UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS BÁSICAS

CLASIFICACIÓN MULTICATEGORÍA PARA DATOS DE DIMENSIÓN ALTA

TESIS PARA OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS MATEMÁTICAS

PRESENTA:

M.C.M. DORILIAN GARCÍA CERINO

BAJO LA DIRECCIÓN DE:

DRA. ADDY MARGARITA BOLÍVAR CIMÉ

EN CODIRECCIÓN DE:

DR. VÍCTOR MANUEL PÉREZ ABREU-CARRIÓN

CUNDUACÁN, TABASCO, SEPTIEMBRE 2024

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán, Tabasco, en el mes de Septiembre del año 2024, el que suscribe Dorilian García Cerino, egresado del programa de Doctorado en Ciencias Matemáticas con número de matrícula 192a22002 adscrito a la División Académica de Ciencias Básicas, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención del grado de Doctor en Ciencias Matemáticas y titulada Clasificación Multicategoría para Datos de Dimensión Alta dirigida por la Dra. Addy Margarita Bolívar Cimé en codirección con el Dr. Víctor Manuel Pérez Abreu-Carrión.

DECLARO QUE:

La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad Juárez Autónoma de Tabasco cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

M.C.M. Dorilian García Cerino

DIRECCIÓN

Cunduacán, Tabasco a 17 de septiembre de 2024

M.C. DORILIAN GARCÍA CERINO
PASANTE DE DOCTORADO EN CIENCIAS MATEMÁTICAS
PRESENTE

Por medio de la presente me dirijo a Usted para hacer de su conocimiento que proceda a la impresión del trabajo titulado "**CLASIFICACIÓN MULTICATEGORÍA PARA DATOS DE DIMENSIÓN ALTA**"; en virtud de que reúne los requisitos para el EXAMEN PROFESIONAL y obtener el grado de DOCTOR EN CIENCIAS MATEMÁTICAS.

Sin más por el momento, reciba un cordial saludo.

ATENTAMENTE

DIVISIÓN ACADÉMICA DE CIENCIAS BÁSICAS

DRA. HERMICENDA PÉREZ VIDAL
DIRECTORA

C.c.p.- Archivo

DIR´DRA.HPV/JP´DRA.EAM/jkal**

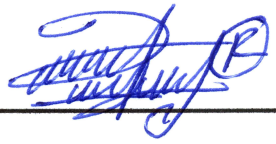
Carta de Cesión de Derechos

Cunduacán, Tabasco, Septiembre 2024

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y realización de la Tesis denominada Clasificación Multicategoría para Datos de Dimensión Alta.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y realización de la Tesis antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entiendo y acepto el alcance del artículo en mención, de que tengo el derecho al reconocimiento como autor de la Tesis, y la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la Tesis en la que colaboré, por lo anterior, cedo el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR



M.C.M. Dorilian García Cerino

Egresado

Dedicatoria

A mi Familia

Agradecimientos

- Estoy profundamente agradecido con el Creador y con la Naturaleza por permitirme obtener el Grado de Doctor en Ciencias Matemáticas.
- De corazón le agradezco a mi Madre, la Sra. Dora María Cerino Hernández, por su gran opoyo incondicional durante toda mi trayectoria académica.
- A mi único hermano, Raúl García Cerino, le agradezco por darme los ánimos para lograr mis metas.
- Muchas gracias a la Lic. Mary Carmen Hernández Vargas, por su apoyo constante y por darme la motivación más importante para lograr ésta y todas mis metas futuras.
- Mis mas sinceros agradecimientos a mis directores de tesis, la Dra. Addy Margarita Bolívar Cimé y al Dr. Víctor Manuel Pérez Abreu-Carrión, por orientarme correctamente, por sus valiosas contribuciones y por ser pacientes conmigo en la realización de este trabajo.
- Agradezco al comité sinodal por sus observaciones y comentarios para la mejora de esta Tesis.
- Gracias a todos aquellos familiares, amigos, conocidos y profesores que me apoyaron con sus palabras de aliento y motivación para lograr este grado.
- Gracias al CONAHCYT-UJAT por el apoyo económico durante mis estudios de Doctorado.

Índice general

Índice de tablas

Índice de figuras

CLASIFICACIÓN MULTICATEGORÍA PARA DATOS DE DIMENSIÓN ALTA

Resumen

Se consideran extensiones multicategoría de métodos de discriminación binaria a través de las metodologías uno contra uno (OVO) o uno contra el resto (OVR), centrándose en las extensiones de los métodos mean difference (MD), support vector machine (SVM), maximal data piling (MDP) y distance weighted discrimination (DWD) a través de OVO, y la extensión multicategoría de MD a través de OVR, en el contexto de datos de alta dimensión alta. Se describe el comportamiento asintótico de los métodos de clasificación multicategoría OVO-MD, OVO-SVM, OVO-MDP y OVO-DWD cuando la dimensión de los datos aumenta y el tamaño de la muestra es fijo, en términos de las probabilidades de clasificación correcta de un nuevo dato. Se encuentran condiciones suficientes para que las probabilidades de clasificación correcta converjan a uno a medida que la dimensión se acerca a infinito. Al igual que en el caso binario, OVO-MD, OVO-SVM y OVO-MDP tienen el mismo comportamiento asintótico, mientras que OVO-DWD podría comportarse de manera diferente. Se aborda también el comportamiento asintótico del método de discriminación multicategoría OVR-MD, donde se proporcionan las condiciones necesarias y suficientes para que un nuevo dato de una clase dada sea correctamente clasificado con probabilidad tendiendo a uno, cuando la dimensión de los datos tiende a infinito y el tamaño de muestra permanece fijo. Se realiza un experimento de simulación para comparar aún más las metodologías, y se consideran los cuatro métodos binarios en el caso OVR. Se evalua el rendimiento de estos ocho métodos de clasificación multicategoría utilizando un conjunto de datos de expresión genética.

Palabras claves: clasificación multicategoría, datos de dimensión alta, uno contra uno, uno contra el resto, máquinas de vector soporte.

Abstract

We consider multicategory extensions of binary discrimination methods via one-versus-one (OVO) or one-versus-rest (OVR) methodologies, focusing on extensions of the binary classification by linear mean difference (MD), support vector machine (SVM), maximal data piling (MDP), and distance weighted discrimination (DWD) via OVO, and the multicategory extension of MD via OVR, in the context of high-dimensional and low sample size (HDLSS) data. The asymptotic behavior of OVO-MD, OVO-SVM, OVO-MDP and OVO-DWD is described when the dimension of the data increases and the sample size is fixed, in terms of the probabilities of correct classification of a new data point, finding sufficient conditions for the correct classification probabilities to converge to one as the dimension approaches infinity. As in the binary case, OVO-MD, OVO-SVM and OVO-MDP have the same asymptotic behavior while OVO-DWD could behave differently. We also consider the asymptotic behavior of the OVR-MD methodology providing necessary and sufficient conditions for a new data point of a given class to be correctly classified with probability tending to one, when the dimension of the data increases and the sample size is fixed. A simulation experiment is conducted to further compare the methodologies, and consider the four binary methods in the OVR case. We evaluate the performance of the considered methods using a microarray data set.

Keywords: multicategory discrimination, high-dimension and low sample size data, one-versus-one, one versus-rest, support vector machine.

Introducción

La clasificación, o discriminación, es una metodología estadística importante en el aprendizaje automático, que es ampliamente usada en problemas actuales tales como el reconocimiento de imágenes (Russakovsky et al. (2015)), el reconocimiento de voz (Yu y Deng (2016)) y el diagnóstico de cáncer (Khan et al. (2001)), entre muchos otros. Esta metodología tiene como propósito determinar una regla de clasificación, o clasificador, que pueda predecir etiquetas de categorías (o clases) en función de un conjunto determinado de variables. Este clasificador es construido con la información proporcionada por un conjunto de datos de entrenamiento, cuyas observaciones tienen etiquetas de categorías conocidas. Los problemas que involucran únicamente a dos clases (clasificación binaria) han sido bien estudiados, véase por ejemplo, Cortes y Vapnik (1995), Izenman (2008), y Hastie et al. (2017).

Actualmente los problemas de clasificación con más de dos categorías (clasificación multicategoría) son de gran interés. Por ejemplo, en el diagnóstico de cáncer de colon, la clasificación precisa de tumores en las categorías: linfoma de Burkitt, sarcoma de Ewing, neuroblastoma y rabdomiosarcoma, en base a sus perfiles de expresión genética (microarreglos ADN), puede proporcionar un mejor diagnóstico del cáncer, Khan et al. (2001). En estos microarreglos ADN se involucran costosas mediciones que producen expresiones simultáneas de miles a decenas de miles de genes, resultando así conjuntos de datos multivariados donde el número de observaciones es mucho menor que la dimensión de estos. A este tipo de datos se les conoce como datos de dimensión alta, y surgen en otras aplicaciones contemporáneas tales como la identificación de huellas dactilares y el reconocimiento facial, entre muchas otras, Wang (2012).

Una forma de abordar problemas de clasificación multicategoría es mediante una de las metodologías one-versus-one (OVO) ó one-versus-rest (OVR), resolviendo una serie de problemas de clasificación binaria. En esta tesis nos enfocamos principalmente en las versiones multicategoría

de los métodos de discriminación binaria mean difference (MD) (Scholkopf y Smola (2002)), support vector machine (SVM) (Cortes y Vapnik (1995)), maximal data piling (MDP) (Ahn y Marron (2010)) y distance weighted discrimination (DWD) (Marron et al. (2007)) vía OVO y en la versión multicategoría de MD vía OVR. Los denotamos por OVO-MD, OVO-SVM, OVO-MDP, OVO-DWD y OVR-MD, respectivamente.

Nuestro propósito principal en esta tesis es abordar teóricamente el comportamiento asintótico de los métodos de clasificación multicategoría OVO-MD, OVO-SVM, OVO-MDP, OVO-DWD y OVR-MD en el contexto de datos de dimensión alta con una representación geométrica cuando la dimensión de los datos crece y el tamaño de la muestra permanece fijo. Hacemos este análisis en términos de las probabilidades de clasificación correcta, análogamente al estudio realizado para algunos métodos de clasificación binaria en el trabajo pionero de esta línea de investigación de Hall et al. (2005). Cabe mencionar aquí que en Nakayama et al. (2017) las propiedades asintóticas de OVO-SVM fueron estudiadas bajo algunas condiciones sobre los momentos de las variables, condiciones diferentes a las consideradas en el presente trabajo.

Un propósito secundario en esta investigación es realizar un estudio de simulación numérico acerca del comportamiento asintótico de los últimos cinco métodos mencionados e incluyendo a los métodos de clasificación multicategoría SVM, MDP y DWD vía OVR, denotados por OVR-SVM, OVR-MDP y OVR-DWD, respectivamente. Hacemos estas simulaciones en el contexto de datos de dimensión alta con representación geométrica asintótica, bajo diferentes escenarios y diversas situaciones dimensionales para obtener las tasas de error de clasificación. Con esto, analizamos y comparamos el desempeño de los ocho clasificadores multicategoría, lo cual pudiera darnos información relevante para estudios teóricos futuros sobre las propiedades asintóticas de los clasificadores multicategoría OVR-SVM, OVR-MDP y OVR-DWD.

Cabe mencionar que para cumplir con el requisito de publicación en una revista científica para la obtención del grado, los resultados de esta tesis fueron publicados en García-Cerino et al. (2024).

Marco Teórico

Muchos métodos estadísticos tradicionales del análisis multivariado, como el análisis discriminante lineal de Fisher, no funcionan para los datos de dimensión alta puesto que la inversa de la matriz de covarianza no existe y deben utilizarse otras metodologías. En esta situación, una representación geométrica asintótica de éstos datos resulta ser muy útil para estudiar el comportamiento asintótico de esas metodologías. Por representación geométrica asintótica se refiere a que, bajo ciertas condiciones, datos de dimensión alta tienden a estar, con probabilidad convergiendo a uno, en los vértices de un simplex cuando la dimensión de los datos tiende a infinito y el tamaño de muestra permanece fijo; véase por ejemplo, Hall et al. (2005), Ahn et al. (2007) y Qiao et al. (2010).

Esta representación geométrica asintótica ha permitido estudiar exitosamente tanto el comportamiento de componentes principales así como el comportamiento asintótico de algunos métodos de clasificación binaria lineales; véase por ejemplo, Jung y Marron (2009), Yata y Aoshima (2012), y Hall et al. (2005). En particular, para los métodos MD, SVM, DWD y MDP recientemente fue mostrado en Bolívar-Cimé (2021) que tienen el mismo comportamiento asintótico en términos del ángulo entre el vector normal al hiperplano separante y el vector óptimo para clasificación, en el sentido que estos ángulos convergen en probabilidad a la misma constante cuando la dimensión de los datos tiende a infinito y el tamaño de muestra permanece fijo. Fue observado también que DWD puede tener un comportamiento diferente a MD, SVM y MDP, en términos de las probabilidades de clasificación correcta. Considerando condiciones basadas sobre los momentos de las entradas de los datos, Nakayama et al. (2017) y Kento et al. (2021) estudiaron las probabilidades de error de clasificación asintóticas de SVM y DWD, así como también de los métodos Bias Corrected SVM (BC-SVM), Bias Corrected DWD (BC-DWD) y Weighted DWD (WDWD).

En el caso de los métodos de clasificación multicategoría existen varios enfoques para extender un método de clasificación binaria al caso multicategoría. No obstante, hay poca investigación acerca del comportamiento asintótico de estos métodos en el contexto de datos de dimensión alta.

Una posible extensión se realiza resolviendo un solo problema de optimización que considera a todas las clases en una sola para encontrar a los clasificadores simultáneamente. En este sentido, el método denominado multiclass support vector machine (MSVM) fue propuesto en Lee et al. (2004) como una generalización de SVM, y sus propiedades asintóticas considerando datos de dimensión alta fueron recientemente obtenidas en Kento (2022). Análogamente, mediante este mismo enfoque, el método denominado multiclass distance weighted discrimination (MDWD) fue desarollado en Huang et al. (2013) como la única generalización de DWD; sin embargo, no se han investigado aún sus propiedades asintóticas en el contexto de datos de dimensión alta.

Justificación

Recientemente se ha visto la importancia del estudio de metodologías estadísticas para datos multivariados con dimensión mayor al tamaño de la muestra (datos de dimensión alta), debido a que aparecen en diversas aplicaciones contemporáneas y en diversos campos de la ciencia, véase por ejemplo Russakovsky et al. (2015), Yu y Deng (2016) y Khan et al. (2001). Es por ello que el estudio y la comprensión de metodologías estadísticas, tales como la clasificación binaria y multicategoría para datos de dimensión alta, han sido de interés para varios autores, véase por ejemplo Hall et al. (2005), Marron et al. (2007), Huang et al. (2013), Bolívar-Cimé y Marron (2013), Bolívar-Cimé y Córdova-Rodríguez (2018), Bolívar-Cimé (2021) y Kento et al. (2021). Estos autores estudian el comportamiento asintótico de algunos métodos de clasificación binaria cuando los datos tienen una representación geométrica asintótica al hacer crecer la dimensión y mantener el tamaño de muestra fijo. Como se menciona en Huang et al. (2013), un problema abierto es el comportamiento asintótico de los métodos de discriminación multicategoría cuando los datos tienen una representación geométrica asintótica. Debido a esto, es de interés llevar a cabo un estudio teórico y por simulaciones para investigar el comportamiento asintótico de diversos métodos de clasificación multicategoría en el contexto de datos de dimensión alta.

Pregunta de Investigación

Se demostró en Hall et al. (2005) y Bolívar-Cimé (2021) que bajo la representación geométrica asintótica de datos de dimensión alta, esto es, bajo algunas condiciones que involucran los tamaños muestrales, las distancias asintóticas entre los datos y sus medias de clase, y la distancia asintótica entre pares de medias de clase, los hiperplanos separantes de los métodos de clasificación binaria MD, SVM y MDP coinciden asintóticamente cuando la dimensión tiende a infinito; y que el hiperplano separante del método de discriminación binaria DWD coincide con aquellos de MD, SVM y MDP únicamente cuando los tamaños muestrales de las dos clases son iguales. Esto conduce a preguntarnos, ¿así como en el caso binario, los métodos de clasificación multicategoría OVO-MD, OVO-SVM y OVO-MDP tienen el mismo comportamiento asintótico cuando la dimensión de los datos tiende a infinito y el tamaño de muestra permanece fijo? y ¿el método de discriminación multicategoría OVO-DWD pudiera tener un comportamiento diferente? En este mismo contexto nos preguntamos, ¿qué relaciones o similitudes hay en el comportamiento asintótico de los métodos de clasificación multicategoría OVR-MD, OVR-SVM, OVR-MDP y OVR-DWD?

Hipótesis o Supuesto

Debido al funcionamiento de la metodología OVO conjeturamos intuitivamente que, bajo la representación geométrica asintótica de datos de dimensión alta, las extensiones multicategoría de MD, SVM, MDP y DWD vía OVO y vía OVR tienen un comportamiento asintótico análogo al del caso binario.

Objetivo General

Determinar, teóricamente y mediante simulaciones, el comportamiento asintótico de algunos métodos de clasificación multicategoría, considerando datos multivariados con representación geométrica asintótica cuando la dimensión de los datos tiende a infinito y el tamaño de la muestra permanece fijo.

Objetivos Específicos

(1) Estudiar los métodos de clasificación binaria MD, SVM, DWD y MDP, y su implementación computacional.

(2) Estudiar las extensiones multicategoría de MD, SVM, DWD y MDP vía OVO y vía OVR, en el contexto de datos de dimensión alta.

(3) Investigar teóricamente el comportamiento asintótico de los métodos de clasificación multicategoría OVO-MD, OVO-SVM, OVO-MDP, OVO-DWD y OVR-MD en términos de probabilidades de error de clasificación, en el contexto de datos multivariados con representación geométrica asintótica cuando la dimensión tiende a infinito y el tamaño de la muestra permanece fijo.

(4) Realizar simulaciones de clasificación multicategoría utilizando métodos de interés con datos multivariados que tengan una representación geométrica asintótica, con el fin de obsevar el comportamiento de los métodos OVO-MD, OVO-SVM, OVO-MDP, OVO-DWD, OVR-MD, OVR-SVM, OVR-MDP y OVR-DWD al variar la dimensión de los datos.

(5) Mostrar un ejemplo de aplicación.

(6) Proponer algunos problemas futuros de investigación.

Metodología

La investigación teórica que realizamos en esta tesis sigue la siguiente metodología. En el Capítulo 1 presentamos un desarrollo de la representación geométrica asintótica de datos estadísticos multivariados con dimensión mayor o igual al tamaño de la muestra (datos de dimensión alta), considerando tanto datos con distribución gaussiana estándar multivariada así como datos con distribuciones probabilistas más generales. Posteriormente, en el Capítulo 2 exponemos una descripción de los métodos de clasificación binaria MD, SVM, DWD y MDP, tanto en un contexto general así como bajo la representación geométrica asintótica. Mostramos además dos importantes resultados acerca del comportamiento asintótico de estos cuatro clasificadores binarios, en términos de las probabilidades de clasificación correcta. En el Capítulo 3 mostramos cómo funcionan las metodologías OVO y OVR al considerar los hiperplanos separantes de MD, SVM, DWD y MDP, para predecir la etiqueta de la clase para un nuevo dato X. Exponemos aquí también el enfoque de resolver un solo problema de optimización (que envuelve a todas las clases en una misma) para extender un método de clasificación binaria al caso multicategoría. Finalizamos este capítulo proporcionando nuestros resultados teóricos acerca del comportamiento asintótico de los métodos de clasificación multicategoría OVO-MD, OVO-SVM, OVO-MDP, OVO-DWD y OVR-MD, en términos de las probabilidades de clasificación correcta. Finalmente, en el Capítulo 4 consideramos tres diferentes clases de datos gaussianos con d características para llevar a cabo un estudio de simulación (usando el software MATLAB) concerniente a estos últimos cinco métodos, y considerando adicionalmente los métodos de discriminación multicategoría OVR-SVM, OVR-MDP y OVR-DWD. Concluimos este último capítulo evaluando el desempeño de estos ocho clasificadores multicategoría al realizar un análisis estadístico de los 64 datos de cáncer de colon proporcionados por Khan et al. (2001), y disponibles en http://bioinf.ucd.ie/people/aedin/R/full_datasets/. Estos datos consisten de 2308 genes (características) para cada uno de los 64 pacientes (observacio-

nes) y son divididos en cuatro clases: 21 observaciones de rabdomiosarcoma, 23 observaciones de sarcoma de Ewing, 12 observaciones de neuroblastoma y 8 observaciones de linfoma de Burkitt. Realizamos un análisis exploratorio para determinar si estas cuatros clases cumplen con la representación geométrica y así observar si hay congruencia con nuestros resultados teóricos al obtener las tasas de error de clasificación de cada método.

Resultados

En lo concerniente a la metodología OVO proporcionamos condiciones suficientes para que las probabilidades de clasificación correcta converjan a uno cuando la dimensión de los datos tiende a infinito. Encontramos que, así como en el caso binario, OVO-MD, OVO-SVM y OVO-MDP tienen las misma propiedades asintóticas (Teorema 3.4.1), mientras que OVO-DWD podría tener propiedades asintóticas diferentes (Teorema 3.4.2). Damos estas condiciones en términos de las distancias asintóticas entre los datos y sus medias de clase, la distancia asintótica entre pares de medias de clase, y los tamaños muestrales de las clases, los cuales no se asumen necesariamente iguales. Respecto al comportamiento asintótico de OVR-MD, proporcionamos condiciones suficientes y necesarias para que un nuevo dato de una clase dada sea correctamente clasificado con probabilidad tendiendo a uno cuando la dimensión de los datos tiende a infinito (Teorema 3.4.4). Logramos esto, obteniendo primero el comportamiento asintótico de las distancias signadas de un nuevo dato de una clase dada a los hiperplanos separantes del método (Teorema 3.4.3). Obtenemos también varias consecuencias particulares de interés, incluyendo el caso cuando los tamaños muestrales de las clases son iguales. En las simulaciones obtuvimos que en el caso de OVO los métodos con mejor desempeño general (en cuanto a clasificar correctamente) son OVO-MD o OVO-DWD, mientras que el comportamiento de OVO-SVM y OVO-MDP es muy similar. En el caso de OVR obtuvimos análogamente al caso OVO que los métodos con mejor desempeño general son OVR-MD o OVR-DWD, mientras que el comportamiento de OVO-SVM y OVO-MDP es casi el mismo. En ambos casos estos resultados son congruentes con el desempeño de los métodos en el caso de clasificación binaria. Finalmente, en el análisis de datos reales obtuvimos que las cuatro clases poseen la representación geométrica asintótica razonablemente y que hay congruencia con nuestros resultados teóricos, pues las tasas de error de clasificación de los ocho métodos son cercanas a cero, habiendo un desempeño ligeramente mejor con la metodología OVR que con OVO.

Capítulo 1

Representación Geométrica Asintótica de Datos de Dimensión Alta

Debido a que los datos multivariados con dimensión mayor o igual al tamaño de la muestra han estado surgiendo de manera relevante en muchos campos de la ciencia, véase por ejemplo Johnstone (2001) y Wang (2012), se vio la necesidad de estudiar este tipo de datos, haciendo crecer la dimensión de los datos a infinito mientras que el tamaño de la muestra permanece fijo. Este análisis asintótico dio como resultado, bajo algunas condiciones de regularidad, la descripción de una estructura geométrica determinista y se le dio el nombre de representación geométrica asintótica de datos de dimensión alta. Esta representación dejó ver que esta clase de datos posee propiedades geométricas diferentes a aquellos datos clásicos donde la dimensión es menor que el tamaño muestral. Por ejemplo, en Hall et al. (2005) fue mostrado que los datos gaussianos estándar multivariados tienden a alejarse de la media poblacional a medida que aumenta la dimensión, lo cual parece paradójico ya que la densidad gaussiana univariada estándar tiene su media en cero.

En este capítulo se expone un desarrollo de la representación geométrica asintótica de datos de dimensión alta, el cual está basado en las referencias Hall et al. (2005), Ahn et al. (2007), Jung y Marron (2009), Qiao et al. (2010), y Yata y Aoshima (2012). Esto se aborda en primer lugar considerando datos gaussianos estándar multivariados para posteriormente considerar datos con distribución no necesariamente gaussiana. Los resultados y definiciones de la teoría de probabilidad y estadística multivariada empleados aquí pueden ser consultados en los capítulos 1 y 5 de Serfling (2002), y Casella y Berger (2002), respectivamente.

1.1. Representación Geométrica Asintótica Gaussiana Estándar

En la Sección 2 de Hall et al. (2005) puede verse que los datos gaussianos estándar de dimensión alta poseen la representación geométrica, y en lo que sigue se expone un desarrollo de los resultados dados allí. Sea $Z_j(d) = \left(Z_{_{1,j}}, Z_{_{2,j}}, ..., Z_{_{d,j}}\right)^\top$ un vector aleatorio d-multivariado de la distribución gaussiana con media 0 y matriz de covarianza $\mathbf{I}_d$, esto es, $Z_j(d) \sim \mathcal{N}_d\left(0, \mathbf{I}_d\right)$. Puesto que para cada r, la variable aleatoria $Z_{r,j}^2$ sigue una distribución chi-cuadrada con 1 grado de libertad y son independientes, entonces el teorema del límite central dice que

$$\frac{d^{1/2}}{2^{1/2}}\left[\left(\frac{1}{d}\sum_{r=1}^{d} Z_{r,j}^2\right) - 1\right] \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \quad \text{cuando} \quad d \to \infty,$$

donde " $\xrightarrow{\mathbb{D}}$ " *denotará la convergencia en distribución*. De manera que,

$$d^{1/2}\left[\left(\frac{1}{d}\sum_{r=1}^{d} Z_{r,j}^2\right) - 1\right] \xrightarrow{\mathbb{D}} \mathcal{N}(0,2) \quad \text{cuando} \quad d \to \infty.$$

Luego, por el método delta,

$$||Z_j(d)|| - d^{1/2} \xrightarrow{\mathbb{D}} \mathcal{N}(0, 1/2) \quad \text{cuando} \quad d \to \infty,$$

y por ende, $||Z_j(d)|| - d^{1/2}$ es una $O_p(1)$. Por lo cual se puede escribir

$$||Z_j(d)|| = d^{1/2} + O_p(1). \tag{1.1}$$

La ecuación (1.1) dice que, conforme la dimensión crece, el vector gaussiano estándar Z_j tiende a estar cerca de la superficie de una esfera explandible.

Una característica más sorprendente de este tipo de datos es revelada llevando acabo un análisis similar: si $Z_i(d), Z_j(d) \sim \mathcal{N}_d(0, \mathbf{I}_d)$ y son independientes, entonces

$$\frac{d^{1/2}}{2^{1/2}}\left\{\left[\frac{1}{2d}\sum_{r=1}^{d}\left(Z_{r,i} - Z_{r,j}\right)^2\right] - 1\right\} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \quad \text{cuando} \quad d \to \infty,$$

ya que $\frac{1}{2}\left(Z_{r,i} - Z_{r,j}\right)^2$ sigue una distribución chi-cuadrada con 1 grado de libertad. Por consiguiente,

$$d^{1/2}\left\{\left[\frac{1}{d}\sum_{r=1}^{d}\left(Z_{r,i} - Z_{r,j}\right)^2\right] - 2\right\} \xrightarrow{\mathbb{D}} \mathcal{N}(0,8) \quad \text{cuando} \quad d \to \infty,$$

y en consecuencia, por el método delta se tiene que

$$||Z_i(d) - Z_j(d)|| - (2d)^{1/2} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \quad \text{cuando} \quad d \to \infty.$$

Por lo que $||Z_i(d) - Z_j(d)|| - (2d)^{1/2}$ es una $O_p(1)$ y así

$$||Z_i(d) - Z_j(d)|| = (2d)^{1/2} + O_p(1). \tag{1.2}$$

Por lo tanto, (1.2) muestra que la distancia entre estos dos vectores gaussianos estándar independientes es aproximadamente constante cuando la dimensión es suficientemente grande.

Análogamente se ve que para el ángulo entre estos dos vectores se tiene que

$$\text{Ang}\,[Z_i(d), Z_j(d)] = \frac{\pi}{2} + O_p(d^{-1/2}). \tag{1.3}$$

En efecto, debido a que $\mathbb{E}[Z_{r,i}Z_{r,j}] = 0$ y $\mathbb{V}ar[Z_{r,i}Z_{r,j}] = 1$, entonces el teorema del límite central dice que

$$d^{1/2}\left(\frac{1}{d}\sum_{r=1}^{d} Z_{r,i}Z_{r,j}\right) = d^{1/2}\left(\frac{1}{d}\langle Z_i(d), Z_j(d)\rangle\right) \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \quad \text{cuando} \quad d \to \infty,$$

donde "$\langle , \rangle$" *denotará al producto punto en* $\mathbb{R}^d$. Así, por el método delta,

$$d^{1/2}\left[\arccos\left(d^{-1}\langle Z_i(d), Z_j(d)\rangle\right) - \arccos(0)\right] \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \quad \text{cuando} \quad d \to \infty,$$

o equivalentemente,

$$d^{1/2}\left(\text{Ang}\,[Z_i(d), Z_j(d)] - \pi/2\right) \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \quad \text{cuando} \quad d \to \infty.$$

En consecuencia, $\text{Ang}[Z_i(d), Z_j(d)] - \pi/2$ es una $O_p(d^{-1/2})$, lo cual implica la igualdad (1.3), diciendo así que el ángulo entre estos dos vectores es aproximadamente ortogonal cuando la dimensión es suficientemente grande.

En general, si se tienen n de estos vectores d-multivariados gaussianos estándar independientes, cuando la dimensión es lo suficientemente grande, todas las distancias entre pares son aproximadamente iguales y todos los ángulos entre pares de vectores son aproximadamente ortogonales. Debido a que todas las distancias entre pares tienden a ser iguales, los n vectores tienden entonces a ser los vértices de un *n-poliedro regular convexo*, es decir, un poliedro con n vértices y con aristas de la misma longitud. Este n-poliedro es llamado *n-simplex*.

Estas propiedades geométricas son ilustradas en la Figura 1 donde el caso $d = n = 3$ es considerado. Los rayos saliendo desde el origen a los puntos respectivos tienen longitud aproximadamente igual a $3^{1/2}$ y los ángulos entre cada par de éstos son casi perpendiculares. Centrando la atención en el *hiperplano generado por los datos*, de dimensión $n - 1 = 2$ en este caso, se observa que al ser todas las distancias entre cada par de puntos casi iguales, los datos esencialmente están en los vértices de un *triángulo equilatero*, el 3-*simplex* en este ejemplo.

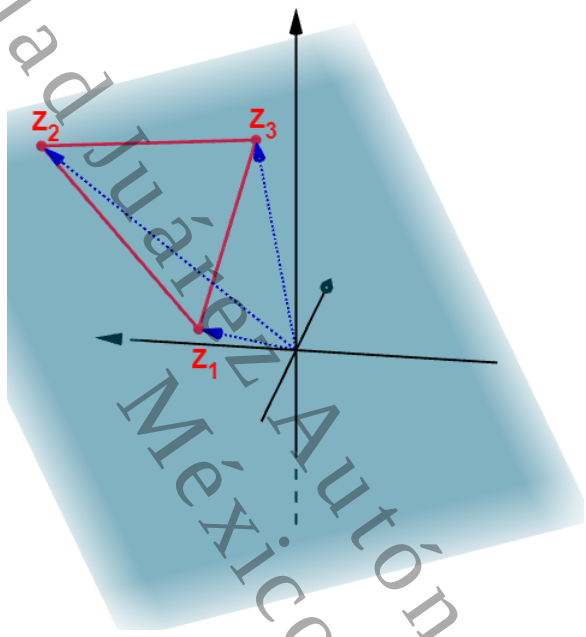


Figura 1: *Representación geométrica asintótica en el caso $d = n = 3$, mostrando el 3-simplex (triángulo equilatero) en el hiperplano generado por los datos.*

En la Figura 2 se proporciona otro ejemplo que aclara un poco más las ideas de la representación geométrica.

Cada panel exhibe diagramas de dispersión superpuestos de 10 muestras, etiquetadas con diferentes figuras, de vectores gaussianos estándar independientes de tamaño $n = 3$ y en dimensiones $d = 2, 20, 200, 20000$ en las figuras 2(a), 2(b), 2(c) y 2(d) respectivamente. Observe que las 10 muestras dan una impresión de la variación muestral, en función de la dimensión, la cual varía para cada uno de los páneles. Para cada muestra y cada dimensión, se encuentra el hiperplano generado por los datos, es decir, en este caso, el plano que es mostrado en la Figura 1, y los datos se proyectan sobre él. En vista de la ecuación (**1.2**) se espera que estos puntos estén cerca de los

vértices del triángulo equilátero, el 3-simplex en este caso, y que esta aproximación se apreciara mejor conforme la dimensión sea más grande. En efecto, la Figura 2 confirma estas conjeturas. Note que para $d = 2$ los puntos parecen ser bastante aleatorios y, de hecho, no todos son fáciles de asociar con el vértice apropiado del triángulo. Sin embargo, para $d = 20$ hay una convergencia razonable a los vértices, lo que sugiere que la representación geométrica ya es informativa. Para $d = 200$, la aproximación es bastante buena, lo que deja en claro que la mayor parte de la variabilidad corresponde a las dos rotaciones que se consideraron previamente. Como era de esperarse, el caso $d = 20000$ muestra una representación geométrica aún más rígida.

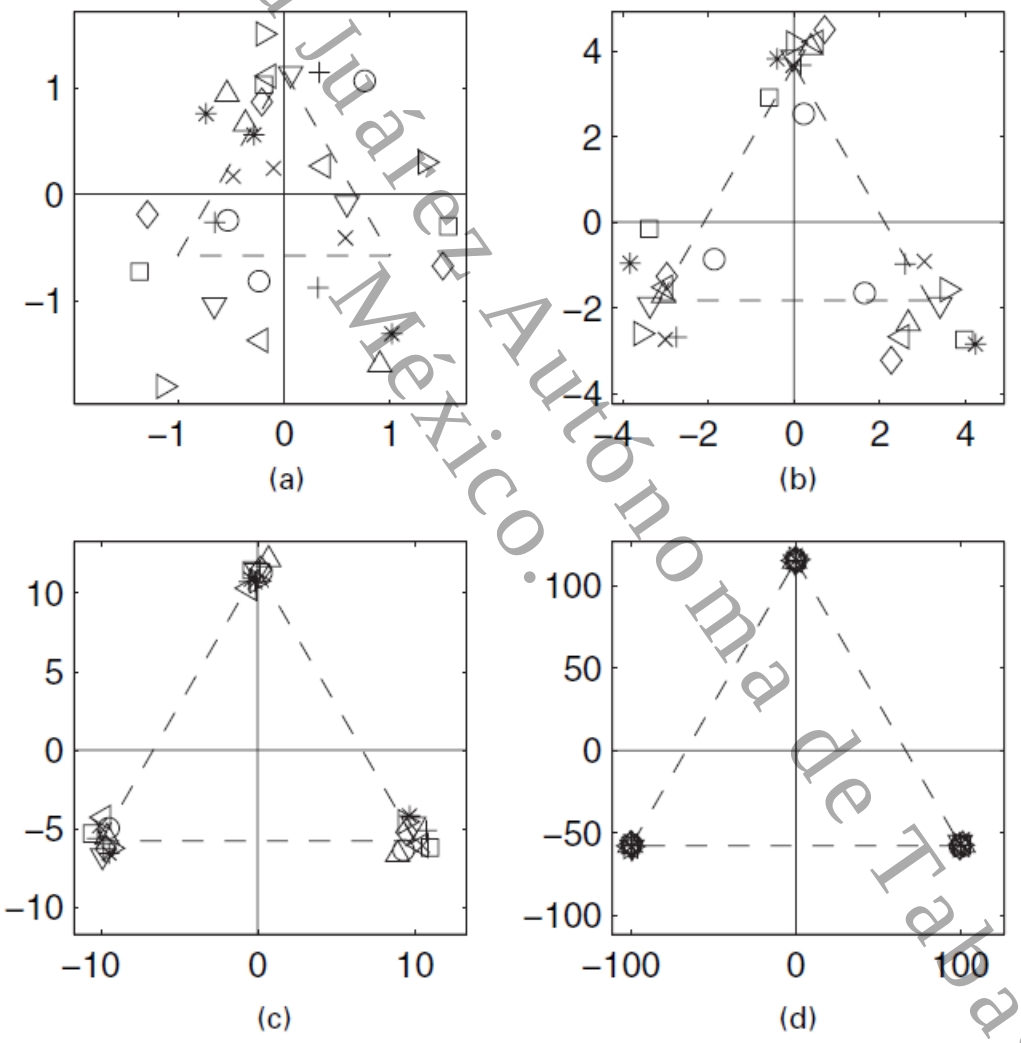


Figura 2: *Convergencia al 3-simplex conforme la dimensión aumenta, (a) $d = 2$; (b) $d = 20$; (c) $d = 200$; (d) $d = 20000$, Hall et al. (2005).*

1.2. Representación Geométrica Asintótica General

La representación geométrica asintótica de datos multivariados, cuando la dimensión crece indefinidamente mientras que el tamaño de muestra está fijo, ha sido estudiada por los autores Hall et al. (2005), Ahn et al. (2007), Jung y Marron (2009), Qiao et al. (2010), y Yata y Aoshima (2012). Estos autores mostraron condiciones bajo las cuales datos de dimensión alta tienden a estar determinísticamente en los vértices de un poliedro regular convexo, así como ocurre con los datos gaussianos estándar multivariados cuando la dimensión de los datos tiende a infinito. Esta estructura geométrica es usada para analizar el comportamiento de algunas metodologías estadísticas para datos multivariados en el contexto de dimensión alta, como se verá en capítulos posteriores de esta tesis.

1.2.1. Representación Geométrica Asintótica de una Clase

El trabajo pionero realizado por Hall et al. (2005) mostró que la estructura aproximada de n-simplex de los datos de dimensión alta gaussianos estándar puede ser observada para datos que siguen una distribución multivariada no necesariamente gaussiana, como se expone en el siguiente resultado.

Teorema 1.2.1. *Sea* $X_j^+(d) = \left(X_{1,j}^+, X_{2,j}^+, ..., X_{d,j}^+\right)^\top$ *un vector aleatorio obtenido al truncar una serie de tiempo infinita* $X^+ = \left(X_{1,j}^+, X_{2,j}^+, ...\right)^\top$, *y considere una clase o muestra aleatoria* $\mathcal{C}_+(d) = \left\{X_1^+(d), X_2^+(d), ..., X_n^+(d)\right\}$ *de vectores aleatorios independientes e idénticamente distribuidos con la misma distribución de* $X_j^+(d)$. *Suponga que:*

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

b) Para una constante $\sigma > 0$,

$$\frac{1}{d}\sum_{r=1}^{d} \mathbb{V}ar\left[X_{r,j}^+\right] \longrightarrow \sigma^2 \quad \textit{cuando} \quad d \to \infty. \tag{1.4}$$

c) La serie de tiempo X^+ *satisface la condición* ρ-mixing *para funciones que son dominadas por cuadráticas, en el sentido que si las funciones* f *y* g *de dos variables satisfacen* $|f(u,v)| + |g(u,v)| \leq Mu^2v^2$ *para* $M > 0$ *fija y todo* u *y* v, *se tiene que*

$$\sup_{1\leq r,s<\infty, |r-s|\geq t} \left|\mathbb{C}orr\left[f\left(U_{r,j}^+, V_{r,j}^+\right), g\left(U_{s,j}^+, V_{s,j}^+\right)\right]\right| \leq \rho(t), \tag{1.5}$$

siendo $(U,V)=(X,X)$ *o* $(U,V)=(X,\hat{X})$*, donde* $\hat{X}$ *es independiente y tiene la misma distribución que* X*, y la función* ρ *satisface* $\rho(t)\longrightarrow 0$ *cuando* $t\to\infty$*.*

Se sigue entonces que la distancia entre $X_i^+(d)$ *y* $X_j^+(d)$*, para* $i\neq j$*, es aproximadamente igual a* $(2d\sigma^2)^{1/2}$ *para* d *suficientemente grande, en el sentido que*

$$\frac{1}{d^{1/2}}\left|\left|X_i^+-X_j^+\right|\right|\xrightarrow{\mathbb{P}}(2\sigma^2)^{1/2}\quad \textit{cuando}\ \ d\to\infty. \tag{1.6}$$

Demostración. Considere los vectores de datos $X_i^+(d):=X_i^+=\left(X_{1,i}^+,X_{2,i}^+,...,X_{d,i}^+\right)^\top$ y $X_j^+(d):=X_j^+=\left(X_{1,j}^+,X_{2,j}^+,...,X_{d,j}^+\right)^\top$, para $i,j\in\{1,2,...,n\}$ fijos. Observe que

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{r=1}^{d}\left\{\left(X_{r,i}^+-X_{r,j}^+\right)^2-\mathbb{E}\left[\left(X_{r,i}^+-X_{r,j}^+\right)^2\right]\right\}\right)^2\right]\\
=\ &\mathbb{E}\left[\left(\sum_{r=1}^{d}\left\{\left(X_{r,i}^+-X_{r,j}^+\right)^2-\mathbb{E}\left[\left(X_{r,i}^+\right)^2\right]+2\mathbb{E}\left[X_{r,i}^+\right]\mathbb{E}\left[X_{r,j}^+\right]-\mathbb{E}\left[\left(X_{r,j}^+\right)^2\right]\right\}\right)^2\right]\\
=\ &\mathbb{E}\left[\left(\sum_{r=1}^{d}\left(X_{r,i}^+-X_{r,j}^+\right)^2-2\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^+\right]\right)^2\right]\\
=\ &\mathbb{E}\left[\left(\sum_{r=1}^{d}\left(X_{r,i}^+-X_{r,j}^+\right)^2\right)^2\right]-4\left\{\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^+\right]\right\}\left\{\sum_{r=1}^{d}\mathbb{E}\left[\left(X_{r,i}^+-X_{r,j}^+\right)^2\right]\right\}\\
+\ &4\left\{\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^+\right]\right\}^2\\
=\ &\mathbb{E}\left[\left(\sum_{r=1}^{d}\left(X_{r,i}^+-X_{r,j}^+\right)^2\right)^2\right]-8\left\{\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^+\right]\right\}^2+4\left\{\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^+\right]\right\}^2\\
=\ &\sum_{r=1}^{d}\mathbb{E}\left[\left(X_{r,i}^+-X_{r,j}^+\right)^4\right]+2\sum_{r<s}\mathbb{E}\left[\left(X_{r,i}^+-X_{r,j}^+\right)^2\left(X_{s,i}^+-X_{s,j}^+\right)^2\right]-4\left\{\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^+\right]\right\}^2\\
=\ &\sum_{r=1}^{d}\mathbb{E}\left[\left(X_{r,i}^+-X_{r,j}^+\right)^4\right]+2\sum_{r<s}\mathbb{C}ov\left[\left(X_{r,i}^+-X_{r,j}^+\right)^2,\left(X_{s,i}^+-X_{s,j}^+\right)^2\right]\\
+\ &2\sum_{r<s}\mathbb{E}\left[\left(X_{r,i}^+-X_{r,j}^+\right)^2\right]\mathbb{E}\left[\left(X_{s,i}^+-X_{s,j}^+\right)^2\right]-4\left\{\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^+\right]\right\}^2\\
=\ &\sum_{r=1}^{d}\mathbb{E}\left[\left(X_{r,i}^+-X_{r,j}^+\right)^4\right]+2\sum_{r<s,|r-s|<t}\mathbb{C}ov\left[\left(X_{r,i}^+-X_{r,j}^+\right)^2,\left(X_{s,i}^+-X_{s,j}^+\right)^2\right]
\end{aligned}$$

$$+ \quad 2 \sum_{r<s,|r-s|\geq t} \mathbb{C}orr\left[\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2},\left(X_{s,i}^{+}-X_{s,j}^{+}\right)^{2}\right]\left\{\mathbb{V}ar\left[\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}\right]\right\}^{1/2}$$

$$\times \quad \left\{\mathbb{V}ar\left[\left(X_{s,i}^{+}-X_{s,j}^{+}\right)^{2}\right]\right\}^{1/2}-4\sum_{r=1}^{d}\mathbb{V}ar^{2}\left[X_{r,i}^{+}\right]. \tag{1.7}$$

El supuesto a) permite elegir constantes positivas M_1, M_2 y M_3 tales que la expresión (1.7) resulta ser menor que

$$\sum_{r=1}^{d} M_1 + 2\left(\sum_{r<s,|r-s|<t} M_2\right) + 2M_3 \sum_{r<s,|r-s|\geq t} \mathbb{C}orr\left[\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2},\left(X_{s,i}^{+}-X_{s,j}^{+}\right)^{2}\right]$$

$$\leq \quad dM_1 + 2M_2(t-1)d + M_3(d-1)d\rho(t), \tag{1.8}$$

donde se ha hecho uso de (1.5), la condición ρ-mixing, y es por esta misma condición que la suma (1.8) dividida por d^2 se puede hacer tan pequeña como se quiera tomando t suficientemente grande y posteriormente d suficientemente grande para así obtener que

$$\frac{1}{d^2}\mathbb{E}\left[\left(\sum_{r=1}^{d}\left\{\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}-\mathbb{E}\left[\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}\right]\right\}\right)^{2}\right] \longrightarrow 0 \quad \text{cuando} \quad d\to\infty.$$

De este modo, al aplicar la desigualdad de Chebyshev se obtiene como resultado que

$$\frac{1}{d}\left(\sum_{r=1}^{d}\left\{\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}-\mathbb{E}\left[\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}\right]\right\}\right)\xrightarrow{\mathbb{P}} 0 \quad \text{cuando} \quad d\to\infty, \tag{1.9}$$

donde " $\xrightarrow{\mathbb{P}}$ " *denotará la convergencia en probabilidad.*

Las convergencias dadas en (1.4) y (1.9) implican la propiedad geométrica (1.6), ya que

$$\frac{1}{d}\sum_{r=1}^{d}\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}-\frac{1}{d}\sum_{r=1}^{d}\mathbb{E}\left[\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}\right]\xrightarrow{\mathbb{P}}0, \quad \text{cuando} \quad d\to\infty,$$

$$\text{es decir,} \quad \frac{1}{d}\sum_{r=1}^{d}\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}-\frac{2}{d}\sum_{r=1}^{d}\mathbb{V}ar\left[X_{r,i}^{+}\right]\xrightarrow{\mathbb{P}}0, \quad \text{cuando} \quad d\to\infty,$$

$$\text{por lo que,} \quad \frac{1}{d^{1/2}}\left\{\sum_{r=1}^{d}\left(X_{r,i}^{+}-X_{r,j}^{+}\right)^{2}\right\}^{1/2}\xrightarrow{\mathbb{P}}\left(2\sigma^{2}\right)^{1/2}, \quad \text{cuando} \quad d\to\infty,$$

demostrando así que (1.6) se cumple. ▲

Un análisis similar a éste revela que también es cierto que

$$\frac{1}{d}\left|\left|X_{i}^{+}-\mathbb{E}[X_{i}^{+}]\right|\right|^{2}\xrightarrow{\mathbb{P}}\sigma^{2} \quad \text{cuando} \quad d\to\infty. \tag{1.10}$$

Por lo tanto, la estructura asintótica n-simplex de los datos gaussianos estándar se puede describir también para datos de dimensión alta que satisfagan las condiciones a)-c) del Teorema 1.2.1. Específicamente, si se trabaja en el espacio $(n-1)$-dimensional, en el cual todos los datos de la clase $\mathcal{C}_+(d)$ pueden ser proyectados sin perder las relaciones intrínsecas entre ellos, entonces la convergencia en probabilidad (**1.6**) revela que, después de reescalar por $d^{-1/2}$, los datos $X_i^+(d)$ están *asintóticamente localizados en los vértices de un poliedro regular convexo* (n-simplex) de aristas de longitud $(2\sigma^2)^{1/2}$. Esto es ilustrado en la Figura 3 considerando el caso en que $n = 4$.

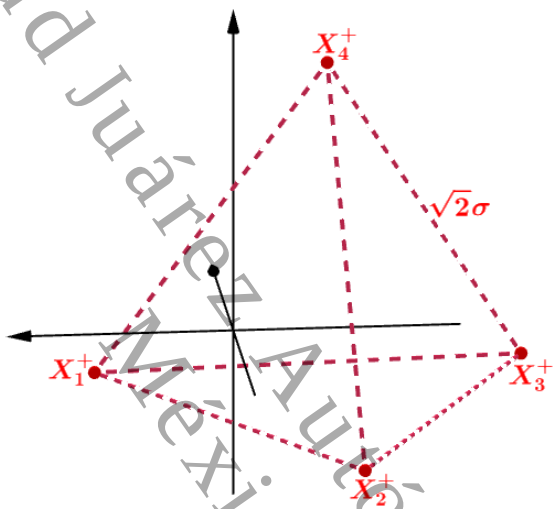


Figura 3: *Representación geométrica asintótica en el caso $n = 4$, mostrando el 4-simplex (tetraedro regular) en el espacio 3-dimensional.*

Representación Geométrica Asintótica bajo Condiciones más Débiles

Puede observarse que el supuesto c) del Teorema 1.2.1 es poco atractivo debido a que éste exige que las variables componentes, consideradas en una serie de tiempo, sean *casi independientes,* ya que deben satisfacer una condición ρ-mixing. Esta condición es también estricta porque es común tener colinealidad fuerte entre las variables componentes, y además, esta condición depende del orden de las variables, el cual puede ser arbitrario en muchas aplicaciones. En Ahn et al. (2007), Jung y Marron (2009), Qiao et al. (2010), y Yata y Aoshima (2012) es demostrado que la representación geométrica de datos de dimensión alta puede ser observada bajo condiciones más débiles, lo cual se presenta en este apartado. Antes de exponer esto, se definen enseguida algunos términos que serán utilizados en el resto de este capítulo.

Sea $\mathcal{C}_+(d) = \{X_1^+(d), X_2^+(d), ..., X_n^+(d)\}$ una clase o muestra con n vectores aleatorios independientes e idénticamente distribuidos de una distribución multivariada d-dimensional con matriz de covarianza *positiva definida* $\mathbf{\Sigma}_d^+$. Asuma sin pérdida de generalidad que cada vector $X_j^+(d) := X_j^+ = \left(X_{1,j}^+, X_{2,j}^+, ..., X_{d,j}^+\right)^\top$ tiene *media cero*, y suponga además que la descomposición en valores propios de $\mathbf{\Sigma}_d^+$ es $\mathbf{\Sigma}_d^+ = \mathbf{V}_d^+ \ \mathbf{\Lambda}_d^+ \ \left(\mathbf{V}_d^+\right)^\top$, donde $\mathbf{\Lambda}_d^+$ es la matriz diagonal de valores propios $\lambda_{1,d}^+ \geq \lambda_{2,d}^+ \geq \cdots \geq \lambda_{d,d}^+ > 0$ y $\mathbf{V}_d^+$ es la matriz de vectores propios correspondientes.

Considere la matriz $\mathbf{X}_d^+ = \left[X_1^+, X_2^+, ..., X_n^+\right]$ de tamaño $d \times n$ con $d > n$ (n siempre fijo). Como se está suponiendo que la media poblacional es el vector cero, entonces la *matriz de covarianza muestral* de $\mathbf{X}_d^+$ resulta ser la matriz de $d \times d$ dada por $\mathbf{S}_d^+ = \dfrac{1}{n}\mathbf{X}_d^+ \ \left(\mathbf{X}_d^+\right)^\top$, y la *matriz de covarianza muestral dual* es la matriz de $n \times n$ definida como $\mathbf{S}_{D,d}^+ = \dfrac{1}{d}\left(\mathbf{X}_d^+\right)^\top \mathbf{X}_d^+$.

Una *matriz factor*, que es esencialmente la raíz cuadrada de $\mathbf{\Sigma}_d^+$, es definida como la matrix $\mathbf{Q}_d^+ = \mathbf{V}_d^+ \left(\mathbf{\Lambda}_d^+\right)^{1/2}$; de manera que $\mathbf{\Sigma}_d^+ = \mathbf{Q}_d^+ \left(\mathbf{Q}_d^+\right)^\top$. Debido a que $\mathbf{\Sigma}_d^+$ es *positiva definida*, se puede escribir $\mathbf{X}_d^+ = \mathbf{Q}_d^+ \ \mathbf{Z}_d^+$, donde $\mathbf{Z}_d^+ = \left(\mathbf{\Lambda}_d^+\right)^{-1/2} \left(\mathbf{V}_d^+\right)^\top \mathbf{X}_d^+$ es una matriz aleatoria $d \times n$ de una distribución multivariada d-dimensional con media cero y matriz de covarianza $\mathbf{I}_d$. Se puede ver que si las columnas de $\mathbf{X}_d^+$ son gaussianas, entonces los elementos de $\mathbf{Z}_d^+$ son variables gaussianas estándar independientes. Debido a que $\left(\mathbf{V}_d^+\right)^\top \mathbf{V}_d^+$ es la matrix identidad de $d \times d$, se tiene que

$$d\mathbf{S}_{D,d}^+ = \left(\mathbf{Z}_d^+\right)^\top \left(\mathbf{Q}_d^+\right)^\top \mathbf{Q}_d^+ \ \mathbf{Z}_d^+ = \left(\mathbf{Z}_d^+\right)^\top \mathbf{\Lambda}_d^+ \ \mathbf{Z}_d^+ = \sum_{r=1}^{d} \lambda_{r,d}^+ \ \mathbf{W}_{r,d}^+, \tag{1.11}$$

con $\mathbf{W}_{r,d}^+ = \left[Z_r^+(d)\right]^\top Z_r^+(d)$, donde $Z_r^+(d) = \left(Z_{r,1}^+, Z_{r,2}^+, ..., Z_{r,n}^+\right)$, $r = 1, 2, ..., d$, son los vectores renglones de $\mathbf{Z}_d^+$. Observe que si $\mathbf{X}_d^+$ es gaussiana, entonces las matrices $\mathbf{W}_{r,d}^+$ son independientes y tienen una distribución Wishart $\mathcal{W}_n(1, \mathbf{I}_n)$.

Con estos supuestos y términos así definidos se tiene a continuación un resultado acerca del comportamiento asintótico de los valores propios muestrales.

Teorema 1.2.2. *Para n fijo, sea $\mathbf{X}_1^+$, $\mathbf{X}_2^+$,...,$\mathbf{X}_d^+$,... una sucesión de matrices aleatorias de $d \times n$, donde cada $\mathbf{X}_d^+$ proviene de una distribución multivariada d-dimensional con matrix de covarianza $\mathbf{\Sigma}_d^+$. Sean $\lambda_{1,d}^+ \geq \lambda_{2,d}^+ \geq \cdots \geq \lambda_{d,d}^+$ los valores propios de $\mathbf{\Sigma}_d^+$, y sea $\mathbf{S}_{D,d}^+$ la matriz de covarianza muestral dual correspondiente de $n \times n$. Asuma lo siguiente:*

a) Cada columna de $\mathbf{X}_d^+$ tiene media cero y matrix de covarianza positiva definida $\mathbf{\Sigma}_d^+$.

b) Los cuartos momentos de las entradas de las columnas de $\mathbf{X}_d^+$ *están uniformemente acotados.*

c) La representación dada en (1.11) se cumple para cada $\mathbf{X}_d^+$.

d) Las entradas de $\boldsymbol{Z}_d^+ = \left(\boldsymbol{\Lambda}_d^+\right)^{-1/2}\left(\mathbf{V}_d^+\right)^\top \boldsymbol{X}_d^+$ *son variables aleatorias independientes.*

e) Los valores propios de $\boldsymbol{\Sigma}_d^+$ *son* suficientemente difusos*, en el sentido de que*

$$\frac{\sum_{r=1}^{d}\left(\lambda_{r,d}^+\right)^2}{\left(\sum_{s=1}^{d}\lambda_{s,d}^+\right)^2} \longrightarrow 0 \quad \textit{cuando} \quad d\to\infty. \tag{1.12}$$

Se sigue entonces que los valores propios muestrales se comportan como si vinieran de la matriz de covarianza identidad, en el sentido que, con $\sigma_d^2 := \dfrac{1}{d}\displaystyle\sum_{r=1}^{d}\lambda_{r,d}^+$,

$$\frac{1}{\sigma_d^2}\mathbf{S}_{D,d}^+ \xrightarrow{\mathbb{P}} \mathbf{I}_n \quad \textit{cuando} \quad d\to\infty. \tag{1.13}$$

Demostración. Debido a (1.11) se tiene que si $Z_r^+(d) = \left(Z_{r,1}^+, Z_{r,2}^+, ..., Z_{r,n}^+\right)$, entonces el j-ésimo elemento de la diagonal de la matriz $\dfrac{1}{\sigma_d^2}\mathbf{S}_{D,d}^+$ puede ser expresado como $\displaystyle\sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\left(Z_{r,j}^+\right)^2$, donde las variables aleatorias $Z_{r,j}^+$, $r=1,2,...,d$, son independientes con media cero y varianza uno y, $\widehat{\lambda}_{r,d}^+ = \lambda_{r,d}^+ / \displaystyle\sum_{s=1}^{d}\lambda_{s,d}^+$ son los *valores propios relativos*. Note que con este último término la condición de esfericidad (1.12) es equivalente a que $\displaystyle\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^+\right)^2 \longrightarrow 0$ cuando $d\to\infty$, y además, se tiene que

$$\mathbb{E}\left[\sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\left(Z_{r,j}^+\right)^2\right] = \sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\,\mathbb{E}\left[\left(Z_{r,j}^+\right)^2\right] = \sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\,\mathbb{V}ar\left[Z_{r,j}^+\right] = \sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+ = 1,$$

lo cual implica que $\mathbb{V}ar\left[\displaystyle\sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\left(Z_{r,j}^+\right)^2\right] = \mathbb{E}\left[\left(\displaystyle\sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\left(Z_{r,j}^+\right)^2 - 1\right)^2\right]$. De este modo, aplicando la Desigualdad de Chebyshev y haciendo uso de la cota uniforme, denotada por M, de los cuartos momentos, se tiene que para cualquier $\epsilon>0$,

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\left(Z_{r,j}^+\right)^2 - 1\right| > \epsilon\right] &\leq \epsilon^{-2}\mathbb{E}\left[\left(\sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+\left(Z_{r,j}^+\right)^2 - 1\right)^2\right] \\
&= \epsilon^{-2}\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^+\right)^2\mathbb{V}ar\left[\left(Z_{r,j}^+\right)^2\right] \\
&\leq M\epsilon^{-2}\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^+\right)^2 \longrightarrow 0 \quad \text{cuando} \quad d\to\infty.
\end{aligned}$$

Así, los elementos de la diagonal de $\frac{1}{\sigma_d^2}\mathbf{S}_{D,d}^{+}$ convergen a 1 en probabilidad cuando $d \to \infty$.

Ahora, nuevamente por la igualdad (1.11) se ve que el (i,j)-ésimo elemento fuera de la diagonal de $\frac{1}{\sigma_d^2}\mathbf{S}_{D,d}^{+}$ es expresado como $\sum_{r=1}^{d} \widehat{\lambda}_{r,d}^{+}\, Z_{r,i}^{+}\, Z_{r,j}^{+}$, donde $Z_{r,i}^{+}$ y $Z_{r,j}^{+}$, $r = 1, 2, ..., d$, son independientes con media cero y varianza uno. De manera que $\mathbb{E}\left[\sum_{r=1}^{d} \widehat{\lambda}_{r,d}^{+}\, Z_{r,i}^{+}\, Z_{r,j}^{+}\right] = 0$, y por consiguiente, $\mathbb{V}ar\left[\sum_{r=1}^{d} \widehat{\lambda}_{r,d}^{+}\, Z_{r,i}^{+}\, Z_{r,j}^{+}\right] = \mathbb{E}\left[\left(\sum_{r=1}^{d} \widehat{\lambda}_{r,d}^{+}\, Z_{r,i}^{+}\, Z_{r,j}^{+}\right)^2\right]$. Luego, considerando que $\mathbb{C}ov\left[Z_{r,i}^{+}, Z_{r,j}^{+}\right] = 0$, la desigualdad de Chebyshev dice que para cualquier $\epsilon > 0$

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{r=1}^{d} \widehat{\lambda}_{r,d}^{+}\, Z_{r,i}^{+}\, Z_{r,j}^{+}\right| > \epsilon\right] &\leq \epsilon^{-2}\mathbb{E}\left[\left(\sum_{r=1}^{d} \widehat{\lambda}_{r,d}^{+}\, Z_{r,i}^{+}\, Z_{r,j}^{+}\right)^2\right] \\
&= \epsilon^{-2}\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^{+}\right)^2 \mathbb{V}ar\left[Z_{r,i}^{+}\, Z_{r,j}^{+}\right] \\
&= \epsilon^{-2}\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^{+}\right)^2 \mathbb{V}ar\left[Z_{r,i}^{+}\right]\mathbb{V}ar\left[Z_{r,j}^{+}\right] \longrightarrow 0 \quad \text{cuando} \quad d \to \infty.
\end{aligned}$$

De modo que los elementos fuera de la diagonal de $\frac{1}{\sigma_d^2}\mathbf{S}_{D,d}^{+}$ convergen a 0 en probabilidad cuando $d \to \infty$, concluyendo así que la convergencia en probabilidad dada en (1.13) es válida. ▲

Una consecuencia del Teorema 1.2.2 es la característica principal que describe a la estructura geométrica n-simplex, y se expone en el resultado siguiente.

Teorema 1.2.3. *Considere la clase $\mathcal{C}_{+}(d) = \{X_1^{+}(d), X_2^{+}(d), ..., X_n^{+}(d)\}$. Bajo las condiciones a), b), d) y e) del Teorema 1.2.2, la distancia al cuadrado entre $X_i^{+}(d)$ y $X_j^{+}(d)$, para $i \neq j$, es aproximadamente igual a $2\sum_{r=1}^{d} \lambda_{r,d}^{+}$ cuando d es suficientemente grande, en el sentido que*

$$\frac{1}{d\sigma_d^2}\left|\left|X_i^{+} - X_j^{+}\right|\right|^2 = \frac{\left|\left|X_i^{+} - X_j^{+}\right|\right|^2}{\sum_{r=1}^{d} \lambda_{r,d}^{+}} \xrightarrow{\mathbb{P}} 2 \quad \textit{cuando} \quad d \to \infty. \tag{1.14}$$

Demostración. Recuerde que $X_j^{+}(d) := X_j^{+} = \left(X_{1,j}^{+}, X_{2,j}^{+}, ..., X_{d,j}^{+}\right)^{\top}$. Teniendo en cuenta que $\mathbf{S}_{D,d}^{+} = \frac{1}{d}\left(\mathbf{X}_d^{+}\right)^{\top}\mathbf{X}_d^{+}$ y notando el hecho de que

$$\frac{\left|\left|X_i^{+} - X_j^{+}\right|\right|^2}{\sum_{r=1}^{d} \lambda_{r,d}^{+}} = \frac{\sum_{r=1}^{d}\left(X_{r,i}^{+}\right)^2}{\sum_{r=1}^{d} \lambda_{r,d}^{+}} + \frac{\sum_{r=1}^{d}\left(X_{r,j}^{+}\right)^2}{\sum_{r=1}^{d} \lambda_{r,d}^{+}} - \frac{2\sum_{r=1}^{d} X_{r,i}^{+}X_{r,j}^{+}}{\sum_{r=1}^{d} \lambda_{r,d}^{+}}, \tag{1.15}$$

se ve que los dos primeros sumandos de la igualdad dada en (1.15) son precisamente las entradas i-ésima y j-ésima en la diagonal de $\frac{1}{\sigma_d^2}\mathbf{S}_{D,d}^{+}$, y el tercer sumando es la (i,j)-ésima entrada de $\frac{1}{\sigma_d^2}\mathbf{S}_{D,d}^{+}$. Así, el Teorema 1.2.2 asegura que la convergencia dada en (1.14) se cumple. ▲

Observe también que del primer sumando de la ecuación dada en (1.15) y, al aplicar el Teorema 1.2.2, se obtiene que

$$\frac{\left|\left|X_i^{+}-0\right|\right|^2}{\sum_{r=1}^{d}\lambda_{r,d}^{+}} \xrightarrow{\mathbb{P}} 1 \quad \text{cuando} \quad d\to\infty. \tag{1.16}$$

Por lo tanto, los n datos tienden a formar un n-simplex cuando la dimensión d crece. En Ahn et al. (2007) se puede ver que la condición e) del Teorema 1.2.2 es más débil que la condición ρ-mixing c) del Teorema 1.2.1. En Jung y Marron (2009) es mostrado que la condición d) del Teorema 1.2.2 puede relajarse suponiendo que las entradas de $\mathbf{Z}_d^{+}$ son ρ-mixing bajo alguna permutación; no obstante, esta condición es aún estricta. Por los resultados de Yata y Aoshima (2012), las convergencias obtenidas en (1.14) y (1.16) se obtienen bajo las condiciones a), b), e) del Teorema 1.2.2 y la nueva condición

$$\frac{\sum_{r,s=1}^{d}\lambda_{r,d}^{+}\,\lambda_{s,d}^{+}\,\mathbb{E}\left[\left\{\left(Z_{r,1}^{+}\right)^2-1\right\}\left\{\left(Z_{s,1}^{+}\right)^2-1\right\}\right]}{Traza\left[\left(\mathbf{\Sigma}_d^{+}\right)^2\right]} \longrightarrow 0 \quad \text{cuando} \quad d\to\infty. \tag{1.17}$$

En Yata y Aoshima (2012) se menciona que (1.17) es más débil que la condición d) del Teorema 1.2.2, o que la condición ρ-mixing en las entradas de $\mathbf{Z}_d^{+}$, debido a que (1.17) se satisface bajo las condiciones d) y e) del Teorema 1.2.2.

1.2.2. Representación Geométrica Asintótica de dos Clases Independientes

En este último apartado de este capítulo se describe una estructura geométrica asintótica que involucra, en cierto sentido, a los dos *simpleces* de dos muestras independientes. Dicha estructura es esencial y de gran utilidad en el análisis de algunos métodos de discriminación binaria, así como también en el análisis de algunas estrategias de clasificación multicategoría en el contexto de datos de dimensión alta. El desarrollo que aquí se expone está basado principalmente de Qiao et al. (2010).

Considere una clase o muestra aleatoria $\mathcal{C}_{-}(d)=\{X_1^{-}(d),X_2^{-}(d),...,X_m^{-}(d)\}$ definida similarmente e independiente a la muestra $\mathcal{C}_{+}(d)=\{X_1^{+}(d),X_2^{+}(d),...,X_n^{+}(d)\}$, dada en la sub-

subsección anterior. En particular, el promedio de los valores propios de la matriz de covarianza positiva definida $\mathbf{\Sigma}_d^-$, queda definido como

$$\tau_d^2 := \frac{1}{d}\sum_{r=1}^{d}\lambda_{r,d}^-.$$

Asuma además que las condiciones correspondientes, dadas en el Teorema 1.2.2, se cumplen para $\mathcal{C}_-(d)$; y por consiguiente, que esta clase posee la estructura geométrica m-simplex:

$$\frac{1}{d\tau_d^2}\left|\left|X_i^- - X_j^-\right|\right|^2 = \frac{\left|\left|X_i^- - X_j^-\right|\right|^2}{\sum_{r=1}^{d}\lambda_{r,d}^-} \xrightarrow{\mathbb{P}} 2 \quad \text{cuando} \quad d\to\infty. \tag{1.18}$$

Ahora, generalice las muestras $\mathcal{C}_+(d)$ y $\mathcal{C}_-(d)$ en el sentido de que las medias poblacionales

$$\mathbb{E}\left[X_i^+(d)\right] := \left(\mathbb{E}\left[X_{1,i}^+\right], \mathbb{E}\left[X_{2,i}^+\right], ..., \mathbb{E}\left[X_{d,i}^+\right]\right)^\top$$

y

$$\mathbb{E}\left[X_j^-(d)\right] := \left(\mathbb{E}\left[X_{1,j}^-\right], \mathbb{E}\left[X_{2,j}^-\right], ..., \mathbb{E}\left[X_{d,j}^-\right]\right)^\top$$

no sean necesariamente cero y que no sean necesariamente iguales, pero que satisfagan que para una constante $c>0$,

$$\frac{1}{d}\left|\left|\mathbb{E}\left[X_i^+(d)\right] - \mathbb{E}\left[X_j^-(d)\right]\right|\right|^2 \longrightarrow c^2 \quad \text{cuando} \quad d\to\infty. \tag{1.19}$$

Suponga además que el límite de los promedios de los valores propios existe, esto es, existen constantes $\sigma>0$ y $\tau>0$ tales que

$$\sigma_d^2 \longrightarrow \sigma^2 \qquad \text{y} \qquad \tau_d^2 \longrightarrow \tau^2, \quad \text{cuando} \quad d\to\infty. \tag{1.20}$$

Antes de presentar el resultado principal y esencial de este primer capítulo, se expone enseguida un lema para demostrar dicho resultado.

Lema 1.2.1. *Suponga que* $\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^+\right)^2 \longrightarrow 0$ *y* $\sum_{s=1}^{d}\left(\widehat{\lambda}_{s,d}^-\right)^2 \longrightarrow 0$, *cuando* $d\to\infty$, *y además que* $\sum_{r=1}^{d}\widehat{\lambda}_{r,d}^+ = \sum_{s=1}^{d}\widehat{\lambda}_{s,d}^- = 1$. *Si* $\mathbf{U} = [U_{r,s}]$ *es una matriz ortogonal de tamaño* $d\times d$, *entonces* $\sum_{r=1}^{d}\sum_{s=1}^{d}\widehat{\lambda}_{r,d}^+\,\widehat{\lambda}_{s,d}^-\,U_{r,s}^2 \longrightarrow 0$ *cuando* $d\to\infty$.

Demostración. Dado que $\mathbf{U}$ es ortogonal, $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_d = \mathbf{U}^\top\mathbf{U}$, y así se deduce que $\sum_{r=1}^{d} U_{r,s}^2 = 1$ para cualquier $s \in \{1, 2, ..., d\}$ fija y $\sum_{s=1}^{d} U_{r,s}^2 = 1$ para cualquier $r \in \{1, 2, ..., d\}$ fijo. Observe que por la desigualdad de Cauchy-Schwarz

$$\begin{aligned}\left|\sum_{r=1}^{d}\sum_{s=1}^{d}\widehat{\lambda}_{r,d}^{+}\,\widehat{\lambda}_{s,d}^{-}\,U_{r,s}^2\right| &= \left|\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^{+}\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^{-}\,U_{r,s}^2\right)\right| \\ &\leq \left[\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^{+}\right)^2\right]^{1/2}\left[\sum_{r=1}^{d}\left(\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^{-}\,U_{r,s}^2\right)^2\right]^{1/2} \\ &\leq \left[\sum_{r=1}^{d}\left(\widehat{\lambda}_{r,d}^{+}\right)^2\right]^{1/2}, \qquad (1.21)\end{aligned}$$

debido a que $\sum_{r=1}^{d}\left(\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^{-}\,U_{r,s}^2\right)^2 \leq 1$. En efecto,

$$\begin{aligned}\sum_{r=1}^{d}\left(\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^{-}\,U_{r,s}^2\right)^2 &= \sum_{r=1}^{d}\left(\sum_{s=1}^{d}\left(\widehat{\lambda}_{s,d}^{-}\right)^2 U_{r,s}^4 + 2\sum_{s<t}\widehat{\lambda}_{s,d}^{-}\,\widehat{\lambda}_{t,d}^{-}\,U_{r,s}^2\,U_{r,t}^2\right) \\ &= \sum_{s=1}^{d}\left[\left(\widehat{\lambda}_{s,d}^{-}\right)^2\sum_{r=1}^{d}U_{r,s}^4\right] + 2\sum_{s<t}\left[\widehat{\lambda}_{s,d}^{-}\,\widehat{\lambda}_{t,d}^{-}\sum_{r=1}^{d}U_{r,s}^2\,U_{r,t}^2\right] \\ &\leq \sum_{s=1}^{d}\left(\widehat{\lambda}_{s,d}^{-}\right)^2 + 2\sum_{s<t}\widehat{\lambda}_{s,d}^{-}\,\widehat{\lambda}_{t,d}^{-} = \left(\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^{-}\right)^2 = 1,\end{aligned}$$

puesto que $\sum_{r=1}^{d} U_{r,s}^2 = 1$ implica que $\sum_{r=1}^{d} U_{r,s}^4 \leq 1$ y $\sum_{r=1}^{d} U_{r,s}^2\,U_{r,t}^2 \leq 1$. Por lo tanto, la convergencia deseada se obtiene de (1.21) haciendo $d \to \infty$. ▲

Se tienen así los elementos para exponer el siguiente resultado, que describe la representación geométrica entre dos clases independientes.

Teorema 1.2.4. *Asuma que las dos clases independientes* $\mathcal{C}_+(d) = \{X_1^+(d), X_2^+(d), ..., X_n^+(d)\}$ *y* $\mathcal{C}_-(d) = \{X_1^-(d), X_2^-(d), ..., X_m^-(d)\}$ *satisfacen las condiciones correspondientes dadas en el Teorema 1.2.2, y que cumplen además las condiciones dadas en (1.19) y (1.20). Entonces, para cualesquiera* (i, j)*, la distancia al cuadrado entre* X_i^+ *y* X_j^-*, dividida por* d*, converge en probabilidad a* $\ell^2 := \sigma^2 + \tau^2 + c^2$ *cuando* $d \to \infty$*; esto es,*

$$\frac{1}{d}\left|\left|X_i^+ - X_j^-\right|\right|^2 \xrightarrow{\mathbb{P}} \ell^2 := \sigma^2 + \tau^2 + c^2 \quad \textit{cuando} \quad d \to \infty. \qquad (1.22)$$

Demostración. Recuerde que $X_i^+(d) := X_i^+ = \left(X_{1,i}^+, X_{2,i}^+, ..., X_{d,i}^+\right)^\top$ es la i-ésima columna de la matriz de datos $\mathbf{X}_d^+$ de tamaño $d \times n$, y $X_j^-(d) := X_j^- = \left(X_{1,j}^-, X_{2,j}^-, ..., X_{d,j}^-\right)^\top$ es la j-ésima columna de la matriz de datos $\mathbf{X}_d^-$ de tamaño $d \times m$. Observe que

$$\begin{aligned}
\left|\left|X_i^+ - X_j^-\right|\right|^2 &= \sum_{r=1}^{d} \left\{\left(X_{r,i}^+ - \mathbb{E}\left[X_{r,i}^+\right]\right) - \left(X_{r,j}^- - \mathbb{E}\left[X_{r,j}^-\right]\right) + \left(\mathbb{E}\left[X_{r,i}^+\right] - \mathbb{E}\left[X_{r,j}^-\right]\right)\right\}^2 \\
&= \sum_{r=1}^{d} \left(\widetilde{X}_{r,i}^+\right)^2 + \sum_{r=1}^{d} \left(\widetilde{X}_{r,j}^-\right)^2 - 2\sum_{r=1}^{d} \left(\widetilde{X}_{r,i}^+\right)\left(\widetilde{X}_{r,j}^-\right) \qquad (1.23) \\
&+ \sum_{r=1}^{d} \left(\mathbb{E}\left[X_{r,i}^+\right] - \mathbb{E}\left[X_{r,j}^-\right]\right)^2 \qquad (1.24) \\
&+ 2\sum_{r=1}^{d} \left(\mathbb{E}\left[X_{r,i}^+\right] - \mathbb{E}\left[X_{r,j}^-\right]\right)\left(\widetilde{X}_{r,i}^+ - \widetilde{X}_{r,j}^-\right), \qquad (1.25)
\end{aligned}$$

donde las variables aleatorias $\widetilde{X}_{r,i}^+ := X_{r,i}^+ - \mathbb{E}\left[X_{r,i}^+\right]$ y $\widetilde{X}_{r,j}^- := X_{r,j}^- - \mathbb{E}\left[X_{r,j}^-\right]$ son las r-ésimas entradas de las columnas i-ésima y j-ésima de las *matrices centradas* $\widetilde{\mathbf{X}}_d^+ = \left[\widetilde{X}_1^+, \widetilde{X}_2^+, ..., \widetilde{X}_n^+\right]$ y $\widetilde{\mathbf{X}}_d^- = \left[\widetilde{X}_1^-, \widetilde{X}_2^-, ..., \widetilde{X}_m^-\right]$, respectivamente.

Teniendo en cuenta que $\widetilde{\mathbf{S}}_{D,d}^+ = \frac{1}{d}\left(\widetilde{\mathbf{X}}_d^+\right)^\top \widetilde{\mathbf{X}}_d^+$ y $\widetilde{\mathbf{S}}_{D,d}^- = \frac{1}{d}\left(\widetilde{\mathbf{X}}_d^-\right)^\top \widetilde{\mathbf{X}}_d^-$ se ve que los dos primeros sumandos en (**1.23**), reescalados por $(d\sigma_d^2)^{-1}$ y $(d\tau_d^2)^{-1}$, son precisamente las i-ésima y j-ésima entradas en la diagonal de $\frac{1}{\sigma_d^2}\widetilde{\mathbf{S}}_{D,d}^+$ y $\frac{1}{\tau_d^2}\widetilde{\mathbf{S}}_{D,d}^-$, respectivamente. De este modo, por la prueba del Teorema 1.2.2 se obtiene que $\frac{1}{d\sigma_d^2}\sum_{r=1}^{d}\left(\widetilde{X}_{r,i}^+\right)^2 \xrightarrow{\mathbb{P}} 1$ y $\frac{1}{d\tau_d^2}\sum_{r=1}^{d}\left(\widetilde{X}_{r,j}^-\right)^2 \xrightarrow{\mathbb{P}} 1$, cuando $d \to \infty$. Por consiguiente, de las condiciones dadas en (**1.20**) se obtiene que $\frac{1}{d}\sum_{r=1}^{d}\left(\widetilde{X}_{r,i}^+\right)^2 \xrightarrow{\mathbb{P}} \sigma^2$ y $\frac{1}{d}\sum_{r=1}^{d}\left(\widetilde{X}_{r,j}^-\right)^2 \xrightarrow{\mathbb{P}} \tau^2$, cuando $d \to \infty$.

Observe que el tercer sumando de (**1.23**) es dos veces el producto interior entre las columnas $\widetilde{X}_i^+$ y $\widetilde{X}_j^-$. Recordando que $\widetilde{\mathbf{X}}_d^+ = \widetilde{\mathbf{V}}_d^+\left(\widetilde{\mathbf{\Lambda}}_d^+\right)^{1/2}\widetilde{\mathbf{Z}}_d^+$ y $\widetilde{\mathbf{X}}_d^- = \widetilde{\mathbf{V}}_d^-\left(\widetilde{\mathbf{\Lambda}}_d^-\right)^{1/2}\widetilde{\mathbf{Z}}_d^-$ se tiene entonces que $\widetilde{X}_i^+ = \widetilde{\mathbf{V}}_d^+\left(\widetilde{\mathbf{\Lambda}}_d^+\right)^{1/2}\widetilde{Z}_i^+$ y $\widetilde{X}_j^- = \widetilde{\mathbf{V}}_d^-\left(\widetilde{\mathbf{\Lambda}}_d^-\right)^{1/2}\widetilde{Z}_j^-$, donde $\widetilde{Z}_i^+ = \left(Z_{1,i}^+, Z_{2,i}^+, ..., Z_{d,i}^+\right)^\top$ $(i = 1, 2, ..., n)$ y $\widetilde{Z}_j^- = \left(Z_{1,j}^-, Z_{2,j}^-, ..., Z_{d,j}^-\right)^\top$ $(j = 1, 2, ..., m)$ siguen una distribución con media cero y matriz de covarianza la identidad. Haciendo $\mathbf{U} = \left(\widetilde{\mathbf{V}}_d^+\right)^\top \widetilde{\mathbf{V}}_d^-$ y teniendo en cuenta

que $(d\sigma_d\tau_d)^{-1} = \left(\sum_{t=1}^{d}\lambda_{t,d}^{+}\right)^{-1/2}\left(\sum_{q=1}^{d}\lambda_{q,d}^{-}\right)^{-1/2}$ se tiene entonces que

$$\begin{aligned}
(d\sigma_d\tau_d)^{-1}\left\langle \widetilde{X}_i^{+},\widetilde{X}_j^{-}\right\rangle &= (d\sigma_d\tau_d)^{-1}\left(Z_{1,i}^{+},...,Z_{d,i}^{+}\right)\left(\widetilde{\mathbf{\Lambda}}_d^{+}\right)^{1/2}\left(\widetilde{\mathbf{V}}_d^{+}\right)^{\top}\widetilde{\mathbf{V}}_d^{-}\left(\widetilde{\mathbf{\Lambda}}_d^{-}\right)^{1/2}\\
&\times \left(Z_{1,j}^{-},...,Z_{d,j}^{-}\right)^{\top}\\
&= (d\sigma_d\tau_d)^{-1}\left(\left[\lambda_{1,d}^{+}\right]^{1/2}Z_{1,i}^{+},...,\left[\lambda_{d,d}^{+}\right]^{1/2}Z_{d,i}^{+}\right)\mathbf{U}\\
&\times\left(\left[\lambda_{1,d}^{-}\right]^{1/2}Z_{1,j}^{-},...,\left[\lambda_{d,d}^{-}\right]^{1/2}Z_{d,j}^{-}\right)^{\top}\\
&= (d\sigma_d\tau_d)^{-1}\left(\sum_{r=1}^{d}\left[\lambda_{r,d}^{+}\right]^{1/2}Z_{r,i}^{+}U_{r,1},...,\sum_{r=1}^{d}\left[\lambda_{r,d}^{+}\right]^{1/2}Z_{r,i}^{+}U_{r,d}\right)\\
&\times\left(\left[\lambda_{1,d}^{-}\right]^{1/2}Z_{1,j}^{-},...,\left[\lambda_{d,d}^{-}\right]^{1/2}Z_{d,j}^{-}\right)^{\top}\\
&= (d\sigma_d\tau_d)^{-1}\sum_{s=1}^{d}\left(\left[\lambda_{s,d}^{-}\right]^{1/2}Z_{s,j}^{-}\sum_{r=1}^{d}\left[\lambda_{r,d}^{+}\right]^{1/2}Z_{r,i}^{+}U_{r,s}\right)\\
&= \left(\sum_{t=1}^{d}\lambda_{t,d}^{+}\right)^{-1/2}\left(\sum_{q=1}^{d}\lambda_{q,d}^{-}\right)^{-1/2}\sum_{s=1}^{d}\sum_{r=1}^{d}\left(\lambda_{s,d}^{-}\lambda_{r,d}^{+}\right)^{1/2}U_{r,s}Z_{s,j}^{-}Z_{r,i}^{+}\\
&= \sum_{s=1}^{d}\sum_{r=1}^{d}\left(\widehat{\lambda}_{s,d}^{-}\widehat{\lambda}_{r,d}^{+}\right)^{1/2}U_{r,s}Z_{s,j}^{-}Z_{r,i}^{+}, \qquad (1.26)
\end{aligned}$$

donde $\widehat{\lambda}_{r,d}^{+} = \lambda_{r,d}^{+}/\sum_{t=1}^{d}\lambda_{t,d}^{+}$ y $\widehat{\lambda}_{s,d}^{-} = \lambda_{s,d}^{-}/\sum_{q=1}^{d}\lambda_{q,d}^{-}$ son los *valores propios relativos* correspondientes. Debido a que el valor esperado de la variable aleatoria obtenida en (1.26) es cero, entonces por la desigualdad de Chebyshev se obtiene que para cualquier $\epsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left[\left|\sum_{s=1}^{d}\sum_{r=1}^{d}\left(\widehat{\lambda}_{s,d}^{-}\widehat{\lambda}_{r,d}^{+}\right)^{1/2}U_{r,s}Z_{s,j}^{-}Z_{r,i}^{+}\right| > \epsilon\right]\\
\leq\ &\epsilon^{-2}\mathbb{E}\left[\left(\sum_{s=1}^{d}\sum_{r=1}^{d}\left(\widehat{\lambda}_{s,d}^{-}\widehat{\lambda}_{r,d}^{+}\right)^{1/2}U_{r,s}Z_{s,j}^{-}Z_{r,i}^{+}\right)^{2}\right]\\
=\ &\epsilon^{-2}\mathbb{V}ar\left[\sum_{s=1}^{d}\sum_{r=1}^{d}\left(\widehat{\lambda}_{s,d}^{-}\widehat{\lambda}_{r,d}^{+}\right)^{1/2}U_{r,s}Z_{s,j}^{-}Z_{r,i}^{+}\right]\\
=\ &\epsilon^{-2}\sum_{s=1}^{d}\sum_{r=1}^{d}\widehat{\lambda}_{s,d}^{-}\widehat{\lambda}_{r,d}^{+}U_{r,s}^{2}\,\mathbb{V}ar\left[Z_{s,j}^{-}Z_{r,i}^{+}\right]\\
=\ &\epsilon^{-2}\sum_{s=1}^{d}\sum_{r=1}^{d}\widehat{\lambda}_{s,d}^{-}\widehat{\lambda}_{r,d}^{+}U_{r,s}^{2}. \qquad (1.27)
\end{aligned}$$

Puesto que $\mathbf{U} = \left(\widetilde{\mathbf{V}}_d^+\right)^\top \widetilde{\mathbf{V}}_d^-$ es el producto de dos matrices ortogonales, entonces $\mathbf{U}$ es una matrix ortogonal, y dado que los valores propios relativos satisfacen las condiciones del Lema 1.2.1 se obtiene de (1.27) que $\sum_{s=1}^{d}\sum_{r=1}^{d}\left(\widehat{\lambda}_{s,d}^- \widehat{\lambda}_{r,d}^+\right)^{1/2} U_{r,s}\, Z_{s,j}^-\, Z_{r,i}^+ \xrightarrow{\mathbb{P}} 0$ cuando $d \to \infty$, es decir, $(d\sigma_d\tau_d)^{-1}\left\langle \widetilde{X}_i^+, \widetilde{X}_j^- \right\rangle \xrightarrow{\mathbb{P}} 0$ cuando $d \to \infty$. Por lo cual se tiene que $\frac{2}{d}\sum_{r=1}^{d}\left(\widetilde{X}_{r,i}^+\right)\left(\widetilde{X}_{r,j}^-\right) \xrightarrow{\mathbb{P}} 0$ cuando $d \to \infty$ ya que $\sigma_d\tau_d \longrightarrow \sigma\tau < \infty$ cuando $d \to \infty$.

Observe que (1.24), dividida por d, converge a c^2 cuando $d \to \infty$ debido a (1.19).

Finalmente, note que haciendo $\delta_r = \mathbb{E}\left[X_{r,i}^+\right] - \mathbb{E}\left[X_{r,j}^-\right]$, el término dado en (1.25) se descompone en los sumandos: $2\sum_{r=1}^{d}\delta_r\widetilde{X}_{r,i}^+$ y $2\sum_{r=1}^{d}\delta_r\widetilde{X}_{r,j}^-$. Note además que por (1.19), $\sum_{r=1}^{d}\delta_r^2 \approx dc^2$. Puede ser mostrado que cada uno de estos sumandos, divididos por d, converge en probabilidad a 0 cuando $d \to \infty$. En fecto, para el primer sumando, dividido por $d\sigma_d$, se tiene que

$$\begin{aligned}
\frac{2}{d\sigma_d}\sum_{r=1}^{d}\delta_r\widetilde{X}_{r,i}^+ &= \frac{2}{d^{1/2}}\left(\frac{1}{d\sigma_d^2}\right)^{1/2}\sum_{r=1}^{d}\delta_r\widetilde{X}_{r,i}^+ \\
&= \frac{2}{d^{1/2}}\sum_{r=1}^{d}\delta_r\left[\sum_{s=1}^{d}\left(\widehat{\lambda}_{s,d}^+\right)^{1/2} V_{r,s}^+\, Z_{s,i}^+\right] \\
&= \frac{2}{d^{1/2}}\sum_{s=1}^{d}\left(\widehat{\lambda}_{s,d}^+\right)^{1/2}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^+\right) Z_{s,i}^+
\end{aligned}$$

dado que $\widetilde{X}_i^+ = \widetilde{\mathbf{V}}_d^+\left(\widetilde{\mathbf{\Lambda}}_d^+\right)^{1/2}\widetilde{Z}_i^+$ y $\widehat{\lambda}_{r,d}^+ = \frac{\lambda_{r,d}^+}{d\sigma_d^2}$. Así, la desigualdad de Chebyshev y el hecho de que $\widetilde{\mathbf{Z}}_d^+$ sigue una distribución con media cero y matriz de covarianza $\mathbf{I}_d$, implican que para cualquier $\epsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{2}{d^{1/2}}\sum_{s=1}^{d}\left(\widehat{\lambda}_{s,d}^+\right)^{1/2}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^+\right) Z_{s,i}^+\right| > \epsilon\right] \\
\leq\ & \epsilon^{-2}\mathbb{E}\left[\left\{\frac{2}{d^{1/2}}\sum_{s=1}^{d}\left(\widehat{\lambda}_{s,d}^+\right)^{1/2}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^+\right) Z_{s,i}^+\right\}^2\right] \\
=\ & \frac{4\epsilon^{-2}}{d}\mathbb{E}\left[\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^+\left(\sum_{r=1}^{d}\delta_r V_{r,s}^+\right)^2\left(Z_{s,i}^+\right)^2\right] \\
+\ & \frac{8\epsilon^{-2}}{d}\mathbb{E}\left[\sum_{s<t}^{d}\left\{\left(\widehat{\lambda}_{s,d}^+\right)^{1/2}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^+\right) Z_{s,i}^+\right\}\left\{\left(\widehat{\lambda}_{t,d}^+\right)^{1/2}\left(\sum_{r=1}^{d}\delta_r V_{r,t}^+\right) Z_{t,i}^+\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{4\epsilon^{-2}}{d}\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^{+}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^{+}\right)^2 \mathbb{E}\left[\left(Z_{s,i}^{+}\right)^2\right] \\
&= \frac{4\epsilon^{-2}}{d}\sum_{s=1}^{d}\widehat{\lambda}_{s,d}^{+}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^{+}\right)^2 \\
&\leq \frac{4\epsilon^{-2}}{d}\operatorname*{máx}_{s}\left\{\widehat{\lambda}_{s,d}^{+}\right\}\sum_{s=1}^{d}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^{+}\right)^2 \\
&\approx \frac{4\epsilon^{-2}}{d}\operatorname*{máx}_{s}\left\{\widehat{\lambda}_{s,d}^{+}\right\}(dc^2) \longrightarrow 0 \quad \text{cuando } d\to\infty,
\end{aligned}
$$

dado que al ser $\widetilde{\mathbf{V}}_d^{+}$ una matriz ortogonal se tiene que $\sum_{s=1}^{d}\left(\sum_{r=1}^{d}\delta_r V_{r,s}^{+}\right)^2 = \sum_{r=1}^{d}\delta_r^2 \approx dc^2$. Por consiguiente, $\frac{2}{d}\sum_{r=1}^{d}\delta_r \widetilde{X}_{r,i}^{+} \xrightarrow{\mathbb{P}} 0$ cuando $d\to\infty$ pues $\sigma_d \longrightarrow \sigma < \infty$. Desde luego, esto deja ver que también $\frac{2}{d}\sum_{r=1}^{d}\delta_r \widetilde{X}_{r,j}^{-} \xrightarrow{\mathbb{P}} 0$ cuando $d\to\infty$. Esto finaliza la demostración de (1.22). ▲

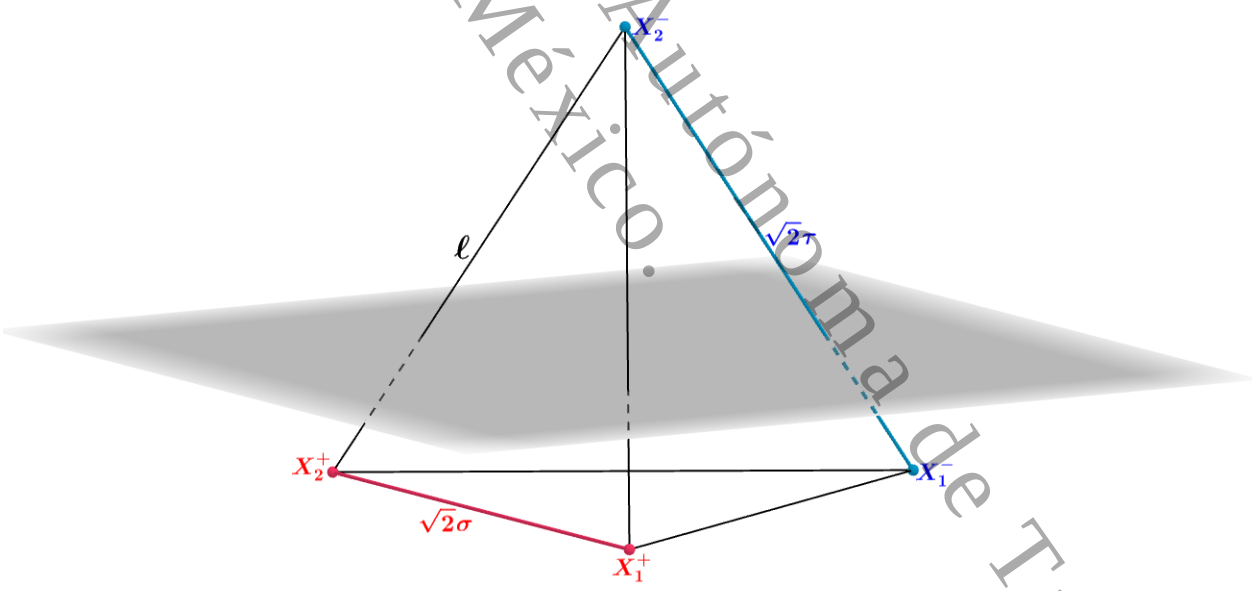


Figura 4: *Representación geométrica asintótica para las clases $\mathcal{C}_+(d)$ y $\mathcal{C}_-(d)$, en el caso $n = m = 2$, mostrando el 4-poliedro (tetraedro irregular) en el espacio 3-dimensional.*

Como es mencionado en Hall et al. (2005), la conclusión que se obtiene de (1.22) es que, después de reescalar cada componente del espacio d-variado por el factor $d^{-1/2}$, los $N := n + m$ vectores de datos de $\mathcal{C}_+(d) \cup \mathcal{C}_-(d)$ están asintóticamente localizados en los vértices de un

N-*poliedro* convexo en el espacio $(N-1)$-dimensional, donde este poliedro tiene N vértices y $N(N-1)/2$ aristas. Justamente n de los vértices son los puntos límites de los n vectores de datos de $\mathcal{C}_+(d)$ y son los vértices de un n-simplex de aristas de longitud $2^{1/2}\sigma$. Los otros m vértices son los puntos límites de los m vectores de datos de $\mathcal{C}_-(d)$ y son los vértices de un m-simplex de aristas de longitud $2^{1/2}\tau$. Las longitudes de las aristas en el N-poliedro que unen un vértice derivado de un dato de $\mathcal{C}_+(d)$ a uno derivado de un dato de $\mathcal{C}_-(d)$ son todas de longitud ℓ. En la Figura 4 se ilustra la estructura geométrica asintótica, considerando el caso en que $n = m = 2$.

Capítulo 2

Métodos de Clasificación Binaria

En este capítulo se aborda una descripción, tanto en el contexto general así como bajo la representación geométrica asintótica, de los métodos de discriminación binaria que son considerados en el tema central de esta tesis. El material que aquí se expone está basado principalmente del trabajo realizado en Bolívar-Cimé y Marron (2013). Para un estudio más profundo de estas metodologías estadísticas se pueden consultar las referencias Scholkopf y Smola (2002), Izenman (2008), Marron et al. (Febrero-2007), y Ahn y Marron (2004).

A continuación se definen los términos y tipos de datos que serán contemplados en las secciones posteriores de este capítulo. Considere pues el siguiente conjunto de *datos de entrenamiento*

$$(X_1, \delta_1),\ (X_2, \delta_2), ...,\ (X_N, \delta_N), \tag{2.1}$$

donde $X_i \in \mathbb{R}^d$ y $\delta_i \in \{-1, 1\}$, para $i = 1, 2, ..., N$. Se tiene así, en particular, dos clases de datos:

$$\mathcal{C}_+ = \{X_1^+, X_2^+, ..., X_n^+\} \quad \text{y} \quad \mathcal{C}_- = \{X_1^-, X_2^-, ..., X_m^-\}, \tag{2.2}$$

correspondientes a los vectores con las etiquetas $\delta_i = 1$ y $\delta_i = -1$, respectivamente, y $N = n+m$.

Sea $\mathbf{X} = [X_1, X_2, ..., X_N]$ la matriz de tamaño $d \times N$, cuyas columnas son precisamente los datos de entrenamiento, y sea $\delta = (\delta_1, \delta_2, ..., \delta_N)^\top$ el vector de las etiquetas correspondientes. Se definen los siguientes términos:

- $\mathbf{X}^+$ y $\mathbf{X}^-$ son las submatrices de $\mathbf{X}$ correspondientes a $\mathcal{C}_+$ y $\mathcal{C}_-$, respectivamente.
- $\mathbf{\Delta}$ es la matriz diagonal de $N \times N$ con los elementos de δ en su diagonal.

- $\mathbf{1}_q$ es el vector q-dimensional de unos.

Definición 2.0.1. *El conjunto de datos dado en (2.1) es linealmente separable si existe un hiperplano tal que todos los datos de la clase $\mathcal{C}_+$ están de un lado y todos los datos de clase $\mathcal{C}_-$ están del otro lado. Un hiperplano que cumple esta propiedad es llamado un hiperplano separante del conjunto de datos de entrenamiento.*

Observación 2.0.1. *De la sección 4.1 de Hall et al. (2005) se obtiene que una condición suficiente para que los datos dados en (2.1) sean linealmente separables casi seguramente es que $d \geq N$ y las distribuciones de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$ tengan una densidad de probabilidad continua.*

- Debido a la Observación 2.0.1, se asume en esta tesis que los conjuntos de datos a utilizar son linealmente separables, y por ende, las descripciones que se dan enseguida de los métodos de clasificación binaria se exponen para el caso de datos linealmente separables.

Suponga pues que existen un vector $U \in \mathbb{R}^d$ y un escalar β tales que cumplen lo siguiente:

$$U^\top X_i + \beta \leq -1 \quad \text{si} \quad \delta_i = -1, \qquad (2.3)$$
$$U^\top X_i + \beta \geq 1 \quad \text{si} \quad \delta_i = 1.$$

Resultando así que el hiperplano

$$U^\top Z + \beta = 0 \qquad (2.4)$$

es un hiperplano separante del conjunto de datos de entrenamiento dado en (2.1).

2.1. Mean Difference (MD)

El hiperplano separante que elige el método mean difference (MD), el cual se puede estudiar con más detalle en la Sección 1.2 de Scholkopf y Smola (2002), es aquel que biseca ortogonalmente al segmento de recta que une a los dos *centroides* o *medias muestrales* de las dos clases. En términos matemáticos, esto quiere decir que si las medias muestrales de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$ están dadas por

$$\overline{X}_+ = \frac{1}{n}\sum_{i=1}^{n} X_i^+ \qquad \text{y} \qquad \overline{X}_- = \frac{1}{m}\sum_{j=1}^{m} X_j^-,$$

respectivamente, entonces el hiperplano MD tiene como *vector normal unitario* e *intercepto*, respectivamente, a

$$U_0 = \frac{\overline{X}_+ - \overline{X}_-}{||\overline{X}_+ - \overline{X}_-||} \qquad \text{y} \qquad \beta_0 = -U_0^\top \left(\frac{\overline{X}_+ + \overline{X}_-}{2} \right); \tag{2.5}$$

y biseca al segmento que une a $\overline{X}_+$ y $\overline{X}_-$. Observe así que el vector normal del hiperplano MD es la diferencia entre dos puntos de las *envolventes convexas* de las dos clases.

Hiperplano MD bajo la representación geométrica asintótica

Bajo la representación geométrica asintótica que fue descrita en el último párrafo del capítulo anterior, resulta que las envolventes convexas de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$ son precisamente el n-simplex y el m-simplex, respectivamente. Desde luego, en este contexto se cumple también, que después de reescalar por $d^{-1/2}$ y hacer $d \to \infty$, las medias muestrales $\overline{X}_+$ y $\overline{X}_-$ convergen en probabilidad a los respectivos centroides de estos simpleces, a los cuales se les denotará como C_+ y C_-, respectivamente. Así por ejemplo, para el caso $n = m = 2$, el hiperplano MD en el 4-poliedro queda definido como se muestra en la Figura 1.

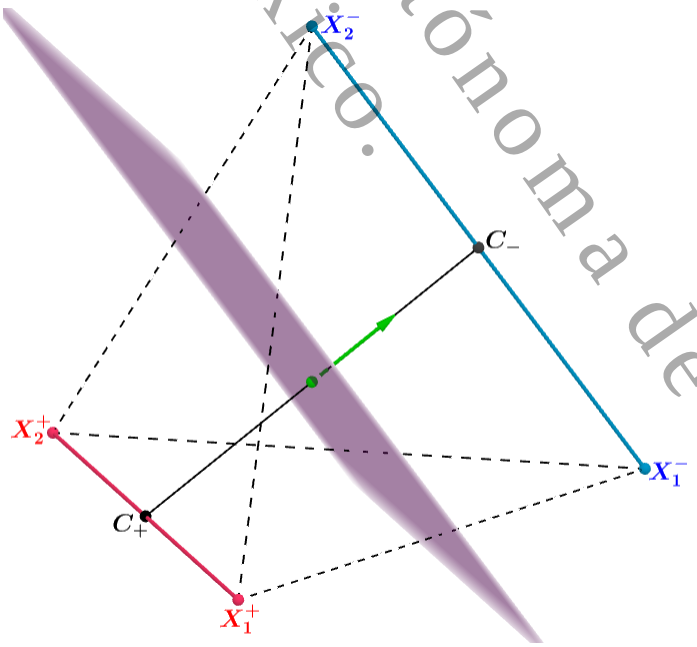


Figura 1: *Hiperplano MD en la representación geométrica asintótica (4-poliedro) de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, en el caso $n = m = 2$.*

2.2. Support Vector Machine (SVM)

El método support vector machine (SVM), propuesto por Cortes y Vapnik (1995), es una de las metodologías más populares que se ha desarrollado para abordar problemas de clasificación. A grosso modo, en este caso linealmente separable, el método SVM consiste en encontrar un hiperplano separante que maximice las distancias del hiperplano a los vectores más cercanos de cada clase. En términos matemáticos esto se expone como sigue.

Observe que las desigualdades dadas en (2.3) se pueden escribir de manera equivalente como

$$\delta_i \left(U^\top X_i + \beta\right) \geq 1, \qquad i = 1, 2, ..., N. \tag{2.6}$$

A los vectores X_i que satisfacen la igualdad en (2.6) se les conoce como *vectores soporte*. Esto es, los vectores soporte son los vectores de entrenamiento que pertenecen a uno de los hiperplanos

$$U^\top Z + \beta = -1 \quad \text{o} \quad U^\top Z + \beta = 1. \tag{2.7}$$

Sean ρ_+ y ρ_- las distancias más cortas entre el hiperplano separante dado en (2.4) y los datos en las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, respectivamente. El *margen* del hiperplano separante (2.4) es definido como $\rho = \rho_+ + \rho_-$. Así, de manera equivalente, este margen queda determinado por la distancia que hay entre los hiperplanos dados en (2.7), la cual está dada por

$$\rho = \frac{2}{||U||}. \tag{2.8}$$

Debido a que se está considerando el caso linealmente separable, entonces el hiperplano separante óptimo o hiperplano SVM

$$U_1^\top Z + \beta_1 = 0$$

es el único hiperplano separante con margen máximo. De este modo, el hiperplano SVM maximiza (2.8) sujeto a la restricción (2.6), o equivalentemente, el hiperplano SVM resuelve el problema de optimización

$$\begin{aligned} &\text{minimizar} && \frac{||U||^2}{2}, \\ &\text{sujeto a} && \delta_i \left(U^\top X_i + \beta\right) \geq 1, \quad i = 1, 2, ..., N. \end{aligned} \tag{2.9}$$

En la Subsección 11.2.1 de Izenman (2008) se ve que el vector óptimo del problema (2.9) puede ser encontrado usando multiplicadores Lagrangianos, y está dado por

$$U_1 = \mathbf{X}\boldsymbol{\Delta}\Theta_1, \tag{2.10}$$

donde $\Theta_1 = \left(\Theta_{1,1}, \Theta_{2,1}, ..., \Theta_{N,1}\right)^\top$ es la solución al problema de optimización

$$\begin{aligned} &\text{maximizar} \quad \mathbf{1}_N^\top \Theta - \frac{1}{2}\left|\left|\mathbf{X}\boldsymbol{\Delta}\Theta\right|\right|, \\ &\text{sujeto a} \quad \Theta \geq 0, \quad \Theta^\top \delta = 0. \end{aligned} \tag{2.11}$$

Allí se muestra además que $\Theta_{i,1} \neq 0$ únicamente si X_i es un vector soporte, y así, de (2.10) se concluye que U_1 es una función lineal de los vectores soportes únicamente. Dado que los vectores soportes satisfacen una de las igualdades dadas en (2.7), se tiene que la ordenada al origen (intercepto) β_1 puede ser tomada como

$$\beta_1 = \delta_i - U_1^\top X_i,$$

para cualquier vector soporte X_i.

De manera equivalente, se ha visto que resolver el problema (2.11) es lo mismo que dar solución al problema de optimización

$$\begin{aligned} &\text{maximizar} \quad \frac{2}{\left|\left|\mathbf{X}\boldsymbol{\Delta}\widehat{\Theta}\right|\right|^2}, \\ &\text{sujeto a} \quad \mathbf{1}_n^\top \widehat{\Theta}^+ = \mathbf{1}_m^\top \widehat{\Theta}^- = 1, \quad \widehat{\Theta}^+, \widehat{\Theta}^- \geq 0, \end{aligned} \tag{2.12}$$

donde $\widehat{\Theta}^+$ y $\widehat{\Theta}^+$ son los subvectores de $\widehat{\Theta}$ correspondientes a las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, respectivamente. Se puede observar que $\mathbf{X}\boldsymbol{\Delta}\widehat{\Theta} = \mathbf{X}^+\widehat{\Theta}^+ - \mathbf{X}^-\widehat{\Theta}^-$; por lo cual resulta claro de (2.12) que se está minimizando la distancia entre puntos de las envolventes convexas de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$. Por lo tanto, si $\widehat{\Theta}_1$ resuelve el problema de optimización (2.12), entonces el vector normal del hiperplano SVM puede ser tomado como

$$\widehat{U}_1 = \mathbf{X}\boldsymbol{\Delta}\widehat{\Theta}_1 = \mathbf{X}^+\widehat{\Theta}_1^+ - \mathbf{X}^-\widehat{\Theta}_1^-,$$

que es la diferencia entre los dos puntos más cercanos de las envolventes convexas de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, y es proporcional al vector U_1 dado en (2.10).

Hiperplano SVM bajo la representación geométrica asintótica

Debido a que, respectivamente, las envolventes convexas de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$ son precisamente el n-simplex y el m-simplex, se tiene entonces que la proyección del hiperplano SVM $(d-1)$-dimensional en el hiperplano $(N-1)$-dimensional que es generado por los datos, es dado asintóticamente por el único hiperplano $(N-2)$-dimensional que biseca cada una de las aristas de longitud ℓ en el N-poliedro (véase la Figura 2).

Observación 2.2.1. *En este contexto asintótico se puede ver que el hiperplano SVM divide exactamente en la misma forma al espacio en que lo hace el hiperplano MD, por lo cual estos hiperplanos satisfacen las mismas propiedades y pueden ser identificados como iguales bajo la representación geométrica asintótica. Esto es ilustrado en las figuras 1 y 2.*

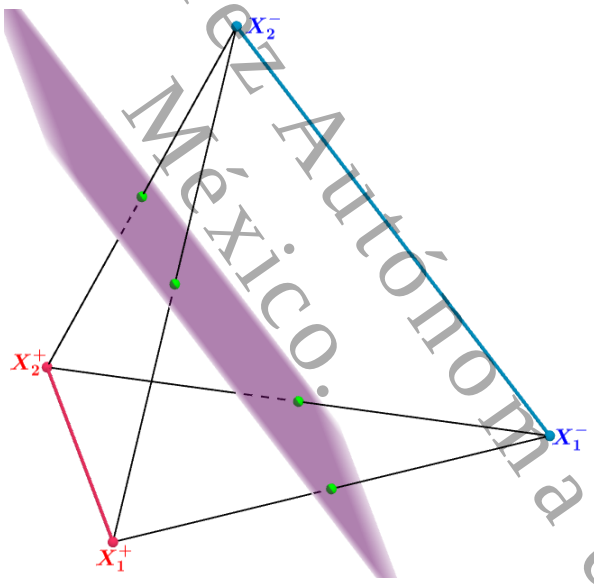


Figura 2: *Hiperplano SVM en la representación geométrica asintótica (4-poliedro) de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, en el caso $n = m = 2$.*

2.3. Distance Weighted Discrimination (DWD)

El método de discriminación binaria, distance weighted discrimination (DWD), fue propuesto por Marron et al. (Febrero-2007), a raíz de que en el contexto de datos de dimensión alta observaron que la proyección de los datos sobre el vector normal del hiperplano SVM produce

acumulación sustancial de datos, lo cual puede afectar el rendimiento de la generalización en el sentido de qué tan bien se pueden discriminar nuevos datos con la misma distribución. El método DWD evita este problema de apilamiento de datos y hace una generalización mejor. Este método encuentra un hiperplano separante que minimiza la suma de los recíprocos de las distancias de los datos de entrenamiento al hiperplano. De este modo todos los datos de entrenamiento contribuyen en la búsqueda del hiperplano DWD, siendo los datos más cercanos a este hiperplano quienes aportan más información. En este caso linealmente separable, este método puede ser descrito como se expone enseguida.

Defina el *residual* del i-ésimo vector X_i como

$$\rho_i = \delta_i \left(U^\top X_i + \beta \right), \tag{2.13}$$

que es precisamente la distancia signada de X_i al hiperplano separante $U^\top Z + \beta = 0$. El *hiperplano DWD*

$$U_2^\top Z + \beta_2 = 0$$

resuelve el problema de optimización

$$\text{minimizar} \quad \sum_{i=1}^{N} \frac{1}{\rho_i},$$

$$\text{sujeto a} \quad ||U|| = 1, \quad \rho_i \geq 0, \quad i = 1, 2, ..., N.$$

Mostrando así que todos los datos de entrenamiento aportan información en la búsqueda del hiperplano DWD, siendo los más cercanos al hiperplano separante (2.4) los que influyen más. Como puede ser visto en Marron et al. (Febrero-2007), el vector óptimo está dado por

$$U_2 = \frac{\mathbf{X\Delta}\Theta_2}{||\mathbf{X\Delta}\Theta_2||},$$

donde $\Theta_2 = \left(\Theta_{1,2}, \Theta_{2,2}, ..., \Theta_{N,2}\right)^\top$ es la solución al problema de optimización

$$\text{maximizar} \quad 2\mathbf{1}_N^\top \sqrt{\Theta} - ||\mathbf{X\Delta}\Theta||, \tag{2.14}$$

$$\text{sujeto a} \quad \Theta \geq 0, \quad \Theta^\top \delta = 0;$$

con $\sqrt{\Theta}$ denotando al vector cuyas componentes son las raíces cuadradas de las componentes de Θ. Note que el problema (2.14) es similar al problema de optimización (2.11) que corresponde al método SVM. Sin embargo, aquí se consideran todos los residuales, los cuales están dados por

$$\rho_i = \frac{1}{\sqrt{\Theta_{i,2}}}, \qquad i = 1, 2, ..., N.$$

De este modo, por la igualdad (2.13), se obtiene que el intercepto β_2 puede ser calculado como

$$\beta_2 = \frac{\delta_i}{\sqrt{\Theta_{i,2}}} - U_2^\top X_i,$$

para cualquier vector X_i.

Análogamente al caso del método SVM, en Marron et al. (Febrero-2007) se muestra que el problema (2.14) es equivalente al problema de optimización

$$\begin{aligned} &\text{maximizar} \quad \frac{\left(\mathbf{1}_n^\top \sqrt{\widehat{\Theta}^+} + \mathbf{1}_m^\top \sqrt{\widehat{\Theta}^-}\right)^2}{\left|\left|\mathbf{X}^+\widehat{\Theta}^+ - \mathbf{X}^-\widehat{\Theta}^-\right|\right|}, \\ &\text{sujeto a} \quad \mathbf{1}_n^\top \widehat{\Theta}^+ = \mathbf{1}_m^\top \widehat{\Theta}^- = 1, \quad \widehat{\Theta}^+, \widehat{\Theta}^- \geq 0. \end{aligned} \tag{2.15}$$

De manera que se está minimizando la distancia entre puntos en las dos envolventes convexas de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, pero divididos por el cuadrado de la suma de las raíces cuadradas de los pesos convexos. Por lo tanto, si $\widehat{\Theta}_2$ es la solución al problema de optimización (2.15), entonces el vector normal del hiperplano DWD es proporcional a $\widehat{U}_2 = \mathbf{X}^+\widehat{\Theta}_2^+ - \mathbf{X}^-\widehat{\Theta}_2^-$.

Hiperplano DWD bajo la representación geométrica asintótica

Observe la Figura 3 para analizar las propiedades del hiperplano DWD bajo este contexto asintótico. Evidentemente, el segmento de recta que une a los centroides C_+ y C_- del n-simplex y del m-simplex, respectivamente, es ortogonal a los subespacios lineales que son generados por estos simpleces. De modo que el hiperplano DWD debe ser ortogonal a este segmento. Sea Q un punto cualquiera en el intervalo C_+ C_-. Se verá qué condición debe cumplir Q para que viva en el hiperplano DWD.

Debido a que el n-simplex es ortogonal al segmento C_+ C_-, todos los vértices de este simplex están a una distancia α del hiperplano que pasa por Q ortogonalmente a C_+ C_-. Análogamente, todos los vértices del m-simplex están a una distancia γ del hiperplano que pasa por Q

ortogonalmente a $C_+\ C_-$. El hiperplano DWD minimiza $\frac{n}{\alpha^q} + \frac{m}{\gamma^q}$, $q > 0$, sujeto a la restricción de que $\alpha + \gamma$ es constante. No es difícil ver que este mínimo satisface la igualdad

$$\frac{\alpha}{\gamma} = \left(\frac{n}{m}\right)^{1/(q+1)}. \tag{2.16}$$

Esto da la ubicación del hiperplano DWD. Es el hiperplano que es ortogonal a $C_+\ C_-$ que pasa a través del punto Q, el cual satisface la condición (2.16). Note aquí que $\alpha = \gamma$ si y solo si $n = m$. En este caso, los hiperplanos SVM y DWD coinciden.

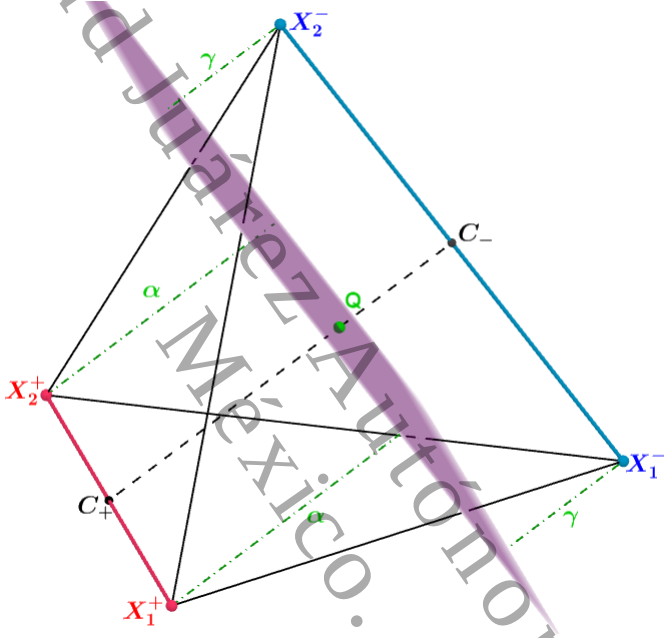


Figura 3: *Ilustración de las propiedades del Hiperplano DWD en la representación geométrica asintótica (4-poliedro) de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, en el caso $n = m = 2$. Aquí $Q = (C_+ + C_-)/2$.*

2.4. Maximal Data Piling (MDP)

En Ahn y Marron (2004) es estudiado con más detalle el método maximal data piling (MDP), el cual fue exclusivamente diseñado para abordar problemas de clasificación binaria que involucren datos de dimensión alta. Es necesario asumir que $d \geq N - 1$ y que el subespacio generado por los datos de entrenamiento tiene dimensión $N - 1$. Bajo estas hipótesis es posible encontrar una dirección (vector) sobre el cual las proyecciones de los datos de entrenamiento caen únicamente en dos puntos distintos, uno por cada clase; esto es, las proyecciones de los vectores

de la clase $\mathcal{C}_+$ resultan ser iguales a un mismo punto, y análogamente para las proyecciones de los vectores de la clase $\mathcal{C}_-$. El vector normal del método MDP es la dirección para el cual la distancia entre estos dos puntos resulta ser máxima.

Para exponer como construir el hiperplano separante que elige este método, considere la matriz $\mathbf{R} = \left[\widetilde{\mathbf{X}}^+, \widetilde{\mathbf{X}}^-\right]$, donde $\widetilde{\mathbf{X}}^+$ y $\widetilde{\mathbf{X}}^-$ son las versiones centradas de $\mathbf{X}^+$ y $\mathbf{X}^-$ respectivamente, esto es,

$$\widetilde{\mathbf{X}}^+ = \mathbf{X}^+ - \overline{X}_+ \mathbf{1}_d^\top \qquad \text{y} \qquad \widetilde{\mathbf{X}}^- = \mathbf{X}^- - \overline{X}_- \mathbf{1}_d^\top.$$

La matriz proyección simétrica sobre el complemento ortogonal del espacio columna de $\mathbf{R}$ está dada por $\mathbf{Q} = \mathbf{I}_d - \mathbf{R}\mathbf{R}^\dagger$, donde $\mathbf{R}^\dagger$ es la *inversa generalizada de Moore-Penrose* de $\mathbf{R}$; véase el Apéndice A de Ben-Israel et al. (2003) para más información sobre $\mathbf{R}^\dagger$.

El *hiperplano MDP*

$$U_3^\top Z + \beta_3 = 0$$

tiene la propiedad de que su vector normal unitario U_3 es la dirección para el cual las proyecciones de las dos medias de las clases tienen distancia máxima, sujeta a la condición de que la proyección de cada uno de los vectores de cada una de las clases cae sobre el mismo punto que la proyección de la media de dicha clase. En términos matemáticos, se tiene que U_3 resuelve el problema de optimización

$$\begin{aligned} &\text{maximizar} && \left(U^\top U_0\right)^2, \\ &\text{sujeto a} && \mathbf{R}^\top U = 0, \quad U^\top U = 1, \end{aligned} \tag{2.17}$$

donde $U_0 = \overline{X}_+ - \overline{X}_-$ es el vector normal del método MD. En Ahn y Marron (2004) puede ser visto que la solución al problema (2.17) está dada por

$$U_3 = \frac{\mathbf{Q}U_0}{||\mathbf{Q}U_0||}, \tag{2.18}$$

lo cual deja ver que el vector normal del método MDP, U_3, es ortogonal al subespacio $(N-2)-$dimensional generado por las columnas de $\mathbf{R}$. Más aún, (2.18) es también equivalente a

$$U_3 = \frac{\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\right)^\dagger U_0}{\left|\left|\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\right)^\dagger U_0\right|\right|},$$

donde $\widetilde{\mathbf{X}}$ es la versión centrada de la matriz de datos inicial $\mathbf{X}$. Por lo tanto, U_3 pertenece al subespacio $(N-1)-$dimensional generado por todos los vectores centrados.

Hiperplano MDP bajo la representación geométrica asintótica

En el contexto de la representación geométrica asintótica, como se puede consultar en Bolívar-Cimé (2021), el vector normal del hiperplano MDP tiende a tener la misma dirección que el vector normal del hiperplano MD cuando $d \to \infty$. De manera que, al tomar el intercepto como $\beta_3 = -U_3^\top \left(\frac{\overline{X}_+ + \overline{X}_-}{2} \right)$, los hiperplanos MD Y MDP coinciden en este contexto asintótico.

2.5. Probabilidades de Clasificación Correcta

Asuma en esta sección que las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, dadas en (2.2), son independientes y que satisfacen las condiciones del Teorema 1.2.4. Esto es, la representación geométrica asintótica se cumple para estas clases. Debido a esta estructura geométrica en los datos, por Hall et al. (2005) y Qiao et al. (2010), se tienen los siguientes resultados acerca del comportamiento asintótico de los métodos de discriminación binaria MD, SVM, DWD y MDP, en términos de las probabilidades de clasificación correcta.

Agregue ahora un nuevo dato al espacio $\mathbb{R}^d$; éste debe ser independiente de los datos en $\mathcal{C}_+ \cup \mathcal{C}_-$ y tener la distribución de la población $\mathcal{C}_+$ o de la población $\mathcal{C}_-$.

Teorema 2.5.1. *Suponga que $\sigma^2/n \geq \tau^2/m$. Si $c^2 > \sigma^2/n - \tau^2/m$, entonces la probabilidad de que un nuevo dato proveniente de la población $\mathcal{C}_+$ o de la población $\mathcal{C}_-$ sea correctamente clasificado por el hiperplano MD, SVM o MDP converge a 1 cuando $d \to \infty$. Si $c^2 < \sigma^2/n - \tau^2/m$, entonces con probabilidad convergiendo a 1 cuando $d \to \infty$ un nuevo dato proveniente de cualquiera de las dos poblaciones $\mathcal{C}_+$ o $\mathcal{C}_-$ será clasificado por el hiperplano MD, SVM o MDP como perteneciente a la población $\mathcal{C}_-$.*

Demostración. Dado que bajo la representación geométrica asintótica los hiperplanos MD, SVM y MDP coinciden, se dará entonces la prueba de este resultado considerando al hiperplano MD. Denote pues al nuevo dato por X, y suponga que tiene la distribución dada en $\mathcal{C}_+$. La representación geométrica asintótica dice que, cuando $d \to \infty$, la distancia de X a cada $X_i^+ \in \mathcal{C}_+$,

reescalada por $d^{-1/2}$, converge en probabilidad a $(2\sigma^2)^{1/2}$; y la distancia de X a cada $X_j^- \in \mathcal{C}_-$, reescalada por $d^{-1/2}$, converge en probabilidad a $\ell = (\sigma^2 + \tau^2 + c^2)^{1/2}$.

Recuerde que los términos n-simplex y m-simplex denotan a los simpleces límites de las clases $\mathcal{C}_+$ y $\mathcal{C}_-$, respectivamente. La distancia al cuadrado de cualquiera de los vértices del n-simplex a su centroide C_+ es igual a

$$(1 - n^{-1})\sigma^2. \tag{2.19}$$

Para ilustrar esto, tome $\sigma^2 = 1$ y represente al n-simplex en el espacio euclideano n-variado a través de sus vértices en los puntos canónicos $(1, 0, ..., 0), (0, 1, ..., 0)..., (0, 0, ..., 1)$. Entonces el centroide del n-simplex es $C_+ = (n^{-1}, ..., n^{-1})$, y así, la distancia al cuadrado de éste a cualquiera de los vértices es igual a $(1 - n^{-1})^2 + (n - 1)n^{-2} = 1 - n^{-1}$.

Sea $X' \in \mathbb{R}^d$ un punto cuya distancia a cada vértice del n-simplex es ρ. Entonces X', cualquier vértice V y el centroide C_+ del n-simplex son los vértices de un triángulo rectángulo cuya hipotenusa es el segmento de recta que une a los vectores X' y V. Luego, por el teorema de Pitagóras, la distancia al cuadrado de X' a C_+ es igual a

$$\rho^2 - (1 - n^{-1})\sigma^2. \tag{2.20}$$

El nuevo dato X es correctamente clasificado si y solo si está más cercano al centroide C_+ del n-simplex que al centroide C_- del m-simplex. Debido a lo obtenido en (2.20), la distancia al cuadrado de X a C_+ y a C_- es igual, respectivamente, a

$$2\sigma^2 - (1 - n^{-1})\sigma^2 = (1 + n^{-1})\sigma^2, \tag{2.21}$$

$$\ell^2 - (1 - m^{-1})\tau^2 = \sigma^2 + m^{-1}\tau^2 + c^2. \tag{2.22}$$

De esta forma, X estará más cercano al n-simplex y por consiguiente correctamente clasificado si $(1 + n^{-1})\sigma^2 < \sigma^2 + m^{-1}\tau^2 + c^2$, es decir, si $c^2 > n^{-1}\sigma^2 - m^{-1}\tau^2$; y estará más cercano al m-simplex, y por lo tanto mal clasificado, si $c^2 < n^{-1}\sigma^2 - m^{-1}\tau^2$.

Hasta aquí no se ha considerado cuál de los dos números σ^2/n y τ^2/m es mayor. Asuma pues que $\sigma^2/n \geq \tau^2/m$. El argumento anterior dice cuándo un nuevo dato de la población $\mathcal{C}_+$ será clasificado correctamente. Para un nuevo dato de la población $\mathcal{C}_-$, el argumento análogo correspondiente dirá que dicho dato será correctamente clasificado si $c^2 > m^{-1}\tau^2 - n^{-1}\sigma^2$. Debido a que el lado derecho de esta última desigualdad es siempre negativa (pues $\sigma^2/n \geq \tau^2/m$), se

tiene entonces que $c^2 > m^{-1}\tau^2 - n^{-1}\sigma^2$ siempre se cumplirá. De este modo, un nuevo dato de la población $\mathcal{C}_-$ será siempre clasificado correctamente. ▲

Como ya se ha observado, cuando los tamaños muestrales n y m son iguales, el hiperplano DWD coincide con el hiperplano SVM. Por lo cual, si $n = m$, entonces el hiperplano DWD satisface también el Teorema 2.5.1. En el caso general, si n es no necesariamente igual a m, se tiene el siguiente resultado.

Teorema 2.5.2. *Sea $q > 0$ y asuma que se cumple que $\sigma^2/n^{(q+2)/(q+1)} \geq \tau^2/m^{(q+2)/(q+1)}$. Si $c^2 > (m/n)^{1/(q+1)}\sigma^2/n - \tau^2/m$, entonces la probabilidad de que un nuevo dato proveniente de la población $\mathcal{C}_+$ o de la población $\mathcal{C}_-$ sea correctamente clasificado por el hiperplano DWD converge a 1 cuando $d \to \infty$. Si $c^2 < (m/n)^{1/(q+1)}\sigma^2/n - \tau^2/m$, entonces con probabilidad convergiendo a 1 cuando $d \to \infty$ un nuevo dato proveniente de cualquiera de las dos poblaciones será clasificado por el hiperplano DWD como perteneciente a la población $\mathcal{C}_-$.*

Demostración. Aplique lo obtenido en (2.20) al caso en que X' es uno de los vértices del m-simplex. La distancia al cuadrado de X' a cualquier vértice del n-simplex es $\ell^2 = \sigma^2 + \tau^2 + c^2$, y así por (2.20), la distancia al cuadrado de X' al centroide C_+ del n-simplex es

$$\sigma^2 + \tau^2 + c^2 - (1 - n^{-1})\sigma^2 = \sigma^2/n + \tau^2 + c^2. \tag{2.23}$$

Esto es cierto para cualquier vértice X' del m-simplex. El mismo análisis dice que el cuadrado de la distancia de C_+ al centroide C_- del m-simplex es

$$\sigma^2/n + \tau^2 + c^2 - (1 - m^{-1})\tau^2 = \sigma^2/n + \tau^2/m + c^2. \tag{2.24}$$

Suponga ahora que el nuevo dato X tiene la distribución dada en $\mathcal{C}_+$ y que es independiente de los datos en $\mathcal{C}_+ \cup \mathcal{C}_-$. Debido al análisis anterior, las distancias entre X y los centroides C_+ y C_- son conocidas, es decir, las longitudes de las aristas del triángulo dado en la Figura 4 son conocidas. En este figura, el punto Q' es la proyección ortogonal de X sobre el segmento de recta C_+C_-, la constante h es la distancia de X a Q', y las constantes α' y γ' son las distancias de Q' a C_+ y C_-, respectivamente. De modo que se obtiene

$$(\alpha')^2 + h^2 = (1 + n^{-1})\sigma^2, \tag{2.25}$$

$$(\gamma')^2 + h^2 = \sigma^2 + \tau^2/m + c^2, \tag{2.26}$$

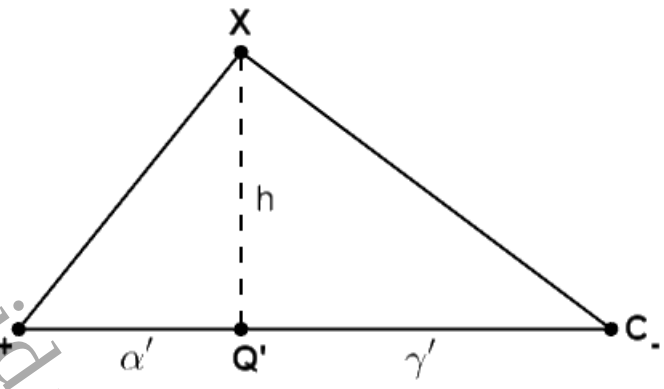


Figura 4: *Proyección ortogonal del nuevo dato X sobre el segmento C_+C_-.*

$$(\alpha' + \gamma')^2 = \sigma^2/n + \tau^2/m + c^2. \tag{2.27}$$

Restando la ecuación (2.26) a la ecuación (2.25) se tiene que

$$(\alpha')^2 - (\gamma')^2 = \sigma^2/n - \tau^2/m - c^2. \tag{2.28}$$

Sumando las ecuaciones (2.27) y (2.28), y posteriormente restando la ecuación (2.27) a la ecuación (2.28) se obtiene, respectivamente que

$$\alpha'(\alpha' + \gamma') = \sigma^2/n,$$

$$\gamma'(\alpha' + \gamma') = \tau^2/m + c^2,$$

de lo cual se deduce que

$$\frac{\alpha'}{\gamma'} = \frac{\sigma^2/n}{\tau^2/m + c^2}.$$

De acuerdo a la condición (2.16), si Q es el punto donde el hiperplano DWD se interseca perpendicularmente con el segmento de recta que une a C_+ con C_- y, si α y γ son las distancias de Q a C_+ y C_-, respectivamente, se debe cumplir entonces que

$$\frac{\alpha}{\gamma} = \left(\frac{n}{m}\right)^{1/(q+1)}.$$

Por consiguiente, el nuevo dato X será clasificado correctamente (en la clase $\mathcal{C}_+$) por el hiperplano DWD si y solo si $\frac{\alpha'}{\gamma'} < \frac{\alpha}{\gamma}$, es decir, si

$$c^2 > \left(\frac{m}{n}\right)^{1/(q+1)} \left(\frac{\sigma^2}{n}\right) - \frac{\tau^2}{m};$$

y en caso contrario será clasificado incorrectamente (en la clase $\mathcal{C}_-$).

El análisis realizado hasta aquí dice cuando un nuevo dato de la población $\mathcal{C}_+$ será correctamente clasificado (o incorrectamente clasificado) sin suponer que

$$\sigma^2/n^{(q+2)/(q+1)} \geq \tau^2/m^{(q+2)/(q+1)}. \tag{2.29}$$

Suponga ahora que (2.29) se cumple y que el nuevo dato X proviene de la población $\mathcal{C}_-$. Note que en este caso el argumento análogo correspondiente dirá que X será correctamente clasificado si

$$c^2 > \left(\frac{n}{m}\right)^{1/(q+1)} \left(\frac{\tau^2}{m}\right) - \frac{\sigma^2}{n}. \tag{2.30}$$

Observe que de (2.29) se obtiene que

$$\frac{\tau^2/m}{\sigma^2/n} \leq \left(\frac{m}{n}\right)^{1/(q+1)},$$

y así, para cualquier $c > 0$ se tiene que

$$\frac{\tau^2/m}{\sigma^2/n + c^2} < \frac{\tau^2/m}{\sigma^2/n} \leq \left(\frac{m}{n}\right)^{1/(q+1)};$$

de lo cual resulta claro que (2.30) siempre se cumple, concluyendo así que cualquier nuevo dato X de la clase $\mathcal{C}_-$ será correctamente clasificado siempre por el hiperplano DWD. ▲

Como es mencionado en Bolívar-Cimé (2021), note que si $n = m$ y $\sigma = \tau$, entonces para cualquier $c > 0$, los métodos MD, SVM, DWD y MDP aseguran una clasificación asintóticamente correcta para cualquier nuevo dato de cualquiera de las dos poblaciones. Por otro lado, si $n \neq m$ ó $\sigma \neq \tau$, entonces los métodos MD, SVM y MDP pueden tener un comportamiento asintótico diferente al método DWD, en términos de las probabilidades de clasificación correcta. Por ejemplo, haciendo

$$M_1 = \frac{\sigma^2}{n} - \frac{\tau^2}{m} \quad \text{y} \quad M_2 = \left(\frac{m}{n}\right)^{1/2} M_1,$$

y suponiendo que $m > n$ y $\sigma^2/n > \tau^2/m$, entonces $M_1 < M_2$ y si $M_1 < c^2 < M_2$ se tiene por el Teorema 2.5.1 que los métodos MD, SVM y MDP logran una clasificación asintóticamente correcta para un nuevo dato de cualquier población, mientras que el método DWD asegura una clasificación asintóticamente perfecta para la población $\mathcal{C}_-$ y una clasificación completamente incorrecta para la población $\mathcal{C}_+$. Lo cual muestra que los métodos MD, SVM y MDP poseen una ventaja sobre el método DWD bajo la representación geométrica asintótica, puesto que los tres primeros métodos clasifican correctamente un nuevo dato de cualquier población para un rango más grande de valores de c.

Capítulo 3

Clasificación Multicategoría

Los métodos de clasificación binaria pueden ser generalizados de varias maneras para resolver problemas de clasificación multiclase o multicategoría. Una estrategia para resolver problemas de clasificación multicategoría consiste en llevar a cabo una serie de problemas de clasificación binaria. Algunos ejemplos son el método *One-Versus-One (OVO)* y el método *One-Versus-The Rest (OVR)*. Una segunda estrategia toma a toda la población de forma simultánea y considera a todas las clases en una sola. Es decir, en los métodos de clasificación multicategoría, se construyen varios clasificadores binarios o bien se resuelve un problema de optimización grande que envuelve a todas las clases al mismo tiempo; esto puede ser consultado por ejemplo en Huang et al. (2010) y Huang et al. (2013).

En un problema de clasificación multicategoría se tiene un conjunto de datos de entrenamiento que consiste de N observaciones

$$(X_1, \delta_1),\ (X_2, \delta_2), ...,\ (X_N, \delta_N).$$

Aquí $X_r \in \mathbb{R}^d$ representa un vector, $\delta_r \in \{1, 2, ..., K\}$ denota el correspondiente índice de clase o categoría, y $K \geq 3$ es el total de categorías. Se asume que los vectores (X_r, δ_r) son vectores aleatorios independientes y distribuidos de acuerdo a alguna función de distribución desconocida $P(X, \delta)$. El objetivo es construir una regla de clasificación $\phi(X) : \mathbb{R}^d \longrightarrow \{1, 2, ..., K\}$, la cual pueda ser usada para predecir el índice de la clase para un nuevo dato X.

3.1. One Versus One (OVO)

Asuma que se tiene un método de clasificación binaria, ya sea MD, SVM, DWD o MDP; entonces para extender este método al caso multicategoría utilizando la estrategia OVO, se construyen $K(K-1)/2$ clasificadores binarios de este tipo, cada uno de los cuales clasifica a los datos de dos de las clases. El clasificador (i,j) es construido con los datos de la i-ésima y j-ésima clase. Existen varios métodos para combinar los resultados de los $K(K-1)/2$ clasificadores, uno de los más utilizados es *la estrategia de votos máximo ganador de Friedman*. Para un nuevo dato X, se calcula primero el total de votos para cada clase, evaluando a X en todos los clasificadores binarios construidos previamente. Si el clasificador (i,j) dice que X pertenece a la i-ésima clase, entonces el total de votos de esta clase se incrementa en uno; de otro modo el total de votos de la j-ésima clase se incrementa en uno. Entonces el nuevo dato X es clasificado en la clase con mayor número de votos.

En Huang et al. (2010) y Huang et al. (2013) se menciona que, en el caso del método DWD, para el (i,j)-ésimo clasificador, que es construido con los datos de la i-ésima y j-ésima clase, se resuelve el problema de clasificación binaria

$$\min_{U_{i,j},\beta_{i,j},\xi^{i,j}} \sum_{r:\delta_r=i \text{ ó } \delta_r=j} \left(\frac{1}{\rho_r} + M\xi_r^{i,j}\right),$$

$$\text{sujeto a} \quad \rho_r = \left(U_{i,j}^\top X_r + \beta_{i,j}\right) + \xi_r^{i,j} \quad \text{para } r \text{ tal que} \quad \delta_r = i,$$

$$\rho_r = -\left(U_{i,j}^\top X_r + \beta_{i,j}\right) + \xi_r^{i,j} \quad \text{para } r \text{ tal que} \quad \delta_r = j,$$

$$U_{i,j}^\top U_{i,j} \leq 1, \;\; \rho_r \geq 0, \;\; \xi_r^{i,j} \geq 0,$$

donde las variables $\xi^{i,j}$ son perturbaciones no negativas y $M > 0$ es un parámetro de penalización.

Si se tiene un nuevo dato X, se calcula el total de votos para cada clase, evaluando el dato en todos los clasificadores construidos. Por ejemplo, si sign $\left\{U_{i,j}^\top X + \beta_{i,j}\right\}$ indica que X está en la i-ésima clase, entonces el total de votos de la i-ésima clase es incrementado en uno; de lo contrario el total de votos de la j-ésima clase es incrementado en uno. Entonces el nuevo dato X es clasificado en la clase con mayor número de votos. En caso de empate de votos para varias clases, se toma la convención de seleccionar la clase con la suma máxima de las distancias de X a los hiperplanos separantes que clasificaron a X en esa clase.

3.2. One Versus The Rest (OVR)

Para extender alguno de los métodos de clasificación binaria MD, SVM, DWD o MDP al caso multicategoría mediante la metodología OVR, se construyen K clasificadores binarios, cada uno construido para distinguir (o clasificar) los datos de una categoría con los datos de todas las categorías restantes. Cuando se desea clasificar un nuevo dato, los K clasificadores son corridos y la clase del clasificador con respuesta más grande es elegido.

Como puede verse en Huang et al. (2010) y Huang et al. (2013), en el caso de DWD, para el i-ésimo clasificador DWD, el cual es construido con los datos de la i-ésima categoría, asignándoles un índice positivo $(+1)$, y a los datos de las categorías restantes, asignándoles un índice negativo (-1), se resuelve el siguiente problema de optimización:

$$\begin{aligned} &\min_{U_i,\beta_i,\xi^i} \sum_{r=1}^{N} \left(\frac{1}{\rho_r} + M\xi_r^i\right), \\ \text{sujeto a} \quad &\rho_r = \left(U_i^\top X_r + \beta_i\right) + \xi_r^i \quad \text{para } r \text{ tal que} \quad \delta_r = i, \\ &\rho_r = -\left(U_i^\top X_r + \beta_i\right) + \xi_r^i \quad \text{para } r \text{ tal que} \quad \delta_r \neq i, \\ &U_i^\top U_i \leq 1, \;\; \rho_r \geq 0, \;\; \xi_r^i \geq 0, \end{aligned}$$

donde las variables ξ^i son perturbaciones no negativas y $M > 0$ es un parámetro de penalización.

Después de resolver el problema de optimización anterior para cada clase, hay K funciones de decisión y entonces se dice que el nuevo dato X está en la clase que tenga el valor más grande de la función de decisión, es decir, la clase de X tiene índice igual a $\underset{i}{\text{Arg máx}} \left\{U_i^\top X + \beta_i\right\}$. Observe que el signo de $U_i^\top X + \beta_i$ determina si X está de un lado o del otro del hiperplano $U_i^\top Z + \beta_i = 0$; si es positivo se clasifica en la clase i y si es negativo se clasifica como perteneciente al resto de las clases. Además, $\left|U_i^\top X + \beta_i\right|$ es análogo a la distancia de X al hiperplano $U_i^\top Z + \beta_i$, la cual está dada por $\dfrac{\left|U_i^\top X + \beta_i\right|}{||U_i||}$. Como puede verse en Marron et al. (2007), si los datos son estrictamente linealmente separables y M es suficientemente grande, entonces la solución óptima de U_i tendrá longitud igual a 1. Por lo tanto, se cumple lo siguiente.

Observación 3.2.1. *En el caso estrictamente separable, si el método OVR clasifica al nuevo dato X en la clase i fue porque la distancia signada de X al hiperplano $U_i^\top Z + \beta_i = 0$ fue la máxima entre todas las distancias signadas de X a los hiperplanos $U_j^\top Z + \beta_j = 0$, para $j = 1, 2, ..., K$.*

3.3. Un solo Problema de Optimización

Como ya ha sido mencionado, una segunda estrategia para generalizar un método de clasificación binaria al caso multicategoría consiste en llevar a cabo un solo problema de optimización que envuelva a todas las clases al mismo tiempo. En Huang et al. (2010) y Huang et al. (2013) se propone cómo usar esta estrategia para generalizar el método DWD al caso multicategoría. A continuación se presenta en que consiste este método denominado Multiclass Distance Weigthed Discrimination (MDWD).

Una de las formas más naturales de representar a un clasificador multicategoría, es introduciendo un vector de funciones de decisión $f = (f_1, f_2, ..., f_K)$, donde cada componente representa a una clase. Para un nuevo dato X, su clase es estimada mediante la regla de clasificación $\widehat{\delta} = \text{Arg}\max_i f_i(X)$, donde $f_i(X) = U_i^\top X + \beta_i$. Se denotará por U a $(U_1, U_2, ..., U_K)$ y se considerará que $||U||^2 = \sum_{i=1}^{K} ||U_i||^2$.

En el caso binario, un dato (X, δ) es clasificado erróneamente si el margen $\delta f(X)$ es menor que 0, donde $f(X) = U^\top X + \beta$. Recuerde que en el caso de clasificación binaria, δ solo toma los valores $+1$ y -1, y $f(X)$ es positivo si el dato X está del mismo lado del hiperplano que la clase $+1$ y es negativo si X está del mismo lado del hiperplano que la clase -1.

En el caso multiclase un punto (X, δ) es clasificado erróneamente por el clasificador f si $\delta \neq \text{Arg}\max_i f_i(X)$, es decir, si alguna componente de $g(f(X), \delta)$ es menor que cero, donde $g(f(X), \delta) = \{f_\delta(X) - f_i(X) : i \neq \delta\}$. Que alguna componente de $g(f(X), \delta)$ sea menor que cero significa que $f_i(X) > f_\delta(X)$ para alguna $i \neq \delta$. Por lo tanto, una extensión natural de un clasificador basado en el margen, del caso binario al caso multiclase, es reemplazar el funcional margen $\delta f(X)$ por el funcional margen general $g(f(X), \delta)$. Una manera natural de hacer esto, es formular el método MDWD en términos del siguiente problema de optimización:

$$\begin{aligned} &\min_{U,\beta,\xi} \sum_{r=1}^{N} \sum_{i \neq j} \left(\frac{1}{\rho_{ij}^r} + M\xi_{ij}^r \right), \\ \text{sujeto a} \quad &\rho_{ij}^r = f_j(X_r) - f_i(X_r) + \xi_{ij}^r \quad \text{para} \;\; \delta_r = j, \;\; i \neq j \\ &\rho_{ij}^r \geq 0, \;\; \xi_{ij}^r \geq 0, \;\; \sum_{i=1}^{K} U_i = 0, \;\; \sum_{i=1}^{K} \beta_i = 0, \;\; ||U||^2 \leq 1. \end{aligned}$$

Note que la contribución del r-ésimo individuo del primer término en la función objetivo

es la suma de los inversos de las diferencias entre $f_{\delta_r}(X)$ y todas las otras funciones discriminantes evaluadas en X_r. El parámetro M en el segundo término controla la penalización en las variables ξ_{ij}^r; las cuales representan la cantidad de violación de clasificación. Similarmente al caso de clasificación binaria, el problema de optimización anterior puede ser reformulado como un problema convexo de segundo orden (SOCP). Si $K = 2$, puede ser mostrado que este problema se reduce al método binario DWD original.

Para SVM, la extensión del caso binario al multicategoría no es única. Algunas formas de hacerla son propuestas por Crammer y Singer (2000), Lee et al. (2004) y Weston y Watkins (1999).

3.4. Comportamiento Asintótico de los Métodos

En esta sección proporcionamos nuestros resultados que explican las propiedades asintóticas de las versiones multicategoría de los métodos MD, SVM, DWD y MDP mediante la metodología OVO, denotados por OVO-MD, OVO-SVM, OVO-DWD y OVO-MDP, respectivamente. Abordamos también el comportamiento asintótico de la versión multicategoría del método MD a través de la metodología OVR, denotada por OVR-MD. Exponemos en primer lugar la terminología y las propiedades que deben satisfacer los datos empleados en los resultados obtenidos.

Sean $n_1, n_2, ..., n_K$ enteros positivos. Sea $\mathcal{C}_j$ la clase de vectores aleatorios d-dimensionales independientes e idénticamente distribuidos X_r^j, para $r = 1, 2, ..., n_j$, con vector de medias μ_j y matriz de covarianzas Σ_j, para $j = 1, 2, ..., K$. Asuma además que las clases $\mathcal{C}_1$, $\mathcal{C}_2$,..., $\mathcal{C}_K$ son independientes, y que se cumplen los siguientes supuestos:

- En cada $\mathcal{C}_j$ se tiene la representación geométrica asintótica: existe $\sigma_j > 0$ tal que

$$\frac{||X_r^j - \mu_j||^2}{d} \xrightarrow{\mathbb{P}} \sigma_j^2 \qquad \text{y} \qquad \frac{||X_r^j - X_s^j||^2}{d} \xrightarrow{\mathbb{P}} 2\sigma_j^2 \tag{3.1}$$

  cuando $d \to \infty$, para toda $r, s = 1, 2, ..., n_j$ con $r \neq s$.

- Las medias poblacionales μ_j satisfacen

$$\frac{||\mu_i - \mu_j||^2}{d} \longrightarrow c_{ij}^2 \tag{3.2}$$

  cuando $d \to \infty$, para algunas constantes $c_{ij} > 0$ y para toda $i, j = 1, 2, ..., K$ con $i \neq j$. Note que $c_{ij} = c_{ji}$.

- Entre cualesquiera $\mathcal{C}_i$ y $\mathcal{C}_j$, con $i \neq j$, se tiene también la representación geométrica asintótica:

$$\frac{||X_r^i - X_s^j||^2}{d} \xrightarrow{\mathbb{P}} \ell_{ij}^2 := \sigma_i^2 + \sigma_j^2 + c_{ij}^2 \tag{3.3}$$

cuando $d \to \infty$, para toda $r = 1, 2, ..., n_i$ y para toda $s = 1, 2, ..., n_j$.

En el Capítulo 1 de esta tesis se expuso que los datos gaussianos estándar multivariados satisfacen (3.1) y también se expusieron condiciones para que otras distribuciones satisfagan (3.1)-(3.3). Similarmente al caso binario, si (3.1)-(3.3) se cumplen, los datos de $\mathcal{C}_j$, $j = 1, 2, \ldots, K$, son asintóticamente los vértices de un n_j-simplex cuando d tiende a infinito. Más aún, los datos de cualesquiera dos clases $\mathcal{C}_i$ and $\mathcal{C}_j$, $i \neq j$, son asintóticamente los vértices de un $(n_i + n_j)$-poliedro cuando d tiende a infinito, donde n_i de los vértices corresponden a un n_i-simplex, y los otros n_j vértices corresponden a un n_j-simplex.

Observe que bajo este contexto asintótico se cumplen los análogos a las expresiones obtenidas por Hall et al. (2005), dadas en (2.19), (2.22) y (2.24). Esto es, si

$$\overline{X}_j = \frac{1}{n_j} \sum_{r=1}^{n_j} X_r^j, \quad j = 1, 2, ..., K,$$

se tiene que:

- $$\frac{\left|\left|X_r^j - \overline{X}_j\right|\right|^2}{d} \xrightarrow{\mathbb{P}} \left(1 - \frac{1}{n_j}\right)\sigma_j^2 \tag{3.4}$$

  cuando $d \to \infty$, para toda $j = 1, 2, ..., K$ y para toda $r = 1, 2, ..., n_j$.

- $$\frac{\left|\left|X_r^i - \overline{X}_j\right|\right|^2}{d} \xrightarrow{\mathbb{P}} \sigma_i^2 + \frac{\sigma_j^2}{n_j} + c_{ij}^2 \tag{3.5}$$

  cuando $d \to \infty$, para toda $i, j = 1, 2, ..., K$ con $i \neq j$ y para toda $r = 1, 2, ..., n_i$.

- $$\frac{\left|\left|\overline{X}_i - \overline{X}_j\right|\right|^2}{d} \xrightarrow{\mathbb{P}} \frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j} + c_{ij}^2 \tag{3.6}$$

  cuando $d \to \infty$, para toda $i, j = 1, 2, ..., K$ con $i \neq j$.

3.4.1. Probabilidades de Clasificación Correcta vía OVO

En Hall et al. (2005) y Bolívar-Cimé (2021) se mostró que, bajo los supuestos (3.1)-(3.3), los hiperplanos separantes de los métodos de clasificación binaria MD, SVM y MDP coinciden asintóticamente cuando la dimensión tiende a infinito; y que el hiperplano separante del método de discriminación binaria DWD coincide con aquellos de MD, SVM y MDP únicamente cuando los tamaños muestrales de las dos clases son iguales. Esto explica intuitivamente nuestros siguientes dos teoremas, los cuales nos dicen que, en la extensión de cada uno de estos cuatro métodos al caso multicategoría vía OVO, bajo las condiciones (3.1)-(3.3), OVO-MD, OVO-SVM y OVO-MDP tienen el mismo comportamiento asintótico cuando la dimensión tiende a infinito, mientras que OVO-DWD pudiera tener un comportamiento diferente.

Teorema 3.4.1. *Considere las clases $\mathcal{C}_j$, $j = 1, 2, ..., K$, como antes junto con los supuestos dados en (3.1)-(3.3). Asuma que $\frac{\sigma_1^2}{n_1} \geq \frac{\sigma_2^2}{n_2} \geq \cdots \geq \frac{\sigma_K^2}{n_K}$; de no ser así, renombre a las clases de tal modo que esto se cumpla.*

a) Suponga que $c_{ij}^2 > \sigma_i^2/n_i - \sigma_j^2/n_j$ para todo par (i, j) con $1 \leq i < j \leq K$. Entonces la probabilidad de que un nuevo dato proveniente de cualquier población $\mathcal{C}_j$, $j = 1, 2, ..., K$, sea correctamente clasificado por el método OVO-MD, OVO-SVM u OVO-MDP converge a 1 cuando $d \to \infty$.

b) Suponga que $c_{ij}^2 < \sigma_i^2/n_i - \sigma_j^2/n_j$ para todo par (i, j) con $1 \leq i < j \leq K$. Asuma además que $c_{ir} > c_{is}$ para toda i, r, s con $i < r < s$, y que $c_{ir} > c_{js}$ para toda i, j, r, s con $i < j$, $i < r$, $j < s$. Entonces con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de cualquier población $\mathcal{C}_j$, $j = 1, 2, ..., K$, será clasificado por el método OVO-MD, OVO-SVM u OVO-MDP en la población $\mathcal{C}_K$.

Demostración. Denote por $H_{r,s}$ al hiperplano MD, SVM o MDP (dependiendo de si se usa el método OVO-MD, OVO-SVM u OVO-MDP) construido con los datos de las clases $\mathcal{C}_r$ y $\mathcal{C}_s$, para $r, s \in \{1, 2, ..., K\}$ con $r < s$. Sea $i \in \{1, 2, ..., K\}$ y sea X un nuevo dato que siga la distribución de la población $\mathcal{C}_i$.

a) Debido al Teorema 2.5.1 se tiene que la probabilidad de que X sea correctamente clasificado por el hiperplano $H_{r,s}$, con r ó s igual a i, converge a 1 cuando $d \to \infty$. Existen $(K - 1)$

hiperplanos $H_{r,s}$, con r ó s igual a i; por lo que el número de votos para la clase $\mathcal{C}_i$ tiende a $(K-1)$ cuando $d \to \infty$. Dado que para cualquier otra clase $\mathcal{C}_j$, con $j \neq i$, se tienen $(K-2)$ hiperplanos $H_{r,s}$, construidos con esta clase $\mathcal{C}_j$ y otra diferente a la clase $\mathcal{C}_i$, a lo más se tienen entonces $(K-2)$ votos para esta clase $\mathcal{C}_j$ cuando $d \to \infty$. De modo que la clase $\mathcal{C}_i$ es la que tiene mayor número de votos, $(K-1)$ en total, cuando $d \to \infty$; por lo cual, la probabilidad de que X sea correctamente clasificado en la clase $\mathcal{C}_i$ converge a 1 cuando $d \to \infty$.

b) En este caso, nuevamente por el Teorema 2.5.1, con probabilidad convergiendo a 1 cuando $d \to \infty$, el nuevo dato X será clasificado por el hiperplano $H_{r,i}$, con $r < i$, como perteneciente a la clase $\mathcal{C}_i$; y será clasificado por el hiperplano $H_{i,s}$, con $i < s$, como perteneciente a la clase $\mathcal{C}_s$.

Se verá que los hiperplanos $H_{r,K}$, con $r < K$, clasifican a X en la clase $\mathcal{C}_K$. Considere pues a los hiperplanos $H_{r,s}$ con $r < s$ y $r \neq i \neq s$. Suponga que estos hiperplanos $H_{r,s}$ son construidos con el método de clasificación binaria MD. De modo que, el nuevo dato X será clasificado por el hiperplano $H_{r,s}$ como perteneciente a la clase $\mathcal{C}_s$ si la distancia de X a la media muestral $\overline{X}_s$ de $\mathcal{C}_s$ es menor que la distancia de X a la media muestral $\overline{X}_r$ de $\mathcal{C}_r$; de lo contrario X será clasificado en la clase $\mathcal{C}_r$.

Como X tiene la misma distribución que X_m^i, por pertenecer a la población $\mathcal{C}_i$, se tiene entonces por (3.5) que

$$\frac{||X - \overline{X}_j||^2}{d} \xrightarrow{\mathbb{P}} \sigma_i^2 + \frac{\sigma_j^2}{n_j} + c_{ij}^2 \tag{3.7}$$

cuando $d \to \infty$, para toda $j = 1, 2, ..., K$ con $j \neq i$.

Así, $H_{r,s}$ clasificará a X como perteneciente a $\mathcal{C}_s$ si $\sigma_i^2 + \sigma_r^2/n_r + c_{ir}^2 > \sigma_i^2 + \sigma_s^2/n_s + c_{is}^2$, o equivalentemente si

$$\sigma_r^2/n_r + c_{ir}^2 > \sigma_s^2/n_s + c_{is}^2; \tag{3.8}$$

de otro modo lo clasificará en $\mathcal{C}_r$.

Dado que $r < s$, entonces $\sigma_r^2/n_r \geq \sigma_s^2/n_s$. Por las hipótesis acerca de las constantes c_{jm}, si $i < r < s$ entonces $c_{ir} > c_{is}$; si $r < i < s$ entonces $c_{ri} > c_{is}$; si $r < s < i$ entonces $c_{ri} > c_{si}$. Así, (3.8) se cumple en todos los casos, y se tiene entonces que con probabilidad convergiendo a 1 cuando $d \to \infty$, el nuevo dato X será clasificado en la clase $\mathcal{C}_s$.

Consecuentemente, todos los hiperplanos $H_{r,K}$ con $r < K$ clasificarán al nuevo dato X como perteneciente a la clase $\mathcal{C}_K$, y esta clase tendrá entonces $(K-1)$ votos, el número máximo

de votos que puede tener una clase, el resto de las clases tendrán a lo más $(K-2)$ votos; por lo tanto con probabilidad tendiendo a 1 cuando $d \to \infty$, el dato X será clasificado por el método OVO-MD como perteneciente a la clase $\mathcal{C}_K$.

Como ya se vio antes, bajo la representación geométrica asintótica de los datos, los hiperplanos SVM Y MDP coinciden con el hiperplano MD cuando $d \to \infty$. Por lo tanto, con probabilidad tendiendo a 1 cuando $d \to \infty$, el nuevo dato X será clasificado por el método OVO-SVM u OVO-MDP como perteneciente a la clase $\mathcal{C}_K$. ▲

Para el método de clasificación multicategoría OVO-DWD obtuvimos lo siguiente.

Teorema 3.4.2. *Considere las clases $\mathcal{C}_j$, $j = 1, 2, ..., K$, como antes junto con los supuestos dados en (3.1)-(3.3). Asuma que $\frac{\sigma_1^2}{n_1^{3/2}} \geq \frac{\sigma_2^2}{n_2^{3/2}} \geq \cdots \geq \frac{\sigma_K^2}{n_K^{3/2}}$, de no ser así, renombre a las clases de tal modo que esto se cumpla.*

a) Suponga que $c_{ij}^2 > (n_j/n_i)^{1/2}\,\sigma_i^2/n_i - \sigma_j^2/n_j$ para todo par (i,j) con $1 \leq i < j \leq K$. Entonces la probabilidad de que un nuevo dato proveniente de cualquier población $\mathcal{C}_j$, $j = 1, 2, ..., K$, sea correctamente clasificado por el método OVO-DWD converge a 1 cuando $d \to \infty$.

b) Suponga que $c_{ij}^2 < (n_j/n_i)^{1/2}\,\sigma_i^2/n_i - \sigma_j^2/n_j$ para todo par (i,j) con $1 \leq i < j \leq K$. Asuma además que $c_{ir} > c_{is}$ para toda i, r, s con $i < r < s$, y que $c_{ir} > c_{js}$ para toda i, j, r, s con $i < j$, $i < r$, $j < s$. Entonces con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de cualquier población $\mathcal{C}_j$, $j = 1, 2, ..., K$, será clasificado por el método OVO-DWD en la población $\mathcal{C}_K$.

Demostración. Denote por $H_{r,s}$ al hiperplano DWD construido con los datos de las clases $\mathcal{C}_r$ y $\mathcal{C}_s$, para $r, s \in \{1, 2, ..., K\}$ con $r < s$. Sea $i \in \{1, 2, ..., K\}$ y sea X un nuevo dato que siga la distribución de la población $\mathcal{C}_i$.

a) Observe que este caso queda demostrado aplicando el Teorema 2.5.2 y los mismos argumentos empleados en la prueba del inciso a) del Teorema 3.4.1.

b) Debido al Teorema 2.5.2, con probabilidad convergiendo a 1 cuando $d \to \infty$, el nuevo dato X será clasificado por el hiperplano $H_{r,i}$, con $r < i$, como perteneciente a la clase $\mathcal{C}_i$; y será clasificado por el hiperplano $H_{i,s}$, con $i < s$, como perteneciente a la clase $\mathcal{C}_s$.

Se probará que los hiperplanos $H_{r,K}$, con $r < K$, clasifican a X en la clase $\mathcal{C}_K$. Considere pues a los hiperplanos $H_{r,s}$ con $r < s$ y $r \neq i \neq s$. El siguiente argumento está basado principalmente de la prueba del Teorema 2.5.2.

Recuerde que la condición (3.1) dice que, después de reescalar por $d^{-1/2}$, los vectores de la clase $\mathcal{C}_j$, para $j = 1, 2, ..., K$, son asintóticamente localizados en los vértices de un n_j-simplex de aristas de longitud $\left(2\sigma_j^2\right)^{1/2}$. Considere así los simpleces límites correspondientes a las clases $\mathcal{C}_r$ y $\mathcal{C}_s$, llamados n_r-simplex y n_s-simplex, respectivamente. Refiérase a la Figura 1. Sean C_r y C_s los centroides del n_r-simplex y del n_s-simplex, respectivamente. Sea Q' la proyección ortogonal de X sobre la recta que une a C_r con C_s. Sea h la distancia de X a Q', y sean α' y γ' las distancias de Q' a C_s y C_r, respectivamente.

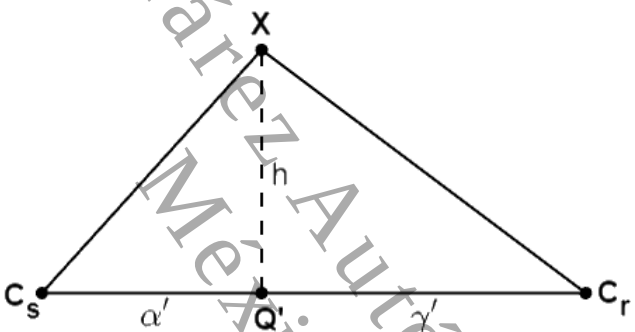


Figura 1: *Proyección ortogonal del nuevo dato X sobre el segmento C_sC_r.*

Observe que por (3.6), (3.7) y el teorema de Pitágoras se tiene que

$$(\alpha')^2 + h^2 = \sigma_i^2 + \sigma_s^2/n_s + c_{is}^2, \tag{3.9}$$

$$(\gamma')^2 + h^2 = \sigma_i^2 + \sigma_r^2/n_r + c_{ir}^2, \tag{3.10}$$

$$(\alpha' + \gamma')^2 = \sigma_r^2/n_r + \sigma_s^2/n_s + c_{rs}^2. \tag{3.11}$$

Restando la ecuación (3.10) a la ecuación (3.9) se obtiene que

$$(\alpha')^2 - (\gamma')^2 = \sigma_s^2/n_s - \sigma_r^2/n_r + c_{is}^2 - c_{ir}^2. \tag{3.12}$$

Sumando las ecuaciones (3.11) y (3.12), y posteriormente restando la ecuación (3.11) a la ecuación (3.12) se obtiene, respectivamente que

$$2\alpha'(\alpha' + \gamma') = 2\sigma_s^2/n_s + c_{rs}^2 + c_{is}^2 - c_{ir}^2,$$

$$2\gamma'(\alpha'+\gamma') = 2\sigma_r^2/n_r + c_{rs}^2 + c_{ir}^2 - c_{is}^2,$$

de lo cual se deduce que

$$\frac{\alpha'}{\gamma'} = \frac{2\sigma_s^2/n_s + c_{rs}^2 + c_{is}^2 - c_{ir}^2}{2\sigma_r^2/n_r + c_{rs}^2 + c_{ir}^2 - c_{is}^2}.$$

Sea Q el punto donde el hiperplano $H_{r,s}$ corta ortogonalmente al segmento de recta que une a C_r con C_s. De modo que si α y γ son las distancias de Q a C_s y C_r, respectivamente, entonces se debe cumplir la condición (**2.16**), esto es,

$$\frac{\alpha}{\gamma} = \left(\frac{n_s}{n_r}\right)^{1/2}.$$

El nuevo dato X será clasificado como perteneciente a la clase $\mathcal{C}_s$ si se encuentra del mismo lado del hiperplano $H_{r,s}$ que C_s, es decir, si $\dfrac{\alpha'}{\gamma'} < \dfrac{\alpha}{\gamma}$, o equivalentemente si

$$\frac{2\sigma_s^2/n_s + c_{rs}^2 + c_{is}^2 - c_{ir}^2}{2\sigma_r^2/n_r + c_{rs}^2 + c_{ir}^2 - c_{is}^2} < \left(\frac{n_s}{n_r}\right)^{1/2}; \qquad (\mathbf{3.13})$$

de lo contrario será clasificado en la clase $\mathcal{C}_r$.

Dado que $r < s$, entonces $c_{rs}^2 < (n_s/n_r)^{1/2}\,\sigma_r^2/n_r - \sigma_s^2/n_s$. Por las hipótesis acerca de las constantes c_{jm}, si $i < r < s$ entonces $c_{ir} > c_{is}$; si $r < i < s$ entonces $c_{ri} > c_{is}$; si $r < s < i$ entonces $c_{ri} > c_{si}$. De aquí se obtiene que

$$0 < 2\left[\left(\frac{n_s}{n_r}\right)^{1/2}\frac{\sigma_r^2}{n_r} - \frac{\sigma_s^2}{n_s}\right] - c_{rs}^2 + \left(\frac{n_s}{n_r}\right)^{1/2} c_{rs}^2 + \left(\frac{n_s}{n_r}\right)^{1/2}\left(c_{ir}^2 - c_{is}^2\right) + \left(c_{ir}^2 - c_{is}^2\right),$$

que es equivalente a

$$\frac{2\sigma_s^2}{n_s} + c_{rs}^2 + c_{is}^2 - c_{ir}^2 < \left(\frac{n_s}{n_r}\right)^{1/2}\left(\frac{2\sigma_r^2}{n_r} + c_{rs}^2 + c_{ir}^2 - c_{is}^2\right);$$

por lo tanto se cumple (**3.13**) y se tiene así que, con probabilidad tendiendo a 1 cuando $d \to \infty$, el nuevo dato X será clasificado en la clase $\mathcal{C}_s$.

Consecuentemente, todos los hiperplanos $H_{r,K}$ con $r < K$ clasificarán al nuevo dato X como perteneciente a la clase $\mathcal{C}_K$, y esta clase tendrá entonces $(K-1)$ votos, el número máximo de votos que puede tener una clase, el resto de las clases tendrán a lo más $(K-2)$ votos; por lo tanto con probabilidad tendiendo a 1 cuando $d \to \infty$, el dato X será clasificado por el método OVO-DWD como perteneciente a la clase $\mathcal{C}_K$. ▲

Note que en el caso donde $n_r \equiv n$ para $r = 1, 2, \ldots, K$, si $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_K^2 > 0$ y $c_{ij}^2 > (\sigma_i^2 - \sigma_j^2)/n$ para todo (i, j) con $1 \leq i < j \leq K$, por a) de los teoremas 3.4.1 y 3.4.2, los métodos OVO-MD, OVO-SVM, OVO-MDP y OVO-DWD aseguran una clasificación asintóticamente correcta de un nuevo dato de cualquier clase, cuando d tiende a infinito. Así, en este caso los cuatro métodos tienen el mismo comportamiento asintótico.

Sin embargo, existen casos donde los parámetros satisfacen las condiciones de a) del Teorema 3.4.1 y al mismo tiempo cumplen las condiciones de b) del Teorema 3.4.2. Entonces, en esta situación los métodos OVO-MD, OVO-SVM, OVO-MDP tienen un comportamiento asintótico diferente de OVO-DWD en términos de las probabilidades de clasificación correcta. Por ejemplo, suponga que $K = 3$, $n_3 > n_2 > n_1$ y $\sigma_1^2/n_1 > \sigma_2^2/n_2 > \sigma_3^2/n_3 > 0$. Sea

$$R_{ij} = \frac{\sigma_i^2}{n_i} - \frac{\sigma_j^2}{n_j}, \quad M_{ij} = \left(\frac{n_j}{n_i}\right)^{1/2} \frac{\sigma_i^2}{n_i} - \frac{\sigma_j^2}{n_j}, \quad \text{para } i < j.$$

Entonces $R_{ij} < M_{ij}$ para todo $i < j$. Además, $R_{23} < R_{13} < R_{12} < M_{12} < M_{13}$ y $M_{23} < M_{13}$. Así, si se eligen constantes positivas c_{12}, c_{13} y c_{23} tales que

$$R_{13} < c_{13}^2 < R_{12} < c_{12}^2 < M_{12}, \quad R_{23} < c_{23}^2 < \text{mín}\{M_{23}, R_{13}\},$$

las condiciones de a) del Teorema 3.4.1 y b) del Teorema 3.4.2 son satisfechas. Por lo tanto, en este caso los métodos OVO-MD, OVO-SVM, OVO-MDP aseguran clasificación asintóticamente correcta de un nuevo dato de cualquier clase, mientras que OVO-DWD aseguran clasificación asintóticamente correcta únicamente para datos nuevo de la clase 3 y clasificación incorrecta para datos nuevos de las clases restantes, cuando d tiende a infinito.

3.4.2. Probabilidades de Clasificación Correcta: MD vía OVR

Abordaremos ahora nuestros resultados acerca de las probabilidades de clasificación correcta del método de discriminación multicategoría OVR-MD, en el contexto de datos de dimensión alta. Antes de ello, introducimos algunos términos que usaremos frecuentemente.

Para $j \in \{1, 2, ..., K\}$, definimos los siguientes términos:

$$N_{-j} := \sum_{r=1, r\neq j}^{K} n_r, \quad \overline{X}_{-j} := \frac{1}{N_{-j}} \sum_{r=1, r\neq j}^{K} n_r \overline{X}_r \quad \text{y} \quad \mathcal{C}_{-j} := \bigcup_{r=1, r\neq j}^{K} \mathcal{C}_r. \tag{3.14}$$

Observe que $\overline{X}_{-j}$ es un promedio ponderado en $\mathcal{C}_{-j}$. Además, denotamos con MD_j al hiperplano construido con los datos de $\mathcal{C}_j$ versus los datos de $\mathcal{C}_{-j}$, a través del método de clasificación binaria MD. De modo que MD_j tiene por ecuación

$$U_j^\top Z + \beta_j = 0,$$

donde por (2.5),

$$U_j = \frac{\overline{X}_j - \overline{X}_{-j}}{||\overline{X}_j - \overline{X}_{-j}||} \quad \text{y} \quad \beta_j = -U_j^\top \left(\frac{\overline{X}_j + \overline{X}_{-j}}{2}\right).$$

Así, la identidad

$$2\langle u - v, u - w\rangle = ||u - v||^2 + ||u - w||^2 - ||v - w||^2, \tag{3.15}$$

para cualesquiera $u, v, w \in \mathbb{R}^d$, implica que

$$\begin{aligned}
U_j^\top Z + \beta_j &= \langle U_j, Z\rangle - \left\langle U_j, \frac{1}{2}\left(\overline{X}_j + \overline{X}_{-j}\right)\right\rangle \\
&= \left\langle U_j, Z - \frac{1}{2}\left(\overline{X}_j + \overline{X}_{-j}\right)\right\rangle \\
&= \frac{1}{2\,||\overline{X}_j - \overline{X}_{-j}||}\left\langle \overline{X}_j - \overline{X}_{-j}, 2Z - \left(\overline{X}_j + \overline{X}_{-j}\right)\right\rangle \\
&= \frac{1}{2\,||\overline{X}_j - \overline{X}_{-j}||}\left[\langle \overline{X}_j - \overline{X}_{-j}, Z - \overline{X}_j\rangle + \langle \overline{X}_j - \overline{X}_{-j}, Z - \overline{X}_{-j}\rangle\right] \\
&= \frac{1}{4\,||\overline{X}_j - \overline{X}_{-j}||}\left[2\left\langle \overline{X}_{-j} - \overline{X}_j, \overline{X}_{-j} - Z\right\rangle - 2\left\langle \overline{X}_j - \overline{X}_{-j}, \overline{X}_j - Z\right\rangle\right] \\
&= \frac{1}{4\,||\overline{X}_j - \overline{X}_{-j}||}\Big[||\overline{X}_{-j} - \overline{X}_j||^2 + ||\overline{X}_{-j} - Z||^2 - ||\overline{X}_j - Z||^2 \\
&\qquad - \left(||\overline{X}_j - \overline{X}_{-j}||^2 + ||\overline{X}_j - Z||^2 - ||\overline{X}_{-j} - Z||^2\right)\Big] \\
&= \frac{1}{4\,||\overline{X}_j - \overline{X}_{-j}||}\left[2\,||\overline{X}_{-j} - Z||^2 - 2\,||\overline{X}_j - Z||^2\right] \\
&= \frac{||Z - \overline{X}_{-j}||^2 - ||Z - \overline{X}_j||^2}{2\,||\overline{X}_j - \overline{X}_{-j}||}.
\end{aligned} \tag{3.16}$$

Note que (3.16) es la *distancia signada* de Z al hiperplano MD_j, y se denotará por $DS(Z, MD_j)$.

Observe que por (3.14) se tiene que para toda $Z \in \mathbb{R}^d$ y para cualquier $j \in \{1, 2, ..., K\}$

$$||Z - \overline{X}_{-j}||^2 = \frac{1}{N_{-j}^2}\left|\left|N_{-j}Z - \sum_{r=1, r\neq j}^{K} n_r \overline{X}_r\right|\right|^2 = \frac{1}{N_{-j}^2}\left|\left|\sum_{r=1, r\neq j}^{K} n_r Z - \sum_{r=1, r\neq j}^{K} n_r \overline{X}_r\right|\right|^2;$$

resultando así que

$$\left|\left|Z-\overline{X}_{-j}\right|\right|^2=\frac{1}{N_{-j}^2}\left|\left|\sum_{r=1,r\neq j}^{K}n_r\left(Z-\overline{X}_r\right)\right|\right|^2. \tag{3.17}$$

Proporcionamos la siguiente igualdad que nos ayudará a determinar la convergencia en probabilidad de las distancias signadas.

Lema 3.4.1. *Sea $m\geq 2$ un número entero y sea $j\in\{1,2,...,m\}$ fijo. Para cualesquiera vectores $Z,Z_1,Z_2,...,Z_m\in\mathbb{R}^d$ y cualesquiera constantes $a_1,a_2,...,a_m$ se satisface que*

$$\begin{aligned}\left|\left|\sum_{r=1}^{m}a_r\left(Z-Z_r\right)\right|\right|^2&=\sum_{r=1}^{m}a_r^2\left|\left|Z-Z_r\right|\right|^2\\&+\sum_{r=2}^{m}\sum_{s=1}^{r-1}a_ra_s\left(\left|\left|Z-Z_r\right|\right|^2+\left|\left|Z-Z_s\right|\right|^2-\left|\left|Z_r-Z_s\right|\right|^2\right).\end{aligned} \tag{3.18}$$

Demostración. Se procederá por inducción. Note que (3.18) se cumple para $m=2$, puesto que (3.15) implica

$$\begin{aligned}\left|\left|\sum_{r=1}^{2}a_r\left(Z-Z_r\right)\right|\right|^2&=\sum_{r=1}^{2}a_r^2\left|\left|Z-Z_r\right|\right|^2+2a_2a_1\left\langle Z-Z_2,Z-Z_1\right\rangle\\&=\sum_{r=1}^{2}a_r^2\left|\left|Z-Z_r\right|\right|^2+\sum_{r=2}^{2}\sum_{s=1}^{r-1}a_ra_s\left(\left|\left|Z-Z_r\right|\right|^2+\left|\left|Z-Z_s\right|\right|^2-\left|\left|Z_r-Z_s\right|\right|^2\right).\end{aligned}$$

Nuevamente por la identidad (3.15), si la igualdad (3.18) se satisface para m, entonces se tiene que para $m+1$

$$\begin{aligned}&\left|\left|\sum_{r=1}^{m+1}a_r\left(Z-Z_r\right)\right|\right|^2\\=&\left|\left|\sum_{r=1}^{m}a_r\left(Z-Z_r\right)\right|\right|^2+\left|\left|a_{m+1}\left(Z-Z_{m+1}\right)\right|\right|^2+2\sum_{r=1}^{m}a_{m+1}a_r\left\langle Z-Z_{m+1},Z-Z_r\right\rangle\\=&\sum_{r=1}^{m}a_r^2\left|\left|Z-Z_r\right|\right|^2+\sum_{r=2}^{m}\sum_{s=1}^{r-1}a_ra_s\left(\left|\left|Z-Z_r\right|\right|^2+\left|\left|Z-Z_s\right|\right|^2-\left|\left|Z_r-Z_s\right|\right|^2\right)\\+&\ a_{m+1}^2\left|\left|Z-Z_{m+1}\right|\right|^2+\sum_{s=1}^{m}a_{m+1}a_s\left(\left|\left|Z-Z_{m+1}\right|\right|^2+\left|\left|Z-Z_s\right|\right|^2-\left|\left|Z_{m+1}-Z_s\right|\right|^2\right)\\=&\sum_{r=1}^{m+1}a_r^2\left|\left|Z-Z_r\right|\right|^2+\sum_{r=2}^{m+1}\sum_{s=1}^{r-1}a_ra_s\left(\left|\left|Z-Z_r\right|\right|^2+\left|\left|Z-Z_s\right|\right|^2-\left|\left|Z_r-Z_s\right|\right|^2\right).\end{aligned}$$

Por lo tanto, (3.18) es válida para toda $m\geq 2$. ▲

El siguiente teorema es uno de los principales resultados de esta tesis. Determinamos allí los límites en probabilidad de las distancias signadas.

Teorema 3.4.3. *Considere las clases $\mathcal{C}_j$, $j = 1, 2, ..., K$, como antes junto con los supuestos dados en (3.1)-(3.3). Suponga que X es un nuevo dato que sigue la distribución de la población $\mathcal{C}_i$, para algún $i \in \{1, 2, ..., K\}$ fijo. Entonces la distancia signada de $X \in \mathcal{C}_i$ al hiperplano MD_j, $DS(X, MD_j)$, $j = 1, 2, ..., K$, satisface*

$$\frac{DS(X, MD_j)}{d^{1/2}} \xrightarrow{\mathbb{P}}$$

$$DS_{ij} := \begin{cases} \dfrac{1}{D_i}\left[\displaystyle\sum_{\substack{r=1\\r\neq i}}^{K} n_r^2 c_{ir}^2 + \sum_{\substack{r=2\\r\neq i}}^{K}\sum_{\substack{s=1\\s\neq i}}^{r-1} n_r n_s \left(c_{ir}^2 + c_{is}^2 - c_{rs}^2\right) - R_i\right] & si \quad j = i, \\ \dfrac{1}{D_j}\left[\displaystyle\sum_{\substack{r=1\\r\neq i,j}}^{K} n_r^2 c_{ir}^2 + \sum_{\substack{r=2\\r\neq i,j}}^{K}\sum_{\substack{s=1\\s\neq i,j}}^{r-1} n_r n_s \left(c_{ir}^2 + c_{is}^2 - c_{rs}^2\right) - N_{-j}^2 c_{ij}^2 - R_j\right] & si \quad j \neq i; \end{cases} \tag{3.19}$$

cuando $d \to \infty$, donde (para $j = 1, 2, ..., K$)

$$R_j := \left(\frac{\sigma_j^2}{n_j}\right) N_{-j}^2 - \sum_{\substack{r=1\\r\neq j}}^{K} n_r \sigma_r^2 \tag{3.20}$$

y

$$D_j := 2N_{-j}\left[\sum_{\substack{r=1\\r\neq j}}^{K} n_r^2 c_{jr}^2 + \sum_{\substack{r=2\\r\neq j}}^{K}\sum_{\substack{s=1\\s\neq j}}^{r-1} n_r n_s \left(c_{jr}^2 + c_{js}^2 - c_{rs}^2\right) + \left(\frac{\sigma_j^2}{n_j}\right) N_{-j}^2 + \sum_{\substack{r=1\\r\neq j}}^{K} n_r \sigma_r^2\right]^{1/2}. \tag{3.21}$$

Demostración. Debido a las igualdades obtenidas en (3.17) y (3.18) se tiene que

$$\begin{aligned} \left|\left|Z - \overline{X}_{-j}\right|\right|^2 &= \frac{1}{N_{-j}^2}\left[\sum_{\substack{r=1\\r\neq j}}^{K} n_r^2 \left|\left|Z - \overline{X}_r\right|\right|^2\right. \\ &+ \left.\sum_{\substack{r=2\\r\neq j}}^{K}\sum_{\substack{s=1\\s\neq j}}^{r-1} n_r n_s \left(\left|\left|Z - \overline{X}_r\right|\right|^2 + \left|\left|Z - \overline{X}_s\right|\right|^2 - \left|\left|\overline{X}_r - \overline{X}_s\right|\right|^2\right)\right], \end{aligned} \tag{3.22}$$

para cualquier $Z \in \mathbb{R}^d$ y para cualquier $j \in \{1, 2, ..., K\}$.

Se abordará en primer lugar la convergencia en probabilidad de la distancia signada de X al hiperplano MD_j, para $j = i$. En efecto, como X tiene la misma distribución que X_m^i, por pertenecer a la población $\mathcal{C}_i$, se tiene entonces por (3.5) que

$$\frac{\left|\left|X - \overline{X}_r\right|\right|^2}{d} \xrightarrow{\mathbb{P}} \sigma_i^2 + \frac{\sigma_r^2}{n_r} + c_{ir}^2, \tag{3.23}$$

cuando $d \to \infty$, para toda $r = 1, 2, ..., K$ con $r \neq i$. De manera que, haciendo uso de las expresiones obtenidas en (3.6), (3.23) y la igualdad (3.22), con $Z = X$, se tiene que

$$\begin{aligned}
\frac{\left|\left|X - \overline{X}_{-i}\right|\right|^2}{d} \xrightarrow{\mathbb{P}} & N_{-i}^{-2} \sum_{\substack{r=1 \\ r\neq i}}^{K} n_r^2 \left(\sigma_i^2 + \frac{\sigma_r^2}{n_r} + c_{ir}^2\right) \\
& + N_{-i}^{-2} \sum_{\substack{r=2 \\ r\neq i}}^{K} \sum_{\substack{s=1 \\ s\neq i}}^{r-1} n_r n_s \left[\sigma_i^2 + \frac{\sigma_r^2}{n_r} + c_{ir}^2 + \sigma_i^2 + \frac{\sigma_s^2}{n_s} + c_{is}^2 - \left(\frac{\sigma_r^2}{n_r} + \frac{\sigma_s^2}{n_s} + c_{rs}^2\right)\right] \\
= & N_{-i}^{-2} \sum_{\substack{r=1 \\ r\neq i}}^{K} n_r^2 \left(\sigma_i^2 + \frac{\sigma_r^2}{n_r} + c_{ir}^2\right) + N_{-i}^{-2} \sum_{\substack{r=2 \\ r\neq i}}^{K} \sum_{\substack{s=1 \\ s\neq i}}^{r-1} n_r n_s \left(2\sigma_i^2 + c_{ir}^2 + c_{is}^2 - c_{rs}^2\right) \\
= & N_{-i}^{-2} \left(\sum_{\substack{r=1 \\ r\neq i}}^{K} n_r^2 + 2 \sum_{\substack{r=2 \\ r\neq i}}^{K} \sum_{\substack{s=1 \\ s\neq i}}^{r-1} n_r n_s\right) \sigma_i^2 + N_{-i}^{-2} \sum_{\substack{r=1 \\ r\neq i}}^{K} n_r \sigma_r^2 \\
& + N_{-i}^{-2} \left[\sum_{\substack{r=1 \\ r\neq i}}^{K} n_r^2 c_{ir}^2 + \sum_{\substack{r=2 \\ r\neq i}}^{K} \sum_{\substack{s=1 \\ s\neq i}}^{r-1} n_r n_s \left(c_{ir}^2 + c_{is}^2 - c_{rs}^2\right)\right] \\
= & N_{-i}^{-2} \left[\sum_{\substack{r=1 \\ r\neq i}}^{K} n_r^2 c_{ir}^2 + \sum_{\substack{r=2 \\ r\neq i}}^{K} \sum_{\substack{s=1 \\ s\neq i}}^{r-1} n_r n_s \left(c_{ir}^2 + c_{is}^2 - c_{rs}^2\right) + \sum_{\substack{r=1 \\ r\neq i}}^{K} n_r \sigma_r^2\right] + \sigma_i^2, \quad (3.24)
\end{aligned}$$

cuando $d \to \infty$, puesto que

$$N_{-j}^2 = \left(\sum_{\substack{r=1 \\ r\neq j}}^{K} n_r\right)^2 = \sum_{\substack{r=1 \\ r\neq j}}^{K} n_r^2 + 2 \sum_{\substack{r=2 \\ r\neq j}}^{K} \sum_{\substack{s=1 \\ s\neq j}}^{r-1} n_r n_s, \qquad \text{para toda } \; j = 1, 2, ..., K. \tag{3.25}$$

Nuevamente, como X pertenece a $\mathcal{C}_i$, se tiene entonces por (3.1) y (3.4) que X, cualquier X_m^i y la media muestral $\overline{X}_i$ tienden a formar un triángulo rectángulo cuya hipotenusa es el segmento de recta que une a X con X_m^i. De manera que, por el teorema de Pitágoras se deduce que

$$\frac{\left|\left|X - \overline{X}_i\right|\right|^2}{d} \xrightarrow{\mathbb{P}} 2\sigma_i^2 - \left(1 - \frac{1}{n_i}\right)\sigma_i^2 = \left(1 + \frac{1}{n_i}\right)\sigma_i^2, \tag{3.26}$$

cuando $d \to \infty$. Así, (3.24) y (3.26) implican que

$$\frac{\left|\left|X-\overline{X}_{-i}\right|\right|^2-\left|\left|X-\overline{X}_{i}\right|\right|^2}{d} \xrightarrow{\mathbb{P}} N_{-i}^{-2}\left[\sum_{\substack{r=1\\r\neq i}}^{K} n_r^2 c_{ir}^2+\sum_{\substack{r=2\\r\neq i}}^{K}\sum_{\substack{s=1\\s\neq i}}^{r-1} n_r n_s\left(c_{ir}^2+c_{is}^2-c_{rs}^2\right)\right] + N_{-i}^{-2}\left(\sum_{\substack{r=1\\r\neq i}}^{K} n_r\sigma_r^2\right)-\frac{\sigma_i^2}{n_i}, \quad \text{cuando } d\to\infty. \tag{3.27}$$

Análogamente, utilizando las convergencias dadas en (3.6) y aplicando una vez más la igualdad obtenida en (3.22), con $Z=\overline{X}_j$, se obtiene que

$$\begin{aligned}\frac{\left|\left|\overline{X}_j-\overline{X}_{-j}\right|\right|^2}{d} &\xrightarrow{\mathbb{P}} N_{-j}^{-2}\sum_{\substack{r=1\\r\neq j}}^{K} n_r^2\left(\frac{\sigma_j^2}{n_j}+\frac{\sigma_r^2}{n_r}+c_{jr}^2\right)\\
&+ N_{-j}^{-2}\sum_{\substack{r=2\\r\neq j}}^{K}\sum_{\substack{s=1\\s\neq j}}^{r-1} n_r n_s\left[\frac{\sigma_j^2}{n_j}+\frac{\sigma_r^2}{n_r}+c_{jr}^2+\frac{\sigma_j^2}{n_j}+\frac{\sigma_s^2}{n_s}+c_{js}^2-\frac{\sigma_r^2}{n_r}-\frac{\sigma_s^2}{n_s}-c_{rs}^2\right]\\
&= N_{-j}^{-2}\left[\sum_{\substack{r=1\\r\neq j}}^{K} n_r^2 c_{jr}^2+\sum_{\substack{r=2\\r\neq j}}^{K}\sum_{\substack{s=1\\s\neq j}}^{r-1} n_r n_s\left(c_{jr}^2+c_{js}^2-c_{rs}^2\right)\right]\\
&+ N_{-j}^{-2}\left(\sum_{\substack{r=1\\r\neq j}}^{K} n_r\sigma_r^2\right)+N_{-j}^{-2}\left(\sum_{\substack{r=1\\r\neq j}}^{K} n_r^2+2\sum_{\substack{r=2\\r\neq j}}^{K}\sum_{\substack{s=1\\s\neq j}}^{r-1} n_r n_s\right)\frac{\sigma_j^2}{n_j}\\
&= N_{-j}^{-2}\left[\sum_{\substack{r=1\\r\neq j}}^{K} n_r^2 c_{jr}^2+\sum_{\substack{r=2\\r\neq j}}^{K}\sum_{\substack{s=1\\s\neq j}}^{r-1} n_r n_s\left(c_{jr}^2+c_{js}^2-c_{rs}^2\right)\right]\\
&+ N_{-j}^{-2}\left(\sum_{\substack{r=1\\r\neq j}}^{K} n_r\sigma_r^2\right)+\frac{\sigma_j^2}{n_j}, \qquad \text{cuando } d\to\infty.\end{aligned} \tag{3.28}$$

Observe que (3.28) es válida para cualquier $j\in\{1,2,...,K\}$.

Note así que, la igualdad

$$\frac{\left|\left|X-\overline{X}_{-j}\right|\right|^2-\left|\left|X-\overline{X}_{j}\right|\right|^2}{2d^{1/2}\left|\left|\overline{X}_j-\overline{X}_{-j}\right|\right|}=\left(\frac{2\left|\left|\overline{X}_j-\overline{X}_{-j}\right|\right|}{d^{1/2}}\right)^{-1}\left(\frac{\left|\left|X-\overline{X}_{-j}\right|\right|^2-\left|\left|X-\overline{X}_{j}\right|\right|^2}{d}\right) \tag{3.29}$$

en conjunto con (3.27) y (3.28) implican la convergencia en probabilidad (3.19), cuando $j=i$.

Ahora, se procederá a ver que la distancia signada de X al hiperplano MD_j, para $j \neq i$, cumple efectivamente (3.19). En efecto, observe que de (3.22), con $Z = X$, resulta que

$$\begin{aligned} \left|\left|X-\overline{X}_{-j}\right|\right|^2 &= N_{-j}^{-2}\left(n_i^2\left|\left|X-\overline{X}_i\right|\right|^2+\sum_{\substack{r=1\\r\neq i,j}}^{K} n_r^2\left|\left|X-\overline{X}_r\right|\right|^2\right) \\ &+ N_{-j}^{-2}\sum_{\substack{s=1\\s\neq i,j}}^{K} n_i n_s\left(\left|\left|X-\overline{X}_i\right|\right|^2+\left|\left|X-\overline{X}_s\right|\right|^2-\left|\left|\overline{X}_i-\overline{X}_s\right|\right|^2\right) \\ &+ N_{-j}^{-2}\sum_{\substack{r=2\\r\neq i,j}}^{K}\sum_{\substack{s=1\\s\neq i,j}}^{r-1} n_r n_s\left(\left|\left|X-\overline{X}_r\right|\right|^2+\left|\left|X-\overline{X}_s\right|\right|^2-\left|\left|\overline{X}_r-\overline{X}_s\right|\right|^2\right), \end{aligned} \tag{3.30}$$

ya que para $j \neq i$,

$$\begin{aligned} \sum_{\substack{r=2\\r\neq j}}^{K}\sum_{\substack{s=1\\s\neq j}}^{r-1}\left\langle n_r\left(X-\overline{X}_r\right), n_s\left(X-\overline{X}_s\right)\right\rangle &= \sum_{\substack{s=1\\s\neq i,j}}^{K}\left\langle n_i\left(X-\overline{X}_i\right), n_s\left(X-\overline{X}_s\right)\right\rangle \\ &+ \sum_{\substack{r=2\\r\neq i,j}}^{K}\sum_{\substack{s=1\\s\neq i,j}}^{r-1}\left\langle n_r\left(X-\overline{X}_r\right), n_s\left(X-\overline{X}_s\right)\right\rangle. \end{aligned}$$

De modo que, las expresiones dadas en (3.6), (3.23), (3.26) y la igualdad (3.30), llevan a que

$$\begin{aligned} \frac{\left|\left|X-\overline{X}_{-j}\right|\right|^2}{d} &\xrightarrow{\mathbb{P}} N_{-j}^{-2}\left[n_i^2\left(1+\frac{1}{n_i}\right)\sigma_i^2+\sum_{\substack{r=1\\r\neq i,j}}^{K} n_r^2\left(\sigma_i^2+\frac{\sigma_r^2}{n_r}+c_{ir}^2\right)\right] \\ &+ N_{-j}^{-2}\sum_{\substack{s=1\\s\neq i,j}}^{K} n_i n_s\left[\left(1+\frac{1}{n_i}\right)\sigma_i^2+\left(\sigma_i^2+\frac{\sigma_s^2}{n_s}+c_{is}^2\right)-\left(\frac{\sigma_i^2}{n_i}+\frac{\sigma_s^2}{n_s}+c_{is}^2\right)\right] \\ &+ N_{-j}^{-2}\sum_{\substack{r=2\\r\neq i,j}}^{K}\sum_{\substack{s=1\\s\neq i,j}}^{r-1} n_r n_s\left[\sigma_i^2+\frac{\sigma_r^2}{n_r}+c_{ir}^2+\sigma_i^2+\frac{\sigma_s^2}{n_s}+c_{is}^2-\left(\frac{\sigma_r^2}{n_r}+\frac{\sigma_s^2}{n_s}+c_{rs}^2\right)\right] \\ &= N_{-j}^{-2}\left[n_i^2\left(1+\frac{1}{n_i}\right)\sigma_i^2+\sum_{\substack{r=1\\r\neq i,j}}^{K} n_r^2\left(\sigma_i^2+\frac{\sigma_r^2}{n_r}+c_{ir}^2\right)\right] \\ &+ N_{-j}^{-2}\left(\sum_{\substack{s=1\\s\neq i,j}}^{K} n_i n_s+\sum_{\substack{r=2\\r\neq i,j}}^{K}\sum_{\substack{s=1\\s\neq i,j}}^{r-1} n_r n_s\right)\left(2\sigma_i^2\right) \\ &+ N_{-j}^{-2}\sum_{\substack{r=2\\r\neq i,j}}^{K}\sum_{\substack{s=1\\s\neq i,j}}^{r-1} n_r n_s\left(c_{ir}^2+c_{is}^2-c_{rs}^2\right) \end{aligned}$$

$$
\begin{aligned}
&= N_{-j}^{-2}\left[\sum_{\substack{r=1\\ r\neq i,j}}^{K} n_r^2 c_{ir}^2 + \sum_{\substack{r=2\\ r\neq i,j}}^{K}\sum_{\substack{s=1\\ s\neq i,j}}^{r-1} n_r n_s\left(c_{ir}^2 + c_{is}^2 - c_{rs}^2\right)\right] \\
&+ N_{-j}^{-2}\left(\sum_{\substack{r=1\\ r\neq j}}^{K} n_r^2 + 2\sum_{\substack{r=2\\ r\neq j}}^{K}\sum_{\substack{s=1\\ s\neq j}}^{r-1} n_r n_s\right)\sigma_i^2 + N_{-j}^{-2}\sum_{\substack{r=1\\ r\neq j}}^{K} n_r\sigma_r^2 \\
&= N_{-j}^{-2}\left[\sum_{\substack{r=1\\ r\neq i,j}}^{K} n_r^2 c_{ir}^2 + \sum_{\substack{r=2\\ r\neq i,j}}^{K}\sum_{\substack{s=1\\ s\neq i,j}}^{r-1} n_r n_s\left(c_{ir}^2 + c_{is}^2 - c_{rs}^2\right) + \sum_{\substack{r=1\\ r\neq j}}^{K} n_r\sigma_r^2\right] + \sigma_i^2,
\end{aligned}
$$

cuando $d \to \infty$, donde se utilizó (3.25) y el hecho de que para $j \neq i$,

$$
\sum_{\substack{r=2\\ r\neq j}}^{K}\sum_{\substack{s=1\\ s\neq j}}^{r-1} n_r n_s = \sum_{\substack{s=1\\ s\neq i,j}}^{K} n_i n_s + \sum_{\substack{r=2\\ r\neq i,j}}^{K}\sum_{\substack{s=1\\ s\neq i,j}}^{r-1} n_r n_s. \tag{3.31}
$$

Consecuentemente, usando de nuevo (3.23), se obtiene que

$$
\begin{aligned}
\frac{\left|\left|X-\overline{X}_{-j}\right|\right|^2 - \left|\left|X-\overline{X}_{j}\right|\right|^2}{d} &\xrightarrow{\mathbb{P}} N_{-j}^{-2}\left[\sum_{\substack{r=1\\ r\neq i,j}}^{K} n_r^2 c_{ir}^2 + \sum_{\substack{r=2\\ r\neq i,j}}^{K}\sum_{\substack{s=1\\ s\neq i,j}}^{r-1} n_r n_s\left(c_{ir}^2 + c_{is}^2 - c_{rs}^2\right)\right] - c_{ij}^2 \\
&+ N_{-j}^{-2}\left(\sum_{\substack{r=1\\ r\neq j}}^{K} n_r\sigma_r^2\right) - \frac{\sigma_j^2}{n_j}, \qquad \text{cuando } d \to \infty. \qquad (3.32)
\end{aligned}
$$

Por lo tanto, (3.28), (3.29) y (3.32) implican que la convergencia en probabilidad (3.19), cuando $j \neq i$, se cumple. Esto finaliza la demostración. ▲

Obtenemos así dos importantes implicaciones estadísticas del Teorema 3.4.3, que explican el comportamiento asintótico del método de clasificación multicategoría OVR-MD.

Teorema 3.4.4. *Considere las clases $\mathcal{C}_j$, $j = 1, 2, ..., K$, como antes junto con los supuestos dados en (3.1)-(3.3). Sea $i \in \{1, 2, ..., K\}$ fijo. Con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de la población $\mathcal{C}_i$ será correctamente clasificado por el método OVR-MD si y sólo si $DS_{ii} \geq DS_{ij}$ para toda $j = 1, 2, ..., K$.*

Demostración. Sean $i \in \{1, 2, ..., K\}$ fijo y X un nuevo dato que siga la distribución de la población $\mathcal{C}_i$. Debido a la Observación 3.2.1, se tiene que en este caso, X será correctamente clasificado (en la clase $\mathcal{C}_i$) por el método OVR-MD si y sólo si la distancia signada de X al hiperplano

MD_i es mayor o igual que la distancia signada de X al hiperplano MD_j para toda $j = 1, 2, ..., K$. Esto es, X será correctamente clasificado si y sólo si $DS(X, MD_i) \geq DS(X, MD_j)$ para toda $j = 1, 2, ..., K$. En este sentido, la convergencia obtenida en (3.19) del Teorema 3.4.3 permite concluir que, con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de la población $\mathcal{C}_i$ será correctamente clasificado por el método OVR-MD si y sólo si $DS_{ii} \geq DS_{ij}$ para toda $j = 1, 2, ..., K$. ▲

Corolario 3.4.1. *Asuma que $n_j = n$, $\sigma_j = \sigma$ y $c_{ij} = c$, para toda $i, j = 1, 2, ..., K$, y para algunas constantes $n \in \mathbb{N}$, $\sigma > 0$ y $c > 0$. Entonces la probabilidad de que un nuevo dato proveniente de cualquier población $\mathcal{C}_j$, $j = 1, 2, ..., K$, sea correctamente clasificado por el método OVR-MD converge a 1 cuando $d \to \infty$.*

Demostración. Suponga que $n_j = n$, $\sigma_j = \sigma > 0$ y $c_{ij} = c > 0$, para toda $i, j = 1, 2, ..., K$. Note que

$$N_{-j} = (K-1)n, \qquad \text{para toda } \ j = 1, 2, ..., K.$$

Luego, para toda $j = 1, 2, ..., K$, se tiene de (3.20) y (3.25) que

$$R_j = (K-2)(K-1)n\sigma^2$$

y

$$\sum_{\substack{r=2 \\ r\neq j}}^{K} \sum_{\substack{s=1 \\ s\neq j}}^{r-1} n_r n_s = \frac{1}{2}\left[(K-1)^2 n^2 - (K-1)n^2\right] = \frac{(K-2)(K-1)n^2}{2}. \tag{3.33}$$

De manera que, al usar (3.33) en (3.21), resulta que

$$D_j = 2(K-1)n\left[(K-1)Kn\left(\frac{nc^2}{2} + \sigma^2\right)\right]^{1/2}, \qquad \text{para toda } \ j = 1, 2, ..., K.$$

y también, usando (3.33) y (3.31) se tiene que

$$\sum_{\substack{r=2 \\ r\neq i,j}}^{K} \sum_{\substack{s=1 \\ s\neq i,j}}^{r-1} n_r n_s = \frac{(K-2)(K-1)n^2}{2} - (K-2)n^2 = \frac{(K-3)(K-2)n^2}{2}$$

para toda $j = 1, 2, ..., K$ con $j \neq i$.

Considere ahora un nuevo dato X que siga la distribución de la población $\mathcal{C}_i$, para algún $i \in \{1, 2, ..., K\}$ fijo. Observe que estos cálculos implican que si $n_j = n$, $\sigma_j = \sigma > 0$ y $c_{rj} = c >$

0, para toda $r, j = 1, 2, ..., K$, entonces

$$\frac{DS(X, MD_j)}{d^{1/2}} \xrightarrow{\mathbb{P}} DS_{ij} = \begin{cases} \dfrac{Knc^2/2 - (K-2)\sigma^2}{2\left[(K-1)Kn\left(nc^2/2 + \sigma^2\right)\right]^{1/2}} & \text{si} \quad j = i, \\ \dfrac{-Knc^2/2 - (K-2)\sigma^2}{2\left[(K-1)Kn\left(nc^2/2 + \sigma^2\right)\right]^{1/2}} & \text{si} \quad j \neq i, \end{cases} \tag{3.34}$$

para toda $j = 1, 2, ..., K$, cuando $d \to \infty$.

De modo que, por el Teorema 3.4.4 se obtiene que X es correctamente clasificado asintóticamente por el método OVR-MD, puesto que claramente

$$DS_{ii} \geq DS_{ij}, \qquad \text{para toda} \quad j = 1, 2, ..., K. \tag{3.35}$$

Dado que (3.35) se cumple para toda $i \in \{1, 2, ..., K\}$, se concluye que, con probabilidad tendiendo a 1 cuando $d \to \infty$, un nuevo dato X proveniente de cualquiera de las clases $\mathcal{C}_i$, $i = 1, 2, ..., K$, es correctamente clasificado por el método OVR-MD. ▲

Otro resultado estadístico respecto a las propiedades asintóticas del método OVR-MD, en una situación particular, es dado a continuación.

Teorema 3.4.5. *Considere las clases $\mathcal{C}_j$, $j = 1, 2, ..., K$, como antes junto con los supuestos dados en (3.1)-(3.3). Suponga que $c_{ij} = c$, para toda $i, j = 1, 2, ..., K$, y para alguna constante $c > 0$. Entonces, con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de cualquier población $\mathcal{C}_j$, $j = 1, 2, ..., K$, será correctamente clasificado por el método OVR-MD si*

$$c^2 \geq \text{máx} \left\{ \max_{1 \leq i \leq K} \left[\frac{R_i}{N_{-i}^2 - S_{-i}} \right], \max_{\substack{1 \leq i, j \leq K \\ j \neq i}} \left[\frac{-R_j}{n_i N_{-j} + S_{-j}} \right] \right\}, \tag{3.36}$$

donde

$$S_{-j} := \sum_{\substack{r=2 \\ r \neq j}}^{K} \sum_{\substack{s=1 \\ s \neq j}}^{r-1} n_r n_s, \qquad j = 1, 2, ..., K. \tag{3.37}$$

Demostración. Sea X un nuevo dato que siga la distribución de la población $\mathcal{C}_i$, para algún $i \in \{1, 2, ..., K\}$ fijo, y suponga que $c_{rs} = c$, para toda $r, s = 1, 2, ..., K$, y para alguna constante $c > 0$. Debido a que

$$N_{-j}^2 = \sum_{\substack{r=1 \\ r \neq j}}^{K} n_r^2 + 2S_{-j}, \qquad \text{para toda} \quad j = 1, 2, ..., K$$

y

$$S_{-j} = \sum_{\substack{s=1 \\ s\neq i,j}}^{K} n_i n_s + \sum_{\substack{r=2 \\ r\neq i,j}}^{K} \sum_{\substack{s=1 \\ s\neq i,j}}^{r-1} n_r n_s, \qquad \text{para toda } \; j = 1, 2, ..., K \;\; \text{con} \;\; j \neq i,$$

por el Teorema 3.4.3 se obtiene que en este caso

$$\frac{DS(X, MD_j)}{d^{1/2}} \xrightarrow{\mathbb{P}} DS_{ij} = \begin{cases} \dfrac{\left(N_{-i}^2 - S_{-i}\right) c^2 - R_i}{2N_{-i}\left[\left(N_{-i}^2 - S_{-i}\right) c^2 + N_{-i}^2 \sigma_i^2 / n_i + \sum_{\substack{r=1 \\ r\neq i}}^{K} n_r \sigma_r^2\right]^{1/2}} & j = i, \\ \dfrac{-\left(n_i N_{-j} + S_{-j}\right) c^2 - R_j}{2N_{-j}\left[\left(N_{-j}^2 - S_{-j}\right) c^2 + N_{-j}^2 \sigma_j^2 / n_j + \sum_{\substack{r=1 \\ r\neq j}}^{K} n_r \sigma_r^2\right]^{1/2}} & j \neq i, \end{cases} \tag{3.38}$$

para toda $j = 1, 2, ..., K$, cuando $d \to \infty$. La Observación 3.2.1 dice que en este caso X será correctamente clasificado (en la clase $\mathcal{C}_i$) por el método OVR-MD si $DS(X, MD_i) \geq DS(X, MD_j)$, para toda $j = 1, 2, ..., K$. Así, por (3.38), X será correctamente clasificado asintóticamente por el método OVR-MD si

$$DS_{ii} \geq DS_{ij}, \qquad \text{para toda } \; j = 1, 2, ..., K. \tag{3.39}$$

Dado que el denominador de DS_{ij}, $j = 1, 2, ..., K$ es positivo, entonces una condición suficiente para que (3.39) se cumpla es que

$$\left(N_{-i}^2 - S_{-i}\right) c^2 - R_i \geq 0$$

y

$$-\left(n_i N_{-j} + S_{-j}\right) c^2 - R_j \leq 0, \qquad \text{para toda} \quad j = 1, 2, ..., K \;\; \text{con} \;\; j \neq i;$$

o equivalentemente,

$$c^2 \geq \frac{R_i}{N_{-i}^2 - S_{-i}}$$

y

$$c^2 \geq \frac{-R_j}{n_i N_{-j} + S_{-j}}, \qquad \text{para toda} \quad j = 1, 2, ..., K \;\; \text{con} \;\; j \neq i.$$

Por consiguiente, es suficiente que

$$c^2 \geq \text{máx}\left\{ \frac{R_i}{N_{-i}^2 - S_{-i}}, \; \underset{\substack{1\leq j\leq K \\ j\neq i}}{\text{máx}} \left\{ \frac{-R_j}{n_i N_{-j} + S_{-j}} \right\} \right\},$$

para que, con probabilidad tendiendo a 1 cuando $d \to \infty$, un nuevo dato de una clase C_i, $i \in \{1, 2, ..., K\}$ fijo, sea correctamente clasificado por OVR-MD.

Por lo tanto, con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de cualquiera de las clases $\mathcal{C}_i$, $i = 1, 2, ..., K$, será correctamente clasificado por el método OVR-MD si

$$c^2 \geq \max_{1 \leq i \leq K} \left\{ \frac{R_i}{N_{-i}^2 - S_{-i}} \right\} \quad \text{y} \quad c^2 \geq \max_{1 \leq i \leq K} \left\{ \max_{\substack{1 \leq j \leq K \\ j \neq i}} \left\{ \frac{-R_j}{n_i N_{-j} + S_{-j}} \right\} \right\},$$

lo cual es equivalente a la condición (3.36). ▲

Observación 3.4.1. *Bajo los supuestos del Teorema 3.4.5 se tiene que para valores fijos de n_i y σ_i, $i = 1, 2, ..., K$, con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de cualquier clase será correctamente clasificado por el método OVR-MD si $c > 0$ pertenece a la región donde se satisfaga que $DS_{ii} \geq DS_{ij}$, para toda $i, j = 1, 2, ..., K$. La condición (3.36) corresponde a un caso particular de los valores de c en esa región.*

Como consecuencia del Teorema 3.4.5, obtenemos el siguiente corolario que expone situaciones donde la condición (3.36), para clasificación asintóticamente correcta, se simplifica.

Corolario 3.4.2. *Asuma sin pérdida de generalidad que en el Teorema 3.4.5 se cumplen las desigualdades $\frac{\sigma_1^2}{n_1}\left(N_{-1}^2 + n_1^2\right) \geq \frac{\sigma_2^2}{n_2}\left(N_{-2}^2 + n_2^2\right) \geq \cdots \geq \frac{\sigma_K^2}{n_K}\left(N_{-K}^2 + n_K^2\right)$; de no ser así, renombre a las clases de tal modo que esto se cumpla.*

a) Suponga que $R_K \geq 0$. Entonces, con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de cualquier población $\mathcal{C}_i$, $i = 1, 2, ..., K$, será correctamente clasificado por el método OVR-MD si

$$c^2 \geq \max_{1 \leq i \leq K} \left\{ \frac{R_i}{N_{-i}^2 - S_{-i}} \right\}.$$

b) Suponga que $R_K < 0$ y que además $n_K \geq \cdots \geq n_2 \geq n_1$. Entonces, con probabilidad convergiendo a 1 cuando $d \to \infty$, un nuevo dato proveniente de cualquier población $\mathcal{C}_i$, $i = 1, 2, ..., K$, será correctamente clasificado por el método OVR-MD si

$$c^2 \geq \max \left\{ \max_{1 \leq i \leq K} \left\{ \frac{R_i}{N_{-i}^2 - S_{-i}} \right\}, \frac{-R_K}{n_1 N_{-K} + S_{-K}} \right\}.$$

Demostración. El hecho de que $\frac{\sigma_i^2}{n_i}\left(N_{-i}^2+n_i^2\right) \geq \frac{\sigma_j^2}{n_j}\left(N_{-j}^2+n_j^2\right)$ para toda $i<j$ implica que $R_i \geq R_j$ para toda $i<j$ y así, en particular,

$$R_j \geq R_K, \qquad \text{para toda } \; j=1,2,...,K. \tag{3.40}$$

Se consideran así los dos casos siguientes:

a) Asuma que $R_K \geq 0$. Note que aquí (3.40) implica que

$$0 \geq \max_{\substack{1\leq j\leq K\\ j\neq i}}\left\{\frac{-R_j}{n_iN_{-j}+S_{-j}}\right\}, \qquad \text{para cualquier } \; i\in\{1,2,...,K\};$$

de lo cual resulta que

$$c^2>0 \geq \max_{1\leq i\leq K}\left\{\max_{\substack{1\leq j\leq K\\ j\neq i}}\left\{\frac{-R_j}{n_iN_{-j}+S_{-j}}\right\}\right\}.$$

Por ende, de la condición (3.36) del Teorema 3.4.5 se obtiene que en este caso, con probabilidad tendiendo a 1 cuando $d\to\infty$, un nuevo dato de cualquier clase será correctamente clasificado por OVR-MD si

$$c^2 \geq \max_{1\leq i\leq K}\left\{\frac{R_i}{N_{-i}^2-S_{-i}}\right\}.$$

b) Asuma que $R_K<0$ y que $n_K\geq\cdots\geq n_2\geq n_1$. Note que

$$N_{-j}=n_K+\sum_{\substack{r=1\\ r\neq j,K}}^{K} n_r \geq n_j+\sum_{\substack{r=1\\ r\neq j,K}}^{K} n_r = N_{-K}$$

y

$$S_{-j}=n_K\sum_{\substack{s=1\\ s\neq j,K}}^{K} n_s+\sum_{\substack{r=2\\ r\neq j,K}}^{K}\sum_{\substack{s=1\\ s\neq j,K}}^{r-1} n_rn_s \geq n_j\sum_{\substack{s=1\\ s\neq j,K}}^{K} n_s+\sum_{\substack{r=2\\ r\neq j,K}}^{K}\sum_{\substack{s=1\\ s\neq j,K}}^{r-1} n_rn_s = S_{-K},$$

para toda $j\neq K$; es decir, $N_{-j}\geq N_{-K}$ y $S_{-j}\geq S_{-K}$ para toda $j\neq K$. Por lo que

$$n_iN_{-j}+S_{-j}\geq n_iN_{-K}+S_{-K}, \quad \text{para toda } \; i,j=1,2,...,K. \tag{3.41}$$

De manera que, (3.40), (3.41) y $-R_K>0$, conducen a que

$$\frac{-R_K}{n_iN_{-K}+S_{-K}} \geq \frac{-R_j}{n_iN_{-j}+S_{-j}}, \quad \text{para toda } \; i,j=1,2,...,K. \tag{3.42}$$

Observe así que de (3.42) se deduce que

$$\frac{-R_K}{n_i N_{-K} + S_{-K}} = \max_{\substack{1\leq j\leq K \\ j\neq i}} \left\{ \frac{-R_j}{n_i N_{-j} + S_{-j}} \right\}, \qquad \text{para cualquier } \ i \in \{1, 2, ..., K-1\},$$

y

$$\frac{-R_K}{n_K N_{-K} + S_{-K}} \geq \max_{\substack{1\leq j\leq K \\ j\neq K}} \left\{ \frac{-R_j}{n_K N_{-j} + S_{-j}} \right\}.$$

Luego, haciendo uso de que $n_i \geq n_1$, para toda $i = 1, 2, ..., K$, se obtiene que

$$\frac{-R_K}{n_1 N_{-K} + S_{-K}} = \max_{1\leq i\leq K-1} \left\{ \max_{\substack{1\leq j\leq K \\ j\neq i}} \left\{ \frac{-R_j}{n_i N_{-j} + S_{-j}} \right\} \right\}$$

y

$$\frac{-R_K}{n_1 N_{-K} + S_{-K}} \geq \max_{\substack{1\leq j\leq K \\ j\neq K}} \left\{ \frac{-R_j}{n_K N_{-j} + S_{-j}} \right\};$$

y así, resulta que

$$\frac{-R_K}{n_1 N_{-K} + S_{-K}} = \max_{1\leq i\leq K} \left\{ \max_{\substack{1\leq j\leq K \\ j\neq i}} \left\{ \frac{-R_j}{n_i N_{-j} + S_{-j}} \right\} \right\}.$$

Por lo tanto, nuevamente por (3.36) del Teorema 3.4.5 se concluye que en este caso, con probabilidad tendiendo a 1 cuando $d \to \infty$, un nuevo dato de cualquier clase será correctamente clasificado por OVR-MD si

$$c^2 \geq \max \left\{ \max_{1\leq i\leq K} \left\{ \frac{R_i}{N_{-i}^2 - S_{-i}} \right\}, \ \frac{-R_K}{n_1 N_{-K} + S_{-K}} \right\}.$$

Esto finaliza la demostración. ▲

Capítulo 4

Estudio de Simulación

En este último capítulo presentamos y analizamos los resultados de un estudio de simulación que hemos realizado con el propósito de complementar nuestros resultados teóricos expuestos en la Sección 3.4, y comparamos, bajo diferentes escenarios, el comportamiento asintótico de los métodos de clasificación multicategoría considerados cuando los tamaños muestrales son fijos y las dimensiones crecen. Llevamos acabo estas simulaciones mediante el software Matlab.

Consideramos las clases $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ de n_j datos gaussianos d-multivariados con matrices de covarianza de la forma $\Sigma_j = \sigma_j^2 \mathbf{I}_d$, $\sigma_j > 0$, y medias poblacionales μ_j, $j = 1, 2, 3$, donde estas medias tienen una de las siguientes dos formas:

i) $\mu_1 = (0, 0, ..., 0)^\top$, $\mu_2 = d^{1/2} c_{12} (1, 0, ..., 0)^\top$, $\mu_3 = d^{1/2} c_{13} (\cos(\pi/3), \sin(\pi/3), 0, ..., 0)^\top$, con $c_{12}, c_{13} > 0$.

ii) $\mu_1 = (0, 0, ..., 0)^\top$, $\mu_2 = c_{12} (1, 1, ..., 1)^\top$, $\mu_3 = c_{13} (1, 1, ..., -1)^\top$, con $c_{12}, c_{13} > 0$.

Estas clases satisfacen las condiciones (3.1)-(3.3), con $c_{23}^2 = (c_{12} - c_{13})^2 + c_{12} c_{13}$ para las medias en i) y con $c_{23} = |c_{12} - c_{13}|$ para las medias en ii), debido al hecho de que los datos gaussianos estándar multivariados cumplen estas condiciones que implican la representación geométrica asintótica. En efecto, note que tanto para i) como para ii) se tiene que

$$\frac{||X_r^1 - \mu_1||^2}{d} = \frac{\sigma_1^2 \left|\left|\frac{X_r^1 - \mu_1}{\sigma_1}\right|\right|^2}{d} \xrightarrow{\mathbb{P}} \sigma_1^2 \quad \text{y} \quad \frac{||X_r^1 - X_s^1||^2}{d} = \frac{\sigma_1^2 \left|\left|\frac{(X_r^1 - \mu_1) - (X_s^1 - \mu_1)}{\sigma_1}\right|\right|^2}{d} \xrightarrow{\mathbb{P}} 2\sigma_1^2$$

cuando $d \hookrightarrow \infty$, para toda $r, s = 1, 2, ..., n_1$ con $r \neq s$, y similarmente se obtiene esto para las clases $\mathcal{C}_2$ y $\mathcal{C}_3$. Observe además que para las medias en i) se cumple que

$$\frac{||\mu_1 - \mu_2||^2}{d} = \frac{\left|\left|-d^{1/2}c_{12}\,(1, 0, ..., 0)\right|\right|^2}{d} = \frac{dc_{12}^2\,||(1, 0, ..., 0)||^2}{d} = c_{12}^2\frac{d}{d} = c_{12}^2,$$

$$\begin{aligned}\frac{||\mu_1 - \mu_3||^2}{d} &= \frac{\left|\left|-d^{1/2}c_{13}\,(cos(\pi/3), sen(\pi/3), 0, ..., 0)\right|\right|^2}{d} \\ &= \frac{dc_{13}^2\,||(cos(\pi/3), sen(\pi/3), 0, ..., 0)||^2}{d} \\ &= c_{13}^2\left[cos^2(\pi/3) + sen^2(\pi/3)\right] = c_{13}^2,\end{aligned}$$

$$\begin{aligned}\frac{||\mu_2 - \mu_3||^2}{d} &= \frac{\left|\left|d^{1/2}c_{12}\,(1, 0, ..., 0) - d^{1/2}c_{13}\,(cos(\pi/3), sen(\pi/3), 0, ..., 0)\right|\right|^2}{d} \\ &= \frac{d\,||(c_{12} - c_{13}cos(\pi/3), -c_{13}sen(\pi/3), 0, ..., 0)||^2}{d} \\ &= [c_{12} - c_{13}cos(\pi/3)]^2 + [-c_{13}sen(\pi/3)]^2 \\ &= c_{12}^2 - 2c_{12}c_{13}cos(\pi/3) + c_{13}^2cos^2(\pi/3) + c_{13}^2sen^2(\pi/3) \\ &= c_{12}^2 + c_{13}^2\left[cos^2(\pi/3) + sen^2(\pi/3)\right] - \frac{2c_{12}c_{13}}{2} \\ &= c_{12}^2 + c_{13}^2 - c_{12}c_{13} \equiv c_{23}^2.\end{aligned}$$

También, para las medias en ii) se satisface que

$$\frac{||\mu_1 - \mu_2||^2}{d} = \frac{||-c_{12}\,(1, 1, ..., 1)||^2}{d} = \frac{c_{12}^2\,||(1, 1, ..., 1)||^2}{d} = c_{12}^2\frac{d}{d} = c_{12}^2,$$

$$\frac{||\mu_1 - \mu_3||^2}{d} = \frac{||-c_{13}\,(1, 1, ..., -1)||^2}{d} = \frac{c_{13}^2\,||(1, 1, ..., -1)||^2}{d} = c_{13}^2\frac{d}{d} = c_{13}^2,$$

$$\begin{aligned}\frac{||\mu_2 - \mu_3||^2}{d} &= \frac{||(c_{12} - c_{13}, ..., c_{12} - c_{13}, c_{12} + c_{13})||^2}{d} = \frac{(c_{12} - c_{13})^2\,(d-1) + (c_{12} + c_{13})^2}{d} \\ &= (c_{12} - c_{13})^2 + \frac{(c_{12} + c_{13})^2 - (c_{12} - c_{13})^2}{d} \\ &\longrightarrow (c_{12} - c_{13})^2 \equiv c_{23}^2, \quad \text{cuando } d \to \infty.\end{aligned}$$

Ahora, para ver que tanto las medias en i) como en ii) satisfacen (3.3) notemos que, para

cualesquiera $i \neq j$ y $r \neq s$, se tiene que

$$\begin{aligned}
\left|\left|X_r^i - X_s^j\right|\right|^2 &= \langle X_r^i - X_s^j, X_r^i - X_s^j\rangle \\
&= \langle X_r^i - \mu_i + \mu_i - X_s^j + \mu_j - \mu_j, X_r^i - \mu_i + \mu_i - X_s^j + \mu_j - \mu_j\rangle \\
&= \left|\left|X_r^i - \mu_i\right|\right|^2 + \langle X_r^i - \mu_i, \mu_i\rangle - \langle X_r^i - \mu_i, X_s^j - \mu_j\rangle - \langle X_r^i - \mu_i, \mu_j\rangle \\
&+ \langle \mu_i, X_r^i - \mu_i\rangle + ||\mu_i||^2 - \langle \mu_i, X_s^j - \mu_j\rangle - \langle \mu_i, \mu_j\rangle - \langle X_s^j - \mu_j, X_r^i - \mu_i\rangle \\
&- \langle X_s^j - \mu_j, \mu_i\rangle + \left|\left|X_s^j - \mu_j\right|\right|^2 + \langle X_s^j - \mu_j, \mu_j\rangle - \langle \mu_j, X_r^i - \mu_i\rangle - \langle \mu_j, \mu_i\rangle \\
&+ \langle \mu_j, X_s^j - \mu_j\rangle + ||\mu_j||^2 \\
&= \left|\left|X_r^i - \mu_i\right|\right|^2 + \left|\left|X_s^j - \mu_j\right|\right|^2 + ||\mu_i - \mu_j||^2 - 2\langle X_r^i - \mu_i, X_s^j - \mu_j\rangle \\
&+ 2\langle X_r^i - \mu_i, \mu_i\rangle - 2\langle X_r^i - \mu_i, \mu_j\rangle - 2\langle \mu_i, X_s^j - \mu_j\rangle + 2\langle X_s^j - \mu_j, \mu_j\rangle. \quad (4.1)
\end{aligned}$$

De modo que,

$$\frac{||X_r^i - X_s^j||^2}{d} \xrightarrow{\mathbb{P}} \sigma_i^2 + \sigma_j^2 + c_{ij}^2, \quad \text{cuando } d \to \infty,$$

puesto que los últimos cinco sumandos de (4.1), divididos por d, convergen en probabilidad a cero cuando d tiende a infinito. En efecto, como $X_t^k \sim \mathcal{N}_d\left(\mu_k, \sigma_k^2 \mathbf{I}_d\right)$, tenemos entonces por la ley de los grandes números que

$$\frac{\langle X_r^i - \mu_i, X_s^j - \mu_j\rangle}{d} = \frac{\sigma_i\sigma_j}{d}\langle \frac{X_r^i - \mu_i}{\sigma_i}, \frac{X_s^j - \mu_j}{\sigma_j}\rangle = \sigma_i\sigma_j \sum_{q=1}^{d} \frac{Z_q^i Z_q^j}{d} \xrightarrow{\mathbb{P}} 0, \quad \text{cuando } d \to \infty,$$

donde $Z^i = \dfrac{X_r^i - \mu_i}{\sigma_i} \sim \mathcal{N}_d\left(0, \mathbf{I}_d\right)$ y $Z^j = \dfrac{X_s^j - \mu_j}{\sigma_j} \sim \mathcal{N}_d\left(0, \mathbf{I}_d\right)$ son independientes. Además, notemos que para ver que los últimos cuatro sumandos de (4.1) cumplen también esta convergencia, es suficiente con observar que para las medias en i) se tiene que $\dfrac{\langle X_r^i - \mu_i, \mu_1\rangle}{d} = 0$,

$$\frac{\langle X_r^i - \mu_i, \mu_2\rangle}{d} = \frac{d^{1/2} c_{12}\sigma_i \left(\frac{X_{r1}^i - \mu_{i1}}{\sigma_i}\right)}{d} = c_{12}\sigma_i \frac{Z_1^i}{d^{1/2}} \xrightarrow{\mathbb{P}} 0, \text{ cuando } d \to \infty \text{ con } Z_1^i \sim \mathcal{N}\left(0,1\right),$$

y

$$\begin{aligned}
\frac{\langle X_r^i - \mu_i, \mu_3\rangle}{d} &= \frac{c_{13}}{d^{1/2}}\sigma_i \left(\frac{X_{r1}^i - \mu_{i1}}{\sigma_i} cos(\pi/3) + \frac{X_{r2}^i - \mu_{i2}}{\sigma_i} sen(\pi/3)\right) \\
&= c_{13}\sigma_i \left(\frac{Z_1^i}{d^{1/2}} cos(\pi/3) + \frac{Z_2^i}{d^{1/2}} sen(\pi/3)\right) \xrightarrow{\mathbb{P}} 0, \\
&\quad \text{cuando } d \to \infty \text{ con } Z_q^i \sim \mathcal{N}\left(0,1\right).
\end{aligned}$$

Similarmente, para las medias en ii) se tiene que $\dfrac{\langle X_r^i - \mu_i, \mu_1\rangle}{d} = 0,$

$$\frac{\langle X_r^i - \mu_i, \mu_2\rangle}{d} = c_{12}\sigma_i \sum_{q=1}^{d} \frac{Z_q^i}{d} \xrightarrow{\mathbb{P}} 0, \text{ cuando } d \to \infty \text{ con } Z^i = \frac{X_r^i - \mu_i}{\sigma_i} \sim \mathcal{N}_d\left(0, \mathbf{I}_d\right), \text{ y}$$

$$\begin{aligned}\frac{\langle X_r^i - \mu_i, \mu_3\rangle}{d} &= c_{13}\sigma_i \frac{d-1}{d} \sum_{q=1}^{d-1} \frac{Z_q^i}{d-1} - \frac{Z_d^i}{d} \xrightarrow{\mathbb{P}} 0, \\ &\quad \text{cuando } d \to \infty \text{ con } Z^i = \frac{X_r^i - \mu_i}{\sigma_i} \sim \mathcal{N}_d\left(0, \mathbf{I}_d\right).\end{aligned}$$

Para las simulaciones consideramos las dimensiones $d = 50, 100, 500, 1000, 1500$. Para cada d, generamos 500 conjuntos de datos de entrenamiento de tamaños n_j por cada $\mathcal{C}_j$, $j = 1, 2, 3$. Calculamos las tasas de error de clasificación promedio por cada clase y globalmente (las tres clases en conjunto) de los métodos multicategoría, tomando 100 datos de prueba por cada clase.

4.1. Simulaciones para la Metodología OVO

En esta sección analizamos mediante simulaciones el comportamiento de los métodos OVO-MD, OVO-SVM, OVO-DWD y OVO-MDP, considerando los casos siguientes:

- **Caso 1**: medias de la forma i), con $n_j \equiv n = 10$, $\sigma_j \equiv \sigma = 1$, $c_{ij} \equiv c = 0.3$, para $i, j = 1, 2, 3$, e $i < j$. Es decir, aquí los valores de los parámetros respectivos de las clases son iguales pero con diferentes formas de las medias. Esto implica que los tres 10-simplex correspondientes a las clases tienen aristas de igual longitud pero con diferentes centroides. En este caso se satisfacen las condiciones de a) de los teoremas 3.4.1 y 3.4.2 .

- **Caso 2**: medias de la forma ii), con $(n_1, \sigma_1) = (10, 1.4)$, $(n_2, \sigma_2) = (8, 1)$, $(n_3, \sigma_3) = (7, 0.7)$, $c_{12} = 0.72$, $c_{13} = 0.42$, $c_{23} = 0.3$. En este caso también se cumplen las condiciones de a) de los teoremas 3.4.1 y 3.4.2, tomando diferentes valores de los parámetros para las clases y diferentes formas de las medias, lo cual implica que los simpleces correspondientes a las clases son diferentes entre sí y se encuentran en diferentes ubicaciones.

- **Caso 3**: medias de la forma ii), con $(n_1, \sigma_1) = (7, 1.3)$, $(n_2, \sigma_2) = (9, 1)$, $(n_3, \sigma_3) = (10, 0.8)$, $c_{12} = 0.3$, $c_{13} = 0.2$, $c_{23} = 0.1$. En este caso se satisfacen las condiciones de b) de los teoremas 3.4.1 y 3.4.2.

Observe los resultados del Caso 1 en la Figura 1. Para cada clase, y globalmente, las tasas de error de clasificación promedio de los cuatro métodos se aproximan a cero cuando d crece, lo cual es congruente con a) de los teoremas 3.4.1 y 3.4.2. Note también que OVO-MD tiene las tasas de error más pequeñas para todas las clases, seguido por OVO-DWD, OVO-SVM y OVO-MDP, no obstante, cuando d es suficientemente grande estos cuatro métodos se comportan casi iguales.

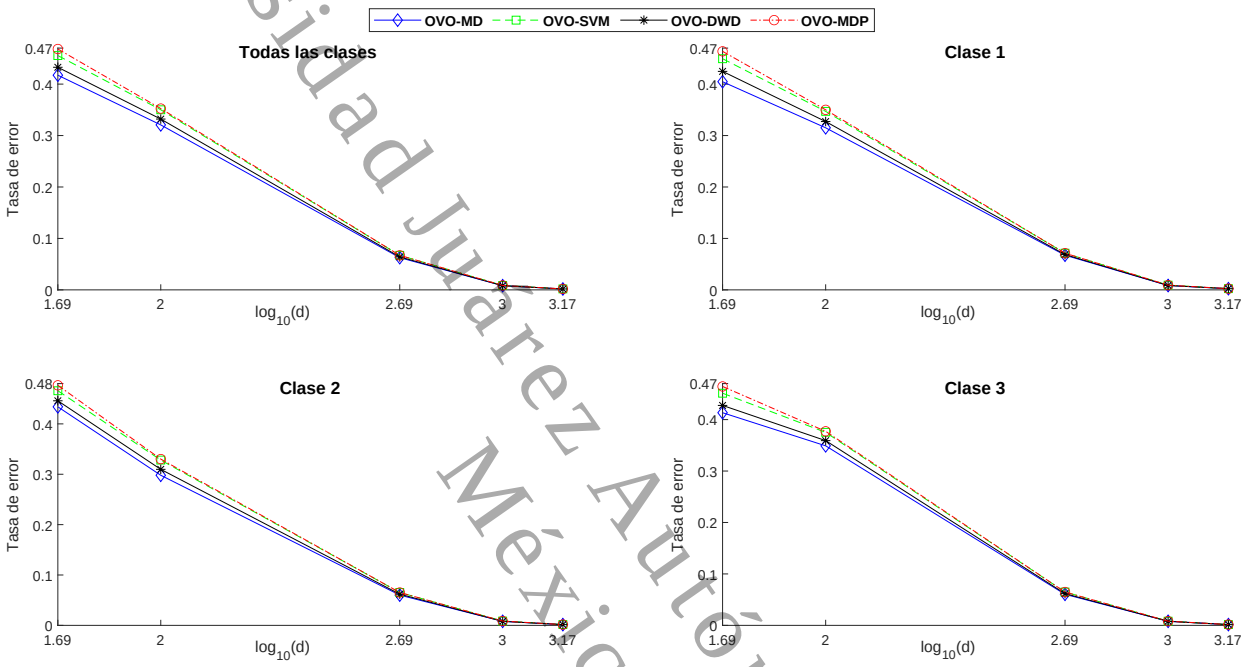


Figura 1: *Tasas de error de clasificación promedio para OVO en el Caso 1.*

En la Figura 2 se ilustran los resultados del Caso 2. Para cada clase, y globalmente, las tasas de error de clasificación promedio de los cuatro métodos tienden a cero conforme d aumenta, lo cual es consistente con los resultados de a) de los teoremas 3.4.1 y 3.4.2. Es importante señalar que aquí se consideraron adicionalmente las dimensiones $d = 2000, 2500, 3000$ para apreciar aún mejor la forma en que estas tasas se aproximan a cero. Note además que la mayor rapidez de esta convergencia ocurre en la clase 3, la segunda mayor rapidez sucede en la clase 2 y la menor ocurre en la clase 1. El método con tasas de error más pequeñas en las clases 1 y 2 es OVO-DWD, y OVO-MD es el que posee esta característica en la clase 3. Para cada clase el comportamiento de OVO-SVM y OVO-MDP es muy similar para casi todos los valores de d considerados.

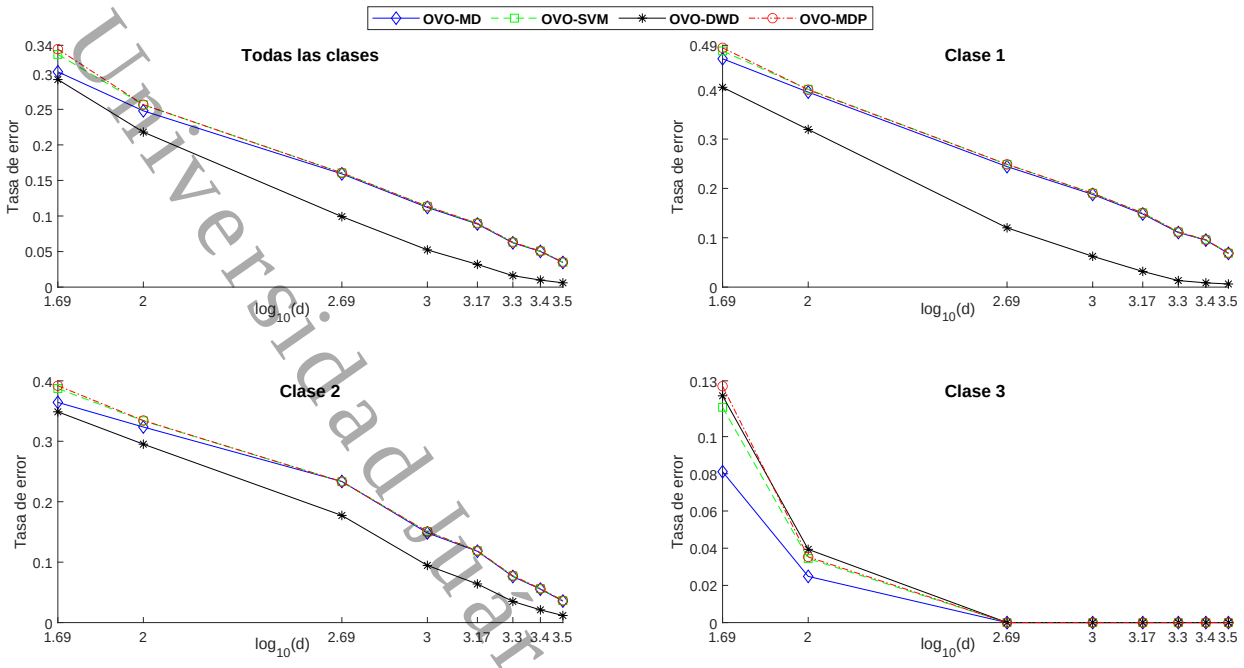


Figura 2: *Tasas de error de clasificación promedio para OVO en el Caso 2.*

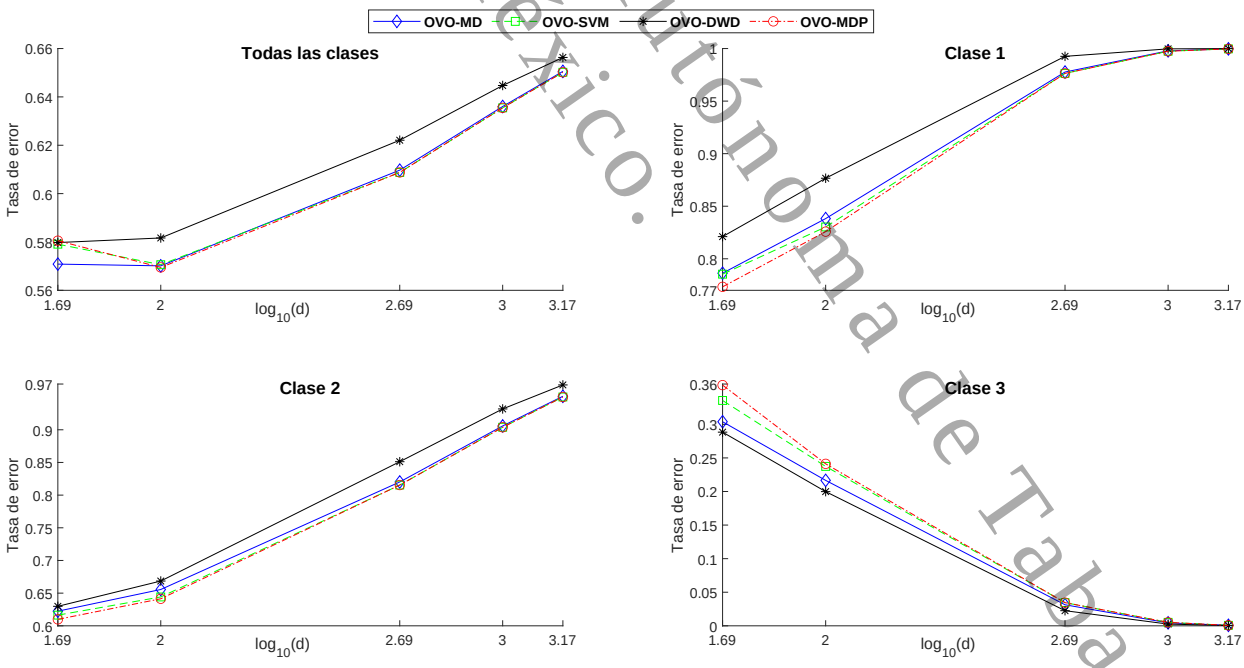


Figura 3: *Tasas de error de clasificación promedio para OVO en el Caso 3.*

La Figura 3 corresponde a los resultados del Caso 3. Observe que las tasas de error de clasificación promedio convergen a cero cuando d crece solo para la clase 3, mientras que para el resto de las clases tienden a uno, lo cual es congruente con los resultados de b) de los teoremas 3.4.1 y 3.4.2. Por ello, globalmente las tasas de error promedio se aproximan a $2/3$ cuando d crece. Note además que, para la clase 3, OVO-DWD posee las tasas de error más pequeñas, seguido de OVO-MD, OVO-SVM y OVO-MDP. El comportamiento de OVO-SVM y OVO-MDP es muy similar en cada clase y en casi todos los valores de d considerados.

Adicionalmente, analizamos el caso de no esfericidad, tomando tres clases de datos gaussianos d-multivariados con matrices de covarianza spiked $\Sigma_i = \sigma_i^2 S$, con $S = \text{diag}(d^\alpha, 1, 1, \ldots, 1)$, $0 < \alpha < 1$ y $\sigma_i > 0$, para $i = 1, 2, 3$. Consideramos medias de la forma ii) con las constantes n_i's, σ_i's and c_{ij}'s como en el Caso 2. Por el Ejemplo 3.2 de Jung y Marron (2009), cada clase satisface nuestra condición (**3.1**). Este nuevo caso satisface también nuestras condiciones (**3.2**) y (**3.3**), y cumple además a) de los teoremas 3.4.1 y 3.4.2. Los resultados de las simulaciones tomando $\alpha = 1/3$ son mostrados en la Figura 4. Se observa un comportamiento similar de las tasas de error de clasificación como en el Caso 2 (cuando $\Sigma_i = \sigma_i^2 \mathbf{I}_d$).

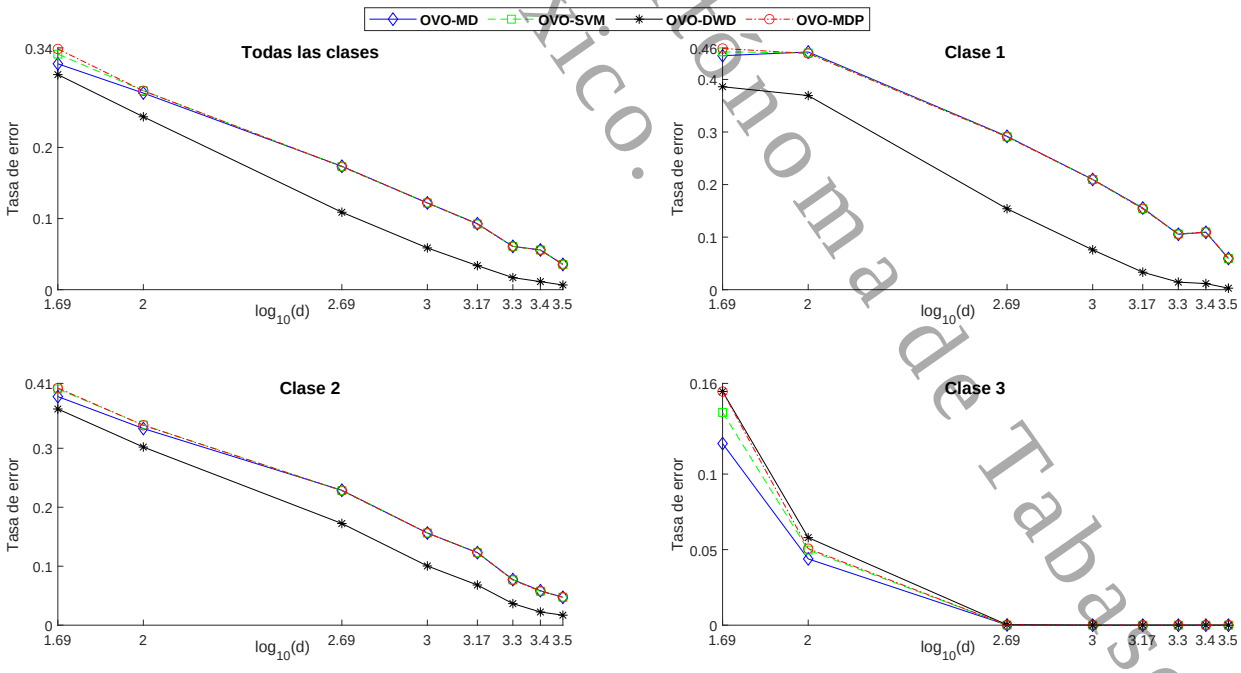


Figura 4: *Tasas de error de clasificación promedio para OVO en el caso de matrices de covarianza spiked.*

4.2. Simulaciones para la Metodología OVR

Analizamos ahora mediante simulaciones el comportamiento del método de discriminación multicategoría OVR-MD, para el cual proporcionamos resultados teóricos, y adicionalmente consideramos los métodos de clasificación multicategoría OVR-SVM, OVR-DWD y OVR-MDP. Esto se realizó en los siguientes casos:

- **Caso 1**: medias de la forma i), con $n_j \equiv n = 10$, $\sigma_j \equiv \sigma = 1$ y $c_{ij} \equiv c = 0.3$, para $i, j = 1, 2, 3$, e $i < j$.
- **Caso 2**: medias de la forma i), con $n_j \equiv n = 10$, $\sigma_j \equiv \sigma = 1$ y $c_{ij} \equiv c = 0.6$, para $i, j = 1, 2, 3$, e $i < j$.
- **Caso 3**: medias de la forma i), con $n_j \equiv n = 10$, $\sigma_j \equiv \sigma = 1.5$ y $c_{ij} \equiv c = 0.3$, para $i, j = 1, 2, 3$, e $i < j$.
- **Caso 4**: medias de la forma i), con $n_j \equiv n = 10$, $\sigma_j \equiv \sigma = 1.5$ y $c_{ij} \equiv c = 0.6$, para $i, j = 1, 2, 3$, e $i < j$. En este caso y en los tres casos anteriores se satisfacen las condiciones del Corolario 3.4.1.
- **Caso 5**: medias de la forma i), con $(n_1, \sigma_1) = (10, 1.2)$, $(n_2, \sigma_2) = (8, 1)$, $(n_3, \sigma_3) = (7, 0.9)$, $c_{12} = 0.3,\ c_{13} = 0.1,\ c_{23} = (0.07)^{1/2}$. En este caso se cumplen las condiciones $DS_{ii} \geq DS_{ij}$, para toda $j = 1, 2, 3$, del Teorema 3.4.4, solamente para $i = 2, 3$.
- **Caso 6**: medias de la forma i), con $(n_1, \sigma_1) = (10, 1.2)$, $(n_2, \sigma_2) = (8, 1)$, $(n_3, \sigma_3) = (7, 0.9)$, $c_{12} = c_{13} = c_{23} \equiv c = 0.33$. Aquí se satisfacen las condiciones de a) del Corolario 3.4.2.
- **Caso 7**: medias de la forma i), con $(n_1, \sigma_1) = (6, 1.3)$, $(n_2, \sigma_2) = (7, 1)$, $(n_3, \sigma_3) = (9, 0.8)$, $c_{12} = c_{13} = c_{23} \equiv c = 0.56$. Aquí se cumplen las condiciones de b) del Corolario 3.4.2.

En las figuras 5, 6, 7 y 8 se muestran los resultados de los casos 1, 2, 3 y 4, respectivamente. Observe que para cada clase, y globalmente, las tasas de error de clasificación promedio del método de clasificación multicategoría OVR-MD tienden a cero conforme d aumenta, lo cual es congruente con los resultados del Corolario 3.4.1. Más aún, observe que los métodos de discriminación multicategoría OVR-SVM, OVR-MDP y OVR-DWD tienen este mismo comportamiento

asintótico. No obstante, note que OVR-MD tiene el mejor comportamiento en todos los casos y en todos los valores de d donde las tasas de error no son cero. El segundo mejor es OVR-DWD, el tercero es OVR-SVM y el cuarto mejor desempeño lo tiene OVR-MDP. Es interesante notar que los resultados de estos casos son muy similares a aquellos del Caso 1 de la metodología OVO (donde los mismos valores de los parámetros respectivos y mismas formas de las medias son considerados) principalmente para d suficientemente grande. Note también que en cualquiera de las situaciones $\sigma_j = 1$ o $\sigma_j = 1.5$, las tasas de error se aproximan más rápidamente a cero cuando $c_{ij} = 0.6$ que cuando $c_{ij} = 0.3$, lo cual tiene sentido pues en esta situación (**3.2**) dice que las clases se separan más al aumentar el valor de c_{ij}, disminuyendo así las posibilidades de clasificación incorrecta. Por otro lado, si $c_{ij} = 0.3$ o $c_{ij} = 0.6$, la tendencia a cero de las tasas de error es más rápida cuando $\sigma_j = 1$ que cuando $\sigma_j = 1.5$, ya que en esta situación (**3.1**) dice que las esferas subyacentes que circunscriben a los 10-simpleces de cada clase se separan más al disminuir el valor de σ_j, implicando así menos posibilidades de clasificar incorrectamente. Es por esto que en el Caso 3 (Figura 7), donde $c_{ij} = 0.3$ y $\sigma_j = 1.5$, se consideraron adicionalmente las dimensiones $d = 2000, 2500, 3000$ para poder observar mejor la forma en que las tasas de error se aproximan a cero.

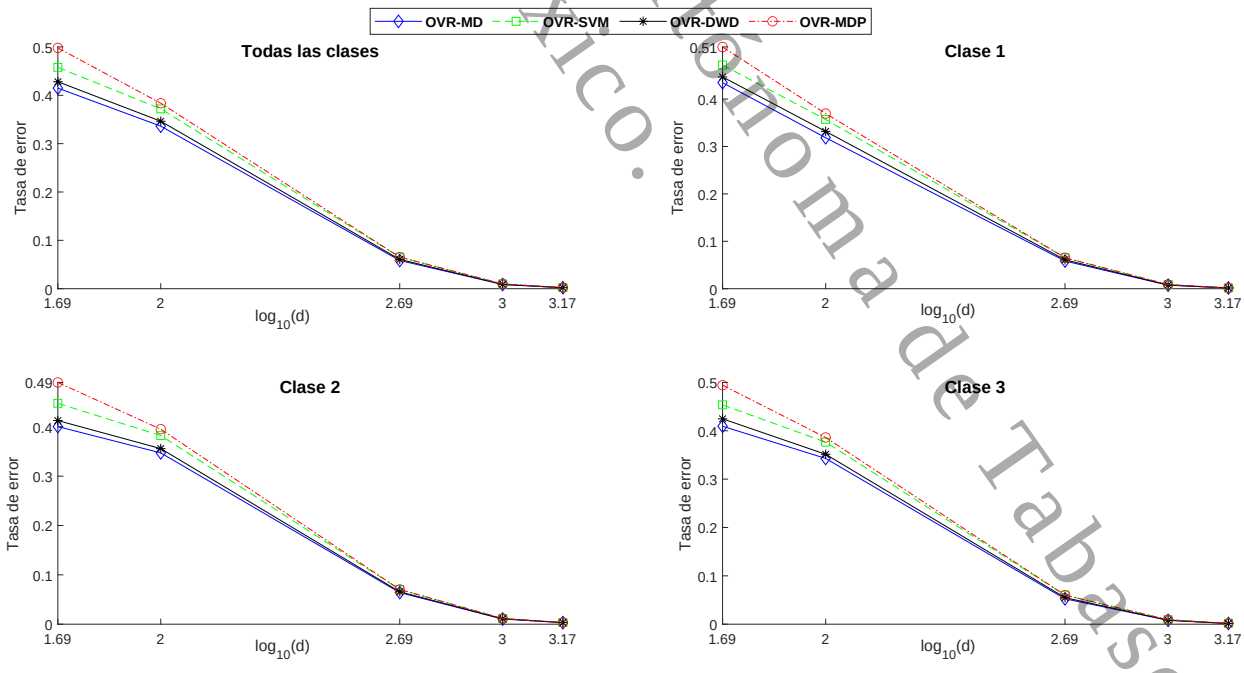


Figura 5: *Tasas de error de clasificación promedio para OVR en el Caso* 1 .

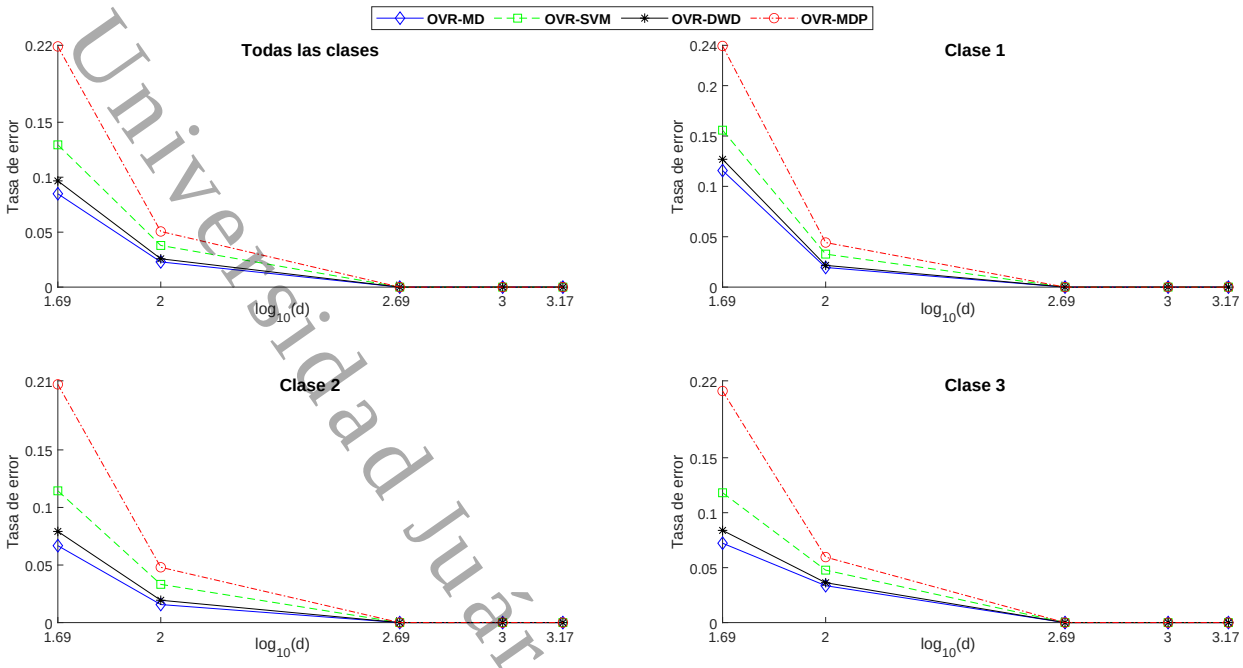


Figura 6: *Tasas de error de clasificación promedio para OVR en el Caso* 2.

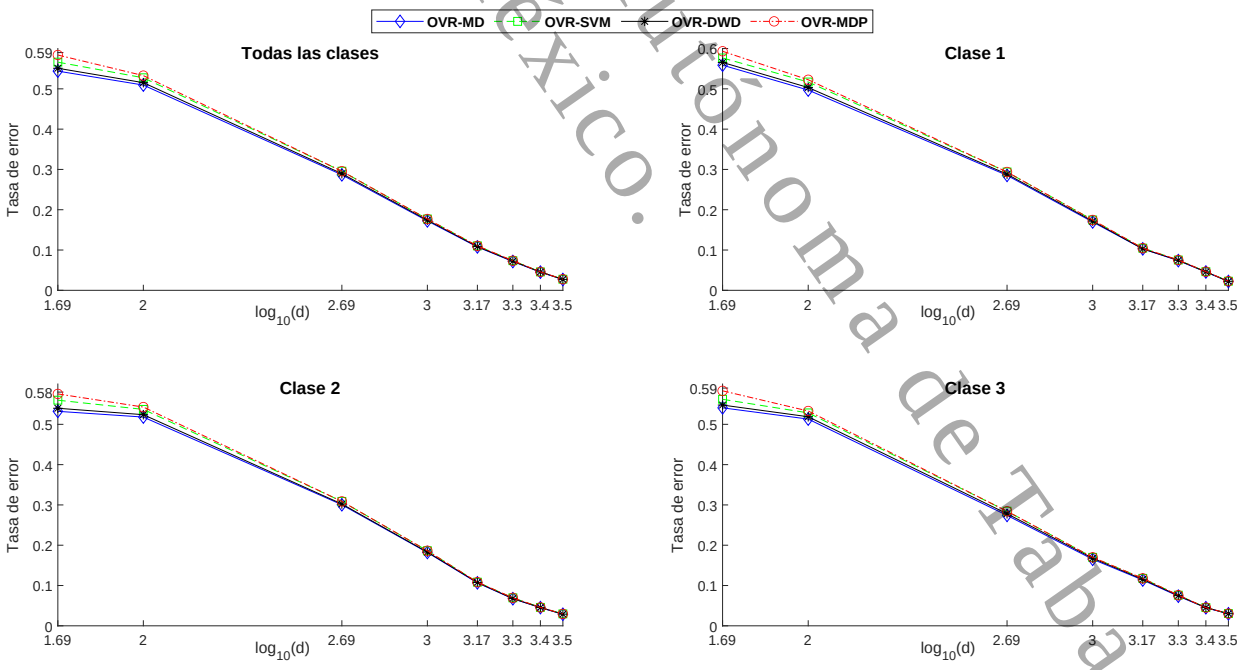


Figura 7: *Tasas de error de clasificación promedio para OVR en el Caso* 3.

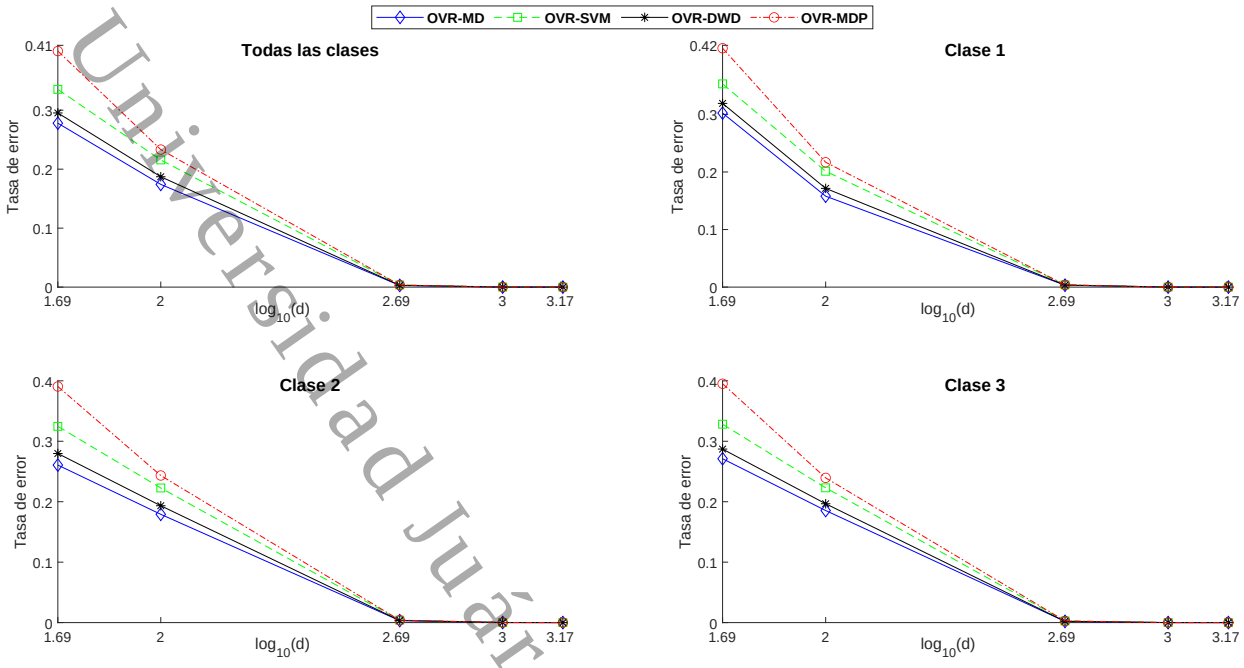


Figura 8: *Tasas de error de clasificación promedio para OVR en el Caso* 4.

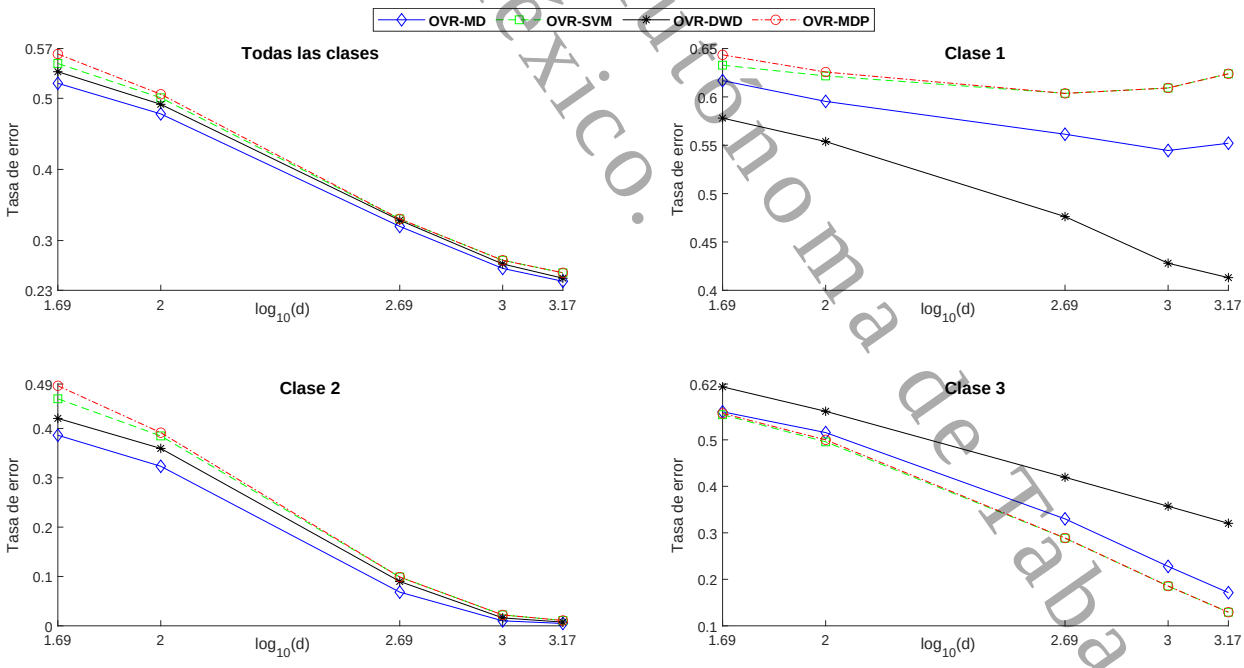


Figura 9: *Tasas de error de clasificación promedio para OVR en el Caso* 5.

En la Figura 9 se ven los resultados del Caso 5. Observe que para las clases 2 y 3 las tasas de error de OVR-MD decrecen a cero conforme d aumenta, mientras que para la clase 1 no ocurre este mismo comportamiento asintótico, lo cual es congruente con el Teorema 3.4.4. Note también que OVR-SVM y OVR-MDP, además de tener resultados muy similares en todas las clases y para casi toda d, tienen la misma tendencia que OVR-MD. Sin embargo, OVR-DWD difiere de esta tendencia en la clase 1, en la cual es el único que se aproxima a cero cuando d crece. Observe también que OVR-MD es el método con tasas de error más pequeñas en la clase 2, no obstante en la clase 3 los métodos con esta característica son OVR-SVM y OVR-MDP.

Los resultados del Caso 6 se presentan en la Figura 10. Note que en cada clase, y globalmente, las tasas de error de clasificación promedio del método OVR-MD se aproximan a cero conforme d aumenta, lo cual está de acuerdo con a) del Corolario 3.4.2. Esta propiedad también se observa para los métodos OVR-SVM, OVR-DWD y OVR-MDP. Note que la convergencia a cero de las tasas de error es ligeramente más rápida para la clase 3 que para las otras clases. OVR-MD es el método con tasas de error más pequeñas en las clases 2 y 3, no obstante en la clase 1 el método con esta propiedad es OVR-DWD. Para casi todos los valores de d considerados los comportamientos de OVR-SVM y OVR-MDP son casi idénticos.

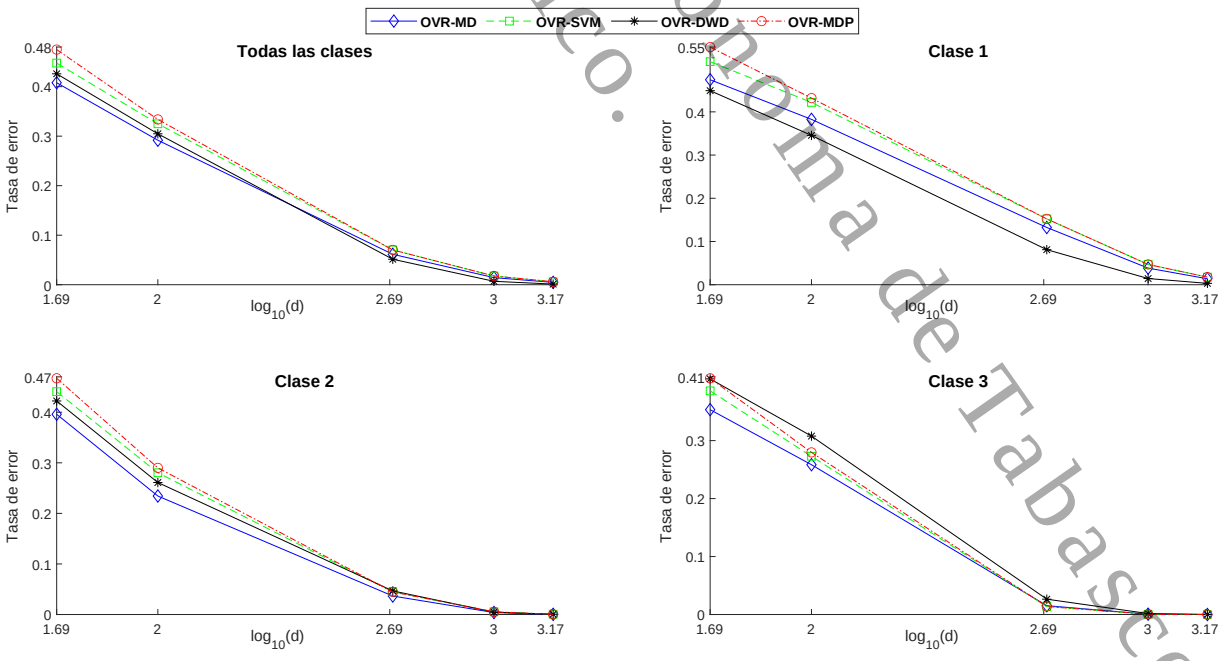


Figura 10: *Tasas de error de clasificación promedio para OVR en el Caso* 6.

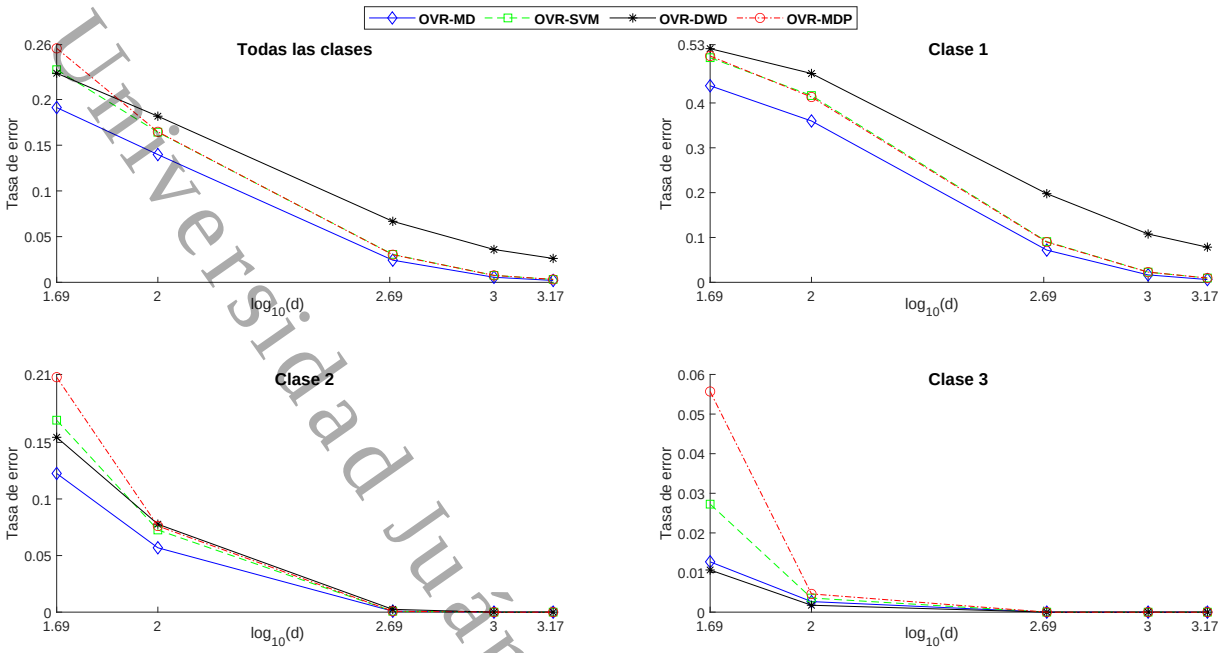


Figura 11: *Tasas de error de clasificación promedio para OVR en el Caso* 7.

En la Figura 11 se ven los resultados del Caso 7. De nuevo que para cada clase, y globalmente, las tasas de error de OVR-MD tienden a cero conforme d aumenta, lo cual es congruente con b) del Corolario 3.4.2. Se ve además que las tasas de OVR-SVM, OVR-DWD y OVR-MDP tienen este mismo comportamiento asintótico, para cada clase, y globalmente. Note que la mayor rapidez de convergencia a cero de las tasas de error sucede en la clase 3, la segunda mayor ocurre en la clase 2 y la más lenta se da en la clase 1. En las clases 1 y 2 OVO-MD posee las tasas de error más pequeñas, y en la clase 3 el método con esta propiedad es OVR-DWD. En este caso los comportamientos de OVR-SVM y OVR-MDP siguen siendo muy similares para cada clase y para casi todos los valores de d considerados.

Adicionalmente, en esta metodología OVR analizamos también el caso de no esfericidad, tomando tres clases de datos gaussianos d-multivariados con matrices de covarianza spiked $\Sigma_i = \sigma_i^2 S$, con $S = \text{diag}(d^\alpha, 1, 1, \ldots, 1)$, $0 < \alpha < 1$ y $\sigma_i > 0$, para $i = 1, 2, 3$. Consideramos medias de la forma i) con los parámetros n_i's, σ_i's and c_{ij}'s como en el Caso 6, y por ende, se satisfacen las condiciones de a) del Corolario 3.4.2. De nuevo, por el Ejemplo 3.2 de Jung y Marron (2009), cada clase satisface nuestra condición (**3.1**), y se verifica además que este nuevo caso satisface también nuestras condiciones (**3.2**) y (**3.3**). Los resultados de las simulaciones tomando $\alpha = 1/3$

son mostrados en la Figura 12. Se observa que análogamente al Caso 6 (cuando $\Sigma_i = \sigma_i^2 \mathbf{I}_d$) las tasas de error de los cuatro métodos tienden a cero, sin embargo lo hacen de una manera más lenta.

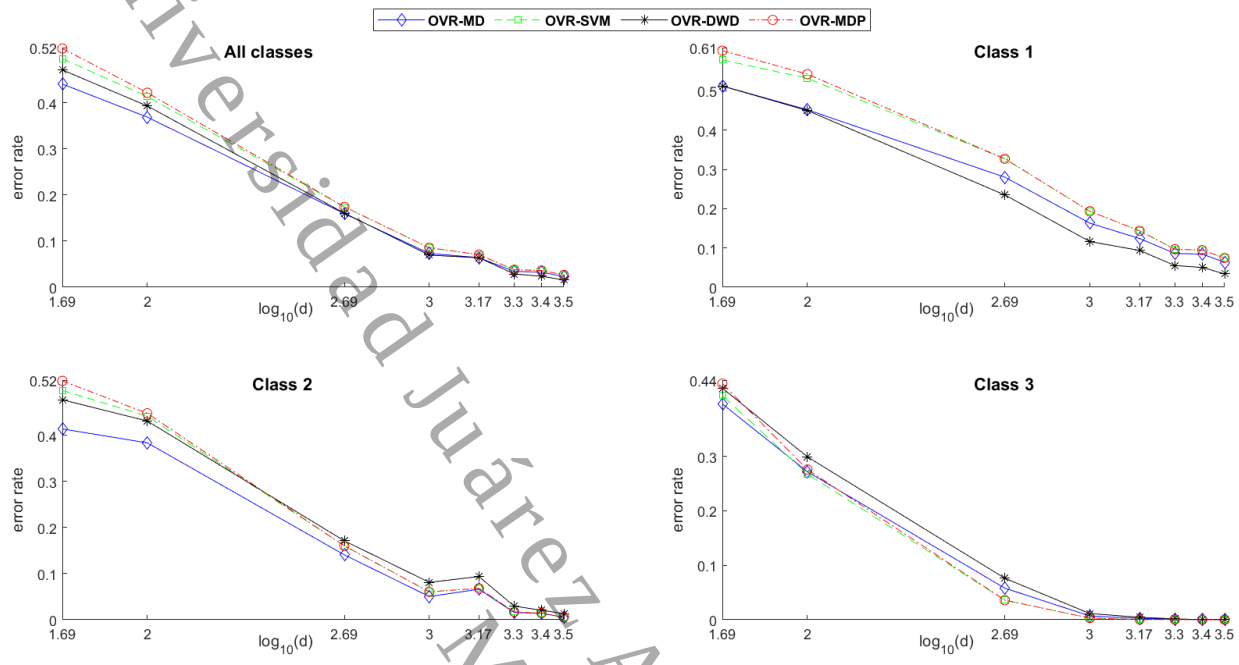


Figura 12: *Tasas de error de clasificación promedio para OVR en el caso de matrices de covarianza spiked.*

4.3. Análisis de Datos Reales

Analizamos datos genéticos empleando MD, SVM, MDP y DWD vía OVO y vía OVR, y evaluamos el desempeño de estos ocho métodos de clasificación multicategoría. Usamos los 64 datos de cáncer de colon proporcionados por Khan et al. (2001), y disponibles en http://bioinf.ucd.ie/people/aedin/R/full_datasets/. Estos datos consisten de 2308 genes para cada uno de los 64 pacientes y son clasificados en: 21 muestras de rabdomiosarcoma (RMS), 23 muestras de sarcoma de Ewing (EWS), 12 muestras de neuroblastoma (NB) y 8 muestras de linfoma de Burkitt (BL). Estos datos fueron analizados en Kento (2022) usando únicamente SVM y DWD vía OVO y vía OVR.

Realizamos primero un análisis exploratorio para determinar si estas cuatro clases poseen la representación geométrica asintótica, verificando que satisfagan las condiciones (3.1)-(3.3). Para hacer esto, tomamos las dimensiones $d = 10, 100, 500, 1000, 1500, 2000, 2308$, y después para

cada d, calculamos el cuadrado de las distancias escaladas (divididas por d) entre los pares de datos de dimensión truncada de cada clase. Análogamente, calculamos también el cuadrado de las distancias escaladas entre los pares de datos de dimensión truncada de dos clases diferentes. Ilustramos los resultados de esto en las figuras 13 y 14, donde también señalamos las medianas (m's) de las distancias para $d = 2308$. Conforme d crece, las gráficas de caja son cada vez más achatadas, indicando que el cuadrado de las distancias escaladas se aproximan a una constante, y sus respectivas medianas muestran una tendencia hacia las m's; por lo cual estas m's se pueden usar para estimar los valores de las distancias asintóticas σ_i's y c_{ij}'s, mediante (**3.1**) y (**3.3**). Calculamos primero las estimaciones de las σ_i^2's de las clases RMS, EWS, NB y BL, que son 0.467, 0.397, 0.233 y 0.211, respectivamente. Luego, calculamos las estimaciones para las c_{ij}^2's de los pares de clases (RMS, EWS), (RMS, NB), (RMS, BL), (EWS, NB), (EWS, BL) y (NB, BL), que son 0.162, 0.189, 0.306, 0.176, 0.249 y 0.208, respectivamente. Esto muestra que los supuestos(**3.1**) y (**3.3**) se cumplen razonablemente para estos datos. El supuesto (**3.2**) es también razonable ya que cuando d crece el cuadrado de las distancias escaladas entre los pares de medias muestrales de las clases se aproximan al valor estimado de las c_{ij}^2's. Concluimos así que estas cuatro clases satisfacen razonablemente las condiciones de la representación geométrica asintótica (**3.1**)-(**3.3**) .

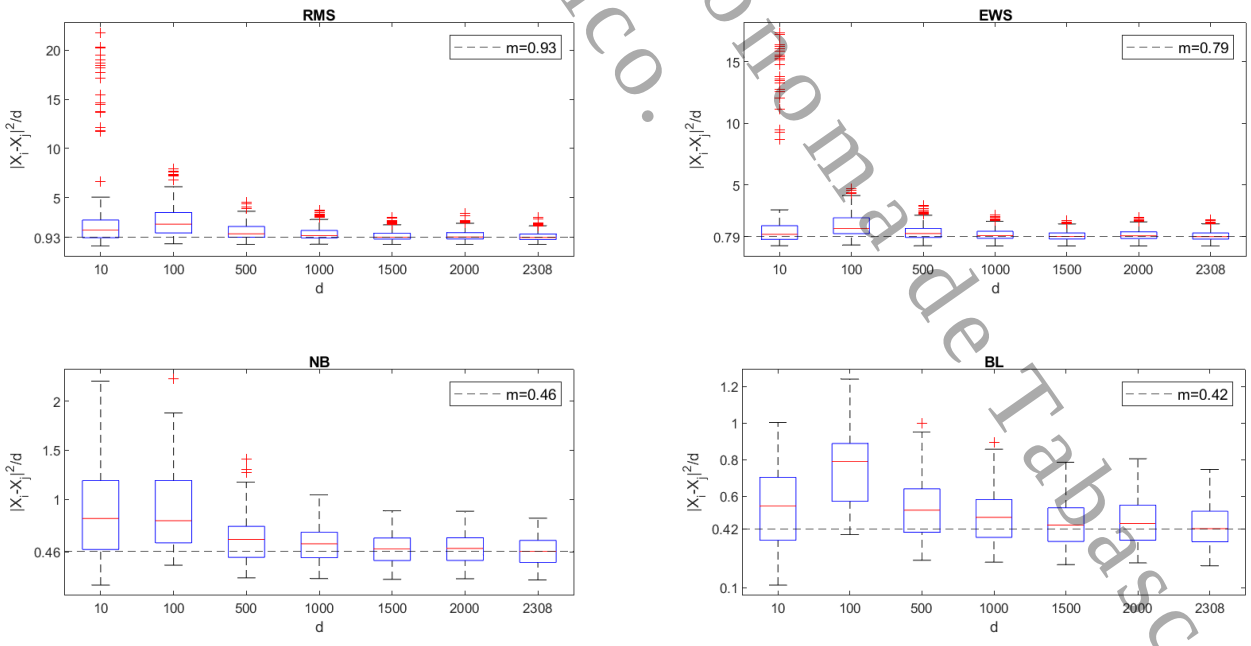


Figura 13: *Gráficas de caja del cuadrado de las distancias escaladas entre pares de datos de dimensión truncada de cada clase, cuando d crece.*

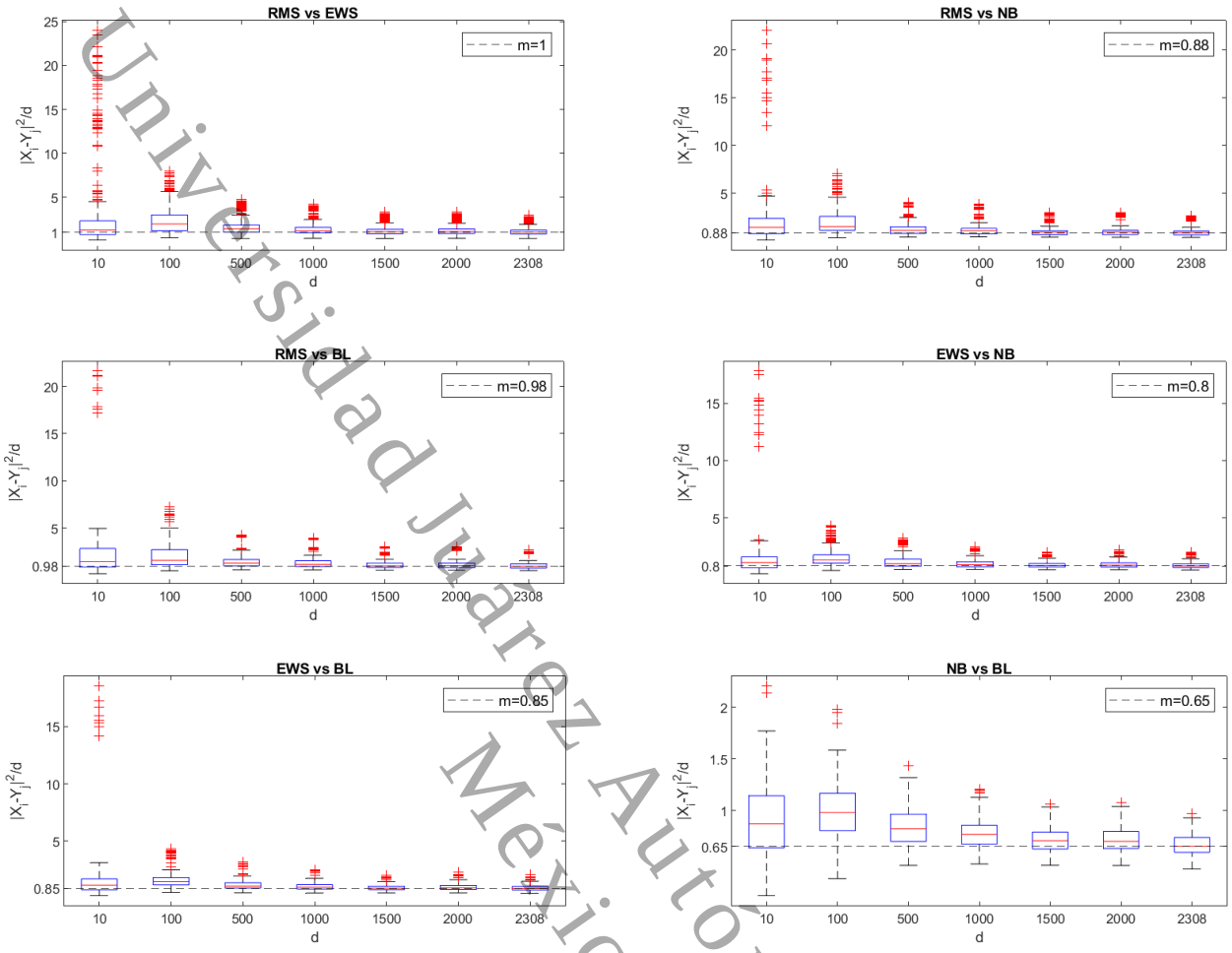


Figura 14: *Gráficas de caja del cuadrado de las distancias escaladas entre pares de datos de dimensión truncada de dos clases diferentes, cuando d crece.*

En segundo lugar, dividimos aleatoriamente los datos de cada clase en datos de entrenamiento y datos de prueba. Luego, construimos los clasificadores usando los datos de entrenamiento, y evaluamos el desempeño de cada método de clasificación multicategoría con los datos de prueba. Repetimos 100 veces este proceso para obtener la tasa de error de clasificación promedio para cada clase y globalmente (para las cuatro clases), y las denotamos por p_1, p_2, p_3, p_4 y p_0, respectivamente. Consideramos dos escenarios.

- **Escenario 1:** Nombrando tal que $\mathcal{C}_1$ ≡RMS, $\mathcal{C}_2$ ≡EWS, $\mathcal{C}_3$ ≡NB y $\mathcal{C}_4$ ≡BL, tomamos $(n_1, n_2, n_3, n_4) = (5, 5, 5, 5)$ datos de entrenamiento y $(21 - n_1, 23 - n_2, 12 - n_3, 8 - n_4)$ datos de prueba.

Vemos en la Tabla 4.1 que globalmente los ocho métodos tienen un buen desempeño, ya que sus tasas de error de clasificación son cercanas a cero. El desempeño de todos los métodos es también muy bueno en cada clase, puesto que las tasas en las clases con menor tamaño muestral, $\mathcal{C}_4 \equiv$ BL y $\mathcal{C}_3 \equiv$ NB, son prácticamente iguales a cero, y las tasas en las clases con mayor tamaño muestral, $\mathcal{C}_1 \equiv$ RMS y $\mathcal{C}_2 \equiv$ EWS, están próximas a cero y son cercanas entre sí, aunque aquellas de OVO-MD y OVR-MD son ligeramente mayores. Para MD, SVM y MDP vía OVO, tenemos que sus tasas son congruentes con a) del Teorema 3.4.1, pues los supuestos $\sigma_i^2/n_i \geq \sigma_j^2/n_j$ y $c_{ij}^2 > \sigma_i^2/n_i - \sigma_j^2/n_j$, para toda $i, j = 1, 2, 3, 4$ con $i < j$, se cumplen con los tamaños muestrales y las estimaciones de las distancias asintóticas, como puede verse en la Tabla 4.2. Similarmente, para OVO-DWD, tenemos que sus tasas están acorde con a) del Teorema 3.4.2, pues las hipótesis $\sigma_i^2/n_i^{3/2} \geq \sigma_j^2/n_j^{3/2}$ y $c_{ij} \geq (n_j/n_i)^{1/2}\sigma_i^2/n_i - \sigma_j^2/n_j$, para toda $i, j = 1, 2, 3, 4$ con $i < j$, también se satisfacen, como vemos nuevamente en la Tabla 4.2. Finalmente, para OVR-MD, tenemos también que sus tasas son congruentes con el Teorema 3.4.4, ya que vemos una vez más en la Tabla 4.2 que la condición necesaria y suficiente $DS_{ii} \geq DS_{ij}$, para toda $j = 1, 2, 3, 4$ con $i \in \{1, 2, 3, 4\}$ fijo, se cumple.

Métodos	p_0	p_1	p_2	p_3	p_4
OVO-MD	0.152	0.328	0.270	0.010	0
OVO-SVM	0.082	0.157	0.161	0.010	0
OVO-MDP	0.081	0.158	0.159	0.010	0
OVO-DWD	0.097	0.188	0.189	0.010	0
OVR-MD	0.130	0.275	0.241	0.007	0
OVR-SVM	0.053	0.099	0.112	0.002	0
OVR-MDP	0.049	0.092	0.102	0.001	0
OVR-DWD	0.077	0.157	0.141	0.010	0

Tabla 4.1: *Tasas de error de clasificación promedio y global para los datos de Khan et al. (2001) en el Escenario 1.*

En cuanto al desempeño entre las metodologías OVO y OVR, vemos que las tasas de error de MD, SVM, MDP y DWD vía OVR son ligeramente menores que sus correspondientes vía OVO,

principalmente en $\mathcal{C}_1 \equiv$ RMS y $\mathcal{C}_2 \equiv$ EWS. Además, mediante OVR y OVO, los métodos MDP y SVM tienen el mejor desempeño, seguidos de DWD y MD.

i	DS_{i1}	DS_{i2}	DS_{i3}	DS_{i4}	n_i	σ_i^2	$\frac{\sigma_i^2}{n_i}$	$\frac{\sigma_i^2}{n_i^{3/2}}$	(i,j)	$\frac{\sigma_i^2}{n_i}$ - $\frac{\sigma_j^2}{n_j}$	c_{ij}^2
1	0.072	-0.144	-0.158	-0.259	5	0.467	0.093	0.041	$(1,2)$	0.014	0.162
2	-0.161	0.062	-0.165	-0.207	5	0.397	0.079	0.035	$(1,3)$	0.046	0.189
3	-0.201	-0.194	0.103	-0.158	5	0.233	0.046	0.020	$(1,4)$	0.051	0.306
4	-0.292	-0.230	-0.147	0.173	5	0.211	0.042	0.018	$(2,3)$	0.032	0.176
									$(2,4)$	0.037	0.249
									$(3,4)$	0.004	0.208

Tabla 4.2: *Valores que verifican las condiciones de a) de los teoremas 3.4.1 y 3.4.2, y del Teorema 3.4.4, en el Escenario* 1.

- **Escenario 2:** Con $\mathcal{C}_1 \equiv$RMS, $\mathcal{C}_2 \equiv$NB, $\mathcal{C}_3 \equiv$BL y $\mathcal{C}_4 \equiv$EWS, tomamos $(n_1, n_2, n_3, n_4) = (10, 5, 5, 10)$ datos de entrenamiento y $(21 - n_1, 12 - n_2, 8 - n_3, 23 - n_4)$ datos de prueba.

En la Tabla 4.3 presentamos las tasas de error de clasificación promedio de los ochos métodos, para cada clase y globalmente. Las conclusiones en este escenario son similares a aquellas obtenidas en el Escenario 1.

Métodos	p_0	p_1	p_2	p_3	p_4
OVO-MD	0.091	0.200	0.007	0	0.159
OVO-SVM	0.024	0.036	0.010	0	0.052
OVO-MDP	0.022	0.034	0.005	0	0.050
OVO-DWD	0.035	0.032	0.061	0	0.048
OVR-MD	0.079	0.150	0.008	0	0.159
OVR-SVM	0.021	0.019	0.008	0	0.059
OVR-MDP	0.019	0.015	0.004	0	0.057
OVR-DWD	0.035	0.024	0.054	0	0.062

Tabla 4.3: *Tasas de error de clasificación promedio y global para los datos de Khan et al. (2001) en el Escenario* 2.

Refiriéndonos a la Tabla 4.4, y así como en el Escenario 1, tenemos que las tasas de error de MD, SVM y MDP vía OVO están acorde con a) del Teorema 3.4.1, y las tasas de error de OVR-MD son consistentes con el Teorema 3.4.4. Para explicar la consistencia de la tasas de error de OVO-DWD mediante a) del Teorema 3.4.2, las clases son renombradas como $\mathcal{C}_1 \equiv$NB, $\mathcal{C}_2 \equiv$BL, $\mathcal{C}_3 \equiv$RMS y $\mathcal{C}_4 \equiv$EWS, y reetiquetando de igual forma a sus correspondientes n_i's, σ_i's y c_{ij}'s.

i	DS_{i1}	DS_{i2}	DS_{i3}	DS_{i4}	n_i	σ_i^2	$\frac{\sigma_i^2}{n_i}$	(i,j)	$\frac{\sigma_i^2}{n_i}$ - $\frac{\sigma_j^2}{n_j}$	c_{ij}^2
1	0.119	-0.194	-0.276	-0.164	10	0.467	0.046	$(1,2)$	0.000	0.189
2	-0.163	0.101	-0.139	-0.154	5	0.233	0.046	$(1,3)$	0.004	0.306
3	-0.253	-0.103	0.176	-0.171	5	0.211	0.042	$(1,4)$	0.007	0.162
4	-0.170	-0.192	-0.223	0.109	10	0.397	0.039	$(2,3)$	0.004	0.208
								$(2,4)$	0.007	0.176
								$(3,4)$	0.002	0.249

Tabla 4.4: *Valores que verifican las condiciones de a) del Teorema 3.4.1, y del Teorema 3.4.4, en el Escenario* 2*.*

Mencionamos por último que las tasas de error de clasificación de los métodos SVM y DWD vía OVO y vía OVR son similares aquellas obtenidas en Kento (2022).

Conclusiones

Asumiendo que K clases de datos multivariados poseen la representación geométrica asintótica cuando la dimensión d de los datos tiende a infinito, mientras que los tamaños de muestra permanecen fijos, obtuvimos condiciones que garantizan clasificación correcta de un nuevo dato de cualquier clase con probabilidad convergiendo a uno cuando d tiende a infinito, para los métodos de clasificación multicategoría OVO-MD, OVO-SVM, OVO-MDP, OVO-DWD y OVR-MD. Dimos estas condiciones en términos de las distancias asintóticas entre los datos y sus medias de clase (σ_i's), la distancia asintótica entre pares de medias de clase (c_{ij}'s) y los tamaños muestrales de las clases (n_i's).

Probamos que los métodos OVO-MD, OVO-SVM y OVO-MDP tienen el mismo comportamiento asintótico, en términos de las probabilidades de clasificación correcta de un nuevo dato de cualquier clase, cuando d tiende a infinito (Teorema 3.4.1), mientras que el método OVO-DWD pudiera tener un comportamiento asintótico diferente, dependiendo de los valores de los parámetros n_i, σ_i y c_{ij}, para $i, j = 1, 2, \ldots, K$ con $i < j$ (Teorema 3.4.2). Estos resultados son generalizaciones del comportamiento asintótico de los métodos de clasificación binaria MD, SVM, MDP y DWD para sus extensiones multicategoría vía la metodología OVO.

Respecto al método OVR-MD proporcionamos condiciones necesarias y suficientes para que un nuevo dato de una clase dada sea correctamente clasificado con probabilidad convergiendo a uno cuando d tiende a infinito (Teorema 3.4.4). Logramos esto determinando en primer lugar el comportamiento asintótico de las distancias signadas de un nuevo dato de una clase dada a cada hiperplano separante del método (Teorema 3.4.3). Obtuvimos también el comportamiento asintótico de OVR-MD, en términos de las probabilidades de clasificación correcta de un nuevo dato de cualquier clase, bajo casos particulares de los valores de los parámetros n_i, σ_i y c_{ij}, para $i, j = 1, 2, \ldots, K$ con $i < j$.

En nuestro estudio de simulación consideramos tres clases de datos gaussianos multivariados y observamos que el comportamiento asintótico cuando d crece de los métodos en cuestión son congruentes con los resultados teóricos proporcionados. Más aún, para la métodología OVO vimos que el método con tasa de error de clasificación promedio más pequeña fue OVO-MD u OVO-DWD, dependiendo del caso en particular y clase. El comportamiento de los métodos OVO-SVM y OVO-MDP fue muy similar en todos los casos y para casi todos los valores de d considerados. En el caso donde los valores de los parámetros respectivos de las tres clases fueron iguales, observamos que las tasas de error de clasificación promedio de los métodos OVO-MD, OVO-SVM, OVO-MDP y OVO-DWD fueron prácticamente iguales para d suficientemente grande. Para la metodología OVR, adicionalmente a OVR-MD consideramos los métodos OVR-SVM, OVR-MDP y OVR-DWD. Similarmente como en la métodología OVO, el método con la tasa de error de clasificación promedio más pequeña fue OVR-MD u OVR-DWD, y los métodos OVR-SVM y OVR-MDP tuvieron un comportamiento muy cercano en todos los casos y para casi todos los valores de d considerados. Además, en el caso donde los valores de los parámetros respectivos fueron los mismos, los resultados fueron muy similares a aquellos de la metodología OVO, principalmente para d grande.

Este trabajo de investigación sugiere que en un futuro se puede abordar lo siguiente:

i) En términos de los parámetros n_i, σ_i y c_{ij}, obtener resultados teóricos que describan las propiedades asintóticas de los métodos de clasificación multicategoría OVR-SVM, OVR-DWD y OVR-MDP, considerando datos multivariados con representación geométrica asintótica, cuando la dimensión d de los datos tiende a infinito y los tamaños muestrales (n_i's) permanecen fijos.

Otro problema en esta línea de investigación que pudiera abordarse en un futuro sería:

ii) Estudiar el comportamiento asintótico del método multiclass distance weigthed discrimination (MDWD), considerando datos multivariados con representación geométrica asintótica, cuando la dimensión de los datos tiende a infinito y los tamaños muestrales permanecen fijos.

Bibliografía

Ash, R. B. (2000). Probability and Measure Theory, Second Edition. Academic Press.

Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis, Third Edition. John Wiley & Sons, Inc.

Ahn, J. & Marron, J. S. (2004). *The Maximal Data Piling in Discrimination*. Manuscrito disponible en http:/ www.cs.unc.edu/Research/MIDAG pubs/papers Biometrika_Ahn_ submit.pdf.

Ahn, J. & Marron, J. S. (2010). *The Maximal Data Piling Direction for Discrimination*. Biometrica, 97(1), 254–259.

Ahn, J., Marron, J. S., Muller, K. M. & Chi, Y. (2007). *The High-Dimension, Low-Sample-Size Geometric Representation Holds Under Mild Conditions*. Biometrika, 94(3), 760–766.

Ben-Israel, Adi & N. E. Greville, Thomas (2003). Generalized Inverses: Theory and Applications. Second Edition. Springer.

Bolívar-Cimé, A. & Marron, J. S. (September-2012). 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.

Bolívar-Cimé, A. & Marron, J. S. (2013). *Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data*. Journal of Multivariate Analysis, 115, 108–121.

Bolívar-Cimé, A. & Córdova-Rodríguez, L. M. (2018). *Binary Discrimination Methods for High-Dimensional Data with a Geometric Representation*. Communications in Statistics - Theory and Methods, 47(11), 2720–2740.

Bolívar-Cimé A. (2021). *More About Asymptotic Properties of Some Binary Classification Methods for High Dimensional Data*. In: Hernández-Hernández D., Leonardi F., Mena R. H., Pardo Millán J. C. (eds) Advances in Probability and Mathematical Statistics. Progress in Probability, vol 79. Birkhäuser, Cham.

Cortes, C. & Vapnik, V. N. (1995). *Support-Vector Networks*. Machine Learning, 20(3), 273–297.

Crammer, K. & Singer, Y. (2000). *On the Learnability and Design of Output Codes for Multiclass Problems*. In Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, 35–46.

Egashira, K., K. Yata, and M. Aoshima. (2021). *Asymptotic Properties of Distance-Weighted Discrimination and Its Bias Correction for High-Dimension, Low-Sample-Size Data*. Japanese Journal of Statistics and Data Science 4: 821–840.

Egashira, K. (2022). *Asymptotic Properties of Multiclass Support Vector Machine under High Dimensional Settings*. Communications in Statistics - Simulation and Computation, 51(6), 1–15.

García-Cerino, D., Bolívar-Cimé, A. & Pérez-Abreu, V. (2024). *Asymptotic Behavior of Some Multicategory Classification Methods for High-Dimensional Data*. Communications in Statistics - Simulation and Computation, 53(4), 1–26.

G. Casella & R. L. Berger. (2002). Statistical Inference, Second Edition. Duxbury.

Hall, P., Marron, J. S. & Neeman, A. (2005). *Geometric Representation of High Dimension, Low Sample Size Data*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(3), 427–444.

Huang, H., Liu, Y., Du, Y., Perou, C. M., Hayes, D. N., Todd, M. J. & Marron, J. S. (2010). *Multiclass Distance Weighted Discrimination with Applications to Batch Adjustment*. Manuscrito disponible en https://people.orie.cornell.edu/ miketodd/multidwd.pdf.

Huang, H., Liu, Y., Du, Y., Perou, C. M., Hayes, D. N., Todd, M. J. & Marron, J. S. (2013). *Multiclass Distance-Weighted Discrimination*. Journal of Computational and Graphical Statistics, 22(4), 953–969.

Hastie, T., R. Tibshirani & J. Friedman. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer.

Izenman, A. J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer.

Johnstone, I. M. (2001). *On the Distribution of the Largest Eigenvalue in Principal Components Analysis*. The Annals of Statistics, 29(2), 295–327.

Jung, S. & Marron, J. S. (2009). *PCA Consistency in High Dimension, Low Sample Size Context*. The Annals of Statistics, 37(6B), 4104–4130.

Khan, J., J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, et al. (2001). *Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks*. Nature Medicine 7(6), 673–679.

Lee, Y., Lin, Y. & Wahba, G. (2004). *Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data*. Journal of the American Statistical Association, 99(465), 67–82.

Muirhead, R. J. (2005). Aspects of Multivariate Statistical Analysis. John Wiley & Sons, Inc.

Marron, J. S., Todd, M. J. & Ahn, J. (February-2007). *Distance-Weighted Discrimination*. Journal of the American Statistical Association. Manuscrito disponible en http://www.stat.uga.edu/ jyahn/DWD/.

Marron, J. S., Todd, M. J. & Ahn, J. (2007). *Distance-Weighted Discrimination*. Journal of the American Statistical Association, 102(480), 1267–1271.

Nakayama, Y., K. Yata, and M. Aoshima. (2017). *Support Vector Machine and Its Bias Correction in High-Dimension, Low-Sample-Size Settings*. Journal of Statistical Planning and Inference 191: 88–100.

Qiao, X., Zhang, H., Liu, Y., Todd, M. J. & Marron, J. S. (2010). *Weighted Distance Weighted Discrimination and Its Asymptotic Properties*. Journal of the American Statistical Association, 105(489), 401–414.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., & Fei-Fei, L. (2015). *Imagenet Large Scale Visual Recognition Challenge*. International Journal of Computer Vision, 115(3), 211–252.

Scholkopf, B. & Smola, A. J. (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. Cambridge, Massachusetts: The MIT Press.

Serfling, R. J. (2002). Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Inc.

Wang, J. (2012). *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer.

Weston, J. & Watkins, C. (1999). *Support Vector Machines for Multi-class Pattern Recognition*. In Proceedings of the Seventh Esann, 219–224.

Yata, K. & Aoshima, M. (2012). *Effective PCA for High-Dimension, Low-Sample-Size Data with Noise Reduction Via Geometric Representations*. Journal of Multivariate Analysis, 105(1), 193–215.

Yu, D. & Deng, L. (2016). Automatic Speech Recognition. Springer.

Anexos

Se muestran a continuación los códigos de programación en MATLAB que fueron empleados para obtener los resultados que son visualizados en las figuras y tablas de las secciones 4.1, 4.2 y 4.3 de esta tesis.

Algoritmos para las Simulaciones

- Se exponen primero los algoritmos de las simulaciones para la metodología OVO.

  Caso 1:

```
n=[10,10,10]; %vector de los tamaños muestrales n_{i}´s
K=length(n); %número total de clases
s =[1,1,1]; %vector de las \sigma_{i}´s
c=[0.3,0.3,0.3]; %c_{ij}´s, c_{12}=c(1),c_{13}=c(2),c_{23}=c(3)
y=[ones(n(1),1);ones(n(2),1)*2;ones(n(3),1)*3]; %etiquetas
d=[50,100,500,1000,1500]; %vector de dimensiones d_{i}
e0=[]; %matrix de tasas de error de clasificación promedio global
       %para cada d_{i} y método
e1=[];
e2=[]; %matrix de tasas de error de clasificación promedio de la
       %clase 2 para cada d_{i} y método
e3=[];
m=100; %número total de datos de prueba
a=500; %total de conjuntos de datos entrenamiento para cada d_{i}
```

```
   for j=1:length(d)
O=zeros(d(j)-2,1); %vector de ceros de dimensión d_{j}-2
mu1=[0; 0; O];
mu2=[c(1)*sqrt(d(j)); 0; O]; %media de la clase 2
mu3=[cos(pi/3)*c(2)*sqrt(d(j)); sin(pi/3)*c(2)*sqrt(d(j)); O];
Id=eye(d(j)); %matrix de covarianza identidad
%matrix de datos de prueba ordenados por clase
X0=[mvnrnd(mu1,Id*(s(1))^2,m);mvnrnd(mu2,Id*(s(2))^2,m);
    mvnrnd(mu3,Id*(s(3))^2,m)];
f=zeros(K,4); %matrix de total de clasificaciones incorrectas por
              %clase y método, $d_{i}$ fijo
    for r= 1:a
%matrix de datos de entrenamiento ordenados por clase
X=[mvnrnd(mu1,Id*(s(1))^2,n(1));mvnrnd(mu2,Id*(s(2))^2,n(2));
   mvnrnd(mu3,Id*(s(3))^2,n(3))];
     for t= 1:K
fmd=0; %contador de clasificaciones incorrectas realizadas por
       %el método MD
fsvm=0;
fdwd=0;
fmdp=0;
      for i= ((t-1)*m+1):(t*m)
%Apartir de aquí usamos la funciones OVO (OVR en su caso), MED,
%SVM, DWD y MDP (códigos de la metodología OVO (OVR) y de los
%métodos de clasificación binaria, en este caso MD ('MED'))

[pclasemd,valfunmd,vectmd,b0md]=OVO(X, y, K, X0(i,:), 'MED');
   if pclasemd ~=t
      fmd=fmd + 1;
   end
```

```
[pclasesvm,valfunsvm,vectsvm,b0svm]=OVO(X, y, K, X0(i,:), 'SVM');
   if pclasesvm ~=t
      fsvm=fsvm + 1;
   end
[pclasedwd,valfundwd,vectdwd,b0dwd]=OVO(X, y, K, X0(i,:), 'DWD');
   if pclasedwd ~=t
      fdwd=fdwd + 1;
   end
[pclasemdp,valfunmdp,vectmdp,b0mdp]=OVO(X, y, K, X0(i,:), 'MDP');
   if pclasemdp ~=t
      fmdp=fmdp + 1;
   end
      end
 f(t,:)=f(t,:) + [fmd,fsvm,fdwd,fmdp];
    end
  end
e0=[e0; sum(f)/(3*a*m)];
e1=[e1; f(1,:)/(a*m)];
e2=[e2; f(2,:)/(a*m)];
e3=[e3; f(3,:)/(a*m)];
   end
%A partir de aquí el código es para visualizar las figuras
g=log10(d);
figure
subplot(2,2,1)
plot(g,e0(:,1),'-db',g,e0(:,2),'--sg',g,e0(:,3),'-*k',g,e0(:,4),
     '-.or')
title('Todas las clases')
xlabel('log_{10}(d)')
ylabel('Tasa de error')
```

```
axis([log10(50) log10(1500) 0 0.47])
subplot(2,2,2)
plot(g,e1(:,1),'-db',g,e1(:,2),'--sg',g,e1(:,3),'-*k',g,e1(:,4),
     '-.or')
title('Clase1')
xlabel('log_{10}(d)')
ylabel('Tasa de error')
legend('OVO-MD','OVO-SVM','OVO-DWD','OVO-MDP')
axis([log10(50) log10(1500) 0 0.47])
subplot(2,2,3)
plot(g,e2(:,1),'-db',g,e2(:,2),'--sg',g,e2(:,3),'-*k',g,e2(:,4),
     '-.or')
title('Clase 2')
xlabel('log_{10}(d)')
ylabel('Tasa de error')
axis([log10(50) log10(1500) 0 0.48])
subplot(2,2,4)
plot(g,e3(:,1),'-db',g,e3(:,2),'--sg',g,e3(:,3),'-*k',g,e3(:,4),
     '-.or')
title('Clase 3')
xlabel('log_{10}(d)')
ylabel('Tasa de error')
axis([log10(50) log10(1500) 0 0.47])
```

Caso 2: Se procedió con el mismo código del Caso 1 pero con los siguientes parámetros y conjunto de medias:

```
n=[10,8,7];
s=[1.4,1,0.7];
c=[0.72,0.42,0.3];
```

```
d=[50,100,500,1000,1500,2000,2500,3000];
mu1=zeros(d(j),1);
mu2=ones(d(j),1)*c(1);
mu3=[ones(d(j)-1,1); -1]*c(2);
```

Caso 3: Se uso el mismo código del Caso 1 pero con `n=[7,9,10];`, `s=[1.3,1,0.8];`, `c=[0.3,0.2,0.1];` y el mismo conjunto de medias del Caso 2.

Caso Spiked: Se realizó como en el Caso 2 pero en vez de usar el conjunto de matrices de covarianza `Id*(s(j))^2, j=1,2,3,` se usaron las matrices de covarianza spiked `diag([d(j)^(1/3);ones(d(j)-1,1)])*(s(j))^2, j=1,2,3.`

- Se expone ahora sobre los algoritmos de las simulaciones para la metodología OVR.

Todos los Casos: En todos los casos se procedió con el mismo código del Caso 1 de OVO, sólo que en vez de emplear la función OVO se utilizó la función OVR, es decir, cambiar en este código OVO por OVR. Desde luego que se usaron los valores de los parámetros n_i´s, σ_i´s y c_{ij}´s que corresponden a cada caso. Además, sólo para el Caso 3 y el Caso Spiked se consideraron las dimensiones `d=[50,100,500,1000,1500,2000,2500,3000].` Por último, el Caso Spiked se hizo como el Caso 6 pero en vez de utilizar las matrices de covarianza `Id*(s(j))^2, j=1,2,3,` se emplearon las matrices de covarianza spiked `diag([d(j)^(1/3);ones(d(j)-1,1)])*(s(j))^2, j=1,2,3.`

Algoritmos para el Análisis de Datos Reales

Código para estimar las distancias asintóticas σ_i's y c_{ij}'s, y ver las gráficas de caja

```
%matrix de todos los datos de tamaño 2308x64
M=readmatrix('khan_training.xlsx','Range','B:BM');
C1=M(:,1:23); %datos EWS
C2=M(:,24:31); %matrix de datos BL de 2308x8
C3=M(:,32:43); %datos NB
C4=M(:,44:64); %datos RMS
```

```
T=[size(C1);size(C2);size(C3);size(C4)];
dim=T(1,1); %dimensión 2308
n=T(:,2); %tamaños muestrales ordenados de las clases
p=[(n(1)-1)*n(1)/2,(n(2)-1)*n(2)/2,(n(3)-1)*n(3)/2,(n(4)-1)*
   n(4)/2]; %vector de número de parejas de datos por cada Ci
d=[10,100,500,1000,1500,2000,dim]; %dimensiones a truncar
%matrix de distancias al cuadrado divididas por d_{i} entre pares
%de datos de C1
D1=zeros(p(1),length(d));
   for r=1:length(d)
s=0;
t=d(r);
    for i=1:(n(1)-1)
     for j=(i+1):n(1)
        s=s+1;
        D1(s,r)=(norm(C1(1:t,i)-C1(1:t,j)))^2/t;
        end
     end
   end
%Distacias al cuadrado reescaladas entre pares de datos de C2
D2=zeros(p(2),length(d));
   for r=1:length(d)
s=0;
t=d(r);
    for i=1:(n(2)-1)
     for j=(i+1):n(2)
        s=s+1;
        D2(s,r)=(norm(C2(1:t,i)-C2(1:t,j)))^2/t;
        end
     end
```

```
    end
%Distacias al cuadrado reescaladas entre pares de datos de C3
D3=zeros(p(3),length(d));
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:(n(3)-1)
      for j=(i+1):n(3)
         s=s+1;
         D3(s,r)=(norm(C3(1:t,i)-C3(1:t,j)))^2/t;
         end
      end
    end
%Distacias al cuadrado reescaladas entre pares de datos de C4
D4=zeros(p(4),length(d));
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:(n(4)-1)
      for j=(i+1):n(4)
         s=s+1;
         D4(s,r)=(norm(C4(1:t,i)-C4(1:t,j)))^2/t;
         end
      end
    end
%vector de número de parejas de datos por cada CiCj, i<j
q=[n(1)*n(2),n(1)*n(3),n(1)*n(4),n(2)*n(3),n(2)*n(4),n(3)*n(4)];
%matrix de distancias al cuadrado divididas por d_{i} entre pares
%de datos de C1 y C2
D12=zeros(q(1),length(d));
```

```
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:n(1)
      for j=1:n(2)
       s=s+1;
       D12(s,r)=(norm(C1(1:t,i)-C2(1:t,j)))^2/t;
       end
     end
    end
%Distacias al cuadrado reescaladas entre pares de datos de C1,C3
D13=zeros(q(2),length(d));
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:n(1)
      for j=1:n(3)
         s=s+1;
         D13(s,r)=(norm(C1(1:t,i)-C3(1:t,j)))^2/t;
         end
      end
    end
%Distacias al cuadrado reescaladas entre pares de datos de C1,C4
D14=zeros(q(3),length(d));
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:n(1)
      for j=1:n(4)
         s=s+1;
```

```
      D14(s,r)=(norm(C1(1:t,i)-C4(1:t,j)))^2/t;
      end
     end
    end
%Distacias al cuadrado reescaladas entre pares de datos de C2,C3
D23=zeros(q(4),length(d));
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:n(2)
      for j=1:n(3)
         s=s+1;
         D23(s,r)=(norm(C2(1:t,i)-C3(1:t,j)))^2/t;
         end
     end
    end
%Distacias al cuadrado reescaladas entre pares de datos de C2,C4
D24=zeros(q(5),length(d));
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:n(2)
      for j=1:n(4)
         s=s+1;
         D24(s,r)=(norm(C2(1:t,i)-C4(1:t,j)))^2/t;
         end
     end
    end
%Distacias al cuadrado reescaladas entre pares de datos de C3,C4
D34=zeros(q(6),length(d));
```

```
    for r=1:length(d)
s=0;
t=d(r);
     for i=1:n(3)
      for j=1:n(4)
         s=s+1;
         D34(s,r)=(norm(C3(1:t,i)-C4(1:t,j)))^2/t;
         end
      end
    end
%mediana de ||X_i^{1}-X_j^{1}||^2/d, i,j=1,2,...,n_{1}
m1=median(D1(:,length(d)));
s1=m1/2; %sigma_{1}^{2}
m2=median(D2(:,length(d)));
s2=m2/2;
m3=median(D3(:,length(d)));
s3=m3/2;
m4=median(D4(:,length(d)));
s4=m4/2;
m12=median(D12(:,length(d)));
c12=m12-s1-s2; %c_rs~|X_i^{r}-X_j^{s}|^2/d - s_r -s_s
m13=median(D13(:,length(d)));
c13=m13-s1-s3;
m14=median(D14(:,length(d)));
c14=m14-s1-s4;
m23=median(D23(:,length(d)));
c23=m23-s2-s3;
m24=median(D24(:,length(d)));
c24=m24-s2-s4;
m34=median(D34(:,length(d)));
```

```
c34=m34-s3-s4;
s=[s1,s2,s3,s4]; %vector de las sigma_{i}^2
c=[c12,c13,c14,c23,c24,c34]; %vector de las c_{i}^2
%Por ejemplo, para ver la gráfica de caja de EWS se utilizó
figure
boxplot([D1(:,1),D1(:,2),D1(:,3),D1(:,4),D1(:,5),D1(:,6),
D1(:,length(d))],{'10','100','500','1000','1500','2000','2308'})
title(´EWS')
xlabel('d')
ylabel('|X_i-X_j|^2/d')
yline(m1,'--k')
legend('m')
% Para visualizar la gráfica de caja de RMS vs EWS  se usó
figure
boxplot([D14(:,1),D14(:,2),D14(:,3),D14(:,4),D14(:,5),D14(:,6),
D14(:,length(d))],{'10','100','500','1000','1500','2000','2308'})
title('RMS vs EWS')
xlabel('d')
ylabel('|X_i-X_j|^2/d')
yline(m14,'--k')
legend('m')
```

Código para ver las tasas de error de clasificación promedio de las tablas 4.1 y 4.3

```
%matrix de todos los datos de tamaño 2308x64
M=readmatrix('khan_training.xlsx','Range','B:BM');
C1=M(:,1:23); %datos EWS
C2=M(:,24:31); %matrix de datos BL de 2308x8
C3=M(:,32:43); %datos NB
C4=M(:,44:64); %datos RMS
T=[size(C1);size(C2);size(C3);size(C4)];
```

```
dim=T(1,1); %dimensión 2308
n=T(:,2); %tamaños muestrales ordenados de las clases
K=length(n); %número total de clases
m=[10,5,5,10]; %vector de números de datos de entrenamiento
%vector de números de datos de prueba
q=[0,n(1)-m(1),n(2)-m(2),n(3)-m(3),n(4)-m(4)];
%etiquetas ordenadas
I=[ones(m(1),1);ones(m(2),1)*2;ones(m(3),1)*3;ones(m(4),1)*4];
e0=[]; %matrix de tasas de error de clasificación promedio global
      %de cada método y d=2308
e1=[];
e2=[];%matrix de tasas de error de clasificación promedio clase 2
        %de cada método y d=2308
e3=[];
e4=[];
a=100; %número de conjunto de datos de entrenamiento para d=2308
%matrix de tasas de error de clasificación por clase y método
f=zeros(K,4);
rng('default')
     for r= 1:a
%division de C1 en datos de entrenamiento y datos de prueba
i1=sort(randsample(n(1),m(1))); %índices datos entrenamiento
E1=C1(:,i1); %datos de entrenamiento de C1
j1=setdiff(1:n(1),i1); %índices de datos de prueba
P1=C1(:,j1); %datos de prueba de C1
%division C2
i2=sort(randsample(n(2),m(2)));
E2=C2(:,i2);
j2=setdiff(1:n(2),i2);
P2=C2(:,j2);
```

```
%division C3
i3=sort(randsample(n(3),m(3));
E3=C3(:,i3);
j3=setdiff(1:n(3),i3);
P3=C3(:,j3);
%division C4
i4=sort(randsample(n(4),m(4)));
E4=C4(:,i4);
j4=setdiff(1:n(4),i4);
P4=C4(:,j4);
E=['E1';E2';E3';E4']; %matrix de datos de entrenamiento
P=['P1';P2';P3';P4']; %matrix de datos de prueba
%Apartir de aquí usamos la funciones OVO (OVR en su caso), MED,
%SVM, DWD y MDP (códigos de la metodología OVO (OVR) y de los
%métodos de clasificación binaria, en este caso MD ('MED'))
       for t= 1:K
fmd=0;
fsvm=0;
fdwd=0;
fmdp=0;
          for i= (sum(q(1:t))+1):(sum(q(1:t+1)))
[pclasemd,valfunmd,vectmd,b0md]=OVO(E, I, K, P(i,:), 'MED');
          if pclasemd ~=t
          fmd=fmd + 1;
          end
[pclasesvm,valfunsvm,vectsvm,b0svm]=OVO(E, I, K, P(i,:), 'SVM');
          if pclasesvm ~=t
          fsvm=fsvm + 1;
          end
[pclasedwd,valfundwd,vectdwd,b0dwd]=OVO(E, I, K, P(i,:), 'DWD');
```

```
          if pclasedwd ~=t
          fdwd=fdwd + 1;
          end
[pclasemdp,valfunmdp,vectmdp,b0mdp]=OVO(E, I, K, P(i,:), 'MDP');
          if pclasemdp ~=t
          fmdp=fmdp + 1;
          end
     end
f(t,:)=f(t,:) + [fmd/q(t+1),fsvm/q(t+1),fdwd/q(t+1),fmdp/q(t+1)];
   end
  end
e0=[e0; sum(f)/(4*a)];
e1=[e1; f(1,:)/a];
e2=[e2; f(2,:)/a];
e3=[e3; f(3,:)/a];
e4=[e4; f(4,:)/a];
```

Algoritmo de las distancias signadas DS_{ij} de los teoremas 3.4.3 y 3.4.4, para determinar si $D_{ii} \geq DS_{ij}$ para toda $j = 1, 2, ..., K$.

```
n=[10,5,5,10]; %vector de los tamaños muestrales n_{i}´s
%vector de las distancias asintóticas sigma_{i}´s
s=[0.467,0.233,0.211,0.397].^(1/2);
%matrix simétrica de las distancias asintóticas c_{ij}, c_{ii}=0
c=[0,0.189,0.306,0.162;0.189,0,0.208,0.176;0.306,0.208,0,0.249;
   0.162,0.176,0.249,0].^(1/2);
K=length(n); %número total de clases
%matrix DS_{ij} para ver que DS_{ii}>=DS_{ij} para toda
%j=1,...,K (´´clasificar correctamente nuevos datos de clase i")
DS=zeros(K,K);
     for i=1:K
```

```
     for j=1:K
nm=n;
nm([i,j])=[]; %n sin las entradas i y j
sm=s;
sm([i,j])=[];
rm=c(i,:);
rm([i,j])=[]; %renglón i de c sin las entradas i y j
cm=c;
cm([i,j],:)=[]; %c sin los renglones i y j
cm(:,[i,j])=[]; %c sin las columnas i y j
          if j==i
          v1=0;
            for q=1:(K-2)
              for r=(q+1):(K-1)
v1=v1 + nm(q)*nm(r)*(rm(q)^2 + rm(r)^2 - cm(q,r)^2);
              end
            end
DS(i,j)=DS(i,j)+(dot(nm.^2,rm.^2)+v1-(s(i)*sum(nm))^(2)/n(i)+
dot(nm,sm.^2))/(2*sum(nm)*(dot(nm.^2,rm.^2)+v1+(s(i)*
sum(nm))^(2)/n(i)+dot(nm,sm.^2))^(1/2));
              else
              v2=0;
              for p=1:(K-3)
               for t=(p+1):(K-2)
v2=v2 + nm(p)*nm(t)*(rm(p)^2 + rm(t)^2 - cm(p,t)^2);
               end
             end
nmj=n;
nmj(j)=[]; %n sin la entrada j
smj=s;
```

```
smj(j)=[];
rmj=c(j,:);
rmj(j)=[]; %renglón j de c sin la  entrada j
cmj=c;
cmj(j,:)=[]; %c sin el renglón j
cmj(:,j)=[]; %c sin la columna j
v3=0;
          for a=1:(K-2)
            for b=(a+1):(K-1)
v3=v3 + nmj(a)*nmj(b)*(rmj(a)^2 + rmj(b)^2 - cmj(a,b)^2);
             end
          end
DS(i,j)=DS(i,j)+(dot(nm.^2,rm.^2)+v2-(sum(nmj)*c(i,j))^2-
(s(j)*sum(nmj))^(2)/n(j)+dot(nmj,smj.^2))/(2*sum(nmj)*
(dot(nmj.^2,rmj.^2)+v3+(s(j)*sum(nmj))^(2)/n(j)+
 dot(nmj,smj.^2))^(1/2));
        end
     end
  end
```

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la Tesis:	Clasificación Multicategoría para Datos de Dimensión Alta
Autor de la Tesis:	Dorilian García Cerino
ORCID:	https://orcid.org/0009-0007-5631-4147
Resumen de la Tesis:	Se consideran las extensiones multicategoría de los métodos mean difference (MD), support vector machine (SVM), maximal data piling (MDP) y distance weighted discrimination (DWD) mediante la metodología one versus one (OVO), y la extensión multicategoría de MD mediante la metodología uno contra el resto (OVR), en el contexto de datos de alta dimensión alta. Se describe el comportamiento asintótico de estos cinco métodos de clasificación multicategoría cuando la dimensión de los datos crece y el tamaño de muestra es fijo, en términos de las probabilidades de clasificación correcta de un nuevo dato. En el caso OVO, se encuentran condiciones suficientes para que estas probabilidades converjan a uno conforme la dimensión tiende a infinito, y se ve que al igual que en el caso binario, OVO-MD, OVO-SVM y OVO-MDP tienen el mismo comportamiento asintótico, mientras que OVO-DWD podría comportarse diferente. Para OVR-MD se proporcionan condiciones necesarias y suficientes para las probabilidades de clasificación correcta converjan a uno. Se realizan simulaciones para comparar aún más los métodos, y se consideran adicionalmente OVR-SVM, OVR-MDP y OVR-DWD. Se evalúa el rendimiento de estos ocho métodos de clasificación multicategoría utilizando un conjunto de datos de expresión genética.
Palabras Claves de la Tesis:	clasificación multicategoría, datos de dimensión alta, uno contra uno, uno contra el resto, máquinas de vector soporte.
Referencias Citadas:	Ash, R. B. (2000). Probability and Measure Theory, Second Edition. Academic Press. Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis, Third Edition. John Wiley & Sons, Inc.

Ahn, J. & Marron, J. S. (2004). The Maximal Data Piling in Discrimination. Manuscrito disponible en http:/ www.cs.unc.edu/Research/MIDAG pubs/papers Biometrika_Ahn_ submit.pdf.

Ahn, J. & Marron, J. S. (2010). The Maximal Data Piling Direction for Discrimination. Biometrica, 97(1), 254--259.

Ahn, J., Marron, J. S., Muller, K. M. & Chi, Y. (2007). The High-Dimension, Low-Sample-Size Geometric Representation Holds Under Mild Conditions. Biometrika, 94(3), 760--766.

Ben-Israel, Adi & N. E. Greville, Thomas (2003). Generalized Inverses: Theory and Applications. Second Edition. Springer.

Bolívar-Cimé, A. & Marron, J. S. (2012). Supplementary Material for Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data. Manuscrito en https://sites.google.com/site/addybolivarcime/publications/supplmentary-files.

Bolívar-Cimé, A. & Marron, J. S. (2013). Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data. Journal of Multivariate Analysis, 115, 108--121.

Bolívar-Cimé, A. & Córdova-Rodríguez, L. M. (2018). Binary Discrimination Methods for High-Dimensional Data with a Geometric Representation. Communications in Statistics – Theory and Methods, 47(11), 2720--2740.

Bolívar-Cimé A. (2021). More About Asymptotic Properties of Some Binary Classification Methods for High Dimensional Data. In: Hernández-Hernández D., Leonardi F., Mena R. H., Pardo Millán J. C. (eds) Advances in Probability and Mathematical Statistics. Progress in Probability, vol 79. Birkhäuser, Cham.

Cortes, C. & Vapnik, V. N. (1995). Support-Vector Networks. Machine Learning, 20(3), 273--297.

Crammer, K. & Singer, Y. (2000). On the Learnability and Design of Output Codes for Multiclass Problems. In Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, 35--46.

Egashira, K., K. Yata, and M. Aoshima. (2021). Asymptotic Properties of Distance-Weighted Discrimination and Its Bias Correction for High-Dimension, Low-Sample-Size Data. Japanese Journal of Statistics and Data Science 4: 821--840.

Egashira, K. (2022). Asymptotic Properties of Multiclass Support Vector Machine under High Dimensional Settings. Communications in Statistics - Simulation and Computation, 51(6), 1--15.

García-Cerino, D., Bolívar-Cimé, A. & Pérez-Abreu, V. (2024). Asymptotic Behavior of Some Multicategory Classification Methods for High-Dimensional Data. Communications in Statistics - Simulation and Computation, 53(4), 1--26.

G. Casella & R. L. Berger. (2002). Statistical Inference, Second Edition. Duxbury.

Hall, P., Marron, J. S. & Neeman, A. (2005). Geometric Representation of High Dimension, Low Sample Size Data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(3), 427--444.

Huang, H., Liu, Y., Du, Y., Perou, C. M., Hayes, D. N., Todd, M. J. & Marron, J. S. (2010). Multiclass Distance Weighted Discrimination with Applications to Batch Adjustment. Manuscrito disponible en https://people.orie.cornell.edu/~miketodd/multidwd.pdf.

Huang, H., Liu, Y., Du, Y., Perou, C. M., Hayes, D. N., Todd, M. J. & Marron, J. S. (2013). Multiclass Distance-Weighted Discrimination. Journal of Computational and Graphical Statistics, 22(4), 953--969.

Hastie, T., R. Tibshirani & J. Friedman. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer.

Izenman, A. J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer.

Johnstone, I. M. (2001). On the Distribution of the Largest Eigenvalue in Principal Components Analysis. The Annals of Statistics, 29(2), 295--327.

Jung, S. & Marron, J. S. (2009). PCA Consistency in High Dimension, Low Sample Size Context. The Annals of Statistics, 37(6B), 4104--4130.

Khan, J., J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, et al. (2001). Classification and Diagnostic Prediction of Cancers Using Gene Expression Profilin and Artificial Neural Networks. Nature Medicine 7(6), 673--679.

Lee, Y., Lin, Y. & Wahba, G. (2004). Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data. Journal of the American Statistical Association, 99(465), 67--82.

Muirhead, R. J. (2005). Aspects of Multivariate Statistical Analysis. John Wiley & Sons, Inc.

Marron, J. S., Todd, M. J. & Ahn, J. (February-2007). Distance Weighted Discrimination. Journal of the American Statistical Association. Manuscrito disponible en http://www.stat.uga.edu/~jyahn/DWD/.

Marron, J. S., Todd, M. J. & Ahn, J. (2007). \emph{Distance Weighted Discrimination}. Journal of the American Statistical Association, 102(480), 1267--1271.

Nakayama, Y., K. Yata, & M. Aoshima. (2017). Support Vector Machine and Its Bias Correction in High-Dimension, Low-Sample Size Settings. Journal of Statistical Planning and Inference 191:88--100.

Qiao, X., Zhang, H., Liu, Y., Todd, M. J. & Marron, J. S. (2010). Weighted Distance Weighted Discrimination and Its Asymptotic Properties. Journal of the American Statistical Association, 105(489), 401--414.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., & Fei-Fei, L. (2015). Imagenet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211--252.

Scholkopf, B. & Smola, A. J. (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. Cambridge, Massachusetts: The MIT Press.

Serfling, R. J. (2002). Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Inc.

Wang, J. (2012). Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Springer.

Weston, J. & Watkins, C. (1999). Support Vector Machines for Multi-class Pattern Recognition. In Proceedings of the Seventh Esann, 219--224.

Yata, K. & Aoshima, M. (2012). Effective PCA for High Dimension, Low-Sample-Size Data with Noise Reduction Via Geometric Representations. Journal of Multivariate Analysis, 105(1), 193--215.

Yu, D. & Deng, L. (2016). Automatic Speech Recognition. Springer.