{ROC}}_{\mathbf{x},h}$ da una versión suavizada de $\widehat{\text{ROC}}_{\mathbf{x}}$, siendo éstos muy parecidos a lo largo de las 1000 replicaciones en todos los modelos considerados. Por último, y como era de esperar, a medida que el tamaño de muestra crece, las curvas estimadas no sólo se parecen más a la verdadera curva, sino que la variabilidad de estimación es mucho menor.



(a) $\mathbf{x}_0 = (0.2488, 0.2875, 0.4636)^{\mathrm{T}}$. (b) $\mathbf{x}_0 = (0.3291, 0.3486, 0.3222)^{\mathrm{T}}$

Figura 5.20: Para cada modelo de las covariables, **MX.1** (a) y **MX.2** (b), el rombo rojo representa el punto elegido para graficar las 1000 estimaciones de la curva ROC condicional.

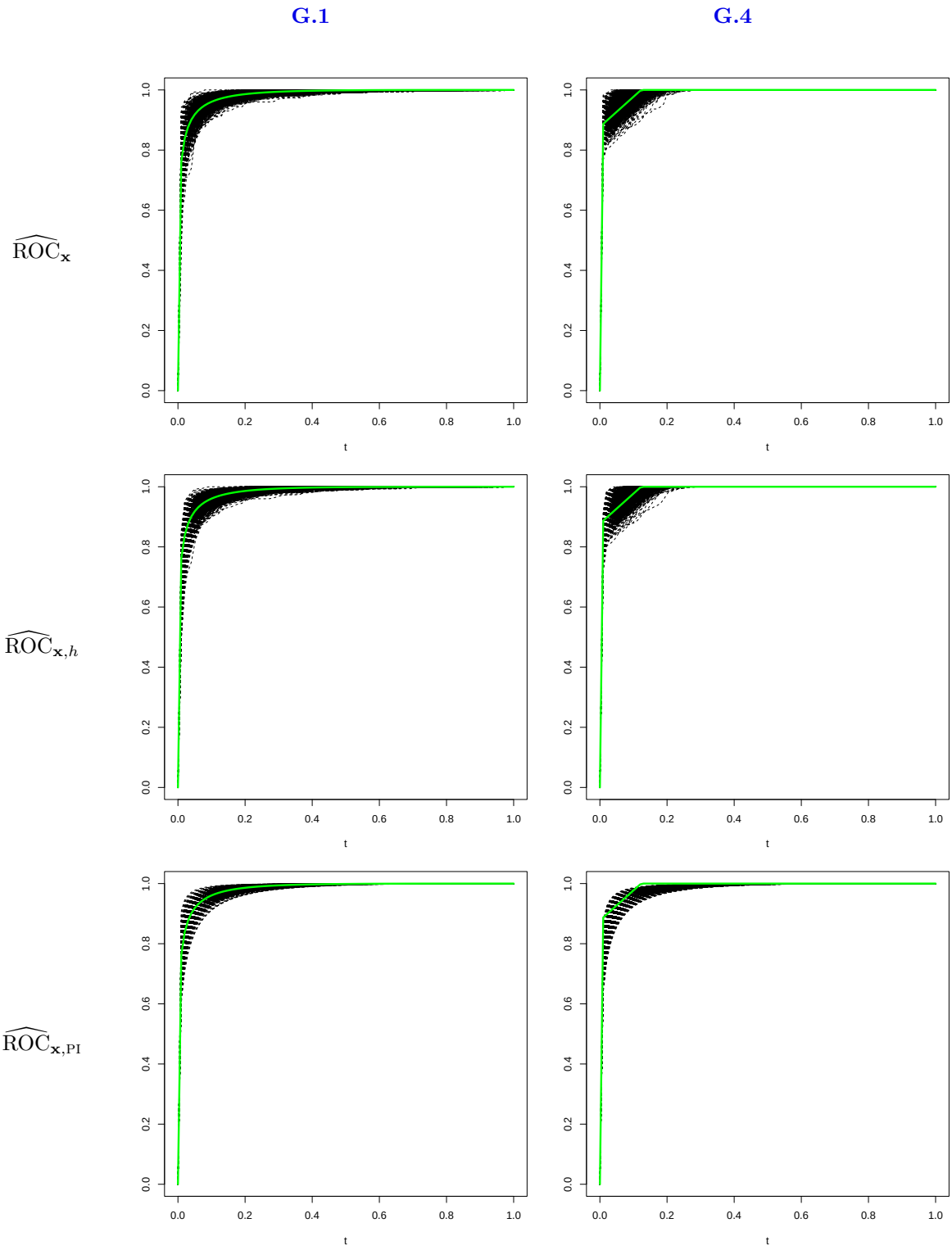


Figura 5.21: Gráfico de las 1000 estimaciones de $\mathrm{ROC}_{\mathbf{x}_0}$ bajo el diseño de covariables **MX.1** y errores con distribución normal estándar (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 300$. La línea punteada corresponde a la verdadera curva ROC condicional a $\mathbf{x}_0$.

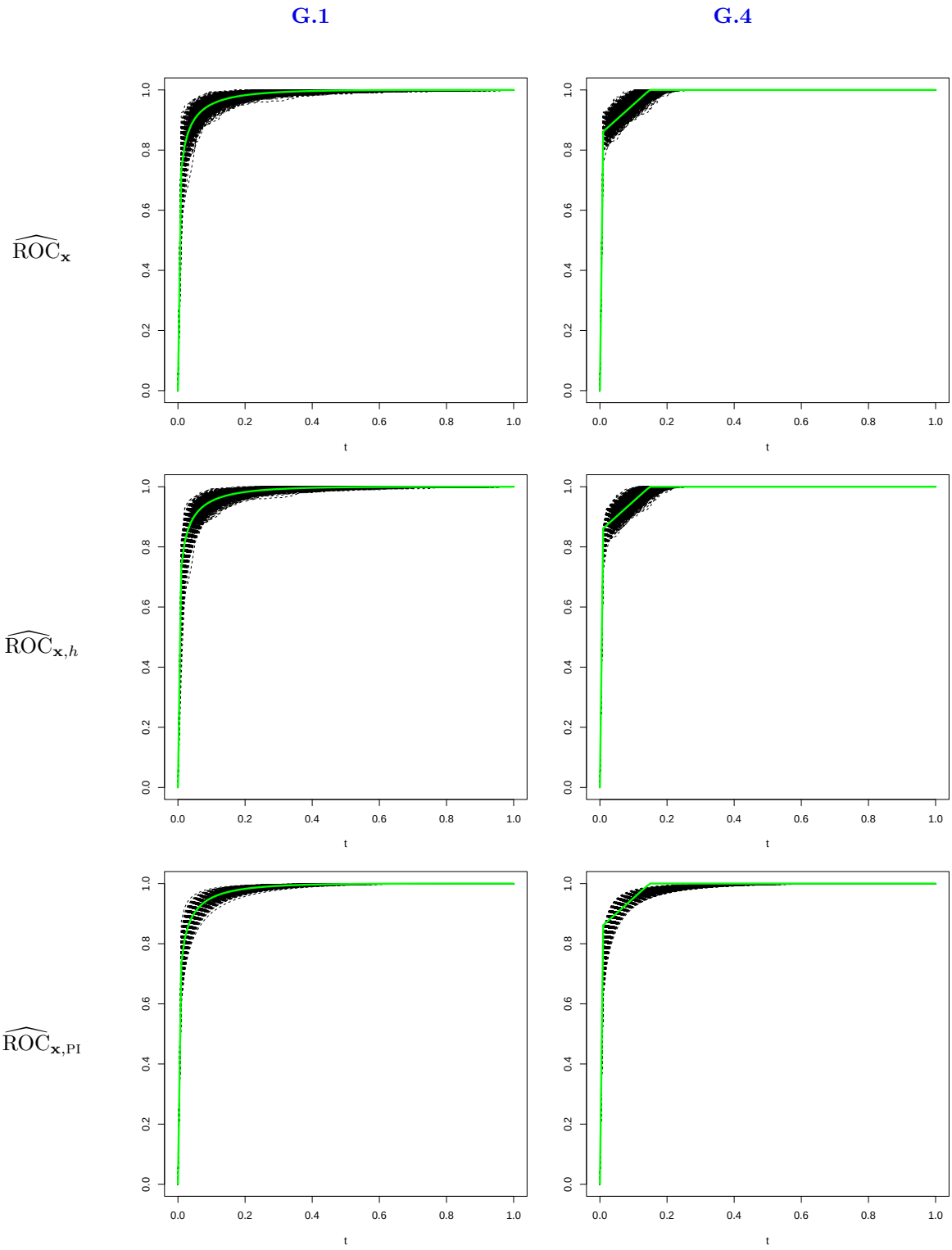


Figura 5.22: Gráfico de las 1000 estimaciones de $\mathrm{ROC}_{\mathbf{x}_0}$ bajo el diseño de covariables **MX.2** y errores con distribución normal estándar (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 300$. La línea punteada corresponde a la verdadera curva ROC condicional a $\mathbf{x}_0$.

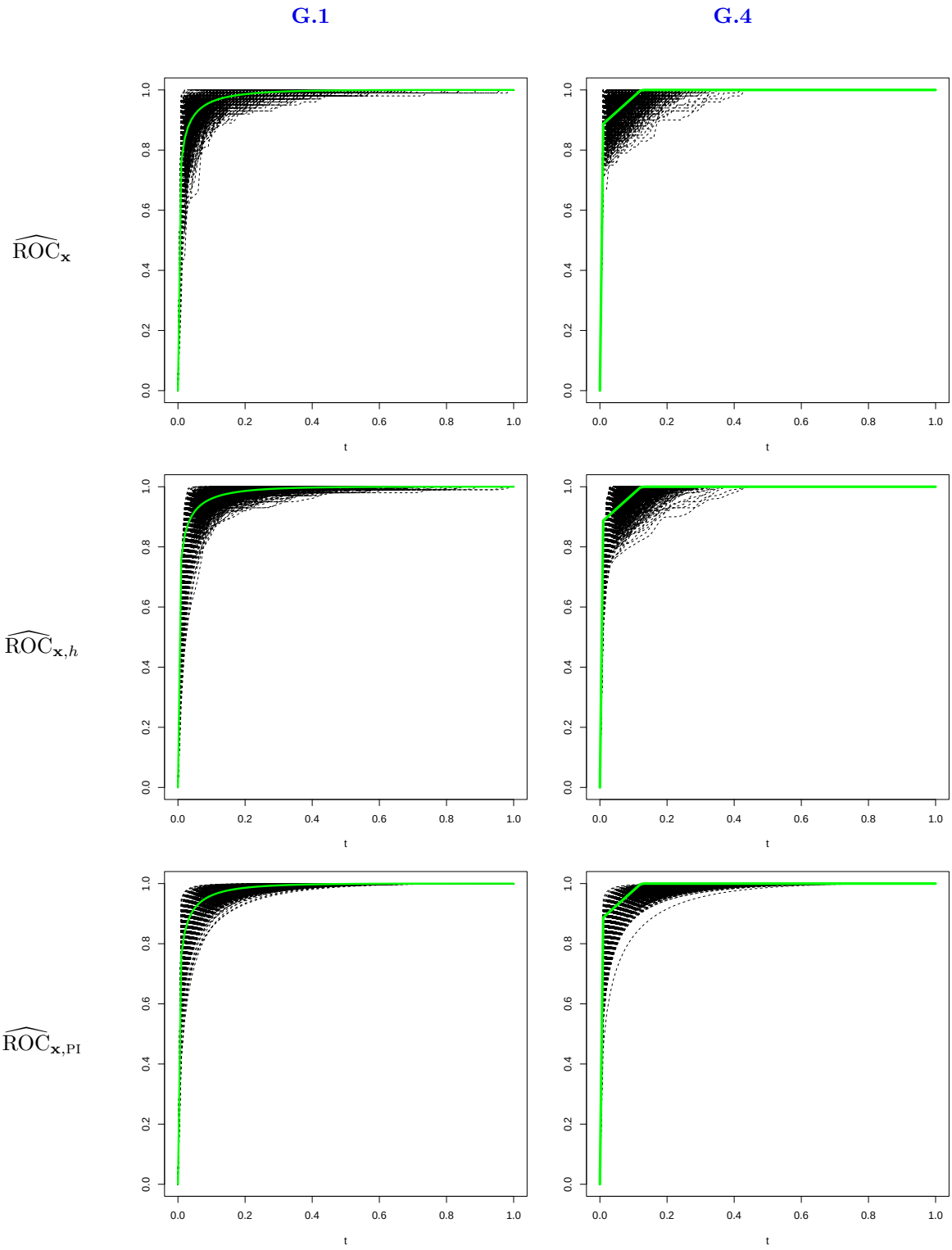


Figura 5.23: Gráfico de las 1000 estimaciones de $\mathrm{ROC}_{\mathbf{x}_0}$ bajo el diseño de covariables **MX.1** y errores con distribución normal estándar (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 100$. La línea punteada corresponde a la verdadera curva ROC condicional a $\mathbf{x}_0$.

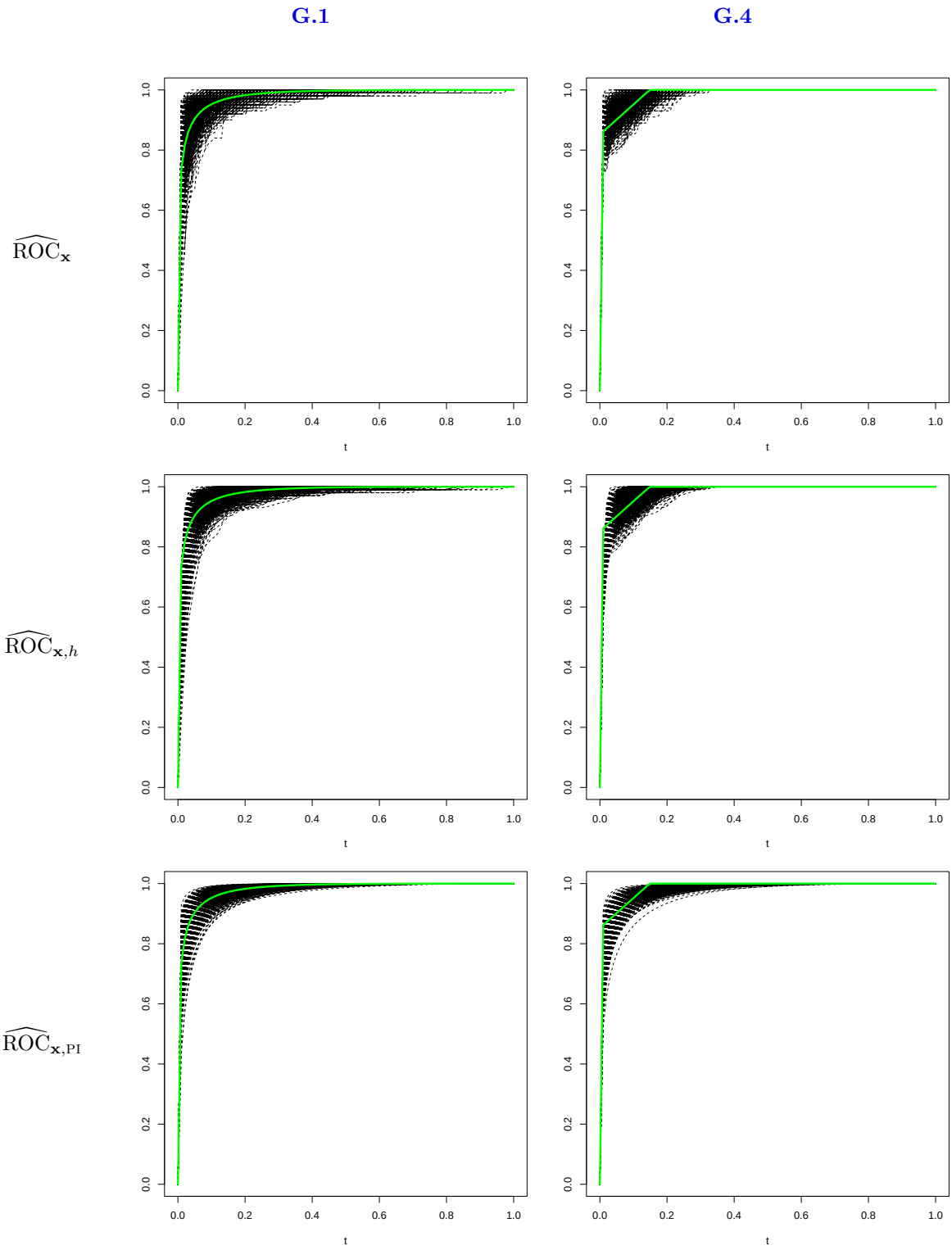


Figura 5.24: Gráfico de las 1000 estimaciones de $\mathrm{ROC}_{\mathbf{x}_0}$ bajo el diseño de covariables **MX.2** y errores con distribución normal estándar (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 100$. La línea punteada corresponde a la verdadera curva ROC condicional a $\mathbf{x}_0$.

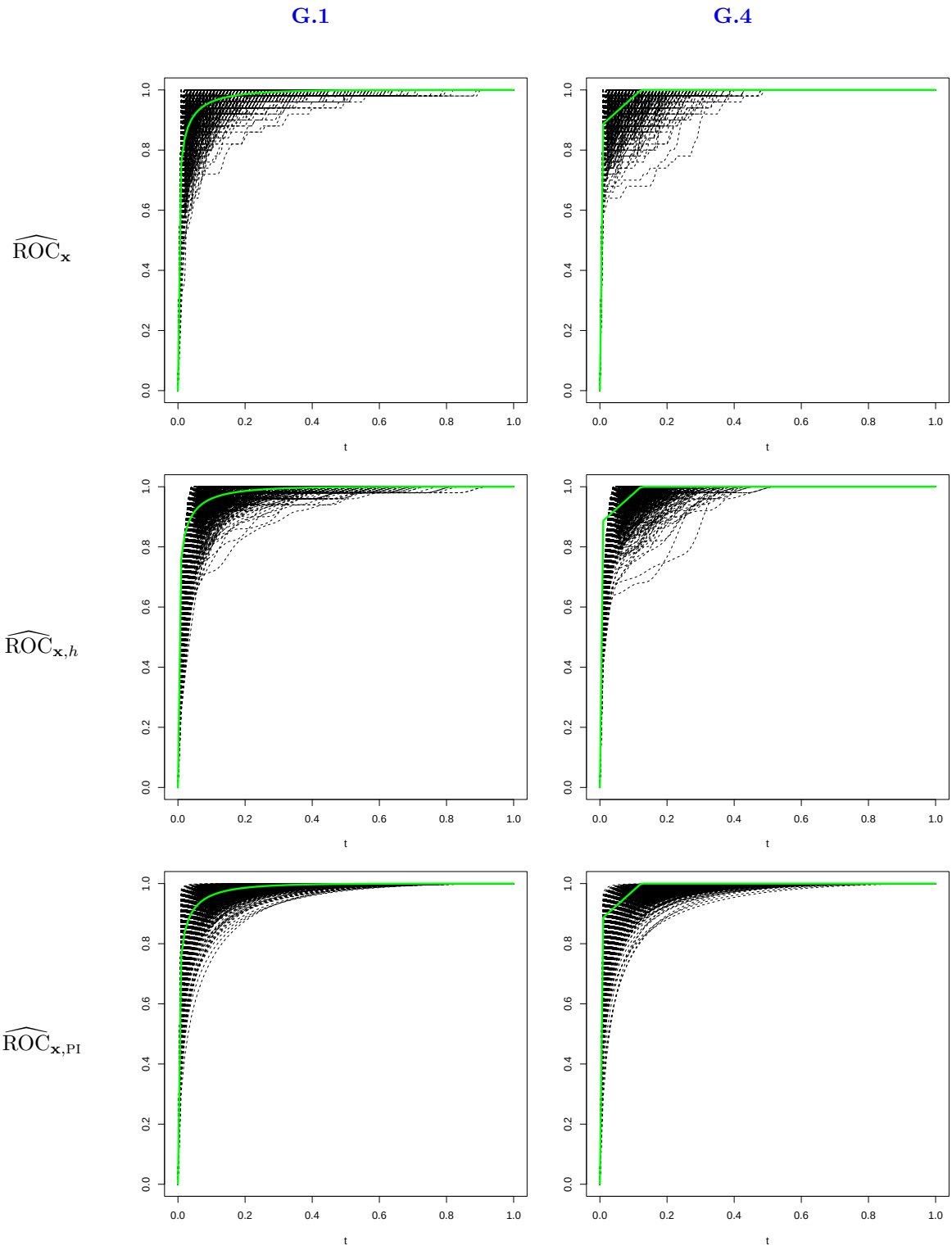


Figura 5.25: Gráfico de las 1000 estimaciones de $\mathrm{ROC}_{\mathbf{x}_0}$ bajo el diseño de covariables **MX.1** y errores con distribución normal estándar (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 50$. La línea punteada corresponde a la verdadera curva ROC condicional a $\mathbf{x}_0$.

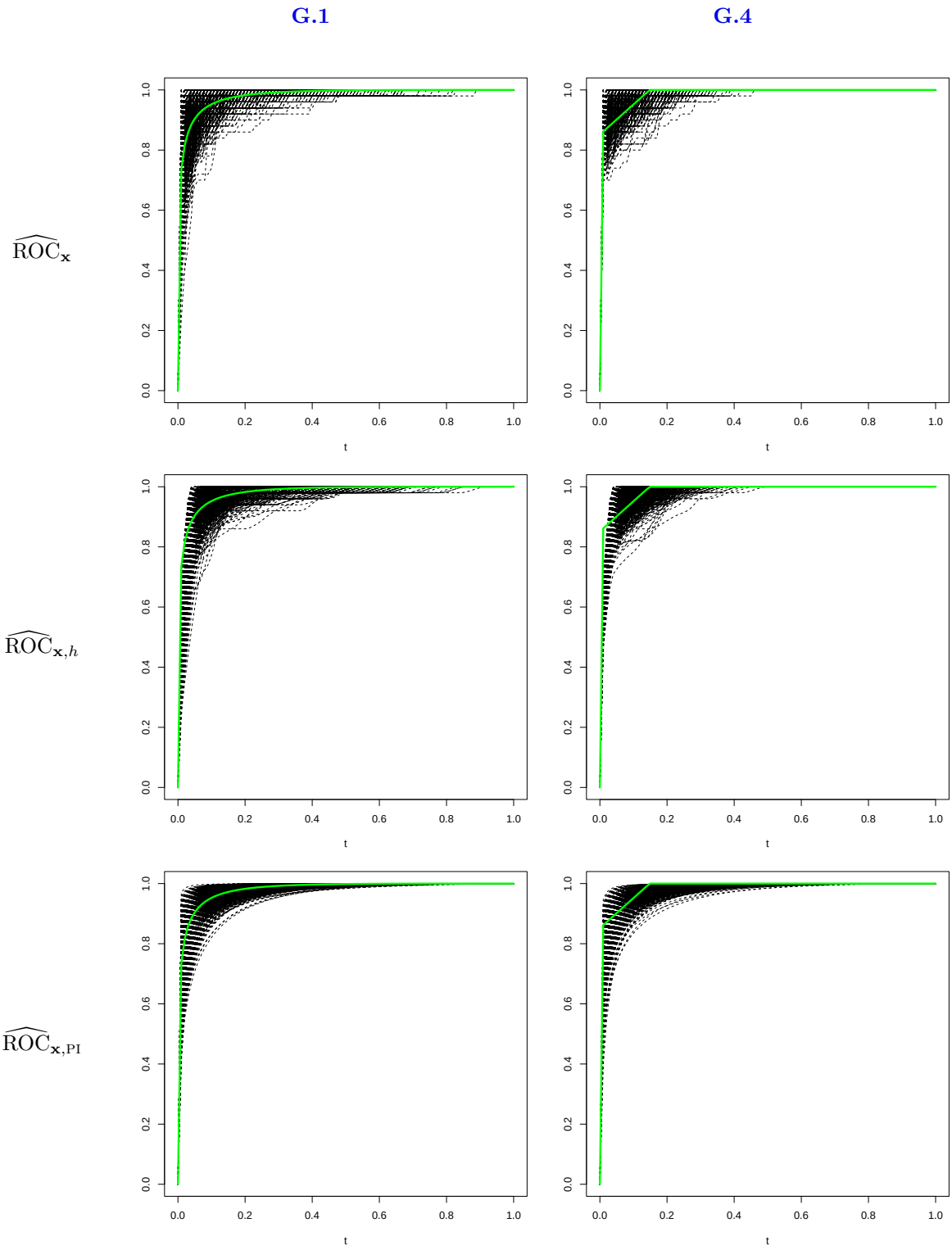


Figura 5.26: Gráfico de las 1000 estimaciones de $\mathrm{ROC}_{\mathbf{x}_0}$ bajo el diseño de covariables **MX.2** y errores con distribución normal estándar (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = n_H = 50$. La línea punteada corresponde a la verdadera curva ROC condicional a $\mathbf{x}_0$.

5.5. Resultados para el caso no balanceado

En este sección reportamos los resultados correspondientes al caso en que las muestras de ambas poblaciones no tienen igual tamaño. Consideramos el caso $n_H = 50$ y $n_D = 100$, así como la situación $n_H = 100$ y $n_D = 50$ y sólo consideraremos el caso en que los errores tienen distribución normal o uniforme, es decir, los modelos **G.1** y **G.4**.

En las Tablas 5.3 y 5.4 se presentan el MSE y KS promedio y sus respectivos desvíos estándar sobre las 1000 iteraciones para cada estimador. Al igual que en el caso balanceado, el escenario en el que las covariables son uniformes da lugar a valores medios del error cuadrático mdeio y de la distancia de Kolmogorov-Smirnov menores. También se puede ver que en todos los casos, el hecho de que el tamaño de muestra en las poblaciones enferma y sana sean diferentes no parece influir en el comportamiento de ninguno de los tres estimadores, pues se obtienen valores de MSE y KS muy similares en ambas situaciones, $n_D = 50, n_H = 100$ y $n_D = 100, n_H = 50$. Esto resulta sorpresivo en especial en el caso del estimador suavizado ya que éste no trata los tamaños de muestra en cada población de forma simétrica.

		G.1					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	100	0.0154 (0.0143)	0.5507 (0.1817)	0.0153 (0.0143)	0.5289 (0.1774)	0.0145 (0.0143)	0.4807 (0.1996)
100	50	0.0158 (0.0146)	0.5660 (0.1848)	0.0156 (0.0146)	0.5416 (0.1720)	0.0149 (0.0146)	0.4952 (0.2035)
		G.4					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	100	0.0148 (0.0140)	0.4852 (0.1960)	0.0149 (0.0140)	0.4981 (0.1740)	0.0155 (0.0138)	0.5244 (0.1788)
100	50	0.0148 (0.0139)	0.4833 (0.1992)	0.0151 (0.0137)	0.5122 (0.1650)	0.0156 (0.0137)	0.5215 (0.1832)

Tabla 5.3: Media y desvío estándar (entre paréntesis en gris) sobre replicaciones de las medidas MSE y KS cuando las covariables tiene distribución **MX.1** y los errores tienen distribución normal estándard (**G.1**) y uniformes (**G.4**).

		G.1					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	100	0.0022 (0.0018)	0.3413 (0.0983)	0.0023 (0.0019)	0.3229 (0.0764)	0.0018 (0.0018)	0.2359 (0.1016)
100	50	0.0022 (0.0018)	0.3469 (0.1044)	0.0025 (0.0019)	0.3488 (0.0604)	0.0018 (0.0018)	0.2405 (0.1011)
		G.4					
		$\widehat{\text{ROC}}_{\mathbf{x}}$		$\widehat{\text{ROC}}_{\mathbf{x},h}$		$\widehat{\text{ROC}}_{\mathbf{x},\text{PI}}$	
n_D	n_H	MSE	KS	MSE	KS	MSE	KS
50	100	0.0021 (0.0018)	0.2546 (0.0813)	0.0025 (0.0018)	0.3209 (0.0564)	0.0027 (0.0015)	0.3328 (0.0804)
100	50	0.0020 (0.0017)	0.2456 (0.0794)	0.0027 (0.0017)	0.3597 (0.0474)	0.0026 (0.0015)	0.3280 (0.0789)

Tabla 5.4: Media y desvío estándar (entre paréntesis en gris) sobre replicaciones de las medidas MSE y KS cuando las covariables tiene distribución **MX.2** y los errores tienen distribución normal estándard (**G.1**) y uniformes (**G.4**).

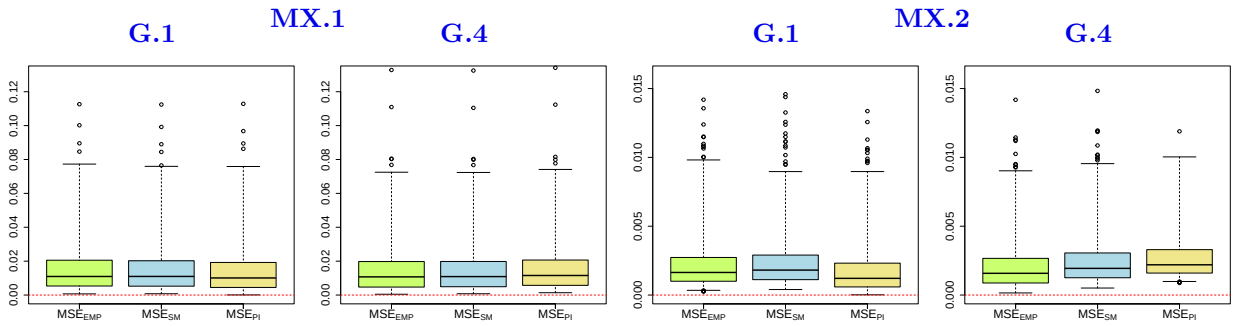


Figura 5.27: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.1** y **MX.2** cuando $\varepsilon_j \sim G_j$ y $G_j \sim \mathcal{N}(0,1)$ (**G.1**) y $G_j \sim \mathcal{U}(-\sqrt{3},\sqrt{3})$ (**G.4**). En todos los casos, $n_D = 50, n_H = 100$.

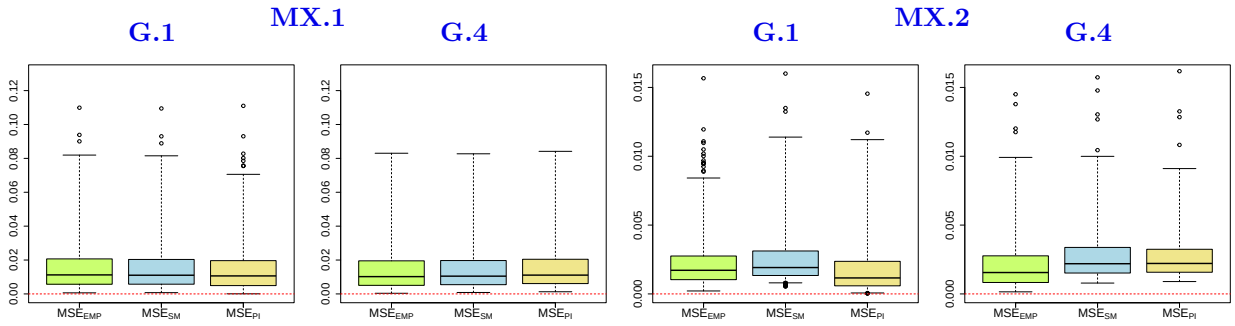


Figura 5.28: Boxplots ajustados del error cuadrático medio (MSE) para los tres estimadores con covariables con distribución **MX.1** y **MX.2** cuando $\varepsilon_j \sim G_j$ y $G_j \sim \mathcal{N}(0,1)$ (**G.1**) y $G_j \sim \mathcal{U}(-\sqrt{3},\sqrt{3})$ (**G.4**). En todos los casos, $n_D = 100, n_H = 50$.

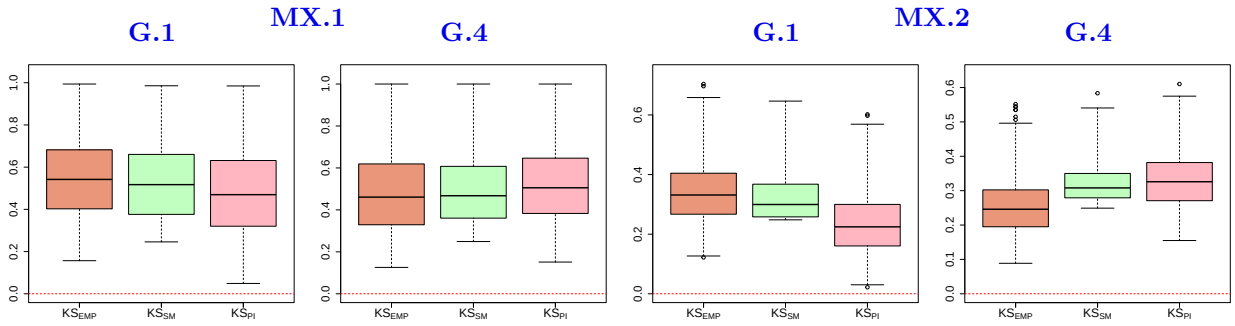


Figura 5.29: Boxplots ajustados de la distancia de Kolmogorov (KS) para los tres estimadores con covariables con distribución **MX.1** y **MX.2** cuando $\varepsilon_j \sim G_j$ y $G_j \sim \mathcal{N}(0,1)$ (**G.1**) y $G_j \sim \mathcal{U}(-\sqrt{3},\sqrt{3})$ (**G.4**). En todos los casos, $n_D = 50, n_H = 100$.

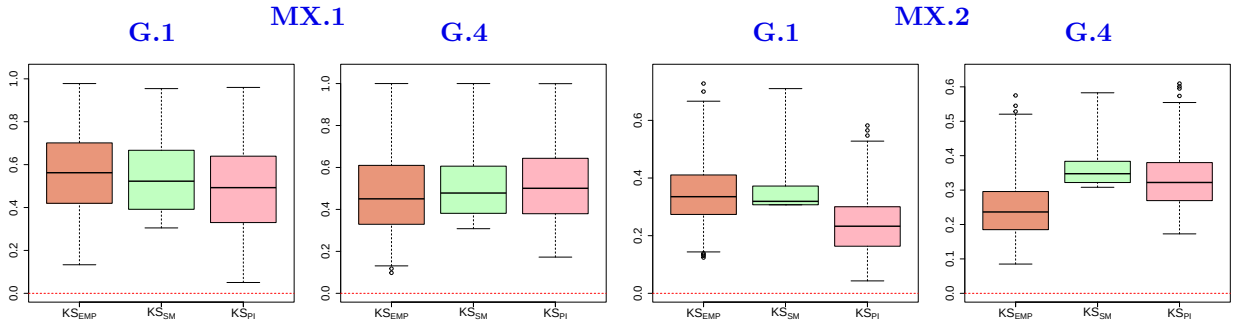


Figura 5.30: Boxplots ajustados de la distancia de Kolmogorov (KS) para los tres estimadores con covariables con distribución **MX.1** y **MX.2** cuando $\varepsilon_j \sim G_j$ y $G_j \sim \mathcal{N}(0,1)$ (**G.1**) y $G_j \sim \mathcal{U}(-\sqrt{3},\sqrt{3})$ (**G.4**). En todos los casos, $n_D = 100, n_H = 50$.

Al igual que antes, para realizar un análisis más detallado de las medidas resumen a lo largo de las 1000 iteraciones, graficamos boxplots ajustados para las dos medidas resumen consideradas, los cuales se presentan en las Figuras 5.27 a 5.30. Observando estas figuras, vemos que, como en el caso balanceado, el estimador basado en el modelo binormal se desempeña igual de bien que sus contendientes semiparamétricos aún cuando la distribución subyacente de los errores se desvía de la normalidad (caso **G.4**). En particular, se observa que cuando los errores son efectivamente normales, el estimador *plug-in* supera a los otros, lo cual era de esperar. Además, el estimador suavizado pareciera presentar valores de KS en general menores que el estimador semiparamétrico basado en empíricas y con menor variabilidad, salvo cuando los errores son uniformes, para ambos escenarios desbalanceados considerados. Una vez más, vemos cómo ambas medidas son necesarias para evaluar globalmente el desempeño de los estimadores.

Como en el caso balanceado, las Figuras 5.31 a 5.34 presentan los boxplots de superficie para los estimadores considerados calculados sobre la grilla de puntos indicada en gris en el diagrama ternario. Nuevamente, la superficie mediana y la verdadera superficie se representan en verde oscuro y verde lima, respectivamente. Estos gráficos permiten apreciar una vez más que las estimaciones basada en los tres procedimientos aplicados no presentan diferencias notables.

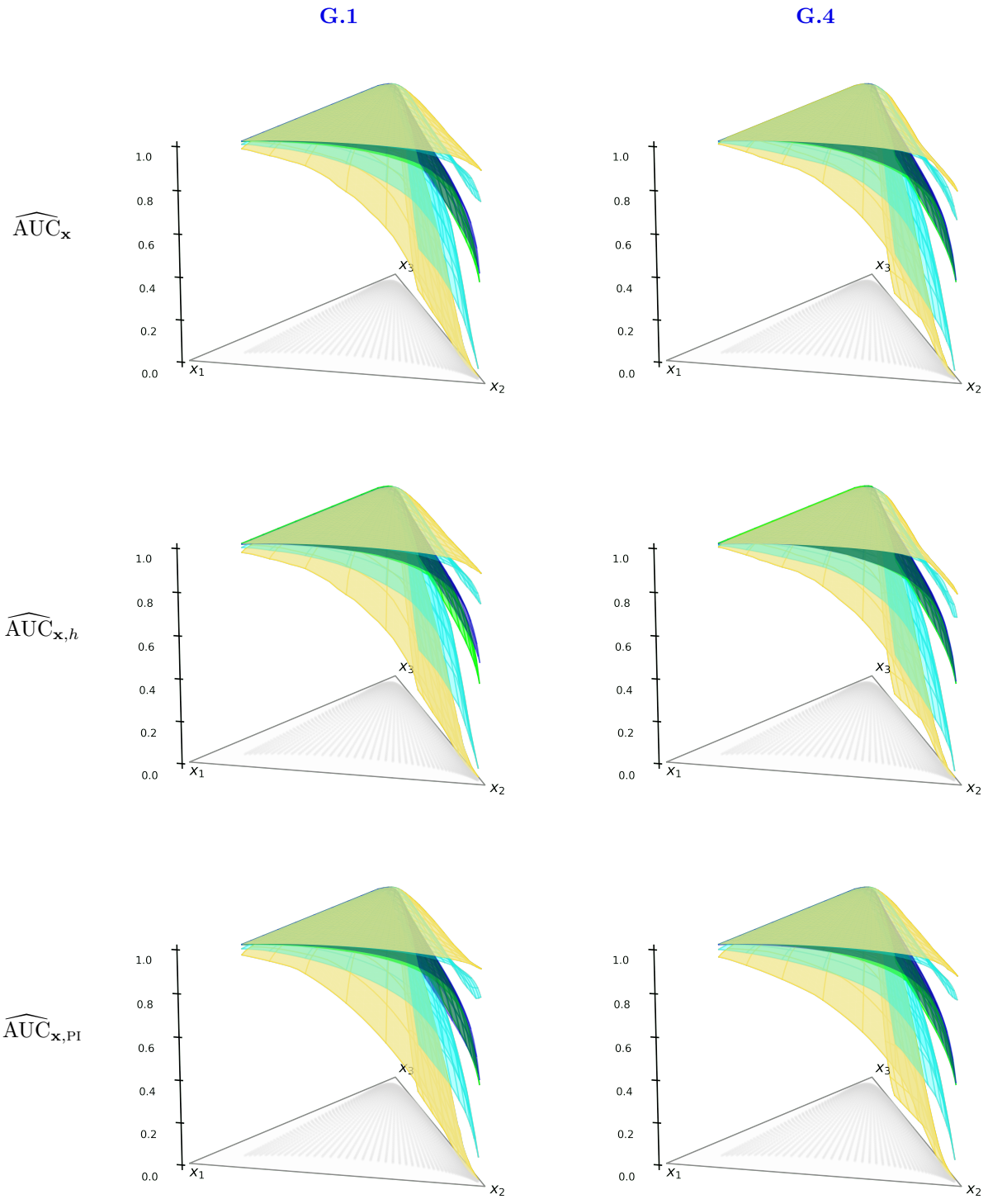


Figura 5.31: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución Dirichlet (**MX.1**) y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = 100$ y $n_H = 50$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

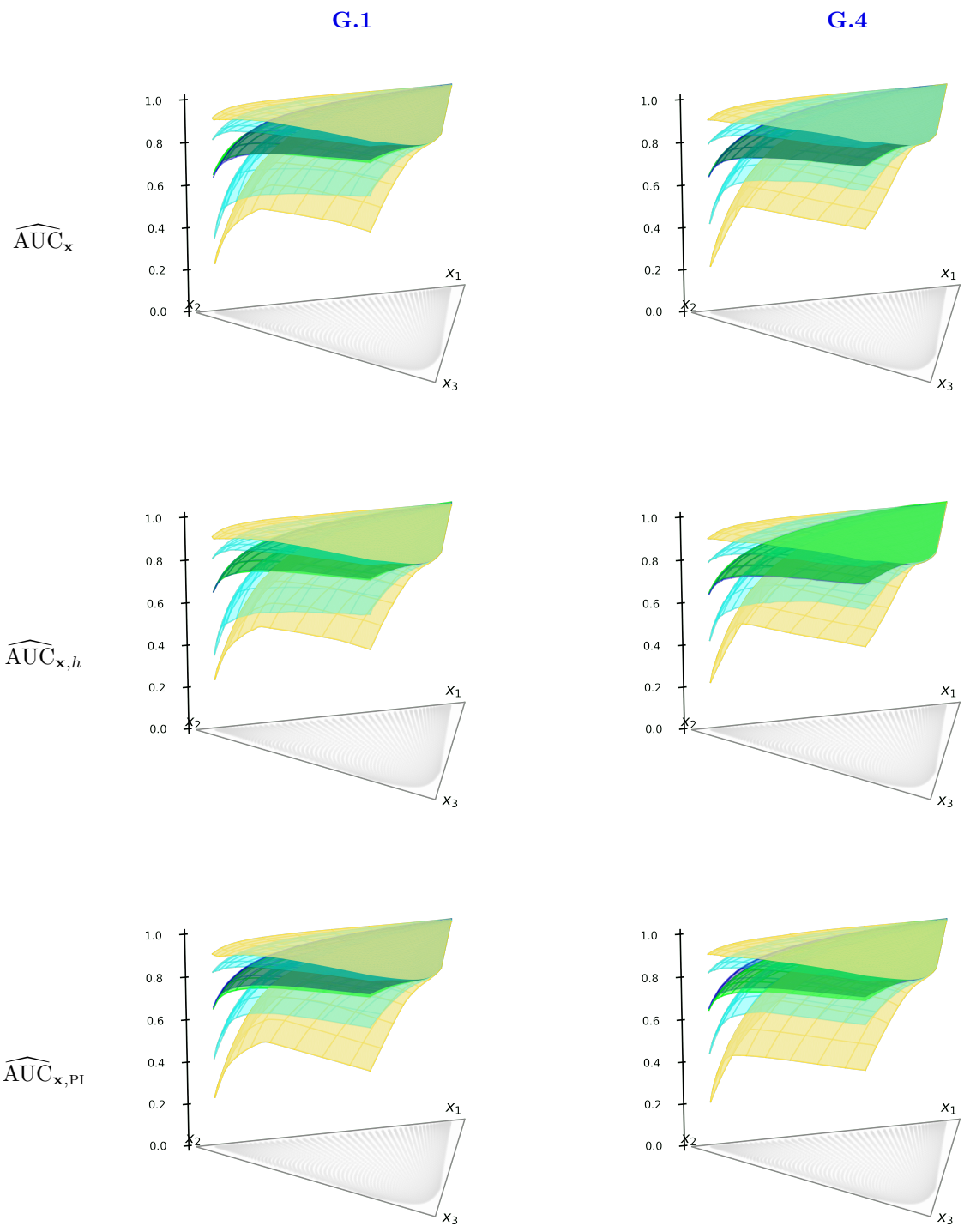


Figura 5.32: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = 100$ y $n_H = 50$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

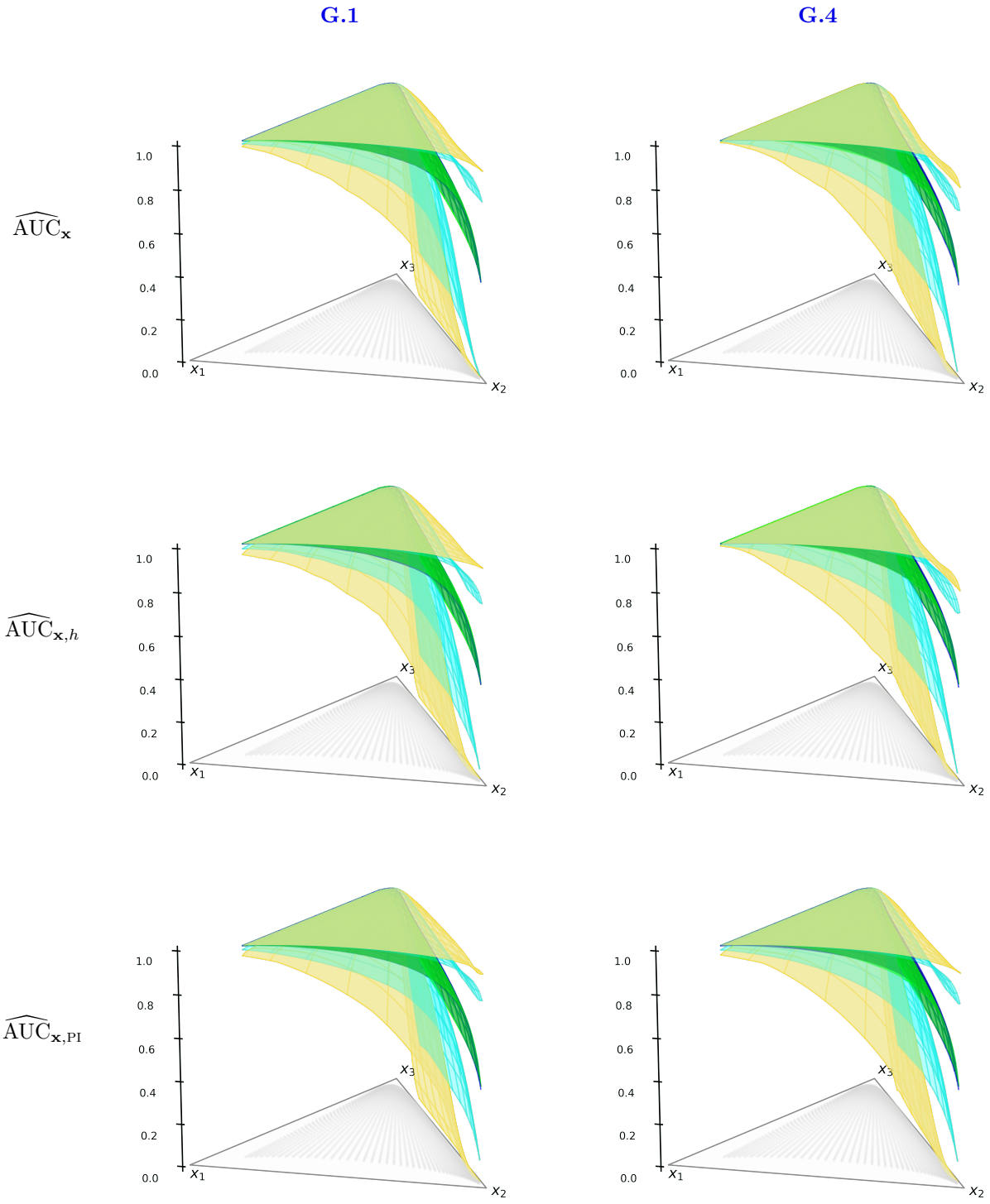


Figura 5.33: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución Dirichlet (**MX.1**) y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = 50$ y $n_H = 100$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

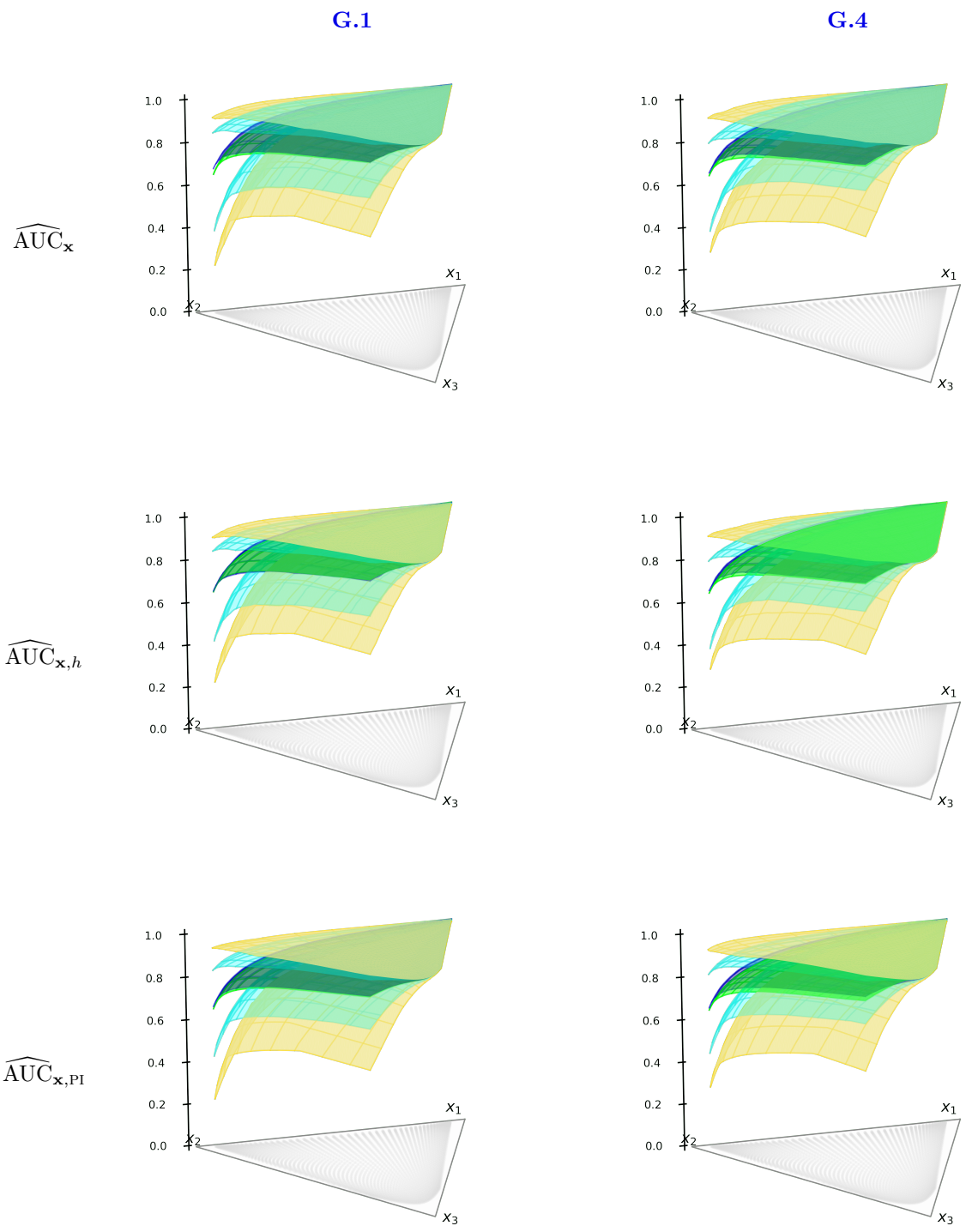


Figura 5.34: Boxplots de superficie para los tres estimadores de $\mathrm{AUC}_{\mathbf{x}}$ con covariables con distribución **MX.2** y errores con distribución normal estándard (**G.1**) y uniforme (**G.4**). En todos los casos, $n_D = 50$ y $n_H = 100$. La superficie mediana se presenta en azul oscuro, mientras que la verdadera superficie se muestra en color verde lima.

Capítulo 6
Aplicación a Datos Reales

En este capítulo, aplicaremos las técnicas desarrolladas a lo largo de la tesis para evaluar la capacidad discriminatoria de un índice glicémico utilizado para distinguir a pacientes con diabetes de aquellos sanos y determinar si el hecho de conocer información adicional sobre la composición de la dieta de los individuos puede mejorar el desempeño de dicho biomarcador.

6.1. El conjunto de datos

Los datos que analizaremos provienen del *Estudio A Estrada de Glicación e Inflamación* (AEGIS), prueba NCT01796184 en www.clinicaltrials.gov, un estudio clínico llevado a cabo en el noroeste de España entre los años 2012 y 2015 que tuvo como objetivo investigar la relación entre los índices de variabilidad glicémica y factores demográficos en una población adulta. En el mismo, participaron un total de 1516 participantes, de los cuales 622 fueron sometidos a un protocolo de monitoreo continuo del nivel de glucosa. De éstos, sólo 580 terminaron dicho protocolo de medición de forma exitosa y proporcionaron datos analizables. Además de estas mediciones, se les pidió a los individuos que, durante seis días consecutivos, registraran su dieta al momento de ingerir alimentos y bebidas. Esto permitió calcular la ingesta de energía y la composición de macronutrientes (carbohidratos, lípidos y proteínas) y de micronutrientes para cada comida del día para cada individuo. Más información acerca del estudio puede consultarse en Gude et al. (2017).

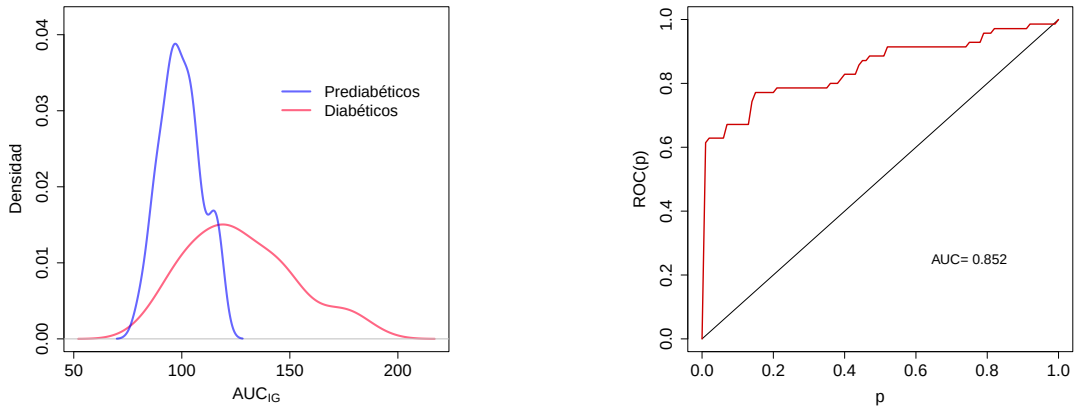
La variable continua que oficia de biomarcador y que utilizaremos como medida de la variación de glucosa es el área bajo la curva (AUC) de las excursiones de glucosa, es decir, el área bajo la curva que resulta de medir las fluctuaciones diarias de glucosa en un paciente. Es importante no confudir esta AUC con el área bajo la curva ROC. Es por eso que denotaremos por AUC_{IG} al área bajo la curva utilizada como índice glicémico. En el estudio descripto anteriormente, también fueron consideradas otras medidas de la variación glicémica como la amplitud media de excursiones glicémicas (MAGE) o la media de diferencias diarias (MODD), pero en este análisis nos centraremos en la variable continua AUC_{IG}.

6.2. Curva ROC de AUC_{IG}

Como la curva ROC permite analizar la capacidad de un biomarcador para clasificar en una de los clases, las dos poblaciones de interés en este capítulo serán los pacientes prediabéticos y los diabéticos, clasificados de acuerdo a la *American Diabetes Association*,

como en Bianco et al. (2024). Los pacientes sanos fueron excluídos del análisis por motivos que serán explicados más adelante. Dado que en promedio, los pacientes diabéticos presentan valores mayores de AUC_{IG}, éstos cumplirán el rol de población enferma, mientras que los pacientes prediabéticos conformarán la población sana.

A continuación, graficamos la curva ROC para la AUC_{IG} como biomarcador, junto con estimaciones por núcleos de las densidades del biomarcador en cada población (utilizando el núcleo gaussiano).



(a) (b)

Figura 6.1: Gráfico de las estimaciones por núcleos de las densidades de la AUC_{IG} en cada población (a) y la curva ROC no condicional para este biomarcador (b).

Como vemos, la capacidad discriminatoria de esta variable en sí es muy alta, y esto se evidencia en un valor de la AUC de la curva ROC de 0.852. En lo que sigue, nos proponemos estudiar si la capacidad discriminatoria mejora o no al considerar como covariable a la composición dietaria del individuo.

6.3. Composición dietaria como covariable

Como mencionamos más arriba, además de medir las fluctuaciones de glucosa, en el estudio se registró la composición dietaria de cada individuo. En este análisis consideraremos la composición de los macronutrientes de la dieta de los pacientes como un vector composicional $\mathbf{X} = (P, L, C)$, dado por la proporción de Proteínas (P), Lípidos (L) y Carbohidratos (C) ingeridas. La Figura 6.2 muestra las covariables composicionales en cada población, graficadas tanto en el simplex a través del diagrama ternario como en el espacio *ilr*. Llamaremos $\mathbf{X}_H$ a la composición de macronutrientes de los prediabéticos y $\mathbf{X}_D$ a la de los diabéticos. Los tamaños de muestra para este conjunto de datos son $n_D = 70$ y $n_H = 79$, habiendo descartado un dato entre los individuos prediabéticos que fue detectado como atípico al considerar el boxplot de los residuos obtenidos mediante un ajuste robusto del modelo de regresión lineal.

Al igual que en el Capítulo 5, las coordenadas *ilr* utilizadas en este análisis son, en nuestro caso particular, las dadas por

$$x_1^* = \sqrt{\frac{1}{2}} \log\left(\frac{P}{L}\right) \qquad x_2^* = \sqrt{\frac{2}{3}} \log\left(\frac{\sqrt{P\,L}}{C}\right) .$$

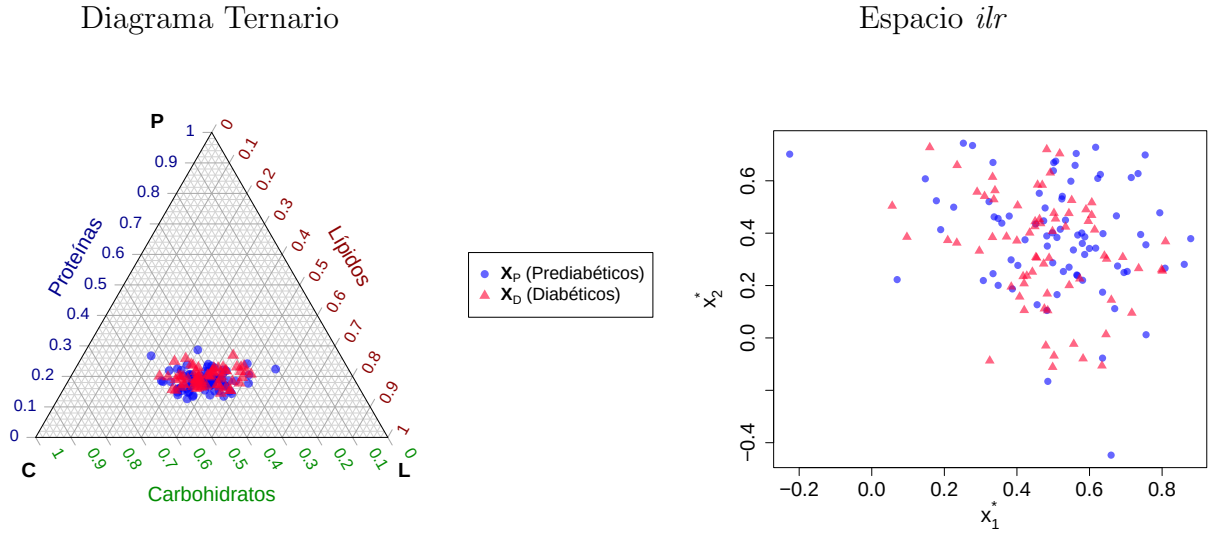


Figura 6.2: Covariables en cada población, graficadas en el diagrama ternario (panel izquierdo) y en el espacio *ilr* (panel derecho).

Por otro lado, para estimar la curva ROC condicional, $\text{ROC}_{\mathbf{x}}$, y siguiendo la metodología inducida, ajustamos a los datos los modelos lineales homoscedásticos

$$\begin{cases} \text{AUC}_{\text{IG, D}} &= \beta_{0,D} + \langle \boldsymbol{\beta}_D, \mathbf{X}_D \rangle_a + \sigma_D \, \varepsilon_D \\ \text{AUC}_{\text{IG, H}} &= \beta_{0,H} + \langle \boldsymbol{\beta}_H, \mathbf{X}_P \rangle_a + \sigma_H \, \varepsilon_H \, , \end{cases} \tag{6.1}$$

donde AUC_{IG} es tomada como variable respuesta y, como antes, los errores se asumen independientes de las covariables y con distribución centrada en 0 y con varianza igual a 1. Los subíndices indican la población, prediabéticos (P) y diabéticos (D).

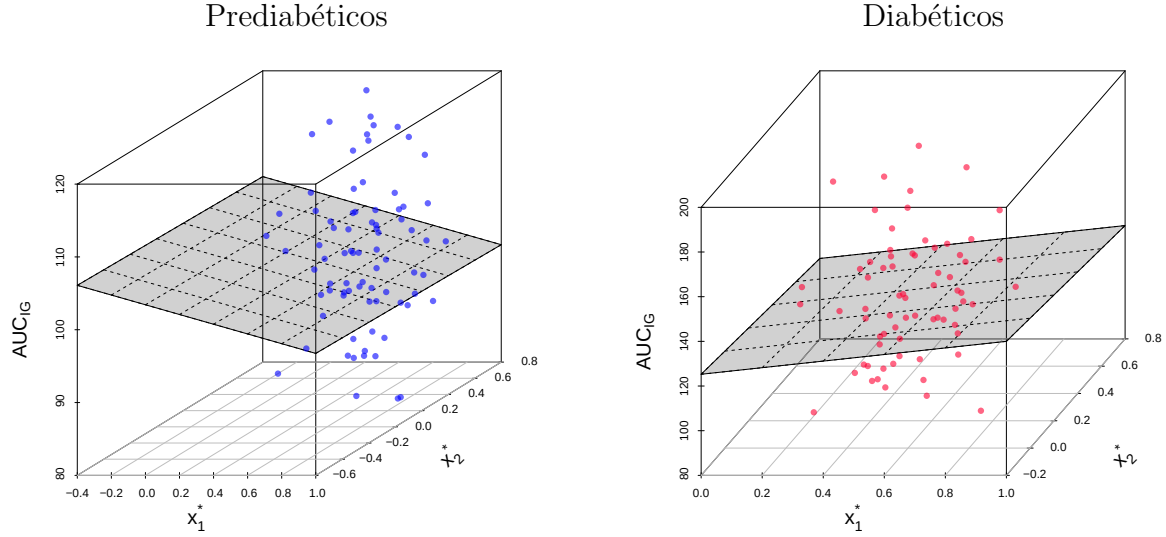


Figura 6.3: Ajuste de los modelos en (6.1) en las población de prediabéticos (panel izquierdo) y diabéticos (panel derecho) en el espacio *ilr* de las covariables.

En la Figura 6.3 se presentan los ajustes de los modelos lineales dados en (6.1) en el espacio de las coordenadas *ilr* de $\mathbf{X}$. El modelo lineal parece una buena forma de modelar

la relación entre el biomarcador y las covariables. Para mejorar la visualización del ajuste, indiquemos por $\widehat{\mathrm{AUC}}_{\mathrm{IG}}$ el predictor de $\mathrm{AUC}_{\mathrm{IG}}$, es decir, $\widehat{\mathrm{AUC}}_{\mathrm{IG,\,j}} = \widehat{\beta}_{0,j} + \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{X}_j \rangle_a$, para $j = D, P$. Para cada población, presentamos en la Figura 6.4 los residuos estandarizados del ajuste, es decir, para cada observación $(\mathrm{AUC}_{\mathrm{IG}} - \widehat{\mathrm{AUC}}_{\mathrm{IG}})/\widehat{\sigma}$, versus las covariables en el espacio *ilr*, mientras que la Figura 6.5 muestra el gráfico de los residuos estandarizados versus los valores predichos $\widehat{\mathrm{AUC}}_{\mathrm{IG}}$. Dichas figuras no permiten observar ninguna estructura en particular, sugiriendo que el ajuste lineal podría ser adecuado. En este punto, vale la pena destacar que considerar modelos de regresión no paramétricos en lugar de modelos paramétricos, como el modelo lineal, podría tal vez ser más adecuado, si se piensa que estos últimos son demasiado restrictivos. Sin embargo, en nuestro análisis mantendremos el nivel de complejidad al mínimo y supondremos la validez del modelo dado en (6.1). Cade destacar que en el grupo de individuos sanos del estudio clínico en cuestión observamos que la relación entre las covariables y la respuesta no parecía ser lineal, por lo que decidimos descartar ese grupo de nuestro análisis y considerar como poblaciones las de los diabéticos y prediabéticos.

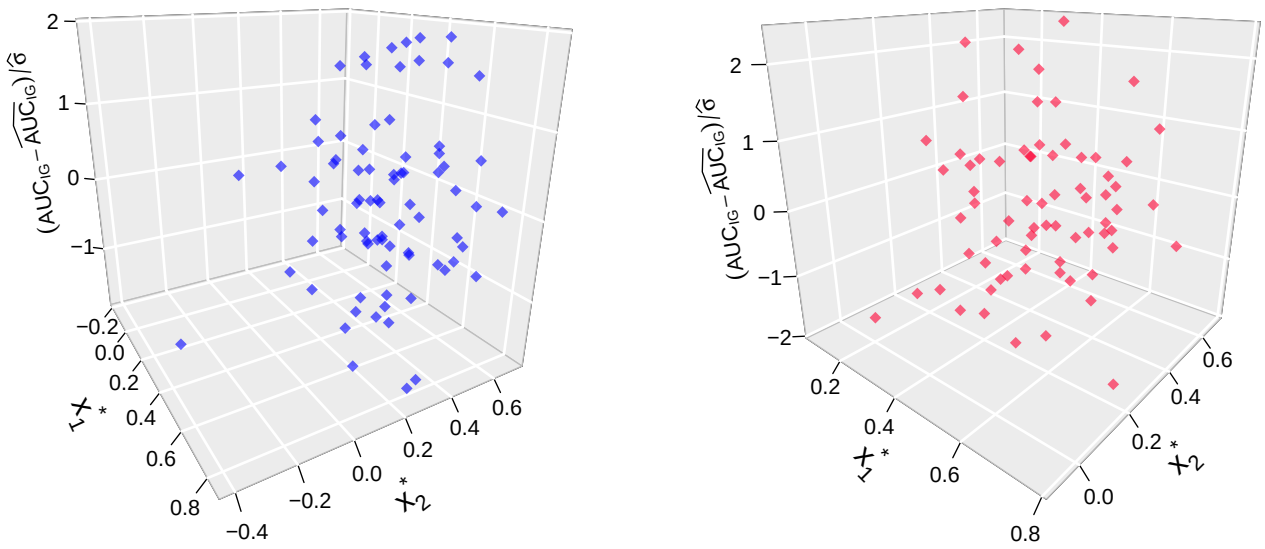


Figura 6.4: Residuos estandarizados resultantes de ajustar los modelos en (6.1) en las poblaciones de prediabéticos (panel izquierdo) y diabéticos (panel derecho) en el espacio *ilr* de las covariables.

Para obtener una apreciación global de la capacidad discriminatoria de la variable respuesta, generamos una grilla en el simplex sobre la cual estimar la curva ROC condicional. Para ello, al igual que en el Capítulo 5, primero generamos una grilla uniforme en el espacio *ilr* en el rectángulo $[0.2, 0.8] \times [-0.1, 0.7]$, a la cual le aplicamos la transformación ilr^{-1}. Ambas grillas se presentan en la Figura 6.6. Cabe destacar que si tomáramos una grilla mucho mayor, podríamos estar contemplando regiones del simplex en las que es poco probable observar valores de las covariables en los dos grupos considerados.

6.3.1. Área bajo la curva condicional

Una vez generada la grilla, estimamos la curva $\mathrm{ROC}_{\mathbf{x}}$ mediante los tres estimadores definidos en el Capítulo 3, el basado en empíricas, $\widehat{\mathrm{ROC}}_{\mathbf{x}}$, la versión suavizada de éste, $\widehat{\mathrm{ROC}}_{\mathbf{x},h}$, y el estimador *plug-in* que asume normalidad de los errores de los modelos de

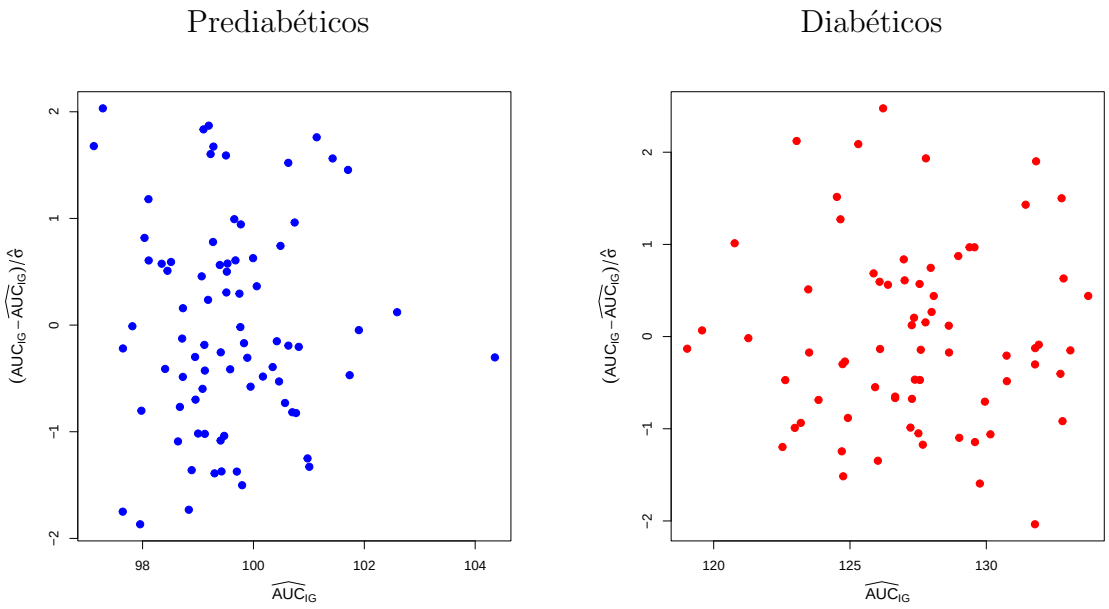


Figura 6.5: Gráficos de los residuos estandarizados versus los valores predichos resultantes de ajustar los modelos en (6.1) en las poblaciones de prediabéticos (panel izquierdo) y diabéticos (panel derecho).

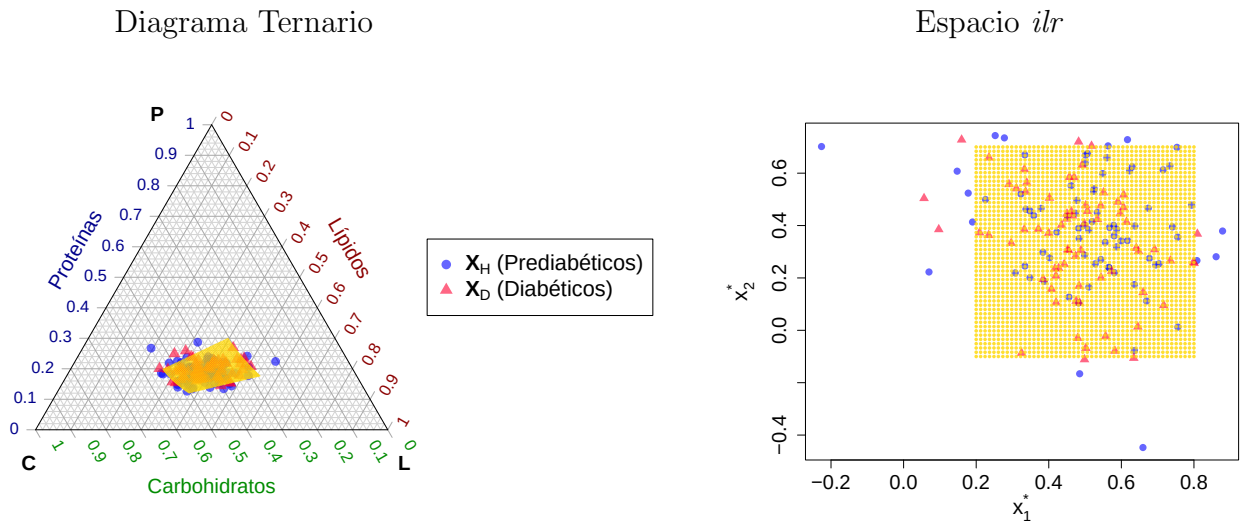


Figura 6.6: Grilla en el simplex sobre la que se estimó la curva ROC condicional (panel izquierdo) y grilla en el espacio *ilr* asociada (panel derecho).

regresión en (6.1), $\widehat{\mathrm{ROC}}_{\mathbf{x},\mathrm{PI}}$. Para el estimador suavizado, al igual que en las simulaciones, se utilizaron el núcleo de Epanechnikov y la ventana definida en (3.8), es decir,

$$h^\star_{n_D}(p) = c_{n_D} \frac{\sqrt{5p(1-p)}}{\sqrt{2n_D}},$$

donde $c_{n_D} = 1 + 1.8\, n_D^{-1/5}$.

A partir de las estimaciones sobre la grilla, podemos graficar el área bajo la curva para cada uno de ellos y compararla con el valor de la AUC de la curva ROC no condicional, que

era de 0.852. Estos gráficos se muestran en la Figura 6.7 donde en el plano horizontal se presenta el diagrama ternario y la grilla de puntos donde se evaluaron los estimadores de $\mathrm{ROC}_{\mathbf{x}}$.

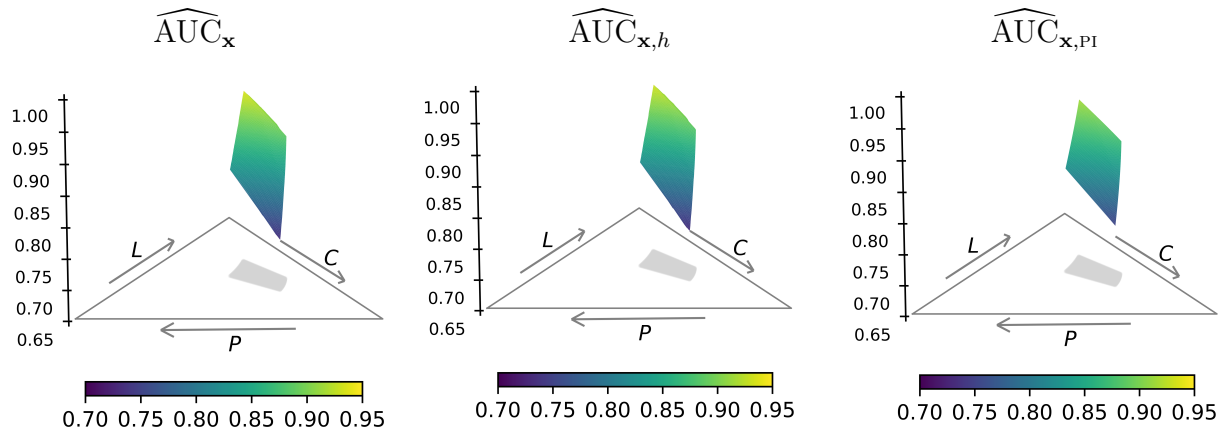


Figura 6.7: Estimadores de la $\mathrm{AUC}_{\mathbf{x}}$ sobre la grilla en el simplex.

Como podemos apreciar, hay una región de la grilla en la que los valores estimados de la $\mathrm{AUC}_{\mathbf{x}}$ se encuentran por encima de 0.852. Dicha región corresponde a pacientes en los que el consumo de macronutrientes es tal que la proporción de carbohidratos es baja y la de lípidos, más alta. Como parámetro, podríamos decir que cuando la proporción de lípidos es mayor al 45 % y la de carbohidratos menor al 35 %, la capacidad discriminatoria de $\mathrm{AUC}_{\mathrm{IG}}$ aumenta. Para visualizar mejor este fenómeno, calculamos las estimaciones de la $\mathrm{AUC}_{\mathbf{x}}$ en todo el simplex. Dichas estimaciones se presentan en la Figura 6.8 para cada uno de los procedimientos considerados. Se grafica también en gris, el plano ternario con ordenada vertical 0.852 que corresponde a la AUC de la curva ROC no condicional y el diagrama ternario con coordenada vertical 0 donde se indican con flechas el sentido de crecimiento de cada variable composicional y la grilla de puntos donde habíamos calculado anteriormente las estimaciones para poder visualizar lo que sucede en dicha región.

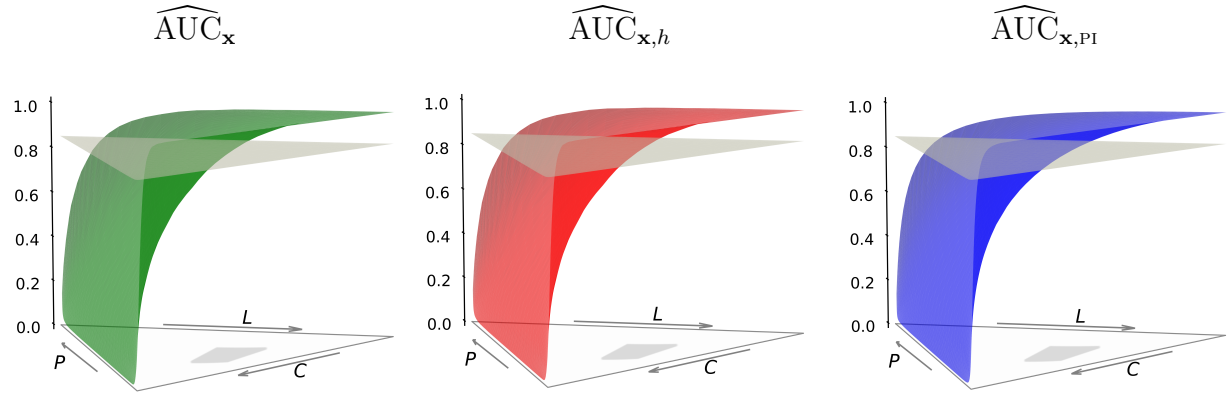


Figura 6.8: Estimaciones de la $\mathrm{AUC}_{\mathbf{x}}$ sobre todo el simplex. El plano horizontal corresponde al valor 0.852, o sea, a la AUC no condicional.

Podemos ver más claramente cómo la capacidad discriminatoria del biomarcador mejora a medida que la dieta está compuesta mayoritariamente por lípidos. También lo hace, aunque en menor medida, cuando las covariables corresponden a valores bajos de carbohidratos. Sin

embargo, tal vez sea muy poco probable que en la dieta se tenga valores tan extremos de estas covariables. Cabe mencionar que los tres estimadores arrojan resultados comparables. Por otro lado, en la siguiente sección, a modo de ejemplo ilustrativo del comportamiento asociado a valores altos o bajos de lípidos, elegimos dos puntos notables de la grilla considerada anteriormente en los que mostraremos las estimaciones de la curva ROC condicional.

6.3.2. $\mathrm{ROC}_{\mathbf{x}}$ en un punto

Observando los gráficos de las Figuras 6.7 y 6.8, detectamos una región de la grilla en la que la capacidad discriminatoria del biomarcador considerado pareciera mejorar. Con el objetivo de obtener una mejor apreciación de esta situación, elegimos dos puntos de la grilla representada en la Figura 6.6 en los que las estimaciones de la $\mathrm{AUC}_{\mathbf{x}}$ se encuentran por encima y por debajo del valor 0.852, que corresponde al área bajo la curva de la curva ROC no condicional. En particular, tomamos, los puntos $\mathbf{x}_1 = (0.1867, 0.3325, 0.4808)^{\mathrm{T}}$ y $\mathbf{x}_2 = (0.1883, 0.4996, 0.3121)^{\mathrm{T}}$. La Tabla 6.1 presenta los valores estimados del área bajo la curva para cada uno de los estimadores en dichos puntos.

	$\mathbf{x}_1$	$\mathbf{x}_2$
$\widehat{\mathrm{AUC}}_{\mathbf{x}}$	0.815	0.916
$\widehat{\mathrm{AUC}}_{\mathbf{x},h}$	0.812	0.913
$\widehat{\mathrm{AUC}}_{\mathbf{x},\mathrm{PI}}$	0.813	0.897

Tabla 6.1: Valores de las estimaciones de $\mathrm{AUC}_{\mathbf{x}}$ para cada uno de los procedimientos considerados en $\mathbf{x}_1 = (0.1867, 0.3325, 0.4808)^{\mathrm{T}}$ y $\mathbf{x}_2 = (0.1883, 0.4996, 0.3121)^{\mathrm{T}}$.

Observemos que para el punto $\mathbf{x}_2$, las estimaciones de la AUC condicional son todos mayores al valor de referencia 0.852. En base a los resultados de la Tabla 6.1, es de esperar que las estimaciones de $\mathrm{ROC}_{\mathbf{x}_2}$ se encuentren mayoritariamente por encima de la estimación de la curva no condicional, indicando que conocer información adicional contenida en las covariables mejora la capacidad de la variable $\mathrm{AUC}_{\mathrm{IG}}$ para distinguir entre prediabéticos y diabéticos. Esto es lo que efectivamente sucede como puede observarse en la Figura 6.9 donde se presentan en línea gris sólida los valores de $\widehat{\mathrm{ROC}}_{\mathbf{x}}$, en línea negra sólida los de $\widehat{\mathrm{ROC}}_{\mathbf{x},h}$ y en línea gris punteada los asociados a $\widehat{\mathrm{ROC}}_{\mathbf{x},\mathrm{PI}}$. En todos los casos, la línea roja corresponde al estimador de la curva ROC no condicional.

La Figura 6.9 permite observar que los valores obtenidos para el estimador empírico y para el suavizado son muy parecidos, siendo las estimaciones obtenidas con este último más suaves. Vale la pena destacar que, en ambos puntos, las estimaciones obtenidas con el estimador *plug-in* están muy cerca de las de los otros estimadores, lo cual podría sugerir que la distribución del biomarcador $\mathrm{AUC}_{\mathrm{IG}}$ condicional a las covariables composicionales $\mathbf{X}$ podría ser normal. Para visualizar el comportamiento de dicha distribución, o equivalentemente, de los errores del modelo (6.1), calculamos el estimador de la densidad de los residuos estandarizados

$$\widehat{\varepsilon}_{j,i} = \frac{y_{j,i} - \widehat{\beta}_{0,j} - \langle \widehat{\boldsymbol{\beta}}_j, \mathbf{x}_{j,i} \rangle_a}{\widehat{\sigma}_j} \qquad 1 \leq i \leq n_j \quad j = P, D\,,$$

$\mathbf{x}_1 = (0.1867, 0.3325, 0.4808)^{\mathrm{T}}$ $\mathbf{x}_2 = (0.1883, 0.4996, 0.3121)^{\mathrm{T}}$

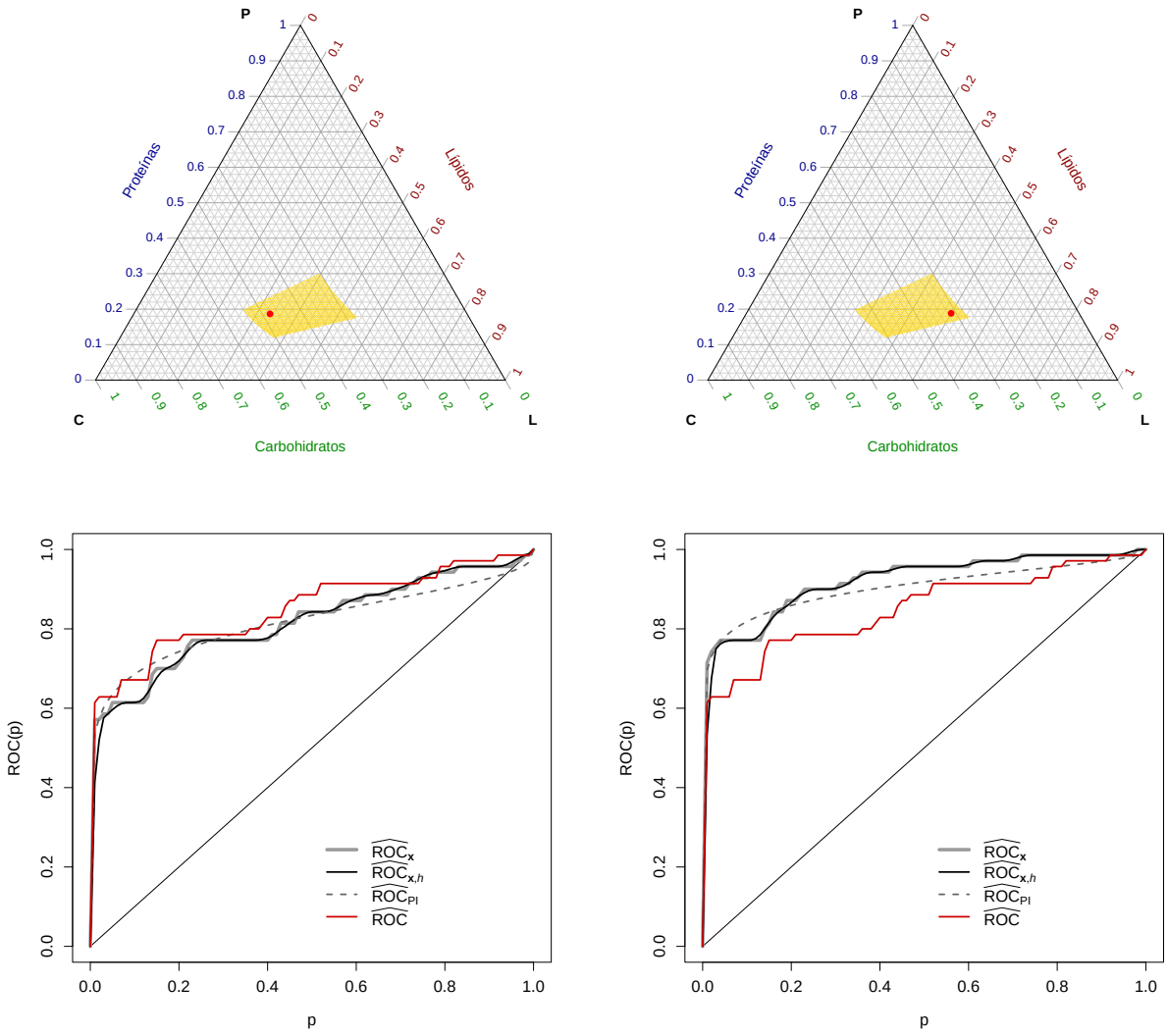


Figura 6.9: En los diagramas ternarios del panel superior, se encuentran graficados en color rojo los dos puntos notables del simplex donde se evaluaron los estimadores de $\mathrm{ROC}_{\mathbf{x}}$. En el panel inferior, se grafican las tres estimaciones de $\mathrm{ROC}_{\mathbf{x}}$. La estimación de la curva ROC no condicional se presenta en rojo oscuro.

definidos en el **Paso 2** del Capítulo 3. Dichas densidades se presentan en la Figura (6.10) indicadas en línea sólida y se parecen mucho a la densidad de una normal estándar que está graficada en línea punteada. La semjenaza de ambas densidades permitiría explicar que el estimador *plug-in* se parezca a los otros dos estimadores. Incluso si la verdadera distribución de los errores no fuera normal, es decir, aún cuando el supuesto de binormalidad sobre el que se basa el estimador *plug-in* no pareciera valer, los resultados del Capítulo 5 indican que dicho estimador da resultados confiables para errores con distribución simétrica con colas más pesadas que las de la normal como la t de Student o la logística. Una observación interesante es que no hemos hecho suposiciones acerca de la distribución de las covariables.

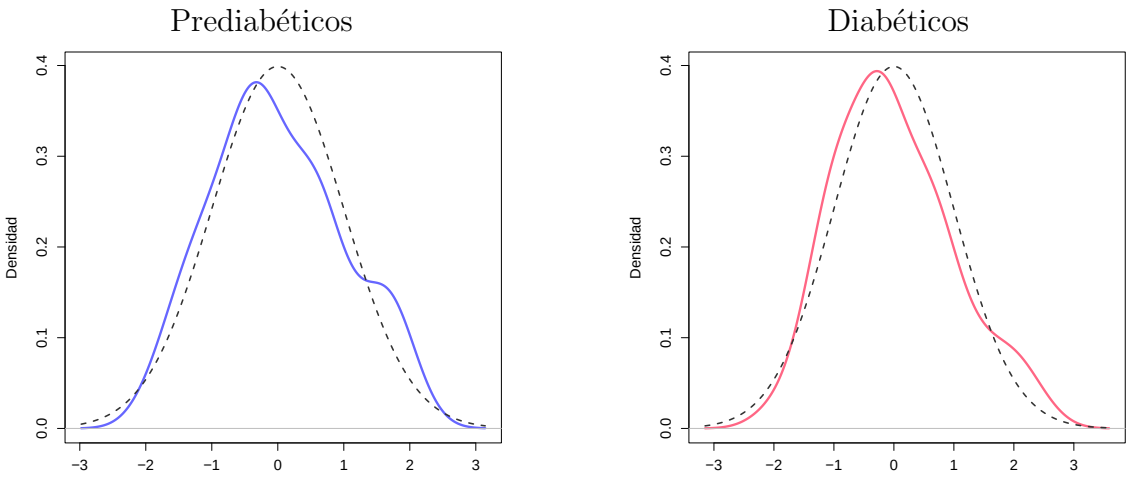


Figura 6.10: Estimación de la densidad de los residuos estandarizados resultantes de ajustar los modelos en (6.1) en las poblaciones de prediabéticos (panel izquierdo) y diabéticos (panel derecho) junto con la densidad de una normal estándar en línea punteada.

Capítulo 7
Conclusiones

Para concluir, hemos logrado adaptar la metodología inducida para la estimación de la curva ROC condicional al caso en el que las covariables son composicionales. Como vimos, fue necesario aplicar a las covariables la transformación *ilr* que, como se trata de una isometría, nos permitió realizar el ajuste de los modelos lineales involucrados a través del ajuste de mínimos cuadrados usual utilizando las coordenadas *ilr* de los puntos. Nos hemos enfocado en modelos lineales, pero podrían plantearse otros tipos de modelos, incluso no paramétricos.

En cuanto a los estimadores presentados en el trabajo, logramos obtener una versión suavizada del estimador basado en funciones de distribución empíricas, lo cual tiene la doble ventaja de ser suave y no imponer condiciones sobre la distribución de los errores de los modelos de regresión. Además, mediante simulaciones, hemos demostrado que aún en el caso desbalanceado, es decir, cuando en la muestra hay más individuos enfermos que sanos, y viceversa, este estimador se comporta de forma satisfactoria. La consistencia débil uniforme demostrada en el Capítulo 4 proporcionó, además, el sustento teórico de este desempeño. Entre los estimadores estudiados, resaltamos la estabilidad presentada por aquel basado en el supuesto de binormalidad, ya que en el estudio de simulación, presentó, en general, resultados comparables a los presentados por los otros estimadores aún cuando la distribución subyacente de los errores era t de Student o logística.

Por último, mediante el análisis de los datos correspondientes al *Estudio A Estrada de Glicación e Inflamación*, hemos logrado identificar en un ejemplo real cómo la capacidad discriminatoria de un biomarcador en particular, el *área bajo la curva* de las excursiones de glucosa en este caso, puede mejorar si se conoce información adicional de los individuos. En particular, los resultados obtenidos mostraron que en individuos cuya dieta está compuesta mayormente por grasas, el desempeño de dicho biomarcador mejora.

Como extensiones de este trabajo, podemos mencionar dos principales líneas de trabajo. La primera tiene que ver con el desarrollo de métodos robustos para los estimadores estudiados, para poder así contrarrestar el efecto que potenciales datos atípicos podrían tener tanto en el ajuste de los modelos de regresión como en la estimación por núcleos de la distribución de la pseudovariable involucrada. En particular, es de interés proveer estimadores suavizados robustos de la curva ROC condicional que sean una alternativa más regular a aquellos definidos en Bianco et al. (2022). La elección de una ventana óptima para el caso en cuestión también es un posible tema a explorar. La segunda línea de trabajo tiene que ver con el *problema de los ceros*, que surge cuando las covariables composicionales presentan observaciones en las que al menos una de las componentes es nula. Como las transformaciones que se aplican a los datos composicionales están definidas a partir de logaritmos, resulta necesario analizar el efecto de éstos en la estimación de la curva ROC y considerar otras distancias adaptadas a esta situación para definir las funciones de regresión y sus estimadores.

Bibliografía

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44:139–160.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. The Blackburn Press.

Aitchison, J. & Shen, S. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67:261–272.

Bianco, A., Boente, G., & González-Manteiga, W. (2022). Robust consistent estimators for ROC curves with covariates. *Electronic Journal of Statistics*, 16:4133–4161.

Bianco, A., Boente, G., González-Manteiga, W., Gude Sampedro, F., & Pérez-González, A. (2024). Robust nonparametric regression for compositional data: the simplicial–real case. Available at `https://arxiv.org/abs/2405.12924`.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35:279–300.

Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied Compositional Data Analysis*. Springer.

Genton, M. G., Johnson, C., Potter, K., Stenchikov, G., & Sun, Y. (2014). Surface boxplots. *Stat*, 3:1–11.

González-Manteiga, W., Pardo-Fernández, J. C., & Van Keilegom, I. (2011). ROC curves in nonparametric location-scale regression models. *Scandinavian Journal of Statistics*, 38:169–184.

Gude, F., Díaz-Vidal, P., R.-P., C., A.-S. M., Ferández-Merino, C., Rey-García, J., Cadarso-Suárez, C., Pazos-Couselo, M., García-López, J. M., & Gonzalez-Quintela, A. (2017). Glycemic variability and its association with demographics and lifestyles in a general adult population. *Journal of Diabetes Science and Technology*, 11:780—-790.

Hubert, M. & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52:5186–5201.

Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F., Alonzo, T., & Cadarso-Suárez, C. (2012). Extending induced ROC methodology to the functional context. *Biostatistics*, 13:594–608.

Jokiel-Rokita, A. & Pulit, M. (2013). Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions. *Statistics and Computing*, 23:703–712.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Pardo-Fernández, J. C., Rodriguez-Álvarez, M. X., & Van Keilegom, I. (2014). A review on ROC curves in the presence of covariates. *REVSTAT*, 12:21–41.

Pawlowsky-Glahn, V., Egozcue, J., & Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Wiley.

Pearson, K. (1897). Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60:489–498. Available at `https://doi.org/10.1098/rspl.1896.0076`.

Peng, L. & Zhou, X. H. (2004). Local linear smoothing for the receiver operating characteristic (ROC) curve. *Journal of Statistical Planning and Inference*, 118:129–143.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer Series in Statistics.

Pulit, M. (2016). A new method of kernel–smoothing estimation of the ROC curve. *Metrika*, 79:603–634.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.