Lecture notes

Contents

1 Statistics in cosmology

Los cosmólogos enfrentan nuevos desafíos en el análisis de datos debido al crecimiento exponencial en la cantidad y calidad de información disponible. Experimentos recientes generan conjuntos de datos masivos que requieren métodos avanzados, ya que algoritmos tradicionales se vuelven ineficientes. Además, al reducirse los errores estadísticos, emergen errores sistemáticos previamente enmascarados, relacionados tanto con artefactos instrumentales como con limitaciones teóricas.

En esta introducción se presentan técnicas para el análisis cosmológico moderno, comenzando con conceptos estadísticos básicos (verosimilitud, previos y posteriores) y su aplicación a espectros de potencia. Introduce herramientas como la *Matriz de Fisher* (para estimación rápida de errores) y métodos *Markov Chain Monte Carlo* (MCMC), que permiten manejar verosimilitudes complejas. Estas metodologías, aunque desarrolladas para cosmología, tienen aplicaciones transversales en diversas áreas científicas que trabajan con grandes volúmenes de datos.

1.1 The Likelihood function

La función de verosimilitud $L(\{d_i\}|\theta)$ representa el núcleo conceptual de los análisis estadísticos modernos en cosmología. Se define formalmente como la probabilidad condicional de observar un conjunto de datos experimentales $\{d_i\}_{i=1}^m$ dado un modelo teórico parametrizado por $\theta = (w, \sigma_w)$. Para mediciones independientes, esta función se expresa como el producto de probabilidades individuales:

$$L(\{d_i\}|\theta) = \prod_{i=1}^{m} P(d_i|\theta), \tag{1}$$

donde cada término $P(d_i|\theta)$ corresponde a la distribución de probabilidad del modelo para cada dato observado. La potencia analítica de este constructo matemático reside en su capacidad para invertir la relación lógica entre datos y teoría mediante la aplicación del teorema de Bayes:

$$P(\theta|\{d_i\}) \propto L(\{d_i\}|\theta)P(\theta), \tag{2}$$

permitiendo inferir los parámetros del modelo a partir de las observaciones. En el contexto del ejemplo pedagógico: -la medición del peso w de un individuo utilizando $m = 100$ básculas independientes con error gaussiano-, cada observación d_i sigue una distribución normal $d_i \sim \mathcal{N}(w, \sigma_w^2)$. La verosimilitud para una sola medición adopta la forma gaussiana clásica:

$$L(d_i|w, \sigma_w) = (2\pi\sigma_w^2)^{-1/2} \exp\left(-\frac{(d_i - w)^2}{2\sigma_w^2}\right). \tag{3}$$

Al extenderse a múltiples observaciones independientes, la verosimilitud conjunta se factoriza en un producto de términos gaussianos, lo que en escala logarítmica se traduce en una suma cuadrática:

$$\ln L(\{d_i\}|w, \sigma_w) = -\frac{m}{2}\ln(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2}\sum_{i=1}^{m}(d_i - w)^2. \tag{4}$$

Los estimadores de máxima verosimilitud (MLE) para los parámetros se obtienen mediante la optimización de esta función. Para el peso w, la condición $\partial \ln L/\partial w = 0$ conduce al estimador:

$$\hat{w} = \frac{1}{m}\sum_{i=1}^{m} d_i, \tag{5}$$

que corresponde a la media muestral. Análogamente, al resolver $\partial \ln L/\partial\sigma_w^2 = 0$ se obtiene el estimador para la varianza

$$\hat{\sigma}_w^2 = \frac{1}{m}\sum_{i=1}^{m}(d_i - \hat{w})^2. \tag{6}$$

Estos resultados ejemplifican cómo los principios de máxima verosimilitud recuperan estimadores intuitivos en casos simples. La transición desde la verosimilitud hacia la distribución posterior de parámetros se realiza mediante el teorema de Bayes:

$$P(w, \sigma_w|\{d_i\}) \propto L(\{d_i\}|w, \sigma_w)P_{\rm prior}(w, \sigma_w), \tag{7}$$

donde $P_{\rm prior}$ incorpora cualquier conocimiento a priori sobre los parámetros. Esta distribución posterior contiene toda la información probabilística sobre los parámetros condicionada a los datos observados. La constante de normalización, denominada evidencia, asegura que la posterior integre a unidad sobre el espacio de parámetros. El análisis de incertidumbres en los estimadores se fundamenta en el examen de la curvatura del logaritmo de la verosimilitud alrededor de su máximo. Para el parámetro w, la varianza del estimador resulta:

$$\mathrm{Var}(\hat{w}) = \frac{\sigma_w^2}{m}, \tag{8}$$

reflejando la reducción clásica de ruido conforme $\sqrt{m}$. En contraste, la estimación de σ_w^2 presenta comportamientos no-gaussianos, con varianza:

$$\mathrm{Var}(\hat{\sigma}_w^2) = \frac{2}{m}\sigma_w^4. \tag{9}$$

Este error en σ_w puede parecer un detalle técnico, pero en cosmología, gran parte de lo que medimos (como fluctuaciones en la densidad de galaxias o la temperatura del CMB) es análogo a σ_w^2. Estas fluctuaciones siguen distribuciones (a menudo gaussianas) cuyos parámetros dependen del modelo cosmológico. Por ello, esta ecuación es fundamental: al estimar la varianza de una distribución, hay una incertidumbre intrínseca proporcional a $\sigma_w^2/\sqrt{m}$, llamada varianza muestral o varianza cósmica.

Por otro lado, es conveniente definir intervalos de credibilidad para los parámetros. En el ejemplo de la verosimilitud gaussiana, los intervalos de confianza al 68% (1-σ) corresponden a la región donde $\ln L$ decrece en 1/2 unidades desde su valor máximo, delimitando así el rango de parámetros consistentes con los datos. De manera general, se definen los valores ω_- y ω_+ tales que:

$$\int_{\omega_-}^{\omega_+} d\omega P(\omega|\{d_i\}) = 0.68, \tag{10}$$

En aplicaciones reales, usualmente hay múltiples parámetros desconocidos. Si algunos son irrelevantes, debemos marginalizar sobre ellos, formalmente esto se implementa mediante la integración multidimensional:

$$P(\theta|\{d_i\}) = \int_\Phi P(\theta, \phi|\{d_i\})d\phi, \tag{11}$$

sobre el espacio completo Φ de parámetros *nuisance*.

1.2 From raw data to parameter constraints

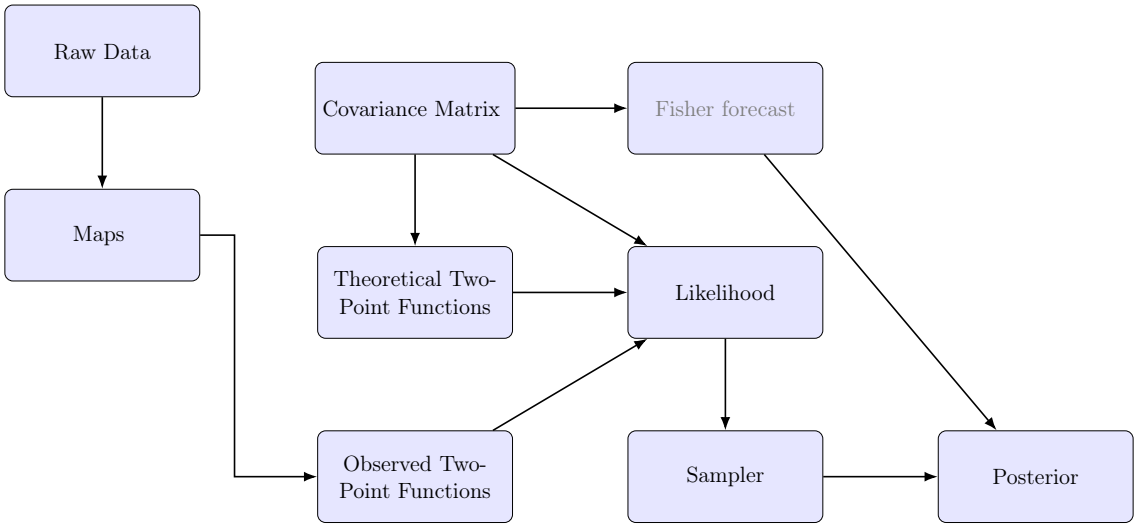

Figure 1: Pipeline desde los datos crudos hasta las restricciones cosmológicas.

El análisis cosmológico moderno sigue una metodología bien establecida que comienza con los mapas observacionales. Estos representan la materia prima del estudio, correspondiendo a representaciones espaciales de los fenómenos cósmicos investigados. En el caso de la radiación cósmica de fondo (CMB), se trabaja con mapas de anisotropías de temperatura; para sondeos galácticos, con distribuciones tridimensionales de densidad; y en estudios de lente gravitacional débil, con campos de elipticidad galáctica.

A partir de estos mapas, se extraen los estadísticos de dos puntos, que capturan las correlaciones fundamentales mediante funciones de dos puntos. Estas pueden expresarse tanto en el espacio espectral como en el espacio de configuraciones, proporcionando así una descripción estadística compacta de las propiedades de los campos observados. La transición de los mapas brutos a estos estadísticos permite reducir considerablemente la dimensionalidad de los datos, reduciendo la complejidad computacional, pero perdiendo información no gaussiana.

Con los estadísticos medidos en mano, el siguiente paso es la construcción de la verosimilitud, que compara estas mediciones con sus contrapartes teóricas. Esta comparación se realiza ponderando las diferencias mediante la matriz de covarianza inversa. Para el caso gaussiano, la función de verosimilitud para las anisotropías del CMB toma la forma:

$$\ln \mathcal{L}(\lambda_\alpha) = -\frac{1}{2}\sum_{ll'}\left(\hat{C}(l) - C_{\text{theory}}(l, \lambda_\alpha)\right) \text{Cov}_{ll'}^{-1}\left(\hat{C}(l') - C_{\text{theory}}(l', \lambda_\alpha)\right) \tag{12}$$

Para explorar el espacio de parámetros cosmológicos λ_α se emplean técnicas de muestreo de parámetros, particularmente algoritmos MCMC (Monte Carlo Markov Chain). Estos algoritmos evalúan sistemáticamente millones de combinaciones paramétricas, permitiendo caracterizar completamente la distribución posterior. Este proceso debe manejar adecuadamente dos desafíos clave: las degeneraciones entre parámetros cosmológicos y la marginalización sobre parámetros molestos no cosmológicos, como el parámetro de bias b_1 en estudios de distribución galáctica.

Finalmente, como complemento al muestreo exhaustivo, las estimaciones analíticas mediante pronósticos de Fisher ofrecen una alternativa eficiente para proyectar barras de error. Estas aproximaciones analíticas son especialmente valiosas en etapas preliminares de diseño experimental, cuando se requiere evaluar rápidamente el potencial científico de diferentes configuraciones instrumentales sin recurrir a costosos análisis completos. El proceso completo se ilustra en la Figura 1.

2 Introducción

2.1 Motivación

La inferencia sin verosimilitud (*likelihood-free inference*) permite estimar parámetros cosmológicos sin asumir una forma analítica para la función de verosimilitud, evitando así sesgos por supuestos incorrectos como Gaussianidad. Esto es especialmente relevante en cosmología, donde procesos no lineales y efectos sistemáticos complican el modelado estadístico. En el contexto del CMB, aunque las fluctuaciones primordiales son aproximadamente Gaussianas, efectos secundarios, como lentes gravitacionales, polarización inducida y *foregrounds* introducen no-Gaussianidades que desafían los análisis tradicionales basados en espectros de potencia.

El CMB es uno de los observables más poderosos para estudiar los parámetros cosmológicos, pero su buena explotación requiere ir más allá de las estadísticas de dos puntos. En este trabajo, comenzaré con el análisis de estadísticas resumidas tradicionales, partiendo de los *angular power spectrum* (APS) de temperatura, que capturan la información gaussiana primordial. Posteriormente, escalaré hacia estadísticas más complejas que incluyan modos de polarización, para finalmente explorar métodos no lineales (de ser posible).

Métodos de aprendizaje profundo, como la compresión neuronal, pueden extraer información no lineal de mapas del CMB, mejorando las restricciones sobre parámetros como Ω_m o σ_8. Además, la inclusión de efectos no ideales como ruido instrumental, cortes en el cielo y *foregrounds* en simulaciones realistas es importante para un análisis robusto. Esta aproximación gradual—desde espectros de potencia hasta estadísticas de alto orden—permitirá validar resultados intermedios y cuantificar ganancias al incorporar información no gaussiana.

2.2 Inferencia Bayesiana

El marco bayesiano permite determinar la distribución posterior de parámetros cosmológicos θ mediante la expresión

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{13}$$

donde x representa los datos observacionales. La determinación precisa de la función de verosimilitud $p(x|\theta)$ presenta varios desafíos en el análisis del CMB. Aunque las fluctuaciones primordiales son aproximadamente gaussianas, efectos secundarios como las lentes gravitacionales, la contaminación por *foregrounds* y las no-linearidades instrumentales introducen complejidades. Estas complicaciones se acentúan al considerar estadísticas más allá del espectro de potencia, como los estudios de no-gaussianidad primordial o los análisis de lentes gravitacionales a pequeña escala.

Los métodos tradicionales basados en supuestos gaussianos para la verosimilitud muestran limitaciones crecientes frente a la precisión de los nuevos experimentos. Particularmente, para análisis que involucran reconstrucción de lentes, estudios de no-gaussianidad o el tratamiento de regiones con máscaras complejas, la distribución exacta de los estadísticos se desvía significativamente de la aproximación gaussiana. Esta discrepancia puede llevar a estimaciones sesgadas de parámetros cosmológicos.

2.3 *Likelihood-free inference*

La inferencia sin verosimilitud (también llamada *Simulation-based inference*) busca realizar inferencia estadística en situaciones donde la función de verosimilitud es intratable o desconocida. Métodos tradicionales como el *Approximate Bayesian Computation* (ABC) abordan este

problema comparando directamente los datos simulados con los datos observados mediante una métrica de distancia. En este enfoque, se generan múltiples simulaciones a partir de diferentes valores del parámetro θ y se conservan aquellos para los cuales los datos simulados x se parecen lo suficiente a los datos observados, según un umbral de tolerancia. Este método depende fuertemente de la elección de estadísticas resumidas y puede requerir un número muy elevado de simulaciones para obtener una aproximación razonable de la distribución posterior. La figura 1 se muestra la pipeline general utilizada para estimar posteriores con el método ABC.

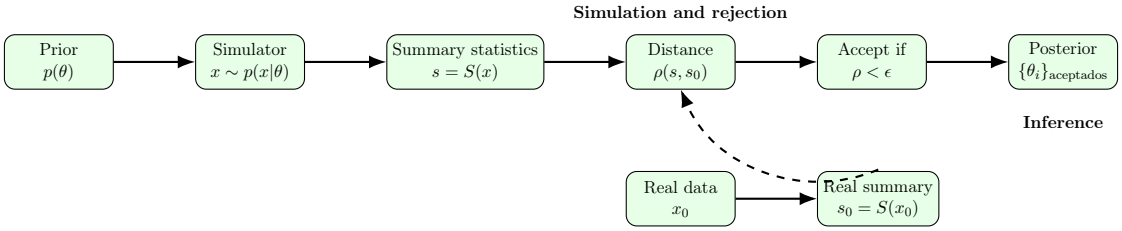

Figure 2: Pipeline general de ABC.

Más recientemente, técnicas modernas basadas en aprendizaje automático han revolucionado este enfoque. En lugar de comparar datos de manera directa, estos métodos reformulan el problema como uno de estimación de densidad: se modela la distribución conjunta de pares (θ, x) simulados, lo que permite aproximar la distribución posterior $p(\theta|x)$. Herramientas como redes neuronales profundas y flujos normalizantes permiten entrenar modelos expresivos que aprenden directamente la relación entre datos y parámetros a partir de muestras sintéticas, evitando el cálculo explícito de la verosimilitud y haciendo la inferencia mucho más eficiente. La figura 2 muestra la pipeline general para técnicas de sbi basadas en aprendizaje automático.

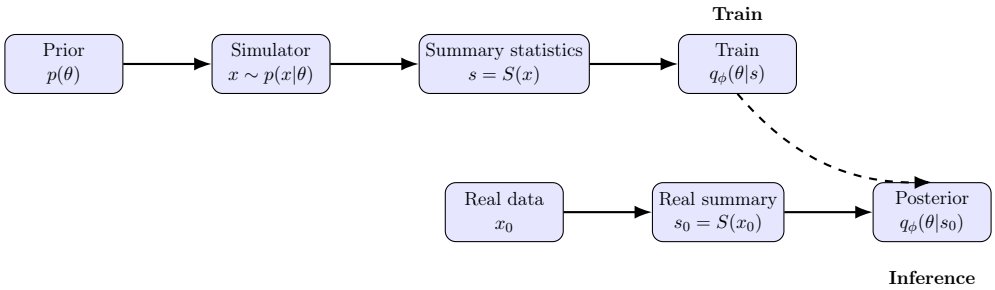

Figure 3: Pipeline general para métodos con aprendizaje automático.

El paquete `sbi`, desarrollado en Python sobre `PyTorch`, implementa este paradigma de manera flexible. Ofrece herramientas para: (1) simular datos bajo distintos parámetros θ, (2) entrenar estimadores de densidad condicional $p(\theta|x)$ usando arquitecturas modernas, y (3) realizar inferencia posterior dado un conjunto de datos observados. La calidad de las simulaciones y de las estadísticas utilizadas continúa siendo muy importante, ya que el modelo entrenado sólo puede aprender lo que está presente en los datos simulados. Sin embargo, al entrenar modelos que generalizan bien, se puede reducir de forma significativa el número de simulaciones necesarias, logrando una inferencia más precisa y escalable que con métodos como ABC.

3 Simulador

3.1 Generador de APS

El generador implementa un cálculo teórico completo de las anisotropías del fondo cósmico de microondas mediante la resolución numérica de las ecuaciones de Boltzmann a través del código

`CAMB`. Parte de un conjunto de parámetros cosmológicos que definen un modelo ΛCDM estándar y devuelve su *A*PS correspondiente. En la actual implementación, el cálculo se restringe al espectro de temperatura, aunque la arquitectura del código permite una extensión futura para incluir los modos de polarización. El resultado final consiste en el APS completo expresado en microkelvin, que cuantifica las fluctuaciones de temperatura en función del multipolo ℓ.

3.2 Ruido instrumental y cobertura parcial del cielo

Para generar simulaciones realistas de los APS, se consideran dos fuentes principales de ruido: el ruido instrumental del experimento y la cobertura parcial del cielo, que introduce varianza cósmica adicional. Primero, se añade el ruido instrumental a los espectros teóricos mediante un modelo basado en la resolución angular del experimento (θ_{fwhm}) y la sensibilidad por píxel (σ_T). El término de ruido instrumental N_ℓ^{TT} que se suma al espectro de potencia teórico C_ℓ es:

$$N_\ell^{\text{TT}} = (\theta_{\text{fwhm}} \cdot \sigma_T)^2 \exp\left[\ell(\ell+1)\frac{\theta_{\text{fwhm}}^2}{8\ln 2}\right], \tag{14}$$

donde θ_{fwhm} se expresa en radianes. Esta fórmula modela el suavizado del cielo debido a la resolución finita del instrumento. Posteriormente, se simula la observación del cielo parcial considerando que solo se mide una fracción $f_{\text{sky}} < 1$ del cielo completo. Como consecuencia, el estimador observable de C_ℓ, denotado $\hat{C}_\ell$, no es determinístico, sino una variable aleatoria cuya dispersión depende de f_{sky}. Para multipolos bajos ($\ell < \ell_{\text{transition}}$), se modela esta varianza como una distribución chi-cuadrado escalada:

$$\hat{C}_\ell = \frac{1}{\nu_\ell}\sum_{i=1}^{\nu_\ell} X_i^2, \quad X_i \sim \mathcal{N}(0, \sqrt{C_\ell}), \tag{15}$$

donde $\nu_\ell = \text{round}(f_{\text{sky}} \cdot (2\ell+1))$ representa el número efectivo de grados de libertad. Esta formulación captura correctamente la dispersión estadística del estimador cuando el número de modos disponibles es pequeño. Para multipolos altos ($\ell \geq \ell_{\text{transition}}$), se asume que el estimador puede aproximarse mediante una distribución normal centrada en C_ℓ con varianza:

$$\hat{C}_\ell \sim \mathcal{N}\left(C_\ell, \frac{2C_\ell^2}{f_{\text{sky}}(2\ell+1)}\right). \tag{16}$$

Este enfoque es válido debido al teorema central del límite, ya que en esta región el número de modos es suficientemente grande como para justificar una aproximación gaussiana.

4 Inference

5 Métodos bayesianos

En las últimas décadas, la cosmología ha experimentado una explosión en la cantidad y calidad de datos observacionales, desde mediciones de anisotropías en el Fondo Cósmico de Microondas (CMB) hasta sondeos de galaxias y lentes gravitacionales a gran escala. Este crecimiento ha impulsado la necesidad de técnicas de análisis más sofisticadas, capaces de manejar conjuntos de datos masivos y modelos teóricos cada vez más complejos.

En este contexto, los métodos bayesianos se han convertido en una herramienta fundamental para la inferencia cosmológica. Estos permiten combinar observaciones con predicciones teóricas mediante el uso de distribuciones de probabilidad, actualizando nuestro conocimiento a través del teorema de Bayes:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)\, P(\theta)}{P(\mathcal{D})}, \tag{17}$$

donde $P(\theta \mid \mathcal{D})$ es la distribución posterior, $P(\mathcal{D} \mid \theta)$ es la verosimilitud, $P(\theta)$ es la distribución previa, y $P(\mathcal{D})$ actúa como evidencia.

Sin embargo, en cosmología, el cálculo de la verosimilitud suele ser computacionalmente costoso o incluso intratable cuando se trabaja con simulaciones numéricas. Esto ha llevado al desarrollo de técnicas de *Simulation-Based Inference* (SBI), que evitan la evaluación explícita de la verosimilitud mediante aproximaciones basadas en simulaciones. En las siguientes secciones, exploraremos cómo estos métodos están revolucionando el análisis de datos cosmológicos.

5.1 Función de verosimilitud

La función de verosimilitud L es fundamental en análisis estadísticos modernos, definida como la probabilidad de observar los datos dado un modelo teórico. Esta herramienta permite estimar parámetros del modelo y sus incertidumbres. Por ejemplo, al medir repetidamente el peso de una persona con múltiples básculas, la verosimilitud modela cómo las mediciones d_i se distribuyen alrededor del peso real w, asumiendo un ruido gaussiano con desviación estándar σ_w.

$$L(w|d_i, \sigma_w) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{(d_i - w)^2}{2\sigma_w^2}\right) \tag{18}$$

Para m mediciones independientes, la verosimilitud conjunta es el producto de las verosimilitudes individuales, resultando en una expresión exponencial que depende de la suma de cuadrados de las diferencias entre datos y modelo.

$$L(w|d_i, \sigma_w) = \prod_i \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{(d_i - w)^2}{2\sigma_w^2}\right) \tag{19}$$