Universidad de Valladolid

Facultad de Ciencias

Trabajo de Fin de Grado

Grado en Matemáticas

Reducción de dimensionalidad y teorema de Johnson-Lindenstrauss

Autor: Daniel González Martín
Tutor: Luis Ángel García Escudero

23 de junio de 2025

Resumen

En Análisis de Datos y Aprendizaje Automático, donde el manejo de datos de alta dimensión es cada vez más común, la reducción de dimensionalidad es una técnica clave para mejorar la eficiencia computacional y reducir los dañinos efectos de la conocida "maldición de la dimensionalidad". El teorema de Johnson-Lindenstrauss es un resultado fundamental, que establece que un conjunto de datos en un espacio de alta dimensión puede ser proyectado en un espacio de menor dimensión con una distorsión mínima de las distancias euclídeas. Estas proyecciones permiten preservar relaciones geométricas esenciales, justificando el uso de técnicas de Análisis de Datos en un espacio de dimensión significativamente menor. Este Trabajo Fin de Grado explora las peculiaridades que el tratamiento de datos de alta dimensión introduce, junto con desigualdades de concentración útiles para conseguir probar el Teorema de Johnson-Lindenstrauss. También se analizarán, mediante simulaciones, estas peculiaridades y las implicaciones prácticas de este resultado.

Abstract

In the fields of Data Analysis and Machine Learning, where handling high-dimensional data has become increasingly common, dimensionality reduction emerges as a key technique for enhancing computational efficiency and mitigating the detrimental effects of the well-known "curse of dimensionality." The Johnson–Lindenstrauss theorem is a foundational result which asserts that a dataset in a high-dimensional space can be projected into a lower-dimensional space with only minimal distortion of Euclidean distances. Such projections preserve essential geometric relationships, thereby justifying the application of data analysis techniques in significantly reduced-dimensional settings.

This undergraduate thesis undertakes a theoretical exploration of the challenges posed by high-dimensional data, along with the concentration inequalities that play a crucial role in proving the Johnson–Lindenstrauss theorem. In addition, these phenomena and the practical implications of the theorem will be examined through simulation-based analysis.

Índice general

Capítulo 1

Introducción

Gran parte de los procedimientos clásicos en Análisis Multivariante de Datos (Peña 2002; Izenman 2008) se basan en la hipótesis de normalidad multivariante. Estos procedimientos basados en normalidad son pioneros en Clasificación, tanto supervisada como no supervisada, y forman parte de pilares metodológicos sobre los que se asienta la Inteligencia Artificial, y en particular el Aprendizaje Automático o "Machine Learning" (Hastie, Tibshirani y Friedman 2009; Bishop 2006; James et al. 2014; Shalev-Shwartz y Ben-David 2014). Así, por ejemplo, el Análisis Discriminante Lineal (LDA por Linear Discriminant Analysis) supone que las observaciones a clasificar han sido generadas por distribuciones normales multivariantes con matrices de varianzas-covarianzas iguales. Este tipo de suposiciones permite clasificar nuevas observaciones mediante la estimación de los parámetros asociados a las distribuciones normales multivariantes que modelan estas clases. Si bien estos métodos gozan de buenas propiedades de optimalidad cuando se verifican esas hipótesis de normalidad, se ha comprobado que estos métodos se enfrentan a serias dificultades cuando la dimensión d, por ejemplo, el número de variables discriminantes, crece. En este punto, cabe notar que las situaciones de alta dimensionalidad son más bien la regla que la excepción en la Ciencia de Datos moderna, debido a la omnipresencia de este tipo de datos (registro de señales de todo tipo, monitorización de imágenes y vídeo, datos no estructurados traducidos a formatos tabulares de alta dimensionalidad, ...).

Una de las dificultades que aparece al trabajar con la distribución normal multivariante en altas dimensiones es el gran número de parámetros asociados a todas las covarianzas entre variables a estimar. Esto hace que la distancia de Mahalanobis, como forma de medir "proximidades", naturalmente asociada a la distribución normal multivariante, deje de ser operativa a nivel práctico cuando la dimensión d sea elevada. Suponer igualdad de varianzas y que las correlaciones entre variables son nulas simplifica el uso de la distancia de Mahalanobis al uso de distancias euclídeas. Así, medir proximidades usando la simple distancia euclídea subyace en muchos e importantes métodos básicos en Ciencia de Datos, especialmente en métodos que podemos considerar de carácter "no paramétrico". Un ejemplo bien conocido es el método de los m-vecinos, que asigna observaciones a clases según la clase mayoritaria entre las m observaciones más cercanas a la que se desea clasificar, dentro del espacio euclídeo.

Desgraciadamente, incluso trabajando en el caso simplificado del uso de distancias euclídeas, gran parte de los métodos habituales en Ciencia de Datos resultan afectados por un bien conocido fenómeno de "espacio vacío" que aparece al incrementarse la di-

mensionalidad de los problemas tratados. Este fenómeno de espacio vacío suele conocerse como la "maldición de la dimensionalidad", término introducido por Richard Bellman en Bellman (1958), y que, de forma simplificada, es atribuible al crecimiento exponencial del volumen del espacio en el que se representan los datos cuando se añaden dimensiones. En este trabajo se formulará detalladamente este fenómeno de espacio vacío y se verá cómo el mismo impacta muy negativamente en muchas de las intuiciones de "baja dimensión" que sustentan muchas técnicas de análisis de datos básicas. Por ejemplo, en trabajos como Aggarwal, Hinneburg y Keim (2001) se muestra cómo, para la clasificación mediante m-vecinos más próximos en espacios de muy alta dimensión, sucede que los datos son tan escasos a nivel "local" que la noción de cercanía o vecino más próximo pierde todo su significado. Adicionalmente, las técnicas basadas en suposiciones de normalidad, aun suponiendo estructuras simplificadas de varianzas-covarianzas, dejan de discriminar correctamente entre observaciones debido a un fenómeno de concentración de medida que se tratará también a lo largo de este trabajo.

No obstante, a pesar de estas dificultades, en este trabajo se verá que, bajo ciertas condiciones bastante generales, la alta dimensión puede, en ocasiones, facilitar sorprendentemente algunos análisis de datos. Por ejemplo, en espacios de alta dimensión, los vectores aleatorios asociados al muestreo de distribuciones normales tienden a ser prácticamente ortogonales entre sí, y sus normas euclídeas se concentran notablemente en torno a sus valores esperados, situándose con alta probabilidad alrededor de dichos valores. Este fenómeno de concentración de medida en alta dimensionalidad, debidamente aprovechado, simplifica algunas tareas estadísticas como, por ejemplo, la separación de clases y permite el establecimiento del Teorema de Johnson-Lindenstrauss. Es por ello que en algunos contextos se hace referencia también al concepto de "blessing" (o bendición) de la dimensionalidad, en oposición a la maldición de la dimensionalidad.

Un ejemplo de la aplicabilidad práctica de estos fenómenos de concentración de medida y la ortogonalidad en espacios de alta dimensión es el Teorema de Johnson-Lindenstrauss. Este teorema, formalizado por primera vez en Johnson y Lindenstrauss (1984), establece que cualquier conjunto de n puntos en un espacio $\mathbb{R}^d$ puede proyectarse a un espacio de dimensión mucho menor sin sacrificar demasiada información en términos de distancias euclídeas. Esta proyección, realizada mediante una matriz aleatoria de tamaño $k \times d$, preserva aproximadamente las distancias euclídeas entre todos los pares de puntos del conjunto original. A lo largo de este trabajo se desarrollarán todos los resultados previos necesarios para llegar a probar ese Teorema de Johnson-Lindenstrauss. Este resultado justifica teóricamente el uso de técnicas de reducción de dimensionalidad que no requieren asumir ninguna estructura particular sobre los datos. En particular, resulta de gran ayuda en métodos como la búsqueda de m-vecinos más próximos y otros métodos intrínsecamente basados en distancias euclídeas. En virtud de este teorema, al reducir la dimensionalidad mediante estas proyecciones aleatorias, se logra que las distancias entre las observaciones se conserven dentro de márgenes pequeños, lo cual permite, por ejemplo, buscar el vecino más cercano en un subespacio de dimensión menor. De esta forma, se mitiga parcialmente el problema de alta dimensionalidad en contextos de alta complejidad en los datos, donde otras técnicas típicas de reducción de la dimensionalidad, como por ejemplo las Componentes Principales, resultarían completamente inviables o muy costosas computacionalmente.

La estructura de este trabajo es la siguiente. En el Capítulo 2 se presenta la maldi-

ción de la dimensionalidad, comenzando con simulaciones sobre la distribución normal multivariante (Sección 2.1) y la distribución uniforme en hipercubos (Sección 2.2). En el Capítulo 3 se estudia la distribución uniforme en la bola unidad en alta dimensión. En el Capítulo 4 se analiza el comportamiento de la distribución normal multivariante en espacios de gran dimensión, profundizando en las ideas de concentración introducidas en el capítulo anterior. Finalmente, en la Sección 5.1 del Capítulo 5 se introduce el Teorema de Johnson-Lindenstrauss y en la Sección 5.2 se muestran algunas de sus aplicaciones, incluyendo su uso en algoritmos como los de m vecinos más próximos, m-medias y LDA para clasificación y estimación paramétrica en distribuciones normales.

Los resultados que se presentan a lo largo de este trabajo están basados principalmente en Wegner (2024) a la vez que se han consultado otras fuentes y elaborado argumentos propios. Asimismo, se ha implementado código estadístico que permite justificar empíricamente varios de estos resultados y facilitar su comprensión y alcance.

Capítulo 2

Maldición de la dimensionalidad

El término "maldición de la dimensionalidad" fue acuñado por Richard Bellman en 1961, en el contexto de la programación dinámica Bellman (1958), para describir cómo ciertos problemas se vuelven intratables a medida que aumenta el número de variables involucradas. En particular, Bellman observó que el volumen del espacio crece exponencialmente con la dimensión, lo que implica que la cantidad de datos necesaria para cubrirlo de manera significativa también crece de forma exponencial. Esto no solo dificulta los cálculos, sino que también debilita la capacidad de generalizar patrones o "aprender" de los datos a partir de conjuntos de entrenamiento finitos.

En este capítulo se abordarán algunos fenómenos que emergen en espacios de alta dimensión y que resultan contraintuitivos respecto a la experiencia habitual cuando se trabaja en espacios de dimensión baja o moderada. Veremos cómo cambian las nociones de distancia, cercanía, ortogonalidad y concentración de la medida y cómo estos efectos motivan la necesidad de adaptar nuestras herramientas y modelos cuando nos enfrentamos a datos en espacios de alta dimensión.

En este capítulo se trabajará sobre el espacio $\mathbb{R}^d$ dotado de la σ-álgebra de Borel denotada por β^d, junto con la medida de Lebesgue en $\mathbb{R}^d$ que se denotará por ℓ_d. Igualmente, se denotará por $\mathbf{0}_d = (0, \ldots, 0) \in \mathbb{R}^d$ y por I_d a la matriz identidad de tamaño $d \times d$.

2.1. Distribución normal multivariante: simulaciones y primeros resultados

En Ciencia de Datos y en Análisis Multivariante de Datos, la distribución normal multivariante juega un papel clave. A continuación se define dicha distribución y la forma de su densidad, en los casos no degenerados.

Definición 2.1. *Sea (Ω, σ, P) un espacio probabilístico. Un vector aleatorio X se dice que tiene una distribución Gaussiana o normal d-variante si para todo vector $\boldsymbol{a} \in \mathbb{R}^d$ se tiene que la variable aleatoria real $\boldsymbol{a}^T X : \Omega \to \mathbb{R}$ sigue una distribución normal univariante ($\boldsymbol{a}^T X$ serían todas las posibles combinaciones lineales de las coordenadas del vector aleatorio X). En particular, si asumimos un vector de medias $\boldsymbol{\mu} \in \mathbb{R}^d$ y una matriz de covarianzas Σ, siendo una matriz $d \times d$ y definida positiva, entonces diremos que*

$X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *y se puede ver que*

$$P(X \in A) = \frac{1}{(2\pi)^{1/2}|\Sigma|^{1/2}} \int_A \exp\Big(-\frac{1}{2}(x-\boldsymbol{\mu})^\top \Sigma^{-1}(x-\boldsymbol{\mu})\Big) dx, \text{ para } A \in \beta^d.$$

A la hora de analizar cómo afecta la dimensionalidad al comportamiento de la normal multivariante, comenzaremos con un pequeño estudio de algunas propiedades básicas que ayudan a comprender su comportamiento al aumentar la dimensión d. Para ello, el siguiente teorema revisa el valor esperado de la norma de X y la varianza de la misma, si $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$.

Proposición 2.2. *Si* $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$ *entonces, para todo* $d \in \mathbb{N}$*, se verifica:*

(i) $E(\|X\|^2) = d$ *y* $Var(\|X\|^2) = 2d$

(ii) $\Big|E(\|X\| - \sqrt{d})\Big| \leq \dfrac{1}{\sqrt{d}}$

(iii) $\mathrm{Var}(\|X\|) \leq 2$

Demostración. Para probar (i) se usa la expresión de la norma euclídea de $X = (X_1, \ldots, X_d)^\top$ como $\|X\|^2 = X_1^2 + \cdots + X_d^2$, que junto a la linealidad de la esperanza, lleva a

$$E(\|X\|^2) = E(X_1^2 + \cdots + X_d^2) = E(X_1^2) + \cdots + E(X_d^2).$$

Si $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$ entonces $X_l \sim \mathcal{N}(0,1)$ para $l = 1, \ldots, d$ (véase la Proposición A.8) y, por tanto

$$E(\|X\|^2) = E[(X_1 - E(X_1))^2] + \cdots + E[(X_d - E(X_d))^2] = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_d) = d. \tag{2.1}$$

Por otro lado, la independencia de las variables X_l, $l = 1, ..., d$ (recuérdese que $\Sigma = I_d$ y que la incorrelación es equivalente a la independencia para distribuciones normales), nos garantiza que

$$\mathrm{Var}(\|X\|^2) = \mathrm{Var}(X_1^2) + \cdots + \mathrm{Var}(X_d^2) = d\,\mathrm{Var}(X_1^2) = d\left[E(X_1^4) - E^2(X_1^2)\right].$$

Siguiendo el argumento en (2.1) sabemos que $E(X_1^2) = 1$. Por otro lado, el momento de orden cuatro de una distribución normal estándar es $E(X_1^4) = 3$, ya que se sabe que la kurtosis de cualquier variable aleatoria distribuida según una normal univariante es igual a 3 y el simple uso de

$$\mathrm{Kurtosis}(X_1) = 3 = \frac{E(X_1^4) - E^4(X_1)}{\mathrm{Var}(X_1)^2} = \frac{E(X_1^4) - 0}{1}.$$

En el Apéndice A.1 se demuestra que la Kurtosis de cualquier distribución normal univariante es 3. Por tanto, (i) queda probado dado que

$$\mathrm{Var}(\|X\|^2) = d\left[E(X_1^4) - E^2(X_1^2)\right] = d \cdot (3-1) = 2d. \tag{2.2}$$

Para probar (ii) debemos tener en cuenta lo probado en (i) y el siguiente desarrollo algebraico:

$$\|X\| - \sqrt{d} = \frac{2\|X\|\sqrt{d} - 2d}{2\sqrt{d}} = \frac{\|X\|^2 - \|X\|^2 + 2\|X\|\sqrt{d} - 2d}{2\sqrt{d}} =$$
$$= \frac{\|X\|^2 - d}{2\sqrt{d}} - \frac{\|X\|^2 + 2\|X\|\sqrt{d} - d}{2\sqrt{d}} = \frac{\|X\|^2 - d}{2\sqrt{d}} - \frac{(\|X\| - \sqrt{d})^2}{2\sqrt{d}} =$$
$$= \frac{\|X\|^2 - d}{2\sqrt{d}} - \frac{(\|X\| - \sqrt{d})^2}{2\sqrt{d}} \cdot \frac{(\|X\| + \sqrt{d})^2}{(\|X\| + \sqrt{d})^2} = \frac{\|X\|^2 - d}{2\sqrt{d}} - \frac{(\|X\|^2 - d)^2}{2\sqrt{d} \cdot (\|X\| + \sqrt{d})^2}$$

Si renombramos los términos como

$$A_d = \frac{\|X\|^2 - d}{2\sqrt{d}} \text{ y } B_d = \frac{(\|X\|^2 - d)^2}{2\sqrt{d} \cdot (\|X\| + \sqrt{d})^2},$$

entonces

$$|E(\|X\| - \sqrt{d})| = |E(A_d) - E(B_d)|.$$

Por lo probado en (2.1) se tiene que $E(A_d) = 0$, para todo $d \in \mathbb{N}$, y como $B_d \geq 0$ para todo $d \in \mathbb{N}$ entonces

$$0 \leq \left|E(\|X\| - \sqrt{d})\right| = E(B_d) \leq \frac{E\left[(\|X\|^2 - d)^2\right]}{2d^{3/2}} = \frac{\mathrm{Var}(\|X\|^2)}{2d^{3/2}} = \frac{1}{\sqrt{d}},$$

donde se ha usado (2.2).

Finalmente, para probar (iii) se hace uso de (ii) y de las propiedades de la varianza, que dicen que $\mathrm{Var}(X + c) = \mathrm{Var}(X)$ para cualquier constante $c \in \mathbb{R}$. Por tanto,

$$\mathrm{Var}(\|X\|) = \mathrm{Var}(\|X\| - \sqrt{d}) \leq E\left[(\|X\| - \sqrt{d})^2\right] = 2\sqrt{d} \cdot E(B_d) \leq 2,$$

donde se ha usado (2.2) y parte de la prueba de (ii). □

Nota 2.3. *Por lo demostrado en el apartado (ii) del teorema anterior, lo que podemos deducir es que para d suficientemente grande, $E(\|X\| - \sqrt{d})$ es prácticamente 0, luego se deduce que $E(\|X\|) \approx \sqrt{d}$. Además, como la varianza de la norma de X al cuadrado está acotada, entonces podemos aplicar la desigualdad de Chebyshev para asegurar que la distribución de $\|X\|$ se concentra alrededor de su media. Es decir, dado $k > 0$ se tiene que*

$$P(|\|X\| - E(\|X\|)| \geq k) \leq \frac{\mathrm{Var}(\|X\|)}{k^2} \leq \frac{2}{k^2}.$$

Esto nos asegura que la probabilidad de que $\|X\|$ se aleje de su media en k unidades está acotada por $\frac{2}{k^2}$.

Por la Proposición 2.2 se sabe que $|E(\|X\| - \sqrt{d})| \leq 1/\sqrt{d}$ y se desea proporcionar una cota para la probabilidad de que $\|X\|$ se aleje de $\sqrt{d}$ en más de $\varepsilon > 0$. Para ello, se puede escribir

$$\begin{aligned} P(|\|X\| - \sqrt{d}| \geq \varepsilon) &= P(|\|X\| - E(\|X\|) + E(\|X\|) - \sqrt{d}| \geq \varepsilon) \ \leq \\ &\leq P(|\|X\| - E(\|X\|)| + |E(\|X\|) - \sqrt{d}| \geq \varepsilon) \leq \\ &\leq P(|\|X\| - E(\|X\|)| \geq \varepsilon - |E(\|X\|) - \sqrt{d}|) \leq \\ &\leq P\left(|\|X\| - E(\|X\|)| \geq \varepsilon - \frac{1}{\sqrt{d}}\right), \end{aligned}$$

y para $\varepsilon > 1/\sqrt{d}$ *se tiene*

$$P\left(\left|\|X\| - \sqrt{d}\right| \geq \varepsilon\right) \leq \frac{2}{\left(\varepsilon - \frac{1}{\sqrt{d}}\right)^2}$$

A medida que d crece, la cota anterior para la probabilidad se acerca cada vez más a $2/\varepsilon^2$. *Luego para* ε *fijo, la probabilidad de que* $\|X\|$ *se aleje de* $\sqrt{d}$ *se mantiene acotada y no crece. Por tanto, para cualquier* X *es de esperar que los valores de* $\|X\|$ *estén cerca de* $\sqrt{d}$ *y esta aproximación mejora conforme se incrementa* d.

Para comprobar experimentalmente los resultados del Teorema 2.2 y las consecuencias de la Nota 2.3, podemos simular mediante `R` una muestra aleatoria simple de la distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ para diferentes valores de d y ver si se cumplen los resultados teóricos que se han demostrado.

Así, se han calculado resúmenes estadísticos de la norma euclídea de n realizaciones aleatorias de la distribución normal d-variante para distintos valores de la dimensión d. En concreto, se ha generado una muestra aleatoria simple de tamaño $n = 8000$ del vector aleatorio $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$, es decir, se han considerado $X_1, \ldots, X_{8000}$ variables aleatorias independientes con $X_i = \boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}_d, I_d)$ para $i = 1, \ldots, 8000$ y se han calculado las normas euclídeas $\|\boldsymbol{x}_1\|, \ldots, \|\boldsymbol{x}_{8000}\|$ de dichas observaciones. Los gráficos de la Figura 2.1 muestran cómo la mayoría de realizaciones de la distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ se sitúan muy próximas a la esfera en $\mathbb{R}^d$ centrada en el origen de radio $\sqrt{d}$ y, a su vez, en la Tabla 2.1 se ve que la dispersión de las normas euclídeas no parece aumentar a medida que se incrementa la dimensión d. Variando la dimensión d en este experimento, la hipótesis que confirmamos es que $E(\|X\|) \approx \sqrt{d}$ y que $\mathrm{Var}(\|X\|)$ está acotada para todo d, si $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$. En la Tabla 2.1 se han recogido las medias y varianzas muestrales de las normas de los individuos de la muestra, así como el valor de $\sqrt{d}$, para cada una de las dimensiones d. Es decir, se han calculado los valores:

$$\overline{\|\boldsymbol{x}_{8000}\|} = \frac{1}{8000}\sum_{i=1}^{8000}\|\boldsymbol{x}_i\| \quad \text{y} \quad S^2_{8000} = \frac{1}{8000}\sum_{i=1}^{8000}\left(\|\boldsymbol{x}_i\| - \overline{\|\boldsymbol{x}_{8000}\|}\right)^2,$$

donde $\overline{\|\boldsymbol{x}_{8000}\|}$ denota la media muestral y S^2_{8000} la varianza muestral de las normas de la muestra aleatoria simple.

	d=1	d=10	d=100	d=500	d=1000	d=10000
$\overline{\|\boldsymbol{x}_{8000}\|}$	0.8024	3.0953	9.9717	22.3490	31.6176	100.0008
$\sqrt{d}$	1.0000	3.1623	10.0000	22.3607	31.6228	100.0000
S^2_{8000}	0.3524	0.4807	0.4889	0.4795	0.4956	0.5004

Tabla 2.1: Resúmenes estadísticos de las normas de una muestra de tamaño $n = 8000$ de normales multivariantes $X_i \sim \mathcal{N}(\mathbf{0}_d, I_d)$ dependiendo de la dimensión d.

Estos resultados ilustran el comportamiento de la normal multivariante al aumentar la dimensión d, observando comportamientos no necesariamente demasiado "intuitivos". Por ejemplo, se ha probado que la distribución normal multivariante está concentrada en torno a la esfera de radio $\sqrt{d}$, mientras que la distribución normal univariante estaba

concentrada en un entorno del cero de amplitud 1 con una probabilidad de 0.6826. No obstante, sí se observa que la varianza de las normas euclídeas se mantiene acotada, como en el caso univariante, independiente de la dimensión d.

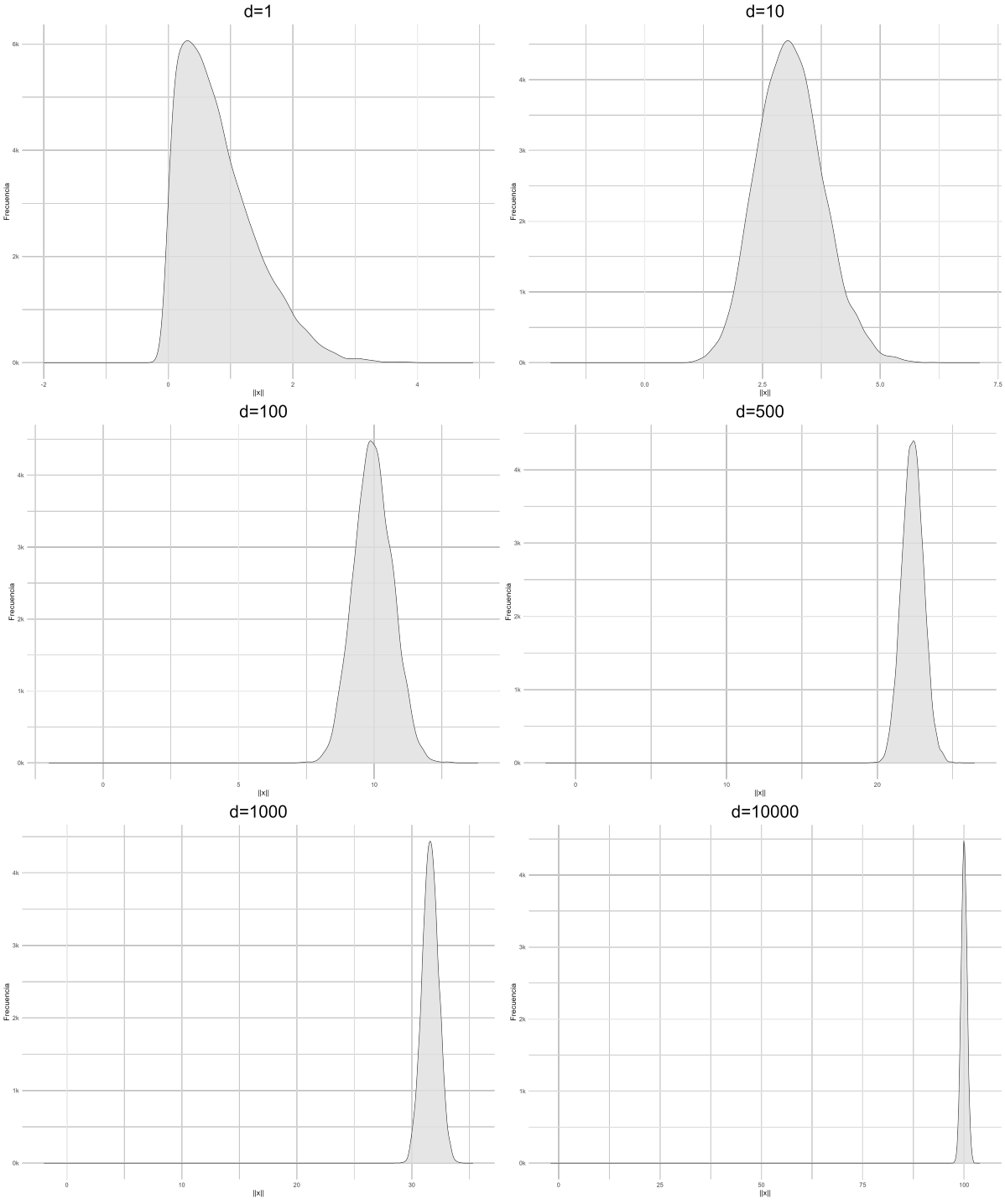

Figura 2.1: Estimación de la función de densidad de $\|X\|$ si X es una distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ para $d = 1, 10, 100, 500, 1.000, 10.000$.

Siguiendo el mismo tipo de argumentos, tiene sentido establecer relaciones entre la

posición relativa de dos variables aleatorias independientes X e Y tales que $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$ e $Y \sim \mathcal{N}(\mathbf{0}_d, I_d)$ mediante $\|X - Y\|$. El resultado análogo a la proposición anterior es el siguiente:

Proposición 2.4. *Si $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$ e $Y \sim \mathcal{N}(\mathbf{0}_d, I_d)$ son vectores aleatorios independientes, entonces se verifica:*

(i) $E(\|X - Y\|^2) = 2d$ *y* $\mathrm{Var}(\|X - Y\|^2) = 8d$

(ii) $|E(\|X - Y\| - \sqrt{2d})| \leq \dfrac{2}{\sqrt{2d}}$

(iii) $\mathrm{Var}(\|X - Y\|) \leq 4$

Demostración. La demostración de (ii) y (ii) se realiza tal como se hizo en la prueba de la Proposición 2.2. En cuanto a (i), se puede presentar una demostración alternativa teniendo en cuenta que la distribución χ^2_k con $k \in \mathbb{N}$ grados de libertad se define como la suma del cuadrado de k variables con distribución $X_l \sim \mathcal{N}_1(0, 1)$ independientes. Es decir, $Z = X_1^2 + \cdots + X_k^2$ sigue una distribución chi-cuadrado, que denotamos como $Z \sim \chi^2_k$. Es conocido que el valor esperado de una distribución $Z \sim \chi^2_k$ satisface $E(Z) = k$ y que $\mathrm{Var}(Z) = 2k$. Teniendo en cuenta que $\|X - Y\|^2 = (X_1 - Y_1)^2 + \cdots + (X_k - Y_k)^2$ y que $X_l - Y_l \sim \mathcal{N}(0 - 0, \sqrt{1^2 + 1^2})$ (recordando las condiciones de independencia asumidas), se llega a

$$\frac{X_l - Y_l}{\sqrt{2}} \sim \mathcal{N}_1(0, 1), \text{ para } l = 1, \ldots, d,$$

por lo que

$$\frac{\|X - Y\|^2}{2} \sim \chi^2_d,$$

y, consecuentemente, $E(\|X - Y\|^2) = 2d$ y $\mathrm{Var}(\|X - Y\|^2) = 8d$, como se deseaba probar. □

Nota 2.5. *Siguiendo lo expuesto en la Nota 2.3, como $E(\|X - Y\|^2) - \sqrt{2d}$ se acerca a 0 para d grande, se tiene que $E(\|X - Y\|^2) \approx \sqrt{2d}$. Además, como la varianza está acotada, se puede probar que $\|X - Y\|$ se concentra alrededor de su media y acotar la probabilidad de que $\|X - Y\|$ se aleje de $\sqrt{2d}$ en más de ε por*

$$P\left(\left|\|X - Y\| - \sqrt{2d}\right| \geq \varepsilon\right) \leq \frac{4}{\left(\varepsilon - \dfrac{2}{\sqrt{2d}}\right)^2}.$$

Juntando las proposiciones 2.2 y 2.4, se garantiza que $E(\|X\|^2) = d$ y $E(\|X - Y\|^2) = 2d$, luego cabe preguntarse si podemos esperar que X e Y sean vectores aleatorios ortogonales (independientes en el caso de normalidad) en virtud del argumento geométrico que proporciona el Teorema de Pitágoras en la Figura 2.2.

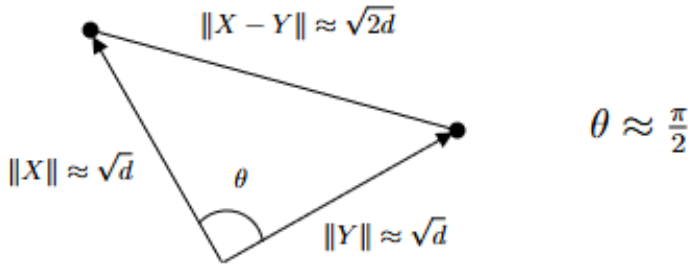

Figura 2.2: Interpretación geométrica usando el Teorema de Pitágoras

Trabajando en la dirección del último comentario, se puede probar lo siguiente:

Proposición 2.6. *Si* $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$ *e* $Y \sim \mathcal{N}(\mathbf{0}_d, I_d)$ *son independientes y* $\mathbf{x} \in \mathbb{R}^d$ *entonces se cumple:*

(i) $E\big(\langle X, Y\rangle\big) = 0$ *y* $\mathrm{Var}\big(\langle X, Y\rangle\big) = d$.

(ii) $E\big(\langle X, \mathbf{x}\rangle\big) = 0$ *y* $\mathrm{Var}\big(\langle X, \mathbf{x}\rangle\big) = \|\mathbf{x}\|^2$.

Demostración. Para la demostración de (i), se usa la definición habitual del producto escalar en $\mathbb{R}^d$ como

$$\langle X, Y\rangle = X_1Y_1 + \cdots + X_dY_d,$$

se tiene que

$$E\big(\langle X, Y\rangle\big) = \sum_{l=1}^{d} E(X_lY_l) = \sum_{l=1}^{d} E(X_l)E(Y_l) = 0,$$

(por la independencia de X_l e Y_l). Aprovechando igualmente las propiedades de independencia tenemos para la varianza que,

$$\begin{aligned}\mathrm{Var}\big(\langle X, Y\rangle\big) &= \sum_{l=1}^{d} \mathrm{Var}(X_lY_l) = d\,\mathrm{Var}(X_1Y_1) = d\left(E(X_1^2)E(Y_1^2) - E^2(X_1)E^2(Y_1)\right)\\ &= d\,E(X_1^2 - E(X_1))\,E(Y_1^2 - E(Y_1)) = d\,\mathrm{Var}(X_1)\mathrm{Var}(Y_1) = d\end{aligned}$$

Para probar (ii) podemos repetir el mismo razonamiento y demostrar fácilmente que $E\big(\langle X, \mathbf{x}\rangle\big) = 0$ y, para la varianza, basta recurrir de nuevo a la independencia obteniendo que

$$\mathrm{Var}\big(\langle X, \mathbf{x}\rangle\big) = \sum_{l=1}^{d} \mathrm{Var}(X_l\,x_l) = \sum_{l=1}^{d} x_l^2\mathrm{Var}(X_l) = \sum_{l=1}^{d} x_l^2 = \|\mathbf{x}\|^2.$$

□

Nota 2.7. *Aunque se ha probado que* $E\big(\langle X, Y\rangle\big) = 0$*, a la hora de comprobar hasta qué punto* X *e* Y *son ortogonales para* d *suficientemente grande, tenemos problemas por el*

hecho de que la varianza $\mathrm{Var}\big(\langle X, Y\rangle\big)$ *no esté acotada, lo que nos hubiera permitido usar directamente la desigualdad de Chebyshev, como en las notas 2.3 y 2.5.*

Sin embargo, si consideramos un punto $\mathbf{x} \in \mathbb{R}^d$ *fijo, se ha obtenido en (ii) que* $\mathrm{Var}\big(\langle X, \mathbf{x}\rangle\big) = \|\mathbf{x}\|^2$*, que sí es finita. Por tanto, usando el argumento de la desigualdad de Chebyshev, se puede probar que* $\langle X, \mathbf{x}\rangle$ *se concentra alrededor de su media* $E\big(\langle X, \mathbf{x}\rangle\big) = 0$ *y, por tanto, cualquier observación es de esperar que sea ortogonal al punto prefijado* $\mathbf{x} \in \mathbb{R}^d$*. Esto nos sugiere que, como las observaciones de* $\|X\|$ *están sobre una esfera de radio* $\sqrt{d}$ *centrada en* $\mathbf{0}_d$*, si elegimos* $\mathbf{x} = (\sqrt{d}, 0, \ldots, 0) \in \mathbb{R}^d$ *entonces que las coordenadas del vector aleatorio* X *estén situadas con alta probabilidad sobre el ecuador de la bola de radio* $\sqrt{d}$*. Es decir, la coordenada* X_1 *de* X *es pequeña (y este argumento es simétrico para cualquier* X_l *con* $l = 1, \ldots, d$*). Esta intuición será formalizada en el Capítulo 4.*

2.2. Distribución uniforme en hipercubos

Definición 2.8. *Sea* (Ω, σ, P) *un espacio probabilístico. Decimos que* X *sigue una distribución uniforme en* $B \in \beta^d$ *si*

$$P(X \in A) = \frac{\ell_d(A \cap B)}{\ell_d(B)},$$

donde $\ell_d(C) = \int_C dx$ *denota la medida de Lebesgue del conjunto* C *para* $C \in \beta^d$*. Una variable con esta distribución será denotada por* $X \sim \mathcal{U}(B)$*.*

Nótese que es importante que tanto el conjunto B sobre el que se define la distribución como los sucesos A que se utilizan sean medibles.

En lo que sigue, se estudiará la distribución uniforme sobre algunos conjuntos especialmente interesantes, como son los hipercubos, y se verá que, al igual que con la distribución normal multivariante, se tienen algunas propiedades interesantes, y no necesariamente demasiado intuitivas, al incrementar la dimensión d.

Definición 2.9. *Se define el hipercubo* d*-dimensional centrado en* $\mathbf{0}_d$ *como el conjunto* $H_d = [-1, 1]^d \subset \mathbb{R}^d$ *y la bola unidad centrada en* $\mathbf{0}_d$ *con radio unitario será denotada como* $B(\mathbf{0}_d, 1) \subset \mathbb{R}^d$*. Estos conjuntos son claramente medibles para la* σ*-álgebra de Borel* β^d*.*

Hay varias propiedades interesantes sobre los hipercubos que, en cierta medida, resultan sorprendentes. Por ejemplo, si se consideran los 2^d vértices de la forma $(v_1, ..., v_d)$, con v_l tomando valores 1 y -1 para $l = 1, \ldots, d$, y se calcula la norma de cualquiera de estos vértices se observa que todos tienen norma $\sqrt{d}$ con $\sqrt{d} \to \infty$. Es decir, cuanto más incremente la dimensión d, más alejados del origen $\mathbf{0}_d$ se encuentran los vértices del hipercubo. Sin embargo, este hecho no es suficiente como para garantizar que la esperanza de la norma de cualquier $X \sim \mathcal{U}([-1, 1]^d)$ crezca con la dimensión d, como ocurría en el caso de la distribución normal multivariante. De hecho, si se consideran los centros de las caras del hipercubo, es decir, los puntos $\mathbf{e}_l = (0, \ldots, 0, 1, 0 \ldots, 0)$ y $-\mathbf{e}_l$ (un 1 o -1 en la posición i y 0 en el resto de las posiciones) se tiene $\|\mathbf{e}_l - (-\mathbf{e}_l)\| = 2$, mientras que la distancia de $\mathbf{e}_l$ a $\mathbf{e}_{l'}$ (o a $-\mathbf{e}_{l'}$) es igual a $\sqrt{2}$ si $l \neq l'$, luego algunas distancias se mantienen invariantes con la dimensión. Esto quiere decir que, conforme aumenta la dimensión d, se evidencia la presencia de picos sobre la superficie del hipercubo, mientras que los centros de las caras

se mantienen a distancia constante. En la Figura 2.3 se pueden ver estas propiedades en dimensión 2 y 3, donde se ha inscrito la bola de radio 1 en el hipercubo correspondiente, mientras que la última representación corresponde a una versión idealizada del hipercubo en dimensión d alta denominada "Equidnaedro", que puede resultar de ayuda para intuir sus características.

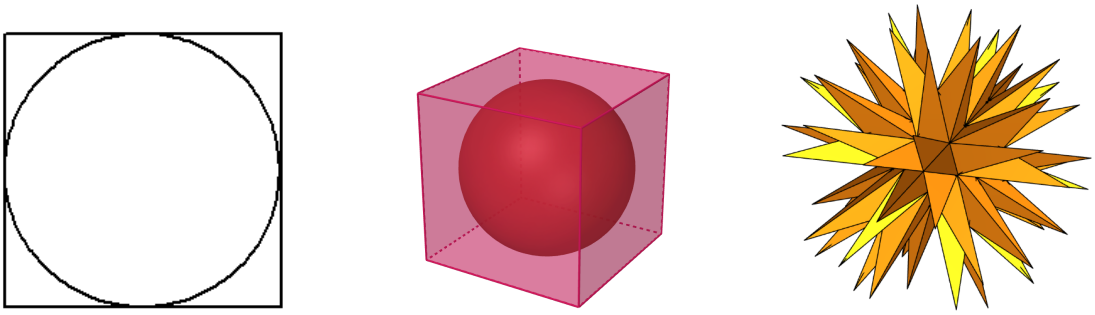

Figura 2.3: Representación de un hipercubo en 2 y 3 dimensiones y de un "Equidnaedro" como una versión idealizada del hipercubo para un d elevado

Puesto que estas distancias parecen no comportarse de manera intuitiva en alta dimensión, se puede esperar que el volumen de H_d, y por tanto, las zonas de alta probabilidad para la distribución uniforme, tampoco se comporten de una forma demasiado intuitiva. De hecho, según se incrementa la dimensión, se puede probar que hay una alta concentración del volumen del hipercubo (masa de probabilidad de la distribución uniforme asociada) cerca de sus caras. Denotando

$$F_{\varepsilon,d} = H_d \setminus (1-\varepsilon)H_d = \{\mathbf{x} \in H_d\} \setminus \{(1-\varepsilon)\cdot \mathbf{x} : \mathbf{x} \in H_d\},$$

se tiene

$$\ell_d(F_{\varepsilon,d}) = \ell_d(H_d) - \ell_d((1-\varepsilon)H_d) = 2^d - (2(1-\varepsilon))^d.$$

Se probará un resultado más general que este en el Lema 3.2. En términos de proporciones, se tendría que

$$\frac{\ell_d(F_{\varepsilon,d})}{\ell_d(H_d)} = 1 - (1-\varepsilon)^d \xrightarrow{d\to\infty} 1,$$

para cualquier $0 < \varepsilon < 1$. Lo que se acaba de probar es que, al incrementar la dimensión, se tiene la mayor concentración de masa de probabilidad en un entorno de la frontera del hipercubo. Este hecho contrasta con la baja probabilidad que se proporciona a una bola $B(\mathbf{0}_d, 1)$ inscrita en el hipercubo, para la que se verifica

$$\ell_d(B(\mathbf{0}_d, 1)) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}. \tag{2.3}$$

Este resultado será probado en el Apéndice B. Usando órdenes de infinitud se obtiene fácilmente que

$$\lim_{d\to\infty} \ell_d(B(\mathbf{0}_d, 1)) = 0.$$

Para evitar el hecho de que el volumen del hipercubo $[-1,1]^d$ no esté acotado, es decir, que suceda que

$$\lim_{d\to\infty} \ell_d([-1,1]^d) = \lim_{d\to\infty} 2^d = \infty,$$

se trabaja con el hipercubo $[-1/2, 1/2]^d$, de tal forma que si $X \sim U([-1/2, 1/2]^d)$ y $C \subset [-1/2, 1/2]^d$ entonces

$$P(X \in C) = \ell_d(C),$$

y, consecuentemente,

$$\lim_{d\to\infty} P\left(\|X\| \leq \frac{1}{2}\right) = \lim_{d\to\infty} \ell_d(B(\mathbf{0}_d, 1/2)) = \lim_{d\to\infty} \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \left(\frac{1}{2}\right)^d = 0. \tag{2.4}$$

La Figura 2.3 hace una representación esquemática del aspecto del hipercubo de lado unidad y la esfera de radio 1/2 inscrita en él. También, se ha calculado mediante `R` el volumen de la d-hiperesfera $B(\mathbf{0}_d, \frac{1}{2})$ al cambiar la dimensión d y se han recogido estos datos en la Tabla 2.2. Por otro lado, si se considera

$$\lim_{d\to\infty} P(X \notin [-1/2+\varepsilon, 1/2-\varepsilon]^d) = \lim_{d\to\infty} 1 - (1-2\varepsilon)^d = 1,$$

concentrándose, de nuevo, la probabilidad en los bordes del hipercubo y no concentrándose para nada en las proximidades del $\mathbf{0}_d$. Estos resultados muestran el tipo de comportamiento contraintuitivo que ya veíamos para la distribución normal multivariante. Nótese que se entiende por "proximidades" a estar a una distancia menor a 1/2, lo que se traduce en el problema de "espacio vacío" descrito en el Capítulo 1 de Introducción.

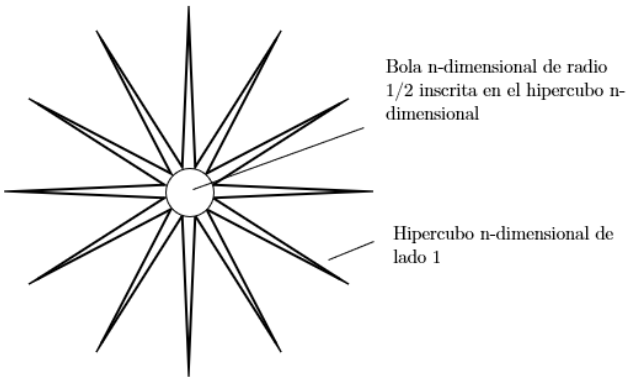

Figura 2.4: Sección de una bola de radio 1/2 inscrita en un hipercubo de lado unidad en dimensión d.

En el caso genérico de un hipercubo centrado en el $\mathbf{0}_d$ de lado L, su volumen será L^d, que coincide con el valor 1 si $L = 1$, tiende a infinito si $L > 1$ o decrece hacia 0 si $L < 1$. No obstante, los razonamientos anteriores son también válidos porque podemos escribir

$$\frac{\ell_d(B(\mathbf{0}_d, \frac{L}{2}))}{\ell_d([-\frac{L}{2}, \frac{L}{2}]^d)} = \frac{L^d \cdot \ell_d(B(\mathbf{0}_d, \frac{1}{2}))}{L^d \cdot \ell_d([-\frac{1}{2}, \frac{1}{2}]^d)} = \ell_d(B(\mathbf{0}_d, \tfrac{1}{2})). \tag{2.5}$$

Luego, según se incrementa la dimensión, la probabilidad de que las observaciones de una distribución uniforme en el hipercubo se encuentren en un entorno del origen tiende a cero.

Dimensión	Volumen
1	1.000000
2	0.785400
3	0.523600
4	0.308430
5	0.164490
6	0.080750
7	0.036910
8	0.015850
9	0.006440
10	0.002490
11	0.000920
12	0.000330
13	0.000110
14	0.000040
15	0.000010

Tabla 2.2: Volumen de la esfera de radio 1/2 al cambiar d.

Estas ideas se pueden formalizar a resultados asintóticos. En los trabajos Anderssen y Bloomfield (1975) y Anderssen, Brent et al. (1976) se demuestra que si $X \sim \mathcal{U}([0,1]^d)$, $E(\|X\|)$ se comporta asintóticamente como $\sqrt{d/3}$, aunque su expresión exacta solo se conoce para pequeños valores de d. Este comportamiento se asemeja al de la distribución normal multivariante (recuérdese el Teorema 2.2). Siguiendo el razonamiento anterior, para el hipercubo $[-1/2, 1/2]^d$, se demuestra en Bailey, Borwein y Crandall (2007) que $E(\|Y\|) = 1/2\, E(\|X\|)$ si $Y \sim \mathcal{U}([-1/2, 1/2]^d)$ y, entonces, $E(\|Y\|)$ se comporta asintóticamente como $1/2\,\sqrt{d/3}$.

Para justificar empíricamente estos comentarios se ha simulado en `R` $n = 8000$ realizaciones aleatorias de la distribución uniforme en $[-1/2, 1/2]^d$ para distintos valores de d. Los gráficos de la Figura 2.5 muestran cómo las observaciones de $\|Y\|$ se encuentran, aproximadamente, a distancia $1/2 \cdot \sqrt{d/3}$ del origen y, a su vez, la dispersión de las mismas no aumenta conforme se incrementa la dimensión. La Tabla 2.3 recoge la media y varianza muestral de las observaciones así como el valor esperado para $E(\|Y\|)$ y la diferencia entre este y la media muestral.

	d=1	d=10	d=100	d=500	d=1000	d=10000
Media muestral	0.2493	0.9018	2.8829	6.4536	9.1288	28.8669
$1/2 \cdot \sqrt{d/3}$	0.2887	0.9129	2.8868	6.4550	9.1287	28.8675
Varianza muestral	0.0208	0.0173	0.0166	0.0166	0.0165	0.0169

Tabla 2.3: Media y varianza muestral de la distribución uniforme en hipercubos para distintos valores de d

Además, en los artículos Bailey, Borwein y Crandall (2007) y Robbins y Bolis (1978), se generaliza el resultado anterior para $\|X - Y\|$ si $X, Y \sim \mathcal{U}([-1/2, 1/2]^d)$, probando que $E(\|X - Y\|)$ está acotada entre $1/2 \cdot \sqrt{d}\,/\,3$ y $1/2 \cdot \sqrt{d/6}$. Esto demuestra que $E(\|X - Y\|)$ tiende a infinito cuando $\sqrt{d}\,/\,3 \to \infty$, extendiéndose el carácter contraintuitivo que

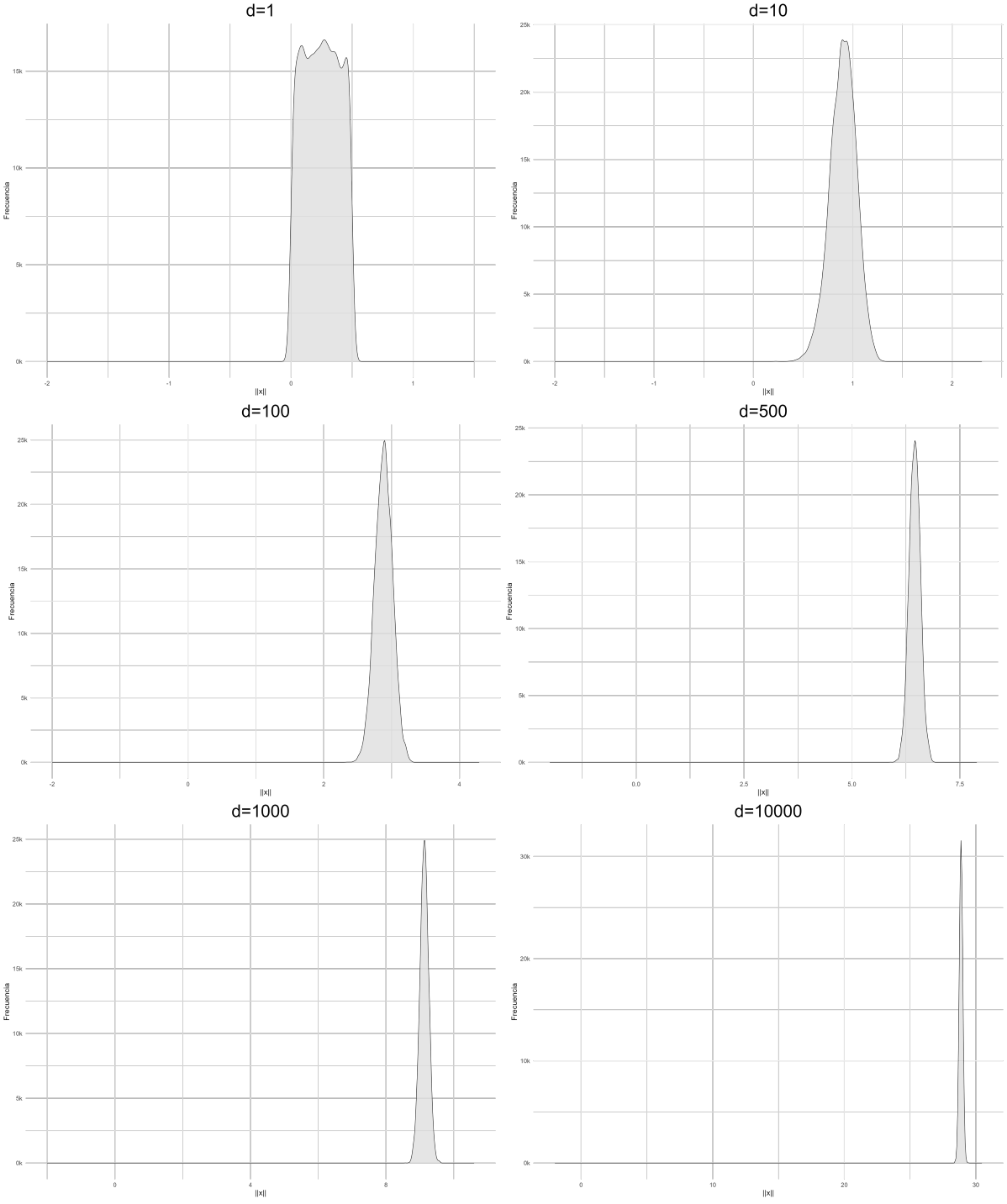

Figura 2.5: Aproximación a la función de densidad de $\|X\|$ si X es una distribución $\mathcal{U}([-1/2, 1/2]^d)$ para $d = 1, 10, 100, 500, 1.000, 10.000$

ya se demostró para la distancia entre dos variables aleatorias independientes $\mathcal{N}(\mathbf{0}_d, I_d)$ en el Teorema 2.4). Por último, las propiedades de independencia de X e Y y que sus coordenadas X_l, Y_l tengan media 0 permiten demostrar que $E(\langle X, Y\rangle) = 0$ para $X, Y \sim \mathcal{U}([-1/2, 1/2]^d)$. De esta manera, se ve que el comportamiento de la distribución uniforme en hipercubos reproduce de alguna forma lo visto para la distribución normal para altas dimensiones.

En esta sección hemos probado teóricamente y comprobado experimentalmente cómo dos de las distribuciones más importantes en Estadística se ven afectadas al aumentar considerablemente la dimensión d. Los resultados obtenidos evidencian un comportamiento inusual en alta dimensión, donde las regiones de alta probabilidad cambian, las distancias aumentan y los vectores tienden a ser ortogonales. Todo esto lleva al bien conocido fenómeno de "espacio vacío", que subyace en la "maldición de la dimensionalidad", y hace que conceptos como utilizar "observaciones más próximas", en problemas de aprendizaje supervisado y no supervisado sea, al menos, cuestionable.

Capítulo 3

Distribución uniforme en la bola unidad en alta dimensión

Como hemos demostrado en el Capítulo 2, a medida que se incrementa la dimensión del espacio, ciertos fenómenos geométricos y probabilísticos empiezan a manifestarse de manera sistemática, como se ha visto en los Teoremas 2.2, 2.4 y 2.6. Uno de los fenómenos que resultaará más relevante en el contexto del análisis de datos de alta dimensión será el fenómeno de concentración de la medida. Este fenómeno describe cómo, en espacios métricos de alta dimensión, la mayoría de las medidas de probabilidad se agrupan en torno a ciertas regiones específicas del espacio. En este capítulo introduciremos formalmente este fenómeno a través del estudio de la distribución uniforme en la bola unidad de $\mathbb{R}^d$, que constituye una de las distribuciones más sencillas e importantes para observar esa concentración de medida.

Previamente a presentar los resultados más importantes en este capítulo, empezaremos presentando un lema técnico que extiende uno de los resultados vistos para distribuciones uniformes en hipercubos.

Definición 3.1. *Se dice que un conjunto $B \subseteq \mathbb{R}^d$ es equilibrado si para todo $\mathbf{z} \in B$ se tiene que $\mu\mathbf{z} \in B$ cuando $|\mu| < 1$. Es decir, $\mu B \subseteq B$ donde $\mu B = \{\mu\mathbf{z} : \mathbf{z} \in B\}$ y $|\mu| < 1$. Por ejemplo, las bolas centradas en el origen o los hipercubos serían conjuntos equilibrados.*

Adicionalmente, dado un conjunto $B \subset \mathbb{R}^d$ medible y equilibrado, llamamos la ε-cáscara de B al conjunto

$$B^\varepsilon = B \setminus (1-\varepsilon)B.$$

De tal forma que B^ε es la parte de B que dista menos que ε de su frontera.

Lema 3.2 (Lema de concentración superficial de medida)**.** *Dado un conjunto $E \subset \mathbb{R}^d$ medible y equilibrado y $\varepsilon > 0$ fijo, se tiene*

$$\frac{\ell_d(E^\varepsilon)}{\ell_d(E)} \to 1, \text{ cuando } d \to \infty.$$

Demostración. Dado que E es equilibrado, podemos considerar el conjunto $(1-\varepsilon)E = \{(1-\varepsilon)\cdot\mathbf{x} : \mathbf{x} \in E\}$, contenido en E y también medible, por serlo E, para todo $\varepsilon > 0$.

Aplicando el cambio de variable dado por el difeomorfismo $\phi : (1-\varepsilon)E \to E$ mediante $\phi(y_i) = y_i/(1-\varepsilon) = x_i$ cuyo jacobiano es $|\det(J_\phi)| = 1/(1-\varepsilon)^d$, se cumple que

$$\ell_d\big((1-\varepsilon)E\big) = \int_{(1-\varepsilon)E} dy_1 \cdots dy_d = \int_E (1-\varepsilon)^d \, dx_1 \cdots dx_d = (1-\varepsilon)^d \ell_d(E),$$

y, consecuentemente,

$$\frac{\ell_d(E^\varepsilon)}{\ell_d(E)} = \frac{\ell_d(E) - \ell_d\big((1-\varepsilon)E\big)}{\ell_d(E)} = 1 - (1-\varepsilon)^d.$$

Si se usa la desigualdad $1 - x \leq e^{-x}$, válida para todo $x \in \mathbb{R}$, se llega a que

$$\frac{\ell_d(E^\varepsilon)}{\ell_d(E)} = 1 - (1-\varepsilon)^d \geq 1 - e^{-\varepsilon d} \xrightarrow{d\to\infty} 1,$$

para todo $\varepsilon > 0$ fijo. □

En particular, con este lema, se ha probado que, para d suficientemente grande, una gran proporción de la masa de probabilidad de una distribución $\mathcal{U}(B(\mathbf{0}_d, 1))$ se concentra cerca de la frontera de la $B(\mathbf{0}_d, 1)$. Además, este efecto de concentración se acrecienta conforme se incrementa la dimensión d. Al aplicar el Lema 3.2 para $\varepsilon = \frac{1}{d}$ se tiene que la $\frac{1}{d}$-cáscara de $B(\mathbf{0}_d, 1)$ recoge una proporción de volumen del $1 - (1 - \frac{1}{d})^d$. Además, tal y como vimos en (2.3), el volumen de la bola unidad tiende a cero conforme aumenta la dimensión d.

Existen más propiedades de la distribución $\mathcal{U}(B(\mathbf{0}_d, 1))$ interesantes. Se probará que, para d suficientemente grande, el volumen de $B(\mathbf{0}_d, 1)$ está concentrado no solo en su ε-cáscara sino que también se concentra alrededor de su "ecuador". Tal y como se demostró en la Nota 2.7, si se elige como punto fijo de $\mathbb{R}^d$ el "polo norte" de una bola centrada en $\mathbf{0}_d$ de radio $\sqrt{d}$ y se generan observaciones aleatorias según la distribución normal multivariante en $\mathbb{R}^d$, es de esperar que estos sean ortogonales al punto fijo predefinido, luego estarían situados sobre el ecuador de $B(\mathbf{0}_d, \sqrt{d})$.

Para hacer riguroso el argumento previo habría que probar que la mayoría del volumen de la bola unidad se concentra alrededor de su ecuador, es decir, en una banda de amplitud pequeña alrededor de su "ecuador", por ejemplo, $|x_1| \leq 1/\sqrt{d}$ (como así se obtendría si se pudiese aplicar la desigualdad de Chebyshev desde la Nota 2.7). Esto será probado en el Lema 3.4 para la banda de $B(\mathbf{0}_d, 1)$ comprendida entre los planos $x_1 = \pm\varepsilon$ cuando $d \geq 3$, para $\varepsilon > 0$ suficientemente pequeño. Previamente expondremos una desigualdad que será útil en la demostración.

Lema 3.3 (Desigualdad de Bernoulli)**.** *Sea $x \leq 1$ y $r \geq 1$ o $r \leq 0$, entonces se cumple*

$$(1-x)^r \geq 1 - rx,$$

La desigualdad de Bernoulli es un resultado bien conocido, basado en la convexidad de la función $(1-x)^r$ y su desarrollo de Taylor de primer orden en torno al 0.

Para el siguiente resultado es necesario definir previamente los conjuntos $H = \{\mathbf{x} \in \mathbb{R}^d : x_1 \geq 0\}$ y $E_\varepsilon = \{\mathbf{x} \in \mathbb{R}^d : |x_1| \leq \varepsilon\}$, comúnmente denominados semiespacios.

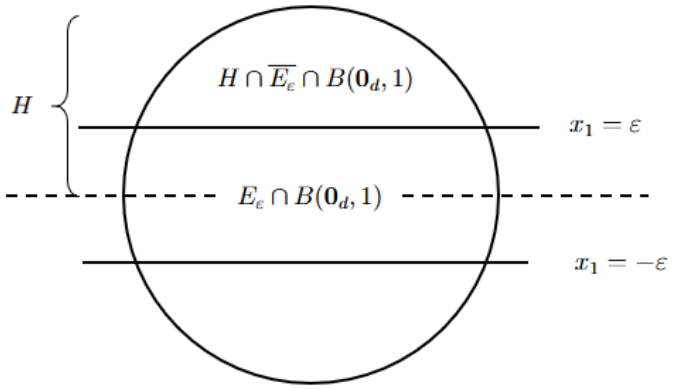

Figura 3.1: Representación de la banda ecuatorial y los hemisferios de $B(\mathbf{0}_d, 1)$

Lema 3.4 (Lema de concentración ecuatorial)**.** *Dado $d \geq 3$ y $\varepsilon > 0$, si $E_\varepsilon \cap B(\mathbf{0}_d, 1)$ es la región de $B(\mathbf{0_d}, 1)$ comprendida entre los planos $x_1 = \varepsilon$ y $x_1 = -\varepsilon$ (véase la Figura 3.1) se cumple que*

$$\ell_d(E_\varepsilon \cap B(\mathbf{0}_d, 1)) \geq \left(1 - \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}}\right) \ell_d(B(\mathbf{0}_d, 1))$$

Demostración. Lo primero que se comprobará es que para $\varepsilon \geq 1$ el resultado es trivial. En ese caso, como $E_\varepsilon \cap B(\mathbf{0}_d, 1) = \{\mathbf{x} \in B(\mathbf{0}_d, 1) \subset \mathbb{R}^d : |x_1| \leq \varepsilon\}$ se tiene que $B(\mathbf{0}_d, 1) \subset E_\varepsilon \cap B(\mathbf{0}_d, 1)$ y usando la monotonía de la integral se cumple que $\ell_d(E_\varepsilon \cap B(\mathbf{0}_d, 1)) \geq \ell_d(B(\mathbf{0}_d, 1))$. Luego basta probar que la constante de la desigualdad del enunciado es positiva y menor o igual que 1. Por otro lado, como $\varepsilon \geq 1$ y $d \geq 3$ y la función e^{-x} es decreciente y positiva, se tiene

$$0 \leq e^{-\frac{\varepsilon^2(d-1)}{2}} \leq e^{-1},$$

y, también, se puede considerar la acotación

$$0 \leq \frac{2}{\varepsilon\sqrt{d-1}} \leq \sqrt{2}.$$

Por tanto,

$$1 \geq c = 1 - \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}} \geq 1 - \sqrt{2} \cdot e^{-1} \geq 0, \tag{3.1}$$

llevando a

$$\ell_d(E_\varepsilon \cap B(\mathbf{0}_d, 1)) \geq c \cdot \ell_d(B(\mathbf{0}_d, 1)).$$

Luego, finalmente, basta probarlo para $0 < \varepsilon < 1$. Por simetría (véase la Figura 3.1), se observa que $\ell_d(E_\varepsilon \cap B(\mathbf{0}_d, 1)) = 2\,\ell_d(H \cap E_\varepsilon \cap B(\mathbf{0}_d, 1))$.

Por tanto, se puede deducir que

$$\begin{aligned}
\frac{\ell_d(E_\varepsilon \cap B(\mathbf{0}_d, 1))}{\ell_d(B(\mathbf{0}_d, 1))} &= \frac{2\,\ell_d(H \cap E_\varepsilon \cap B(\mathbf{0}_d, 1))}{2\,\ell_d(H \cap B(\mathbf{0}_d, 1))} = \\
&= \frac{\ell_d(H \cap B(\mathbf{0}_d, 1)) - \ell_d(H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1))}{\ell_d(H \cap B(\mathbf{0}_d, 1))} = \\
&= 1 - \frac{\ell_d(H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1))}{\ell_d(H \cap B(\mathbf{0}_d, 1))} \quad .
\end{aligned} \tag{3.2}$$

Con este razonamiento, se ve que bastaría demostrar que al menos una c fracción (para c en 3.1) del volumen del hemisferio superior satisface que $x_1 \leq \varepsilon$. Para ello se probará que la proporción del volumen de $H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)$ sobre el volumen de $H \cap B(\mathbf{0}_d, 1)$ tiende a cero calculando una cota superior del volumen de $\ell_d(H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1))$ y una cota inferior para el volumen de $\ell_d(H \cap B(\mathbf{0}_d, 1))$.

En el cálculo del volumen de $H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)$ se usará el principio de Cavalieri, de tal forma que, para cada $x_1 \in [\varepsilon, 1]$, la x_1-sección de $H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)$, que se denotará por D_{x_1}, será una bola centrada en cero en dimensión $d-1$ y de radio $\sqrt{1-x_1^2}$ puesto que

$$\begin{aligned} D_{x_1} =& \{(x_2, \ldots, x_d) \in \mathbb{R}^{d-1} : (x_1, \ldots, x_d) \in H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)\} = \\ =& \{(x_2, \ldots, x_d) \in \mathbb{R}^{d-1} : \|(x_1, \ldots, x_d)\| \leq 1\} = \\ =& \{\mathbf{y} \in \mathbb{R}^{d-1} : x_1^2 + \|\mathbf{y}\|^2 \leq 1\} = B\Big(\mathbf{0}_{d-1}, \sqrt{1-x_1^2}\Big) \subset \mathbb{R}^{d-1}. \end{aligned}$$

Luego,

$$\begin{aligned} \ell_d(H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)) &= \int_\varepsilon^1 \ell_{d-1}(D_{x_1})\, dx_1 = \int_\varepsilon^1 \ell_{d-1}\Big(B\big(\mathbf{0}_{d-1}, \sqrt{1-x_1^2}\big)\Big)\, dx_1 = \\ &= \int_\varepsilon^1 \ell_{d-1}\Big(\sqrt{1-x_1^2} \cdot B(\mathbf{0}_{d-1}, 1)\Big)\, dx_1 = \\ &= \int_\varepsilon^1 (1-x_1^2)^{\frac{d-1}{2}} \cdot \ell_{d-1}\big(B(\mathbf{0}_{d-1}, 1)\big)\, dx_1 = \\ &= \ell_{d-1}\big(B(\mathbf{0}_{d-1}, 1)\big) \cdot \int_\varepsilon^1 (1-x_1^2)^{\frac{d-1}{2}}\, dx_1. \end{aligned} \tag{3.3}$$

Para encontrar una cota superior se usa que $1 - x \leq e^{-x}$, luego:

$$\int_\varepsilon^1 (1-x_1^2)^{\frac{d-1}{2}}\, dx_1 \leq \int_\varepsilon^1 (e^{-x_1^2})^{\frac{d-1}{2}}\, dx_1$$

y ahora como $x_1 \geq \varepsilon$ entonces $\frac{x_1}{\varepsilon} \geq 1$ luego se puede multiplicar la integral anterior por esta cantidad manteniendo la desigualdad y, entonces:

$$\begin{aligned} \int_\varepsilon^1 (e^{-x_1^2})^{\frac{d-1}{2}} \frac{x_1}{\varepsilon}\, dx_1 =& \frac{1}{\varepsilon} \int_\varepsilon^1 (e^{-x_1^2})^{\frac{d-1}{2}} x_1\, dx_1 = \frac{-1}{\varepsilon(d-1)} \int_\varepsilon^1 -(d-1) \cdot x_1 \cdot (e^{-x_1^2})^{\frac{d-1}{2}}\, dx_1 = \\ =& \frac{-1}{\varepsilon(d-1)} \Big[e^{-x^2(\frac{d-1}{2})}\Big]_\varepsilon^1 = \frac{1}{\varepsilon(d-1)} \Big[e^{-x^2(\frac{d-1}{2})}\Big]_1^\varepsilon \leq \frac{1}{\varepsilon(d-1)} e^{-\frac{\varepsilon^2(d-1)}{2}} \end{aligned} \tag{3.4}$$

Por tanto, juntando (3.3) y (3.4) obtenemos que:

$$\ell_d(H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)) \leq \frac{1}{\varepsilon(d-1)} e^{-\frac{\varepsilon^2(d-1)}{2}} \cdot \ell_{d-1}\big(B(\mathbf{0}_{d-1}, 1)\big). \tag{3.5}$$

Ahora falta acotar inferiormente el volumen en dicho hemisferio. En realidad, el volumen de la región del hemisferio por debajo de cualquier plano $x_1 = a$ para $0 < a < 1$ ya

es una cota inferior de todo el volumen. En particular, si se considera $a = \frac{1}{\sqrt{d-1}}$ se tiene que:

$$\ell_d(H \cap B(\mathbf{0}_d, 1)) = \int_0^1 \ell_{d-1}\Big(B\big(\mathbf{0}_{d-1}, \sqrt{1-x_1^2}\big)\Big)\, dx_1 \geq$$
$$\geq \ell_{d-1}\big(B(\mathbf{0}_{d-1}, 1)\big) \int_0^{\frac{1}{\sqrt{d-1}}} (1-x_1^2)^{\frac{d-1}{2}} dx_1, \tag{3.6}$$

y si $x_1 \leq \frac{1}{\sqrt{d-1}}$ entonces $x_1^2 \leq \frac{1}{d-1}$. Por lo cual, se puede seguir acotando inferiormente la última integral de (3.6) como

$$\int_0^{\frac{1}{\sqrt{d-1}}} (1-x_1^2)^{\frac{d-1}{2}}\, dx_1 \geq \int_0^{\frac{1}{\sqrt{d-1}}} (1-\frac{1}{d-1})^{\frac{d-1}{2}}\, dx_1 = (1-\frac{1}{d-1})^{\frac{d-1}{2}} \cdot \frac{1}{\sqrt{d-1}},$$

donde se puede usar la desigualdad de Bernoulli (Lema 3.3) ya que $\frac{d-1}{2} \geq 1$ si $d \geq 3$. Entonces se tiene

$$(1-\frac{1}{d-1})^{\frac{d-1}{2}} \cdot \frac{1}{\sqrt{d-1}} \geq (1-\frac{d-1}{2}\cdot\frac{1}{d-1}) \cdot \frac{1}{\sqrt{d-1}} = \frac{1}{2\sqrt{d-1}}. \tag{3.7}$$

Juntando (3.6) y (3.7) se ha obtenido una cota inferior para el volumen del hemisferio que está relacionada con el volumen de la bola unidad centrada en $\mathbf{0}_{d-1}$ como se conseguía con la cota superior del volumen de $H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)$ de forma que:

$$\ell_d(H \cap B(\mathbf{0}_d, 1)) \geq \frac{1}{2\sqrt{d-1}} \cdot \ell_{d-1}\big(B(\mathbf{0}_{d-1}, 1)\big). \tag{3.8}$$

Finalmente, con (3.5) y (3.8) se puede obtener una cota para la proporción de volumen de $H \cap B(\mathbf{0}_d, 1)$ que está recogida en $H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)$:

$$\frac{\ell_d(H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1))}{\ell_d(H \cap B(\mathbf{0}_d, 1))} \leq \frac{\frac{1}{\varepsilon(d-1)} e^{-\frac{\varepsilon^2(d-1)}{2}} \cdot \ell_{d-1}\big(B(\mathbf{0}_{d-1}, 1)\big)}{\frac{1}{2\sqrt{d-1}} \cdot \ell_{d-1}\big(B(\mathbf{0}_{d-1}, 1)\big)} = \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}} \tag{3.9}$$

Esto prueba que el volumen de $H \cap \overline{E_\varepsilon} \cap B(\mathbf{0}_d, 1)$ tiende a cero para $\varepsilon > 0$ fijo cuando $d \to \infty$ y juntando la desigualdad que se acaba de obtener con la relación obtenida en (3.2) se tiene que:

$$\frac{\ell_d(E_\varepsilon \cap B(\mathbf{0}_d, 1))}{\ell_d(B(\mathbf{0}_d, 1))} \geq 1 - \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}},$$

que es lo que se quería probar. □

Pese a que el resultado que se acaba de probar es muy poderoso en algunas situaciones, se debe tener cuidado con su correcta aplicación.

Nota 3.5. *Denotando la constante que aparece en (3.1) el Lema 3.4 por*

$$\mathrm{C}_{\varepsilon,d} = 1 - \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}} \in (-\infty, 1),$$

se tiene que para $\varepsilon > 0$ fijo se ha visto que $\mathrm{C}_{\varepsilon,d} \xrightarrow{d\to\infty} 1$ y esta es la forma en la que el teorema debe interpretarse. Sin embargo, no es posible interpretar el Lema 3.4 fijando la dimensión d ya que, en ese caso, se cumple que $\mathrm{C}_{\varepsilon,d} \xrightarrow{\varepsilon\to 0} -\infty$. Si se considera $\varepsilon > 0$ y la dimensión d fija, entonces la cota que se ha obtenido del Lema 3.4 solo es útil si $\mathrm{C}_{\varepsilon,d} > 0$ puesto que en cualquier otro caso se estaría acotando inferiormente el volumen de la banda $E_\varepsilon \cap B(\mathbf{0}_d, 1)$ por una cantidad negativa, lo cual es trivial para la medida de Lebesgue. De hecho, si se considera $x = \varepsilon\sqrt{d-1}$, la función $f(x) = 1 - \frac{2}{x}e^{-\frac{x^2}{2}}$ tiene una raíz x_0 en el intervalo $(1, 2)$ por el Teorema de Bolzano. Por lo tanto, si $x > x_0$ entonces $\varepsilon > \frac{x_0}{\sqrt{d-1}}$ y se tiene que $\mathrm{C}_{\varepsilon,d} > 0$. Lo que se acaba de ver es que, fijada la dimensión d, siempre se va a poder encontrar una banda de amplitud $\varepsilon > 0$ que verifique el Lema 3.4. Por sí solo, esto no resulta útil a no ser que se pueda considerar ε arbitrariamente pequeño y, como se ha visto, esto solo sucede en una dimensión d suficientemente grande.

Nótese que al aplicar el Lema 3.2 a $B(\mathbf{0}_d, 1)$ se obtiene una cota para la proporción de volumen que recoge la ε-cáscara de $B(\mathbf{0}_d, 1)$, y la constante $1 - e^{-\varepsilon d}$, que es dependiente de ε y d pero siempre positiva. A su vez, la utilidad de este lema residía justo en que la mayor parte del volumen de la bola se encuentra recogido cerca de su frontera para dimensiones altas, lo cual está fundamentado en que la constante $1 - e^{-\varepsilon d}$ esté cerca de 1 y, por tanto, para d grande es necesario que ε sea lo suficientemente pequeño como para que $e^{-\varepsilon d}$ sea prácticamente 0 mientras que la distancia ε a la frontera de la bola se mantenga pequeña. Esta consideración, unida a la Nota 3.5, remarca la posibilidad de elegir ε lo suficientemente pequeño en casos en los que la dimensión d sea muy alta, proporcionando, de forma conjunta entre ambos lemas, resultados útiles en casos de alta dimensionalidad. Sin embargo, para dimensiones bajas los resultados que se obtienen al aplicar ambos lemas pueden no ser los deseados.

El Lema 3.2, Si X es una variable aleatoria tal que $X \sim \mathcal{U}(B(\mathbf{0}_d, 1))$, prueba que

$$P[\,\|X\| \geq 1 - \varepsilon\,] = \frac{\ell_d(B(\mathbf{0}_d, 1) \setminus B(\mathbf{0_d}, 1 - \varepsilon))}{\ell_d(B(\mathbf{0}_d, 1))} \geq 1 - e^{-\varepsilon d}, \tag{3.10}$$

para cualquier $\varepsilon > 0$. El siguiente resultado extiende este tipo de afirmación a muestras aleatorias simples (m.a.s.) (variables aleatorias independientes e igualmente distribuidas) generadas desde $X \sim \mathcal{U}(B(\mathbf{0}_d, 1))$.

Teorema 3.6. *Dado un espacio probabilístico (Ω, σ, P). Se considera, para $n \geq 2$, una m.a.s. $X_1, \ldots, X_n$ con $X_i : \Omega \to \mathbb{R}^d$ para $d \geq 3$ y tal que $X_i \sim \mathcal{U}(B(\mathbf{0}_d, 1))$. Entonces se cumple que*

$$P\Big[\, \|X_i\| \geq 1 - \frac{2\log n}{d} \text{ para todo } i = 1, \ldots, n \,\Big] \geq 1 - \frac{1}{n}.$$

Demostración. El resultado es consecuencia del Lema 3.2 eligiendo $\varepsilon = \frac{2\log n}{d} > 0$, para obtener

$$P[\|X_i\| < 1 - \varepsilon\,] = \frac{\ell_d(B(\mathbf{0}_d, 1 - \varepsilon))}{\ell_d(B(\mathbf{0}_d, 1))} \leq e^{-\varepsilon d} = e^{-\frac{2\log n}{d}\cdot d} = \frac{1}{n^2}.$$

El resultado final se deduce de la acotación

$$P\left[\|X_i\| \geq 1 - \frac{2\log n}{d} \text{ para todo } i = 1, \ldots, n\right] = 1 - P\left[\bigcup_{i=1}^{n}\left\{\|X_i\| < 1 - \frac{2\log n}{d}\right\}\right]$$

$$\geq 1 - \sum_{i=1}^{n} P\left(\|X_i\| < 1 - \frac{2\log n}{d}\right) \geq 1 - \sum_{i=1}^{n} \frac{1}{n^2} = 1 - \frac{1}{n}.$$

□

La importancia de este último teorema reside en comprobar que el resultado de una muestra aleatoria de tamaño n de una distribución uniforme en $B(\mathbf{0}_d, 1)$ se sitúa con alta probabilidad cerca de la frontera de la bola, extendiendo lo que se había probado para una única observación. De hecho, el resultado proporciona una cota para la probabilidad de que todas las observaciones de la muestra se alejen más que la distancia $\frac{2\log n}{d}$ de la superficie de $B(\mathbf{0}_d, 1)$. Y, puesto que esta distancia es variable según la dimensión d y el tamaño de la muestra n, es necesario hacer algunas consideraciones adicionales para poder aplicar correctamente el resultado.

Nota 3.7. *Las condiciones del Teorema 3.6 sobre d y n son necesarias, pues si $d = 1$ entonces se calcularía $P[\|X_i\| < 1 - 2\log n]$ lo cual es evidentemente igual a 1 (con $n = 1$) o igual a 0 (para $n > 1$) si $X_i \sim \mathcal{U}((-1, 1))$. En el caso de $d = 2$ entonces $\log 2 < 1$ y necesitaríamos que $n > 2$ para que la cantidad $1 - \frac{2\log n}{d}$ sea positiva. Puesto que este caso no es muy interesante al ser una dimensión lo suficientemente baja como para que otras técnicas estadísticas nos proporcionen resultados más precisos, nos podemos limitar al caso $d \geq 3$ con $n \geq 2$.*

La cantidad $\frac{2\log n}{d}$ disminuye con la dimensión d si n es fijo, luego, para el mismo tamaño de muestra, cuanto mayor sea la dimensión sobre la que se esté trabajando, menor puede ser la distancia a la superficie que se considere para obtener la misma probabilidad. Sin embargo, la distancia aumenta logarítmicamente conforme aumenta el tamaño de la muestra n y si $\log n > \frac{d}{2}$ entonces el Teorema 3.6 carece de sentido, ya que $\|X\|$ es trivialmente siempre mayor o igual que 0. A su vez, para d fijo, la probabilidad es dependiente de $1/n$ luego esta mejorará conforme aumente el tamaño de la muestra n más rápido de lo que empeorará la distancia a la superficie $\frac{2\log n}{d}$ que estamos considerando.

En vista de estas consideraciones, es evidente que el Teorema 3.6 proporciona resultados valiosos cuando d sea grande y el tamaño muestral n sea moderadamente grande en comparación con d, de tal manera que las cantidades $1 - \frac{2\log n}{d}$ y $1 - \frac{1}{n}$ sean aproximadamente 1.

Teniendo ahora en cuenta el Teorema 3.4, nos interesa estudiar bajo qué condiciones cabe esperar que todo par de observaciones $\{X_i, X_{i'}\}$ extraídas de una m.a.s. $X_1, ..., X_n$ generada por una distribución con $X_i \sim \mathcal{U}(B(\mathbf{0}, 1))$ sean aproximadamente ortogonales. Nótese que este resultado también mantiene conexiones con lo visto en la Proposición 2.6 (y en la Nota 2.7) para el caso de distribuciones normales en alta dimensionalidad. El caso de una m.a.s. de la distribución uniforme en la bola unidad se presentará en el Teorema 3.8.

Teorema 3.8. *Dado un espacio probabilístico (Ω, σ, P). Se considera, para $n \geq 2$, una m.a.s. $X_1, \ldots, X_n$ con $X_i : \Omega \to \mathbb{R}^d$ donde $d \geq 3$ y tal que $X_i \sim \mathcal{U}(B(\mathbf{0}, 1))$. Entonces se cumple que*

$$P\left[|\langle X_i, X_{i'}\rangle| \leq \frac{\sqrt{6\log n}}{\sqrt{d-1}} \text{ para todo par } \{i, i'\} \text{ con } i \neq i'\right] \geq 1 - \frac{1}{n}. \tag{3.11}$$

Demostración. Sean $X, Y \sim \mathcal{U}(B(\mathbf{0}_d, 1))$ variables aleatorias independientes y se fija $\varepsilon = \frac{\sqrt{6 \cdot \log n}}{\sqrt{d-1}}$. Sea $\omega \in \Omega$ tal que $X(\omega) = \mathbf{x} \neq \mathbf{0}_d$. Se puede construir una base ortonormal $\{\mathbf{u}_1, \ldots, \mathbf{u}_d\}$ del espacio vectorial $\mathbb{R}^d$ con $\mathbf{u_1} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$, siguiendo el proceso de ortogonalización de Gram-Schmidt iniciado en $\mathbf{u}_1$. Entonces, existirá una única aplicación lineal ortonormal $f : \mathbb{R}^d \to \mathbb{R}^d$ que convierte la base $\{\mathbf{u}_1, \ldots, \mathbf{u}_d\}$ en la base estándar $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ con $\mathbf{e}_l = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^d$, con un 1 en la posición l. Es decir, $f(\mathbf{u}_l) = \mathbf{e}_l$ para $l = 1, \ldots, d$. Como f es una aplicación lineal ortogonal que conserva productos escalares, se tiene $|\langle X(\omega), Y(\omega)\rangle| = |\langle f\big(X(\omega)\big), f\big(Y(\omega)\big)\rangle|$ y, como $X(\omega) = \mathbf{x} = \|\mathbf{x}\| \cdot \mathbf{u}_1$, entonces

$$\begin{aligned}|\langle X(\omega), Y(\omega)\rangle| =& |\langle f\big(X(\omega)\big), f\big(Y(\omega)\big)\rangle| = |\langle \|\mathbf{x}\| \cdot f(\mathbf{u}_1), f\big(Y(\omega)\big)\rangle| = \\ =& \|\mathbf{x}\| \cdot |\langle \mathbf{e}_1, f\big(Y(\omega)\big)\rangle| = \|X(\omega)\| \cdot |\langle \mathbf{e}_1, f\big(Y(\omega)\big)\rangle|.\end{aligned}$$

Dado que la esfera es invariante por transformaciones ortogonales, se tiene

$$\begin{aligned}P\big[|\langle \mathbf{e}_1, f(Y)\rangle| > \varepsilon\big] =& \frac{\ell_d(\{y \in B(\mathbf{0}_d, 1) : |\langle \mathbf{e}_1, f(y)\rangle| > \varepsilon\})}{\ell_d(B(\mathbf{0}_d, 1))} = \\ =& \frac{\ell_d(\{f^{-1}(\mathbf{z}) \in B(\mathbf{0}_d, 1) : |\langle \mathbf{e}_1, \mathbf{z}\rangle| > \varepsilon\})}{\ell_d(B(\mathbf{0}_d, 1))} = \\ =& \frac{\ell_d(\{(z_1, \ldots, z_d) \in B(\mathbf{0}_d, 1) : |z_1| > \varepsilon\})}{\ell_d(B(\mathbf{0}_d, 1))},\end{aligned}$$

y se puede usar ahora el Lema 3.4 para acotar esta probabilidad de la forma

$$P\big[|\langle \mathbf{e}_1, f(Y)\rangle| > \varepsilon\big] = \frac{\ell_d(\{(z_1, \ldots, z_d) \in B(\mathbf{0}, 1) : |z_1| > \varepsilon\})}{\ell_d(B(\mathbf{0}, 1))} \leq \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}},$$

donde, si

$$\varepsilon = \frac{\sqrt{6 \cdot \log n}}{\sqrt{d-1}} \text{ y } \frac{2}{\sqrt{6 \cdot \log n}} \leq 1,$$

se llega a

$$P\big[|\langle \mathbf{e}_1, f(Y)\rangle| > \varepsilon\big] \leq \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}} = \frac{2}{\sqrt{6 \cdot \log n}} \cdot e^{-\frac{6 \cdot \log n}{2}} \leq n^{-3}. \tag{3.12}$$

En realidad, la elección de la aplicación lineal ortonormal f depende de la elección inicial de $\omega \in \Omega$ puesto que se necesita $X(\omega) \neq 0$. Para eliminar esta dependencia, se puede escribir

$$\begin{aligned}P\big[|\langle X, Y\rangle| > \varepsilon\big] =& P(\{\omega \in \Omega : |\langle X(\omega), Y(\omega)\rangle| > \varepsilon\}) = \\ =& P(\{\omega \in \Omega : \|X(\omega)\| \cdot |\langle \mathbf{e}_1, f\big(Y(\omega)\big)\rangle| > \varepsilon\}) \leq \\ \leq& P(\{\omega \in \Omega : |\langle \mathbf{e}_1, f\big(Y(\omega)\big)\rangle| > \varepsilon\}) \leq \frac{1}{n^3},\end{aligned} \tag{3.13}$$

donde se ha usado la cota en (3.12) y que si $X(\omega) \in B(\mathbf{0}_d, 1)$ entonces $\|X(\omega)\| \leq 1$.

Se está ahora en condiciones de repetir un razonamiento similar al de prueba del Teorema 3.6, teniendo en cuenta que para una m.a.s. de tamaño n hay $\binom{n}{2}$ pares de

productos escalares $\langle X_i, X_{i'}\rangle$, y que

$$P\left[\,|\langle X_i, X_{i'}\rangle| \leq \varepsilon \text{ para todo } i \neq i'\right] = 1 - P\left[\bigcup_{i\neq i'}\left\{|\langle X_i, X_{i'}\rangle| > \varepsilon\right\}\right] \geq$$
$$\geq 1 - \binom{n}{2}\frac{1}{n^3} \geq 1 - \frac{n^2-n}{n^3} \geq 1 - \frac{1}{n}.$$

□

Nota 3.9. *El Teorema 3.8 depende, como es obvio, del tamaño de la muestra n y de la dimensión d, lo que produce cambios en la cota asociada de que la probabilidad de que $\langle X_i, X_{i'}\rangle$ esté alejada de 0. Ese tipo de consideraciones son análogas a las formuladas en la Nota 3.7 para el Teorema 3.6. Si $n = 2$, el Teorema 3.8 acota inferiormente la probabilidad de que el producto escalar entre dos variables aleatorias $\langle X, Y\rangle$ sea menor o igual que una constante con probabilidad $1/2$, por lo que no parece que el resultado así formulado tenga excesiva relevancia. No obstante, en la demostración del teorema, se prueba también que*

$$P[|\langle X, Y\rangle| \leq \varepsilon\,] \geq 1 - \frac{2}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{2}},$$

para cualquier $\varepsilon > 0$. Por lo tanto, incluso para tamaños muestrales n reducidos, se puede mejorar la cota si d es fijo y hacer la acotación relevante, como fue ya explicado en la Nota 3.5.

Combinando los teoremas 3.10 y 3.8, se deduce que para cualquier m.a.s $X_1, ..., X_n$, generada por una distribución con $X_i \sim \mathcal{U}(B(\mathbf{0}_d, 1))$, todas las observaciones se encuentran con gran probabilidad sobre la superficie de $B(\mathbf{0}_d, 1)$ y son prácticamente ortogonales, ya que las parejas de productos escalares $\langle \frac{X_i}{\|X_i\|}, \frac{X_{i'}}{\|X_{i'}\|}\rangle$ también verifican el Teorema 3.12. En vista de estas consideraciones, argumentos como los expuestos en la Nota 2.5 prueban intuitivamente que la distancia entre cualquier par de observaciones $X, Y \sim \mathcal{U}(B(\mathbf{0}_d, 1))$ dada por $\|X - Y\|$ debe ser próxima a $\sqrt{2}$. Este es el resultado que se presenta a continuación.

Teorema 3.10. *Dado un espacio probabilístico (Ω, σ, P), si consideramos $X, Y : \Omega \to \mathbb{R}^d$ tal que $X, Y \sim \mathcal{U}(B(\mathbf{0}_d, 1))$, se cumple:*

$$P[\big|\,\|X - Y\| - \sqrt{2}\big| \leq \varepsilon\,] \geq 1 - 2d^{-\varepsilon d/5} - \frac{9}{\varepsilon\sqrt{d-1}} e^{-\frac{\varepsilon^2(d-1)}{36}},$$

si $0 < \varepsilon < 1$ y $d \geq 3$.

Demostración. Se tiene que $V = \|X - Y\| + \sqrt{2} \geq \sqrt{2}$ y que $U\,V \geq \sqrt{2}\,U$ si $U = \big|\,\|X - Y\| - \sqrt{2}\,\big|$. Por tanto, $U > \varepsilon$ si y solo si $\sqrt{2}\,U > \sqrt{2}\,\varepsilon$, implicando que $U\,V > \sqrt{2}\,\varepsilon$. Entonces se tiene la siguiente contención de sucesos:

$$\left(\big|\,\|X - Y\| - \sqrt{2}\,\big| > \varepsilon\right) \subset \left(\big|\,\|X - Y\| - \sqrt{2}\,\big| \cdot \left(\|X - Y\| + \sqrt{2}\right) > \varepsilon\right),$$

y, la propiedad de monotonía de la probabilidad, garantiza

$$P\left(\big|\,\|X - Y\| - \sqrt{2}\,\big| > \varepsilon\right) \leq P\left(\big|\,\|X - Y\| - \sqrt{2}\,\big| \cdot \left(\|X - Y\| + \sqrt{2}\right) > \sqrt{2}\,\varepsilon\right). \quad (3.14)$$

Ahora, escribiendo la probabilidad que aparece en el enunciado en términos de su suceso complementario y usando (3.14) se obtiene

$$\begin{aligned} P\left[\,\left|\|X-Y\|-\sqrt{2}\,\right|\leq\varepsilon\,\right] &= 1-P\left[\,\left|\|X-Y\|-\sqrt{2}\,\right|>\varepsilon\,\right]\geq \\ &\geq 1-P\left(\,\left|\,\|X-Y\|-\sqrt{2}\,\right|\cdot(\|X-Y\|+\sqrt{2})>\sqrt{2}\,\varepsilon\,\right)= \\ &= 1-P\left(\,\left|\,\|X-Y\|^2-2\,\right|>\sqrt{2}\,\varepsilon\,\right)= \\ &= P\left(-\sqrt{2}\,\varepsilon<\|X-Y\|^2-2<\sqrt{2}\,\varepsilon\,\right)= \\ &= P\left(-\sqrt{2}\,\varepsilon<\|X\|^2+\|Y\|^2-2\,\langle X,Y\rangle-2<\sqrt{2}\,\varepsilon\,\right), \end{aligned} \tag{3.15}$$

Y aquí es donde intervienen los Teoremas 3.2 y 3.8 al aparecer $\|X\|^2$, $\|Y\|^2$ y $\langle X,Y\rangle$ en 3.15.

En primer lugar, se agrupan $\|X\|^2+\|Y\|^2-2\,\langle X,Y\rangle-2$ como $(\|X\|^2-1)+(\|Y\|^2-1)-2\,\langle X,Y\rangle$ y se desea que esta suma de tres términos sea en valor absoluto menor que $\sqrt{2}\varepsilon$ luego se definen

$$A=\left(\,\left|(\|X\|^2-1)+(\|Y\|^2-1)-2\,\langle X,Y\rangle\right|<\sqrt{2}\varepsilon\,\right)$$

y

$$B=\left\{\left\{\left|\|X\|^2-1\right|<\frac{\sqrt{2}\,\varepsilon}{3}\right\}\bigcap\left\{\left|\|Y\|^2-1\right|<\frac{\sqrt{2}\,\varepsilon}{3}\right\}\bigcap\left\{\left|\langle X,Y\rangle\right|<\frac{\sqrt{2}\,\varepsilon}{6}\right\}\right\},$$

y se tiene $A\subset B$ y, consecuentemente,

$$P(A)\leq P(B). \tag{3.16}$$

Los sucesos que constituyen B son independientes y se pueden usar los resultados previos para su acotación.

En primer lugar, para poder usar el Teorema 3.2 debemos aplicar la desigualdad de Bernoulli (ver Lema 3.3) y así conseguir una expresión como la de (3.10). De esta forma, se puede escribir

$$\begin{aligned} P\left[\,\left|\|X\|^2-1\right|<\frac{\sqrt{2}\,\varepsilon}{3}\,\right] &= P\left[\,1-\|X\|^2<\frac{\sqrt{2}\,\varepsilon}{3}\,\right]= \\ &= P\left[\,\|X\|^2\geq 1-\frac{\sqrt{2}\,\varepsilon}{3}\,\right]= \\ &= P\left[\,\|X\|\geq\left(1-\frac{\sqrt{2}\,\varepsilon}{3}\right)^{1/2}\,\right]\geq \\ &\geq P\left[\,\|X\|\geq\left(1-\frac{1}{2}\cdot\frac{\sqrt{2}\,\varepsilon}{3}\right)\,\right]\geq \\ &\geq 1-e^{-\frac{\sqrt{2}\,\varepsilon}{6}d}, \end{aligned} \tag{3.17}$$

donde directamente se ha aplicado (3.10) del Lema 3.2. La desigualdad de Bernoulli es aplicable ya que $0<1/2<1$ y $\sqrt{2}\varepsilon/3<1$ si $0<\varepsilon<1$. Como Y está igualmente distribuido que X y son independientes, también se tiene

$$P\left[\,\left|\|X\|^2-1\right|<\frac{\sqrt{2}\,\varepsilon}{3}\,\right]=P\left[\,\left|\|Y\|^2-1\right|<\frac{\sqrt{2}\,\varepsilon}{3}\,\right]$$

Y, por último, para la última de las probabilidades se hace uso del Teorema 3.8 y entonces

$$P\left[\,\left|\langle X,Y\rangle\right| < \frac{\sqrt{2}\,\varepsilon}{6}\,\right] \geq 1 - \frac{6\sqrt{2}}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{36}}. \tag{3.18}$$

Ahora, juntando los resultados de (3.15), (3.16), (3.17) y (3.18) se obtiene que

$$\begin{aligned} P\left[\,\left|\|X-Y\| - \sqrt{2}\,\right| \leq \varepsilon\,\right] \geq P\left[\,\left|\|X\|^2 - 1\,\right| \leq \frac{\sqrt{2}\,\varepsilon}{3}\,\right]^2 \cdot P\left[\,\left|\langle X,Y\rangle\right| < \frac{\sqrt{2}\,\varepsilon}{6}\,\right] \geq \\ \geq \left(1 - e^{-\frac{\sqrt{2}\,\varepsilon}{6}d}\right)^2 \cdot \left(1 - \frac{6\sqrt{2}}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{36}}\right) \geq \\ \geq 1 - 2e^{-\frac{\sqrt{2}\,\varepsilon}{6}d} - \frac{6\sqrt{2}}{\varepsilon\sqrt{d-1}} \cdot e^{-\frac{\varepsilon^2(d-1)}{36}}, \end{aligned} \tag{3.19}$$

y el resultado buscado se consigue tras aplicar las desigualdades $\frac{\sqrt{2}}{6} \geq \frac{1}{5}$ y $6\sqrt{2} \leq 9$. □

De nuevo, la cota que se proporciona en el Teorema 3.10 depende de manera conjunta de ε y d de manera análoga a cómo sucedía en las notas 3.7 y 3.9. Luego, para d fijo, se puede conseguir una acotación relevante haciendo ε lo suficientemente pequeño.

En conclusión, en esta sección se han obtenido resultados relevantes para distribuciones uniformes en $B(\mathbf{0}_d, 1)$ en alta dimensión y se ha demostrado cómo cambian las regiones de alta probabilidad asociadas, cómo se comportan las distancias entre observaciones y cómo los vectores son prácticamente ortogonales. Estos resultados generalizan las intuiciones de la Sección 2.2 y demuestran que para distribuciones uniformes X e Y en $B(\mathbf{0}_d, 1)$ se tiene que $E(\|X\|) \approx 1$, $E(\|X-Y\| \approx \sqrt{2})$ y que $E(\langle X,Y\rangle) \approx 0$. Estas características tienen paralelismo en gran medida con los resultados obtenidos en la Sección 2.1 para distribuciones normales multivariantes.

Capítulo 4

Distribución normal multivariante en alta dimensión

En este capítulo, el objetivo será completar las pruebas realizadas en la Sección 2.1 para los teoremas 2.2, 2.4 y 2.6 dando resultados que nos aseguren que los valores obtenidos para las esperanzas de $\|X\|$, $\|X-Y\|$, e $\langle X, Y\rangle$, si $X, Y \sim \mathcal{N}(\mathbf{0}_d, I_d)$ e independientes, son también aplicables a muestras aleatorias de estas distribuciones. Se proporcionarán resultados similares a los que fueron obtenidos en el Capítulo 3 para la distribución uniforme en $B(\mathbf{0}_d, 1)$.

El Teorema 2.2 establecía que $|E(\|X\|) - \sqrt{d}| \leq \frac{1}{\sqrt{d}}$ y que $\mathrm{Var}(\|X\|) \leq 2$ con lo que se probó, mediante la desigualdad de Chebyshev, en la Nota 2.3 que

$$P(|\|X\| - \sqrt{d}| \geq \varepsilon) \leq \frac{2}{\left(\varepsilon - \frac{1}{\sqrt{d}}\right)^2}.$$

Lo cual, junto a las simulaciones, nos permite pensar que, para m.a.s. $X_1, ..., X_n$ de $X_i \sim \mathcal{N}(\mathbf{0}_d, I_d)$ para $i = 1, \ldots, n$, las observaciones muestrales X_i se encuentran localizadas en torno a la superficie de una esfera de centro $\mathbf{0}_d$ y radio $\sqrt{d}$. Y, como vimos en el Capítulo 2, sabemos que el volumen de cualquier bola centrada en $\mathbf{0}_d$ se concentra en las coronas esféricas más próximas a su superficie en virtud de lo probado en (3.10). Sin embargo, esa primera desigualdad de concentración es mejorable con el uso de técnicas más avanzadas. Para ello, mejoraremos esa cota mediante el uso del denominado Teorema del Anillo Gaussiano, para lo cual es necesario demostrar una serie de resultados previos.

Lema 4.1 (Desigualdad de Bernstein)**.** *Sea (Ω, σ, P) un espacio probabilístico y $X_1, \ldots, X_d : \Omega \to \mathbb{R}$ variables aleatorias independientes dos a dos con $E(X_l) = 0$ y momentos de orden k acotados por $|E(X_l^k)| \leq \frac{k!}{2}$ para todo $l = 1, \ldots, d$ y $k \geq 2$. En estas condiciones, para todo $\varepsilon > 0$, se tiene que:*

$$P\left[\,|X_1 + \cdots + X_d| \geq \varepsilon\,\right] \leq 2e^{-\frac{1}{4}\min(\frac{\varepsilon^2}{d}, \varepsilon)}$$

Demostración. Se toma $X = X_1 + \cdots + X_d$ y se considera $0 < t \leq 1/2$. Usando el crecimiento estricto de la función e^x en todo $x \in \mathbb{R}$, se tiene que $tX \geq t\varepsilon$ si y solo si $e^{tX} \geq e^{t\varepsilon}$, luego

$$P(X \geq \varepsilon) = P(e^{tX} \geq e^{t\varepsilon}), \tag{4.1}$$

donde se puede usar ahora la desigualdad de Markov para probar

$$P(e^X \geq e^\varepsilon) \leq \frac{E(e^{tX})}{e^{t\varepsilon}} = e^{-t\varepsilon}\, E(e^{t(X_1+\cdots+X_d)}) = e^{-t\varepsilon} \prod_{l=1}^{d} E(e^{tX_l}), \tag{4.2}$$

ya que las X_i son independientes dos a dos. Entonces, usando el desarrollo en serie de Taylor de la exponencial se tiene

$$E(e^{tX_l}) = E\Big(1 + tX_l + \sum_{k=2}^{\infty} \frac{(tX_l)^k}{k!}\Big) \leq 1 + \sum_{k=2}^{\infty} \frac{t^k E(X_l)^k}{k!},$$

y como hipótesis se ha asumido que $|E(Y_l^k)| \leq \frac{k!}{2}$, luego podemos continuar acotando la función generadora de momentos $E(e^{tX_l})$ como

$$\begin{aligned} E(e^{tX_l}) \leq & 1 + \sum_{k=2}^{\infty} \frac{t^k E(X_l)^k}{k!} \leq 1 + \sum_{k=2}^{\infty} \frac{t^k}{2} = 1 - \frac{1}{2}\Big(1 + t - \sum_{k=0}^{\infty} t^k\Big) = \\ = & 1 - \frac{1}{2}\Big(1 + t - \frac{1}{1-t}\Big) = 1 + \frac{1}{2}\frac{t^2}{1-t} \leq 1 + \frac{1}{2}2t^2 = 1 + t^2 \leq e^{t^2}, \end{aligned} \tag{4.3}$$

donde se ha usado la fórmula de sumación de una serie geométrica con razón $0 < t \leq 1/2$ y desigualdades elementales basadas en que $0 < t \leq 1/2$. Tomando el inferior en (4.1) con $0 < t \leq 1/2$ y juntándolo con lo obtenido en (4.2) y (4.3) se llega a

$$P(X \geq \varepsilon) \leq \inf_{0<t\leq 1/2} e^{-t\varepsilon} \prod_{l=1}^{d} E(e^{tX_l}) \leq \inf_{0<t\leq 1/2} e^{-t\varepsilon} \prod_{l=1}^{d} e^{t^2} = \inf_{0<t\leq 1/2} e^{-t\varepsilon + dt^2} \tag{4.4}$$

De hecho, esta cota se alcanza para algún $t \in (0, \frac{1}{2}]$. Si consideramos la función $f(x) = e^{-\varepsilon x + dx^2}$ (en $\mathcal{C}^\infty(\mathbb{R})$) que alcanza sus extremos relativos en sus puntos críticos:

$$\frac{d}{dx}(e^{-\varepsilon x + dx^2}) = (-\varepsilon + 2dx) \cdot e^{-\varepsilon x + dx^2} = 0 \text{ si y solo si } x = \frac{\varepsilon}{2d}.$$

Como $f(x) \to \infty$ para $d \to \pm\infty$, entonces este es el único extremo relativo de la función f y, además, es un mínimo absoluto. Por tanto, si $x_0 = \frac{\varepsilon}{2d} \leq \frac{1}{2}$ entonces la restricción de f al intervalo $(0, \frac{1}{2}]$ encuentra el mínimo en x_0. En cambio, si $x_0 > \frac{1}{2}$ entonces f debe ser decreciente en $(0, \frac{1}{2}]$ y, por tanto, el mínimo se encuentra en $x = \frac{1}{2}$. Para encontrarse en el primer caso es necesario que $\varepsilon > d$ mientras que cualquier otra situación lleva al segundo caso. En definitiva, se tiene que si $\varepsilon > d$ entonces el valor mínimo de f en $(0, \frac{1}{2}]$ es $e^{-\varepsilon x_0 + dx_0^2} = e^{-\frac{\varepsilon}{4d}}$, y si $\varepsilon \leq d$ entonces el valor mínimo f en $(0, \frac{1}{2}]$ es $e^{-\varepsilon\frac{1}{2} + d(\frac{1}{2})^2} < e^{-\frac{\varepsilon}{4}}$. Añadiendo esta información a (4.4) se obtiene

$$\begin{aligned} P(X \geq \varepsilon) \leq & \inf_{0<t\leq 1/2} e^{-t\varepsilon + dt^2} = \min_{0<t\leq 1/2} e^{-t\varepsilon + dt^2} = e^{-t\varepsilon + dt^2}\Big|_{t=\min(\frac{\varepsilon}{2d}, \frac{1}{2})} \leq e^{\min(-\frac{\varepsilon^2}{4d}, -\frac{\varepsilon}{4})} = \\ = & e^{-\max(\frac{\varepsilon^2}{4d}, \frac{\varepsilon}{4})} \leq e^{-\min(\frac{\varepsilon^2}{4d}, \frac{\varepsilon}{4})}. \end{aligned} \tag{4.5}$$

Entonces se acaba de demostrar que

$$P(X \geq \varepsilon) \leq e^{-\frac{1}{4}\min(\frac{\varepsilon^2}{d}, \varepsilon)}.$$

Para probar el resultado pedido, bastaría reemplazar X con $-X$ y seguir el mismo razonamiento para probar que

$$P(-X \geq \varepsilon) \leq e^{-\frac{1}{4}\,\text{mín}(\frac{\varepsilon^2}{d},\varepsilon)}$$

De donde se obtiene la desigualdad del enunciado

$$P\left[\,|X_1 + \cdots + X_d| \geq \varepsilon\,\right] \leq 2e^{-\frac{1}{4}\,\text{mín}(\frac{\varepsilon^2}{d},\varepsilon)}$$

□

Por lo visto en el Teorema 2.2 se sabe que $E(\|X\|) \approx \sqrt{d}$ para d suficientemente grande, si $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$. Lo que se pretenderá demostrar en el Teorema del Anillo Gaussiano, que se formulará en el Teorema 4.2, es que las observaciones de una muestra de esta distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ deben estar concentradas alrededor de $\sqrt{d}$ con alta probabilidad. A este valor, $\sqrt{d}$, se le conoce como "radio" de la distribución normal multivariante (ver Figura 4.1).

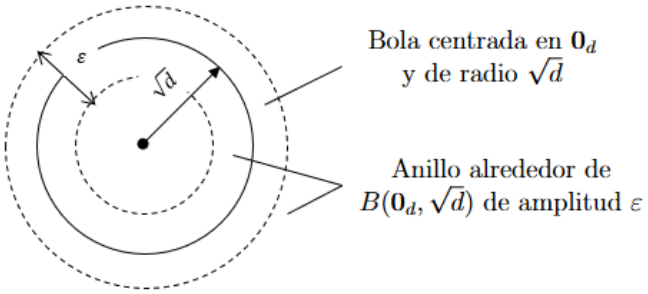

Figura 4.1: Representación del "anillo" alrededor de $B(\mathbf{0}_d, \sqrt{d})$

Para probar este resultado es necesario hacer uso de la desigualdad para las colas de la suma de variables aleatorias independientes con momentos acotados que fue obtenida en el Lema 4.1.

Teorema 4.2 (Teorema del Anillo Gaussiano)**.** *Sean (Ω, σ, P) un espacio probabilístico, $X : \Omega \to \mathbb{R}^d$ una variable aleatoria con distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ y $0 \leq \varepsilon \leq \sqrt{d}$. Entonces se cumple que*

$$P\left[\,\left|\|X\| - \sqrt{d}\,\right| \leq \varepsilon\,\right] \geq 1 - 2e^{\frac{-\varepsilon^2}{16}}.$$

Demostración. Puesto que $\|X\| + \sqrt{d} \geq \sqrt{d}$ entonces multiplicando por la cantidad $\|X\| - \sqrt{d}$ se tiene que $\left|\|X\| - \sqrt{d}\,\right| \geq \varepsilon$ si y solo si $\left|\|X\| - \sqrt{d}\,\right| \cdot \left(\|X\| + \sqrt{d}\right) \geq \varepsilon\sqrt{d}$. Por tanto, se tiene

$$\begin{aligned}
P\left[\,\left|\|X\| - \sqrt{d}\,\right| \geq \varepsilon\,\right] &\leq P\left[\,\left|\|X\| - \sqrt{d}\,\right| \cdot (\|X\| + \sqrt{d}) \geq \varepsilon\sqrt{d}\,\right] = \\
&= P\left[\,\left|X_1^2 + \cdots + X_d^2 - d\,\right| \geq \varepsilon\sqrt{d}\,\right] = \\
&= P\left[\,\left|(X_1^2 - 1) + \cdots + (X_d^2 - 1)\,\right| \geq \varepsilon\sqrt{d}\,\right] = \\
&= P\left[\left|\frac{1}{2}(X_1^2 - 1) + \cdots + \frac{1}{2}(X_d^2 - 1)\right| \geq \frac{\varepsilon\sqrt{d}}{2}\right] \qquad (4.6)
\end{aligned}$$

De esta manera, se ha transformado la probabilidad original en una probabilidad basada en sumas de variables aleatorias $Y_l = \frac{1}{2}(X_l^2 - 1)$ para $l = 1, \ldots, d$ a las cuales se les puede aplicar la desigualdad de Bernstein del Lema 4.1. Para ello, hay que comprobar que verifican todas las hipótesis. En primer lugar, si $X \sim \mathcal{N}(\mathbf{0}_d, I_d)$ y $X = (X_1, \ldots, X_d)$ entonces $X_l \sim \mathcal{N}(0, 1)$ con $l = 1, \ldots, d$ e independientes entre sí (véase la Proposición A.8). Por tanto, las variables aleatorias Y_l serán independientes por pares, por serlo las X_l, y cumplen que $E(Y_l) = 0$, ya que $E(X_l^2) = \mathrm{Var}(X_l) = 1$. Consecuentemente, bastaría con demostrar que los momentos orden $k \geq 2$ de las Y_l están acotados por $\frac{k!}{2}$. Como las Y_l están igualmente distribuidas, para todo $l = 1, \ldots, d$, basta comprobarlo para una de ellas. Es decir, se necesita probar que

$$\left| E\left(\frac{X_l^2 - 1}{2}\right)^k \right| \leq \frac{k!}{2}.$$

Se puede reescribir el término de la izquierda como

$$\frac{1}{2^k} E(|X_l^2 - 1|^k),$$

y, para acotar esta última esperanza, se puede usar la desigualdad $|x^2 - 1|^k \leq 1 + x^{2k}$, que se cumple para todo $x \in \mathbb{R}$ porque si $|x| \leq 1$ entonces $|x^2 - 1|^k \leq 1$ y si $|x| \geq 1$ se tiene $|x^2 - 1|^k \leq x^{2k}$ y, en ningún caso, estas cantidades exceden la cota proporcionada. Aplicando esta desigualdad y la definición de esperanza se tiene

$$\begin{aligned}
\left| E\Big(\frac{X_l^2 - 1}{2}\Big)^k \right| = \frac{1}{2^k} E(|X_l^2 - 1|^k) \leq & \frac{1}{2^k} E(1 + X_l^{2k}) = \frac{1}{2^k} \int_{\mathbb{R}} (1 + x^{2k}) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \\
= & \frac{1}{2^k} \left(1 + \int_{\mathbb{R}} x^{2k} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx\right) = \frac{1}{2^k} \left(1 + \frac{(2k)!}{2^k k!}\right) = \\
= & \frac{1}{2^k} \left(1 + \frac{(2k) \cdot (2k-1) \cdots 2 \cdot 1}{2^k \cdot k \cdot (k-1) \cdots 2 \cdot 1}\right) = \\
= & \frac{1}{2^k} \Big(1 + (2k-1) \cdot (2k-3) \cdots 3 \cdot 1\Big) \qquad (4.7)
\end{aligned}$$

En el Apéndice A.1 se proporciona la expresión para los momentos de una distribución normal estándar que se ha usado. Falta solo comprobar que

$$\frac{1}{2^k} \Big(1 + (2k-1) \cdot (2k-3) \cdots 3 \cdot 1\Big) \leq \frac{k \cdot (k-1) \cdots 2 \cdot 1}{2},$$

lo cual puede demostrarse por inducción. Esta desigualdad puede escribirse, de una forma más sencilla, como

$$\big(1 + (2k-1) \cdot (2k-3) \cdots 3 \cdot 1\big) \leq 2^{k-1} \cdot k \cdot (k-1) \cdots 2 \cdot 1 = (2k) \cdot (2k-2) \cdots 4 \cdot 2,$$

y, por tanto, lo que debemos probar es que el producto de los k primeros números pares es mayor que el producto de los k primeros impares más 1. Para $k = 2$ se tiene que $1 + (4-1) \cdot 1 = 4 \leq 8 = (2 \cdot 2) \cdot 2$ luego se cumple la desigualdad. Para $k = n$ se supone la "hipótesis de inducción"

$$\big(1 + (2n-1) \cdot (2n-3) \cdots 3 \cdot 1\big) \leq (2n) \cdot (2n-2) \cdots 4 \cdot 2, \qquad (4.8)$$

y, para completar la inducción, se debe probar que para $k = n + 1$ se tiene

$$\big(1 + (2n+1)\cdot(2n-1)\cdots 3\cdot 1\big) \leq (2n+2)\cdot(2n)\cdots 4\cdot 2.$$

Sacando factor común al término $(2n+1)$ en el lado izquierdo de la desigualdad se ve que

$$\begin{aligned}(2n+1)\left(\frac{1}{2n+1} + (2n-1)\cdot(2n-3)\cdots 3\cdot 1\right) \leq & (2n+1)\Big(1 + (2n-1)\cdot(2n-3)\cdots 3\cdot 1\Big)\\ \leq & (2n+1)\cdot(2n)\cdot(2n-2)\cdots 4\cdot 2\\ \leq & (2n+2)\cdot(2n)\cdot(2n-2)\cdots 4\cdot 2,\end{aligned}$$

que es justo lo que se quería probar. En el proceso se ha usado que $\frac{1}{2n+1} \leq 1$, que $2n+1 \leq 2n+2$, y la hipótesis de inducción de (4.8).

Por lo tanto, ya se ha probado que

$$\big|E(Y_l^k)\big| = \left| E\left(\frac{X_l^2 - 1}{2}\right)^k \right| \leq \frac{k!}{2},$$

para todo $k \geq 2$, y se está en condiciones de acotar (4.6) con el uso de la desigualdad de Bernstein del Lema 4.1, de tal forma que

$$\begin{aligned}P\left[\,\big|\|X\| - \sqrt{d}\,\big| \geq \varepsilon\,\right] \leq & P\left[\,\Big|\frac{1}{2}(X_1^2 - 1) + \cdots + \frac{1}{2}(X_d^2 - 1)\Big| \geq \frac{\varepsilon\sqrt{d}}{2}\,\right]\\ \leq & \, 2\, e^{-\frac{1}{4}\min\left(\frac{(\varepsilon\sqrt{d}/2)^2}{d}, \frac{\varepsilon\sqrt{d}}{2}\right)} \leq\\ \leq & 2\, e^{-\frac{1}{4}\min\left(\frac{\varepsilon^2}{4}, \frac{\varepsilon^2}{2}\right)} = 2\, e^{-\frac{1}{16}\varepsilon^2}, \qquad (4.9)\end{aligned}$$

donde se usa que $\frac{\varepsilon\sqrt{d}}{2} \geq \frac{\varepsilon^2}{2}$ para todo d y para todo $\varepsilon \leq \sqrt{d}$. □

Una de las cuestiones a tener en cuenta para aplicar el Teorema 4.2 es que la cantidad ε considerada. El teorema es trivial si la cota proporcionada a la derecha es menor o igual que 0. Este caso se alcanza cuando $\varepsilon = 4 \cdot \sqrt{\log 2} \approx 3.3$, mientras que la cota $1 - 2\, e^{-\frac{1}{16}\varepsilon^2}$ se encuentra entre 0 y 1 si ε es mayor que esta cantidad, que es justo cuando tiene interés el Teorema 4.2. Esto remarca la concentración de la distribución normal multivariante alrededor de su radio $\sqrt{d}$ y esta concentración se hace especialmente interesante cuando d es muy grande en comparación con ε, ya que el teorema proporciona una cota para la probabilidad de que $\|X\|$ se aleje de su esperanza $\sqrt{d}$ más que ε. El hecho de que no se pueda elegir ε convenientemente para cada dimensión d significa que la anchura del anillo gaussiano en la Figura 4.1 es constante en contraposición a lo que sucedía en el Capítulo 3, donde se podía elegir ε convenientemente con la dimensión, de manera que los resultados que daban los teoremas mejoraban cuando d aumentaba, disminuyendo ε lo suficiente. El siguiente corolario es un breve resultado que se demuestra con el Teorema 4.2 y será necesario en posteriores demostraciones.

Corolario 4.3. *Sean (Ω, σ, P) un espacio probabilístico, $X : \Omega \to \mathbb{R}^d$ una variable aleatoria con distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ y $0 \leq \varepsilon \leq d$. Entonces se cumple que:*

$$P\left[\,\big|\|X\|^2 - d\,\big| \leq \varepsilon\,\right] \geq 1 - 2\, e^{-\frac{1}{8}\min\left(\frac{\varepsilon^2}{2d}, \varepsilon\right)}$$

Demostración. La demostración es análoga a la del Teorema 4.2 ya que se puede escribir

$$P\left[\,\left|\|X\|^2-d\,\right|\geq\varepsilon\,\right]=P\left[\left|\frac{1}{2}(X_1^2-1)+\cdots+\frac{1}{2}(X_d^2-1)\right|\geq\frac{\varepsilon}{2}\right]$$

y, como se probó en la demostración del Teorema 4.2, las variables aleatorias $Y_i=\frac{1}{2}(X_i^2-1)$ cumplen las hipótesis del Lema 4.1 y, consecuentemente, se tiene

$$P\left[\,\left|\|X\|^2-d\,\right|\geq\varepsilon\,\right]=P\left[\,\left|\frac{1}{2}(X_1^2-1)+\cdots+\frac{1}{2}(X_d^2-1)\,\right|\geq\frac{\varepsilon}{2}\right]\leq$$
$$\leq 2\,e^{-\frac{1}{4}\min(\frac{\varepsilon^2}{4d},\frac{\varepsilon}{2})}=2\,e^{-\frac{1}{8}\min(\frac{\varepsilon^2}{2d},\varepsilon)}$$

□

Siguiendo el mismo desarrollo realizado en el Capítulo 3, a continuación, se considerará el ángulo entre dos variables aleatorias independientes, ambas, con distribución $\mathcal{N}(\mathbf{0}_d,I_d)$. Recuérdese que, en virtud de lo probado en el Teorema 2.6, dos variables aleatorias independientes distribuidas normalmente en $\mathbb{R}^d$ satisfacen que $E(\langle X,Y\rangle)=0$. El objetivo de los siguientes resultados será demostrar que $\langle X,Y\rangle$ está también próximo a 0 con alta probabilidad.

Teorema 4.4 (Teorema de Ortogonalidad Gaussiana)**.** *Sean (Ω,σ,P) un espacio probabilístico, y $X,Y:\Omega\to\mathbb{R}^d$ variables aleatorias con $X,Y\sim\mathcal{N}(\mathbf{0}_d,I_d)$ y $\varepsilon>0$. Entonces se cumple que:*

$$P\left[\left|\left\langle\frac{X}{\|X\|},\frac{Y}{\|Y\|}\right\rangle\right|\leq\varepsilon\right]\geq 1-\frac{2/\varepsilon+7}{\sqrt{d}}$$

(en esta formulación, se trabaja condicionalmente a que $\|X\|$ e $\|Y\|$ son distintos de 0, evento que sucede con probabilidad 1 por tratar con la distribución continua normal multivariante).

Demostración. Como se ha comentado, se puede considerar directamente que X e Y no toman el valor 0. Además, usando la Ley de Probabilidad Total sobre la variable aleatoria Y se tiene que

$$P\left[\left|\left\langle\frac{X}{\|X\|},\frac{Y}{\|Y\|}\right\rangle\right|\leq\varepsilon\right]=E_Y\left[P_X\left(\left|\left\langle\frac{X}{\|X\|},\frac{Y}{\|Y\|}\right\rangle\right|\leq\varepsilon\,\middle|\,Y=\mathbf{y}\right)\right].\tag{4.10}$$

Se puede ver que la probabilidad condicionada es la misma para todos los puntos $\mathbf{y}\in\mathbb{R}^d$ con $\mathbf{y}\neq\mathbf{0}_d$, de tal manera que entonces la esperanza de la probabilidad condicionada de (4.10) será simplemente la probabilidad condicionada en un punto $\mathbf{u}=\frac{\mathbf{y}}{\|\mathbf{y}\|}\in\mathbb{R}^d$. Para ello, dado cualquier otro vector unitario $\mathbf{v}\in\mathbb{R}^d$ se sabe que, por la estructura de espacio vectorial de $\mathbb{R}^d$, existe una matriz R ortogonal de rotación de tamaño $d\times d$ tal que $R\mathbf{v}=\mathbf{u}$ y

$$\left\langle\frac{X}{\|X\|},\mathbf{u}\right\rangle=\left\langle\frac{X}{\|X\|},R\mathbf{v}\right\rangle=\left\langle R^\top\frac{X}{\|X\|},\mathbf{v}\right\rangle.\tag{4.11}$$

Pero la distribución normal multivariante es invariante por rotaciones (véase la Proposición A.9). Entonces $R^\top X\sim X$, y $R^\top\frac{X}{\|X\|}$ tiene la misma distribución que $\frac{X}{\|X\|}$. Por tanto, usando (4.11) se tiene

$$P_X\left(\left|\left\langle\frac{X}{\|X\|},\mathbf{u}\right\rangle\right|\leq\varepsilon\right)=P_X\left(\left|\left\langle\frac{X}{\|X\|},\mathbf{v}\right\rangle\right|\leq\varepsilon\right)\tag{4.12}$$

para cualesquiera vectores $\mathbf{u}$ y $\mathbf{v}$ unitarios. Como la probabilidad condicionada es la misma para cualquier $\mathbf{y} \neq \mathbf{0}_d$, el promedio sobre todos los posibles valores es igual a la probabilidad para un solo vector unitario $\mathbf{v}$ fijo, luego

$$P\left[\left|\left\langle \frac{X}{\|X\|}, \frac{Y}{\|Y\|}\right\rangle\right| \leq \varepsilon\right] = E_Y\left[P_X\left(\left|\left\langle \frac{X}{\|X\|}, \frac{Y}{\|Y\|}\right\rangle\right| \leq \varepsilon \,\middle|\, Y = \mathbf{y}\right)\right] = \\ = P_X\left(\left|\left\langle \frac{X}{\|X\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|}\right\rangle\right| \leq \varepsilon\right). \tag{4.13}$$

Si $\lambda_l = \frac{y_l}{\|\mathbf{y}\|}$, para $l = 1, \ldots, d$, se tiene que $\lambda_1^2 + \cdots + \lambda_d^2 = 1$, y se puede definir una nueva variable aleatoria $U : \Omega \to \mathbb{R}$ dada por

$$U = \left\langle X, \frac{\mathbf{y}}{\|\mathbf{y}\|}\right\rangle = \sum_{l=1}^{d} \frac{y_l}{\|\mathbf{y}\|} X_l = \sum_{l=l}^{d} \lambda_l X_l.$$

Como las variables $X_l \sim \mathcal{N}(0,1)$ entonces $\lambda_l X_l \sim \mathcal{N}(0, \lambda_l^2)$ y $U \sim \mathcal{N}(0, \lambda_1^2 + \cdots + \lambda_d^2) \sim \mathcal{N}(0,1)$ (véase el Corolario A.4). Para calcular la probabilidad de (4.13), vamos a usar esta nueva variable aleatoria junto con la Ley de Probabilidad Total, usando el suceso $\|X\| \leq \frac{\sqrt{d}}{2}$ y su complementario, y sabiendo que ese suceso ocurre con baja probabilidad por el Teorema 4.2. Se tiene

$$\begin{aligned} P_X\left(\left|\left\langle \frac{X}{\|X\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|}\right\rangle\right| \leq \varepsilon\right) &= P_X\left(\left|\left\langle \frac{X}{\|X\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|}\right\rangle\right| \leq \varepsilon \,\middle|\, \|X\| \leq \frac{\sqrt{d}}{2}\right) \cdot P_X\left(\|X\| \leq \frac{\sqrt{d}}{2}\right) + \\ &+ P_X\left(\left|\left\langle \frac{X}{\|X\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|}\right\rangle\right| \leq \varepsilon \,\middle|\, \|X\| \geq \frac{\sqrt{d}}{2}\right) \cdot P_X\left(\|X\| \geq \frac{\sqrt{d}}{2}\right) \\ &\geq P_X\left(|U| \leq \varepsilon \cdot \|X\| \,\middle|\, \|X\| \geq \frac{\sqrt{d}}{2}\right) \cdot P_X\left(\left|\|X\| - \sqrt{d}\right| \geq \frac{\sqrt{d}}{2}\right) \\ &\geq P_X\left(|U| \leq \varepsilon \cdot \frac{\sqrt{d}}{2}\right) \cdot \left(1 - e^{-\frac{d}{4 \cdot 16}}\right), \end{aligned} \tag{4.14}$$

donde el último término es la cota proporcionado por el Teorema 4.2 con $\varepsilon = \sqrt{d}/2$. Para completar la demostración se debe tratar la última probabilidad que aparece en (4.14). Como $U \sim \mathcal{N}(0,1)$, se tiene

$$\begin{aligned} P_X\left(|U| \leq \varepsilon \cdot \frac{\sqrt{d}}{2}\right) &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{\varepsilon\sqrt{d}}{2}}^{\frac{\varepsilon\sqrt{d}}{2}} e^{-t^2/2}\, dt \;\geq 1 - \frac{2}{\sqrt{2\pi}} \int_{-\frac{\varepsilon\sqrt{d}}{2}}^{\infty} \frac{1}{t^2}\, dt \\ &= 1 - \frac{2}{\sqrt{2\pi}} \cdot \left(-\frac{1}{t}\Big|_{-\frac{\varepsilon\sqrt{d}}{2}}^{\infty}\right) \geq 1 - \frac{2}{\varepsilon\sqrt{d}}, \end{aligned} \tag{4.15}$$

donde se ha usado que $e^x \geq 1 + \frac{1}{x}$ y que $\frac{2}{\sqrt{2\pi}} \leq 1$ con el fin de acotar inferiormente esa integral y, después, mejorar la cota tomando un límite de integración mayor para la integral que se está restando.

Finalmente, juntando (4.13), (4.14) y (4.15) se tiene

$$P\left[\left|\left\langle \frac{X}{\|X\|}, \frac{Y}{\|Y\|}\right\rangle\right| \leq \varepsilon\right] \geq \left(1 - \frac{2}{\varepsilon\sqrt{d}}\right) \cdot \left(1 - e^{-\frac{d}{4\cdot 16}}\right) = 1 - \frac{2}{\varepsilon\sqrt{d}} - e^{-\frac{d}{4\cdot 16}} + \frac{4e^{-\frac{d}{64}}}{\varepsilon\sqrt{d}} \geq$$
$$\geq 1 - \frac{2}{\varepsilon\sqrt{d}} - \frac{7}{\sqrt{d}},$$

que coincide con lo que se deseaba probar. □

Como consideración adicional, el Teorema 4.4 de Ortogonalidad Gaussiana demuestra que, con alta probabilidad, dos variables aleatorias independientes con distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ son prácticamente ortogonales en alta dimensión mediante una cota inferior para la probabilidad de que $\langle \frac{X}{\|X\|}, \frac{Y}{\|Y\|} \rangle$ se aleje en más de ε de 0, y esa acotación es dependiente tanto de $\sqrt{d}$ como de ε. Esta cota es diferente a la que obteníamos en el Teorema 4.2 del Anillo Gaussiano, de tal manera que aquí sí que es posible elegir un ε más pequeño conforme se aumente la dimensión d y el teorema aún proporciona resultados útiles, lo cual remarca la hipótesis de que cualquier par de vectores unitarios generados desde una distribución normal multivariante son esencialmente ortogonales en dimensiones altas. Este resultado extiende lo ya visto al tratar el valor esperado del producto escalar en la Proposición 2.6.

Por último, se considera la distancia entre dos variables aleatorias independientes con distribución $\mathcal{N}(\mathbf{0}_d, I_d)$. Intuitivamente, por lo probado en los teoremas 4.2 y 4.4, se sabe que, con alta probabilidad, $\|X\| \approx \sqrt{d}$, $\|Y\| \approx \sqrt{d}$ y los vectores X e Y son prácticamente ortogonales. Por tanto, usando el Teorema de Pitágoras, se puede esperar que $\|X - Y\| \approx \sqrt{2d}$, tal y como se ilustra en la Figura 2.2.

Otra consecuencia de estos resultados es el proporcionado en el Teorema 4.5 relativo a $\|X - Y\|$

Teorema 4.5. *Sean (Ω, σ, P) un espacio probabilístico, $X, Y : \Omega \to \mathbb{R}^d$ variables aleatorias independientes con $X, Y \sim \mathcal{N}(\mathbf{0}_d, I_d)$ y $0 \leq \varepsilon \leq \sqrt{d}$. Entonces se cumple que:*

$$P\left[\left|\|X - Y\| - \sqrt{2d}\right| \leq \varepsilon\right] \geq 1 - 4e^{-\varepsilon^2/72} - \frac{6\sqrt{2}}{\varepsilon d} - \frac{7}{\sqrt{d}}.$$

Demostración. Para probar este resultado se siguen las ideas de las demostraciones del Teorema 4.2 y del Teorema 3.10. Como $\|X - Y\| + \sqrt{2d} \geq \sqrt{2d}$, se tiene que $\Big|\|X - Y\|^2 -$

$2d\Big| \geq \varepsilon\sqrt{2d}$ si $\Big|\|X-Y\| - \sqrt{2d}\Big| \geq \varepsilon$. Por tanto,

$$\begin{aligned}
P\left[\Big|\|X-Y\| - \sqrt{2d}\Big| \leq \varepsilon\right] &= 1 - P\left[\Big|\|X-Y\| - \sqrt{2d}\Big| \geq \varepsilon\right] = \\
&= 1 - P\left[\Big|\|X-Y\| - \sqrt{2d}\Big| \cdot \left(\|X-Y\| + \sqrt{2d}\right) \geq \varepsilon\sqrt{2d}\right] = \\
&= 1 - P\left[\Big|\|X-Y\|^2 - 2d\Big| \geq \varepsilon\sqrt{2d}\right] = \\
&= P\left[-\varepsilon\sqrt{2d} \leq \|X-Y\|^2 - 2d \leq \varepsilon\sqrt{2d}\right] = \\
&= P\left[-\varepsilon\sqrt{2d} \leq \|X\|^2 + \|Y\|^2 - 2\langle X,Y\rangle - 2d \leq \varepsilon\sqrt{2d}\right] \geq \\
&\geq P\left[\Big|\|X\|^2 - d\Big| \leq \frac{\varepsilon\sqrt{2d}}{3}\right]^2 P\left[\Big|\langle X,Y\rangle\Big| \leq \frac{\varepsilon\sqrt{2d}}{6}\right].
\end{aligned} \tag{4.16}$$

Para la primera probabilidad en el último término, se puede usar el Corolario 4.3 que permite la acotación

$$P\left[\Big|\|X\|^2 - d\Big| \leq \frac{\varepsilon\sqrt{2d}}{3}\right] \geq 1 - 2e^{-\frac{1}{8}\min(\frac{\varepsilon^2}{9}, \frac{\varepsilon\sqrt{2d}}{3})} \geq 1 - 2e^{-\varepsilon^2/72}, \tag{4.17}$$

y, por el Teorema 4.4, se tiene también la acotación

$$P\left[\Big|\langle X,Y\rangle\Big| \leq \frac{\varepsilon\sqrt{2d}}{6}\right] \geq 1 - \frac{\frac{12}{\varepsilon\sqrt{2d}} + 7}{\sqrt{d}} = 1 - \frac{6\sqrt{2}}{\varepsilon d} - \frac{7}{\sqrt{d}}. \tag{4.18}$$

Juntando (4.16), (4.17) y (4.18) se llega a

$$\begin{aligned}
P\left[\Big|\|X-Y\| - \sqrt{2d}\Big| \leq \varepsilon\right] &\geq P\left[\Big|\|X\|^2 - d\Big| \leq \frac{\varepsilon\sqrt{2d}}{3}\right]^2 P\left[\Big|\langle X,Y\rangle\Big| \leq \frac{\varepsilon\sqrt{2d}}{6}\right] \geq \\
&\geq \left(1 - 2e^{\frac{-\varepsilon^2}{72}}\right)^2 \left(1 - \frac{6\sqrt{2}}{\varepsilon d} - \frac{7}{\sqrt{d}}\right) = \\
&= \left(1 + 4e^{\frac{-\varepsilon^2}{36}} - 4e^{\frac{-\varepsilon^2}{72}}\right) \cdot \left(1 - \frac{6\sqrt{2}}{\varepsilon d} - \frac{7}{\sqrt{d}}\right) \geq \\
&\geq 1 - 4e^{\frac{-\varepsilon^2}{36}} \cdot \left(\frac{6\sqrt{2}}{\varepsilon d} + \frac{7}{\sqrt{d}}\right) - 4e^{\frac{-\varepsilon^2}{72}} \\
&\geq 1 - \frac{6\sqrt{2}}{\varepsilon d} - \frac{7}{\sqrt{d}} - 4e^{\frac{-\varepsilon^2}{72}},
\end{aligned}$$

ya que los términos positivos se pueden acotar inferormente por 0 mientras que $4e^{\frac{-\varepsilon^2}{36}} \geq 1$. □

La interacción de los Teoremas del Anillo Gaussiano y de la Ortogonalidad Gaussiana (teoremas 4.2 y 4.4) en la demostración del Teorema 4.5 provoca que la cota obtenida tenga

características propias de las cotas propias de ambos teoremas. Es decir, en el Teorema del Anillo Gaussiano se consigue una cota con términos únicamente dependientes del parámetro ε, lo cual esta indicando la presencia de otro "anillo" en la distribución de la variable $\|X - Y\|$. La aparición también términos dependientes de ε y d sugiere que la cota final que se puede alcanzar es mejorable para cada dimensión d, según el ε que sea elegido. Como cabe esperar, la situación óptima se alcanza para ε suficientemente pequeño en comparación con d pero, por supuesto, respetando un valor mínimo de ε que evite que el término $4e^{-\varepsilon^2/72}$ convierta la acotación global en una cantidad negativa, lo que inutilizaría el interés del resultado.

Capítulo 5

Teorema de Johnson-Lindestrauss

5.1. Enunciado y prueba del teorema de Johnson-Lindestrauss

En los capítulos anteriores se han presentado algunas características no directamente esperables de las distribuciones más importantes en Estadística cuando se trabaja en espacios de alta dimensión. Sin embargo, algunos fenómenos, como la concentración de medida o la ortogonalidad de vectores, pueden resultar de utilidad en Análisis de Datos. En este capítulo presentaremos el Teorema de Johnson-Lindestrauss, que permitirá conservar las buenas propiedades geométricas de los conjuntos de datos a la vez que se proyectan a un espacio de dimensión menor donde las técnicas de Análisis de Datos clásicas se vuelven más fiables. A su vez, se analizarán algunas aplicaciones prácticas de este resultado y se proporcionarán simulaciones que apoyen las conclusiones obtenidas.

En primer lugar, se introducirá el concepto de proyecciones aleatorias. Sea (Ω, σ, P) un espacio probabilístico, en este capítulo se considerarán matrices aleatorias del tipo

$$U : \Omega \to \mathbb{R}^{k\times d}, \quad U = \begin{bmatrix} U_{11} & \cdots & U_{1d} \\ \vdots & & \vdots \\ U_{k1} & \cdots & U_{kd} \end{bmatrix},$$

con $U_{ij} \sim \mathcal{N}(0,1)$, $i = 1, \ldots, k$ y $j = 1, \ldots, d$, variables aleatorias normales unidimensionales independientes. Esta matriz aleatoria la será denotada por $U \sim \mathcal{N}(\mathbf{0}_{k\times d}, 1)$.

Algunas propiedades que se pueden extraer de esta matriz aleatoria están basadas en el comportamiento de sus filas como vectores normales multivariantes. Si se denotan las filas de U como $U_i = [U_{i1}, \ldots, U_{id}]^T$, para $i = 1, \ldots, k$, entonces todas ellas son vectores aleatorios independientes con distribución $\mathcal{N}(\mathbf{0}_d, I_d)$ (ver Proposición A.8) y se pueden explotar las propiedades vistas en la Sección 2.1 del Capítulo 2 y, a su vez, los resultados del Capítulo 4, cuando d sea elevado. Por ejemplo, si usamos el Teorema 4.4 de Ortogonalidad Gaussiana, se tiene que las filas de la matriz U normalizadas son "aproximadamente" ortogonales unas a otras. Entonces, normalizando la matriz aleatoria U, se obtendría una matriz aleatoria V tal que $VV^\top \approx I_k$. Esta propiedad recuerda a la propiedad de ortogonalidad en las matrices de proyección usadas en Álgebra Lineal. Una de las aplicaciones más conocidas de las proyecciones es que permiten transformar conjuntos de datos en $\mathbb{R}^d$ en el subespacio de $\mathbb{R}^k$ generado por las k filas de la matriz ortogonal de proyección. Se

debe tener en cuenta que si $k < d$ entonces la proyección es una aplicación lineal, que cuenta con esta buena propiedad de ortonormalidad que evita la "duplicidad" de información en los datos proyectados, pero que no necesariamente conservará las distancias entre puntos del espacio original $\mathbb{R}^d$. Esta conservación de las distancias originales sería una propiedad muy útil en contextos de clasificación, tanto supervisada como no supervisada, por ejemplo, en métodos como los m vecinos más próximos. El objetivo de este capítulo será probar que las proyecciones basadas en matrices aleatorias del tipo $U \sim \mathcal{N}(\mathbf{0}_{k\times d}, 1)$, además de tener la característica de la ortonormalidad, de forma aproximada, mantienen también, de forma aproximada, las distancias entre datos en el espacio original.

Teorema 5.1 (Teorema de la Proyección aleatoria)**.** *Sea (Ω, σ, P) un espacio probabilístico y $k < d$. Se considera $U : \Omega \to \mathbb{R}^{k\times d}$ una matriz aleatoria con $U \sim \mathcal{N}(\mathbf{0}_{k\times d}, 1)$. Entonces, para todo $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$ y $0 < \varepsilon \leq 1$ se tiene*

$$P\left[\left|\|U\mathbf{x}\| - \sqrt{k}\|\mathbf{x}\|\right| \leq \varepsilon\sqrt{k}\|\mathbf{x}\|\right] \geq 1 - 2e^{-k\varepsilon^2/16}.$$

Demostración. En primer lugar, se debe tener en cuenta que el producto matriz por vector de U por $\mathbf{x} \in \mathbb{R}^d$ da como resultado un vector en $\mathbb{R}^k$ que contiene los productos escalares entre cada vector fila de la matriz U, denotados por U_i, con el vector $\mathbf{x}$. Es decir,

$$U\mathbf{x} = \left[\langle U_1, \mathbf{x}\rangle, \ldots, \langle U_k, \mathbf{x}\rangle\right]^T.$$

Ahora, fijado $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$, se puede considerar el vector $U\frac{\mathbf{x}}{\|\mathbf{x}\|}$, cuya coordenada i-ésima es

$$\left(U\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)_i = \left\langle U_i, \frac{\mathbf{x}}{\|\mathbf{x}\|}\right\rangle = \sum_{j=1}^{d} U_{ij}\frac{x_i}{\|\mathbf{x}\|}.$$

Usando que $U_{ij} \sim \mathcal{N}(0,1)$ para todo (i,j), se tiene que $\left(U\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)_i \sim \mathcal{N}(0,1)$ (ver Proposición A.4 luego, junto a la independencia de los U_{ij}, se tiene $U\frac{\mathbf{x}}{\|\mathbf{x}\|} \sim \mathcal{N}(\mathbf{0}_d, I_d)$ (ver Proposición A.8). Se puede utilizar ahora el Teorema 4.2 del Anillo Gaussiano para tener

$$P\left[\left|\|U\mathbf{x}\| - \sqrt{k}\|\mathbf{x}\|\right| \leq \varepsilon\sqrt{k}\cdot\|\mathbf{x}\|\right] = P\left[\left|\left\|U\frac{\mathbf{x}}{\|\mathbf{x}\|}\right\| - \sqrt{k}\right| \leq \varepsilon\sqrt{k}\right] \geq 1 - 2e^{\varepsilon^2 k/16}.$$

En esta acotación se ha aplicado dicho teorema con $d = k$ y $\varepsilon = \varepsilon\sqrt{k}$. Nótese que se cumple la hipótesis $0 < \varepsilon\sqrt{k} \leq \sqrt{k}$ dado que $0 < \varepsilon \leq 1$.

□

Puesto que $\sqrt{k}$ es fijo a lo largo de todo la prueba del Teorema 5.1, una reformulación más habitual de este resultado es

$$P\left[\left|\left\|\frac{1}{\sqrt{k}}U\frac{\mathbf{x}}{\|\mathbf{x}\|}\right\| - 1\right| \leq \varepsilon\right] \geq 1 - 2\,e^{\varepsilon^2 k/16}. \tag{5.1}$$

Definición 5.2. *Sea (Ω, σ, P) un espacio probabilístico y $k < d$ y $U : \Omega \to \mathbb{R}^{k\times d}$ una matriz aleatoria con $U \sim \mathcal{N}(\mathbf{0}_{k\times d}, 1)$. Si $\omega \in \Omega$, se denomina proyección de Johnson-Lindestrauss a la aplicación $T_{U(\omega)} : \mathbb{R}^d \to \mathbb{R}^k$ dada por*

$$T_{U(\omega)} = \frac{1}{\sqrt{k}}U(\omega).$$

Estrictamente hablando, no puede hablarse de $T_{U(\omega)}$ como una "proyección" en el sentido de Álgebra Lineal porque no necesariamente son matrices ortonormales. Con esta notación, otra reformulación del Teorema 5.1 es que para todo $\mathbf{x} \in \mathbb{R}^d \backslash \{\mathbf{0}_d\}$, si $0 < \varepsilon \leq 1$ y $U \sim \mathcal{N}(\mathbf{0}_{k\times d}, 1)$, se tiene que

$$P\left[\left|\|T_U\mathbf{x}\| - \|\mathbf{x}\|\right| \leq \varepsilon\|\mathbf{x}\|\right] \geq 1 - 2e^{\varepsilon^2 k/16}, \tag{5.2}$$

o, dicho de otra forma, se tiene que con alta probabilidad

$$\frac{\|T_U x\|}{\|x\|} \approx 1.$$

Es decir, en cierta manera, la proyección de Jonhson-Lindestrauss mantiene prácticamente invariante la norma de cada $\mathbf{x} \in \mathbb{R}^p$.

El siguiente resultado, el Teorema de Johnson-Lindestrauss, finalmente establece que es lo que sucede cuando consideramos las distancias entre pares de puntos tras usar la proyección T_U.

Teorema 5.3 (Teorema de Johnson-Lindestrauss)**.** *Sea (Ω, σ, P) un espacio probabilístico, $0 < \varepsilon < 1$ y $k \geq \frac{48}{\varepsilon^2}\log n$ y $U : \Omega \to \mathbb{R}^{k\times d}$ una matriz aleatoria con $U \sim \mathcal{N}(\mathbf{0}_{k\times d}, 1)$. Entonces, para cada cualquier conjunto de n puntos $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ se verifica que*

$$P\left[(1-\varepsilon)\cdot\|\mathbf{x}_i - \mathbf{x}_{i'}\| \leq \|T_U\mathbf{x}_i - T_U\mathbf{x}_{i'}\| \leq (1+\varepsilon)\cdot\|\mathbf{x}_i - \mathbf{x}_j\| \text{ para todo } i, i'\right] \geq 1 - \frac{1}{n}.$$

Demostración. Para cada $i \neq i'$ se define $\mathbf{x} = \mathbf{x}_i - \mathbf{x}_{i'}$ y se aplica el Teorema 5.1 para acotar

$$\begin{aligned}
P_{i,i'} &= P\left[(1-\varepsilon)\cdot\|\mathbf{x}_i - \mathbf{x}_{i'}\| \leq \|T_U\mathbf{x}_i - T_U\mathbf{x}_{i'}\| \leq (1+\varepsilon)\cdot\|\mathbf{x}_i - \mathbf{x}_{i'}\|\right] \\
&= P\left[(1-\varepsilon)\cdot\|\mathbf{x}\| \leq \|T_U\mathbf{x}\| \leq (1+\varepsilon)\cdot\|\mathbf{x}\|\right] \\
&= P\left[\left|\|T_U\mathbf{x}\| - \|\mathbf{x}\|\right| \leq \varepsilon\cdot\|\mathbf{x}\|\right] \\
&\geq 1 - 2e^{\varepsilon^2 k/16}.
\end{aligned} \tag{5.3}$$

Puesto que hay n puntos en el conjunto $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, existen $\binom{n}{2} \leq \frac{n^2}{2}$ pares distintos de puntos, y se puede desarrollar la probabilidad a acotar en función de las $P_{i,i'}$ (acotadas en 5.3) de la forma

$$\begin{aligned}
&P\Big[(1-\varepsilon)\cdot\|\mathbf{x}_i - \mathbf{x}_{i'}\| \leq \|T_U\mathbf{x}_i - T_U\mathbf{x}_{i'}\| \leq (1+\varepsilon)\cdot\|\mathbf{x}_i - \mathbf{x}_{i'}\| \text{ para todo } i, i'\Big] = \\
&= 1 - P\Big[\bigcup_{i,i'}\Big\{\Big|\|T_U\mathbf{x}_i - T_U\mathbf{x}_{i'}\| - \|\mathbf{x}_i - \mathbf{x}_{i'}\|\Big| > \varepsilon\cdot\|\mathbf{x}_i - \mathbf{x}_{i'}\|\Big\}\Big] = \\
&= 1 - \sum_{i\neq i'}(1 - P_{i,i'}) \geq 1 - \sum_{i\neq i'} 2e^{-k\varepsilon^2/16} \geq 1 - \frac{n^2}{2}2e^{-k\varepsilon^2/16} \geq \\
&\geq 1 - n^2\cdot e^{-\left(\frac{48}{\varepsilon^2}\log n\right)\varepsilon^2/16} \geq 1 - n^2\cdot e^{\log(n^{-3})} = 1 - \frac{1}{n}.
\end{aligned} \tag{5.4}$$

□

Nota 5.4. *Para aplicar el Teorema 5.3 (de Johnson-Lindestrauss) correctamente se debe interpretar la relación que existe entre las distintas constantes que aparecen en el enunciado: n tamaño de la muestra en consideración, d dimensión original del conjunto de puntos y k dimensión a la que se pretende proyectar los puntos. En particular, $k \geq \frac{48}{\varepsilon^2} \log n$ luego crece de forma logarítmica con respecto a n e inversamente proporcional al cuadrado de ε, lo cual quiere decir que conforme más grande sea el tamaño muestral n y más pequeño sea ε más grande será la dimensión k a la que se debe proyectar. La interacción entre n y ε en el k necesario es inversamente proporcional y, además, k crece más rápidamente con una disminución de ε que con el aumento del tamaño muestral n.*

La clave está en que la dimensión proyectada k no depende de la dimensión original d, sino solo de n y ε. El Teorema de Johnson–Lindenstrauss garantiza que, al proyectar un conjunto de n puntos desde un espacio de alta dimensión a un subespacio euclídeo de dimensión $k \geq \frac{48}{\varepsilon^2} \log n$, es posible conservar las distancias entre todos los pares de puntos con un error de a lo sumo ε, y esto implica que la estructura geométrica del conjunto se preserva casi idénticamente por la proyección de Johnson-Lindestrauss T_U. Nótese que si d es muy pequeña en comparación a n y ε es lo suficientemente pequeño, se puede conseguir una dimensión k de proyección que sea superior a la dimensión d original lo carecería de interés a nivel práctico. El Teorema 5.3 proporciona resultados relevantes en el caso de que d sea de una dimensión moderadamente grande en comparación con k que es del orden de $\log n$. Esto nos permite incluso considerar conjuntos de puntos más grandes que la dimensión del espacio d que se podrán proyectar a un espacio de dimensión menor sin perder su geometría.

Además, los resultados del Teorema 5.3 son totalmente independientes del conjunto original de puntos $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ o de la distribución que los generó (si la tuviesen). De hecho, únicamente son dependientes del tamaño del conjunto sobre el que se quieran aplicar. Por ejemplo, considerando un conjunto de 10 puntos en dimensión $d = 10^6$, entonces, por el Teorema 5.3 (de Johnson-Lindestrauss) se puede proyectar los datos a un espacio de dimensión k menor tal que las distancias entre los proyectados y las distancias entre los puntos originales difieren en menos de $\varepsilon = 0.15$ con una probabilidad mayor que $1 - 1/10 = 0.99$. La dimensión del espacio de proyección debería ser mayor a $\frac{48}{(0.15)^2} \log 10 = 4912.182$, es decir, bastaría elegir $k = 4913$, lo cual supondría una reducción del espacio original a otro de dimensión unas 203 veces más pequeña, conservando las distancias con fidelidad. La dimensión de este nuevo espacio de proyección, pese a seguir siendo alta, se reduce considerablemente haciendo más manejable el conjunto de datos original.

Nota 5.5. *El enunciado propuesto en el Teorema 5.3 es una versión reducida de la formulación original del teorema propuesta en Johnson y Lindenstrauss (1984) y se puede generalizar cuando la matriz U del Teorema 5.3 tiene entradas que son variables aleatorias independientes, isotrópicas (los vectores fila de U, U_i, cumplen que su matriz de covarianzas $E(U_i U_i^\top)$ es I_d) y subgaussianas: las colas de las variables U_{ij} decrecen tan rápido como las colas de la distribución normal, es decir, cumplen $P(|U_{ij}| > t) \leq Ce^{-ct^2}$ para ciertas constantes C, c y para todo $t > 0$. Algunas distribuciones con estas propiedades son la normal, la uniforme en un intervalo centrado o las variables con distribución de Rademacher que toman los valores $\{1, -1\}$ con probabilidad $1/2$. Esta demostración se puede encontrar en Vershynin (2018) y se basa en versiones más generales de la desigualdad de Bernstein para distribuciones subgaussianas.*

Pese a que la proyección al espacio de dimensión k en el ejemplo traslada los datos a una dimensión considerablemente inferior, esta aún se puede mejorar. Es por ello que en trabajos como Dasgupta y Gupta (2003), se consigue mejorar la acotación inferior de k demostrando que basta elegir

$$k \geq \frac{4 \log n}{\varepsilon^2/2 - \varepsilon^3/3}, \tag{5.5}$$

para que se cumpla el Teorema 5.3. Volviendo al ejemplo anterior y, usando esta nueva acotación de k, bastaría elegir k mayor o igual que 909.66 lo cual consigue reducir la dimensión de forma aún más considerable.

Para comprobar experimentalmente los resultados del Teorema 5.3 se ha simulado en `R` una muestra de tamaño $n = 150$ de la distribución $X \sim \mathcal{U}([0,1]^d)$ con $d = 100.000$. Se ha querido comprobar la distorsión promedio entre las distancias proyectadas y las originales para una muestra de 500 pares distintos de observaciones. Para ello se ha escogido $\varepsilon = 0.1$ y $k = 4295$ según (5.5) y se ha generado una matriz aleatoria de tamaño $k \times d$ con entradas gaussianas. Lo que se representa en la Figura 5.1 es el ratio entre los 500 pares de distancias proyectadas y originales que se puede comprobar que es aproximadamente igual a 1, como establece el Teorema 5.3. Además, el Teorema de Johnson-Lindestrauss proporciona una cota inferior para esta probabilidad de $1 - 1/n \approx 0.99$. Esto se ve representado en la Figura 5.1 mediante las líneas discontinuas horizontales $y = 1 \pm \varepsilon$ que encierran la región donde se encuentran los pares de individuos que cumplen el Teorema 5.3.

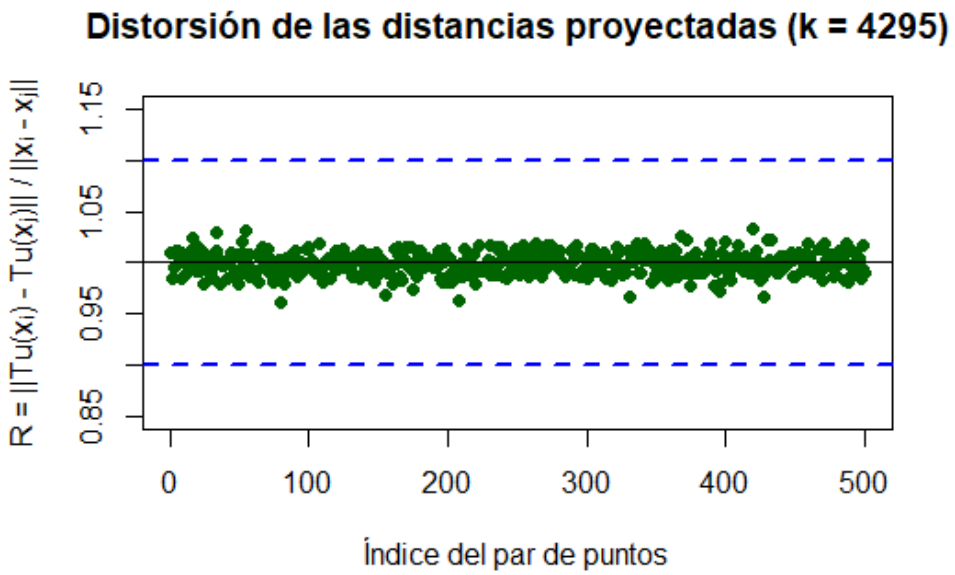

Figura 5.1: Distorsión de 500 pares de puntos de una muestra de 150 observaciones de la distribución uniforme en $[0,1]^d$ con $d = 100.000$

Para generalizar esta idea, se ha calculado el promedio de observaciones de esa misma muestra que cumplen el Teorema 5.3 escogiendo una muestra representativa de pares de observaciones. Además, se ha variado ε y se ha considerado el k dado por el Teorema 5.3 y el dado en (5.5) para probar con diferentes situaciones. Estos resultados se han recogido en

la Tabla 5.1 donde se puede observar cómo en todos los casos la estimación promedio del porcentaje de pares de puntos cuya distorsión está próxima a 1 con una tolerancia de ε es del 100 % lo cual confirma los resultados teóricos del Teorema 5.3 de forma experimental. Se ha considerado la k dada por el Teorema 5.3 para apreciar la diferencia en la dimensión de proyección de los datos entre esta y la dada en (5.5) y se ha considerado un ε mayor para no sobreexceder el límite de memoria al construir las matrices de proyección.

ε	k (con Teorema 5.3)	%	k (con cota de (5.5))	%
0.20	6013	100	1157	100
0.17	8323	100	1565	100
0.15	10690	100	1980	100

Tabla 5.1: Comprobación del Teorema de Johnson-Lindestrauss para distintos ε y k

Resumiendo, en esta sección se ha presentado el Teorema de Johnson-Lindestrauss y sus conclusiones que son de gran interés en Ciencia de Datos. El resultado está basado en el comportamiento de las matrices aleatorias como método de proyección de un conjunto de puntos a una dimensión menor conservando de forma aproximada las normas y las distancia entre los puntos. La dimensión del espacio de proyección es del orden del tamaño del conjunto de puntos e independiente de la dimensión original. Estas buenas características han hecho del Teorema de Johnson-Lindenstrauss una herramienta fundamental en técnicas de reducción de la dimensión, especialmente en contextos donde se manejan grandes volúmenes de datos. Su aplicación permite, en muchas ocasiones, mejorar la eficiencia computacional sin perder información geométrica relevante, lo que lo convierte en un pilar teórico clave en algoritmos de aprendizaje automático, recuperación de información y análisis de datos de alta dimensión.

5.2. Aplicaciones Teorema de Jonhson-Lindestrauss

En esta sección presentaremos algunas de las principales aplicaciones del Teorema de Jonhson-Lindestrauss que se aprovechan de las consideraciones anteriores para proporcionar mejores resultados en contextos de alta dimensionalidad.

5.2.1. Efecto de la dimensionalidad en m-vecinos más próximos

Uno de los problemas más extendidos en clasificación supervisada es la obtención de reglas que, basadas en un conjunto de datos de entrenamiento $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ donde cada punto $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^\top$, con $x_{i1}, \ldots, x_{id}$ los valores que cada observación toma en d variables predictivas, tiene asignada una etiqueta $y_i \in \mathcal{Y}$, sean capaces de, dado un nuevo individuo $\boldsymbol{x} \in \mathbb{R}^d$, precedir su etiqueta. El método de clasificación supervisada más sencillo es la clasificación mediante los m-vecinos más próximos o m-NN (*nearest neighbors*). Este método asigna a $\boldsymbol{x}$ la etiqueta mayoritaria entre los m individuos en $\mathcal{X}$ más cercanos a $\boldsymbol{x}$ con la norma euclídea. Este método es ampliamente utilizado por su simplicidad y por su capacidad de adaptarse a características locales de los datos, siendo especialmente efectivo en problemas de clasificación supervisada donde no se presupone una estructura global o distribución generadora en los datos.

A pesar de su utilidad, el método m-NN se ve severamente afectado en espacios de alta dimensión. Por ejemplo, consideremos un conjunto de datos $\mathcal{X} \subset [0,1]^d$ con n puntos elegidos uniformemente en $[0,1]^d$ (es decir, observaciones generadas por una distribución $\mathcal{U}([0,1]^d)$, y sea $\boldsymbol{x} \in [0,1]^d$ un nuevo punto también elegido aleatoriamente. Como demostramos en la Sección 2.2, la distancia esperada al vecino más cercano tiende a aumentar con d. Este fenómeno tiene efectos prácticos devastadores puesto que los vecinos "más cercanos" dejan de ser informativos, lo que degrada el rendimiento de los clasificadores basados en distancias, como m-NN. Además, los algoritmos exactos de búsqueda requieren $O(nd)$ operaciones, lo cual resulta prohibitivo cuando tanto n como d son suficientemente elevados.

El Teorema 5.3 puede ser de utilidad en este problema. Por ejemplo, en el artículo Ailon y Chazelle (2009) se propone una solución eficiente a este problema que usa como argumento principal el Teorema de Johnson-Lindestrauss. Al proyectar los datos a un espacio de dimensión reducida $\mathbb{R}^k$ con $k = O(\varepsilon^{-2} \log n)$, se puede realizar la búsqueda de vecinos con un coste mucho menor, manteniendo un control probabilístico sobre la distorsión en las distancias. Esto permite implementar algoritmos de búsqueda de vecinos más cercanos aproximados con garantías teóricas de precisión y un coste computacional muy inferior. Bastará encontrar la etiqueta de la proyección del punto deseado $\boldsymbol{x}$ sobre el espacio $\mathbb{R}^k$ y, con alta probabilidad, esta será la etiqueta que le corresponderá a $\boldsymbol{x}$ en el espacio original.

5.2.2. Efecto de la dimensionalidad en m-medias

Dado un conjunto de puntos $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^d$, el objetivo del algoritmo de k-medias es encontrar una partición $\{C_1, \ldots, C_m\}$ de $\mathcal{X}$ (es decir, $C_1 \cup \cdots \cup C_m = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ y $C_j \cap C_{j'} = \emptyset$ para $j \neq j'$) que minimice la función objetivo

$$\min_{\{C_1,\ldots,C_m\}} \sum_{j=1}^{m} \sum_{\boldsymbol{x_i} \in C_j} \|x_i - \boldsymbol{\mu}_j\|^2,$$

donde $\boldsymbol{\mu}_j$ es el centroide del clúster C_j tal que

$$\boldsymbol{\mu}_j = \frac{1}{\#C_j} \sum_{\boldsymbol{x}_i \in C_j} \boldsymbol{x}_i,$$

con $\#C_j$ el cardinal del conjunto C_j. El algoritmo de m-medias es un algoritmo ampliamente utilizado en clasificación no supervisada o Análisis Clúster (Everitt et al. 2011) que pretende subdividir el conjunto original $\mathcal{X}$ en m subgrupos o clúster de observaciones con valores "similares" en las d variables analizadas (las d variables numéricas analizadas se represetan en el vector d-dimensional $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^\top$). La formulación teórica y el algoritmo típicamente utilizado para implementar m-medias se basa en el cálculo de las distancias euclídeas entre las observaciones y los centroides de los grupos que, en alta dimensión, ya sabemos que puede incurrir en problemas. Además, la alta dimensionalidad dificulta la interpretación de los clústeres dado que las distancias entre puntos pierden el carácter discriminatorio entre subgrupos, y que el costo computacional se vuelve prohibitivo.

En trabajos como Boutsidis, Zouzias y Drineas (2010) se demuestra cómo, con una combinación entre proyecciones con matrices aleatorias de entradas con distribución de Rademacher, se puede reducir la dimensión original d de los datos a una dimensión $k = O(k\,\varepsilon^2)$ para calcular las distancias $\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2$, preservando el coste óptimo y disminuyendo el coste computacional, mientras que el error que se comete en las aproximaciones es del orden de ε con alta probabilidad.

5.2.3. Efecto de la dimensionalidad en LDA para separación y estimación paramétrica de distribuciones normales

Los ejemplos anteriores presentan problemas basados en la "maldición" de la dimensionalidad debida al cálculo de distancias entre puntos cuando d es muy grande. Sin embargo, este no es el único enfoque de aplicación del Teorema de Johnson-Lindenstrauss. Por ejemplo, el Teorema 5.3 puede ayudar en situaciones como el LDA.

Si se considera un nuevo problema de clasificación supervisada con m clases (y un total de n observaciones) donde se supone que las clases están generadas según una distribución normal multivariante:

$$X \,|\, Y = j \sim \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j), \text{ para todo } j = 1, \ldots, m,$$

con Y una variable aleatoria de etiquetas que toma valores en $\{1, \ldots, m\}$ y donde, en principio, los valores $\boldsymbol{\mu}_j$ y Σ_j son desconocidos. La técnica de clasificación más sencilla en este caso es el LDA (Linear Discriminant Analysis) que asume $\Sigma_j = \Sigma$ común a todas las clases, y busca asignar un nuevo individuo $\boldsymbol{x}$ a la clase que maximiza la probabilidad a posteriori (también se supone la presencia de unas probabilidades "a priori" π_j conocidas o estimables). Lo que se consigue es separar los conjuntos de datos originales por medio de hiperplanos basados en la regla de Bayes y la distancia de Mahalanobis, maximizando la dispersión entre grupos y minimizando la dispersión en cada clase. Para ello es necesario calcular (y posteriormente invertir) la matriz de covarianzas estimador común, $\hat{\Sigma}$, mediante una estimación "plug-in" que se basa en sustituir los parámetros desconocidos de cada grupo por sus estimaciones muestrales. Estas técnicas son muy eficaces en baja dimensión, pero en escenarios donde n y d son muy elevados, la matriz de covarianzas estimada $\hat{\Sigma}$ suele ser singular, lo cual impide el cálculo de su inversa a la vez que eleva el coste computacional de su obtención. Bajo estas condiciones el modelo suele sobreajustar en la mayoría de los casos, produciendo errores de clasificación altos que imposibilitan la aplicación de esta técnica en contextos de alta dimensión.

En trabajos como Cannings y Samworth (2017) se propone un enfoque del problema utilizando proyecciones aleatorias de dimensión k con entradas subgaussianas, a la vez que se estima la matriz combinada $\hat{\Sigma}$ en el espacio proyectado, evitando conseguir una matriz no invertible. Los autores también demuestran cómo la tasa de mala clasificación debida a la clasificación en el espacio proyectado se aproxima a la tasa de mala clasificación de la regla de Bayes para un LDA clásico. Este enfoque mejora considerablemente la eficiencia y el coste computacional cuando la dimensión d y el número de observaciones n son elevadas.

Se puede ampliar el enfoque de este problema a la estimación de las matrices de dispersión Σ_j distintas. En artículos como Marzetta, Tucci y Simon (2010) se demuestra que las proyecciones mediante matrices aleatorias con entradas gaussianas ayudan a estimar los parámetros en un espacio reducido donde la estimación es más estable. Variando la matriz

aleatoria de proyección, se puede obtener una estimación promediada de los parámetros desconocidos en el espacio proyectado que se puede usar para reconstruir la estimación en el espacio original con las garantías de concentración similares a las expuestas en este trabajo. Estos métodos permiten reconstruir la distribución gaussiana original que ha generado los datos, evitando la singularidad de las matrices de covarianzas y mejorando las estimaciones "plug-in" para recuperar los parámetros poblacionales en contextos donde la dimensión d impide el uso de técnicas convencionales.

Capítulo 6

Conclusiones

A lo largo de este trabajo se han seguido tres líneas principales. En primer lugar, se ha estudiado el efecto de la alta dimensión sobre algunas de las distribuciones más comunes en Estadística. Esto conduce al conocido fenómeno del "espacio vacío" o "maldición de la dimensionalidad". Por otro lado, se ha observado que, bajo ciertas condiciones, esa misma alta dimensión puede resultar beneficiosa gracias a propiedades probabilísticas particulares, como la concentración de norma y propiedades de ortogonalidad. Estas propiedades conforman lo que algunos autores denominan la "bendición de la dimensionalidad". Estas buenas propiedades quedan ejemplificadas en el Teorema de Johnson-Lindenstrauss, que permite reducir la dimensión de los datos sin perder apenas información sobre las distancias euclídeas entre observaciones. Estos fenómenos explican, por ejemplo, por qué ciertas tareas como la separación de clases pueden llegar a simplificarse en espacios de gran dimensión bajo estructuras probabilísticas adecuadas. Es por ello que el Teorema de Johnson-Lindestrauss se ha propuesto como una alternativa para combatir algunos de los efectos indeseables de la alta dimensionalidad. A diferencia de técnicas como el Análisis de Componentes Principales, las proyecciones aleatorias justificadas por el Teorema de Johnson-Lindenstrauss no requieren supuestos estructurales sobre los datos ni estimación de parámetros complejos, y además presentan una menor carga computacional. Por ello, resulta de utilidad en algoritmos de clasificación, cuando se trabaja con datos que recogen información de numerosas variables, que son directamente dependientes de las distancias entre las observaciones, tanto métodos no paramétricos como los m-vecinos más próximos y las m medias o aquellos que dependen de la suposición de normalidad como el LDA.

Finalmente, el desarrollo del trabajo ha permitido no solo comprender mejor los desafíos de la alta dimensión, sino también identificar métodos y estrategias que pueden mitigar sus efectos. En particular, las simulaciones, los resultados analíticos y las aplicaciones discutidas demuestran que el conocimiento del comportamiento probabilístico en espacios de gran dimensión es esencial para el diseño de algoritmos estadísticos eficaces en Ciencia de Datos.

Nuevas direcciones de trabajo podrían ser explorar el uso del Teorema de Johnson-Lindestrauss, o formulaciones equivalentes, a metodologías como el SVM, la estimación de densidades, etc.

Apéndice A

Resultados adicionales sobre la distribución normal

A.1. Distribución normal univariante

En este anexo se presentarán los resultados sobre distribuciones normales univariantes que han sido utilizados en las demostraciones proporcionadas en este trabajo de fin de grado.

Definición A.1. *Una variable aleatoria X se dice que tiene una distribución normal univariante de media μ y varianza σ^2 si verifica:*

$$P(X \in A) = \frac{1}{\sigma\sqrt{2\pi}} \int_A \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big) dx, \text{ para } A \in \beta,$$

y se denota como $X \sim \mathcal{N}(\mu, \sigma^2)$.

Proposición A.2. *Si $X \sim \mathcal{N}(\mu, \sigma^2)$ y $a, b \in \mathbb{R}$ entonces $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.*

Proposición A.3. *Si $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ son variables aleatorias independientes, entonces $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_x^2 + \sigma_Y^2)$ y $X - Y \sim \mathcal{N}(\mu_X - \mu_Y, \sigma_x^2 + \sigma_Y^2)$.*

Corolario A.4. *Dadas $X_1, \ldots, X_d$ variables aleatorias mutuamente independientes tales que $X_l \sim \mathcal{N}(\mu_l, \sigma_l^2)$ para $l = 1, \ldots, d$. Entonces, toda combinación lineal $\lambda_1 X_1 + \cdots + \lambda_l X_l$ es una variable aleatoria con distribución normal de media $\lambda_1\mu_1 + \cdots + \lambda_l\mu_l$ y varianza $\lambda_1^2\sigma_1^2 + \cdots + \lambda_l^2\sigma_l^2$.*

Proposición A.5. *El momento de orden k de una variable aleatoria X de media μ se define como $E[(X-\mu)^k]$, y si $X \sim \mathcal{N}(0,1)$ se cumple que*

$$E(X^k) = \begin{cases} 0 & \text{si } k \text{ es impar} \\ \dfrac{(2n)!}{2^n n!} & \text{si } k = 2n \end{cases}$$

Demostración. Usando la densidad de X se tiene que:

$$E[X^k] = \int_{-\infty}^{\infty} x^k \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

En primer lugar, si k es impar, se observa que x^k es una función impar mientras que $\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ es par luego su producto es una función impar. Como la integral de una función impar sobre un intervalor simétrico respecto del origen es cero, se tiene

$$E[X^{2n+1}] = \int_{-\infty}^{\infty} x^{2n+1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0$$

para $k = 2n + 1$.

En el caso de que k sea par con $k = 2n$, usando integración por partes, se demuestra que:

$$E[Z^{2n}] = (2n-1)E[Z^{2n-2}].$$

Para probar esto último, sea

$$I_n = \int_{-\infty}^{\infty} x^{2n} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx,$$

e integramos por partes con $u = x^{2n-1}$ y $dv = x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dz$, para tener, $du = (2n-1)x^{2n-2}dx$ y $v = -\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ con lo que se prueba

$$I_n = (2n-1)I_{n-1}.$$

Por la formula de recurrencia tenemos que:

$$E[Z^{2n}] = (2n-1)(2n-3)\cdots 3 \cdot 1 = \frac{(2n)(2n-1)\cdots 2 \cdot 1}{(2n)(2n-2)\cdots 4 \cdot 2} = \frac{(2n)!}{2^n n!}$$

□

Proposición A.6. *La Kurtosis de una variable aleatoria X se define como*

$$\text{Kurtosis}(X) = \frac{E\left[(X-\mu)^4\right]}{\left(E\left[(X-\mu)^2\right]\right)^2},$$

y si $X \sim \mathcal{N}(\mu, \sigma^2)$ se cumple que $\text{Kurtosis}(X) = 3$.

Demostración. En primer lugar se considera $Z \sim N(0,1)$, de tal forma que $E(Z) = 0$ y $E(Z^2) = 1$. Para una distribución normal estándar se sabe que $E(Z^4) = 3$ (por la Proposición A.5) luego:

$$\text{Kurtosis}(Z) = \frac{E(Z^4)}{[E(Z^2)]^2} = 3.$$

En el caso general $X \sim \mathcal{N}(\mu, \sigma^2)$ se define la variable tipificada $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ y entonces se tiene

$$E[(X-\mu)^4] = E[(\sigma Z)^4] = \sigma^4 E[Z^4] = 3\sigma^4,$$

y, además,

$$E[(X-\mu)^2] = \sigma^2.$$

Por lo tanto,

$$\text{Kurtosis}(X) = \frac{E[(X-\mu)^4]}{(E[(X-\mu)^2])^2} = \frac{3\sigma^4}{(\sigma^2)^2} = 3$$

□

A.2. Distribución normal multivariante

Se proporcionan a continuación algunos resultados relativos a distribuciones normales multivariantes que han sido también usados en el desarrollo de este trabajo.

Recordamos aquí la definición de distribución normal multivariante, totalmente análoga a la expuesta en la Definición 2.1.

Definición A.7. *Sea (Ω, σ, P) un espacio probabilístico. Un vector aleatorio $X : \Omega \to \mathbb{R}^d$ se dice que tiene una distribución Gaussiana o normal d-variante si para todo vector $a \in \mathbb{R}^d$ se tiene que la variable aleatoria real $a^T X : \Omega \to \mathbb{R}$ sigue una distribución normal univariante ($a^T X$ serían todas las posibles combinaciones lineales de las coordenadas del vector aleatorio X). En particular, si asumimos un vector de medias $\boldsymbol{\mu} \in \mathbb{R}^d$ y una matriz de covarianzas Σ de tamaño $d \times d$ definida positiva, entonces se dice que $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ y admite la densidad*

$$P(X \in A) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \int_A \exp\Big(-\frac{1}{2}(x-\boldsymbol{\mu})^\top \Sigma^{-1}(x-\boldsymbol{\mu})\Big) dx, \text{ para } A \in \beta^d. \quad \text{(A.1)}$$

Se podría ver que $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ si y solo si $\Sigma^{-1/2}(X - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}_d, I_d)$. La densidad de $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ en la expresión A.7 surge del correspondiente cambio de variable y de la densidad de $\mathcal{N}(\mathbf{0}_d, I_d)$ que (usando la bien conocida densidad de la normal estándar y la factorización de la densidad conjunta que proporciona la independencia de sus componentes) es

$$f(x_1, \ldots, x_d) = \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-x_l^2/2} \text{ para todo } (x_1, \ldots, x_n)^\top \in \mathbb{R}^d.$$

Proposición A.8. *Sea (Ω, σ, P) un espacio probabilístico y $X : \Omega \to \mathbb{R}^d$ un vector aleatorio de componentes $X_1, \ldots, X_d$. Se tiene que, $X \sim N(\mathbf{0}_d, I_d)$ si y solo si $X_l \sim N(0, 1)$ para $l = 1, \ldots, d$ y $X_1, \ldots, X_d$ son independientes.*

La demostración de este resultado se basa, entre otros aspectos, en que la incorrelación es equivalente a la independencia para distribuciones normales. Una prueba rigurosa se puede encontrar en Gut (2005).

Proposición A.9. *Sea (Ω, σ, P) un espacio probabilístico, $X : \Omega \to \mathbb{R}^d$ un vector aleatorio tal que $X \sim N(\mathbf{0}_d, I_d)$ y $f : \mathbb{R}^d \to \mathbb{R}^d$ una rotación. Entonces, el vector aleatorio $f(X)$ sigue una distribución $X \sim N(\mathbf{0}_d, I_d)$. Es decir, las distribuciones $N(\mathbf{0}_d, I_d)$ son invariantes bajo rotaciones.*

Demostración. Si f es una rotación de $\mathbb{R}^d$ entonces existe una matriz ortogonal $R \in \mathbb{R}^{d\times d}$ tal que $f(X) = RX$, donde RX es el vector: $RX = (\langle R_1, X\rangle, \ldots, \langle R_d, X\rangle)$ y R_l es la fila l de la matriz R para $l = 1 \ldots, d$. Entonces RX es un vector de combinaciones lineales de las variables aleatorias $X_1, \ldots, X_d$ que cumplen que $X_l \sim \mathcal{N}(0, 1)$ luego es un vector aleatorio cuyas componentes $\langle R_l, X\rangle$ son variables aleatorias con distribución normal.

Además, como R es una matriz ortogonal, cumple que $R^\top R = I_d$ luego $E(f(X)) = E(RX) = R\,E(X) = R\,\mathbf{0}_d = \mathbf{0}_d$ y, a su vez, $\mathrm{Var}(f(X)) = \mathrm{Var}(RX) = R^\top \mathrm{Var}(X) R = R^\top I_d R = I_d$.

□

Apéndice B

Volumen de la bola unidad

El siguiente anexo presenta el desarrollo para el cálculo del volumen de la bola unitaria $B(\mathbf{0}_d, 1)$, para una dimensión d genérica.

Para calcular el volumen $\ell_d(B(\mathbf{0}_d, 1))$ en $\mathbb{R}^d$, se puede integrar en coordenadas cartesianas o esféricas generalizadas. En este caso es más sencillo integrar usando coordenadas esféricas. El volumen viene dado por:

$$\ell_d(B(\mathbf{0}_d, 1)) = \int_{\mathbb{S}^{d-1}} \int_{r=0}^{1} r^{d-1} dr \, d\Omega_d,$$

donde $\mathbb{S}^{d-1}$ es la esfera en dimensión d y $d\Omega_d$ el diferencial de superficie. Como esta es una integral de variables separadas,

$$\ell_d(B(\mathbf{0}_d, 1)) = \int_{\mathbb{S}^{d-1}} d\Omega_d \int_0^1 r^{d-1} dr = \frac{1}{d} \int_{\mathbb{S}^{d-1}} d\Omega_d = \frac{\ell_{d-1}(\mathbb{S}^{d-1})}{d}$$

Para calcular el volumen de $\ell_{d-1}(\mathbb{S}^{d-1})$ se parte de la integral

$$I(d) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-(x_1^2 + x_2^2 + \cdots + x_d^2)\right) dx_1 \cdots dx_d.$$

Esta integral $I(d)$ puede calcularse integrando tanto en coordenadas cartesianas como esféricas generalizadas. En coordenadas cartesianas, la integral se puede escribir como el producto de d integrales unidimensionales:

$$I(d) = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right)^d = (\sqrt{\pi})^d = \pi^{d/2}, \tag{B.1}$$

ya que la integral $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$. Por otro lado, en coordenadas esféricas, la integral $I(d)$ se convierte en

$$I(d) = \int_{\mathbb{S}^{d-1}} d\Omega_d \int_0^{\infty} e^{-r^2} r^{d-1} dr.$$

donde la integral de $\mathbb{S}^{d-1}$ sobre Ω_d da el área superficial $\ell_{d-1}(\mathbb{S}^{d-1})$, mientras que la integral sobre r se puede calcular mediante el cambio de variable $u = r^2$, $du = 2rdr$ obteniendo

$$\int_0^{\infty} e^{-r^2} r^{d-1} dr = \frac{1}{2} \int_0^{\infty} e^{-u} u^{(d/2)-1} du = \frac{1}{2} \Gamma\left(\frac{d}{2}\right). \tag{B.2}$$

Por lo tanto, si igualamos (B.1) y (B.2) se tiene

$$\pi^{d/2} = \ell_{d-1}(\mathbb{S}^{d-1}) \cdot \frac{1}{2}\Gamma\left(\frac{d}{2}\right)$$

Despejando el área de la esfera de dimensión d, se llega a

$$\ell_{d-1}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}.$$

Finalmente, se tiene

$$\ell_d(B(\mathbf{0}_d, 1)) = \frac{\ell_{d-1}(\mathbb{S}^{d-1})}{d},$$

y, consecuentemente,

$$\ell_d(B(\mathbf{0}_d, 1)) = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)},$$

donde también se ha usado la propiedad de la función Gamma $\Gamma(z+1) = z\,\Gamma(z)$ para todo z real.

Apéndice C

Programas

C.1. Código R usado para la generación de la Tabla 2.2

```r
dimensiones <- 1:15
radio <- 0.5

volumen_esfera <- function(n, R) {
  2*pi^(n / 2) / (n * gamma(n / 2)) * R^n
}

volumen <- round(sapply(dimensiones, volumen_esfera, R = radio)
   ,5)

tabla_resultado <- data.frame(
  Dimension = dimensiones,
  Volumen = volumen
)

print(tabla_resultado)
```

Listing C.1: Código R para calcular el volumen de la bola unidad en dimensión d

C.2. Código R usado para la generación de la Figura 2.1 y la Tabla 2.1

```r
library("MASS")
library(ggplot2)
library("doParallel")
library(foreach)

cl <- makeCluster(3)
registerDoParallel(cl)
```

```r

n <- 8000
i <- c(1, 10, 100, 500, 1000, 10000)
tamano_bloques <- 5000

res <- foreach(d = i, .combine = 'list', .packages =
c("MASS", "ggplot2", "scales")) %dopar% {

  if (d <= 1000) {
    X <- matrix(rnorm(n * d), nrow = n, ncol = d)
    Y <- sqrt(rowSums(X^2))
  } else {
    Y <- numeric(n)
    bloques <- seq(1, d, by = tamano_bloques)

    for (j in bloques) {
      end_j <- min(j + tamano_bloques - 1, d)
      X_bloque <- matrix(rnorm(n * (end_j - j + 1)), nrow = n,
      ncol = (end_j - j + 1))
      Y <- Y + rowSums(X_bloque^2)
      rm(X_bloque)
      gc()
    }
    Y <- sqrt(Y)
  }

  Media <- mean(Y)
  Varianza <- var(Y)

  p <- ggplot(data.frame(valores = Y), aes(x = abs(valores))) +
    geom_density(aes(y = ..count..), color = "black",
             linewidth = 0.3, fill = "gray", alpha = 0.4) +
    scale_x_continuous(limits = c(-2, max(Y) + 1)) +
    scale_y_continuous(labels = label_number(scale = 1e-3,
             suffix = "k")) +
    labs(title = paste0("d=", d), x = "||x||", y = "Frecuencia")
       +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5, size = 25))

  ggsave(paste0("grafico", match(d, i), ".png"), plot = p,
        width = 10, height = 8)

  rm(X, Y, p)
  cat(paste("Iteracion con d =", d, "completada.\n"))

  c(round(Media, 4), round(sqrt(d), 4), round(Varianza, 4))
}

```

```r
Tc <- data.frame(res, row.names = c("Media", "Sqrt(d), "Varianza"
   ))
colnames(Tc) <- paste0("d=", i)

stopCluster(cl)
Tc
```

Listing C.2: Código R para estimar las densidades de la norma de una normal multivariante.

C.3. Código R usado para la generación de la Figura 2.5 y la Tabla 2.3

```r
library("MASS")
library(ggplot2)
library("doParallel")
library(foreach)

cl <- makeCluster(3)
registerDoParallel(cl)

n <- 8000
i <- c(1, 10, 100, 500, 1000, 10000)
tamano_bloques <- 5000

res <- foreach(d = i, .combine = 'list', .packages = c("MASS",
                                          "ggplot2"))%dopar% {
  if (d <= 1000) {
    X <- matrix(runif(n * d), nrow = n, ncol = d)
    Y <- sqrt(rowSums((X-1/2)^2))
  } else {
    Y <- numeric(n)
    bloques <- seq(1, d, by = tamano_bloques)

    for (j in bloques) {
      end_j <- min(j + tamano_bloque - 1, d)
      X_bloque <- matrix(runif(n * (end_j - j + 1)), nrow = n,
                                           ncol = (end_j - j + 1))
      Y <- Y + rowSums((X_bloque-1/2)^2)
      rm(X_bloque)
      gc()
    }
    Y <- sqrt(Y)
  }

  Media <- mean(Y)
  Varianza <- var(Y)
```

```
35
36    p <- ggplot(data.frame(valores = Y), aes(x = abs(valores))) +
37      geom_density(aes(y = ..count..), color = "black",
38               linewidth = 0.3, fill = "gray", alpha = 0.4) +
39      scale_x_continuous(limits = c(-2, max(Y) + 1)) +
40      scale_y_continuous(labels = label_number(scale = 1e-3,
41               suffix = "k")) +
42      labs(title = paste0("d=", d), x = "||x||", y = "Frecuencia")
          +
43      theme_minimal() +
44      theme(plot.title = element_text(hjust = 0.5, size = 25))
45
46    ggsave(paste0("graficounif", match(d, i), ".png"),
47                  plot = p, width = 10, height = 8)
48
49    rm(X, Y, p)
50    cat(paste("Iteracion con d =", d, "completada"))
51
52    c(round(Media, 4), round(1/2*sqrt(d/3), 4), round(Varianza, 4))
53  }
54
55  Tc <- data.frame(res, row.names = c("Media", "1/2 Sqrt(d/3)", "
      Varianza"))
56  colnames(Tc) <- paste0("d=", i)
57
58  stopCluster(cl)
59  Tc
```

Listing C.3: Código R para estimar las densidades de la norma de una uniforme en hipercubos.

C.4. Código R para la generación de la Figura 5.1

```
1
2   library(stats)
3
4   n <- 150
5   d <- 100000
6
7   X <- matrix(runif(n * d, min = 0, max = 1), nrow = n, ncol = d)
8
9   epsilon <- 0.1
10  k <- ceiling(4 * log(n) / ( (epsilon^2)/2-(epsilon^3)/3 ) )
11  k
12
13  sample_pairs <- combn(n, 2)
14
```

```r
Phi <- matrix(rnorm(k * d, mean = 0, sd = 1 / sqrt(k)), nrow = k,
    ncol = d)
Z <- X %*% t(Phi)
rm(Phi);gc()

selected_pairs <- sample(ncol(sample_pairs), 500)
pairs <- sample_pairs[, selected_pairs]


D_original <- apply(pairs, 2, function(pair) {
    i <- pair[1]; j <- pair[2]
    sqrt(sum((X[i, ] - X[j, ])^2))
})


D_proj <- apply(pairs, 2, function(pair) {
  i <- pair[1]; j <- pair[2]
  sqrt(sum((Z[i, ] - Z[j, ])^2))
})

R <- D_proj / D_original

cumplen <- (R >= (1 - epsilon)) & (R <= (1 + epsilon))
porcentaje_cumple <- mean(cumplen) * 100

cat(sprintf("Porcentaje de pares proyectados a la dimension %d
   que cumplen la cota: %.3f%%\n",k, porcentaje_cumple))

plot(R,
  pch = 19, col = ifelse((R >= (1 - epsilon)) & (R <= (1 +
     epsilon)), "darkgreen", "red"),
  main = sprintf("Distorsion de las distancias proyectadas (k = %
     d)", k),
  xlab = "Indice del par de puntos", ylab = "R = ||Tu(xi) - Tu(xj
     )|| / ||xi - xj||",
  ylim = c(min(R, 1 - epsilon) - 0.05, max(R, 1 + epsilon) +
     0.05))

abline(h = 1 - epsilon, col = "blue", lty = 2, lwd = 2)
abline(h = 1 + epsilon, col = "blue", lty = 2, lwd = 2)
abline(h = 1, col = "black", lty = 1, lwd = 1)
```

Listing C.4: Código `R` para calcular la distorsión promedio de 500 pares de puntos de una distribución uniforme en $[0,1]^d$ al proyectarlos a un subespacio de dimensión $k = 4295$ cuando $d = 100000$

C.5. Código `R` para la generación de la Tabla 5.1

```r
library(foreach)
library(doParallel)

# Valores de epsilon a probar
epsilon_vals <- c(0.2, 0.17, 0.15)

n <- 150
d <- 100000

X <- matrix(runif(n * d, min = 0, max = 1), nrow = n, ncol = d)
sample_pairs <- combn(n, 2)

num_cores <- parallel::detectCores() - 1
cl <- makeCluster(num_cores)
registerDoParallel(cl)

resultados <- foreach(epsilon = epsilon_vals, .combine = rbind, .
   packages = "stats") %dopar% {

k <- c(ceiling(48 * log(n) / (epsilon^2)),ceiling(4 * log(n) / ((
   epsilon^2)/2 - (epsilon^3)/3)))
media_porcentaje_cumple <- numeric(2)

for (ki in k) {

  Phi <- matrix(rnorm(ki * d, mean = 0, sd = 1 / sqrt(ki)), nrow
     = ki, ncol = d)
  Z <- X %*% t(Phi)

porcentaje_cumple <- numeric(500)

for (contador in 1:10){

selected_pairs <- sample(ncol(sample_pairs), 500)
pairs <- sample_pairs[, selected_pairs]

D_original <- apply(pairs, 2, function(pair) {
  i <- pair[1]; j <- pair[2]
  sqrt(sum((X[i, ] - X[j, ])^2))
})


D_proj <- apply(pairs, 2, function(pair) {
  i <- pair[1]; j <- pair[2]
  sqrt(sum((Z[i, ] - Z[j, ])^2))
})

R <- D_proj / D_original

```

```
47 cumplen <- (R >= (1 - epsilon)) & (R <= (1 + epsilon))
48 porcentaje_cumple[contador] <- mean(cumplen) * 100
49
50 }
51
52 media_porcentaje_cumple[match(ki,k)] <- mean(porcentaje_cumple)
53 cat(sprintf("Porcentaje de pares proyectados a la dimension %d
       que cumplen la cota: %.3f%%\n",
54             ki, media_porcentaje_cumple[match(ki, k)]))
55 rm(Phi)
56 rm(Z)
57 gc()
58 }
59
60 c(epsilon, k[1], media_porcentaje_cumple[1], k[2], media_
       porcentaje_cumple[2])
61 }
62
63 stopCluster(cl)
64
65 colnames(resultados) <- c("epsilon", "k1", "cumplen_k1(%)", "k2",
       "cumplen_k2(%)")
66 print(resultados)
```

Listing C.5: Código `R` para estimar la probabilidad del Teorema de Johnson-Lindestrauss para distintos valores de ε y dimensiones de proyección k

Bibliografía

Aggarwal, C. C., A. Hinneburg y D. A. Keim (2001). "On the Surprising Behavior of Distance Metrics in High Dimensional Space". En: *Database Theory — ICDT 2001*. Lecture Notes in Computer Science. Springer.

Ailon, N. y B. Chazelle (2009). "The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors". En: *SIAM Journal on Computing*.

Anderssen, R. S. y P. Bloomfield (1975). "Properties of the Random Search in Global Optimization". En: *Journal of Optimization Theory and Applications* 16.5-6, págs. 383-398.

Anderssen, R. S., R. P. Brent et al. (1976). "Concerning $\int_0^1 \cdots \int_0^1 (x_1^2+\cdots+x_k^2)^{1/2}\, dx_1 \cdots dx_k$ and a Taylor Series Method". En: *SIAM Journal on Applied Mathematics* 30.1, págs. 22-30.

Bailey, D. H., J. M. Borwein y R. E. Crandall (2007). "Box Integrals". En: *Journal of Computational and Applied Mathematics* 206.1, págs. 196-208.

Bellman, R. (1958). *Dynamic Programming*. Princeton University Press, Princeton.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Boutsidis, C., A. Zouzias y P. Drineas (2010). "Random Projections for k-Means Clustering". En: *Advances in Neural Information Processing Systems* 23.

Cannings, T. I. y R. J. Samworth (2017). "Random-Projection Ensemble Classification". En: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.

Dasgupta, S. y A. Gupta (2003). "An Elementary Proof of Theorem of Johnson and Lindenstrauss". En: *Random Structures & Algorithms*.

Everitt, B. et al. (2011). *Cluster analysis*. English. 5th. Wiley.

Gut, A. (2005). *Probability: A Graduate Course*. Springer.

Hastie, T., R. Tibshirani y J. Friedman (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.

Izenman, A. J. (2008). *Modern multivariate statistical techniques : regression, classification, and manifold learning*. Springer.

James, G. et al. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Johnson, W. y J. Lindenstrauss (1984). "Extensions of Lipschitz Maps into a Hilbert Space". En: *Contemporary Mathematics* 26, págs. 189-206.

Marzetta, T. L., G. H. Tucci y S. H. Simon (2010). "A Random Matrix–Theoretic Approach to Handling Singular Covariance Estimates". En: *IEEE Transactions on Information Theory*.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw Hill,

Robbins, David P. y Theodore S. Bolis (1978). "E2629". En: *The American Mathematical Monthly* 85.4, págs. 277-278.

Shalev-Shwartz, S. y S. Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wegner, S. A. (2024). *Mathematical Introduction to Data Science*. Springer.