UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS BÁSICAS

COMPARACIÓN DE MÉTODOS DE DISCRIMINACIÓN BINARIA PARA DATOS DE DIMENSIÓN ALTA: UN ESTUDIO POR SIMULACIÓN

TESIS

QUE PARA OBTENER EL TÍTULO DE

MAESTRO EN CIENCIAS
EN MATEMÁTICAS APLICADAS

PRESENTA

LUIS MIGUEL CÓRDOVA RODRÍGUEZ

DIRECTORES DE TESIS:

DRA. ADDY MARGARITA BOLÍVAR CIMÉ
DR. AROLDO PÉREZ PÉREZ

CUNDUACÁN, TAB, MEX. 2016

UNIVERSIDAD JUÁREZ
AUTÓNOMA DE TABASCO

"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE CIENCIAS BÁSICAS
Dirección

11 de mayo de 2016

Mat. Luis Miguel Córdova Rodríguez
Pasante de la Maestría en Ciencias
en Matemáticas Aplicadas
P r e s e n t e.

Por medio del presente y de la manera más cordial, me dirijo a Usted para hacer de su conocimiento que proceda a la impresión del trabajo titulado *"Comparación de métodos de discriminación binaria para datos de dimensión alta: un estudio por simulación"*, en virtud de que reúne los requisitos para el EXAMEN PROFESIONAL para obtener el grado de Maestro en Ciencias en Matemáticas Aplicadas.

Sin otro particular, reciba un cordial saludo.

A t e n t a m e n t e.

DIVISIÓN ACADÉMICA DE CIENCIAS BÁSICAS
DIRECCIÓN.

Dr. Gerardo Delgadillo Piñón
Director

C.c.p.- Archivo
Dr'GDP/Dr'JLSC/emt

CARTA DE AUTORIZACIÓN

El que suscribe, autoriza por medio del presente escrito a la Universidad Juárez Autónoma de Tabasco para que utilice tanto física como digitalmente la tesis de grado denominada *"Comparación de métodos de discriminación binaria para datos de dimensión alta: un estudio por simulación"*, de la cual soy autor y titular de los Derechos de Autor.

La finalidad del uso por parte de la Universidad Juárez Autónoma de Tabasco de la tesis antes mencionada, será única y exclusivamente para difusión, educación y sin fines de lucro; autorización que se hace de manera enunciativa más no limitativa para subirla a la Red Abierta de Bibliotecas Digitales (RABID) y a cualquier otra red académica con las que la Universidad tenga relación institucional.

Por lo antes manifestado, libero a la Universidad Juárez Autónoma de Tabasco de cualquier reclamación legal que pudiera ejercer respecto al uso y manipulación de la tesis mencionada y para los fines estipulados en éste documento.

Se firma la presente autorización en la ciudad de Villahermosa, Tabasco a los 23 días del mes de Mayo del año 2016.

AUTORIZO

LUIS MIGUEL CORDOVA RODRIGUEZ

Índice general

Dedicado a
mis padres, esposa e hijos.

Agradecimientos

Agradezco profundamente a Dios por brillar en mi camino y dirigir con su sabiduría mi intelecto y mi espíritu en la elaboración de esta tesis.

A mi Universidad Juárez Autónoma de Tabasco que me permitió alcanzar la meta de realizar un estudio de posgrado, conservando los más altos ideales de ética y justicia, procurando elevar el nivel educativo de nuestro estado.

Al proyecto PRODEP (UJAT-PTC-178) por el apoyo brindado para llevar a cabo la investigación de este trabajo.

A mis directores de tesis que me brindaron su apoyo, respaldo y conocimientos, para poder convertirme en un profesionista completo y deseoso de superarme constantemente, renunciando al conformismo, a ustedes estimados doctores, gracias por su paciencia y compromiso.

Introducción

Los datos multivariados de dimensión alta, es decir, datos multivariados cuya dimensión es mayor que el tamaño de la muestra, han tenido aplicaciones en estudios de genómica, análisis de imágenes médicas, climatología, finanzas, entre otros. En genómica, se analizan microarreglos para medir expresiones de genes. En el estudio de imágenes médicas, se analizan poblaciones de imágenes tridimensionales, en las que se representan numéricamente muestras de órganos de interés en vectores de parámetros con dimensiones muy altas. Sin embargo, al ser investigaciones muy costosas, los tamaños de las muestras son muy pequeñas, resultando datos multivariados de dimensión alta.

En este trabajo, se analiza una generalización de resultados asintóticos que consideran únicamente datos multivariados Gaussianos, utilizando datos multivariados más generales que tendrán una representación geométrica asintótica cuando la dimensión de los datos tiende a infinito y el tamaño de la muestra permanece fijo (datos de dimensión alta). El estudio que realizaremos está enfocado en el comportamiento asintótico de los métodos de discriminación binaria Mean Difference(MD), Support Vector Machine (SVM), Distance Weighted Discrimination (DWD) y Maximal Data Piling (MDP), los cuales se basan en hiperplanos separantes. En [5] se presentan resultados del comportamiento de estos métodos considerando datos multivariados Gaussianos, en términos del ángulo entre los vectores normales de los hiperplanos separantes de estos métodos y la dirección óptima. Mediante simulaciones, en este trabajo se hace un análisis para determinar si los resultados de [5] se pueden generalizar considerando datos multivariados con una representación geométrica asintótica.

En el Capítulo 1 presentamos un estudio de algunos conceptos y resultados de la teoría de probabilidad que necesitaremos para el desarrollo de esta tesis. Se inicia con la construcción de un espacio de probabilidad y finalizamos con el estudio de las formas de convergencia de variables aleatorias, sin embargo, asumiremos familiaridad con el concepto de integral de Lebesgue y sus propiedades fundamentales. En el Capítulo 2 nos enfocamos en el estudio de la distribución normal multivariada, y a partir de ésta, describiremos cuando un conjunto de datos multivariados cumple con la representación geométrica asintótica, que es la característica que tendrán los datos tratados en las simulaciones de este trabajo. La descripción de los cuatro métodos de discriminación binaria que consideramos en nuestro estudio lo daremos en el Capítulo 3. La

comparación mediante simulaciones de estos métodos se realizó con tres tipos de datos multivariados que cumplen con la representación geométrica y los resultados obtenidos son presentados en el Capítulo 4.

Capítulo 1

Teoría de Probabilidad

En este capítulo se muestran los conceptos y resultados de la teoría de probabilidad necesarios para la comprensión de los resultados asintóticos que serán presentados en esta tesis.

1.1. Espacios de probabilidad

Iniciamos esta sección con las definiciones básicas que nos llevan a la construcción de un espacio de probabilidad, el cual es el modelo matemático adecuado para un experimento aleatorio. En este contexto, Ω será un conjunto no vacío que representa al conjunto de todas las posibles respuestas del experimento, adecuadamente "codificadas".

Por lo general, una vez realizado el experimento, es usual que se esté interesado en saber si el resultado pertenece o nó a cierto subconjunto de interés del espacio muestral. De esta manera, aunque no de forma estricta, asociado a Ω debe haber una familia $\mathcal{F}$ consistente de los subconjuntos A de Ω para los cuales pueda decidirse si A ocurre o nó. En concordancia con esta idea intuitiva, se tiene el siguiente concepto.

Definición 1.1.1 *Sea $\Omega \neq \emptyset$. Una **σ-álgebra** sobre Ω es una familia $\mathcal{F}$ de subconjuntos de Ω tal que*

(i) $\Omega \in \mathcal{F}$.

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.

(iii) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Observemos que si $\mathcal{F}$ es una σ-álgebra

(iv) $\emptyset \in \mathcal{F}$.

(v) $A_1, ..., A_k \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{k} A_i \in \mathcal{F}, \quad \bigcap_{i=1}^{k} A_i \in \mathcal{F}$.

(vi) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

(vii) A_1 y $A_2 \in \mathcal{F} \Rightarrow A_1 - A_2 = A_1 \cap A_2^c \in \mathcal{F}$.

Al par $(\Omega, \mathcal{F})$ se le llama **espacio medible** y a los elementos de $\mathcal{F}$ se les llama **conjuntos $\mathcal{F}$-medibles**.

Definición 1.1.2 *Sea $\mathcal{A}$ una familia de subconjuntos de Ω. La σ-álgebra **generada** por $\mathcal{A}$ es la intersección de todas las σ-álgebras que contienen a $\mathcal{A}$, la cual denotaremos por $\sigma(\mathcal{A})$.*

Tal definición tiene sentido por que la intersección (arbitraria) de σ-álgebras es una σ-álgebra. De esta manera $\sigma(\mathcal{A})$ está contenida en toda σ-álgebra que contiene a $\mathcal{A}$ y se dice entonces que $\sigma(\mathcal{A})$ es la menor σ-álgebra que contiene a $\mathcal{A}$.

En las aplicaciones, frecuentemente nos encontramos con la σ-álgebra generada por todos los conjuntos abiertos en $\mathbb{R}^n$ (con la métrica euclidiana), $\mathcal{B}(\mathbb{R}^n)$, a la cual se le conoce como la **σ-álgebra de Borel**. A los elementos de $\mathcal{B}(\mathbb{R}^n)$ les llamaremos **conjuntos de Borel** en $\mathbb{R}^n$.

Dado un espacio medible $(\Omega, \mathcal{F})$, con cada $A \in \mathcal{F}$ se asociará una probabilidad de ocurrencia, $P(A)$. En la siguiente definición se establecen las propiedades que esta asignación de probabilidades debe poseer.

Definición 1.1.3 *Sea $(\Omega, \mathcal{F})$ un espacio medible. Una **medida de probabilidad** P es una función $P : \mathcal{F} \longrightarrow [0, 1]$ que cumple*

(i) $P(\Omega) = 1$.

(ii) *Para cada sucesión $A_1, A_2, ... \in \mathcal{F}$ con $A_i \cap A_j = \emptyset$ para $i \neq j$,*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Toda medida de probabilidad P cumple las siguientes propiedades

(iii) $P(\emptyset) = 0$.

(iv) $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ si $A_i \cap A_j = \emptyset$, $i \neq j$.

(v) $P(A^c) = 1 - P(A)$.

(vi) Si $A \subset B$, entonces $P(A) \leq P(B)$.

(vii) (Desigualdad de Boole) $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$.

(viii) Si $A_1, A_2, ... \in \mathcal{F}$ es una sucesión creciente, es decir, $A_n \subset A_{n+1}$ para todo n, entonces $lim_{n\to\infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i\right)$.

(ix) Si $A_1, A_2, ... \in \mathcal{F}$ es una sucesión decreciente, es decir, $A_n \supset A_{n+1}$ para todo n, entonces $lim_{n\to\infty} P(A_n) = P\left(\bigcap_{i=1}^{\infty} A_i\right)$.

A la terna $(\Omega, \mathcal{F}, P)$ se le llama **espacio de probabilidad**. En este contexto, a los elementos de $\mathcal{F}$ se les llamará **eventos**; y diremos que un evento A ocurre casi seguramente (abreviado c.s.) si $P(A) = 1$.

Presentamos ahora el Lema de Borell-Cantelli, un resultado útil en algunas demostraciones.

Teorema 1.1.1 *(Lema de Borel-Cantelli). Si* $A_1, A_2, ... \in \mathcal{F}$ *y* $\sum_{n=1}^{\infty} P(A_n) < \infty$, *entonces* $P(\limsup_n A_n) = 0$.

Demostración. Recordemos que $\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$, así

$$\begin{aligned} P(\limsup_n A_n) &\leqslant P\left(\bigcup_{k=n}^{\infty} A_k\right) \\ &\leqslant \sum_{k=n}^{\infty} P(A_k) \to 0 \text{ cuando } n \to \infty. \end{aligned}$$

■

1.2. Variables y vectores aleatorios

Supongamos que el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ es un modelo para cierto experimento aleatorio, y que $X : \Omega \to \mathbb{R}$ representa una característica numérica asociada con las posibles respuestas, de manera que para cada $\omega \in \Omega$, $X(\omega)$ es el valor de la característica de interés para el resultado ω. Consideremos ahora la situación antes de efectuar el experimento. En general, no es posible predecir la respuesta ω que se obtendrá, y entonces tampoco puede determinarse de antemano el correspondiente valor $X(\omega)$. Sin embargo nos gustaría poder contestar algunas preguntas básicas acerca de X, como por ejemplo

¿Cual es la probabilidad de que $X \in B$, donde $B \in \mathcal{B}(\mathbb{R})$?

Para contestar a esta pregunta empezamos determinando

$$A = \{\omega \in \Omega : X(\omega) \in B\}$$

y entonces la respuesta es $P(A)$. En este punto es importante recordar que la medida de probabilidad P está definida en la σ-álgebra $\mathcal{F}$, por lo que $P(A)$ tiene sentido sólo cuando $A \in \mathcal{F}$. Luego, para poder evaluar $P(A)$ es necesario que $A = \{\omega \in \Omega : X(\omega) \in B\}$ sea un elemento de $\mathcal{F}$, es decir, un evento. Así, $\mathcal{F}$ debe contener a todos lo conjuntos de la forma $X^{-1}(B)$, donde B es un conjunto de Borel en $\mathbb{R}$.

Definición 1.2.1 *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad . Una función $X : \Omega \longrightarrow \mathbb{R}$ es una **variable aleatoria** si $X^{-1}(B) \in \mathcal{F}$ para todo $B \in \mathcal{B}(\mathbb{R})$.*

Si $X : \Omega \longrightarrow \mathbb{R}$ es una variable aleatoria, entonces P_X definida por

$$P_X(B) = P(X \in B), \quad B \in \mathcal{B}(\mathbb{R})$$

es una medida de probabilidad en $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A P_X se le llama la **distribución** de X.

Los números $P_X(B), B \in \mathcal{B}(\mathbb{R})$, caracterizan completamente a la variable aleatoria X en el sentido de que dan las probabilidades de todos los eventos que involucran a X. Es útil, conocer que esta información puede ser capturada por una función $F : \mathbb{R} \to \mathbb{R}$ única. Pero antes, recordemos los siguientes conceptos y resultados (cuya demostración puede ser consultada, por ejemplo, en [3], Sección 1.4).

Definición 1.2.2 *(i) Una **medida de Lebesgue - Stieltjes** sobre $\mathbb{R}$ es una medida $\mu : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ tal que $\mu(I) < \infty$ para cada intervalo acotado I.*

*(ii) Una **función de distribución** sobre $\mathbb{R}$ es un mapeo $F : \mathbb{R} \longrightarrow \mathbb{R}$ que es creciente y continuo por la derecha.*

Los siguientes dos teoremas establecen una correspondencia uno a uno entre las medidas de Lebesgue - Stieltjes y las funciones de distribución, donde dos funciones de distribución que difieren por una constante son identificadas.

Teorema 1.2.1 *Sea μ una medida de Lebesgue - Stieltjes sobre $\mathbb{R}$. Si $F : \mathbb{R} \longrightarrow \mathbb{R}$ está definida, salvo una constante aditiva, por $F(b) - F(a) = \mu(a, b]$, entonces F es una función de distribución.*

Teorema 1.2.2 *Sea F una función de distribución sobre $\mathbb{R}$. Si $\mu(a, b] = F(b) - F(a), a < b$, entonces existe una única extensión de μ a una medida de Lebesgue - Stieltjes sobre $\mathbb{R}$.*

Definición 1.2.3 *La **función de distribución** de una variable aleatoria X es la función $F_X \equiv F : \mathbb{R} \longrightarrow [0, 1]$ definida por*

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

Como, para $a < b$,

$$F(b) - F(a) = P(a < X \leqslant b) = P_X\,(a, b]\,,$$

entonces por el Teorema 1.2.1, F es una función de distribución correspondiente a la medida de Lebesgue - Stieltjes P_X. En particular, F es creciente y continua por la derecha. Usando las propiedades $(viii)$ y (ix) de una medida de probabilidad P es fácil ver que

$$F(x) \to 1 \text{ cuando } x \to \infty \qquad \text{y} \qquad F(x) \to 0 \text{ cuando } x \to -\infty.$$

Así entre todas las funciones de distribución correspondientes a P_X, elegimos la que cumple

$$F(\infty) \equiv \lim_{x\to\infty} F_X(x) = 1 \quad \text{y} \quad F(-\infty) \equiv \lim_{x\to-\infty} F_X(x) = 0.$$

Es muy frecuente decir lo siguiente

"Sea X una variable aleatoria con función de distribución F",

donde $F : \mathbb{R} \longrightarrow [0, 1]$ es una función creciente y continua por la derecha con $F(\infty) = 1$ y $F(-\infty) = 0$. No se hace referencia al espacio $(\Omega, \mathcal{F}, P)$, y de hecho la naturaleza de este espacio no importa. La función de distribución F determina la medida de probabilidad P_X, la cual a su vez determina la probabilidad de todos los eventos que involucran a X. Lo único que debe garantizarse es la existencia de al menos un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ sobre el cual una variable aleatoria X con función de distribución F pueda ser definida. De hecho, puede **siempre** suministrarse un espacio de probabilidad en una forma canónica; tome $\Omega = \mathbb{R},\ \mathcal{F} = \mathcal{B}(\mathbb{R}),\ P$ la medida de Lebesgue-Stieltjes correspondiente a F y defina $X(\omega) = \omega,\ \omega \in \Omega$.

Como

$$P_X(B) = P\,(\omega : X(\omega) \in B) = P(B),$$

X tiene medida de probabilidad inducida (distribución) P y por lo tanto función de distribución F.

Tenemos así demostrado lo siguiente.

Proposición 1.2.1 *Si F es una función de distribución con $F(\infty) = 1$ y $F(-\infty) = 0$, existe una variable aleatoria X tal que $F = F_X$.*

Definición 1.2.4 *Una variable aleatoria X es **discreta** si y sólo si $X(\Omega)$ es contable, es decir, X toma un conjunto finito o numerable de valores.*

Si X es discreta y los valores $\{x_1, x_2, ...\}$ de X pueden ser ordenados de manera que $x_n < x_{n+1},\ n = 1, 2, ...$, es claro que su función de distribución F es una función escalonada con una discontinuidad en cada x_n de magnitud $p_n = P(X = x_n)$; F es constante entre x_n y $x_{n+1}, n = 1, 2, ...$ y toma el valor supremo en cada discontinuidad.

Si X es una variable aleatoria discreta, las probabilidades de X son completamente determinadas por la **función de probabilidad** p_X, definida por

$$p_X(x) = P(X = x), \quad x \in \mathbb{R}.$$

Explícitamente,

$$P_X(B) = \sum_{x \in B} p_X(x), \qquad B \in \mathcal{B}(\mathbb{R}).$$

Esta es una suma contable ya que p_X es cero excepto en los $x_n's$.

Definición 1.2.5 *La variable aleatoria X es **continua** si y sólo si su función de distribución F es continua en $\mathbb{R}$.*

Notemos que X es continua si y sólo si $P(X = x) = 0$ para toda x.

Definición 1.2.6 *Una variable aleatoria continua X es **absolutamente continua** si y sólo si existe una función $f : \mathbb{R} \to [0, \infty)$ Borel medible ($f^{-1}(B) \in \mathcal{B}(\mathbb{R})$ para todo $B \in \mathcal{B}(\mathbb{R})$) tal que*

$$F(x) = \int_{-\infty}^{x} f(t)dt, \qquad x \in \mathbb{R}.$$

En este caso a f se le llama la **función de densidad** (o simplemente **densidad**) de X.

Si X es absolutamente continua con función de distribución F y una función de densidad f, entonces debido a que $F(x) \to 1$, cuando $x \to \infty$,

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Ahora, debido a que la medida μ definida por $\mu(B) = \int_B f(x)dx,\ B \in \mathcal{B}(\mathbb{R})$ cumple

$$\mu\,(a, b] = F(b) - F(a), \qquad a < b$$

se sigue del Teorema 1.2.2 que μ es la medida de Lebesgue-Stieltjes correspondiente a F y por lo tanto $\mu = P_X$, esto es

$$P_X(B) = \int_B f(x)dx, \qquad B \in \mathcal{B}(\mathbb{R}).$$

Notemos que toda función de Borel $f : \mathbb{R} \to [0, \infty)$ con $\int_{-\infty}^{\infty} f(x)dx = 1$ es la función de densidad de alguna variable aleatoria X. En efecto, la función $F(x) = \int_{-\infty}^{x} f(t)dt,\ x \in \mathbb{R}$

es claramente creciente y continua con $F(\infty) = 1$ y $F(-\infty) = 0$, y así, por la Proposición 1.2.1, es la función de distribución de alguna variable aleatoria X. Ahora, por la definición de esta F, X tiene densidad f.

Consideremos ahora situaciones en las que se asocia más de una variable aleatoria con el mismo experimento.

Definición 1.2.7 *Un **vector aleatorio** de dimensión n sobre un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ es una función $X : \Omega \to \mathbb{R}^n$ tal que $X^{-1}(B) \in \mathcal{F}$ para todo $B \in \mathcal{B}(\mathbb{R}^n)$.*

Si $X : \Omega \to \mathbb{R}^n$ y $X_i = \pi_i X$, donde π_i es la proyección de $\mathbb{R}^n$ sobre el i-ésimo espacio coordenado, entonces X es un vector aleatorio si y sólo si cada X_i es una variable aleatoria. Por lo tanto un vector aleatorio puede ser considerado como una n-áda $(X_1, ..., X_n)$ de variables aleatorias.

Definamos la medida de probabilidad P_X por

$$P_X(B) = P(X \in B), \qquad B \in \mathcal{B}(\mathbb{R}^n).$$

Veremos, igual que en el caso de variables aleatorias, que P_X está determinada por una función $F : \mathbb{R}^n \to \mathbb{R}$ única. Para ello recordemos los siguientes conceptos y resultados (véase [3], Sección 1.4).

Notación. Sean $a = (a_1, ..., a_n), b = (b_1, ..., b_n) \in \mathbb{R}^n$.

(i) $(a, b] \equiv \{x = (x_1, ..., x_n) \in \mathbb{R}^n : a_i < x_i \leqslant b_i, i = 1, ..., n\}$,
$(a, \infty) \equiv \{x = (x_1, ..., x_n) \in \mathbb{R}^n : x_i > a_i, i = 1, ..., n\}$,
$(-\infty, b] \equiv \{x = (x_1, ..., x_n) \in \mathbb{R}^n : x_i \leq b_i, i = 1, ..., n\}$,
los demás tipos de intervalos se definen en forma similar.

(ii) $a \leq b$ si y sólo si $a_1 \leq b_1, ..., a_n \leq b_n$.

Definición 1.2.8 *Una **medida de Lebesgue-Stieltjes** sobre $\mathbb{R}^n$ es una medida $\mu : \mathcal{B}(\mathbb{R}^n) \to \mathbb{R}$ tal que $\mu(I) < \infty$ para todo intervalo acotado $I \subset \mathbb{R}^n$.*

Ahora, dada una función $G : \mathbb{R}^n \to \mathbb{R}$ y $a = (a_1, ..., a_n), b = (b_1, ..., b_n) \in \mathbb{R}^n$, definamos

$$\Delta_{b_i a_i} G(x_1, \ldots, x_n) = G(x_1, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_n) - G(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n).$$

Teorema 1.2.3 *Sea $\mu : \mathcal{B}(\mathbb{R}^n) \to \mathbb{R}$ una medida finita y definamos $F : \mathbb{R}^n \to \mathbb{R}$ por*

$$F(x) = \mu(-\infty, x], \qquad x \in \mathbb{R}^n.$$

Si $a, b \in \mathbb{R}^n, a \leq b$, entonces

$$\mu(a, b] = \Delta_{b_1 a_1} \cdots \Delta_{b_n a_n} F(x).$$

Definición 1.2.9 *Dada una función* $F : \mathbb{R}^n \to \mathbb{R}$ *y* $a, b \in \mathbb{R}^n$ *con* $a \leq b$, *sea* $F(a, b] := \Delta_{b_1 a_1} \cdots \Delta_{b_n a_n} F(x)$.

(i) La función F *es* ***creciente*** *si y sólo si* $F(a, b] \geq 0$, *para* $a \leq b$.

(ii) La función F *es* ***continua por la derecha*** *si y sólo si para cualquier sucesión* $\{x^k\}$ *en* $\mathbb{R}^n$ *con* $x^k \geq x^{k+1}, k = 1, 2, ...,$ *se tiene que*

$$F(x^k) \to F(x) \qquad \text{cuando} \quad x^k \to x.$$

Notemos que cuando $n = 1$, estos conceptos coinciden con los correspondientes en $\mathbb{R}$.

Definición 1.2.10 *Una* ***función de distribución*** *sobre* $\mathbb{R}^n$ *es un mapeo* $F : \mathbb{R}^n \to \mathbb{R}$ *que es creciente y continuo por la derecha.*

Notemos que si F surge a través de una medida μ como en el Teorema 1.2.3, entonces F es una función de distribución. Recíprocamente (véase [3], Sección 1.4), si F es una función de distribución sobre $\mathbb{R}^n, \mu(a, b] = F(a, b]$ tiene una extensión única a una medida de Lebesgue-Stieltjes en $\mathbb{R}^n$.

Definición 1.2.11 *Sea* X *un vector aleatorio de dimensión* n. *La* ***función de distribución*** *de* X *es la función* $F_X \equiv F : \mathbb{R}^n \to [0, 1]$ *definida por*

$$F(x) = P_X(-\infty, x] = P(X_i \leq x_i, i = 1, ..., n).$$

A F se le llama también la **función de distribución conjunta** de $X_1, ..., X_n$.

Notemos que F es creciente y continua por la derecha en $\mathbb{R}^n$, y que P_X es la medida de Lebesgue-Stieltjes determinada por F.

Por propiedades de la medida de probabilidad P, tenemos también que

$$F(x_1, ..., x_n) \to \begin{cases} 1 & \text{cuando } x_i \uparrow \infty \text{ para toda } i, \\ 0 & \text{cuando } x_i \downarrow -\infty \text{ para alguna } i \\ & \text{(con todas las otras coordenadas} \\ & \text{fijas).} \end{cases} \tag{1.1}$$

Si F es una función de distribución sobre $\mathbb{R}^n$ que satisface (1.1), entonces F es la función de distribución de algún vector aleatorio X. El espacio de probabilidad subyacente puede ser construido en la siguiente forma canónica:
Tomar $\Omega = \mathbb{R}^n, \mathcal{F} = \mathcal{B}(\mathbb{R}^n)$, P la medida de Lebesgue-Stieltjes determinada por F y X la función identidad sobre Ω.

El hecho de que P es una medida de probabilidad se sigue de la hipótesis de que

$F(x) \to 1$ cuando $x \uparrow \infty$.

El vector aleatorio X se dice ser **discreto** si y sólo si el conjunto de valores de X es contable, o equivalentemente, si y sólo si las variables aleatorias componentes $X_1, ..., X_n$ son todas discretas. En este caso, las propiedades de X son determinadas por la **función de probabilidad** p_X, definida por

$$p_X(x) = P(X = x), \qquad x \in \mathbb{R}^n.$$

Explícitamente,

$$P_X(B) = \sum_{x \in B} p_X(x), \qquad B \in \mathcal{B}(\mathbb{R}^n).$$

El vector aleatorio X se dice ser **absolutamente continuo** si y sólo si existe una función $f : \mathbb{R}^n \to [0, \infty)$ Borel medible, llamada la **función de densidad** de X, tal que

$$F(x) = \int_{(-\infty, x]} f(t)dt, \qquad x \in \mathbb{R}^n.$$

Dado que $\mu(B) = \int_B f(x)dx, \quad B \in \mathcal{B}(\mathbb{R}^n)$ cumple

$$\mu(a, b] = \int_{(a,b]} f(x)dx = F(a, b],$$

μ es la medida de Lebesgue-Stieltjes determinada por F; por lo tanto $\mu = P_X$, esto es

$$P_X(B) = \int_B f(x)dx, \qquad B \in \mathcal{B}(\mathbb{R}^n).$$

Igual que para variables aleatorias, es fácil ver que toda función $f : \mathbb{R}^n \to [0, \infty)$ Borel medible con $\int_{\mathbb{R}^n} f(x)dx = 1$ es la función de densidad de algún vector aleatorio absolutamente continuo.

1.3. Independencia

Intuitivamente, la independencia de $X_1, X_2, ..., X_n$ significa que un enunciado sobre una o más de las X_i's no afecta las probabilidades concernientes al resto de las X_i's. La definición formal es como sigue.

Definición 1.3.1 *Sean $X_1, X_2, ..., X_n$ variables aleatorias sobre $(\Omega, \mathcal{F}, P)$. Las variables $X_1, X_2, ..., X_n$ son **independientes** si y sólo si para todos los conjuntos $B_1, ..., B_n \in \mathcal{B}(\mathbb{R})$,*

$$P\left(X_1 \in B_1, ..., X_n \in B_n\right) = P\left(X_1 \in B_1\right) \cdots P\left(X_n \in B_n\right).$$

Si $X_1, X_2, ..., X_n$ son variables aleatorias independientes y $B_1, ..., B_k \in \mathcal{B}(\mathbb{R})$, $k < n$, entonces

$$\begin{aligned} P(X_1 \in B_1, ..., X_k \in B_k) &= P(X_1 \in B_1, ..., X_k \in B_k, X_{k+1} \in \mathbb{R}, ..., X_n \in \mathbb{R}) \\ &= P(X_1 \in B_1) \cdots P(X_k \in B_k). \end{aligned}$$

Luego, la independencia de $X_1, X_2, ..., X_n$ implica la independencia de $X_1, X_2, ..., X_k$. En general, si $X_1, X_2, ..., X_n$ son independientes, toda subcolección de ellas es también independiente.

Ahora, si $X_1, X_2, ..., X_n$ son vectores aleatorios en $(\Omega, \mathcal{F}, P)$. Entonces $X_1, X_2, ..., X_n$ son independientes si y sólo si

$$P(X_1 \in B_1, ..., X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n)$$

para todo $B_1, ..., B_n \in \mathcal{B}(\mathbb{R}^n)$.

Definición 1.3.2 *Sean $\{X_i, i \in I\}$ una familia arbitraria de variables aleatorias (o vectores aleatorios). La familia $\{X_i, i \in I\}$ es* ***independiente*** *si y sólo si $X_{i_1}, ..., X_{i_n}$ son independientes para cada subconjunto finito $\{i_1, ..., i_n\}$ de indices distintos en I.*

Si $g_i : \mathbb{R}^n \to \mathbb{R}^m$, $i \in I$ es una familia de funciones medibles respecto a las σ-álgebras $\mathcal{B}(\mathbb{R}^n)$ y $\mathcal{B}(\mathbb{R}^m)$ $(g_i^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$ para todo $B \in \mathcal{B}(\mathbb{R}^m)$) y $\{X_i, i \in I\}$ es una familia de vectores aleatorios independientes de dimensión n, entonces claramente la familia $\{g_i \circ X_i, i \in I\}$ es una familia independiente de vectores aleatorios de dimensión m. Esto se sigue directamente de la definición de independencia y el hecho de que

$$\{g_i \circ X_i \in B_i\} = \{X_i \in g_i^{-1}(B_i)\}, \quad \forall\, B_i \in \mathcal{B}(\mathbb{R}^m).$$

Tenemos así que "funciones de vectores aleatorios independientes son independientes".

Las demostraciones de las siguientes caracterizaciones de independencia pueden ser consultadas en [3], Sección 4.8.

Teorema 1.3.1 *Sean $X_1, ..., X_n$ variables aleatorias sobre $(\Omega, \mathcal{F}, P)$ con funciones de distribución $F_1, ..., F_n$ respectivamente. Si F es la función de distribución de $X = (X_1, ..., X_n)$, entonces $X_1, ..., X_n$ son independientes si y sólo si*

$$F(x_1, ..., x_n) = F_1(x_1) \cdots F_n(x_n),$$

para todo $x_1, ..., x_n \in \mathbb{R}$.

Antes de caracterizar la independencia en términos de densidades, recordemos lo siguiente.

Si el vector aleatorio $X = (X_1, ..., X_n)$ tiene una densidad (conjunta) f, entonces cada X_i tiene una densidad (marginal) f_i dada por

$$f_i(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

Teorema 1.3.2 *Si el vector aleatorio $X = (X_1, ..., X_n)$ tiene una densidad f, entonces $X_1, ..., X_n$, son variables aleatorias independientes si y sólo si*

$$f(x_1, ..., x_n) = f_1(x_1) \cdots f_n(x_n)$$

para todos los puntos $(x_1, ..., x_n)$ excepto posiblemente para un subconjunto de Borel de $\mathbb{R}^n$ con medida de Lebesgue cero, donde f_i es la densidad de X_i, $i = 1, ..., n$.

Notemos ahora que si $X_1, ..., X_n$ son variables aleatorias independientes y X_i tiene densidad $f_i, i = 1, ..., n$, entonces $X = (X_1, ..., X_n)$ tiene densidad f dada por

$$f(x_1, ..., x_n) = f_1(x_1) \cdots f_n(x_n).$$

En efecto, por el Teorema 1.3.1

$$F(x_1, ..., x_n) = F_1(x_1) \cdots F_n(x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_1(t_1) \cdots f_n(t_n) dt_1 \cdots dt_n.$$

Luego, si $g(x_1, ..., x_n) = f_1(x_1) \cdots f_n(x_n)$, entonces

$$P_X(B) = \int_B g(x) dx, \qquad B \in \mathcal{B}(\mathbb{R}^n).$$

Pero $P_X(B) = \int_B f(x) dx$, y se sigue entonces que $f = g$ excepto posiblemente en un conjunto de medida de Lebesgue cero.

1.4. Esperanza

Sea X una variable aleatoria sobre $(\Omega, \mathcal{F}, P)$ con $X(\Omega) = \{x_1, ..., x_n\}$ y sea $p_i = P(X = x_i),\ i = 1, ..., n$. Si se realiza un número grande, N, de repeticiones independientes del experimento aleatorio, X tomaría el valor x_i aproximadamente Np_i veces, de manera que el promedio aritmético de los valores de X en las N observaciones es aproximadamente

$$\frac{1}{N} (Np_1x_1 + Np_2x_2 + \cdots + Np_nx_n) = \sum_{i=1}^{n} p_i x_i.$$

Esta es una cifra razonable para el valor medio de X. Denotando por $1_B, B \in \mathcal{F}$ a la función indicadora del evento B, esto es,

$$1_B(\omega) = \begin{cases} 1 & \text{si } \omega \in B, \\ 0 & \text{si } \omega \notin B, \end{cases}$$

tenemos que X puede ser representada por

$$\sum_{i=1}^{n} x_i 1_{B_i},$$

donde $B_i = X^{-1}(\{x_i\})$, $i = 1, ..., n$, son eventos disjuntos en $\mathcal{F}$. Así, el valor medio de X puede ser expresado como

$$\sum_{i=1}^{n} x_i P(B_i),$$

lo cual es $\int_\Omega X dP$. Ya que las variables aleatorias arbitrarias son construídas a partir de este tipo de variables, llamadas variables simples, es razonable tomar a $\int_\Omega X dP$ como la definición del valor medio (llamado "esperanza") de X.

Definición 1.4.1 *Si X es una variable aleatoria sobre $(\Omega, \mathcal{F}, P)$ tal que $\int_\Omega X dP$ existe, la **esperanza** o **valor esperado** de X es*

$$E(X) = \int_\Omega X dP.$$

Es frecuentemente inconveniente calcular $E(X)$ integrando sobre Ω; el siguiente resultado cuya demostración puede ser consultada en [3], Sección 4.10, expresa $E(X)$ como una integral con respecto a la medida de probabilidad inducida P_X, la cual, a su vez, es determinada por la función de distribución F.

Teorema 1.4.1 *Sean X un vector de dimensión n sobre $(\Omega, \mathcal{F}, P)$ y $g : \mathbb{R}^n \to \mathbb{R}$ una función Borel medible. Si $Y = g \circ X$, entonces*

$$E(Y) = \int_{\mathbb{R}^n} g(x) dF(x)$$

en el sentido de que si una integral existe la otra también y las dos son iguales.

La integral $\int_{\mathbb{R}^n} g(x) dF(x)$ significa $\int_{\mathbb{R}^n} g dP_X$, no es una integral de Riemann - Stieltjes.

Se tienen los siguientes casos especiales:

(i) Si $X : \Omega \to \mathbb{R}^n$ es un vector aleatorio absolutamente continuo con densidad f, entonces

$$\int_{\mathbb{R}^n} g(x) dF(x) = \int_{\mathbb{R}^n} g(x) f(x) dx$$

en el sentido de que si una integral existe la otra también y las dos son iguales.

(ii) Si X es un vector aleatorio discreto con función de probabilidad p, entonces

$$\int_{\mathbb{R}^n} g(x)dF(x) = \sum_x g(x)p(x),$$

en el sentido de que la integral existe si y sólo si la serie existe y son iguales.

Definición 1.4.2 *Sea X una variable aleatoria sobre $(\Omega, \mathcal{F}, P)$ y $k > 0$.*

*(i) El **k-ésimo momento** de X es $E(X^k)$.*

*(ii) Si $E(X)$ es finita, el **k-ésimo momento central** de X es $E\left[(X - E(X))^k\right]$.*

El primer momento, $E(X)$, es frecuentemente llamado la **media** de X. Al segundo momento central $\sigma^2 = E[(X - E(X))^2]$ se le llama la **varianza** de X y es algunas veces denotada por $VarX$; a la raíz cuadrada positiva de σ^2 se le llama la **desviación estándar** de X.

Notemos que $E(X^k)$ es finita si y sólo si $E(|X|^k)$ es finita. Además, si $0 < j < k$, entonces

$$\begin{aligned} E(|X|^j) &= \int_\Omega |X|^j dP = \int_{\{|X|^j<1\}} |X|^j dP + \int_{\{|X|^j\geqslant 1\}} |X|^j dP \\ &\leqslant P(|X|^j < 1) + \int_\Omega |X|^k dP = P(|X|^j < 1) + E(|X|^k). \end{aligned}$$

Tenemos así el siguiente resultado.

Proposición 1.4.1 *Si $k > 0$ y $E(X^k)$ es finita, entonces $E(X^j)$ es finita para todo $j \in (0, k)$.*

Usando el teorema del binomio y la aditividad de las integrales es fácil probar el siguiente resultado.

Proposición 1.4.2 *Si n es un entero positivo mayor que 1, $E(X^{n-1})$ es finita y $E(X^n)$ existe, entonces*

$$E\left[(X - E(X))^n\right] = \sum_{k=0}^{n} \binom{n}{k} \left[-E(X)\right]^{n-k} E(X^k).$$

Como $X^2 \geqslant 0$, $E(X^2)$ siempre existe y así, si $E(X)$ es finita, se tiene de la Proposición 1.4.2 que

$$VarX = E(X^2) - \left[E(X)\right]^2.$$

Ahora, si X es una variable aleatoria no negativa, $0 < p < \infty$ y $0 < \varepsilon < \infty$, entonces

$$E(X^p) = \int_\Omega X^p dP \geqslant \int_{\{X\geqslant\varepsilon\}} X^p dP \geqslant \varepsilon^p P(X \geqslant \varepsilon).$$

Tenemos así el siguiente resultado.

Proposición 1.4.3 *(Desigualdad de Chebyshev) Si X es una variable aleatoria no negativa, $0 < p < \infty$ y $0 < \varepsilon < \infty$, entonces*

$$P(X \geqslant \varepsilon) \leqslant \frac{E(X^p)}{\varepsilon^p}.$$

Así, si X es una variable aleatoria con media $E(X)$ finita, varianza σ^2 finita y $0 < k < \infty$, entonces

$$P(|X - E(X)| \geqslant k\sigma) \leqslant \frac{1}{k^2}.$$

Esto muestra que para una variable aleatoria con varianza pequeña, sus valores son cercanos a la media.

Una variable aleatoria con distribución normal tiene la útil propiedad de que la distribución es completamente deteminada por la media y la varianza. Específicamente, si X tiene la densidad normal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

entonces, usando que $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$ y $\int_{-\infty}^{\infty} x^2 e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$ es fácil ver que $m = E(X)$ y $\sigma^2 = VarX$.

El siguiente resultado sobre la esperanza de un producto de variables alatorias independientes es una consecuencia directa del teorema de Fubini.

Teorema 1.4.2 *Sean $X_1, ..., X_n$ variables aleatorias independientes sobre $(\Omega, \mathcal{F}, P)$. Si $X_i \geqslant 0$, $i = 1, ..., n$ o si $|E(X_i)| < \infty$, $i = 1, ..., n$, entonces $E(X_1 \cdots X_n)$ existe y*

$$E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n).$$

Definición 1.4.3 *Sean X y Y variables aleatorias para las cuales $E(X)$, $E(Y)$ y $E(XY)$ son finitas. La **covarianza** de X y Y es*

$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right].$$

Es fácil ver que

$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

Así, si X y Y son independientes, del Teorema 1.4.2 se sigue que

$$Cov(X, Y) = 0.$$

El recíproco, sin embargo, no es en general cierto (considérese por ejemplo, $X = cos\theta$ y $Y = sen\theta$, donde θ es una variable aleatoria con distribución uniforme en $[0, 2\pi]$).

Definición 1.4.4 *Sean X y Y variables aleatorias para las cuales $Var(X) > 0$ y $Var(Y) > 0$. La* ***correlación*** *de X y Y es*

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Si $X_1, ..., X_n$ son variables aleatorias con esperanza finita y $E(X_iX_j)$ es finita para todo $i, j = 1, ..., n$ con $i \neq j$, entonces

$$\begin{aligned} Var(X_1 + \cdots + X_n) &= E\left[(X_1 + \cdots + X_n - E(X_1) - \cdots - E(X_n))^2\right] \\ &= \sum_{i=1}^{n} E\left[(X_i - E(X_i))^2\right] + 2\sum_{\substack{i,j=1 \\ i<j}}^{n} E\left[(X_i - E(X_i))(X_j - E(X_j))\right]. \end{aligned} \tag{1.2}$$

Es también fácil ver que si $a_1, ..., a_n, b \in \mathbb{R}$,

$$Var(a_iX_i + b) = a_i^2 VarX_i \quad y \quad Cov(a_iX_i, a_jX_j) = a_ia_jCov(X_i, X_j), \tag{1.3}$$

$i, j = 1, \ldots, n$. De (1.2) y (1.3) es fácil probar lo siguiente.

Teorema 1.4.3 *Si $X_1, ..., X_n$ son variables aleatorias con esperanza finita y $E(X_iX_j)$ es finita para todo $i, j = 1, ..., n$ con $i \neq j$, entonces para $a_1, ..., a_n \in \mathbb{R}$,*

$$Var(a_1X_1 + \cdots + a_nX_n) = \sum_{i=1}^{n} a_i^2 VarX_i + 2\sum_{\substack{i,j=1 \\ i<j}}^{n} a_ia_jCov(X_i, X_j).$$

Tenemos así que si $X_1, ..., X_n$ son variables aleatorias incorrelacionadas, es decir, $Cov(X_i, X_j) = 0$ para $i, j = 1, ..., n$ con $i \neq j$, entonces

$$Var(X_1 + \cdots + X_n) = \sum_{i=1}^{n} VarX_i. \tag{1.4}$$

En particular (véase el Teorema 1.4.2), (1.4) se cumple si $X_1, ..., X_n$ son independientes.

1.5. Formas de convergencia

En este trabajo estudiaremos el comportamiento asintótico de los métodos de discriminación binaria para datos de dimensión alta. Por lo que incluimos esta sección donde presentamos diferentes tipos de convergencia de variables aleatorias, haciendo referencia también a vectores aleatorios.

Para considerar los casos de vectores aleatorios en las definiciones que presentamos haremos uso de la norma euclidiana, la cual denotaremos por $|\cdot|$.

Definición 1.5.1 *Sean $X, X_1, X_2, ...$ variables aleatorias (o vectores aleatorios) definidas sobre $(\Omega, \mathcal{F}, P)$. Diremos que*

(i) $X_n \to X$ ***uniformemente****, si dado $\varepsilon > 0$ existe $N = N(\varepsilon) \in \mathbb{N}$ tal que para cualquier $\omega \in \Omega$,*

$$|X_n(\omega) - X(\omega)| < \varepsilon, \quad \forall\ n \geqslant N.$$

(ii) $X_n \to X$ ***puntualmente****, si dados $\varepsilon > 0$ y $\omega \in \Omega$ existe $N = N(\varepsilon, \omega) \in \mathbb{N}$ tal que*

$$|X_n(\omega) - X(\omega)| < \varepsilon, \quad \forall\ n \geqslant N.$$

(iii) $X_n \to X$ ***casi seguramente*** *(c.s.), si existe un conjunto $E \in \mathcal{F}$ con $P(E) = 0$ tal que $X_n \to X$ puntualmente en $\Omega \setminus E$.*

Notemos que $(i) \Rightarrow (ii) \Rightarrow (iii)$.

En efecto, para $(i) \Rightarrow (ii)$ dado $\varepsilon > 0$ podemos tomar $N(\varepsilon, \omega) = N(\varepsilon)$. Para $(ii) \Rightarrow (iii)$ basta tomar a $E = \emptyset$.

Necesitamos de la siguiente definición para mostrar otro tipo de convergencia.

Definición 1.5.2 *Sea $p > 0$. Definimos el espacio $L^p = L^p(\Omega, \mathcal{F}, P)$ como la colección de todas las variables aleatorias (o vectores aleatorios) X tales que $E(|X|^p) = \int_\Omega |X|^p dP < \infty$.*

Ahora podemos dar la siguiente definición de convergencia.

Definición 1.5.3 *Sean $X, X_1, X_2, ...$ variables aleatorias (o vectores aleatorios) definidas sobre $(\Omega, \mathcal{F}, P)$. Diremos que $X_n \to X$ en* ***norma*** *$L^p, 0 < p < \infty$, si dado $\varepsilon > 0$ existe $N = N(\varepsilon)$ tal que*

$$E(|X_n - X|^p) < \varepsilon, \quad \forall\ n \geqslant N.$$

La demostración del siguiente resultado sobre la convergencia en L^p con respecto a diferentes valores de p, puede verse en [18], Sección 6.2.

Teorema 1.5.1 *La convergencia en L^p implica la convergencia en L^q, para cualquier q tal que $0 < q \leq p$.*

El siguiente resultado es utilizado para demostrar la completez de L^p.

Lema 1.5.1 *Si $Y_1, Y_2, ... \in L^p,\ p > 0$ y*

$$E(|Y_k - Y_{k+1}|^p) < \left(\frac{1}{4}\right)^k, \quad k = 1, 2, ...,$$

entonces $\{Y_k\}$ converge casi seguramente.

Demostración. Sea $A_k = \{|Y_k - Y_{k+1}| \geq 2^{-k}\}$. Entonces por la desigualdad de Chebyshev (Proposición 1.4.3)

$$P(A_k) \leq 2^{kp} E(|Y_k - Y_{k+1}|^p) < 2^{-kp}.$$

De aquí usando el Lema de Borel-Cantelli (Teorema 1.1.1) obtenemos que $P(\limsup_n A_n) = 0$. Pero si $\omega \notin \limsup_n A_n$, entonces $|Y_k(\omega) - Y_{k+1}(\omega)| < 2^{-k}$ para k grande, de manera que $\{Y_k(\omega)\}$ es una sucesión de Cauchy y por lo tanto convergente. ■

Definición 1.5.4 *Diremos que* $\{X_n\}$ *es de* ***Cauchy en norma*** L^p *si dado* $\varepsilon > 0$ *existe* $N = N(\varepsilon)$ *tal que*

$$E(|X_n - X_m|^p) < \varepsilon, \quad \forall\ n, m \geqslant N.$$

Teorema 1.5.2 *(Completez de* L^p, $1 \leqslant p < \infty$*) Si* $X_1, X_2, \ldots$ *es una sucesión de Cauchy en norma* L^p*, entonces existe* $X \in L^p$ *tal que* $E(|X_n - X|^p) \to 0$ *cuando* $n \to \infty$.

Demostración. Sea n_1 tal que $E(|X_n - X_m|^p) < \dfrac{1}{4}$ para $n, m \geq n_1$ y definamos $Y_1 = X_{n_1}$. En general, teniendo elegidas $Y_1, \ldots, Y_k$, y $n_1, \ldots, n_k$, consideremos $n_{k+1} > n_k$ de manera que $E(|X_n - X_m|^p) < \left(\dfrac{1}{4}\right)^{k+1}$ para $n, m \geq n_{k+1}$ y definamos $Y_{k+1} = X_{n_{k+1}}$. Por el Lema 1.5.1, tenemos que $\{Y_k\}$ converge casi seguramente a una variable aleatoria (o vector aleatorio) límite X. Dado $\varepsilon > 0$, elegimos N de manera que $E(|X_n - X_m|^p) < \varepsilon$ para $m, n \geq N$. Fijemos $n \geq N$ y hagamos $m \to \infty$ sobre los valores en la subsucesión, es decir, sea $m = n_k$, $k \to \infty$. Entonces, por el lema de Fatou

$$\begin{aligned} \varepsilon &\geq \liminf_{k\to\infty} E(|X_n - X_{n_k}|^p) = \liminf_{k\to\infty} \int |X_n - X_{n_k}|^p dP \\ &\geq \int \liminf_{k\to\infty} |X_n - Y_k|^p dP = E(|X_n - X|^p). \end{aligned}$$

Así, $E(|X_n - X|^p) \to 0$ y ya que $X = (X - X_n) + X_n$, tenemos que $X \in L^p$. ■

Un tipo de convergencia también de interés en este estudio es la convergencia en probabilidad.

Definición 1.5.5 *Sean* $X, X_1, X_2, \ldots$ *variables aleatorias (o vectores aleatorios) definidas sobre* $(\Omega, \mathcal{F}, P)$*. Diremos que* $X_n \to X$ *en* ***probabilidad*** *si y sólo si para cada* $\varepsilon > 0$ *y* $\eta > 0$*, existe* $N = N(\varepsilon, \eta) \in \mathbb{N}$ *tal que*

$$P(|X_n - X| > \varepsilon) < \eta, \quad \forall\ n \geqslant N. \tag{1.5}$$

La expresión (1.5) es equivalente a que $P(|X_n - X| > \varepsilon) \to 0$ cuando $n \to \infty$. El siguiente resultado muestra que la convergencia en L^p es más fuerte que la convergencia en probabilidad.

Teorema 1.5.3 *Sean $X, X_1, X_2, ... \in L^p$, $0 < p < \infty$. Si $X_n \to X$ en L^p, entonces $X_n \to X$ en probabilidad.*

Demostración. Dado $\varepsilon > 0$, tenemos por la desigualdad de Chebyshev, que

$$P(|X_n - X| \geq \varepsilon) \leq \frac{1}{\varepsilon^p} E(|X_n - X|^p).$$

Ahora, dado $\eta > 0$ tenemos por hipótesis que existe $N = N(\varepsilon, \eta)$ tal que

$$E(|X_n - X|^p) < \varepsilon^p \eta, \qquad \forall\, n \geqslant N.$$

Obtenemos así que

$$P(|X_n - X| > \varepsilon) < \eta \qquad \forall\, n \geqslant N.$$

■

Definición 1.5.6 *Diremos que $\{X_n\}$ es de* ***Cauchy en probabilidad*** *si para cada $\varepsilon > 0$ y $\eta > 0$, existe $N = N(\varepsilon, \eta) \in \mathbb{N}$ tal que*

$$P(|X_n - X_m| > \varepsilon) < \eta, \qquad \forall\, n, m \geqslant N.$$

De la desigualdad

$$P(|X_n - X_m| > \varepsilon) \leq P(|X_n - X| > \frac{\varepsilon}{2}) + P(|X_m - X| > \frac{\varepsilon}{2}), \quad n, m \geq N,$$

se observa que si $X_n \to X$ en probabilidad, entonces $\{X_n\}$ es de Cauchy en probabilidad.

De los teoremas 1.5.2 y 1.5.3 se obtiene que si $\{X_n\}$ es una sucesión de Cauchy en L^p, entonces $\{X_n\}$ es de Cauchy en probabilidad.

Veremos en el siguiente resultado que la convergencia casi segura es más fuerte que la convergencia en probabilidad.

Teorema 1.5.4 *Si $X_n \to X$ casi seguramente, entonces $X_n \to X$ en probabilidad.*

Demostración. Dado $\varepsilon > 0$ definamos

$$E_n(\varepsilon) = \{|X_n - X| > \varepsilon\}, \qquad n \geq 1.$$

Sea

$$D = \{\omega \in \Omega : X_n(\omega) \nrightarrow X(\omega) \text{ cuando } \ n \to \infty\}.$$

Luego

$$\begin{aligned} D &= \bigcup_{\varepsilon>0} \limsup_n E_n(\varepsilon) \\ &= \bigcup_{\varepsilon>0} \left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_m(\varepsilon) \right) \\ &= \bigcup_{k=1}^{\infty} \left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_m\left(\frac{1}{k}\right) \right). \end{aligned}$$

Se sigue que

$$X_n \to X \text{ c.s.} \Leftrightarrow P(D) = 0 \Leftrightarrow P\left(\limsup_n E_n(\varepsilon) \right) = 0, \quad \forall\, \varepsilon > 0.$$

Ya que $\bigcup_{m=n}^{\infty} E_m(\varepsilon) \downarrow \limsup_n E_n(\varepsilon)$, tenemos que

$$P\left(\limsup_n E_n(\varepsilon) \right) = \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} E_m(\varepsilon) \right).$$

Así que

$$X_n \to X \text{ c.s.} \Leftrightarrow \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} \{|X_m - X| \geq \varepsilon\} \right) = 0, \quad \forall\, \varepsilon > 0. \tag{1.6}$$

Y ya que para cualquier $n \geqslant 1$

$$\{|X_n - X| > \varepsilon\} \subset \bigcup_{m=n}^{\infty} \{|X_m - X| > \varepsilon\},$$

el resultado se sigue. ■

En el siguiente ejemplo mostramos que el recíproco del Teorema 1.5.3 no es en general válido.

Ejemplo 1 *Consideremos el espacio de probabilidad* $([0,1], \mathcal{B}([0,1]), P)$, *donde* P *es la medida de Lebesgue y definamos*

$$X_n(x) = \begin{cases} e^n, & si\ 0 \leqslant x \leqslant \dfrac{1}{n}, \\ 0, & en\ cualquier\ otra\ parte. \end{cases}$$

Entonces $X_n \to 0$ *c.s., y por lo tanto, en probabilidad, pero para cada* $p \in (0,\infty)$, $\{X_n\}$ *no converge en* L^p. *En efecto, para* $0 < p < \infty$

$$E(|X_n|^p) = \frac{1}{n} e^{np} \to \infty, \quad cuando\ n \to \infty.$$

Definición 1.5.7 *Una variable aleatoria (o vector aletorio) X es* ***acotada casi seguramente****, si existe una constante positiva $c < \infty$, tal que $P(|X| \leqslant c) = 1$.*

En el siguiente teorema presentamos un recíproco parcial del Teorema 1.5.3.

Teorema 1.5.5 *La convergencia en probabilidad para variables aleatorias (o vectores aleatorios) acotadas casi seguramente implica la convergencia en L^p, $p > 0$.*

Demostración. Supongamos que $X_n - X \to 0$ en probabilidad y que $P(|X_n - X| \leqslant c) = 1$, para algún $0 < c < \infty$. Entonces dado $\varepsilon > 0$ tenemos que

$$\begin{aligned} E\left(|X_n - X|^p\right) &= \int_{|x|\leq c} |x|^p dP_{|X_n - X|}(-\infty, x] \\ &= \int_{|x|\leq \varepsilon} |x|^p dP_{|X_n - X|}(-\infty, x] + \int_{\varepsilon < |x| \leq c} |x|^p dP_{|X_n - X|}(-\infty, x] \\ &\leqslant \varepsilon^p P(|X_n - X| \leqslant \varepsilon) + c^p P(|X_n - X| > \varepsilon). \end{aligned}$$

De aquí el resultado se sigue debido al hecho de que, para cualquier $p > 0$, la parte derecha puede hacerse arbitrariamente pequeña eligiendo ε pequeño y n suficientemente grande. ■

Teorema 1.5.6 *Si una sucesión de variables aleatorias es de Cauchy en probabilidad, existe una subsucesión que converge casi seguramente (y desde luego, en probabilidad) a una variable aleatoria.*

La demostración de este resultado puede ser consultada, por ejemplo, en [13], Sección 1.3.

El siguiente resultado nos muestra la completez de la convergencia en probabilidad.

Teorema 1.5.7 *(Completez de la Convergencia en Probabilidad) La sucesión $\{X_n\}$ converge en probabilidad a una variable aletoria X si y sólo si, $\{X_n\}$ es de Cauchy en probabilidad.*

Demostración. Sólo necesitamos probar que, si $\{X_n\}$ es de Cauchy en probabilidad, entonces converge en probabilidad. Del Teorema 1.5.6 se sigue que $\{X_n\}$ tiene una subsucesión $\{X_{n_k}\}$ que converge casi seguramente a una variable aleatoria X. Por lo tanto $X_{n_k} \to X$ en probabilidad. Claramente para cualquier $\varepsilon > 0$

$$P\left(|X_n - X| > \varepsilon\right) \leq P\left(|X_n - X_{n_k}| > \frac{\varepsilon}{2}\right) + P\left(|X_{n_k} - X| > \frac{\varepsilon}{2}\right).$$

Ya que $\{X_n\}$ es de Cauchy en probabilidad, el resultado deseado se sigue. ■

Mostramos a continuación más resultados sobre convergencia en probabilidad.

Teorema 1.5.8 *Sea $\{X_n\}$ una sucesión de variables aleatorias, tal que para alguna constante positiva c, $X_n \to c$ en probabilidad o casi seguramente. Si $g : \mathbb{R} \to \mathbb{R}$ es una función continua en $x = c$, entonces $g(X_n) \to g(c)$ en el mismo tipo de convergencia que $X_n \to c$.*

Demostración. Por la continuidad de g en c, tenemos que para cada $\varepsilon > 0$, existe un $\delta = \delta(\varepsilon) > 0$, tal que

$$|x - c| \leq \varepsilon \Rightarrow |g(x) - g(c)| \leq \delta. \tag{1.7}$$

Así, para cualquier n (fija) y $\varepsilon > 0$,

$$P\left(|X_n - c| \leq \varepsilon\right) \leq P\left(|g(X_n) - g(c)| \leq \delta\right); \tag{1.8}$$

$$P\left(\bigcap_{N \geq n} |X_N - c| \leq \varepsilon\right) \leq P\left(\bigcap_{N \geq n} |g(X_N) - g(c)| \leq \delta\right). \tag{1.9}$$

Así, si $X_n \to c$ en probabilidad, entonces $P(|X_n - c| \leq \varepsilon) \to 1$ cuando $n \to \infty$, por lo que el resultado se sigue de (1.8); Si $X_n \to c$ c.s., entonces (véase (1.6)) $P\left(\bigcap_{N \geq n} |X_n - c| \leq \varepsilon\right) \to 1$, cuando $n \to \infty$, por lo que el resultado se sigue de (1.9). ■

Podemos tener una generalización del Teorema 1.5.8, reemplazando c por una variable aleatoria X, para lo que se requerirá la continuidad uniforme de g. Esto es expresado en el siguiente teorema.

Teorema 1.5.9 *Si $X_n \to X$ en probabilidad o casi seguramente, y si g es uniformemente continua, entonces $g(X_n) \to g(X)$ en el mismo tipo de convergencia que $X_n \to X$.*

Demostración. Si g es uniformemente continua, entonces tenemos que (1.7) se cumple reemplazando c y $g(c)$ por X y $g(X)$, respectivamente, por lo que el resto de la demostración se sigue como en la demostración del Teorema 1.5.8.

■

Analizaremos, por último, un tipo de convergencia más, la convergencia en distribución de variables aleatorias.

Definición 1.5.8 *Sean $X, X_1, X_2, \ldots$ variables aleatorias (o vectores aleatorios) definidas sobre $(\Omega, \mathcal{F}, P)$, con funciones de distribución correspondientes $F, F_1, F_2, \ldots$. Diremos que $X_n \to X$ en **distribución** si para cada punto x en donde F es continua, dado $\varepsilon > 0$ existe un $N = N(\varepsilon, x)$, tal que*

$$|F_n(x) - F(x)| < \varepsilon, \qquad \forall\ n \geqslant N.$$

La convergencia en distribución es llamada también convergencia débil y se dice que $F_n \to F$ débilmente. El siguiente resultado nos muestra otra implicación de los tipos de convergencia.

Teorema 1.5.10 *La convergencia de variables aleatorias en probabilidad implica la convergencia en distribución.*

Demostración Supongamos que $X_n \to X$ en probabilidad y que X tiene función de distribución F. Sea x un punto de continuidad de F, y sea $\varepsilon > 0$ un número arbitrariamente pequeño. Entonces

$$\begin{aligned} F(x-\varepsilon) &= P(X \leqslant x-\varepsilon) \\ &= P(X \leqslant x-\varepsilon, |X_n - X| \leqslant \varepsilon) + P(X \leqslant x-\varepsilon, |X_n - X| > \varepsilon) \\ &\leqslant P(X_n \leqslant x) + P(|X_n - X| > \varepsilon) \end{aligned} \tag{1.10}$$

y

$$F(x+\varepsilon) \geqslant P(X_n \leqslant x) + P(|X_n - X| > \varepsilon). \tag{1.11}$$

Dado que los últimos términos del lado derecho de (1.10) y (1.11) convergen a 0 cuando $n \to \infty$, tenemos que

$$F(x-\varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon). \tag{1.12}$$

Ya que x es un punto de continuidad de F, $F(x+\varepsilon) - F(x-\varepsilon)$ puede hacerse arbitrariamente pequeño haciendo $\varepsilon \to 0$, obteniéndose así el resultado deseado. ■

Ahora, si X es una variable aleatoria degenerada en un punto c, es decir, $P(X = c) = 1$, entonces su función de distribución es

$$F(x) = \begin{cases} 0, & \text{si } x < c, \\ 1, & \text{si } x \geq c. \end{cases} \tag{1.13}$$

Por lo tanto si $X_n \to X$ en distribución, entonces $F_n(x) \to 0$ si $x < c$ y $F_n(x) \to 1$ si $x > c$, de manera que $X_n \to X$ en probabilidad. Esto conduce al siguiente resultado que será utilizado para demostrar una de las versiones de la Ley Débil de los Grandes Números.

Teorema 1.5.11 *Sean $X, X_1, X_2, \ldots$ variables aleatorias (o vectores aleatorios). Si X_n converge en distribución a una variable aleatoria (o vector aleatorio) X, degenerada en un punto c, entonces se tiene que $X_n \to c$ en probabilidad.*

Definimos ahora la función característica que nos será útil para el estudio de la Ley Débil de los Grandes Números.

Definición 1.5.9 *Para una función de distribución F, definida en $\mathbb{R}$, la* ***función característica*** $\phi_F(t),\ t \in \mathbb{R}$, *es*

$$\phi_F(t) = \int_{\mathbb{R}} \exp(itx) dF(x), \qquad t \in \mathbb{R}.$$

En general, la función característica toma valores complejos. Si la función de distribución F es degenerada en el punto $c \in \mathbb{R}$ (satisface (1.13)), entonces $\phi_F(t) = \exp(itc)$, para cualquier $t \in \mathbb{R}$. El teorema de Lévy clásico dice que si dos funciones de distribución F y G tienen la misma función característica $\phi_F(t)$, $t \in \mathbb{R}$, entonces F y G son iguales, y el recíproco es también válido. Un resultado más general en este contexto, que establecemos aquí sin demostración, es el siguiente.

Teorema 1.5.12 *(Teorema de Lévy-Cramér). Sea $\{F_n\}$ una sucesión de funciones de distribución en $\mathbb{R}$ con funciones características correspondientes $\{\phi_{F_n}\}$. Entonces $\{F_n\}$ converge débilmente a una función de distribución F si y sólo si, para cualquier $t \in \mathbb{R}$, $\{\phi_{F_n}(t)\}$ converge a un límite $\phi_F(t)$, el cual es una función continua en $t = 0$ y es la función característica de la función de distribución F.*

Presentamos ahora tres versiones de la Ley Débil de los Grandes Números, llamada así por considerar convergencia en probabilidad.

Teorema 1.5.13 *(Ley Débil de los Grandes Números) Sean $X_1, X_2, \ldots$ variables aleatorias independientes (no necesariamente con la misma distribución), cada una con media y varianza finitas. Supongamos que las varianzas son uniformemente acotadas por $M < \infty$. Sea $S_n = X_1 + X_2 + \cdots + X_n$. Entonces $\dfrac{S_n - E(S_n)}{n}$ converge en probabilidad a 0, es decir, dado $\varepsilon > 0$*

$$P\left[\left|\frac{S_n - E(S_n)}{n}\right| > \varepsilon\right] \to 0, \quad \textit{cuando } n \to \infty.$$

Demostración. El resultado se obtiene aplicando la desigualdad de Chebyshev, en efecto

$$0 \leq P\left[\left|\frac{S_n - E(S_n)}{n}\right| > \varepsilon\right] \leq \frac{1}{\varepsilon^2} E\left[\left(\frac{S_n - E(S_n)}{n}\right)^2\right] = \frac{1}{\varepsilon^2 n^2} \sum_{k=1}^{n} Var(X_k) \to 0,$$

cuando $n \to \infty$. ■

Definición 1.5.10 *Si las variables aleatorias $X_1, X_2, \ldots$ tienen todas la misma distribución, diremos que son* ***idénticamente distribuidas****.*

La frase "independientes e idénticamente distribuidas" será abreviado i.i.d.

Teorema 1.5.14 *(Ley Débil de los Grandes Números de Khintchine) Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. con $E(X_1) = \theta$, y sea $\overline{X}_n = n^{-1}\sum_{k=1}^{n} X_k$, $n = 1, 2, \ldots$. Entonces $\overline{X}_n \to \theta$ en probabilidad.*

Demostración. Sea $F(x)$, $x \in \mathbb{R}$, la función de distribución de X_1 y sea $\phi_F(t)$, $t \in \mathbb{R}$ la función característica de X_1. Notemos que cuando $n \to \infty$,

$$\begin{aligned}\phi_{\overline{X}_n}(t) &= E\left(\prod_{k=1}^{n} e^{it\overline{X}_k/n}\right) = [\phi_F(t/n)]^n \\ &= \left\{1 + itn^{-1}E(X_1) + o(n^{-1})t\right\}^n \to \exp(it\theta).\end{aligned}$$

Ya que $\exp(it\theta)$ es la función característica de una variable Z, que es degenerada en el punto θ, el resultado se obtiene aplicando el Teorema 1.5.11. ■

El siguiente resultado lo damos sin demostración, pero puede ser consultada, por ejemplo, en [18], Sección 6.3.

Teorema 1.5.15 *(Ley Débil de los Grandes Números de Markov) Sean $X_1, X_2, \ldots$ variables aleatorias independientes, tales que $E(X_k), k = 1, 2, \ldots$ existe, y que para algún δ con $0 < \delta \leq 1$, $E(|X_k - E(X_k)|^{1+\delta})$ existe. Supongamos que*

$$n^{-1-\delta}\sum_{k=1}^{n} E(|X_k - E(X_k)|^{1+\delta}) = \rho_n(\delta) \to 0, \quad \text{cuando } n \to \infty. \tag{1.14}$$

Entonces $\overline{X}_n - E(\overline{X}_n) \to 0$ en probabilidad.

Los siguientes dos resultados (cuya demostración puede consultarse en [3], Sección 6.1) son clave para justificar la Ley Fuerte de los Grandes Números.

Lema 1.5.2 *(Lema de Kronecker) Sea $\{b_n\}$ una sucesión creciente de números reales positivos con $b_n \to \infty$ y sea $\{x_n\}$ una sucesión de números reales con $\sum_{n=1}^{\infty} x_n = x$ (finito). Entonces*

$$\frac{1}{b_n}\sum_{j=1}^{n} b_j x_j \to 0, \qquad \text{cuando } n \to \infty.$$

Teorema 1.5.16 *Sean $X_1, X_2, \ldots$ variables aleatorias independientes con esperanza finita. Si $\sum_{n=1}^{\infty} Var(X_n) < \infty$, entonces $\sum_{n=1}^{\infty} [X_n - E(X_n)]$ converge casi seguramente.*

Presentamos a continuación la Ley Fuerte de los Grandes Números en dos de sus versiones, llamada así por considerar la convergencia casi segura.

Teorema 1.5.17 *(Ley Fuerte de los Grandes Números de Kolmogorov) Sean $X_1, X_2, ...$ variables aleatorias independientes, cada una con media y varianza finitas. Si $\sum_{n=1}^{\infty} \frac{Var(X_n)}{n^2} < \infty$, entonces (con $S_n = X_1 + \cdots X_n$)*

$$\frac{S_n - E(S_n)}{n} \to 0 \quad \textit{casi seguramente.}$$

Demostración. Por hipótesis obtenemos que

$$\sum_{n=1}^{\infty} Var\left(\frac{X_n - E(X_n)}{n}\right) = \sum_{n=1}^{\infty} \frac{Var\,(X_n)}{n^2} < \infty.$$

Por el Teorema 1.5.16, $\sum_{n=1}^{\infty} \frac{X_n - E(X_n)}{n}$ converge casi seguramente. Pero

$$\frac{S_n - E(S_n)}{n} = \frac{1}{n} \sum_{k=1}^{n} k \left(\frac{X_k - E(X_k)}{k}\right),$$

lo cual, por el Lema de Kronecker (Lema 1.5.2), tiende a 0 casi seguramente. ■

En particular, si $X_1, X_2, ...$ son variables aleatorias independientes cada una con media finita m y varianza finita σ^2, entonces $\frac{S_n}{n} \to m$ casi seguramente.

Los siguientes dos resultados (cuya demostración puede consultarse en [18], Sección 7.2), junto con el Teorema 1.5.8, constituyen una herramienta importante en la búsqueda de la distribución asintótica de transformaciones de estadísticos con distribuciones asintóticas conocidas.

Teorema 1.5.18 *(Sverdrup) Sea $\{X_n\}$ una sucesión de variables aleatorias tal que $X_n \to X$ en distribución y $g : \mathbb{R} \to \mathbb{R}$ una función continua. Entonces $g(X_n) \to g(X)$ en distribución.*

Teorema 1.5.19 *(Cramér-Wold). Sean $X, X_1, X_2, ...$ vectores aleatorios en $\mathbb{R}^p$; entonces $X_n \to X$ en distribución, si y sólo si, para cada $\lambda \in \mathbb{R}^p$ fijo, tenemos que $\lambda' X_n \to \lambda' X$ en distribución.*

Mostramos a continuación los siguientes dos resultados, los cuales son conocidos como **teoremas del límite central** y aunque son sólo casos especiales de resultados más generales, son suficientes para situaciones de interés en este trabajo.

Teorema 1.5.20 *(Teorema del Límite Central) Sean $X_1, X_2, ...$ variables aleatorias i.i.d. con media μ y varianza finita σ^2. Definimos*

$$Z_n = (S_n - n\mu)/\sigma\sqrt{n}.$$

Entonces, $Z_n \to Z$ en distribución, donde $Z \sim N(0, 1)$.

Demostración. Sea $\phi_{Z_n}(t) = E[\exp(itZ_n)]$ y observemos que

$$\begin{aligned}\phi_{Z_n}(t) &= E\left\{\exp\left[\frac{it}{\sigma\sqrt{n}}\left(\sum_{k=1}^{n} X_k - n\mu\right)\right]\right\} \\ &= \prod_{k=1}^{n} E\left\{\exp\left[\frac{it}{\sigma\sqrt{n}}\left(X_k - \mu\right)\right]\right\} \\ &= \left\{E\left[\exp\left(\frac{it}{\sigma\sqrt{n}}U\right)\right]\right\}^n = \left[\phi_U\left(\frac{t}{\sqrt{n}}\right)\right]^n,\end{aligned}$$

donde $U = (X_1 - \mu)/\sigma$ y $\phi_U(t) = E[\exp(itU)]$. Considerando una expansión de Taylor de $\phi_U(t/\sqrt{n})$ y dado que $E(U) = 0$ y $E(U^2) = 1$, tenemos que

$$\begin{aligned}\phi_U\left(\frac{t}{\sqrt{n}}\right) &= 1 + iE(U)\frac{t}{\sqrt{n}} + i^2\frac{t^2}{2n}E(U^2) + o\left(\frac{t^2}{n}\right) \\ &= 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).\end{aligned}$$

Así, cuando $n \to \infty$, tenemos

$$\phi_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \to \exp\left(\frac{-t^2}{2}\right),$$

que es la función característica de una distribución normal estándar, lo que completa la demostración. ■

Una extensión del Teorema 1.5.20 en que se considera la suma de variables aleatorias indepedientes pero no necesariamente idénticamente distribuidas, es dado por el siguiente teorema, cuya demostración puede consultarse en [18], Sección 7.3.

Teorema 1.5.21 *(Liapounov) Sean $X_1, X_2, \ldots$ variables aleatorias independientes, tales que $E(X_k) = \mu_k$ y $Var(X_k) = \sigma_k^2$ y para algún $0 < \delta \leq 1$,*

$$\nu_{2+\delta}^{(k)} = E(\mid X_k - \mu_k \mid^{2+\delta}) < \infty, \quad k \geq 1.$$

Sean también $S_n = \sum_{k=1}^n X_k$, $\xi_n = E(S_n) = \sum_{k=1}^n \mu_k$, $\tau_n^2 = Var(S_n) = \sum_{k=1}^n \sigma_k^2$, $Z_n = (S_n - \xi_n)/\tau_n$ y $\rho_n = \tau_n^{-(2+\delta)} \sum_{k=1}^n \nu_{2+\delta}^{(k)}$. Entonces, si $\lim_{n\to\infty} \rho_n = 0$, tenemos que, $Z_n \to Z$ en distribución, donde $Z \sim N(0,1)$.

A continuación mostramos otras definiciones y resultados que utilizaremos en nuestro trabajo.

Definición 1.5.11 *Sea $\{X_n\}$ una sucesión de variables aleatorias (o vectores aleatorios). Diremos que $X_n = O_p(1)$, si para cada $\eta > 0$ existe una constante positiva finita $K \equiv K(\eta)$ y $N \equiv N(\eta) \in \mathbb{N}$ tal que*

$$P(|X_n| \leqslant K) \geqslant 1-\eta, \qquad \forall\ n \geqslant N.$$

Más generalmente, si $\{Y_n\}$ es otra sucesión de variables aleatorias (o vectores aleatorios), diremos que $X_n = O_p(Y_n)$, si para cada $\eta > 0$ existe una constante positiva finita $K \equiv K(\eta)$ y $N \equiv N(\eta) \in \mathbb{N}$ tal que

$$P(|\tfrac{X_m}{Y_m}| \leqslant K) \geqslant 1-\eta, \qquad \forall\ n \geqslant N.$$

En ocaciones, para indicar que $X_n = O_p(Y_n)$ se dice que $\left\{\frac{X_n}{Y_n}\right\}$ es una $O_p(1)$. Tenemos el siguiente resultado mencionado en [21], Sección 1.2.

Teorema 1.5.22 *Si $X_n \to X$ en distribución, entonces $\{X_n\}$ es una $O_p(1)$.*

Definición 1.5.12 *Diremos que $X_n = o_p(1)$, si para cualesquiera $\varepsilon > 0$ y $\eta > 0$, existe $N \equiv N(\varepsilon, \eta) \in \mathbb{N}$, tal que*

$$P(|X_n| > \varepsilon) < \eta, \qquad \forall\ n \geqslant N.$$

En otras palabras, $X_n = o_p(1)$ es equivalente a decir que $X_n \to 0$ en probabilidad.

Además, diremos que $X_n = o_p(Y_n)$, si para cualesquiera $\varepsilon > 0$ y $\eta > 0$, existe $N \equiv N(\varepsilon, \eta) \in \mathbb{N}$, tal que

$$P(|\tfrac{X_n}{Y_n}| > \varepsilon) < \eta, \qquad \forall\ n \geqslant N.$$

En el siguiente teorema, cuya demostración puede ser consultada en [8], Sección 5.5, $X_n \to N(a, b)$ en distribución indica que $\{X_n\}$ converge a una variable aleatoria normal con media a y varianza b.

Teorema 1.5.23 *(Método Delta) Sea $\{Y_n\}$ una sucesión de variables aleatorias tales que $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ en distribución. Para una función g y un valor específico de θ, suponga que $g'(\theta)$ existe y es distinto de 0. Entonces*

$$\sqrt{n}\,[g(Y_n) - g(\theta)] \to N(0, \sigma^2[g'(\theta)]^2)$$

en distribución.

Capítulo 2

La Normal Multivariada y Representación Geométrica para Datos de Dimensión Alta

La distribución normal multivariada es la principal y más básica en el Análisis Multivariado. Esto se debe a que en la mayoría de los casos en que se consideran observaciones multivariadas éstas presentan, aproximadamente, una distribución normal, especialmente si nos referimos a la media muestral, debido al efecto del Teorema de Límite Central. Este efecto también es usado cuando las observaciones pueden ser consideradas como sumas de vectores aleatorios independientes.

Las definiciones y resultados que se presentan en las primeras cuatro secciones de este capítulo son extraídas de [16], por lo cual las demostraciones que no sean presentadas aquí pueden ser consultadas en la Sección 1.2 del mismo.

2.1. Preliminares

Mencionaremos primero algunas propiedades de los vectores aleatorios, necesarias para proceder a definir la distribución normal multivariada.

Definición 2.1.1 ***La media o esperanza*** *de un vector aleatorio* $X = (X_1, ..., X_m)'$ *es definida como el vector de esperanzas*

$$E(X) = \begin{pmatrix} E(X_1) \\ \cdot \\ \cdot \\ \cdot \\ E(X_m) \end{pmatrix}.$$

Generalizando aún más, si $Z = (Z_{ij})$ es una matriz aleatoria $p \times q$, entonces $E(Z)$, la esperanza de Z, es la matriz cuyo $ij-$ésimo elemento es $E(Z_{ij})$. Además, si B, C, y D son matrices de constantes $m \times p$, $q \times n$ y $m \times n$, entonces

$$E(BZC + D) = BE(Z)C + D. \tag{2.1}$$

Definición 2.1.2 ***La matriz de covarianza*** *de un vector aleatorio $X = (X_1, ..., X_m)'$, con media μ, es definida como*

$$\Sigma \equiv Cov(X) = E[(X - \mu)(X - \mu)'].$$

El $ij-$ésimo elemento de Σ es

$$\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)],$$

la covarianza entre X_i y X_j, y el $ii-$ésimo elemento es

$$\sigma_{ii} = E[(X_i - \mu_i)^2],$$

la varianza de X_i, por lo que los elementos de la diagonal de Σ deben de ser no negativos. Obviamente Σ es simétrica, es decir, $\Sigma = \Sigma'$.

Frecuentemente tenemos transformaciones lineales de vectores aleatorios y deseamos conocer sus matrices de covarianza. Para ver esto, supongamos que X es un vector aleatorio $m \times 1$ con media μ_x y matriz de covarianza Σ_x y sea $Y = BX + b$, donde B es una matriz $k \times m$ y b es $k \times 1$. La media de Y es, por (2.1), $\mu_y = B\mu_x + b$, y la matriz de covarianza de Y es

$$\Sigma_y = B\Sigma_x B'. \tag{2.2}$$

Otra definición que será utilizada es la función característica de un vector aleatorio, la cual se da a continuación.

Definición 2.1.3 *Para X un vector aleatorio $m \times 1$, la función característica es*

$$\phi_X(t) = E\left[\exp(it'X)\right], \quad \textit{donde } t \in \mathbb{R}^m.$$

2.2. Definición y propiedades

Antes de definir la distribución normal multivariada daremos el siguiente resultado, cuya demostración puede consultarse en [16].

Teorema 2.2.1 *Si X es un vector aleatorio $m \times 1$, entonces su distribución está determinada únicamente por las distribuciones de las funciones lineales $\alpha'X$, para cualquier $\alpha \in \mathbb{R}^m$.*

Procederemos a definir la distribución normal multivariada.

Definición 2.2.1 *Sea X un vector aleatorio $m \times 1$. Decimos que X tiene una distribución normal m-variada si, para cualquier $\alpha \in \mathbb{R}^m$, la distribución de $\alpha' X$ es normal univariada.*

Teorema 2.2.2 *Si X tiene una distribución normal m-variada, entonces $\mu \equiv E(X)$ y $\Sigma \equiv Cov(X)$ existen y la distribución de X es determinada por μ y Σ.*

Demostración. Si $X = (X_1, \ldots, X_m)'$, entonces, para cada $i = 1, 2, \ldots, m$, X_i es normal univariada (por la Definición 2.2.1), así que $E(X_i)$ y $Var(X_i)$ existen y son finitas. Así, $Cov(X_i, X_j)$ existe (puesto que $|Cov(X_i, X_j)|$ está acotada por $\sqrt{Var(X_i)Var(X_j)}$). Definiendo $\mu = E(X)$ y $\Sigma = Cov(X)$, de (2.1) y (2.2), tenemos

$$E(\alpha' X) = \alpha' \mu$$

y

$$Var(\alpha' X) = \alpha' \Sigma \alpha$$

así que la distribución de $\alpha' X$ es $N(\alpha' \mu, \alpha' \Sigma \alpha)$ para cada $\alpha \in \mathbb{R}^m$. Ya que estas distribuciones univariadas son determinadas por μ y Σ, así lo es también la distribución de X, debido al Teorema 2.2.1. ■

La distribución normal m-variada de un vector aleatorio X será denotada por $N_m(\mu, \Sigma)$ y escribiremos $X \sim N_m(\mu, \Sigma)$. Ahora mostraremos algunas de las propiedades de la distribución normal multivariada.

Teorema 2.2.3 *Si $X \sim N_m(\mu, \Sigma)$ entonces la función característica de X es*

$$\phi_X(t) = \exp(it'\mu - \frac{1}{2} t'\Sigma t). \tag{2.3}$$

Demostración. Observemos que

$$\phi_X(t) = E\left[\exp(it'X)\right] = \phi_{t'X}(1),$$

donde el lado derecho denota la función característica de la variable aleatoria $t'X$ evaluada en 1. Ya que $X \sim N_m(\mu, \Sigma)$ entonces $t'X \sim N_m(t'\mu, t'\Sigma t)$ por lo que

$$\phi_X(t) = \phi_{t'X}(1) = \exp(it'\mu - \frac{1}{2} t'\Sigma t),$$

completando la demostración. ■

Podríamos dudar sobre la existencia de vectores aleatorios que cumplan la Definición 2.2.1, puesto que no se ha demostrado que la colección es no vacía. Esto podemos hacerlo mostrando que la función dada por (2.3) es ciertamente la función característica de

algún vector aleatorio. Para esto, consideremos a Σ una matriz de covarianza $m \times m$ de rango r y sean $U_1, ..., U_r$ variables aleatorias normal estándar independientes. Entonces el vector $U = (U_1, ..., U_r)'$ tiene función característica

$$\phi_U(t) = E\left[\exp(it'U)\right] = \exp(-\frac{1}{2}t't).$$

Ahora, definamos

$$X = CU + \mu \tag{2.4}$$

donde C es una matriz $m \times r$ de rango r tal que $\Sigma = CC'$, y $\mu \in \mathbb{R}^m$, entonces X tiene función característica (2.3), puesto que

$$E\left[\exp(it'X)\right] = E\left[\exp(it'CU)\right]\exp(it'\mu) = \phi_U(C't)\exp(it'\mu)$$

$$= \exp(-\frac{1}{2}t'CC't)\exp(it'\mu) = \exp(it'\mu - \frac{1}{2}t'\Sigma t).$$

Luego, observemos que podemos tener definida la distribución normal multivariada $N_m(\mu, \Sigma)$ por medio de la transformación lineal (2.4) de variables aleatorias independientes normal estándar.

Otra de las propiedades de la distribución normal multivariada es que cualquier transformación lineal de un vector normal tiene también distribución normal, tal como lo muestra el siguiente resultado, el cual es verificado directamente de la Definición 2.2.1.

Teorema 2.2.4 *Si $X \sim N_m(\mu, \Sigma)$ y B es una matriz $k \times m$, b es un vector $k \times 1$ entonces*

$$Y = BX + b \sim N_k(B\mu + b, B\Sigma B').$$

Demostración. El hecho de que Y es una normal k-variada es una consecuencia directa de la Defición 2.2.1, ya que todas las funciones lineales de las componentes de Y son funciones lineales de las componentes de X y éstas son normales. La media y la matriz de covarianza son claramente las que se indican. ∎

Ahora nuestro problema consistirá en encontrar la función de densidad de un vector aleatorio X con distribución $N_m(\mu, \Sigma)$. Para esto, hablaremos de algunos conceptos que involucran a X y Σ. Es bien conocido que la clase de matrices de covarianza coincide con la clase de matrices **no negativas definidas** (ver Definición en A.2).

En lo referente a esto, un resultado nos dice que si la matriz covarianza Σ de un vector aleatorio X no es **positiva definida** (ver Definición en A.2) entonces, con

probabilidad 1, las componentes X_i de X están linealmente relacionadas. Es decir, existe $\alpha \in \mathbb{R}^m$, $\alpha \neq 0$, tal que

$$Var(\alpha' X) = \alpha' \Sigma \alpha = 0$$

así que, con probabilidad 1, $\alpha' X = k$, donde $k = \alpha' E(X)$ de tal manera que X vive en un hiperplano, así que una función de densidad para X (con respecto a la medida de Lebesgue en $\mathbb{R}^m$) puede no existir. En este caso se dice que X tiene una distribución normal **singular**. Sin embargo, si Σ es positiva definida la función de densidad existe y es fácil encontrarla si se usa la representación (2.4) de X en términos de variables aleatorias independientes normal estándar.

Teorema 2.2.5 *Si $X \sim N_m(\mu, \Sigma)$ y Σ es positiva definida, entonces la función de densidad de X es*

$$f_X(x) = (2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu) \right]$$

(donde $|A|$ se usa para denotar al determinante de una matriz A).

Demostración. Sea $\Sigma = CC'$ donde C es un matriz $m \times m$ no singular y definamos

$$X = CU + \mu,$$

donde U es un vector $m \times 1$ de variables aletorias normal estándar independientes, es decir, $U \sim N_m(0, I_m)$. La función de densidad conjunta de $U_1, \ldots, U_m$ es

$$f_U(u) = \prod_{i=1}^{m} (2\pi)^{-1/2} \exp\left(-\frac{1}{2} u_i^2 \right) = (2\pi)^{-m/2} \exp\left(-\frac{1}{2} u'u \right), \quad \text{con } u \in \mathbb{R}^m.$$

La transformación inversa es $U = B(X - \mu)$, con $B = C^{-1}$, y el Jacobiano de esta tranformación es

$$\begin{vmatrix} \frac{\partial u_1}{\partial x_1} & \cdots & \frac{\partial u_1}{\partial x_m} \\ \vdots & \vdots & \vdots \\ \frac{\partial u_m}{\partial x_1} & \cdots & \frac{\partial u_m}{\partial x_m} \end{vmatrix} = \begin{vmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mm} \end{vmatrix} = |B| = \left|C^{-1}\right| = |C|^{-1} = |CC'|^{-1/2}$$
$$= |\Sigma|^{-1/2}$$

así que la función de densidad de X es

$$f_X(x) = (2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)'(C^{-1})' C^{-1} (x-\mu) \right];$$

y ya que $\Sigma^{-1} = (C^{-1})'C^{-1}$, tenemos el resultado. ■

En particular, la función de densidad normal bivariada "estándar" se escribe como

$$f_Z(z_1, z_2) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left[-\frac{1}{2(1-\rho^2)}(z_1^2 + z_2^2 - 2\rho z_1 z_2)\right],$$

donde $Var(X_1) = \sigma_1^2$, $Var(X_2) = \sigma_2^2$, ρ es la correlación entre X_1 y X_2 y $Z_i = (X_i - \mu_i)/\sigma_i$, $i = 1, 2$.

2.3. Distribuciones marginales y condicionales

Otra propiedad de mucha importancia es que todas las distribuciónes marginales son normales.

Teorema 2.3.1 *Si $X \sim N_m(\mu, \Sigma)$, entonces la distribución marginal de cualquier subconjunto de $k(< m)$ componentes de X es normal k-variada.*

Demostración. Particionemos X, μ, y Σ como

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

donde X_1 y μ_1 son vectores $k \times 1$ y Σ_{11} es una submatriz $k \times k$. Tomamos $B = [I_k : 0]$ que es una matriz $k \times m$ y $b = 0$, en el Teorema 2.2.4. De aquí, se obtiene inmediatamente que X_1 es $N_k(\mu_1, \Sigma_{11})$. Similarmente, la distribución marginal de cualquier subvector de k componentes de X es normal k-variada, donde la media y la matriz de covarianza se obtienen escogiendo el subvector y la submatriz de una manera evidente. ■

Una consecuencia directa de este teorema es que la distribución marginal de cada componente de X es normal univariada. El inverso en general no es cierto; es decir, el hecho de que cada componente de un vector aleatorio es (marginalmente) normal no implica que el vector tenga una distribución normal multivariada. Sin embargo, este hecho es cierto, si las componentes de X son todas normales e independientes.

Recordemos que la independencia de dos variables aleatorias implica que la covarianza entre ellas, si existe, es cero, pero que el inverso no es cierto en general. Sin embargo, para la distribución normal multivariada, este resultado es cierto en los dos sentidos como se muestra en el siguiente resultado.

Teorema 2.3.2 *Si $X \sim N_m(\mu, \Sigma)$ y X, μ, y Σ son particionados como*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

donde X_1 y μ_1 son vectores $k \times 1$ y Σ_{11} es una matriz $k \times k$, entonces los subvectores X_1 y X_2 son independientes si y sólo si $\Sigma_{12} = 0$.

Demostración. Σ_{12} es la matriz de covarianzas entre las componentes de X_1 y las componentes de X_2, así, por la independencia de X_1 y X_2, tenemos que $\Sigma_{12} = 0$. Sean Y_1 y Y_2 vectores aleatorios independientes donde Y_1 es $N_k(\mu_1, \Sigma_{11})$ y Y_2 es $N_{m-k}(\mu_2, \Sigma_{22})$ y definamos $Y = (Y_1', Y_2')'$. Entonces, tanto X como Y son $N_m(\mu, \Sigma)$, donde

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix},$$

por lo que están idénticamente distribuidos. Por lo tanto, X_1 y X_2 son independientes. ■

El siguiente teorema muestra que la distribución condicional de un subvector de un vector con distribución normal, dadas las componentes restantes, es también normal.

Teorema 2.3.3 *Si $X \sim N_m(\mu, \Sigma)$ y X, μ, y Σ son particionados como*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

donde X_1 y μ_1 son vectores $k \times 1$ y Σ_{11} es una matriz $k \times k$. Sea Σ_{22}^- la inversa generalizada de Σ_{22}, es decir, una matriz que satisface

$$\Sigma_{22}\Sigma_{22}^-\Sigma_{22} = \Sigma_{22}$$

y sea $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21}$. Entonces

(i) $X_1 - \Sigma_{12}\Sigma_{22}^- X_2 \sim N_k(\mu_1 - \Sigma_{12}\Sigma_{22}^-\mu_2, \Sigma_{11\cdot 2})$
y es independiente de X_2, y

(ii) La distribución condicional de X_1 dado X_2 es

$$N_k(\mu_1 + \Sigma_{12}\Sigma_{22}^-(X_2 - \mu_2), \Sigma_{11\cdot 2}).$$

La media de la distribución condicional de X_1 dado X_2

$$E(X_1 \mid X_2) = \mu_1 + \Sigma_{11}\Sigma_{22}^-(X_2 - \mu_2)$$

es llamada la función regresión de X_1 sobre X_2 con coeficiente la matriz de regresión $\Sigma_{11}\Sigma_{22}^-$.

Una caracterización de las más importantes de la distribución normal univariada se presenta en el siguiente lema, la cual se sigue teniendo para el caso de la distribución normal multivariada.

Lema 2.3.1 *Si X y Y son variables aleatorias independientes cuya suma $X+Y$ tiene distribución normal, entonces tanto X como Y tienen distribución normal.*

Teorema 2.3.4 *Si los vectores aleatorios X y Y de $m \times 1$ son independientes y $X+Y$ tiene ditribución normal m-variada, entonces X y Y son normales.*

Demostración. Para cada $\alpha \in \mathbb{R}^m$, $\alpha'(X+Y) = \alpha'X + \alpha'Y$ es normal (por la Definición 2.2.1, ya que $X+Y$ es normal). Dado que $\alpha'X$ y $\alpha'Y$ son independientes, el Lema 2.3.1 implica que ambos son normales, y así, tanto X como Y tienen distribución normal. ■

2.4. Media y covarianza muestral

Otra propiedad de la distribución normal univariada que también se puede extender a la normal m-variada es que cualquier combinación lineal de variables aleatorias normales independientes es normal, como se describe en el siguiente teorema.

Teorema 2.4.1 *Si $X_1, ..., X_N$ son todos independientes y $X_i \sim N_m(\mu_i, \Sigma_i)$ para $i = 1, ..., N$, entonces para cualesquiera constantes fijas $\alpha_1, ..., \alpha_N$,*

$$\sum_{i=1}^{N} \alpha_i X_i \sim N_m\left(\sum_{i=1}^{N} \alpha_i \mu_i, \sum_{i=1}^{N} \alpha_i^2 \Sigma_i\right)$$

El siguiente corolario nos muestra la distribución del vector media muestral.

Corolario 2.4.1 *Si $X_1, ..., X_N$ son todos independientes y cada uno de ellos tiene distribución $N_m(\mu, \Sigma)$, entonces la distribución del* ***vector media muestral***

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

es $N_m(\mu, \frac{1}{N}\Sigma)$.

Este corolario nos dice que

$$\sqrt{N}(\bar{X} - \mu) = \frac{\sum_{i=1}^{N}(X_i - \mu)}{\sqrt{N}} \sim N_m(0, \Sigma).$$

Mas aún, cuando los vectores $X_1, ..., X_N$ no son normales, pero tenemos que $E(\bar{X}) = \mu$ y $Cov(\bar{X}) = \frac{1}{N}\Sigma$, podemos hablar de la distribución asintótica de $\bar{X}$, como se enuncia en la siguiente versión del Teorema del Límite Central.

Teorema 2.4.2 *Sea $X_1, X_2, ...$ una sucesión de vectores aleatorios independientes e identicamente distribuidos con media μ y matriz de covarianza Σ y sea*

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad N \geqslant 1.$$

Entonces, cuando $N \to \infty$, la distribución asintótica de

$$\sqrt{N}(\bar{X}_N - \mu) = \frac{\sum_{i=1}^{N}(X_i - \mu)}{\sqrt{N}}$$

es $N_m(0, \Sigma)$.

Mencionaremos ahora una aplicación de este teorema. Sea $X_1, ..., X_N$ una muestra aleatoria de cualquier distribución m-variada y supongamos que

$$E(X_i) = \mu \qquad \text{y} \qquad Cov(X_i) = \Sigma \qquad (i = 1, ..., N).$$

Definamos

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})',$$

donde $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. A S^2 se le conoce como la **matriz de covarianza muestral** y es un estimador insesgado de Σ, es decir, $E(S^2) = \Sigma$. Ahora, si T es una matriz $p \times q$, entonces $vec(T)$ representará el vector $pq \times 1$ formado por el apilamiento de las columnas de T, una bajo la otra. Es decir, si

$$T = [t_1 \; t_2 \cdots t_q],$$

donde t_i es un vector $p \times 1$, para $i = 1, \cdots q$, entonces

$$vec(T) = \begin{pmatrix} t_1 \\ t_2 \\ . \\ . \\ . \\ t_q \end{pmatrix}.$$

Ahora, enunciemos el siguiente resultado.

Teorema 2.4.3 *Sean $X_1, X_2, \ldots$ una sucesión de vectores aleatorios $m \times 1$ i.i.d. con cuartos momentos finitos, media μ y matriz de covarianza Σ y sea*

$$A(N) = \sum_{i=1}^{N}(X_i - \bar{X}_N)(X_i - \bar{X}_N)'$$

donde $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$. *Entonces, la distribución asintótica de*

$$T(N) = \frac{1}{\sqrt{N}}\left[A(N) - N\Sigma\right]$$

es normal con media 0 *y matriz de covarianza*

$$V = Cov\left[vec((X_i - \mu)(X_i - \mu)')\right].$$

El siguiente corolario expresa este resultado asintótico en términos de la matriz de covarianza muestral.

Corolario 2.4.2 *Sea* $n = N - 1$ *y definamos* $S^2(n) = \frac{1}{n}A(N)$. *Bajo las condiciones del Teorema 2.4.3 la distribución asintótica de* $U(n) = \sqrt{n}[S^2(n) - \Sigma]$ *es normal con media* 0 *y matriz de covarianza* V.

2.5. Representación geométrica Gaussiana

La representación geométrica de datos Gaussianos de dimensión alta ha sido útil para comprender el comportamiento de metodologías estadísticas para estos datos. Esta característica de los datos Gausianos de dimensión alta es planteado en [11], Sección 2 y 3 y se presenta a continuación. Sea $Z(d) = (Z^{(1)}, ..., Z^{(d)})$ un vector aleatorio de dimensión d de distribución normal con media cero y matriz de covarianza la identidad. Ya que el cuadrado de cada entrada tiene distribución $\mathcal{X}^2$ con 1 grado de libertad y son independientes, tenemos que la suma de estas tienen distribución $\mathcal{X}^2$ con d grados de libertad. Entonces, por el Teorema del Límite Central (Teorema 1.5.20)

$$\frac{d^{1/2}\left[\frac{\sum_{i=1}^{d}(Z^{(i)})^2}{d} - 1\right]}{\sqrt{2}} \to N(0,1)$$

en distribución cuando $d \to \infty$. Luego

$$d^{1/2}\left[\frac{\sum_{i=1}^{d}(Z^{(i)})^2}{d} - 1\right] \to N(0,2)$$

en distribución cuando $d \to \infty$. Aplicando el Método Delta (Teorema 1.5.23) a $Y_d = \frac{\sum_{k=1}^{d}(Z^{(k)})^2}{d} = \frac{|Z(d)|^2}{d}$ y tomando $g(x) = x^{1/2}$, tenemos que

$$d^{1/2}\left[\frac{\left(\sum_{i=1}^{d} Z^{(i)^2}\right)^{1/2}}{d^{1/2}} - \sqrt{1}\right] = |Z(d)| - d^{1/2} \to N\left(0, \frac{1}{2}\right)$$

en distribución cuando $d \to \infty$. Así, por el Teorema 1.5.22, $|Z(d)| - d^{1/2}$ es una $O_p(1)$. Por lo tanto se tiene que la distancia Euclidiana de $Z(d)$ tiene la propiedad

$$|Z(d)| = \left(\sum_{k=1}^{d} Z^{(k)^2} \right)^{1/2} = d^{1/2} + O_p(1), \quad \text{cuando } d \to \infty.$$

Esto prueba en cierto sentido, que los datos están cerca de la superficie de una esfera expandible. El resultado se extiende para el caso de dos vectores independientes de la normal estándar d-variada, digamos $Z_1(d)$ y $Z_2(d)$. Puesto que $Z_1^{(i)} - Z_2^{(i)} \sim N(0,2)$, entonces

$$\frac{d^{1/2}\left[\frac{\sum_{i=1}^{d}(Z_1^{(i)} - Z_2^{(i)})^2}{2d} - 1\right]}{\sqrt{2}} \to N(0,1)$$

en distribución cuando $d \to \infty$. Luego

$$d^{1/2}\left[\frac{\sum_{i=1}^{d}(Z_1^{(i)} - Z_2^{(i)})^2}{d} - 2\right] \to N(0,8)$$

en distribución cuando $d \to \infty$. Ahora, aplicando el Método Delta a $Y_d = \frac{\sum_{i=1}^{d}(Z_1^{(i)} - Z_2^{(i)})^2}{d} = \frac{|Z_1 - Z_2|^2}{d}$ con $g(x) = x^{1/2}$ y $\theta = 2$, obtenemos que

$$|Z_1(d) - Z_2(d)| - (2d)^{1/2} \to N(0,1)$$

en distribución. Por lo que $|Z_1(d) - Z_2(d)| - (2d)^{1/2}$ es una $O_p(1)$ y así

$$|Z_1(d) - Z_2(d)| = (2d)^{1/2} + O_p(1), \quad \text{cuando } d \to \infty. \tag{2.5}$$

De esta misma forma al considerar el ángulo, en el origen, entre los vectores Z_1 y Z_2 independientes de la normal estándar d-variada, tenemos por el Teorema del Límite Central que

$$d^{1/2}\left(\frac{\sum_{i=1}^{d} Z_1^{(i)} Z_2^{(i)}}{d}\right) \to N(0,1)$$

en distribución cuando $d \to \infty$. Aplicando el Método Delta a $Y_d = \frac{\sum_{i=1}^{d} Z_1^{(i)} Z_2^{(i)}}{d}$, con $\theta = 0$ y $g(x) = \cos^{-1}(x)$, obtenemos que $d^{1/2}\left(\cos^{-1}(Z_1 \cdot Z_2) - \cos^{-1}(0)\right) \to N(0,1)$ en distribución. Es decir $d^{1/2}\left(ang(Z_1, Z_2) - \frac{\pi}{2}\right) \to N(0,1)$ en distribución cuando $d \to \infty$. Por lo que

$$ang(Z_1, Z_2) = \frac{\pi}{2} + O_p(d^{-1/2}), \quad \text{cuando } d \to \infty, \tag{2.6}$$

donde $ang(Z_1, Z_2)$ denota el ángulo, en radianes, entre los vectores Z_1 y Z_2.

De hecho, (2.5) y (2.6) se tienen para una muestra aleatoria $Z_1, ..., Z_n$, deduciendo que todas las distancias entre los pares de puntos en la muestra son aproximadamente iguales y que todos los pares de ángulos son aproximadamente perpendiculares.

En la Figura 2.1, se ilustra un ejemplo, considerando el caso $d = 3$ y $n = 3$. Todas las distancias del origen a los puntos respectivos son aproximadamente iguales (de longitud $d^{1/2} = 3^{1/2}$), cuyos rayos del origen a los puntos forman ángulos casi ortogonales. Si nos enfocamos en el hiperplano (de dimensión 2 en este caso) generado por los datos, al ser todas las distancias de cada par de puntos casi iguales, los datos viven esencialmente en los vértices de un triángulo equilátero (que llamaremos el 3-simplex). En general el n-simplex es un poliedro de n vértices con todos sus bordes de igual longitud, por ejemplo, para $n = 3$ tenemos un triángulo equilátero como el que se muestra en la Figura 2.1.

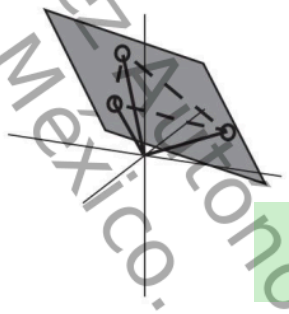

Figura 2.1: Representación geométrica en el caso n=3 y d=3, mostrando el 3-simplex en el hiperplano generado por los datos.

2.6. Representación geométrica general

La representación geométrica de los datos multivariados Gaussianos puede observarse también en escenarios más generales. Consideremos un vector $X(d) = (X^{(1)}, ..., X^{(d)})'$ que se obtiene al truncar una serie de tiempo infinita $X = (X^{(1)}, X^{(2)}, ...)'$. La estructura aproximada de n-simplex como la de los datos multivariados Gaussianos, se seguirá del comportamiento límite entre las distancias de los pares de puntos en una muestra $\mathcal{X}(d) = \{X_1(d), ..., X_n(d)\}$, donde los vectores $X_i(d)$ son tomados i.i.d. con la misma distribución de $X(d)$. Suponemos lo siguiente:

(a) Los cuartos momentos de las entradas de los vectores están uniformemente acotados.

(b) Para una constante $\sigma^2 > 0$,

$$\frac{1}{d}\sum_{k=1}^{d} Var(X^{(k)}) \to \sigma^2, \quad \text{cuando } d \to \infty. \tag{2.7}$$

(c) La serie de tiempo X satisface la **condición ρ mixing para funciones dominadas por cuadráticas**, es decir, para funciones de dos variables f y g que satisfacen $|f(u,v)| + |g(u,v)| \leq Cu^2v^2$ para $C > 0$ fija y para todo u y v, se tiene que

$$\sup_{1\leq k,l<\infty, |k-l|\geq r} |corr\left\{f(U^{(k)}, V^{(k)}), g(U^{(l)}, V^{(l)})\right\}| \leq \rho(r),$$

con $(U,V) = (X,X)$ ó $(U,V) = (X,\tilde{X})$, donde $\tilde{X}$ tiene la misma distribución que X y es independiente de ella, y la función ρ satisface $\rho(r) \to 0$ cuando $r \to \infty$.

Entonces, se sigue (ver A.3.1) que la distancia entre $X_i(d)$ y $X_j(d)$, para cualquier $i \neq j$, es aproximadamente igual a $(2\sigma^2 d)^{1/2}$ cuando $d \to \infty$, en el sentido que

$$\frac{1}{d^{1/2}}\left\{\sum_{k=1}^{d}(X_i^{(k)} - X_j^{(k)})^2\right\}^{1/2} \to (2\sigma^2)^{1/2} \tag{2.8}$$

en probabilidad cuando $d \to \infty$. Concluimos a partir de lo obtenido en (2.8) que después de reescalar por $d^{-1/2}$, los puntos $X_i(d)$ están asintóticamente localizados en los vértices de un $n-$simplex donde cada borde tiene longitud $(2\sigma^2)^{1/2}$.

En los resultados obtenidos, las condiciones sobre las variables son muy restrictivas, ya que la condición ρ-mixing pide algo "cercano a la independencia" de las variables, lo cual en la práctica no siempre se tiene, además de que esta condición depende del orden de las variables, y en la práctica este orden es comunmente aleatorio. El siguiente teorema (propuesto en [2] y [19]) implica esta misma representación geométrica en forma más general, debilitando específicamente la condición ρ-mixing. Para presentarlo, consideremos $\mathbf{X} = [X_1, \ldots, X_n]$ la matriz $d \times n$ con $d > n$, donde los $X_j = (X_j^{(1)}, \ldots, X_j^{(d)})'$, $j = 1, \ldots, n$, son i.i.d. de una distribución multivariada de dimensión d con media 0 y matriz de covarianza Σ. La descomposición en valores propios de Σ es $\Sigma = V\Lambda V'$, donde Λ es una matriz diagonal de valores propios $\lambda_1 \geqslant \cdots \geqslant \lambda_d \geqslant 0$ y V es la matriz de vectores propios correspondiente. Una **matriz factor**, que es esencialmente la raíz cuadrada de Σ es definida como $F = V\Lambda^{1/2}$, por lo que $\Sigma = FF'$. Si Σ es positiva definida (y por tanto sus valores propios son todos positivos, ver A.2), podemos entonces escribir $\mathbf{X} = F\mathbf{Z}$, donde $\mathbf{Z} = \Lambda^{-1/2}V'\mathbf{X}$ es una matriz aleatoria $d \times n$ de una distribución con matriz de covarianza la identidad. Note que si $\mathbf{X}$ es Gaussiana, los elementos de $\mathbf{Z}$ son variables normal estándar independientes.

Consideremos la matriz de covarianza muestral $S^2 = \frac{1}{n}\mathbf{X}\mathbf{X}'$. Definimos la **matriz de covarianza muestral dual** como $S_D^2 = \frac{1}{n}\mathbf{X}'\mathbf{X}$. Así, S_D^2 tiene los mismos valores propios, distintos de cero, que S^2. Luego, usando el hecho de que $V'V$ es la identidad, tenemos que

$$nS_D^2 = (\mathbf{Z}'F')(F\mathbf{Z}) = \mathbf{Z}'\Lambda\mathbf{Z} = \sum_{i=1}^{d} \lambda_i W_i, \tag{2.9}$$

con $W_i = Z_i' Z_i$, donde $Z_i = (Z_i^{(1)}, Z_i^{(2)}, \ldots, Z_i^{(n)})$, $i = 1, \ldots, d$, son los vectores fila de $\mathbf{Z}$. Notar que si $\mathbf{X}$ es Gaussiana, W_i tiene distribución Wishart $W_n(1, I_n)$.

Teorema 2.6.1 *Para n fijo, sean $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \ldots$ matrices aleatorias $d \times n$ de distribuciones multivariadas de dimensión d con media 0 y matriz de covarianza $\Sigma_1, \ldots \Sigma_d, \ldots$, respectivamente. Asumimos que los cuartos momentos de cada variable están uniformemente acotados, que la condición en (2.9) se tiene para cada $\boldsymbol{X}_i$ y que las entradas de $\boldsymbol{Z}$ son independientes. Sean $\lambda_{1,d} \geqslant \cdots \geqslant \lambda_{d,d}$ los valores propios de la matriz de covarianza Σ_d y sea $S_{D,d}^2$ la matriz de covarianza dual muestral correspondiente. Suponga que los valores propios de Σ_d están suficientemente difusos, en el sentido que*

$$\frac{\sum_{i=1}^{d} \lambda_{i,d}^2}{(\sum_{i=1}^{d} \lambda_{i,d})^2} \to 0, \quad \textit{cuando } d \to \infty. \tag{2.10}$$

Entonces los valores propios muestrales se comportan como si vinieran de la matriz de covarianza la identidad, en el sentido que $c_d^{-1} S_{D,d}^2 \to I_n$, cuando $d \to \infty$, donde $c_d = \frac{1}{n}\sum_{i=1}^{d} \lambda_{i,d}$.

Demostración. Debido al resultado en (2.9), si $Z_i = (Z_i^{(1)}, Z_i^{(2)}, \ldots, Z_i^{(n)})$ cualquier elemento de la diagonal de nS_D^2 puede ser expresado como $\sum_{i=1}^{d} \lambda_{i,d}(Z_i^{(k)})^2$, donde los $Z_i^{(k)}$, $i = 1, 2, \ldots, d$, son independientes y tienen media cero y varianza uno. Definimos los **valores propios relativos** como $\tilde{\lambda}_{i,d} = \lambda_{i,d} / \sum_{i=1}^{d} \lambda_{i,d}$. Así, la condición en (2.10) es equivalente a $\sum_{i=1}^{d} \tilde{\lambda}_{i,d}^2 \to 0$ cuando $d \to \infty$. Observemos que $E\left(\sum_{i=1}^{d} \tilde{\lambda}_{i,d}(Z_i^{(k)})^2\right) = \sum_{i=1}^{d} \tilde{\lambda}_{i,d} E\left[(Z_i^{(k)})^2\right] = \sum_{i=1}^{d} \tilde{\lambda}_{i,d} Var(Z_i^{(k)}) = \sum_{i=1}^{d} \tilde{\lambda}_{i,d} = 1$, por lo que $Var\left(\sum_{i=1}^{d} \tilde{\lambda}_{i,d}(Z_i^{(k)})^2\right) = E\left([\sum_{i=1}^{d} \tilde{\lambda}_{i,d}(Z_i^{(k)})^2 - 1]^2\right)$. Luego, por la desigualdad de Chebyshev, para cualquier $\tau > 0$ y la cota uniforme M de los cuartos momentos, $P\left(|\sum_{i=1}^{d} \tilde{\lambda}_{i,d}(Z_i^{(k)})^2 - 1| > \tau\right) \leqslant \tau^{-2} E\left(|\sum_{i=1}^{d} \tilde{\lambda}_{i,d}(Z_i^{(k)})^2 - 1|^2\right)$ $= \tau^{-2} Var\left(\sum_{i=1}^{d} \tilde{\lambda}_{i,d}(Z_i^{(k)})^2\right) \leqslant \tau^{-2} M \sum_{i=1}^{d} \tilde{\lambda}_{i,d}^2 \to 0$ cuando $d \to \infty$. Por lo que los elementos de la diagonal de $c_d^{-1} S_{D,d}^2$ convergen a 1 en probabilidad cuando $d \to \infty$. Los elementos fuera de la diagonal de nS_D^2 pueden ser expresados como $\sum_{i=1}^{d} \lambda_{i,d} Z_i^{(j)} Z_i^{(k)}$, donde $Z_i^{(j)}$ y $Z_i^{(k)}$ son independientes. Entonces, similarmente, $P\left(|\sum_{i=1}^{d} \tilde{\lambda}_{i,d} Z_i^{(j)} Z_i^{(k)}| > \tau\right)$

$\leqslant \tau^{-2} Var\left(\sum_{i=1}^{d} \tilde{\lambda}_{i,d} Z_i^{(j)} Z_i^{(k)}\right) = \tau^{-2} \sum_{i=1}^{d} \tilde{\lambda}_{i,d}^2 \to 0$ cuando $d \to \infty$. Así los elementos fuera de la diagonal de $c_d^{-1} S_{D,d}^2$ convergen a 0 en probabilidad cuando $d \to \infty$. ■

Ahora, usando el Teorema 2.6.1, estableceremos la representación geométrica. Sea $X_j = (X_j^{(1)}, \ldots, X_j^{(d)})'$, $j = 1, \ldots, n$, la j-ésima columna de la matriz $\mathbf{X}$, cuya distribución satisface las condiciones del Teorema 2.6.1. La distancia al cuadrado entre X_k y X_l es

$$|X_k - X_l|^2 = \sum_{i=1}^{d} (X_k^{(i)})^2 + \sum_{i=1}^{d} (X_l^{(i)})^2 - 2 \sum_{i=1}^{d} X_k^{(i)} X_l^{(i)}. \tag{2.11}$$

Entonces, dado que los dos primeros términos del lado derecho en (2.11) son las entradas k-ésima y l-ésima de la diagonal de nS_D^2, por el Teorema 2.6.1, para d suficienteme grande, los primeros dos términos son similares a $\sum_{i=1}^{d} \lambda_{i,d}$.

También, ya que el tercer término es la kl-ésima entrada de nS_D^2, éste es cercano a cero cuando d es grande. Así, para d suficientemente grande, (2.11) es aproximadamente $2\sum_{i=1}^{d} \lambda_{i,d}$, en el sentido de que $\frac{|X_k - X_l|^2}{2\sum_{i=1}^{d} \lambda_{i,d}} \to 1$ en probabilidad cuando $d \to \infty$. Así que las distancias entre los pares de puntos de los n vectores son aproximadamente las mismas y los datos forman un n-simplex en $\mathbb{R}^d$. Por otro lado, en [2] se verifica que la condición (2.10) es más general que la condición ρ-mixing.

Notemos que si Σ_d es positiva definida, entonces la condición (2.9) se satisface ya que $\mathbf{Z}$ está bien definida, debido a que $\lambda_1 \geqslant \cdots \geqslant \lambda_d > 0$. Por lo tanto, otras condiciones para que datos multivariados con media 0 tengan la representación geométrica son:

(i) Los cuartos momentos de cada variable están uniformemente acotados.

(ii) Σ_d es positiva definida.

(iii) La condición (2.10).

(iv) Las entradas de $\mathbf{Z}$ son independientes

Estas son las condiciones que serán verificadas para los datos utilizados en las simulaciones del Capítulo 4 de esta tesis.

Capítulo 3

Métodos de Discriminación Binaria

En la práctica existen muchos problemas en los que tenemos datos multivariados de dimensión alta (datos multivariados en los cuales la dimensión es mayor que el tamaño de la muestra) y es necesario llevar a cabo una clasificación binaria de estos. En este capítulo describiremos cuatro métodos de discriminación binaria que son utilizados en este tipo de datos.

3.1. Discriminación binaria

En esta sección daremos la notación necesaria y el tipo de datos que se utilizan en la implementación de los métodos de discriminación binaria.

Supongamos que tenemos el siguiente conjunto de datos de entrenamiento

$$(x_1, w_1), (x_2, w_2), ..., (x_N, w_N) \tag{3.1}$$

donde $x_i \in \mathbb{R}^d$ y $w_i \in \{-1, 1\}$, para $i = 1, 2, ..., N$. Tenemos dos clases de datos, C_+ y C_-, correspondientes a los vectores x_i con w_i igual a 1 y -1, respectivamente.

Sea $\mathbf{x} = [x_1, x_2, ..., x_N]$ la matriz de $d \times N$ cuyas columnas corresponden a los datos vectoriales y $w = (w_1, w_2, ..., w_N)'$ el vector con las etiquetas de los x_i's. Se usará la siguiente notación

- W es la matriz diagonal de $N \times N$ con los elementos de w en su diagonal.
- $\mathbf{x}_+$ y $\mathbf{x}_-$ son las submatrices de $\mathbf{x}$ correspondientes a la clase C_+ y C_-, respectivamente.
- m y n son las cardinalidades de las clases C_+ y C_-, respectivamente.
- 1_k es el vector k-dimensional de unos.

Un conjunto de datos como (3.1) es **linealmente separable** si existe un hiperplano tal que todos los datos de la clase C_+ están de un lado y todos los de la clase C_- están del otro lado. Un hiperplano que cumple esto es llamado un **hiperplano separante** del conjunto de datos. En el contexto de datos con dimensión mayor al tamaño de la muestra, se ha visto que cuando los datos tienen función de densidad continua son linealmente separables casi seguramente. En este trabajo nos restringiremos únicamente al caso linealmente separable.

A continuación describimos de forma concisa los métodos de discriminación binaria para datos de dimensión alta que emplearemos en este trabajo.

3.2. Support Vector Machine

El método Support Vector Machine (SVM) es uno de los más populares de los métodos de discriminación binaria (para un estudio detallado ver por ejemplo [7], [9], [10], [12], [22] y [23]). En el caso linealmente separable, este método consiste en encontrar un hiperplano separante que maximice las distancias entre el hiperplano y los vectores más cercanos de cada clase. Los puntos más cercanos al hiperplano separante se llaman **vectores soporte**. La descripción del método la presentamos a continuación.

Supongamos que existen un vector v y un escalar b tales que cumplen las siguientes desigualdades

$$\begin{aligned} v^{'}x_i + b \geq 1 \quad &\text{si } w_i = 1 \\ v^{'}x_i + b \leq -1 \quad &\text{si } w_i = -1. \end{aligned} \tag{3.2}$$

En este caso el hiperplano $v^{'}x_i + b = 0$ es un hiperplano separante para los datos.

Notar que las desigualdades de (3.2) se pueden escribir como

$$w_i(v^{'}x_i + b) \geq 1, \quad i = 1, 2, ..., N. \tag{3.3}$$

Los vectores que satisfacen la igualdad en (3.3) son los vectores soporte. Es decir, los vectores soporte son los que pertenecen a uno de los hiperplanos

$$v^{'}x + b = -1 \quad \text{ó} \quad v^{'}x + b = 1. \tag{3.4}$$

Al conjunto de vectores soporte lo denotaremos por SV.

Sean d_+ y d_- las distancias más cortas del hiperplano $v^{'}x + b = 0$ a los vectores en C_+ y C_-, respectivamente. El **margen** del hiperplano separante se define como $d_0 = d_+ + d_-$. Es sabido que si $v^{'}x + b = 0$ es un hiperplano y $x_0 = (x_1, ..., x_d)$ es un punto, la distancia entre ellos está dada por

$$d = \frac{|v^{'}x_0 + b|}{|v|}.$$

Sea x_+ un punto soporte sobre la recta $v^{'}x + b = 1$. Por lo que

$$d_+ = \frac{|v^{'}x_+ + b|}{|v|} = \frac{1}{|v|}.$$

Análogamente

$$d_- = \frac{1}{|v|}.$$

Por lo tanto

$$d_0 = d_+ + d_- = \frac{2}{|v|}.$$

En el caso linealmente separable, el hiperplano separante $v_0^{'}x + b_0 = 0$ del método SVM es el único hiperplano separante con máximo margen. Así, el hiperplano separante SVM maximiza $\frac{2}{|v|}$ sujeto a la condición (3.3).

Equivalentemente este hiperplano resuelve el problema de optimización

$$\begin{aligned} &\text{minimizar} \quad \frac{|v|^2}{2} \\ &\text{sujeto a} \quad w_i(v^{'}x_i + b) \geq 1, \quad i = 1, 2, ..., N. \end{aligned}$$

Se concluye que el vector óptimo es dado por

$$v_0 = \mathbf{x}W\widehat{\alpha} \tag{3.5}$$

donde $\widehat{\alpha}$ resuelve el problema de optimización

$$\begin{aligned} &\text{maximizar } 1_N^{'}\alpha - \frac{1}{2}|\mathbf{x}W\alpha|, \\ &\text{sujeto a } \alpha \geq 0,\ \alpha^{'}w = 0. \end{aligned} \tag{3.6}$$

Además $\widehat{\alpha_i} \neq 0$ sólo si x_i es un vector soporte, por lo tanto, por (3.5), v_0 es una función lineal de los vectores soporte únicamente. Ya que los vectores soporte satisfacen la igualdad de (3.4), la ordenda al origen b_0 puede ser calculada como

$$b_0 = w_i - v_0^{'}x_i,$$

para cualquier $x_i \in \text{SV}$.

El problema de optimización (3.6) es equivalente a

$$\text{maximizar } \frac{2}{|\mathbf{x}W\alpha^*|^2}$$

$$\text{sujeto a } 1_m^{'}\alpha_+^* = 1_n^{'}\alpha_-^* = 1, \quad \alpha_+^*, \quad \alpha_-^* \geqslant 0 \tag{3.7}$$

donde α_+^* y α_-^* son subvectores de α^* correspondientes a las clases C_+ y C_-, respectivamente. Note que $\mathbf{x}W\alpha^* = \mathbf{x}_+\alpha_+^* - \mathbf{x}_-\alpha_-^*$; así estamos minimizando la distancia entre los puntos de las envolventes convexas de las clases C_+ y C_-. Por lo que si $\widehat{\alpha^*}$ resuelve el problema (3.7), el vector normal al hiperplano separante de SVM puede ser tomado como

$$v_0^* = \mathbf{x}W\widehat{\alpha^*}$$

el cual es la diferencia entre el par de vectores más cercanos de las envolentes convexas de C_+ y C_- y es proporcional al vector v_0 que dimos en la ecuación (3.5).

3.3. Mean Difference

En el método Mean Difference (MD), el cual puede verse con más detalle en [20], el hiperplano separante que se toma es el que bisecta ortogonalmente al segmento que une a las medias de las dos clases, es decir, si las medias de las clases C_+ y C_- están dadas por

$$\overline{x}_+ = \frac{1}{m}\sum_{x_i \in C_+} x_i \quad y \quad \overline{x}_- = \frac{1}{n}\sum_{x_i \in C_-} x_i,$$

respectivamente, entonces el hiperplano MD tiene vector normal

$$u = \overline{x}_+ - \overline{x}_-. \tag{3.8}$$

Notar que u es la distancia entre dos puntos de las envolventes convexas de las dos clases.

3.4. Distance Weighted Discrimination

La idea del método Distance Weighted Discrimination (DWD) es encontrar un hiperplano separante que maximice la suma de los recíprocos de las distancias de los datos al hiperplano (ver [14] y [15]). Así, todos los datos son tomados en cuenta para encontrar el hiperplano; en este caso, los datos más cercanos al hiperplano hacen una contribución mayor que los más lejanos.

Definamos el **residual** del i-ésimo vector x_i como la distancia signada de x_i al hiperplano $v^{'}x + b = 0$ dada por

$$r_i = w_i(v^{'}x_i + b). \tag{3.9}$$

En la clase C_+ a los vectores x_i les corresponde $w_i = 1$, en la clase C_- a los vectores x_i les corresponde $w_i = -1$. El hiperplano separante DWD

$$v^{'}x + b = 0$$

resuelve el problema de optimización

$$\begin{aligned} &\text{minimizar } \textstyle\sum_{i=1}^{N} \dfrac{1}{r_i} \\ &\text{sujeto a } \ |v| = 1, r_i \geqslant 0, \ i = 1, 2, ..., N. \end{aligned}$$

Así, todos los vectores contribuyen en la búsqueda del hiperplano DWD, siendo los más cercanos al hiperplano separante los que influyen más. Puede verse que el vector óptimo v_1 está dado por

$$v_1 = \frac{\mathbf{x}W\widehat{\beta}}{|\mathbf{x}W\widehat{\beta}|},$$

donde $\widehat{\beta}$ resuelve el problema de optimización

$$\begin{aligned} &\text{maximizar } 2\ 1_N^T\sqrt{\beta} - |\mathbf{x}W\beta| \\ &\text{sujeto a } \beta \geqslant 0, \ \beta^{'}w = 0, \end{aligned} \tag{3.10}$$

donde $\sqrt{\beta}$ denota el vector cuyas entradas son las raíces cuadradas de las entradas de β. Notemos que el problema de optimización (3.10) es parecido al de (3.6) del método SVM. Se puede ver que los residuales están dados por

$$r_i = \frac{1}{\sqrt{\hat{\beta}_i}}, \quad i = 1, 2, ..., N.$$

Así, por la ecuación (3.9) tenemos que

$$b_1 = \frac{w_i}{\sqrt{\hat{\beta}_i}} - v_1^{'}x_i,$$

para cualquier vector x_i.

El problema de optimización (3.10) es equivalente a

$$\begin{aligned} &\text{maximizar } \frac{(1_m^{'}\sqrt{\beta_+^*} + 1_n^{'}\sqrt{\beta_-^*})^2}{|\mathbf{x}_+\beta_+^* - \mathbf{x}_-\beta_-^*|} \\ &\text{sujeto a } 1_m^{'}\beta_+^* = 1_n^{'}\beta_-^* = 1, \ \beta_+^*, \beta_-^* \geq 0. \end{aligned} \tag{3.11}$$

Notar que se está minimizando la distancia entre puntos de las envolventes convexas, pero divididos por el cuadrado de la suma de las raíces cuadradas de los pesos convexos. Por lo que si $\hat{\beta}^*$ resuelve el problema (3.11) el vector normal al hiperplano DWD es proporcional a

$$v_1^* = \mathbf{x}_+\hat{\beta}_+^* - \mathbf{x}_-\hat{\beta}_-^*,$$

que es la diferencia de dos puntos de las envolventes convexas de las dos clases, análogamente a como sucedió en MD y SVM.

3.5. Maximal Data Piling

El método Maximal Data Piling (MDP) está diseñado para datos de dimensión mayor al tamaño de la muestra (ver [1]). Se necesita suponer que $d \geqslant N-1$ y que el subespacio generado por los datos tiene dimensión N (ó $N-1$ si $d = N-1$). Bajo este supuesto existe una dirección (vector) sobre la cual la proyección de los datos cae en dos puntos únicamente, uno por cada clase. Todos los datos de una clase caen sobre un punto y todos los de la otra clase caen sobre el otro punto.

Para encontrar el hiperplano MDP, sea $u = \bar{x}_+ - \bar{x}_-$ el vector normal dado por (3.8). Definamos $C = \left[\mathbf{x}_C^+, \mathbf{x}_C^-\right]$, donde $\mathbf{x}_C^+$ y $\mathbf{x}_C^-$ son las versiones centradas de $\mathbf{x}_+$ y $\mathbf{x}_-$ respectivamente, es decir

$$\mathbf{x}_C^+ = \mathbf{x}_+ - \bar{x}_+ 1_d^{'}, \quad \mathbf{x}_C^- = \mathbf{x}_- - \bar{x}_- 1_d^{'}.$$

La matriz proyección simétrica en el complemento ortogonal del espacio columna C, está dado por $Q = I_d - CC^{\dagger}$, donde $C^{\dagger}$ es la inversa generalizada de Moore-Penrose de C.

El hiperplano MDP

$$v_2^{'}x + b_2 = 0$$

tiene la propiedad de que su vector normal unitario v_2 es la dirección para el cual la proyección de las medias de clase tienen distancia máxima, sujeta a la condición de que la proyección de los vectores de las clases caen sobre el mismo punto que su media de clase. En otras palabras, tenemos que v_2 resuelve el problema de optimización

$$\begin{gathered}\text{maximizar } (v^{'}u)^2 \\ \text{sujeto a } C^{'}v = 0,\ v^{'}v = 1\end{gathered}$$

Es posible demostrar (consultar en [1]) que la solución de este problema de optimización es

$$v_2 = \frac{Qu}{|Qu|}. \tag{3.12}$$

Esto significa que v_2 es ortogonal al subespacio de dimensión $N-2$ generado por las columnas de C. Además la fórmula (3.12) es equivalente a

$$v_2 = \frac{(\mathbf{x}_C \mathbf{x}_C^{'})^{\dagger} u}{|(\mathbf{x}_C \mathbf{x}_C^{'})^{\dagger} u|},$$

donde $\mathbf{x}_C$ es la versión centrada de la matriz $\mathbf{x}$, es decir, $\mathbf{x}_C = \mathbf{x} - \bar{x}1_d^{'}$. Por lo tanto v_2 petenece al subespacio de dimensión $N-1$ generado por todos los vectores centrados. Finalmente el intercepto b_2 puede ser calculado como

$$b_2 = \frac{-v_2^{'}(m\bar{x}_+ + n\bar{x}_-)}{N}.$$

Algunas veces se toma $b_2 = -v_2^{'}\left(\frac{\bar{x}_+ + \bar{x}_-}{2}\right)$.

3.6. Comportamiento asintótico de los métodos en el caso Gaussiano

En el siguiente teorema (ver [5]) se consideran dos clases de datos, C_- y C_+, donde los datos de la clase C_- tienen distribución $N_d(0, \Sigma_d)$ y los datos de la clase C_+ tienen distribución $N_d(v_d, \Sigma_d)$, por lo que se toma la dirección de v_d como la óptima para el vector normal al hiperplano separante. El resultado que se obtiene es en base al ángulo que forma la dirección de la media v_d con el vector normal al hiperplano separante de los métodos de discriminación MD, SVM, DWD y MDP. Cuando $|v_d|$ es grande, la clasificación de los datos se puede obtener con facilidad, en cambio, si es muy pequeña, realizar la clasificación es complicado. En el capítulo anterior, vimos que bajo ciertas condiciones, los datos multivariados de dimensión alta están ubicados a una distancia $d^{1/2}$ de la media cuando d es grande. Así que podemos decir que tenemos una frontera para realizar satisfactoriamente la clasificación de los datos que es precisamente cuando $|v_d| \approx d^{1/2}$, tal como se confirma a continuación en el siguiente teorema.

Teorema 3.6.1 *Supongamos que los vectores aleatorios en C_+ y C_- son independientes y provienen de las distribuciones normales multivariadas $N_d(v_d, \Sigma_d)$ y $N_d(0, \Sigma_d)$ respectivamente, donde $\Sigma_d = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$ y $\{\sigma_k\}_{k=1}^{\infty}$ es una sucesión acotada de números positivos tales que $\sum_{k=1}^{d} \frac{\sigma_k^2}{d} \to \sigma^2$, cuando $d \to \infty$, para algún $\sigma > 0$. Suponemos que $|v_d| d^{-1/2} \to c$, cuando $d \to \infty$, con $0 \leqslant c \leqslant \infty$. Si v representa el vector normal del hiperplano MD, SVM, DWD o MDP del conjunto de datos, entonces*

$$ang(v, v_d) \to \begin{cases} 0, & si\ c = \infty, \\ \frac{\pi}{2}, & si\ c = 0, \\ \arccos\left(\dfrac{c}{(\gamma\sigma^2 + c^2)^{1/2}}\right), & si\ 0 < c < \infty; \end{cases}$$

en probabilidad cuando $d \to \infty$, *donde* $\gamma = \frac{1}{m} + \frac{1}{n}$.

Observamos, según el resultado del Teorema 3.6.1, que cuando $|v_d| \gg d^{1/2}$ (que corresponde al caso cuando $c = \infty$) los ángulos entre los vectores normales a los hiperplanos de los cuatro métodos y el vector óptimo v_d convergen a cero y en este sentido se dice que los métodos son **consistentes**; sin embargo, cuando $|v_d| \ll d^{1/2}$ (que corresponde al caso cuando $c = 0$) los vectores normales no convergen a la dirección óptima, de hecho los ángulos son asintóticamente ortogonales, por lo que en este caso se dice que los métodos son **fuertemente inconsistentes**.

En [5] se muestra que los datos del Teorema 3.6.1 satisfacen la representación geométrica descrita en el capítulo anterior. En el siguiente capítulo se presentan los resultados de las simulaciones realizadas con el propósito de explorar una generalización de este teorema, considerando datos multivariados más generales que cumplan con esta representación geométrica.

Capítulo 4

Estudio por Simulación para Comparar los Métodos de Discriminación Binaria

En este capítulo presentamos los resultados que se obtuvieron en las simulaciones para comparar los métodos de discriminación binaria MD, SVM, DWD y MDP descritos en el Capítulo 3. Con estas simulaciones se pretende realizar una exploración de una generalización del Teorema 3.6.1 considerando datos multivariados más generales que los de ese teorema.

El estudio se realizó con tres tipos de datos que se especifican a continuación y que en cada caso deben cumplir con la representación geométrica que fue descrita en el Capítulo 2.

- Caso 1. Datos normales multivariados con matriz de covarianza Σ_d no diagonal.
- Caso 2. Datos multivariados no normales con matriz de covarianza Σ_d diagonal.
- Caso 3. Datos multivariados no normales con matriz de covarianza Σ_d no diagonal.

Se consideraron dos clases de datos, C_- y C_+, donde los datos de la clase C_- tienen distribución con la representación geométrica con media cero y los datos de la clase C_+ con la misma distribución que los de C_- pero con media v_d. Debido a lo anterior, v_d es la dirección óptima del vector normal al hiperplano separante. Denotamos por $\Sigma_d = (\sigma_{ij})$ a la matriz de covarianza común de las dos clases. Las simulaciones para comparar los métodos se realizaron en base al ángulo entre la dirección óptima v_d y el vector normal al hiperplano separante. Los conjuntos de datos de entrenamiento C_- y C_+ se generaron cada uno con tamaño 20. Para cada caso se consideraron vectores media v_d de la forma $v_d = (d^\delta, 0, \cdots, 0)'$ (con $\delta > 0$) y $v_d = \beta 1_d$ (donde $\beta = \beta_d \to c$, cuando $d \to \infty$, con $0 < c < \infty$). Se tomó $\delta = 0.2, 0.5$ y 0.8 que corresponden a los casos $|v_d| d^{-1/2} \to 0, 1$ e ∞, respectivamente, y se tomó $\beta = 0.5$ y 10 que corresponden

a los casos $|u_d|d^{-1/2} \to 0.5$ y 10, respectivamente. Los valores que se tomaron para las dimensiones son $d = 10, 30, 100, 300, 1000, 2000$ y para cada valor de d se generaron 500 conjuntos de datos de entrenamiento. Para cada conjunto de datos de entrenamiento se calcularon los ángulos entre v_d y los vectores normales a los hiperplanos separantes de los cuatro métodos.

4.1. Caso normal con Σ_d no diagonal

Sean $Z_1(d), Z_2(d), \ldots, Z_n(d)$ datos multivariados normales independientes con media cero y matriz de covarianza $\Sigma_d = toeplitz(1, s, s^2, \ldots, s^{d-1})$ con $0 < s < 1$ (ver A.1.1); para las simulaciones de este trabajo tomamos $s = 0.5$. Demostraremos a continuación que estos datos cumplen con la representación geométrica. Para ello verificaremos las condiciones $(i) - (iv)$ de la Sección 2.6 que implican esta representación geométrica.

(i) Verifiquemos primeramente que los cuartos momentos de cada variable están uniformemente acotados. Dado que $Z_i^{(k)} \sim N(0,1)$, entonces $(Z_i^{(k)})^2 \sim \mathcal{X}^2_{(1)}$. Así, $E\left[(Z_i^{(k)})^2\right] = 1$ y $Var\left[(Z_i^{(k)})^2\right] = 2$. Por lo tanto

$$E\left[(Z_i^{(k)})^4\right] = Var\left[(Z_i^{(k)})^2\right] + \left(E\left[(Z_i^{(k)})^2\right]\right)^2 = 3.$$

Por lo que los cuartos momentos son uniformemente acotados, ya que se mantienen contantes al variar k y al variar d.

(ii) En A.3.2 se demuestra por inducción que todos los menores principales $|B_l|$, $l = 1, 2, \ldots, d$, son positivos, donde $B_l = (\sigma_{ij})$, $i, j = 1, 2, \ldots, l$, es la submatriz de Σ_d correpondiente a las primeras l filas y columnas, de esta forma se tiene por el Criterio A.2.1 que Σ_d es positiva definida. De hecho se demuestra que $|B_l| = (1 - s^2)^{l-1} > 0, \ \forall\ l = 1, 2, \ldots, d$.

(iii) Veamos que la condición (2.10) se cumple. Observemos que

$$\begin{aligned}\sum_{i,j=1}^{d} \sigma_{ij}^2 &= d + 2[s^2(d-1) + s^4(d-2) + \cdots + s^{2(d-2)}(d-(d-2)) \\ &+ s^{2(d-1)}(d-(d-1))] \\ &= d + 2\sum_{k=1}^{d-1} k s^{2(d-k)} = 2\sum_{k=1}^{d} k s^{2(d-k)} - d,\end{aligned}$$

luego

$$\frac{\sum_{i=1}^{d} \lambda_{i,d}^2}{(\sum_{i=1}^{d} \lambda_{i,d})^2} = \frac{\sum_{i,j=1}^{d} \sigma_{ij}^2}{(tr(\Sigma_d))^2} = \frac{2\sum_{k=1}^{d} k s^{2(d-k)}}{d^2} - \frac{1}{d} = \frac{2s^{2d}\sum_{k=1}^{d} k s^{-2k}}{d^2} - \frac{1}{d}.$$

Tenemos que (ver A.3.3)

$$\sum_{k=1}^{d} ks^{-2k} \leq \frac{s^{-2}}{\ln(s^{-2})}\left[ds^{-2d} + \left(1 - \frac{1}{\ln(s^{-2})}\right)\left(s^{-2d} - 1\right)\right]. \tag{4.1}$$

Por lo que

$$0 \leq \frac{s^{2d}\sum_{k=1}^{d} ks^{-2k}}{d^2} \leq \frac{s^{-2}}{\ln(s^{-2})}\left[\frac{1}{d} + \left(1 - \frac{1}{\ln(s^{-2})}\right)\left(\frac{1}{d^2} - \frac{s^{2d}}{d^2}\right)\right] \to 0,$$

cuando $d \to \infty$. Por lo tanto

$$\frac{\sum_{i=1}^{d}\lambda_{i,d}^2}{(\sum_{i=1}^{d}\lambda_{i,d})^2} = \frac{2s^{2d}\sum_{k=1}^{d} ks^{-2k}}{d^2} - \frac{1}{d} \to 0, \quad \text{cuando} \quad d \to \infty.$$

(iv) Esta condición se satisface debido a que en este caso se utilizan datos Gaussianos.

4.1.1. Vector media $v_d = (d^\delta, 0, \cdots 0)'$ con $\delta > 0$

Para el primer conjunto de simulaciones consideramos el vector media $v_d = (d^\delta, 0, \cdots 0)'$, donde δ toma los valores de 0.2, 0.5 y 0.8 que corresponden a los casos $|v_d|d^{-1/2} \to 0$, $|v_d|d^{-1/2} \to 1$ y $|v_d|d^{-1/2} \to \infty$ cuando $d \to \infty$, respectivamente.

En las figuras 4.1, 4.2 y 4.3 se muestran las medias de los ángulos formados por v_d y el vector normal de los cuatro métodos considerados variando la dimensión d, correspondientes a los valores de $\delta = 0.2, 0.5$ y 0.8, respectivamente. En la Figura 4.1 observamos que cuando la dimensión crece las medias de los ángulos de todos los métodos tiende a un valor cercano a $\pi/2 = 1.5707$. Si se cumpliera el Teorema 3.6.1 para datos con representación geométrica, para $\delta = 0.5$ el límite de los ángulos estaría dado por $\arccos(c/(\gamma^2\sigma^2 + c^2)^{1/2}) = 0.3062$ cuando la dimensión tiende a infinito, donde $\gamma = 1/m + 1/n = 1/10$ y $\sigma^2 = \lim_{d\to\infty}\frac{\sum_{k=1}^{d}\sigma_{kk}}{d} = \lim_{d\to\infty}\frac{tr(\Sigma_d)}{d} = 1$ y $c = 1$, lo cual está de acuerdo con lo observado en la Figura 4.2. En la Figura 4.3 la media de los ángulos de todos los métodos tiende a cero. Por lo que podemos decir que para este caso los resultados del Teorema 3.6.1 siguen valiéndose. Observemos que para este caso, MD y DWD tienen el mejor compotamiento entre los cuatro métodos, ya que los ángulos son los más pequeños al incrementar la dimensión de los datos.

4.1.2. Vector media $v_d = \beta 1_d$ con $\beta > 0$

Tomamos también el vector media $v_d = \beta 1_d$ donde $\beta = 0.5$ y 10, los cuales corresponden a los casos $|v_d|d^{-1/2} \to 0.5$ y $|v_d|d^{-1/2} \to 10$ cuando $d \to \infty$, respectivamente. En las figuras 4.4 y 4.5 se muestran las medias de los ángulos formados por v_d y los vectores normales de los cuatro métodos considerados variando la dimensión d. Si se cumpliera

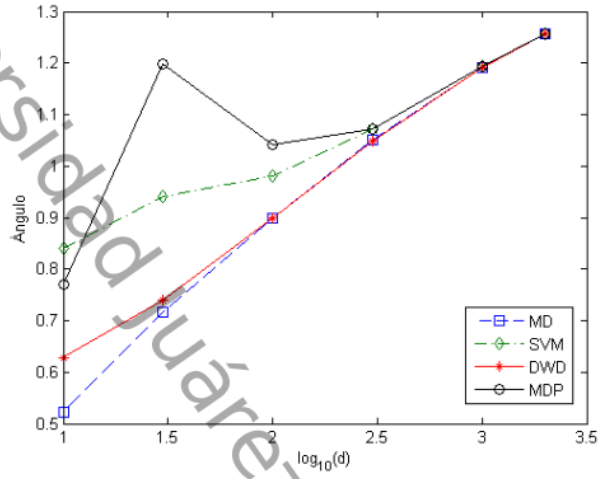

Figura 4.1: Medias de los ángulos entre $v_d = (d^\delta, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.2$. Caso normal con Σ_d no diagonal.

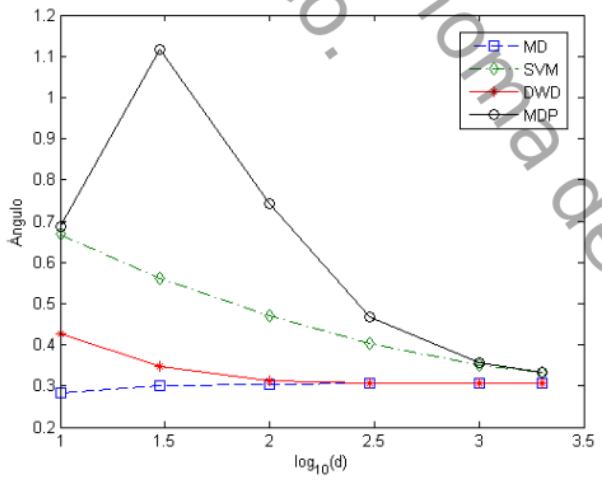

Figura 4.2: Medias de los ángulos entre $v_d = (d^\delta, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.5$. Caso normal con Σ_d no diagonal.

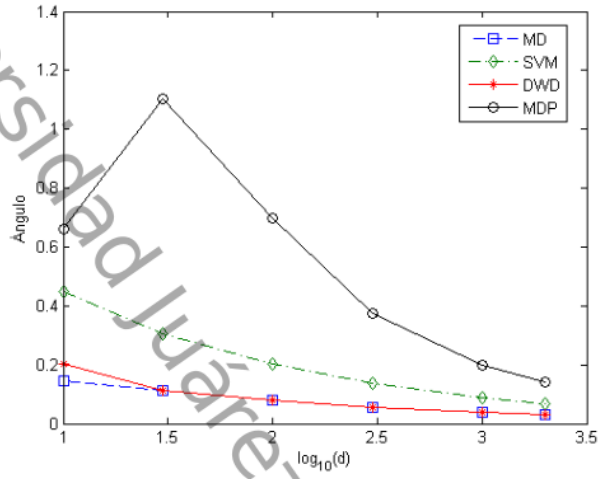

Figura 4.3: Medias de los ángulos entre $v_d = (d^{\delta}, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.8$. Caso normal con Σ_d no diagonal.

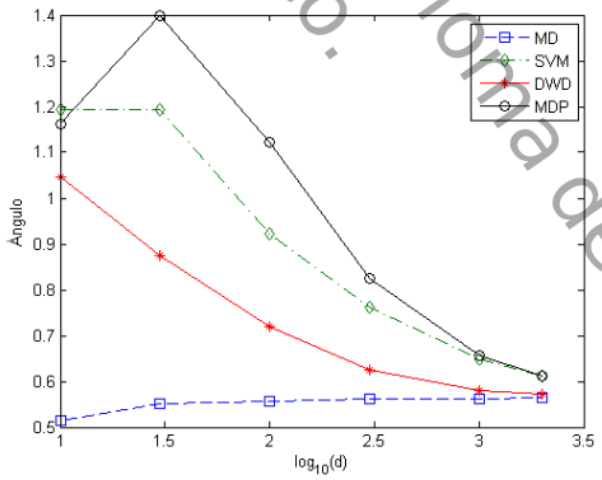

Figura 4.4: Medias de los ángulos entre $v_d = \beta 1_d$ y el vector normal del hiperplano separante; $\beta = 0.5$. Caso normal con Σ_d no diagonal.

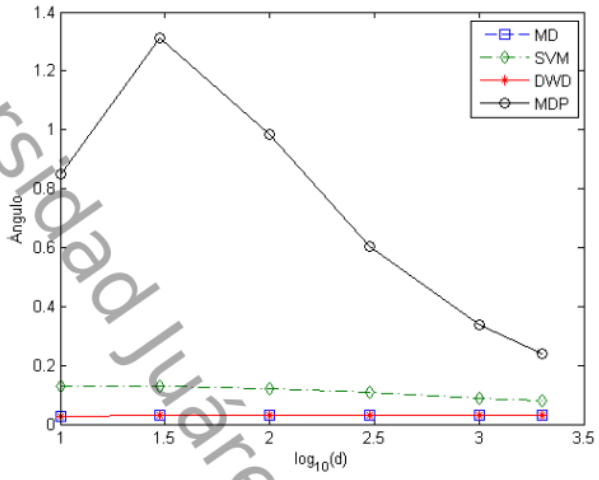

Figura 4.5: Medias de los ángulos entre $v_d = \beta 1_d$ y el vector normal del hiperplano separante; $\beta = 10$. Caso normal con Σ_d no diagonal.

el Teorema 3.6.1 para datos con representación geométrica, los límites de los ángulos cuando $d \to \infty$ serían $\arccos\left(\frac{\beta}{(\gamma\sigma^2+\beta^2)^{1/2}}\right) = 0.5639$ y 0.0316 para $\beta = 0.5$ y 10, respectivamente. En las figuras se observa la tendencia de la medias de los ángulos a estos valores. Observemos que en ambos casos MD presenta el mejor comportamiento, el método DWD es el segundo mejor método y prácticamente coincide con MD cuando $\beta = 10$.

4.2. Caso no normal con Σ_d diagonal

Sea $Z(d) = (Z^{(1)}, \cdots, Z^{(d)})'$ un vector aleatorio donde las $Z^{(i)}$'s son independientes con distribución $(\mathcal{X}^2_{(1)} - 1)/\sqrt{2}$. Tomemos $Z_1(d), Z_2(d), \ldots, Z_n(d)$ datos multivariados i.i.d. con la misma distribución que la de $Z(d)$, la cual tiene media cero y matriz de covarianza $\Sigma_d = I_d$. Verifiquemos a continuación que estos datos cumplen con la representación geométrica. Para ello verifiquemos, al igual que en el caso anterior, las condiciones $(i) - (iv)$ de la Sección 2.6 que implican esta representación geométrica.

(i) Dado que las entradas de cada vector son i.i.d. y que su distribución no depende de d, tenemos que los cuartos momentos de las entradas son uniformemente acotados.

(ii) Tenemos también que $\Sigma_d = I_d$ es positiva definida.

(iii) Ahora, dado que $\lambda_{i,d} = 1$, $i = 1, \cdots, d$, tenemos que

$$\frac{\sum_{i=1}^{d} \lambda_{i,d}^2}{(\sum_{i=1}^{d} \lambda_{i,d})^2} = \frac{d}{d^2} = \frac{1}{d} \to 0, \quad \text{cuando } d \to \infty.$$

(iv) Por último, la independencia de las entradas de $\mathbf{Z}$ se satisface debido a que $\Sigma_d = I_d$, por lo que $\mathbf{Z} = [Z_1(d), Z_2(d), \ldots, Z_n(d)]$, donde las $Z_i(d)$'s son independientes y tienen entradas independientes.

4.2.1. Vector media $v_d = (d^\delta, 0, \cdots 0)'$ con $\delta > 0$

En este caso, también tomamos para δ los valores de 0.2, 0.5 y 0.8, considerando $v_d = (d^\delta, 0, \cdots 0)'$ como el vector media para el primer conjunto de simulaciones.

En las figuras 4.6, 4.7 y 4.8 se muestran las medias de los ángulos formados por v_d y el vector normal de los cuatro métodos considerados variando la dimensión d, correspondientes a los valores de $\delta = 0.2, 0.5$ y 0.8, respectivamente. En la Figura 4.6 observamos que cuando la dimensión crece las medias de los ángulos de todos los métodos tiende a un valor cercano a $\pi/2 = 1.5707$. Si se cumpliera el Teorema 3.6.1 para datos con representación geométrica, para $\delta = 0.5$ el límite de los ángulos estaría dado por $\arccos(c/(\gamma^2\sigma^2 + c^2)^{1/2}) = 0.3062$ cuando la dimensión tiende a infinito, donde $\gamma = 1/m + 1/n = 1/10, \sigma^2 = \lim_{d\to\infty} \frac{\sum_{k=1}^{d} \sigma_{kk}}{d} = \lim_{d\to\infty} \frac{tr(\Sigma_d)}{d} = 1$ y $c = 1$, lo cual está de acuerdo con lo observado en la Figura 4.7. En la Figura 4.8 la media de los ángulos de todos los métodos tiende a cero. Podemos decir que el resultado del Teorema 3.6.1 se sigue cumpliendo. Observemos que, al igual que en el caso anterior, MD y DWD tienen el mejor compotamiento entre los cuatro métodos, ya que los ángulos son los más pequeños al incrementar la dimensión de los datos.

4.2.2. Vector media $v_d = \beta 1_d$ con $\beta > 0$

Ahora consideremos el vector media $v_d = \beta 1_d$ con $\beta = 0.5, 10$. En las figuras 4.9 y 4.10 se muestran las medias de los ángulos formados por v_d y los vectores normales de los cuatro métodos considerados variando la dimensión d. Si el Teorema 3.6.1 fuera válido, los límites de los ángulos cuando $d \to \infty$ serían $\arccos\left(\frac{\beta}{(\gamma\sigma^2+\beta^2)^{1/2}}\right) = 0.5639$ y 0.0316 para $\beta = 0.5$ y 10, respectivamente, lo cual está de acuerdo con lo observado en las figuras 4.9 y 4.10. Observemos que, al igual que en el caso anterior, en ambos casos, MD presenta el mejor comportamiento, el método DWD es el segundo mejor método y prácticamente coincide con MD cuando $\beta = 10$.

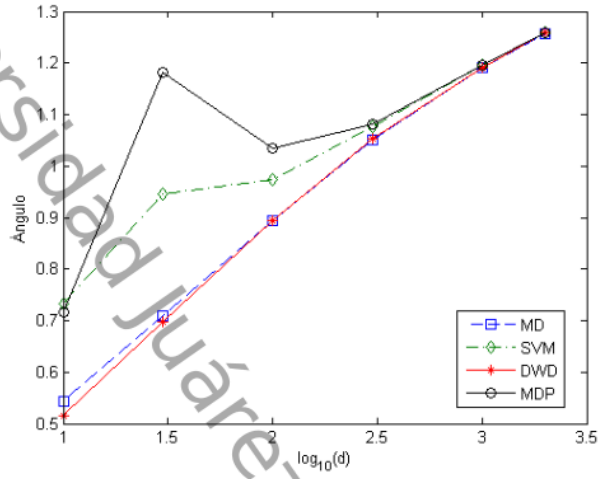

Figura 4.6: Medias de los ángulos entre $v_d = (d^\delta, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.2$. Caso no normal con Σ_d diagonal.

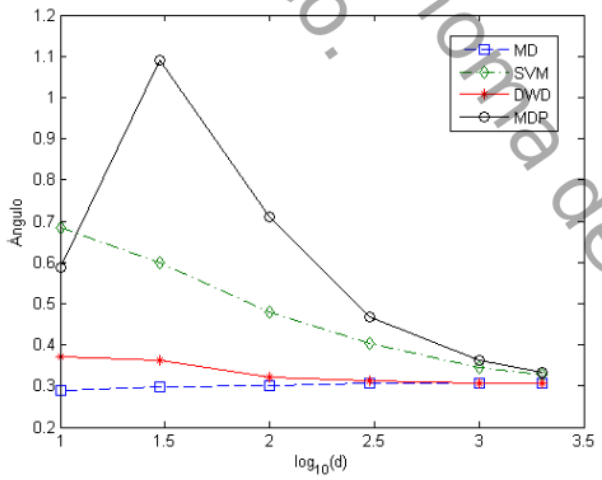

Figura 4.7: Medias de los ángulos entre $v_d = (d^\delta, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.5$. Caso no normal con Σ_d diagonal.

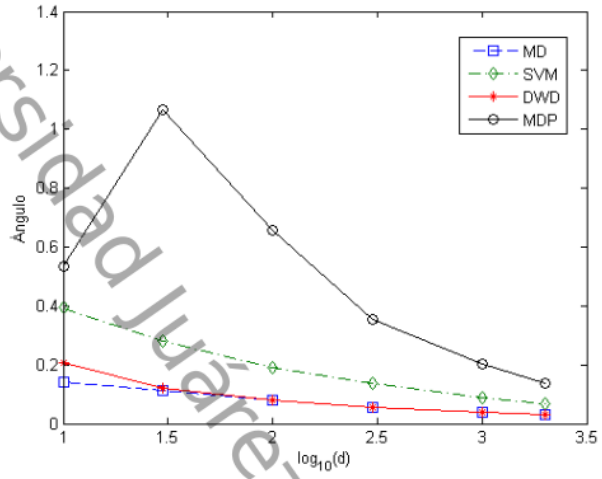

Figura 4.8: Medias de los ángulos entre $v_d = (d^{\delta}, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.8$. Caso no normal con Σ_d diagonal.

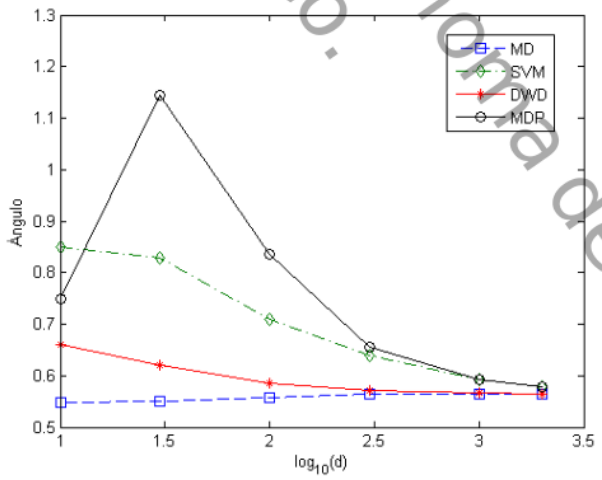

Figura 4.9: Medias de los ángulos entre $v_d = \beta 1_d$ y el vector normal del hiperplano separante; $\beta = 0.5$. Caso no normal con Σ_d diagonal.

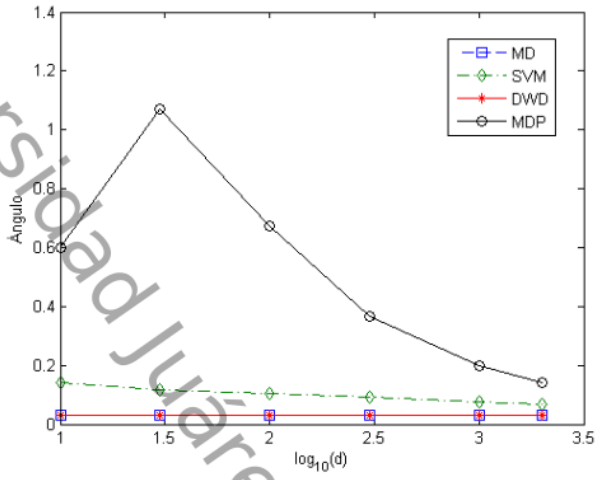

Figura 4.10: Medias de los ángulos entre $v_d = \beta 1_d$ y el vector normal del hiperplano separante; $\beta = 10$. Caso no normal con Σ_d diagonal.

4.3. Caso no normal con Σ_d no diagonal

Sea d par. Supongamos que $Z^{(1)}, Z^{(2)}, \ldots, Z^{(d/2)}$ son independientes con distribución $N(0,1)$ y sea $Y^{(k)} = (Z^{(k)})^2 + Z^{(k)} - 1$, para $i = 1, 2, \ldots, d/2$. Definamos el vector $Z(d) = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(d/2)}, Y^{(1)}, Y^{(2)}, \cdots, Y^{(d/2)})'$. Notemos que $E(Y^{(k)}) = 0$, así que $E(Z(d)) = 0$.

Además $Var(Z^{(k)}) = 1$, para $k = 1, \ldots, d/2$. Luego

$$\begin{aligned} Var(Y^{(k)}) &= Var((Z^{(k)})^2 + Z^{(k)} - 1) = E[((Z^{(k)})^2 + Z^{(k)} - 1)^2] \\ &= E[((Z^{(k)})^2 + Z^{(k)})^2 - 2((Z^{(k)})^2 + Z^{(k)}) + 1)] \\ &= E[((Z^{(k)})^2 + Z^{(k)})^2] - 2E[(Z^{(k)})^2 + Z^{(k)}] + 1 \\ &= E[(Z^{(k)})^4 + 2(Z^{(k)})^3 + (Z^{(k)})^2] - 1 \\ &= E[((Z^{(k)})^2)^2] = Var[(Z^{(k)})^2] + [E((Z^{(k)})^2)]^2 = 3, \end{aligned}$$

para $k = 1, \ldots, d/2$, ya que $(Z^{(i)})^2$ tiene distribución $\mathcal{X}^2$ con 1 grado de libertad y son independientes. Observemos también que $Cov(Z^{(k)}, Y^{(k)}) = 1$ para $k = 1, \ldots, d/2$ y $Cov(Z^{(i)}, Y^{(j)}) = 0$, para $i \neq j$. Por lo que la matriz de covarianza es

$$\Sigma_d = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 & 3 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 3 \end{pmatrix}.$$

Tomemos $Z_1(d), Z_2(d), \ldots, Z_n(d)$ datos multivariados i.i.d. con la misma distribución que $Z(d)$. En este caso, veremos que estos datos cumplen con la representación geométrica cuando la dimensión d crece, sin verificar las condiciones $(i) - (iv)$ que se especifican en la Sección 2.6 del Capítulo 2, puesto que a pesar de que estos datos satisfacen las condiciones $(i) - (iii)$, resulta que (iv) no se tiene.

Ya que $E\left[(Z_i^{(k)})^2\right] = 1$ y $E\left[\left((Z_i^{(k)})^2 + Z_i^{(k)} - 1\right)^2\right] = 3$, para $i = 1, 2, \ldots, n$ y $k = 1, 2, \ldots, d/2$, por la Ley Débil de los Grandes Números (Teorema 1.5.14) tenemos

$$\frac{|Z_i|^2}{d} = \frac{1}{2}\frac{\sum_{k=1}^{d/2}(Z_i^{(k)})^2}{d/2} + \frac{1}{2}\frac{\sum_{k=1}^{d/2}\left((Z_i^{(k)})^2 + Z_i^{(k)} - 1\right)^2}{d/2} \to \frac{1}{2} + \frac{3}{2} = 2$$

en probabilidad cuando $d \to \infty$. Análogamente, dado que $E[(Z_i^{(k)} - Z_j^{(k)})^2] = 2$ y $E[((Z_i^{(k)})^2 + Z_i^{(k)} - (Z_j^{(k)})^2 - Z_j^{(k)})^2] = 6$, para $i \neq j$ y $k = 1, 2, \ldots, d/2$, nuevamente por la Ley Débil de los Grandes Números tenemos

$$\begin{aligned} \frac{|Z_i - Z_j|}{d} &= \frac{1}{2}\frac{\sum_{k=1}^{d/2}(Z_i^{(k)} - Z_j^{(k)})^2}{d/2} + \frac{1}{2}\frac{\sum_{k=1}^{d/2}[(Z_i^{(k)})^2 + Z_i^{(k)} - 1 - ((Z_j^{(k)})^2 + Z_j^{(k)} - 1)]^2}{d/2} \\ &\to \frac{1}{2}Var(Z_i^{(k)} - Z_j^{(k)}) + \frac{1}{2}Var\left[(Z_i^{(k)})^2 + Z_i^{(k)} - (Z_j^{(k)})^2 - Z_j^{(k)}\right] = \frac{2}{2} + \frac{6}{2} = 4 \end{aligned}$$

en probabilidad cuando $d \to \infty$. Por lo que los datos tienen una representación geométrica asintótica.

4.3.1. Vector media $v_d = (d^\delta, 0, \cdots 0)'$ con $\delta > 0$

Tomamos $v_d = (d^\delta, 0, \cdots 0)'$, con $\delta = 0.2, 0.5$ y 0.8, como el vector media para el primer conjunto de simulaciones.

En las figuras 4.11, 4.12 y 4.13 se muestran las medias de los ángulos formados por v_d y el vector normal de los cuatro métodos considerados variando la dimensión d, correspondientes a los valores de $\delta = 0.2, 0.5$ y 0.8, respectivamente. En la Figura 4.11 observamos que cuando la dimensión crece las medias de los ángulos de todos los métodos tienden a un valor cercano a $\pi/2 = 1.5707$. Si se cumpliera el Teorema 3.6.1 para datos con representación geométrica, para $\delta = 0.5$ el límite de los ángulos estaría dado por $\arccos(c/(\gamma^2\sigma^2 + c^2)^{1/2}) = 0.4205$ cuando la dimensión tiende a infinito, donde $\gamma = 1/m + 1/n = 1/10$, $\sigma^2 = \lim_{d\to\infty} \frac{\sum_{k=1}^{d} \sigma_{kk}}{d} = \lim_{d\to\infty} \frac{tr(\Sigma_d)}{d} = 2$ y $c = 1$, lo cual está de acuerdo con lo observado en la Figura 4.12. En la Figura 4.13 las medias de los ángulos de todos los métodos tienden a cero. Nuevamente, observemos que el Teorema 3.6.1 sigue valiéndose para este caso. Notememos que en este caso, al igual que en los casos anteriores, MD y DWD tienen también el mejor comportamiento cuando $\delta = 0.5$ y 0.8, ya que las medias de los ángulos son las más pequeñas al incrementar la dimensión de los datos. Para $\delta = 0.2$, los dos mejores métodos fueron DWD y SVM.

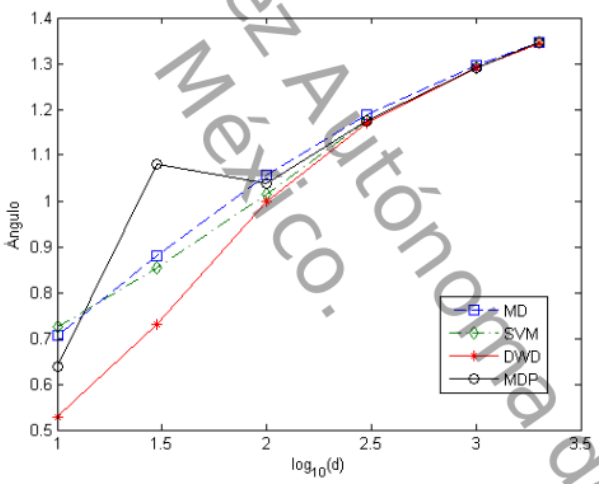

Figura 4.11: Medias de los ángulos entre $v_d = (d^\delta, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.2$. Caso no normal con Σ_d no diagonal.

4.3.2. Vector media $v_d = \beta 1_d$ con $\beta > 0$

Ahora consideremos el vector media $v_d = \beta 1_d$ con $\beta = 0.5, 10$. En las figuras 4.14 y 4.15 se muestran las medias de los ángulos formados por v_d y los vectores normales de los cuatro métodos considerados variando la dimensión d. Si el Teorema 3.6.1 fuera válido, los límites de los ángulos cuando $d \to \infty$ serían $arccos\left(\frac{\beta}{(\gamma\sigma^2+\beta^2)^{1/2}}\right) = 0.7297$ y 0.0447 para $\beta = 0.5$ y 10, respectivamente, lo cual coincide con lo que se observa

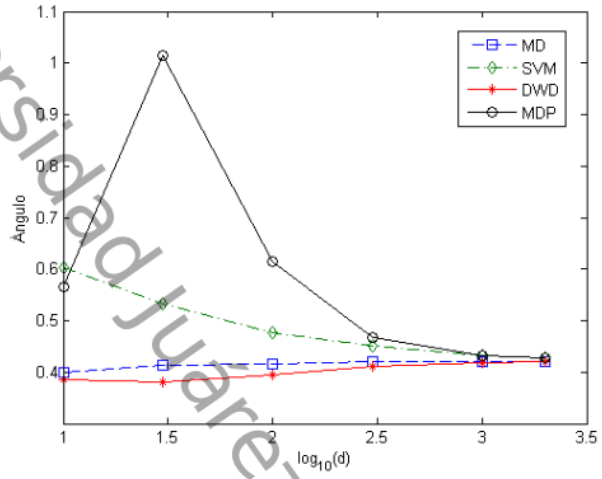

Figura 4.12: Medias de los ángulos entre $v_d = (d^\delta, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.5$. Caso no normal con Σ_d no diagonal.

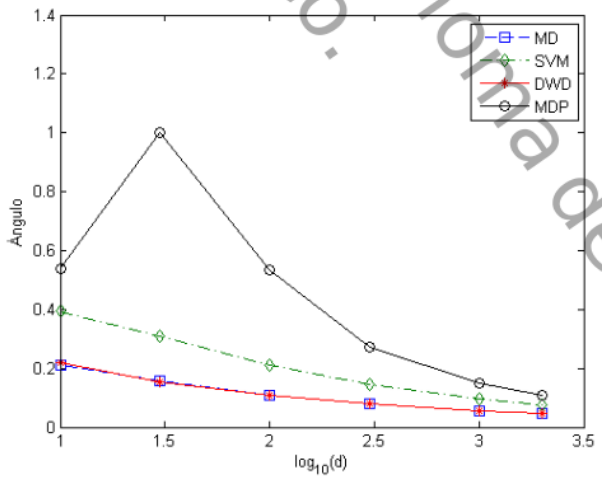

Figura 4.13: Medias de los ángulos entre $v_d = (d^\delta, 0, \cdots 0)'$ y el vector normal del hiperplano separante; $\delta = 0.8$. Caso no normal con Σ_d no diagonal.

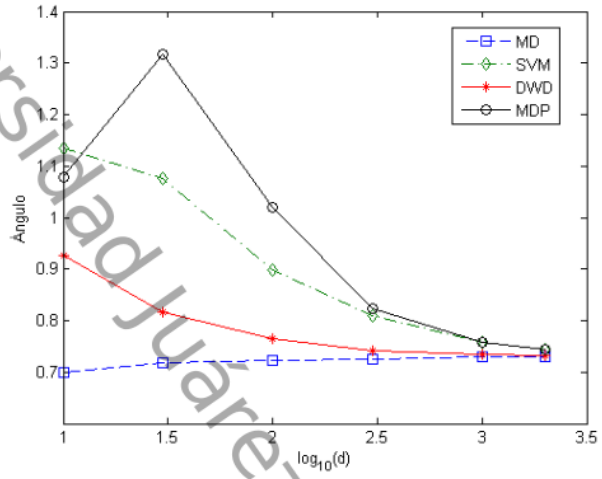

Figura 4.14: Medias de los ángulos entre $v_d = \beta 1_d$ y el vector normal del hiperplano separante; $\beta = 0.5$. Caso no normal con Σ_d no diagonal.

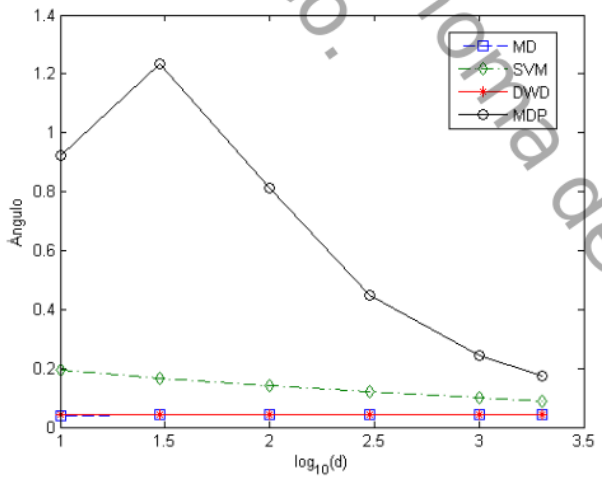

Figura 4.15: Medias de los ángulos entre $v_d = \beta 1_d$ y el vector normal del hiperplano separante; $\beta = 10$. Caso no normal con Σ_d no diagonal.

en las figuras 4.14 y 4.15. Observemos que en ambos casos MD vuelve a presentar el mejor comportamiento, el método DWD es el segundo mejor método y prácticamente coincide con MD cuando $\beta = 10$.

Conclusiones

Para realizar la exploración de la generalización del Teorema 3.6.1, demostrado en [5], que considera únicamente datos multivariados Gaussianos con matriz de covarianza diagonal, en esta tesis se hicieron simulaciones considerando datos multivariados más generales, ya que lo único que se pide es que tengan una representación geométrica asintótica cuando la dimensión tiende a infinito, como la de los datos multivariados Gaussianos. Con los resultados obtenidos en las simulaciones se concluye que el Teorema 3.6.1 podría generalizarce considerando datos que cumplan la representación geométrica. Por lo que se tendría para esta clase de datos que, bajo ciertas condiciones, los métodos MD, SVM, DWD y MDP tienen el mismo comportamiento asintótico cuando la dimensión crece.

Por otra parte observamos que los métodos MD y DWD tienen el mejor comportamiento entre los cuatro métodos, ya que en casi todos los casos y escenarios considerados presentan siempre las menores medias de los ángulos entre sus vectores normales y la dirección óptima. Conforme la dimensión de los datos crece se nota un comportamiento muy similar de estos dos métodos en algunos escenarios. El tercer mejor método es SVM y el método con el peor comportamiento en casi todos los escenarios es el MDP.

Cabe mencionar que los resultados de las simulaciones que son presentados en [4], considerando únicamente datos multivariados Gaussianos con matriz de covarianza diagonal para comprobar los resultados teóricos del Teorema 3.6.1, tienen similitud a los que hemos obtenido en este trabajo. De hecho se observó en las simulaciones que cuando la distancia entre las clases es cada vez más grande (caso cuando $|v_d|d^{-1/2} \to \infty$) los cuatro métodos tienden a ser consistentes cuando $d \to \infty$, en el sentido que los ángulos entre los vectores normales de los hiperplanos separantes de los métodos y la dirección óptima tienden a cero; mientras que cuando esta distancia es cada vez más pequeña (caso cuando $|v_d|d^{-1/2} \to 0$) los cuatro métodos tienden a ser fuertemente inconsistentes, en el sentido que los ángulos mencionados anteriormente tienden a $\pi/2$. Cuando la distancia entre las clases es aproximadamente constante (caso cuando $|v_d|d^{-1/2} \to c$, con $0 < c < \infty$) los métodos tienden a ser inconsistentes, en el sentido que los ángulos convergen a un valor en el intervalo $(0, \pi/2)$.

Apéndice A

Algunas Definiciones y Demostraciones

A.1. Matriz Toeplitz

La siguiente definición puede consultarse en [6].

Definición A.1.1 *Una* ***matriz Toeplitz*** $A = [a_{i,j}]$ *es una matriz* $d \times d$*, tal que* $a_{i,j} = a_{i+1,j+1}$*,* $i, j = 1, \ldots, d-1$.

Por ejemplo, para $d = 5$, la matriz Toeplitz presenta la forma

$$A = \begin{pmatrix} a & b & c & d & e \\ f & a & b & c & d \\ g & f & a & b & c \\ h & g & f & a & b \\ i & h & g & f & a \end{pmatrix}$$

Notemos que si la matriz A es simétrica, entonces la matriz Toeplitz está determinada por la primera fila y por ejemplo, para $d = 5$, se usa la notación $A = toeplitz(a, b, c, d, e)$ y tiene la forma

$$A = \begin{pmatrix} a & b & c & d & e \\ b & a & b & c & d \\ c & b & a & b & c \\ d & c & b & a & b \\ e & d & c & b & a \end{pmatrix}.$$

Así, la matriz Toeplitz $\Sigma_d = toeplitz(1, s, s^2, \ldots, s^{d-1})$ considerada en la Sección 4.1 es

$$\Sigma_d = \begin{pmatrix} 1 & s & s^2 & s^3 & \cdots & s^{d-1} \\ s & 1 & s & s^2 & \cdots & s^{d-2} \\ s^2 & s & 1 & s & \cdots & s^{d-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s^{d-1} & s^{d-2} & s^{d-3} & s^{d-4} & \cdots & 1 \end{pmatrix}.$$

A.2. Matriz positiva definida

A continuación presentamos algunas definiciones y enunciamos dos criterios (que pueden ser consultados por ejemplo en [17], Sección 4.8, y en [16], Sección A8) que son utilizados para demostrar que Σ_d satisface la condición (ii) de la representación geométrica en la Sección 4.1.

Definición A.2.1 *Una matriz A simétrica $d \times d$ es* ***positiva (negativa) definida*** *si $\alpha' A\alpha > 0$ ($\alpha' A\alpha < 0$), $\forall\ \alpha \in \mathbb{R}^d,\ \alpha \neq 0$. La matriz A es llamada* ***no negativa (no positiva) definida*** *si $\alpha' A\alpha \geqslant 0$ ($\alpha' A\alpha \leqslant 0$), $\forall\ \alpha \in \mathbb{R}^d$.*

Criterio A.2.1 *Sea $A = (a_{pq})$, $p, q = 1, 2, \ldots, d$, una matriz simétrica. El l-ésimo menor principal de A es $|B_l| = det(B_l)$, donde $B_l = (a_{pq})$, $p, q = 1, 2, \ldots, l$. La matriz A es positiva definida si y sólo si todos sus menores principales son positivos.*

Criterio A.2.2 *Una matriz simétrica A es positiva definida (no negativa definida) si y sólo si todos sus valores propios son positivos (no negativos).*

A.3. Detalles técnicos

A.3.1. Representación geométrica con condición ρ-mixing

En este apartado se detalla cómo la condición ρ-mixing es usada para justificar (2.8).

Observemos que para la muestra $\mathcal{X}(d) = \{X_1(d), ..., X_n(d)\}$, tenemos que

$$
\begin{aligned}
& E\left[\sum_{k=1}^{d}\left\{\left(X_i^{(k)} - X_j^{(k)}\right)^2 - E\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2\right]\right\}\right]^2 \\
&= E\left[\sum_{k=1}^{d}\left(X_i^{(k)} - X_j^{(k)}\right)^2 - \sum_{k=1}^{d}\left(E\left[\left(X_i^{(k)}\right)^2\right] - 2E(X_i^{(k)})E(X_j^{(k)}) + E\left[\left(X_j^{(k)}\right)^2\right]\right)\right]^2 \\
&= E\left[\sum_{k=1}^{d}\left(X_i^{(k)} - X_j^{(k)}\right)^2 - 2\sum_{k=1}^{d} VarX_i^{(k)}\right]^2 \\
&= E\left[\left\{\sum_{k=1}^{d}\left(X_i^{(k)} - X_j^{(k)}\right)^2\right\}^2\right] - 4\left[\sum_{k=1}^{d} VarX_i^{(k)}\right]\left[\sum_{k=1}^{d} E\left(X_i^{(k)} - X_j^{(k)}\right)^2\right] \\
&+ 4\left(\sum_{k=1}^{d} VarX_i^{(k)}\right)^2 \\
&= E\left[\left\{\sum_{k=1}^{d}\left(X_i^{(k)} - X_j^{(k)}\right)^2\right\}^2\right] - 8\left[\sum_{k=1}^{d} VarX_i^{(k)}\right]^2 + 4\left(\sum_{k=1}^{d} VarX_i^{(k)}\right)^2.
\end{aligned}
$$

Notemos que la expresión anterior es igual a lo siguiente

$$\sum_{k=1}^{d} E\left[\left(X_i^{(k)} - X_j^{(k)}\right)^4\right] + 2\sum_{k<l} E\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2 \left(X_i^{(l)} - X_j^{(l)}\right)^2\right]$$
$$- 4\left(\sum_{k=1}^{d} Var X_i^{(k)}\right)^2$$
$$= \sum_{k=1}^{d} E\left[\left(X_i^{(k)} - X_j^{(k)}\right)^4\right] + 2\sum_{k<l} Cov\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2, \left(X_i^{(l)} - X_j^{(l)}\right)^2\right]$$
$$+ 2\sum_{k<l} E\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2\right] E\left[\left(X_i^{(l)} - X_j^{(l)}\right)^2\right] - 4\left(\sum_{k=1}^{d} Var X_i^{(k)}\right)^2$$
$$= \sum_{k=1}^{d} E\left[\left(X_i^{(k)} - X_j^{(k)}\right)^4\right] + 2\sum_{k<l,|k-l|\leqslant r} Cov\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2, \left(X_i^{(l)} - X_j^{(l)}\right)^2\right]$$
$$+ 2\sum_{k<l,|k-l|>r} Corr\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2, \left(X_i^{(l)} - X_j^{(l)}\right)^2\right] \sqrt{Var X_i^{(k)}}\sqrt{Var X_i^{(l)}}$$
$$+ 8\sum_{k<l} Var X_i^{(k)} Var X_i^{(l)} - 4\left(\sum_{k=1}^{d} Var X_i^{(k)}\right)^2. \tag{A.1}$$

Debido a que los cuartos momentos de las entradas de X_i están uniformemente acotadas, se tiene que existen constantes positivas M_1, M_2 y M_3 tales que

$$E\left[\left(X_i^{(k)} - X_j^{(k)}\right)^4\right] \leqslant M_1,$$
$$Cov\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2, \left(X_i^{(l)} - X_j^{(l)}\right)^2\right] \leqslant M_2 \quad \text{y}$$
$$Var X_i^{(k)} \leqslant M_3$$

para todo $k, l = 1, 2, \ldots, d,\ k \neq l$. Así, la expresión (A.1) es menor o igual a

$$\sum_{k=1}^{d} M_1 + 2\left(\sum_{k<l,|k-l|\leqslant r} M_2\right) + 2\left(\sum_{k<l,|k-l|>r} M_3 Corr\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2, \left(X_i^{(l)} - X_j^{(l)}\right)^2\right]\right)$$
$$- 4\left(\sum_{k=1}^{d} Var^2 X_i^{(k)}\right)$$
$$\leqslant dM_1 + 2r(d-1)dM_2 + 2\frac{(d-r-1)(d-r)}{2} M_3\rho(r) - 4dM_3^2. \tag{A.2}$$

Debido a la condición ρ-mixing, la expresión (A.2), dividida entre d^2 se puede hacer tan pequeña como se quiera tomando r suficientemente grande y posteriormente d

suficientemente grande, por lo que

$$\frac{1}{d^2}E\left[\sum_{k=1}^{d}\left\{\left(X_i^{(k)}-X_j^{(k)}\right)^2-E\left[\left(X_i^{(k)}-X_j^{(k)}\right)^2\right]\right\}\right]^2\to 0,\quad \text{cuando } d\to\infty.$$

Así, por la desigualdad de Chebyshev (Teorema 1.4.3), tenemos que

$$\frac{1}{d}\left[\sum_{k=1}^{d}\left\{\left(X_i^{(k)}-X_j^{(k)}\right)^2-E\left[\left(X_i^{(k)}-X_j^{(k)}\right)^2\right]\right\}\right]\to 0 \tag{A.3}$$

en probabilidad cuando $d\to\infty$. Lo obtenido en (A.3) junto con la condición (2.7), nos implica el resultado en (2.8). En efecto

$$\frac{1}{d}\left[\sum_{k=1}^{d}\left\{\left(X_i^{(k)}-X_j^{(k)}\right)^2-E\left[\left(X_i^{(k)}-X_j^{(k)}\right)^2\right]\right\}\right]^2\to 0$$

$$\Rightarrow \frac{1}{d}\sum_{k=1}^{d}\left(X_i^{(k)}-X_j^{(k)}\right)^2-\frac{1}{d}\sum_{k=1}^{d}E\left[\left(X_i^{(k)}-X_j^{(k)}\right)^2\right]\to 0$$

$$\Rightarrow \frac{1}{d}\sum_{k=1}^{d}\left(X_i^{(k)}-X_j^{(k)}\right)^2-\frac{2}{d}\sum_{k=1}^{d}VarX_i^{(k)}\to 0$$

$$\Rightarrow \frac{1}{\sqrt{d}}\left\{\sum_{k=1}^{d}\left(X_i^{(k)}-X_j^{(k)}\right)^2\right\}^{1/2}\to\left(2\sigma^2\right)^{1/2}$$

en probabilidad cuando $d\to\infty$, como se deseaba mostrar.

A.3.2. Demostración de que Σ_d de la Sección 4.1 es positiva definida

Demostraremos por inducción que todos los menores principales $|B_l|$, $l=1,2,\ldots,d$, son positivos, donde $B_l=(\sigma_{ij})$, $i,j=1,2,\ldots,l$. Así, por el Criterio A.2.1, Σ_d es positiva definida.

Para $l=1$, tenemos que $|B_1|=1=(1-s^2)^0$. Supongamos que $|B_l|=(1-s^2)^{l-1}$. Veamos ahora que $|B_{l+1}|=(1-s^2)^l$. Por propiedades de los determinantes tenemos que

$$
\begin{aligned}
|B_{l+1}| &= \begin{vmatrix}
1 & s & s^2 & s^3 & \cdots & s^l \\
s & 1 & s & s^2 & \cdots & s^{l-1} \\
s^2 & s & 1 & s & \cdots & s^{l-2} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
s^{l-1} & s^{l-2} & s^{l-3} & s^{l-4} & \cdots & s \\
s^l & s^{l-1} & s^{l-2} & s^{l-3} & \cdots & 1
\end{vmatrix} \\
&= 1|B_l| - s \begin{vmatrix}
s & s^2 & s^3 & \cdots & s^{l-1} & s^l \\
s & 1 & s & \cdots & s^{l-3} & s^{l-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
s^{l-1} & s^{l-2} & s^{l-3} & \cdots & s & 1
\end{vmatrix} \\
&\quad + s^2 \begin{vmatrix}
s & s^2 & \cdots & s^{l-2} & s^{l-1} & s^l \\
1 & s & \cdots & s^{l-3} & s^{l-2} & s^{l-1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
s^{l-1} & s^{l-2} & \cdots & s^2 & s & 1
\end{vmatrix} \\
&\quad + \cdots + (-1)^{l+1} s^l \begin{vmatrix}
s & s^2 & \cdots & s^{l-2} & s^{l-1} & s^l \\
1 & s & \cdots & s^{l-3} & s^{l-2} & s^{l-1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
s^{l-2} & s^{l-3} & \cdots & s^3 & s^2 & s
\end{vmatrix}.
\end{aligned}
$$

Notemos que a partir del segundo término de la última expresión, podemos factorizar s del primer renglón de la matriz en el determinante. A partir del tercer término, cada determinante es igual a cero debido a que la primera y la segunda fila de la matriz son iguales. Así,

$$|B_{l+1}| = |B_l| - s^2|B_l| = |B_l|(1 - s^2) = (1 - s^2)^l,$$

por la hipótesis de inducción. De esta forma, $|B_l| = (1 - s^2)^{l-1} > 0$ para toda $l \geqslant 1$.

A.3.3. Desigualdad (4.1) de la Sección 4.1

A continuación justificaremos la desigualdad utilizada en (iii) de la Sección 4.1. Dada $0 < s < 1$, se tiene que la función $f(x) = xs^{-2x}$ es creciente para valores positivos de x, ya que $f'(x) = s^{-2x} + x\left[-\ln(s^2)s^{-2x}\right] = s^{-2x}\left[1 - x\ln(s^2)\right]$, y

$$f'(x) > 0 \Leftrightarrow 1 - x\ln(s^2) > 0 \Leftrightarrow x\ln(s^2) < 1 \Leftrightarrow x > \frac{1}{\ln(s^2)}.$$

Por lo tanto, por las propiedades de las sumas de Riemann es válida la desigualdad

$$\sum_{k=1}^{d} ks^{-2k} \leq \int_1^{d+1} xs^{-2x}dx.$$

Luego

$$\begin{aligned}
\int_1^{d+1} xs^{-2x}dx &= \left[\frac{xs^{-2x}}{\ln(s^{-2})} - \frac{s^{-2x}}{(\ln(s^{-2}))^2}\right]\Bigg|_1^{d+1} = \left[\frac{s^{-2x}}{\ln(s^{-2})}\left(x - \frac{1}{\ln(s^{-2})}\right)\right]\Bigg|_1^{d+1} \\
&= \frac{s^{-2(d+1)}}{\ln(s^{-2})}\left(d+1-\frac{1}{\ln(s^{-2})}\right) - \frac{s^{-2}}{\ln(s^{-2})}\left(1-\frac{1}{\ln(s^{-2})}\right) \\
&= \frac{ds^{-2d}s^{-2}}{\ln(s^{-2})} + \frac{s^{-2d}s^{-2}}{\ln(s^{-2})}\left(1-\frac{1}{\ln(s^{-2})}\right) - \frac{s^{-2}}{\ln(s^{-2})}\left(1-\frac{1}{\ln(s^{-2})}\right) \\
&= \frac{s^{-2}}{\ln(s^{-2})}\left[ds^{-2d} + \left(1-\frac{1}{\ln(s^{-2})}\right)\left(s^{-2d}-1\right)\right].
\end{aligned}$$

Por lo que

$$\sum_{k=1}^{d} ks^{-2k} \leq \frac{s^{-2}}{\ln(s^{-2})}\left[ds^{-2d} + \left(1-\frac{1}{\ln(s^{-2})}\right)\left(s^{-2d}-1\right)\right].$$

Bibliografía

[1] J. Ahn and J. S. Marron. The Maximal Data Piling Direction for Discrimination. *Biometrika*, 97(1):254–259, 2010.

[2] J. Ahn, J. S. Marron, K. M. Muller, and Y. Chi. The High-dimension, Low-sample-size Geometric Representation Holds Under Mild Conditions. *Biometrika*, 94(3):760–766, 2007.

[3] R. B. Ash. *Probability and Measure Theory*. Harcourt Science and Technolog Company, 2000.

[4] A. Bolivar-Cime and J. S. Marron. Supplementary material for "Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data". Manuscrito disponible en `https://sites.google.com/site/addybolivarcime/publications/supplementary-files`, Septiembre 2012.

[5] A. Bolivar-Cime and J. S. Marron. Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data. *J. Multivar. Anal.*, 115:108–121, 2013.

[6] A. Bottcher and S. M. Grudsky. *Toeplitz Matrices, Asymtotic Linear Algebra, and Functional Analysis*. Birkhauser, 2012.

[7] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Minng. Knowl. Disc.*, 2:121–167, 1998.

[8] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, 2002.

[9] C. Cortes and V. N. Vapnik. Support-Vector Networks. *Mach. Learn.*, 20:273–297, 1995.

[10] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, U.K., 2000.

[11] P. Hall, J. S. Marron, and A. Neeman. Geometric Representation of High Dimension, Low Sample Size Data. *J. R. Statist. Soc. B*, 67(3):427–444, 2005.

[12] A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. Springer Texts in Statistics. Springer, New York, 2008.

[13] R. G. Laha and V. K. Rohatgi. *Probability Theory*. State University, 2000.

[14] J. S. Marron, M. J. Todd, and J. Ahn. Distance-Weighted Discrimination. Manuscript available at http://www.stat.uga.edu/∼jyahn/DWD/, February 2007.

[15] J. S. Marron, M. J. Todd, and J. Ahn. Distance-Weighted Discrimination. *J. Am. Statist. Ass.*, 102(480):1267–1271, 2007.

[16] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., 2005.

[17] N. Mukhopadhyay. *Probability and Statistical Inference*. Marcel Dekker, 2000.

[18] K. S. Pranab, M. S. Julio, and P. de Lima Antonio C. *From Finite Sample to Asymptotic Methods in Statistics*. Cambridge University Press, 2010.

[19] X. Qiao, H. H. Zhang, Y. Liu, M. J. Todd, and J. S. Marron. Weighted Distance Weighted Discrimination and its Asymptotic Properties. *J. Am. Statist. Ass.*, 105(489):401–414, 2010.

[20] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. The MIT press, Cambridge, Massachusetts, 2002.

[21] R. J. Serfling. *Approximation Theorems of Mathematical Statics*. John Wiley and Sons, 1980.

[22] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer-Verlag, Berlin, 1982.

[23] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, 1995.

COMPARACIÓN DE MÉTODOS DE DISCRIMINACIÓN BINARIA PARA DATOS DE DIMENSIÓN ALTA: UN ESTUDIO POR SIMULACIÓN

ORIGINALITY REPORT

6%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.geocities.ws Internet	254 words — 1%
2	www.amestad.mx Internet	132 words — 1%
3	www.matematicas.unam.mx Internet	98 words — 1%
4	hdl.handle.net Internet	55 words — < 1%
5	www.ugr.es Internet	50 words — < 1%
6	kupdf.net Internet	41 words — < 1%
7	dokumen.tips Internet	36 words — < 1%
8	www.eecs.tufts.edu Internet	36 words — < 1%
9	www.planetcup.com Internet	36 words — < 1%

10	www.dynamics.unam.edu Internet	34 words — < 1%
11	laberintos.itam.mx Internet	24 words — < 1%
12	www.slideshare.net Internet	24 words — < 1%
13	www.coursehero.com Internet	20 words — < 1%
14	b-ok.org Internet	18 words — < 1%
15	cms.dm.uba.ar Internet	18 words — < 1%
16	followscience.com Internet	17 words — < 1%
17	www.wiwi.uni-saarland.de Internet	15 words — < 1%
18	Jiandong Nie, Qiaoling Chen, Zhidong Teng, Yihan Zhang, Ramziya Rifhat. "Dynamics of a Stochastic Measles Model with General Incidence Rate and Black–Karasinski Process", Bulletin of the Malaysian Mathematical Sciences Society, 2024 Crossref	14 words — < 1%
19	Timothy C. Haas. "Probability and Simulation", Wiley, 2013 Crossref	14 words — < 1%
20	www.stat.wisc.edu Internet	

		13 words — < 1%
21	Kenneth Baclawski. "Introduction to Probability with R", Chapman and Hall/CRC, 2019 Publications	12 words — < 1%
22	documen.site Internet	12 words — < 1%
23	iugspace.iugaza.edu.ps Internet	12 words — < 1%
24	universidadesdemexico.mx Internet	12 words — < 1%
25	www.cs.famaf.unc.edu.ar Internet	12 words — < 1%
26	www.mathematik.uni-ulm.de Internet	12 words — < 1%
27	Kuwaranancharoen, Kananart. "Analyses and Scalable Algorithms for Byzantine-Resilient Distributed Optimization", Purdue University, 2023 ProQuest	11 words — < 1%
28	cimat.repositorioinstitucional.mx Internet	11 words — < 1%
29	qdoc.tips Internet	11 words — < 1%
30	www.math.uaic.ro Internet	11 words — < 1%

31	Alok Goswami, B.V. Rao. "Measure Theory for Analysis and Probability", Springer Science and Business Media LLC, 2025 Crossref	10 words — < 1%
32	aprenderly.com Internet	10 words — < 1%
33	pdfcookie.com Internet	10 words — < 1%
34	repositorio.cinvestav.mx Internet	10 words — < 1%
35	www.ecap.uab.es Internet	10 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 10 WORDS