Universidad de Cantabria

Facultad
de
Ciencias

Análisis de Componentes Principales aplicado a la eliminación de *foregrounds* en observaciones simuladas de la línea de 21 cm con *SKA*

(Principal Component Analysis applied to foreground removal in simulated observations of the 21 cm line with SKA)

Trabajo de Fin de Grado
para acceder al

GRADO EN FÍSICA

Autor: Fernando Pardo Santiago

Director: Marcos Cruz Rodríguez

Fecha: 04/09/2025

A mis padres y a mi hermana por haberme apoyado tanto en estos 4 años y a lo largo de toda mi vida, y a ese grupo de compañeros, LOS CHIPIS, que tantas alegrías y buenos momentos hemos pasado en esta etapa.

Resumen

Este trabajo se centra en el estudio de la línea de 21 cm del hidrógeno neutro, una de las herramientas más prometedoras para investigar el universo temprano, en particular la Época de Reionización (EoR). El principal reto para su detección es la presencia de *foregrounds*, señales astrofísicas e instrumentales mucho más intensas que la señal cosmológica buscada. Para afrontarlo, se han utilizado datos simulados del *Science Data Challenge 3a* del radiotelescopio *SKA-Low*, aplicando un proceso de reconvolución y la técnica de Análisis de Componentes Principales (PCA) con el fin de eliminar los *foregrounds* y obtener espectros de potencia de la señal de 21 cm. Los resultados muestran que la técnica de PCA permite recuperar parcialmente la señal de 21 cm y estimar espectros de potencia cercanos a los esperados, aunque con limitaciones asociadas a los errores sistemáticos instrumentales difíciles de cuantificar y la dependencia del número de componentes eliminadas. El trabajo concluye que la PCA es una herramienta eficaz, pero que deberá combinarse con otras técnicas para garantizar una extracción más eficaz de la señal en futuros experimentos.

Abstract

This work focuses on the study of the 21 cm line of neutral hydrogen, one of the most promising tools to investigate the early universe, particularly the Epoch of Reionization (EoR). The main challenge for its detection is the presence of foregrounds, astrophysical and instrumental signals much stronger than the cosmological signal of interest. To address this, simulated data from the *Science Data Challenge 3a* of the SKA-Low radio telescope were used, applying a reconvolution process and the Principal Component Analysis (PCA) technique to remove the foregrounds and obtain power spectra of the 21 cm signal. The results show that the PCA technique allows partial recovery of the 21 cm signal and the estimation of power spectra close to the expected ones, although with limitations associated with difficult-to-quantify instrumental systematics and the dependence on the number of removed components. The study concludes that PCA is an effective tool, but it should be combined with other techniques to ensure a more reliable extraction of the signal in future experiments.

Índice

1 Introducción

1.1. Evolución del Universo

El Universo se originó en un estado de energía extraordinariamente caliente y denso hace aproximadamente 14 mil millones de años (14 Gyr), evento que recibe el nombre de *Big Bang*. En torno a los $10^{-43} - 10^{-36}$ s desde el inicio del Universo, las fuerzas fundamentales estaban unificadas excepto la gravedad. Alrededor de los 10^{-35} s tuvo lugar una rápida expansión exponencial y para los 10^{-32} s posteriores al *Big Bang* se produjo la separación de la fuerza fuerte y la unificación de las fuerzas débil y electromagnética. Hasta 10^{-6} s el Universo estaba formado por quarks y gluones, por lo que menos de 1 s después del inicio del Universo, las primeras partículas subatómicas ya se habían formado.

Desde este momento hasta aproximadamente 380 mil años, el Universo estaba formado principalmente por fotones que interactuaban constantemente con partículas cargadas. Unos 380 mil años después del Big Bang, a medida que el universo continuaba expandiéndose y enfriándose, su temperatura descendió a unos 3000 K. Esto permitió que los electrones y los núcleos se combinaran para formar los primeros átomos neutros estables en lo que se conoce como *recombinación*. Esto hizo que el universo se volviera transparente a la radiación, ya que los fotones ya no se dispersaban constantemente por los electrones libres. Esta radiación remanente es lo que observamos hoy y que se denomina el Fondo Cósmico de Microondas *(CMB)*, un campo de radiación de cuerpo negro casi perfectamente isótropo con una temperatura actual de aproximadamente 3 K. El período que siguió a la formación de átomos neutros, pero antes de que se encendieran las primeras estrellas, a menudo se conoce como *Dark Ages (Época Oscura)*. Esta etapa recibe el nombre de *Época Oscura* debido a que aun no existían estrellas ni galaxias, y la única radiación presente era el CMB que se iba debilitando y desplazando a longitudes de ondas más largas a medida que el Universo se expandía.

Las *Época Oscura* concluyó alrededor de 200 Myr después del *Big Bang* con la formación de las primeras estrellas. La radiación de estas primeras estrellas reionizaron el hidrógeno neutro en el universo, iniciándose así una nueva etapa que recibe el nombre de la *Época de Reionización (EoR)*. Posteriormente, durante la *Época galáctica* (desde 200 Myr hasta aproximadamente 3 Gyrs después del *Big Bang*), se formaron las estructuras a gran escala y la mayoría de las galaxias. El recorrido

por las distintas etapas del Universo quedan reflejadas en la Figura 1.1.

Desde entonces, el Universo ha continuado evolucionando y vive una etapa caracterizada por la formación continua de estrellas dentro de las galaxias y la aparición de planetas y vida. Actualmente, el universo también se encuentra en una era de la energía oscura, donde esta es el componente dominante del cosmos, impulsando su expansión acelerada. [1]

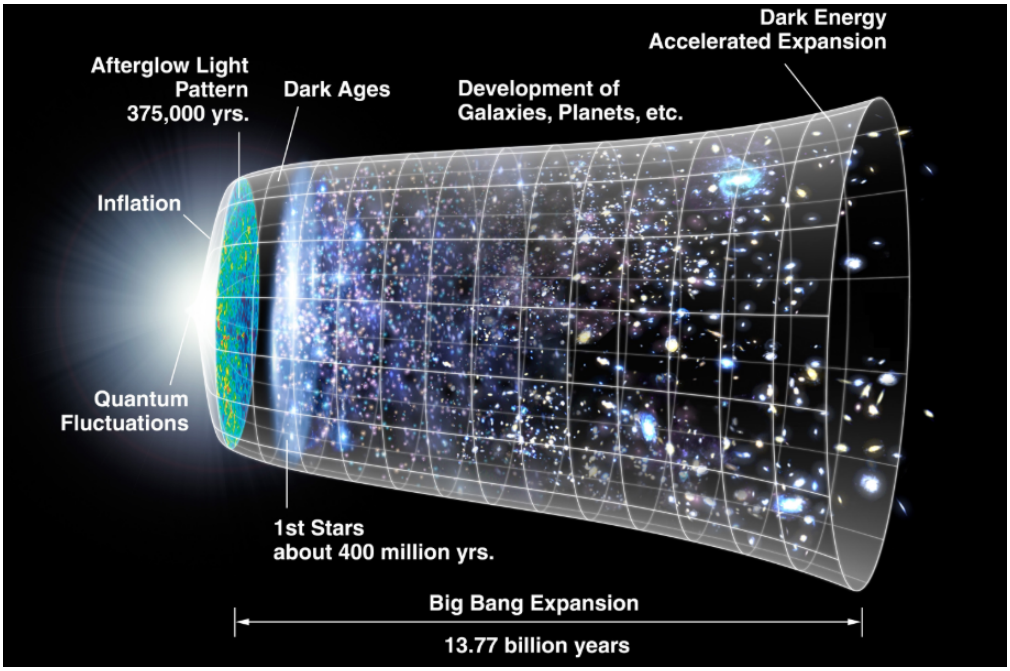

Figura 1.1: Etapas de la evolución del Universo desde el *Big Bang* hace aproximadamente 14 Gyr hasta el presente. [2]

A pesar de tener una visión bastante generalizada de las etapas del Universo desde el *Big Bang*, existen determinadas épocas de las cuales no tenemos suficiente información, como es el caso de las *Dark Ages* o la *Época de Reionización*. El *CMB* y los sondeos de galaxias permiten conocer cómo era el Universo en sus etapas iniciales y posteriores, sin embargo, existen etapas entre medias que no han sido completamente estudiadas, por lo que existe una brecha de entre 400 mil y 1.5 Gyr después del *Big Bang*, correspondientes a *redshifts* $1100 < z < 3$. [3]
El estudio de la radiación del hidrógeno neutro, del cual el Universo en estas épocas estaba formado casi en su totalidad permite obtener información de cómo estaba conformado el Universo.

1.2. Línea de 21 cm del hidrógeno neutro

El hidrógeno es el elemento más abundante del universo, y su estudio a nivel astronómico permite conocer cómo estaba conformado el Universo en las distintas etapas desde su formación, es decir, desde el Big Bang hace casi 14 Gyr [4]. Una de las propiedades que ayuda a recabar información acerca de una región del espacio en una determinada etapa del Universo es la transición *spin-flip* del átomo de hidrógeno neutro. Esta transición hiperfina del hidrógeno neutro H_I se produce cuando los espines del protón y el electrón pasan de estar paralelos a antiparalelos, emitiendo un fotón con

una frecuencia ν_{10} que viene dada por:

$$\nu_{10} = \frac{8}{3} g_I \left(\frac{m_e}{m_p}\right) \alpha^2 (R_M c) \approx 1420.406 \text{ MHz} \tag{1.1}$$

donde g_I es el factor g nuclear ($g_I \approx 5.59$), m_e y m_p son las masas del electrón y del protón respectivamente, α es la constante de estructura fina y $R_M c$ es la frecuencia de Rydberg para el hidrógeno ($R_M c = 3.28 \cdot 10^{15}$ Hz) [5].

Esta frecuencia corresponde a una longitud de onda del fotón emitido de $\lambda = 21$ cm. El estudio de esta radiación recibe el nombre de *cosmología de 21 cm* y en los últimos años ha ganado una vital importancia ya que es una herramienta muy útil para obtener mapas de cómo el gas de hidrógeno neutro ha evolucionado en distintas etapas del universo.

Durante la *Época Oscura*, la señal del hidrógeno neutro (HI) era casi uniforme, pero comenzó a presentar estructuras cuando las primeras fuentes astronómicas crearon burbujas de gas ionizado a su alrededor. A medida que estas burbujas crecían y se fusionaban, la señal de HI adquiría variaciones en frecuencia asociadas a la línea de 21 cm desplazada al rojo. Hacia $z \approx 6$, el tamaño característico de las burbujas alcanzó unos 10 Mpc, produciendo señales con escalas angulares de varios minutos de arco y anchuras de varios MHz, que contienen información única sobre la formación de las primeras estructuras cósmicas. Sin embargo, la detección de estas señales resulta extremadamente difícil: son muy débiles, aparecen en frecuencias bajas (≈100 MHz) afectadas por interferencias de radio, y están enmascaradas por un *foreground* mucho más intenso proveniente de fuentes de radio extragalácticas. [5]

1.2.1. Cosmología de 21 cm

Tras el Big Bang, el Universo se fue expandiendo y enfriando al mismo tiempo, hasta que unos 370 mil años después del Big Bang, los primeros átomos de hidrógeno se fueron formando debido a la recombinación de protones y electrones presentes en los primeros años del Universo. Con la formación de estos primeros átomos de hidrógeno, el Universo comenzó una nueva etapa que recibe el nombre de *Dark Age*. El Universo estaba compuesto principalmente por hidrógeno neutro, hasta que las primeras estrellas y galaxias comenzaron a formarse, iniciándose así una nueva época llamada *Cosmic Dawn*, aproximadamente 100 millones de años después del Big Bang. [6]
La radiación emitida por estas nuevas estrellas y galaxias comenzaron a reionizar el gas neutro de los alrededores, volviendo a encontrarse el Universo en un estado ionizado. Esta época recibe el nombre de *Época de Reionización (EoR)*. La transición entre estas dos etapas del Universo puede entenderse entonces a través de las señales que emite el hidrógeno, en concreto la línea de emisión de 21 cm, que va a permitir conocer cómo estaba distribuido el gas de hidrógeno neutro en el espacio intergaláctico (IGM).

1.2.2. Características de la línea de 21 cm/Foregrounds

La línea de 21 cm del hidrógeno emitida por la transición *spin-flip* sufre un corrimiento al rojo (*redshift)* a medida que se propaga por el espacio. La emisión del fotón de frecuencia $\nu = 1420$ MHz se encuentra en el rango de las microondas, sin embargo, los *redshifts* z correspondientes a las

épocas del *Cosmica Dawn* y *Época de Reionización*, se encuentran en un rango de $30 \leq z \leq 6$ [7] aproximadamente, por lo que esta señal se observará en la Tierra en el rango de las ondas de radio del espectro electromagnético.

Todos los objetos astronómicos y cuerpos celestes emiten radiación electromagnética dentro del rango de las ondas de radio. Solo existen algunos fuentes de radio que no llegan a la Tierra, pero en su gran mayoría penetran la atmósfera ya que las ondas de radio pueden penetrar nubes de polvo en el espacio intergaláctico. La atmósfera terrestre permite que las ondas de radio pasen en un rango de frecuencia que va desde los 10 MHz hasta 1 Thz aproximadamente. [5]
En el espacio intergaláctico (IGM) existen múltiples fuentes de ondas de radio además de la línea de 21 cm que se quiere obtener. Todas estas señales de radio que no interesan y es preciso eliminar para obtener solo la señal de radio de la línea de 21 cm reciben el nombre de *foreground*. Estos *foregrounds* pueden ser galácticos o extragalácticos, incluso terrestres, y deben ser tenidos en cuenta. No todas estas fuentes de radio emiten con la misma intensidad en todo el rango de frecuencias de radio, sino que algunas son más intensas a frecuencias bajas y a frecuencias mayores son despreciables y viceversa, por lo que existe una relación muy marcada entre los *foregrounds* y la frecuencia.
Aunque los *foregrounds* sean científicamente interesantes por sí mismas, para la cosmología de 21 cm deben ser eliminados. Para frecuencias inferiores a los 20 GHz, la emisión de ondas de radio está dominada por la radiación de sincrotón. A medida que se observan frecuencias más bajas, la radiación de sincrotón se vuelve cada vez más intensa. Es cierto que existen otras fuentes de radio tales como la *free-free emission* o fuentes puntuales brillantes de radio. Sin embargo, aunque estas sean considerablemente menos intensas que la radiación de sincrotón, pueden llegar a tener intensidades iguales o superiores a la de la señal de 21 cm del hidrógeno, por lo que también se deben tener en cuenta y deben ser eliminadas.

Por una parte, las observaciones de la línea de 21 cm son similares a las del CMB, sobre todo en el hecho de que en ambas se debe eliminar el *foreground* para tener medidas precisas. En el caso del CMB, la señal cosmológica de interés es la principal fuente de emisión en el plano galáctico, por lo que la eliminación del *foreground* es solo necesaria en ciertas regiones. Sin embargo, para el caso de la cosmología de 21 cm, los *foreground* dominan en todas las regiones del cielo por lo que técnicas de sustracción eficientes del *foreground* son indispensables para obtener la señal. [3]

Debido al *redshift* de las ondas de radio que nos interesa medir, $(30 \leq z \leq 6)$, se encontrarán en un rango de frecuencias de 30 a 200 MHz [8]. Por ejemplo para $z = 8.5$, corresponde una frecuencia de 150 MHz [7]. Es por esto que el hecho de observar diferentes frecuencias en el rango de las ondas de radio equivale a observar la radiación de la línea de 21 cm a distintos corrimientos al rojo, y por lo tanto a distintas distancias respecto a nosotros en la dirección paralela a la línea de visión.

Introducidas las posibles fuentes de emisión de ondas de radio más importantes del Universo y que más nos conciernen a la hora de estudiar la línea espectral de 21 cm del hidrógeno neutro, se debe conocer cual va a ser el dispositivo experimental necesario para medir estas ondas de radio, y a partir de ellas, obtener la línea cosmológica de 21 cm.

1.3. Radiotelescopios

Para detectar la línea de 21 cm del hidrógeno neutro H_I se deben usar radiotelescopios de baja frecuencia. Debido a que la banda de radio es muy ancha en frecuencia, un único telescopio no va a poder cubrir todo este rango eficazmente, por lo que se necesita una combinación de telescopios e interferómetros para abarcar todo el rango y poder tener una mayor precisión y resolución. Para ello, existen dos grandes dispositivos de medida: las antenas únicas (*single dish*) y los interferómetros.
Los radiotelescopios deben poseer grandes diámetros de apertura D para obtener una buena resolución angular $\theta \approx \lambda/D$ en radianes, y λ la longitud de ondas de radio. Es por esta razón que dispositivos de medida que involucren una gran cantidad de interferómetros, que logren un diámetro de apertura del orden de $D \approx 10^4$ km son de suma utilidad.
[5]

1.3.1. Interferómetros

La interferometría es una técnica que usa aperturas pequeñas distribuidas para sintetizar una apertura mayor. Los interferómetros usados para medir ondas de radio reciben el nombre de *aperture-synthesis telescopes* y permiten inferir propiedades del emisor, en este caso ondas de radio, a partir de ciertas características del campo eléctrico recibido.
El principio fundamental de la inteferometría está basado en la relación que existe entre las señales de voltaje recibidas en dos antenas. La diferencia de tiempo en la que frentes de ondas de radio llega a cada antena se conoce como *time delay* τ . Este retraso produce un patrón de interferencias. Un interferómetro formado por dos elementos mide la distribución de brillo en el cielo para una determinada frecuencia espacial, que viene determinada por la longitud y orientación de las *baselines*, que son las líneas de base que unen las antenas y que se miden en coordenadas u, v. Así pues, la relación entre la distribución de brillo de una fuente en el cielo $T(l, m)$, medida en coordenadas l, m, y la visibilidad compleja $V(u, v)$ que mide el interferómetro están relacionadas por una transformada de Fourier. [5]

$$V(u,v) = \int\int T(l,m)\exp(-i2\pi(ul+vm))\,dl\,dm$$
$$T(l,m) = \int\int V(u,v)\exp(i2\pi(ul+vm))\,du\,dv$$

Si en vez de tener solo un par de antenas, se tienen N antenas, al medir la visibilidad $V(u, v)$ en una amplia gama de *baselines* se logra sintetizar una apertura mucho mayor que la que podría conseguirse con un radiotelescopio, simulando de esta forma un telescopio gigante. Esta colección de visibilidades medidas permite hacer una reconstrucción de la imagen de la distribución de brillo del cielo mediante la transformada de Fourier inversa. Concretamente, la transformada de Fourier de las visibilidades genera la distribución del brillo del cielo convolucionada con la *point spread function (PSF)*, que se define como la respuesta de un instrumento de medida a una fuente puntual [9]. En nuestro caso la *PSF* también va a ser nuestro *primary beam* (haz primario), aunque el haz que se recibe el cual es incompleto ya que faltan *baselines* es el *diry beam* (haz sintetizado). De esta forma esta técnica de interferometría va a permitir mapear la distribución del hidrógeno neutro (línea de 21 cm) con una gran resolución angular y sensibilidad.

Sin embargo, debido a que no se tiene un continuo de antenas, sino que se tiene un número de antenas N con unas determinadas *baselines*, el plano uv no está totalmente completo, es decir, en el plano uv los puntos no están uniformemente distribuidos, y sin embargo como se ha visto, para obtener $T(l, m)$ se necesita tener un *grid* uniforme. El hecho de tener un mayor número de *baselines* hace que se tenga una mayor cobertura del plano uv. Cabe destacar que se debe dejar un tiempo de medida para que la rotación de la Tierra una cobertura lo más completa posible. Debido a este hecho de que no se tiene un *grid* uniforme, se deben introducir los *weightings*. Existen varios tipos de *weightings*, aunque en este caso nos interesan solo dos: *natural weighting* y *uniform weighting*: [10]

- **Natural weighting**: maximiza la sensibilidad. En este caso, el pesado de la visibilidad es inversamente proporcional a la variancia del ruido de la visibilidad. Este *weighting* tiene una peor resolución, aunque maximiza la señal ante el ruido.

- **Uniform weighting**: minimiza los *sidelobes* de la PSF (*primary beam*), tiene una menor sensibilidad que el *natural weighting* pero tiene una mayor resolución angular. En este caso el pesado es inversamente proporcional a una función de densidad local.

1.3.2. SKA-Low

El SKA *Square Kilometre Array* es un radiotelescopio que a su vez está dividido en dos radiotelescopios: el SKA-Low y el SKA-Mid. El SKA-Low cubre frecuencias de entre 50 a 350 MHz, mientras que el SKA-Mid cubre frecuencias mayores, llegando a alcanzar los 15.4 GHz.
El SKA-Low es un radiotelescopio situado al oeste de Australia, que utilizando la técnica de interferometría es capaz de medir señales en un rango de frecuencias de 50 a 350 MHz. Con 512 estaciones, con 256 antenas por estación, las 131072 antenas ocupan alrededor de 419000 m^2. El hecho de que esté localizado en estas zonas es debido a que son áreas libres de ondas de radio terrestres, por lo que solo analiza ondas de radio extraterrestres, tanto galácticas como extragalácticas. [11]
Debido al gran área que ocupa y el hecho de que las *baselines* sean muy grandes, se obtiene una resolución angular mayor que cualquier radiotelescopio jamás creado debido a la técnica anteriormente mencionada de *aperture synthesis*.

Debido a la alta sensibilidad del *SKA-Low* y del gran rango de frecuencias que abarca, este radiotelescopio es de una utilidad tremenda para obsevar la línea de 21 cm del hidrógeno neutro. Puesto que se miden visibilidades en un rango de 50 a 250 MHz, va a ser posible observar estas líneas de emisión para *redshifts* de en torno a 6-10, correspondientes a la Época de Reionización.

Uno de los retos que plantea el grupo de investigadores del *SKA-Low* es el *Science Data Challenge 3a (SDC3a)*, que consiste en la eliminación del *foreground* de una señal perteneciente a la Época de Reionización (EoR). El objetivo principal de este reto es el de que diferentes grupos de investifación de todo el mundo se familiaricen con los datos que se reciban del SKA y que al mismo tiempo implementen e ideen técnicas para limpiar los *foregrounds*.
Para ello, se proporciona un conjunto de datos simulados de la Época de Reionización del *SKA-Low* que incluye la señal de reionización, es decir la línea de 21 cm del hidrógeno neutro que se quiere obtener, *foregrounds* cuyo origen es variado y está presente para distintas frecuencias además de ruido producido por el dispositivo experimental. A partir de este conjunto de datos se desea obtener un espectro de potencias cilíndrico de la señal de 21 cm, libre de todo tipo de contaminación, es

decir, libre de cualquier *foreground*.

Con respecto a las simulaciones, el telescopio simulado utiliza la configuración del SKA-Low de 512 estaciones con un diseño llamado *"Vogel"*, se utiliza un modelo del cielo a partir de dos componentes distintas, una externa y otra interna, y por último se utiliza un modelo de error que incluye errores instrumentales para hacer la simulación lo más realista posible tales como la adición de ruido térmico o la atenuación de fuentes del modelo de cielo externo.

A partir de esta simulación se obtuvieron un conjunto de visibilidades, de las cuales se obtuvo una imagen, concretamente un cubo de datos. Con este cubo de datos que puede tener un *uniform weighting* o un *natural weighting* se debe eliminar los *foregrounds* con alguna técninca específica, para una vez obtenida la señal de 21 cm de la Época de Reionización, obtener el espectro de potencias cilíndrico. De esta forma, el principal objetivo del *SDC3a* es el de fomentar nuevas técnicas de eliminación de *foregrounds* para simulaciones acordes a las medidas que se esperan, de manera que cuando el radiotelescopio *SKA-Low* obtenga las primeras señales de la *Época de Reionización*, se obtengan buenas estimaciones de la línea de 21 cm.

Las simulaciones utilizados en el *SDC3a* cubren un rango de frecuencias de 106 a 196 MHz y se proporcionan tanto las visibilidades como los cubos de datos. A partir de las visibilidades, se logra obtener una imagen en dos dimensiones del cielo. Esta imagen se forma para una determinada frecuencia, por lo que si se obtienen diferentes imágenes para distintas frecuencias, se acaba por conseguir un cubo de datos que contiene todas las imágenes para distintas frecuencia en la que uno de los ejes es la frecuencia que va de 106 a 196 MHz, y los otros dos ejes son coordenadas espaciales de una imagen del cielo. [12]

1.4. Espectros de potencias

En el estudio de la señal de 21 cm, el análisis de los espectros de potencias constituye una herramienta fundamental para extraer información cosmológica. La señal de 21 cm se cree que es 10^5 menos intensa que la intensidad de los *foregrounds*. Debido a su debilidad y al fuerte nivel de contaminación por *foregrounds*, además del ruido instrumental, no siempre es posible recuperar directamente la imagen espacial de la señal, por lo que el espectro de potencia se convierte en una herramienta clave, ya que resume de manera estadística la distribución espacial de la intensidad de la señal $T(l, m)$ en función de la escala $(k_\perp, k_\parallel)$, sin requerir una reconstrucción precisa de la imagen. Para describir las fluctuaciones de la intensidad del brillo en el espacio se utilizan los números de onda $k_\perp$ y $k_\parallel$, que son los números de onda perpendiculares y paralelos a la línea de visión respectivamente.
Los números de onda $k_\perp$ están asociados con las direcciones angulares en el cielo. En un interferómetro, la distribución de las *baselines* (la distancia entre pares de antenas) determina qué números de onda angulares $k_\perp$ se pueden acceder. Las *baselines* más largas son sensibles a escalas espaciales angulares más finas (mayores valores de $k_\perp$).
Por otro lado, los números de onda $k_\parallel$ están relacionados con la información espectral, concretamente de la frecuencia, ya que la dirección de la línea de visión es equivalente al eje espectral de un interferómetro. En este caso, los modos de $k_\parallel$ más altos están limitados por la resolución espectral del instrumento, mientras que los modos más bajos de $k_\parallel$ están limitados por el ancho de banda total de los datos recolectados.
Un espectro de potencias (*power spectrum*) es una medida estadística que cuantifica la varianza de

las fluctuaciones espaciales en un campo, como el campo de temperatura de brillo $T(l, m)$ de 21 cm, en función de diversas escalas de longitud ($k_\perp$, $k_\parallel$). En otras palabras, lo que nos dice un espectro de potencias es cómo de intensas son las fluctuaciones para diferentes escalas. Por ejemplo, un valor alto del espectro de potencias para un determinado k indica que hay muchas estructuras (regiones de alta o baja temperatura de brillo) de un tamaño correspondiente a $1/k$ en el Universo.
Partiendo de la distribución del brillo del cielo $T(\vec{r})$, se define una transformada de Fourier del cielo $\tilde{T}(\vec{k})$. El espectro de potencias $P(\vec{k})$ se define por la relación: [3]

$$< \tilde{T}(\vec{k})\tilde{T}(\vec{k'})^* >= (2\pi)^3\delta^D(\vec{k} - \vec{k'})P(\vec{k}) \tag{1.2}$$

En resumen, los espectros de potencia nos van a servir de extrema utilidad a la hora de visualizar la relación entre la señal cosmológica y los *foregrounds*. Sin embargo, se debe prestar especial atención a los *foregrounds* y su relación con los dinstintos números de onda $k_\perp$ y $k_\parallel$ ya que a la hora de la eliminación de estos se pueden cometer graves errores. Puesto que los *foregrounds* no son uniformes en el espacio, para distintos $k_\perp$ y $k_\parallel$ se van a tener diferentes propiedades. Esto se muestra gráficamente en la Figura 1.2.

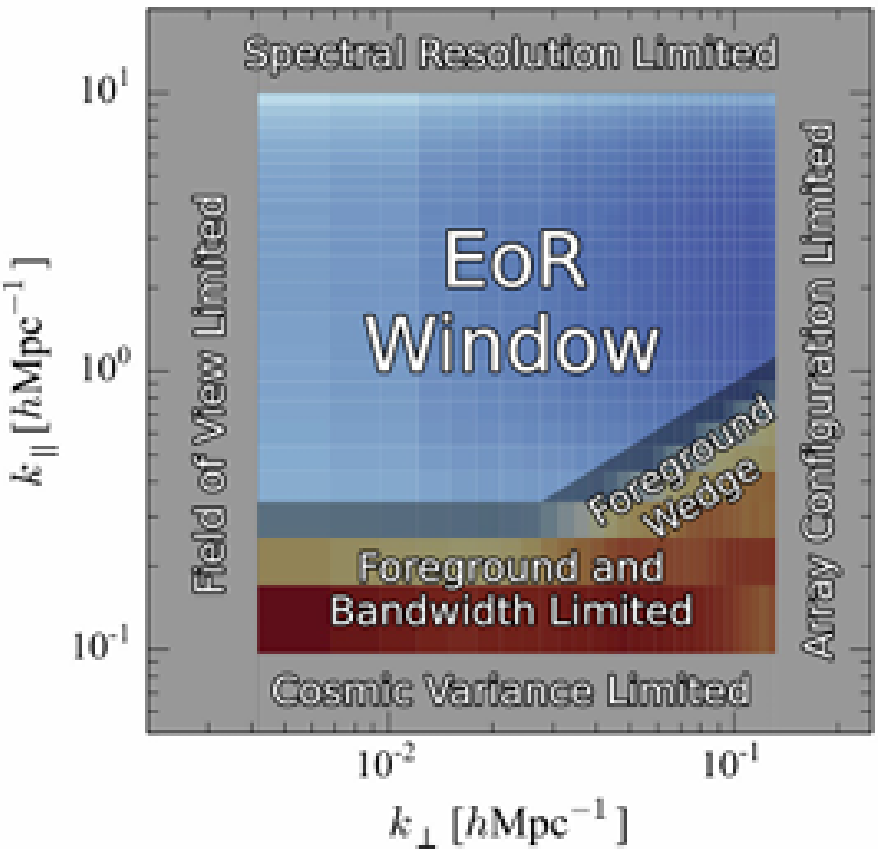

Figura 1.2: Esquema de la ventana de la EoR en el plano de Fourier $k_\perp k_\parallel$. A bajos $k_\perp$, los errores aumentan por el campo de visión limitado; a altos $k_\perp$, la sensibilidad cae por la longitud de los baselines. En $k_\parallel$, los foregrounds contaminan las bajas frecuencias y se extienden en forma de cuña (*foreground wedge*) hacia valores más altos. Fuera de esta región, en la ventana de la EoR, el espectro puede medirse de forma más limpia, dominado solo por ruido térmico. [13]

La Figura 1.2 es un diagrama esquemático del espectro de potencias en el plano de Fourier cilíndrico $k_\perp$, $k_\parallel$. Puesto que nosotros vamos a trabajar con números de onda $k_\perp$ y $k_\parallel$ que van desde 0.05 hMpc^{-1} hasta 0.5 hMpc^{-1}, la región de interés va a ser el *foreground wedge*. Esta cuña (*wedge*) está caracterizada por la presencia de *foregrounds*, que poseen una intensidad 10^5 veces mayor que la de la línea de 21 cm del hidrógeno. Por otra parte, debido a la cromaticidad inherente de los interferómetros, la información de los *foregrounds* se mezcla desde valores bajos de $k_\parallel$ a valores altos de $k_\parallel$. Además de estos factores, existe un sesgo aditivo en las estimaciones del espectro de potencia, donde la contaminación se extiende a $k_\parallel$ a medida que $k_\perp$ aumenta, de manera que se forma una cuña. Otra de las consecuencias de la *foreground wedge* es que las mediciones en esta región tendrán barras de error más elevadas. [13]

En definitiva, la Figura 1.2 ilustra el dilema de la búsqueda de la señal de 21 cm, ya que la *foreground wedge* restringe el acceso a los modos de k bajos, donde la relación señal-ruido del espectro de potencia cosmológico es máxima. Se deduce entonces que el espectro de potencias es una herramienta muy útil pero que debe tomarse con precaución y analizarlo con detalle ya que podemos incurrir en estimaciones que se alejan de la realidad.

2 Metodología

En este capítulo se describe el proceso para, partiendo de unas simulaciones proporcionadas por el *Science Data Challenge 3a*, eliminar el *foreground* y obtener los espectros de potencia. En el desafío original, 17 equipos participaron, aplicando cada uno de ellos distintas técnicas, tales como la sustracción de fuentes puntuales, la PCA (*Principal Component Analysis*) o SVD (*Single-Value Decomposition*) u otras técnicas que implementan *deep learning*. En nuestro caso, vamos a replicar la técnica empleada por el grupo de investigación HIMALAYA, fundamentada en la PCA. [14]

2.1. Simulación y cubo de datos

La simulación de las señales recibidas por el *SKA-Low* se encuentra en concordancia con el radiotelescopio real, ya que el modelo de telescopio usado consistía en 512 estaciones. Por otro lado, el modelo del cielo empleado es una combinación de diferentes fuentes de emisión de ondas de radio, entre las que se encuentra los *foregrounds*, que proceden tanto de fuentes galácticas como extragalácticas, errores tanto instrumentales como sistemáticos y la emisión de la línea de 21 cm del hidrógeno neutro que es la señal de interés que se debe recuperar.
La simulación cubre un rango de 106 a 196 MHz, con un muestreo de frecuencia de 100 kHz. El resultado de esta simulación son las visibilidades $V(u, v)$, a partir de las cuales se generaron los cubos de imágenes. Estos cubos de imágenes se presentan tanto con una ponderación natural *(natural weighting)*, como con una ponderación uniforme *(uniform weighting)*. Estos cubos que contienen la suma de la emisión simulada de la línea de 21 cm, *foregrounds* además de efectos instrumentales y sistemáticos conforman un mapa de coordenadas espaciales y de frecuencia. Los cubos tienen una dimensión original de 2048×2048 píxeles que equivalen a un campo de 9.12×9.12 grados. Además de estos cubos, se proporciona otro cubo que contiene la *point spread function* para cada canal de frecuencia.
Aparte de los cubos de datos simulados, se incluye también un conjunto de datos de prueba que recibe el nombre de *"test dataset"*, que consiste en una simulación de la línea de 21 cm y ruido, para poder obtener los espectros de potencia corregidos. [14]

Debido a las especificaciones del desafío, antes de proceder con la eliminación del *foreground* del

cubo de datos, el cubo debe recortarse a 256 ×256 píxeles. Esto debe realizarse tanto para los cubos de datos con el *uniform* y *natural weightings* como para los cubos de datos que incluyen la *point spread function*.
Una vez se tienen recortados los cubos, estos se dividen en 6 intervalos de frecuencia, en el que cada intervalo abarca 15 MHz de frecuencia, y todos ellos cubren el rango pedido por el *SDC3a* de 106-196 MHz. Para dividir el nuevo cubo de datos recortado en los seis intervalos de frecuencia demandados, se usó un *pipeline* creado por el grupo *HIMALAYA* [15].

2.2. Sustracción del *foreground*

Los *foregrounds*, como ya se ha explicado con anterioridad son fuentes de radio que emiten en el mismo rango de frecuencias que la señal de interés, la línea de 21 cm del hidrógeno. Para su eliminación, es de vital importancia conocer cómo dependen estos *foregrounds* con la frecuencia, ya que a la hora de eliminarlos va a ser de suma importancia tener clara su dependencia suave con la frecuencia. Esto se debe principalmente a que los *foregrounds* generalmente tienen una dependencia suave con la frecuencia, a diferencia de la señal cosmológica de 21 cm, que no posee una relación tan suave con la frecuencia sino más compleja, debido a la evolución cosmológica entre otros factores.
Una vez se tiene que los *foregrounds* dependen suavemente con la frecuencia, el siguiente paso es eliminarlos. Para ello existen diferentes técnicas, siendo las más conocidas el ajuste polinomial, el análisis de componentes principales (PCA) y el análisis de componentes independientes (ICA). En nuestro caso, nos vamos a centrar en la técnica de PCA, y aplicándola vamos a observar qué resultados se obtienen.
Por último, la eliminación de *foregrounds* conlleva una pérdida de información, además de que puede inducir algún sesgo en la obtención de los espectros de potencia, tanto a la alza como a la baja, ya que se pueden haber eliminado *foregrounds* de forma que se pierde información de la señal de 21 cm. [16] A continuación se detalla la metodología seguida para la correcta eliminación del *foreground* y de la estimación del espectro de potencia cilíndrico en dos dimensiones para los cubos de datos recortados proporcionados por el *Science Data Challenge 3a*.

2.2.1. Reconvolución

El factor fundamental para poder eliminar los *foregrounds* satisfactoriamente es el de conseguir que estos dependan suavemente con la frecuencia. Esto cualitativamente implica que para un ligero cambio de valor en la frecuencia, se espera que la intensidad del *foreground* varíe también ligeramente. Sin nigún tipo de efecto instrumental, los *foregrounds* muestran un alto grado de suavidad con respecto a la frecuencia, por lo que en un primer momento, facilita enormemente su eliminación con respecto a la señal cosmológica de 21 cm. Sin embargo, cuando se recibe la señal en el interferómetro, en este caso el *SKA-Low*, debido a los efectos instrumentales, los *foregrouds* dejan de depender suavemente con la frecuencia, por lo que se debe corregir. [14]

Es conveniente recordar lo que se conoce como el plano *uv*, el cual representa las frecuencias espaciales, que son las coordenadas conjugadas de las coordenadas espaciales (x, y) en el plano de la imagen. Estas coordenadas miden distancias en el frente de onda incidente. El problema de la dependencia no suave del *foreground* se explica debido a que no se cubre totalmente el plano *uv*. Con el interferómetro del *SKA-Low*, los frentes de las ondas de radio que se reciben, a la hora de medirlos y recopilarlos, este muestreo no es continuo y depende de cómo estén configuradas las antenas. El

interferómetro mide una visibilidad por cada *baseline*, por lo que el plano uv está fundamentalmente vacío. [17] Debido a que las *baselines* pueden ser muy largas, existe otro efecto que provoca que la dependencia de los *foregrounds* no sea suave con la frecuencia, el cual recibe el nombre de *mode-mixing effect*. Este efecto unido al de la incompleta cobertura del plano uv dan como resultado una imagen sucia (*dirty beam*), en la que los *foregrounds* dejan de exhibir una relación suave con la frecuencia, y por tanto no se puede distinguir bien la señal cosmológica de la línea de 21 cm del hidrógeno neutro.

Al observar con el interferómetro tiene lugar una convolución natural entre el *primary beam* que es la radiación de ondas de radio que dependen suavemente con la frecuencia y el *dirty beam* causado por la incompleta cobertura del plano uv. Más específicamente, se trata de una imagen que está convolucionada con el *dirty beam* y modulada por el *primary beam*. Para volver a tener una relación suave con la frecuencia, podría deshacerse la convolución natural entre el *primary beam* y el *dirty beam*, es decir una desconvolución. Sin embargo, esta desconvolución no garantiza que se tenga una suavidad con la frecuencia de los *foregrounds* por lo que no es un método factible. [14].
Un método ideado por el equipo de *HIMALAYA* [14] es el *método de reconvolución*. En lugar de intentar desconvolucionar la imagen, se realiza una nueva convolución, de ahí el nombre de *método de reconvolución*, con una nueva función para recuperar la suavidad de los *foregrounds* con la frecuencia. Esta función no debe ser escogida al azar, siendo las mejores posibilidades un *beam kernel* de la *PSF* o un *beam kernel* Gaussiano. Concretamente, este equipo realizó la reconvolución con el *beam kerneal* de la *PSF* que viene proporcionado por el propio *Science Data Challenge 3a*.
Este proceso de reconvolución elimina las visibilidades $V(u, v)$ de las *baselines* más largas del interferómetro, manteniendo en todo momento una resolución espacial suficiente para obtener los modos de Fourier requeridos para posteriormente estimar los espectros de potencia. De esta forma, para cada uno de los seis intervalos de frecuencia dentro de los rangos requeridos por el *challenge* de 106 MHz a 196 MHz, se realiza una reconvolución entre el cubo de datos que contiene el *dirty beam* con el *beam kernel* de la *PSF*. Tras realizar este proceso de reconvolución, se logra eliminar la *sidelobe* de la *PSF* además de reducir el *mode-mixing effect* que complejizaba la dependencia de los *foregrounds* con la frecuencia. [14]
Una vez se ha realizado el proceso de reconvolución, y se ha recuperado la suavidad de los *foregrounds* con la frecuencia, tal como ellos dependen de ella si no existe ningún efecto instrumental, el siguiente paso a realizar es la eliminación de los *foregrounds* de forma que se obtenga finalmente la señal de interés, que no es otra que la línea de 21 cm del hidrógeno.

2.2.2. Técnica de PCA

La técnica que se va a emplear para eliminar los *foregrounds* recibe el nombre de *PCA (Principal Component Analysis)*. La principal función de esta técnica es la de reducir un conjunto de datos voluminoso y complejo a una dimensión inferior para revelar estructuras simplificadas. Así pues, la PCA es un método para extraer información relevante, en nuestro caso la línea de 21 cm, de un conjunto de datos complejos, en nuestro caso, todos los *foregrounds*. [18]
El objetivo principal de la PCA es identificar la base más significativa para reexpresar un conjunto de datos. Una de las asunciones que toma la PCA es la consideración del problema a resolver como un problema lineal, siendo su resolución un cambio de base lineal. Esto indica que lo que se busca es reexpresar los datos iniciales como combinaciones lineales de sus vectores base. Esta nueva representación de los datos permite observar con mayor claridad la dependencia de una determinada magnitud con otra, y por tanto conocer la varianza de los datos, en nuestro caso con la frecuencia. Otro de los principios fundamentales de la PCA es que las direcciones en el espacio de los datos que

muestran la mayor varianza son las que contienen la información de interés. Esto se fundamenta en que la señal que se quiere medir tiene una varianza alta, mientras que el ruido tiene una varianza menor. De esta forma, la PCA busca minimizar la redundancia, que se mide por la magnitud de covarianza a la par que maximizar la señal de interés que es medida por la varianza. Esto se logra diagonalizando la matriz de covarianza de los datos.
La PCA se puede resumir entonces en varios pasos: [18]

- Organizar los datos en una matriz
- Restar el valor medio a cada medición realizada
- Cálculo de la matriz de covarianza frecuencia-frecuencia en nuestro caso.
- Diagonalización de la matriz de covarianza que implica el cálculo de autovectores y autovalores
- Esos autovalores ordenados de mayor a menor valor reciben el nombre de componentes principales.

Para la aplicación de la PCA en la eliminación de los *foregrounds*, en la matriz de covarianza frecuencia-frecuencia obtenida tras reducir la dimensionalidad de los datos proporcionados por el *SDC3a*, los *foregrounds* se manifiestan como las componentes de mayor varianza, es decir, los primeros autovalores (ordenados de mayor a menor). Esto se debe básicamente a que la información principal de los *foregrounds* se concentra en un pequeño conjunto de autovalores muy grandes.
Puesto que los *foregrounds* tienen una dependencia suave con la frecuencia, esta correlación se manifiesta como una gran varianza, lo que implica altos valores en la matriz de covarianza, que implican altos autovalores. [16] Por otro lado, para la línea de 21 cm, además del ruido instrumental, se sabe que no se tiene una relación tan suave con la frecuencia, por lo que la varianza será menor, y los autovalores serán menores. De esta forma, si se consigue identificar las direcciones de mayor varianza y estas son eliminadas, lo que se está eliminando realmente son los *foregrounds* dejando solamente la señal cosmológica de 21 cm, además del ruido.
En el caso que nos ocupa, tras aplicar el método de la reconvolución, la distribución de los autovalores en la matriz de covarianza se asemeja al caso en el que no están involucrados los efectos instrumentales, lo que va a facilitar la eliminación de los autovalores más altos, y por tanto la eliminación del *foreground*. El objetivo entonces será obtener el número de componentes principales que se deben descartar para obtener los valores pertenecientes a la línea de 21 cm. Este será el momento en el que se han eliminado los *foregrounds* y por tanto se ha obtenido la línea de 21 cm.
Finalmente, tras utilizar la técnia de PCA se podrá estimar el espectro de potencias para los distintos modos de Fourier.

2.2.3. Corrección del espectro de potencia

El método de reconvolución y la posterior eliminación del *foreground* tras la utilización de la técnica de PCA, permite generar un espectro de potencias. Sin embargo, este espectro de potencias no es del todo correcto y se deben tener en cuenta varias consideraciones para corregirlo. En primer lugar, existen hasta 4 factores que pueden alterar la amplitud de los espectros de potencias obtenidos: [14]

- El efecto del haz primario *(primary beam effect)*: el interferómetro no tiene la misma sensibilidad en todas las direcciones que provienen del cielo, por lo que el haz puede variar con la frecuencia.

- La *PSF* del haz sintetizado (*dirty beam*)
- Proceso de reconvolución: puesto que al convolucionar la imagen obtenida con la *PSF* se modifica la original, esto irremediablemente altera el espectro de potencias, y por lo tanto debe tenerse en cuenta.
- Pérdida de la señal de 21 cm: al aplicar la técnica de PCA para eliminar los *foregrounds* es posible que se hayan eliminado componentes que corresponden a la señal cosmológica de interés, la línea de 21 cm, por lo que también se debe tener en cuenta en el espectro obtenido.

Debido a estos 4 factores, el espectro de potencias obtenido no debe considerarse como verdadero ni definitivo, sino que se debe corregir. Estas correciones se pueden considerar a la hora de analizar las imágenes proporcionadas por el *SDC3a* o, por otro lado, se pueden considerar en el propio espectro de potencias obtenido. Este espectro de potencias obtenido en primera instancia lo denominamos $\tilde{P}(k_\perp, k_\parallel)$.
Para realizar esta corrección del espectro de potencias real, se define una función de transferencia (*transfer function*) $T(k_\perp, k_\parallel)$, que relaciona el espectro de potencias obtenido en primer lugar, con el verdadero espectro de potencias de la forma siguiente:

$$T(k_\perp, k_\parallel) = \left\langle P(k_\perp, k_\parallel)/\tilde{P}(k_\perp, k_\parallel) \right\rangle \tag{2.1}$$

De esta forma, si se poseen dos espectros de potencia se puede obtener la función de transferencia $T(k_\perp, k_\parallel)$, ya que esta no depende de la forma precisa del espectro de potencias obtenido $\tilde{P}(k_\perp, k_\parallel)$, sino que solo depende de la relación entre el *input* y el *output*. Puesto que el *SDC3a* proporciona además del cubo de datos simulado, un *test dataset* que incluye un espectro de potencias, se va a obtener el espectro de potencias en ambos casos, con ello la función de transferencia, y por último, se podrá estimar el espectro de potencias real $P(k_\perp, k_\parallel)$.

La *transfer function* se va a obtener a partir de la relación entre el espectro de potencias verdadero proporcionado por el *test dataset* y el espectro de potencias obtenido tras realizar simulaciones a partir del espectro de potencias verdadero. Estas simulaciones consisten en la generación de un determinado número de realizaciones gaussianas a partir del espectro del *test dataset*. Estas simulaciones crean mapas del cielo con solo la señal del hidrógeno, ya que el *test dataset* no contiene *foregrounds* ni efectos instrumentales. Una vez se han obtenido estas imágenes, se les añade un *foreground* y se repite la metodología seguida en los dos primeros pasos de la eliminación del *foreground*, es decir, el proceso de reconvolución, la eliminación de este por la técnica de PCA y la estimación del espectro de potencias. Con este espectro de potencia procedente de las simulaciones realizadas y con el espectro proporcionado por el *test dataset* se obtiene la *transfer function* $T(k_\perp, k_\parallel)$. A partir de aquí, multiplicando la *transfer function* por el espectro de potencias obtenido en primera instancia $\tilde{P}(k_\perp, k_\parallel)$ se obtiene el espectro de potencias $P(k_\perp, k_\parallel)$, el cual es el que el *Science Data Challenge 3a* demanda. [14]

2.2.4. Evaluación de los resultados: *score*

Una de las maneras para evaluar los espectros de potencia $P(k_\perp, k_\parallel)$ que cada equipo envía a *SDC3a* es elaborando una métrica que analiza la precisión de la señal de 21 cm recuperada tras la mitigación del *foreground* y que analiza también la exactitud de las barras de error $\Delta P(k_\perp, k_\parallel)$ obtenidas. Esta métrica tiene la forma de sistema de puntuación, de manera que una mejor puntuación implicará una mejor estimación del espectro de potencia con respecto a los valores reales y unas barras de error

más pequeñas. Esta puntuación recibe el nombre de *score* o $SDC3a_{FOM}$ y se calcula asumiendo que los valores del espectro de potencia son independientes y que la distribución de la incertidumbre es gaussiana. En esta aproximación gaussiana, si se toma un índice j que recorra todos los espectros de potencia obtenidos para cada $(k_\perp, k_\parallel)$, la probabilidad del valor real P'_j dada la medida $P_j \pm \Delta P_j$ viene dada por $\Pr(P'_j)$, donde la suma de todas estas probabilidades será el $SDCa_{FOM}$: [14]

$$\Pr(P'_j) = \frac{1}{\sqrt{2\pi}\Delta P_j} \exp(-(P'_j - P_j)^2/2\Delta P_j) \tag{2.2}$$

$$SDC3a_{FOM} = \sum_j \Pr(P'_j) \tag{2.3}$$

De esta forma, una puntuación final más alta indica un mejor rendimiento, lo que significa que las estimaciones del equipo están más cerca de los valores verdaderos y sus barras de error reflejan con una mayor precisión la incertidumbre. Además de esta razón, este sistema de puntuación permite estimar el número de componentes que se deben eliminar en la técnica de PCA para obtener mejores resultados, es decir, una mejor eliminación del *foreground*. Por conisiguiente, es de utilidad calcular esta puntuación para distinto número de componentes empleado, de forma que se observe con mayor claridad para qué número de componentes se obtienen mayores puntuaciones, de manera que sirva de apoyo para que estimaciones de futuras mediciones obtengan resultados más satisfactorios ya conociendo cuál es el número idóneo de componentes que se deben eliminar para obtener la señal cosmológica de la línea de 21 cm, totalmente limpia de *foregrounds*.

3 Resultados

En este capítulo se quieren replicar los resultados obtenidos en [14]. Concretamente, puesto que se utiliza el código de HIMALAYA, el objetivo de esta sección va a ser el de conseguir los resultados de este equipo, principalmente espectros de potencia, además de estudiar diferentes vías para conseguir resultados más óptimos, como por ejemplo variando el número de componentes a eliminar al usar la técnica de PCA. Para la eliminación del *foreground* se ha utilizado mayoritariamente un *natural weighting*, aunque a modo de comparación se presentan algunos resultados con un *uniform weighting*. Concretamente, todos los resultados y gráficas que aparecen en primer lugar se han realizado con un *natural weighting* mientras que la última sección de resultados se estudiarán los espectros de potencia habiendo usado un *uniform weighting*.

3.1. *Score* vs. componentes

Puesto que se está trabajando con el mismo *Test Dataset* del *SDC3a* original, se desean obtener resultados similares a los de [14]. De esta forma, se busca obtener los mejores resultados posibles, concretamente los mejores espectros de potencia posibles que sean lo más parecido posible a los espectros de potencia sin ningún tipo de *foreground*, lo que indicará que los resultados obtenidos son satisfactorios. La forma de evaluar qué tan parecidas son las estimaciones realizadas con los valores reales proporcionados por el *SDC3a* es calculando el *score* de la expresión 2.3. Por tanto, lo que se va a realizar en primer lugar es el cálculo del *score* tras aplicar la técnica de PCA para distinto número de componentes. De esta forma, se podrá conocer cuál es el número de componentes eliminadas con la técnica de PCA que mejores resultados proporciona, favoreciendo así futuras mediciones y estimaciones. Los resultados numéricos del *paper* del desafío indican que el equipo que mejores resultados obtuvo fue *DOTSS-21cmAdvancedML-GPR* con un *score* de 240226, mientras que para el caso de HIMALAYA, se obtuvo 134752.

Tras calcular el *score* para distinto número de componentes, en los 6 rangos de frecuencia y realizando 50 realizaciones gaussianas, se obtienen los resultados de la Tabla 3.1.

ν/MHz	20	30	40	50	60	70	80	90	100
106 - 121	666	1037	1409	2727	1555	1733	397	6	41
121 - 136	57	468	2077	2988	8337	15311	27442	29508	22907
136 - 151	65	413	748	1100	2127	10530	23412	24167	15234
151 - 166	347	2114	14420	40516	37483	39170	34851	8285	7379
166 - 181	11697	17736	30512	35068	24314	14208	14554	4188	859
181 - 196	7568	21765	41716	58876	40950	45499	44594	31943	31231
Score	20400	43533	90882	141275	114766	126451	145250	98132	77616

Tabla 3.1: Resultados del *score* para distinto número de componentes utilizando la técnica de PCA para eliminar *foregrounds* y para cada rango de frecuencias y en la línea inferior el valor del *score* total sumando todas las contribuciones.

Para un mejor entendimiento de estos resultados y lo que conllevan, se realiza una representación gráfica del *score* frente al número de componentes para una mejor visualización. Cabe destacar que a pesar de que el cálculo del *score* nos dice cuánto se acercan a los valores reales tras la eliminación del *foreground*, estos valores no se deben tomar como una verdad absoluta, ya que en el cálculo del *score* intervienen los espectros de potencia estimados tras realizar una serie de simulaciones. Estas simulaciones gaussianas son aleatorias, por lo que un mayor o menor número de ellas puede arrojar resultados que puedan diferir bastante. Es de esperar que a mayor número de realizaciones, resultados más acordes a la realidad se obtendrán, aunque también implicaría un gran volumen de trabajo. Es por esta razón que se han tomado 50 realizaciones gaussianas como un buen indicador, y a partir de aquí se utilizarán los resultados obtenidas a través de ellas. Esta representación gráfica se muestra en la Figura 3.1.

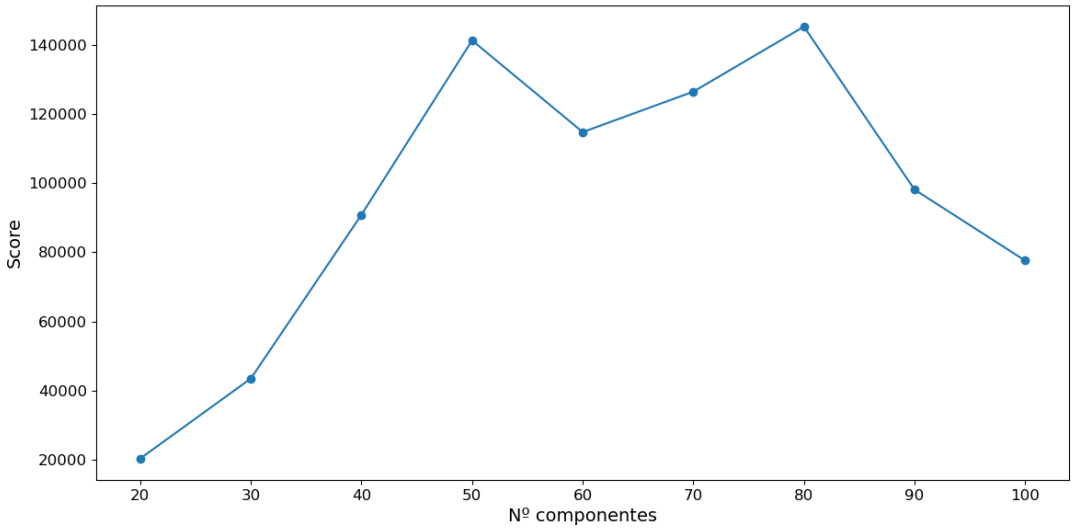

Figura 3.1: Representación gráfica del *score* obtenido para distinto número de componentes.

En el estudio del grupo HIMALAYA, se obtuvo un máximo *score* inferior (134752) al que hemos obtenido (145250), además de que mencionan que los mejores resultados los consigue eliminando alrededor de 20-30 componentes con la técnica de PCA. Sin embargo, siguiendo nuestros resultados, se observa que en el rango de 50-80 componentes es donde mejor se elimina el *foreground*. Esta diferencia se puede explicar atendiendo a los pasos seguidos para eliminar el *foreground*, ya que el *score* está totalmente correlacionado con las barras de error en el espectro de potencia, que son estimadas según el número de realizaciones gaussianas que se hacen cuando se obtiene la *transfer function*.

3.2. Diagonal con *foreground*

Tras obtener el número de componentes que mejor eliminación del *foreground* produce basándonos en la sección anterior (80 componentes), es de interés intentar representar los espectros de potencia obtenidos tras la eliminación del *foreground* junto a los espectros de potencia que lo tienen, es decir, representar la señal recibida en el *SKA-Low* con la señal de 21 cm del hidrógeno neutro libre del *foreground*. Puesto que los resultados que se obtienen son los espectros de potencia con sus barras de error correspondientes $P \pm \Delta P$ en función de los números de onda $k_\perp$ y $k_\parallel$, una buena forma para visualizar los resultados es tomando la diagonal, es decir, utilizar los espectros de potencia cuando $k_\perp = k_\parallel$ para observar la diferencia en órdenes de magnitud de la señal con o sin *foreground* y de paso ver qué tan bien se acercan los resultados obtenidos a los resultados reales proporcionados por el *SDC3a*. Esta representación gráfica se muestra en la Figura 3.2.

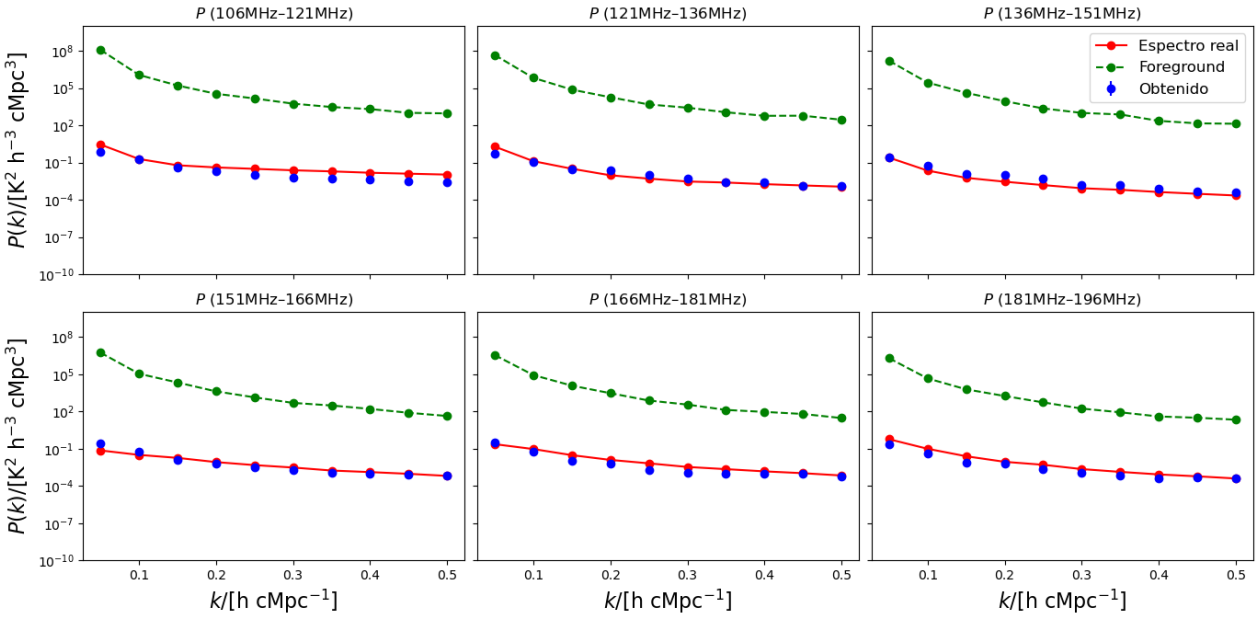

Figura 3.2: Representación gráfica de los espectros de potencia para valores de $k = k_\perp = k_\parallel$ para la señal con *foreground* (verde), los valores reales dados por *SDC3a* (rojo) y los obtenidos tras eliminar 80 componentes (azul) en los 6 rangos de frecuencia y con un *natural weighting*.

Al observar la Figura 3.2 se observa que los espectros de potencia de la señal recibida por los interferómetros y que contiene todos los *foregrounds* son varios órdenes de magnitud más grandes que la señal limpia de ellos, es decir la señal de la línea de 21 cm del hidrógeno. Concretamente la señal con *foreground* es aproximadamente 6 órdenes de magnitud mayor que la señal de 21 cm de hidrógeno, lo que da una idea de la importancia de eliminar los *foregrounds* ya que sin quitarlos no se puede obtener información acerca de la señal del hidrógeno. Por otro lado, se observa que a medida que k aumenta, el valor de $P(k)$ va disminuyendo, ya que, como se explicaba en la sección 1.4, para mayores valores de $k_{\parallel}$, nos acercamos cada vez más a la ventana de la EoR, en la que predomina la señal cosmológica en vez de los *foregrounds* y por tanto se tiene una potencia menor. Además, la distancia entre la potencia $P(k)$ con *foreground* y sin él para un mismo k va disminuyendo a medida que k aumenta por el mismo hecho de que a mayores $k_{\parallel}$, se espera que el *foreground* disminuya y no sea tan dominante con respecto a la señal cosmológica de la línea de 21 cm. Por otra parte, a pesar de que se trabaje con una escala logarítmica y la Figura 3.2 abarque un rango de hasta 20 órdenes de magnitud, se observa que los valores obtenidos tras la eliminación de 80 componentes se acercan bastante a los esperados.

El *score* calculado en la sección *Score vs. componentes* da una idea de qué tan buenos son los espectros de potencia calculados en relación con los espectros reales dados por el *challenge*. A pesar de que el número óptimo de componentes sea 80, ya que es el número de componentes para el cual se obtiene el mejor *score*, si se observa detenidamente la Tabla 3.1, se observa que a bajas frecuencias, los tres mejores *scores* se obtienen aproximadamente para 80 componentes, mientras que a mayores frecuencias, se obtienen mejores resultados para 50 componentes. Esto indica que los resultados más satisfactorios para todos los distintos rangos de frecuencia no se obtienen eliminando el mismo número de componentes, sino que dependiendo en qué rango de frecuencia nos encontremos, se deben eliminar un determinado número de componentes para obtener los mejores espectros de potencia posibles. Sin embargo, en este caso en el que ya se conocen los resultados del *SDC3a*, es más fácil ajustar los espectros de potencia que se calculan a los reales, pero en un supuesto caso en el que no se tengan estos espectros reales, lo más óptimo será eliminar 80 componentes.

Para estudiar más detenidamente cómo de bien se ajustan los valores obtenidos a los esperados, se puede representar los mismos términos de la diagonal pero sin tener en cuenta el *foreground*. De esta forma más exhaustiva se obtienen resultados más concisos que nos permite saber con mayor precisión qué tan buenos son los espectros recuperados, si son mayores o menores que los esperados, etc. Esto queda reflejado en la Figura 3.3, donde se representan los términos de la diagonal de los espectros de potencia cuando se eliminan 20, 50 y 80 componentes.

Tras observar la Figura 3.3 se pone de manifiesto el hecho de que eliminar el mismo número de componentes para los diferentes rangos de frecuencia arroja resultados algo diversos. En primer lugar, al eliminar 80 componentes, especialmente para los tres rangos de frecuencia mayores, los espectros obtenidos son menores que los reales, es decir, se tiene una subestimación, aunque por el contrario, en el rango de 121 a 136 MHz y de 136 a 151 MHz se obtiene una sobreestimación. Sin embargo, a rasgos generales, los espectros obtenidos, si bien varios valores no coinciden con los reales ni con sus respectivas barras de error, el comportamiento que siguen en su totalidad concuerda con los valores esperados.

Por otro lado, para el caso en el que se eliminan 50 componentes, también se observa que para frecuencias mayores, se obtienen resultados muy cercanos a los esperados, mientras que a frecuencias más bajas, donde la señal del *foreground* es más intensa, se alejan más de los valores esperados. Para

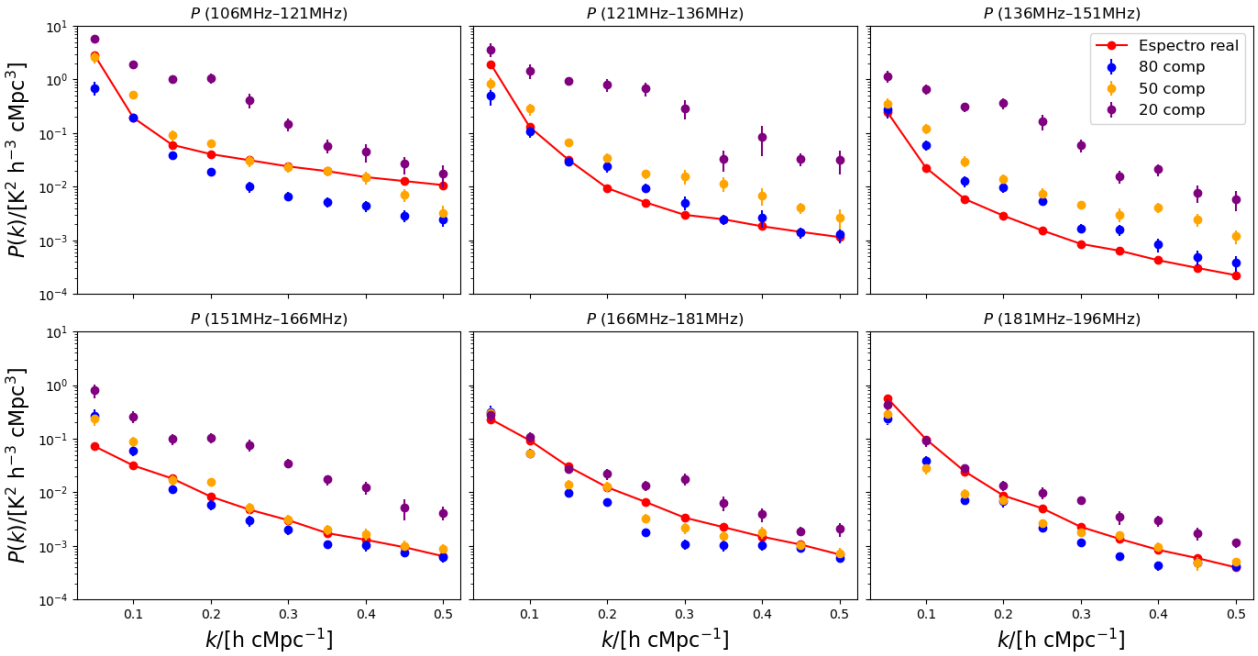

Figura 3.3: Representación gráfica de los espectros de potencia obtenidos tras la eliminación de distinto número de componentes para valores de $k_\perp = k_\parallel$ y los espectros de potencia reales dados por *SDC3a*(rojo) para los 6 rangos de frecuencia y con un *natural weighting*.

el caso en el que se eliminan 20 componentes, los resultados para las distintas frecuencias difieren considerablemente de los esperados, especialmente a frecuencias bajas. Concretamente, al observar las figuras para las distintas frecuencias, se tiene que los espectros al eliminar 20 componentes son superiores a los esperados ya que todavía existe una gran cantidad de componentes de intensidades similares a las del *foreground*.

3.3. Espectros de potencia

El siguiente paso para visualizar los resultados y compararlos con los valores reales de forma más rigurosa es representando los espectros de potencia en 2 dimensiones, para todos los valores de $k_\perp$ y $k_\parallel$ a diferencia de las secciones anteriores en las que solo atendíamos a cuando $k_\perp$ era igual a $k_\parallel$.

Para ello, una buena forma de representar estos resultados es mostrando el espectro real, el cual queremos reproducir lo más parecidamente posible, junto a los espectros obtenidos tras eliminar un distinto número de componentes. De esta forma, se espera que para las componentes con las que se ha obtenido un mayor *score* se ajusten más a los valores esperados, mientras que a medida que nos alejamos de este número de componentes, los espectros obtenidos serán peores y no se parecerán tanto a los esperados. Puesto que los mejores resultados se obtuvieron tras eliminar 80 y 50 componentes, es de gran utilidad representar ambos y ver cuánto se parecen a los espectros reales prestando especial atención a cómo se comportan para distintos rangos de frecuencia. Además de esto, se representan también los espectros tras eliminar 20 componentes para ver qué tanto se acercan o alejan de los esperados. Todos estos espectros obtenidos se muestran en la Figura 3.4.

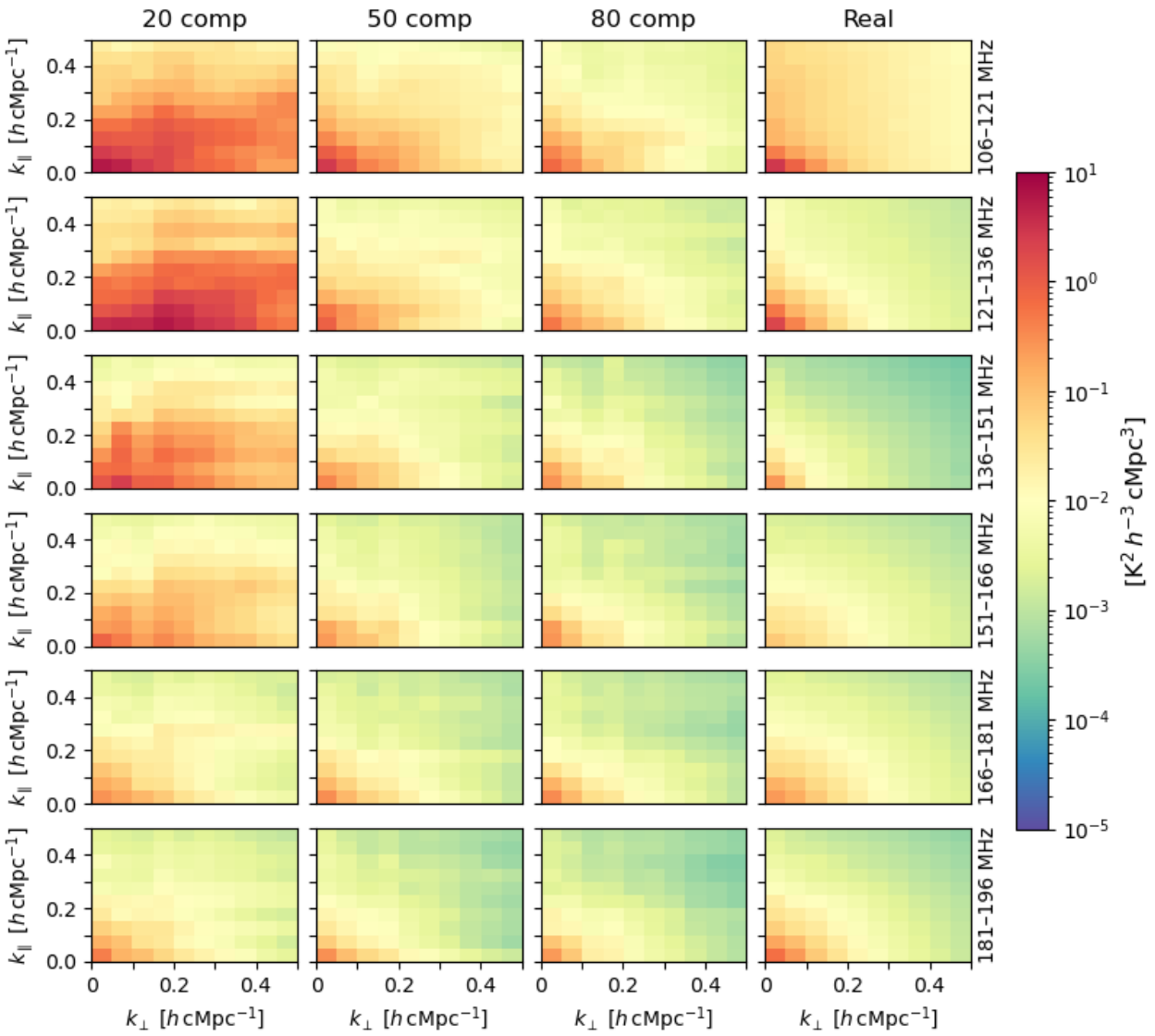

Figura 3.4: Comparación de distintos espectros obtenidos tras eliminar distinto número de componentes con los esperados (última columna) para los 6 rangos de frecuencia estudiados con *natural weighting*.

Al comparar la columna de las 80 componentes con las del espectro real en la Figura 3.4, hay algunos rangos de frecuencia en los que los espectros de potencia son más similares a los esperados, como es el caso de los espectros que van desde los 166 a 196 MHz. Esto en primer lugar indica que a la hora de recuperar la señal de 21 cm tras eliminar el *foreground*, a medida que la frecuencia aumenta, la emisión del *foreground* se vuelve algo más débil, por lo que las componentes que restan tras eliminar 80 se aproximan más a las de la señal del hidrógeno, incluso llegando a ser valores menores que la señal de 21 cm. Para frecuencias más bajas, los espectros de potencia no son tan parecidos por esta misma razón, ya que aun eliminando 80 componentes, todavía se obtienen valores de la potencia mayores que los esperados.

Al variar el número de componentes, los espectros de potencia varían consecuentemente, por lo que si en vez de tomar 80 componentes, tomamos 50, siguiendo el segundo mejor resultado del *score*, los espectros de potencia de mayores frecuencias no eliminarán tantas componentes y por tanto no se perderá información de la señal de 21 cm, es decir se tendrá una sobreestimación, mientras que los de bajas frecuencias ajustarán algo peor.
En el caso, en el que se han eliminado 50 componentes, para los tres mayores rangos de frecuencia, los espectros se asemejan bastante a los espectros esperados, mientras que para las frecuencias más bajas, ya no se parecen tanto. Esto se debe a que a frecuencias bajas, la emisión del *foreground* es más intensa, por lo que para ajustar mejor a la señal de 21 cm se deben eliminar más componentes. El hecho de eliminar menos componentes hace que la señal que se obtiene todavía tiene *foregrounds*, como se observa en los altos valores de la potencia (rojos) especialmente para valores altos de $k_\perp$ y $k_\parallel$.

Por otro lado, el número de componentes para el cual se ha obtenido un bajo *score*, es de esperar que el espectro de potencia que se obtenga diferirá más del esperado. Si se toma el caso de 20 componentes, los espectros resultantes se muestran en la primera columna 3.4.

En este caso, al eliminar 20 componentes se obtienen resultados que ya se alejan de los esperados, especialmente a frecuencias bajas, en las que se sigue teniendo una presencia del *foreground* bastante fuerte como se observa en los valores altos (rojos) de la potencia en los espectros. En cualquier caso, para los 6 rangos de frecuencia, se tiene una clara sobreestimación del espectro de frecuencia para todos los $k_\perp$ y $k_\parallel$. Al eliminar tan pocas componentes, a pesar de eliminar los *foregrounds* más intensos, ya que estos se concentran en las primeras componentes, todavía quedan restos comparables e incluso superiores a los de la señal de 21 cm, de manera que se sobreestiman los valores de la potencia, indicándonos que la eliminación de 20 componentes no arroja resultados tan satisfactorios como en los anteriores casos con 50 y 80 componentes.

Una vez se han representado los espectros de potencia obtenidos tras eliminar distinto número de componentes, y que estos se encuentran evidentemente en concordancia con los valores obtenidos del *score* es conveniente centrarse ahora en los mejores resultados que se han obtenido, es decir para 80 componentes, y estudiarlos más a fondo analizando sus barras de error y comparándolos de una manera más cuantitativa con los espectros reales para cada rango de frecuencia. En cuanto a los errores cometidos en la estimación de los espectros de potencia, en su mayoría son errores sistemáticos, no errores aleatorios. Cabe destacar que quizá lo más complicado del desafío sea proporcionar barras de error lo más acercadas a la realidad posible, lo cual es complejo y depende de las técnicas utilizadas para eliminar el *foreground*. En las secciones anteriores se han representado los errores como barras de error, aunque quizá no es la mejor forma de visualizarlo. Como se ha explicado anteriormente, los errores vienen determinados por las desviaciones entre los espectros de potencia generados por las realizaciones gaussianas y una buena forma de visualizarlos es representando los espectros junto a su error. Esto queda reflejado en las Figuras 3.5.

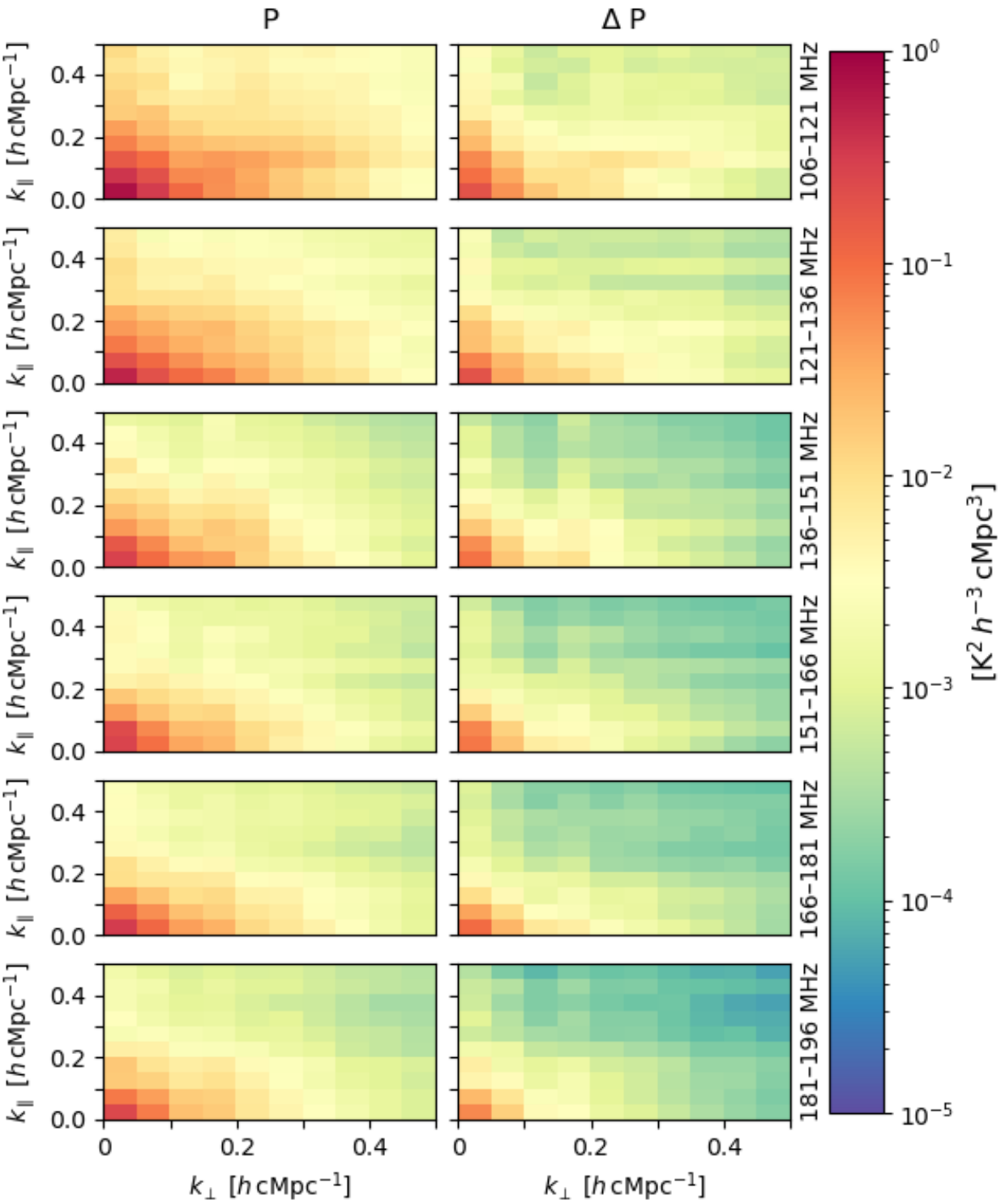

Figura 3.5: Espectros de potencia obtenidos $P(k_\parallel, k_\perp)$ junto a sus errores ΔP para los 6 rangos de frecuencia tras eliminar 80 componentes con *natural weighting*.

Comparando las figuras de los espectros obtenidos y los errores obtenidos, ambas siguen una distribución similar, en el sentido de que para valores pequeños de $k_\perp$ y $k_\parallel$ se obtienen valores del mismo

orden de magnitud prácticamente tanto en los espectros como en su error, si bien en los errores decrece más rápidamente a medida que se aumentan $k_{\parallel}$ y $k_{\perp}$. En cualquier caso, los valores de la potencia en la gráfica de los errores es de alrededor de un orden de magnitud inferior a los de los espectros obtenidos, haciéndose esta diferencia mayores para valores de k mayores.
El hecho de que para valores pequeños de $k_{\parallel}$ y $k_{\perp}$ se obtengan errores en el mismo orden de magnitud se puede explicar nuevamente con el hecho de que la intensidad del *foreground* en esta región es mucho más fuerte, por lo que a la hora de estimar la *transfer function* que se obtenía a partir de varias simulaciones en las que se añadían *foregrounds* aleatorios, es más complejo el obtener errores pequeños ya que la señal fluctúa con mayor intensidad que en zonas donde el *foreground* es menos intenso y se aprecia más la señal de 21 cm.

3.4. Uniform weighting

Los resultados mostrados con anterioridad utilizaron un *natural weighting* y a pesar de la dificultad del desafío se obtienen resultados satisfactorios. A continuación, a modo de comparación, se van a mostrar resultados de los espectros de potencia obtenidos $P(k_{\perp}, k_{\parallel}) \pm \Delta P(k_{\perp}, k_{\parallel})$ pero en este caso las visibilidades han sido ponderadas a través de un *uniform weighting*. Puesto que los mejores resultados se obtuvieron con 80 componentes, se van a representar los términos de la diagonales de los espectros además de los espectros en su totalidad tras eliminar 80 componentes, 50 y 20, para el *uniform weighting* en las Figura 3.6 y 3.7.

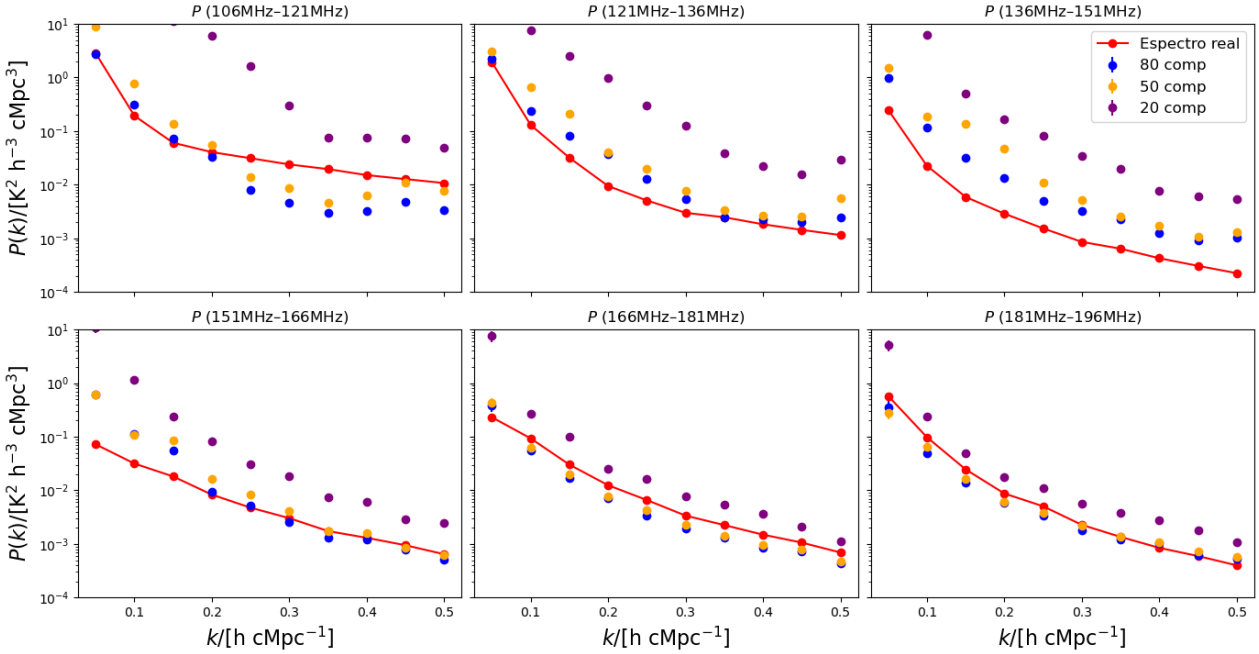

Figura 3.6: Representación gráfica de los espectros de potencia obtenidos tras la eliminación de distinto número de componentes para valores de $k_{\perp} = k_{\parallel}$ y los espectros de potencia reales dados por *SDC3a*(rojo) para los 6 rangos de frecuencia y con un *uniform weighting*.

En primer lugar se han utilizado 20, 50 y 80 componentes a modo de comparación con las componentes utilizadas cuando se usó el *natural weighting*. De nuevo se observa que a mayor número de componentes utilizadas, en este caso 80, es donde mejores espectros de potencia se obtienen, es decir que son los más parecidos a los esperados. Cabe mencionar que al igual que en el caso anterior, si se eliminan más componentes, los espectros estarán subestimados. Para darle un valor numérico a estos espectros con respecto a qué tan bien se ajustan a los esperados se han calculado los *scores* para cada rango de frecuencia y para cada número de componentes, los cuales se presentan en la Tabla 3.2.

ν/MHz	20	50	80
106 - 121	338	3104	260
121 - 136	8	1851	13866
136 - 151	0	0	0
151 - 166	0	61276	45494
166 - 181	2002	11134	4493
181 - 196	149	59468	70397
Score	2497	126833	134510

Tabla 3.2: Resultados del *score* para distinto número de componentes utilizando la técnica de PCA para eliminar *foregrounds* y para cada rango de frecuencias y en la línea inferior el valor del *score* total sumando todas las contribuciones para el *uniform weighting*.

Tras visualizar los valores numéricos que ayudan a cuantificar mejor los resultados, ver Tabla 3.2 se observa que en efecto los espectros obtenidos que mejor se ajustan a los esperados son al eliminar 80 componentes, con un *score* de 134510. Esto está en concordancia con los resultados anteriores del *natural weighting*, si bien son valores menores que indican que con el *uniform* weighting no se obtienen resultados tan satisfactorios como con el *natural*.
Atendiendo a los espectros de la Figura 3.7, para el caso en el que se eliminan 20 componentes se ve claramente que no se recupera el espectro esperado ya que aun habiéndose eliminado 20 componentes, todavía quedan señales con una intensidad mucho mayor a la de la señal de 21 cm, lo que quiere decir que todavía quedan muchas componentes relacionadas con el *foreground* por eliminar. Además, como en el caso del *natural weighting*, para frecuencias mayores se obtienen mejores resultados, ya que en estas frecuencias la emisión del *foreground* es menos intensa, por lo que al eliminar el mismo número de componentes en todos los rangos, se ajustará mejor para frecuencias mayores.

Tras observar la Figura 3.6, de nuevo se observa que para frecuencias mayores, los espectros obtenidos se ajustan más a los esperados, mientras que a frecuencias bajas se obtienen valores del espectro especialmente para el caso de las 20 componentes. Un caso llamativo y que explica los valores del *score* igual a 0, en el rango de frecuencias de 136 a 151 MHz de la Tabla 3.2 es el hecho de que en este rango, para distinto número de componentes, los espectros recuperados no se acercan al espectro real, lo que indica que para obtener un mejor espectro en este rango de frecuencias será necesario eliminar un mayor número de componentes.

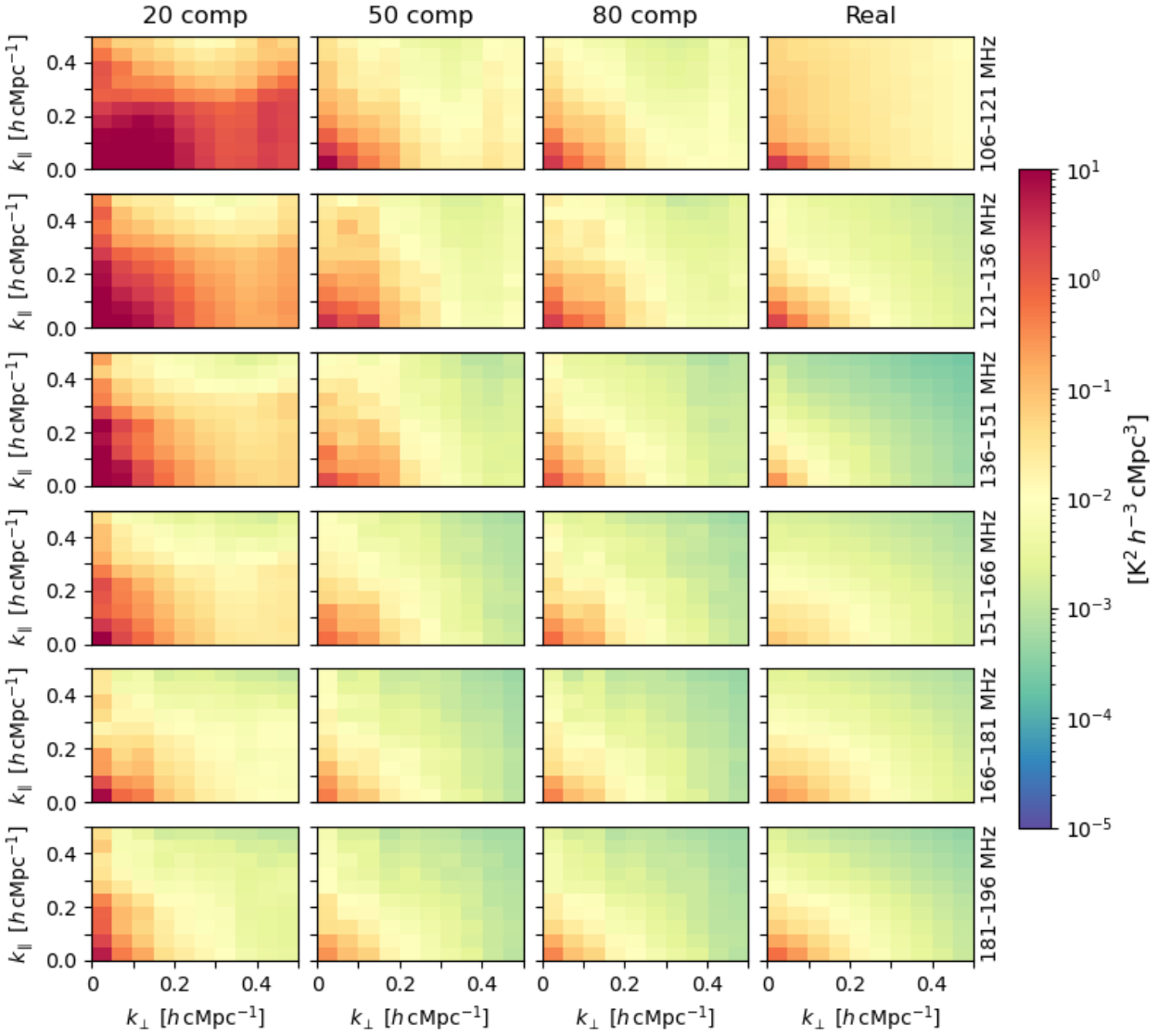

Figura 3.7: Comparación de distintos espectros obtenidos tras eliminar distinto número de componentes con los esperados (última columna) para los 6 rangos de frecuencia estudiados habiendo utilizado un *uniform weighting*.

4 Conclusiones

En este trabajo se ha abordado el problema de la detección de la señal de 21 cm del hidrógeno neutro, cuya observación representa una herramienta fundamental para estudiar la historia del universo temprano, en particular la época de reionización y etapas anteriores. La principal dificultad de la observación de esta señal reside en que se encuentra oculta por *foregrounds* de origen galáctico y extragaláctico de gran intensidad y por el ruido instrumental, lo que dificulta su extracción directa.

Con el objetivo de mitigar este problema, se ha implementado un método basado en el Análisis de Componentes Principales (PCA) aplicado a imágenes de radiofrecuencia. Los resultados obtenidos muestran que el PCA constituye una técnica eficaz para eliminar los *foregrounds* especialmente para frecuencias de ondas de radio que van desde los 151 a los 196 MHz, revelando información compatible con la señal cosmológica subyacente. Sin embargo, también se han identificado limitaciones: la separación no es perfecta y existe el riesgo de eliminar parcialmente la señal de 21 cm junto con los *foregrounds* al eliminar más componentes de las necesarias, lo que requiere una evaluación más precisa del número de componentes eliminados y del rango de frecuencia en el que nos encontramos.

El estudio ha permitido comprobar que el PCA, pese a ser un método relativamente sencillo y computacionalmente eficiente ya que con un portátil de 8 GB de RAM se ha podido solventar, debe complementarse con otras técnicas quizá más sofisticadas para alcanzar un mayor grado de robustez en la recuperación de la señal. A pesar de haber recuperado los espectros con buena precisión, el reto todavía consiste en estimar con mayor detalle las barras de error cuyo origen reside en los errores instrumentales mayoritariamente, así como en el *weighting* usado, las imágenes con las que el cubo de datos original es reconvolucionado, etc. Cabe destacar que en este estudio se utilizó la *PSF (Point Spread Function)* para reconvolucionar la imagen, aunque otras opciones como intentar reconvolucionar con una distribución gaussiana en el futuro pueden ser de gran utilidad.

En conclusión, este trabajo ha evidenciado tanto el potencial del PCA en la eliminación del *foreground* en cosmología de 21 cm como las posibles limitaciones existentes a la hora de recuperar los espectros de potencia con errores detallados. Para el futuro, mejores simulaciones de las que se conozcan mejor sus posibles errores sistemáticos harán que con ayuda de diferentes técnicas se logre recuperar con mayor precisión la señal de 21 cm del hidrógeno neutro.

Referencias

[1] Steve McMillan Eric Chaisson. *Astronomy Today*. Pearson, 8th edition, 2014.

[2] Wikipedia contributors. Chronology of the universe. `https://en.wikipedia.org/wiki/Chronology_of_the_universe`, 2025.

[3] J. Richard Shaw Adrian Liu. Data analysis for precision 21 cm cosmology. *Publications of the Astronomical Society of the Pacific*, 132(062001), 2020.

[4] The Planck Collaboration. Planck2013 results. i. overview of products and scientific results. *Astronomy amp; Astrophysics*, 571:A1, October 2014.

[5] Scott M. Ransom James J. Condon. *Essential Radio Astronomy*. Princeton University Press, 2016.

[6] Ghara R. Chatterjee A. et al Bera, A. Studying cosmic dawn using redshifted hi 21-cm signal: A brief review. *J Astrophys Astron 44, 10*, 2023.

[7] Steven R. Furlanetto, S. Peng Oh, and Frank H. Briggs. Cosmology at low frequencies: The 21cm transition and the high-redshift universe. *Physics Reports*, 433(4–6):181–301, October 2006.

[8] Jonathan R Pritchard and Abraham Loeb. 21 cm cosmology in the 21st century. *Reports on Progress in Physics*, 75(8):086901, July 2012.

[9] David Wilner. Radio astronomy and interferometry fundamentals. Presentación PowerPoint, 2015. Harvard-Smithsonian Center for Astrophysics.

[10] R. J. Sault and T. A. Oosterloo. Imaging algorithms in radio interferometry, 2007.

[11] Square Kilometre Array Observatory. Ska-low telescope, 2024.

[12] Square Kilometre Array Observatory. The sdc3a foregrounds: Data product descriptions. `https://docs.google.com/document/d/1UZCsztjZDlbGbz3uqvEbPiIRXkct3rJDVtR_EiVM_so`, 2021.

[13] Adrian Liu, Aaron R. Parsons, and Cathryn M. Trott. Epoch of reionization window. i. mathematical formalism. *Phys. Rev. D*, 90:023018, Jul 2014.

[14] A. Bonaldi, P. Hartley, and R. et al. Braun. Square kilometre array science data challenge 3a: foreground removal for an eor experiment. arXiv:2503.11740v1, 2025.

[15] L. Zhang. Himalaya sdc3a software. `https://github.com/553445316/HIMALAYA.git`, 2023.

[16] David Alonso, Philip Bull, Pedro G. Ferreira, and Mário G. Santos. Blind foreground subtraction for intensity mapping experiments. *Monthly Notices of the Royal Astronomical Society*, 447(1):400–416, December 2014.

[17] Damien Ségransan. Observability and uv coverage. *New Astronomy Reviews*, 51(8):597–603, 2007. Observation and Data Reduction with the VLT Interferometer.

[18] Jonathon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.