Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Matemáticas

APLICACIONES DE LA TEORÍA DE LA INFORMACIÓN EN EL APRENDIZAJE AUTOMÁTICO

Autor: Miguel Santos Pascual
Tutores: Eustasio del Barrio Tellado
Año 2024

Índice general

RESUMEN

Paralelo al desarrollo del aprendizaje automático, aparece un concepto llamado **overfitting**. Esta idea hace referencia a aquellos modelos que se ajustan demasiado bien a los datos de entrenamiento pero no dan buenos resultados de cara a predicciones posteriores y siempre ha supuesto un reto complicado de evitar. Además, con la aparición de las redes neuronales, los avances que se habían hecho en relación a este problema parecen llevar a contradicciones. Surge así el **Cuello de Botella de la Información**, un intento de dar solución a estos nuevos descubrimientos.

Este nuevo modelo, que busca eliminar la información irrelevante de los datos de entrenamiento, se basa en la matemática que **la teoría de la información** de Shannon proporciona. A su vez, esta nos permite desarrollar analíticamente casos mas concretos como es el caso Gaussiano y finalmente aplicar estos nuevos avances de forma numérica y computacional para entender más todos estos fenómenos mencionados.

Palabras clave: Cuello de Botella de la Información, doble descenso, desigualdad de potencias de entropía, desigualdad de procesado de datos y regresión en alta dimensión.

ABSTRACT

Parallel to the development of machine learning, a concept called **overfitting** emerges. This idea, which refers to those models that fit prefectly the training data but do not yield good results for future predictions, have always been a challenging issue to avoid. Moreover, with the advent of neural networks, the advances made concerning this problem seem to lead to contradictions. Thus, the **IB** arises, an attempt to address these new discoveries.

This new model, which aims to eliminate irrelevant information from the training data, is based on the mathematics provided by **Shannon's information theory**. In turn, this theory allows us to analytically develop more specific cases such as the Gaussian case and ultimately apply these new advances numerically and computationally to better understand all these mentioned phenomena.

Key words: Information Bottleneck, Double Descent, Entropy Power Inequality, Data Processing Inequality and high-dimensional regression.

Capítulo 1

Introducción

Actualmente, la inteligencia artificial se ha vuelto notablemente importante en gran variedad de áreas, como por ejemplo el procesamiento del lenguaje natural, la visión por ordenador, el reconocimiento de voz, el filtrado de correo electrónico, la agricultura de precisión o la medicina personalizada, entre otros. Este campo se centra en el desarrollo de algoritmos y modelos que permiten a los ordenadores aprender sin necesidad de ser explícitamente programadas para una tarea en concreto. Es decir, en lugar de seguir instrucciones estrictas paso por paso, estas máquinas buscan un tipo diferente de algoritmos que les permiten aprender de los datos existentes para poder generalizar y predecir ciertos comportamientos.

Dentro del aprendizaje automático, encontramos dos principales tipos de problemas a los que enfrentarse: el aprendizaje supervisado y el no supervisado. En el aprendizaje supervisado se parte de en un algoritmo que recibe un conjunto de datos etiquetados, es decir, datos que incluyen unos inputs o atributos junto con sus correspondientes outputs o etiquetas. El objetivo es entonces encontrar una función que pueda trasformar las entradas en las salidas de la forma más acertada posible. Un ejemplo de este tipo de problemas es la clasificación de imágenes, donde cada una de ellas está acompañada de una etiqueta que indica a qué categoría pertenece o el objeto que representa. Así, el algoritmo aprende a partir de estos ejemplos etiquetados, pudiendo hacer predicciones precisas sobre nuevas entradas.

Por otro lado, en el aprendizaje no supervisado, el algoritmo recibe un conjunto de datos sin etiquetar y debe encontrar patrones o estructuras intrínsecas en los datos por sí mismo. En este caso, el objetivo es explorar la estructura oculta de los datos sin la guía de etiquetas externas. Por ejemplo, en un conjunto de datos de clientes de una tienda, el aprendizaje no supervisado podría utilizarse para identificar grupos de clientes con comportamientos de compra similares, sin tener etiquetas que indiquen cuáles son esos grupos de antemano.

A partir de aquí, nos centraremos en el aprendizaje supervisado, es decir, encontrar esas funciones predictoras a partir de unos datos de manera óptima [9, 15, 25].

1.1. El formalismo del aprendizaje supervisado

Sin asumir ninguna estructura sobre los datos de entrenamiento sería difícil o imposible establecer una serie de criterios para evaluar matemáticamente la calidad de una regla predictora. Es habitual asumir que el conjunto de entrenamiento es una realización de vectores aleatorios (X_i, Y_i) independientes e idénticamente distribuidos (i.i.d.), $1 \leq i \leq n$. Estos X_i son las entradas de nuestra función y reciben el nombre de atributos, mientras que los Y_i son los valores del espacio cuyo comportamiento se desea predecir y reciben el nombre de etiquetas. Además, de aquí en adelante, asumiremos que (X, Y) es un vector aleatorio con la misma distribución que (X_i, Y_i).

Generalmente se tiene que los x_i pertenecen a un espacio $\mathcal{X}$ que suele asociarse con $\mathcal{X} = \mathbb{R}^d$, mientras que las etiquetas y_i dependen del caso en el que se este trabajando. Opciones frecuentes son $\mathcal{Y} = \{0, 1\}$ (clasificación binaria), $\mathcal{Y} = \{0, 1, ..., k\}$ (clasificación muticlase) o $\mathcal{Y} = \mathbb{R}^s$ (regresión).

Intuitivamente, el problema de predecir Y a partir de X estaría resuelto si se conociera la distribución de probabilidad conjunta (X, Y). Sin embargo, al no ser así, el objetivo es entonces encontrar una función $h : \mathcal{X} \to \mathcal{Y}$ que se adapte bien a los datos que tenemos y que de esta forma nos ayude a predecir nuevos casos lo mejor posible. Con esta notación, se busca $h \in \mathcal{H}$ donde $\mathcal{H}$ es el conjunto de funciones de $\mathcal{X}$ a $\mathcal{Y}$ que cumple una serie de condiciones que dependerán del problema que se este estudiando o del procedimiento con el que este se aborde.

El siguiente paso será buscar una forma de medir cuánto de buena es una función $h \in \mathcal{H}$ con respecto a las otras. Aparece así lo que se conoce como **función de pérdida** (*loss function, cost function o error function* en la literatura). Esta función asocia a cada evento o valores de una o más variables aleatorias un número real que representa intuitivamente un coste asociado con ese evento.

Formalmente, es una función. $l : X \times Y \to \mathbb{R}$ que depende de la función h que elijamos. Por tanto, en muchas ocasiones se usa la siguiente notación, $l : X \times Y \times \mathcal{H} \to \mathbb{R}$. Algún ejemplo de las funciones más comunes que se adoptan como funciones de pérdida son:

$$l(x, y, h) = (y - h(x))^2 \text{ pérdida cuadrática,} \tag{1.1}$$

$$l(x, y, h) = \left(1 - \mathbb{I}_{(h(x) \neq y)}\right) \text{ pérdida 0-1,} \tag{1.2}$$

$$l(x, y, h) = |y - h(x)| \text{ pérdida absoluta.} \tag{1.3}$$

Como $l(X, Y, h)$ es un valor aleatorio, es conveniente trabajar con el **riesgo**, que se define como el valor esperado de la función de pérdida para h fijo:

$$R(h) = E\left(l(X, Y, h)\right) = E\left(l(Y, \hat{Y})\right), \tag{1.4}$$

donde se define $\hat{Y} = h(X)$ como la predicción sobre X que se está estudiando.

En resumen, el objetivo del aprendizaje supervisado es en principio encontrar esa función $\hat{h}$ que minimice el riesgo

$$\bar{h} = \arg\min_{h \in \mathcal{H}} R(h). \tag{1.5}$$

Este planteamiento cumple que para ciertas funciones de pérdida, como es el caso de la pérdida cuadrática, este problema tiene solución exacta dada por $h_B(x) = E\left(Y | X = x\right)$ ya que $E(h_B(X)) =$

$E(E(Y|X)) = E(Y)$ y por tanto se cumpliría que:

$$R(h_B) = E(l(Y, \hat{Y})) = E\left(l(Y, Y)\right) = E(0) = 0. \tag{1.6}$$

Sin embargo, es imposible conocer la función $h_B(x)$ ya que esto implicaría conocer $f_{X,Y}\left(x, y\right)$ y por tanto el problema ya estaría resuelto desde un principio.

Por esta misma razón, conocer esta función de riesgo para un h fijo es imposible ya que el calculo de la esperanza requiere conocer la función de densidad $f_{X,Y}\left(x, y\right)$. Por eso, en la mayoría de casos se tiende a usar esta misma función pero en su forma empírica. Recibe el nombre de **función de riesgo empírica.**

$$R_n(h) = \frac{1}{n}\sum_n l(x_i, y_i, h). \tag{1.7}$$

Que según la ley de los grandes números se tiene que

$$R_n(h) \xrightarrow{n\to\infty} R(h). \tag{1.8}$$

Por tanto, el problema se transforma de $\hat{h} = \arg\min_{h\in\mathcal{H}} R(h)$ a $\hat{h} = \arg\min_{h\in\mathcal{H}} R_n(h)$. Esto se conoce como **principio de minimización del riesgo empírico** y al no ser exactamente el mismo resultado, tiene una serie de problemas asociados que se conocen como **overfitting** y **underfitting**.

Para visualizar correctamente en qué consisten estos problemas, definimos el **exceso de riesgo** como la diferencia existente entre nuestro resultado $\hat{h}$ obtenido y el mejor resultado posible que denotamos por h_B.

$$\left|R(\hat{h}) - R(h_B)\right|. \tag{1.9}$$

Para estudiar el comportamiento de este exceso de riesgo, se escribe de la siguiente manera:

$$\left|R(\hat{h}) - R(h_B)\right| = \left|R(\hat{h}) - R_n(\hat{h}) + R_n(\hat{h}) - R(h) + R(h) - R(h_B)\right|. \tag{1.10}$$

donde h es un elemento de $\mathcal{H}$ cualquiera y por tanto cumple que $R_n(\hat{h}) \leq R_n(h)$:

$$\begin{aligned}
& \left|R(\hat{h}) - R_n(\hat{h}) + R_n(\hat{h}) - R(h) + R(h) - R(h_B)\right| \leq \\
\leq\ & \left|R(\hat{h}) - R_n(\hat{h})\right| + \left|R_n(\hat{h}) - R(h)\right| + |R(h) - R(h_B)| \leq \\
\leq\ & \left|R(\hat{h}) - R_n(\hat{h})\right| + |R_n(h) - R(h)| + |R(h) - R(h_B)| \leq \\
\leq\ & \sup_{h\in\mathcal{H}} |R(h) - R_n(h)| + \sup_{h\in\mathcal{H}} |R_n(h) - R(h)| + |R(h) - R(h_B)| = \\
=\ & 2\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| + |R(h) - R(h_B)|.
\end{aligned} \tag{1.11}$$

Y como está expresión es válida para cualquier h, el último término se puede sustituir por el ínfimo de forma que se llega a la siguiente desigualdad

$$\left|R(\hat{h}) - R(h_B)\right| \leq 2\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| + \inf_{h\in\mathcal{H}} |R(h) - R(h_B)|. \tag{1.12}$$

Aparece así que el exceso de riesgo para nuestra candidata $\hat{h}$, obtenida minimizando el riesgo empírico, está acotada por dos tipos de errores:

- **Error de estimación** ($\sup_{h\in\mathcal{H}} |R(h) - R_n(h)|$). Recoge el error debido a minimizar el riesgo empírico y no el riesgo teórico. Tiene la peculiaridad de que conforme el conjunto de las funciones posibles se hace más grande, aumenta.

- **Error de aproximación** ($\inf_{h\in\mathcal{H}} |R(h) - R(h_B)|$). Describe el error que hay entre las posibles elecciones de h y el resultado exacto de la minimización del riesgo. Al contrario que el error anterior, este disminuye al ir haciéndose más grande el conjunto $\mathcal{H}$, ya que conforme este aumente, las posibles elecciones de h se parecen más a h_B. Además, si se tiene que $h_B \in \mathcal{H}$, este error será nulo.

Frecuentemente en el aprendizaje supervisado se maneja no una única clase $\mathcal{H}$, sino una colección de modelos $\mathcal{H}_k$ tal que $\mathcal{H}_k \subset \mathcal{H}_{k+1}$. De esta forma, elegir un modelo de alta complejidad ($\mathcal{H}_k$ "grande") produce un error de aproximación pequeño y un error de estimación alto, entonces los resultados del ajuste sobre la muestra serán bueno pero la capacidad de generalización será pobre. Este es el fenómeno conocido como **sobreajuste** u **overfitting** en inglés.

En el otro extremo, elegir un modelo muy sencillo generará poca capacidad de aprender de la muestra. Esto se conoce como **infraajuste** o **underfitting**.

La idea gráfica de estos dos fenómenos se puede comprobar mediante las gráficas de la Figura 1.1.

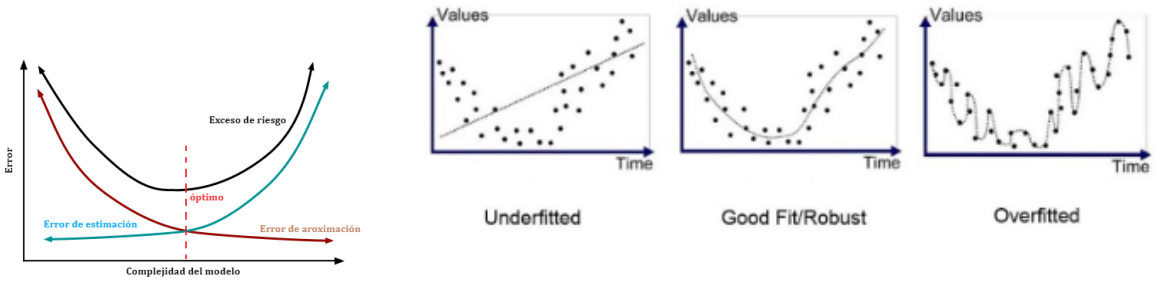


Figura 1.1: **(a)** Descripción gráfica del exceso de riesgo en función de los dos tipos de errores, **(b)** Visualización del problema sobre la complejidad del modelo. *Imagen obtenida en* [9]

1.2. Fenómeno del doble descenso

La gráfica de la Figura 1.1(a) parece indicar que el comportamiento de un modelo de aprendizaje no puede ser bueno en modelos de alta complejidad. Sin embargo, gran parte de las aplicaciones recientes en Inteligencia Artificial se basan en modelos fuertemente sobreparametrizados como son las redes neuronales. Estas son funciones predictivas muy complejas y que dependen de una cantidad elevada de parámetros que se obtienen a partir de los datos de entrenamiento. Sorprendentemente, a pesar de su complejidad y de lo que uno podría esperar según lo visto anteriormente, estas redes presentan muy buen comportamiento en lo que respecta a predecir nuevos datos [16].

Esta contradicción recibe el nombre de **doble descenso** y aparece como resultado del estudio del exceso de riesgo ($R(\hat{h}) - R(h_B)$) en este tipo de modelos de excesiva complejidad y parametrización [18].

En la Figura 1.2 se puede ver cómo llega un momento en el que la complejidad del modelo deja de ajustarse a la cota superior presentada en (1.12) y comienza a descender. Es por esto que nos referimos como **doble descenso** a este fenómeno.

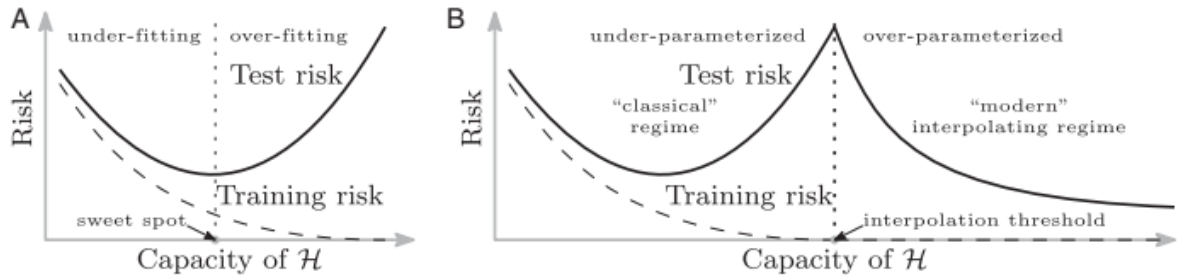


Figura 1.2: Representación intuitiva del doble descenso en una gráfica del exceso de riesgo. *Imagen obtenida en* [18]

Esta contradicción da lugar a nuevos enfoques para el estudio del **sobreajuste** y los diferentes mecanismos para hacerle frente. Consecuentemente, el objetivo de este trabajo es estudiar uno de los planteamientos que parece tener capacidad de explicar este fenómeno previo. Este se conoce como **Cuello de Botella de la Información** o **IB** por las siglas en inglés de **Information Bottleneck**.

1.3. El Cuello de Botella de la Información

La idea parte de conceptos propios de la Teoría de la Información, en concreto de la información mutua. Esta es una medida de la asociación entre variables aleatorias de manera que cuando son independientes la información mutua es nula y cuanto más se parecen entre sí, la información será cada vez mayor.

En general, el modelo de **Cuello de Botella** trata de conseguir que los algoritmos de aprendizaje mantengan una información controlada sobre la muestra de entrenamiento maximizando la información sobre el mecanismo de interés generador de los datos. A su vez busca desechar la máxima información residual, es decir, aquella que se "repite" en los datos de entrenamiento para mejorar la eficacia de aprendizaje.

Para poder explicar el **Cuello de Botella de la Información** de forma adecuada es necesario cuantificar de alguna forma esa idea de "información". Por ello, en este trabajo se dedica el capitulo 2 a desarrollar los conceptos básicos con la Teoría de la Información como la entropía, la información mutua o la divergencia de Kullback-Leibler.

El objetivo final es analizar este metodo para un cierto modelo de regresión lineal normal. Para que esto sea técnicamente posible es necesario estudiar con detalle el procedimiento del **Cuello de**

Botella de la Información en el contexto gaussiano. Y para ello, hay un resultado fundamental que reduce el problema a uno de autovalores y autovectores de ciertas matrices de covarianza. Probar ese resultado implica bastante complejidad técnica y requiere el desarrollo de algunas desigualdades en Teoría de la Información. A estas desigualdades se dedica el Capítulo 3 del documento.

Una vez se hayan comprendido estos preliminares, el método del **Cuello de Botella de la Información** se presenta en el capítulo 4 con atención especial al caso Gaussiano.

Y finalmente, el trabajo se completa con el capitulo 5 en el que se analiza un método de regresión lineal con las herramientas mencionadas a lo largo del trabajo. También se exponen en el ultimo capítulo una serie de conclusiones que ayudan al lector a entender globalmente el trabajo.

Capítulo 2

Conceptos básicos de la Teoría de la Información

La teoría de la información [11] es un campo de las Matemáticas que surgió en la década de 1940 gracias al trabajo pionero de Claude Shannon. En esencia, esta disciplina se ocupa del estudio cuantitativo de la transmisión, el almacenamiento y el procesamiento de la información, proporcionando así un marco conceptual poderoso para entender cómo se maneja la información en el mundo moderno.

En el corazón de la teoría de la información se encuentra el concepto de "información", que es una medida de la incertidumbre reducida o eliminada por un mensaje. Este concepto se relaciona estrechamente con ideas como la entropía o la divergencia de Kullback-Leibler, que veremos a continuación. También abarca otros aspectos como la codificación, la representación de la información de manera que sea más fácil de almacenar, transmitir y procesar entre multitud de aplicaciones.

Clásicamente, el marco de la teoría de la información se desarrolla principalmente para funciones de probabilidad discretas. Esto se debe a que en el tratamiento de la información se trabaja con mensajes finitos, por lo que los espacios muestrales que lo componen han de serlo también.

Sin embargo, la mayoría de conceptos que se van a presentar en este capítulo se manejan de forma habitual en Estadística Matemática, ampliando así su aplicación en el mundo real donde generalmente los datos siguen distribuciones de probabilidad continuas o, debido a la cantidad de datos existentes, se pueden aproximar por probabilidades con densidades continuas.

Con esta intención de generalizar estos conceptos a cualquier distribución de probabilidad, la teoría de la medida cobra un papel imprescindible presentándonos conceptos como la continuidad absoluta o derivada de Radon-Nikodym, una herramienta fundamental para el uso de la información mutua en cualquier situación. Los detalles de estos detalles vienen explicados convenientemente en [Appendix A.1].

2.1. Entropía

Definición 2.1.1 (**Entropía**). Sean P una probabilidad definida en $\{x_1, x_2, ..., x_n\}$. La **entropía de** P se define como:

$$H(P) = -\sum_{i=1}^{n} p_i \log(p_i) = \sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right), \tag{2.1}$$

donde se cumple siempre que $0 \leq p_i \leq 1$, y por tanto se tiene que $\log\left(\frac{1}{p_i}\right) \geq 0$ para todo i y por tanto $H(P) \geq 0$.

Otra forma en la que se puede expresar entropía es como la esperanza del logaritmo cambiada de signo, es decir, $H(P) = E_P(\log(\frac{1}{P(X)})) = -E_P(\log(P(X)))$. Y en cuánto a la notación utilizada, dado que la distribución de probabilidad utilizada P siempre está asociada a una variable aleatoria X, esta entropía u otros conceptos posteriores que veremos, se suele denotar por $H(P) = H(X)$

Generalmente, esta entropía viene definida para el logaritmo en base 2, ya que de esta forma, es posible demostrar que el mínimo número de preguntas binarias necesarias para determinar una variable aleatoria X se encuentra entre $H(X)$ y $H(X)+1$. [11] Sin embargo, debido a las propiedades del logaritmo $\log_b(p) = \log_b(a)\log_a(p)$, se tiene que $H_b(X) = \log_b(a)H_a(X)$ y esto nos permite trabajar en la base que uno prefiera. De aquí en adelante se usara el logaritmo natural.

De forma intuitiva, la entropía es una medida de la incertidumbre de una variable, se convierte en autoinformación de ella misma. Esta definición justifica la indiferencia de la base logarítmica utilizada ya que el hecho de tener una variable con entropía igual a 5 no nos dice nada. En cambio, si nos dicen que otra variable tiene entropía igual a 10, implica una mayor incertidumbre en este segundo caso.

Para entender este término veamos un ejemplo muy sencillo. Consideremos un dado justo, es decir, tal que la probabilidad de obtener cada uno de los seis resultados posibles es 1/6.

$$H(X) = \sum_{i=1}^{6} \frac{1}{6} \log(6) = \log(6) = 1{,}79175. \tag{2.2}$$

En cambio, si se tiene un dado trucado en el que por ejemplo la probabilidad de sacar un 1 y un 2 es 1/12 y la de sacar un 6 es ahora 1/3 dejando intactas el resto de probabilidades, ahora la entropía será:

$$H(X') = \frac{3}{6}\log(6) + \frac{2}{12}\log(12) + \frac{1}{3}\log(3) = 1{,}67623. \tag{2.3}$$

El hecho de que la entropía sea mayor en el primer caso implica una mayor incertidumbre en el resultado de la variable aleatoria. Esto se debe a que al ser equiprobables, no existe ningún comportamiento predecible, no existe un orden. En cambio, en el segundo caso, existe un mayor orden y por tanto la entropía es menor. El caso extremo es aquel en el que solo hay un resultado posible con probabilidad 1 y por tanto su entropía es nula, no hay ningún tipo de desorden.

Una curiosidad importante acerca de la entropía consiste en caracterizar el punto en el que, para una serie de sucesos $\{x_1, x_2, ..., x_n\}$, esta entropía se hace máxima.

Proposición 2.1.2 (**Máxima Entropía**)**.** Dados una serie de sucesos $\{x_1, x_2, ..., x_n\}$ posibles, la entropía máxima se alcanza cuando estos son equiprobables, es decir, $p_i = P(x_i) = \frac{1}{n}$ para todo $1 \leq i \leq n$.

Demostración. Maximizar la entropía $-\sum_{i=1}^{n} x_i \log(x_i)$ equivale a minimizar $G(x_1, ..., x_n) = \sum_{i=1}^{n} x_i \log(x_i)$ sujeto a $\sum x_i = 1$. Por tanto, aplicando el método de los multiplicadores de Lagrange,

$$H(x_1, ..., x_n) = G(x_1, ..., x_n) - \lambda(x_1 + ... + x_n), \tag{2.4}$$

$$\frac{\partial H}{\partial x_i} = 1 + \log(x_i) - \lambda = 0. \tag{2.5}$$

Lo cuál implica que $1 + \log(x_i) = \lambda$ para todo i y por tanto, al tener $x_i = x_j$ y $\sum x_i = 1$, se concluye que $x_i = \frac{1}{n}$. □

Una vez visto el caso discreto para el cual se define la entropía inicialmente, es necesario expandir el concepto a distribuciones continuas. Este no coincide exactamente con lo visto previamente y por eso recibe el nombre de **entropía diferencial**.

Definición 2.1.3 (**Entropía Diferencial**)**.** La entropía diferencial $h(X)$ de una variable aleatoria continua X con densidad $f(x)$ se define como

$$h(X) = -\int_{\mathbb{R}^n} f(x) \log f(x)\, dx, \tag{2.6}$$

donde se extiende la continuidad de $f \log f$ al cero por $\lim_{x \to 0} x \log x = 0$. También es importante darse cuenta de que aunque el logaritmo no se anule, la integral podría diverger y por lo tanto decirse que su entropía es infinita.

Al igual que en el caso discreto, la entropía diferencial depende únicamente de la densidad de probabilidad de la variable aleatoria, y por lo tanto, la entropía diferencial puede verse también escrita como $h(f)$ en lugar de $h(X)$.

A continuación, a partir de la definición de **Entropía**, se pueden definir algunos conceptos.

Definición 2.1.4 (**Entropía Conjunta**)**.** Sean P y Q dos probabilidades definidas en $\mathcal{X}$ y $\mathcal{Y}$ respectivamente. La entropía conjunta se define como:

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log \left(\frac{1}{p_{X,Y}(x, y)} \right). \tag{2.7}$$

Esta es básicamente la entropía asociada a la distribución conjunta de las variables X e Y, $H(X, Y)$.

Definición 2.1.5 (**Entropía Condicionada**)**.** Sean P y Q dos probabilidades definidas en $\mathcal{X}$ y $\mathcal{Y}$ respectivamente. La entropía condicionada de P respecto a Q se define como:

$$\begin{aligned} H(X|Y) &= \sum_{x \in \mathcal{X}} p_X(x) H(Y|X = x) = \sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p(y|X = x) \log \left(\frac{1}{p(y|X = x)} \right) = \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log \left(\frac{1}{p(y|X = x)} \right). \end{aligned} \tag{2.8}$$

Que tienen también su análogo para el caso continuo.

Definición 2.1.6 (**Entropía Diferencial Conjunta**)**.** La entropía diferencial de un conjunto $X_1, X_2, \ldots, X_n$ de variables aleatorias con densidad $f(x_1, x_2, \ldots, x_n)$ se define como

$$h(X_1, X_2, \ldots, X_n) = -\int f(x_1, x_2, \ldots, x_n) \log f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 ... dx_n. \tag{2.9}$$

Definición 2.1.7 (**Entropía Diferencial Condicionada**)**.** Si X e Y tienen una función de densidad conjunta $f(x, y)$, podemos definir la entropía diferencial condicional $h(X|Y)$ como

$$h(X|Y) = -\iint f(x, y) \log f(x|y)\, dx\, dy. \tag{2.10}$$

Aviso sobre notación: en cuanto a esta entropía condicionada, ya sea en el caso discreto o continuo, uno tiene que tener en cuenta una serie de aspectos. Cuando en Estadística se usa la notación $H(X|Y = y) = -\sum_x P_{x|Y=y} \log(P_{x|Y=y})$, se trata de una función de y, $H(X|Y = y) = g(y)$. Teniendo en cuenta este criterio, cuando se escribe de la forma, $H(X|Y) = g(Y)$, se esta hablando entonces de un variable aleatoria. Por el contrario, en Teoría de la Información, cuando se escribe $H(X|Y)$, se esta haciendo referencia exactamente a la esperanza de esa variable aleatoria $g(Y)$, es decir, $H(X|Y) = Eg(Y)$.

A partir de las definiciones previas surgen una serie de propiedades válidas tanto para el caso continuo como para el caso discreto. La demostración de estas es análoga para ambos caso y por simplicidad las llevaremos a cabo para uno solo.

Proposición 2.1.8 (**Relación entre entropías**)**.** La relación entre las diferentes entropías vistas previamente es la siguiente: $H(X, Y) = H(Y|X) + H(X)$. Que para el caso continuo será, $h(X, Y) = h(Y|X) + h(X)$.

Demostración. La demostración se deduce a partir de la propia definición de entropía.

$$\begin{aligned}
H(X, Y) &= -\sum_{i,j} P(x_i, y_j) \log P(x_i, y_j) \\
&= -\sum_{i,j} P(y_j|x_i) \cdot P(x_i) \log \left(P(y_j|x_i) \cdot P(x_i)\right) \\
&= -\sum_{i,j} P(y_j|x_i) \cdot P(x_i) \left(\log P(y_j|x_i) + \log P(x_i)\right) \\
&= -\sum_{i,j} P(y_j|x_i) \cdot P(x_i) \log P(y_j|x_i) - \sum_{i,j} P(y_j|x_i) \cdot P(x_i) \log P(x_i) \\
&= H(Y|X) + H(X).
\end{aligned} \tag{2.11}$$

□

La relación entre la entropía conjunta $H(X, Y)$, la entropía marginal $H(X)$ y la entropía condicional $H(Y|X)$ proporciona una comprensión profunda de cómo relacionar las incertidumbres entre múltiples variables aleatorias. La fórmula $H(X, Y) = H(Y|X) + H(X)$ establece que la incertidumbre conjunta de X e Y es la suma de la incertidumbre de X más la incertidumbre residual que proporciona Y una vez conocido X. Lo cuál en términos de incertidumbre es bastante lógico.

Además, de la proposición previa es posible deducir otra propiedad importante conocida como la regla de la cadena para entropias.

Proposición 2.1.9 (**Regla de la cadena para le entropía diferencial**)**.**

$$h(X_1, ..., X_n) = \sum_i h(X_i|X_{i-1}, ..., X_1). \tag{2.12}$$

Demostración. Por recurrencia usando la propiedad $h(X, Y) = h(X) - h(Y|X)$. □

Proposición 2.1.10. La entropía es invariante bajo traslación $h(X + c) = h(X)$, donde X representa un vector aleatorio. La demostración de esta afirmación se deduce de la propia definición.

Proposición 2.1.11. $h(AX) = h(X) + \log(|\det(A)|)$ donde X es un vector aleatorio.

Demostración. La demostración se basa únicamente en un cambio de variable [Theorem A.4.1] de determinante jacobiano $\det(A)$ y tal que $y = Ax$

$$\begin{aligned} h(AX) &= -\int f_Y(y) \log f_Y(y)\, dy && (2.13)\\ &= -\int \frac{1}{\det(A)} f_X\left(A^{-1}y\right) \log\left(\frac{1}{\det(A)} f_X\left(A^{-1}y\right)\right) dy && (2.14)\\ &= -\int f_X(x) \log f_X(x)\, dx + \log(\det(A)) && (2.15)\\ &= h(X) + \log(\det(A)). && (2.16) \end{aligned}$$

□

2.2. Divergencia de Kullback-Leibler

La entropía es un valor en el que participa una única variable ya sea condicionada o no. Además, como se vio en los ejemplos, para sacar conclusiones a partir de la entropía, es necesario calcular esta para varios casos y así poder llevar a cabo una comparación. Esto no es del todo cómodo y como solución surge la **La Divergencia de Kullback-Leibler**, un parámetro en el que participan dos probabilidades distintas y nos permite esa comparación que se mencionaba.

Definición 2.2.1 (**Divergencia de Kullback-Leibler**)**.** La entropía relativa o divergencia de Kullback-Leibler $D(P||Q)$ entre dos medidas de probabilidad P y Q con respecto a una medida μ tal que $P \ll \mu$ y $Q \ll \mu$ se define como:

$$D(P||Q) = \int f_P log \frac{f_P}{f_G} d\mu, \tag{2.17}$$

donde $f_P = \frac{dP}{d\mu}$ y $f_G = \frac{dQ}{d\mu}$. Además, esta suele escribirse como $D(X||Y)$ haciendo referencia a los vectores aleatorios asociados a las probabilidades P y Q.

Para que la definición sea válida, en primer lugar observamos que la existencia de una tercera medida μ no es una hipótesis restrictiva ya que esta siempre puede ser elegida como $\mu = P + Q$ de forma que se cumpla $P \ll \mu$ y $Q \ll \mu$. Por otro lado, esta misma idea impone un condición no implícita en la definición. Esta consiste en que el valor de esta divergencia debe ser independiente de la medida μ utilizada.

Propiedad 2.2.2. La diferencia de Kullback-Leibler es independiente de la medida utiliza.

Demostración. Sean μ y ν dos medidas distintas tales que $f_P = \frac{dP}{d\mu}$, $f_G = \frac{dQ}{d\mu}$, $h_P = \frac{dP}{d\nu}$, $h_G = \frac{dQ}{d\nu}$. Como $\mu \ll \mu + \nu$ y $\mu \ll \mu + \nu$ se puede considerar $g_\mu = \frac{d\mu}{d(\mu+\nu)}$ y $g_\nu = \frac{d\nu}{d(\mu+\nu)}$. Y haciendo uso de la regla de la cadena:

$$\hat{r}_P = \frac{dP}{d(\mu+\nu)} = \frac{dP}{d\mu}\frac{d\mu}{d(\mu+\nu)} = f_P g_\mu\,, \tag{2.18}$$

$$\hat{r}_Q = \frac{dQ}{d(\mu+\nu)} = \frac{dQ}{d\mu}\frac{d\mu}{d(\mu+\nu)} = f_Q g_\mu\,, \tag{2.19}$$

$$\hat{r}_P = \frac{dP}{d(\mu+\nu)} = \frac{dP}{d\nu}\frac{d\nu}{d(\mu+\nu)} = h_P g_\nu\,, \tag{2.20}$$

$$\hat{r}_P = \frac{dQ}{d(\mu+\nu)} = \frac{dQ}{d\nu}\frac{d\nu}{d(\mu+\nu)} = h_Q g_\nu. \tag{2.21}$$

Y por tanto

$$(\mu):\ D(P||Q) = \int f_P log\frac{f_P}{f_G}d\mu = \int f_P \log\frac{f_P g_\mu}{f_G g_\mu}g_\mu d(\mu+\nu) = \int \hat{r}_P \log\frac{\hat{r}_P}{\hat{r}_Q}d(\mu+\nu)\,, \tag{2.22}$$

$$(\nu):\ D(P||Q) = \int h_P log\frac{h_P}{h_G}d\mu = \int h_P \log\frac{h_P g_\nu}{h_G g_\nu}g_\nu d(\mu+\nu) = \int \hat{r}_P \log\frac{\hat{r}_P}{\hat{r}_Q}d(\mu+\nu). \tag{2.23}$$

Llegando al mismo resultado por ambos lados y probando así lo que se quería. □

De está forma, al ser totalmente independiente de la medida, si trabajamos en $(\mathbb{R}^n, \mathcal{B}^n)$ con la medida de Lebesgue, la divergencia de Kullback-Leibler, se puede escribir como:

$$D(P||Q) = \int \log\frac{dP}{dQ}\,dP = \int f(x)\log\frac{f(x)}{g(x)}\,dx, \tag{2.24}$$

donde $f(x)$ y $g(x)$ son las funciones de densidad habituales.

Y de manera paralela, si la medida μ utilizada, se corresponde con la medida de conteo con la que las integrales se convierten en sumatorio, queda definida la **Divergencia de Kullback-Leibler** para el caso discreto.

Es importante notar que $D(P||Q)$ (también escrito como $D(f||g)$ para el caso en $(\mathbb{R}^n, \mathcal{B}^n)$) tiene sentido únicamente si ambos densidades están definidos en el mismo conjunto, lo cuál es lógico ya que nuestra intención es comparar dos distribuciones de probabilidad.

Alguna de sus propiedades más importantes se muestran a continuación y por simplicidad, y al igual que se hizo para la entropía, las demostraciones se harán únicamente para uno de los casos, ya sea discreto o continuo.

Proposición 2.2.3. : $\mathbf{D(P||Q) \geq 0}$

Demostración. Para probar esto, usaremos que $\log(x) \leq x - 1$ para todo $x > 0$.

$$\begin{aligned} -D(P||Q) = -\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{n} p_i \log\left(\frac{q_i}{p_i}\right) \leq \\ \leq \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_{i=1}^{n} (q_i - p_i) = \\ = \left(\sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i\right) = (1-1) = 0. \end{aligned} \tag{2.25}$$

□

Proposición 2.2.4. $\mathbf{D(P||Q)} = 0$ si y solo si $P = Q$.

Demostración. Para que en 2.25 se de la igualdad, se tiene que cumplir que:

$$\begin{aligned} \sum_{i=1}^{n} p_i \log\left(\frac{q_i}{p_i}\right) = \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) \\ \sum_{i=1}^{n} p_i \left(-\log\left(\frac{q_i}{p_i}\right) + \left(\frac{q_i}{p_i}\right) - 1\right) = 0. \end{aligned} \tag{2.26}$$

El término $\left(-\log\left(\frac{q_i}{p_i}\right) + \left(\frac{q_i}{p_i}\right) - 1\right)$ siempre es mayor o gial que cero por tanto para que el sumatorio se anule, se tiene que dar que $\left(-\log\left(\frac{q_i}{p_i}\right) + \left(\frac{q_i}{p_i}\right) - 1\right) = 0$ para todo i. y por tanto $p_i = q_i$ para todo i. Esto nos lleva a la conlución buscada, $P = Q$ si y solo si $\mathbf{D(P||Q)} = \mathbf{0}$. □

Proposición 2.2.5. $\mathbf{D(P||Q) \neq D(Q||P)}$

Esta propiedad se deduce a simple vista a partir de la definición y de esta forma, la divergencia de Kullback-Leibler deja de ser una distancia

De forma intuitiva, la divergencia de Kullback-Leibler, a pesar de no ser una distancia, se convierte en una medida de cuanto se diferencia la distribución de probabilidad Q de nuestra distribución de referencia P.

Por ejemplo, retomando el caso de los dados veamos, cuanto se parece el dado trucado al dado justo.

$$D(X||X') = -\frac{1}{6}(\log(6/12) + \log(6/12) + \log(6/3)) = 0{,}1155245301. \tag{2.27}$$

Para compararlo, sea P'' el dado totalmente trucado en el que la probabilidad de obtener un 6 es $1/2$ y la del resto de resultados es $1/10$. En este caso

$$D(X||X'') = -\frac{1}{6}(5 * \log(6/10) + \log(6/2)) = 0{,}242585. \tag{2.28}$$

En este último caso, la divergencia a aumenta, esto implica que el segundo caso difiere más del dado equiprobable que el primero de los casos.

2.3. Información mutua

Esta información mutua se trata de otro concepto similar a la **Divergencia de Kullback-Leibler** y cuyo objetivo es también permitirnos comparar diferentes probabilidades.

Definición 2.3.1 (**Información mutua**)**.** Consideremos dos variables aleatorias X e Y. La entropía mutua se define como $I(X;Y) = h(X) - h(X|Y)$ o $I(X;Y) = H(X) - H(X|Y)$ para el caso discreto.

Intuitivamente esta definición surge de medir cuanto desorden nos quita conocer la variable Y a la hora de evaluar X. En otras palabras, nos dice que la información que una variable Y nos da de otra variable aleatoria X es igual al desorden que tiene X si le quitas la incertidumbre que conocer Y nos daría de X. Aún así, en la literatura existen otra forma diferentes de definir inicialmente esta información mutua. la cual se muestra a continuación como proposición.

Proposición 2.3.2. La información mutua $I(X;Y)$ entre dos variables aleatorias con densidad conjunta $f_{X,Y}(x,y)$ se puede escribir también como:

$$\begin{aligned} I(X;Y) &= D(P_{X,Y}||P_X \otimes P_Y) \\ &= \iint \log \frac{dP_{X,Y}}{dP_X \otimes dP_Y}\, dP_{X,Y} \\ &= \iint f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy. \end{aligned} \tag{2.29}$$

Demostración. La idea de esta demostración es comprobar que ambas definiciones son equivalentes.

$$\begin{aligned} I(X;Y) &= \int_X \int_Y f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \\ &= \int_{X,Y} f_{X,Y}(x,y) \log \frac{f(x|y)}{f_X(x)} \\ &= -\int_{X,Y} f_{X,Y}(x,y) \log f_X(x) + \int_{X,Y} f_{X,Y}(x,y) \log f_{X|Y=y}(x|y) \\ &= -\int_X f_X(x) \log f_Y(x) - \left(-\int_{X,Y} f_{X,Y}(x,y) \log f_{X|Y=y}(x|y) \right) \\ &= h(X) - h(X|Y). \end{aligned} \tag{2.30}$$

□

Definición 2.3.3. Consideremos dos variables aleatorias X e Y con una función de probabilidad conjunta $p(x, y)$ y funciones de probabilidad marginales $p(x)$ y $p(y)$. La información mutua $I(X;Y)$ es la entropía relativa entre la probabilidad conjunta y la probabilidad conjunta si ambas variables fueran independientes, es decir, el producto de las distribuciones $p(x)p(y)$. Que, equivalentemente, se corresponde con la divergencia de Kullback-Leibler.

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = D(P_{X,Y}||P_X P_Y). \tag{2.31}$$

Al venir ahora escrita como una **Divergencia de Kullback-Leibler**, se pueden extraer directamente una serie de propiedades acerca de esta información.

Proposición 2.3.4. $\mathbf{I(X;Y) = I(Y;X)}$.

A diferencia de **la Divergencia de Kullback-Leibler**, la información mutua si que permite intercambiar los papeles de las variables aleatorias implicadas. Esto a su vez nos permite escribir reescribir la Definición 2.3.1 de dos maneras distintas $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$.

Intuitivamente, estas igualdades nos permiten entender que la información de dos variables, depende totalmente del desorden remanente de una de ellas al conocer la otra. Es decir, si la entropía de X conociendo Y el muy alta, significa que conocer Y no ha aportado apenas información sobre los posibles resultados de X y por tanto, al ir con signo menos, la información de Y respecto de X deberá ser baja. Por el contrario, si esta entropía condicionada es muy baja, significa que conocer Y ha eliminado prácticamente toda la incertidumbre de X y por tanto, la información de X e Y debe ser alta.

Proposición 2.3.5. $\mathbf{I(X;Y) \geq 0}$.

Proposición 2.3.6. $\mathbf{I(X;Y) = 0}$ si y solo si son variables independientes.

Demostración.

$$I(X;Y) = 0 \iff D(P_{X,Y}||P_X P_Y) = 0 \iff P_{X,Y} = P_X P_Y. \tag{2.32}$$

□

De forma intuitiva, la información mutua se interpreta como una medida de cuánta información comparten dos variables aleatorias. Es decir, si son totalmente independientes, la información que nos da conocer el resultado de X de cara a conocer Y es nula y por tanto así lo será la información. En el lado contrario, tenemos el caso de $I(X;X)$ en el que conocer el resultado de X, nos desvela totalmente el resultado de esta segunda X. Es curioso porque calcular esto para el caso discreto se puede hacer llegando a $I(X;X) = \sum_{x\in X}\sum_{y\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \sum_{x\in X} p(x)\log\frac{p(x)}{p(x)} = H(X)$. Esto es lógico ya que la información que X nos da de si misma esta determinada por su propia desinformación. Sin embargo, para el caso continuo, al no poder definirse correctamente la función de densidad conjunta $f_{X,X}(x,y)$ al estar solamente definida en una subvariedad de dimensión inferior, la integral deja de tener sentido y la información mutua diverge.

A su vez, debido a la no negatividad de H y de I, se obtiene un resultado importante.

Propiedad 2.3.7. Dadas dos variables aleatorias X e Y,la entropía de la variable X es mayor igual que la entropía de esa misma variable conocido Y.

Demostración.

$$0 \leq I(X;Y) = h(X) - h(X|Y) \implies h(X) \geq h(X|Y). \tag{2.33}$$

□

La idea intuitiva se basa en que hecho de que la incertidumbre de una variable siempre será mayor que el incertidumbre de esta conocido Y. La igualdad se cumple cuando $I(X;Y) = 0$, cuando son independientes, es decir, cuando conocer una no nos aporta nada sobre la otra.

Propiedad 2.3.8. Dadas dos variables aleatorias X e Y, la información muta es siempre menor o igual que el mínimo entre la entropía de X y la de Y.

Demostración. Dado que la entropía es no negativa, se tiene que $H(X|Y) \geq 0$ para cualquier par de variables (X, Y).

$$I(X;Y) = h(X) - h(X|Y) \leq h(X) \tag{2.34}$$
$$I(X;Y) = h(Y) - h(Y|X) \leq h(Y). \tag{2.35}$$

Esto implica que $0 \leq I(X;Y) \leq \text{mín}(h(X), h(Y))$, donde recordemos que para el caso continuo, tanto $h(X)$ como $h(Y)$ pueden diverger y por tanto se cumpla que $0 \leq I(X;Y) \leq \infty$. □

Esta idea no es tan fácil de ver a simple vista. Sin embargo, razonando como se vio anteriormente, la información máxima se obtiene cuando $X = Y$ y esta será igual al desorden de la propia variable ya que conocer X, nos determina totalmente $Y = X$, es decir, la información que X nos aporta es igual a la desinformación que teníamos de esa variable.

Corolario 2.3.9. Sean X e Y dos vectores independientes, entonces se tiene que:

$$H(X|Y) \leq H(X). \tag{2.36}$$

Demostración.. La demostración es inmediata tras aplicar a $I(X;Y) = H(X) - H(X|Y)$ que $H(X;Y)$ es siempre mayor o igual que cero. □

Proposición 2.3.10. Sean X e Y dos vectores aleatorios independientes, entonces se tiene que:

$$H(X|Y) = H(X)\,, \tag{2.37}$$
$$H(Y|X) = H(Y)\,. \tag{2.38}$$

Demostración.. La demostración es inmediata tras aplicar a la definición de entropía condicionada que $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ y por tanto $f_{X|Y=y}(x) = f_X(x)$. □

Proposición 2.3.11. Sean X_i vectores independientes, entonces se tiene que:

$$H(X_1, X_2, ..., X_n) \leq \sum_i H(X_i). \tag{2.39}$$

Y la igualdad se da únicamente si son vectores independientes. Además, en el caso continua, hay que tener cuidado si se da que alguna la entropía de la distribución conjunta diverge.

Demostración.. La demostración es inmediata aplicando las Proposiciones 2.3.9 y 2.3.10 junto con la regla de la cadena para entropías. □

Proposición 2.3.12. Haciendo uso de la ecuación 2.11 nos permite escribir la información de la siguiente manera.

$$I(X;Y) = h(X) + h(Y) - h(X,Y). \tag{2.40}$$

Esta igualdad de manera intuitiva se puede interpretar como que la información que una variable tiene de la otra, es igual a la suma de las entropías individuales, es decir, del desorden que estas tienen que restándole el desorden que ellas tienen en conjunto.

Por otro lado, al igual que se hizo para la entropía, existe un concepto de información mutua condicionada. Esto a su vez nos permite desarrollar un análogo a la regla de la cadena de la entropía para la **Información Mutua**.

Definición 2.3.13 (**Información Mutua Condicionada**)**.** La información mutua condicionada $I(X;Y|Z)$ entre dos variables aleatorias condicionada de Z se define como

$$\begin{aligned} I(X;Y|Z) &= E_Z\left(D(P_{X,Y|Z=z}||P_{X|Z=z} \otimes P_{Y|Z=z})\right) \\ &= \int f_Z(z) \int f(x,y) \log \frac{f_{X,Y|Z=z}(x,y)}{fX|Z=z(x)fY|Z=z(y)}\, dx\, dy \\ &= \iint f_{X,Y,Z}(x,y,z) \log \frac{f_{X,Y|Z=z}(x,y)}{fX|Z=z(x)fY|Z=z(y)}\, dx\, dy. \end{aligned} \tag{2.41}$$

Proposición 2.3.14. Regla de la cadena para información mutua:

$$I(X_1, ..., X_n; Y) = \sum_i I(X_i; Y|X_{i-1}, ..., X_1). \tag{2.42}$$

Que aplicado a $n = 2$, se escribe como $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$.

Demostración. Haciendo uso de la regla de la cadena para la entropía,

$$\begin{aligned} I(X_1, ..., X_n; Y) &= h(X_1, ..., X_n) - h(X_1, ..., X_n|Y) \\ &= \sum_i h(X_i|X_{i-1}, ..., X_1) - \sum_i h(X_i|X_{i-1}, ..., X_1, Y) \\ &= \sum_i \left(h(X_i|X_{i-1}, ..., X_1) - h(X_i|X_{i-1}, ..., X_1, Y)\right) \\ &= \sum_i I(X_i; Y|X_{i-1}, ..., X_1). \end{aligned} \tag{2.43}$$

□

Corolario 2.3.15. Si X e Y son condicionalmente independientes dado Z, entonces se tiene que $I(X;Y|Z) = 0$.

Demostración. La demostración se deduce tras aplicar la definición de información mutua condicionada junto con $f_{X,Y|Z} = f_{X|Z} f_{Y|Z}$.

□

Y finalmente para terminar de entender estos conceptos la Figura 2.1 presenta un gráfico en el que queden reflejados esta información y desinformación de dos variables.

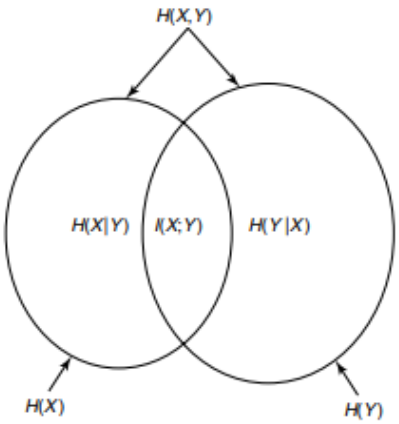


Figura 2.1: Esquema que refleja la relación entre entropía e información mutua [11].

En resumen, la información mutua es la cantidad de información que una variable proporciona sobre la otra. Si la información mutua es alta, significa que las variables están fuertemente correlacionadas y proporcionan mucha información una de la otra. Por otro lado, si la información mutua es baja, las variables están poco correlacionadas y conocer una de ellas no nos enclarece a penas nada sobre la otra.

2.4. Conceptos de la teoría de la información aplicados a distribuciones Gaussianas.

Debido a la importancia que esta distribución tiene y sobre todo debido que va a ser la principal protagonista a lo largo del documento, vamos a hacer el cálculo explicito de los conceptos previos para distribuciones Gaussianas multivariantes junto con alguna propiedad que las caracteriza.

Proposición 2.4.1. Sea $X \sim \mathcal{N}(\mu, \Sigma)$ un variable normal en $\mathbb{R}^n$. Entonces el valor de su entropía es $h(X) = \frac{1}{2}\log((2\pi e)^n |\Sigma_X|$.

Demostración. La función de densidad de probabilidad de la distribución normal multivariante se

escribe:

$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}. \tag{2.44}$$

Por tanto, la entropía tiene la siguiente forma:

$$\begin{aligned}
h(X) &= \int f(x)\log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}\right)dx \\
&= \int f(x)\log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\right)dx - \frac{1}{2}\int f(x)(x-\mu)^T\Sigma^{-1}(x-\mu)dx \\
&= \frac{1}{2}\log((2\pi)^n|\Sigma_X| - \frac{1}{2}\int f(x)\mathrm{tr}\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx \\
&= \frac{1}{2}\log((2\pi)^n|\Sigma_X| - \frac{1}{2}\int f(x)\mathrm{tr}\left(\Sigma^{-1}(x-\mu)^T(x-\mu)\right) \\
&= \frac{1}{2}\log((2\pi)^n|\Sigma_X| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma\right) = \frac{1}{2}\log((2\pi)^n|\Sigma_X| - \frac{n}{2} \\
&= \frac{1}{2}\log((2\pi)^n|\Sigma_X| - \frac{n}{2} = \frac{1}{2}\log((2\pi)^n|\Sigma_X| + \frac{1}{2}\log(e^{-n}) = \frac{1}{2}\log((2\pi e)^n|\Sigma_X|.
\end{aligned} \tag{2.45}$$

Donde se ha usado el intercambio del orden matricial en el interior de la traza. □

Proposición 2.4.2. Sea X una variable aleatoria cualquiera con media μ y matriz de covarianzas Σ. Sea por otro lado Z una variable normal con misma media y covarianza, $Z_X \sim \mathcal{N}(\mu, \Sigma)$. Entonces se tiene siempre que $h(X) \geq h(Z_X)$ cumpliéndose la igualdad si y solo si X es normal, es decir, $Z_X = X$.

Demostración. Calculamos en primer lugar la divergencia de Kullback-Leibler de f_X respecto de f_Z. (recordemos que el orden en el que se mencionan estas importa):

$$\begin{aligned}
D(f_X||f_Z) &= \int f_X\log(\frac{f_X}{f_Z}) = \int f_X\log(f_X) - \int f_X\log(f_Z) \\
&= h(X) - \int f_X\log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\right)dx - \frac{1}{2}\int f_X\mathrm{tr}((x-\mu)^T\Sigma^{-1}(x-\mu))dx \\
&= h(X) - \frac{1}{2}\log((2\pi)^n|\Sigma_X| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma\right) = h(X) - h(Z_X),
\end{aligned} \tag{2.46}$$

donde se ha usado que $\int f_X\log(f_Z) = \int f_Z\log(f_Z)$ al tener igual covarianza y media.

Finalmente, debido a la no negatividad de la divergencia de Kullback-Leibler. $0 \leq D(X||Z_X) = h(X) - h(Z_X) \implies h(X) \geq h(Z_X)$. De este mismo razonamiento y teniendo en cuenta que $0 = D(X||Z_X)$ exclusivamente si las distribuciones son iguales, se deduce la última afirmación de la propiedad. □

Este teorema es el análogo al caso discreto en el que se demuestra que la entropía alcanza el máximo para una probabilidad equiprobable en todos sus posibles sucesos.

Y además, la propia demostración nos proporcionan además dos corolarios que nos serán útiles en capítulos próximos.

Corolario 2.4.3. Sea X una variable aleatoria cualquiera con media μ y matriz de covarianzas Σ. Sea por otro lado Z una variable normal con misma media y covarianza, $Z_X \sim \mathcal{N}(\mu, \Sigma)$. Entonces se tiene siempre que $D(X||Z_X) = h(X) - h(Z_X)$.

El segundo de ellos nos dice que para cualquier variable con una determinada media y covarianza, la distribución normal es siempre la que tiene menor entropía y nos da una cota inferior para las entropías.

Corolario 2.4.4. Para cualquier variable aleatoria X con matriz de covarianzas Σ, su entropía está siempre acotada inferiormente por $\frac{1}{2}\log((2\pi e)^n|\Sigma_X|$.

Para terminar el capítulo incluimos el cálculo explícito de la **Divergencia de Kullback-Leibler** y de la e**Información Mutua** en el caso Gaussiano que serán utilizados de forma reiterada en capítulos posteriores.

Teorema 2.4.5. Sean dos distribuciones normales multivariantes $X_1 \sim N(\mu_1, \Sigma_1)$ y $X_1 \sim N(\mu_2, \Sigma_2)$ definidas ambas en $X_1, X_2 \in \mathbb{R}^n$, entonces su divergencia de Kullback-Leibler se puede escribir como sigue:

$$D_{\mathrm{KL}}(X_1||X_2) = \frac{1}{2}\left(\log\left(\frac{\det\Sigma_2}{\det\Sigma_1}\right) + \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1) - n\right). \tag{2.47}$$

Demostración. Sean $f_i(x) = \frac{1}{(2\pi)^{n/2}\det(\Sigma_i)}e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}$ las funciones de onda de las variables X_1 y X_2.

$$\begin{aligned} D_{\mathrm{KL}}(X_1||X_2) &= \int_{\mathbb{R}^n} f_1(x)\log\left(\frac{f_1(x)}{f_2(x)}\right)dx \\ &= \frac{1}{2}\int_{\mathbb{R}^n} f_1(x)\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right)dx - \frac{1}{2}\int_{\mathbb{R}^n} f_1(x)(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)dx + \\ &+ \frac{1}{2}\int_{\mathbb{R}^n} f_1(x)(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)dx). \end{aligned} \tag{2.48}$$

Analizamos los términos por separado:

$$\int_{\mathbb{R}^n} f_1(x)\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right)dx = \log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right)\int_{\mathbb{R}^n} f_1(x)dx = \log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right). \tag{2.49}$$

$$\begin{aligned} \int_{\mathbb{R}^n} f_1(x)(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)dx &= \int_{\mathbb{R}^n} f_1(x)\mathrm{Tr}\left((x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)dx \\ &= \int_{\mathbb{R}^n} f_1(x)\mathrm{Tr}\left(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T\right)dx \\ &= \mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_1\right)dx = \mathrm{Tr}\left(\mathbf{I}\right)dx = n, \end{aligned} \tag{2.50}$$

donde se ha usado que $E\left((x-\mu_1)(x-\mu_1)^T\right)=\Sigma_1$.

$$\begin{aligned}\int_{\mathbb{R}^n} f_1(x)(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)dx &= \int_{\mathbb{R}^n} f_1(x)((x-\mu_1)+(\mu_1-\mu_2))^T\Sigma_2^{-1}((x-\mu_1)+(\mu_1-\mu_2))dx\\ &= \int_{\mathbb{R}^n} f_1(x)(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)dx + \int_{\mathbb{R}^n} f_1(x)(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)dx -\\ &- \int_{\mathbb{R}^n} f_1(x)(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)dx - \int_{\mathbb{R}^n} f_1(x)(\mu_1-\mu_2)^T\Sigma_2^{-1}(x-\mu_1)dx =\\ &= \int_{\mathbb{R}^n} f_1(x)\mathrm{Tr}\left((x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)\right)dx + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\int_{\mathbb{R}^n} f_1(x)dx - 0 - 0\\ &= \int_{\mathbb{R}^n} f_1(x)\mathrm{Tr}\left(\Sigma_2^{-1}(x-\mu_1)(x-\mu_1)^T\right)dx + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) =\\ &= \mathrm{Tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2), \end{aligned} \tag{2.51}$$

donde se ha usado,

$$\int_{\mathbb{R}^n} f_1(x)(\mu_1-\mu_2)^T\Sigma_2^{-1}(x-\mu_1)dx = (\mu_1-\mu_2)^T\Sigma_2^{-1}\int_{\mathbb{R}^n} f_1(x)(x-\mu_1)dx = (\mu_1-\mu_2)^T\Sigma_2^{-1}\mathbf{0} = 0, \tag{2.52}$$

y análogamente para $\int_{\mathbb{R}^n} f_1(x)(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)dx = 0$.

Teniendo todo esto en cuenta y reemplazando en (4.13), se llega a lo que se quería probar:

$$D_{\mathrm{KL}}(X_1||X_2) = \frac{1}{2}\left(\log\left(\frac{\det\Sigma_2}{\det\Sigma_1}\right) + \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2-\mu_1)^T\,\Sigma_2^{-1}\,(\mu_2-\mu_1) - n\right). \tag{2.53}$$

□

Si lo aplicamos esta fórmula para calcular la información mutua se obtiene el siguiente corolario.

Corolario 2.4.6. Sean dos distribuciones normales $X_1 \sim N(\mu_1,\Sigma_1)$ y $X_1 \sim N(\mu_2,\Sigma_2)$ definidas ambas en $X_1, X_2 \in \mathbb{R}^n$, entonces la información mutua se escribe

$$I(X_1;X_2) = D_{KL}(P_{X_1,X_2}||P_{X_1}P_{X_2}) = \frac{1}{2}\log\left(\frac{\det(\Sigma_{X_1})\det(\Sigma_{X_2})}{\det(\bar{\Sigma}_{X_1,X_1})}\right). \tag{2.54}$$

Demostración.

$$P_{X_1,X_2} \sim N(\mu,\Sigma_A) \text{ donde } \mu_A = \begin{bmatrix}\mu_{X_1}\\ \mu_{X_2}\end{bmatrix} \text{ y } \Sigma_A = \begin{bmatrix}\Sigma_{X_1} & \Sigma_{X_1X_2}\\ \Sigma_{X_1X_2}^T & \Sigma_{X_2}\end{bmatrix} = \bar{\Sigma}_{X_1,X_2}. \tag{2.55}$$

$$P_{X_1}P_{X_2} \sim N(\mu_B,\Sigma_B) \text{ donde } \mu_B = \begin{bmatrix}\mu_{X_1}\\ \mu_{X_2}\end{bmatrix} \text{ y } \Sigma_B = \begin{bmatrix}\Sigma_{X_1} & 0\\ 0 & \Sigma_{X_2}\end{bmatrix}. \tag{2.56}$$

Teniendo esto en cuenta y usando que:

$$\Sigma_B^{-1} = \begin{bmatrix}\Sigma_{X_1}^{-1} & 0\\ 0 & \Sigma_{X_2}^{-1}\end{bmatrix} \quad \Longrightarrow \quad Tr(\Sigma_B^{-1}\Sigma_A) = Tr(\mathbb{I}) = n. \tag{2.57}$$

$$I(X_1;X_2) = \frac{1}{2}\left(\log\left(\frac{\det(\Sigma_{X_1})\det(\Sigma_{X_2})}{\det(\bar{\Sigma}_{X_1,X_2})}\right) + \mathrm{Tr}(\mathbb{I}) + (\mu-\mu)^T\,\Sigma_2^{-1}\,(\mu-\mu) - n\right) =$$
$$= \frac{1}{2}\log\left(\frac{\det(\Sigma_{X_1})\det(\Sigma_{X_2})}{\det(\bar{\Sigma}_{X_1,X_2})}\right). \tag{2.58}$$

□

Capítulo 3

Desigualdades importantes en la Teoría de la Información

En capítulos posteriores se presentará una serie de modelos en relación al aprendizaje supervisado. Para abordar estos problemas y comprenderlos ampliamente es necesario conocer previamente unos resultados relacionado con la información mutua y la entropía diferencial conocidos como **Desigualdad de Procesado de Datos** y **Desigualdad de potencias de entropía**.

A diferencia de las propiedades y resultados obtenidos en el capítulo previo que se deducen de las propias definiciones de los diferentes conceptos, en esta parte del documento se van a explicar dos teorema que necesitan de un desarrollo de mayor complejidad y que además permiten llegar a conclusiones importantes con lo que respecta al aprendizaje automático o incluso otros ámbitos de las Matemáticas [5, 6, 14, 22, 23].

3.1. Desigualdad de procesado de datos

Esta desigualdad, conocida en inglés como **Data Processing Inequality**, se basa en la idea intuitiva de que cualquier trasformación que se haga a una serie de datos, reducirá la información que estos tenían inicialmente. En otras palabras y en relación al aprendizaje supervisado, si uno parte de una serie de datos $\{(x_i)\}_{i=1}^n$ con los que predecir $\{(y_i)\}_{i=1}^n$, cualquier trasformación que se haga cambiando esos, va a afectar a la cantidad de información que estos proporcionaban previamente.

Para dar paso a la expresión formal de la **Desigualdad de Procesado de Datos**, en primer lugar hay que entender el escenario en el que se esta trabajando que es muy similar al del aprendizaje supervisado. Este parte entonces de unos datos de entrenamiento que siguen una distribución conjunta desconocida descrita por los vectores aleatorios (X, Y). La idea es entonces ver lo que ocurre con la información disponible en caso de hacer una trasformación de los atributos, es decir, del vector aleatorio X.

Esta trasformación será otro vector aleatorio al que denotaremos por T y que no dependerá de Y en caso de que X sea conocido. Esto matemáticamente se expresa como $P(T|X, Y) = P(T|X)$ y da lugar a lo que formalmente se conocen como Cadenas de Markov y que veremos a continuación.

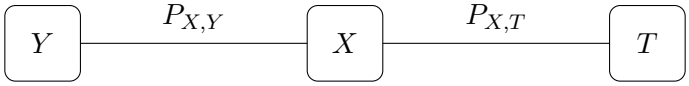
Definición 3.1.1. Una Cadena de Markov comúnmente se conoce como un proceso estocástico, es decir, una colección de variables aleatorias $(X_t)_{t\in T}$ indexadas por un conjunto $T = 0, 1, 2, ..., n$ y definidas en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, que cumple la siguiente condición: para cualquier entero $n \geq 0$ y para cualesquiera $x_0, x_1, \ldots, x_{n+1} \in \Omega$ se tiene que

$$P[X_{n+1} = x_{n+1}|X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n] = P[X_{n+1} = x_{n+1}|X_n = x_n]. \tag{3.1}$$

En nuestro caso, el concepto de cadena de Markoz se expande a una secuencia de tres variable, no necesariamente definidas en al mismo espacio que cumplen la siguiente propiedad:

$$P(T|X, Y) = P(T|X)\,, \tag{3.2}$$
$$P(Y|X, T) = P(T|X). \tag{3.3}$$

Es decir, son independientes con respecto a la variable más alejada de ellas.



Esto nos permite llegar a una serie de conclusiones importantes:

$$P(Y, T|X)P(X) = P(X)P(T)P(Y|T, X) = P(X)P(T)P(Y|X) = P(T)P(Y, X)\,, \tag{3.4}$$
$$P(Y, T|X)P(X) = P(X)P(Y)P(T|Y, X) = P(X)P(Y)P(T|X) = P(Y)P(T, X). \tag{3.5}$$

Es importante darse cuenta que esta condición no es equivalente a decir que Y y T son independientes, es más, el hecho de que sean independientes es justo lo que no se desea ya que en ese caso $I(Y; T) = 0$ y esto implica que la trasformación realizada no aporta información útil para la predicción.

Una vez se ha entendido el escenario en el que estamos trabajando, pasamos ya a enunciar y explicar correctamente la **Desigualdad de Procesado de Datos**.

Como se ha mencionado previamente, la información que conocer T nos da de Y, no puede ser mayor que la que nos da X. Formalmente esto se escribe como enuncia el siguiente teorema.[11]

Teorema 3.1.2. Dados tres variables aleatorias X, Y y T que forman la siguiente cadena de Markov $Y \leftrightarrow X \leftrightarrow T$, se cumple la siguiente desigualdad conocida **Desigualdad de Procesado de Datos**.

$$\mathbf{I}(\mathbf{Y}; \mathbf{X}) \geq \mathbf{I}(\mathbf{T}; \mathbf{Y}). \tag{3.6}$$

Esto es resultado directo de que Y, X y T formen una cadena de Markov ya que nos permite usar las Proposición 2.3.14 y 2.3.15.

Demostración. Usando la regla de la cadena para la información mutua se tiene que:

$$I(Y; X, T) = I(Y; X) + I(Y; T|X)\,, \tag{3.7}$$
$$I(Y; X, T) = I(Y; T) + I(Y; X|T)\,, \tag{3.8}$$

Si además, le aplicas las Proposición 2.3.15, se tiene que $I(Y;T|X) = 0$ y por tanto:

$$I(Y;X) = I(Y;T) + I(Y;X|T) \geq I(Y;T), \tag{3.9}$$

donde se ha usado que $I(Y;X|T) \geq 0$. □

Intuitivamente y para una comprensión de la desigualdad en la dirección en la que la vamos a aplicar, esta cantidad de información que estemos dispuesto a perder al aplicar la trasformación de X a T viene dada por $I_p := I(Y;X) - I(Y;T) \geq 0$, donde I_p representa esa perdida de información que se asume.

3.2. Desigualdad de potencias de entropía

Esta desigualdad generalmente es conocida por sus siglas en inglés **Entropy Power Inequality** como **EPI** y previo a su presentación, es necesario conocer la definición de **potencia de entropía**.

Definición 3.2.1 (**Potencia de Entropía**)**.** Este concepto de *entropia potencia* o *entropy power* en inglés se define como:

$$N(X) = \frac{1}{2\pi e} e^{\frac{2}{n}h(X)}, \tag{3.10}$$

que como su propio nombre indica, se trata de la potencia en base natural de la entropía diferencial.

Teorema 3.2.2 (**EPI**)**.** La **EPI** afirma que dados dos vectores aleatorios X e Y independientes entre sí y con funciones de densidad en $\mathbb{R}^n$. Entonces la desigualdad es la siguiente:

$$N(X+Y) \geq N(X) + N(Y). \tag{3.11}$$

y la igualdad se cumple si y solo si X e Y son distribuciones Gaussianas con matrices de covarianza proporcionales.[1]

A lo largo de la historia se han dado numerosas versiones equivalentes para este teorema. Estas permiten demostrar y aplicar la desigualdad de diferentes formas. En un intento de resumir estas, la siguiente proposición engloba tres de las formas más comunes y las que se utilizarán a continuación.

Proposición 3.2.3. Dadas $\{X_i\}_{i=1}^n$ variables aleatorias independientes con funciones de densidad en $\mathbb{R}^n$, las siguientes desigualdades son equivalentes entre sí, siendo la primera de ellas la **EPI**:

(a) $N(\sum_i a_i X_i) \geq \sum_i a_i^2 N(X_i) = \sum_i N(a_i X_i)$.

(b) $h(\sum_i a_i X_i) \geq h(\sum_i a_i \hat{X}_i)$ donde $\hat{X}_i$ son variables normales independientes con $h(\hat{X}_i) = h(X_i)$ con matrices de covarianzas Σ_i proporcionales entre sí. Nótese que esto es posible eligiendo $X_i \sim N(0, \sqrt[n]{\frac{e^{h(X_i)}}{(2\pi e)^n}} I)$ ya que en ese caso $h(\hat{X}_i) = \frac{1}{2} log(2\pi e)^n |\Sigma_i| = h(X_i)$ y todas ellas son proporcionales a la identidad.

[1] Esta condición de trabajar en $\mathbb{R}^n$ no es totalmente necesaria, pero debido a que el tipo de distribuciones que se suelen manejar y las que van a aparecer a lo largo del documento son de este tipo, reduciremos el teorema a este único caso.

(c) $h(\sum_i a_i X_i) \geq \sum_i a_i^2 h(X_i)$ siempre que $\sum_i a_i^2 = 1$.

Demostración. ($a \implies b$) Como el logaritmo es una función creciente podemos aplicar este a (a) de forma que se siga manteniendo la igualdad:

$$\log(N(\sum_i a_i X_i) = -\log(2\pi e) + \frac{2}{n} h(\sum_i a_i X_i). \tag{3.12}$$

Sean ahora $\hat{X}_i$ las variables que se enuncian en (b).

$$a_i^2 N(X_i) = \sum_i N(a_i X_i) = \sum_i N(a_i \hat{X}_i) = N(\sum_i a_i \hat{X}_i). \tag{3.13}$$

Entonces aplicando logaritmos:

$$\log(a_i^2 N(X_i)) = \log(N(\sum_i a_i \hat{X}_i)) = -\log(2\pi e) + \frac{2}{n} h(\sum_i a_i \hat{X}_i). \tag{3.14}$$

Y comparando los términos a los que hemos llegado se termina la demostración.

($b \implies a$) Análogamente al caso anterior pero aplicando en este caso la exponencial que es también una función creciente.

$$N(\sum_i a_i X_i) = \frac{1}{2\pi e} e^{\frac{2}{n} h(\sum_i a_i X_i)} \geq \frac{1}{2\pi e} e^{\frac{2}{n} \sum_i h(a_i \hat{X}_i)} = N(\sum_i a_i \hat{X}_i) = \sum_i a_i^2 N(X_i). \tag{3.15}$$

($a \implies c$) Se va a ser usar de la concavidad del logaritmo $f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$, a variables variables ya que se cumple $\sum a_i^2 = 1$, es decir,$log(\sum_i a_i^2 N(X_i)) \geq \sum_i a_i^2 \log(N(X_i))$

$$\begin{aligned} \log(N(\sum_i a_i X_i)) &\geq \log(\sum_i a_i^2 N(X_i)) \\ -\log(2\pi e) + \frac{n}{2} h(\sum a_i X_i) \geq \sum_i a_i^2 \log(N(X_i)) &= -\sum_i a_i^2 \log(2\pi e) + \sum_i \frac{n}{2} a_i^2 h(X_i). \end{aligned} \tag{3.16}$$

De forma que al tener $\sum_i a_i^2 = 1$ y reorganizando términos:

$$h(\sum a_i X_i) \geq \sum_i a_i^2 h(X_i). \tag{3.17}$$

($c \implies a$) En primer lugar, remarcamos que aunque (a) no imponga que $\sum_i a_i^2 = 1$, podemos demostrarlo bajo esta hipótesis ya que si se cumple en este único caso, se pueden elegir b_i de forma que

$$N(\sum_i a_i X_i) = N(\sum_i b_i \frac{a_i}{b_i} X_i) \geq \sum_i b_i^2 N(\frac{a_i}{b_i} X_i) = \sum_i b_i^2 \frac{a_i^2}{b_i^2} N(X_i) = \sum_i a_i^2 N(X_i). \tag{3.18}$$

y así se cumpla para cualquier caso.

Este mismo razonamiento, nos permite demostrarlo para las variables $Y_i = N(X_i)^{-n/2}X_i$:

$$N(\sum_i a_i X_i) = N(\sum_i \frac{a_i}{N(X_i)^{-n/2}} N(X_i)^{-n/2} X_i) \geq \sum_i \frac{a_i^2}{N(X_i)^{-n}} N(N(X_i)^{-n/2} X_i)) = \sum_i \frac{a_i^2}{N(X_i)^{-n}} N(X_i)^{-n} N(X_i)) = \sum_i a_i^2 N(X_i)). \tag{3.19}$$

El caso para las variables Y_i es muy sencillo ya que $h(Y_i) = h(X_i) - \frac{n}{2}\log(N(x_i)) = h(X_i) - \frac{n}{2}\log(2\pi e) - h(X_i) = -\frac{n}{2}\log(2\pi e)$ que es constante e igual para todo i y por tanto aplicando exponenciales a ambos lados de (c) y usando $\sum_i a_i^2 = 1$:

$$e^{\frac{2}{n}h(\sum_i a_i X_i)} \geq e^{\frac{2}{n}\sum_i h(a_i X_i)} = e^{\frac{2}{n}h(X_1)} = 2\pi e N(X_1) = 2\pi e \sum_i a_i^2 N(X_i). \tag{3.20}$$

□

Aunque no se ha demostrado aún la desigualdad, vamos a demostrar previamente la última afirmación del Teorema 3.2.2. Es decir, suponiendo que es cierta la **EPI** y por tanto sus formas equivalentes, veamos que la igualdad se cumple si y solo si X_i son variables normales con matrices de covarianza proporcionales.

Lema 3.2.4. La igualdad del Teorema 3.2.2 se cumple si y solo si las variables $\{X_i\}_{i=1}^n$ son variables normales con matrices de covarianza proporcionales.

Demostración. Supongamos que la **EPI** es cierta y por tanto se cumplen cualquiera de los tres enunciados equivalentes.

En primer lugar veamos que si son variables normales independientes $\{X_i\}_{i=1}^p$ en dimensión n con matrices de covarianza proporcionales ($\Sigma_i = b_i\Sigma_0$) se cumple la igualdad. Para ello, observamos que al ser suma de variables normales independientes $a_i X_i \sim \mathcal{N}(a_i\mu_i, a_i^2\Sigma_i)$ y por tanto $Y := \sum_i a_i X_i \sim \mathcal{N}(\sum_i a_i\mu_i, \sum_i a_i^2\Sigma_i)$

$$N(\sum_i a_i X_i) = N(Y) = |\sum_i a_i^2\Sigma_i|^{\frac{1}{n}} = |(\sum_i a_i^2 b_i)\Sigma_0|^{\frac{1}{n}} = (\sum_i a_i^2 b_i)^{\frac{n}{n}}|\Sigma_0|^{\frac{1}{n}} = (\sum_i a_i^2 b_i)|\Sigma_0|^{\frac{1}{n}}. \tag{3.21}$$

$$\sum_i a_i^2 N(X_i) = \sum_i a_i^2 b_i^{\frac{n}{n}}|\Sigma_0|^{\frac{1}{n}} = (\sum_i a_i^2 b_i)|\Sigma_0|^{\frac{1}{n}}. \tag{3.22}$$

donde se ha usado la propiedad de los determinantes de orden n, $|cA| = c^n|A|$.

Veamos ahora la implicación contraria y para ello vamos a razonar por reducción al absurdo. Supongamos que existen $\{X_i\}_{i=1}^p$ cualesquiera e independientes tales que cumplen la igualdad. Por tanto cumplen (b): $h(\sum_i a_i X_i) = h(\sum_i a_i \hat{X}_i)$ donde $\hat{X}_i$ son variables normales independientes con $h(\hat{X}_i) = h(X_i)$ con matrices de covarianzas Σ_i proporcionales entre si. Según el Corolario 3.7.2, esta igualdad solo es posible en caso de que ambas sean la misma distribución Gaussiana de matriz de covarianza total $\sum_i a_i^2\Sigma_i = (\sum_i a_i^2 b_i)\Sigma_0$. Esto, añadiendo el hecho de que su determinante tiene que ser el mismo para tener misma entropía diferencial, implica que el determinante de la suma sea la suma de los determinantes lo cual se cumple exclusivamente para matrices proporcionales. □

3.2.1. Resultados previos a la demostración de la EPI

Previa a la demostración de la **EPI**, es necesario conocer una serie de resultados esenciales que se presentan en los lemas a continuación.

Lema 3.2.5. Sea $X \sim \mathcal{N}(\mu, \Sigma)$, entonces $N(X) = |\Sigma|^{\frac{1}{n}}$.

Demostración. Se deduce directamente de la definición partiendo del cálculo hecho para la entropía diferencial, $h(X) = \frac{1}{2}\log((2\pi e)^n|\Sigma_X|$.

$$N(X) = \frac{1}{2\pi e}e^{\frac{2}{n}h(X)} = \frac{1}{2\pi e}e^{\frac{2}{n}\frac{1}{2}\log((2\pi e)^n|\Sigma|} = \frac{1}{2\pi e}e^{\frac{2}{n}\frac{1}{2}\log((2\pi e)^n|\Sigma|} = |\Sigma|^{\frac{1}{n}}. \tag{3.23}$$

□

Lema 3.2.6. Sea X una variable aleatoria cualquiera y $a \in \mathbb{R}$, entonces se tiene $N(aX) = a^2N(X)$

Demostración. Se deduce directamente de la definición partiendo del cálculo hecho para la entropía diferencial, $h(aX) = h(X) + n\log(|a|)$.

$$N(aX) = \frac{1}{2\pi e}e^{\frac{2}{n}h(aX)} = \frac{1}{2\pi e}e^{\frac{2}{n}h(X)+\log(a^2)} = a^2N(X). \tag{3.24}$$

□

Lema 3.2.7. Dado X e Y variables aleatorias independientes, la información mutua de $X+Y$ y X se puede escribir como:

$$I(X+Y;X) = h(X+Y) - h(Y). \tag{3.25}$$

Este hecho y teniendo en cuenta una de las propiedades fundamentales de la información mutua $I(U;V) = h(U) - h(U|V)$, equivale a decir que $h(X+Y|Y) = h(X)$. Esto es bastante lógico ya que la entropía de $X+Y$ conocido Y, se reduce únicamente a la entropía de X.

Demostración. Llevando a cabo el siguiente cambio de variable $(X;Y) \to (X;X+Y) := (U,V)$, donde el valor absoluto del determinante del jacobiano es 1, nos permite escribir la función de densidad conjunta como $f_{X,X+Y}(u,v) = f_{X,Y}(x,y) = f_X(x)f_Y(y) = f_X(u)f_Y(v-u)$.[Theorem A.4.1] Por tanto:

$$\begin{aligned} I(X+Y;X) &= \int\int f_X(u)f_Y(v-u)\log\left(\frac{f_X(u)f_Y(v-u)}{f_X(u)f_{X+Y}(v)}\right)dudv \\ &= \int f_X(u)\int f_Y(v-u)\log(f_Y(v-u)dvdu - \int \log(f_{X+Y}(v))\int f_X(u)f_Y(v-u)dudv \\ &= -\int f_X(u)h(Y)du - \int f_{X+Y}(v)\log(f_{X+Y}(v))dv = -h(Y) + h(X+Y). \end{aligned} \tag{3.26}$$

Donde se ha usado que $f_{X+Y} = \int f_X(u)f_Y(v-u)du$, es decir, calculo de la distribución marginal. □

Lema 3.2.8. Dado X e Y variables aleatorias independientes, la entropía diferencial de $aX+Y$ e conocido aX es igual a la la entropía diferencial de $aX+Y$ conocido X.

$$h(aX+Y|X) = h(aX+Y|X) = h(Y). \tag{3.27}$$

Esto nos permite deducir inmediatamente haciendo uso del lema previo que $I(aX+Y;X) = \\ = h(aX+Y) - h(Y) = I(aX+Y|aX)$

Demostración. Se va a usar que $f_{W+T|W}(u) = \frac{f_{W+T,W}(u,v)}{f_W(u)} = \frac{f_W(u)f_T(u-v)}{f_W(u)} = f_T(u-v)$ [Theorem A.4.1]

$$h(aX+Y|aX) = \int\int f_{aX}(u)f_Y(v-u)\log f_Y(v-u)dudv = \int f_{aX}(u)h(Y)du = h(Y). \tag{3.28}$$

$$h(aX+Y|X) = \int\int \frac{1}{|a|}f_X(\frac{u}{a})f_Y(v-\frac{u}{a})\log f_Y(v-\frac{u}{a})dudv = \int \frac{1}{|a|}f_X(\frac{u}{a})h(Y)du = h(Y). \tag{3.29}$$

□

Lema 3.2.9 (Desigualdad de Sato). Dados $(X_i)_i^n$ variables aleatorias independientes entre sí y de $Z \sim \mathcal{N}(0,1)$:

$$I((X_i+Z)_i;Z) \leq \sum_i I(X_i+Z;Z). \tag{3.30}$$

Demostración. Denotamos $Y_i = X_i + Z$ y haciendo uso de la regla de la cadena para la información mutua,

$$\begin{aligned} I(Y_1,...,Y_n;Z) &= \sum_i I(Y_i;Z|Y_{i-1},...,Y_1) \leq \\ &\leq \sum_i I(Y_i;Z|Y_{i-1},...,Y_1) + I(Y_i;Y_{i-1},...,Y_1) \\ &= \sum_i I(Y_i;Z,Y_{i-1},...,Y_1) \\ &= \sum_i I(Y_i;Z) + I(Y_i;Y_1,...,Y_{i-1}|Z) \\ &= \sum_i I(Y_i;Z). \end{aligned} \tag{3.31}$$

Donde en el primer paso se ha usado la regla de la cadena, posteriormente se ha añadido un término positivo y de ahí la desigualdad. A continuación, se reagrupan los términos usando también la regla de la cadena pero para $n=2$ y luego se vuelven a desagrupar pero invirtiendo las variables. Finalmente, en el último paso $I(Y_i;Y_1,...,Y_{i-1}|Z)$ ya que conocido Z, $Y_i = X_i + Z$ es independiente de Y_j al serlo X_i y X_j con $i \neq j$. □

Lema 3.2.10. Si X y Z son vectores aleatorios independientes entonces:

$$\lim_{t\to 0^+} I(X+\sqrt{t}Z;Z)=0. \tag{3.32}$$

Si, además, $I(X+\sqrt{t}Z;Z)$ es diferenciable en $t=0$, entonces para cualquier constante real a

$$I(X+a\sqrt{t}Z;Z)=a^2 I(X+\sqrt{t}Z;Z)+o(t). \tag{3.33}$$

donde $o(t)$ es una función definida para todo $t\geq 0$ tal que $\frac{o(t)}{t}\to 0$ cuando $t\to 0^+$.

Demostración. Dado que X y $\sqrt{t}Z$, son independientes la función característica de la suma $X_t := X+\sqrt{t}Z$ es el producto de las funciones características. Además, la función característica de la distribución normal ($\psi_Z(u)=e^{-\frac{\sigma^2 u^2}{2}}e^{i\mu u}$) cumple que $\psi_{\sqrt{t}Z}(u)=\psi_Z(\sqrt{t}u)$.

$$\psi_{X_t}(u)=\psi_X(u)\psi_{\sqrt{t}Z}(u)=\psi_X(u)\psi_Z(\sqrt{t}u)\xrightarrow[t\to 0^+]{}\psi_X(u)\psi_Z(0)=\psi_X(u)e^0=\psi_X(u). \tag{3.34}$$

Por tanto, como la función característica de X_t convergen la de X, tenemos que X_t converge en distribución a X.[Theorem A.5.3]

Se define $X^*\sim\mathcal{N}(\mu_x,\Sigma_X)$ donde (μ_x,Σ_X) son la media y covarianza de X respectivamente y también $X_t^* := X^*+\sqrt{t}Z$ y de la misma forma que antes, se prueba que X_t^* converge en distribución hacia X^*.

$$D(X_t||X_t^*)-D(X||X^*)=h(X_t^*)-h(X_t)+h(X)-h(X^*)=h(X^*+\sqrt{t}Z)-h(X^*)-I(X_t;Z). \tag{3.35}$$

donde se ha usado el Corolario 3.7.3 y el Lema 5.2.3.

Ahora, teniendo en cuenta que la información mutua es siempre positiva y que $\lim_{t\to 0^+} h(X^*+\sqrt{t}Z)=\lim_{t\to 0^+}\frac{1}{2}\log(2\pi e)^n|\Sigma_X+tI|=\frac{1}{2}\log(2\pi e)^n|\Sigma_X|=h(X^*)$:

$$\begin{aligned}
0 &\geq \lim_{t\to 0^+}\sup(-I(X+\sqrt{t}Z;Z))=\\
&=\lim_{t\to 0^+}\sup(D(X_t||X_t^*)-D(X^*||X^*))-(h(X^*)-h(X^*+\sqrt{t}Z))=\\
&=\lim_{t\to 0^+}\sup(D(X_t||X_t^*)-D(X^*||X^*))\geq\\
&\geq\lim_{t\to 0^+}\text{ínf}(D(X_t||X_t^*))-D(X^*||X^*).
\end{aligned} \tag{3.36}$$

El siguiente paso se basa en la semicontnuidad inferior de la divergencia de Kullback que afirma que si $X_n\xrightarrow{d}X$ y $Y_n\xrightarrow{d}Y$, entonces $\lim_{n\to\infty}\text{ínf}\, D(X_n||Y_n)\geq D(X||Y)$. Léase [Section A.6]y [22] para la demostración detallada de esta semicontinuidad inferior. Aplicado a nuestro caso, se tiene $\lim_{n\to 0^+}\text{ínf}\, D(X_t||X_t^*)\geq D(X||X^*)$ y por tanto,

$$0\geq\lim_{t\to 0^+}\sup(-I(X+\sqrt{t}Z;Z))\geq\lim_{t\to 0^+}\text{ínf}(D(X_t||X_t^*))-D(X^*||X^*)=0. \tag{3.37}$$

Concluyendo la primera parte de lo que se quería probar, $\lim_{t\to 0^+} I(X+\sqrt{t}Z;Z)=0$ ya que también se ha de cumplir que $\lim_{t\to 0^+}\text{ínf}(I(X+\sqrt{t}Z;Z))\geq 0$ por la no negatividad de la información mutua y por tanto, los limites superior e inferior coinciden.

Para la segunda parte del lema, supongamos que $I(X+\sqrt{t}Z;Z)$ es diferenciable en $t=0$ (al ser en una variable implica que es derivable) y podemos escribir la información como:

$$I(X+\sqrt{t}Z;Z) = I(X;Z) + \left.\frac{\partial I(X+\sqrt{t}Z;Z)}{\partial t}\right|_{t=0} + o(t) = \left.\frac{\partial I}{\partial t}\right|_{t=0} + o(t). \tag{3.38}$$

donde se ha usado que $I(X;Z)=0$ al ser independientes.

Análogamente,

$$I(X+a\sqrt{t}Z;Z) = I(X;Z) + \left.\frac{\partial I(X+a\sqrt{t}Z;Z)}{\partial a^2 t}\right|_{t=0} t + o(a^2t) = \left.\frac{\partial I}{\partial a^2 t}a^2 t\right|_{t=0} + o(t). \tag{3.39}$$

Igualando con ambas expresiones $\left.\frac{\partial I(X+a\sqrt{t}Z;Z)}{\partial a^2 t}\right|_{t=0} = \left.\frac{\partial I(X+\sqrt{t}Z;Z)}{\partial t}\right|_{t=0}$, se obtiene el resultado al que se quería llegar:

$$I(X+a\sqrt{t}Z;Z) = a^2 I(X+\sqrt{t}Z;Z) + o(t). \tag{3.40}$$

□

A continuación, se van a presentar una serie de resultados característicos de las distribuciones Gaussianas y de la distribución de $X+\sqrt{t}Z$ necesarios para demostrar la diferenciabilidad de $h(X+\sqrt{t}Z)$ en $t>0$ y posterior aplicabilidad en el estudio de la derivabilidad de $I(X+\sqrt{t}Z;Z)$ en $t=0$ [5]. Estos resultados hacen uso del teorema de derivación bajo el signo integral, el cual se enuncia debidamente en [Appendix A.2]. Además, se van a presentar y demostrar únicamente para dimensión uno para evitar una notación excesiva y demostraciones engorrosas que nos impidan ver el objetivo principal de estas. Sin embargo, si uno quisiera demostrarlo para dimensiones mayores, simplemente habría que ajustar las constantes de la distribución normal multivariante de $\sqrt{2\pi t}$ a $(2\pi\sigma^2)^{n/2}$, las derivadas primeras por $\frac{\partial}{\partial x_i}$, y las derivadas segundas por $\frac{\partial^2}{\partial x_i^2}$ y ∇^2.

Lema 3.2.11 (**Ecuación de trasmisión del calor para distribuciones Gaussianas**)**.** La función de densidad de una distribución normal centrada y de varianza t, $\phi_t(x) = \frac{1}{\sqrt{2\pi t}}e^{-\frac{x^2}{2t}}$ cumple la siguiente ecuación para todo $t>0$ y para todo $x\in\mathbb{R}$:

$$\frac{\partial \phi_t}{\partial t} = \frac{1}{2}\frac{\partial^2 \phi_t}{\partial x^2}. \tag{3.41}$$

Esta ecuación diferencial en derivadas parciales escrita, $\frac{\partial}{\partial t}\phi(x,t) = \alpha\frac{\partial^2}{\partial x^2}\phi(x,t)$, se conoce como la ecuación de difusión o ecuación del calor. Esta representa un ecuación fundamental en diversos campos de las matemáticas pero también en la física ya que describe la evolución temporal de la temperatura, donde la constante α determina la velocidad a la que evoluciona el sistema.

Demostración. La demostración se deduce simplemente de derivar:

$$\frac{\partial \phi_t}{\partial t} = \left(-\frac{1}{2\sqrt{2\pi}t^{3/2}} + \frac{x^2}{2\sqrt{2\pi}t^{5/2}}\right)e^{-\frac{x^2}{2t}} = \left(-\frac{1}{2t} + \frac{x^2}{2t^2}\right)\frac{1}{2\sqrt{2\pi}}e^{-\frac{x^2}{2t}} = \frac{1}{2t}\left(\frac{x^2}{t}-1\right)\phi_t(x). \tag{3.42}$$

$$\frac{\partial \phi_t}{\partial x} = \frac{1}{\sqrt{2\pi t}} \frac{-x}{t} e^{-\frac{x^2}{2t}} = -\frac{x}{t}\phi_t(x). \tag{3.43}$$

$$\frac{\partial^2 \phi_t}{\partial x^2} = -\frac{1}{t}\phi_t(x) - \frac{x}{t}\frac{\partial \phi_t}{\partial x} = \frac{1}{t}\left(\frac{x^2}{t} - 1\right)\phi_t(x). \tag{3.44}$$

Hallándose la desigualdad que se quería obtener.

Antes de pasar a otros lemas, al estar ya tratando con esta función normal centrada, y de cara a aplicar en los siguientes apartados el teorema de la derivación de integrales parametricas [Appendix A.2], veamos una serie de acotaciones que nos serán útiles:

$$|\frac{\partial \phi_t}{\partial x}| = |\frac{x}{t}\phi_t(x)| \leq \frac{1}{\sqrt{2\pi e t^2}}. \tag{3.45}$$

$$|\frac{\partial \phi_t}{\partial x}| = |\frac{x}{t}\phi_t(x)| \leq 2(et)^{1/2}\phi_{2t}(x). \tag{3.46}$$

$$|\frac{\partial \phi_t}{\partial t}| = |\frac{1}{2}\frac{\partial^2 \phi_t}{\partial x^2}| = \frac{1}{\sqrt{8\pi t^3}} \leq \frac{1}{\sqrt{8\pi t^3}}. \tag{3.47}$$

$$|\frac{\partial \phi_t}{\partial t}| = |\frac{1}{2}\frac{\partial^2 \phi_t}{\partial x^2}| = |\frac{1}{2t}\left(\frac{x^2}{t} - 1\right)\phi_t(x)| \leq 2\sqrt{2}(et)^{-2}\phi_{2t}(x). \tag{3.48}$$

La demostración de estas cuatro propiedades se deduce simplemente del estudio de máximos y mínimos de las funciones presentes. A su vez, estas funciones que dominan a las derivadas parciales, son funciones Gaussianas y por tanto integrables. □

Si ahora definimos como $g_t(y)$ a la función de densidad de de $X + \sqrt{t}Z$, al ser suma de variables independientes, se puede escribir según la convolución

$$g_t(y) = \int f_X(x)\phi_t(y-x)dx = E_X(\phi_t(y - X)). \tag{3.49}$$

Lema 3.2.12. La función de densidad $g_t(y)$ también cumple la ecuación de difusión.

$$\frac{\partial g_t(y)}{\partial t} = \frac{1}{2}\frac{\partial^2 g_t(y)}{\partial y^2}. \tag{3.50}$$

Demostración. La idea principal para la demostración se basa en demostrar la dominación de las derivadas parciales por una función integrable en $\mathbb{R}$ par así poder aplicar el teorema derivación de integrales paramétricas.

En primer lugar, se tiene que $|\frac{\partial \phi_t}{\partial t}| \leq \frac{1}{\sqrt{8\pi t^3}}$ y por tanto, $|\frac{\partial f_x \phi_t}{\partial t}| = f_x(x)|\frac{\partial \phi_t}{\partial t}| \leq f_x(x)\frac{1}{\sqrt{8\pi t^3}}$. Para que se cumple en un entorno de t, trabajamos en $0 < a < t < b$ y así se tiene $|\frac{\partial f_x \phi_t}{\partial t}| \leq f_x(x)\frac{1}{\sqrt{8\pi a^3}}$ que es integrable al ser f_X una función de densidad. Podemos así aplicar el teorema

$$\frac{\partial g_t(y)}{\partial t} = \frac{\partial}{\partial t}\int f_X(x)\phi_t(y-x)dx = \int f_X(x)\frac{\partial}{\partial t}\phi_t(y-x)dx = E_X(\frac{\partial}{\partial t}\phi_t(y - X)). \tag{3.51}$$

Razonando análogamente para la primera y segunda derivada respecto de y:

$$\left|\frac{\partial f_x(x)\phi_t(y-x)}{\partial y}\right| = f_x(x)\left|\frac{\partial \phi_t(y-x)}{\partial y}\right| \leq f_x(x)\frac{1}{2\pi e a^2}. \tag{3.52}$$

$$\left|\frac{\partial^2 f_x(x)\phi_t(y-x)}{\partial y^2}\right| = f_x(x)\left|\frac{\partial^2 \phi_t(y-x)}{\partial y^2}\right| \leq f_x(x)\frac{1}{\sqrt{8\pi a^3}}. \tag{3.53}$$

Y por tanto, aplicando el teorema de derivación bajo el signo integral dos veces consecutivas:

$$\frac{\partial^2 g_t(y)}{\partial y^2} = \frac{\partial^2}{\partial y^2}\int f_X(x)\phi_t(y-x)dx = \int f_X(x)\frac{\partial^2}{\partial y^2}\phi_t(y-x)dx = E_X(\frac{\partial^2}{\partial y^2}\phi_t(y-X)). \tag{3.54}$$

Finalmente, aplicando la ecuación del calor para el caso Gaussiano se llega a lo que se quería probar inicialmente:

$$\frac{\partial g_t(y)}{\partial t} = E_X(\frac{\partial}{\partial t}\phi_t(y-X)) = E_X(\frac{1}{2}\frac{\partial^2}{\partial y^2}\phi_t(y-X)) = \frac{1}{2}\frac{\partial^2 g_t(y)}{\partial y^2}. \tag{3.55}$$

□

El siguiente lema, que prueba finalmente la derivabilidad de la entropía diferencial $h(X+\sqrt{t}Z)$ necesita también de una serie de acotaciones necesarias para aplicar el teorema de derivación bajo el signo integral

Lema 3.2.13 (**Derivabilidad de** $h(X+\sqrt{t}Z)$)**.** La entropía $h(X+\sqrt{t}Z)$ es diferenciable con respecto a t y

$$\frac{\partial}{\partial t}h(X+\sqrt{t}Z) = -\int\left(\frac{\partial}{\partial t}g_t(y)\right)\log g_t(y)\,dy. \tag{3.56}$$

Demostración. Debido a que $h(X+\sqrt{t}Z) = \int g_t(y)\log g_t(y)dy$, queremos demostrar el paso de derivación bajo el signo integral, para ello es necesario la acotación de

$$\frac{\partial}{\partial t}g_t(y)\log g_t(y) = \frac{\partial g_t(y)}{\partial t}\log g_t(y) + \frac{\partial g_t(y)}{\partial t}. \tag{3.57}$$

En primer lugar, tomaremos para hacer esta acotación un entorno $0 < a < t < b$ de t y en todo $\mathbb{R}$ respecto a y.

Como se vio previamente, $\left|\frac{\partial}{\partial t}\phi_t(y-X)\right| \leq 2\sqrt{2}(et)^{-2}\phi_{2t}(x)$. Es posible dominar las función Gaussiana con una de mayor varianza y ajustando las constantes previas para que se cumple en todo $\mathbb{R}$:

$$\left|\frac{\partial}{\partial t}\phi_t(y-X)\right| \leq c\,\phi_{2b}(y-X) \quad \text{con} \quad c = 2(ea)^{-1}(2b/a)^{1/2}. \tag{3.58}$$

Aplicando esto a la función g_t

$$\left|\frac{\partial}{\partial t}g_t(y)\right| = \left|E_X(\frac{\partial}{\partial t}\phi_t(y-X)\right| \leq E\left|\frac{\partial}{\partial t}\phi_y(y-X)\right| = E_X(c\,\phi_{2b}(y-X))) \leq c\,g_{2b}(y).. \tag{6.12}$$

Por lo tanto, $\frac{\partial}{\partial t} g_t(y)$ está dominada, uniformemente en un entorno de t, por una función integrable al ser g_{2b} una función de densidad.

El siguiente paso es acotar $|\log g_t(y)|$. Debido a que toda distribución Gaussiana ϕ_t esta acotada por $\frac{1}{\sqrt{2\pi t}}$

$$g_t(y) = E_X(\phi_t(y - X)) \leq E_X(\frac{1}{\sqrt{2\pi t}}) = \frac{1}{\sqrt{2\pi t}} \leq \frac{1}{\sqrt{2\pi a}}. \tag{3.59}$$

Debido a que se trata del valor absoluto del logaritmo, es necesario acotar inferiormente g_t para cuando $g_t(y) < 1$. Esto se hace detalladamente en la sección 5 de [17].

$$g_t(y) \geq c_t \phi_{t/2}(y) \quad \text{con} \quad c_t = (3/4\sqrt{2})e^{-4/t}. \tag{3.60}$$

Entonces

$$\begin{aligned} |\log g_t(y)| &= \log g_t(y) I_{(0,1)}(g_t(y)) - \log g_t(y) I_{(1,\infty)}(g_t(y)) \\ &\leq \log \frac{1}{\sqrt{2\pi a}} - \log \left(\frac{c_a}{\sqrt{\pi b}} e^{-y^2/a} \right) = \frac{1}{a}(y^2 + c'). \end{aligned} \tag{3.61}$$

donde $c' = 4 + a \log(4/3)(b/a)^{1/2}$.

Usando Las desigualdades previas, la derivada se puede acotar uniformemente en un entorno de t por

$$\left| \frac{\partial}{\partial t}(g_t(y) \log g_t(y)) \right| \leq \frac{c}{a}(y^2 + c' + a) g_{2b}(y). \tag{3.62}$$

Esta función, al trabajar con funciones de densidad exige tener momento de orden dos finito, es decir, varianza correctamente definida, lo cual ocurre para las funciones de densidad con las que se suele trabajar. Por lo tanto, esta cota nos permite aplicar el teorema de derivación bajo el signo integral

$$\begin{aligned} \frac{\partial}{\partial t} h(X + \sqrt{t}Z) &= -\frac{\partial}{\partial t} \int g_t(y) \log g_t(y)\, dy = \int \frac{\partial g_t(y)}{\partial t} \log g_t(y)\, dy + \int \frac{\partial g_t(y)}{\partial t}\, dy = \\ &= \frac{\partial}{\partial t} \int g_t(y) dy = \int \frac{\partial g_t(y)}{\partial t} \log g_t(y)\, dy + 0 = \int \frac{\partial g_t(y)}{\partial t} \log g_t(y)\, dy. \end{aligned} \tag{3.63}$$

□

Una vez se ha probado la derivabilidad de $h(X + \sqrt{t}Z)$ en $t > 0$, se deduce inmediatamente la derivabilidad de $I(X + \sqrt{t}Z; Z) = h(X + \sqrt{t}Z) - h(x)$ en $t > 0$. Sin embargo, lo que interese y en lo que se centra el siguiente lema, es probar la derivabilidad de estas funciones en $t = 0$

Lema 3.2.14 (Derivabilidad de $I(X + \sqrt{t}Z; Z)$)**.** La información mutua, $I(X + \sqrt{t}Z; Z)$, es diferenciable en $t = 0$

Demostración. Definimos una variable auxiliar $X_u^* := X + \sqrt{u}Z'$ donde Z' es también una distribución normal estándar independiente de Z.

$$\begin{aligned} I(X_u^* + \sqrt{t}Z; Z) &= I(X + \sqrt{u}Z' + \sqrt{t}Z; Z) = \\ &= h(X + \sqrt{u}Z' + \sqrt{t}Z) - h(X + \sqrt{u}Z') \\ &= h(X + \sqrt{u}Z' + \sqrt{t}Z) - h(X) + h(X) - h(X + \sqrt{u}Z') = \\ &= I(X + \sqrt{u}Z' + \sqrt{t}Z; \sqrt{u}Z' + \sqrt{t}Z) - I(X + \sqrt{u}Z'; \sqrt{u}Z') = \\ &= I(X + \sqrt{u+t}Z; \sqrt{u+t}Z) - I(X + \sqrt{u}Z'; \sqrt{u}Z'). \end{aligned} \tag{3.64}$$

Donde se ha usado una propiedad de las distribuciones Gaussianas en la suma de dos variables aleatorias normales independientes con desviación típica $\sqrt{u}$ y $\sqrt{t}$ es igual a una variable aleatoria normal con desviación típica $\sqrt{t+u}$. Retomando la igualdad anterior y haciendo uso del Lema 5.2.4

$$\begin{aligned} I(X_u^* + \sqrt{t}Z; Z) &= I(X + \sqrt{u+t}Z; \sqrt{u+t}Z) - I(X + \sqrt{u}Z'; \sqrt{u}Z') = \\ &= I(X + \sqrt{u+t}Z; Z) - I(X + \sqrt{u}Z'; Z'). \end{aligned} \tag{3.65}$$

De esta forma, como la igualdad se cumple para todo $u > 0$, $I(X_u^* + \sqrt{t}Z; Z)$ se escribe como suma de dos funciones diferenciables en $t = 0$

$$\frac{\partial}{\partial t} I(X_u^* + \sqrt{t}Z; Z) = \frac{\partial}{\partial t} I(X + \sqrt{u+t}Z; Z) = \frac{\partial}{\partial u+t} I(X + \sqrt{u+t}Z; Z)\Big|_{u+t=u} \frac{\partial u+t}{\partial t}\Big|_{t=0}. \tag{3.66}$$

Nótese que se ha probado para X^*, pero es intercambiable con cualquier variable X. □

3.2.2. Demostración de la EPI (Proposición 3.2.2)

Con ayuda de estos lemas, se demostrará el siguiente teorema, conocido también como la **Convexidad de la Información Mutua** o **MII** por sus siglas en inglés **Mutual Information Inequality**. Este es el paso final antes de demostrar **EPI**.

Teorema 3.2.15 (MII)**.** Para un número finito de vectores aleatorios independientes n-dimensionales $(X_i)_i$ con covarianzas finitas, cualquier conjunto de coeficientes reales $(a_i)_i$ normalizados tales que $\sum_i a_i^2 = 1$, y cualquier vector Gaussiano n-dimensional Z independiente de $(X_i)_i$,

$$I\left(\sum_i a_i X_i + Z; Z\right) \leq \sum_i a_i^2 I(X_i + Z; Z). \tag{3.67}$$

Demostración.

$$I(\sum_i a_i X_i + Z; Z) = I(\sum_i a_i X_i + \sum_i a_i^2 Z; Z) = I(\sum_i a_i (X_i + a_i Z); Z). \tag{3.68}$$

Llamando $\hat{X}_i := X_i + a_i Z$, $Y = \sum_i a_i \hat{X}_i = \sum_i a_i X_i + Z$ y $W = (\hat{X}_1, ..., \hat{X}_n)$, se tiene la siguiente cadena de Markov $Z \leftrightarrow W \leftrightarrow Y$ ya que $P(Z|W,Y) = P(Z|W)$ y $P(Y|W,Z) = P(Z|W,Y)$. Esto nos permite aplicar **Desigualdad de Procesado de Datos**:

$$I(\sum_i a_i (X_i + a_i Z); Z) = I(Y; Z) \leq I(W; Z) = I(\hat{X}_1, ..., \hat{X}_n; Z). \tag{3.69}$$

Aplicando ahora el Lema 5.2.5

$$I(\hat{X}_1, ..., \hat{X}_n; Z) \leq \sum_i I(X_i + a_i Z; Z). \tag{3.70}$$

Encadenando estas desigualdades aplicando a $\sqrt{t}Z$ y usando el Lema 5.2.6

$$I(\sum_i a_i X_i + \sqrt{t}Z; Z) \leq \sum_i I(X_i + a_i\sqrt{t}Z; Z) = \sum_i a_i^2 I(X_i + \sqrt{t}Z; Z) + o(t). \tag{3.71}$$

Definimos ahora $X_i^* = X_i + \sqrt{u}Z_i^*$ siendo Z_i^* normales estándar e independiente con $Z = \sum_i a_i Z_i^* \sim \mathcal{N}(0,1)$.

$$\begin{aligned} I(\sum_i a_i X_i + \sqrt{u}Z^* + \sqrt{t}Z; Z) &= I(\sum_i a_i X_i + a_i\sqrt{u}Z_i^* + \sqrt{t}Z; Z) \\ &= I(\sum_i a_i(X_i + \sqrt{u}Z_i^*) + \sqrt{t}Z; Z) \\ &= I(\sum_i a_i X_i^* + \sqrt{t}Z; Z) \leq \sum_i I(X_i^* + a_i\sqrt{t}Z; Z) \\ &= \sum_i I(X_i + \sqrt{u}Z_i + \sqrt{t}Z; Z) + o(t). \end{aligned} \tag{3.72}$$

Por otro lado, haciendo uso de la ecuación (5.54), $(I(X+\sqrt{u}Z^*+\sqrt{t}Z; Z) = I(X+\sqrt{u+t}Z; Z) - I(X + \sqrt{u}Z'; Z'))$ del Lema 5.3.4. a ambos lados de la desigualdad:

$$I(\sum_i a_i X_i + \sqrt{u}Z^* + \sqrt{t}Z; Z) \leq \sum_i I(X_i + \sqrt{u}Z_i + \sqrt{t}Z; Z) + o(t)$$

$$I(\sum_i a_i X_i + \sqrt{u+t}Z; Z) - I(\sum_i a_i X_i + \sqrt{u}Z; Z) \leq \sum_i a_i^2 \left(I(X_i + \sqrt{u+t}Z; Z) - I(X_i + \sqrt{u}Z; Z)\right) + o(t). \tag{3.73}$$

Definiendo:

$$f(t) := I(\sum_i a_i X_i + \sqrt{t}Z; Z) - \sum_i a_i^2 I(X_i + \sqrt{t}Z; Z). \tag{3.74}$$

La desigualdad (5.62), se traduce en $f(u+t) \leq f(u) + o(t)$. Además, $f(u)$ es diferenciable en $u \geq 0$ al serlo $I(X + \sqrt{u}Z; Z)$ Está función cumple dos propiedades:

(i) $f(0) = 0$ ya que X_i y Z son independientes

(ii) Es una función decreciente

$$\left.\frac{\partial f(t)}{\partial t}\right|_{t=u} = \lim_{t\to 0^+} \frac{f(u+t) - f(u)}{t} \leq \lim_{t\to 0^+} \frac{o(t)}{t} = 0. \tag{3.75}$$

Por estos dos aspectos

$$0 = f(0) \geq f(1) = I\left(\sum_i a_i X_i + Z; Z\right) - \sum_i a_i^2 I(X_i + Z; Z). \tag{3.76}$$

□

Proposición 3.2.16. (EPI) El teorema 5.3.1 (**MII**), es equivalente a **(EPI)** al ser equivalente a la desigualdad (c) de la proposición 5.1.3.

Demostración. $(MII \implies (c))$Igualando los siguientes términos,

$$I(X+Z;Z) = h(Z+X) - h(X)\,, \tag{3.77}$$
$$I(X+Z;X) = h(Z+X) - h(Z). \tag{3.78}$$

se obtiene, $I(X+Z;Z) = I(X+Z;X) + h(Z) - h(X)$ y aplicándolo a $X = \sum_i a_i X_i$ con $\sum_i a_i^2 = 1$ y X_i individualmente se obtiene las siguientes igualdades.

$$(i) \quad I(\sum_i a_i X_i + Z; Z) = I(\sum_i a_i X_i + Z; X) + h(Z) - h(\sum_i a_i X_i). \tag{3.79}$$
$$(ii) \quad I(X_i + Z; Z) = I(X_i + Z; X) + h(Z) - h(X_i). \tag{3.80}$$

Si ahora se aplican estas desigualdades a ambos lados de la **MII**

$$I(\sum_i a_i X_i + Z; Z) \leq \sum_i a_i^2 I(X_i + Z; Z)$$
$$I(\sum_i a_i X_i + Z; \sum_i a_i X_i) + h(Z) - h(\sum_i a_i X_i) \leq \sum_i a_i^2 \left(I(X_i + Z; X_i) + h(Z) - h(X_i)\right)$$
$$I(\sum_i a_i X_i + Z; \sum_i a_i X_i) - \sum_i a_i I(X_i + Z; X_i) \leq h(\sum_i a_i X_i) - \sum_i a_i^2 h(X_i). \tag{3.81}$$

Si finalmente, sustituimos Z, por $\sqrt{t}Z$:

$$I(X_i; X_i + \sqrt{t}Z) = I(\frac{1}{\sqrt{t}}X_i; \frac{1}{\sqrt{t}}X_i + Z) = I(X_i; \frac{1}{\sqrt{t}}X_i + Z). \tag{3.82}$$

$$\lim_{t\to\infty} I(X_i; \frac{1}{\sqrt{t}}X_i + Z) = I(X_i; Z) = 0 \quad \text{al ser independientes.} \tag{3.83}$$

Por lo tanto, tomando limite $t \to \infty$ en (5.70), se llega a:

$$0 \leq h(\sum_i a_i X_i) - \sum_i a_i^2 h(X_i). \tag{3.84}$$

O equivalentemente

$$h(\sum_i a_i X_i) \geq \sum_i a_i^2 h(X_i). \tag{3.85}$$

Quedando demostrada la **EPI**. □

3.2.3. Aplicaciones de la EPI

Como se indicó al inicio del capítulo, la **EPI**, el objetivo principal por el que se desarrolla esta desigualdad junto a sus demostraciones es únicamente aplicar esta en la solución analítica de diferentes modelos englobados en el aprendizaje automático. Sin embargo, tanto en el desarrollo realizado

previamente como en las conclusiones en sí, se ha llegado a una serie de resultados importantes que se alejan del objetivo principal.

Aplicaciones en la Física: la ecuación del calor.

Se ha desarrollado en las demostraciones previas una expresión que se escribe como:

$$\frac{\partial \phi_t}{\partial t} = \frac{1}{2}\frac{\partial^2 \phi_t}{\partial x^2}, \tag{3.86}$$

donde $\phi_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$. Esta ecuación en derivadas parciales es conocida como la **ecuación del calor**, que suele escribirse comúnmente como $\frac{\partial u}{\partial t} - \alpha \nabla^2 u = 0$.

Esta simple idea, permite escribir cualquier solución de la ecuación del calor a partir de esta solución Gaussiana junto con la operación de convolución.

Aplicaciones en Álgebra.

Aplicando la **EPI**, $N(X+Y) \geq N(X) + N(Y)$, a dos distribuciones Gaussianas X e Y independientes de matrices de covarianza Σ_X y Σ_Y respectivamente y dado que para una distribución Gaussiana se tiene que $N(X) = |\Sigma_X|$, se tiene el siguiente lema.

Lema 3.2.17. Desigualdad de Minkowski para determinantes Sean A y B dos matrices cuadradas de orden n definidas positivas. Entonces se tiene que:

$$|A+B|^{\frac{1}{n}} \geq |B|^{\frac{1}{n}} + |A|^{\frac{1}{n}}, \tag{3.87}$$

donde la igualdad de cumple únicamente para matrices proporcionales.

Y si aplicamos el lema primero de forma recurrente (o aplicamos el enunciado de la Proposición 3.2.3 de la **EPI**) se obtiene el siguiente lema.

Lema 3.2.18. Sean $\{A_i\}_{i=1}^n$ matrices cuadradas de orden n definidas positivas. Entonces se tiene que:

$$\left|\sum_i A_i\right|^{\frac{1}{n}} \geq \sum_i |A_i|^{\frac{1}{n}}, \tag{3.88}$$

donde la igualdad de cumple únicamente para matrices proporcionales.

Aplicaciones en Análisis.

De manera similar a la que se hizo en los lemas previos y sin realizar una demostración puesto que este apartado final supone simplemente una serie de aplicaciones de la **EPI**, se enuncia el siguiente teorema.

Teorema 3.2.19 (Teorema de Brunn-Minkowski). Sea $n \geq 1$ y sea μ la medida de Lebesgue en $\mathbb{R}^n$. Sean A y B dos subconjuntos compactos no vacíos de $\mathbb{R}^n$. Entonces se cumple la siguiente desigualdad:

$$[\mu(A+B)]^{1/n} \geq [\mu(A)]^{1/n} + [\mu(B)]^{1/n}, \tag{3.89}$$

donde $A + B$ denota la suma de Minkowski:

$$A + B := \{a + b \in \mathbb{R}^n \mid a \in A, b \in B\}.$$

Capítulo 4

Cuello de Botella de la Información

El concepto de **Cuello de Botella de la Información**, al que denotaremos por **IB** por sus siglas en inglés **IB** [13, 26], fue desarrollado por Naftali Tishby en 1999 y constituye un marco teórico crucial en el ámbito del aprendizaje supervisado y en la teoría de la información. Su objetivo principal es buscar un análisis de la información de entrada que nos permita predecir la salida de la forma mas precisa y eficiente posible, de forma que se minimice la redundancia y se mantengan únicamente los datos que aporten una mayor relevancia del modelo.

Un aspecto fundamental de este enfoque es su capacidad para abordar tanto el desafío del sobreajuste u *overfitting*, ya que ese exceso de información se adapta excesivamente a los datos de entrenamiento, perdiendo así su capacidad de generalización a nuevos datos como se vio previamente, como el *underfitting*, al evitar que se pierda demasiada información que las futuras predicciones no sean fiables.

Resumiendo, busca encontrar este equilibrio óptimo entre la compresión de la información de entrada y la preservación de la información relevante para la predicción, lo que mejora significativamente su capacidad para generalizar a nuevos datos y mejorar su desempeño predictivo en la práctica.

4.1. Descripción del IB

Situándonos en el marco del aprendizaje supervisado en el que se tiene un espacio $\mathcal{X}$ de atributos que corresponde a los datos extraíbles y un espacio $\mathcal{Y}$ de etiquetas que son los datos que se desean predecir, todo ello junto a una distribución conjunta $P_{X,Y}$ desconocida, el objetivo es entonces extraer la información relevante que una variable aleatoria $\mathcal{X}$ contiene sobre esa variable $\mathcal{Y}$. Para ello, se busca encontrar la distribución de probabilidad condicionada de un espacio auxiliar $\mathcal{T}$, que dependa exclusivamente de X, es decir, $P(T|X) = P(T|X,T)$ y que de esta forma sea capaz de contener de la mejor manera posible la información que X contiene sobre Y, quedadnos con la mínima cantidad de información que T tiene de X evitando así la redundancia.

En términos de información mutua, el objetivo es encontrar $\mathcal{T}$ tal que $I(X;T)$ sea mínimo, pero sin perder de vista la información que T nos da de Y, esto es, manteniendo un cierto umbral en

la $I(Y;T) \geq \alpha$. de esta forma, el resultado obtenido para $P(X,T)$ dependerá de la cantidad de información que estaremos dispuestos a sacrificar que, según la notación utilizada, viene dada por $I(Y;X) - I(Y;T) \leq I(Y;X) - \alpha = I_0$, donde I_0 representa esa perdida de información que se ha de asumir.

La idea es encontrar un $P_{T|X}$ que extraiga información sobre Y, es decir, de $I(Y;T)$, mientras comprime al máximo X, lo cual se cuantifica manteniendo $I(X;T)$ pequeño. Dado que a la representación comprimida T no puede transmitir más información que la señal original, el problema de **IB** se formula a través de la optimización restringida:

$$\inf_{P_{T|X}:I(Y;T)\geq\alpha} I(X;T). \tag{4.1}$$

donde la distribución conjunta subyacente es $P_{X,Y,T} = P_{X,Y}P_{T|X,Y} = P_{X,Y}P_{T|X}$.

Alternativamente, el problema de minimización anterior puede venir descrito por la siguiente función objetivo:

$$L_\beta(P_{T|X}) := I(X;T) - \beta I(T;Y). \tag{4.2}$$

Que según esta definición, el problema de **IB** se puede reformular como la minimización de $L_\beta(P_{T|X})$ sobre todos las distribuciones de probabilidad $P_{T|X}$. Aquí, β controla cuanta información estamos dispuestos a perder sobre la variable Y con respecto a la simplificación que se hace de X. Entonces, un β pequeño implica un mayor sacrificando de información, mientras que un β más grande empuja hacia una representación más fina que favorece la información que T preserva sobre X, siendo este caso una simplificación más pobre que el anterior. En un principio, podemos pensar que β varia en $\beta \in [0, +\infty)$ sin embargo, teniendo en cuenta la **Desigualdad de Procesado de Datos** [1], tenemos que $I(X;T) - \beta I(T;Y) \geq (1-\beta)I(T;Y)$, y por lo tanto, para todos los valores de $\beta \leq 1$, la solución óptima del problema de minimización es degenerada, es decir, $I(T;X) = I(T;Y) = 0$. Esto se debe a que $I(T;Y) \geq 0$ para cualquier T, por lo tanto exigiremos que β pertenezca al intervalo $(1, +\infty)$.

Sin embargo, esta no es la única forma de expresar la idea y el comportamiento del **IB**. Haciendo uso de la regla de la cadena para la información mutua:

$$I(X;T \mid Y) = I(X,Y;T) - I(Y;T). \tag{4.3}$$

$$I(X,Y;T) = I(X;T) - I(Y;T \mid X) = I(X;T) - \left(h(T \mid X) - h(T|X,Y) = I(X;T)\right), \tag{4.4}$$

donde se ha usado que $h(T|X,Y)h) = h(T|X)$ al ser una cadena de Markov.

Juntando estas dos ideas, se puede escribir $I(X;T|Y) = I(X;T) - I(W;T)$ y por tanto:

$$L_\beta(P_{T|X}) := I(X;T|Y) - (\beta - 1)I(T;Y). \tag{4.5}$$

Esta nueva representación es importante ya que el término $I(X;T|Y) = I(X;T) - I(Y;T)$ recibe el nombre de información residual, es básicamente, la información que T sigue manteniendo de X si ya se conoce el resultado Y, es decir, información que sobra. Y de forma paralela, $I(T;Y)$ representa la información relevante que se sigue manteniendo del problema inicial. Se busca así un equilibrio

[1] Cabe remarcar que en esta ocasión se está usando la desigualdad cambiando los papales de Y y de T al contrario que se vió en la demostración.

que dependerá del factor β, en el que se intenta maximizar la información relevante mientras se elimina la mayor información redundante posible.

Otro aspecto que hay que tratar consiste en el procedimiento que hay que seguir para encontrar esa distribución $P_{T|X}$ solución de nuestro problema. Esto abarca variedad de posibilidades. Por ejemplo se puede intentar resolver analíticamente como se hará a continuación para el caso Gaussiano o también es posible llevar a cabo un método iterativo que converja a la distribución buscada como se desarrolla en [10, 26].

Teniendo esto en cuenta, la forma de atacar un problema dependerá tanto de él mismo como de las distribuciones que estén involucradas. Además, dentro de estas diferentes formas habrá unas más generales de forma que se puedan aplicar a gran variedad de problemas, y otras que se puedan aplicar únicamente a un tipo de problema. Este es el caso que nos concierne, **IB** aplicado a distribuciones Gaussianas que tendrá una posterior aplicación en regresión lineal en altes dimensiones.

4.2. IB aplicado a distribuciones Gaussianas

A lo largo de esta sección, se usarán propiedades bien conocidas de las distribución normal multivariante como por ejemplo que las distribuciones condicionadas en un vector Gaussiano siguen siendo Gaussianas o diferentes formas de expresar las matrices de covarianza Gaussianos en estos casos. Las demostraciones de estos pueden verse detalladamente en [2].

La descripción del escenario de trabajo es la siguiente. Sea $X \sim \mathcal{N}(\mu_x, \Sigma_X)$ e $Y \sim \mathcal{N}(\mu_y, \Sigma_Y)$ vectores columnas multivariantes Gaussianos de dimensiones d_x y d_y, respectivamente. De forma que conjuntamente forman una distribución Gaussiana $(X, Y)^T \sim \mathcal{N}(\mu, \Sigma)$ donde $\mu = (\mu_x, \mu_y)^T$ y $\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{bmatrix}$ donde $\Sigma_{XY} := E[(X - \mu_X)(Y - \mu_Y)] \in \mathbb{R}^{d_x \times d_y}$. El objetivo de este capítulo consiste en encontrar la solución analítica del **IB** para este escenario [10, 13, 14].

El primer paso para caracterizar analíticamente el valor óptimo de $L_\beta(P_T|X)$ es demostrar que el valor se alcanza si se trabaja también con una distribución T normal multivariante de forma que (X, Y, T) sean conjuntamente Gaussianos. Una demostración general de este hecho puede consultarse en [7, 14] Esta demostración es muy larga y técnica y hace uso de la desigualdad de potencia de entropía (*EPI)*. Por tanto, en lugar de reproducir el argumento aquí, hemos optado por desarrollar nuestra propia demostración en el caso de que X, Y y T sean unidimensionales

Teorema 4.2.1. Dadas (X, Y) una distribución Gaussianas bidimensional $\mathcal{N}(\mu, \sigma)$ tal que $\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ y $\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$. El valor óptimo de $L_\beta(P_{T|X}) = I(X;T) - \beta I(T;Y)$ se alcanza para T Gaussiano unidimensional, es decir, la solución del **IB** se obtiene para el caso en el que $T \sim \mathcal{N}$.

Demostración. Si inicialmente trabajamos con distribución conjunta $(X, Y) \sim \mathcal{N}((\mu_x, \mu_y)^T, \Sigma)$, podemos escribir la distribución condicionada de $(Y|X)$ de forma general como sigue:

$$f_{Y|X=x}(y) \sim \mathcal{N}\left(\bar{\mu} = \mu_y + \Sigma_{XY}\Sigma_{xx}^{-1}(x - \mu_x), \bar{\Sigma} = \Sigma_{yy} - \Sigma_{XY}\Sigma_{xx}^{-1}\Sigma_{XY}\right). \tag{4.6}$$

Que, aplicándolo a nuestro caso particular unidimensional:

$$f_{Y|X=x}(y) \sim \mathcal{N}\left(\mu_y + \frac{\sigma_{xy}}{\sigma_{xx}^2}(x-\mu_x), \sigma_{yy}^2 - \frac{\sigma_{xy}^2}{\sigma_{xx}^2}\right). \tag{4.7}$$

Esto nos permite escribir la variable aleatoria $Y = \frac{\sigma_{xy}}{\sigma_{xx}^2}X + Z$ donde $Z \sim \mathcal{N}(\mu_y - \frac{\sigma_{xy}}{\sigma_{xx}^2}\mu_x, \sigma_{yy}^2 - \frac{\sigma_{xy}^2}{\sigma_{xx}^2})$ independiente de X. Por simplicidad de la notación definimos $a = \frac{\sigma_{xy}}{\sigma_{xx}^2}$, $b = \mu_y - a\mu_x$ y $c^2 = \sigma_{yy}^2 - \frac{\sigma_{xy}^2}{\sigma_{xx}^2})$ de forma que $Y = aX + Z$ y $Z \sim \mathcal{N}(b, c^2)$

El siguiente paso consiste en aplicar **EPI** a la variable $Y|T$ de la siguiente forma:

$$N(Y|T) = N(aX + N|T) \geq N(aX|T) + N(N|T) = N(aX|T) + N(Z). \tag{4.8}$$

donde en el último paso se ha usado que Z es independiente de T, entonces $h(Z|T) = h(Z)$.

Ahora aplicando la definición de power entropy:

$$\begin{aligned} e^{2h(Y|T)} &\geq e^{2h(aX|T)} + e^{2h(Z)} \\ &= e^{2h(X|T)+2\log|a|} + 2\pi e\, c^2 \\ &= e^{2(-I(X;T)+h(X))}\, a^2 + 2\pi e\, c^2 \\ &= e^{-2I(X;T)} 2\pi e\, \sigma_{xx}^2\, a^2 + 2\pi e\, c^2. \end{aligned} \tag{4.9}$$

Aplicando logaritmos a ambos lados, se obtiene

$$h(Y|T) \geq \log(2\pi e) + \log(e^{-2I(X;T)}\, \sigma_{xx}^2\, a^2 + c^2). \tag{4.10}$$

Para llegar a la desigualdad buscada, basta con aplicar esta desigualdad a la función a minimizar:

$$\begin{aligned} L_\beta(P_{T|X}) &= I(X;T) - \beta I(T;Y) = I(X;T) - \beta(h(Y) - h(Y|T)) \\ &\geq I(X;T) + \beta \log(e^{-2I(X;T)}\, \sigma_{xx}^2\, a^2 + c^2) - \beta h(Y) + \beta \log(2\pi e). \end{aligned} \tag{4.11}$$

donde el único término dependiente de T es $I(T;X)$. Esto nos da así una acotación inferior de la función a minimizar. Además, la **EPI** nos asegura que la igualdad se alcanza si y solo si las distribuciones implicadas $f_{X|T=t}$ y $f_{N|T=t} = f_N$ son distribuciones normales con matrices proporcionales, cosa que, al estar en dimensión uno, ocurre siempre.

Por todo lo anterior, es posible afirmar que para todas las posibles distribuciones de $f_{X,T}$ con información $I(X;T)$ fija, aquellas que alcancen la cota mínima de L_β son aquellas en las que la distribución condicionada $f_{X|T=t}$ es Gaussiana. Lo cual ocurre para el caso en el que $f_{X,T}$ y f_T lo sean.

El último paso para terminar la demostración consiste en probar que para todo valor fijo de $\alpha := I(X;T)$, existe una distribución Gaussiana conjunta (X,T) que alcanza ese valor α. Esto nos asegura llegar a la cota mínima para cualquier valor de $I(X;T)$.

La demostración de esto consiste simplemente en aplicar el Teorema del valores intermedio a la función $g(t) := I(X; tX + (1-t)Z)$ donde $Z \sim \mathcal{N}(0,1)$ independiente de X. Se tiene así que $g(t)$ es continua al venir dada en función del logaritmo de las matrices de covarianza de distribuciones

normales todas ellas con determinantes estrictamente positivos (Corolario 2.4.6). Además $g(0) = 0$ al ser variables independientes y $g(1)$ diverge al no existir función de densidad conjunta de $f_{X,X}$.

Concluimos así con la demostración al afirmar que siempre existe T tal que $I(X;T) = \alpha$ para todo α.

□

Es importante darse cuenta que este teorema previo no afirma que ese mínimo se alcanza únicamente para el caso de T Gaussiano, es decir, es posible que existan otras distribuciones que alancen también el mínimo pero como la normal es una que lo cumple, podemos reducir nuestro estudio a este tipo de distribuciones Gaussianas.

Y como se mencionó al inicio de la sección, a pesar de solo exponerse la demostración para el caso unidimensional, este es cierto para cualquiera que sea la dimensión de trabajo. Se enuncia así el correspondiente teorema.

Teorema 4.2.2 (**Generalización del Teorema 4.2.1**)**.** Dadas (X, Y) una distribución Gaussiana $\mathcal{N}(\mu, \sigma)$ tal que $\mu = \begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}$ y $\Sigma = \begin{bmatrix}\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_{YY}\end{bmatrix}$. El valor óptimo de $L_\beta(P_{T|X}) = I(X;T) - \beta I(T;Y)$ se alcanza para T Gaussiano, es decir, la solución del **IB** se obtiene para el caso en el que $T \sim \mathcal{N}(\mu_T, \Sigma_T)$.

Para encontrar la forma analítica particularizada al caso Gaussiano de la expresión a minimiza, $L_\beta(P_{T|X}) = I(X;T) - \beta I(T;Y)$, se va a usar la forma que toma la divergencia de Kullback-Leibler para distribuciones normales vista en el Capítulo 2. Aplicamos esta fórmula a nuestro caso, $I(X;T) = D_{KL}(P_{X,T}||P_X P_T)$:

$$P_{X,T} \sim N(\mu, \Sigma_1) \text{ donde } \mu = \begin{bmatrix}\mu_X \\ \mu_T\end{bmatrix} \text{ y } \Sigma_1 = \begin{bmatrix}\Sigma_X & \Sigma_{XT} \\ \Sigma_{XT}^T & \Sigma_T\end{bmatrix} = \bar{\Sigma}_{X,T}. \tag{4.12}$$

$$P_X P_T \sim N(\mu, \Sigma_2) \text{ donde } \mu = \begin{bmatrix}\mu_X \\ \mu_T\end{bmatrix} \text{ y } \Sigma_2 = \begin{bmatrix}\Sigma_X & 0 \\ 0 & \Sigma_T\end{bmatrix}. \tag{4.13}$$

Teniendo esto en cuenta y usando que:

$$\Sigma_2^{-1} = \begin{bmatrix}\Sigma_X^{-1} & 0 \\ 0 & \Sigma_Y^{-1}\end{bmatrix} \quad \Longrightarrow \quad Tr(\Sigma_2^{-1}\Sigma_1) = Tr(\mathbb{I}). \tag{4.14}$$

$$\begin{aligned} I(X;T) &= \frac{1}{2}\left(\log\left(\frac{\det(\Sigma_X)\det(\Sigma_T)}{\det(\bar{\Sigma}_{X,T})}\right) + \mathrm{Tr}(\mathbb{I}) + (\mu-\mu)^T \Sigma_2^{-1}(\mu-\mu) - n\right) = \\ &= \frac{1}{2}\log\left(\frac{\det(\Sigma_X)\det(\Sigma_T)}{\det(\bar{\Sigma}_{X,T})}\right). \end{aligned} \tag{4.15}$$

Análogamente para (Y, T):

$$I(Y;T) = \frac{1}{2}\log\left(\frac{\det(\Sigma_T)\det(\Sigma_Y)}{\det(\bar{\Sigma}_{TY})}\right). \tag{4.16}$$

Y entonces tendríamos la versión analítica de nuestra función a minimizar.[2]

$$L_\beta(P_{T|X}) = I(X;T) - \beta I(T;Y) = \log\left(\frac{\det(\Sigma_X)\det(\Sigma_T)}{\det(\bar{\Sigma}_{XT})}\right) - \beta\log\left(\frac{\det(\Sigma_T)\det(\Sigma_Y)}{\det(\bar{\Sigma}_{TY})}\right). \quad (4.17)$$

En primer lugar observamos que esta cantidad a minimizar es independiente de las medias de las distintas distribuciones, por ello sin perder generalidad, podemos suponer que todas ellas están centradas en el $\mathbf{0}$.

Como último paso antes de llegar a nuestra solución y relacionarlo con nuestro problema, falta aplicar el hecho de que $Y \leftrightarrow X \leftrightarrow T$ sea una cadena de Markov. Esto nos permite escribir la variable que buscamos como $T = AX + Z$, donde $Z \sim \mathcal{N}(0, \Sigma_Z)$ es independiente de (X, Y).

Lema 4.2.3. Sea (X, T) una distribución Gaussiana conjunta, entonces es posible encontrar una matriz A tal que $T \overset{d}{=} AX + Z$, donde Z es una distribución Gaussiana independiente de X.

Demostración. Como $P_{T|X,Y} = P_{T|X}$ entonces podemos contemplar únicamente estas variables sin involucrar a Y. Así, suponiendo que $(X,T)^T \sim \mathcal{N}((\mu_x, \mu_t)^T, \begin{bmatrix} \Sigma_X & \Sigma_{XT} \\ \Sigma_{TX} & \Sigma_T \end{bmatrix})$ llamando $A = \Sigma_{XT}\Sigma_X^{-1}$ y $Z \sim \mathcal{N}(\mu_x - \Sigma_{XT}\Sigma_X^{-1}\mu_t, \Sigma_T - \Sigma_{XT}\Sigma_X^{-1}\Sigma_{TX})$, se obtiene el resultado deseado □

Teniendo en cuenta está definición de T, el problema de optimización se reduce a encontrar esas matrices A y Σ_Z, es decir:

$$\inf_{A,\Sigma_Z} I(X;T) - \beta I(T;Y). \quad (4.18)$$

Reescribimos ahora al información teniendo en cuenta una serie de propiedades características de las matrices de covarianzas de la distribución normal multivariante y del caso concreto de cadena de Markov en el que se está trabajando [2]:

(i) $\Sigma^T = \Sigma$.

(ii) $\det(\bar{\Sigma}_{XT}) = \det(\Sigma_{T|X})\det(\Sigma_X)$.

(iii) $\det(\bar{\Sigma}_{YT}) = \det(\Sigma_{T|Y})\det(\Sigma_Y)$.

(iv) $\Sigma_{TX} = A\Sigma_X$.

(v) $\Sigma_{TY} = A\Sigma_{XY}$.

(vi) $\Sigma_T = A\Sigma_X A^T + \Sigma_Z$

(vii) $\Sigma_{T|X} = \Sigma_T - \Sigma_{TX}\Sigma_X^{-1}\Sigma_{TX}^T = A\Sigma_X A^T + \Sigma_Z - A\Sigma_X\Sigma_X^{-1}\Sigma_X^T A^T = A\Sigma_X A^T + \Sigma_Z - A\Sigma_X A^T = \Sigma_Z$.

(viii) $\Sigma_{T|Y} = \Sigma_T - \Sigma_{TY}\Sigma_Y^{-1}\Sigma_{TY}^T = A\Sigma_X A^T + \Sigma_Z - A\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T A^T = A\left(\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}\right)A^T + \Sigma_Z = A\Sigma_{X|Y}A^T + \Sigma_Z$.

[2] Omitimos el factor $\frac{1}{2}$ ya que no afecta a la optimización.

Aplicando esto, a la información que tenemos:

$$L_\beta(P_{T|X}) = (1-\beta)\log(|A\Sigma_X A^T + \Sigma_Z|) - \log(|\Sigma_Z|) + \beta\log(|A\Sigma_{X|Y}A^T + \Sigma_Z|). \tag{4.19}$$

Llegados este punto tendríamos una función a optimizar dependiente de dos matrices (A, Σ_Z). Para solventar esta mayor dificultad de cálculo, el Lema A.1 en el Apéndice A muestra que para cada par (A, Σ_Z), existe otra proyección $\tilde{A}$ tal que el par $(\tilde{A}, I)$ obtiene el mismo valor de L, es decir, alcanza el mismo mínimo de manera que solo hay que buscar el mínimo en función $\tilde{A}$.

Demostración. Denotemos por V la matriz que diagonaliza Σ_Z, es decir, $\Sigma_Z = VDV^T$ donde $D = \text{diag}(d_1, ..., d_k)$. Al ser definida positiva admite raíz cuadrada de la forma $\Sigma_Z^{1/2} = VD^{1/2}V^T$, donde $D^{1/2} = \text{diag}(d_1^{1/2}, ..., d_k^{1/2})$.

Llamamos ahora $A = \Sigma_Z^{-1/2}A$ de forma que se tiene el siguiente resultado:

$$\begin{aligned}
L(\tilde{A}, I) &= (1-\beta)\log(|\tilde{A}\Sigma_X\tilde{A}^T + I_d|) - \log(|I_d|) + \beta\log(|\tilde{A}\Sigma_{X|Y}\tilde{A}^T + I_d|) \\
&= (1-\beta)\log(|\Sigma_Z^{1/2}\left(\tilde{A}\Sigma_X\tilde{A}^T + I_d\right)\Sigma_Z^{1/2}|) - (1-\beta)\log(|\Sigma_Z|) + \\
&\quad + \beta\log(|\Sigma_Z^{1/2}\left(\tilde{A}\Sigma_{X|Y}\tilde{A}^T + I_d\right)\Sigma_Z^{1/2}|) - \beta\log(|\Sigma_Z|) \\
&= (1-\beta)\log(|A\Sigma_X A^T + \Sigma_Z|) - \log(|\Sigma_Z|) + \beta\log(|A\Sigma_{X|Y}A^T + \Sigma_Z|) \\
&= L(A, \Sigma_Z).
\end{aligned} \tag{4.20}$$

donde la primera igualdad se deriva del hecho de que el determinante de un producto de matrices es el producto de los determinantes. □

Esto nos permite simplificar los cálculos reemplazando la matriz de covarianza Σ_Z con la matriz identidad Id.

$$L_\beta(P_{T|X}) = (1-\beta)\log(|A\Sigma_X A^T + I|) + \beta\log(|A\Sigma_{X|Y}A^T + I|). \tag{4.21}$$

Para identificar el mínimo de L, diferenciamos L con respecto a la matriz A. Este concepto de diferenciabilidad matricial viene detalladamente explicado en [21]. En esencia, se trabaja de igual forma que en cualquier búsqueda de máximos en varias dimensiones pero reagrupando los términos de tal forma que se obtengan identidades matriciales. Haremos uso de la identidad (53) de [21] $\frac{\delta\log(|ACA^T|)}{\delta A} = (ACA^T)^{-1}2AC$, que se cumple para cualquier matriz simétrica C:

$$\frac{\delta L_\beta}{\delta A} = (1-\beta)(A\Sigma_X A^T + I)^{-1}2A\Sigma_X + \beta(A\Sigma_{X|Y}A^T + I)^{-1}2A\Sigma_{X|Y}. \tag{4.22}$$

Igualando esta derivada a cero y reorganizando, obtenemos condiciones necesarias que debe cumplir la matriz A para que se alcance un mínimo interno de L. Para encontrar ese A, en primer lugar, vamos a tratar el caso más sencillo en el que t sea un escalar para después generalizar a cualquier dimensión.

4.2.1. Caso de una única proyección escalar

En este caso, $t \in \mathbb{R}$, $x \in \mathbb{R}^{d_x}$ e $y \in \mathbb{R}^{d_y}$, de forma que la trasformación que buscamos es $t = Ax+z$ con $A \in \mathcal{M}_{1\times d_x}(\mathbb{R})$ y $z \sim \mathcal{N}(0,1)$.

$$\frac{\delta L_\beta}{\delta A} = (1-\beta)(A\Sigma_X A^T + 1)^{-1} 2A\Sigma_X + \beta(A\Sigma_{X|Y}A^T + 1)^{-1} 2A\Sigma_{X|Y}. \tag{4.23}$$

Y reordenando términos se llega a:

$$\frac{\beta-1}{\beta}\frac{A\Sigma_{X|Y}A^T+1}{A\Sigma_X A^T+1}A = A\left(\Sigma_{X|Y}\Sigma_x^{-1}\right). \tag{4.24}$$

Observamos que es un problema de valores propios en el cual los autovalores dependen de A. Tiene dos tipos de soluciones dependiendo del valor de β. Primero, sin necesidad de ningún tipo de cálculo, A puede ser idénticamente nulo y en este caso $L_\beta = 0$ De lo contrario, A debe ser el vector propio de $\Sigma_{X|Y}\Sigma_x^{-1}$, con un valor propio igual a

$$\lambda = \frac{\beta-1}{\beta}\frac{A\Sigma_{X|Y}A^T+1}{A\Sigma_X A^T+1}. \tag{4.25}$$

Para caracterizar los valores de β para los cuales la solución óptima no degenera, definimos en primer lugar $r = \frac{A\Sigma_X A^T}{\|A\|^2}$.

Cuando A es un vector propio de $\Sigma_{X|Y}\Sigma_x^{-1}$, el siguiente lema, muestra que r es positivo y que

$$A\Sigma_{X|Y}\Sigma_x^{-1}\Sigma_X A^T = \lambda r\|A\|^2. \tag{4.26}$$

Lema 4.2.4. Denotemos el conjunto de autovalores por la izquierda de $\Sigma_{x|y}\Sigma_X^{-1}$ normalizados como v_i ($||v_i|| = 1$) y sus correspondientes autovalores como λ_i. Entonces se cumple,

1. Todos los autovalores son reales y satisfacen $0 \leq \lambda_i \leq 1$.
2. Existen $r_i > 0$ tal que $v_i^T\Sigma_X v_j = \delta_{ij}r_i$.
3. $v_i^T\Sigma_{x|y}v_j = \delta_{ij}\lambda_i r_i$.

Demostración:. 1. Las matrices $\Sigma_{X|Y}\Sigma_X^{-1}$ y $\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1}$ son semidefinidas positivas al tratarse de matrices covarianza, inversas o producto de ellas y por lo tanto, sus autovalores son positivos. Por otro lado, y razonando a partir de la forma que tiene la matriz de covarianza de una distribución condicionada se tiene que $\Sigma_{X|Y}\Sigma_X^{-1} = I - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1}$, y por tanto, los autovalores de $\Sigma_{X|Y}\Sigma_X^{-1}$ están acotados entre 0 y 1.

2. Denotemos por V la matriz cuyas filas son v_i^T. La matriz $V\Sigma_X^{1/2}$ [3] es la matriz de autovectores de $\Sigma_X^{-1/2}\Sigma_X|y\Sigma_X^{-1/2}$ ya que $V\Sigma_X^{1/2}\Sigma_X^{-1/2}\Sigma_X|y\Sigma_X^{-1/2} = V\Sigma_{X|Y}\Sigma_X^{-1/2} = V\Sigma_{X|Y}\Sigma_X^{-1}\Sigma_X^{1/2} = DV\Sigma_X^{1/2}$. siendo D la matriz diagonal de autovalores. Dado que $\Sigma_X^{-1/2}\Sigma_{X|Y}\Sigma_X^{-1/2}$ es simétrica, $V\Sigma_X^{1/2}$ es

[3] Al tratarse de matrices positivas, tiene sentido definir la raíz cuadrada de esta de forma única.

ortogonal, y al ser toda matriz de covarianza simétrica, $\Sigma_X^T = \Sigma_X$, se tiene que $V\Sigma_X^{1/2}\left(V\Sigma_X^{1/2}\right)^T = V\Sigma_X^{1/2}\Sigma_X^{1/2}V^T = V\Sigma_X V^T$ es diagonal. Esto nos permite deducir que existen tal que $v_i^T\Sigma_X v_j = \delta_{ij}r_i$. El hecho de que estos sean positivos se deduce de que se esta trabajando siempre con matrices semidefinidas positivas.

3. Del punto anterior se deduce que $v_i^T\Sigma_{X|Y}\Sigma_X^{-1}\Sigma_X v_j = \lambda_i v_i^T\Sigma_X v_j = \lambda_i\delta_{ij}r_i$.

Por lo tanto, teniendo en cuenta que A es un autovector de $\Sigma_{X|Y}\Sigma_X^{-1}$, es decir, $A\Sigma_{X|Y}\Sigma_X^{-1}\Sigma_X A^T = A\Sigma_{X|Y}A^T$ podemos reescribir el valor propio como 4.25 y aislando $\|A\|^2$, tenemos

$$\|A\|^2 = \frac{A\Sigma_{X|Y}A^T}{r\lambda} = \frac{\frac{\beta\lambda}{\beta+1}\left(A\Sigma_x A^T + 1\right)}{r\lambda} = \frac{\frac{\beta\lambda}{\beta+1}\left(r\|A\|^2 + 1\right)}{r\lambda}, \tag{4.27}$$

$$0 \leq \|A\|^2 = \frac{\beta(1-\lambda)-1}{r\lambda}. \tag{4.28}$$

□

De forma que al ser $r\lambda > 0$, nos da una condición que se debe cumplir entre el valor de β en relación a los autovalores para salir fuera del caso degenerado:

$$\beta(1-\lambda) - 1 \geq 0\,, \tag{4.29}$$

$$\beta^c(\lambda) = \frac{1}{1-\lambda} \leq \beta\,. \tag{4.30}$$

Para $\beta \geq \beta_c(\lambda)$, el peso de la preservación de información es lo suficientemente grande, y la solución óptima para A es un autovector no nulo de $\Sigma_X|y\Sigma_X^{-1}$.

De esta forma, fijado β, es posible que sean varios autovectores los que satisfagan la condición para soluciones no degeneradas. Sin embargo, para identificar el autovalor óptimo, substituimos el valor de $||A||^2$ de $A\Sigma_X|yA^T = r\lambda||A||^2$ y $A\Sigma_X A^T = r||A||^2$ en la función objetivo L y obtenemos:

$$\begin{aligned} L_\beta &= (1-\beta)\log(r+1) - \log(1)\beta\log(r\lambda+1)) \\ &= (1-\beta)\log\left(\frac{(1-\lambda)(\beta-1)}{\lambda}\right) + \beta\log(\beta(1-\lambda)). \end{aligned} \tag{4.31}$$

Dado que $\beta \geq 0$, esta función es monótonamente creciente en λ y se minimiza con el autovalor más pequeño de $\Sigma_{X|Y}\Sigma_X^{-1}$.

Concluimos que para proyecciones escalares, la solución analítica de **IB** es la siguiente:

$$A(\beta) = \begin{cases} \sqrt{\frac{\beta(1-\lambda)-1}{r\lambda}}v_\lambda & 0 < \lambda \leq \frac{\beta-1}{\beta}. \\ 0 & \text{en otro caso.} \end{cases} \tag{4.32}$$

donde v_λ es un autovector de $\Sigma_{X|Y}\Sigma_X^{-1}$ de norma uno asociado al autovalor λ más pequeño.

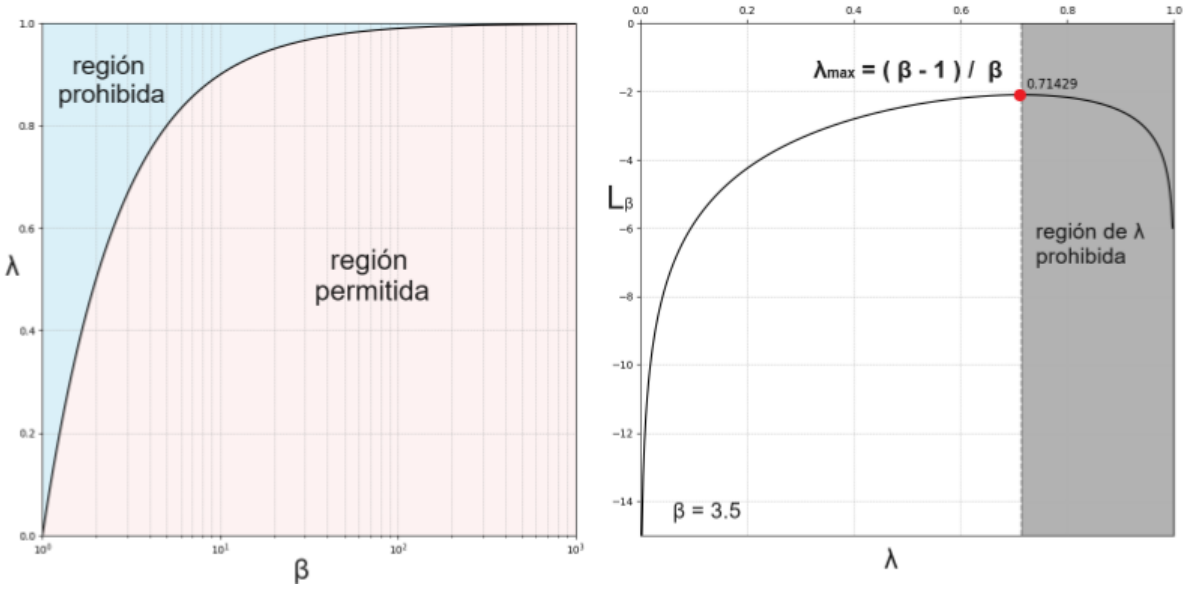


Figura 4.1: A la izquierda, representación de la región permitida de los valores de λ y β para una solución no degenerada que cumple $\lambda \leq \frac{\beta-1}{\beta}$. A la derecha, representación de la función de coste para $\beta = 0{,}35$ fijo en función deλ de forma que es posible visualizar ese comportamiento creciente hasta un máximo por encima del cuál está únicamente la solución degenerada.

Esta idea de de trabajar exclusivamente con los autovalores de $\Sigma_{X|Y}\Sigma_X^{-1}$ que cumplan que $\lambda \leq \frac{\beta-1}{\beta}$ se ve de forma intuitiva en la Figura 4.1. Además, también es posible ver el hecho de que, dentro de la región permitida, cuánto menor es el valor de λ, menor es la función L_β a minimizar. Es importante también darse cuenta de que esta función objetivo toma valores negativos y que por tanto, la solución degenerada $L_\beta = 0$ no es la solución buscada.

4.2.2. Caso general

En este caso, $t \in \mathbb{R}^{d_t}$, $x \in \mathbb{R}^{d_x}$ e $y \in \mathbb{R}^{d_y}$, de forma que la trasformación que buscamos es $t = Ax + z$ con $A \in \mathcal{M}_{d_t \times d_x}(\mathbb{R})$ y $z \sim \mathcal{N}(0, I_{d_t})$.

Se busca entonces la solución a la siguiente igualdad.

$$\frac{\beta-1}{\beta}\left(A\Sigma_{X|Y}A^T + Id\right)\left(A\Sigma_X A^T + Id\right)^{-1} A = A\Sigma_{X|Y}\Sigma_X^{-1}. \tag{4.33}$$

Como se tiene el producto $BA = B\begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_{d_t} \end{bmatrix} = \begin{bmatrix} \dots \\ \sum_k b_{ik}v_k \\ \dots \end{bmatrix}$ que consiste en una matriz cuyas filas son combinaciones lineales de las filas de A, es decir, es un espacio generado por los vectores que conforman A, a pesar de que los valores b_{ik} dependen del propio A.

Esto significa que $A\Sigma_{X|Y}\Sigma_X^{-1}$ debería estar formado por esa combinación lineal. De esta forma, podemos representar la proyección A como un producto $A = WV$, donde las filas de V son autovec-

tores por la izquierda de $\Sigma_X|y\Sigma_X^{-1}$ normalizados y W esa matriz de mezcla que nos permite llegar a la igualdad.

El siguiente lema, nos permite encontrar la forma de esa matriz W.

Lema 4.2.5. El óptimo de la función de costo se obtiene con una matriz de mezcla diagonal W de la forma

$$W = \text{diag}\left(\sqrt{\frac{\beta(1-\lambda_1)-1}{r_1\lambda_1}}, \sqrt{\frac{\beta(1-\lambda_2)-1}{r_2\lambda_2}}, ..., \sqrt{\frac{\beta(1-\lambda_k)-1}{r_k\lambda_k}}, 0, 0, .., 0\right). \tag{4.34}$$

donde $\{\lambda_1, \ldots, \lambda_k\}$ son $k \leq d_t$ autovalores de $\Sigma_{X|Y}\Sigma_X^{-1}$ con valores críticos $\lambda_1, \ldots, \lambda_k \leq \frac{\beta-1}{\beta}$ siguiendo 4.30 y $r_i \equiv v_i^T \Sigma_X v_i$.

Demostración:. Escribimos $V\Sigma_{X|Y}\Sigma_X^{-1} = DV$ donde D es una matriz diagonal cuyos elementos son los autovalores correspondientes, y denotamos por R la matriz diagonal cuyo elemento i-ésimo es r_i.

Analizaremos en primer lugar el caso en el que $k = n_x$, substituimos $A = WV$ en la ecuación 4.33 y usando que y $V\Sigma_{X|Y}V^T = VDRV^T = VRDV^T$ al ser D y R diagonales, y $V\Sigma_x V^T = VRV^T$, se llega a,

$$\beta^{-1}\beta(WDRW^T + Id)(WRW^T + Id)^{-1}WV = WDV. \tag{4.35}$$

Al ser autovectores normalizados se tiene que $VV^T = I_{d_t}$ entonces se puede multiplicar en primer lugar la expresión anterior por V^T.

$$\beta^{-1}\beta(WDRW^T + Id)(WRW^T + Id)^{-1}W = WD.. \tag{4.36}$$

Además al ser $k = d_t$, entonces W es de rango máximo y podemos multiplicar a la izquierda por su inversa y a la derecha por $W^{-1}(WRWT + Id)W$.

$$\beta^{-1}\beta(DRWTW + Id) = D(RWTW + Id). \tag{4.37}$$

Reordenando, se llega a la condición que debe cumplir nuestra matriz W

$$W^TW = [\beta(I - D) - I](DR)^{-1}. \tag{4.38}$$

que es una matriz diagonal (W no tiene que serlo pero si W^TW). Esto no caracteriza de manera única la matriz W.

Ahora si nos fijamos $W^TWRW^T = \left(W^TWR\right)W^T$ siendo W^TWR diagonal, lo que nos indica que W^T son los autovectores de WRW^T con autovalores los elementos de la matriz diagonal W^TWR. Esto nos permite escribir el determinante como el producto de los autovalores de la matriz que son $||w_i^T||^2 r_i$ donde $||w_i^T||^2$ es el i-ésimo elemento de la diagonal de W^TW. Para la descripción completa, sea Q la matriz invertible que nos permite escribir WRW^T en su forma diagonal de autovalores ($QWRW^TQ^{-1} = W^TWR$).

$$\begin{aligned}|A\Sigma_X A^T + I_{d_t}| = |WV\Sigma_X V^T W^T + I_{d_t}| = |WRW^T + I_{d_t}| = |Q||WRW^T + I_{d_t}||Q^{-1}| = \\ = |QWRW^TQ^{-1} + Id| = |W^TWR + I_{d_t}| = \prod_{i=1}^{d_t}\left(\|w_i^T\|^2 r_i + 1\right).\end{aligned} \tag{4.39}$$

Razonando análogamente:

$$|A\Sigma_{X|Y}A^T + I_{d_t}| = |WV\Sigma_{X|Y}V^TW^T + I_{d_t}| = |WRDW^T + I_{d_t}| = |Q||WDRW^T + I_{d_t}||Q^{-1}| =$$
$$= |QWRDW^TQ^{-1} + Id| = |W^TWRD + I_{d_t}| = \prod_{i=1}^{d_t} \left(\|w_i^T\|^2 r_i\lambda_i + 1\right). \quad (4.40)$$

Reemplazamos esta información en la función objetivo:

$$L = (1-\beta)\sum_{i=1}^{n} \log\left(||w_i^T||^2 r_i + 1\right) + \beta\sum_{i=1}^{n} \log\left(||w_i^T||^2 r_i\lambda_i + 1\right). \quad (4.41)$$

Esto muestra que L depende solo de la norma de las filas de W^T, es decir de la norma de las columnas de W. De esta forma podemos optar por tomar W como la matriz diagonal que según $W^2 = W^TW = [\beta(I-D) - I]\,(DR)^{-1}$, se tiene que:

$$W = \sqrt{[\beta(I-D) - I]\,(DR)^{-1}}. \quad (4.42)$$

$$w_{ii} = \sqrt{\frac{\beta(1-\lambda_i) - 1}{r_i\lambda_i}}. \quad (4.43)$$

□

Esto completa la prueba para rango máximo. Para probarlo para $k < d_t$, el primer paso es probar que cualquier matriz de rango bajo es equivalente en términos del valor de la función objetivo a una matriz de rango bajo que tiene solo k filas no nulas.

Para ello consideremos una matriz W con rango $k < d_t$, pero sin filas nulas. Sea U el conjunto de autovectores por la izquierda de WW^T (es decir, $WW^T = U\Lambda U^T$). Entonces, dado que WW^T es simétrica, sus autovectores pueden ser elegidos ortonormales ($U^TU = UU^T = I$), por lo tanto $(UW)(UW)^T = \Lambda$ con $d_t - k$ autovalores nulos.

Esto nos permite ver que la matriz $W_0 = UW$, está compuesto por k filas no nulas y $d_t - k$ columnas nulas. Veamos que W_0 obtiene el mismo valor de la función objetivo y por tanto podemos suponer lo mencionado.

$$\begin{aligned}
L_\beta &= (1-\beta)\log\left(|W_0RW_0^T + I|\right) + \beta\log\left(|W_0DRW_0^T + I|\right) \\
&= (1-\beta)\log\left(|UWRW^TU^T + UU^TI|\right) + \beta\log\left(|UWDRW^TU^T + UU^TI|\right) \\
&= (1-\beta)\log\left(|U||WRW^T + I||U^T|\right) + \beta\log\left(|U||UWDRW^TU^T + I||U^T|\right) \\
&= (1-\beta)\log\left(|WRW^T + I|\right) + \beta\log\left(|WDRW^T + I|\right).
\end{aligned}$$

donde hemos utilizado el hecho de que U es ortonormal y, por lo tanto, $|U| = 1$.

Notamos además que las filas no nulas de W_0 también tienen $d_t - k$ columnas nulas y, por lo tanto, definen una matriz cuadrada de rango k, para la cual se aplica la prueba anterior del caso de rango completo pero esta vez proyectando a una dimensión k en lugar de d_t.

Finalmente, queda un único paso que consiste en probar que el mínimo global cuales de los autovalores y autovectores de $\Sigma_{X|Y}\Sigma_X^{-1}$ participara finalmente en esa matriz W diagonal.

Del Lema 4.1. se deduce que para cada autovalor que participe en la matriz W, se tiene que cumplir que el interior de la raíz cuadrada sea positivo y, al ser λ_i y r_i positivos, necesariamente se ha de cumplir que $\beta \geq \frac{1}{1-\lambda_i}$ para todo i, entonces solo aquellos autovalores que cumplan esta condición podrán contribuir. Ahora bien, para elegir con cuales de estos nos quedamos, escribimos en primer lugar la función de coste en función de los λ_i:

$$L = \sum_{i=1}^{k} \left((\beta - 1) \log \lambda_i + \log(1 - \lambda_i) \right) + f(\beta). \tag{4.44}$$

El comportamiento de esta función es independiente para cada autovalor y se trata de la misma función que para el caso de un único autovalor. Concluimos así que han de participar todos aquellos autovectores y autovalores que cumplan la condición 4.30 siendo estos lo más pequeños posibles.

Como conclusión de los cálculos anteriores, enunciamoss el resultado que nos da la solución del problema **IB** en el caso Gaussiano.

Teorema 4.2.6 (**IB en el caso Gaussiano**)**.** Sea $X \sim \mathcal{N}(\mu_x, \Sigma_X)$ e $Y \sim \mathcal{N}(\mu_y, \Sigma_Y)$ vectores columnas multivariantes Gaussianos de dimensiones d_x y d_y, respectivamente. De forma que conjuntamente forman una distribución Gaussiana $(X, Y)^T \sim \mathcal{N}(\mu, \Sigma)$ donde $\mu = (\mu_x, \mu_y)^T$ y $\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{bmatrix}$. La solución T, del **IB** que minimiza la función objetivo $L_\beta(P_{T|X}) = I(X;T) - \beta I(T;Y)$ en este caso Gaussiano, tiene la siguiente forma:

$$T \overset{d}{=} AX + Z, \tag{4.45}$$

donde $Z \sim \mathcal{N}(0, I_{d_t})$ independiente de X y la matriz A es la siguiente:

$$A = \begin{bmatrix} \sqrt{\frac{\beta(1-\lambda_1)-1}{r_1\lambda_1}} v_1 \\ \sqrt{\frac{\beta(1-\lambda_2)-1}{r_2\lambda_2}} v_1 \\ \cdots \\ \sqrt{\frac{\beta(1-\lambda_k)-1}{r_k\lambda_k}} v_k \\ 0 \\ 0 \end{bmatrix} \in \mathcal{M}_{d_t \times d_x}(\mathbb{R}). \tag{4.46}$$

Donde $k \leq d_t$ y $\lambda_1, \lambda_2, ..., \lambda_k$ son los primeros k autovalores de $\Sigma_{X|Y}\Sigma_X^{-1}$ ordenados de menor a mayor y que cumplan $\lambda_i \leq \frac{\beta-1}{\beta}$ para todo i. Por otro lado, $v_1, v_2, ..., v_k$ son los autovectores asociados.

Capítulo 5

Aplicación del IB en regresión en alta dimensión

Una vez visto en que consiste exactamente el **IB** y su aplicación a las distribuciones Gaussianas, el objetivo siguiente es ver como este método se comporta en algún caso concreto [19, 20]

Con esta idea en mente, lo que haremos en este capítulo es estudiar lo que ocurre para una regresión lineal en dimensiones altas. Para ello, vamos a atacar el problema desde dos puntos de vista distintos, el primero, que se corresponde con el **IB**, y el segundo, que recibe el nombre de **Regresión de Gibbs**.

El enfoque que vamos a llevar a cabo consiste en tomar el resultado que se extraiga del **IB** como el resultado ideal, es decir, aquel que presenta la menor información residual posible ($I(Y;T|X)$) mientras que mantiene una mayor información de la variable de estudio ($I(T;Y)$) y de esta forma, se va a estudiar el resto de enfoques de forma comparativa.

Además, como el objetivo principal por el que surge el **IB** es explicar el **doble descenso** característico de los modelos sobreparametrizados, se va a estudiar esta comparativa descrita anteriormente para diferentes modelos de complejidad variable. Para ello, vamos a pasar de modelos donde el número de datos de entrenamiento es superior al número de parámetros implicados, a el caso totalmente contrario, es decir, el caso sobreparametrizado.

5.1. Modelo a estudiar: regresión lineal

Consideremos una muestra de entrenamiento dada por N pares independientes e idénticamente distribuidos, $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, donde x_i es un vector en $\mathbb{R}^P$ y que representa el atributo mientras que la etiqueta y_i es una respuesta escalar.

Suponemos además una relación lineal entre el atributo y la etiqueta:

$$y_i = W \cdot x_i + \epsilon_i \quad \text{y} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{5.1}$$

donde $W \in \mathbb{R}^P$ representa una regresión lineal desconocida y ϵ_i es cierto ruido Gaussiano escalar

con media cero y varianza σ^2.

De esta forma, si definimos $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$ y $X = (x_1, \ldots, x_N) \in \mathbb{R}^{P \times N}$, de forma que agrupamos todos los datos de entrenamiento disponibles, es posible describir la relación que presentan los *outputs* y los *inputs* según la siguiente distribución Gaussiana:

$$Y \mid X, W \sim \mathcal{N}(X^T W, \sigma^2 I_N), \tag{5.2}$$

que si uno se fija bien, implica intuir la independencia entre los pares de los datos de entrenamiento al consistir en una matriz de covarianzas diagonal.

Por otro lado, vamos a adoptar un escenario en donde los atributos o *inputs*, X, son deterministas y solo las respuestas Y son variables aleatorias. En otras palabras, vamos a suponer que X es una matriz $X \in \mathbb{R}^{P \times N}$ fija y las únicas variables que tienen una distribución conjunta son Y y W, de forma que $Y|X, W = Y|W$.

Este acercamiento nos permite escribir la información mutua entre los datos de entrenamiento $S = (X, Y)$ y cualquier variable aleatoria A como $I(A; S) = I(A; X, Y) = I(A; Y)$. Además, en los análisis siguientes, utilizaremos Y para hacer referencia a los datos de entrenamiento S.

Consideremos un modelo de **efectos aleatorios**, es decir, vamos a asumir que el vector W que representa esa trasformación o esa regresión lineal sigue una distribución también Gaussiana centra en cero y con matriz de covarianzas igual a $\frac{\omega^2}{P} I_P$. Esta forma de trabajar también es conocida en inglés como **random effects**.

$$W \sim \mathcal{N}(0, \frac{\omega^2}{P} I_P), \tag{5.3}$$

donde se ha definido la covarianza de esa manera para que la fuerza de la desviación de este vector no dependa de la dimensión P en la que se trabaje ya que así $E\|W\|^2 = \omega^2$.

Es importante darse cuenta de que la notación ha cambiado respecto a la que se utilizó en la presentación y análisis del **IB**. Ahora, nuestra variable intermedia es Y, mientras que la variable a predecir es W. La cadena de Markov que se tiene es la siguiente:

$$W \longleftrightarrow Y \longleftrightarrow T_{IB}, \tag{5.4}$$

donde T_{IB} representa la solución del **IB**. Además, puesto que W e $Y|W$ son variables Gaussianas, la distribución conjunta también lo será. Y por tanto, estamos en el caso Gaussiano tratado en el capítulo 4.

5.2. Eficiencia de Información

Cómo se ha dicho en la introducción del capítulo, se va a comparar el comportamiento que dos aproximaciones distintas llevan a cabo para enfrentarse a este problema. Esta comparativa que se va a hacer está directamente relacionada con la información residual que los dos puntos de vista mantienen, es decir, $I(T; Y|W)$.

La razón por la que se hace de esta forma se debe a que T_{IB} es la representación óptima que minimiza esta información residual para un valor fijo de $I(T : W)$. Esto viene como consecuencia

de que el **IB** es la solución de minimizar

$$L_\beta(P_{T|Y}) := I(Y;T) - \beta I(T;W) = I(Y;T|W) - (\beta - 1)I(T;W), \tag{5.5}$$

de forma que fijando el último término, el mínimo de $I(Y;T|W)$ lo alcanza T_{IB}.

Sea ahora T_I la solución al problema que el **IB** propone y sea $\tilde{T}$ la solución propuesta por otro método. Como hemos dicho, se fija inicialmente la información que T nos da de W ($I(W;T)$). Para hacer esto, debido a que $I(W;Y)$ es independiente de T y a la desigualdad que nos proporciona la *Desigualdad de Procesado de Datos*: $I(Y;W) \geq I(T;W)$, definimos la constante,

$$\mu = \frac{I(T_{IB};W)}{I(W;Y)} = \frac{I(\tilde{T};W)}{I(W;Y)}, \tag{5.6}$$

que representa esa pérdida de información. Se tiene así que $0 \leq \mu \leq 1$ y que $I(T_{IB};W) = I(\tilde{T};W) = \mu I(W;Y)$ es contante para μ fijo.

Una vez definida esta constante que da cuenta de la información relevante que se mantiene, definimos la **eficacia de información** a para un valor fijo de información μ como el cociente entre $I(T_{IB};Y|W)$ e $I(\tilde{T};Y|W)$.

$$\eta_\mu = \frac{I(T_{IB};Y|W)}{I(\tilde{T};Y|W)}. \tag{5.7}$$

Este valor representa cuan bueno es $\tilde{T}$ ya que, como se dijo al inicio de la sección, para μ fijo, T_{IB} es la representación que reduce al mínimo la información residual. Es por esto que este factor η_μ esta también $0 \leq \eta_\mu \leq 1$.

5.3. El método IB

Una vez visto el procedimiento comparativo que se va a llevar a cabo, vamos a resolver el problema desde el punto de vista que el **IB** proporciona.

Al tratarse de distribuciones Gaussianas, este desarrollo ya se hizo en el capítulo 6 en el que se llegaba a la expresión analítica de T_{IB} encontrando esa matriz A que permite definir $T_{IB} = AY + Z$. Sin embargo, debido al desarrollo comparativo que se va a hacer, no necesitamos conocer explícitamente A, sino que es suficiente con escribir el valor de las informaciones de $I(Y;T_{IB}|W)$ e $I(W;T_{IB})$.

Para ello, se van a hacer uso de una serie de propiedades ya mostradas en capítulos previos y relacionadas con las matrices de covarianzas junto con los resultados de distribuciones Gaussianas e **IB** del capítulo previo.

Lema 5.3.1. La información mutua entre T y W para el resultado que el **IB** propone en la situación estudiada en este capítulo se escribe como

$$I(T_{IB};W) = \frac{1}{2}\sum_i \log\left(\frac{1 - 1/\beta}{\lambda_i}\right), \tag{5.8}$$

donde i recorre todos los autovalores λ_i de $\Sigma_{Y|W}\Sigma_Y^{-1}$ que cumplen la condición $\lambda_i \leq \frac{\beta-1}{\beta}$.

Demostración.

$$\begin{aligned} I(T_{IB};W) &= h(T_{IB}) - h(T_{IB}|W) \\ &= \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\left(|\Sigma_{T_{IB}}|\right) - \frac{n}{2}\log(2\pi e) - \frac{1}{2}\log\left(\left|\Sigma_{T_{IB}|W}\right|\right) \\ &= \frac{1}{2}\log\left(|\Sigma_{T_{IB}}|\left|\Sigma_{T_{IB}|W}\right|^{-1}\right) \\ &= \frac{1}{2}\log\left(\frac{\left|A\Sigma_Y A^T + I_N\right|}{\left|\Sigma_{T_{IB}|W}\right|}\right), \end{aligned} \tag{5.9}$$

donde en el último paso se ha usado que $\Sigma_{T_{IB}} = A\Sigma_Y A^T + \Sigma_Z$. El siguiente paso consiste en aplicar las fórmulas (6.37) y (6.38) junto con (6.41), estas formulas son las que caracterizan la solución del **IB** para el caso Gaussiano.

$$\begin{aligned} I(T_{IB};W) &= \frac{1}{2}\log\left(\frac{\prod_i\left(\frac{\beta(1-\lambda_i)-1}{\lambda_i}+1\right)}{\prod_i\left(\beta(1-\lambda_i)\right)}\right) \\ &= \frac{1}{2}\sum_i \log\left(\frac{\beta(1-\lambda_i)-1+\lambda_i}{\beta\lambda_i(1-\lambda_i)}\right) \\ &= \frac{1}{2}\sum_i \log\left(\frac{1-1/\beta}{\lambda_i}\right). \end{aligned} \tag{5.10}$$

Para todos aquellos autovalores λ_i de $\Sigma_{Y|W}\Sigma_Y^{-1}$ que cumplen la condición $\lambda_i \leq \frac{\beta-1}{\beta}$. □

Lema 5.3.2. La información mutua entre T e Y dado W, es decir, la información residual para el resultado que el **IB** propone en la situación estudiada en este capítulo se escribe como

$$I(T_{IB};W) = \frac{1}{2}\sum_i \log\left(\beta(1-\lambda_i)\right), \tag{5.11}$$

donde i recorre todos los autovalores λ_i de $\Sigma_{Y|W}\Sigma_Y^{-1}$ que cumplen la condición $\lambda_i \leq \frac{\beta-1}{\beta}$.

Demostración. La demostración es análoga al caso previo. Sin embargo, si es necesario hacer uso del hecho de que $W \longleftrightarrow Y \longleftrightarrow T_{IB}$ sea una cadena de Markov. Esto implica que $P_{T_{IB}|W;Y} = P_{T_{IB}|Y}$ y por tanto $\Sigma_{T_{IB}|W;Y} = \Sigma_{T_{IB}|Y}$.

$$\begin{aligned} I(T_{IB};Y|W) &= \frac{1}{2}\log\left(\left|\Sigma_{T_{IB}|W}\right|\left|\Sigma_{T_{IB}|W;Y}\right|^{-1}\right) \\ &= \frac{1}{2}\log\left(\left|\Sigma_{T_{IB}|W}\right|\left|\Sigma_{T_{IB}|Y}\right|^{-1}\right) \\ &= \frac{1}{2}\log\frac{\prod_i\left(\beta(1-\lambda_i)\right)}{|\Sigma_Z|} = \frac{1}{2}\log(\beta(1-\lambda_i)), \end{aligned} \tag{5.12}$$

donde se ha usado $\left|\Sigma_{T_{IB}|Y}\right| = |\Sigma_Z|$ y las relaciones (6.38) y (6.41). □

El siguiente paso será entonces caracterizar los autovalores λ_i de $\Sigma_{Y|W}\Sigma_Y^{-1}$ en función de las distribuciones de W e Y.

Dado que $W \sim \mathcal{N}(0, \frac{\omega^2}{P} I_P)$ e $Y \mid W \sim \mathcal{N}(X^T W, \sigma^2 I_N)$ y teniendo en cuenta que para una distribución normal multivariante (W, Y),

$$\Sigma_{Y|W} = \Sigma_Y - \Sigma_{YW} \Sigma_W \Sigma_{YW}^T, \tag{5.13}$$

se puede escribir que:

$$\Sigma_{Y|W} \Sigma_Y^{-1} = \left(I_N + \frac{1}{\lambda^* N} X^T X \right), \tag{5.14}$$

donde $\lambda^* = \frac{\sigma^2}{\omega^2} \frac{P}{N}$.

Podemos ahora escribir los autovalores λ_i en función de los autovalores de $\frac{X^T X}{N}$ a los que denotaremos por ψ_i.

Propiedad 5.3.3. Sea A una matriz y a uno de sus autovalores de forma que $Av = aV$. Entonces se tiene que:

$$(I_N + \alpha A)v = (1 + \alpha a)v \tag{5.15}$$

$$A^{-1} v = \frac{1}{a} v \tag{5.16}$$

Y por tanto, el nuevo autovalor de $(I_N + \alpha A)^{-1}$ es $\frac{1}{1+\alpha a}$.

Aplicando esto, se puede escribir la relación entre los ψ_i y los λ_i como:

$$\lambda_i = \frac{1}{1 + \frac{\psi_i}{\lambda^*}}. \tag{5.17}$$

Otro aspecto que hay que tener en cuenta, es la condición que los λ_i debían cumplir para formar parte del sumatorio de las expresiones (7.8) y (7.9). Esta condición es $\lambda_i \leq \frac{\beta-1}{\beta}$ y es por tanto necesario trasformarla para que acote los ψ_i.

$$\begin{aligned} \lambda_i &\leq \frac{\beta - 1}{\beta} \\ \frac{\beta}{\beta - 1} &\leq 1 + \frac{\psi_i}{\lambda^*} \\ \frac{\lambda^* \beta}{\beta - 1} - \lambda^* &\leq 1 + \psi_i \\ \left(\frac{\lambda^*}{\beta - 1} \right) &\leq \psi_i. \end{aligned} \tag{5.18}$$

Denotamos por $\psi_c = \frac{\lambda^*}{\beta - 1}$ a esa cota inferior que limita la participación de los autovalores de $\frac{X^T X}{N}$ en las informaciones mutuas.

De esta forma, podemos escribir las informaciones (7.8) y (7.9) como sigue:

$$I(T_{IB};W) = \frac{1}{2}\sum_{\psi_i\geq\psi_c}\log\left(\frac{1-1/\beta}{\lambda_i}\right) = \frac{1}{2}\sum_{\psi_i\geq\psi_c}\log\left(1+\frac{\psi_i-\psi_c}{\psi_c+\lambda^*}\right). \tag{5.19}$$

$$\begin{aligned} I(T_{IB};Y\mid W) &= \frac{1}{2}\sum_{\psi_i\geq\psi_c}\log\left(\beta(1-\lambda_i)\right) = \frac{1}{2}\sum_{\psi_i\geq\psi_c}\log\left(\frac{\psi_i(\lambda^*+\psi_c)}{\psi_c(\lambda^*+\lambda^*)}\right) \\ &= \frac{1}{2}\sum_{\psi_i\geq\psi_c}\log\left(\frac{\psi_i}{\psi_c}\right) - \frac{1}{2}\sum_{\psi_i\geq\psi_c}\log\left(\frac{\lambda^*+\psi_c}{\lambda^*+\psi_i}\right) \\ &= \frac{1}{2}\sum_{\psi_i\geq\psi_c}\log\left(\frac{\psi_i}{\psi_c}\right) - I(T_{IB};W). \end{aligned} \tag{5.20}$$

Estas serian las expresiones con las que poder calcular las diferentes informaciones mutuas para el caso Gaussiano. Es importante fijarse que estas depende en primer lugar del vector X elegido, también dependen de las varianzas ω^2 y σ^2 y también de las dimensiones en las que se este trabajando P. Pero no solo eso, también depende número de datos de entrenamiento N y de la constante β que mide el **IB**.

Por otro lado, dado que vamos a trabajar en altas dimensiones, el número de autovalores va a ser muy grande y por tanto, vamos a sustituir el sumatorio por una integral.

Para ello se va a usar la función de distribución empírica asociada a los autovalores de $\Psi := \frac{X^TX}{N}$ a la que denotaremos por $dF^{\Psi}(\psi) = f^{\Psi}(\psi)d\psi$. Aplicando esta densidad y no olvidándonos de que solo aquellos $\phi_i \geq \phi_c$ participan en el sumatorio, es posible escribir el resultado final al que buscábamos llegar:

$$I(T_{IB};W) = \frac{P}{2}\int_{\psi\geq\psi_c}\log\left(1+\frac{\psi-\psi_c}{\psi_c+\lambda^*}\right)dF^{\Psi}(\psi). \tag{5.21}$$

$$I(T_{IB};Y\mid W) = \frac{P}{2}\int_{\psi\geq\psi_c}\log\left(\frac{\psi}{\psi_c}\right)dF^{\Psi}(\psi) - I(T_{IB};W). \tag{5.22}$$

La idea principal de aplicar este cambio a una integral es encontrar una densidad continua aproximada que sigan los autovalores de $\Psi := \frac{X^TX}{N}$ y cuando las dimensiones de este vector tiendan a infinito, simplificar los cálculos.

Finalmente y antes de pasar al estudio del problema desde el punto de vista de **Regresión de Gibbs**, como la información mutua de $I(W;T)$ la íbamos a tomar como un valor fijo obtenida a partir de un cierto μ entre 0 y 1 de forma que $\mu = I(W;T)/I(Y;W)$, necesitamos también caracterizar esa $I(W;T)$.

Lema 5.3.4. La información mutua entre W e Y, es decir, la información mutua que limita $I(W;T)$ se escribe como

$$I(Y;W) = \frac{1}{2}\sum_i \log\left(\beta(1-\lambda_i)\right), \tag{5.23}$$

donde i recorre todos los autovalores λ_i de $\Sigma_{Y|W}\Sigma_Y^{-1}$ que cumplen la condición $\lambda_i \leq \frac{\beta-1}{\beta}$.

Demostración. La demostración es aún más sencilla que los lemas anteriores ya que $W \sim \mathcal{N}(0, \frac{\omega^2}{P} I_P)$ e $Y \mid W \sim \mathcal{N}(X^T W, \sigma^2 I_N)$.

$$\begin{aligned} I(Y;W) = h(Y) - h(Y|W) &= \frac{1}{2}\log\left(|\Sigma_Y||\Sigma_{Y|T}|^{-1}\right) \\ &= \frac{1}{2}\log\left(\frac{|\Sigma_Y||\Sigma_Y^{-1}|}{|\Sigma_{Y|T}\Sigma_Y^{-1}|}\right) = \frac{1}{2}\log\left(\frac{1}{|\Sigma_{Y|T}\Sigma_Y^{-1}|}\right) \\ &= \frac{1}{2}\sum_i \log\left(\frac{1}{\lambda_i}\right) = \frac{1}{2}\sum_i \log\left(1 + \frac{\psi_i}{\lambda^*}\right) \end{aligned} \tag{5.24}$$

□

En términos de la distribución F^Ψ, la conlcusión del Lema 5.3.4

$$I(Y;W) = \frac{P}{2}\int_{\psi \geq 0} \log\left(1 + \frac{\psi}{\lambda^*}\right) dF^\Psi(\psi). \tag{5.25}$$

Es importante darse cuenta de que esta integral no depende de T, es exactamente igual para está solución y para la próxima que veamos. Además, el intervalo de integración no está restringido por ψ_c como ocurría con las informaciones mutuas previas.

5.4. Regresión de Gibbs

Se va a analizar una variante de la **Regresión ridge** a la luz de la teoría **IB**. Esta solución que este problema propone recibe el nombre de **Regresión de Gibbs** y se basa en el algoritmo de regresión lineal por mínimos cuadrados, uno de los más conocidos y utilizados en la práctica. Además, este ha demostrado ser especialmente adecuado para el análisis del aprendizaje en el régimen sobreparametrizado, el sobreajuste benigno y el doble descenso en modelos sobreparametrizados.

Para inferir un modelo a partir de los datos se requiere una suposición sobre una clase de modelos, que define el espacio de hipótesis. La regresión lineal restringe la clase de modelos a una aplicación lineal, parametrizado por $T \in \mathbb{R}^P$, entre una entrada $x_i^T \in \mathbb{R}^P$ y una respuesta $\hat{y}_i \in \mathbb{R}$ cuyo objetivo es predecir los datos de entrenamiento $y_i \in \mathbb{R}$:

$$\hat{y}_i = T \cdot x_i. \tag{5.26}$$

Para caracterizar está solución se busca aquella que minimice el error cuadrático medio:

$$L(T,Y) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 = \frac{1}{N}\|Y - X^T T\|_2^2 \tag{5.27}$$

La solución a este problema es muy conocida y adopta la siguiente forma:

$$T^* = (XX^T)^{-1}XY. \tag{5.28}$$

Sin embargo, esta solución determinista requiere que $XX^T \in \mathbb{R}^{P\times P}$ sean invertibles y por lo tanto no funciona en el régimen sobreparametrizado ($P \geq N$), donde el rango máximo es igual al mín(P, Q) puesto que rango(XX^T) =rango(X) y por tanto, existen infinitas soluciones con error cuadrático medio nulo.

Con el objetivo de romper esta degeneración, aparece la **regresión de *ridge*** en inglés [16, 24], que añade a ese error cuadrático medio un término adicional proporcional al módulo al cuadrado de T de forma que de entre todas las soluciones posibles para T, se elija la que tenga norma L_2 más pequeña. El problema se reduce entonces a encontrar T tal que minimice la siguiente función de pérdida:

$$L(T, Y) = \frac{1}{N}\|Y - X^T T\|_2^2 + \lambda\|T\|_2^2, \tag{5.29}$$

donde $\lambda > 0$ controla la fuerza de la regresión. Minimizar esta función de pérdida conduce a una solución única incluso en el caso sobreparametrizado.

$$T_\lambda^* = (XX^T + \lambda N I_P)^{-1} XY. \tag{5.30}$$

ya que si $P > N$, el término distinto de cero que se añade con la identidad I_P, elimina la degeneración.

Sin embargo, este modelo que calcula T de forma determinista no es válido para el estudio comparativo que se va a llevar a cabo en relación a esas informaciones mutuas explicadadas en secciones previas. Esto se debe a que si tenemos una función determinista $T = g(Y)$, aunque g esté debidamente definido, la información mutua $I(Y;T) = I(Y; g(Y))$ diverge al no existir función de densidad conjunta de (Y, T).

Para hacer frente a este problema, entra en juego la **inferencia Bayesiana**, en concreto la **distribución a posteriori**. Esta se basa parte de una estimación **a priori** de la función de distribución de un parámetro a estimar, $p(\theta)$. Además, se tiene una serie de observaciones X que dependen de esa variable según $p(X|\theta)$. Entonces, la distribución asociada a θ **a posteriori** se define como $p(\theta|X) \propto p(\theta)p(X|\theta)$.

En nuestro caso, el parámetro a estimar es T en la regresión lineal

$$y_i = \cdot x_i + \epsilon_i \quad \text{donde} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{5.31}$$

y el conjunto de datos que se tienen para la estimación de T es Y, (recordemos que X es determinista). De esta forma, se tiene que $Y|T \sim \mathcal{N}(X^T T; \sigma^2 I_N)$. Además, como se menciono al principio del capítulo, la distribución **a priori** de T es $T \sim \mathcal{N}(0, \frac{\omega^2}{P} I_P)$.

Aplicando finalmente la definición de **distribución a posteriori** se tiene que:

$$\begin{aligned} P_{T|Y} &\propto P_T P_{Y|T} \propto e^{-\frac{1}{2\omega^2}||T||^2} e^{-\frac{1}{2\sigma^2}||Y - X^T T||^2} \\ &\propto e^{-\frac{1}{2\omega^2}\left(||Y-X^T T||^2 + \lambda||T||^2\right)} \propto e^{-\alpha L(T,Y)} \end{aligned} \tag{5.32}$$

Donde se ha llevado a cabo una agrupación y renombramiento de constantes.

Esta distribución $P_{T|Y} \propto e^{-\alpha L(T,Y)}$ es conocida como **distribución de Gibbs** o **distribución de Boltzmann** [4]. Esta distribución por primera vez en el ámbito de la física en el estudio de la

mecánica estadística sobre sobre la población de los niveles energéticos discretos a una temperatura dada. Esta probabilidad se basa en el hecho de que cuanto menor sea la energía de este estado, mayor será la probabilidad de ocupación. Por tanto, se puede escribir como sigue:

$$p_i = \frac{e^{-\varepsilon_i/kT}}{\sum_{i=1}^{M} e^{-\varepsilon_i/kT}}, \tag{5.33}$$

donde p_i es la probabilidad del estado i, ε_i es la energía del estado i, k es la constante de Boltzmann, T es la temperatura del sistema y M es el número de todos los estados accesibles al sistema.

Aplicando está idea pero de forma continua a la **regresión ridge**, si tomamos como "energía"la función de pérdida, aquellos candidatos de T con menor pérdida tendrán una probabilidad mayor de ser elegidos.

$$P_{T|Y} \propto e^{-\alpha L(T,Y)}, \tag{5.34}$$

donde α mide la la fuerza con la que actúa la exponencial negativa. Si uno se da cuenta, el candidato T con mayor probabilidad es exactamente el que menor perdida tenga, es decir, la solución T_λ^* que además se obtiene en el límite de $\alpha \to \infty$.

Por otro lado y para seguir con el estudio de este modelo, si sustituimos la función de pérdida en la exponencial de Boltzmann, agrupamos términos y normalizamos, se demuestra que la distribución de T condicionada de Y sigue una distribución normal multivariante centrada en T_λ^* y con matriz de covarianzas igual a $\Sigma_{T|Y} = \frac{1}{2\beta}\left(\Psi + \lambda I_P\right)^{-1}$ donde $\Psi = \frac{XX^T}{N}$.

$$T \mid Y \sim \mathcal{N}(\frac{1}{N}\left(\Psi + \lambda I_P\right)^{-1} XY, \frac{1}{2\beta}\left(\Psi + \lambda I_P\right)^{-1}) \tag{5.35}$$

Una vez tenemos como se comporta nuestro candidato de estudio T dado Y, es posible escribir la distribución conjunta de (W, Y, T):

$$P_{W,Y,T} = P_W \otimes P_{T,Y|W} = P_W \otimes P_{Y|W} \otimes P_{T|Y,W} = P_W \otimes P_{Y|W} \otimes P_{T|Y}, \tag{5.36}$$

donde en la última igualdad se ha usado que $W \longleftrightarrow Y \longleftrightarrow T_{IB}$, es una cadena de Markov. Las tres distribuciones del último término de la igualdad las tenemos ya descritas y son Gaussianas:

$$W \sim \mathcal{N}(0, \frac{\omega^2}{P} I_P) \tag{5.37}$$

$$Y \mid W \sim \mathcal{N}(X^T W, \sigma^2 I_N) \tag{5.38}$$

$$T \mid Y \sim \mathcal{N}(\frac{1}{N}\left(\Psi + \lambda I_P\right)^{-1} XY, \frac{1}{2\beta}\left(\Psi + \lambda I_P\right)^{-1}) \tag{5.39}$$

Además, al ser todas ellas distribuciones Gaussianas, podemos afirmar que las distribuciones marginales y condicionadas lo serán también. Llevando este cálculo a cabo de las distribuciones marginales ($f_U(u) = \int f_{U,V}(u,v)dv$), agrupando términos y normalizando, se llegan a las siguientes conclusiones:

$$T \mid W \sim \mathcal{N}\left(\Psi(\Psi + \lambda I_P)^{-1} W, \frac{1}{2\beta}\left(\Psi + \lambda I_P\right)^{-1} + \frac{\sigma^2}{N}(\Psi + \lambda I_P)^{-1}\Psi\right) \tag{5.40}$$

$$T \sim \mathcal{N}\left(0, \frac{1}{2\beta}\left(\Psi + \lambda I_P\right)^{-1} + \frac{\sigma^2}{N}(\Psi + \lambda I_P)^{-1}\Psi + \frac{\omega^2}{P}((\Psi + \lambda I_P)^{-1})^2\Psi^2\right) \tag{5.41}$$

Esto nos permite al igual que se hizo para el **IB**, aplicar las fórmulas de información mutua del caso Gaussiano y escribir estas en función de los determinantes de las matrices de covarianza.

Lema 5.4.1. La información mutua entre T y W para el resultado que **Regresión de Gibbs** propone en la situación estudiada en este capítulo se escribe como

$$I(T;W) = \frac{1}{2}\sum_i \log\left(1 + \frac{\psi_i^2/\lambda^*}{\psi_i + \frac{N}{2\beta\sigma^2}(\psi_i + \lambda)}\right), \tag{5.42}$$

donde i recorre todos los autovalores ψ_i de $\Psi = \frac{XX^T}{N}$.

Demostración. Al no ser este el caso del **IB**, la demostración es algo diferente ya que no se puede hacer uso de las igualdades que aparecen en el capítulo 6. En su lugar, simplemente se va a aplicar que como toda matriz simétrica definida positiva, como son las matrices de covarianza, es diagonalizable ortogonalmente, el determinante de estas matrices es igual al producto de sus autovalores.

$$\begin{aligned} I(T;W) &= h(T) - h(T|W) = \frac{1}{2}\log(|\Sigma_T|) - \frac{1}{2}\log(|\Sigma_{T|W}|) \\ &= \frac{1}{2}\sum_i \log\left(\frac{\frac{1}{2\beta(\psi_i+\lambda)} + \frac{\sigma^2\psi_i}{N(\psi_i+\lambda)} + \frac{\omega^2\psi_i^2}{P(\psi_i+\lambda)^2}}{\frac{1}{2\beta(\psi_i+\lambda)} + \frac{\sigma^2\psi_i}{N(\psi_i+\lambda)}}\right) \\ &= \frac{1}{2}\sum_i \log\left(1 + \frac{\psi_i^2/\lambda^*}{\psi_i + \frac{N}{2\beta\sigma^2}(\psi_i + \lambda)}\right) \end{aligned} \tag{5.43}$$

□

Lema 5.4.2. La información mutua entre T e Y condicionado de W para el resultado que **Regresión de Gibbs** propone en la situación estudiada en este capítulo se escribe como

$$I(T;W) = \frac{1}{2}\sum_i \log\left(1 + \frac{2\beta\sigma^2}{N}\frac{\psi_i}{\psi_i + \lambda}\right), \tag{5.44}$$

donde i recorre todos los autovalores ψ_i de $\Psi = \frac{XX^T}{N}$.

Demostración. Analogamente a como se hizo en el lema previo,

$$\begin{aligned} I(T;Y|W) &= h(T|W) - h(T|W;Y) = h(T|W) - h(T|Y) \\ &= \frac{1}{2}\log\left(|\Sigma_{T|W}|\right) - \frac{1}{2}\log\left(|\Sigma_{T|Y}|\right) \\ &= \frac{1}{2}\sum_i \log\left(\frac{\frac{1}{2\beta(\psi_i+\lambda)} + \frac{\sigma^2\psi_i}{N(\psi_i+\lambda)}}{\frac{1}{2\beta(\psi_i+\lambda)}}\right) \\ &= \frac{1}{2}\sum_i \log\left(1 + \frac{2\beta\sigma^2}{N}\frac{\psi_i}{\psi_i + \lambda}\right). \end{aligned} \tag{5.45}$$

□

Aparte de ser diferentes fórmulas, la gran diferencia entre estas radica en que no existe una cota inferior a partir de la cual iniciar ese sumatorio. Además, al ser XX^T una matriz simétrica tal que $w^T XX^T w = (X^T w)^T X^T w = \|X^T w\|_2^2 > 0$ para todo $w \in \mathbb{R}^P - \{0\}$, podemos afirmar que es definida positiva y por tanto todos sus autovalores serán positivos.

En segundo lugar, aplicamos a estos resultados la misma idea que se aplicó previamente al caso del **IB** para reescribir los sumatorios como integrales y así llegar a los resultados finales.

$$I(T;W) = \frac{P}{2} \int_{\psi>0} \log\left(1 + \frac{\psi^2/\lambda^*}{\psi + \frac{N}{2\beta\sigma^2}(\psi+\lambda)}\right) dF^{\Psi}(\psi). \tag{5.46}$$

$$I(T;Y|W) = \frac{P}{2} \int_{\psi>0} \log\left(1 + \frac{2\beta\sigma^2}{N}\frac{\psi}{\psi+\lambda}\right) dF^{\Psi}(\psi). \tag{5.47}$$

5.5. Implementación numérica

Una vez se tiene la forma de calcular las diferentes informaciones mutuas asociadadas tanto al **IB** como al **Regresión de Gibbs**, ya se puede analizar como se comportan estas y comenzar con ese punto de vista comparativo que se comentó al inicio del capítulo. En el contexto que se va a trabajar, se mencionó que $x \in \mathbb{R}^P$ iban a ser fijos, deterministas. Además, se van a tomar inicialmente de forma aleatoria e independiente siguiendo una distribución normal estándar, es decir, $X \sim \mathcal{N}(0, \Sigma_X = I_P)$. Y como el número de muestras aleatorias es N, nuestra matriz determinista que engloba todos los datos de entrenamiento sera $X \in \mathbb{R}^{N\times P}$.

A partir de aquí, denotamos como n al cociente entre el número de datos de entrenamiento N y el número de parámetros implicados en el modelo P, es decir, $n = \frac{N}{P}$. De esta forma, se va a poder caracterizar el caso sobreparametrizado, $n \ll 1$, o el caso $n \gg 1$ y sacar conclusiones de como estos se comportan respecto al caso $n = 1$.

Por otro lado y antes de comenzar a la resolución numérica de estas integrales, vamos a enunciar un Teorema de vital importancia para poder sustituir las distribuciones empíricas de $\Psi = \frac{XX^T}{N}$ por una distribución continua.

Proposición 5.5.1. Teorema de Marchenko-Pastur sea $X \in \mathbb{R}^{N\times P}$ la matriz que representa los datos de entrenamiento deterministas y cuyas entradas son variables aleatorias independientes que siguen una distribución Gaussiana de media cero y varianza unidad. Entonces, fijado $n = N/P$ perteneciente al intervalo $(0,\infty)$ de forma que $P = P(N) = N/n$ y denotando por F_N^{Ψ} a la distribución empírica sobre los autovalores de $\Psi = \frac{XX^T}{N}$. Se tiene que

$$F_N^{\Psi} \xrightarrow{d} F^{\Psi} \quad \text{cuando} \quad N \to \infty, \tag{5.48}$$

siendo F^{Ψ} de la siguiente forma:

$$\frac{dF^{\Psi}(\psi)}{d\psi} = n\frac{\sqrt{(\psi_+ - \psi)(\psi - \psi_-)}}{2\pi\psi}. \tag{5.49}$$

Donde $\psi_{\pm} = \left(1 \pm \frac{1}{\sqrt{n}}\right)^2$. Es importante darse cuenta de que el dominio de definición de está función es $\psi_- < \psi < \psi_+$ que depende a su vez del valor de n.

La demostración de esta proposición excede los objetivos de este trabajo y se puede encontrar detalladamente en [1] de forma que nosotros simplemente la asumiremos como cierta. De manera visual, en la Figura 5.1 se muestra el comportamiento de esta aproximación asociada a los autovalores de Ψ. En ella se representan las densidades de **Marchenko-Pastur** frente a las densidades estimadas con estimadores núcleo para diferentes valores de n y con $N = 1000$.

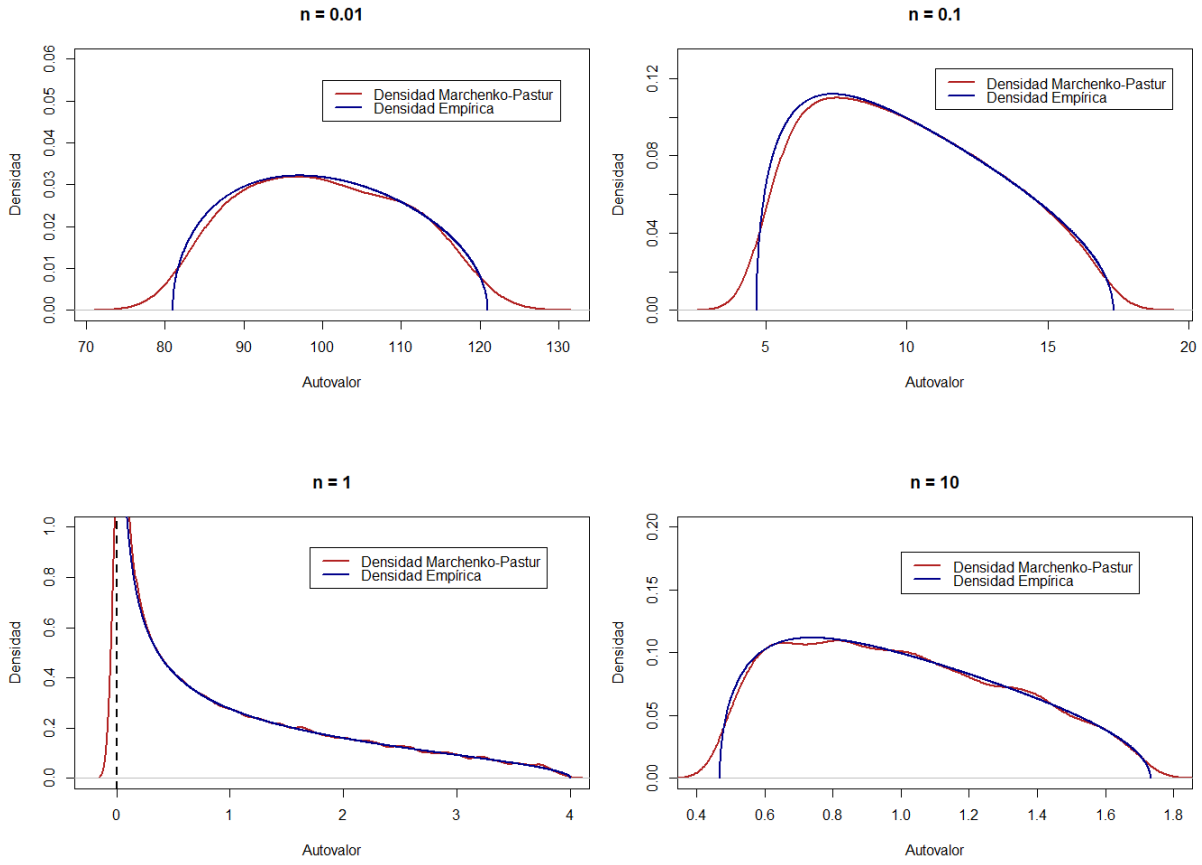


Figura 5.1: Comparación gráfica del comportamiento de la distribución empírica de los autovalores de Ψ con la distribución continua aproximada que eproporciona la ley de Marchenko-Pastur.

En la Figura 5.1, es importante darse cuenta que el dominio de definición de la integral depende notablemente del valor de n. En concreto, cuanto más cerca esta ese parámetro del 0, es decir $n \ll 1$, mayor es el intervalo de definición de la integral y más se aleja este del 1. Por el contrario, si $n > 1$, el intervalos de definición de dF^{Ψ}, se hace cada vez más pequeño en un entorno del 1. Esto nos será importante a tener en cuenta a la hora de analizar resultados.

5.5.1. Estudio individual del IB

Entendida esta aproximación ya nos podemos centrar en el cálculo de las informaciones mutuas. Para ello, vamos a aproximar la información relevante y residual de las expresiones 5.21 y 5.22 con

ayuda del **Teorema de Marchenko-Pastur** en el caso del **IB**:

$$I(T_{IB};W) \approx \frac{P}{2}\int_{\psi\geq\psi_c} dF^{\Psi}(\psi)\log\left(1+\frac{\psi-\psi_c}{\psi_c+\lambda^*}\right). \tag{5.50}$$

$$I(T_{IB};Y\mid W) \approx \frac{P}{2}\int_{\psi\geq\psi_c} dF^{\Psi}(\psi)\log\left(\frac{\psi}{\psi_c}\right) - I(T_{IB};W). \tag{5.51}$$

$$I(Y;W) \approx \frac{P}{2}\int_{\psi\geq 0} dF^{\Psi}(\psi)\log\left(1+\frac{\psi}{\lambda^*}\right). \tag{5.52}$$

De forma que 5.50 representa la **información relevante**, 5.51 la **información residual** y 5.52 la información total disponible.

Lo primero en lo que reparamos es que estas integrales dependen de una serie de parámetros ψ_c y λ^*. Esta dependencia es muy complicada de entender a partir de las expresiones analíticas, por ello, vamos a representar para diferentes valores de $\psi_c/\lambda^* = \frac{1}{\beta-1}$ los valores de $I(T_{IB};W)$ y $I(T_{IB};Y\mid W)$. Estos resultados se han llevado a cabo para diferentes valores de n y además se ha tomado en todo caso $\frac{\sigma^2}{\omega^2}=1$.

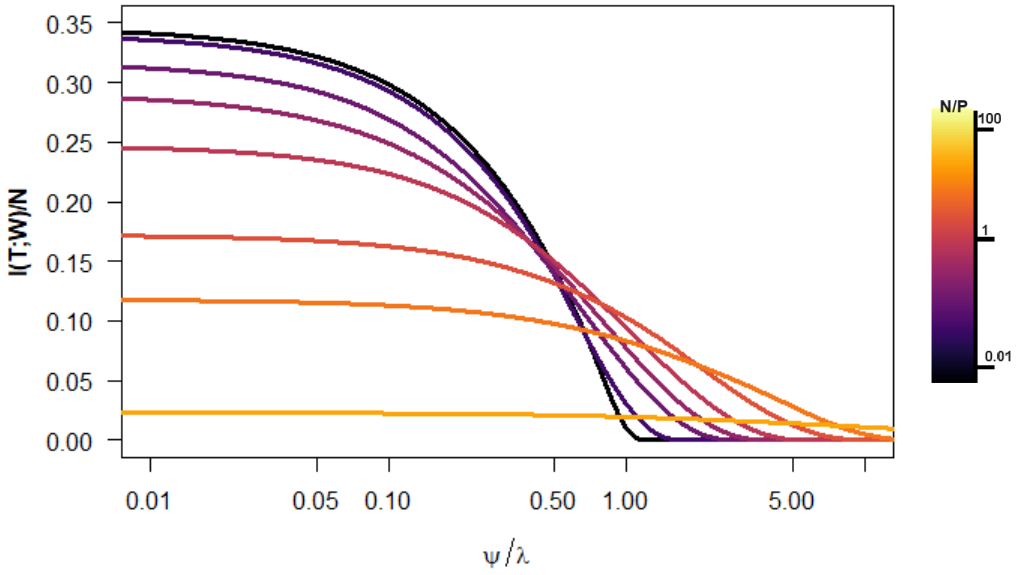


Figura 5.2: Representación logarítmica de la información relevante $I(T_{IB};W)$ relevante normalizada $I(T_{IB};W)/N$, de esta forma se pierde la dependencia que aparece al multiplicar por el factor P) en función del cociente de los parámetros que afectan al valor de la integral para diferentes valores de n.

Para analizar la Figura 5.2, vamos a suponer λ^* fijo, de esta forma, estudiamos como se comporta está información mutua que T_{IB} tiene de W variando ψ_c, que representa ese valor límite que deben cumplir los autovalores para formar parte de T_{IB} a partir de Y. Cuando $\psi_c \to 0$ se observan

rápidamente esa especia de mesetas en las que variando este parámetro ψ_c, la información mutua apenas varía. Esto se debe a que en este rango cercano a cero, la restricción de autovalores es demasiado baja y por tanto no afecta al limite inferior de esa integral.

Por otro lado, cuando esta constante ψ_c alcanza un determinado valor, la información decrece bruscamente. Esto significa que el número de autovalores que colaboran en esa integral es cada vez menor. Y además, como es lógico, cuando este parámetro se hace muy grande, la información se vuelve cero ya que existe un punto en el que ψ_c supera a todos los autovalores de XX^T y la integral se vuelve nula.

En cuanto al comportamiento cualitativo de estas gráficas, observamos que cuanto más cerca estemos del caso sobreparametrizado, más pronunciado será esa caída que se mencionaba en los párrafos previos.

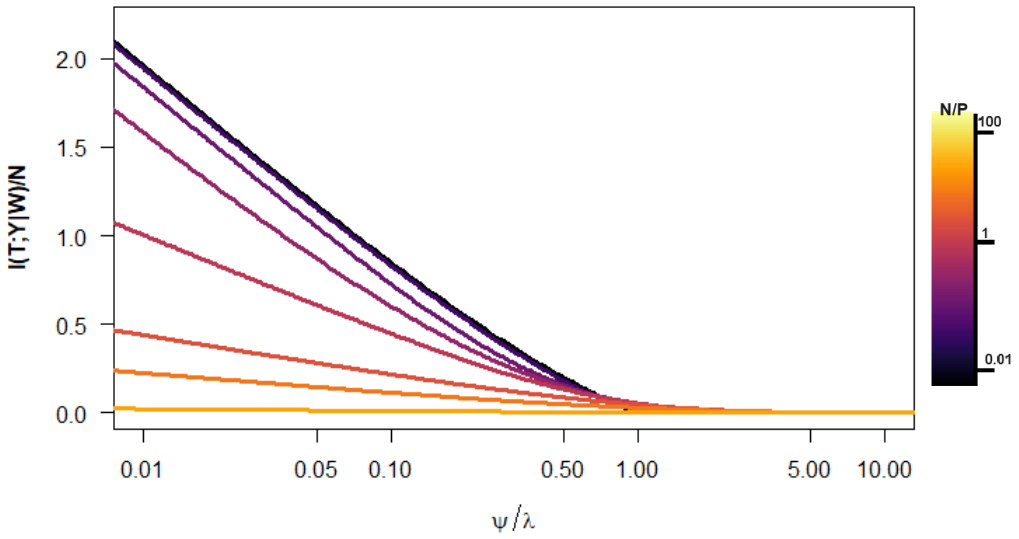


Figura 5.3: Representación logarítmica de la información mutua $I(T_{IB}; Y \sim W)$ normalizada $(I(T_{IB}; W)/N$, de esta forma se pierde la dependencia que aparece al multiplicar por el factor $P)$ en función del cociente de los parámetros que afectan al valor de la integral para diferentes valores de n.

Para analizar la Figura 5.3, vamos a proceder igual que en el caso previo, suponemos λ^* fijo, y así estudiamos el comportamiento de la información residual $I(T; Y \mid W)$ variando ψ_c, que representa ese valor límite que deben cumplir los autovalores para formar parte de T_{IB} a partir de Y. Cuando $\psi_c \to 0$ se observan que ya no aparecen esas mesetas, simplemente cuanto menor es ese valor psi_c, la información residual aumenta. Esto se debe a que conforme participan un mayor número de autovalores en la trasformación de Y a T, estas variables se parecen cada vez más.

Por otro lado, cuando esta constante ψ_c alcanza un determinado valor, la información residual se vuelve cero. La razón de esto es la misma que la del caso previo: el número de autovalores

que colaboran en esa integral es cada vez menor y cuando este parámetro se hace muy grande, la información mutua $I(T;W)$ se vuelve cero ya que existe un punto en el que ψ_c supera a todos los autovalores de XX^T y nuestro estadístico $T = 0$, la información residual será nula también.

Es importante también saber relacionar estas gráficas con el β característico de la función de coste del **IB**. Para ello, recordemos que $\psi_c/\lambda^* = \frac{1}{\beta-1}$ de forma que cuando $\beta \to \infty$, entonces $\psi_c \to 0$. En este caso la función a minimizar $L_\beta(P_{T|X}) := I(X;T|Y) - (\beta-1)I(T;Y)$, se le está dando mucha más importancia a mantener la información relevante $I(T;Y)$ que a reducir la información residual. Por el contrario, cuando $\beta \to 1$, $\psi_c \to \infty$, se tiene que la importancia que se le da a reducir la información residual $I(Y;T|W)$ aumenta con respecto al caso anterior. Por ello, esta se reduce notablemente, pero conlleva una disminución similar de la información relevante.

Entendido como se comportan la información que T guarda de W y la información residual que posee T, vamos a representar el cociente $\mu = \frac{I(T;W)}{I(T;Y)}$ que recordemos que es el porcentaje de información que T preserva de W respecto a la información que había inicialmente en Y, en función de la información residual $I(T;Y|W)$. Esto se muestra en la Figura 5.4.

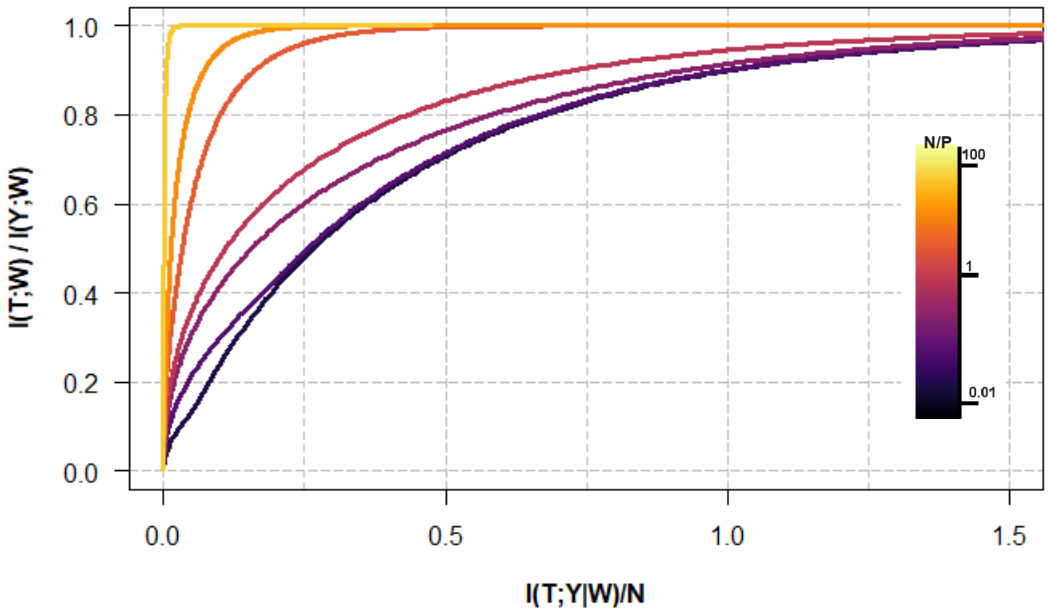


Figura 5.4: Representación de $\mu = \frac{I(T;W)}{I(T;Y)}$, que consiste en una medida de cuanta información preserva T de W de la que había inicialmente, en función de la información residual $I(T;Y|W)$. Todo ello para diferentes valores de n.

Se observa que, independientemente del valor de n, cuanto mayor es la información residual, mayor es la información que T mantiene de W. Esto se debe a que cuanto más es esa información residual significa que más se parece T a Y y por tanto mayor será el porcentaje μ.

Por otro lado, en cuanto al comportamiento de la representación en función de n, se puede apreciar que el caso sobreparametrizado, $P \gg N$ o $n \ll 1$, si uno quiere mantener un elevado valor de $I(T;W)$ cercano a $I(Y;W)$, tiene que ceder también mucha información residual. Por el contrario, el caso en el que el número de datos de entrenamiento N, es superior al número de parámetros P, es decir $n \gg 1$, se tiene que para mantener una información $I(T;W)$ elevada, apenas tiene que mantenerse la información residual.

5.5.2. Estudio individual de la Regresión de Gibbs

De cara a abordar una comparación entre modelos, en primer lugar, vamos a ver como se comporta individualmente este método. Para ellos y de la misma forma que se hicieron las aproximaciones en el caso del **IB**, las integrales a resolver son las siguientes:

$$I(T;W) \approx \frac{P}{2} \int_{\psi>0} dF^{\Psi}(\psi) \log\left(1 + \frac{\psi^2/\lambda^*}{\psi + \frac{N}{2\beta\sigma^2}(\psi+\lambda)}\right). \tag{5.53}$$

$$I(T;Y|W) \approx \frac{P}{2} \int_{\psi>0} dF^{\Psi}(\psi) \log\left(1 + \frac{2\beta\sigma^2}{N}\frac{\psi}{\psi+\lambda}\right). \tag{5.54}$$

$$I(Y;W) \approx \frac{P}{2} \int_{\psi\geq 0} dF^{\Psi}(\psi) \log\left(1 + \frac{\psi}{\lambda^*}\right). \tag{5.55}$$

Donde las integrales representan **la información residual, la información residual y la información total disponible**.

Es evidente que la última información es idéntica al caso del **IB**, y además, ninguna integral tiene el dominio de definición restringido y esta integral depende únicamente de λ^* y de $N/2\alpha\sigma^2$

Por otro lado, vamos a trabajar para $n = 1$, dado que las conclusiones acerca del comportamiento de estos términos que se extraerían sería idénticas, es decir, cambiarían los ordenes de magnitud, pero no la formas que describe ese comportamiento.

También se va a trabajar para $\lambda^* = 1$ fijo y de esta forma variamos únicamente $\kappa := N/2\alpha\sigma^2$ donde α representa esa factor inverso de la temperatura característico de la distribución de Gibbs ($P_{T|Y} \propto e^{-\alpha L(T,Y)}$). Se va a trabajar a su vez, para diversos valores de λ que representa el peso relativo $||T||^2$ en esa función pérdida $L(T,Y) = \frac{1}{N}\|Y - X^T T\|_2^2 + \lambda\|T\|_2^2$.

En la Figura 5.5, se observa que para λ fijo, conforme aumenta el valor de β, es decir $\kappa \to 0$, la $I(T;W)$, se acerca un valor fijo. Este valor, se corresponde con la $I(W;Y)$ ya que en este caso, al tender $\alpha \to \infty$, $P_{T|Y} \propto e^{-\alpha L(T,Y)}$ se convierte en una distribución cada vez más cerca de la solución determinista propuesta por la **regresión ridge** al dar cada vez más peso a la función pérdida.

Por el contrario, cuando $\alpha \to 0$, $\kappa \to \infty$, la información que preserva T de W se vuelve nula. Esto es consecuencia de que al hacer α muy pequeño, $P_{T|Y} \propto e^{-\alpha L(T,Y)}$, se vuelve prácticamente constante, y la distribución de $P_{T|Y}$, se convierte en uniforme independientemente de la función de pérdida, es decir, sin tener en cuenta ese Y, serán así, totalmente independientes.

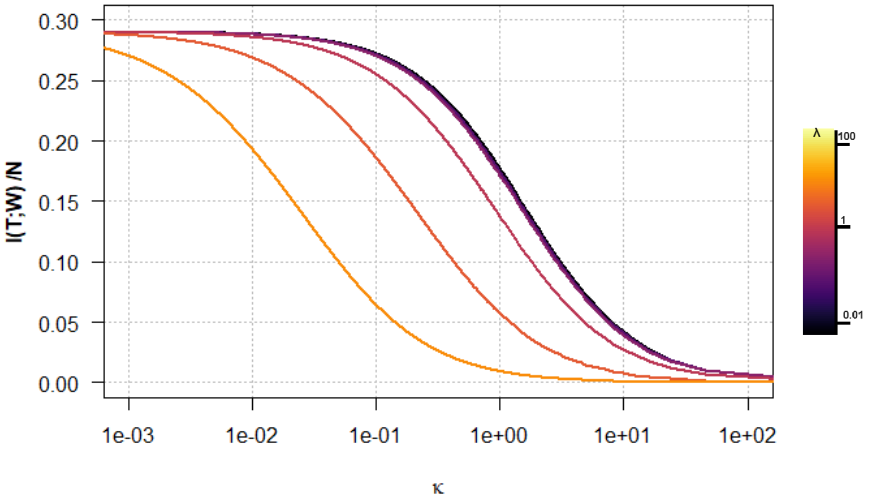


Figura 5.5: Representación logarítmica de la información $I(T_{GP};W)$ normalizada ($I(T_{GP};W)/N$, de esta forma se pierde la dependencia que aparece al multiplicar por el factor P), en función del cociente de los parámetros que afectan al valor de la integral $\kappa := N/2\alpha\sigma^2$ para diferentes valores de λ.

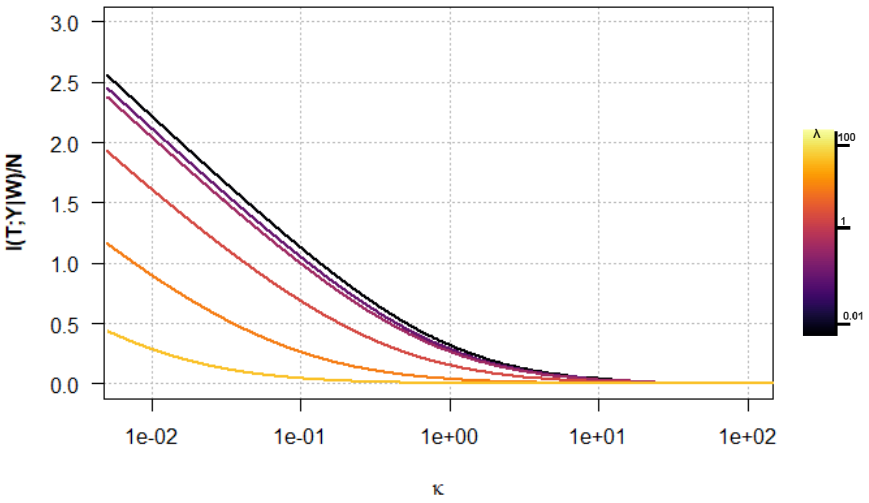


Figura 5.6: Representación logarítmica de la información $I(T_{GP};Y|W)$ normalizada, de esta forma se pierde la dependencia que aparece al multiplicar por el factor P, y en función del cociente de los parámetros $\kappa := N/2\alpha\sigma^2$ que afectan al valor de la integral para diferentes valores de λ.

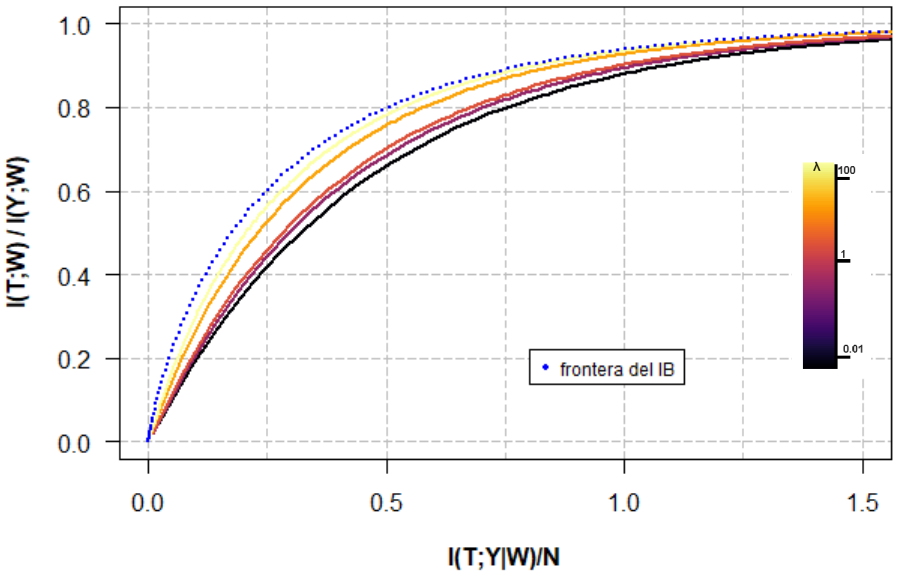


Figura 5.7: Representación del cociente de la información $I(T_{GP};W)$ entre,$I(W;Y)$, en función de la información residual $I(Y;T_{GP}|W)$ para diferentes valores de λ. Además, se muestra también el comportamiento del resultado obtenido para el IB.

En cuanto al comportamiento de la información residual, este se observa en la Figura 5.6.

El razonamiento es análogo al que se hizo previamente para la Figura 5.5, cuando $\beta \to \infty$, el modelo se convierte en determinista y por tanto, la información residual se corresponde exactamente al información que se tendría en el caso $Y = T$. Por el contrario, cuando $\beta \to 0$, el modelo deja de recibir influencia de lo que es la función pérdida y por tanto Y y T, se vuelven independientes y la información mutua se vuelve nula.

En cuanto a la influencia que tiene el valor λ, no se observa ningún comportamiento diferente, simplemente esos crecimientos o decrecimientos se producen para valores mayores de κ. Aun así, si que es posible notar que a mayor λ, la información residual será menor, pero por el contrario, esto supone una perdida mayor de información $I(W;T)$.

Por otro lado, sin cambiar las constantes de trabajo mencionadas en esta sección y al igual que se hizo para el caso del **IB**, en la Figura 5.7 se observa ese comportamiento que tiene la constante μ, medida de la cantidad de información que T retiene de W del total disponible, frente a la información residual que este suponen.

Como se ha mencionado en resultados previos, la curva presenta un comportamiento más ventajoso siempre que tenga una mayor curvatura, es decir, lo que uno busca es que $I(W;T)$ sea lo mayor posible mientras se mantiene la información residual lo más baja posible. con este razonamiento y junto con los resultados observados en la Figura 5.7, es posible deducir que, cuanto mayor es λ, es decir, ese peso que se le asocia al módulo de T en la función de pérdida, mejor se comporta la

regresión de Gibbs. Sin embargo, esto no siempre ocurre así, en [19] (sec 4.2) se trata con mayor detalle este tema y se concluye que la máxima eficacia se obtiene para un valor de λ intermedio que depende de la estructura de los datos de entrenamiento así como de las densidades involucradas. Aún así, por muy eficiente que sea este parámetro λ, nunca se podrá igualar el comportamiento del IB.

La razón de porque ocurre esto es evidente, el **IB** supone la solución óptima que para una información fija $I(W;T)$, minimiza la información residual $I(Y;T|W)$ y por tanto, cualquier otro modelo se encontrará por debajo en este tipo de curvas. Esta es la razón además por la que esa constante η_μ esta correctamente definida en el intervalo $(0,1)$.

5.5.3. Comparativa de ambos modelos

Finalmente, con la intención de realizar una comparativa de modelos para los diferentes casos de n, y así poder estudiar el caso sobreparametrizado, vamos a llevar a cabo la representación de la constante η_μ en función de $n = N/P$. Recordemos que esta constante η_μ representa el cociente entre informaciones residuales asociadas a T_{IB} entre a T_{GP} para un valor fijo de información $I(T_{IB};W) = I(T_{GP};W) = \mu I(Y;W)$. Además, se va a trabajar para $\sigma^2/\omega^2 = 1$ y para $\lambda = 1$

Es importante destacar que mientras las figuras previas se obtenían mediante integración, el proceso de obtención las figuras 5.8 y 5.9 es más complejo ya que conlleva el uso de los resultados previos obtenidos, en concreto de las Figuras 5.7 y 5.4, junto con una interpolación que fijado μ te permita obtener la información residual.

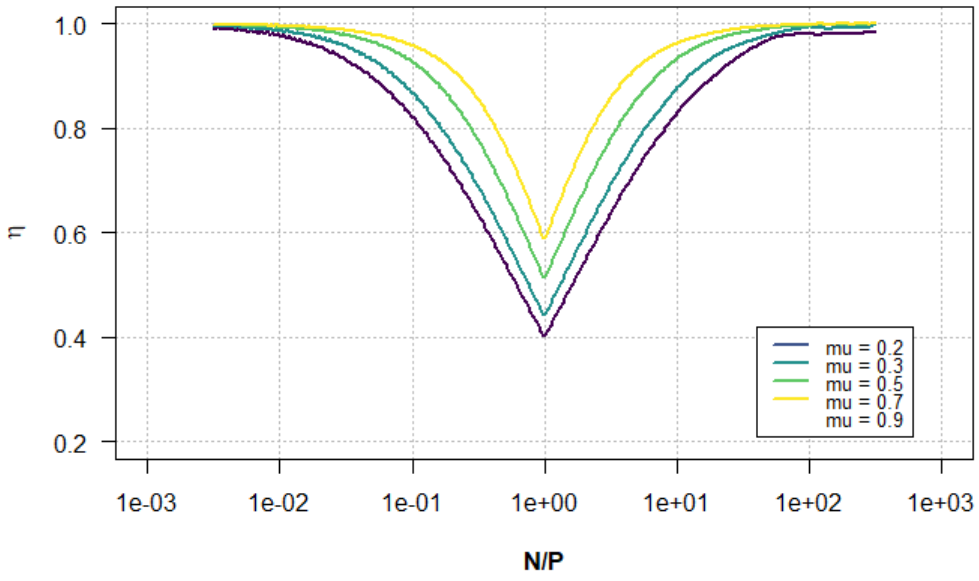


Figura 5.8: Representación logarítmica del cociente $\eta_\mu = \frac{I(T_{IB};Y|W)}{I(T_{GB};Y|W)}$ en función de $n = N/P$ para diferentes valores del grado de información mutua de $I(W;T)$, μ.

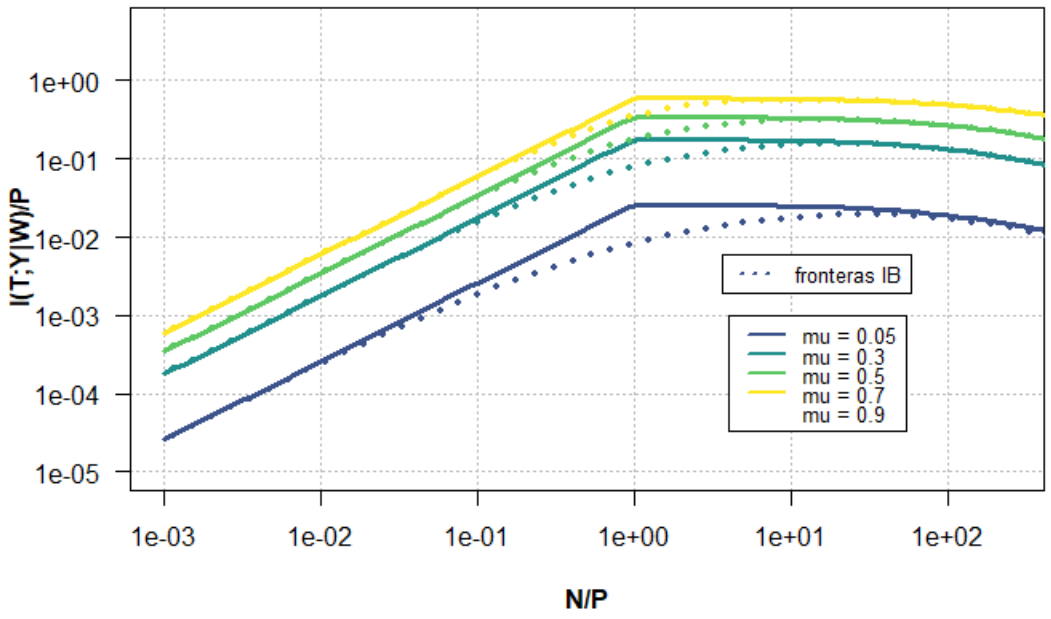


Figura 5.9: Representación doble-logarítmica de la información residual $I(T_{IB};Y|W)$ y $I(T_{GB};Y|W)$ en función de $n = N/P$ para diferentes valores del grado de información mutua de μ.

De cara a analizar esta Figura 5.8, simplemente hay que tener en cuenta que, partiendo de que el modelo del **IB** es el modelo ideal, cuánto más cerca este esa constante η_μ de 1, mejor será el resultado que T_{GP} proporcione ya que implica que para el mismo valor fijo de $I(T_{IB};W) = I(T_{GP};W) = \mu I(Y;W)$, as informaciones residuales son muy parecidas.

Se observa que para un valor fijo de μ, el modelo de Gibbs tiene un comportamiento muy bueno en el caso $N/P \gg 1$. Esto es lógico ya que al tener muchos más datos de entrenamiento que los parámetros implicados, el acercamiento será muy bueno independientemente de cuál se tome. A partir de aquí, conforme aumenta la complejidad del modelo (P) para un mismo número de datos de entrenamiento (N), el rendimiento del modelo de Gibbs empeora respecto al modelo **IB**.

Sorprendentemente, cuando n llega a 1, es decir, que el número de parámetros pasa a ser superior al número de datos de entrenamiento, la relación η_μ vuelve a aumentar de forma simétrica.

Además, en el caso sobreparametrizado en el que el número de parámetros es 3 ordenes de magnitud superior a los datos de entrenamiento, $n = 10^{-3}$, el modelo de Gibbs se comporta, en términos de información mutua, exactamente igual que el **IB**.

Es aquí donde se refleja la idea del **doble descenso**. Pese a que en un principio es lógico pensar que cuanto mayor sea el número de datos de entrenamiento frente al número de parámetros mejor funcionan estos modelos (primer descenso), llega un punto en el que si nos alejamos del paradigma en el que se suele trabajar en el aprendizaje supervisado, $n \approx 1$, el modelo utilizado vuelve a presentar un buen comportamiento con lo que a la información mutua respecta.

Por otro lado, el comportamiento mencionado previamente no depende de μ, este factor solo afectará a cuanto de pronunciado es este pico. De forma que cuanto más cerca este μ de la unidad, menor será el pico de $n = 1$. Esto es lógico ya que cuanto más cerca este μ de 1, significa que $I(T_{IB}; W) = I(T_{GP}; W) = \mu I(Y; W)$ están más cerca de $I(W; Y)$ y por tanto más similares son T_{IB} y T_{GP} a Y, lo que implica informaciones residuales más parecidas.

Otra forma de obtener estas misma conclusiones es representar en lugar del cociente η_μ, la información residual $I(T; Y|W)/P$. Esta normalización que surge de dividir entre P nos permite interpretar la gráfica de la Figura 5.9 como aumentar el número de datos experimentales que se poseen N, mientras la complejidad del modelo permanece constante.

Además haciendo esto también para **IB**, nos permite comparar la eficacia del modelo **Gibbs Posterior**. Se observa que cuando el número de datos es menor que el número de parámetros a estimar, ambas informaciones residuales crecen de igual forma y de manera lineal. Cuando nos acercamos a igualdad de datos de entrenamiento y parámetros, la información residual del **IB**, se comporta mucho mejor que la otra (esto es la razón de los picos de la Figura 5.8). Es esta franja, la que se corresponde con el doble descenso del caso sobreparametrizado.

Y de la misma forma que ocurría para la Figura 5.8, tenemos que para el caso con mayor número de datos de entrenamiento, la diferencia entre lo ideal, **IB**, y lo que está ocurriendo para la **Regresión de Gibbs** vuelve a reducirse notablemente.

Capítulo 6

Conclusiones

Tras un primer análisis de los problemas a los que se enfrenta actualmente el aprendizaje supervisado, en el que hemos podido comprender el **sobreajuste** desde un punto de vista más técnico al que estamos acostumbrados a escuchar, se ha tenido también un primer contacto con un concepto llamado **doble descenso**. Este es un término menos conocido que se ha empezado a usar recientemente y además es la causa por la que se busca desde un primer momento un acercamiento al aprendizaje supervisado basado en la Teoría de la Información.

Por todo ello, esta forma de analizar el problema nos ha permitido un análisis de los diferentes conceptos que este campo engloba como son la entropía, la divergencia de Kullback-Leibler o la información mutua. Además, el hecho de tener que desarrollar unas herramientas como son la **desigualdad de potencias de entropía** o la **desigualdad de procesado de datos**, nos ha exigido una comprensión más profunda y una mayor soltura a la hora de manejar los conceptos previos.

Una vez se comprendió correctamente la base del aprendizaje supervisado y la visión que la Teoría de la Información nos ofrece sobre este tema, se introduce ya el método de **Cuello de Botella de la Información**. Entender este una vez se tiene clara la idea de información mutua es sencillo, sin embargo, su resolución para el caso Gaussiano requiere especial atención a conceptos relacionados con el álgebra matricial o la distribución normal multivariante estudiados durante la carrera.

Por otro lado, el último apartado nos ha permitido ver la idea del **Cuello de Botella de la Información** no solo como una herramienta aplicada a casos concretos, sino que también nos ha dado la posibilidad de medir de forma comparativa cómo se comportan otros modelos y mecanismos de aprendizaje. Este modelo comparativo se basa en la idea de contrastar cualquier otro método de aprendizaje en términos de información mutua con el comportamiento que el **IB** nos ofrece. Esto se debe a que que este último se ha presentado desde un principio como el mecanismo que optimiza la información relevante. En concreto, lo hemos aplicado a la **regresión de Gibbs**, un nuevo acercamiento al aprendizaje supervisado basado en ideas como la estadística Bayesiana o la regresión ridge.

Finalmente, el desarrollo numérico computacional nos ha permitido entender cómo se comportan el **IB** y la **regresión de Gibbs** en los diferentes casos infraparametrizados y sobreparametrizados

y cómo se comportan las informaciones residuales y relevantes en ambos casos. Por último, con ayuda de las información mutua y en función de la complejidad del modelo, es decir, del número de parámetros, se ha podido comprender también de forma más profunda y cuantitativa la idea del **doble descenso**.

Como comentario final del trabajo, si uno quisiera continuar con la investigación en lo que respecta a este tema, se podría en primer lugar buscar métodos más generales de resolución del **IB**, ya que nosotros nos hemos centrado únicamente en el caso Gaussiano, que tiene una forma analítica concreta. Sin embargo, la variedad de problemas de aprendizaje existentes es enorme dependiendo de los datos de entrenamiento, las distribuciones implicadas o los mecanismos de entrenamiento, por lo que encontrar su solución sería muy ventajoso de cara a resolver problemas como el **sobreajsute** o **doble descenso**.

Bibliografía

[1] G. W. Anderson, A. Guionnet, O. Zeitouni, *An Introduction to Random Matrices*, (Cambridge Univesity Press, 2009).

[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, (Wiley New York, 1958).

[3] T. M. Apostol, *Análisis Matemático*, (Reverte, 2020).

[4] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, y G. Wornell, An Exact Characterization of the Generalization Error for the Gibbs Algorithm, *Advances in Neural Information Processing Systems*, **34** 8106–8118, (2021). Enlace.

[5] A. R. Barron, *Monotonic central limit theorem for densities* (Department of Statistics, Stanford University, California, Tech. Rep, 1984).

[6] A. R. Barron, Entropy and the central limit theorem, *Ann. Probab.* **1** 336-342, (1986). Enlace.

[7] T. Berger y R. Zamir, A semi-continuous version of the Berger-Yeung problem, *IEEE Trans. Inf. Theory* **45** 1520-1526 (1999). Enlace.

[8] P. Billingsley, *Convergence of probability measures*, (John Wiley & Sons, 2013).

[9] O. Bousquet, S. Boucheron y G. Lugosi, *Introduction to statistical learning theory*, (Springer, 2003).

[10] G. Chechik, A. Globerson, N. Tishby y Y. Weiss, IB for Gaussian variables, *J. Mach. Learn. Res.* **6** 165-188 (2003). Enlace.

[11] T. M. Cover, *Elements of information theory*, (John Wiley & Sons, 1999).

[12] F. Galindo, J. Sanz y L. A. Tristán, *Guía práctica de cálculo infinitesimal en varias variables*, (Ediciones Paraninfo, SA, 2005).

[13] Z. Goldfeld y Y. Polyanskiy, The IB Problem and its Applications in Machine Learning, *IEEE J. Sel. Areas Inf. Theory* **1** 19-38 (2020). Enlace.

[14] A. Globerson, The minimum information principle in learning and neural data analysis, PhD Thesis, (Hebrew University of Jerusalem, 2005).

[15] T. Hastie, R. Tibshirani y J. H. Friedman, *The elements of statistical learning: data mining, inference*, and prediction, (Springer, 2009).

[16] T. Hastie, A. Montanari, S. Rosset y R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, *Ann. Stat.* **50**, 949 (2022). Enlace.

[17] G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh, *Statistics and Probability: Essays in Honor of C.R. Rao*, (North-Holland Publishing Company, 1982).

[18] M. Belkina, H. Daniel, M. Siyuan y M. Soumik, Reconciling modern machine-learning practice and the classical bias–variance trade-off, *PNAS*, **116**, 15849-15854. Enlace.

[19] V. Ngampruetikorn, J. S. David, IB theory of high-dimensional regression: relevancy, effciency and optimality, (*NeurIPS*, 2022). Enlace.

[20] V. Ngampruetikorn, J. S. David, Supplementary Material: IB theory of high-dimensional regression: relevancy, effciency and optimality, (*NeurIPS*, 2022). Enlace.

[21] K. B. Petersen, M. S. Pedersen y otros, *The matrix cookbook*, (Technical University of Denmark, 2012).

[22] E. Posner, Random coding strategies for minimum entropy, *IEEE Trans. Inf. Theory* **21** 388-391, (1975).

[23] O. Rioul, Information theoretic proofs of entropy power inequalities, *IEEE Trans. Inf. Theory* **57** 33-53, (2010). Enlace.

[24] D. Richards, J. Mourtada y L. Rosasco, Asymptotics of Ridge(less) Regression under General Source Condition, in Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, **130** 3889–3897 (2021). Enlace.

[25] S. Shalev-Shwartz, S. Ben-David, *Understanding machine learning: From theory to algorithms*, (Cambridge university press, 2014).

[26] N. Tishby, F. C. Pereira y W. Bialek, The IB method, *The 37th annual Allerton Conference on Communication, Control, and Computing* 368–377. Enlace.

Apéndice A

Teoremas y conceptos auxiliares

A.1. Continuidad absoluta

Definición A.1.1. Sea (X, Σ, μ) un espacio de probabilidad y ν una medida compleja definida en el mismo espacio medible. Decimos que ν **es absolutamente continua respecto de** μ y escribimos $\nu \ll \mu$ si se cumple

$$\nu(M) = 0 \text{ para todo } M \in \Sigma \text{ tal que } \mu(M) = 0. \tag{A.1}$$

Propiedad A.1.2. si μ y ν son medidas sobre el mismo espacio medible, entonces son absolutamente continuas respecto de $\mu + \nu$, es decir, $\mu \ll \mu + \nu$ y $\nu \ll \mu + \nu$.

Demostración. se extrae de la propia definición ya que si se tiene $M \in \sigma$ tal que $(\mu + \nu)(M) = 0 = \mu(M) + \nu(M)$ y $\mu \geq 0$, $\nu \geq 0$, necesariamente se tiene $\nu(M) = \mu(N) = 0$ □

Teorema A.1.3. Teorema de Radon-Nikodym.[8] Sea (X, Σ, μ) un espacio de medida σ-finito, ν una medida compleja definida en el espacio medible (X, Σ) y absolutamente continua respecto de μ. Existe una función integrable $f \in L^1(\mu, \mathbb{C})$ que cumple:

$$\nu(A) = \int_A d\nu = \int_A f \, d\mu, \quad \text{para todo } A \text{ medible.} \tag{A.2}$$

Se dice que f es la **derivada de Radon-Nikodym** de ν respecto de μ, cualquier otra derivada es igual en casi todo X a esta y se le denota por $\frac{d\nu}{d\mu}$.

En estas condiciones y al trabajar con funciones de probabilidad es útil las siguientes afirmaciones.

(a) Si ν es real, podemos tomar f real.

(b) Si ν es positiva, podemos tomar $f \geq 0$.

A.2. Teorema de derivación de integrales paramétricas

Teorema A.2.1. [3][12] Sean A un abierto de $\mathbb{R}^m$, $E \subset \mathbb{R}^p$ medible, y F una función real definida en $A \times E \subset \mathbb{R}^{m+p}$. Supongamos que:

1. Para cada $x \in A$ la función F_x, definida por $F_x(y) = F(x, y)$, es integrable en E.
2. Para casi todo $y \in E$ la función F_y, definida por $F_y(x) = F(x, y)$, admite derivada parcial continua respecto de x_j en A.
3. Para cada $x \in A$ la función $y \mapsto \frac{\partial F}{\partial x_j}(x, y)$ es medible en E y existe una función g_j, integrable en E, tal que

$$|D_j F_y(x)| = \left| \frac{\partial F}{\partial x_j}(x, y) \right| \leq g_j(y). \tag{A.3}$$

   para todo $x \in A$ y casi todo $y \in E$.

Entonces la función f, definida en A por $f(X) = \int_E F(x, y)dy$, admite derivada parcial con respecto de x_j en , y se tiene que:

$$D_j F_y(x) = \frac{\partial}{\partial x_j} \int_E F(x, y)dy = \int_E \frac{\partial}{\partial x_j} F(x, y)dy. \tag{A.4}$$

A.3. Regla de la cadena para medidas de probabilidad

Propiedad A.3.1. Sean P y Q dos medidas de probabilidad en el mismo espacio medible $(\mathbb{R}, \mathcal{B}, m)$, de forma que siguen funciones de densidad f_P y f_Q respectivamente. En otras palabras:

$$\begin{aligned} f_P &= \frac{dP}{dx}, \\ f_Q &= \frac{dQ}{dx}. \\ &. \end{aligned} \tag{A.5}$$

Cúmplase también $P \ll Q$. Entonces se tiene

$$\frac{dP}{dQ} = \frac{f_P(x)}{f_Q(x)}. \tag{A.6}$$

Demostración.

$$P(E) = \int_E dP = \int_E f(x)dx = \int_E \frac{g(x)}{g(x)} f(x)dx = \int_E \frac{f(x)}{g(x)} g(x)dx = \int_E \frac{f(x)}{g(x)} dQ.$$

Y de esta forma, y aplicando se llega a lo que se quería probar. □

A.4. Teorema del cambio de variable para el cálculo integral

Teorema A.4.1. Sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ una variable aleatoria n-dimensional con función de densidad conjunta $f_{\mathbf{X}}(\mathbf{x})$. Supongamos que $\mathbf{Y} = g(\mathbf{X}) = (g_1(\mathbf{X}), g_2(\mathbf{X}), \ldots, g_m(\mathbf{X}))$ es una transformación de $\mathbf{X}$ tal que g es una función biyectiva y diferenciable con inversa g^{-1} continua y diferenciable, es decir, sea un difeomorfismo. Entonces, la función de densidad conjunta de $\mathbf{Y}$, denotada como $f_{\mathbf{Y}}(\mathbf{y})$, está dada por:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))\left|\det\left(\mathcal{J}g(y)\right)\right|. \tag{A.7}$$

donde $\mathcal{J}g(y)$ es la matriz Jacobiana de la inversa de g evaluada en $\mathbf{y}$. De esta forma.

$$\int f = \int (f \circ g)\left|\det\left(\mathcal{J}g(y)\right)\right|. \tag{A.8}$$

A.5. Convergencia en distribución

Definición A.5.1. Sea X una variable aleatoria. La función característica de X, denotada por $\varphi_X(t)$, es una función definida para todo número real t y se obtiene tomando la esperanza del exponencial complejo de tX. Matemáticamente, la función característica se define como:

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right]. \tag{A.9}$$

Definición A.5.2. Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias y X una variable aleatoria. Se dice que X_n converge en distribución a X o $X_n \xrightarrow{d} X$ si

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{X}_n \in A) = \mathbb{P}(\mathbf{X} \in A). \tag{A.10}$$

para todo $A \in \mathbb{R}^n$ que sea un conjunto continuo, es decir, cuya frontera tenga medida nula.

Teorema A.5.3. Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias y X una variable aleatoria. Denotemos por $\varphi_{X_n}(t)$ y $\varphi_X(t)$ las funciones características de X_n y X, respectivamente. Si

$$\varphi_{X_n}(t) \to \varphi_X(t) \quad \text{para todo } t \in \mathbb{R}, \tag{A.11}$$

entonces X_n converge en distribución a X, es decir,$X_n \xrightarrow{d} X$.

A.6. Semicontinuidad inferior

Definición A.6.1. Una función $f : \mathbb{R}^n \to \mathbb{R}$ es semicontinua inferior en un punto $x_0 \in \mathbb{R}^n$ si, para todo $\epsilon > 0$, existe un $\delta > 0$ tal que para todo x en el entorno $\|x - x_0\| < \delta$, se cumple que:

$$f(x_0) \leq f(x) + \epsilon. \tag{A.12}$$

En otras palabras, f es semicontinua inferior en x_0 si:

$$\liminf_{x\to x_0} f(x) \geq f(x_0). \tag{A.13}$$

Apéndice B

Códigos utilizados

El código utilizado para analizar el comportamiento del **IB** y de la **Regresión de Gibbs** se ha realizado en ***RStudio***.

B.1. Códigos aplicados al Teorema de Marchenko-Pastur

Código asociado a la Figura 5.1

```
    integrand <- function(phi) {
   n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

n <- 1
N <- 1000  # numero de filas
P <- N/n   # numero de columnas

x <- seq(from = (1-1/sqrt(n))^2, to = (1+1/sqrt(n))^2, length.out = 1000)
y <- integrand(x)/n

# Generar la matriz N x P con distribucion normal (0,1)
X <- matrix(rnorm(N * P, mean = 0, sd = 1), nrow = N, ncol = P)

# Calcular XX^T
XXt <- X %*% t(X)/N

# Calcular los autovalores
eigenvalues <- eigen(XXt)$values

# Calcular la funcion de densidad
density_eigenvalues <- density(eigenvalues, bw =0.05)

# Crear la grafica de la funcion de densidad con limites en los ejes
plot(density_eigenvalues, main="n = 1",
     xlab="Autovalor", ylab="Densidad", col="#B22222", lwd=2, ylim=c(0, 1)
```

```
         , xlim = c(-0.2,4))

# Anadir la segunda linea a la grafica
lines(x, y, col="#00008B", lwd=2)

# Anadir la leyenda
legend(y = 0.92,x=1.7, legend=c("Densidad Marchenko-Pastur", "Densidad Empirica"),
       col=c("#B22222", "#00008B"), lwd=2)

abline(v = 0, col = "black", lwd = 2, lty = 2)
```

B.2. Códigos aplicados al IB

Código asociado a la Figura 5.4

```
# Generar matriz A de tamano p x N con distribucion uniforme
# Definir la funcion a integrar
integrand <- function(phi) {
   n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
   integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
   integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
   integrand(phi)*log(1+phi/lambda_c)
}

n_aux <- c(0.01,0.1,0.5,1,5,10,100)

matriz_I_TW <- matrix(nrow = 0, ncol = length(n_aux))
matriz_I_TY_W <- matrix(nrow = 0, ncol = length(n_aux))
matriz_I_YW <- matrix(nrow = 0, ncol = length(n_aux))

b <- seq(from = -100, to = 100, length.out = 10000)
b <- exp(b)

N <- 1000

library(viridis)
colores_inferno <- inferno(length(n_aux)+2)

for (i in 1:length(b)) {
   row_aux1 <- matrix(nrow = 1, ncol = 0)
```

```
35    row_aux2 <- matrix(nrow = 1, ncol = 0)
36    row_aux3 <- matrix(nrow = 1, ncol = 0)
37
38    for (j in 1:length(n_aux)) {
39       # Parametros
40       n <- n_aux[j]
41       p <- N/n
42       lambda_c <- 1/n
43       phi_c <- b[i]
44
45       if (phi_c > (1+1/sqrt(n))^2) {
46          I_TW <- 0.000
47          I_TY_W <- 0.000
48          I_YW <- 0.000
49       } else if (phi_c < (1-1/sqrt(n))^2) {
50          I_TW <- integrate(information_TW, lower = (1-1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
51          I_TY_W <- (integrate(information_TY_W, lower = (1-1/sqrt(n))^2,
               upper = (1+1/sqrt(n))^2)$value)*p/2-I_TW
52          I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
53       } else {
54          I_TW <- integrate(information_TW, lower = phi_c, upper = (1+1/
               sqrt(n))^2)$value*p/2
55          I_TY_W <- (integrate(information_TY_W, lower = phi_c, upper =
               (1+1/sqrt(n))^2)$value)*p/2-I_TW
56          I_YW <- integrate(information_YW, lower = phi_c, upper = (1+1/
               sqrt(n))^2)$value*p/2
57       }
58
59       row_aux1 <- cbind(row_aux1, I_TW)
60       row_aux2 <- cbind(row_aux2, I_TY_W)
61       row_aux3 <- cbind(row_aux3, I_YW)
62    }
63
64    matriz_I_TW <- rbind(matriz_I_TW, row_aux1)
65    matriz_I_TY_W <- rbind(matriz_I_TY_W, row_aux2)
66    matriz_I_YW <- rbind(matriz_I_YW, row_aux3)
67 }
68
69 par(las = 1)
70 plot(matriz_I_TY_W[,1]/N, matriz_I_TW[,1]/matriz_I_YW[,1], type = "l", lwd
       = 3,
71       main = "", col = colores_inferno[2], xlab = "",
72       ylab = "", pch = 11, xlim = c(0, 1.5), ylim = c(0, 1), panel.first =
            {
73          # Configurar la cuadrilla con mas lineas
74          grid(col = "gray", lty = 3, lwd = 0.5)
75
76          # Agregar lineas en el fondo del grafico
77          abline(h = seq(0, 1, by = 0.2), col = "gray", lty = 2)
78          abline(v = seq(0, 1.5, by = 0.25), col = "gray", lty = 2)
79       })
```

```
# Ajustar titulos de los ejes
title(xlab = expression(bold("I(T;Y|W)/N")), ylab = expression(bold("I(T;W)␣/␣I(Y;W)")))

for (i in 2:length(n_aux)) {
   lines(matriz_I_TY_W[,i]/N, matriz_I_TW[,i]/matriz_I_YW[,i], col = colores_inferno[i+1], lwd = 3)
}
```

Código asociado a la Figura 5.2

```
integrand <- function(phi) {
   n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
   integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
   integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
   integrand(phi)*log(1+phi/lambda_c)
}

n_aux <- c(0.01,0.1,0.5,1,2,5,10,100)

matriz_I_TW <- matrix(nrow = 0, ncol = length(n_aux))
matriz_I_TY_W <- matrix(nrow = 0, ncol = length(n_aux))
matriz_I_YW <- matrix(nrow = 0, ncol = length(n_aux))

N <- 100

b <- seq(from = -10, to = 10 , length.out = 10000)
b <- exp(b)

for (i in 1:length(b)){

   row_aux1 <- matrix(nrow = 1, ncol = 0)
   row_aux2 <- matrix(nrow = 1, ncol = 0)
   row_aux3 <- matrix(nrow = 1, ncol = 0)

   for (j in 1:length(n_aux)){
      # Parametros
      n <- n_aux[j]
      p <- N/n
      lambda_c <-1/n
      phi_c <- b[i]

      if (phi_c>(1+1/sqrt(n))^2){
```

```
42          I_TW <- 0.000
43          I_TY_W <- 0.000
44          I_YW <- integrate(information_YW, lower = (1+1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
45       } else if (phi_c<(1-1/sqrt(n))^2){
46          I_TW <- integrate(information_TW, lower = (1-1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
47          I_TY_W <- (integrate(information_TY_W, lower = (1-1/sqrt(n))^2,
               upper = (1+1/sqrt(n))^2)$value)*p/2-I_TW
48          I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
49       } else{
50          I_TW <- integrate(information_TW, lower = phi_c,upper = (1+1/sqrt
               (n))^2)$value*p/2
51          I_TY_W <- (integrate(information_TY_W, lower = phi_c,upper = (1+1
               /sqrt(n))^2)$value)*p/2-I_TW
52          I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
53
54       }
55
56       row_aux1 <- cbind(row_aux1, I_TW)
57       row_aux2 <- cbind(row_aux2, I_TY_W)
58       row_aux3 <- cbind(row_aux3, I_YW)
59
60    }
61
62    matriz_I_TW <- rbind(matriz_I_TW, row_aux1)
63    matriz_I_TY_W <- rbind(matriz_I_TY_W, row_aux2)
64    matriz_I_YW <- rbind(matriz_I_YW, row_aux3)
65 }
66
67 library(viridis)
68 colores_inferno <- inferno(length(n_aux)+3)
69
70 par(las = 1)
71 plot(b*n_aux[1],matriz_I_TW[,1]/N,log = "x",
72      type = "l",lwd = 3,
73      main = "",col = colores_inferno[1],xlab = "",
74      ylab = "", pch = 11,xlim = c(0.01, 10), ylim = c(0, 0.35))
75
76 # Ajustar titulos de los ejes
77 title(ylab = expression(bold("I(T;W)/N")), xlab = expression(bold(psi/
    lambda)))
78
79 for (i in 2:length(n_aux)) {
80    lines(b*n_aux[i],matriz_I_TW[,i]/N, col = colores_inferno[i+1], lwd =
       3)
81 }
```

Código asociado a la Figura 5.3

```
integrand <- function(phi) {
   n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
   integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
   integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
   integrand(phi)*log(1+phi/lambda_c)
}

n_aux <- c(0.01,0.1,0.5,1,2,5,10,100)

matriz_I_TW <- matrix(nrow = 0, ncol = length(n_aux))
matriz_I_TY_W <- matrix(nrow = 0, ncol = length(n_aux))
matriz_I_YW <- matrix(nrow = 0, ncol = length(n_aux))

N <- 100

b <- seq(from = -10, to = 10 , length.out = 1000)
b <- exp(b)

for (i in 1:length(b)){

   row_aux1 <- matrix(nrow = 1, ncol = 0)
   row_aux2 <- matrix(nrow = 1, ncol = 0)
   row_aux3 <- matrix(nrow = 1, ncol = 0)

   for (j in 1:length(n_aux)){
      # Parametros
      n <- n_aux[j]
      p <- N/n
      lambda_c <-1/n
      phi_c <- b[i]

      if (phi_c>(1+1/sqrt(n))^2){
         I_TW <- 0.000
         I_TY_W <- 0.000
         I_YW <- 0.000
      } else if (phi_c<(1-1/sqrt(n))^2){
         I_TW <- integrate(information_TW, lower = (1-1/sqrt(n))^2, upper
            = (1+1/sqrt(n))^2)$value*p/2
         I_TY_W <- (integrate(information_TY_W, lower = (1-1/sqrt(n))^2,
            upper = (1+1/sqrt(n))^2)$value)*p/2-I_TW
         I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper
            = (1+1/sqrt(n))^2)$value*p/2
      } else{
```

```
50          I_TW <- integrate(information_TW, lower = phi_c,upper = (1+1/sqrt
               (n))^2)$value*p/2
51          I_TY_W <- (integrate(information_TY_W, lower = phi_c,upper = (1+1
               /sqrt(n))^2)$value)*p/2-I_TW
52          I_YW <- integrate(information_YW, lower = (1+1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
53
54        }
55
56        row_aux1 <- cbind(row_aux1, I_TW)
57        row_aux2 <- cbind(row_aux2, I_TY_W)
58        row_aux3 <- cbind(row_aux3, I_YW)
59
60     }
61
62     matriz_I_TW <- rbind(matriz_I_TW, row_aux1)
63     matriz_I_TY_W <- rbind(matriz_I_TY_W, row_aux2)
64     matriz_I_YW <- rbind(matriz_I_YW, row_aux3)
65  }
66
67  library(viridis)
68  colores_inferno <- inferno(length(n_aux)+3)
69
70  par(las = 1)
71  plot(b*n_aux[1],matriz_I_TY_W[,1]/N,log = "x",
72       type = "l",lwd = 3,
73       main = "",col = colores_inferno[1],xlab = "",
74       ylab = "", pch = 11,xlim = c(0.01, 10), ylim = c(0, 2.2))
75
76  # Ajustar titulos de los ejes
77  title(ylab = expression(bold("I(T;Y|W)/N")), xlab = expression(bold(psi/
       lambda)))
78
79  for (i in 2:length(n_aux)) {
80     lines(b*n_aux[i],matriz_I_TY_W[,i]/N, col = colores_inferno[i+1], lwd =
          3)
81  }
```

B.3. Códigos aplicados a Regresión de Gibbs

Código asociado a la Figura 5.7

```
integrand <- function(phi) {
   n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
   integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
   integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
   integrand(phi)*log(1+phi/lambda_c)
}

informationGibbs_TW <- function(phi){
   integrand(phi)*log(1+((phi**2/lambda_c)/(phi+(N*(phi+lambda)/2/alpha/
      sigma**2))))
}

informationGibbs_TY_W <- function(phi){
   integrand(phi)*log(1+2*alpha*sigma**2/N*(phi/(phi+lambda)))
}

n <- 1
N <-1
p <- N/n
lambda_c <-100/n
phi_c <- 10
lambda_aux <- c(0.01,0.05,0.1,1,10,100)

matriz_IGibbs_TW <- matrix(nrow = 0, ncol = length(lambda_aux))
matriz_IGibbs_TY_W <- matrix(nrow = 0, ncol = length(lambda_aux))
matriz_I_YW <- matrix(nrow = 0, ncol = length(lambda_aux))

sigma <- 1
alpha_aux <- seq(from = 0.01, to = 100, length.out = 1000)

for (i in 1:length(alpha_aux)) {

   alpha <- alpha_aux[i]

   row_aux1 <- matrix(nrow = 1, ncol = 0)
   row_aux2 <- matrix(nrow = 1, ncol = 0)
   row_aux3 <- matrix(nrow = 1, ncol = 0)

```

```
47    for (j in 1:length(lambda_aux)){
48
49       lambda <- lambda_aux[j]
50
51       IGibbs_TW <- integrate(informationGibbs_TW, lower = (1-1/sqrt(n))^2,
             upper = (1+1/sqrt(n))^2)$value*p/2
52       IGibbs_TY_W <- (integrate(informationGibbs_TY_W, lower = (1-1/sqrt(n
            ))^2, upper = (1+1/sqrt(n))^2)$value)*p/2
53       I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper =
            (1+1/sqrt(n))^2)$value*p/2
54
55       row_aux1 <- cbind(row_aux1, IGibbs_TW)
56       row_aux2 <- cbind(row_aux2, IGibbs_TY_W)
57       row_aux3 <- cbind(row_aux3, I_YW)
58    }
59
60    matriz_IGibbs_TW <- rbind(matriz_IGibbs_TW, row_aux1)
61    matriz_IGibbs_TY_W <- rbind(matriz_IGibbs_TY_W, row_aux2)
62    matriz_I_YW <- rbind(matriz_I_YW, row_aux3)
63 }
64
65 library(viridis)
66 colores_inferno <- inferno(length(lambda_aux))
67
68 par(las = 1)
69 plot(matriz_IGibbs_TY_W[,1]/N,matriz_IGibbs_TW[,1]/matriz_I_YW[,1], type =
       "l",lwd = 2,
70      main = "",col = colores_inferno[1],xlab = "",
71      ylab = "", pch = 11,xlim = c(0, 1.5), ylim = c(0, 1),panel.first = {
72         # Configurar la cuadrilla con mas lineas
73         grid(col = "gray", lty = 3, lwd = 0.5)
74
75         # Agregar lineas en el fondo del grafico
76         abline(h = seq(0, 1, by = 0.2), col = "gray", lty = 2)
77         abline(v = seq(0, 1.5, by = 0.25), col = "gray", lty = 2)
78      })
79 # Ajustar titulos de los ejes
80 title(xlab = expression(bold("I(T;Y|W)/N")), ylab = expression(bold("I(T;W
       )␣/␣I(Y;W)")))
81
82 for (i in 2:length(lambda_aux)) {
83    lines(matriz_IGibbs_TY_W[,i]/N,matriz_IGibbs_TW[,i]/matriz_I_YW[,i],
         col = colores_inferno[i+1], lwd = 2)
84 }
85
86 matriz_I_TW <- matrix(nrow = 0, ncol = 1)
87 matriz_I_TY_W <- matrix(nrow = 0, ncol = 1)
88 matriz_I_YW <- matrix(nrow = 0, ncol = 1)
89
90 b <- seq(from = -9, to = 10 , length.out = 500)
91 b <- exp(b)
92
93 for (i in 1:length(b)){
```

```

   phi_c <- b[i]

   if (phi_c>(1+1/sqrt(n))^2){
      I_TW <- 0.000
      I_TY_W <- 0.000
      I_YW <- integrate(information_YW, lower =  (1-1/sqrt(n))^2, upper =
         (1+1/sqrt(n))^2)$value*p/2
   } else if (phi_c<(1-1/sqrt(n))^2){
      I_TW <- integrate(information_TW, lower = (1-1/sqrt(n))^2, upper =
         (1+1/sqrt(n))^2)$value*p/2
      I_TY_W <- (integrate(information_TY_W, lower = (1-1/sqrt(n))^2,
         upper = (1+1/sqrt(n))^2)$value)*p/2-I_TW
      I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper =
         (1+1/sqrt(n))^2)$value*p/2
   } else{
      I_TW <- integrate(information_TW, lower = phi_c,upper = (1+1/sqrt(n)
         )^2)$value*p/2
      I_TY_W <- (integrate(information_TY_W, lower = phi_c,upper = (1+1/
         sqrt(n))^2)$value)*p/2-I_TW
      I_YW <- integrate(information_YW, lower =  (1-1/sqrt(n))^2, upper =
         (1+1/sqrt(n))^2)$value*p/2
   }

   matriz_I_TW <- rbind(matriz_I_TW, I_TW)
   matriz_I_TY_W <- rbind(matriz_I_TY_W, I_TY_W)
   matriz_I_YW <- rbind(matriz_I_YW, I_YW)
}

points(matriz_I_TY_W/N,matriz_I_TW/matriz_I_YW, col = "blue", pch = 19,
   cex = 0.01)

# Agregar una leyenda al grafico
legend(x=0.8,y=0.22, legend = "frontera del IB",
       col = "blue", pch = 19,pt.cex = 0.5,cex =.75)
```

Código asociado a la Figura 5.5

```
integrand <- function(phi) {
   n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
   integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
   integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
   integrand(phi)*log(1+phi/lambda_c)
```

```
}

informationGibbs_TW <- function(phi){
   integrand(phi)*log(1+((phi**2/lambda_c)/(phi+(N*(phi+lambda)/2/alpha/
      sigma**2))))
}

informationGibbs_TY_W <- function(phi){
   integrand(phi)*log(1+2*alpha*sigma**2/N*(phi/(phi+lambda)))
}

n <- 1
N <- 1
p <- N/n
lambda_c <-1/n
phi_c <- 10
lambda_aux <- c(0.01,0.05,0.1,1,10,100)

matriz_IGibbs_TW <- matrix(nrow = 0, ncol = length(lambda_aux))
matriz_IGibbs_TY_W <- matrix(nrow = 0, ncol = length(lambda_aux))
matriz_I_YW <- matrix(nrow = 0, ncol = length(lambda_aux))

sigma <- 1
alpha_aux <- seq(from = 0.001, to = 100, length.out = 10000)

for (i in 1:length(alpha_aux)) {

   alpha <- alpha_aux[i]

   row_aux1 <- matrix(nrow = 1, ncol = 0)
   row_aux2 <- matrix(nrow = 1, ncol = 0)
   row_aux3 <- matrix(nrow = 1, ncol = 0)

   for (j in 1:length(lambda_aux)){

      lambda <- lambda_aux[j]

      IGibbs_TW <- integrate(informationGibbs_TW, lower = (1-1/sqrt(n))^2,
          upper = (1+1/sqrt(n))^2)$value*p/2
      IGibbs_TY_W <- (integrate(informationGibbs_TY_W, lower = (1-1/sqrt(n
         ))^2, upper = (1+1/sqrt(n))^2)$value)*p/2
      I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper =
         (1+1/sqrt(n))^2)$value*p/2

      row_aux1 <- cbind(row_aux1, IGibbs_TW)
      row_aux2 <- cbind(row_aux2, IGibbs_TY_W)
      row_aux3 <- cbind(row_aux3, I_YW)
   }

   matriz_IGibbs_TW <- rbind(matriz_IGibbs_TW, row_aux1)
   matriz_IGibbs_TY_W <- rbind(matriz_IGibbs_TY_W, row_aux2)
   matriz_I_YW <- rbind(matriz_I_YW, row_aux3)
}
```

```

library(viridis)
colores_inferno <- inferno(length(lambda_aux)+2)

par(las = 1)
plot(N/2/alpha_aux/sigma**2,matriz_IGibbs_TY_W[,1]/N, type = "l",lwd = 2,
    log ="x",
      main = "",col = colores_inferno[1],xlab = "",
      ylab = "", pch = 11,xlim = c(0.007, 100), ylim = c(0, 3),panel.first
         = {
         # Configurar la cuadrilla con mas lineas
         grid(col = "gray", lty = 3, lwd = 0.5)
      })
# Ajustar titulos de los ejes
title(xlab = expression(bold(kappa)), ylab = expression(bold("I(T;Y|W)/N")
    ))

for (i in 2:length(lambda_aux)) {
    lines(N/2/alpha_aux/sigma**2,matriz_IGibbs_TY_W[,i]/N, col = colores_
       inferno[i+1], lwd = 2)
}
```

Código asociado a la Figura 5.6

```
integrand <- function(phi) {
    n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
    integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
    integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
    integrand(phi)*log(1+phi/lambda_c)
}

informationGibbs_TW <- function(phi){
    integrand(phi)*log(1+((phi**2/lambda_c)/(phi+(N*(phi+lambda)/2/alpha/
       sigma**2))))
}

informationGibbs_TY_W <- function(phi){
    integrand(phi)*log(1+2*alpha*sigma**2/N*(phi/(phi+lambda)))
}

n <- 1
N <- 1
p <- N/n
```

```
lambda_c <-1/n
phi_c <- 10
lambda_aux <- c(0.01,0.05,0.1,1,10,100)

matriz_IGibbs_TW <- matrix(nrow = 0, ncol = length(lambda_aux))
matriz_IGibbs_TY_W <- matrix(nrow = 0, ncol = length(lambda_aux))
matriz_I_YW <- matrix(nrow = 0, ncol = length(lambda_aux))

sigma <- 1
alpha_aux <- seq(from = 0.001, to = 100, length.out = 10000)

for (i in 1:length(alpha_aux)) {

   alpha <- alpha_aux[i]

   row_aux1 <- matrix(nrow = 1, ncol = 0)
   row_aux2 <- matrix(nrow = 1, ncol = 0)
   row_aux3 <- matrix(nrow = 1, ncol = 0)

   for (j in 1:length(lambda_aux)){

      lambda <- lambda_aux[j]

      IGibbs_TW <- integrate(informationGibbs_TW, lower = (1-1/sqrt(n))^2,
          upper = (1+1/sqrt(n))^2)$value*p/2
      IGibbs_TY_W <- (integrate(informationGibbs_TY_W, lower = (1-1/sqrt(n
         ))^2, upper = (1+1/sqrt(n))^2)$value)*p/2
      I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper =
         (1+1/sqrt(n))^2)$value*p/2

      row_aux1 <- cbind(row_aux1, IGibbs_TW)
      row_aux2 <- cbind(row_aux2, IGibbs_TY_W)
      row_aux3 <- cbind(row_aux3, I_YW)
   }

   matriz_IGibbs_TW <- rbind(matriz_IGibbs_TW, row_aux1)
   matriz_IGibbs_TY_W <- rbind(matriz_IGibbs_TY_W, row_aux2)
   matriz_I_YW <- rbind(matriz_I_YW, row_aux3)
}

library(viridis)
colores_inferno <- inferno(length(lambda_aux)+2)

par(las = 1)
plot(N/2/alpha_aux/sigma**2,matriz_IGibbs_TY_W[,1]/N, type = "l",lwd = 2,
   log ="x",
     main = "",col = colores_inferno[1],xlab = "",
     ylab = "", pch = 11,xlim = c(0.007, 100), ylim = c(0, 3),panel.first
        = {
        # Configurar la cuadrilla con mas lineas
        grid(col = "gray", lty = 3, lwd = 0.5)
     })
# Ajustar titulos de los ejes
```

```
title(xlab = expression(bold(kappa)), ylab = expression(bold("I(T;Y|W)/N")
    ))

for (i in 2:length(lambda_aux)) {
    lines(N/2/alpha_aux/sigma**2,matriz_IGibbs_TY_W[,i]/N, col = colores_
        inferno[i+1], lwd = 2)
}
```

B.4. Códigos aplicados a la comparación de modelos

Código asociado a la Figura 5.8

```
integrand <- function(phi) {
    n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
    integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
    integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
    integrand(phi)*log(1+phi/lambda_c)
}

informationGibbs_TW <- function(phi){
    integrand(phi)*log(1+((phi**2/lambda_c)/(phi+(N*(phi+lambda)/2/alpha/
        sigma**2))))
}

informationGibbs_TY_W <- function(phi){
    integrand(phi)*log(1+2*alpha*sigma**2/N*(phi/(phi+lambda)))
}

n_aux <- seq(from = -2.5, to = 2.5, length.out = 500)
n_aux <- 10**n_aux
mu_aux <- c(0.2,0.3,0.5,0.7,0.9)

matriz_nu <- matrix(nrow = length(n_aux) , ncol = length(mu_aux))

for (k in 1:length(n_aux)){
    n <- n_aux[k]
    N <-1000
    p <- N/n
    lambda_c <-100/n
    phi_c <- 10
```

```
37    lambda <- 10**(-6)
38    sigma <- 1
39
40    matriz_IGibbs_TW <- matrix(nrow = 0, ncol = 1)
41    matriz_IGibbs_TY_W <- matrix(nrow = 0, ncol = 1)
42    matriz_IGibbs_YW <- matrix(nrow = 0, ncol = 1)
43    alpha_aux <- seq(from = 0.01, to = 10000, length.out = 200)
44
45    for (i in 1:length(alpha_aux)) {
46
47       alpha <- alpha_aux[i]
48
49       IGibbs_TW <- integrate(informationGibbs_TW, lower = (1-1/sqrt(n))^2,
             upper = (1+1/sqrt(n))^2)$value*p/2
50       IGibbs_TY_W <- (integrate(informationGibbs_TY_W, lower = (1-1/sqrt(n
            ))^2, upper = (1+1/sqrt(n))^2)$value)*p/2
51       IGibbs_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2,
            upper = (1+1/sqrt(n))^2)$value*p/2
52
53       matriz_IGibbs_TW <- rbind(matriz_IGibbs_TW, IGibbs_TW)
54       matriz_IGibbs_TY_W <- rbind(matriz_IGibbs_TY_W, IGibbs_TY_W)
55       matriz_IGibbs_YW <- rbind(matriz_IGibbs_YW, IGibbs_YW)
56    }
57
58    muGIBBS <- matriz_IGibbs_TW/matriz_IGibbs_YW
59    # Remover duplicados en y y sus correspondientes x
60    unique_indices <- !duplicated(muGIBBS)
61    matriz_IGibbs_TY_W <- matriz_IGibbs_TY_W[unique_indices]
62    muGIBBS <- muGIBBS[unique_indices]
63
64    matriz_I_TW <- matrix(nrow = 0, ncol = 1)
65    matriz_I_TY_W <- matrix(nrow = 0, ncol = 1)
66    matriz_I_YW <- matrix(nrow = 0, ncol = 1)
67    b <- seq(from = -9, to = 10 , length.out = 200)
68    b <- exp(b)
69
70    for (i in 1:length(b)){
71
72       phi_c <- b[i]
73
74       if (phi_c>(1+1/sqrt(n))^2){
75          I_TW <- 0.000
76          I_TY_W <- 0.000
77          I_YW <- integrate(information_YW, lower =  (1-1/sqrt(n))^2, upper
               = (1+1/sqrt(n))^2)$value*p/2
78       } else if (phi_c<(1-1/sqrt(n))^2){
79          I_TW <- integrate(information_TW, lower = (1-1/sqrt(n))^2, upper
              = (1+1/sqrt(n))^2)$value*p/2
80          I_TY_W <- (integrate(information_TY_W, lower = (1-1/sqrt(n))^2,
              upper = (1+1/sqrt(n))^2)$value)*p/2-I_TW
81          I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper
              = (1+1/sqrt(n))^2)$value*p/2
82       } else{
```

```
83          I_TW <- integrate(information_TW, lower = phi_c,upper = (1+1/sqrt
               (n))^2)$value*p/2
84          I_TY_W <- (integrate(information_TY_W, lower = phi_c,upper = (1+1
               /sqrt(n))^2)$value)*p/2-I_TW
85          I_YW <- integrate(information_YW, lower =  (1-1/sqrt(n))^2, upper
                = (1+1/sqrt(n))^2)$value*p/2
86       }
87
88       matriz_I_TW <- rbind(matriz_I_TW, I_TW)
89       matriz_I_TY_W <- rbind(matriz_I_TY_W, I_TY_W)
90       matriz_I_YW <- rbind(matriz_I_YW, I_YW)
91    }
92
93    muIB <- matriz_I_TW/matriz_I_YW
94    # Remover duplicados en y y sus correspondientes x
95    unique_indices <- !duplicated(muIB)
96    matriz_I_TY_W <- matriz_I_TY_W[unique_indices]
97    muIB <- muIB[unique_indices]
98
99    for (j in 1:length(mu_aux)){
100      mu <- mu_aux[j]
101      nu <- approx(muIB,matriz_I_TY_W, xout = mu)$y/approx(muGIBBS,matriz_
            IGibbs_TY_W, xout = mu)$y
102
103      matriz_nu[k,j] <- nu
104   }
105   print(n)
106}
107
108library(viridis)
109colores_inferno <- viridis(length(lambda_aux))
110
111par(las = 1)
112plot(n_aux,matriz_nu[,1], type = "l",lwd = 2, log ="x",
113     main = "",col = colores_inferno[1],xlab = "",
114     ylab = "", pch = 11,xlim = c(0.001, 1000), ylim = c(0.2, 1),panel.
           first = {
115        # Configurar la cuadrilla con mas lineas
116        grid(col = "gray", lty = 3, lwd = 0.5) })
117# Ajustar titulos de los ejes
118title(xlab = expression(bold(eta)), ylab = expression(bold("N/P")))
119
120for (i in 2:length(mu_aux)) {
121   lines(n_aux,matriz_nu[,i], col = colores_inferno[i+1], lwd = 2)
122}
123
124# Agregar la leyenda
125legend_labels <- paste("mu", mu_aux, sep = "␣=␣")
126legend(x=40,y=0.42, legend = legend_labels, col = colores_inferno[2:(
   length(mu_aux) + 1)], lty = 1, lwd = 2, cex = 0.8, pt.cex = 0.8)
```

Código asociado a la Figura 5.9

```
integrand <- function(phi) {
   n*sqrt(((1+1/sqrt(n))^2-phi)*(phi - (1-1/sqrt(n))^2))/2/phi/pi
}

information_TW <- function(phi){
   integrand(phi)*log(1+(phi-phi_c)/(phi_c+lambda_c))
}

information_TY_W <- function(phi){
   integrand(phi)*log(phi/phi_c)
}

information_YW <- function(phi){
   integrand(phi)*log(1+phi/lambda_c)
}

informationGibbs_TW <- function(phi){
   integrand(phi)*log(1+((phi**2/lambda_c)/(phi+(N*(phi+lambda)/2/alpha/
      sigma**2))))
}

informationGibbs_TY_W <- function(phi){
   integrand(phi)*log(1+2*alpha*sigma**2/N*(phi/(phi+lambda)))
}

n_aux <- seq(from = -3, to = 3, length.out = 100)
n_aux <- 10**n_aux
mu_aux <- c(0.05,0.3,0.5,0.7,0.9)

matriz_I1 <- matrix(nrow = length(n_aux) , ncol = length(mu_aux))
matriz_I2 <- matrix(nrow = length(n_aux) , ncol = length(mu_aux))

for (k in 1:length(n_aux)){
   n <- n_aux[k]
   N <-1000
   p <- N/n
   lambda_c <-100/n
   phi_c <- 10
   lambda <- 10**(-6)
   sigma <- 1

   matriz_IGibbs_TW <- matrix(nrow = 0, ncol = 1)
   matriz_IGibbs_TY_W <- matrix(nrow = 0, ncol = 1)
   matriz_IGibbs_YW <- matrix(nrow = 0, ncol = 1)
   alpha_aux <- seq(from = 0.01, to = 10000, length.out = 200)

   for (i in 1:length(alpha_aux)) {

      alpha <- alpha_aux[i]

      IGibbs_TW <- integrate(informationGibbs_TW, lower = (1-1/sqrt(n))^2,
          upper = (1+1/sqrt(n))^2)$value*p/2
```

```
51      IGibbs_TY_W <- (integrate(informationGibbs_TY_W, lower = (1-1/sqrt(n
            ))^2, upper = (1+1/sqrt(n))^2)$value)*p/2
52      IGibbs_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2,
            upper = (1+1/sqrt(n))^2)$value*p/2
53
54      matriz_IGibbs_TW <- rbind(matriz_IGibbs_TW, IGibbs_TW)
55      matriz_IGibbs_TY_W <- rbind(matriz_IGibbs_TY_W, IGibbs_TY_W)
56      matriz_IGibbs_YW <- rbind(matriz_IGibbs_YW, IGibbs_YW)
57  }
58
59  muGIBBS <- matriz_IGibbs_TW/matriz_IGibbs_YW
60  # Remover duplicados en y y sus correspondientes x
61  unique_indices <- !duplicated(muGIBBS)
62  matriz_IGibbs_TY_W <- matriz_IGibbs_TY_W[unique_indices]
63  muGIBBS <- muGIBBS[unique_indices]
64
65  matriz_I_TW <- matrix(nrow = 0, ncol = 1)
66  matriz_I_TY_W <- matrix(nrow = 0, ncol = 1)
67  matriz_I_YW <- matrix(nrow = 0, ncol = 1)
68  b <- seq(from = -9, to = 10 , length.out = 200)
69  b <- exp(b)
70
71  for (i in 1:length(b)){
72
73      phi_c <- b[i]
74
75      if (phi_c>(1+1/sqrt(n))^2){
76          I_TW <- 0.000
77          I_TY_W <- 0.000
78          I_YW <- integrate(information_YW, lower =  (1-1/sqrt(n))^2, upper
                = (1+1/sqrt(n))^2)$value*p/2
79      } else if (phi_c<(1-1/sqrt(n))^2){
80          I_TW <- integrate(information_TW, lower = (1-1/sqrt(n))^2, upper
                = (1+1/sqrt(n))^2)$value*p/2
81          I_TY_W <- (integrate(information_TY_W, lower = (1-1/sqrt(n))^2,
                upper = (1+1/sqrt(n))^2)$value)*p/2-I_TW
82          I_YW <- integrate(information_YW, lower = (1-1/sqrt(n))^2, upper
                = (1+1/sqrt(n))^2)$value*p/2
83      } else{
84          I_TW <- integrate(information_TW, lower = phi_c,upper = (1+1/sqrt
                (n))^2)$value*p/2
85          I_TY_W <- (integrate(information_TY_W, lower = phi_c,upper = (1+1
                /sqrt(n))^2)$value)*p/2-I_TW
86          I_YW <- integrate(information_YW, lower =  (1-1/sqrt(n))^2, upper
                = (1+1/sqrt(n))^2)$value*p/2
87      }
88
89      matriz_I_TW <- rbind(matriz_I_TW, I_TW)
90      matriz_I_TY_W <- rbind(matriz_I_TY_W, I_TY_W)
91      matriz_I_YW <- rbind(matriz_I_YW, I_YW)
92  }
93
94  muIB <- matriz_I_TW/matriz_I_YW
```

```
95      # Remover duplicados en y y sus correspondientes x
96      unique_indices <- !duplicated(muIB)
97      matriz_I_TY_W <- matriz_I_TY_W[unique_indices]
98      muIB <- muIB[unique_indices]
99
100     for (j in 1:length(mu_aux)){
101        mu <- mu_aux[j]
102        I1 <- approx(muIB,matriz_I_TY_W, xout = mu)$y
103        I2 <- approx(muGIBBS,matriz_IGibbs_TY_W, xout = mu)$y
104
105        matriz_I1[k,j] <- I1/p
106        matriz_I2[k,j] <- I2/p
107     }
108     print(n)
109 }
110
111 library(viridis)
112 colores_inferno <- viridis(length(mu_aux))
113
114 par(las = 1)
115 plot(n_aux,matriz_I2[,1], type = "l",lwd = 3, log ="xy",
116      main = "",col = colores_inferno[2],xlab = "",
117      ylab = "", pch = 11,xlim = c(0.001, 250), ylim = c(0.00001, 5),panel.
            first = {
118         # Configurar la cuadrilla con mas lineas
119         grid(col = "gray", lty = 3, lwd = 0.5)
120      })
121 # Ajustar titulos de los ejes
122 title(ylab = expression(bold("I(T;Y|W)/P")), xlab = expression(bold("N/P")
    ))
123
124 for (i in 2:length(mu_aux)) {
125    lines(n_aux,matriz_I2[,i], col = colores_inferno[i+1], lwd = 3)
126 }
127
128 # Filtrar los datos para mantener solo los indices pares
129 indices_pares <- seq(2, length(n_aux), by = 2)
130
131 for (i in 1:length(mu_aux)) {
132    # Filtrar los datos para mantener solo los indices pares
133    vec <- matriz_I1[,i]
134    vec <- vec[indices_pares]
135    points(n_aux[indices_pares],vec, col = colores_inferno[i+1], pch = 19,
          cex = 0.5)
136 }
137 legend_labels <- paste("mu", mu_aux, sep = "␣=␣")
138 legend(x=4,y=0.001, legend = legend_labels, col = colores_inferno[2:(
    length(mu_aux) + 1)], lty = 1, lwd = 2, cex = 0.8, pt.cex = 0.8)
139
140 legend(x=3.6,y=0.006, legend = "fronteras␣IB", col = colores_inferno[2:(
    length(mu_aux) + 1)], lty = 3, lwd = 2, cex = 0.8, pt.cex = 0.8)
```