Universidad Nacional Autónoma de México

Facultad de Ciencias

Regresión con procesos gaussianos en la reconstrucción de funciones cosmológicas

T E S I S

que para obtener el título de:

Licenciado en Física

presenta:

José de Jesús Velázquez Ugalde

Tutor:

Dr. José Alberto Vázquez González

Ciudad Universitaria, Cd. Mx.
2024

CONSTANCIA DE PRESENTACIÓN DE EXAMEN PROFESIONAL

En la Universidad Nacional Autónoma de México, en el AULA SOTERO PRIETO I de la FACULTAD DE CIENCIAS siendo las 13:00 horas del día 15 de noviembre de 2024, el alumno:

JOSE DE JESUS VELAZQUEZ UGALDE

de nacionalidad MEXICANA con número de cuenta 313334175 se presentó con el fin de sustentar el examen oral para obtener el título de:

FÍSICO

en su modalidad de titulación por TESIS con el trabajo titulado: Regresión con procesos gaussianos en la reconstrucción de funciones cosmológicas. El alumno cursó sus estudios en el periodo 2016-1 a 2024-1 habiendo obtenido un promedio de 8.86 y cumpliendo con los requisitos académicos señalados en el plan de estudios 1081 aprobado por el H. Consejo Universitario.

El Jurado designado por el Comité Académico integrado por:

PRESIDENTE: DR. DARIO NUÑEZ ZUÑIGA
SECRETARIO: DR. JOSE ALBERTO VAZQUEZ GONZALEZ
VOCAL: DR. JOSE ANTONIO DE DIEGO ONSURBE
SUPLENTE: DR. AXEL RICARDO DE LA MACORRA PETTERSSON MORIEL
SUPLENTE: DRA. MARIANA VARGAS MAGAÑA

Tras el interrogatorio y deliberación resolvió otorgarle la calificación de:

Aprobado

Procediendo a informarle el resultado, tomarle la Protesta Universitaria y dar por concluido el acto.

PRESIDENTE DEL JURADO | **SECRETARIO DEL JURADO** | **VOCAL DEL JURADO**

El Titular de la entidad académica hace constar que las firmas electrónicas que anteceden son válidas y corresponden a los miembros del jurado.

"POR MI RAZA HABLARÁ EL ESPÍRITU"
DR. VÍCTOR MANUEL VELÁZQUEZ AGUILAR
DIRECTOR DE LA ENTIDAD ACADÉMICA

QR de verificación

Cadena de verificación digital
c592e9ee2b9e78fa5cfe77d9f1b64bcef128a5c920c6d96ea02568cd8dc1b0a2540e1777eb76d8a5bb2fc070ae251feffffee19e8bfbc3b01a56c013498edbab

A mi madre y a todas las personas que me apoyaron durante mi trayectoria académica.

Agradecimientos

Agradezco a la Universidad Nacional Autónoma de México y a la Facultad de Ciencias por brindarme la oportunidad de acceder al conocimiento necesario para realizar este trabajo a lo largo de la licenciatura, así como a todos los profesores que participaron en mi formación.

A mi tutor, Dr. José Alberto Vázquez, por su dedicación y compromiso al asesorarme durante los últimos años, así como a mis sinodales por su retroalimentación.

Al Instituto de Ciencias Físicas y al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM por su apoyo durante la realización de mi servicio social.

Resumen

El objetivo de este trabajo es utilizar una técnica de aprendizaje de máquina en la reconstrucción independiente del modelo de funciones cosmológicas, como el parámetro de Hubble o la ecuación de estado de la energía oscura, a partir de mediciones en cronómetros cósmicos, con el propósito de ofrecer una alternativa a otra clase de mediciones como supernovas Tipo Ia y de esclarecer el comportamiento del Universo. Se hace énfasis en la determinación *a posteriori* del modelo que mejor representa la relación entre variables, mediante la selección del kernel según sus métricas.

Así mismo, se obtiene un valor para el parámetro de Hubble en el tiempo actual de $H_0 = 68.798 \pm 6.340\ \mathrm{km} \cdot s^{-1} \cdot \mathrm{Mpc}^{-1}$, el cual es un resultado que apoya las observaciones de Pantheon sobre el Fondo Cósmico de Microondas.

Abstract

The aim of this work is to use a machine learning technique in the model-independent reconstruction of cosmological functions, such as the Hubble parameter or the dark energy equation of state, from measurements in cosmic chronometers, with the purpose of offering an alternative to other kind of measurements such as supernovae of Type Ia and clarify the behavior of the Universe. Emphasis is placed on determining *a posteriori* the model that best represents the relationship between variables, by selecting the kernel according to its metrics.

Likewise, a value for the Hubble parameter is obtained at the current time of $H_0 = 68,798 \pm 6,340$ km $\cdot s^{-1} \cdot \text{Mpc}^{-1}$, which is a result that supports Pantheon's observations on the Cosmic Microwave Background.

Notación

- **AI**: Inteligencia artificial (Artificial Intelligence).
- **ANN**: Redes neuronales artificiales (Artificial Neural Network).
- **DE**: Energía oscura (Dark Energy).
- **DL**: Aprendizaje profundo (Deep Learning).
- **EoS**: Ecuación de estado (Equation of State).
- **GA**: Algoritmos genéticos (Genetic Algorithms).
- **GP**: Procesos gaussianos (Gaussian Processes).
- **GPR**: Regresión con Procesos Gaussianos (Gaussian Processes Regression).
- **ML**: Aprendizaje de máquina (Machine Learning).
- **PDF**: Función de densidad de probabilidad (Probability Density Function).
- **PMF**: Función de masa de probabilidad (Probability Mass Function).

Introducción

En 1929, el astrónomo Edwin Hubble confirmó que el Universo que habitamos se está expandiendo usando mediciones del corrimiento al rojo de galaxias espirales hechas por Milton Humason. Cuanto más grande es el corrimiento al rojo, más rápido se aleja un objeto de nosotros y, en conjunto, descubrieron que cuerpos celestes como galaxias se alejan de la Tierra más rápido cuanto más lejos están de ella. Este fenómeno se conoce hoy en día como la Ley de Hubble-Lemaître.

Las investigaciones de Hubble y Humason confirmaron que el Universo no es estático como asumía Einstein en su Teoría de la Relatividad General (1915 y 1916), sino dinámico como predecían las soluciones de las ecuaciones de campo de Alexander Friedmann, publicadas en 1922.

El consenso en el siglo XX era que la expansíon del Universo se ralentizaba con el pasar del tiempo debido a las interacciones gravitatorias, pues así lo predecía la Relatividad General. Pero todo cambió a partir de 1998, cuando las observaciones en supernovas Tipo Ia realizadas por los proyectos *Supernova Cosmology Project* y *High-Z Supernova Search Team*, mostraron que la expansión del Universo se acelera, contradiciendo lo que se esperaba teóricamente hasta entonces.

Utilizando el brillo de las supernovas se determinó su distancia a la Tierra y, a partir del espectro de potencias, se pudo obtener su corrimiento al rojo con lo que es posible conocer indirectamente cuán rápido se alejan dichos objetos de nosotros.

Además, los investigadores encontraron que las supernovas no estaban tan cerca como se esperaba, sino que se alejaban más rápido de lo predicho, concluyendo así que el Unverso se expande más rápido con el pasar del tiempo. En años posteriores se siguieron acumulando evidencias de este fenómeno, el cual pasó a llamarse *aceleración cósmica*.

Aquello que incrementa el ritmo de expansión del Universo se conoce como *energía oscura* y existen varios candidatos que han sido puestos a prueba, como la energía del vacío o la quintaesencia, sin embargo, éstas siguen siendo hipotesis de la verdadera naturaleza de la energía oscura. Actualmente, se desconoce qué es lo que provoca la aceleración cósmica y es uno de los temas de investigación abiertos más intrigantes en cosmología [1].

Los modelos más simples de energía oscura, como ΛCDM, la consideran una constante cosmológica con presión negativa p (ya que genera un incremento de volumen en el Universo) de igual magnitud a su densidad de energía ρ obteniendo la ecuación de estado $w = p/\rho = -1$. Sin embargo, aún existe debate sobre si la energía oscura es una constante cosmológica o no.

A lo largo de este trabajo se emplea una técnica de aprendizaje de máquina, conocida como Regresión con Procesos Gaussianos (GPR por sus siglas en inglés), en la reconstrucción no paramétrica de la ecuación de estado de la energía oscura y de otras funciones cosmológicas, a partir de observaciones experimentales, con el propósito de determinar la forma de w en la naturaleza y compararla con las predicciones teóricas, así como obtener una descripción de dichas funciones que no asuma una expresión explícita.

Al principio, se proporciona una explicación de los fundamentos matemáticos de la Regresión con Procesos Gaussianos, así como el procedimiento para implementarla. A continuación, se profundiza en la teoría cosmológica que motiva la reconstrucción y, finalmente, se presentan los resultados de la reconstrucción y su comparación con el modelo cosmológico ΛCDM.

Índice general

Agradecimientos II

Resumen III

Abstract IV

Notación V

Introducción VI

1 Inteligencia artificial 1

§1.1 ¿Qué es AI? . . . 1

§1.2 Formas de aprendizaje . . . 4

§1.3 Técnicas comunes en AI . . . 7

§1.3.1 Árboles de decisión . . . 7

§1.3.2 Redes neuronales artificiales . . . 8

§1.3.3 Algoritmos genéticos . . . 11

§1.3.4 Modelos no paramétricos . . . 12

2 Regresión con Procesos Gaussianos 14

§2.1 Preliminares matemáticos . . . 15

§2.1.1 Probabilidad . . . 15

§2.1.2 Distribuciones de probabilidad . . . 19

§2.1.3 Distribuciones de probabilidad multidimensionales. . . . 25

§2.2 Reconstrucción de una función y sus derivadas con GPR . . . 31

§2.2.1 Definición de GP . . . 31

§2.2.2 Inferencia Bayesiana en GPR 32
§2.2.3 Estimación de la derivada de una función 36
§2.3 Selección del kernel 37
§2.3.1 Prueba de χ^2 y Error Cuadrático Medio. 39
§2.4 Ejemplos de aplicación 40

3 Cosmología **44**
§3.1 El Universo en expansión 45
§3.1.1 Ley de Hubble 46
§3.1.2 La ecuación de Friedmann 49
§3.1.3 La ecuación de fluido 50
§3.1.4 La ecuación de aceleración 51
§3.1.5 La constante cosmológica 52
§3.2 Parámetros cosmológicos 53
§3.2.1 Parámetros de densidad 53
§3.2.2 Parámetro de desaceleración 57
§3.2.3 El modelo ΛCDM 58
§3.3 Medición de distancias en el Universo 59
§3.4 Cronómetros cósmicos 62

4 Reconstrucción de funciones cosmológicas **64**
§4.1 Parámetro de Hubble 64
§4.2 Derivadas de la distancia comóvil normalizada 66
§4.3 Parámetro de desaceleración 67
§4.4 Ecuación de estado de la energía oscura 69

5 Conclusiones **71**
§5.1 Sobre la GPR 71
§5.2 Sobre las funciones cosmológicas 72

Índice de figuras

1.1 Relación entre Inteligencia Artificial, Machine Learning y Deep Learning. . 2
1.2 Diferencia entre la programación clásica y el aprendizaje de máquina. . . . 3
1.3 Estructura simple para el algoritmo de un árbol de decisión. 7
1.4 Modelo de una neurona abstracta. 9

2.1 Diferentes funciones de densidad de probabilidad gaussianas. 23
2.2 Ejemplos de los diferentes tipos de correlación entre dos variables aleatorias. 29
2.3 Reconstrucción de modelo lineal con GPR a partir de observables con varianzas nulas y no nulas. 41
2.4 Comparación de los valores de MSE y χ^2 para modelos con diferentes kernels. 42
2.5 Reconstrucción de una función sinusoidal de prueba y sus derivadas. 43

3.1 Valores de H_0 determinados usando diferentes técnicas. 62

4.1 Reconstrucción del parámetro de Hubble con cronómetros cósmicos. 65
4.2 Reconstrucción de la primera derivada de la distancia comóvil normalizada. 66
4.3 Reconstrucción de la segunda derivada de la distancia comóvil normalizada. 67
4.4 Reconstrucción del parámetro de desaceleración. 68
4.5 Comparación del modelo GPR para $q(z)$ con un conjunto de datos simulados basados en ΛCDM. 69
4.6 Reconstrucción de la energía oscura. 70

5.1 Valores de χ^2 para los modelos desarrollados con diferentes funciones de covarianza. 71

Capítulo 1

Inteligencia artificial

Desde principios del siglo XX, la ciencia ficción imaginaba robots que pudieran pensar por sí mismos para realizar tareas que a los humanos les parecían tediosas o aburridas. Fue el deseo de que las máquinas pensaran por sí mismas lo que originó aquello que hoy conocemos como Inteligencia Artificial (AI).

Sólo hasta la década de 1950, Alan Turing sugirió que las máquinas podrían aprender de una forma similar a como lo hacen los humanos en su artículo *Computing Machinery and Intelligence*, dando inicio a una rama de la computación que rápidamente tomaría aceptación por su gran cantidad de aplicaciones [2].

Este capítulo proporciona un panorama general sobre lo que es la inteligencia artificial y sus aplicaciones haciendo énfasis en aquellas relacionadas con la ciencia. Particularmente, en cosmología, la AI se ha convertido en una herramienta muy útil, por su practicidad y resultados prometedores. Posteriormente, profundizaremos en una técnica particular de AI que se utilizará constantemente a lo largo de este trabajo para fundamentar su aplicación en la reconstrucción de funciones en cosmología.

1.1. ¿Qué es AI?

La Inteligencia Artificial nacio en la década de 1950, cuando varios pioneros de las ciencias de la computación, entre ellos Alan Turing, se preguntaban si las máquinas podrían pensar por sí mismas, con el propósito de automatizar tareas intelectuales que normalmen-

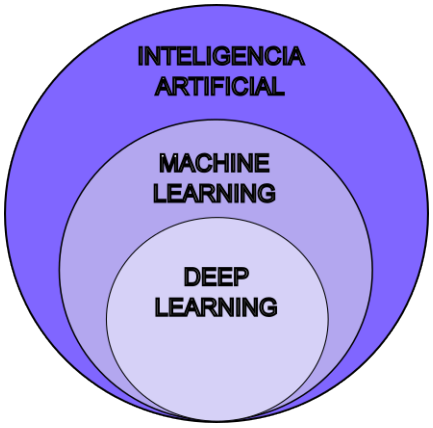

Figura 1.1: Relación entre Inteligencia Artificial, Machine Learning y Deep Learning.

te sólo podían realizar los humanos. Como tal, la AI es un campo general que engloba el aprendizaje automático (Machine Learning o ML) y el aprendizaje profundo (Deep Learning o DL), pero también incluye algunos temas que no involucran ningún aprendizaje, por ejemplo, las reglas utilizadas en los primeros programas de computadora que jugaban partidas de ajedrez o la robótica (Figura 1.1).

Durante mucho tiempo, se creyó que se podría emular por completo la inteligencia humana si los programadores elaboraban un conjunto de reglas lo suficientemente sofisticadas para una tarea específica. A este enfoque se le conoce como *inteligencia artificial simbólica* y fue el paradigma dominante hasta la década de 1980.

A medida que los problemas se volvían más complejos, las reglas necesarias para resolverlos también lo hacían, hasta que fue imposible encontrar reglas explicitas para tratar ciertos tópicos, como la traducción de idiomas o la clasificación de imágenes. Por lo anterior, fue necesario introducir un nuevo enfoque a la AI que sería el aprendizaje de máquina.

El aprendizaje de máquina o ML surge de la pregunta "¿puede una máquina ir más allá de lo que sabemos para descubrir por sí misma cómo realizar una tarea específica?". De esta manera, el enfoque del ML está en hacer que una computadora aprenda independientemente, a partir de datos, aquellas reglas que en AI simbólica debían ser conocidas extensamente por los programadores.

En la programación clásica, las entradas son reglas y datos para obtener respuestas. En cambio, en ML, las entradas son datos y respuestas de entrenamiento para encontrar

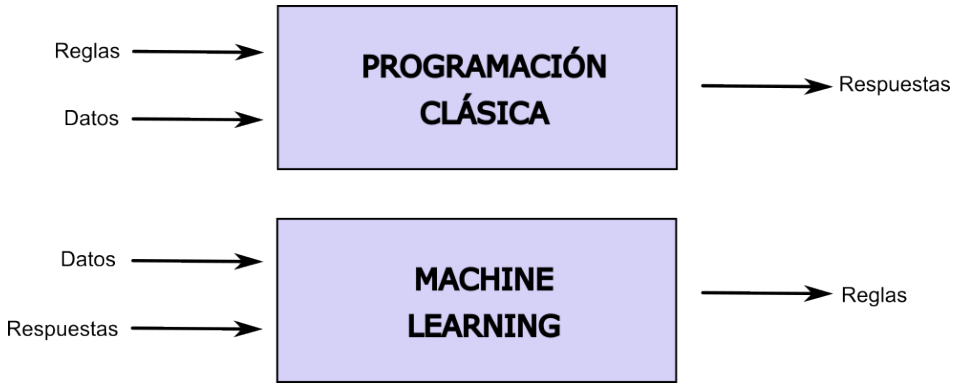

Figura 1.2: Diferencia entre la programación clásica y el aprendizaje de máquina.

las reglas que los relacionan, de este modo, el algoritmo de ML es capaz de utilizar las reglas que determinó para hallar respuestas sobre nuevos datos de prueba (Figura 1.2). Se dice que un sistema de ML es entrenado después de ser programado, ya que se somete a diferentes ejemplos representativos de la tarea que debe realizar y encuentra una estructura estadística que eventualmente permite al sistema determinar las reglas con las que automatiza la ejecución.

El aprendizaje de máquina está íntimamente relacionado con la estadística, pero difiere con ella en que el primero es capaz de tratar con conjuntos de datos mucho más grandes, con los que la segunda resultaría impráctica. Además, el ML presenta menor cantidad de teoría matemática en comparación.

Por otro lado, el Deep Learning o DL es un subcampo específico del ML. Es una nueva visión del aprendizaje a partir de datos que pone énfasis en diferentes capas de aprendizaje que incrementan su importancia gradualmente. Cuantas más capas tiene el modelo, se dice que mayor es su profundidad. Las técnicas modernas de DL involucran decenas o hasta centenas de capas sucesivas de representación.

A pesar de que los orígenes del DL son relativamente antiguos, no fue hasta la década de 2010 que alcanzó gran relevancia. Actualmente, algunos ejemplos de campos en que el DL ha tenido buenos resultados son: clasificación de imágenes, reconocimiento de diálogo, mejoras en la traducción de idiomas, desarrollo de asistentes digitales, conducción autónoma, mejoras en las búsquedas web, entre muchas otras.

Aún a día de hoy se está indagando en todo lo que el DL es capaz de hacer utilizándolo

en la mayor variedad de problemas posibles y tratando de llevarlo al extremo. Si se tiene éxito, podría significar una nueva era en la investigación científica [3].

1.2. Formas de aprendizaje

Se le llama *agente* a cualquier cosa que pueda considerarse que percibe su entorno a través de sensores e interactúa con dicho entorno usando actuadores. Por ejemplo, un agente humano tiene sensores como oídos, ojos y otros órganos, además de actuadores como manos, piernas y cuerdas vocales. Del mismo modo, un agente de software tiene pulsaciones de teclas, contenidos de archivo y paquetes de red a modo de sensores, mientras que cuenta con despliegue en pantalla, sonido y escritura de archivo como actuadores.

Un agente aprende si mejora su desempeño en tareas futuras después de hacer observaciones en su entorno. Una de las principales razones por las que es relevante que un agente de software aprenda es que el programador no siempre puede anticipar todas las posibles situaciones con las que el agente tendrá que lidiar. Además, las condiciones no siempre serán las mismas a lo largo de toda su vida útil esperada. Así mismo, existen situaciones en las que el programador no puede conocer las reglas necesarias para resolver determinado problema.

Cualquier componente de un agente se puede mejorar aprendiendo de los datos. Las mejoras y técnicas usadas para lograrlo, dependen mayormente de cuatro factores: qué componente se va a mejorar, qué conocimiento tiene el agente *a priori*, qué representaciones tienen los datos y con qué retroalimentación cuenta el agente.

Es necesario tener en cuenta cuál componente se busca mejorar porque no todos los componentes en un agente de software realizan la misma función y podría suceder que mejorar uno en específico no tenga relevancia en el desempeño deseado.

El conocimiento previo o *prior knowledge* se refiere a toda la información disponible sobre el problema además de los datos. Su importancia radica en la búsqueda y la optimización. Se sabe matemáticamente que todos los algoritmos de búsqueda presentan en promedio el mismo desempeño, lo que significa que para optimizarse es necesario usar un algoritmo especializado que incluya algo de conocimiento previo sobre el problema.

La representación de los datos se refiere al tipo de datos que un algoritmo de aprendizaje recibirá como entrada. La mayoría de los programas de machine learning y la investigación se centran en la representación factorizada, donde las entradas son vectores de valores de atributos y las salidas pueden ser tanto un valor numérico continuo como un valor discreto.

La retroalimentación o *feedback* es aquello que le dice al algoritmo de aprendizaje si un resultado resuelve el problema correctamente o no. De acuerdo al tipo de *feedback* que recibe un agente, las formas de aprendizaje se clasifican simplificadamente en tres grupos: no supervisado, por refuerzo y supervisado. En la práctica, la línea divisoria entre estas categorías no siempre es tan definida, pues algunos algoritmos hacen uso de varios aprendizajes a la vez.

Aprendizaje no supervisado

En este enfoque, el agente aprende patrones en los datos de entrada a pesar de que no se le proporciona retroalimentación explícitamente.

El ejemplo más común de aprendizaje no supervisado es el *clustering*, que consiste en agrupar items en categorías con características similares. Entre sus aplicaciones se encuentran la clasificación de elementos y la detección de patrones. Por ejemplo, un agente taxi podría organizar los días de trabajo agrupándolos como días de poco tráfico o de mucho tráfico y esta tarea la puede realizar sin que nadie nunca haya etiquetado alguno de los días.

Aprendizaje por refuerzo

Inspirado en la psicología conductista, se basa en que el agente aprenda a partir de una serie de refuerzos, que pueden ser recompensas o castigos. Se trata de un aprendizaje empírico, ya que el agente busca constantemente aquellas decisiones que le premien de algún modo, evitando las acciones que lo castiguen.

Por ejemplo, recibir puntos al ganar una partida de ajedrez le dicen al agente que algo hizo bien o no recibir ninguno le dice que algo hizo mal. Corresponde al agente decidir cuál de las acciones previas al refuerzo fueron los principales responsables de ello.

Aprendizaje supervisado

En este tipo de aprendizaje, el agente observa algún ejemplo de par (o pares) entrada-salida y desarrolla una función que mapea entradas en salidas. Es importante notar que un conjunto de salidas siempre está disponible para la percepción del agente, por lo que éste aprende directamente del entorno.

El objetivo del aprendizaje supervisado es:

> *"Dado un conjunto de entrenamiento con N parejas ordenadas de entrada-salida $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, donde cada y_i fue generada a partir de una función desconocida $y = f(x)$, determinar una función h que aproxime a la función real f."*

Aquí, x e y pueden ser cualquier valor, no necesariamente números, y la función h se conoce como *hipótesis*. Decimos que una hipótesis hace una buena *generalización* si predice correctamente el valor de y para nuevas pruebas. Algunas veces, la función f es estocástica, es decir que se representa por una distribución de probabilidad condicional[1] $P(Y|x)$ en lugar de una función explícita.

Cuando la salida y consta de un conjunto finito de valores (como soleado, lluvioso y seco), el problema de aprendizaje se llama **clasificación** y se trata de una clasificación binaria si y toma sólo dos posibilidades.

De otro modo, cuando y es un número (como en el tema que trataremos a lo largo de este trabajo), el problema se conoce como **regresión**. Técnicamente, resolver un problema de regresión es encontrar una distribución de probabilidad para cada valor de y, ya que encontrar el valor exacto de y para algún x no es posible.

Se dice que un problema es relalizable si el espacio de funciones contiene a la función buscada, pero no siempre se puede saber si un problema es realizable o no porque se desconoce a dicha función.

En algunos casos, vale la pena hacer distinciones más detalladas sobre el espacio de hipótesis $\mathcal{H}$ y no solamente determinar si una hipótesis es posible o imposible, sino más bien cuán probable es. El aprendizaje supervisado se puede hacer eligiendo la hipótesis h^* que es más probable dados los datos [4]:

[1] En el capítulo siguiente se definirá con precisión a una distribución de probabilidad condicional.

$$h^* = \arg\max_{h \in \mathcal{H}} P(h|\text{data}) \tag{1.1}$$

1.3. Técnicas comunes en AI

1.3.1. Árboles de decisión

Se trata de una de las formas más sencillas y exitosas de ML. Un árbol de decisión o *decision tree* representa una función que toma un vector de atributos como entrada y devuelve una decisión. Tanto las entradas como las salidas pueden ser discretas o continuas; si la salida tiene dos valores, hablamos de una clasificación *booleana*, donde cada entrada será clasificada como verdadera (positiva) o falsa (negativa).

Este tipo de algoritmos toman una decisión a partir de una serie de pruebas. Cada nodo en el árbol corresponde a una pregunta a la que será sometido el valor de una de las entradas y las ramificaciones representan las etiquetas que puede adquirir dicho valor al responderse la pregunta del nodo. Finalmente, cada hoja del árbol representa una de las categorías en las que será colocado el valor. Por lo tanto, el árbol de decisión es una función que toma valores y devuelve categorías (Figura 1.3).

Este algoritmo sólo es útil cuando el número de preguntas necesario no es tan grande, ya que la complejidad del problema puede aumentar exponencialmente de lo contrario. Sin embargo, para una amplia variedad de casos, este formato produce resultados favorables y concisos [4].

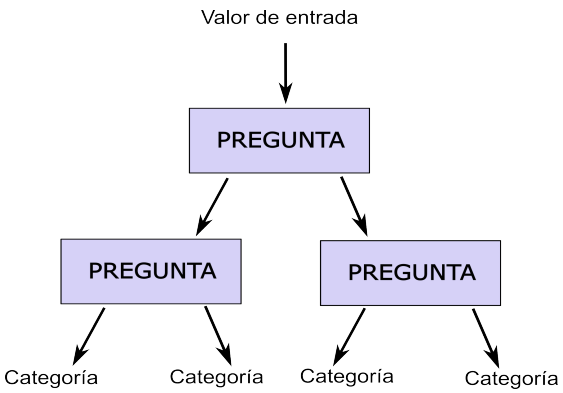

Figura 1.3: Estructura simple para el algoritmo de un árbol de decisión.

Los problemas que funcionan bien con el aprendizaje por árboles de decisión se identifican porque tienen un pequeño número de valores en todos los atributos (por ejemplo, un atributo "salsa" con valores "rojo" y "verde") y porque las salidas de la función tienen un número pequeño de valores discretos (como "sí" y "no"). Sin embargo, estas son condiciones ideales y algunas de estas limitaciones pueden evitarse [5].

Una colección de árboles de decisión organizados con un mismo objetivo es conocida como bosque aleatorio o *random forest* y casi siempre son el algoritmo preferido para cualquier tarea superficial de ML por su simplicidad y gran variedad de aplicaciones [3].

En cosmología, tanto los árboles de decisión como los bosques aleatorios se han aplicado exitosamente en la investigación. Por ejemplo, en la estimación de masas de galaxias [6], en el cálculo del corrimiento al rojo de objetos con el fin de encontrar la distancia a la que se encuentran de nosotros [7] o en la simulación de la formación de galaxias y la evolución del Universo [8], entre muchos otros. Además, constantemente se publican a día de hoy nuevos artículos de investigación que ponen a prueba estas técnicas.

1.3.2. Redes neuronales artificiales

Parece natural que, si se desea diseñar sistemas artificialmente inteligentes, podemos comenzar por analizar el sistema más inteligente que conocemos: el cerebro humano.

El cerebro humano consta de 10 a 100 mil millones de neuronas que están estrechamente relacionadas entre sí, algunas de las cuales pueden comunicarse con hasta cientos de otras neuronas a su alrededor. Tomando inspiración en este paradigma natural, se han desarrollado las redes neuronales artificiales (ANN) a lo largo de décadas pasadas y sus aplicaciones han abarcado desde el mercado de valores hasta el manejo autónomo de vehículos, pasando por la investigación científica.

El aprendizaje llevado a cabo por el cerebro humano es adaptativo debido a que las fuerzas de conexiones entre neuronas se modifican de acuerdo a reacciones con el entorno. De la misma forma las ANN deben contar con ciertos "pesos" que cambien para emular esta adaptabilidad.

El primer modelo de una neurona artificial se remonta a 1943 y se atribuye a McCulloch y Pitts, quienes buscaban entender y simular el comportamiento del sistema nervioso en

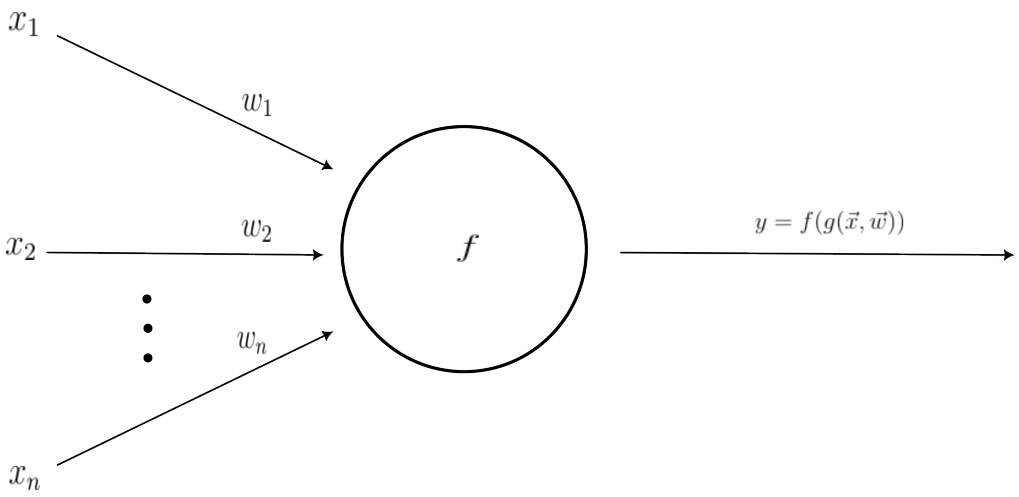

Figura 1.4: Modelo de una neurona abstracta.

animales. Hoy en día sabemos que las neuronas se comunican entre sí a través de dendritas, filamentos en forma de cabellos, que procesan información en el cuerpo o soma de la célula y que, cuando se presenta suficiente excitación, envían una señal eléctrica a lo largo de una prolongación conocida como axón. Una neurona usualmente tiene sólo un axón y varias dendritas; la señal fluye a través del axón hasta alcanzar la parte final llamada bulbo. El contacto axón-dendrita (o axón-soma o axón-axón) entre un bulbo terminal y una célula que influencia se llama *sinapsis*.

Hay cuatro elementos que la inteligencia artificial ha adoptado de este modelo biológico: el cuerpo de la célula, el axón, las dendritas y la sinápsis. Los equivalentes en ANN son, respectivamente, el cuerpo de una neurona, los canales de salida, los canales de entrada y los pesos. Los pesos son valores reales que reflejan el nivel conductivo de una sinápsis biológica que determina el grado de influencia que tiene una neurona sobre otra.

La entrada de una neurona es un vector con valores reales $\vec{x} = (x_1, x_2, \ldots, x_n)$ de n componentes y se define también un vector de pesos $\vec{w} = (w_1, w_2, \ldots, w_n)$. El cuerpo de una neurona desarrolla una función primitiva f para calcular las salidas $\vec{y}$ aplicándola sobre el producto punto de $\vec{x}$ con $\vec{w}$ o, más generalmente, sobre una función g de $\vec{x}$ y $\vec{w}$ (Figura 1.4). Se conoce a f como *función de activación* y a g como *función de entradas*.

Una ANN es una colección de neuronas abstractas correlacionadas en un determinado arreglo [5]. Una ANN se caracteriza por el patrón de conexiones entre sus neuronas (ar-

quitectura) el método con el que se modifican los pesos en las conexiones (entrenamiento o algoritmo de aprendizaje) y la función de activación.

Entre las arquitecturas típicas, se encuentra el arreglo en capas, donde las neuronas de una misma capa hacen uso, típicamente, de la misma función de activación y el mismo patrón de conexiones con otras neuronas, ya sea completamente interconectadas o sin interconectar. En este sentido, las ANN se pueden clasificar como multicapa o de capa simple.

Por otro lado, la forma en la que se modifican los pesos también determina la operación de una ANN. Se puede hacer uso tanto de entrenamiento supervisado como no supervisado; además hay redes cuyos pesos son fijos, es decir, no presentan un proceso de entrenamiento iterativo. Generalmente, se usa entrenamiento no supervisado cuando se desconocen las respuestas correctas y la ANN debe encontrarlas por sí misma.

En cuanto a la función de activación, elegirla es crucial en el comportamiento de la red neuronal y pueden ser desde una función identidad hasta combinaciones no lineales de las entradas. Algunos ejemplos de funciones de activación de uso común son: identidad, paso binario (Heavyside), sigmoide, tangente hiperbólica, entre otras. Cada función tiene aplicaciones donde se desempeña mejor y la elección dependerá del problema que se desea resolver [9].

Una de las más grandes desventajas de las ANN es que no explican cómo obtuvieron sus resultados y, actualmente, la pregunta de cómo podrían hacerlo sigue siendo un tema de investigación abierto.

Por supuesto que el objetivo principal que se busca con las ANN es diseñar una red que pueda emular las capacidades que tiene el cerebro humano, el cual aún está lejos de conseguirse. A pesar de que se han obtenido resultados prometedores en áreas como el procesamiento de señales, el reconocimiento de patrones, la producción y reconocimiento de lenguaje o los negocios y la medicina, las ANN presentan diferencias significativas con el pensamiento humano.

En la cosmología, las redes neuronales viven actualmente un auge de productividad en investigación. Constantemente, se publican resultados que usan ANN en temas como pruebas de la Relatividad General [10], mediciones de la constante de Hubble [11], la

emulación de fenómenos físicos en el Universo [12] o, como el que compete a este trabajo, en la reconstrucción de funciones cosmológicas [13, 14, 15], entre muchas otras aplicaciones. Las ANN son de las técnicas más populares de AI que se utilizan en la investigación cosmológica.

1.3.3. Algoritmos genéticos

En 1859, el naturalista Charles Darwin publicó el trabajo por el que es más reconocido: "El Origen de las Especies", donde se resumen sus investigaciones en la evolución de la flora y la fauna. Darwin concluye que las poblaciones de seres vivos se adaptan, es decir, que las características que hacen que un organismo responda más adecuadamente a las condiciones del medio se propagan con mayor frecuencia a lo largo de varias generaciones, en un proceso que se llama selección natural. Esta selección natural puede verse también como un tipo de aprendizaje en el que las especies en conjunto "aprenden" la mejor forma de responder al entorno en el que viven para garantizar su supervivencia.

El desarrollo de lo que hoy se conoce como Algoritmos Genéticos (GA) se atribuye a John Holland en la Universidad de Michigan a finales de la década de 1960, inspirado en los trabajos de Darwin.

En GA, se representan poblaciones como arreglos de cadenas, donde cada cadena es un cromosoma y cada entrada de una cadena es un gen. El proceso comienza con una población que representa una solución al problema. Se asigna a cada individuo de la población (cromosoma) una puntuación de aptitud que evalúa su habilidad para resolver el problema; se buscan aquellos individuos que tengan la mayor puntuación de aptitud, por lo que se les da mayor probabilidad de reproducirse que a otros. Después, algunos de los individuos mueren y se reemplazan por la nueva generación. Se espera que los individuos de generaciones posteriores tengan mejores genes que los primeros. Para modificar las poblaciones con el paso de las generaciones, existen tres operadores genéticos populares: selección, recombinación y mutación.

En la selección, se da preferencia a los individuos que formaran la nueva generación según su puntuación de aptitud. En la recombinación, se toman partes de dos o más individuos al azar y a partir de ellas se crea a la nueva generación emulando la reproducción

en la naturaleza. En la mutación, se introduce un gen cambiado a individuos al azar para mantener la diversidad en la población. Se puede hacer uso de uno o varios de estos operadores genéticos, según convenga en el problema a resolver [5].

Entre las ventajas de los GA se encuentra que no se ven afectados por el ruido, pues los individuos se eligen según su puntuación de aptitud, además de que el rango de posibles aplicaciones es muy amplio. Por otro lado, sus desventajas son que convergen muy lentamente (hacen falta varias generaciones) y que es muy sensible a los cambios en parámetros como los tamaños de población o la tasa de mutación.

Se han utilizado exitosamente en el procesamiento de señales e imágenes, en el modelado financiero, en operaciones logísticas, entre muchas otras. En cosmología se han usado para estudiar la inflación cósmica [16], en la clasificación de ondas gravitacionales [17] o en el modelado de la energía oscura [18], por mencionar algunas aplicaciones.

1.3.4. Modelos no paramétricos

Hasta ahora hemos visto que, por ejemplo, las redes neuronales usan datos de entrenamiento para estimar un conjunto de parámetros llamados pesos, por lo que, después del entrenamiento, el arreglo de pesos representa toda la información de los datos. Un modelo de este tipo, en el que un conjunto de parámetros representa a los datos de entrenamiento se conoce como *modelo paramétrico*. No importa qué tan grande sea la cantidad de datos que se aporten al aprendizaje de un modelo paramétrico, el número de parámetros necesarios no se verá afectado.

Por otro lado, un *modelo no paramétrico* es aquel que no puede ser caracterizado por un número finito de valores reales. Es un tipo de modelo que no asume ninguna forma específica para la relación entre las entradas y las salidas [4]. Son útiles cuando la densidad de probabilidad de la hipótesis o la forma explícita de la relación entre variables son desconocidas.

El más simple de los modelos no paramétricos es el método del histograma, en el que el espacio muestral de una variable de salida se divide en varios bins y, en cada bin, se aproxima la densidad de probabilidad por una constante proporcional al número de datos de entrenamiento que caen en el bin. Tiene la desventaja de que la forma y el tamaño

óptimo de los bins no siempre son tan claros, además de que las densidades de probabilidad se vuelven discontínuas al dividirlas en un número discreto de bins [19].

Otro método no paramétrico muy común es el de los k vecinos más cercanos (k-nearest neighbors o k-NN), que se utiliza principalmente en problemas de clasificación. Consiste en encontrar los puntos más cercanos a un determinado dato central utilizando una métrica particular, de modo que se pueda asignar una etiqueta a todos los puntos que caen dentro de un cierto radio de cada punto central. Existen diferentes métricas y su elección depende del tipo de problema y de datos que se tengan a disposición; como algunos ejemplos se encuentran la distancia euclidiana, la Manhattan y la de Minkowski.

En regresión, el método k-NN se usa como mejora para la técnica de "conectar los puntos" en una gráfica de dispersión, que es útil cuando el ruido de los datos es muy poco, lo cual se puede interpretar como que ciertas mediciones son muy parecidas a un valor certero. Usando k-NN, en lugar de simplemente conectar puntos, se eligen aquellos más cercanos a un centro de acuerdo a la métrica considerada para darle forma a la función hipótesis $h(x)$ [4]. Este enfoque es intuitivo y fácil de implementar, pero es muy suceptible a los valores anómalos (*outliers*).

Existe una amplia variedad de modelos no paramétricos, pero los mencionados anteriormente son considerablemente representativos.

Algunos modelos de AI no paramétricos han sido aplicados exitosamente en el reconocimiento de patrones, la minería de datos, predicciones en el mercado financiero, la clasificación de infromación, entre muchas otras áreas. En cosmología, los modelos no paramétricos se han vuelto una de las herramientas favoritas al momento de reconstruir funciones como ecuaciones de estado [20, 21] por la ventaja de ser aplicables cuando se desconoce la distribución de probabilidad de las hipótesis; además, algunas de estas técnicas se usan en conjunto con otros métodos de clasificación en AI para la creación de catálogos de galaxias [22] que se han vuelto necesarios debido al incremento de información observacional.

Capítulo 2

Regresión con Procesos Gaussianos

Los Procesos Gaussianos (GP) son la generalización de una distribución de probabilidad gaussiana conjunta para un número infinito de variables (esta definición quedará más clara posteriormente), que se usan en problemas de regresión y clasificación con AI. Los algoritmos que hacen uso de GP son modelos no paramétricos de aprendizaje supervisado y su uso se ha popularizado en la predicción de funciones desconocidas a partir de un conjunto de datos, ya que no requieren asumir una forma explícita de la relación entre las variables. La Regresión con Procesos Gaussianos (GPR) está basada en un *prior knowledge* y tiene la ventaja de que provee una incertidumbre sobre sus predicciones.

Desde su introducción detallada en 2005 por Carl Edward Rasmussen y Christopher K. I. Williams [23], se ha desarrollado una amplia variedad de código abierto para resolver problemas utilizando GP en diferentes lenguajes de programación, lo cual facilita su implementación al no tener que realizar el procedimiento desde cero.

A lo largo de este capítulo, se explicará a detalle en qué consiste la GPR, comenzando por los fundamentos matemáticos necesarios para comprender su forma de operación y se proporcionarán ejemplos de modelos de regresión programados en el lenguaje `Python` en su versión 3.12 con ayuda de algunas bibliotecas predefinidas. El objetivo principal es comprender por qué los GP se han convertido en una herramienta tan útil en la investigación cosmológica actual.

2.1. Preliminares matemáticos

En esta sección, se introducen las definiciones y teoremas que constituyen la base probabilística y estadística de la GPR.

2.1.1. Probabilidad

Cuando se lanza un dado de seis caras, existen seis posibles resultados $\{1, 2, 3, 4, 5, 6\}$ y ningún otro. Cada una de los posibles resultados se conoce como punto muestral (*sample point*) y el conjunto de todos los resultados posibles es el espacio muestral (*sample space*), denotado usualmente como Ω.

Llamamos *evento* a un subconjunto del espacio muestral. Por ejemplo, el evento donde se obtienen todos los números impares en el lanzamiento de un dado se denota como $A = \{1, 3, 5\}$. El evento que no contiene ningún punto muestral es un *evento vacío* ($\varnothing$), el que contiene sólo un punto es un *evento elemental*, el que contiene varios puntos es un *evento compuesto* y el que contiene todos los puntos muestrales es el *evento completo*. Dado que los eventos son conjuntos, heredan todas sus propiedades y operaciones del álgebra.

Se dice que dos o más eventos son *independientes* si la ocurrencia de uno no influye sobre la del otro, como sacar una carta de un mazo y después sacar otra habiendo regresado la primera. Los eventos son *dependientes* si no son independientes, como sacar una segunda carta sin devolver la primera.

La **probabilidad** es una medida de la certidumbre de que ocurra un evento y la probabilidad de un evento A se denota como $P(A)$ o, algunas veces, como $\Pr(A)$. El matemático ruso Andréi Kolmógorov estableció en 1933 los axiomas que rigen a la teoría de la probabilidad, los cuales son:

1. **No negatividad**. La probabilidad de ningún evento puede ser negativa.

2. **Unitaridad**. La probabilidad del espacio muestral completo es 1.

3. **Aditividad**. La probabilidad de la unión de un conjunto de eventos excluyentes (cuya intersección es vacía) es la suma de las probabilidades individuales.

Estas son las condiciones mínimas que debe verificar una función para ser considerada una probabilidad y, de ellas se derivan otras propiedades [19]. Cualquier función P cuyo dominio sean eventos, que devuelva números reales y que satisfaga los axiomas anteriores puede ser llamada una probabilidad.

Si Ω es un espacio muestral finito y si cada resultado es igualmente certero, definimos la probabilidad de A como la razón entre el número de puntos muestrales en A y el número de puntos muestrales en Ω, es decir, la fracción de la cardinalidad de A (denotada como $|A|$) entre la cardinalidad de Ω ($|\Omega|$):

$$P(A) = \frac{|A|}{|\Omega|}. \tag{2.1}$$

A esta ecuación se le conoce como definición clásica de la probabilidad y restringe su valor al intervalo $0 \leq P(A) \leq 1$.

Existen algunas limitaciones en la definición clásica, por ejemplo, que se debe modificar de alguna manera cuando la cantidad de resultados posibles es infinita y que no es aplicable cuando los resultados no son igualmente certeros, como cuando una moneda está trucada para dar una preferencia a obtener cara. Por lo tanto, se debe ampliar esta definición para llevar a la teoría casos como los anteriores.

Un *experimento aleatorio* es la reproducción controlada de un fenómeno cuyo resultado es indeterminado. Si un experimento aleatorio que puede dar lugar al evento A o a su complemento se repite un número de veces N, y el evento A ocurre n veces, el cociente entre n y N converge a la probabilidad de A cuando N es lo suficientemente grande, es decir:

$$P(A) = \lim_{N\to\infty} \frac{n}{N}. \tag{2.2}$$

Cuantas más veces se repita el experimento ($N \to \infty$) y, si los resultados son equitativamente certeros, más se acercará el cociente n/N a la probabilidad (2.1), recuperando la definición clásica.

En la aplicación de la teoría probabilística a problemas prácticos, es común encontrarse con la necesidad de predecir qué sucederá dado que tal o cual cosa ha sucedido ya. Por

ejemplo, uno puede estar interesado en determinar la probabilidad de que una bombilla pueda permanecer 100 horas más encendida, dado que ya lo ha estado por 24 horas. Este tipo de cuestiones son tratadas usando *probabilidad condicional*.

Sean dos eventos A y B, la probabilidad condicional de que ocurra A dado que ya ha ocurrido B, denotada como $P(A|B)$, se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2.3}$$

Para aclarar el uso de esta ecuación, veamos un ejemplo. Supongamos que, en un grupo de 50 personas, la mitad son hombres y la mitad son mujeres y que 10 personas en el grupo son mujeres menores de 18 años, ¿cuál es la probabilidad de que una persona tenga menos de 18 años sabiendo que es mujer?

En este caso, el evento A es que una persona sea menor de 18 y el evento B es que una persona sea mujer. Tenemos que la probabilidad de ambos eventos ocurriendo simultáneamente es $P(A \cap B) = 10/50$ y que la probabilidad de que una persona del grupo sea mujer es $P(B) = 1/2$. Entonces, sustituyendo estos valores en (2.3), se obtiene:

$$P(A|B) = \frac{10/50}{1/2} = \frac{2}{5} = 40\,\%. \tag{2.4}$$

Nótese que este valor es equivalente al que resulta de calcular la probabilidad considerando que hay $50/2 = 25$ mujeres en el grupo y que 10 de éstas son menores de 18 años, por lo que $P(A|B) = 10/25 = 2/5 = 40\,\%$. Además, la definición (2.3) es compatible con el enfoque frecuentista porque si se observa un número grande N de ocurrencias en un experimento aleatorio, entonces $P(A|B)$ representa la proporción de ocurrencias simultáneas de A y B con respecto al número de ocurrencias de B [24].

La probabilidad de dos eventos A y B independientes ocurriendo uno después del otro se calcula como el producto de las probabilidades de cada evento:

$$P(A \text{ y } B) = P(A) \cdot P(B). \tag{2.5}$$

Por otro lado, si A y B son dependientes, se tiene:

$$P(A \text{ y } B) = P(A) \cdot P(B|A). \tag{2.6}$$

Se le conoce como *variable aleatoria* a una función $X : \Omega \to \mathbb{R}$ que asigna un valor real a cada evento elemental $\omega \in \Omega$, es decir, que transforma los elementos del espacio muestral en números reales. Por ejemplo, el espacio muestral del lanzamiento de dos monedas es:

$$\Omega = \{(\text{cara}, \text{cara}), (\text{cara}, \text{cruz}), (\text{cruz}, \text{cara}), (\text{cruz}, \text{cruz})\}. \tag{2.7}$$

En este caso, una variable aleatoria X podría ser el número de caras obtenidas, de modo que cada punto muestral de Ω se mapea a 0, 1 o 2, según corresponda, al aplicarse X. Los valores en el rango de X pueden ser discretos o continuos.

Un *proceso estocástico* es una familia de variables aleatorias $\{X(t)|t \in T\}$, donde t denota generalmente el tiempo. Es decir, para cada valor de tiempo t en el conjunto T, se observa un número aleatorio $X(t)$. Se dice que un proceso estocástico es discreto si el conjunto de tiempos T es finito o contable (como $T = \{0, 1, 2, \dots\}$). En cambio, el proceso es continuo si T no es finito o contable.

Una *distribución de probabilidad* es una función que asigna a cada valor de una varibale aleatoria su correspondiente probabilidad. En el ejemplo anterior, $X = 2$ tiene 1 caso favorable (pues sólo hay un caso donde se obtienen 2 caras) de los 4 posibles en el espacio muestral, por lo que su probabilidad de ocurrencia es $P(X = 2) = 1/4$; en cambio, $X = 1$ tiene 2 casos favorables, por lo que su probabilidad es $P(X = 1) = 2/4 = 1/2$. Tal como las variables aleatorias, las distribuciones pueden ser discretas o continuas.

Si la probabilidad para cada valor de la variable aleatoria está dado por una función f, de tal forma que $P(X = x) = f(x)$[1] y la distribución es discreta, se le conoce a f como la función de masa de probabilidad (*probability mass function* o PMF) y debe cumplir que

[1]Cabe resaltar la diferencia de notación entre X que representa la variable aleatoria como función y x que representa uno de los valores que puede tomar dicha función.

$\sum_x f(x) = 1$ para satifacer el axioma de unitaridad. Cuando la distribución es continua, se conoce a f como función de densidad de probabilidad (*probability density function* o PDF) y debe satisfacer que $\int f(x)dx = 1$.

2.1.2. Distribuciones de probabilidad

Antes de discutir propiedades de las distribuciones de probabilidad, es importante definir algunas herramientas que las identifiquen. Las medidas de tendencia central son valores estadísticos que buscan resumir un conjunto de datos en uno sólo. En cambio, las medidas de dispersión evalúan qué tanto varía un conjunto de valores entre sí.

Medidas de tendencia central

El *valor esperado* $E[X]$ es el resultado que debería tener una variable aleatoria si un experimento se repite un número suficientemente grande de veces. Se define como el promedio de una variable aleatoria pesado con respecto a la función de probabilidad:

$$\begin{aligned} \text{Discreto:} \quad E[X] &= \sum_x x f(x), \\ \text{Continuo:} \quad E[X] &= \int x f(x) dx. \end{aligned} \tag{2.8}$$

Algunas veces el valor esperado se conoce simplemente como "promedio" y se denota con μ. El valor esperado puede no ser lo suficientemente representativo en presencia de *outliers*. En dichos casos, suele ser más útil la *media*, que se define como el valor b para el cual $P(x \leq b) = 1/2$. Es decir, la media es el "centro" de una distribución de probabilidad en el sentido de que es el punto del 50 % tanto desde el lado izquierdo como desde el lado derecho. Como generalización de la media, se define el α-cuantil para $0 \leq \alpha \leq 1$, como la constante b tal que $P(x \leq b) = \alpha$.

Medidas de dispersión

La *varianza* de una variable aleatoria se define como

$$V[X] = E[(X - E[X])^2], \tag{2.9}$$

la cual se puede expandir equivalentemente a

$$V[X] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - E[X]^2. \tag{2.10}$$

La raíz cuadrada de la variancia se conoce como *desviación estándar*:

$$D[X] = \sqrt{V[X]}. \tag{2.11}$$

Convencionalmente, la varianza y la desviación estándar se denotan como σ^2 y σ, respectivamente [19]. Más adelante, veremos que cualquiera de estos dos valores junto al promedio μ pueden representar completamente ciertas distribuciones de probabilidad particulares.

A continuación, revisaremos algunas distribuciones de probabilidad de uso común, mencionando sus valores característicos. Las derivaciones detalladas de ellos escapan del propósito de este trabajo, pero pueden consultarse en cualquier texto de estadística, como [24].

Ejemplos de distribuciones discretas

- **Distribución binomial**.

  Se conoce como *ensayos de Bernoulli* a las repeticiones independientes de experimentos con dos posibles resultados (por ejemplo: éxito y fracaso). Si p es la probabilidad de éxito, entonces $q = 1 - p$ es la probabilidad de fracaso. La distribución binomial, denotada por $Bi(n, p)$ es la distribución de probabilidad del número x de éxitos obtenidos en n ensayos de Bernoulli.

  La probabilidad de tener x éxitos está dada por p^x y la de tener $n - x$ fracasos es q^{n-x}. El número de combinaciones de x éxitos con $n - x$ fracasos está dado por el *coeficiente binomial* de n en x:

  $$\binom{n}{x} = \frac{n!}{x!(n-x)!}, \tag{2.12}$$

  donde $n!$ denota el factorial de n dado por $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$.

Utilizando los axiomas de probabilidad, la PMF de la distribución binomial es:

$$f(x) = p^x q^{n-x} \binom{n}{x}, \quad \text{para } x = 0, 1, 2, \ldots N. \tag{2.13}$$

El valor esperado y la varianza de la distribución binomial están dados por

$$E[X] = np \quad \text{y} \quad V[X] = npq. \tag{2.14}$$

- **Distribución de Poisson**.

  La distribución de Poisson, denotada como $Po(\lambda)$, es una generalización de la distribución binomial. Es la distribución de probabilidad cuya PMF está dada por

  $$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \tag{2.15}$$

  donde $\lambda = pn$. Surge por la necesidad de representar los posibles éxitos cuando la probabilidad p en una distribución binomial es muy baja y que cobra importancia cuando el número de repeticiones es muy grande.

  El valor esperado y la varianza de esta distribución son

  $$E[X] = \lambda \quad \text{y} \quad V[X] = \lambda. \tag{2.16}$$

- **Distibución geométrica**.

  En un conjunto de ensayos de Bernoulli con probabilidad de éxito p, el número de fracasos x hasta que se obtiene el primer éxito siguen una distribución geométrica, denotada por $Ge(p)$. Su PMF está dada por

  $$f(x) = p(1-p)^x, \tag{2.17}$$

  donde la probabilidad disminuye exponencialmente conforme x aumenta. El valor

esperado y la varianza de esta distribución son

$$E[X] = \frac{1-p}{p} \quad \text{y} \quad V[X] = \frac{1-p}{p^2}. \tag{2.18}$$

Ejemplos de distribuciones continuas

- **Distribución uniforme continua**.

  La distribución uniforme continua tiene una densidad de probabilidad constante en un intervalo finito $[a, b]$ y nula en el resto del espacio. Su PDF está dada por

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{en otro caso.} \end{cases} \tag{2.19}$$

  Esta distribución se denota como $U(a, b)$ y su valor esperado y varianza son

$$E[X] = \frac{a+b}{2} \quad \text{y} \quad V[X] = \frac{(b-a)^2}{12}. \tag{2.20}$$

- **Distribución normal**. También llamada *distribución Gaussiana*, se trata posiblemente de la distribución continua más importante. Una variable aleatoria continua x está normalmente distribuida si su PDF está dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right\}, \tag{2.21}$$

  donde los parámetros $-\infty < \mu < \infty$ y $\sigma > 0$ resultan ser el valor esperado (o promedio) y la desviación estándar de la distribución, respectivamente. Cualquier distribución definida por la PDF en (2.21) es una distribución normal.

  Los parámetros μ y σ^2 en conjunto, definen por completo a una distribución normal, por lo que, cuando una variable aleatoria X sigue una distribución Gaussiana, se denota como $X \sim N(\mu, \sigma^2)$ y se dice que X es una variable aleatoria normal. Si, adicionalmente, $X \sim N(0, 1)$, entonces X es una *variable aleatoria normal estándar*. La curva Gaussiana (Figura 2.1) es positiva, continua y simétrica alrededor de $x = \mu$

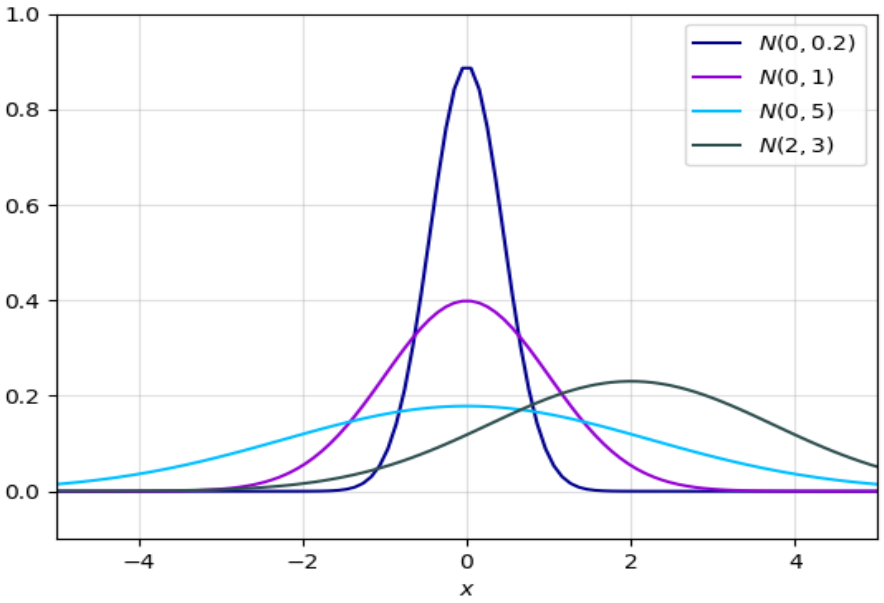

Figura 2.1: Diferentes funciones de densidad de probabilidad gaussianas.

y la varianza (o la desviación estándar) determina el ancho de la gráfica con forma de campana que tiene la distribución.

La distribución normal es un modelo razonable del comportamiento de una gran cantidad de fenómenos aleatorios, por lo que aparece constantemente en ciencias naturales y sociales. Su importancia radica en el *Teorema del Límite Central*, el cual establece que la distribución de n repeticiones de cualquier variable aleatoria, sin importar su densidad de probabilidad propia, converge a una distribución normal en el límite de infinitas repeticiones ($n \to \infty$) [24].

Por lo tanto, las cantidades físicas que son producto de la repetición de uno o varios procesos independientes, tal como los errores en las mediciones, se aproximan a ser normalmente distribuidas. Este hecho es fundamental en la experimentación científica, pues establece la base de una metodología en el tratamiento de datos, como podremos ver a lo largo de este trabajo.

Integrando la PDF (2.21), se puede encontrar que

$$\begin{aligned} P(\mu - 1\sigma \leq x \leq \mu + 1\sigma) &\approx 68.27\,\%, \\ P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) &\approx 95.45\,\%, \\ P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) &\approx 99.73\,\%, \end{aligned} \tag{2.22}$$

entonces, los múltiplos enteros de la desviación estándar determinan intervalos en los que se espera que el valor de la variable aleatoria normal caiga con una cierta probabilidad. A esta relación se le conoce como *regla empírica*.

- **Distribución de Cauchy** Se define, a partir de dos variables aleatorias normales estándar z y z', la distribución de Cauchy como aquella que sigue la variable que es el cociente entre z y z':

$$x = \frac{z}{z'}. \tag{2.23}$$

La PDF de una distribución de Cauchy es:

$$f(x) = \frac{1}{\pi(x^2 + 1)}. \tag{2.24}$$

El valor esperado se puede calcular a partir de (2.8) como

$$E[X] = \int_{-\infty}^{+\infty} \frac{x}{\pi(x^2+1)} dx = \frac{1}{2\pi}[\log(1+x^2)]_{-\infty}^{+\infty} = \frac{1}{2\pi} \lim_{\alpha \to +\infty,\ \beta \to -\infty} \log \frac{1+\alpha^2}{1+\beta^2}, \tag{2.25}$$

lo cual significa que depende de qué tan rápido incrementa el parámetro α respecto al decremento de β. Como el valor esperado no está determinado, esta distribución no tiene varianza.

Bajo ciertas condiciones [25], la varianza de una distribución de Cauchy se puede aproximar como

$$V[X] = V\left[\frac{Z}{Z'}\right] \approx \frac{\mu_Z^2}{\mu_{Z'}^2}\left(\frac{\sigma_Z^2}{\mu_Z^2} + \frac{\sigma_{Z'}^2}{\mu_{Z'}^2}\right). \tag{2.26}$$

La distribución de Cauchy es un tipo particular de distribución racional, la cual tiene la forma

$$z = \frac{x}{y}, \tag{2.27}$$

donde x e y son valores de variables aleatorias no necesariamente gaussianas.

2.1.3. Distribuciones de probabilidad multidimensionales.

Cuando se habla de múltiples variables aleatorias, la PDF o PMF dan lugar a distribuciones de probabilidad multidimensionales. En estos casos, es importante estudiar la dependencia de una variable sobre otra. Algunos tipos de relación entre dos variables x e y se discuten a continuación.

Distribución de probabilidad conjunta

Si x e y son valores de variables aleatorias discretas, la probabilidad de que tomen un determinado valor en un conjunto contable se denota $\Pr(x, y)$ [2]. La función que devuelve la probabilidad de obtener algún valor de las variables se llama *distribución de probabilidad conjunta* (*joint probability distribution*) y la PMF que la describe $f(x, y)$ es la función de masa de probabilidad conjunta, de modo que $\Pr(x, y) = f(x, y)$.

Las PMF de x e y por separado, $g(x)$ y $h(y)$, se relacionan con $f(x, y)$ como

$$g(x) = \sum_y f(x, y) \quad \text{y} \quad h(y) = \sum_x f(x, y). \tag{2.28}$$

Estas funciones se conocen como *distribuciones de probabilidad marginales* y el proceso de obtenerlas a partir de $f(x, y)$ se llama *marginalización*.

Este razonamiento se puede extender al caso continuo para definir la función de densidad de probabilidad conjunta como

$$\Pr(a \leq x \leq b, c \leq y \leq d) = \int_c^d \int_a^b f(x, y) dx dy. \tag{2.29}$$

[2] En este texto reservaremos la notación $\Pr(x, y)$ para referirnos exclusivamente a distribuciones de probabilidad multivariable, con el fin de resaltar su distinción con la distribución de probabilidad univaluada $P(x)$, pero en la práctica se usan indistintamente.

Las PDF de x e y por separado están dadas por

$$\begin{aligned} P(a \leq x \leq b) &= \int_a^b \int f(x,y)dydx = \int_a^b g(x)dx, \\ P(c \leq y \leq d) &= \int_c^d \int f(x,y)dxdy = \int_c^d h(y)dy, \end{aligned} \tag{2.30}$$

donde $g(x)$ y $h(y)$ son las marginalizaciones de $f(x,y)$.

La distribución de probabilidad conjunta se interpreta como la distribución de probabilidad de que dos eventos (determinados por los valores de la variables aleatorias) ocurran de forma simultánea.

Distribución de probabilidad condicional

Para variables discretas, se define, extendiendo el razonamiento de la ecuación (2.3), la *distribución de probabilidad condicional* (*conditional probability distribution*) de x dada y, denotada por $\Pr(x|y)$, como la probabilidad de que ocurra x después de que ocurre y:

$$\Pr(x|y) = \frac{\Pr(x,y)}{P(y)}. \tag{2.31}$$

La PMF de esta distribución es

$$g(x|y) = \frac{f(x,y)}{h(y)}. \tag{2.32}$$

Cuando x e y son contínuas, $P(y) = 0$ y no se puede definir la probabilidad condicional con (2.31), pero la PDF condicional se puede obtener usando (2.32).

Teorema de Bayes

A partir de la definición de probabilidad condicional (Eq. 2.3), se puede obtener

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{y} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{2.33}$$

Resolviendo para $P(A \cap B)$ e igualando, se tiene la siguiente ecuación:

$$P(A|B)P(B) = P(B|A)P(A) \tag{2.34}$$

o, equivalentemente,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.35}$$

A este resultado se le conoce como *teorema de Bayes* y tiene implicaciones trascendentales en la forma en que se calcula la probabilidad, ya que permite encontrar la probabilidad de un evento "causa" A dado que se observa un evento "efecto" B si se conocen las probabilidades de los eventos independientes, $P(A)$ y $P(B)$, y la probabilidad de que ocurra B dado que ocurre el evento A, $P(B|A)$. También se dice que el evento A es una *hipótesis* y el evento B es la *evidencia* del fenómeno.

Como $P(A)$ es la probabilidad de una causa antes de observar el efecto, se conoce como probabilidad previa o *prior*; se puede entender como el conocimiento pre-experimental que se tiene sobre la posibilidad de ocurrencia de la causa A. Por otro lado, $P(A|B)$ es la probabilidad de una causa después de que se ha observado un efecto, por lo que se llama probabilidad posterior o, simplemente, *posterior*. El término $P(B|A)$ es la probabilidad de que ocurra un efecto B dado que ha ocurrido la causa A por lo que se le da el nombre de verosimilitud o *likelihood*. Finalmente, $P(B)$ es la probabilidad marginal (tambien llamada *marginal likelihood*) y representa la posibilidad de ocurrencia del efecto bajo todas las hipótesis posibles.

El teorema de Bayes (también llamado *regla de Bayes*) puede ser extendido para explicar la relación de dos variables aleatorias x e y, utilizando la definición (2.31), como

$$\Pr(x|y) = \frac{\Pr(y|x)P(x)}{P(y)}. \tag{2.36}$$

Para variables continuas, usando las PDF g y h, el teorema se escribe

$$g(x|y) = \frac{h(y|x)g(x)}{h(y)}. \tag{2.37}$$

La importancia del teorema de Bayes en la estadística (y, particularmente, en la investigación en ciencias naturales) es que permite encontrar la probabilidad de una hipótesis según la evidencia que se tenga, por lo que se actualiza según las observaciones.

Covarianza y correlación

Se define la *covarianza* entre dos variables aleatorias X e Y como

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])]. \tag{2.38}$$

Este número mide el grado de variación de X e Y en conjunto respecto a sus medias. Si $\text{Cov}[X, Y] > 0$, un incremento en X tiende a presentar un incremento en Y; si $\text{Cov}[X, Y] < 0$, un incremento en X tiende a presentar un decremento en Y; si $\text{Cov}[X, Y] \approx 0$, las variables X e Y no están relacionadas entre sí. Es importante notar que la covarianza de una variable aleatoria con sí misma es igual a su varianza, al recuperar la ecuación (2.9).

La matriz que resume la varianza y covarianzas entre variables se llama matriz de covarianza:

$$\mathbf{\Sigma} = \begin{pmatrix} V[X] & \text{Cov}[X, Y] \\ \text{Cov}[Y, X] & V[Y] \end{pmatrix}, \tag{2.39}$$

y se puede extender su definición para resumir la comparación entre N variables aleatorias $X_1, X_2, \ldots X_N$ como:

$$\mathbf{\Sigma} = \begin{pmatrix} V[X_1] & \text{Cov}[X_1, X_2] & \ldots & \text{Cov}[X_1, X_N] \\ \text{Cov}[X_2, X_1] & V[X_2] & \ldots & \text{Cov}[X_2, X_N] \\ \vdots & \vdots & \ldots & \vdots \\ \text{Cov}[X_N, X_1] & \text{Cov}[X_N, X_2] & \ldots & V[X_N] \end{pmatrix}. \tag{2.40}$$

Dado que $\text{Cov}[X, Y] = \text{Cov}[Y, X]$, la matriz de covarianza es simétrica.

Se define el *coeficiente de correlación de Pearson* entre X e Y como su covarianza normalizada por el producto de las desviaciones estándar de cada variable, es decir:

$$\rho[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{V[X]}\sqrt{V[Y]}}. \tag{2.41}$$

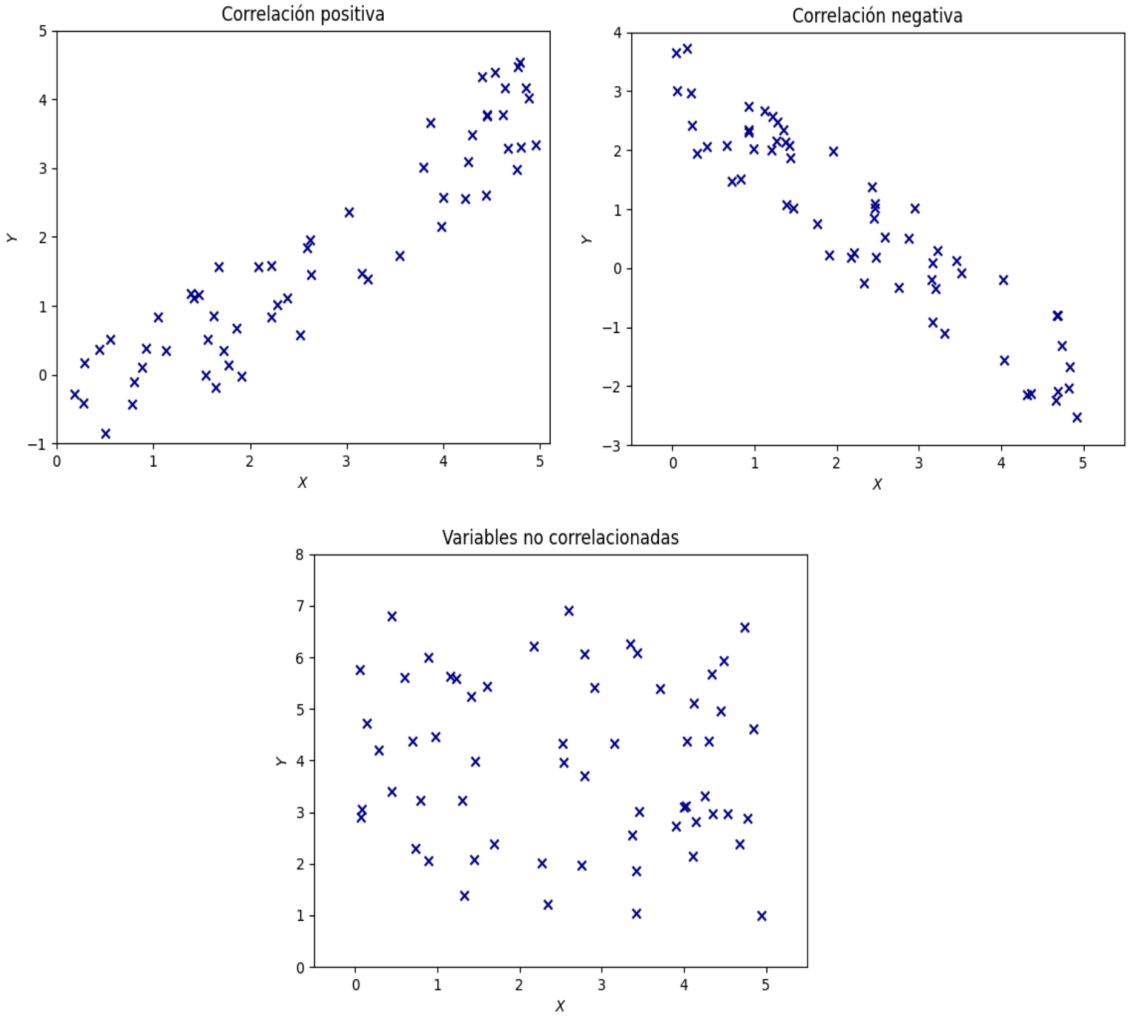

Figura 2.2: Ejemplos de los diferentes tipos de correlación entre dos variables aleatorias.

Este número tiene los mismos usos que la covarianza, pero su valor está limitado al intervalo $-1 \leq \rho \leq 1$. Si $\rho > 0$, se dice que X e Y están *positivamente correlacionadas*; si $\rho < 0$, X e Y están *negativamente correlacionadas* y, si $\rho = 0$, las variables no están correlacionadas. Los diferentes tipos de correlación se ilustran en la Figura 2.2.

Distribución normal multivariada

Un conjunto de n variables aleatorias normales estándar se representa en un vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$, donde $y_i \sim N(0, 1)$ para $i = 1, 2, \ldots,$ conocido como vector aleatorio

normal estándar. La PDF conjunta de las variables está dada por:

$$g(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{\left(-\frac{y_i^2}{2}\right)\right\} = \frac{1}{(2\pi)^{n/2}} \exp\left\{\left(-\frac{1}{2}\mathbf{y}^\top \mathbf{y}\right)\right\}. \tag{2.42}$$

El valor esperado y la matriz de covarianza de $\mathbf{y}$ son:

$$E[\mathbf{y}] = \mathbf{0} \quad \text{y} \quad V[\mathbf{y}] = \mathbf{I}, \tag{2.43}$$

donde $\mathbf{0}$ es el vector cero e $\mathbf{I}$ es la matriz identidad de tamaño n.

Si $\mathbf{x}$ es una transformación de $\mathbf{y}$ tal que

$$\mathbf{x} = \mathbf{T}\mathbf{y} + \boldsymbol{\mu}, \tag{2.44}$$

donde $\mathbf{T}$ es una matriz invertible de dimensión $n \times n$ y $\boldsymbol{\mu}$ es un vector de n entradas, la PDF conjunta de las variables en $\mathbf{x}$ está dada por

$$f(\mathbf{x}) = \frac{g(\mathbf{y})}{|\det(\mathbf{T})|} = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{2.45}$$

donde $\boldsymbol{\Sigma} = \mathbf{T}\mathbf{T}^\top$.

La ecuación (2.45) es la forma general de una *distribución normal multivariada* y se denota por $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. El valor esperado y la matriz de covarianza de $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ están dados por

$$\begin{aligned} E[\mathbf{x}] &= \mathbf{T}E[\mathbf{y}] + \boldsymbol{\mu} = \boldsymbol{\mu} \\ V[\mathbf{x}] &= V[\mathbf{T}\mathbf{y} + \boldsymbol{\mu}] = \mathbf{T}V[\mathbf{y}]\mathbf{T}^\top = \boldsymbol{\Sigma}. \end{aligned} \tag{2.46}$$

Las líneas de contorno de una distribución normal de dos dimensiones son elipses y sus ejes principales coinciden con los ejes de coordenadas cuando la matriz de covarianza es diagonal [19].

Si $\mathbf{x} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ e $\mathbf{y} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ son dos vectores de distribuciones normales multivariadas, entonces su distribución conjunta es también una distribución normal multivariada

definida por

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y\end{bmatrix}\right), \tag{2.47}$$

donde $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^{\top}$ es la matriz de correlación entre las variables que conforman a $\mathbf{x}$ e $\mathbf{y}$. En este caso, la distribución marginal de $\mathbf{x}$ es $\mathbf{x} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ y ocurre lo mismo para $\mathbf{y}$. Además, la distribución condicional de $\mathbf{x}$ dada $\mathbf{y}$ es [26]

$$\mathbf{x}|\mathbf{y} \sim N\left(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}\right). \tag{2.48}$$

2.2. Reconstrucción de una función y sus derivadas con GPR

2.2.1. Definición de GP

En la sección anterior, vimos que un conjunto de variables aleatorias pueden representarse con un vector aleatorio, digamos $\mathbf{h} = (h_1, h_2, \ldots, h_N)$. Del mismo modo, de manera más general, podemos decir que el conjunto de variables son el rango de una función $h : i \to h(i) = h_i$ con dominio $i = 1, 2, \ldots N$. Si el valor de N no es finito, el mapeo de la función $h : x \to h(x)$ se vuelve, intuitivamente, un vector de longitud infinita, donde $h(x)$ es una variable aleatoria para cada valor de x.

Si todas las variables en $\mathbf{h}$ siguen una distribución Gaussiana y el vector aleatorio es finito, entonces tenemos una distribución normal multivariada. Por otro lado, si el vector es infinito, se obtiene un proceso estocástico conocido como **Proceso Gaussiano** (*Gaussian Process* o GP). Por lo tanto, un GP difiere de una distribución Gaussiana en la cantidad de variables involucradas.

Más formalmente, un GP es una colección de variables aleatorias (proceso estocástico), donde cualquier subconjunto finito de ellas consiste de una distribución Gaussiana conjunta.

Así como el promedio μ y la varianza σ^2 definen por completo a una distribución normal, o como un vector de promedios $\boldsymbol{\mu}$ y una matriz de covarianza $\boldsymbol{\Sigma}$ determinan a una

distribución Gaussiana multivariada, un GP queda totalmente descrito por una función de promedios $m(x)$ y una función de covarianzas, también llamada **kernel**

$$K(x, x') = E[(h(x) - m(x))(h(x') - m(x'))] = \text{Cov}(x, x'), \tag{2.49}$$

la cual especifica la covarianza entre cualesquiera dos variables $h(x)$ y $h(x')$. Denotamos un conjunto de variables que siguen una distribución tipo GP como $f \sim GP$ o $f \sim GP(m(x), K(x, x'))$ [27].

2.2.2. Inferencia Bayesiana en GPR

Se conoce como *inferencia bayesiana* al proceso de ajustar una distribución de probabilidad (*prior*) a un conjunto de datos, utilizando la regla de Bayes (Eq. (2.37)), y resumir el resultado en otras distribuciones de probabilidad (*posterior*) sobre cantidades no incluidas en las observaciones con la finalidad de hacer predicciones. Se trata de un tipo de análisis estadístico en la categoría inferencial, la cual se basa en hacer generalizaciones sobre una población a partir de una muestra [28].

En el proceso, primero debe especificarse un prior que, en el caso de GPR, se trata de una distribución del tipo GP.

También se requiere de datos (usualmente llamados observaciones) que tomarán el rol de conjunto de entrenamiento ya que la regresión se trata de un aprendizaje supervisado. Los datos deben caracterizar muy bien la relación que hay entre entradas y salidas, pues el modelo dependerá estrechamente de este conjunto.

Finalmente, se encuentra la distribución posterior que refina al prior incorporando evidencia de las observaciones.

Prior

Para establecer una distribución de probabilidad $h \sim GP(m(x), K(x, x'))$ como prior, necesitamos caracterizarla mediante sus funciones de promedio y covarianza; a menos que se especifique lo contrario, por simplicidad, se utiliza $m(x) = 0$. Por otro lado, el kernel es una función que debe cumplir ciertas características para ser válido, pero uno de los

más comunes es el *exponencial cuadrático*. Profundizaremos en las diferentes funciones disponibles más adelante, ya que determinar el mejor kernel para un problema particular no siempre es tan inmediato y las características del modelo varían según cuál se elija.

Muchas de las funciones de covarianza son caracterizadas por parámetros ajustables, los cuales serán inferidos o aprendidos durante el entrenamiento mediante los datos utilizando algoritmos bayesianos. Cuando $m(x)$ no se considera igual a cero, también puede ser una función caracterizada por valores ajustables. Como $m(x)$ y $K(x, x')$ definen por completo a un GP y estos a su vez se caracterizan por parámetros, los parámetros de la función de promedios y de covarianzas se conocen como *hiperparámetros*.

El objetivo principal del aprendizaje en GPR es determinar los hiperparámetros adecuados para la distribución posterior. Los hiperparámetros son considerados los pesos del modelo y se agrupan en un vector $\boldsymbol{\theta}$.

Observaciones

Asumimos que las observaciones $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ están distribuidas alrededor del modelo de regresión f, donde los valores de y difieren del modelo evaluado en los correspondientes valores de x por un ruido aditivo Gaussiano $\varepsilon \sim N(0, \sigma_N^2)$,

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}) + \varepsilon \\ \mathrm{Cov}(\mathbf{y}) &= K(\mathbf{x}, \mathbf{x}) + \sigma_N^2 I, \end{aligned} \tag{2.50}$$

donde $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, I representa la matriz identidad de tamaño N y $K(\mathbf{x}, \mathbf{x})$ es la matriz de covarianzas obtenida al evaluar el kernel en los diferentes valores de x, es decir $[K(\mathbf{x}, \mathbf{x})]_{ij} = K(x_i, x_j)$. Las salidas $\mathbf{y}$ de los datos de entrenamiento también se denotan como $\mathbf{f}$.

Posterior

Se buscan las salidas $\mathbf{f}_*$ del modelo evaluado en puntos de prueba $\mathbf{x}_*$ diferentes a las observaciones en $\mathbf{x}$, de modo que $\mathbf{f}_* = f(\mathbf{x}_*) \sim N(\mathbf{0}, K(\mathbf{x}_*, \mathbf{x}_*))$. Nótese que $\mathbf{f}$ y $\mathbf{f}_*$ son distribuciones normales multivariadas, por lo que se puede escribir su distribución conjunta

mediante la relación (2.47) y la matriz de covariaza de las observaciones (2.50)

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{x},\mathbf{x}) + \sigma_N^2 I & K(\mathbf{x},\mathbf{x}_*) \\ K(\mathbf{x}_*,\mathbf{x}) & K(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix}\right). \tag{2.51}$$

Si hay N puntos de entrenamiento y N_* puntos de prueba, entonces $K(\mathbf{x},\mathbf{x}_*)$ denota la matriz de tamaño $N \times N_*$ de la función de covarianza evaluada en todas las parejas de entrenamiento y prueba, de manera similar para el resto de entradas $K(\mathbf{x},\mathbf{x})$, $K(\mathbf{x}_*,\mathbf{x}_*)$ y $K(\mathbf{x}_*,\mathbf{x})$.

Para obtener la distribución posterior, necesitamos restringir la distribución conjunta (2.51) para que contenga sólo aquellas funciones que coincidan con las observaciones. Lo anterior se consigue al condicionar la distribución sobre las observaciones usando la relación (2.48), con lo que se tiene

$$\mathbf{f}_*|\mathbf{f} \sim N(K(\mathbf{x}_*,\mathbf{x})[K(\mathbf{x},\mathbf{x}) + \sigma_N^2 I])^{-1}\mathbf{f}, K(\mathbf{x}_*,\mathbf{x}_*) - K(\mathbf{x}_*,\mathbf{x})[K(\mathbf{x},\mathbf{x}) + \sigma_N^2 I]^{-1}K(\mathbf{x},\mathbf{x}_*)). \tag{2.52}$$

La diagonal de la matriz de covarianzas de la distribución anterior corresponde a la varianza de una sola predicción f_*.

Se introduce la notación simplificada $K = K(\mathbf{x},\mathbf{x})$ y $K_* = K(\mathbf{x},\mathbf{x}_*)$, además, para el caso de un sólo punto prueba x_*, el vector de covarianzas entre el punto de prueba y los N puntos de entrenamiento se denota como $\mathbf{k}_*$. Por lo tanto, la predicción para el punto de prueba x_* sigue una distribución normal cuyo promedio y varianza están dados por [23]

$$\begin{aligned} E[f_*] &= \mathbf{k}_*^\top (K + \sigma_N^2 I)^{-1}\mathbf{y} \\ V[f_*] &= K(x_*,x_*) - \mathbf{k}_*^\top \left(K + \sigma_N^2 I\right)^{-1}\mathbf{k}_*. \end{aligned} \tag{2.53}$$

Estimación de máxima verosimilitud para hiperparámetros

Antes de poder usar la ecuación (2.53) para hacer predicciones, hace falta determinar los hiperparámetros $\boldsymbol{\theta}$ de un cierto kernel.

La probabilidad posterior para los hiperparámetros dadas las observaciones puede ser

calculada a partir del teorema de Bayes (2.37)

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{x})}, \tag{2.54}$$

aquí, $p(\boldsymbol{\theta})$ es conocida como el *hiper-prior*. En este caso, el *marginal likelihood* es la constante de normalización dada por

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.55}$$

La integral anterior no tiene solución analítica en la gran mayoría de los casos, por lo cual se hace una estimación $\hat{\boldsymbol{\theta}}$ de los hiperparámetros . Si los datos en el conjunto de entrenamiento son independientes, la densidad de probabilidad de las observaciones dados los hiperparámetros, que corresponde al *likelihood* en (2.54), se puede expresar como un producto de densidades individuales

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|x_i,\boldsymbol{\theta}). \tag{2.56}$$

Sustituyendo en (2.55) y usando el hecho de que el producto de distribuciones Gaussianas es otra distribución Gaussiana no normalizada, la integral se transforma en el *log marginal likelihood*

$$\log p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}\mathbf{y}^T(K+\sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K+\sigma_n^2 I| - \frac{n}{2}\log 2\pi. \tag{2.57}$$

El vector de los hiperparámetros buscados $\hat{\boldsymbol{\theta}}$ es aquel que maximiza el *log marginal likelihood*, es decir,

$$\hat{\boldsymbol{\theta}} \in \arg\max_{\theta} \log p(\mathbf{y}|\mathbf{x}). \tag{2.58}$$

El proceso de encontrar los valores que maximizan (2.57) se llama *maximización de evidencia*, *maximum likelihood estimation* o *estimación de máxima verosimilitud* y puede realizarse mediante algoritmos de optimización como Cadenas de Markov-Montecarlo (MCMC) o el Descenso de Gradiente [27].

2.2.3. Estimación de la derivada de una función

Por las propiedades de la distribución normal, la derivada de un GP también es un GP, lo cual permite extender la GPR para reconstruir derivadas de modelos.

La función de promedios de la primera derivada del modelo f' está dada por $m'(x)$ y su matriz de covarianzas se puede encontrar a partir de las derivadas del kernel de f como

$$\text{Cov}\left(f_i, \frac{\partial f_j}{\partial x_j}\right) = \frac{\partial K(x_i, x_j)}{\partial x_j}. \tag{2.59}$$

Derivando nuevamente, ahora respecto a x_i, se tiene

$$\text{Cov}\left(\frac{\partial f_i}{\partial x_i}, \frac{\partial f_j}{\partial x_j}\right) = \frac{\partial^2 K(x_i, x_j)}{\partial x_i \partial x_j}. \tag{2.60}$$

Por lo tanto, la distribución de f' es

$$f' \sim GP\left(m'(x), \frac{\partial^2 K(x_i, x_j)}{\partial x_i \partial x_j}\right) \tag{2.61}$$

y, utilizando la ecuación (2.47), la distribución conjunta de las observaciones $\mathbf{f}$ y las predicciones $\mathbf{f'}_*$ es

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f'}_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma_N^2 I & K'(\mathbf{x}, \mathbf{x}_*) \\ K'(\mathbf{x}_*, \mathbf{x}) & K''(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right), \tag{2.62}$$

donde $K'(\mathbf{x}, \mathbf{x}_*) = K'(\mathbf{x}_*, \mathbf{x})^\top$ y los elementos de cada matriz son [29]

$$[K'(\mathbf{x}, \mathbf{x}_*)]_{ij} = \frac{\partial K(x_i, x_{j*})}{\partial x_{j*}} \quad \text{y} \quad [K''(\mathbf{x}_*, \mathbf{x}_*)]_{ij} = \frac{\partial^2 K(x_{i*}, x_{j*})}{\partial x_{i*} \partial x_{j*}}. \tag{2.63}$$

Entonces, repitiendo el procedimiento usado para el modelo f, el promedio y la varianza de la predicción para un punto prueba x_* resulta ser

$$\begin{aligned} E[f_*] &= \mathbf{k'}_*^\top (K + \sigma_N^2 I)^{-1} \mathbf{y} \\ V[f_*] &= K''(x_*, x_*) - \mathbf{k'}_*^\top \left(K + \sigma_N^2 I\right)^{-1} \mathbf{k'}_*. \end{aligned} \tag{2.64}$$

El razonamiento puede ser extendido de manera análoga para obtener las derivadas subsecuentes del modelo, pero a lo largo de este trabajo sólo se hará uso de la primera.

Cabe destacar que para poder encontrar las derivadas con este método es necesario que el kernel sea derivable, entonces se prefieren kernels infinitamente diferenciables y con derivadas sencillas de obtener.

2.3. Selección del kernel

Como se ha podido ver hasta ahora, la función de covarianzas es un ingrediente crucial en un GP, ya que contiene información sobre la relación que siguen las variables aleatorias en el proceso.

En general, no cualquier función con entradas $\mathbf{x}$ y $\mathbf{x}'$ puede ser un kernel válido. Para que una función represente las covarianzas entre dos variables, ésta debe ser positiva semi-definida, lo que significa cualquier matriz $K_{ij} = K(x_i, x_j)$ que produzca debe ser simétrica e invertible. Además, cualquier escalar obtenido como $z^\top Kz$, donde z es un vector con dimensión adecuada, debe ser positivo o cero.

Se dice que un kernel es *estacionario* si es una función de $\mathbf{x}-\mathbf{x}'$, pues es invariante ante translaciones en el espacio de entradas. Si además, es una función solamente de $|\mathbf{x}-\mathbf{x}'|$ (la distancia entre las entradas), se le llama *isotrópico* por ser invariante ante transformaciones rígidas.

El proceso de crear un nuevo kernel no siempre es trivial, por lo que se suele elegir uno de un conjunto de funciones predefinidas para resolver un problema. A continuación revisaremos algunos kernels de uso común y sus hiperparámetros.

Exponencial cuadrático (RBF)

Se trata de un kernel isotrópico definido por la función

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2l^2}\right), \tag{2.65}$$

donde $l > 0$ es el hiperparámetro conocido como la longitud característica. Esta función es infinitamente diferenciable, por lo que los GP con este kernel tienen derivadas de todos los órdenes y por eso es la mejor opción para reconstruir las derivadas de un modelo.

Como K depende solamente de la distancia radial $r = |\mathbf{x}-\mathbf{x}'|$, se le suele llamar también *radial basis function* o RBF. Algunas veces se denota en términos de un parámetro θ que depende de la longitud característica

$$K(\mathbf{x}, \mathbf{x}') = e^{-\theta|\mathbf{x}-\mathbf{x}'|^2}. \tag{2.66}$$

Producto punto

Si una función de covarianza depende sólo de $\mathbf{x}$ y $\mathbf{x}'$ a través de $\mathbf{x} \cdot \mathbf{x}'$, hablamos de una función producto punto (*dot product*). Es un kernel sencillo y no estacionario dado por

$$K(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x} \cdot \mathbf{x}', \tag{2.67}$$

donde σ_0 es el hiperparámetro que controla la inhomogeneidad de la función. Este kernel es invariante ante rotaciones de coordenadas alrededor del origen, pero no ante traslaciones.

Matern

Es una generalización del kernel RBF, por lo que también es un kernel estacionario. Se define a través de la función

$$K(\mathbf{x}, \mathbf{x}') = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{l} |\mathbf{x} - \mathbf{x}'| \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{l} |\mathbf{x} - \mathbf{x}'| \right), \tag{2.68}$$

con parámetros positivos ν y l, donde K_ν es la función Bessel modificada y Γ es la función gamma. Cuanto más pequeño es el valor de ν, menor es la suavidad que presenta la función resultante. Como K depende de $|x - x'|^\nu$, es anisotrópico.

Racional cuadrático (RQ)

Se puede interpretar como una suma infinita de kernels RBF con longitudes características diferentes, por lo que también es estacionario. Se define como

$$K(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{|\mathbf{x} - \mathbf{x}'|^2}{2\alpha l^2} \right)^{-\alpha}, \tag{2.69}$$

donde $l > 0$ es la longitud característica global y α es el hiperparámetro de mezcla de escala.

Exp-Sine-Squared (kernel periódico o ESS)

El kernel periódico o *Exp-Sine-Squared* permite modelar funciones que se repiten periódicamente de manera exacta. Está dado por

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{2\sin^2(\pi|\mathbf{x} - \mathbf{x}'|/p)}{l^2}\right), \tag{2.70}$$

donde $l > 0$ se conoce como la longitud de escala y $p > 0$ como la periodicidad del kernel [23].

2.3.1. Prueba de χ^2 y Error Cuadrático Medio.

En la mayoría de problemas de GPR, no es posible determinar *a priori* el mejor kernel, ya que se requiere conocer la relación que existe entre las entradas y las salidas, pero eso no siempre es posible. En estos casos, la alternativa disponible es obtener una serie de modelos con diferentes kernels y determinar cuál de ellos representa mejor a la función buscada.

Se puede utilizar cualquier técnica estadística para saber cuál modelo se ajusta de manera más adecuada al conjunto de observables, pero una de las más prácticas y populares es la *prueba de* χ^2. El método consiste en definir una función objetivo

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i)]^2}{\sigma_{y_i}^2}, \tag{2.71}$$

donde (x_i, y_i) son los datos en el conjunto de entrenamiento (observables), $\sigma_{y_i}^2$ son sus varianzas y $f(x_i)$ son las predicciones del modelo evaluado en las mismas variables independientes que las observables. Es claro que todos los datos deben tener varianzas no nulas para poder emplear esta técnica. Cuando se usa GPR en un problema real, las observables frecuentemente son mediciones reproducibles que siguen una distribución normal por el teorema del límite central, por lo que suelen tener una varianza asociada.

El valor χ^2 se interpreta como una medida de las distancias entre los puntos de entrenamiento y los valores que el modelo arroja en los mismos puntos. Por lo tanto, el modelo más adecuado será aquel que minimice a χ^2.

Nótese que este método no es aplicable cuando al menos una de las varianzas en las observables es cero, pues la ecuación (2.71) se indetermina. Para este caso, se puede utilizar una métrica diferente conocida como Error Cuadrático Medio (*Mean Squared Error* o MSE), la cual es la suma de la distancias euclideanas cuadradas entre evaluaciones del modelo $f(x_i)$ y puntos en algún modelo esperado o de prueba y_i, escalada por el número de puntos N. El MSE se escribe como:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i)]^2. \tag{2.72}$$

De manera similar a χ^2, el modelo que minimice el valor de MSE se considera como el más adecuado.

2.4. Ejemplos de aplicación

Como se mencionó al principio de este capítulo, a día de hoy existe una gran variedad de códigos computacionales pre-programados de uso libre para aplicar la GPR a diferentes problemas, entre ellos se encuentran `GPy` [30], `GPflow` [31], `GPyTorch` [32], `scikit-learn` [33] and `GaPP` [34]. A continuación, presentamos ejemplos de la reconstrucción de modelos sencillos utilizando GPR, mediante los proyectos de `scikit-learn` y `GaPP`; el desarrollo completo puede verse en el repositorio público [35].

Modelo lineal

Para comenzar, se construye un conjunto de datos simulados aleatorios obtenidos a partir de una relación lineal que tomará el papel de observables. La salida de cada punto de entrenamiento es resultado de evaluar la ecuación de una línea recta $Y = mX + b$, con $m = 3$ y $b = -4$, en alguna entrada $X \in [0, 10]$ y agregar o sustraer un número aleatorio entre -15 y 15. El objetivo es determinar el modelo de la línea recta que representa la

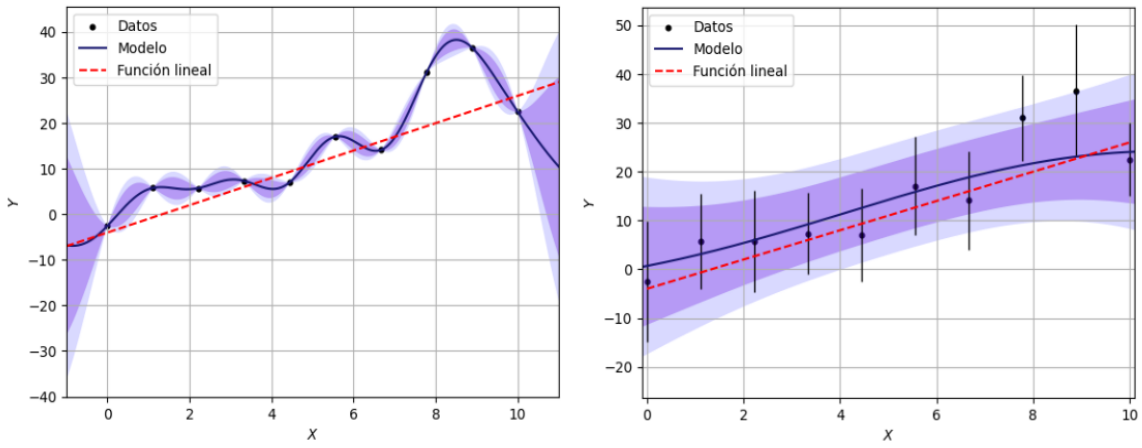

Figura 2.3: Reconstrucción de modelo lineal con GPR a partir de observables con varianzas nulas y no nulas.

relación entre los puntos usando solamente GPR.

El caso más simple se obtiene ignorando las varianzas que pudieran presentar los datos, es decir, haciendo $\sigma_{yi}^2 = 0$. Esto sería equivalente a asumir que las mediciones de una prueba son exactas, por lo que no hay barras de error asociadas.

En el código, el objeto `GaussianProcessRegressor()` inicializa el *prior* con el kernel que se especifique y el método `fit()` realiza el ajuste de los hiperparámetros mediante el *maximum likelihood estimation* a partir de los observables proporcionados como parámetros en dos listas `x` e `y`. Finalmente, el método `predict()` regresa los promedios y desviaciones estándar de las predicciones hechas sobre otra lista de entradas, usando las ecuaciones (2.53).

La Figura 2.3 (izquierda) muestra el conjunto de datos simulados, las zonas de confianza correspondientes a 2σ y 3σ, la gráfica de la relación lineal de la cual se obtuvieron los observables y el modelo de la regresión a partir de un kernel RBF (se eligió este kernel de manera ilustrativa, pues es muy parecido a los modelos a partir de otros kernels, los cuales pueden observarse a detalle en [35]). Nótese que las varianzas de las predicciones se reducen a cero cuando el modelo es evaluado en las observables, lo que significa que sobreajustan las predicciones en estos puntos durante el aprendizaje y por eso no se obtiene el modelo lineal esperado.

Cuando las observables tienen errores no nulos, sus varianzas deben agregarse a la

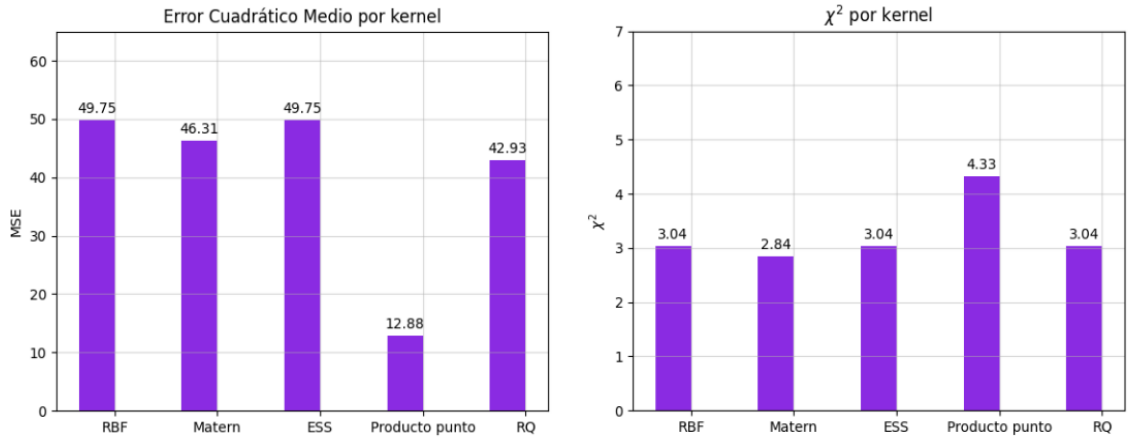

Figura 2.4: Comparación de los valores de MSE y χ^2 para modelos con diferentes kernels.

diagonal de la matriz de covarianzas definida por el kernel, como sugiere la ecuación (2.50). El objeto `GaussianProcessRegressor()` puede tomar como atributo a un arreglo `alpha` de longitud igual a las de las listas donde se encuentran las observables y que contiene las varianzas asociadas a cada dato.

Repitiendo el proceso con un arreglo de varianzas aleatorias, se obtiene el modelo de la Figura 2.3 (derecha), el cual representa mucho mejor la relación lineal buscada.

Es posible usar el MSE (Eq. 2.72) y la función objetivo χ^2 (Eq. (2.71)) para encontrar el kernel que genera el modelo más óptimo, cuando se tienen varianzas nulas y no nulas, respectivamente. Para esto, se crean modelos de prueba con diferentes kernels y se calcula su valor correspondiente de MSE o χ^2. La Figura (2.4) muestra los resultados y se puede concluir que la regresión que representa mejor a las observables es el que se obtuvo con un kernel producto punto con varianzas nulas y un kernel Matern con varianzas no nulas.

Modelo de una función y sus derivadas

La librería que permite encontrar derivadas de una función con GPR es `GaPP` publicada en 2012 por [29]. El código es capaz de calcular hasta la tercera derivada de una función y no es posible elegir un kernel diferente al RBF para asegurar su propiedad de diferenciabilidad.

Para poner a prueba este software, se creó un conjunto de datos simulados, como

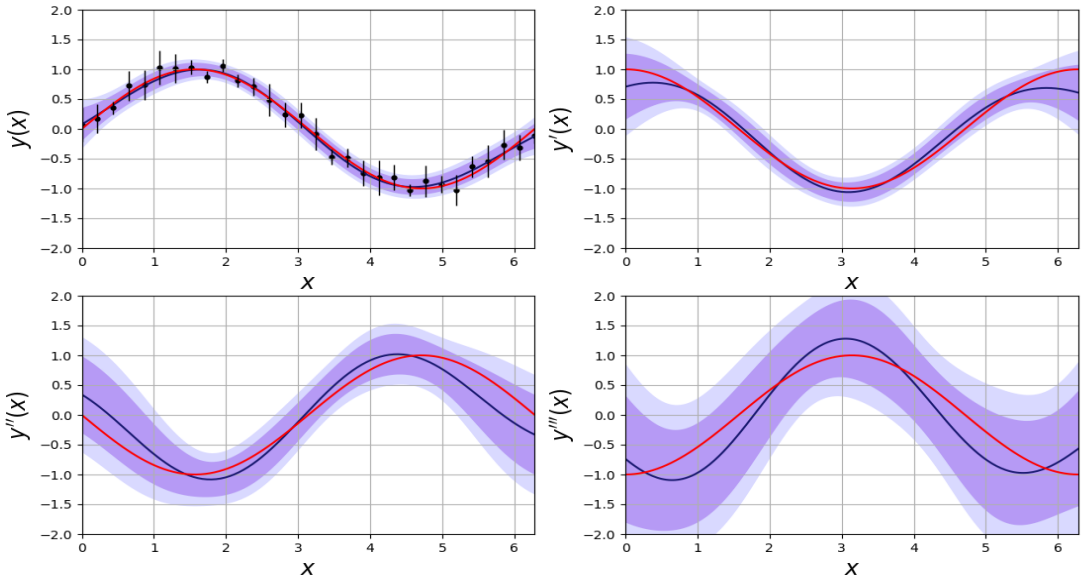

Figura 2.5: Reconstrucción de una función sinusoidal de prueba y sus derivadas.

en el caso del modelo lineal, pero esta vez dispersos alrededor de una función sinusoidal $y = \sin(x)$, ya que encontrar sus derivadas analíticas es muy sencillo.

La Figura (2.5) muestra la regresión de la función y sus derivadas (líneas azules), así como la curva sinusoidal que se usó para crear el conjunto de observables y sus derivadas analíticas (líneas rojas). Cuando se calcula la derivada de la relación entre los observables, no existen (y tampoco es necesario tener) datos que representen directamente a la función derivada, porque el modelo se obtiene a partir de la función original.

Si bien los modelos no son idénticos a la función buscada, se puede ver claramente que las predicciones representan fielmente a las funciones analíticas correspondientes porque, en todos los casos, la línea roja está dentro del intervalo delimitado por 2σ.

Capítulo 3

Cosmología

La Cosmología se encarga de estudiar la estructura dinámica del Universo a gran escala. Las bases de esta disciplina se encuentran en la Relatividad General, a pesar de que es posible desarrollarla con un enfoque Newtoniano utilizando varias suposiciones, pero la interpretación de algunos resultados es diferente. De cualquier forma, la Relatividad General ha sido puesta a prueba en numerosas ocasiones demostrando su validez, como en la medición de Eddington, por lo que el enfoque Newtoniano no se ha desarrollado en profundidad.

En los primeros años de la Cosmología moderna (1916-1917), se creía que el Universo tenía la misma estructura siempre, es decir, que era estático. Sin embargo, en décadas más recientes, ha surgido un modelo como la mejor opción disponible para explicar el comportamiento y origen del Universo: la Teoría del Big Bang.

En este modelo, se asume que, hace unos 13.8 millones de años, ocurrió un evento cataclísmico llamado "Big Bang" cuando el Universo comenzó a expandirse a partir de un punto singular. En las fases más tempranas, había sólo radiación a temperaturas y densidades muy altas y, a medida que el Universo se expandía, éstas disminuyeron para dar lugar a los electrones y protones a partir del baño de radiación. Una vez que el sistema se enfrió lo suficiente, las partículas se relacionaron formando primero átomos de los elementos más simples como hidrógeno o helio y, posteriormente, produciendo materiales más pesados. A medida que el espacio se expandía, las condiciones se volvían favorables para que la materia primitiva se condensara en nebulosas, estrellas y galaxias.

En el desarrollo de esta teoría, se ha hecho uso de las leyes de la física a nivel local porque estamos limitados a ellas y, aún así, hasta ahora han probado ser exitosas en numerosas ocasiones proporcionando información sobre la estructura y comportamiento del Universo. Desafortunadamente, no debemos descartar la posibilidad de que existan interacciones adicionales que sólo son evidentes en escalas cosmológicas. Además, la Cosmología tiene la particularidad de que los fenómenos que estudia no pueden reproducirse a voluntad ni aislarse de otros. A pesar de esto, se ha podido entender el Universo hasta los 10^{-43} segundos después de su nacimiento[36].

A lo largo de este capítulo describiremos algunas de las definiciones, resultados y desafíos actuales que constituyen a la Cosmología moderna, con el propósito de utilizar la GPR en un problema real.

3.1. El Universo en expansión

Todo el conocimiento cosmológico actual está basado en un principio de simplicidad[1] conocido como *el principio cosmológico*. Éste es, en escencia, una generalización del principio Copernicano de que la Tierra no es el centro del Sistema Solar, por lo que, siguiendo este pensamiento, no podríamos esperar que ningún cuerpo en el Universo ocupara alguna posición especial o favorecida. El principio cosmológico se establece formalmente como:

"En cada época, el universo presenta el mismo aspecto a partir de cada punto, excepto por irregularidades locales."

Podemos pensar en el Universo como una superficie (más propiamente, una variedad) de 3 dimensiones espaciales y una temporal en la que no hay ninguna posición privilegiada para cada corte temporal, esto significa que es *homogéneo*. Técnicamente, una variedad es homogéna si existe un grupo de isometrías (transformaciones que preservan distancias) que mapean cada punto en cualquier otro punto. La homogeneidad del universo debe ser comprendida a gran escala (10^8 a 10^9 años luz) para despreciar las irregularidades locales, sin embargo, comparada con su tamaño ($\sim$ 92 mil millones de años luz), esta distancia es corta.

[1]Un principio de simplicidad o de parsimonia es la idea de que explicaciones u observaciones más simples son preferidas por encima de aquellas más complejas [37].

El principio requiere que no sólo cada corte carezca de posiciones privilegiadas, sino que tampoco tenga direcciones preferenciales alrededor de cualquier punto. Se dice que una variedad es *isotrópica* si cumple esta característica y es globalmente isotrópica si es isotrópica con respecto a todos sus puntos. Se puede probar que si una variedad es globalmente isotrópica, necesariamente también es homogénea.

Por lo tanto, el principio cosmológico alude a la necesidad de que el universo debe ser homogéneo e isotrópico y existen algunas evidencias que nos hacen pensar que así es. Por ejemplo, el mayor sustento para la isotropía global se encontró en 1965 con el descubrimiento del fondo cósmico de microondas (*Cosmic Microwave Background* o CMB) por Arno Penzias y Robert Wilson; encontraron que el universo está totalmente impregnado con radiación a una temperatura de 2.7 K y que esta radiación es isotrópica a fracciones de uno por ciento. Se sabe que el CMB es un remanente del Big Bang que proporciona información sobre el universo primordial [36].

3.1.1. Ley de Hubble

La razón principal para desechar el modelo de un universo estático es que su existencia es incompatible con la ley del incremento de entropía; un universo de este estilo no podría existir para siempre en un mismo estado. Sin embargo, la prueba real está en que, desde la primera década del siglo XX, los astrónomos ya hacían observaciones de las velocidades radiales entre galaxias utilizando el efecto Doppler en sus líneas espectrales, esperando que dichas velocidades fueran aleatorias. Vesto Slipher, en 1914, fue el primero en encontrar que esto no era así, ya que la mayoría de los espectros de las galaxias analizadas presentaban un corrimiento al rojo, lo que indicaba que se alejan rápidamente de la Tierra y, posteriormente, se determinó que también se alejan entre sí. A partir de este descubrimiento, se comenzó a hablar en términos de una expansión cosmológica.

Algunos años más tarde, en 1925, el astrónomo Edwin Hubble combinó las velocidades calculadas por Slipher con sus mediciones de distancias a diferentes galaxias y encontró que la velocidad de recesión v (aquella con la que todos los objetos astronómicos se alejan entre sí) es proporcional a la distancia d con respecto al observador, es decir:

$$v = H_0 d, \tag{3.1}$$

donde H_0 es conocida como la constante de Hubble. Actualmente, esta ecuación se referencia como la *Ley de Hubble* y, después de su descubrimiento, Hubble y su asistente Humason la verificaron con 32 galaxias diferentes, con lo que la expansión del universo se volvió un hecho real [38].

Ignorando los movimientos particulares causados por irregularidades locales, los sistemas confinados gravitacionalmente se alejan unos de otros uniformemente. Si tres cuerpos forman un triángulo, en un tiempo posterior formarán un triángulo más grande conservando su estructura; el principio cosmológico indica que el movimiento debe ser homogéneo e isotrópico, por lo que ambos triangulos deben ser similares. En otras palabras, la distancia física l (más correctamente denominada "distancia propia") en un determinado tiempo t entre dos objetos incrementa de acuerdo a la relación

$$l(t) = l_0 \cdot a(t), \tag{3.2}$$

donde l_0 es una constante correspondiente a la distancia inicial y $a(t)$ es una función adimensional que incrementa con el tiempo conocida como factor de escala (o factor de expansión universal) y su valor en el tiempo inicial es $a(t_0) = 1$. El factor de escala mide cuánto crecen las separaciones en el espacio con el pasar del tiempo; por ejemplo, si entre los tiempos t_1 y t_2, el factor de escala duplica su valor, eso nos dice que el Universo se ha expandido en tamaño por un factor de dos, y nos tomará dos veces más tiempo llegar de un objeto a otro.

La derivada con respecto al tiempo de la ecuación (3.2) es la velocidad de recesión de un objeto con respecto al otro:

$$v = \dot{l} = l_0 \dot{a} = l \cdot \frac{\dot{a}}{a} = Hl, \tag{3.3}$$

que corresponde a la Ley de Hubble (Eq. (3.1)) y donde se ha definido

$$H(t) \equiv \frac{\dot{a}}{a} \tag{3.4}$$

como el *parámetro de Hubble* al tiempo t, entendiendo que el parámetro de Hubble en el tiempo inicial (o acutal) es H_0.

Como se mencionó brevemente al principio de esta sección, la recesión entre dos objetos astronómicos produce un corrimento al rojo en las líneas espectrales de uno observado con respecto al otro, el cual puede ser determinado por comparación de las líneas transicionales de niveles energéticos con las de una fuente estacionaria. Para velocidades de recesión pequeñas, el índice

de corrimiento al rojo z se obtiene con la relación del efecto Doppler a primer orden:

$$z \equiv \frac{\lambda_o - \lambda_e}{\lambda_e} = \frac{\lambda_o}{\lambda_e} - 1 = \sqrt{\frac{1 + v/c}{1 - v/c}} - 1 \approx \frac{v}{c} = \frac{Hl}{c}, \tag{3.5}$$

siendo λ_o la longitud de onda percibida por el observador y λ_e la longitud de onda emitida por la fuente [39]. Como z es proporcional a la distancia entre dos objetos, la ley de Hubble es una herramienta muy importante para calcular distancias a galaxias lejanas conociendo únicamente su corrimiento al rojo.

El corrimento al rojo se relaciona con el factor de escala de la siguiente manera:

$$1 + z = \frac{\lambda_o}{\lambda_e} = \frac{a(t_o)}{a(t_e)}, \tag{3.6}$$

donde $a(t_o)$ y $a(t_e)$ son los factores de escala cuando la radiación fue observada y emitida, respectivamente. Por convención, se establece que el factor de escala en el tiempo de observación (o tiempo actual) es igual a 1, por lo que la relación es equivalente a

$$a = \frac{1}{1 + z}, \tag{3.7}$$

donde a es el factor de escala en el tiempo de emisión de la radiación.

Desafortunadamente, el valor de H_0 no ha sido tan sencillo de determinar. Hasta finales del siglo XX sólo se sabía que estaba entre 50 y 100 km·s^{-1}·Mpc^{-1}. El problema radica en que medir distancias de manera precisa a objetos lejanos es muy difícil.

En 2001, el *Hubble Space Telescope Key Project* midió el valor usando variables cefeidas como marcadores de distancias. Las variables cefeidas son un tipo de estrella que pulsa regularmente cambiando su luminosidad y la relación entre el periodo pulsar con los cambios de luminosidad permite determinar su distancia a partir de la magnitud absoluta. El programa concluyó que el valor de la constante de Hubble era 72 km·s^{-1}·Mpc^{-1} y se ha refinado hasta 73.5 km·s^{-1}·Mpc^{-1} con un 2 % de error.

Sin embargo, más recientemente se adoptó un enfoque diferente basado en las observaciónes del satélite Planck sobre el fondo cósmico de microondas. Las mediciones del proyecto han permitido analizar cómo debió haber evolucionado el universo temprano y se encontró un valor para H_0 de 67.4 km·s^{-1}·Mpc^{-1}, lo cual difiere considerablemente con el resultado de las estrellas cefeidas. Se

cree que la diferencia no se debe a errores sistemáticos porque los valores siguen sin concordar conforme las mediciones se vuelven más precisas. Las pruebas posteriores no parecen favorecer ninguno de los dos resultados, pero se alínean más con los de Planck [40].

Esta discrepancia de valores se conoce como *la tensión de Hubble*. En el capítulo siguiente desarrollaremos un método basado en GPR donde se determina un valor para H_0 de manera indirecta utilizando datos de cronómetros cósmicos.

3.1.2. La ecuación de Friedmann

Se conoce como ecuación de Friedmann a la relación matemática que describe la expansión del universo, por lo que se trata de la ecuación más importante en cosmología. Por simplicidad, presentaremos la deducción Newtoniana de la ecuación, pero se puede obtener de manera relativista a partir de las ecuaciones de Einstein y su deducción detallada está disponible para su consulta en cualquier libro de Relatividad General, como [36].

Consideremos un observador en un medio con densidad de masa ρ que se expande uniformemente. Por el principio cosmológico, podemos elegir cualquier punto como su centro. Consideremos también un pequeño volumen que contiene una masa m, es decir, una partícula a una distancia r del centro.

De acuerdo al teorema del cascarón, la partícula experimenta una atracción gravitacional por la materia en radios más pequeños como si toda estuviera concentrada en el centro. La masa total al interior del cascarón de radio r está dada por $M = 4\pi r^3\rho/3$ y la fuerza gravitacional sobre la partícula es

$$F = \frac{GMm}{r^2} = \frac{4\pi G\rho r m}{3}. \tag{3.8}$$

Por otro lado, la energía potencial gravitatoria de la partícula es

$$V = -\frac{GMm}{r} = -\frac{4\pi G\rho r^2 m}{3} \tag{3.9}$$

y su energía cinética se escribe como

$$T = \frac{1}{2}m\dot{r}^2, \tag{3.10}$$

donde $\dot{r}$ denota la derivada respecto al tiempo de la posición.

Por la ley de conservación de energía, la suma de (3.9) y (3.10) debe ser igual a una constante

U:

$$U = T + V = \frac{1}{2}m\dot{r}^2 - \frac{4\pi G\rho r^2 m}{3}. \tag{3.11}$$

Dado que la distribución de materia es homogénea, se puede cambiar el sistema de coordenadas a uno que se mueva junto a la expansión, conocido como *coordenadas comóviles*. Como la expansión es uniforme, la relación entre la posición $\vec{r}$ de la partícula con respecto al centro y sus coordenadas comóviles $\vec{x}$ está dada por

$$\vec{r} = a(t)\vec{x}, \tag{3.12}$$

donde se ha usado la homogeneidad del universo para asegurar que el factor de escala $a(t)$ es sólo una función del tiempo y no del espacio. Los objetos permanecen en posiciones fijas en el sistema comóvil $\vec{x}$ y el sistema de posiciones reales $\vec{r}$ se conoce como *coordenadas físicas*.

Como las posiciones no cambian en el sistema comóvil, $\dot{x} = 0$ y $\dot{r} = \dot{a}x$, por lo que la ecuación (3.11) se escribe

$$U = \frac{1}{2}m\dot{a}^2x^2 - \frac{4\pi}{3}G\rho a^2x^2 m. \tag{3.13}$$

Multiplicando ambos lados por $2/ma^2x^2$ y simplificando se obtiene

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2}, \tag{3.14}$$

donde $k = -2U/mc^2x^2$ es una constante conocida como la *curvatura* del universo porque proporciona información sobre su geometría y c es la velocidad de la luz. La ecuación (3.14) es la forma estándar de la **ecuación de Friedmann**.

3.1.3. La ecuación de fluido

La ecuación de Friedmann no es suficiente para modelar el comportamiento del Universo sin una ecuación que describa cómo evoluciona la densidad de material ρ con el tiempo, la cual se conoce como **ecuación de fluido** e involucra la presión p de la materia.

De acuerdo a la primera ley de la termodinámica, la suma del cambio de energía interna del sistema dE y el trabajo realizado $pd\mathcal{V}$, donde $d\mathcal{V}$ es el cambio de volumen, es igual al calor producido TdS, donde T es la temperatura y dS el cambio de entropía del proceso, es decir

$$dE + pd\mathcal{V} = TdS, \tag{3.15}$$

considerando que $\mathcal{V}$ es un volumen obtenido a partir de un radio en coordenadas comóviles. De acuerdo a la ecuación (3.12), el radio físico es a, por lo que la energía, usando la ecuación de Einstein $E = mc^2$, es

$$E = \frac{4\pi}{3} a^3 \rho c^2. \tag{3.16}$$

El cambio instantáneo de energía se puede obtener diferenciando la ecuación anterior:

$$\frac{dE}{dt} = 4\pi a^2 \rho c^2 \dot{a} + \frac{4\pi}{3} a^3 \dot{\rho} c^2, \tag{3.17}$$

mientras que el cambio en volumen es

$$\frac{d\mathcal{V}}{dt} = \frac{d}{dt}\left(\frac{4\pi}{3} a^3\right) = 4\pi a^2 \dot{a}. \tag{3.18}$$

Si la expansión es un proceso reversible, $dS = 0$, entonces, sustituyendo (3.17) y (3.18) en (3.15) y simplificando, se tiene

$$\dot{\rho} + 3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right) = 0. \tag{3.19}$$

Esta ecuación, junto a la ecuación de Friedmann son todo lo que se necesita para describir la evolución del Universo.

3.1.4. La ecuación de aceleración

Existe también una ecuación que describe la aceleración del factor de escala. Derivando (3.14) con respecto al tiempo se obtiene

$$2\left(\frac{\dot{a}}{a}\right)\left(\frac{a\ddot{a} - \dot{a}^2}{a^2}\right) = \frac{8\pi G}{3}\dot{\rho} + 2\frac{kc^2 \dot{a}}{a^3}. \tag{3.20}$$

Sustituyendo la expresión para ρ que se obtiene de (3.19) y simplifcando, la ecuación se transforma en

$$\frac{\ddot{a}}{a} - \left(\frac{\dot{a}}{a}\right)^2 = -4\pi G\left(\rho + \frac{p}{c^2}\right) + \frac{kc^2}{a^2}. \tag{3.21}$$

Utilizando nuevamente (3.14), se llega a la **ecuación de aceleración**

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right). \tag{3.22}$$

Esta relación se conoce también como segunda ecuación de Friedmann. Nótese que si la materia

en el Universo tiene alguna presión, esta provoca una desaceleración en la expansión.

3.1.5. La constante cosmológica

Durante el desarrollo de la Relatividad General, Einstein creía que el Universo era estático, pero encontró que la teoría no lo permitía porque toda la matería se atrae gravitacionalmente. Para solucionar este problema, introdujo en 1917 un valor conocido como la *constante cosmológica* que compensara el colapso gravitacional. Este valor se denota con la letra griega Λ y es uno de los objetos más intrigantes en cosmología actualmente.

Su efecto en la ecuación de Friedmann es la aparición de un término extra

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3}. \tag{3.23}$$

Las unidades de Λ, en este caso, son [longitud]$^{-2}$ y su valor puede ser tanto positivo como negativo, pero se considera más frecuentemente un valor positivo.

La consecuencia de la introducción de la constante cosmológica Λ puede verse más claramente en la ecuación de aceleración, la cual se transforma en

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda c^2}{3}. \tag{3.24}$$

Aquí, si $\Lambda > 0$, entonces actúa como una fuerza repulsiva que aumentaría la aceleración de la expansión, llegando a contrarrestar la atracción gravitatoria debida al primer término del lado derecho si Λ es lo suficientemente grande.

También es útil describir la constante cosmológica como si fuera un fluido con densidad de energía ρ_Λ y presión p_Λ. Definiendo

$$\rho_\Lambda \equiv \frac{\Lambda c^2}{8\pi G}, \tag{3.25}$$

la ecucación (3.23) se escribe como

$$H^2 = \frac{8\pi G}{3}(\rho + \rho_\Lambda) - \frac{kc^2}{a^2}. \tag{3.26}$$

La ecuación (3.19) para este fluido es

$$\dot{\rho}_\Lambda + 3\frac{\dot{a}}{a}\left(\rho_\Lambda + \frac{p_\Lambda}{c^2}\right) = 0 \tag{3.27}$$

y, como ρ_Λ es constante por definición, se tiene que

$$p_\Lambda = -\rho_\Lambda c^2. \tag{3.28}$$

Entonces, la constante cosmológica tiene una presión negativa, lo que significa que, a medida que el Universo se expande, se realiza trabajo sobre el fluido de Λ. Esto permite que su densidad de energía permanezca constante aunque el volumen del Universo incremente.

Se interpreta a Λ como la densidad de energía del espacio vacío, llamada **energía oscura** debido al desconocimiento que se tiene de ella, la cual permanece incluso si no hay partículas presentes. Podría ser que la constante cosmológica se tratara solamente de un fenómeno transitorio que desaparezca en el futuro. Otra posibilidad teórica, conocida como *quintaesencia*, es que Λ no sea una constante perfecta, sino que exhiba variaciones muy lentas, proponiendo una ecuación de estado del tipo

$$p = w\rho c^2, \tag{3.29}$$

donde w es una constante. En unidades naturales, donde $c^2 = 1$, se simplifica a

$$w = \frac{p}{\rho}. \tag{3.30}$$

El caso $w = -1$ corresponde a Λ constante, aunque es posible una expansión más general acelerada siempre que $w < -1/3$. La ecuación (3.30) se conoce como *ecuación de estado de la energía oscura*.

3.2. Parámetros cosmológicos

Existen varios modelos que se obtienen al resolver las ecuaciones de Friedmann y de fluido, por lo que es usual caracterizar cada uno con ciertos parámetros para decidir posteriormente cuál versión describe mejor a nuestro universo.

3.2.1. Parámetros de densidad

Combinando las ecuaciones (3.4) y (3.14) se tiene

$$H^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2}. \tag{3.31}$$

Es importante notar que, para un valor fijo de H, existe un valor de densidad que hace que la curvatura del Universo sea cero $k = 0$, es decir, que sea plano. Dicho valor se conoce como densidad crítica ρ_c y está dada por

$$\rho_c(t) = \frac{3H^2}{8\pi G}. \tag{3.32}$$

La cantidad ρ_c no es necesariamente la densidad real del Universo, ya que éste no es necesariamente plano. Debido a esto, es útil comparar la densidad real con la densidad crítica definiendo una cantidad adimensional llamada *parámetro de densidad*

$$\Omega(t) \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}, \tag{3.33}$$

donde ρ es la densidad real del Universo. El parámetro de densidad en el tiempo actual se denota Ω_0 y, en un universo plano, $\Omega = 1$.

De la misma forma en que es útil expresar la densidad comparada con la densidad crítica, es conveniente definir un *parámetro de densidad de energía oscura* como

$$\Omega_\Lambda \equiv \frac{\Lambda c^2}{3H^2}. \tag{3.34}$$

Utilizando Ω y Ω_Λ en la ecuación (3.23), se obtiene

$$\Omega + \Omega_\Lambda - 1 = \frac{kc^2}{a^2H^2}, \tag{3.35}$$

por lo que, en este caso, la condición de planitud se escribe

$$\Omega + \Omega_\Lambda = 1. \tag{3.36}$$

Se define también el *parámetro de densidad de curvatura* como

$$\Omega_k \equiv -\frac{kc^2}{a^2H^2}, \tag{3.37}$$

el cual puede ser positivo o negativo y permite escribir la ecuación (3.31) en términos de los

parámetros de densidad [41]

$$\Omega + \Omega_k = 1 \tag{3.38}$$

Además de constante cosmológica, también tenemos radiación en el Universo. Tanto los fotones como cualquier partícula relativista pueden contribuir a la cantidad de radiación. La mayor parte de la densidad de energía para fotones proviene del CMB y el único candidato claro de partícula que aporta a la radiación es el neutrino.

En la mayoría de modelos cosmológicos se considera a los neutrinos como partículas sin masa, pero no está tan claro que esto sea así. Si se descubriera que los neutrinos tienen masa lo suficientemente grande, se habrían vuelto no relativistas hoy y contribuirían más a la componente de materia que a la de radiación. Además, es posible que existan partículas relativistas aún no detectadas que contribuyan a la radiación [42].

De cualquier manera, se definen el *parámetro de densidad de materia* Ω_m y el *parámetro de densidad de radiación* Ω_r como

$$\Omega_m \equiv \frac{8\pi G\rho_m}{3H^2} \quad \text{y} \quad \Omega_r \equiv \frac{8\pi G\rho_r}{3H^2}, \tag{3.39}$$

donde ρ_m y ρ_r son las densidades de materia y radiación del Universo respectivamente.

La ecuación de estado de la radiación, incluyendo neutrinos y otras partículas que se volvieron no relativistas con el tiempo es $w = 1/3$.

La ecuación de Friedmann en términos de los parámetros cosmológicos

Combinando (3.19) en unidades naturales (con $c^2 = 1$) y (3.30), se obtiene

$$\dot{\rho} + 3\rho\left(\frac{\dot{a}}{a}\right)(1 + w) = 0. \tag{3.40}$$

Multiplicando ambos lados por $a^3/\dot{a}$:

$$\frac{\dot{\rho}}{\dot{a}}\left(a^3\right) + 3\rho a^2\,(1 + w) = 0. \tag{3.41}$$

Desarrollando y reordenando los términos se tiene

$$\frac{\dot{\rho}}{\dot{a}}\left(a^3\right) + 3\rho a^2 = -3w\rho a^2. \tag{3.42}$$

Nótese, usando la regla de la cadena, que el lado izquierdo de la ecuación corresponde a la derivada de ρa^3 con respecto al factor de escala a, entonces

$$\frac{d(\rho a^3)}{da} = -3w\rho a^2. \tag{3.43}$$

La anterior es una ecuación diferencial cuya solución es

$$\rho \propto a^{-3(w+1)}, \tag{3.44}$$

la cual describe la evolución de la densidad de energía del Universo. Para materia no relativista, la presión es pequeña comparada con su densidad, por lo que $w_m \approx 0$ y la ecuación (3.44) nos da $\rho_m \propto a^{-3}$. Por otro lado, para la radiación, $w_r = 1/3$, entonces $\rho_r \propto a^{-4}$ y, para el caso Λ constante, la ecuación de estado de energía oscura es $w_\Lambda = -1$, por lo tanto $\rho_\Lambda \propto a^0$.

Separando la densidad total del Universo en radiación y materia, tal que $\rho = \rho_r + \rho_m$, la ecuación (3.23) se reescribe como

$$H^2 = \frac{8\pi G}{3}(\rho_r + \rho_m) - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3}. \tag{3.45}$$

De este modo, podemos escribir una contribución fraccional para cada componente del Universo en el tiempo actual

$$H^2 = \frac{8\pi G}{3}\left(\frac{\rho_{r,0}}{a^4} + \frac{\rho_{m,0}}{a^3}\right) - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3}. \tag{3.46}$$

Utilizando las definiciones de parámetros de densidad en el tiempo actual, se tiene

$$H^2 = H_0^2\left(\frac{\Omega_{r,0}}{a^4} + \frac{\Omega_{m,0}}{a^3} + \frac{\Omega_{k,0}}{a^2} + \Omega_{\Lambda,0}\right). \tag{3.47}$$

En términos del corrimiento al rojo,

$$H^2 = H_0^2[\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + \Omega_{k,0}(1+z)^2 + \Omega_{\Lambda,0}]. \tag{3.48}$$

Además, de la ecuación (3.45) en el tiempo actual, se obtiene

$$1 = \frac{8\pi G\rho_{r,0}}{3H_0^2} + \frac{8\pi G\rho_{m,0}}{3H_0^2} - \frac{kc^2}{a^2H_0^2} + \frac{\Lambda c^2}{3H_0^2}, \tag{3.49}$$

lo que significa que los parámetros están relacionados como [43]

$$\Omega_{r,0} + \Omega_{m,0} + \Omega_{k,0} + \Omega_{\Lambda,0} = 1 \tag{3.50}$$

3.2.2. Parámetro de desaceleración

Como se ha descubierto, no sólo el tamaño del Universo crece, también la velocidad con la que lo hace. El *parámetro de desaceleración* es una forma de cuantificar este fenómeno. Recibió ese nombre por razones históricas, ya que se esperaba que el colapso gravitacional frenara la expansión del Universo, pero posteriormente se descubrió que no es así.

Desarrollando una expansión de Taylor del factor de escala $a(t)$ alrededor del tiempo actual t_0 se tiene

$$a(t) = a(t_0) + \dot{a}(t_0)(t-t_0) + \frac{1}{2}\ddot{a}(t_0)(t-t_0)^2 + \dots \tag{3.51}$$

y, dividiendo por $a(t_0)$,

$$\frac{a(t)}{a(t_0)} = 1 + H_0(t-t_0) - \frac{q_0}{2}H_0^2(t-t_0)^2 + \dots, \tag{3.52}$$

donde se definió el parámetro de desaceleración en el tiempo actual q_0 como

$$q_0 \equiv -\frac{\ddot{a}(t_0)}{a(t_0}\frac{1}{H_0^2} = -\frac{a(t_0)\ddot{a}(t_0)}{\dot{a}(t_0)^2} \tag{3.53}$$

y, extendiendo la definición a un tiempo posterior,

$$q \equiv -\frac{a\ddot{a}}{\dot{a}^2}. \tag{3.54}$$

Se dice que la expansión del Universo se está acelerando si $q < 0$ y cuanto mayor sea el valor

absoluto de q, más rápida es la aceleración [41].

Derivando la ecuación (3.4) con respecto al tiempo se puede encontrar la relación entre el parámetro de desaceleración y el de Hubble

$$\frac{\dot{H}}{H^2} = -(1+q). \tag{3.55}$$

3.2.3. El modelo ΛCDM

Sabemos que un componente importante del Universo es la materia, la cual puede ser cualquier colección de partículas no relativistas; sin embargo, la materia que se descubre por su inflencia gravitacional puede no ser del mismo tipo de aquella que conocemos en la Tierra. Llamamos *materia ordinaria*[2] a todo lo que está compuesto de átomos o sus partículas constituyentes.

Hasta hace aproximadamente 30 años, se creía que el Universo contenía únicamente materia bariónica, pero existe evidencia reciente de que puede existir otro tipo de materia. Este tipo de materia no ordinaria se conoce como *materia oscura*.

Una de las pocas cosas que conocemos acerca de la materia oscura es que debe ser fría; debe ser no relativista y debió serlo desde hace mucho tiempo. Si la materia oscura fuera caliente, habría fluido libremente desde regiones excesivamente densas, impidiendo la formación de galaxias. Otra cosa que se conoce sobre la materia oscura fría (*Cold Dark Matter* o CDM) es que interactúa muy débilmente con la materia ordinaria por todo el tiempo que se ha retrasado su detección [42].

Una de las pruebas más claras de la existencia de la materia oscura es que la masa de algunas galaxias, calculada a partir de su velocidad de rotación, resulta casi 10 veces mayor que la que se puede asociar a las estrellas, gas y polvo que las conforman.

El parámetro de densidad de materia puede ser dividido así en su componente bariónica Ω_b y de materia oscura Ω_c, de modo que

$$\Omega_m = \Omega_b + \Omega_c. \tag{3.56}$$

Se conoce como ΛCDM al modelo cosmológico basado en energía oscura y materia oscura fría. También es referido a veces como el modelo cosmológico estándar por ser el modelo más simple que explica varios fenómenos en cosmología, como la expansión acelerada del Universo. En él se asume que la ecuación de estado de la energía oscura es una constante cosmológica $w = -1$

[2] También conocida como materia bariónica, porque los bariones aportan la mayor parte de masa comparados con otras partículas como electrones.

y que el Universo es plano, es decir, $\Omega_k = 0$.

El modelo queda totalmente determinado por seis parámetros principales y de ellos se obtienen algunos derivados. Los parámetros principales no son necesariamente los mismos, pero se suele considerar a H_0, Ω_b, Ω_m, el camino óptico hasta la reionización τ, la amplitud de fluctuación escalar A_s y el índice espectral escalar n_s[3]. Dado que el valor de Ω_r es muy pequeño, se suele despreciar en este modelo, dejando a Ω_Λ como un parámetro derivado.

El satélite Planck de la Agencia Espacial Europea, encargado de determinar los valores de parámetros cosmológicos analizando el CMB, encontró en 2018 [44] que $\Omega_b \approx 0.04$, $\Omega_c \approx 0.26$ y $\Omega_\Lambda \approx 0.679$, por lo que aproximadamente 67 % del Universo es energía oscura, 26 % es materia oscura y solamente alrededor de 4 % es materia bariónica.

3.3. Medición de distancias en el Universo

Se le llama *evento* a cualquier punto en el espacio-tiempo del Universo y, como hemos podido ver hasta ahora, es muy útil describir las distancias entre eventos. Sin embargo, medir la distancia real es muy difícil físicamente, es por esto que se definen diferentes formas de medir distancias a escalas cosmológicas y se utilizan las que un determinado problema requiera. A continuación revisaremos algunas de ellas.

Distancia y tiempo de Hubble

De acuerdo a la ecuación (3.1), la constante H_0 debe tener unidades de [tiempo]$^{-1}$. Por lo que el recíproco de H_0 es un tiempo, conocido como *tiempo de Hubble*

$$t_H \equiv \frac{1}{H_0}. \tag{3.57}$$

Representa la edad del Universo si su expansión fuera lineal.

Se define la *distancia* o *longitud de Hubble* como la velocidad de la luz multiplicada por el tiempo de Hubble

$$d_H \equiv ct_H = \frac{c}{H_0}. \tag{3.58}$$

Se trata de la distancia a un objeto basada solamente en la expansión del Universo. Frecuentemente, se usan unidades geométricas, donde $c = t_H = d_H = 1$, para simplificar los cálculos.

[3] Para el trabajo que se desarrollará más adelante, sólo será necesario conocer los primeros 3 parámteros.

Distancia comóvil

A partir de la ecuación (3.48), se define el *parámetro de Hubble adimensional* como

$$E(z) \equiv \frac{H(z)}{H_0} = \sqrt{\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + \Omega_{k,0}(1+z)^2 + \Omega_{\Lambda,0}}. \tag{3.59}$$

Utilizando $E(z)$ se puede escribir al parámetro de Hubble en función del corrimiento al rojo

$$H(z) = H_0 E(z), \tag{3.60}$$

la cual es una forma muy común de trabajar con el parámetro H en cosmología.

La razón $dz/E(z)$ es proporcional al tiempo de viaje de un fotón a lo largo del intervalo de corrimiento al rojo dz dividido por el factor de escala en ese tiempo. Como el fotón se mueve a la velocidad de la luz, que es constante, esta es su distancia propia dividida por el factor de escala, la cual es la definición de *distancia comóvil*. Integrando a lo largo de todo el recorrido, se tiene

$$d_C \equiv d_H \int_0^z \frac{dz'}{E(z)} = c \int_0^z \frac{dz'}{H(z)}. \tag{3.61}$$

Como vimos al momento de derivar la ecuación de Friedmann, la distancia en coordenadas comóviles mide la separación entre dos puntos del Universo en las condiciones del tiempo actual, por lo que su valor no cambia con el tiempo debido a la expansión. d_C es un tipo de distancia fundamental, porque otras distancias están relacionadas con ella.

Distancia comóvil transversa

La distancia comóvil entre dos objetos a la misma separación de un observador, pero cuya posición difiere sólo por un ángulo pequeño $d\theta$, está dada por $d_M d\theta$, donde d_M es la *distancia comóvil transversa* y se relaciona con la separación sobre la línea de observación de la siguiente manera:

$$d_M = \begin{cases} \frac{d_H}{\sqrt{\Omega_k}} \sinh\left(\frac{\sqrt{\Omega_k} d_C(z)}{d_H}\right), & \Omega_k > 0 \\ d_C(z), & \Omega_k = 0 \\ \frac{d_H}{\sqrt{-\Omega_k}} \sinh\left(\frac{\sqrt{-\Omega_k} d_C(z)}{d_H}\right), & \Omega_k < 0. \end{cases} \tag{3.62}$$

Las funciones trigonométricas sin y sinh surgen de tomar en cuenta la curvatura del Universo, determinada por el parámetro Ω_k.

Distancia de luminosidad

Se define la *distancia de luminosidad* como la relación entre el flujo F y la luminosidad L, integrados sobre todas las frecuencias, de una fuente

$$d_L \equiv \sqrt{\frac{L}{4\pi F}}. \tag{3.63}$$

Se relaciona con la distancia comóvil transversa mediante la ecuación

$$d_L = (1+z)d_M \tag{3.64}$$

y se usa frecuentemente para relacionar la luminosidad de un objeto con su distancia, ya que se puede calcular midiendo el flujo de un objeto con luminosidad fija y conocida. Dichos objetos se conocen como *standard candles* y son muy importantes para determinar distancias cosmológicas desde la Tierra [45].

Además de las anteriores, se define también la *distancia comóvil normalizada* como

$$D(z) = \left(\frac{H_0}{c}\right)\left(\frac{1}{1+z}\right) d_L(z), \tag{3.65}$$

la cual se reduce a

$$D(z) = \int_0^z \frac{H_0}{H(z')}dz' = \int_0^z \frac{dz'}{E(z')} \tag{3.66}$$

para el caso de un universo plano $\Omega_k = 0$. Por lo tanto, se puede encontrar fácilmente que la derivada de la distancia comóvil normalizada es

$$D'(z) = \frac{H_0}{H(z)}. \tag{3.67}$$

Diferenciando la ecuación anterior y combinando las derivadas de $D(z)$ con (3.55), se encuentra una relación para el parámetro de desaceleración en términos del corrimiento al rojo.

$$q(z) = -\frac{D'(z)}{D''(z)}(1+z) - 1. \tag{3.68}$$

3.4. Cronómetros cósmicos

Para solucionar la tensión de Hubble vista anteriormente, es muy útil hacer uso de diferentes fenómenos físicos para el cálculo de H_0, con el fin de compararlos con aquellos que generan el conflicto.

En la Figura 3.1, se muestran valores de H_0 calculados en diferentes proyectos actualizados a enero del 2022, utilizando técnicas como la determinación de escaleras de distancias cósmicas (HST Key Project, Cepheids+SNIa y TRGB Dist Ladder), mediciones indirectas sobre el CMB (WMAP9++ y Planck_PR3++), oscilaciones acústicas de bariones (BAO+D/H), el efecto térmico Sunyaev–Zeldovich (CHANDRA+tSZ, XMM+Planck tsZ), lentes gravitacionales fuertes (Gravlens Time Delay) y ondas gravitacionales (LIGO+Virgo+KAGRA) [46]. Salta a la vista el hecho

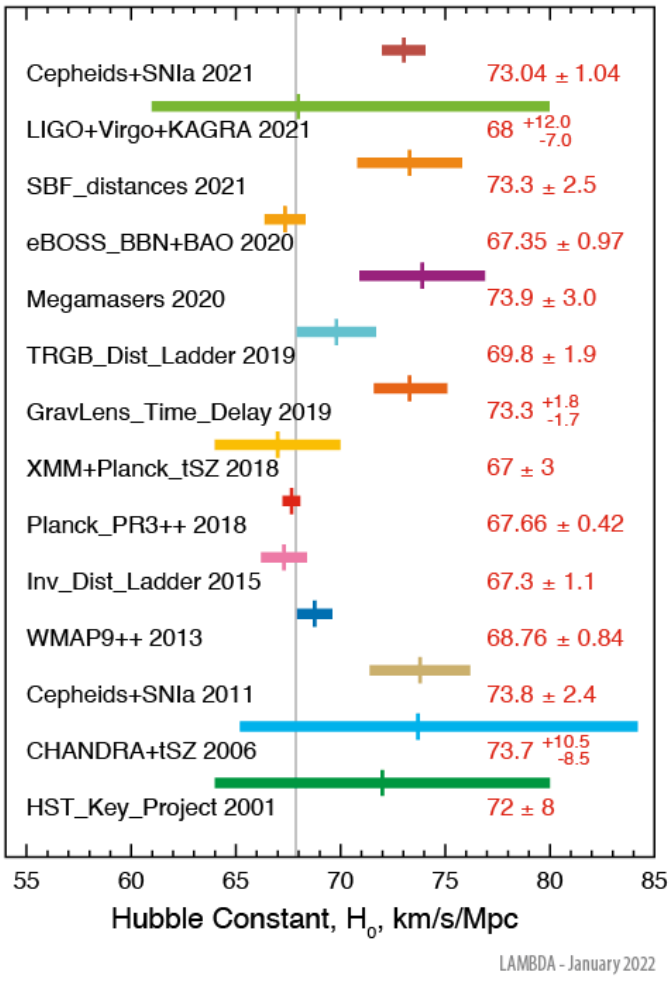

Figura 3.1: Valores de H_0 determinados usando diferentes técnicas.
(Imagen extraida de [46]).

de que los resultados se concentran en dos valores diferentes, uno alrededor de 68 $\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}$ y otro alrededor de 73 $\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}$. Además, algunas mediciones tienen intervalos de error tan bajos, que la discrepancia difícilmente podría asociarse a problemas con la medición.

Recientemente, los métodos que hacen uso de ML se han popularizado como herramienta para resolver este conflicto por las ventajas que se mencionaron en el primer capítulo de este texto. Entre ellos, se encuentra la reconstrucción de H como una función del corrimiento al rojo y evaluándola para $z = 0$, para lo cual se necesitan algunos datos de $H(z)$ en diferentes corrimientos al rojo.

Una manera de obtener datos de $H(z)$ es mediante la observación de cronómetros cósmicos (CC). Los CC son objetos astronómicos de los que se conoce muy bien las características que presentan en diferentes etapas de su evolución temporal. A partir del corrimiento al rojo, podemos determinar la distancia a la que varios de estos objetos se encuentran de nosotros y, analizando las diferencias entre sus estados de evolución, se puede deducir el tiempo que ha transcurrido entre los diferentes valores de z que tienen. El intervalo de tiempo nos dice cuán rápido se ha expandido el Universo entre los corrimientos al rojo de los objetos, lo cual es precisamente $H(z)$. Esta manera de determinar el parámetro de Hubble se conoce como *differential age method* o método de etapas diferenciales.

Generalmente, se usa como CC a galaxias que evolucionan pasivamente, las cuales han agotado casi toda su materia en forma de gas y forman muy pocas estrellas nuevas. Como las estrellas azules mueren más rápido que las rojas, estas galaxias se vuelven más rojas al pasar el tiempo, con lo que se puede deducir desde cuando se detuvo su formación de estrellas. Comparando dos galaxias que se formaron al mismo tiempo, pero que están en diferentes corrimientos al rojo, se encuentra el tiempo que ha pasado entre sus valores de z [47].

Se han publicado varios artículos donde se obtiene $H(z)$ en diferentes corrimientos al rojo, pero muchos autores han recopilado un conjunto de 31 valores en sus artículos propios, como [48] y [49]. Estos datos son muy útiles en la reconstrucción de modelos no paramétricos y se hará uso de ellos en el siguiente capítulo.

Capítulo 4

Reconstrucción de funciones cosmológicas

Una vez conocidos los fundamentos de la regresión con procesos gaussianos y la teoría cosmológica, comenzamos con la aplicación de GPR en la reconstrucción de funciones de interés. Entre los objetivos principales de este trabajo se encuentran: probar la viabilidad de la GPR en un problema real y obtener información física relevante a partir del modelo reconstruido para, posteriormente, contrastarlo con resultados de otros autores que usan ya sea el mismo método o algunos diferentes.

4.1. Parámetro de Hubble

Sustituyendo los valores de parámetros de densidad y de Hubble en el tiempo actual calculados a partir de las mediciones del satélite Planck [44] y $\Omega_k = 0$ en la ecuación (3.59), se obtiene una expresión para el parámetro de Hubble como función del corrimiento al rojo en el modelo ΛCDM. Se utilizaron los parámetros de Planck, pues se asumió un universo tipo ΛCDM para su obtención.

Buscamos comparar esta función con el modelo de predicciones que se encuentra usando GPR con las observables de $H(z)$ a partir de cronómetros cósmicos [47]. Además, evaluar el modelo en $z = 0$ para determinar el ratio de expansión del Universo en el tiempo actual.

Utilizando un procedimiento similar al de la reconstrucción de una línea recta cuando las varianzas son no nulas, discutido en el capítulo 2, se construyen modelos predictivos con *priors* de diferentes kernels. Se encontró que el mejor de ellos es nuevamente el que usa un kernel de tipo

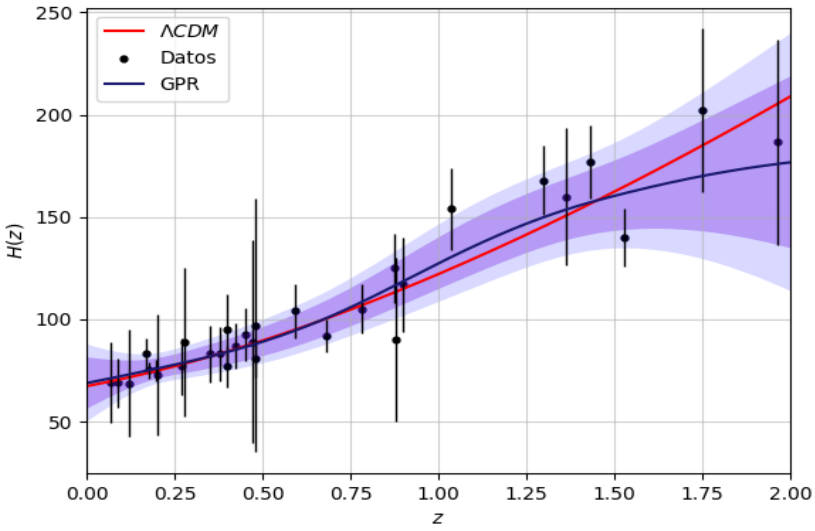

Figura 4.1: Reconstrucción del parámetro de Hubble con cronómetros cósmicos.

Matern.

El parámetro de Hubble como función de z en ΛCDM, el modelo de regresión y las observables de cronómetros cósmicos se presentan en la Figura 4.1. Es interesante que el modelo es consistente con ΛCDM, especialmente en valores de z cercanos a cero.

La predicción para H_0 dada por la regresión y su desviación estándar es

$$H_0 = 68.798 \pm 6.340 \text{ km} \cdot s^{-1} \cdot \text{Mpc}^{-1}, \tag{4.1}$$

lo cual coincide con las observaciones del CMB por Planck y difiere del resultado calculado con variables cefeidas.

Hay que destacar que las mediciones de $H(z)$ no dependen de ningún modelo cosmológico, es decir, no se asumen características del Universo, como su curvatura. Por otro lado, las mediciones del CMB sí lo hacen y aún así, los modelos son muy semejantes, lo cual podría también confirmar el hecho de que el Universo es aproximadamente plano, por lo menos en el tiempo actual.

Además, la GPR es un procedimiento no paramétrico (no es necesario asumir una relación específica entre variables) y aprende únicamente de las observables, por lo que mantiene su fidelidad a los datos. Entonces, la GPR sobre datos de cronómetros cósmicos tiene la ventaja de permitir obtener información directa de $H(z)$, comparada con otras técnicas.

4.2. Derivadas de la distancia comóvil normalizada

Como sugiere la ecuación (3.67), en un universo plano, existe una expresión sencilla para la derivada de la distancia comóvil normalizada en términos del corrimiento al rojo que depende del parámetro de Hubble. De esta manera, es posible encontrar a $D'(z)$ en el modelo ΛCDM con los parámetros adecuados.

A partir de los datos de $H(z)$ para las mediciones de cronómetros cósmicos y del valor predicho para H_0 en (4.1), se pueden calcular observables para $D'(z)$. Las varianzas de estos nuevos datos se pueden aproximar usando la relación (2.26). Estos valores serán las herramientas de entrenamiento para un modelo de predicciones con GPR que proporcione el valor de $D'(z)$ para diferentes corrimientos al rojo.

En la Figura (4.2), se presentan las observables obtenidas y sus varianzas aproximadas, el modelo de regresión (utilizando un *prior* con kernel Matern) y la curva de $D'(z)$ para ΛCDM. Como en el caso del parámetro de Hubble, el modelo parece coincidir con ΛCDM a lo largo de todo el intervalo $0 \leq z \leq 2$.

Sustituyendo (3.59) en (3.67) y diferenciando con respecto a z, se encuentra una expresión analítica para la segunda derivada de la distancia comóvil normalizada en el modelo ΛCDM usando los parámetros de densidad correspondientes.

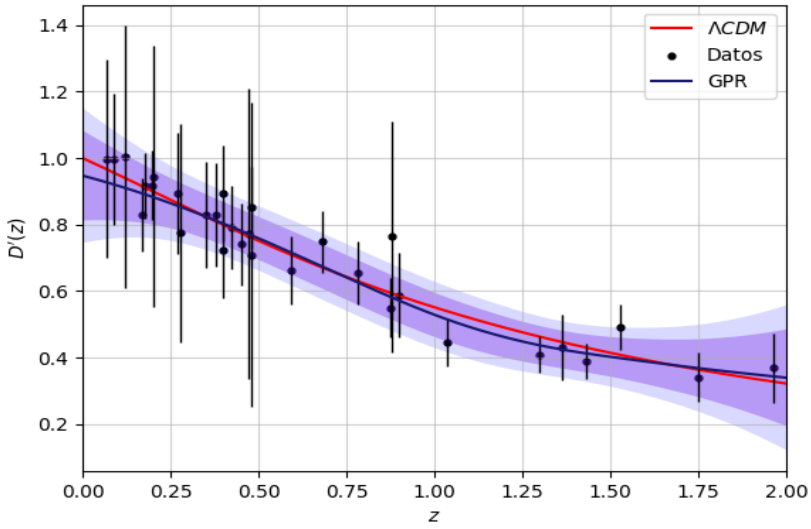

Figura 4.2: Reconstrucción de la primera derivada de la distancia comóvil normalizada.

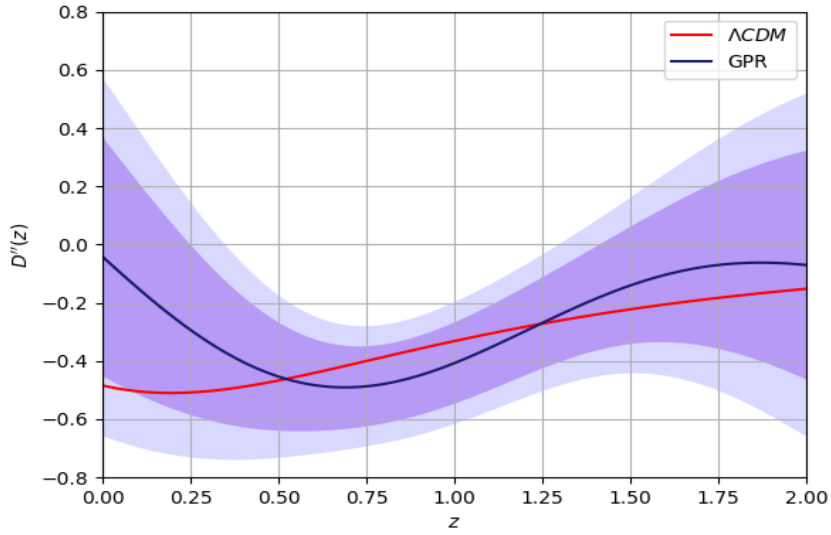

Figura 4.3: Reconstrucción de la segunda derivada de la distancia comóvil normalizada.

Por otro lado, a partir de las observables de $D'(z)$, es posible reconstruir la derivada de esta función, lo que corresponde a $D''(z)$, de la misma forma en que se encontró la derivada de una relación sinusoidal en el capítulo 2. Para realizar el proceso no es necesario generar datos directos de $D''(z)$, por lo que la Figura (4.3) muestra solamente el modelo de regresión y la función en ΛCDM.

En este caso, la curva en ΛCDM no es tan similar al promedio de las predicciones, como en las reconstrucciones anteriores, sin embargo, permanece en el intervalo delimitado por 2σ en la mayoría de valores de $0 \leq z \leq 2$.

Desafortunadamente, no es posible saber con certeza si la diferencia se debe al proceso de ajuste o a que las mediciones de cronómetros cósmicos resultan en discrepancias con ΛCDM para $D''(z)$, pero las pruebas sobre otras funciones (como la Figura 2.5) indican que el modelo de la derivada suele estar muy cercano a la derivada analítica. Por lo tanto, la experiencia apunta a que las mediciones no concuerdan exactamente con ΛCDM.

4.3. Parámetro de desaceleración

Las derivadas de la distancia comóvil normalizada no son lo realmente importante al momento de interpretar físicamente los resultados. Estas funciones se usan de manera auxiliar para encon-

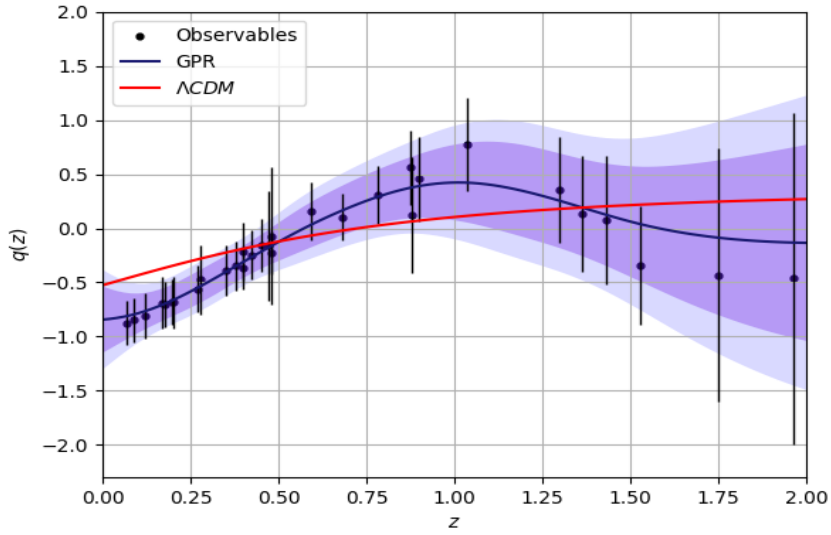

Figura 4.4: Reconstrucción del parámetro de desaceleración.

trar, mediante la ecuación (3.68), al parámetro de desaceleración en función del corrimiento al rojo, el cual sí aporta información sobre los cambios en la tasa de expansión del Universo.

Como no existen datos observables para $D''(z)$, es necesario evaluar el modelo de regresión en los mismos valores de corrimiento al rojo que los datos de $D'(z)$ para construir observables de $q(z)$ con los promedios y varianzas de esas evaluaciones y los datos que ya se tenían para $D'(z)$. Por otro lado, las varianzas de los datos de q se aproximan con la relación (2.26). En la Figura 4.4 se presentan las observables cosntruidas para $q(z)$, el modelo de regresión y su comparación con ΛCDM.

En el modelo se puede observar que el parámetro de desaceleración en el tiempo actual $q_0 = q(z = 0)$ es negativo, lo cual confirma que la expansión del Universo se está acelerando en la actualidad. Sin embargo, en corrimientos al rojo posteriores (en el pasado), el parámetro es positivo, por lo que la aceleración en la expansión es algo reciente en eras cosmológicas.

Como en el caso anterior, es difícil determinar si la diferencia entre la regresión y ΛCDM se debe al proceso de ajuste o a discrepancias entre ΛCDM y las mediciones, pero las pruebas inclinan a concluir la segunda opción.

Si las observables hubieran sido obtenidas de un universo representado por ΛCDM, el modelo de $q(z)$ se vería más como una distribución de probabilidad para cada valor de corrimiento al rojo, diferente a la curva roja en la Figura 4.4. Para comparar adecuadamente la regresión con ΛCDM,

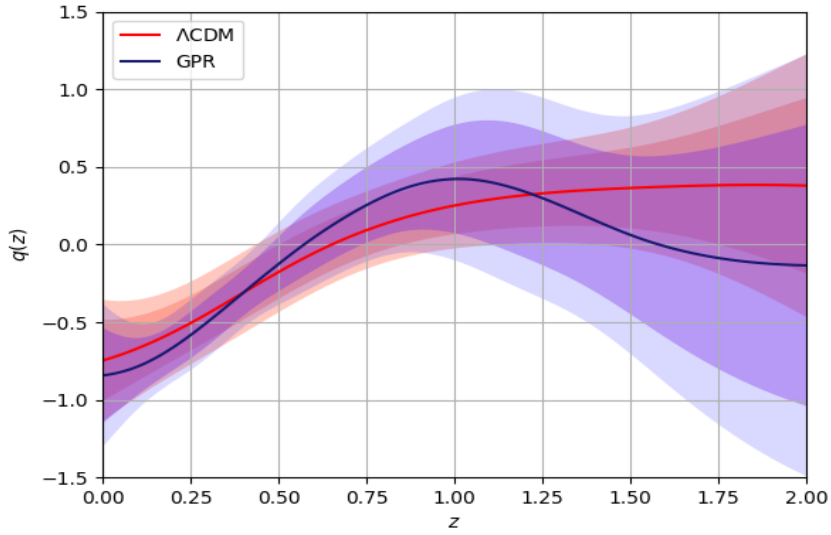

Figura 4.5: Comparación del modelo GPR para $q(z)$ con un conjunto de datos simulados basados en ΛCDM.

se construyó un conjunto de datos observables simulados dispersos alrededor de los valores de H que predice ΛCDM y se repitieron todas las reconstrucciones explicadas a lo largo de este capítulo hasta obtener un modelo de regresión para q que sea producido únicamente por valores teóricos del modelo cosmológico (no experimentales).

El resultado y su comparación con la reconstrucción GPR experimental con cronómetros cósmicos se muestran en la Figura 4.5. En este caso, se observa más claramente cómo ΛCDM representa mejor a las mediciones de cronómetros cósmicos que en la Figura 4.4, por lo menos en corrimientos al rojo cercanos a cero. A pesar de que la curva roja y la azul no coinciden exactamente, las zonas de confianza delimitadas por 2σ se superponen en casi todo el intervalo $0 \leq z \leq 2$.

4.4. Ecuación de estado de la energía oscura

Para reconstruir la EoS de la energía oscura, se han usado diferentes técnicas tanto paramétricas [50, 51, 52] como no-paramétricas [53, 54]. En este trabajo, utilizamos la GPR como verificación de otro método independiente del modelo cosmológico.

En [55], se utilizan algoritmos de interpolación lineal y de bines sobre parametrizaciones de la EoS para encontrar los valores que mejor se ajustan a observaciones del parámetro de Hubble

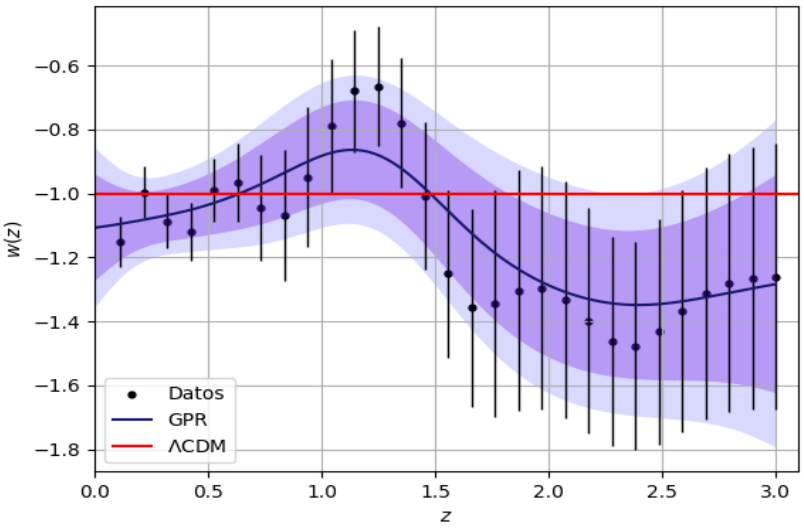

Figura 4.6: Reconstrucción de la energía oscura.

provenientes de Supernovas Tipo Ia y Oscilaciones Acústicas de Bariones.

De estos modelos se obtienen datos simulados para $w(z)$ en un cierto intervalo de z, que se pueden usar como conjunto de entrenamiento para una GPR. Si ambos modelos son similares, podemos decir que son consistentes.

Los datos de entrenamiento están disponibles en [35] y su represetnación gráfica, así como el modelo GPR con un *prior* de kernel Matern y su comparación con la EoS de energía oscura en ΛCDM, dada por $w = -1$, se presentan en la Figura 4.6.

Se puede observar que la EoS para ΛCDM se encuentra en la zona de confianza 2σ a lo largo de todo el intervalo. Sin embargo, la forma de la curva de regresión sugiere que una ecuación de estado constante $w = -1$ podría ser demasiado reduccionista a la hora de representar mediciones reales, pues no proporciona información detallada.

Capítulo 5

Conclusiones

5.1. Sobre la GPR

En la Figura 5.1 presentamos los valores de χ^2 para los modelos obtenidos a partir de *priors* con diferentes kernel para la reconstrucción del parámetro de Hubble y la EoS de la energía oscura.

Se encontró que el mejor modelo para el parámetro de Hubble es el obtenido a partir de un kernel Matern y para energía oscura, el de un kernel racional cuadrático. Por lo tanto, esas funciones de covarianza se eligieron para los modelos presentados en (4.1) y (4.6).

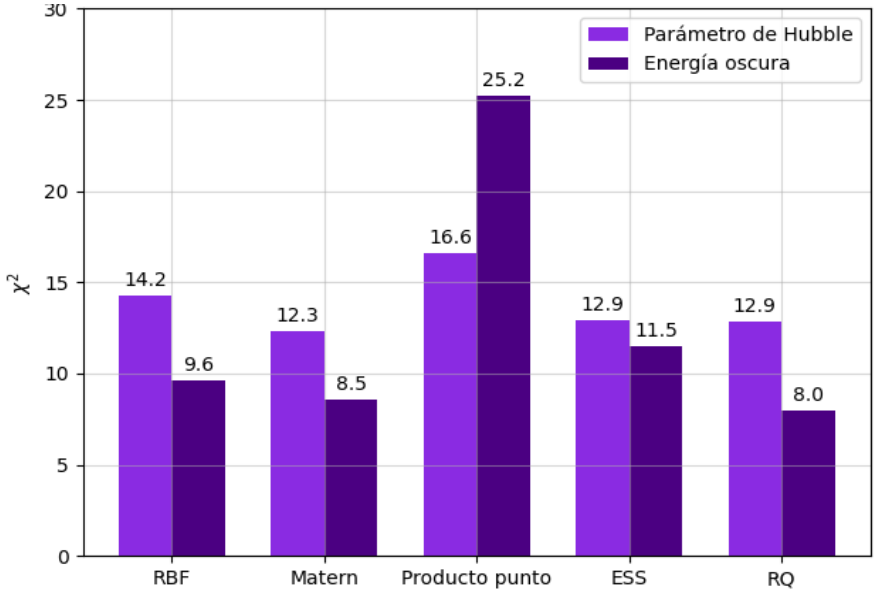

Figura 5.1: Valores de χ^2 para los modelos desarrollados con diferentes funciones de covarianza.

En [35], se pueden observar los modelos con diferentes kernels y uno puede confirmar a ojo que las regresiones con Matern y Racional Cuadrático son las más representativas, por lo que concluimos que la prueba de χ^2 se trata de una herramienta útil para determinar la mejor función de covarianza *a posteriori*. Como se pudo ver, los mejores modelos no siempre se obtienen a partir del mismo kernel, por eso es importante contar con una manera práctica y justificable para determinarlo.

Entre las principales ventajas de la GPR, se encuentra que las funciones de la regresión no necesitan asumir ninguna información sobre la relación entre las variables, por lo que la GPR representa totalmente a los datos de entrenamiento. Desafortunadamente, por esta razón, la técnica es muy dependiente de la fiabilidad de las mediciones; se debe asegurar que el conjunto de observables sea veraz. Además, vimos que es importante considerar varianzas no nulas en las observaciones para evitar el sobreajuste.

En general, no parece que la GPR sea particularmente suceptible a *outliers*, pues el aprendizaje con *maximum likelihood estimation* toma en cuenta a todos los puntos de entrenamiento. Sin embargo, un valor anómalo muy alejado de la relación "real" debería ser tratado con atención.

Otra ventaja es que, gracias a toda la teoría que existe alrededor de GPR, se ha vuelto un proceso relativamente sencillo. Sólo es necesario contar con un conjunto de datos y la GPR podrá encontrar un buen ajuste en la mayoría de las situaciones. Es por esto que las aplicaciones de este método no se limitan a las ciencias naturales; cualquier disciplina que requiera conocer la relación entre dos variables podría utilizarla.

Por otro lado, la GPR permite calcular la distribución Gaussiana de una predicción en cualquier valor de la variable independiente dentro del intervalo en el que se encuentran las observables. Por ejemplo, a lo largo de este trabajo, se modelaron predicciones en corrimientos al rojo $0 \leq z \leq 2$ porque los datos de cronómetros cósmicos sólo están contemplados en ese rango. Si se intenta extrapolar la relación a valores donde no hay datos, la regresión ya no representa adecuadamente a los observables.

5.2. Sobre las funciones cosmológicas

A grandes rasgos, concluimos que ΛCDM modela adecuadamente las mediciones experimentales del parámetro de Hubble con cronómetros cósmicos y todas las funciones cosmológicas que

se derivan de él, principalmente a *redshifts* bajos. Este fenómeno podría ser consecuencia del aumento directo en la dificultad de medir la tasa de expansión del Universo con el incremento en la distancia a nosotros.

Sin embargo, no podemos descartar el hecho de que el modelo cosmológico podría tratarse de una sobre-simplificación, ya que las regresiones resultantes de mediciones reales no coinciden exactamente con la teoría de ΛCDM.

Encontramos que la GPR con CC da lugar a un valor del parámetro de Hubble en el tiempo actual $H_0 = 68.798 \pm 6.340(1\sigma)$ km $\cdot\, s^{-1} \cdot$ Mpc^{-1}, que coincide con las mediciones del satélite Planck sobre el CMB, lo cual es otro aporte que lleva a pensar que las mediciones sobre variables cefeidas (y aquellas que concuerdan con éstas) podrían estar en un error o que hay algo aún desconocido causando el conflicto. Otros autores que han realizado reconstrucciones, incluso utilizando técnicas diferentes a GPR sobre mediciones de CC, tales como [56, 57, 58], que coinciden en el resultado aquí obtenido.

Analizando el parámetro de desaceleración encontrado, podemos determinar que la tasa expansión del Universo está incrementándose en el tiempo actual ($q(z = 0) < 0$), es decir, vivimos una época en la que domina la energía oscura, pero que en el pasado no fue así. Para ΛCDM, hay un punto en $z \sim 0.7$ a partir del cual la expansión se desacelera ($q(z > 0.7) > 0$), que corresponde a la temporada donde la materia dominó sobre la radiación. Durante este periodo, fue posible la formación de galaxias y grandes estructuras que vemos hoy en día. En el modelo de regresión, el punto final de la época de materia dominante es un poco más adelante: $z \sim 0.6$.

La curva de ΛCDM predice que la expansión se ralentizaba para mayores corrimientos al rojo, lo que se conoce como la época dominada por la radiación. Sin embargo, la regresión muestra un punto en el pasado ($z \sim 1.6$) donde la expansión fue acelerada.

Como se mencionó previamente, es difícil determinar si este detalle surge por el proceso de regresión o si aporta información real. Por lo anterior, es muy útil comparar las reconstrucciones con las de otros autores llevando a cabo trabajos similares, como [29], quienes concluyen que los resultados deben interpretarse con cuidado debido a la alta sensibilidad de la GPR a datos erróneos.

Sabemos que es difícil obtener datos confiables a grandes distancias (corrimientos al rojo) y que los modelos resultantes presentan mayores inconsistencias con la teoría precisamente cuanto

más grande es el valor de z, lo que indica que las discrepancias podrían deberse a problemas con las mediciones.

En cuanto a la EoS de la energía oscura, se observa en la Figura 4.6 que el ajuste representa correctamente a los datos y que su gráfica se concentra alrededor de la curva analítica para ΛCDM, pero las variaciones podrían indicar que la DE es más bien algo cuyo comportamiento cambia.

Como indica la ecuación (3.30), la ecuación de EoS de la DE proporciona información sobre la relación entre la presión y la densidad del fluido de constante cosmológica. Por lo tanto, encontrar que w cambia su valor en función del corrimiento al rojo sería una razón más para considerar modelos donde la ecuación de estado varía en términos de z, como $wCDM$.

Como se mencionó a lo largo de este trabajo, la GPR se ha utilizado previamente en la reconstrucción de las mismas funciones cosmológicas, encontrando resultados similares a los aquí desarrollados, por lo que se buscó hacer hincapié en algunos diferenciadores como la selección del kernel óptimo mediante el uso de métricas de comparación. El método desarrollado se puede aplicar a problemas futuros donde sea necesario determinar la mejor función de covarianzas para resolver un problema determinado.

La GPR demostró ser muy útil cuando se busca información sobre cómo cambia una variable en función de otra, pero es también una técnica que debe ser contrastada con otras para asegurar su fiabilidad, debido a su gran sensibilidad a datos erróneos. Aún así, encontramos que para corrimientos al rojo cercanos a cero, donde las mediciones de $H(z)$ a partir de cronómetros cósmicos es más confiable, ΛCDM sigue siendo una teoría que describe adecuadamente los fenómenos que observamos en el Universo.

Por lo anterior, se concluye que un conjunto de datos de entrenamiento adecuado es escencial para el éxito en la reconstrucción de funciones con GPR y se espera que con el avance en las formas de realizar observaciones o con el lanzamiento de nuevas misiones que tengan como objetivo medir el parámetro de Hubble a grandes distancias, se pueda mejorar la fiabilidad de esta técnica.

Bibliografía

[1] C. Gohd, "What is dark energy? Inside our accelerating, expanding Universe." NASA, 2024.

[2] R. Anyoha, "The History of Artificial Intelligence." Harvard, 2017.

[3] F. Chollet, *Deep Learning with Python*. Manning, 2017.

[4] S. Russell, P. Norvig, and E. Davis, *Artificial Intelligence: A Modern Approach*. Prentice Hall series in artificial intelligence, Prentice Hall, 2010.

[5] S. Lucci, S. Musa, and D. Kopec, *Artificial Intelligence in the 21st Century*. Mercury Learning & Information, 2022.

[6] E. Carlesi, Y. Hoffman, and N. I. Libeskind, "Estimation of the masses in the local group by gradient boosted decision trees," *Monthly Notices of the Royal Astronomical Society*, vol. 513, p. 2385–2393, Apr. 2022.

[7] A. Momtaz, M. H. Salimi, and S. Shakeri, "Estimating the Photometric Redshifts of Galaxies and QSOs Using Regression Techniques in Machine Learning," 2022.

[8] H. M. Kamdar, M. J. Turk, and R. J. Brunner, "Machine learning and cosmological simulations I: Semi-analytical models," *Monthly Notices of the Royal Astronomical Society*, vol. 455, p. 642–658, Nov. 2015.

[9] L. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall international editions, Prentice-Hall, 1994.

[10] J.-Y. Ran and J.-J. Wei, "Extragalactic Test of General Relativity from Strong Gravitational Lensing by using Artificial Neural Networks," 2024.

[11] T. Liu, X. Yang, Z. Zhang, J. Wang, and M. Biesiada, "Measurements of the hubble constant from combinations of supernovae and radio quasars," 2023.

[12] S. S. Sethuram, R. K. Cochrane, C. C. Hayward, V. Acquaviva, F. Villaescusa-Navarro, G. Popping, and J. H. Wise, "Emulating Radiative Transfer with Artificial Neural Networks," 2023.

[13] K. F. Dialektopoulos, P. Mukherjee, J. L. Said, and J. Mifsud, "Neural network reconstruction of cosmology using the pantheon compilation," *arXiv preprint arXiv:2305.15499*, 2023.

[14] J.-Z. Qi, P. Meng, J.-F. Zhang, and X. Zhang, "Model-independent measurement of cosmic curvature with the latest H(z) and SNe Ia data: A comprehensive investigation," *Physical Review D*, vol. 108, sep 2023.

[15] T. Liu, S. Cao, S. Zhang, X. Gong, W. Guo, and C. Zheng, "Revisiting the cosmic distance duality relation with machine learning reconstruction methods: the combination of HII galaxies and ultra-compact radio quasars," *The European Physical Journal C*, vol. 81, oct 2021.

[16] A. Kamerkar, S. Nesseris, and L. Pinol, "Machine learning cosmic inflation," *Physical Review D*, vol. 108, Aug. 2023.

[17] D. S. Deighan, S. E. Field, C. D. Capano, and G. Khanna, "Genetic-algorithm-optimized neural networks for gravitational wave classification," *Neural Computing and Applications*, vol. 33, p. 13859–13883, Apr. 2021.

[18] S. Nesseris and J. García-Bellido, "A new perspective on dark energy modeling via genetic algorithms," *Journal of Cosmology and Astroparticle Physics*, vol. 2012, p. 033–033, Nov. 2012.

[19] M. Sugiyama, *Introduction to Statistical Machine Learning*. Elsevier Science, 2015.

[20] R. Essick, I. Legred, K. Chatziioannou, S. Han, and P. Landry, "Phase Transition Phenomenology with Nonparametric Representations of the Neutron Star Equation of State," 2023.

[21] M.-Z. Han, Y.-J. Huang, S.-P. Tang, and Y.-Z. Fan, "Plausible presence of new state in neutron stars with masses above $0.98\mathrm{M}_{TOV}$," $Science Bulletin, vol.$ 68, 2023.

[22] X. Xu, S. Ho, H. Trac, J. Schneider, B. Poczos, and M. Ntampaka, "A first look at creating mock catalogs with machine learning techniques," *The Astrophysical Journal*, vol. 772, p. 147, July 2013.

[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Adaptive computation and machine learning, MIT Press, 2006.

[24] A. Mood, F. Graybill, and D. Boes, *Introduction to the Theory of Statistics*. McGraw-Hill international editions: Statistics series, McGraw-Hill, 1974.

[25] Howard Seltman, "Approximations for Mean and Variance of a Ratio." https://www.stat.cmu.edu/~hseltman/files/ratio.pdf, 2018.

[26] R. von Mises and H. Geiringer, *Mathematical Theory of Probability and Statistics*. Elsevier Science, 2014.

[27] M. P. Deisenroth, "Efficient Reinforcement Learning using Gaussian Processes," 2012.

[28] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2013.

[29] M. Seikel, C. Clarkson, and M. Smith, "Reconstruction of dark energy and expansion dynamics using gaussian processes," *Journal of Cosmology and Astroparticle Physics*, vol. 2012, p. 036–036, June 2012.

[30] GPy, "GPy: A gaussian process framework in python." <http://github.com/SheffieldML/GPy>, since 2012.

[31] M. van der Wilk, V. Dutordoir, S. John, A. Artemev, V. Adam, and J. Hensman, "A framework for interdomain and multioutput Gaussian processes," *arXiv:2003.01115*, 2020.

[32] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration," in *Advances in Neural Information Processing Systems*, 2018.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] M. Seikel, C. Clarkson, and M. Smith, "Reconstruction of dark energy and expansion dynamics using gaussian processes," *Journal of Cosmology and Astroparticle Physics*, vol. 2012, p. 036–036, June 2012.

[35] J. Ugalde, "Gp_in_cosmology." <https://github.com/JesusUg2497/GP_in_Cosmo logy>, 2023.

[36] R. D'Inverno, *Introducing Einstein's Relativity*. Clarendon Press, 1992.

[37] J. Feldman, "The simplicity principle in perception and cognition," *WIREs Cognitive Science*, vol. 7, pp. 330–340, 2016.

[38] B. Carroll and D. Ostlie, *An Introduction to Modern Astrophysics*. Cambridge University Press, 2017.

[39] P. Peebles, *Principles of Physical Cosmology*. Princeton Series in Physics, Princeton University Press, 1993.

[40] C. Andreoli, "New Hubble Constant Measurement Adds to Mystery of Universe's Expansion Rate." NASA, 2019.

[41] A. Liddle, *An Introduction to Modern Cosmology*. Wiley, 2003.

[42] S. Carroll, *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge University Press, 2019.

[43] F. F. Dirirsa, *Dark Energy and the Coincidence Problem*. PhD thesis, 06 2012.

[44] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J.-P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J.-F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J.-M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli,

K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J.-M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y.-Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M.-A. Miville-Deschênes, D. Molinari, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J.-L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A.-S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca, "Planck2018 results: Vi. cosmological parameters," *Astronomy amp; Astrophysics*, vol. 641, p. A6, Sept. 2020.

[45] D. W. Hogg, "Distance measures in cosmology," 2000.

[46] The NASA / LAMBDA Archive Team, "Hubble Constant." NASA, 2024.

[47] Astrobites, "Measuring the Curvature of the Universe with Cosmic Clocks." AAS Nova, 2020.

[48] S. Vagnozzi, A. Loeb, and M. Moresco, "Eppur è piatto? the cosmic chronometers take on spatial curvature and cosmic concordance," *The Astrophysical Journal*, vol. 908, p. 84, Feb. 2021.

[49] A. Gómez-Valent and L. Amendola, "H0from cosmic chronometers and type ia supernovae, with gaussian processes and the novel weighted polynomial regression method," *Journal of Cosmology and Astroparticle Physics*, vol. 2018, p. 051–051, Apr. 2018.

[50] A. Errahmani, A. Bouali, S. Dahmani, I. E. Bojaddaini, and T. Ouali, "Constraining dark energy equations of state in $f(r,t)$ gravity," 2024.

[51] M. Koussour, N. Myrzakulov, A. H. A. Alfedeel, E. I. Hassan, D. Sofuoğlu, and S. M Mirgani, "Square-root parametrization of dark energy in f(q) cosmology," *Communications in Theoretical Physics*, vol. 75, p. 125403, Dec. 2023.

[52] A. Tripathi, A. Sangwan, and H. Jassal, "Dark energy equation of state parameter and its evolution at low redshift," *Journal of Cosmology and Astroparticle Physics*, vol. 2017, p. 012–012, June 2017.

[53] B. R. Dinda and N. Banerjee, "A comprehensive data-driven odyssey to explore the equation of state of dark energy," 2024.

[54] T. Holsclaw, U. Alam, B. Sansó, H. Lee, K. Heitmann, S. Habib, and D. Higdon, "Nonparametric reconstruction of the dark energy equation of state," *Physical Review D*, vol. 82, Nov. 2010.

[55] L. A. Escamilla and J. A. Vazquez, "Model selection applied to reconstructions of the dark energy," *The European Physical Journal C*, vol. 83, no. 3, p. 251, 2023.

[56] C. Bengaly, M. A. Dantas, L. Casarini, and J. Alcaniz, "Measuring the hubble constant with cosmic chronometers: a machine learning approach," *The European Physical Journal C*, vol. 83, June 2023.

[57] M. Moresco, "Raising the bar: new constraints on the hubble parameter with cosmic chronometers at z $\sim$ 2," *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 450, p. L16–L20, Apr. 2015.

[58] G.-J. Wang, X.-J. Ma, S.-Y. Li, and J.-Q. Xia, "Reconstructing functions and estimating parameters with artificial neural networks: A test with a hubble parameter and sne ia," *The Astrophysical Journal Supplement Series*, vol. 246, p. 13, Jan. 2020.